Forecasting Origin Ramp to Destination Ramp Pairs for Peak Shipping Period for a Major Global Courier Service Company

2015 MAA PIC Math[1]

MAA 100
MATHEMATICAL ASSOCIATION OF AMERICA
CELEBRATING A CENTURY OF ADVANCING MATHEMATICS

Daniel King, Andrew Lash, Joseph Molisani, Jamie Rossi, Norman Sivi, Zachary Stuck, Andrew Wearden

Jeffrey Paul Wheeler (Faculty)

Department of Mathematics, the University of Pittsburgh

Pittsburgh, PA 15260, USA

jwheeler@pitt.edu

Abstract

Positioning aircraft in the most effective locations during the peak shipping period is critical in order to maximize available resources while maintaining high service levels. Therefore, an accurate volume forecast is essential for planning purposes. Forecasting monthly airport-to-airport raw shipping poundage is addressed and as well as a very user-friendly Excel Macro deliverable.

[1]Funding for this project is provided by NSF grant **DMS-1345499** through the **MAA** ***Preparation for Industrial Careers in Mathematical Sciences Program*** **(PIC Math)**, `www.maa.org/picmath`.

Contents

1 The Problem

The University of Pittsburgh PIC Math Team was tasked with the following problem from the Peak Planning team of a major US global courier service and the service initially submitted the following to the team:

"Positioning aircraft in the most effective locations during the peak shipping period is critical in order to maximize available resources while maintaining high service levels. Therefore, an accurate volume forecast is essential for planning purposes.

The following will be helpful in understanding the statement of the problem:

Industry Definitions.

- *"ramp" means airport;*
- *"facility" will by default mean "ramp" unless otherwise specified;*
- *"origin ramp forecast" would be a forecast which includes: customer name, number of packages, the services by which the packages are being shipped, the date they are to be shipped, and the ramp out of which they will be shipped;*
- *"service" means either overnight shipping, 2-day shipping, or 3-day shipping.*

Given historical shipping data and considering other economic factors and trends, create an origin-to-destination-ramp forecast (by date, account, volume, origin ramp/facility, service, and destination ramp). That is,

The Problem.
for a given date:

- *forecast the volume of packages a top-peak-shipping customer will wish to send,*
- *forecast the origin ramps/facilities and respective quantities of these packages,*
- *forecast the destination ramps and their respective quantities,*
- *and forecast by which service they will be shipped.*

Data Provided to the University of Pittsburgh will be historical shipping data for 30 top peak accounts from Jan 2012 - Dec 2014 will be provided.

Data 1.1.
Fields included will be:

- *coded customer ID*
- *origin ramp*

- *origin facility (if applicable)*
- *destination ramp*
- *service*
- *packages*
- *ship date*

Note: this will be big data as each of the 30 customers is a major US corporation."

2 Legal Issues and Workaround

Despite the process beginning in early December, the University of Pittsburgh's Office of General Counsel and the Senior Counsel at the courier service could not come to an agreement regarding a contract. As such, the team was not provided the materials stated in Data 1.1. Wanting to address the problem, the team instead acquired public raw poundage shipped airport to airport for 153 airports from 2010 – 2014 for FedEx from the `US Department of Transportation` [3]. The team used this as data to prepare a deliverable for the fictitious courier service "FedUPS". The deliverable is a very user friendly Excel Macro that runs multiple SARIMA models in the statistics software *R* [1] and Matlab. The models developed use the monthly data from top 30 airports but can be easily extended to all 153 airports.

3 Introduction

Winter of 2013 proved to be one of the most unexpectedly difficult seasons for shipping companies. Due to unforeseen weather conditions and a sudden increase in demand from online retailers, numerous shipping companies found themselves suddenly overworked with a glut of packages for which they were unprepared. This lead to packages not being delivered, which lead to unhappy customers and missed profits, including ample backlash from the media, the public, and the government [4], [5]. Clearly, better forecasting methods for both packages shipped and their overall distribution is needed. Major shipping companies already use numerous mathematical models for both forecasting and logistical services and the 2015 University of Pittsburgh PIC Math Team hopes to improve upon this important and necessary work.

