Weather Station Data Analysis

By Livio Aziz

Supervisor: Dr. Matthias Maischak

2021/22

Acknowledgment

First of all, I'd like to express my sincere gratitude towards my supervisor and personal tutor Dr. Matthias Maischak. With his continuous support throughout the year and his expert advice I was able to complete this project.

I would also like to thank my family for helping me get through this very stressful period and not letting me give up.

Abstract

Dealing with a lot of data can pose a challenge when it comes to extracting information. Because of this, techniques such as regression analysis are used in order to extract meaningful information from the data that will allow effective data analysis to occur. This project will be discussing the methods in which data from a weather station can be extracted as well as outlining the quality of the data . Furthermore, the project will aim to find the relationship between temperature and humidity based on the data that's given from the weather station.To do this, multiple regression analysis will be used in order to find coefficients that will enable useful information to be plotted such as the mean temperature throughout the year.

Lastly, the project will delve into the programming language MATLAB, showing the different ways in which this language can manipulate and sort out data.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The aim of this project is to analyse data from a weather station. By using this data we can see how temperature and humidity varies throughout the day and night and extract information such as when the temperature is usually highest/lowest.

To begin with we are given 4 spreadsheets, containing timestamps and a reading for temperature and humidity every 5 minutes. That is a lot of data to be dealing with. 104482 to be exact. Without doing anything to data, there is no information that can be extracted as well as no way of knowing if the results given are accurate. In order to tackle this we must find a way in order to extract information as well as comment on the quality of data. To do this we must use the programming language MATLAB. MATLAB is a powerful programming language that allows to manipulate large data sets in order to gather useful information that can be extracted. This will be essential since other data processing applications such as Microsoft excel would be ineffective.

In order to effectively use MATLAB we must also understand the fundamentals of statistical analysis. Regression analysis methods let us forecast and predict observations in the data.These forecasts can help in finding correlations between temperature and humidity or determining whether the data follows forecasted trends.

Chapter 2

Statistical mathematics

2.1 Regression analysis

Since we're given a lot of data from the weather station its important to understand the statistical analysis methods in order to extract information effectively.

Simple linear regression

Regression analysis has two main objectives. First it attempts to establish if there are any statistical relationships between two variables. Secondly it attempts to forecast/predict new observations in the data.

In regression models the different variables each have a role. The dependent variable is the variable whose values we need in order to explain or forecast. Theses values are dependent on something else, meaning it can change because of something else. The dependent variable is denoted as the letter y. The independent variable, denoted by the letter x, has values that are independent and are used to explain the dependent variables. An example of a common regression model is simple linear regression. In the case of linear regression a linear equation can be used in order to plot a line of best fit for the data using the equations:

$$y = a + bx$$

$$y = \beta_0 + \beta_1 x$$

Since it is linear we use the equation of line but we replace the original equation with statistical notations β_0 being the intercept, β_1 being the x coefficient with x as the independent variable and y being the dependent variable. To calculate a line of best fit, we follow these steps: [3]

- Step 1: for each point of variable x and y calculate the value of x^2 and xy [3].
- Step 2: find the sum of all x, y, x^2 and xy ($\sum x$, $\sum y$, $\sum x^2$ and $\sum xy$).
- Step 3: Calculate the slope (β_1:

$$\beta_1 = \frac{N\sum(xy) - \sum x \sum y}{N\sum(x^2) - (\sum x)^2}$$

  Where N is the number of points.

- Step 4: Calculate the intercept β_0:

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{N}$$

- Step 5: Use the values gathered in order to create a linear equation of a line in the form:

$$y = \beta_0 + \beta_1 x$$

However, since the world isn't linear, there are bound to be errors and because of that it needs to be added to the equation which gives us:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

2.2 Linear regression

Linear regression is a parametric regression method. A parametric learning method uses a fixed equation form that establishes the relationship between the dependent variable y and the independent co-variants.These parametric regression models are expressed with the equation:

$$y = f(X) + \varepsilon$$

Where y is the predicted output, and ϵ is the error term independent of X.[2]

