A MULTIGENE GENETIC PROGRAMMING APPROACH ON WEATHER FORECASTING

NESLİHAN ÇEVİK

Master's Thesis

Graduate School

Izmir University of Economics

Izmir

2020

A MULTIGENE GENETIC PROGRAMMING APPROACH ON WEATHER FORECASTING

NESLİHAN ÇEVİK

A Thesis Submitted to

The Graduate School of Izmir University of Economics

Industrial Engineering Program

İzmir

2020

ABSTRACT

A MULTIGENE GENETIC PROGRAMMING APPROACH ON WEATHER FORECASTING

Çevik, Neslihan

M.S. in Industrial Engineering

Advisor: Prof. Dr. Ahmet Sermet ANAGÜN

June, 2020

Weather and precision of weather forecasts have a very important role in our daily lives especially in the field of transportation since it directly affects the quality and the safety of the service. In this study, the aim was to compare the forecast errors executed by different forecasting approaches. The data has been provided by Republic of Turkey Ministry of Agriculture and Forestry, General Directorate of Meteorology for Izmir Adnan Menderes Airport with eight independent variables and the daily average temperature and daily average wind speed as the dependent variables for the years 2015-2017. Results show that Multi-Gene Genetic Programming Approach and Gaussian Regression with kernels; Rational Quadratic and Squared Exponential models have lower RMSE values compared with the SVR and ANN.

KEYWORDS: weather forecasting, support vector regression, multiple regression, nonlinear regression, data mining, multi-gene genetic programming, artificial neural network

ÖZET

ÇOKLU GEN GENETİK PROGRAMLAMA YAKLAŞIMI İLE HAVA TAHMİNİ

Çevik, Neslihan

Endüstri Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Prof. Dr. Ahmet Sermet Anagün

Haziran, 2020

Hava durumu ve hava durumu tahminlerinin kesinliği, özellikle yolcu taşımacılığı alanında ve günlük yaşamımızda çok önemli bir role sahiptir, çünkü hizmetin kalitesini ve güvenliğini doğrudan etkiler. Bu çalışmada amaç, farklı tahmin yaklaşımları analizler sonucu elde edilen tahmin hatalarını kullanarak karşılaştırmaktır. Veriler, TC Tarım ve Orman Bakanlığı, İzmir Meteoroloji Genel Müdürlüğü Adnan Menderes Havalimanı tarafından 2015-2017 yılları için, sekiz bağımsız değişken ile günlük ortalama sıcaklık ve rüzgar hızı iki bağımlı değişken olacak şekilde sağlanmıştır. Çoklu Gen Genetik Programlama Yaklaşımı ve Gaussian Regresyonunun, SVR ve ANN ile karşılaştırıldığında daha düşük RMSE değerleriyle daha başarılı bir performansa sahip olduğunu göstermiştir.

ANAHTAR KELİMELER: hava durumu tahmini, destek vektör regresyonu, çoklu regresyon analizi, doğrusal olmayan regresyon, veri madenciliği, çoklu gen genetik programlama, yapay sinir ağı.

Dedicated to my father

ACKNOWLEDGMENTS

First and foremost, I am most grateful to my supervisor Prof. Dr. Ahmet Sermet Anagün for his endless guidance, insights, encouragement and kindness during the course of this thesis.

I want to thank Prof. Dr. Gözde Yazgı Tütüncü and Asst. Prof. Fehmi Burçin Özsoydan for accepting to be part of thesis jury.

Many thanks to IEU family, both academic staff and my precious colleagues for all their guidance and support.

I would like to thank my friend, Begüm Kanat for always being there to share her support and guidance through every stage of my life.

Lastly, I would like to thank my family for always being there to support me and making everything possible. I am grateful to have my parents Mehmet Çevik and Gülfem Çevik for their endless support and love. My twin, Aslıhan Çevik for being the best companion a person could ask for.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANN	Artificial Neural Network
MGGP	Multi Gene Genetic Programming
SVR	Support Vector Regression
SVM	Support Vector Machine
RMSE	Root Mean Squared Error
GPR	Gaussian Process Regression
MLR	Multiple Linear Regression
ADB	Izmir Adnan Menderes Airport
GP	Genetic Programming
GEP	Gene Expression Programming
RT	Regression Trees
MLP	Multilayer Perceptron
BR	Bayesian Regularization Backpropagation
LM	Levenberg-Marquardt Backpropagation
RP	Resilient Backpropagation
SCG	Scaled Conjugate Gradient Backpropagation
CGF	Fletcher-Powell Conjugate Gradient Backpropagation
BFG	BFGS Quasi-Newton Backpropagation

CHAPTER 1: INTRODUCTION

1.1 Forecasting

Accuracy of the forecast is vital for many cases in everyday life since it has huge impact on costly and sensitive areas like transportation, industry, and environment. As the number and complexity of the data increases each day, the sensitivity of the precision gets more attention and extra care. Forecasting and the precision are not only some extensions of the new developments but they already are the focus. Today there are studies on creating new methods which can improve the level of precision compared to the already used methods. Linear, nonlinear regression and Artificial Neural Network (ANN) methods have been used excessively in the field of prediction. Support Vector Regression and Multi Gene Genetic Programming (MGGP) are the trending methods which are used especially when the data used for the study has a nonlinear behavior. Instead of using the stand-alone Genetic Programming (GP), studies show that MGGP has better performance with the nonlinear data. The MGGP approach has been used especially in the fields of weather, streamflow, gas consumption forecasting.

1.2 Problem Statement

Weather forecasting is one of the areas which should be strong on the point of precision and has room for improvement. The importance of the precision of the forecast depends on the data and analyzed system. This study focuses on daily average temperature and wind speed forecasting for Izmir Adnan Menderes Airport since the forecasting performance has an important role in the aviation sector and it is the reason for choosing this site for this study. ADB is the largest airport in the Aegean region in Turkey and has the total number of 7.10 million passengers for domestic and international flights for the first half of 2019 (DHMI, 2018). Another reason for choosing ADB as the site of study is the meteorology recording accuracy being a priority for the facility; leading to having high quality data for a statistical analysis. The impact of the performance of this study is also an important reason for choosing ADB because of possible improvements on cost and security of the service.

The focus of the prediction is on the daily average temperature and wind speed for this study, but it can be also be conducted for the other metrological parameters which are especially important for aviation such as daily average cloudiness, daily average atmospheric pressure. There are examples in the literature about the impacts of climate change on the takeoff performance of aircrafts. A recent research mentions that increase in temperature causes longer takeoff distances and lower climb rates. The average takeoff distance is expected to increase by 0.95-6.5% from the period (1976-2005) to the period (2021-2050) whereas the climb rate is expected to decrease by 0.68-3.4% for the previously mentioned periods (Zhou et al., 2018). A good forecasting system can be used to take preventive action and especially for the aviation, new regulations can be considered for the future.

The data which has been provided for this study is the meteorological data from Republic of Turkey Ministry of Agriculture and Forestry, General Directorate of Meteorology for Izmir Adnan Menderes Airport between years 2015-2017. The reason why data for 2018 hasn't been used is that the quality of the data was not appropriate for any analysis. In case of improvement on 2018 data, it can be added to assess the success of this study. For all of the methods, 2015 and 2016 were used for training the data whereas 2017 data was used for testing purpose. After making the regression analysis it was clear that the data used for this analysis was nonlinear and that for further analysis the methods should be applied by taking this information into consideration.

1.3 Purpose of the Study

The aim of this thesis is to minimize the forecasting errors for the weather data on predicting the daily average temperature and wind speed by comparing different regression and forecasting approaches. The performance comparisons and discussions on different methods are also included in this study.

There have been many examples of studies conducted in the area of weather forecasting. One of the studies was on predicting the fog at Canberra International Airport using artificial neural network (Fabbian, De Dear, and Lellyett, 2007) which had a very critical impact on the area of aviation security and on reducing possible risks. There have been separate studies on predicting the rainfall data using MGGP

approach and ANN (Alweshah, Ababneh, and Alshareef, 2017). There are also studies on minimizing the forecasting errors on wind speed prediction using both ANN and MGGP.

Among other studies conducted in the forecasting field, this study contributes to other works in literature on following aspects:

- MGGP approach has not been used enough in temperature forecasting considering the reachable literature
- SVR, ANN and MGGP performances on a large time series data set haven't been studied yet.
- SVR which is not common in forecasting since SVM was used more on classification cases, has been extensively analyzed.

1.4. Structure of the Thesis

The remaining chapters of this thesis are organized as follows. In Chapter 2, impact of weather parameters on aviation and the previous studies on forecasting with SVR, ANN, MGGP and other methods commonly used in the literature are reviewed. The regression methods used in this study and the implementation of the methodology are discussed in Chapter 3. Performance evaluation regarding the forecast methods are given in Chapter 4. Lastly, summary of the study and future plan regarding this analysis are provided in Chapter 5.

CHAPTER 2: LITERATURE RESEARCH

2.1 Impacts of Weather Forecasting on Aviation

Impacts of weather parameters on aviation has to be studied carefully in order to determine the focus of this study. There have been many works on determining the specific safety risks of a possible weather extreme. The following part of this section provides brief information on scientific deductions on the impacts of different weather parameters.

The importance of forecasts arises in the aspect of economy and it can only be seen by the influence on the individuals and organizations (Fabbian, De Dear, and Lellyett, 2007). The most affected party on this subject is surely the airlines and the airports considering the costs and the damaged reputations besides the possible number of people that can be affected in the case of an accident. Wind, turbulence, high density altitude, temperature extremes, lightning, visibility problem, thermal lift are some of the possible threats for the aviation that may cause accidents (Gultepe et al., 2019).

As it can be observed from Figure 2.1, wind caused aircraft accidents has the highest value of 1149 accidents between the year 2003 and 2007 followed by visibility and high-density altitude.

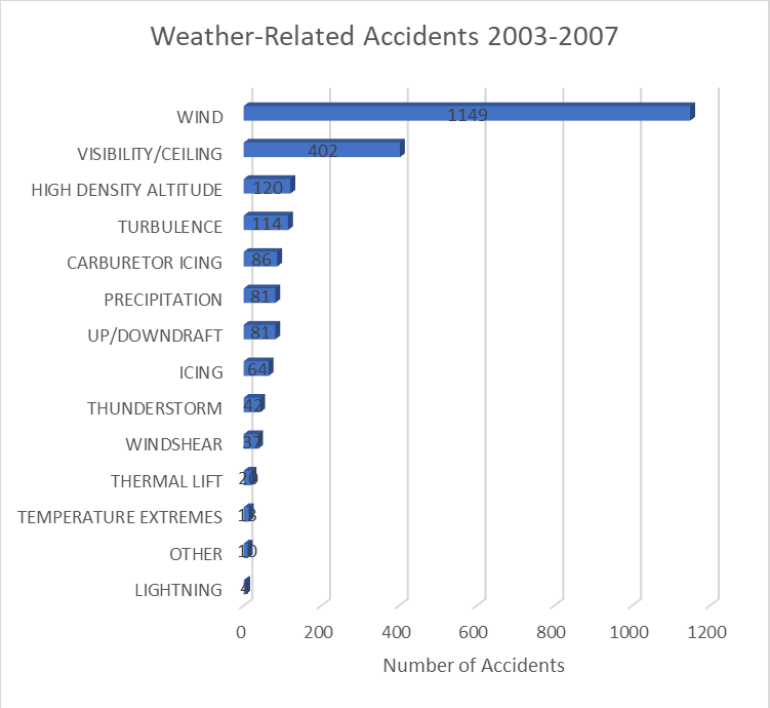


Figure 2-1 Weather-related Accidents by category (Source: ASIAS, 2010)

Also, for the same study, the phase of the flight for wind related aircraft accidents have been studied. According to the results, compared to other phases among all wind caused accidents; landing and takeoff phases have greater number of incidents with 663 and 216 incidents respectively (ASIAS, 2010). Appendix 1 shows the number of accidents between 2003 and 2007 for all of the categories during landing and takeoff actions. These two phases are very important for this study since they are connected directly with the airport, the site. Both landing and takeoff take place when the aircraft is near the airport so, an accurate weather forecasting is highly significant for many airports especially for the landing and takeoff phases possessing high risks during weather extremes. Figure 2.2 displays the accidents caused by the wind speed between the years 2003 and 2007. As it can be observed from the Figure 2.1, takeoff and landing phases of flights possess the highest risk of accidents due to the wind speed.

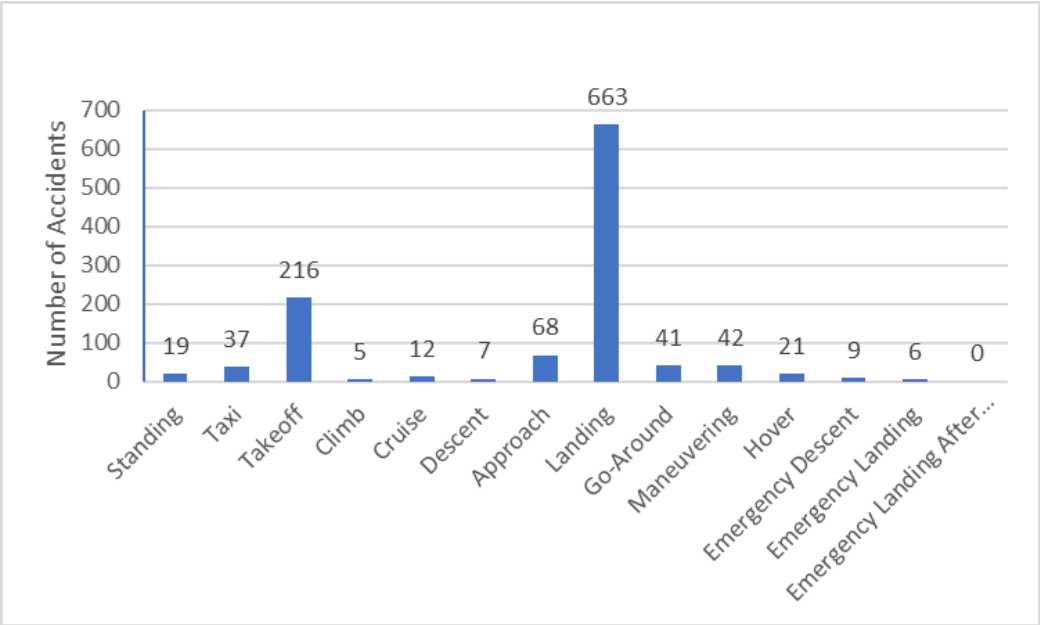


Figure 2-2 Wind Accidents Phase of Flight (Source: ASIAS, 2010)

Although not having many accidents compared to the other categories in this study, temperature extremes also have an impact on the aircraft performance. Extreme high temperatures can create high density altitudes and can even cause heating of fuels by the engines whereas extreme low temperatures can cause system to not operate as a consequence of water being present in the system and the freezing (ASIAS, 2010). As a result of an extreme low temperature having even 0.1mm of ice for 2 minutes on the wing surface can increase the drag and can reduce the airplane lift by 25-30% which can be considered as a very high risk (ASIAS, 2010).

Considering the above stated safety risks, meteorological parameters such as wind, temperature, atmospheric pressure, and relative humidity should be measured carefully with high quality, sensitive equipment and prediction tools should be improved aiming highest accuracy. For example, visibility is a very hard parameter to be predicted because it depends highly o local conditions (Fabbian, De Dear, and Lellyett, 2007).

Nowcasts and forecasts are not only important for short term period but also has an impact on the long term. In a study conducted by Zhaou et al. (2018) showed that the increase in temperature over the years can affect the takeoff performance. It is known that increasing temperature and decreasing pressure altitude decreases the

takeoff performance and this has been supported by the decrease of takeoff performance of 30 airports, data provided by the Global Surface Summary of Day Data produced by the US National Climatic Data Center due the to the climate change between the year 1976 and 2005. Increase in temperature requires longer takeoff distance and lower climb rate to meet airline safety standards (Zhou et al., 2018).

The cost aspect of effects of weather on aviation should not be neglected. The main use of Terminal Aerodrome Forecasts, TAFs by airlines and aircraft operators is for flight planning, both for before and during the flight since alternative fuel decision should be made according to the forecasts for the destination. To carry additional fuel or to encounter unexpected conditions can even turn the profit into a loss because of extra flight time or diversion to an alternate landing site (Fabbian, De Dear, and Lellyett, 2007).