4 Initial Approaches

4.1 Markov Chains, and Seasonality

Once overcoming the lack of a proper data set from the client and obtaining a data base of $10,048$ data points from a US government website [3], the team addressed what to do with the January 2010 to November 2014 airport-to-airport shipping data for FedUPS for the top 30 major airports in the United States. The first step was to sort the month by month data for each airport into more easily digestible sets, which was done using a Python program. The team's initial forecasting approach for both volume and placement of freight was one based on Markov chains, namely the data in the form that it appeared lent itself very naturally to Markov processes. If shipping totals are from one month to the next to be a random process, one can easily build a transition matrix, and using one month as the time step, it is easy to achieve a volume distribution from month to month. Unfortunately, the outputs projected by this model proved to be highly inaccurate. While seemingly a promising lead, the team did not consider the fundamental importance of the seasonality conditions on the data that was being examined.

The high level of inaccuracy was due to the type of data being modeled does not obey the Markov property, namely, the likelihood of distribution of freight was not dependent on the previous timestep (month), but rather on what month in the year it was. The team realized the problem with trying to project later months in the year just using data from that year, specifically it becomes necessary to account for trends both for January through October and from what November and December shipping information looked like in previous years. Additionally, as the transition matrix was compounded, non-zero probabilities began to crop up for shipping routes that did not exist.

4.2 Regression

Frustrated by initial attempts at producing a viable model with Markov chains, success ultimately came from considering another approach. Seemingly at a dead end, the team decided to look at the problem from the simplest angle possible. With a few hundred data points and our goal: forecast future months based on the shipping totals from previous months and years, the team took the individual airport shipping totals and summed them in to monthly totals, giving a reasonably small set of data points with which to work. A simple linear regression on 58 monthly totals and compared the 59th month's (November 2014) known data led to a surprising result.

Not expecting a particularly successful output given the complexity and seasonal trends inherent to the data, the team was surprised to find that the linear regressive model gave a 95% accurate predictor of the 59th month. This was exceedingly counterintuitive as the possibility that the trends we were trying to model could be accurately encapsulated by such a simple test seemed highly unlikely. The team proceeded to try a few higher order regressive models, namely cubic and quadratic regressions, but both offered similar accuracy levels.

Unfortunately there was an inherent limitation to these models in that they produced a very accurate model for predicting a final month but proved to be downright awful at interpolating in any fashion. This, of course, was caused by the seasonality aspect of the kind of data that was being considered. A suggestion from another classmate lead to the team's most successful approach to this problem. Realizing the seasonal nature of the data the team researched and developed an SARIMA time series forecast to project what shipping totals would be, first in aggregate, and following that, on an airport by airport basis.

5 Time Series

A *time series* refers to a sequence of successive data points, measured periodically for some length of time, which fit perfectly our publicly available FedEx data set. *Time series* forecasting broadly refers to the practice of using statistical inference and models to project future data based on preexisting data points.

Most time series data models can be considered one of three major types: *auto-regressive*, *integrated*, and *moving averages*. *Auto-regressive models* seek to project future data points as a function of an early points, plus some kind of random error term, while *moving average models* seek to project later values of the series based on strictly on randomness. *Integrated models* refer to an order of differencings, the practice of attempting to eliminate seasonality from a given data by subtracting the observation in the current time period to the corresponding point in early time periods. The order of differencing refers to the number of times this process occurs. A combination of all three of these models, called ARIMA, was what the team decided to use as its model.

5.1 ARIMA

Having settled on time series as the most useful possible tool to forecast the given data, the team chose from many available models to use an SARMIA forecast, which is a special type of ARIMA forecast. ARIMA stands forAuto-Regressive Integrated Moving Average, which considers an integrated series, some number of lags of the stationized series (auto-regressive terms), and some number of lags of forecast errors (moving average terms).

ARIMA models have a general forecasting equation as

$$\hat{y}_t = \mu + \Phi_1 + y_{t-1} + ... + \Phi p y_{t-p} - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{1}$$

and ARIMA models are ultimately summarized by 3 parameters:

$$\begin{aligned} ARIMA(p, d, q) \text{ where} p &:= \text{the number of auto regressive terms;} \\ d &:= \text{the number of non-seasonal differences; and} \\ q &:= \text{the number of moving averages.} \end{aligned} \tag{2}$$