$f(X)$ can be used in different forms depending on the statistical method.In the case of With one co-variant X in linear regression, the equation takes the form of :

$$f(X) = \beta_0 + \beta_1 X$$

This equation reduces the problem of finding a relationship between the co-variants X and the response variable y to determining the 2 coefficients β_0 and β_1. What linear regression does is it estimates the values of these coefficients. So the goal of linear regression is to find estimate values for the coefficients based on the given data. These estimates are represented by the hat.

$$\widehat{y} = \widehat{f(X)}$$

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} X$$

Since this equation is the linear regression estimate, there is no ϵ term.As mentioned before ϵ is independent of X which means it cannot be determined by regression analysis so no matter how perfect a model is, it will still have an error ϵ. Because of this, ϵ is known as the irreducible error.

In order to estimate the coefficients accurately, we must minimise the the difference between the actual value and the predicted/estimated value. This is called the residual error.

$$e = y - \widehat{y}$$

$$e = \beta_0 + \beta_1 X + \varepsilon - (\widehat{\beta_0} + \widehat{\beta_1} X)$$

$$e = (\beta_0 - \widehat{\beta_0}) + (\beta_1 - \widehat{\beta_1}) X + \varepsilon$$

If there are n sample data points where the residual error of the i sample is e_i, our goal would be to minimise the sum of square residuals (RSS) because the magnitude of the error is important no matter if the model outputs a higher or lower prediction.

$$e_i = y_i - \widehat{y_i}$$

$$RSS = \sum_{i=1}^{n} e_i^2$$

We rewrite the equation using arg min. What this does is it remits the values of β_0 and β_1 that minimises the residual sum of squares.

$$argmin \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

We then substitute the label estimate of $\hat{y}$ and expand the brackets.

$$argmin \sum_{i=1}^{n} (y_i - (\widehat{\beta_0} + \widehat{\beta_1} X_i))^2$$

$$argmin \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$$

In order to find the stationary points we differentiate with respect to β_0 and β_1.

First we differentiate with respect to β_0.

$$\sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$$

$$2 \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)(-1) = 0$$

$$\sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i) = 0$$

We then equate everything to $\widehat{\beta_0}$ and divide the sum by the total sample n.

$$\widehat{\beta_0} = \frac{\sum_{i=1}^{n} (y_1 - \widehat{\beta_1} X_i)}{n}$$

$$\widehat{\beta_0} = \sum_{i=1}^{n} (\frac{y_i}{n} - \widehat{B_1} \frac{X_i}{n})$$

$$\widehat{\beta_0} = \sum_{i=1}^{n} \frac{y_i}{n} - \widehat{B_1} \sum_{i=1}^{n} \frac{X_i}{n}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} \bar{X}$$

Where $\bar{y}$ and $\bar{X}$ are the sample means.

Now we differentiate with respect to β_1 using partial differentiation rules.

$$2\sum_{i-1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}X_i)(X_i) = 0$$

$$\sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1})(X_i) = 0$$

$$\sum_{i=1}^{n} X_i y_i - \widehat{B_0}\sum_{i=1}^{n} X_i - \widehat{\beta_1}\sum_{i=1}^{n} X_i^2 = 0$$

We then substitute the value of $\widehat{\beta_0}$ that we found before.

$$\sum_{i=1}^{n} X_i y_i - (\sum_{i=1}^{n}\frac{y_i}{n} - \widehat{\beta_1}\sum_{i=1}^{n}\frac{X_i}{n})\sum_{i=1}^{n} X_i - \widehat{\beta_1}\sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i y_i - \sum_{i=1}^{n}\frac{y_i}{n}\sum_{i=1}^{n} X_i + \widehat{\beta_1}\sum_{i=1}^{n}\frac{X_i}{n}\sum_{i=1}^{n} X_i - \widehat{\beta_1}\sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i-1}^{n} X_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} X_i + \widehat{\beta_1} * \frac{1}{n}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i - \widehat{\beta_1}\sum_{i=1}^{n} X_i^2 = 0$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i y_i - n\bar{X}\bar{y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

replacing the sample size with the respected means.After completely simplifying everything we get an equation for $\widehat{\beta_1}$ estimates.