2.2 Adaptations and Implementations of Current Techniques

As an example of predictive forecasting, there have been multiple studies on predicting the rainfall with different methods. There has been a survey study on rainfall predictions, and it has been observed that ANN is the most popular method used by the researchers followed by MLR, SVM and BPNN (Hirani and Mishra, 2016). This has guided the roadmap for this study as the alternative methods before implementing MGGP. Another study conducted on forecasting the rainfall has compared different kernels for SVR combined with a special data preprocessing technique and achieved to develop a successful prediction model (Hasan, Nath, and Rasel, 2016). A study on forecasting the fog occurrences for Canberra International Airport using data for 44 years using the ANN architectures aimed to improve the forecasting performance of The Australian Bureau of Meteorology (BoM) for the National Fog Project (Fabbian, De Dear, and Lellyett, 2007). There is also a study that came up with a very accurate rainfall prediction system with Multi Gene Genetic Programming (MGGP) which shapes the focus of this thesis (Alweshah, Ababneh, and Alshareef, 2017).

Another study on predicting the monthly rainfall data in Tenggarong Station, East Klimantan, Indonesia had successfully performed using Backpropagation Neural Network (BPNN) algorithm since methods such as Simple Regression, Exponential Smoothing and autoregressive integrated moving average (ARIMA) didn’t have good

performance because of data being nonlinear (Mislan et al., 2015). They have compared three different cases compared with their Mean Squared Error (MSE) performance (Mislan et al., 2015). There is also a study on predicting river flow using FFNN where they have compared AR and FFNN performances (Thota, 2018).

There are also hybrid algorithms developed by researchers that help improving the accuracy of predictions. Combining RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm gave very successful results on rainfall prediction (Wu, Long, and Liu, 2015).

Galashani et al. used both ANN and MGGP in order to predict the bond strength of GFPR bars in concrete and they both outperformed MLR methods (Golafshani, Rahai, and Sebt, 2015).

A study on predicting the dynamic travel time predictions to improve the traffic information systems in United States used the MGGP comparing the performances of different number of clusters (Elhenawy, Chen, and Rakha, 2014). Another study conducted by Faris et al on 2014 compared the prediction performances of ANN, GA, PSO and MGGP for the temperature of a metal cutting tool (Faris and Sheta, 2016). As a result of their study, MGGP outperformed all of the other methods (Faris and Sheta, 2016). There are other applications of MGGP such as predicting the student failure rates based on the previous academic performance of the students (Orove, Osegi, and Eke, 2015). This study created an application with the MGGP logic behind and they have provided the codes for any contributions (Orove, Osegi, and Eke, 2015).

Mehr et al have connected the Moving Average method with MGGP for the streamflow prediction while they have used the Pareto optimal solutions provided by the GPTIPS on their study on Şenöz Stream on 2017 (Danandeh and Kahya, 2017). This hybrid method had a better performance compared to stand-alone GP, MGGP and MLR methods but for comparing Pareto-optimal MAMGGP and MAMGGP, there were no significant difference on the accuracy of the predictions (Danandeh and Kahya, 2017). The idea behind combining MA with MGGP is reducing the complexity of the models by applying a smoothing step in the beginning (Danandeh and Kahya, 2017). One of the most significant research using the MGGP focused on predicting the global solar irradiance to utilize the solar energy (Pan, Pandey, and Das, 2013). They have observed that with the application of MGGP, they have decreased

the error of 4% with the already used ANN method to 3% with the MGGP approach (Pan, Pandey, and Das, 2013). The studies mentioned above are clear examples of MGGP approach being a popular and successful method which can be employed almost on every field.

Table 2.1 contains the list methods such as ANN, GP, MGGP, and hybrid methods used in the field of forecasting especially weather forecasting. It can be observed that MGGP has not been used extensively in the reachable literature. As a contribution to the performance comparison for MGGP, ANN has been also analyzed in the forecasting problems that use MGGP in many other studies. This concept lead this study to include ANN as a regression method. There are also hybrid implementations of MGGP and ANN which promise better forecasting performances.

Table 2-1 Literature Review

Method	Area	Relevant Literature
FFNN	Black River Flow	(Thota, 2018)
ANN/MGGP	Bond Strength	(Golafshani, Rahai, and Sebt, 2015)
Hybrid MA-MGGP	Daily Streamflow	(Danandeh and Kahya, 2017)
BPNN	Daily Temperature	(Narvekar and Fargose, 2015)
FFNN	Fog Forecasting	(Pasini, Pelino, and Potestà, 2001)
ANN	Global Solar Energy	(Khatib, Mohamed, and Mahmoud, 2012)
MGGP	Global Solar Irradiation	(Pan, Pandey, and Das, 2013)
GP	Global Solar Irradiation	(Demirhan and Kayhan Atilgan, 2015)
SVR	Global Solar Radiation	(Olatomiwa et al., 2015)
SVR	Global Solar Radiation	(Ramedani et al., 2014)
GP	Seasonal Forecasts	(Neill et al., 2012)
ANN, GP	Surface Air Temperature	(Ramesh, Anitha, and Ramalakshmi, 2015)
AR/GP	Nile River Flow	(Sheta and Mahmoud, 2012)
GEP	Ozone Level	(Samadianfard et al., 2013)
FFNN	Sea Level Variability	(Roshni, Sajid, and Samui, 2017)

CHAPTER 3: METHODOLOGY

3.1 Data Collection

The raw data was obtained for each parameter for three years in Microsoft Office Excel sheets from the General Directorate of Meteorology for Izmir Adnan Menderes Airport. The data contain ten different meteorological parameters. The parameters being used as inputs are given in Table 3.1. The remaining parameters being used as outputs, namely daily average temperature (˚C) and daily average wind speed (m/s) are given in Table 3.2.

Table 3-1 Input Variables

Inputs		Units
x_1	Daily Maximum Atmospheric Pressure	hPa
x_2	Daily Maximum Wind Direction	(°)
x_3	Daily Maximum Wind Speed	(m/s)
x_4	Daily Average Wind Direction	(°)
x_6	Daily Maximum Temperature	(°C)
x_7	Daily Average Atmospheric Pressure	(hPa)
x_8	Daily Average Cloudiness	(8 Okta)
x_9	Daily Average Relative Humidity	(%)

Table 3-2 Output Variables

Outputs		Units
x_5	Daily Average Wind Speed	(m/s)
x_{10}	Daily Average Temperature	(°C)

There were several data points which were not been able to recorded for certain reasons, so those were filled with one of the conventional methods; simple arithmetic average. Since the missing data were only for one day for the inputs, and that it was not frequent to affect the consistency of the data, simple arithmetic average has been used (Yozgatligil et al., 2013). Since more than 60% of the data for daily total rainfall ($mm=kg\div m^2$), an input provided by General Directorate of Meteorology was missing,

it had to be removed from the study. In fact, one of the possible forecast targets was the daily total rainfall for this thesis since it is known that rainfall has a high effect on aviation especially on decreasing the visibility. In a study conducted by Yihua Cao, Zhenlong Wu and Zhengyu Xu in 2014, effects of rain for the aircraft performance has been discussed. As a result of their study they have listed possible negative effects of rain to the aircraft and takeoff performance. These are; decreased visibility, poor accuracy of measurement instruments on an aircraft, possible engine flameout because of the standing water on the runway splashing from wheels to undercarriage, water vapor condensation cloud occurrence in low-pressure regions etc (Cao, Wu, and Xu, 2014). These effects are not still extensively matched with the engineering concepts, but the risk of possible effects is too high to take (Cao, Wu, and Xu, 2014). As a result of the above reasons, in case of obtaining usable data, this study can be conducted for the total daily rainfall prediction instead of daily average temperature and daily average wind speed.

After filling the missing input data points and organizing the data in the order of the years 2015 to 2017, the data was ready to be analyzed.

Figure 3.1 and Figure 3.2 show the daily average temperature and daily average wind speed patterns for the years 2015-2017, respectively.

The correlation coefficients among the input and output variables are given in Table 3.3. Highest correlation is between the input variable daily maximum temperature and daily average temperature which was expected before the analysis since they have similar time series pattern. It can be said that there is no correlation between the input variable daily maximum wind direction and the daily average wind speed. This supported the subset selection step conducted before the actual method execution of all possible regression methods.

As it can be observed from the Figures 3.1 and 3.2, both daily average temperature and wind speed data have the time series pattern, but no time series regression method have been applied in this study. The reason behind this is the nonlinear relationship between the input and output variables. Appendix C includes the relationship between the input variables and the target variable daily average temperature and only daily maximum temperature have a linear relationship with the output variables. This fact shaped the methodology of this study in the aspect of not

using time series regression methods such as ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal Autoregressive Integrated Moving Average), etc. Methods known with their superior performances on nonlinear data have been chosen.

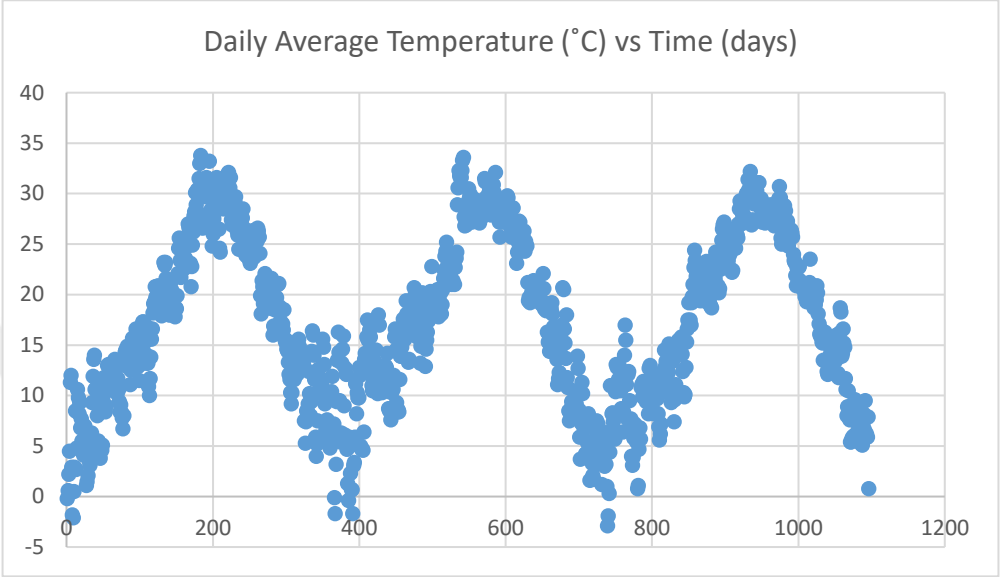


Figure 3-1 Time Series Graph for Daily Average Temperature

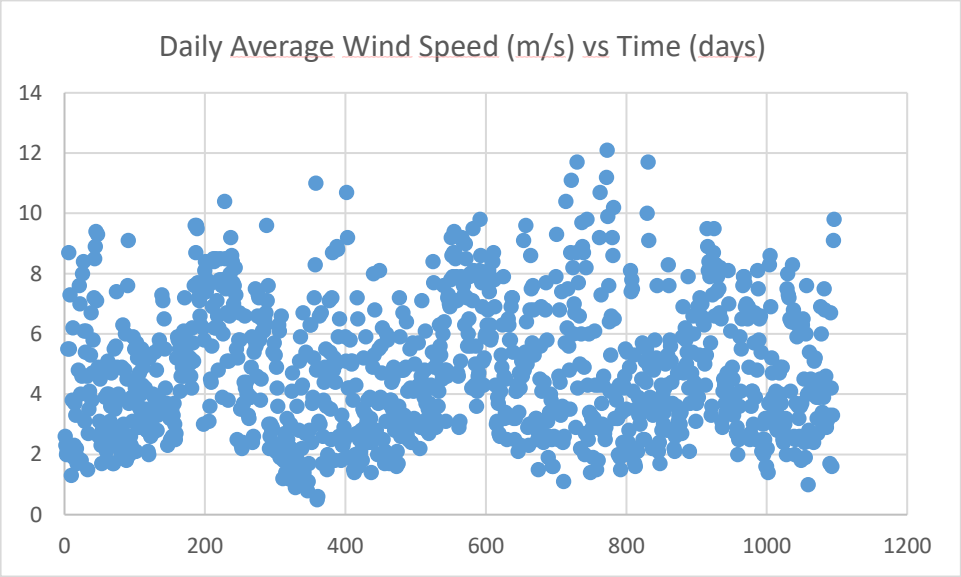


Figure 3-2 Time Series Graph for Daily Average Wind Speed

Table 3-3 Correlation Coefficients Among the Input and Output Variables

Attributes	x_7	x_8	x_9	x_4	x_1	x_6	x_2	x_3	x_5	x_{10}
x_7	1	-0,14	0,18	-0,01	0,98	-0,55	-0,08	-0,28	-0,16	-0,63
x_8	-0,14	1	0,55	0,00	-0,07	-0,41	0,05	0,08	-0,13	-0,32
x_9	0,18	0,55	1	0,00	0,244	-0,62	-0,02	-0,31	-0,51	-0,62
x_4	-0,01	0,00	0,00	1	-0,02	0,00	0,03	0,03	0,01	0,00
x_1	0,98	-0,07	0,24	-0,02	1	-0,63	-0,08	-0,25	-0,17	-0,70
x_6	-0,55	-0,41	-0,62	0,00	-0,63	1	0,034	0,074	0,11	0,98
x_2	-0,08	0,047	-0,02	0,03	-0,08	0,03	1	-0,01	0	0,04
x_3	-0,28	0,08	-0,31	0,03	-0,25	0,07	-0,01	1	0,76	0,18
x_5	-0,16	-0,13	-0,51	0,01	-0,17	0,11	0	0,76	1	0,23
x_{10}	-0,63	-0,32	-0,62	-0,00	-0,7	0,98	0,04	0,18	0,23	1

3.2 Multiple Regression Analysis

For the forecasting method, Multiple Linear Regression has been used as the first method for the study since the forecasting system focuses on the impact of 9 inputs on the output variable. There are many regression method alternatives provided by different softwares. In this study, the regression methods were applied provided by the Regression Learner Toolbox of MATLAB R2017a.

3.2.1. Multiple Linear Regression

Linear regression, known as the simplest method, models relationship between the input and the output variables. In the presence of multiple input variables used for the prediction of the output variables, the method is called as Multiple Linear Regression (MLR). This may not be the case for many data set just as our data set used in this thesis. We still applied linear regression to prove that the data set does not contain linear relationship between input and output variables. The first step in this study was to detect the linear relation between the variables and the responses. The study and the methodology were to evolve around the characteristics of the data. As it can be observed from the figures in Appendix Figure A-1 to C-7, all input data have a nonlinear relationship with one of the targets, daily average temperature except the daily maximum temperature. This was conducted to see the characteristics of the data.

This step helps to the analysis of the impact of each variable on the response and therefore to the construction of the best subset for the rest of the study. The correlation between the variables have also been studied and as a result of the correlation matrix in the Table 3.3 and the Variance Inflation Factor (VIF) being greater than 5 and 10 for some of the variables it has been decided that the data didn't include the multicollinearity. The determination for the use of the nonlinear regression tools have been done with the linear regression assumptions; linearity, no or little multicollinearity, no auto-correlation, homoscedasticity and multivariate normality.

After the application of linear regression methods, it has been clear that the meteorological data is nonlinear and that linear regression methods are not suitable for predicting temperature and wind speed. This led the study to search for nonlinear regression and machine learning methods.

3.2.2. Subset Selection

In order to determine the best subset, Minitab® 17.3.1 Best Subsets tool has been used. According to the results for the prediction of daily average temperature, x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9 were suggested as the best subset but since both Stepwise Selection and Backward Elimination methods neglected x_4, daily average wind direction, x_4 has been removed from the data set for the rest of the study. The results of the Stepwise Selection, Backward Elimination and Forward Selection are provided in the Appendix Table B-1 and B-2 for Daily Average Temperature and Daily Average Wind Speed, respectively. The result table includes the S, R-squared and Mallow's C_P measures. The final decision for the subset selection was made considering these performance measures. As a result, the final subset of input parameters has been chosen as x_1, x_3, x_5, x_6, x_7, x_8, and x_9 for the prediction of daily average temperature.

For the prediction of daily average wind speed, the best subset feature selection has also been applied. As a result, although having all of the inputs for backward elimination, stepwise and forward selection method, both daily average and maximum wind direction has been removed from the set of best subsets with the guidance of Best Subsets tool. For the rest of this thesis, the analysis is conducted for the all data including the daily average and maximum wind direction, and the best subset

separately, and the performance evaluation will be provided for both cases for comparison for daily average temperature and daily average wind speed.