5.2 A Few Basic Examples

$ARIMA(1, 0, 0)$ describes a model that is first order in the auto regressive term and can be predicted by multiples of itself plus a constant term:

$$\hat{Y}_t = \mu + \Phi_1 Y_{t-1}. \tag{3}$$

$ARIMA(0, 1, 0)$ is a simple random walk:

$$\hat{Y}_t = \mu + Y_{t-1}. \tag{4}$$

$ARIMA(1, 1, 0)$ describes a first-order autoregressive model with one order of nonseasonal differencing and a constant term:

$$\hat{Y}_t = \mu + Y_{t-1} + \Phi_1(Y_{t-1} - Y_{t-2}). \tag{5}$$

Of highest importance to the use of the base ARIMA model is that the data needs to be stationary, that is, non-seasonal, with its statistical properties having no discernable trends with respect to time. Given the yearly nature of shipping data, a base ARIMA model would be fundamentally inappropriate for this problem. Hence we consider a Seasonal ARIMA (i.e., SARIMA) model.

5.3 SARIMA

A Seasonal Auto-Regressive Integrated Moving Average(SARIMA) is a further sophistication of ARIMA forecasting which accounts for non-stationarity (seasonality) in the data. Similar to the standard ARIMA model, SARIMA models can be summarized by the basic ARIMA parameters (p, q, d) together with 4 additional parameters:

$$
\begin{aligned}
&SARIMA(p, q, d)\times(P, Q, D)_S \\
&\qquad\qquad \text{where } p :=&&\text{the number of auto regressive terms;} \\
&\qquad\qquad d :=&&\text{the number of non-seasonal differences;} \\
&\qquad\qquad q :=&&\text{the number of moving averages;} \\
&\qquad\qquad P :=&&\text{the number of seasonal autoregressive terms;} \\
&\qquad\qquad D :=&&\text{the number of seasonal differences;} \\
&\qquad\qquad Q :=&&\text{the number of seasonal moving averages; and} \\
&\qquad\qquad S :=&&\text{the seasonal period.}
\end{aligned}
\tag{6}
$$

Note that the S term in our case will always be 12 as the data points are monthly and seasonal in years. In common practice this term is almost always either 12 (for monthly data) or 4 (for quarterly data).

5.4 Many Models

Having determined the model, the team wrote a program in R that would run an SARIMA forecast on the data set for a preselected number of combination of parameters, totaling 24 different possible models. The R program also determines the Bayesian Information Criterion (BIC) for each possible model and then selects the model with the lowest BIC, giving the most accurate model from amongst the possible combinations. Using this method, the team obtained an accuracy level of 98.2% for the monthly totals. A specific forecasting example can be seen in the team's video presentation.

6 Probability Matrix

Having successfully figured out a way to project month to month shipping totals, the team was left with considering how to accurately forecast volume distributions. Fortunately, using some of the methods adapted from the work incorporating Markov chains proved to be relevant for this portion of our solution. While the transition matrices could not accurately project totals from month to month, these probability matrices could be used to project distributions based on our forecasting totals for the appropriate months. Using the previous year's totals a probability matrix is created which is then used to forecast a given airport's distribution to other airports.

We note that a natural question to consider is whether the probabilities should be modeled using an SARIMA forecast. This was considered by the team and the result was negligibly different from the un-SARIMA probabilities.

7 Final Approach and Results

Our final approach to the forecasting problem is two-fold. As first expected, to project total monthly shipping tallies for a given airport an SARIMA model is used. Given these forecasted results we then predict their destination using a probability transition matrix derived from the previous year's data. This technique projects both the total output and specific raw-poundage destinations on an airport-to-airport basis. Running the model on the 2010-2013 data and comparing it with the known 2014 data, we found the model has a high degree of accuracy generally in in the range of 90% or better[2].

7.1 Deliverable

Having answered the clients chief question using *R*, the team had one last task to complete: make the model readily usable to the analysts at FedUPS. As typical business practice uses Excel and Access, the team faced one more challenge in adapting the model to a final deliverable that would operate entirely within an Excel program.

This was accomplished by converting the *R* program to a similar program in Matlab and creating an Excel macro which could be implemented with relative ease. In this macro, the client inputs a particular data set into Excel and the macro returns not only a projection of the total freight coming out of a particular airport, but a projection for which other airports the freight will be shipped as well. An example of this can be seen in our accompanying materials in the presentation video and the team did provide the instruction manual it prepared for FedUPS in the appendix.

