Using Data Mining Technique to Predict Seasonal Climate Change

Bing Gao, Saad Malik, Yudi Santoso, Zhanxue Zhu
Department of Computer Science,
University of Victoria
May 12, 2016

This report was written as part of the requirements for SENG 474 / CSC 578D: Data Mining, offered at the University of Victoria, taught by Dr. Alona Fyshe.

***Abstract*—In this paper, we use data mining techniques and statistical methods to analyse weather data. We collected and processed records from six weather stations located at various locations in the southern part of Vancouver Island. Weather data is an example of time series data which requires a special approach in data mining. In particular, we studied on how to use ARIMA to build models for our dataset. We discuss how to use the Dickey-Fuller method to test data stationarity. We also discuss how to determine the optimum ARIMA model parameters, by using ACF/PACF, and AIC/BIC. Finally, we show some examples of our analyses and discuss the results.**

Keywords – Climate; Weather Prediction; Data Mining; Time Series; ARIMA;

I. Introduction

Weather and climate prediction play an important role for agriculture and industry. A great proportion of agricultural and industrial activities are strongly affected by climate conditions.

From the view of short-term, the temperature and precipitation are basic condition for crop growth and yield in agriculture. Each crop has its own minimum, optimal, maximum temperature for growing. The crop stops growing when the temperature goes below the minimum temperature. The crop growth increased as the temperature goes up from the minimum to the maximum temperature. However the crop growth decreases as the temperature goes beyond its optimal temperature to the maximum temperature. The crop growth stops again when the temperature reaches its maximum temperature. Warmer temperature may favor some crops to grow more quickly and increase their yields, but it could also reduce growth and yields for other kinds of crops. So accurate prediction of future weather condition could help farmers select the proper crops in order to increase growth and yield as well as economic incomes. The fast growth of crops such as grains may reduce the amount of time that seeds need to grow and mature [1]. The crop growth not only depends on the temperature, but also on soil water and nutrient elements such as nitrogen, phosphorus and potassium. The soil nutrients are adsorbed by crops with soil water adsorption. The proper soil moisture is strongly related to the precipitation. The seasonal distribution of precipitation also affects the crop growth and yields. Similar to responding to the temperature, some crops favor the wet weather condition, others favor dry weather condition. So accurate prediction of future precipitation could help farmers select the crops too in order to maximum crop growth yields as well as economic income.

Aviation, airports, and airlines are strongly affected by the weather condition. A bad weather condition may result in flight cancellations and substantial losses for both consumers and the airline industry. Some business losses caused by the

cancellation of flight due to bad weather condition are not compensated. Business traveling to meetings and conferences are major revenue to many airlines and regional carriers, but canceled events are often not rescheduled. As a result, revenue lost due to bad weather is unlikely to be recovered [2]. Industries such as constructions need to consider weather conditions for planning and the operations. Heavy wind and extreme low temperature not only affect the construction quality, but also injury the workers, and damage the equipment and machines [3]. Extreme weather condition is the main reasons of natural disasters. Flooding and storms were the main cause of property damages. The extreme high temperature often causes people die and more common to affect adversely human life. An accurate prediction for future weather could help prevent the loss of property and life.

The earth system and climate condition has been changed by the human activities. The third report of the Intergovernmental Panel on Climate Change (ICPP) stated: "There is now new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities." The global warming is referred to the increase the mean air temperature as a result of increased atmospheric loading of greenhouse gases such as carbon dioxide from fossil fuel combustion. So the prediction of long-term impacts of human activities on global weather condition in terms of temperature and precipitation is critical to maintain a health global living environment for human being [4].

The short and long term weather simulation and forecast have been a scientific research area for meteorologists. With the development of computer technology, the integration of complicated mathematical models with computer algorithms has been used to effectively simulate and predict the short- and long-term climate and weather condition for local,regional, and global scale. Although the mathematical models and computer algorithms have been developed, the accuracy of simulation and prediction are still needed to be improved. The objectives of this course project was to predict seasonal climate change using temperature as a indictor based on the data collected from weather stations located in Vancouver Island. These weather stations are operated by the Environment Canada. To achieve the objective, we used the time series model ARIMA in this project. The ARIMA was trained and tested by the temperature data from different weather stations.

II. Related Works

Weather forecast could be back up to a century ago when the forecast was imprecise and unreliable. The reasons are that the weather stations were not enough to collect enough data for building mathematical and statistical models that can be used to simulate and forecast the weather conditions. In addition, the observations in those weather stations were often irregular. Especially, the observations for the upper atmospheric layers and over the ocean were extremely difficult. It was impossible to calculate massive atmospheric physics data by hands. The weather forecast had significantly breakthrough as computer technology advanced. The first time of computer used in meteorology study was in 1950 when the Electronic Numerical Integrator and Computer (ENIAC) were used in weather condition simulation in Aberdeen, Maryland of the Unites States [5], [6]. Using advanced computer technology, meteorological scientists could study the complex atmospheric systems by integrating the numerical models with computational algorithms to simulate and predict short- and long-term climate and weather changes.