3.2.3. Gaussian Process Regression

Gaussian Process Regression (GPR) is a non-parametric approach which uses kernel-based probability models, making the method itself different from other models. The non-parametric property of GPR removes the limitation of fitting over a specific function and to have different probability distributions for all possible functions. The model doesn't provide specific values for each parameter, it only computes the posterior probabilities using the trained data and the prior distribution $p(w)$ with parameter w (Heimann et al., 2018). The aim of this system is to find the f^*, prediction distribution with the provided test data x^* assuming that both prior and likelihoods are following a Gaussian distribution as it can be seen from Equation 3.2.3.3 (Heimann et al., 2018).

$$p(w|y,X) = \frac{p(y|X,w)p(w)}{p(y|X)} \tag{3.2.3.1}$$

$$posterior = \frac{likelihood \times prior}{marginal\ likelihood} \tag{3.2.3.2}$$

$$p(f^*|x^*,y,X) = \int_w p(f^*|x^*,w)p(w|y,X)dw \tag{3.2.3.3}$$

The first step is to create the prior probabilities, with the mean and kernel (covariance) functions in Equation 3.2.3.4.

$$f(x) \sim GP(m(x), k(x,x')) \tag{3.2.3.4}$$

With the existence of prior with the Gaussian distribution we have the sufficient knowledge about the space of functions (Heimann et al., 2018). The noise factor in this model is shown as follows;

$$\varepsilon \sim N(0,\sigma^2) \tag{3.2.3.5}$$

$$f(x) \sim GP\big(m(x), k(x,x')\big) + \delta_{ij}\sigma_n^2 \tag{3.2.3.6}$$

In order to add the training and testing data to the model, there should be the covariance matrices constructed in Equation 3.2.3.7 (Heimann et al., 2018). The model in Equation 3.2.3.7, containing the test data, can be used for prediction.

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{3.2.3.7}$$

The notation K is used for the covariance kernel functions and for GPR, there are several types of kernel functions. For the GPR, the kernels are used to express the situation of similar inputs variables x_i having similar target values y_i. Hence, these different types of kernels have been used in this thesis for comparison of the performances of different methods. The most popular kernels used for GPR are; constant, linear, squared exponential and rational quadratic each having different parameters (Heimann et al., 2018). Equations 3.2.3.8 -3.2.3.10 show the kernel functions of rational quadratic, exponential and squared exponential kernels, respectively.

$$k(x_i, x_j|\theta) = \sigma_f^2 \left(1 + \frac{r^2}{2a\sigma_l^2}\right)^{-a} \tag{3.2.3.8}$$

where σ_l is the length scale and a is the positive-valued scale-mixture parameter and where $r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$ is the Euclidean distance between x_i and x_j.

$$k(x_i, x_j|\theta) = \sigma_f^2 \exp\left(-\frac{r}{\sigma_l}\right) \tag{3.2.3.9}$$

where $r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$ is the Euclidean distance between x_i and x_j.

$$k(x_i, x_j|\theta) = \sigma_f^2 \exp\left[-\frac{1}{2}\frac{(x_i - x_j)^T (x_i - x_j)}{\sigma_l^2}\right] \tag{3.2.3.10}$$

where σ_f is the signal standard deviation.

The MATLAB Regression Learner tool optimizes the hyper parameters of the kernel functions as an option and the hyper parameter optimization tool have been used in this study. Table 3.4 and 3.5 show the model training results for GPR with 4 different kernels for two different targets. According to the results, GPR with the Exponential kernel outperformed compared with the other three kernels for daily average temperature whereas, Exponential kernel had the best performance in training for daily average wind speed.

Table 3-4 GPR Training Results for Daily Average Temperature

Method	Trained RMSE		Trained R-Sq		Trained MSE		Trained MAE	
	All	Best	All	Best	All	Best	All	Best
GPR- Exponential	0,005	0,006	1	1	0	0	0,004	0,005
GPR- Rational Quadratic	1,04	0,79	0,98	0,86	1,07	0,63	0,80	0,60
GPR-Squared Exponential	1,04	1,04	0,98	0,98	1,07	1,08	0,80	0,80

Table 3-5 GPR Training Results for Daily Average Wind Speed

Method	Trained RMSE		Trained R-Sq		Trained MSE		Trained MAE	
	All	Best	All	Best	All	Best	All	Best
GPR- Exponential	0,20	0,26	0,99	0,99	0,04	0,07	0,15	0,20
GPR- Rational Quadratic	0,89	0,79	0,83	0,86	0,78	0,63	0,67	0,60
GPR-Squared Exponential	0,98	0,83	0,79	0,85	0,95	0,69	0,74	0,62

3.2.4. Regression Trees

Decision Trees are used both in the fields of classification and regression. When the dataset has nonlinear property, it is very hard to compute a function that fits for all of the dataset and it is recommended to partition the dataset into subgroups (Fox,

2012). The method uses the trees to express the recursive patterns. The terminal nodes or the leaves of each tree represent the specific cell in a partition (Fox, 2012). The structure starts with a root node and it assigns the values to leaves as it goes through questions on the inner nodes and the questions are determined according to the answers given to the previous questions (Fox, 2012). This part is the recursive pattern and after finding these patterns, one should try to understand the logic of simple local models.

The simple local model is actually the sample mean of dependent variables predicted from the constant estimation of samples of Y (Fox, 2012). After the reaching a certain stage, the model has to stop creating new nodes, or assigning values for leaves. That certain stage is called the Information Gain (IG). The aim of the information gain is to get the most informative features by splitting the nodes, in other words the aim of the regression trees is to maximize the Information Gain (IG) at each split (Li, 2019). The following Equation 3.2.3.1 shows the IG for binary decision trees;

$$IG(D_p, f) = I(D_p) - \left(\frac{N_{left}}{N_p} I(D_{left}) + \frac{N_{right}}{N_p} I(D_{right}) \right) \quad (3.2.3.1)$$

Here, the f is the feature of the specific split whereas D_p, D_{left}, D_{right} are the dataset of the parent and the child nodes. I is the impurity measure, N_p is the total number of samples in the parent node, and similarly N_{left} and N_{right} are the total number of samples in the child nodes denoted as left and right (Li, 2019). As it can be observed from the above equation, the IG is the difference between the impurity of the parent node and the child nodes showing that as the impurity of the child nodes decrease, the information gain increases (Li, 2019).

Advantages of this tree-like structure enables simple calculations and distinguishing the important variables for the prediction. While using this method, the maximum depth should be selected very carefully since it can lead to the overfitting of the model. The most important part about the construction of these decision trees is to determine the optimal maximum depth. Tables 3.6 and 3.7 show the training result of both output variables and fine tree had the lowest RMSE values for both all and best subset data.

Table 3-6 Regression Trees Training Results for Daily Average Temperature

Method	Trained RMSE		Trained R-Sq		Trained MSE		Trained MAE	
	All	Best	All	Best	All	Best	All	Best
Tree-Medium Tree	1,23	1,23	0,98	0,98	1,53	1,53	0,94	0,94
Tree-Fine Tree	0,82	0,84	0,99	0,99	0,68	0,70	0,63	0,65
Ensemble-Boosted Trees	1,30	1,30	0,98	0,98	1,68	1,69	1,07	1,08
Ensemble-Bagged Trees	1,30	1,25	0,98	0,98	1,68	1,56	0,96	0,89
Tree-Coarse Tree	1,79	1,79	0,95	0,95	3,22	3,22	1,42	1,42

Table 3-7 Regression Trees Training Results for Daily Average Wind Speed

Method	Trained RMSE		Trained R-Sq		Trained MSE		Trained MAE	
	All	Best	All	Best	All	Best	All	Best
Tree-Medium Tree	0,93	0,93	0,81	0,81	0,86	0,86	0,70	0,70
Tree-Fine Tree	0,69	0,71	0,90	0,89	0,48	0,50	0,50	0,52
Ensemble-Boosted Trees	0,89	0,88	0,83	0,83	0,78	0,78	0,68	0,68
Ensemble-Bagged Trees	0,82	0,82	0,86	0,86	0,66	0,67	0,62	0,62
Tree-Coarse Tree	1,16	1,16	0,71	0,71	1,34	1,34	0,88	0,88

3.3 Support Vector Regression

Support Vector Algorithm is a nonlinear generalization algorithm which has been developed by Vapnik and Chervonenkis in 1963 and improved until present day (Smola and Schölkopf, 2004). The motivation behind this algorithm was first the classification but with the developments and improvements, it has also been used for regression and time series forecasting. Support Vector Machine (SVM) was widely used and continued to be improved with the work of Vladimir N. Vapnik and his team especially on their work on optical character recognition (OCR), a real industrial subject and many other areas due to its success on recognition tasks (Smola and Schölkopf, 2004).

Support Vector Regression (SVR) on the other hand was first introduced by Alex J. Smola and Bernhard Schölkopf (2004). The way it differs from SVM is that it

aims to find the optimal hyperplane that fits for all the training set instead of a classification problem. Allowing certain amount of errors, SVR tries to approximate the target using the provided training data. Since the algorithm uses training data, this method can be classified as a supervised learning method.

The aim for this method is to find the function $f(x)$ which leads to obtaining at most ε deviation from the target values y_i for the training data set. For training data $\{(x_1, y_1), \dots, (x_\ell, y_\ell)\} \subset X \times \mathbb{R}$ where X denotes the input patterns which are meteorological parameters for this study (Smola and Schölkopf, 2004).

Besides finding the function $f(x)$ that has at most ε deviation, the desired $f(x)$ is expected to be as flat as possible. The Flatness in the case of Equation 3.3.1 is provided with smallest w possible.

$$f(x) = \langle \mathrm{w}, \mathrm{x} \rangle + \mathrm{b} \; with \; \mathrm{w} \in \mathrm{X}, \mathrm{b} \in \mathbb{R} \tag{3.3.1}$$

In order to find the smallest w, an optimization problem in Equation 3.3.2 is constructed. The optimization problem in Equation 3.3.2 assumes that this optimization problem is feasible, in other words that function f approximates all pairs of $(x_\mathrm{i}, y_\mathrm{i})$ with ε precision (Smola and Schölkopf, 2004). Considering the fact that there can be errors, the concept of soft margins is brought to the SVR just like the works of Vapnik and Cortes in 1992 by adding the slack variables ξ_i, ξ_i^* (Smola and Schölkopf, 2004) to solve the infeasibility problem in Equation 3.3.2.

$$minimize \; \frac{1}{2}\|w\|^2$$
$$subject \; to \; \begin{cases} y_i - \langle w | x_i \rangle - b \leq \varepsilon \\ \langle w | x_i \rangle + b - y_i \leq \varepsilon \end{cases} \tag{3.3.2}$$

The optimization problem in Equation 3.3.3 contains the constant $C > 0$ representing the trade-off between the flatness of f and the ε being larger than tolerated amount (Smola and Schölkopf, 2004).

$$minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_{i+}\ \xi_i^*)$$

$$subject\ to \begin{cases} y_i - \langle w|x_i\rangle - b \le \varepsilon + \xi_i \\ \langle w|x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \quad \ge 0 \end{cases} \tag{3.3.3}$$

The trade-off leads to creation of an ε–insensitive loss function $|\xi|_\varepsilon$ shown in Equation 3.3.4.

$$|\xi|_\varepsilon \coloneqq \begin{cases} 0 & if\ |\xi| \le \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{3.3.4}$$

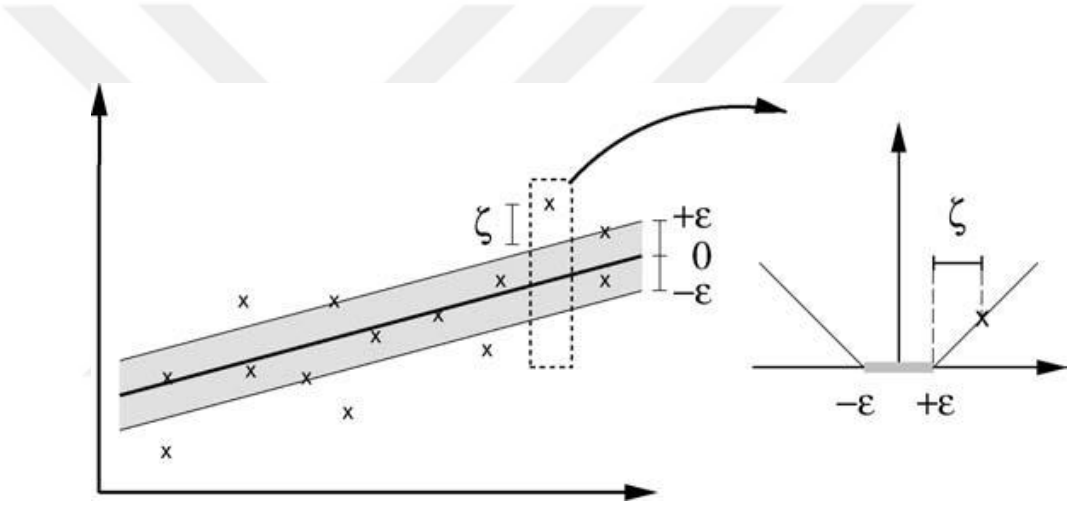


Figure 3-3 The soft margin loss for linear SVR (Source: Hirani and Mishra, 2016).

Figure 3.3 clearly shows the impact of having the points outside the tolerated shaded area, adding as the cost representing the trade-off.

The optimization problem Equation 3.3.3 appeared to be solved in the dual from more easily with the Lagrange function method (Smola and Schölkopf, 2004). L being the Lagrangian and $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ being the Lagrange multipliers, Equation 3.3.5 depicts the dual problem for Equation 3.3.3.

$$L := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) - \sum_{i=1}^{\ell}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{\ell}\alpha_i(\varepsilon + \xi_i - y_i + \langle w|x_i\rangle + b) - \sum_{i=1}^{\ell}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w|x_i\rangle - b) \quad (3.3.5)$$

The non-negativity constraints for the Lagrange multipliers are shown in Equation 3.3.6, $\eta_i^{(*)}$ referring η_i and η_i^*.

$$\alpha_i^{(*)}, \eta_i^{(*)} \geq 0 \quad (3.3.6)$$

The partial derivate for Lagrange with respect to the primal variables (w, b, ξ_i, ξ_i^*) are shown in Equations 3.3.7-3.3.9 (Smola and Schölkopf, 2004).

$$\partial_b L = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i) = 0 \quad (3.3.7)$$

$$\partial_\omega L = w - \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)x_i = 0 \quad (3.3.8)$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \quad (3.3.9)$$

As a result of substituting the Equations 3.3.7-3.3.9 to Equation 3.3.5, the optimization problem in Equation 3.3.10 has been constructed (Smola and Schölkopf, 2004).

$$maximize \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i|x_j\rangle \\ -\varepsilon\sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{\ell}y_i(\alpha_i - \alpha_i^*) \end{cases}$$
$$subject\ to \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C] \quad (3.3.10)$$

Through the construction of the optimization problem in Equation 3.3.10, the slack variables have been removed because of the new formulation of (3.3.9) being $\eta_i^{(*)} = C - \alpha_i^{(*)}$.

As a result of substitutions, Equation 3.3.8 is transformed into the following form;

$$w = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i \,, \text{thus } f(x) = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \langle x_i | x \rangle + b \tag{3.3.11}$$

The process of creating a linear combination of the training data x_i is an example of Support Vector Expansion concept (Smola and Schölkopf, 2004) and the aim is to make the complexity of a function depend on the number of Support Vector's instead of the dimensionality of the input space X.

The b variable which has been used, hasn't been discussed up to this point. Karush–Kuhn–Tucker (KKT) conditions are used to compute the b (Smola and Schölkopf, 2004). According to these conditions, the products between dual variables and constraints are not included.

$$\begin{gathered} \alpha_i(\varepsilon + \xi_i - y_i + \langle w | x_i \rangle + b) = 0 \\ \alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w | x_i \rangle - b = 0 \\ (C - \alpha_i)\, \xi_i = 0 \end{gathered} \tag{3.3.12}$$

$$(C - \alpha_i^*)\xi_i^* = 0 \tag{3.3.13}$$

The implementation of SVR has been made on MATLAB R2017a, Regression Learner Toolbox. The data for the years 2015 and 2016 were used to train the regression models, where 2017 data was used to test the predicted values which was generated using the trained SVR model. Table 3.8 and 3.9 show the training model RMSE, R-squared, MSE, and MAE values for both all data and best subset for six SVR models with different kernels for daily average temperature and wind speed, respectively. As it can be observed from the tables, SVR with Cubic and Medium Gaussian kernels had the best training performances with the RMSE 0.98 for the prediction of daily average temperature for the best subset. The performance of the daily average wind speed forecast is also the same with the temperature since it also has the Cubic and the Medium Gaussian kernels with lower RMSE 0.88 and 0.85, respectively.