8 Conclusion

Throughout our semester of work on this project, we gained a vast amount of knowledge that will surely be invaluable as we go on to look for careers in mathematical industries. Most obviously, the forecasting techniques we researched

[2]The team did consider attaching printouts of the Excel files, but they were too large. We will be happy to provide them upon request.

and applied to real-life data was a spectacular opportunity for all of us. This sort of forecasting technique was something that none of us had any particular experience with in earlier classes or internships, and our participation in this program provided the unique circumstances to learn practical forecasting methods in the setting of a mathematics course. Additionally, the work we put in to research and find model that would accurately forecast all of the client's requests proved to be a tremendous portion of the work. Unlike the normal type of learning about applications of math in a classroom - where problems and examples are used to teach about a particular method - we were given a problem and had to figure out what to do on our own. This was both an enormous amount of freedom and room for error. Ultimately, we felt this would reflect better reflect a project in industry and are grateful for the chance to gain such experience.

While the mathematical portions of this of the project were certainly important to our overall experience, surprisingly, we believe that the non-mathematical aspects of our work will be most useful in our future careers. Learning about mathematical models and writing code are the sort of thing one can do in other classes. Navigating the murky waters of intellectual property law and the general legal difficulties that accompanied this project, on the other hand, were completely foreign to us. This setbacks, while frustrating, taught us a lesson about how important the non-technical aspects of industrial work can be. Additionally unique to this class was the challenge of finding and learning a model that worked. The greatest takeaway from this project, though, was the development of the deliverable. In all our other courses our work is being evaluated by people with strong backgrounds in mathematics. In this case however, we had to be able to explain our methods to people that did not have our background in mathematics. This challenge demonstrated the importance of communicating mathematics: ultimately, no matter how accurate or useful our model was, it was not helpful to our client unless they understood how to use it. This meant keeping the advanced programming and mathematical jargon at a minimum. Our team was fortunate enough to have the opportunity to present our work to both an analyst at FedEx Services and the World-wide Sales Manager of FedEx Services. ***The value of being able to communicate our results proved to be the most important skill we learned throughout our work as PIC competitors.***

9 Summary of Skills Learned

In completing this project, the team members learned:

- Graph Theory (Modeling)
- Big Data
- Holt-Winters Method (not used)

- Markov Chains
- Probability
- Moving Averages
- Exponential Smoothing
- Linear Algebra (Matrix Operations)
- Time Series Forecasting; ARIMA and SARIMA
- Independent Learning

and also the software:

- Microsoft Excel
- Writing Excel Macros
- Microsoft PowerPoint
- Python through IDLE
- *R* Programming
- Matlab

10 Student Testimonial

Graduating Mathematics major Andrew Lash wrote the following in his evaluation of the course:

"*I literally knew absolutely nothing about any of this stuff before this class, so to reiterate, I wouldn't possibly trade it for anything. Pretty sure it was largely responsible for me getting my job too* (which starts in July). I will be working for Aon Hewitt (Radford division) as an Equity Valuation Consultant. Aon is an enormous risk and reinsurance company and in the job I will be utilizing VBA (Excel Coding) to create macros for manipulating big data. Some of the methods I will be using, I'm told, is the Monte Carlo for simulating (and forecasting) stocks so that way we can create an optimal stock-based compensation packages for C-level executives. If the stock values are forecasted incorrectly, the executive's pay won't necessarily be indicative of their successes or failures (which stockholders would not appreciate).

Clearly, from this class, I will be able to harness my newly acclaimed big data experience with statistical methods (which Monte Carlo is one). Additionally, they do much of their analysis through excel and VBA macros, which I created to run the deliverable for this class. Lastly, it's heavy in consulting and dealing with clients, as such the experience I gained presenting in front of several 'executives' and classmates will be invaluable going forward."

11 Thank You

The University of Pittsburgh PIC Math team would like to express its deepest gratitude to the directors of the PIC Math Program, the MAA, and the NSF for giving us this incredible opportunity.

12 Appendix

The following pages contain the instruction manual for the Excel Macro running R code in Matlab the University of Pittsburgh PIC Math team prepared for FedUPS.

References

[1] *The R Project for Statitical Computing*, http://www.r-project.org/.

[2] R.H. Shumway and D.S. Stoffer, *Time Series Analysis and Its Applications: With R Examples (Edition 2)*, Springer Texts in Statistics, 2006.