For the long term, various general circulation models (GCMs) of the atmosphere and ocean have been developed based on the human activities such as increased greenhouse emission in terms of carbon dioxide caused by the combustion of fissile fuel [4], [6]. The GCMs were trained using the weather data from past centuries and predict the climate change over the next couple of centuries based on the assumed scenarios of greenhouse mission in the future. The Canadian Centre for Climate Modelling and Analysis has developed the Canadian Middle Atmosphere Model (CMAM) and the Canadian Regional Climate Model (CRCM) to simulate and predict regional climate changes in Canada and the world over the past and future centuries based on the different scenarios of greenhouse emission [7]. Hawkins and Rowan [8] studied the importance of narrowing uncertainty on regional climate predictions. They also introduced several approaches in order to reduce prediction uncertainty. Giorgi [9]

provided an example about using Limited Area Model (LAM) to simulate and predict January climate over the western US. Through this example, he introduced other three models (T42, R15 and MM4).

For the short term weather forecast, Radhika and Shashi [10] used support vector machine (SVM) for weather prediction for a chosen location. The SVM was used to predict the next day maximum temperature based on the maximum temperature of the previous days. The results of the SVM prediction were compared with the Multi Layer Perception (MLP), the SVM performed better than the MLP. The mean square error of the MLP prediction varies from 8.07 to 10.2 and the SVM from 7.07 to 7.56. Patel and Christian [11] used fuzzy set theory to predict temperature for inland cities of India based on the mean sea level pressure and relative humidity. The results showed that the fuzzy set theory reproduced the observed temperature well with root mean square error equal to 2.59. De and Debnath [12] used an artificial neural network (ANN) to predict the maximum and minimum temperature for the summer monsoon months, i.e., June, July, and August, in India. The ANN-based predicted mode was a three-layered feed forward neural net. In both cases, maximum and minimum temperature were greatly predicted in the month of August. In the largest part of the cases, prediction error lies below 5%. Pal et al. used a hybrid neural network to predict atmospheric temperature [13]. The self-organizing feature map (SOFM) and MLPs were used to build the hybrid network named SOFMMLP. The SOFMMLP has been compared with other local and global predictors and has been found to produce a much better prediction than others.

Besides the temperature, the data mining techniques were also used to predict recipitation and wind speed. Guhathakurta used the deterministic artificial neural network model to predict the monsoon rainfall in the next year based on the past years of monsoon rainfall data [14] . The deterministic artificial neural network model predicted monsoon rainfall well for the 36 meteorological sub-divisions. The Mohandes et al used the SVM to predict wind speed. The results were compared with the MLP. It showed that the SVM performance was better than the MLP [15].

III. Data Resources

The climate data were retrieved from the Environment Canada historical online climate database. The Environment Canada has operated over eight thousand weather stations across the country to record weather conditions since 1840. The data attributes include: maximum, minimum and average temperature; total rain, snow and precipitation; heating and cooling degree day, dew point temperature, relative humidity, wind direction and speed, visibility, atmosphere pressure, and cloudy condition in hourly, daily and monthly time step. The data also include the station geographical information, in terms of latitude, longitude, and elevation. These data can be directly downloaded from the Environment Canada website [16]. The climate data varies from station to station in details, time duration and attributes. Some stations have more details in weather records than the others.

Missing data often occur due to the failure of the equipment. The missing data was filled using Lagrange polynomial interpolation [17]. Its mathematical equation is expressed as

$$\begin{aligned} P(X) = &\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \\ &\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \\ &\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2) \end{aligned} \quad (1)$$

where $f(x_0)$, $f(x_1)$and $f(x_2)$ are the observed temperature value at the time x_0, x_1, and x_2, respectively.

Six weather stations were selected for this project. They are located in Pender island, Victoria international airport, Shawnigan lake, Saanichton town, Victoria Hightland, and the University of Victoria (Fig. 1). These stations were chosen mainly because the weather data were relatively consistence.

The weather station in Pender Island is located at latitude N $48°76'0''$ and longitude W $123°29'0''$ and data covered from 1972 to 2007; Shawnigan Lake at N $48°56'0''$ and W $123°29'0''$ from 1911 to 2007; Saanichton Town at 48 o N $48°62'0''$ and W $123°42'0''$ from 1914 to 2007; the University of Victoria at N $48°46'0''$ and W $123°30'0''$ from