The trained models were exported to the MATLAB code to be used in the prediction with the input data of 2017 provided. All target values for 365 days have been constructed with the code and they were used in order to detect the difference with the actual 2017 target values and using the differences the RMSE for test data

have been obtained. Detailed comparison and discussion on test data is provided in Chapter 4, Results.

Table 3-8 SVR Trained Model Analysis for Daily Average Temperature

Method	Trained RMSE		Trained R-Squared		Trained MSE		Trained MAE	
	All	Best	All	Best	All	Best	All	Best
Quadratic SVR	1,07	1,09	0,98	0,90	1,14	1,18	0,84	0,85
Cubic SVR	0,95	0,98	0,99	0,99	0,91	0,96	0,78	0,78
Linear SVR	1,21	1,21	0,98	0,98	1,46	1,47	0,94	0,95
Coarse Gaussian SVR	1,14	1,13	0,98	0,98	1,29	1,28	0,89	0,89
Medium Gaussian SVR	0,94	0,98	0,99	0,99	0,88	0,95	0,78	0,80
Fine Gaussian SVR	1,49	1,22	0,97	0,98	2,21	1,48	1,16	0,99

Table 3-9 SVR Trained Model Analysis for Daily Average Wind Speed

Method	Trained RMSE		Trained R-Squared		Trained MSE		Trained MAE	
	All	Best	All	Best	All	Best	All	Best
Quadratic SVR	0,95	0,95	0,80	0,81	0,91	0,89	0,70	0,69
Cubic SVR	0,87	0,88	0,84	0,83	0,75	0,77	0,62	0,64
Linear SVR	1,02	1,01	0,78	0,78	1,04	1,01	0,76	0,75
Coarse Gaussian SVR	1,09	1,08	0,74	0,75	1,19	1,17	0,82	0,82
Medium Gaussian SVR	0,85	0,85	0,84	0,84	0,73	0,72	0,61	0,62
Fine Gaussian SVR	0,53	0,49	0,94	0,95	0,28	0,24	0,33	0,32

3.4 Artificial Neural Networks

Artificial Neural Network (ANN) is an extensively used method since 1950's but the ANN application on atmosphere science was first introduced by Rumelhart and McClelland in 1986 (Fabbian, De Dear, and Lellyett, 2007). The first neural network was simple perceptron used in linear models. After the need of a model for nonlinear cases, multi-layer perceptron has been proposed (Fabbian, De Dear, and Lellyett, 2007). If the training procedure is applied correct, ANN provides the link between the

input and the output variables. There are many studies in the literature about weather forecasting especially with the ANN method and some hybrid methods developed with different studies. Specifically, ANN has been used in tornado detection, predicting precipitation and temperature over the years (Fabbian, De Dear, and Lellyett, 2007). The reason behind the popularity of ANN is its capability to solve complex problems where the knowledge in advance is not required. Besides its popularity, the reason why ANN makes a good fit as a forecasting method is existence of the hidden layer enabling the ANN architecture to include nonlinear features of the system and with that the performance of the prediction gets better (Fabbian, De Dear, and Lellyett, 2007).

Many studies conducted in the rainfall prediction show that Backpropagation Neural Network (BPNN) is working better compared to the other methods (Mislan et al., 2015). Backpropagation is an iterated search algorithm adjusting the form of the output layer back to the input layer for every run up to the point where no longer improvements can be achieved (Golafshani, Rahai, and Sebt, 2015). In cases of data not following a certain pattern, BPNN would face difficulties on having good performance because of flawed networks. Fabbian and De Dear had this problem in their work in 2006, so as a solution they have tried to adjust the ratio of fog events to no-fog events but this did not give a better training system for the ANN (Fabbian, De Dear, and Lellyett, 2007).

A simple ANN architecture has three main components. These are; input, output and hidden layers. Input layer provides the data to the system and enables its flow through the other layers, the first hidden layer for the learning procedure. Output layer is the final point which includes the values that assesses the network's learning capability. Lastly, hidden layers are the layers which adjusts and transforms the input into something output unit can use with the support of activation functions (Demirhan and Kayhan Atilgan, 2015). The emergence of hidden layers has improved the accuracy performance of MLP for nonlinear models (Fabbian, De Dear, and Lellyett, 2007). On these previously defined layers, there are different number of neurons. These neurons are the training and testing data provided in the beginning of the procedure. The neurons operate and transmit the information between the layers until it leaves the network from the output layer.

Also, there are weights between the layers, representing the effects of previous layer on any layer element and these weights determine the information transfer

between neurons (Khatib, Mohamed, and Mahmoud, 2012). Lastly there is the bias factor between the neurons in hidden and output layers. Figure 3.4 shows the network structure including the bias and the weights in system.

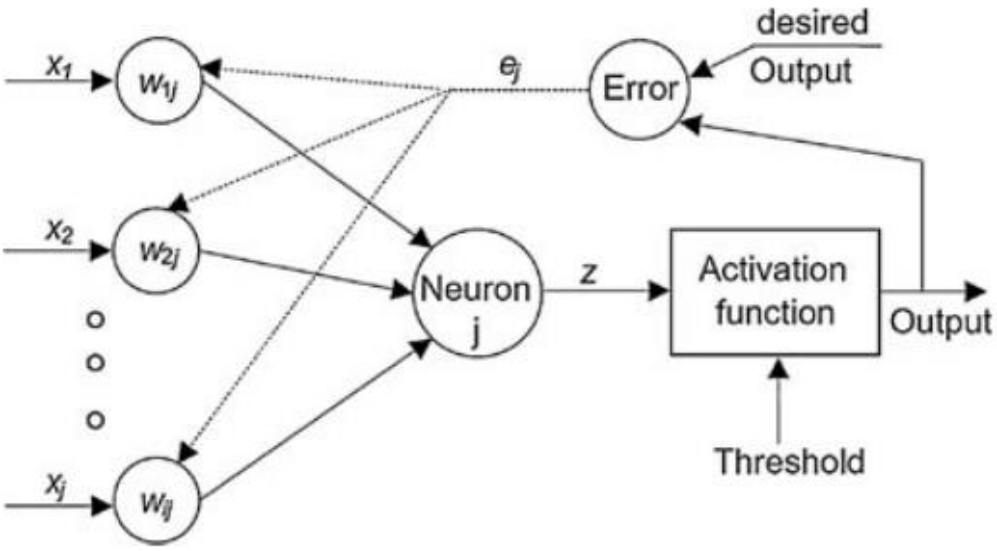


Figure 3-4 Network Structure (Source: Khatib, Mohamed, and Mahmoud, 2012)

For the hidden and output layers there are similar processes. The first one is computation of the sum of inputs that is received by a neuron. This sum is named as the net input and Equation 3.4.1 depicts the calculation.

$$net_k = \sum_{i=1}^{n} w_{ik}.x_i + bias_k \qquad (3.4.1)$$

In the Equation 3.4.1, net_k is the weighted sum of the k^{th} neuron, w_{ik} is the weight between the i^{th} and k^{th} neurons and lastly x_i is the output of i^{th} neuron in the preceding layer (Golafshani, Rahai, and Sebt, 2015).

The next step is to calculate the output of the k^{th} neuron which is named as out_k in the Equation 3.4.2 below using the sigmoid function (Golafshani, Rahai, and Sebt, 2015).

$$out_k = f(net_k) = \frac{1}{1 + e^{-(net_k)}} \qquad (3.4.1)$$

Most common type of ANN networks is the Feed Forward Neural Networks (FFNN), where the information flows through single direction forward (Thota, 2018). Multilayer feedforward neural network uses gradient descent operators such as backpropagation where the simple logic is to start from the most general possible solution to the most specific by increasing the threshold unit (Banzhaf et al., 1998).

After entering the data to the ANN system, it tries to create a relationship between the inputs and the target variables in order to come up with a model working with the new data to be used for prediction. The structures of ANN depend on the following aspects;

- Method used for training
- Number of hidden layers
- Learning algorithm
- Type of error function
- Direction of information flow
- Number of neurons in layers
- Activation function

Number of neurons should be sufficiently low to ensure successful generalization. The determination of these aspects can be based on experimentation and experience of the user (Golafshani, Rahai, and Sebt, 2015). The user should consider these before constructing the network structure.

There are number of different learning algorithms used to train the network. We have used Levenberg-Marquardt Backpropagation (LM), Resilient Backpropagation (RP), Bayesian Regularization Backpropagation (BR), Fletcher-Powell Conjugate Gradient Backpropagation (CGF), Scaled Conjugate Gradient Backpropagation (SCG), BFGS Quasi-Newton Backpropagation (BFG) learning algorithms for this thesis. LM is the most common learning algorithm known for its ability to process the large data sets. LM is known for its robustness but as a disadvantage it needs memory.

3.4.1. Application of ANN

The application of ANN on the weather data set was done on MATLAB from an source code named as Neural Network Training Code. Six different learning algorithms have been tried, keeping the rest of the parameters same as the default settings. The comparison of these learning algorithms is provided in Section 4 on Table 4.1. Starting with the construction of the network structure we have tried using 1, 2, 3 and 5 number of hidden layers. The number of neurons in hidden layer has been set to 10.

All of the learning algorithms ended with lower RMSE with number of hidden layers set to 1, so the rest of the analysis have continued with 1 hidden layer. Results show that for prediction of daily average temperature Bayesian regularization backpropagation and for daily average wind speed LM backpropagation learning algorithm showed the best performance.

The Figure 3.5 shows the MSE epoch graph for the training, validation and testing stages. The graph indicates that the best training performance is achieved at the 8th epoch. Figure 3.6 shows the error histogram for three stages of BPNN with LM learning algorithm. Models constructed under ANN appeared to be successful for all three stages.

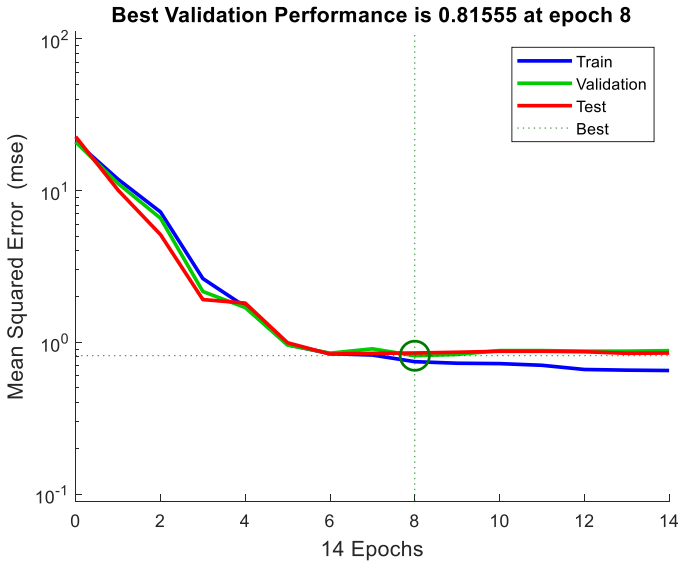


Figure 3-5 Error-Epoch Graph for Daily Average Wind Speed (Best Subset-LM)

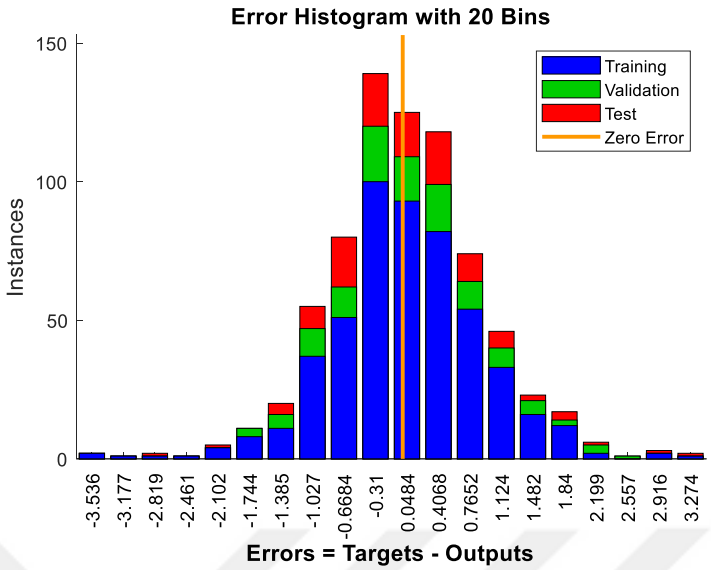


Figure 3-6 Error Histogram for Daily Average Wind Speed (Best Subset- LM)

3.5 Multi-Gene Genetic Programming

Genetic Programming (GP) is a biologically inspired Machine Learning method which is one of the most popular and successful branches of Evolutionary Algorithms (Morrison, Searson, and Willis, 2010). GP uses the Darwinian Theory where each computer program is represented with tree structures also known as genes with varying lengths searching the candidate solutions by using the natural selection and the evolution logic (Orove, Osegi, and Eke, 2015). GP can be applied for the following purposes which makes the model itself very popular compared to the other methods used in areas mentioned;

- Classification
- Regression
- Clustering
- Problem Solving
- Capturing Solutions for any type of problem (learning, optimization, game playing, etc.)

The reason why it is popular is, unlike the other regression method, GP handles both the tree structure and the regression parameters (Searson, 2009). The researchers

or users who don't have any background information on both the dataset and the algorithm can easily analyze and interpret the model output. Without the need of predefining the system structure or the estimates of the regression coefficients, GP provides the relationship between the dependent and the independent variables (Elhenawy, Chen, and Rakha, 2014). Another important feature of GP which leads the users to prefer over ANN and SVR especially in this field of study is that ANN and SVR has very long training processes (Elhenawy, Chen, and Rakha, 2014). GP has the following features which makes it effective; heuristic nature of search, symbolic program representation, input sensitivity, inductive nature, being comprehensive and the allowance of unconstrained data types (Krawiec, 2010). GP representation of any problem is a superset of other machine learning representation since it can include Boolean operators, threshold functions that are used in ANN, conditional branching structures and case-based structures like K-nearest neighbor systems (Banzhaf et al., 1998). Besides the capability of including different representations of concepts, GP is also superior because of not having the fixed size programs to evolve which actually limits the performance of the machine learning techniques (Banzhaf et al., 1998).

As in the other machine learning processes, GP also has a learning and a testing procedure where it has the learning domain including set of features as inputs and the anticipated classes as results (Banzhaf et al., 1998). The training stage of the process is basically the process of creating a computer program to predict the outputs of the training set using the inputs provided (Banzhaf et al., 1998). The inputs provided for the training stage are a part of the terminal stage which helps building the branches as a starting point. Besides the input variable there are also constants such as random ephemeral constants that are chosen randomly from the population in the beginning of the run and they do not change during the whole process.

After the creation of the training system; the system performance, the prediction success is determined by the inclusion of the test data. GP creates initial populations in the form of tree structures and evolves them by using mutations and crossovers operators for transformation until reaching the best performing population that fits the objective. This process can also be observed in the hill climbing concept where it continuously searches for the best solution but GP is actually a type of beam search where it first searches for the most promising solutions (population) and then have some transformations (Banzhaf et al., 1998). This is the reason why we can say that this

method is exhaustive but successful performance-wise. Here, the use of the operators has the most important role since for example crossover makes it sure that the best combination of exchanged portions of parents is built to create better solutions.

3.5.1. MGGP Process

MGGP is seeking to minimize the mean squares error of the fitted data set by evolving multiple solution just like the GP logic (Orove, Osegi, and Eke, 2015). The feature that differs MGGP from the stand-alone GP is the multi gene (tree structures) used to create the candidate solutions. This feature enables multiple lower depth trees and as a result provides simpler, easily interpretable models compared to classical GP (Danandeh and Kahya, 2017). Each tree structure contains functional sets each presented on the nodes which connect the child and the parent nodes. Depending on the complexity of a system, the functional set may even contain, sin, tan, exp besides standard mathematical operators (Danandeh and Kahya, 2017). This feature is also a very significant feature of GP, since it provides complex but accurate regression models compared to other traditional regression methods. The first step of the GP is to initialize a population with the provided training data. The individuals on the initial population are the G_{max} number of randomly generated trees, where the number is between 1 to the maximum number predefined by the user (Faris and Sheta, 2016). This step is followed by assessing the fitness values of each individual. The fitness function is by default the root mean squared error, it is the performance measure determining the system solution with the actual solution obtained from the training and testing data. Starting from the genes having the best solution among the others in the population, evolution of the genes is performed. Figure 3.7 shows the MGGP process diagram.