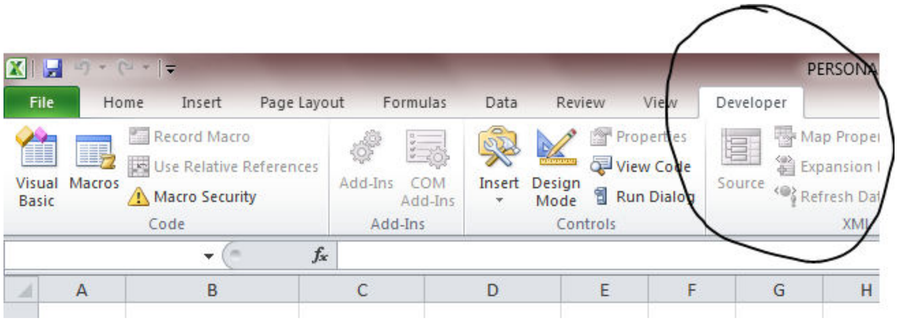
[3] United States Department of Transportation, http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=258&DB_Short_Name=Air%20Carriers

[4] US News, *UPS, FedEx scramble to deliver delayed Christmas packages*, online http://usnews.nbcnews.com/_news/2013/12/26/22058674-ups-fedex-scramble-to-deliver-delayed-christmas-packages

[5] USA Today, *UPS hopes for a merrier Christmas is 2014*, online http://www.usatoday.com/story/money/business/2014/12/14/ups-fedex-holiday-deliveries/20209311/

Re-creating the Macro in Excel:

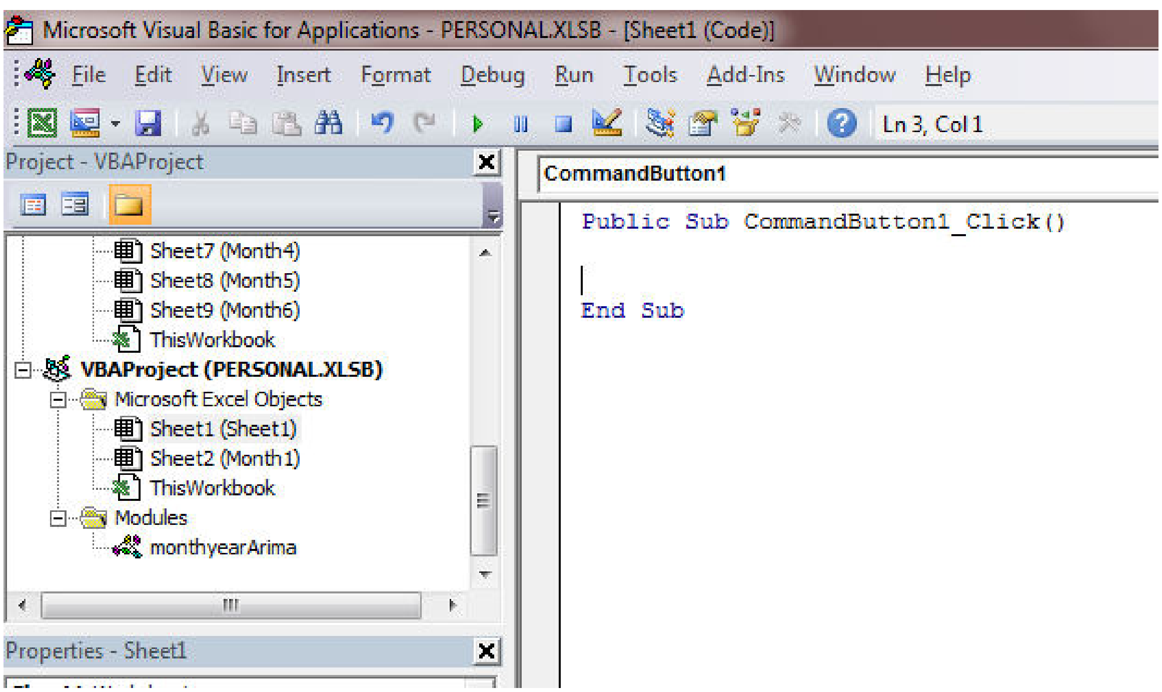
In your "Personal" Workbook, access the developer tab (instructions on getting the developer tab are readily available through a Google search if not already active).