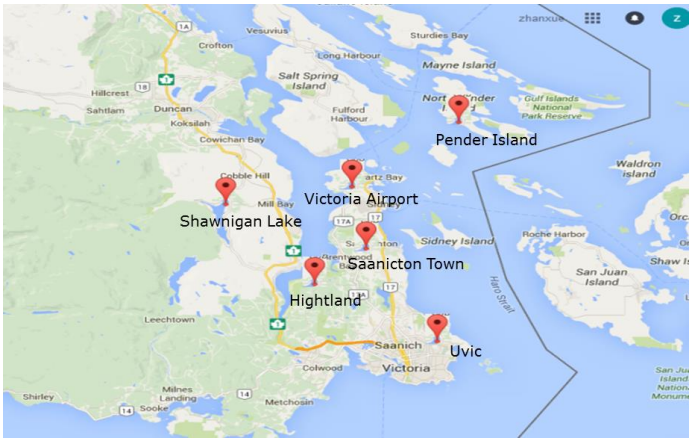

Fig. 1: Geographic location of six weather station in Vancouver Island

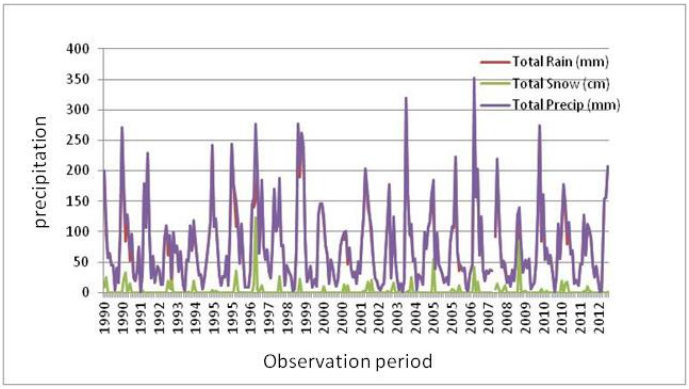

Fig. 3: Rain, snow and total precipitation in the University of Victoria

1992 to 2007; Victoria Hightland at N $48°51'0''$ and W $123°52'0''$ from 1961 to 2007; and Victoria International Airport at N $48°65'0''$ and W $123°43'0''$ from 1940 to 2016. A typical temperature variations for the University of Victoria are shown in Fig. 2. The figure shows that the extreme maximum and minimum temperature was around 35^oC and -7.5^oC, respectively.

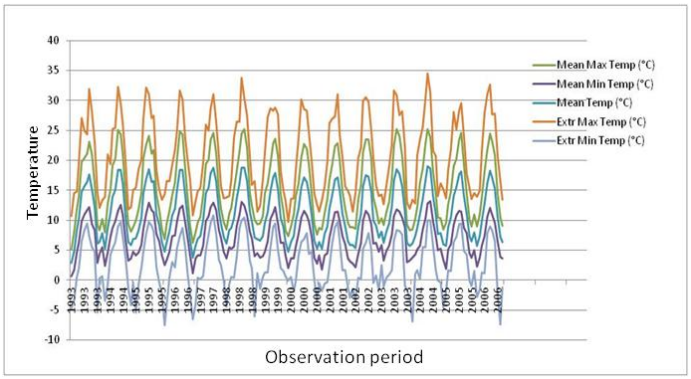

Fig. 2: Variation of temperature data in the University of Victoria

The extreme temperature changed around 8^oC from year to year. Average mean temperature range from 4^oC to 18^oC. The precipitation major consists of rain and a little snow (Fig. 3).

IV. Modeling

A. Time Series

Data based on time intervals is called time series (TS), and there are many ways and methods for analysis, process and forecast. One of them, which deals with time based data is Time Series Modeling. As the name suggests, it involves working on time (yearly, monthly, daily, etc) based data, to unveil hidden aspects in order to make informed decision. Time series models are useful when we have serially correlated data. Most businesses implement some kind of time series analysis, such as to analyze sales number for the next year, predict peek website traffic, and market competitive position, just to name a few.

Now we know Time series (TS) is a collection of data points collected at constant time intervals. These are analyzed to determine the long term trend so as to forecast the future or perform some other form of analysis. But what makes a TS different from say a regular regression problem? There are two reasons:

Time dependent: The basic assumption of a linear regression model is that the observations are independent, which does not hold in this case. The value of an attribute at an instant might depend on the value at the time before.

Seasonality trends: Along with an increasing or decreasing trend, most TS have some form of seasonality trends, i.e. variations specific to a particular time frame. For example, if you see the sales of a woolen jacket over time, you will invariably find higher sales in winter.

Solution to such a problem of time series data is autoregressive method. In an autoregressive model,

we forecast the variable of interest using a linear combination of past values of the variable. The term autoregressive indicates that it is a regression of the variable against itself. Thus an autoregressive model of order P can be written as

$$y_t = c + \phi_1 y_{t1} + \phi_2 y_{t2} + \ldots + \phi_p y_{tp} + e_t \qquad (2)$$

where c is a constant value and e_t is white noise. This is like a multiple regression but with lagged values of y_t as predictors. We refer to this as an AR(P) model. Autoregressive models are remarkably flexible at handling a wide range of different time series patterns.