There are three different evolutionary operators. These are crossover, mutation and reproduction. The crossover operation is the interchange of the genetic material among genes on a randomly chosen crossover point to improve the fitness of an individual (Elhenawy, Chen, and Rakha, 2014). Mutation on the other hand is the replacement of a randomly selected subexpression of a gene with a randomly generated subexpression (Krawiec, 2010). In other words, changing a part of genetic material with a randomly generated subtree which may even include addition or deletion (Orove, Osegi, and Eke, 2015). Some studies like the study performed by Mehr and

Kahya proved that higher crossover fraction than mutation fraction on a model performs better, so users should take this information into consideration (Danandeh and Kahya, 2017).

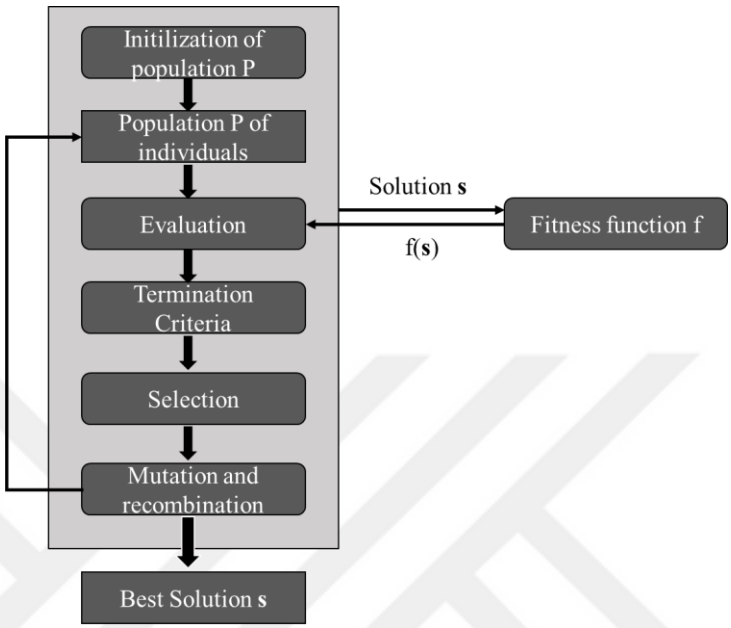


Figure 3-7 MGGP Procedure

Lastly, reproduction which is not used as much as crossover and mutation, is copying the genetic material of the older best performing individuals to the next generations (Elhenawy, Chen, and Rakha, 2014). As a result of these operations, new generations which are the candidate solutions, are obtained. The first set of solutions cannot be the best solutions since the process must undergo an evolution method until it reaches to the best solution (Orove, Osegi, and Eke, 2015). The cycle shown in the Figure 3.8 proves this rule. The cycle stops when it reaches the termination criteria which is predefined by the user which can be reaching to the ideal solution or a predefined runtime.

The models created throughout the cycle are the weighted linear combinations of each gene, where the optimal weights used in the combinations are obtained by the Least Squares (LS) method (Orove, Osegi, and Eke, 2015). This means that every prediction of the target variable is the sum of weighted value of the trees of the multigene individuals and the bias term. This mentioned relation is depicted in the Equation 3.5.1.

The trees used in the models are the functions of zero or more N number of variables. The GP procedure is depicted in the Table 3.10.

$$y = d_0 + d_1 * tree_1 + \cdots + d_m * tree_m \ldots$$
$$where\ d_0\ is\ the\ bias\ term$$
$$d_1, \ldots, d_m are\ the\ gene\ weights\ with\ m\ number\ of\ g\text{enes} \tag{3.5.1}$$

Table 3-10 GP Procedure (Source: Searson, 2015)

Basic steps describing the GP
1: procedure GP
2: begin GP
3: Generate initial population of n individuals.
4: Initialize the GP parameters.
5: Calculate the fitness of each individual.
6: while (t < Max Generation) or (stop criterion not met)
7: Select pair of individuals using Tournament Selection Mechanism.
8: Produce a new offspring using crossover, mutation and elitism.
9: Evaluate the population.
10: Replace current population with newly created one.
11: Update the generation counter.
12: end while
13: end GP

The algorithm for genetic programming has been demonstrated in the Figure 3.8 on a flowchart. MGGP applies the previously stated steps to achieve the best performing population.

The determination of the maximum depth parameter (MDP), which is the minimum depth which can be tolerated between the terminals and the root node. The size of a tree structure can be determined by 2^{MDP}.

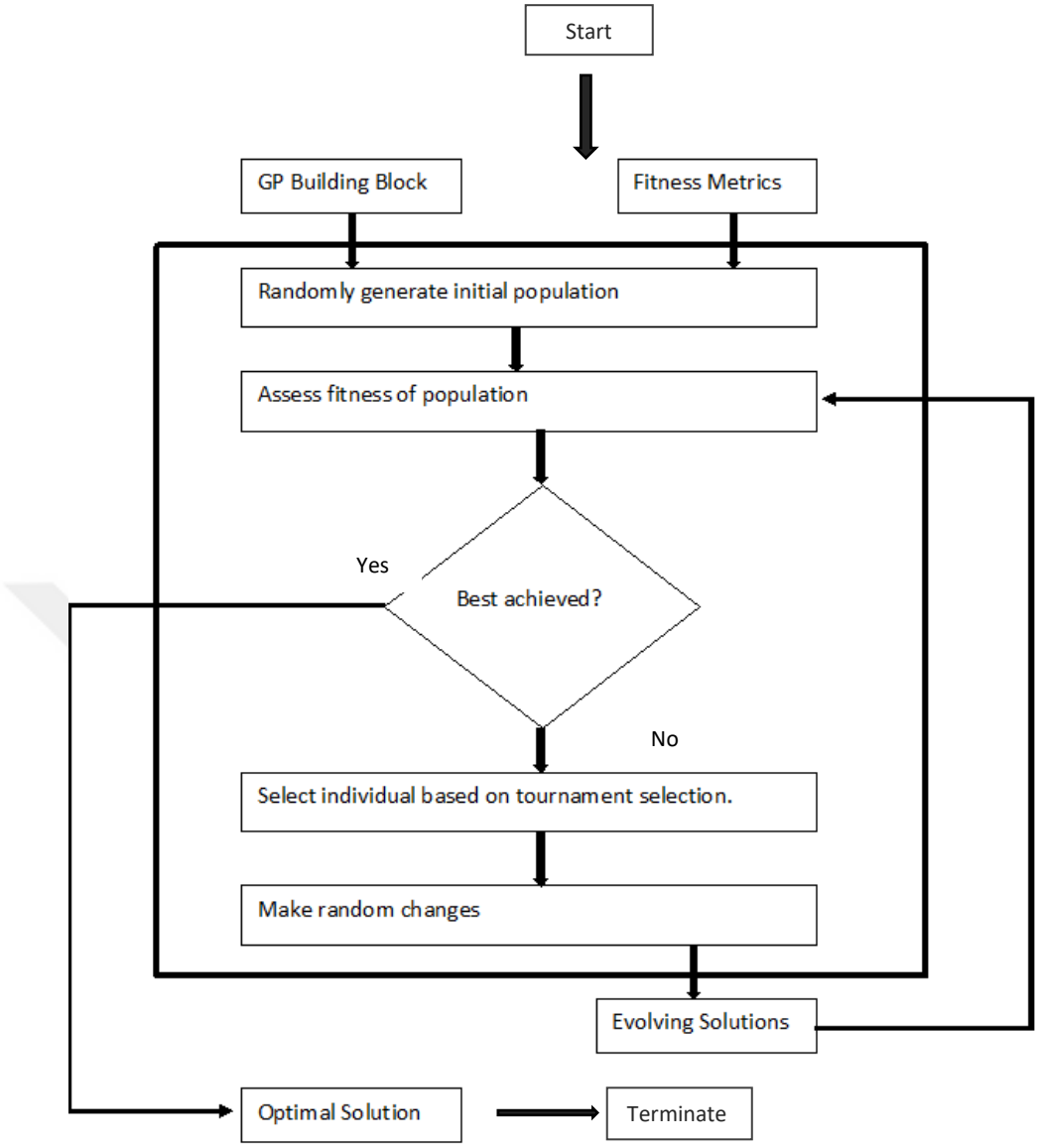


Figure 3-8 GP Process Flowchart

The initial parameters change/evolve over each iteration in order to reach the best tree structure which is much easier for a user since they don't have to think about setting the best condition initially, GP evolves the best structures and parameters automatically. Fitness in the case of determining the success of the model is reduced sum of squared errors with respect to the provided data set.

3.5.2. Application of MGGP

The analysis on MGGP has been executed on MATLAB R2017a, GPTIPS® Toolbox. The GPTIPS provide the following features (Searson, Leahy, and Willis, 2010);

- Multiple tree (multi-gene) individuals
- Tournament selection & lexicographic tournament selection
- Standard sub-tree crossover operator
- Elitism
- Early run termination criterion
- Graphical population browser showing best and non-dominated individuals (fitness & complexity).
- Graphical summary of fitness over GP run.
- 6 different mutation operators.

The analysis has been done on 3 different number of genes. The default number of genes in GPTIPS is 0, but 4, 5 and 6 are the ones which show the best performance considering both complexity and RMSE. The GP parameters used for the 5 number of genes for the best subsets of daily average temperature and daily average wind speed are shown in Table 3.11 The GP parameters except the maximum number of genes have been left in their default values. Since the default settings for GPTIPS have been used, and only number of genes have been modified for each attempt, the parameter table is common for daily average temperature and daily average wind speed.

Figure 3.9 provides the gene weights with the bias factor and the p values for the genes for the application with maximum number of genes set to 5. The toolbox provides tree structures for the best performing model and the tree structures for model of daily average wind speed with the best subset data having 5 genes is shown below on Figure 3.10. The toolbox also provides the regression function of the best performing model and it can also be observed in the Equation 3.5.2 and 3.5.3 for daily average temperature and wind speed with 5 genes, respectively.

Table 3-11 MGGP GPTIPS Parameters

Run parameter	Value
Population size	100
Max. generations	150
Tournament size	2
Elite fraction	0.05
Selection Method	Lexicographic selection pressure tournament selection
Probability of pareto tournament	0.7
Max. genes	5
Max. tree depth	4
Crossover probability	0.84
Mutation probabilities	0.14

$$y = 0.851x3 + 0.844x4 - 0.129x5 + 0.00685\tanh(x5) + 0.00685x6x7 - 1.53e^{-4}x2x3^3 - 5.87e^{-4}x3x6x7 + 123 \quad (3.5.2)$$

GP algorithms are assessed by their percentage of runs ended with success, the time needed to achieve the success and the difference between the actual and best solution generated by the GP (Krawiec, 2010). The RMSE shouldn't be the only performance evaluation metric assessing the GP, so the user should also take the complexity and runtime into consideration.

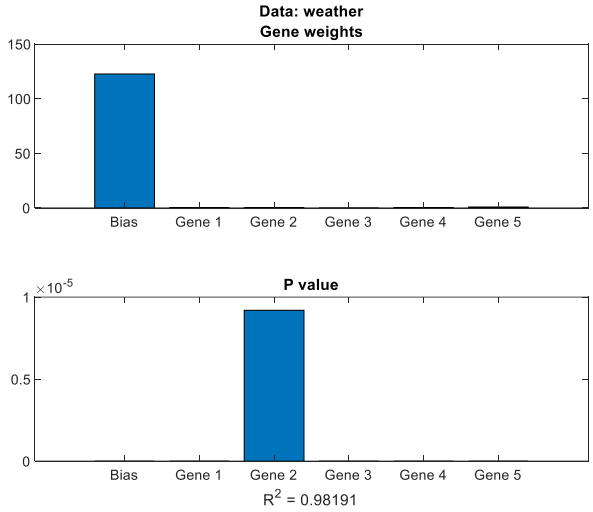


Figure 3-9 Gene Weights for 5 Gene Daily Average Temperature Prediction

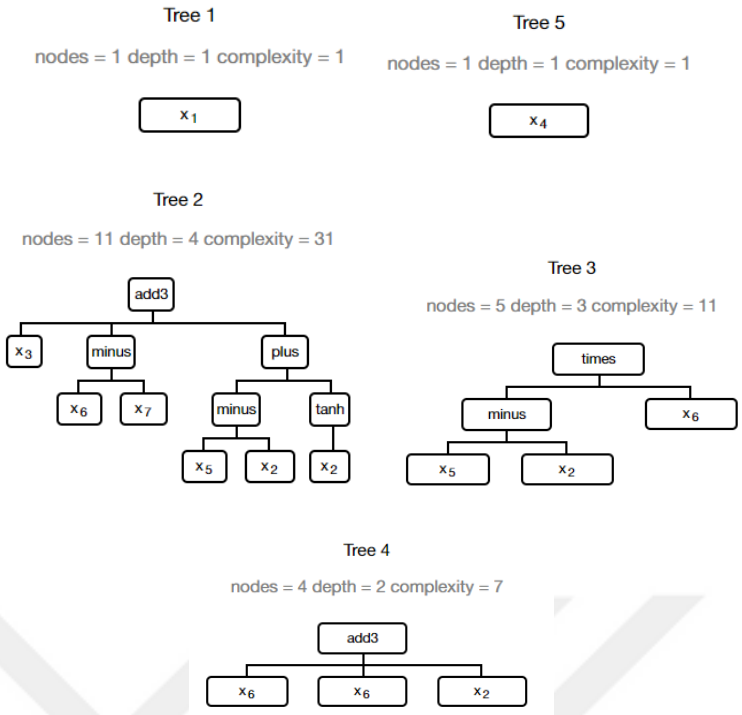


Figure 3-10 5 Genes Tree Structure for Best Subset Wind Speed

$$y = 0.563x2 - 0.0746x1 - 0.383x3 + 0.0974x4 - 0.383x5 - 0.023x6 + 0.383x7 - 0.383\tanh(x2) - 0.00322x6(x2 - 1.0x5) - 17.9 \quad (3.5.3)$$

Table 3-12 Comparison of RMSE and Complexity on different number of genes

No of Genes	RMSE				Complexity			
	Temperature		Wind Speed		Temperature		Wind Speed	
	All	Best	All	Best	All	Best	All	Best
4	1,06	1,04	1,07	1,06	50	42	36	48
5	1,04	1,02	1,03	1,03	75	69	47	51
6	1,03	1,05	1,06	1,03	64	60	94	91

CHAPTER 4: RESULTS

We can come to the inference from previous tables (Table 3.3-3.7) representing the training RMSE results for all of the multiple regression methods and Table 4.1 that the trained model had immensely better performance compared to the predicted value performance. For example, for the prediction of the daily average temperature SVR with the Fine Gaussian kernel had the 1.22 RMSE value for model training and 5.12 RMSE value for the testing procedure. We can say that the SVR with Fine Gaussian kernel wasn't able to create a successful generalization of the training data. The proceeding 4.1 and 4.2 parts include the detailed analysis of the regression results for daily average temperature and daily average wind speed, respectively.

The following Table 4.1 is the collection of RMSE result for all of the regression methods for best subset and all data for the prediction of daily average temperature and daily average wind speed.