Then click on the **Insert** dropdown pictured above, a couple icons to the left of the circle. Select and add a Command Button from the ActiveX portion.

Click the **Design Mode** icon directly to the right of the Insert Icon from the previous step.

Right click on the Command Button that has appeared on your worksheet and select **"View Code**" from the drop down. Next you'll see:



At this point you just input the following code. You may copy paste the next couple of lines, or manually replicate what the result should be in the following picture. **NOTE THE BOLDED LINE BELOW:**

Public Sub CommandButton1_Click()

Dim myValue As Variant
Dim myotherValue As Variant

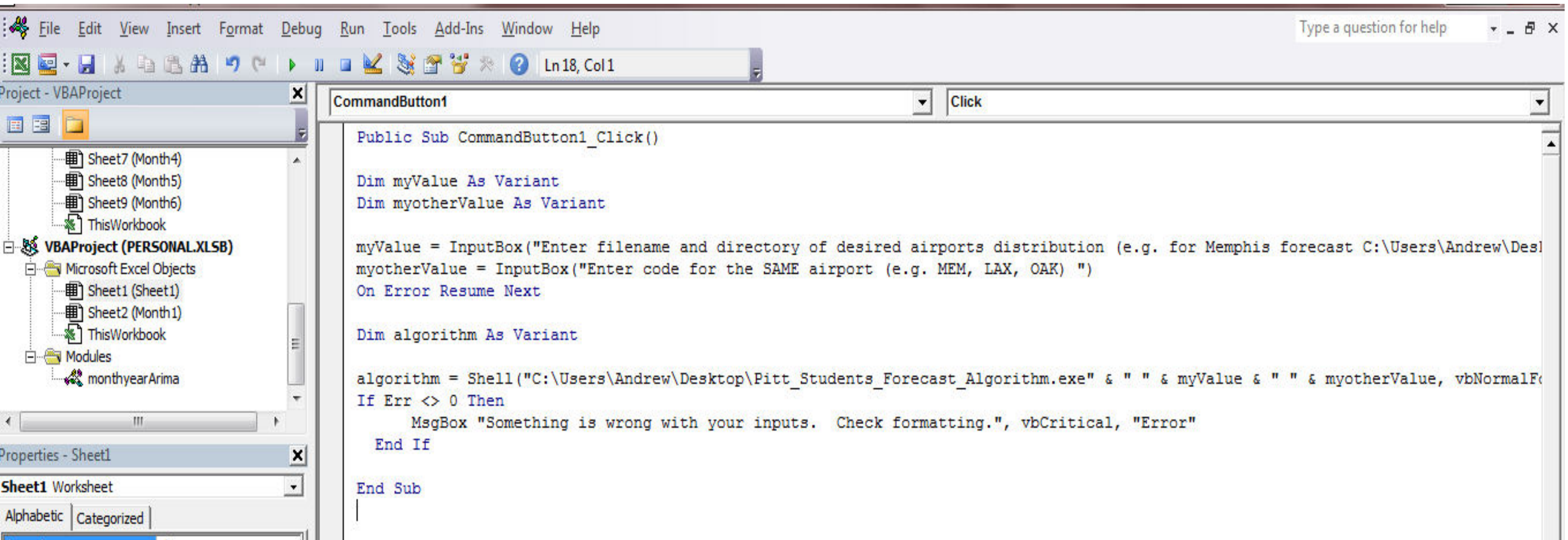
myValue = InputBox("Enter filename and directory of desired airports distribution (e.g. for Memphis forecast C:\Users\Andrew\Desktop\MEM.txt)")
myotherValue = InputBox("Enter code for the SAME airport (e.g. MEM, LAX, OAK) ")
On Error Resume Next