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{X}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i y_i - n\bar{X}\bar{y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$\widehat{\beta_0}$ and $\widehat{\beta_1}$ together define the least square coefficients estimates, so given some data points we can determine the coefficients by plugging the values into the regression equation which can allow a value of y to be computed with an unseen X value, thus a prediction can be made. [2]

2.3 Multiple regression

Again like linear regression, multiple regression has the general parametric equation:

$$y = f(X) + \varepsilon$$

However with the case of multiple regression, $f(X)$ takes the form of the sum of products p co-variants and coefficients.

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

The estimated value of the response variable y is given by the equation:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + ... + \widehat{\beta_p} X_p$$

for n samples, number of operations $= n * (p-1)^2$

In order to compute large numbers, we use matrices in order to make it easier and quicker for computers to compute. So, multiple regression is represented in a matrix form.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_1,1 & X_1,2 & . & . & X_1,p \\ 1 & X_2,1 & X_2,2 & . & . & . \\ 1 & X_3,1 & X_3,2 & . & . & . \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & X_n,1 & X_n,2 & . & . & X_n,p \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

From our multiple regression equations, we see that $y = X\beta + \varepsilon$ and $\hat{y} = X\hat{\beta}$.

The matrix of y is (n**x**1) and X is a n**x**$(p+1)$.It's $p+1$ because of the additional matrix multiplicand which in this case is one.β is a $(p+1)$ x 1 dimensional column vector and ϵ is a n**x**1 dimensional column vector.

To prove the relationship is correct we substitute the values into the matrix with multiplication and addition.

$$Y = X\beta + \varepsilon \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} + ... + \beta_p X_{1,p} + \varepsilon_1 \\ \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} + ... + \beta_p X2, p + \varepsilon_2 \\ \beta_0 + \beta_1 X_{3,1} + \beta_2 X_{3,2} + ... + \beta_p X_{3,p} + \varepsilon_3 \\ . \\ . \\ . \\ \beta_0 + \beta_1 X_{n,1} + \beta_2 X_{n,2} + ... + \beta_p X_{n,p} + \varepsilon n \end{bmatrix}$$

We can see that each entry of the vector takes the multiple regression form $Y = X\beta + \varepsilon$.

To compute these coefficients we minimise the least square criteria like linear regression, however we apply this to the matrices. So we need to minimise the sum of squares of residuals(RSS),this is also known as sum squared error. To do this we consider the matrix form of the residuals e.

$$e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \widehat{y_1} \\ y_2 - \widehat{y_2} \\ y_3 - \widehat{y_3} \\ . \\ . \\ . \\ y_n - \widehat{y_n} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} - \begin{bmatrix} \widehat{y_1} \\ \widehat{y_2} \\ \widehat{y_3} \\ . \\ . \\ . \\ \widehat{y_n} \end{bmatrix} = y - \widehat{y}$$

We see that residuals of all samples is equivalent to the difference between vectors of y and $\hat{y}$.

$$RSS = \sum_{i=1}^{n} e_i^2$$

The residual sum of squares can be re written as:

$$RSS = e^T e$$

$RSS = e^T e$
We then replace e with (y - $\hat{y}$)
$RSS = (y - \hat{y})^T (y - \hat{y})$
Then replace $\hat{y}$ with $X\hat{\beta}$
$RSS = (y - X\hat{\beta})^T (y - X\hat{\beta})$
$RSS = (y^T - \widehat{\beta^T} X^T)(y - X\hat{\beta})$
Finally we expand to get the equation:
$RSS = y^T y - y^T X\hat{\beta} - \hat{\beta} X^T y + \widehat{\beta^T} X^T X\hat{\beta}$

Matrix differentiation
$x = m\mathbf{x}1$ matrix
$A = n\mathbf{x}m$

If x and A satisfy these equations, we can use the following formulae:

$y = A \rightarrow \frac{\delta y}{\delta x} = 0$

$y = Ax \rightarrow \frac{\delta y}{\delta x} = A$

$y = xA \rightarrow \frac{\delta y}{\delta x} = A^T$

$y = x^T Ax \rightarrow \frac{\delta y}{\delta x} = 2x^T A$
We'll use these formulas in order to minimise the RSS. [1]

$$RSS = y^T y - y^T X\widehat{\beta} - \widehat{\beta} X^T y + \widehat{\beta^T} X^T X\widehat{\beta}$$

To minimise the RSS we must find the stationary points, and to do this we equate the differential of the equation to 0. We do this in order to find the value of the estimate beta (β) for which the partial differential of RSS with respect to $\widehat{\beta}$ is 0.

$$\frac{\delta(RSS)}{\delta\widehat{\beta}} = \frac{\delta(y^T y - y^T X\widehat{\beta} - \widehat{\beta} X^T y + \widehat{\beta} X^T X\widehat{\beta})}{\delta\widehat{\beta}} = 0$$

We then expand the partial derivatives to every term.

$$\frac{\delta(y^T y)}{\delta\widehat{\beta}} - \frac{\delta(y^T X\widehat{\beta})}{\delta\widehat{\beta}} - \frac{\delta(\widehat{\beta} X^T y)}{\delta\widehat{\beta}} + \frac{\delta(\widehat{\beta^T} X^T X\widehat{\beta})}{\delta\beta} = 0$$

$$0 - y^T X(X^T y)^T + 2\widehat{\beta^T} X^T X = 0$$

The first term is independent of X so their derivative is 0, the second term has the second formula form $y = Ax$ so the derivative is simply $y^T X$. The third term is of the form $y = xA$ so the derivative is $(X^T y)^T$ and the last term has the form $y = x^T Ax$ so the derivative is $2\widehat{\beta^T} X^T X$.

$$0 - y^T X - y^T X + 2\widehat{\beta^T} X^T X = 0$$

$$2\widehat{\beta^T} X^T X = 2y^T X$$

We expand the transpose and put the negative terms on the other side.

$$\widehat{\beta^T} - y^T X(X^T X)^{-1}$$

$$\widehat{\beta^T} X^T X = y^T X$$

We now cancel the two from both sides and apply $(X^T X)^-1$ on both sides.

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

Finally we get the transpose on both sides in order to get the $\widehat{\beta}$ term on the left [2].

This equation can also be written without the hat as [5]:

$$\beta = (X^T X)^{-1} X^T y$$

Chapter 3

Dealing with data

3.1 Matlab code

In order to begin analysis from the spreadsheets given, a method must be found in order to do effective analysis.We are given 4 spreadsheets from April 2019 to March 2020. The spreadsheets have 4 columns containing the time stamps, the timezone/time, the temperature and the humidity. The values of temperature and humidity are taken every 5 minutes which is a lot of data. Overall there are roughly more than 100000 readings of temperature and humidity throughout the year. Because of the large number of readings, using excel would be ineffective and insufficient so MATLAB is used since its more effective at dealing with large data sets. As well as this, we face many challenges with the way the spreadsheet is formatted.In the case of the spreadsheets used, timestamps are used which without correct manipulation, no useful information can be extracted, thus MATLAB can be used in order to re sample the format of the data allowing more information to be extracted. Matlab is a powerful tool for data analysis which is able to quickly and effectively handle most challenges that come with data analysis.

3.1.1 Importing and combining all spreadsheets

Listing 3.1: Importing spreadsheets into MATLAB

```
1 clc;
2 which extract
```

```
3 AprMayJun2019 = readtable("Outdoor_AprMayJun_2019.xls")
4 JulAugSep2019 = readtable("Outdoor_JulAugSep_2019.xls")
5 OctNovDec2019 = readtable("Outdoor_OctNovDec_2019.xls")
6 JanFebMar2020 = readtable("Outdoor_JanFebMar_2020.xls")
7
8 yeardata = [AprMayJun2019; JulAugSep2019; OctNovDec2019;
      JanFebMar2020]
```

In order to import the four spreadsheets into MATLAB, we use the readtable function in order for the program to create a table using the .csv spreadsheet file. To make dealing with the data easier we combine all the spreadsheets together and call this spreadsheet "yearsdata".From this code we have combined all spreadsheets into one which contains all readings throughout the year.