4.1 Performance Evaluation

As the performance measure, the RMSE measurement have been used to determine the accuracy of the prediction of the regression methods. The reason of choosing the RMSE was its extensive use in the literature on prediction models for weather. The RMSE calculation is given by the Equation 4.1.1. The Y is the actual target value and Y^* is the predicted value generated by the regression methods for testing data size of n.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y - Y^*)^2} \tag{4.1.1}$$

Table 4-1 RMSE for Both Targets for All Methods

Model	RMSE			
	Temperature		Wind Speed	
	All	Best	All	Best
ANN-Bayesian regularization backpropagation	1,13	1,05	1,08	1,11
ANN-BFGS Quasi-Newton	1,28	1,13	1,08	1,07
ANN-Fletcher-Powell Conjugate Gradient	1,65	1,56	1,12	1,16
ANN-Levenberg-Marquardt backpropagation	1,34	1,27	1,13	1,03
ANN-Resilient backpropagation	2,13	1,79	1,17	1,17
ANN-Scaled conjugate gradient backpropagation	1,68	1,75	1,25	1,22
Coarse Gaussian SVR	1,19	1,19	1,21	1,19
Cubic SVR	1,03	1,03	1,15	1,11
Ensemble w/ Bagged Trees	1,80	1,80	1,29	1,27
Ensemble w/ Boosted Trees	1,47	1,47	1,29	1,28
Fine Gaussian SVR	5,12	5,12	1,94	1,75
Gaussian Process Regression- Exponential GPR	1,13	1,13	1,08	1,04
Gaussian Process Regression- Rational Quadratic GPR	1,01	1,01	0,99	0,97
Gaussian Process Regression-Squared Exponential GPR	1,01	1,01	1,04	0,97
GP-4 genes	1,06	1,04	1,07	1,06
GP-5 genes	1,04	1,02	1,03	1,03
GP-6 genes	1,03	1,05	1,06	1,03
Linear Regression	1,18	1,18	1,04	1,03
Linear Regression-Interactions Linear	1,03	1,03	1,03	1,00
Linear Regression-Robust Linear	1,17	1,17	1,03	1,00
Linear SVR	1,17	1,17	1,07	1,05
Medium Gaussian SVR	1,49	1,49	1,14	1,10
Quadratic SVR	1,01	1,01	0,99	0,96
Stepwise Linear Regression	1,14	1,07	1,00	1,00
Tree-Coarse Tree	1,85	1,85	1,47	1,47
Tree-Fine Tree	1,59	1,59	1,47	1,50
Tree-Medium Tree	1,61	1,61	1,41	1,38

4.1 Daily Average Temperature

As a result of the study, the forecasting procedure with the regression methods applied for predicting the daily average temperature had the following performance shown in the Table 4.1. Gaussian Process Regression with the kernels Rational Quadratic and Squared Exponential and the SVR with Quadratic kernel and MGGP with 5 and 6 number of genes had the best RMSE results among the regression methods.

The reason behind the outstanding training performance for the several regression methods for daily average temperature was the effect of the input variable, daily maximum temperature. The only linear relation in this analysis was between the daily average and maximum temperature which also can be observed in the Appendix Figure A-1.

The focus of this thesis, the application MGGP on weather data set had a very successful performance with the GPTIPS default parameters with varying number of genes of 4, 5 and 6. In the presence of optimal parameters for the MGGP it is possible that the method outperforms the other regression methods. The predicted and actual daily average temperature graph can be seen in the following Figure 4.4. The Pareto optimal solution can be obtained from the Figure 4.3. We can say that with respect to the MGGP structured models predicting the daily average wind speed (see Figure 4.8), for daily average temperature we have a very high percent of solutions with lower RMSE and complexity measures. The solution represented with red dot is the best performing solution with its level of accuracy and complexity. The Figures 4.1 and 4.2 presents the RMSE for training and testing data.

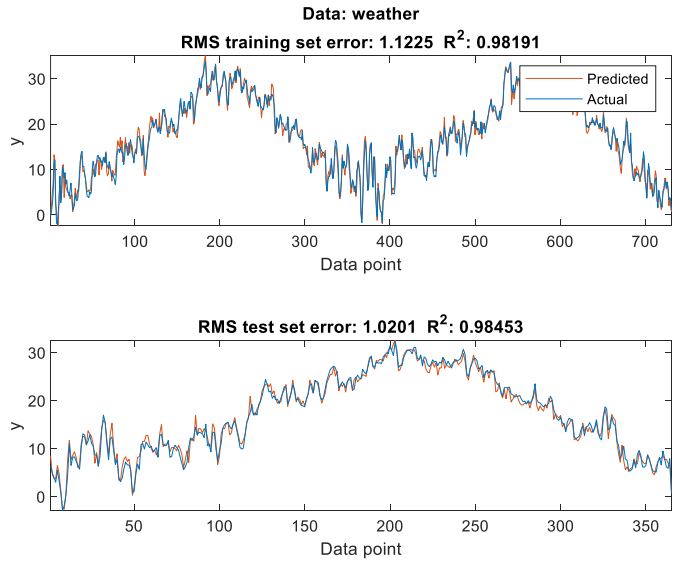


Figure 4-1 Train/Test RMSE (Best Subset Daily Average Temperature–MGGP)

Figure 4-2 Actual vs. Predicted Scatterplot for training and test data

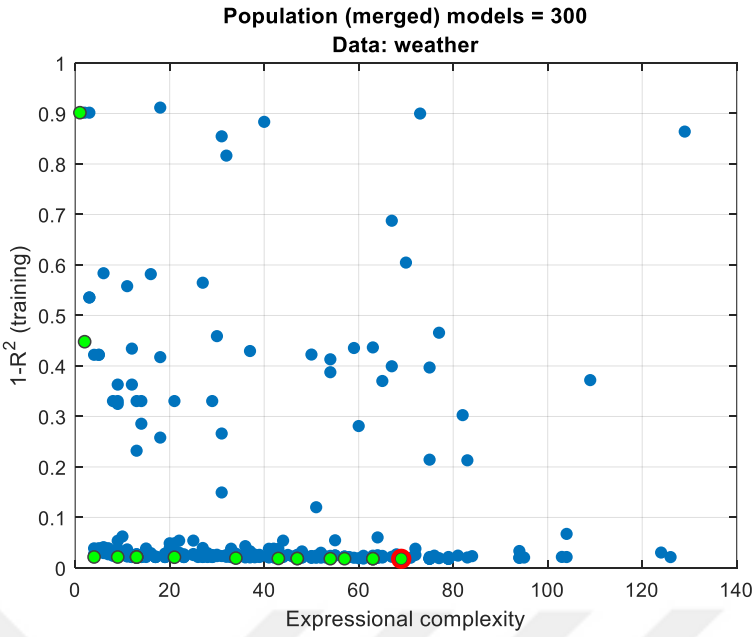


Figure 4-3 Pareto Front Graph (Best Subset Daily Average Temperature-MGGP)

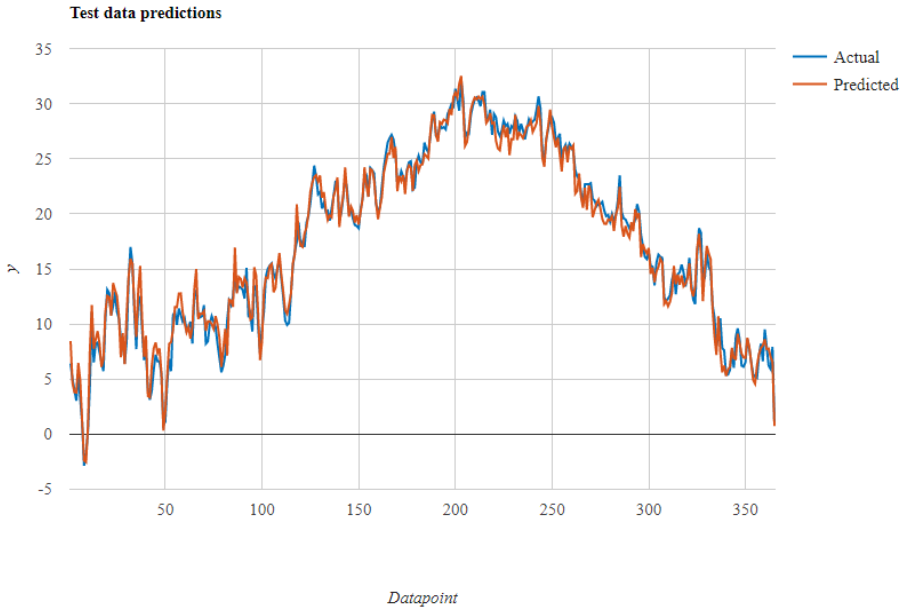


Figure 4-4 Predicted /Actual Daily Average Temperature (Best Subset-5 genes)

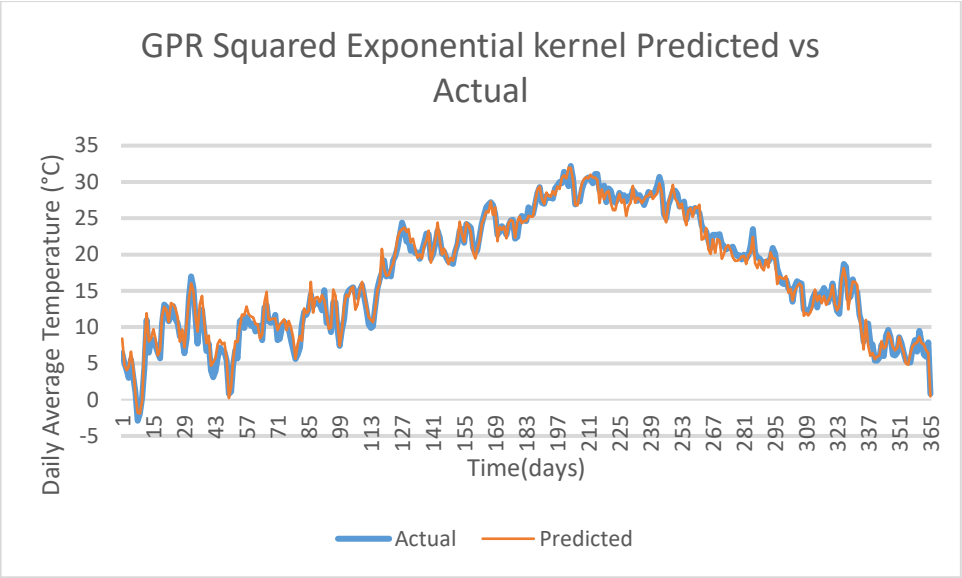


Figure 4-5 Daily Average Temperature (Best Subset GPR-Squared Exponential)

4.2 Daily Average Wind Speed

The forecasting performance is similar to the performance of daily average temperature for testing step but not for the training. The possible reason has been explained in the Section 4.1. Especially the Gaussian Process Regression for all 3 kernels had the best prediction performance compared both for the multiple regression methods and the ANN and MGGP approaches. The comparison of the actual and predicted 2017 daily average wind speed can be seen in the Figure 4.10 with GPR method with the Squared Exponential kernel.

As a result of the MGGP approach, we have observed the following Figures 4.6-4.8. The scatterplot for the predicted target value and the actual daily wind speed and the RMSE of both training and testing steps can be seen in the Figures 4.6 and 4.7.

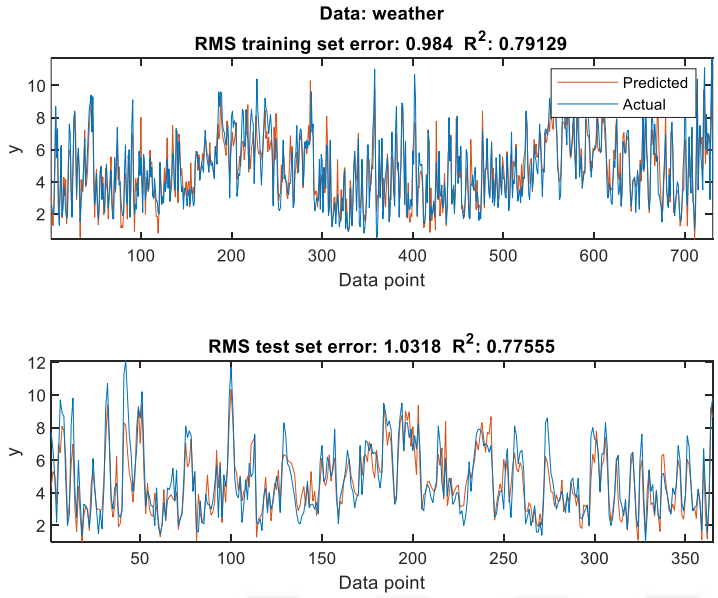


Figure 4-6 Train/Test RMSE (Best Subset Daily Average Wind Speed–MGGP)

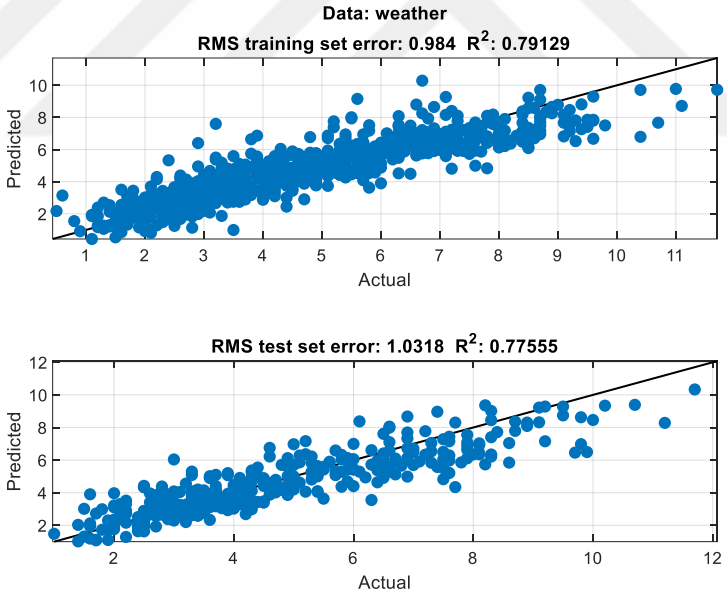


Figure 4-7 Actual vs. Predicted Scatterplot for training and test data

Figure 4.8 is very useful for visually recognizing all of the multigene regression models. This graph contains the final population of the MGGP process representing the relationship between the accuracy of the fitness of the model and the complexity (Searson, 2015). The model highlighted with the red dot is the best performing model in the population. The performances of these models are based on the R^2 values of the

training data. Green dots on the plot represent the non-dominated solutions which are also classified as the Pareto front of the models. The other solutions which either have a higher fitness or a lower complexity, they have higher fitness or lower complexity, respectively (Searson, 2015). The models represented with blue dots are the dominated models, these are non-Pareto models.

The best solution can be obtained from both the Pareto-optimal solution and the solution with highest accuracy and lowest complexity (Searson, 2015).