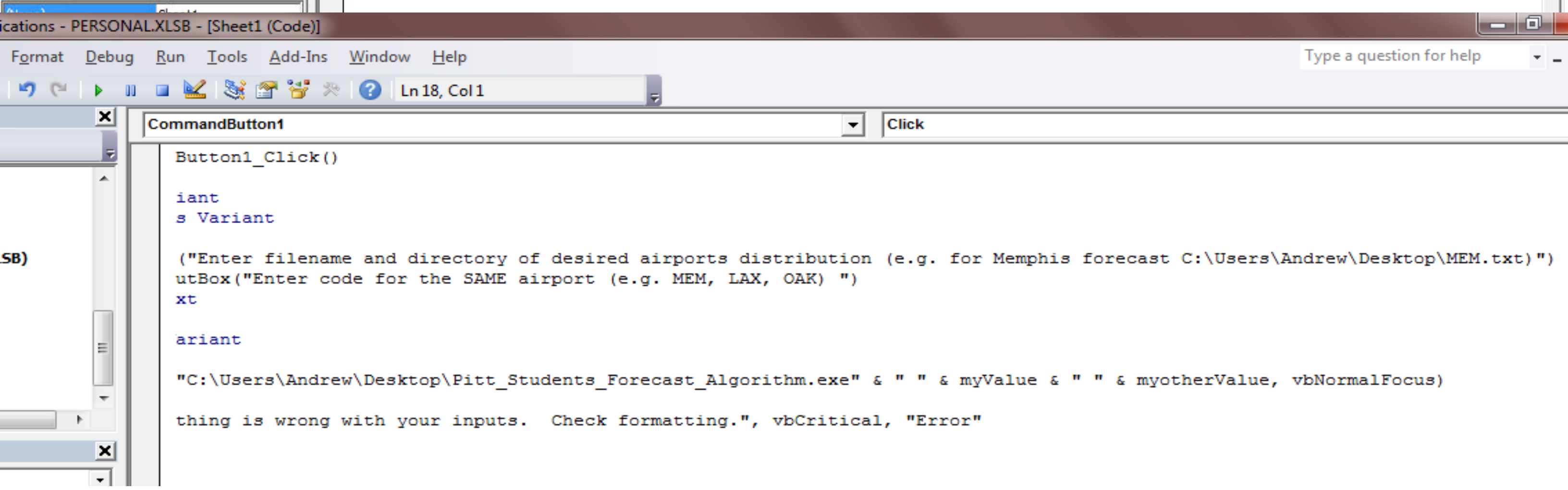
Dim algorithm As Variant

algorithm = Shell("**C:\Users\Andrew\Desktop\Pitt_Students_Forecast_Algorithm.exe**" & " " & myValue & " " & myotherValue, vbNormalFocus)
If Err <> 0 Then
MsgBox "Something is wrong with your inputs. Check formatting.", vbCritical, "Error"
End If

End Sub

This line is bolded because this must be edited manually. You must input the path/directory of the Pitt_Students_Forecast_Algorithm. Example is as seen.