`Sklearn` [18] is an open source machine learning library for the Python. Unfortunately, at this moment, it does not have a complete set of tools for time series analysis. In this project we mainly used `pandas` [19] and `statsmodels` [20]. `Pandas` is a python library written for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. `Pandas` introduces two new data structures to Python **Series** and **DataFrame**, both of which are built on top of `NumPy` [21]. `Pandas` also has dedicated libraries for handling time series objects, particularly the `datetime64[ns]` class which stores time information and allows us to perform some operations really fast.

The key library is `statsmodels`, which is a Python library that allows users to explore data, estimate statistical models, and perform statistical tests. An extensive list of descriptive statistics, statistical tests, plotting functions, and result statistics are available for different types of data and each estimator. This module provides a comprehensive autoregressive integrated moving average (ARIMA) model for time series analysis. We also used `matplotlib.pylab` [22] for making graphs.

B. Statistical stationarity

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc., are all constant over time. Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary (e.g.,"stationarized") through the use of mathematical transformations. Stationarizing a time series through differencing (where needed) is an important part of the process of fitting an ARIMA model.

Dickey Fuller Test of Stationarity This is one of the statistical tests for checking stationarity. Here the null hypothesis is that the TS is non-stationary. The test results comprise of a Test Statistic and some Critical Values for difference confidence levels. If the Test Statistic is less than the Critical Value, we can reject the null hypothesis and say that the series is stationary.

ARMA Model ARMA models are commonly used in time series modeling. In ARMA model, AR stands for auto-regression and MA stands for moving average. And ARMA model is built not to be applied on non-stationary time series so it is important to always have stationary data for ARMA model.The primary difference between an AR and MA model is based on the correlation between time series objects at different time points. The correlation between x(t) and x(t-n) for $n >$ order of MA is always zero. This directly flows from the fact that covariance between $x(t)$ and $x(t-n)$ is zero for MA models. The correlation of $x(t)$ and $x(t-n)$ gradually declines with n becoming larger in the AR model.

ARIMA Model ARIMA is a generalization of ARMA where we have integrated (I) part.

Cross-validation for time series Cross-validation is primarily a way of measuring the predictive performance of a statistical model. One way to measure the predictive ability of a model is to test it on a set of data not used in estimation. Data miners call this a "test set" and the data used for estimation is the "training set". However, there is often not enough data to allow some of it to be kept back for testing. A more sophisticated version of training/test sets is leave-one-out cross-validation (LOOCV).One of Variations on cross-validation is k-fold cross-validation (where the original sample is randomly partitioned into k subsamples and one is left out in each iteration). When the data are not independent, cross-validation becomes more difficult as leaving out an observation does not remove all the associated information due to the correlations with other observations. A

cross-validation method would works as follows [24] :

1) Fit the model to the data $y_1, \ldots, y_t$ and let $\hat{y}_{t+1}$ denote the forecast of the next observation. Then compute the error $(e^*_{t+1} = y_{t+1} - \hat{y}_{t+1})$ for the forecast observation.

2) Repeat step 1 for $t = m, \ldots, n1$ where m is the minimum number of observations needed for fitting the model.

3) Compute the MSE from $(e^*_{m+1}, \ldots, e^*_n)$.

Time-series (or other intrinsically ordered data) can be problematic for cross-validation. If some pattern emerges in year 3 and stays for years 4-6, then your model can pick up on it, even though it wasn't part of years 1 & 2.

V. Analyses and Results

For the analysis we take samples from our data collection. The first sample that we looked at is the monthly mean temperature at the Shawnigan Lake station from January 1991 to December 2000. We use the python library `pandas` to extract and label the data (using the months). The dataset that we have is a times series, with one temperature data for each month. Naturally, we do not have many data points in our dataset. Our preliminary study had shown the difficulty of using simple linear regression with polynomial features to built a model; namely, on determining the degree of the polynomials, and the related problem of overfitting/underfitting.

We then studied data mining methods that are suited for time series [1]. This lead us to the ARIMA which was described above. The python library `statsmodels.tsa` (tsa stands for time series analysis) has ARIMA class that we use for our computations. To use ARIMA we need to check the stationarity and invertibility of our dataset. The AR requires that the dataset is stationer, while the MA requires that the dataset is within the invertibility boundary.

Our first step in the analysis is to visualize the dataset in order to get a preliminary understanding of its properties. This is shown in Fig. 4 (the blue line). We have also computed the rolling average (red line) and the rolling standard deviation (gray line). These, at any instant of time, are calculated for the previous twelve months; hence the first twelve months are undefined.

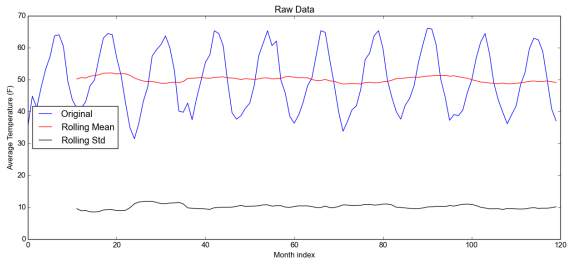

Fig. 4: Monthly mean temperature at Shawnigan Lake from 1991 to 2000, with rolling average and standard deviation.