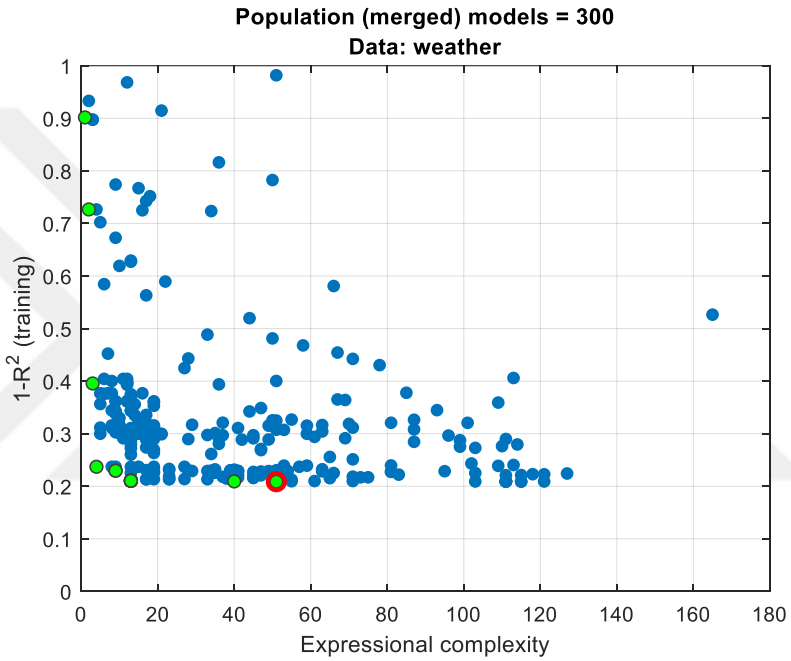


Figure 4-8 Pareto Front Graph (Best Subset Daily Average Wind Speed-MGGP)

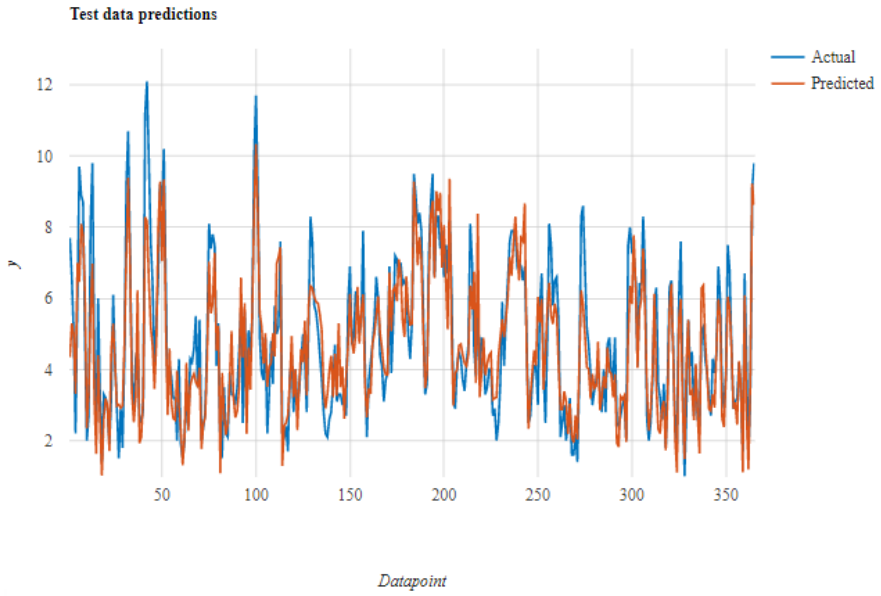


Figure 4-9 Predicted/Actual Daily Average Wind Speed (Best Subset-5 genes)

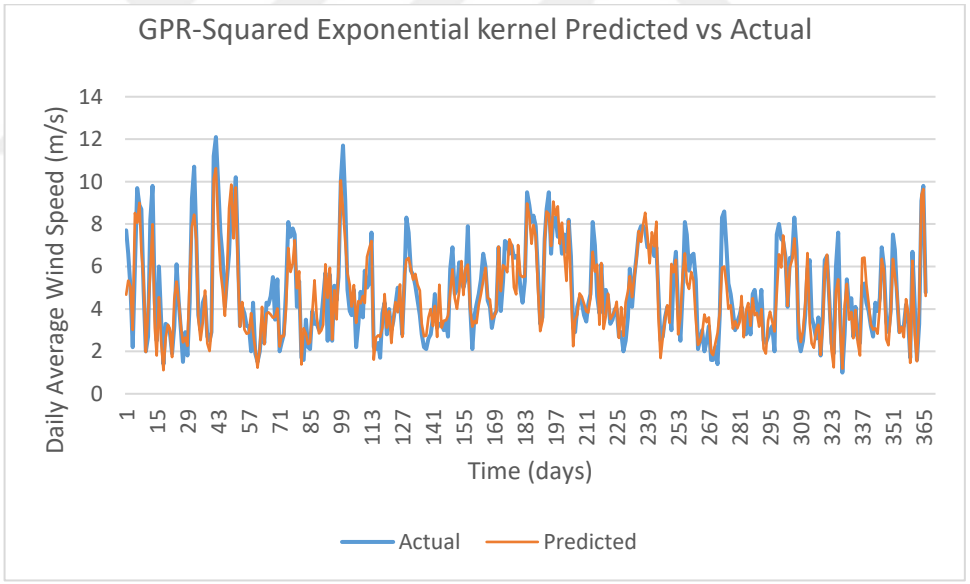


Figure 4-10 Daily Average Wind Speed (Best Subset GPR-Squared Exponential)

CHAPTER 5: CONCLUSION AND FUTURE WORK

In conclusion, in this study there have been the application of many different regression approaches on a single data set. We have provided the comparison of these methods both within the model with different parameters and among different methods. Firstly, the data was collected from General Directorate of Meteorology for Izmir Adnan Menderes Airport and then the data was processed to make it suitable for regression analysis. When the data was ready to be used, the subset selection to avoid overfitting has been done. For the rest of the analysis, both the data including all parameters and the best subset have been used for comparison. After selecting the best subset, linear regression analysis has been constructed to show that there is no linear relationship between the variables. Application of nonlinear regression tools such as multiple regression analysis has been done as a result of the data having nonlinear relationship. These multiple regression methods included, Gaussian Process Regression, Regression Trees, Support Vector Regression and Multiple Linear Regression. The remaining two methods, ANN and MGGP were the focus of this thesis. Finally, the RMSE performance evaluations of each method has been compared.

As a result of the study the aim was to achieve best performing regression model with various regression methods including the MGGP approach. In the end, some multiple regression methods such as Gaussian Process Regression have competed with the MGGP on predicting the target variables with their ability to catch and to generalize a nonlinear relation. The MGGP application has only been done using 3 different number of genes which is an option that the toolbox, GPTIPS provides for the user to change. There are other parameters which can also be adjusted to achieve the best performing regression model for predicting both daily average temperature and daily average wind speed for Izmir Adnan Menderes Airport. The use of SVR with the other regression tools and achieving successful prediction performance was also aimed in this study. The SVR is a commonly used classification tool and we have supported the concept of using SVR as a regression tool on predicting weather related parameters as a contribution to the literature.

Prediction of wind speed and temperature using MGGP was not common in the reachable literature with this much comprehensive approach using 6 different

regression methods. As a result of this study, we have provided a wide range of alternatives of methods which can be used as a guide for future studies. There is no significant difference of MGGP performance with respect to the other methods, but this performance can be improved by optimizing the other parameters in the future.

As a future work, an experimental design on changing the MGGP parameters and determining the most efficient will be conducted. Also, the dataset will be extended with the 2018 and 2019 data to improve the model training and generalization ability of the models.

The aviation sector values the precision of weather forecasts since it holds the importance of cost and safety. Any improvement to the currently used forecasting techniques will have huge impact to the airlines and airport managements.

REFERENCES

Weather-releated Aviation Accident Study (2010). [Online]. Available at https://www.asias.faa.gov/apex/f?p=100:8:::NO::P8_STDY_VAR:2.

Alweshah, M., Ababneh, M. and Alshareef, A. (2017) *Multi-Gene Genetic Programming for predict rainfall data*. *The International Arab Conference on Information Technology*. Yassmine Hammamet, Tunisia. 22-24 December 2017.

T.C. Ulaştırma ve Altyapı Bakanlığı Devlet Hava Meydanları İşletmesi Genel Müdürlüğü (2018) *Yolcu Trafiği (Gelen-Giden) 2018 Yılı Temmuz Sonu Havalimanları* [Online]. Available at https://www.dhmi.gov.tr/sayfalar/istatistik.aspx. (Accessed 15 May 2020).

Banzhaf, W., Nordin, P., Keller, R. E., Francone, F. (1998) *Genetic Programming: An Introduction on the Automatic Evolution of computer programs and its Applications*. San Francisco: Morgan Kaufmann.

Cao, Y., Wu, Z. and Xu, Z. (2014) *Effects of rainfall on aircraft aerodynamics*, Progress in Aerospace Sciences. Elsevier, 71, pp. 85–127. doi: 10.1016/j.paerosci.2014.07.003.

Danandeh Mehr, A. and Kahya, E. (2017) *A Pareto-optimal moving average multigene genetic programming model for daily streamflow prediction*, Journal of Hydrology. Elsevier B.V., 549, pp. 603–615. doi: 10.1016/j.jhydrol.2017.04.045.

Demirhan, H. and Kayhan Atilgan, Y. (2015) *New horizontal global solar radiation estimation models for Turkey based on robust coplot supported genetic programming technique*, Energy Conversion and Management. Elsevier Ltd, 106, pp. 1013–1023. doi: 10.1016/j.enconman.2015.10.038.

Elhenawy, M., Chen, H. and Rakha, H. A. (2014) *Dynamic travel time prediction using data clustering and genetic programming*, Transportation Research Part C:

Emerging Technologies. Elsevier Ltd, 42, pp. 82–98. doi: 10.1016/j.trc.2014.02.016.

Fabbian, D., de Dear, R. and Lellyett, S. (2007) *Application of Artificial Neural Network Forecasts to Predict Fog at Canberra International Airport*, Weather and Forecasting, 22(2), pp. 372–381. doi: 10.1175/WAF980.1.

Faris, H. and Sheta, A. (2016) *A comparison between parametric and non-parametric soft computing approaches to model the temperature of a metal cutting tool*, International Journal of Computer Integrated Manufacturing. Taylor & Francis, 29(1), pp. 64–75. doi: 10.1080/0951192X.2014.1002809.

Fox, J. (2000) *Multiple and Generalized Nonparametric Regression*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-131. Thousand Oaks, CA: Sage OR.

Golafshani, E. M., Rahai, A. and Sebt, M. H. (2015) *Artificial neural network and genetic programming for predicting the bond strength of GFRP bars in concrete*, Materials and Structures/Materiaux et Constructions, 48(5), pp. 1581–1602. doi: 10.1617/s11527-014-0256-0.

Gultepe, I., Sharman, R., Williams, P.D., Zhou, B., Ellrod, G., Minnis, P., Trier, S., Griffin, S., Yum, S. S., Gharabaghi, B., Feltz, W., Temimi, M., Pu, Z., Storer, L. N., Kneringer, P., Weston, M.J., Chuang, H., Thobois, L., Dimri, A.P., Dietz, S.J., Françca, G.B., Almeida, M.V., Neto, F.L.A. (2019) *A Review of High Impact Weather for Aviation Meteorology*, Pure and Applied Geophysics 176, pp. 1869–1921. doi: 10.1007/s00024-019-02168-6.

Hasan, N., Nath, N. C. and Rasel, R. I. (2016) *A support vector regression model for forecasting rainfall*, *2nd International Conference on Electrical Information and Communication Technologies, EICT 2015*, Khulna, Bangladesh 10-12 December 2015, pp. 554–559. doi: 10.1109/EICT.2015.7392014.

Heimann, P. and Isaacs, S., Klein M., Riviere, J. (2018) *Regression. Developments in Psychoanalysis*. New York: Routledge.

Hirani, M. D. and Mishra, D. N. (2016) *A Survey On Rainfall Prediction Techniques*, International Journal of Computer Application, 6(2), pp. 1797–2250. doi: 10.3389/fnhum.2014.00445.

Khatib, T., Mohamed, A. and Mahmoud, M. (2012) *Estimating Global Solar Energy Using Multilayer Perception Artificial Neural Network*, International Journal of Energy, 6(1), pp. 25–33. [Online]. Available at: http://www.naun.org/journals/energy/17-377.pdf. (Accessed 15 May 2020).

Li, L. (2019) *Classification and Regression Analysis with Decision Trees, towards data science*. [Online]. Available at: https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054. (Accessed 15 May 2020).

Lones, M. (2004). *Genetic Programming* [Online]. Available at: http://www.macs.hw.ac.uk/~ml355/common/thesis/c6.html. (Accessed 15 May 2020).

Mislan, Haviluddin, Hardwinarto, S., Sumaryono, Aipassa, M. (2015) *Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggarong Station, East Kalimantan - Indonesia*, Procedia Computer Science. Elsevier Masson SAS, 59(Iccsci), pp. 142–151. doi: 10.1016/j.procs.2015.07.528.

Morrison, G., Searson, D. and Willis, M. (2010) *Using Genetic Programming to Evolve a Team of Data Classifiers*, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 4, pp. 1–4. [Online]. Available at: https://pdfs.semanticscholar.org/b17d/5d6b998590cf6fc6dd97d3d0125df88579cf.pdf. (Accessed 15 May 2020).

Narvekar, M. and Fargose, P. (2015) *Daily Weather Forecasting using Artificial Neural Network*, International Journal of Computer Applications, 121(22), pp. 9–13. doi: 10.5120/21830-5088.

Neill, O., Agapitos, A., Brabazon, A. (2012) *Genetic Programming for the Induction of Seasonal Forecasts: A Study on Weather-derivatives*, Financial Decision Making using Computational Intelligence, Series in Optimisation and its Applications, 70.

Olatomiwa, L., Mekhilef, S., Shamshirband, S., Mohammadi, K., Petkovic, D.,Sudheer, C. (2015) *A support vector machine-firefly algorithm-based model for global solar radiation prediction*, Solar Energy, 115, pp. 632–644. doi: 10.1016/j.solener.2015.03.015.

Orove, J. O., Osegi, N. E. and Eke, B. O. (2015) *A Multi-Gene Genetic Programming Application for Predicting Students Failure at School*, African Journal of Computing & ICT *arXiv:1503.03211 [cs]*, pp. 21–34.

Pan, I., Pandey, D. S. and Das, S. (2013) *Global solar irradiation prediction using a multi-gene genetic programming approach*, Journal of Renewable and Sustainable Energy, 5(6), pp. 1–31. doi: 10.1063/1.4850495.

Potesth, S. (2001) *A neural network model for visibility nowcasting from surface observations: Results and sensitivity to physical input variables*, Journal of Geophysical Research, 106, pp. 951–959.

Ramedani, Z., Omid, M., Keyhani, A., Shamshirband, S., Khoshnevisan, B. (2014) *Potential of radial basis function-based support vector regression for global solar radiation prediction*, Renewable and Sustainable Energy Reviews. Elsevier, 39, pp. 1005–1011. doi: 10.1016/j.rser.2014.07.108.

Ramesh, K., Anitha, R. and Ramalakshmi, P. (2015) *Prediction of lead seven day minimum and maximum surface air temperature using neural network and genetic programming*, Sains Malaysiana, 44(10), pp. 1389–1396. doi: 10.17576/jsm-2015-4410-03.

Roshni, T., Md. Sajid, K. and Samui, P. (2017) *Potential of regression models in projecting sea level variability due to climate change at Haldia Port, India*, Ocean Systems Engineering, 7(4), pp. 319–328. doi: 10.12989/ose.2017.7.4.319.

Samadianfard, S., Delirhassania, R., Kişi, Ö., Agirre-Basurko, E. (2013) *Comparative analysis of ozone level prediction models using gene expression programming and multiple linear regression*, Geofizika, 30, pp. 43–74.

Searson, D. (2009) *Genetic Programming & Symbolic Regression for MATLAB User Guide*, (November), pp. 1–26.

Searson, D. P. (2015) *GPTIPS 2 - an open-source software platform for symbolic data mining*, Handbook of Genetic Programming Applications, (c), pp. 1–26. doi: 10.1007/978-3-319-20883-1_22.

Searson, D., Leahy, D. and Willis, M. (2010) *GPTIPS: An open source genetic programming toolbox for multigene symbolic regression*, *Proceedings of the International multi conference on engineering computer science*, pp. 77–80. Hong Kong. 17-19 March 2010. doi: 10.1080/14783363.2011.611358.

Sheta, A. F. and Mahmoud, A. (2001) *Forecasting using genetic programming*, *Proceedings of the Annual Southeastern Symposium on System Theory*, 2001-January (January 2015), pp. 343–347. doi: 10.1109/SSST.2001.918543.

Smola, A. J. and Schölkopf, B. (2004) *A tutorial on support vector regression*, Statistics and Computing, 14, pp. 199–222.