3.1.2 New arrays for Time, Temperature and humidity

Next we need to assign a name for each column of the spreadsheet.This is done so we are able to easily call specific columns of the spreadsheets with ease. We must also tackle the issue of dealing with timestamps.

Listing 3.2: Creating arrays

```
1 yeardata=[AprMayJun2019; JulAugSep2019; OctNovDec2019;
      JanFebMar2020]
2
3 Times =table2array(yeardata(:,1));
4 Times = Times-Times(1)+17;
5 Temp =table2array(yeardata(:,3));
6 Humidity = table2array(yeardata(:,4));
7 TimeAndTemp = [Times,Temp];
8 TimeAndHumidity = [Times,Humidity]
```

In order to manipulate the years data spreadsheet more effectively we use the function table2array in order to covert the table into a homogeneous array. In the case of time, we convert the first row of yearsdata into an array and assign it to the name "Times". Since the times column starts from 00:17 we have to change it so it starts from zero add 17 seconds since it doesn't start exactly at midnight.We do this by subtracting the "Times" array by itself in order to get 0 and add 17.We then assign this new array to "Times". Line 5

of the code assigns the 3rd column of the spreadsheet to the name "Temp" which represents the temperature column. Similarly, we do the same with the 4th column and assign it to "humidity". Since we are analysing how temperature and humidity changes with time, we create tables with Time and temperature, and also time and humidity in order to compare them directly with each other.

3.1.3 Extract function

Now to extract data from each day we must create a function that will do this effectively.

Listing 3.3: Function daytemp

```
1 function daytemp = extractTemp(TimeAndTemp,day)
2 C =(TimeAndTemp(:,1)>=day*86400)&(TimeAndTemp(:,1)<(day
      +1)*86400);
3 daytemp = TimeAndTemp(C,:);
4 daytemp(:,1)=daytemp(:,1)-day*86400;
```

By using the function term, we create a new array and assign it to the new fucntion "extractTemp" which extracts data from the TimeAndTemp array for a day. Now in order to make sure it is an accurate extraction of data, we give the code a set of conditions, denoted by the letter "C" in line 2. Firstly, the first column of the TimeandTemp array must be greater than or equal to the day number multiplied by 86400 (which is the amount of seconds in a day) so that we only analyse the time and temperature for that day and not for the other day which would be less than 86400*day. The other condition states that the Time column must be less than the day+1 multiplied by 86400, this is done to ensure that the next days time and temperature isn't analysed. Once the conditions have been established, we assign day temp to the combined array between TimeandDay and the conditions C which gives an array of temperature and time for a specific day. We then subtract day*86400 from the time column in order to get readings that start from zero.

3.1.4 For loop for each day

Listing 3.4: For loop for everyday

```
for day = 0:365
daytemp = extractTemp(TimeAndTemp,day)
plot(daytemp(:,1),daytemp(:,2))
hold on
```

To obtain data for each of the 365 days, we must create a for loop in order to make the code short and concise and allow corrections and easier manipulation. So we create a for loop from 0 to 365 and use the function we created from /cite code 3 in order to get data for time and temperature for each day. From that we can plot a a temperature time graph for each day, as well as plot all days on a single plot.

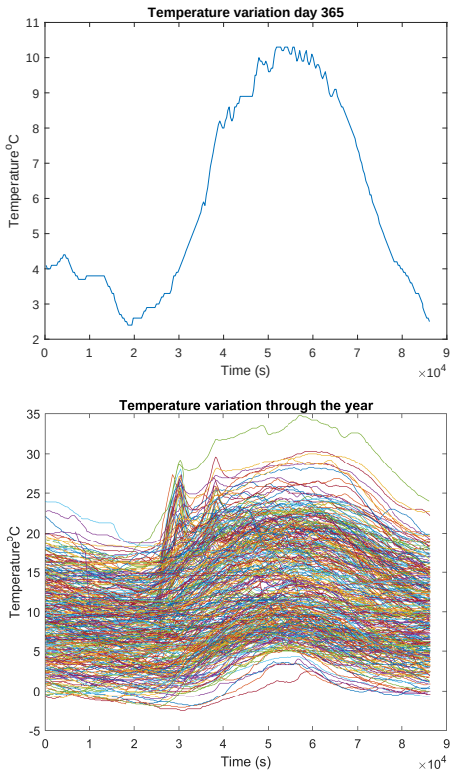

Figure 3.1:

After running the code with the hold on command we get all of the days plotted onto one day, if we remove the hold on we get a plot for one

specific day, in this case day 365. From the graphs plotted there isn't a lot of information that can be extracted so we have to develop our code further to fix this.

3.1.5 Smoothing procedure, multiple least squares regression

What we can see from the graphs is that they are time series graphs, this means we must use multiple regression in order to work out estimate coefficients needed to smooth the graph and create a predicted curve on the graph.Multiple regression will allow us to predict an output using correlations we get from the inputs. To do this we must first model the graph in order to use multiple regression to find the coefficients. So the equation of the model since its a time series graph is:

$$y = \beta_0 + \beta_1 \cos(x) + \beta_2 \sin(x) + \beta_3 \cos(x) + \beta_4 \sin(x)$$

$$T(t) = a + b\cos(k_1 t) + c\sin(k_1 t) + d\cos(k_2 t) + e\sin(k_2 t)$$

What we do now is implement the derived formula of multiple regression into Matlab to find the coefficients and use them to create a temperature time curve.

Listing 3.5: Multiple least squares regression

```
k1 = 2*pi./(24*60*60);
k2 = 2*pi./(12*60*60);
X = [ones(size(daytemp(:,1))) cos(daytemp(:,1)*(k1))
    sin(daytemp(:,1)*(k1)) cos(daytemp(:,1)*(k2)) sin(
    daytemp(:,1)*(k2)) ] ;

Beta = inv(X'*X)*X'*daytemp(:,2); % derived formula

SSE = (daytemp(:,2)-X*Beta)'*(daytemp(:,1)-X*Beta); %
    sum squared error
SStot = sum((daytemp(:,2)-mean(daytemp(:,2))).^2); %
    total variance in the system

R2 = 1-SSE/SStot; %how good our model fits

```

```
12 plot(daytemp(:,1),X*Beta,'.')
13 hold on
14 plot(daytemp(:,1),daytemp(:,2))
15
16 legend('Real reading','Least squares fitting')
```

Initially, we define the values of k1 and k2 which represent the time period, k1 is the 24 hour cycle and k2 is the 12 hour daily cycle. Secondly we create a matrix as seen in chapter 1 under multiple regression. We have the first column which is just ones that is the same size as daytemp(:,1) (the dependent variable time) and the other 4 co-variants. This whole matrix is assigned to the variable X. In line 5 of the code, we assign the term Beta to the equation we derived in chapter 1.

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

What this code does is it find the values of the 5 coefficients in the model formula. In line 7 of the code we calculate the sum squared error or the residual square error which was derived in chapter 1 again with the formula:

$$RSS = (y^T - \widehat{\beta^T} X^T)(y - X\hat{\beta})$$

In the next line we calculate the total variance in the system and in line 10 we see how good our model fits.
Now that we have the 5 coefficients of the model we are able to plot our new predicted y variable against the time. To see how the new predicted plot looks we plot both the real graph and the predicted graph together using the hold on code from line 13.