We see a periodicity, of period twelve months, which is as expected for weather data. However, there is also year to year variation, which although is not relatively large is significant. This is exactly the problem that we need to deal with in building our model.

Visual presentation is not reliable for checking stationarity. To check the stationarity of the dataset we employ the Dickey-Fuller test, which is a statistical measure on the correlation among values at different times. The `statsmodels` library also provides a function to do this test. For this raw data, the result is shown in Table I.

TABLE I: The Dickey-Fuller test result for the raw dataset.

Test Statistic	-2.466812
p-value	0.123760
Critical Value (5%)	-2.889217
Critical Value (1%)	-3.493602
Critical Value (10%)	-2.581533

The test statistic for this dataset is less negative than the critical values, even at 10% (which corresponds to the 90% confidence level). Alternatively, we can also look at the p-value, and for this one it is greater than 0.1. Therefore, this raw dataset does not satisfy the stationarity requirement.

[1]We are indebted to Aarshay Jain [23] for giving us insight on how to do time series analysis using python.

Taking a log on the data can alleviate the stationarity problem as it penalize the high values. Indeed this is what we got with the Dickey-Fuller test, as shown in Table II.

TABLE II: The Dickey-Fuller test result for the log dataset.

Test Statistic	-2.651400
p-value	0.082832
Critical Value (5%)	-2.889217
Critical Value (1%)	-3.493602
Critical Value (10%)	-2.581533

In fact, the log dataset stationarity is acceptable at the 10% level, so we can use this dataset for some ARIMA models, but let us go a bit further. There are many ways to stationarize a dataset, which can be categorized as detrending, differencing, or seasonal decomposition. Here we use a differencing method, in which we take differences of the log of a data point and the previous one. We will refer this as Log-Diff from here on.

The Dickey-Fuller test result for our Log-Diff data set is shown in Table III.

TABLE III: The Dickey-Fuller test result for the Log-Diff dataset.

Test Statistic	-7.976118e+00
p-value	2.699565e-12
Critical Value (5%)	-2.888697e+00
Critical Value (1%)	-3.492401e+00
Critical Value (10%)	-2.581255e+00

We can see that now the p-value is tiny, which means that the dataset is stationer. Thus, we can then proceeds to the modeling using ARIMA. First, we need to determine the values of the ARIMA parameters (p, d, q) that we should use. One method is by using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). We computed ACF and PACF for this dataset, and the result is shown in Fig. 4.

The ACF graph is gradually decreasing, and this suggests (roughly) that we do not need MA, so $q = 0$. However, we will discuss more on this later. The PACF graph is decaying fast (although with some chaotic behaviour later on), and crossing the threshold at around 1 or 2. This suggests that the

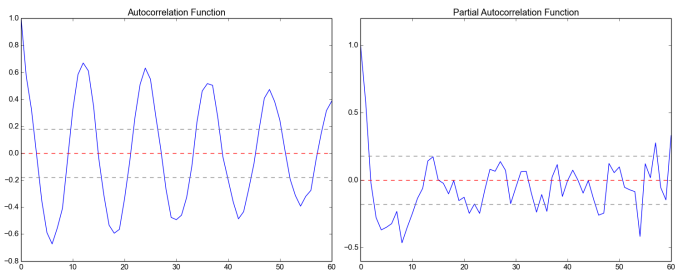

Fig. 5: The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the Log-Diff dataset for the Shawnigan Lake 1991-2000 data (Fig. 4).

value of p is either 1 or 2. It turns out (when we tried both) that $p = 1$ yields a (slightly) better fit. So, we choose $p = 1$. Since the differencing that we do is first order, we take $d = 1$ (but we can also take a look at the log dataset instead, and set $d = 0$). Thus, we take $(p, d, q) = (1, 1, 0)$, and we proceed with fitting the ARIMA model with this set of parameters. The result is shown in Fig. 6.

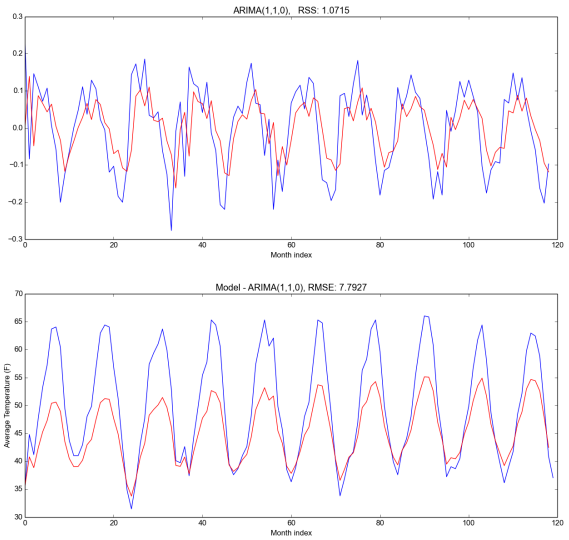

Fig. 6: ARIMA model/fit on the Shawnigan Lake dataset for $(p, d, q) = (1, 1, 0)$ (top), and the fit in temperature (bottom) . Note: the fit at the top is shifted one time step to the right.