Thota, S. (2018) *Forecasting Black River Flow Using Feedforward Artificial Neural Networks.* Master Thesis. Corpus Christi, Texas A&M University

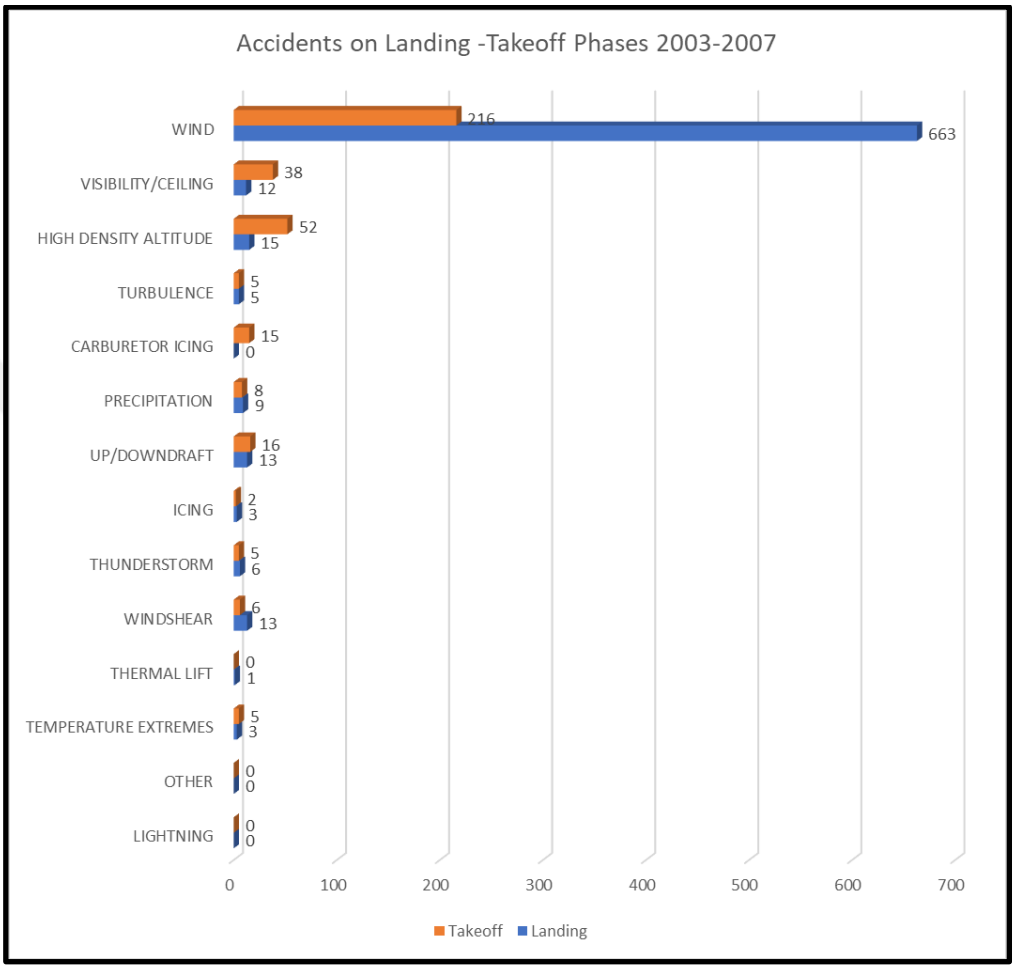
Wu, J., Long, J. and Liu, M. (2015) *Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm*, Neurocomputing, 148, pp. 136–142. doi: 10.1016/j.neucom.2012.10.043.

Yozgatligil, C., Aslan, S., İyigün, C., Batmaz, İ. (2013) *Comparison of missing value imputation methods in time series: The case of Turkish meteorological data*, Theoretical and Applied Climatology, 112(1–2), pp. 143–167. doi: 10.1007/s00704-012-0723-x.

Zhou, Y., Zhang, N., Li, C., Liu, Y., Huang, P. (2018) *Decreased takeoff performance of aircraft due to climate change*, Climatic Change. Climatic Change, 151(3–4), pp. 463–472. doi: 10.1007/s10584-018-2335-7.

APPENDICES

Appendix A.



Appendix Figure A-1 Number of Accidents During Landing-Takeoff Phases (data from NTSB Aviation Accident and Incident Database 2010)

Appendix B.

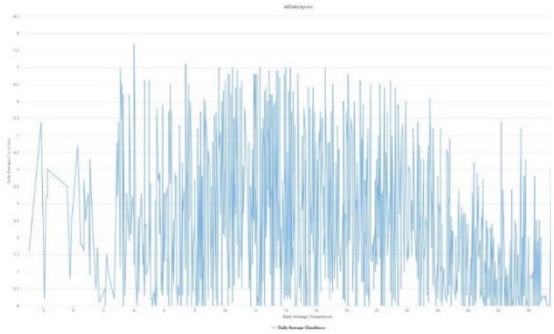
Appendix Table B-1 Regression Results for Daily Average Temperature

Method	Inputs	S	R-Sq	R-sq(adj)	R-sq(pred)	Mallow's Cp	Regression Function
Stepwise Selection	x_3, x_5, x_6, x_7, x_8, x_9	1,19	97,97%	97,95%	97,93%	7,87	x_{10}=107.13-0.0790x_3+0.6343 x_5 + 0.88165 x_6 -0.11442x_7 +0.2552x_8+ 0.02214 x_9
Forward Selection	x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9	1,19	97,97%	97,96%	97,93%	8,48	x_{10}=110.80-0.0568x_1-0.0729x_3-0.0646x_4+0.6273x_5+ 0.87766x_6-0.0611x_7+0.2571x_8+ 0.02209 x_9
Backward Elimination	x_3, x_5, x_6, x_7, x_8, x_9	1,19	97,97%	97,95%	97,93%	7,87	x_{10} = 107.13 - 0.0790x_3+0.6343x_5+ 0.88165x_6-0.11442x_7+0.2552x_8+0.02214 x_9

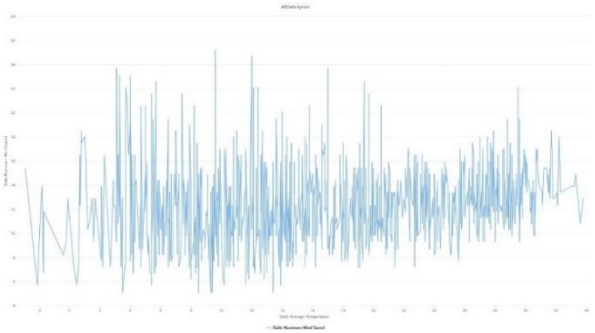
Appendix Table B-2 Regression Results for Daily Average Wind Speed

Method	Inputs	S	R-Sq	R-sq(adj)	R-sq(pred)	Mallow's Cp	Regression Function
Stepwise Selection	x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_9, x_{10}	0,1	78,79%	78,65%	78,40%	6,45	x_5=-18.40 - 0.1107 x_1+ 0.3517 x_3- 0.4472 x_6+ 0.1364 x_7- 0.1602x_8- 0.05611 x_9+ 0.4454 x_{10}
Forward Selection	x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_9, x_{10}	0,1	78,79%	78,65%	78,40%	6,45	x_5=-18.40 - 0.1107 x_1+ 0.3517 x_3- 0.4472x_6 + 0.1364 x_7- 0.1602 x_8- 0.05611 x_9+ 0.4454 x_{10}
Backward Elimination	x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_9, x_{10}	0,1	78,79%	78,65%	78,40%	6,45	x_5=-18.40 - 0.1107x_1 + 0.3517 x_3- 0.4472 x_6+ 0.1364 x_7- 0.1602 x_8- 0.05611 x_9+ 0.4454 x_{10}

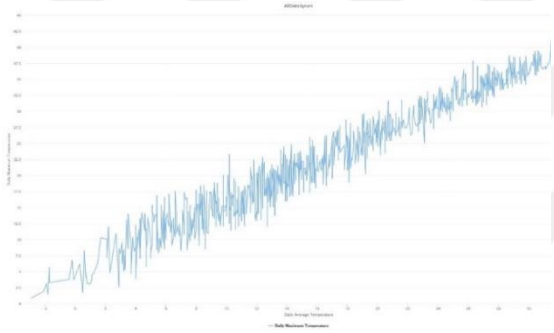
Appendix C.



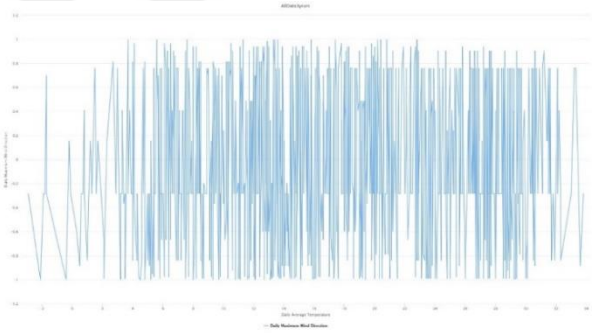
Appendix Figure C-1 Daily Average Cloudiness-Temperature



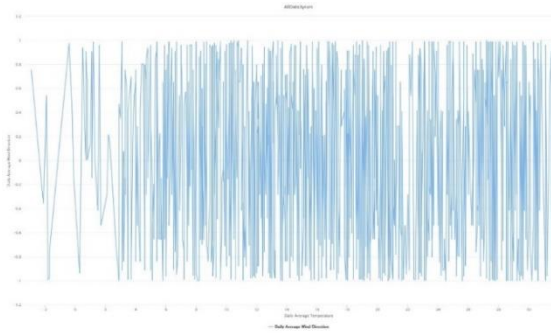
Appendix Figure C-4 Daily Maximum Wind Speed – Temperature



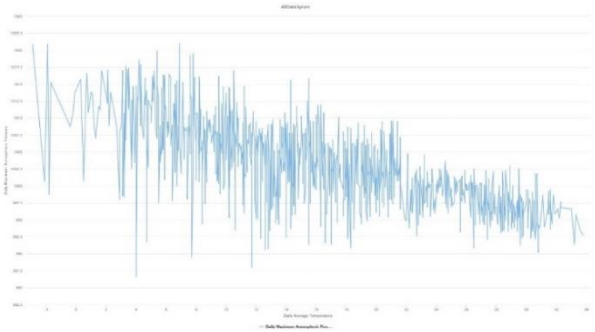
Appendix Figure C-2 Daily Maximum Temperature – Average Temperature



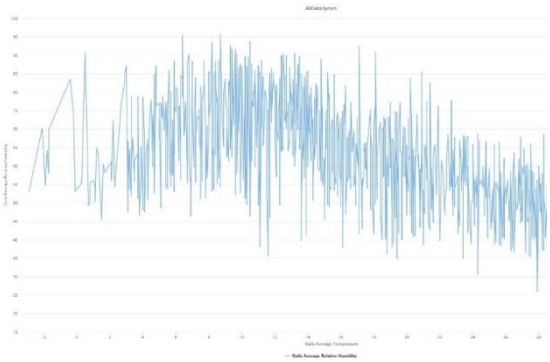
Appendix Figure C-5 Daily Maximum Wind Direction -Average Temperature



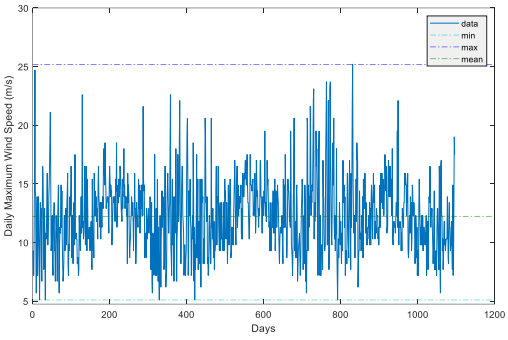
Appendix Figure C-3 Daily Average Wind Direction - Temperature



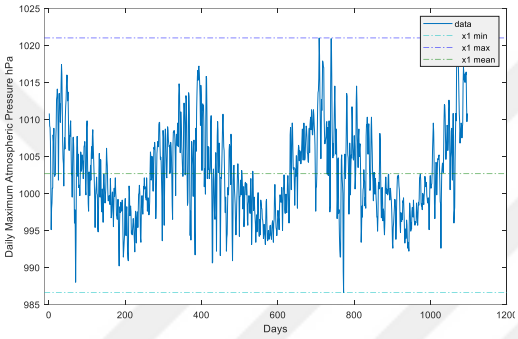
Appendix Figure C-6 Daily Maximum Atm. Pressure – Average Temperature



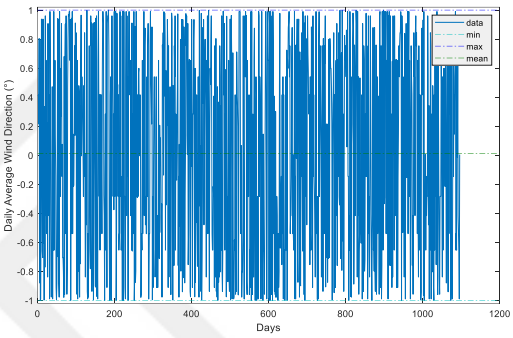
Appendix Figure C-7 Daily Average Relative Humidity- Temperature



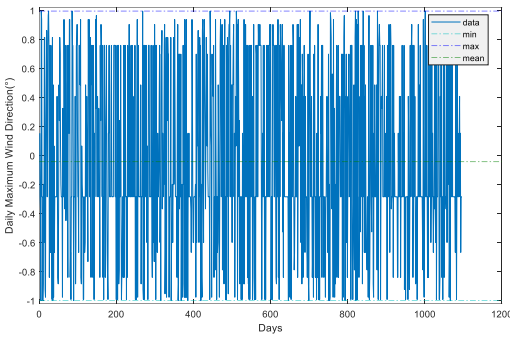
Appendix Figure C-10 Daily Maximum Wind Speed Time Series



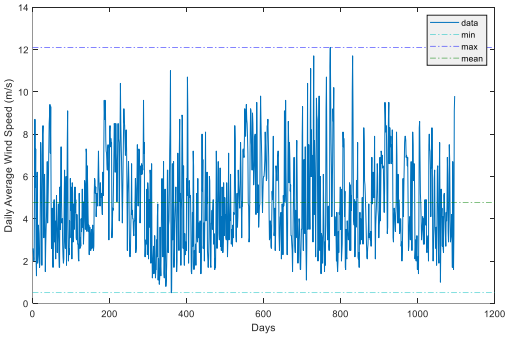
Appendix Figure C-8 Daily Maximum Atm. Pressure Time Series



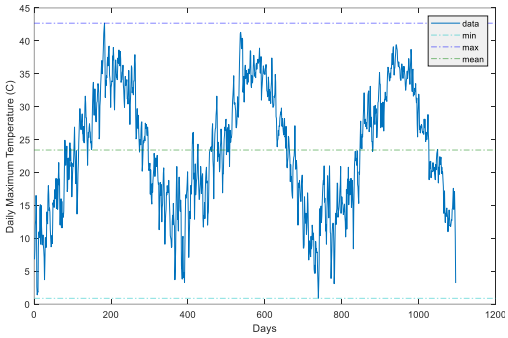
Appendix Figure C-11 Daily Average Wind Direction Time Series



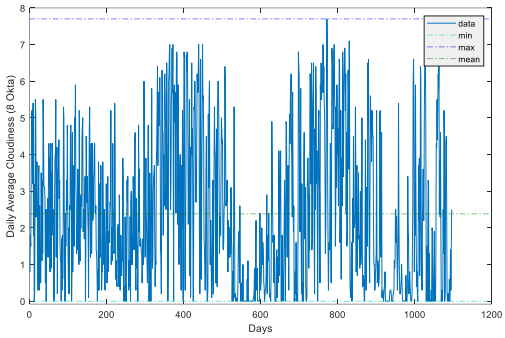
Appendix Figure C-9 Daily Maximum Wind Direction Time Series



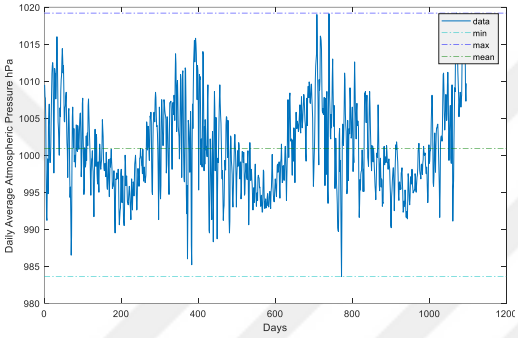
Appendix Figure C-12 Daily Average Wind Speed Time Series



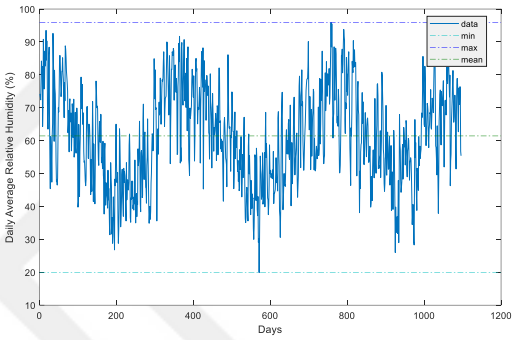
Appendix Figure C-13 Daily Maximum Temperature Time Series



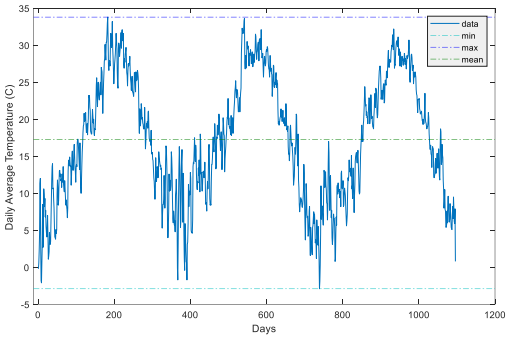
Appendix Figure C-15 Daily Average Cloudiness Time Series



Appendix Figure C-14 Daily Average Atm. Pressure Time Series



Appendix Figure C-16 Daily Average Relative Humidity Time Series



Appendix Figure C-17 Daily Average Temperature Time Series