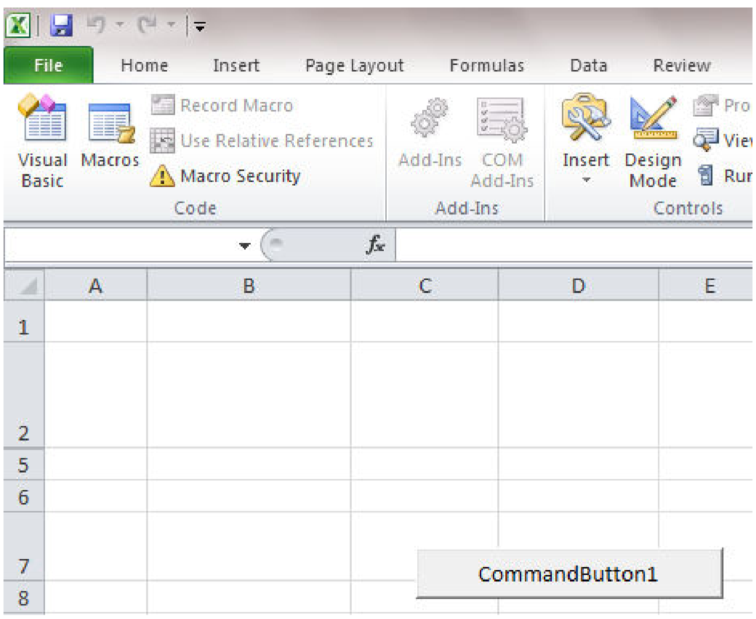


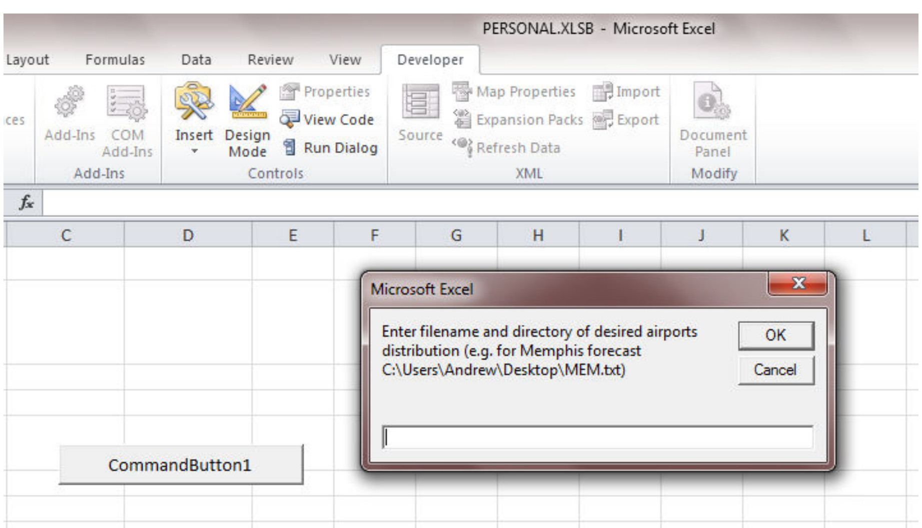
What you need to run the forecasting Macro:

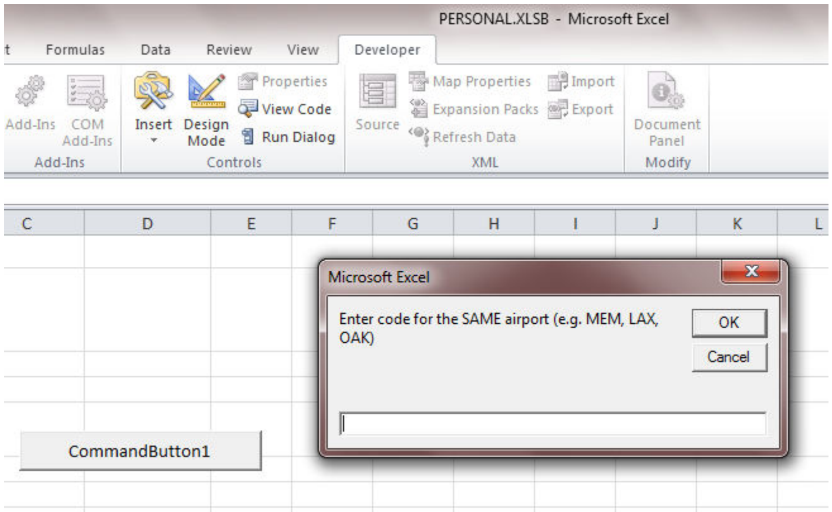
1) An Excel or Text file containing a single column of **monthly totals** (without any column headers- just the numbers).

2) You'll need to know the path/ directory of that same file above, e.g. *C:\Users\Andrew\Desktop\FILE.TXT*

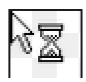
3) Click the button of the macro given in your Excel worksheet and provide inputs as instructed. Note that due to the data we independently found on the internet (from government website), there are a limited amount of airport "codes" that you may use. A full list of the airport codes is given on the **last page of this document.**







4) After about 4-5 minutes of runtime, the resulting file is called MonthlyForcasts.xlsx. The default location will be in your documents, but if it doesn't appear there, simply run a search on your computer. *Note that there isn't a popup indicating the program is finished. It's noticeable in the sense that your mouse will no longer be showing a "loading graphic"*



5) Open the excel file and view the results! A sheet called "Forecasted Monthly Totals" holds the next 12 months of predicted volume through the airport chosen in step 1. The remaining 12 sheets hold each months predicted distribution to the *other* airports. So if in step 1 you input LAX.xls and then the code LAX, you'll get forecasted monthly totals for LAX itself, and in the other sheets you'll find the amount of the totals that is distributed to other airports, such as OAK or MEM. Shipping totals are all in **pounds.**

Airport Codes:

AFW
ANC
ATL
BOS
DEN
DFW
DTW
EWR
HNL
IAD
IAH
IND
JFK
LAX
MEM
MIA
MSP
OAK
ONT
ORD
PDX
PHL
PHX
SAN
SAT
SEA
SLC
TPA