Also, we got a model that is shrank from the original data. When shown in the original temperature scale, the model yields lower numbers compared to the original data. Turns out, the difference can be

alleviated by a multiplication (adjustment) factor. Multiplying by 1.4 gives Fig. 7.

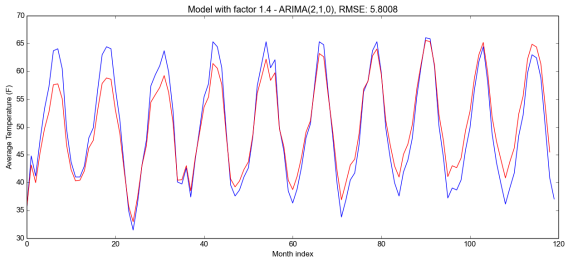

Fig. 7: ARIMA(1,1,0) fit in temperature as in (Fig. 6) with adjustment factor of 1.4.

The prediction, on the other hand, is quite good. The in-sample prediction, for the year 1994, is shown in Fig. 8. Although not a perfect match, the prediction (red line) is tagging along the original data. For out-sample prediction, we note that the model can only predict one or two months into the future; afterward the prediction becomes flat, hence meaningless. But, perhaps, this is the nature of weather prediction in general, that it is impossible to predict far into the future (unless we have an oracle and then we would be rich).

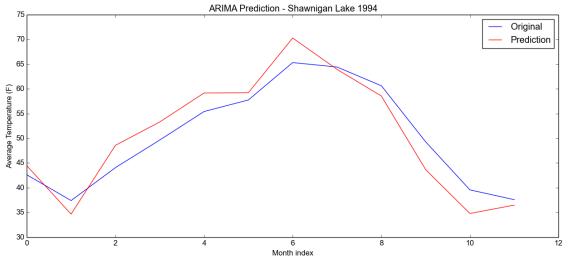

Fig. 8: In-sample prediction of ARIMA(1,1,0) for the monthly mean temperature at Shawnigan Lake in 1994 (red), compared with the real data.

Now, going back to Fig. 7, notice, however, that even with adjustment the fit is still not perfect. The ARIMA plot (green) is a bit tilted, lower in the past and higher in the future. Also, this adjustment factor is difficult to explain. We would have to dig up the ARIMA modelling mechanism in detail.

Yet another way to determine the optimum ARIMA parameters is by using the AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria), which are some kinds of statistical measures using likelihood. We should choose parameters that give us minimum AIC and BIC.

We computed the AIC and BIC on our log dataset, for the range of $0 \leq p \leq 5$, $0 \leq d \leq 2$ and $0 \leq q \leq 3$. We cannot compute the AIC and BIC for some sets of the parameters because the dataset is not invertible for those. Table IV is a partial list of the results. We include only those that are essential for illustration and for our discussion here.

TABLE IV: The AIC and BIC values for some ARIMA (p, d, q) parameters for the Shawnigan Lake 1991-2000 (Log) dataset.

p	d	q	AIC	BIC
0	0	0	-44.74	-39.17
0	0	1	-151.76	-143.40
0	1	0	-171.68	-166.12
0	1	1	-204.49	-196.16
1	0	0	-179.70	-171.34
1	0	1	-216.81	-205.66
1	1	0	-218.46	-210.13
2	0	1	-320.48	-306.54
2	1	0	-216.51	-205.40
3	0	1	-324.28	-307.55
3	1	0	-224.88	-210.98
3	2	0	-191.50	-177.65
4	0	1	-325.40	-305.89
4	0	2	-356.37	-334.07
4	0	3	-353.51	-328.42

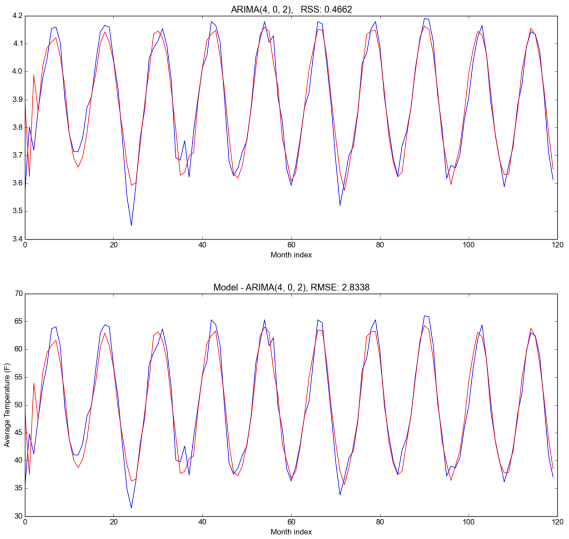

Fig. 9: ARIMA model/fit on the Shawnigan Lake dataset for $(p, d, q) = (4, 0, 2)$ (top), and the fit in temperature (bottom) .

We can see that $(p, d, q) = (1, 1, 0)$ is actually

not the minimum. The set of the parameters that minimizes AIC and BIC turns out to be $(p, d, q) = (4, 0, 2)$. Thus, we tried ARIMA model with these parameters on the log dataset (without the diff as $d = 0$). The result is shown in Fig. 9. We can see that this model gives a better fit with lower RMSE. Moreover, we do not need any adjustment parameter.

Note, however, that with $(p, d, q) = (4, 0, 2)$ we got a warning from python about convergence. We need to understand this before we could be confidence with our result. Unfortunately, due to the time constraint of the project we have not been able to look into this in more detail.

Similarly, for the other stations we searched the optimum ARIMA (p, d, q) parameters by scanning AIC and BIC values. We then use those parameters to build the model. The results are plotted as in Fig. 10. The North Pender Island station is not as reliable as the other stations. There are many missing data, and we do not have long enough continuous data segment. For this reason, we decided to drop analysing data from this station.

The next question that we looked into, is whether more data (i.e. longer period of time) in the dataset would help. To explore this problem, we sample the Shawnigan Lake data from 1951 to 2000. We found that, from stationarity point of view, more data does help. Table V is the Dickey-Fuller test result for the raw data. We can see that the p-value is already small, hence the time series is stationer. Compare this to Table I above.

TABLE V: The Dickey-Fuller test result for Shawnigan Lake 1951-2000 raw data.

Test Statistic	-4.914344
p-value	0.000033
Critical Value (5%)	-2.866493
Critical Value (1%)	-3.441578
Critical Value (10%)	-2.569408

The fitting is also a bit better. With $(p, d, q) = (3, 0, 2)$ which is slightly different from the one we used for the shorter time series above, we get RMSE = 2.5, as compared to 2.8 above. However, one might argue that the improvement is not much. The model is plotted in Fig. 11. We notice that the

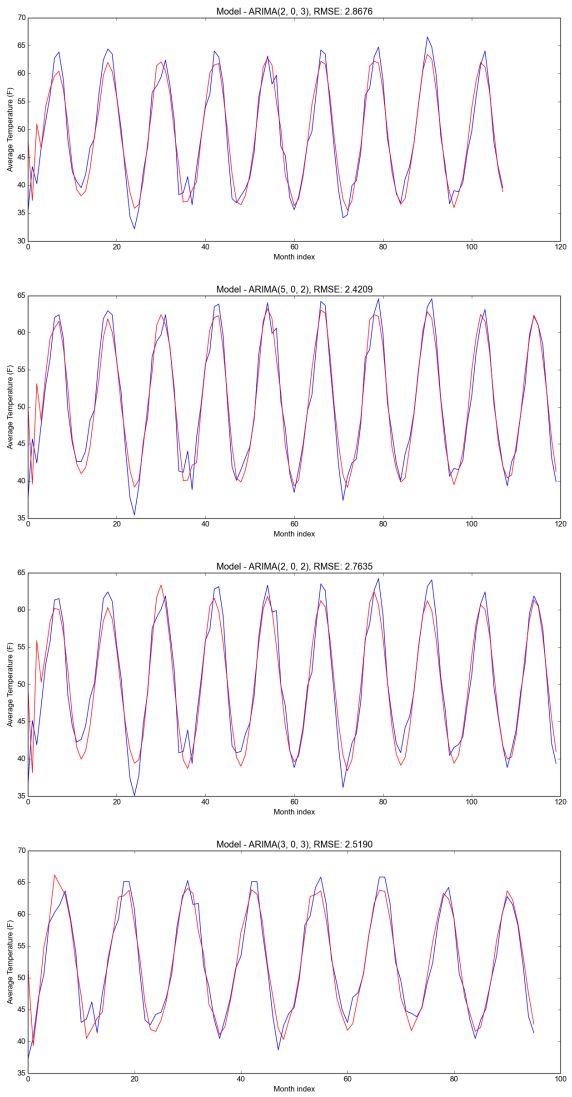

Fig. 10: ARIMA model/fit for (a) Victoria Highland 1991-1999 using ARIMA(2,0,3), (b) Saanichton 1991-2000 using ARIMA(5,0,2), (c) Victoria International Airport 1991-2000 using ARIMA(2,0,2), and University of Victoria 1993-2002 using ARIMA(3,0,3).

outlier maximum and minimum points are still not caught by the model.

We choose to first analyze the mean temperature because it intuitively should be easier to handle than the extreme temperature. Now that we have a working program, it should work in the same way with extreme temperature. If given more time, we can expand our research on the other weather attributes as well.

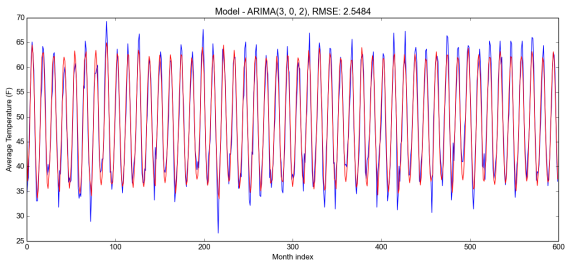

Fig. 11: ARIMA model/fit for Shawnigan Lake from 1951 to 2000 using ARIMA(3,0,2).

We do not include cross validation here due to the time and space restriction of the project.

VI. Conclusion

In conclusion, we have implemented a time series analysis tool to forecast local temperature using Python libraries. The result shows that the `statsmodels` ARIMA model works quite well. The key in using ARIMA is how to determine the best parameters. We found that visual interpretation (using ACF/PACF graphs) is not reliable. Statistical criterias such as AIC and BIC should be used instead.

We have learned a lot on time series analysis in this project; a topic which was not covered in the class. Nonetheless, it is a valuable subject. As computer scientists, we might continue studying this subject on our own.

The time series models such as ARIMA allow for the inclusion of information from the past observations of a series, but not for the inclusion of other information that may be relevant. In the future, we plan to extend ARIMA models to allow other information to be included in the models. The idea is to combine regression models and ARIMA models to give regression with ARIMA errors, which is known as dynamic regression models.

References

[1] Semenov, Mikhail A., and J. R. Porter. "Climatic variability and the modelling of crop yields." Agricultural and forest meteorology 73.3 (1995): 265-283.

[2] Cook, Andrew, and Graham Tanner. "Modelling the airline costs of delay propagation." AGIFORS Airline Operations Conference, London, UK. 2011.

[3] Alshebani, Mohammed N., and Gayan Wedawatta. "Making the construction industry resilient to extreme weather: lessons from construction in hot weather conditions." Procedia Economics and Finance 18 (2014): 635-642.

[4] Weaver, Andrew J. "The science of climate change." Hard choices climate change in Canada. Wilfred Laurier University Press, Waterloo, Canada (2004): 13-14.

[5] Charney, Jules G., Ragnar Fjrtoft, and J. von Neumann. "Numerical integration of the barotropic vorticity equation." Tellus 2.4 (1950): 237-254.

[6] Lynch, Peter. "The origins of computer weather prediction and climate modeling." Journal of Computational Physics 227.7 (2008): 3431-3444.

[7] CCCma. the Canadian Middle Atmosphere Model. Accessed at April 2, 2016. http://www.ec.gc.ca/ccmac-cccma/default.asp?lang=En&n=4A642ED

[8] Hawkins, Ed, and Rowan Sutton. "The potential to narrow uncertainty in regional climate predictions." Bulletin of the American Meteorological Society 90.8 (2009): 1095-1107.

[9] Giorgi, Filippo. "Simulation of regional climate using a limited area model nested in a general circulation model." Journal of Climate 3.9 (1990): 941-963.

[10] Radhika, Y., and M. Shashi. "Atmospheric temperature prediction using support vector machines." International Journal of Computer Theory and Engineering 1.1 (2009): 55.

[11] Patel, Dipi A., and R. A. Christian. "Ambient atmospheric temperature prediction using fuzzy knowledge-rule base for inland cities in India." World Appl. Sci. J 20 (2012): 1448-1452.

[12] De, S. S., and A. Debnath. "Artificial neural network based prediction of maximum and minimum temperature in the summer monsoon months over India." Applied Physics Research 1.2 (2009): 37.

[13] Pal, Nikhil R., et al. "SOFM-MLP: A hybrid neural network for atmospheric temperature prediction." Geoscience and Remote Sensing, IEEE Transactions on 41.12 (2003): 2783-2791.

[14] Guhathakurta, P. "Long lead monsoon rainfall prediction for meteorological sub-divisions of India using deterministic artificial neural network model." Meteorology and Atmospheric Physics 101.1-2 (2008): 93-108.

[15] Mohandes, M. A., et al. "Support vector machines for wind speed prediction." Renewable Energy 29.6 (2004): 939-947.

[16] Environment Canada. Historical Climate Data, accessed January 20, 2016, http://climate.weather.gc.ca/advanceSearch/searchHistoricData_e.html

[17] Chapra Steven, C., and P. Canale Raymond. Numerical methods for engineers. Tata McGraw-Hill Publishing Company, 2008.

[18] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[19] http://pandas.pydata.org/

[20] http://statsmodels.sourceforge.net/

[21] http://www.numpy.org/

[22] http://matplotlib.org/

[23] http://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

[24] http://robjhyndman.com/hyndsight/crossvalidation/