Listing 3.6: Calculating yearly coefficients

```
1
2 function Gamma = SFit(daytemp)
3
4 k1 = 2*pi./(24*60*60);
5 k2 = 2*pi./(12*60*60);
6
7
8 X = [ones(size(daytemp(:,1))) cos(daytemp(:,1)*(k1))
    sin(daytemp(:,1)*(k1)) cos(daytemp(:,1)*(k2)) sin(
    daytemp(:,1)*(k2)) ];
```

```
9
10
11 Beta = inv(X'*X)*X'*daytemp(:,2); % derived formula
12
13
14
15 Gamma(1)=Beta(1);
16 [Gamma(3),Gamma(2)]=cart2pol(Beta(2),Beta(3));
17 [Gamma(5),Gamma(4)]=cart2pol(Beta(4),Beta(5));
18
19 delta1 = Gamma(2);
20 phi1 = Gamma(3);
21 delta2 = Gamma(4);
22 phi2 = Gamma(5);
23
24 end
```

To calculate the yearly coefficients, we create a new function labelled as Sfit. Similarly to the least squares fitting is listing 2.5, the first 11 lines are the same, however we create a new array called Gamma and assign it to the beta coefficients. Gamma(1) is equal to beta(1) however with the other gammas, the cart2pol function needs to be used in order to convert the Cartesian beta into polar coordinates for Gamma. This will allow us to determine the coefficients of delta1, phi1, delta2 and phi2.

Chapter 4

Understanding Temperature and humidity

In order to conduct effective data analysis, we must first understand the fundamentals of temperature and humidity.

4.1 Day and night cycles of temperature

In nature, all things have daily patterns because they change throughout the day. This is known as diurnal patterns [4]. Diurnal in most cases often refers the the change in temperature between the highest temperature throughout the day and the lowest temperature at night.

The process of reaching the highest temperature is a gradual one [4]. It begins each morning when the sun rises and its ray hit the earths surface. This radiation from the sun directly heats the earths surface. But since the land has a high heat capacity, the temperature doesn't increase straight away and will only start to increase once its absorbed a certain amount of energy. As the surface warms, it heats a thin layer of air through the process of conduction and this in turn heats a column of cool air above it.

As the earth spins, the sun continues to move across the sky, where at high noon it reaches its peak height where its sunlight is at its most concentrated strength. However since the ground must store heat in order to radiate energy, the surrounding areas maximum temperature hasn't been reached yet.

This is called diurnal temperature lag [4]. The daily high temperature occurs when the incoming solar radiation is equal to the outgoing radiation. After noon, the surface temperature gradually begins to decrease as the suns intensity continually declines. Once the sun is no longer providing solar radiation, the minimum temperature is reached providing there is more heat energy being lost than incoming heat energy.

The range of the diurnal temperature can differ depending on a number of conditions. [4]

- **Clouds**. Clouds affect the diurnal temperature as they shield the surface from the suns solar radiation but also trap heat as well. For example on a cloudy day, the surface is shielded from solar radiation because these solar waves are reflected back out to space meaning there is less incoming heat which therefore causes a decrease in the diurnal temperature range. However on a cloudy night, heat gained throughout the day is trapped near the ground which allows the air temperature to remain constant rather than cool.

- **Humidity**. Humidity is the amount of water vapor in the air. Similar to clouds, humidity (water vapour) is good at absorbing and releasing solar radiation this reduces the amount of radiation reaching the earths surface. This causes the daily high to be lower in humid environments than in dry environments. This affects the diurnal temperature range.
  .

Chapter 5

Analysis of collected data

After initially computing and plotting all days onto one graph we get the following data.

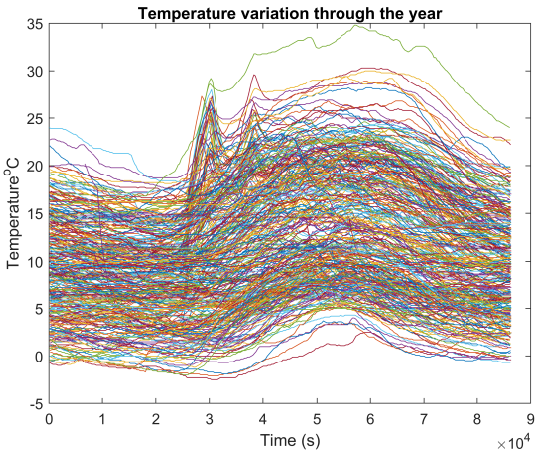

Figure 5.1: Everyday temperature variation

From the following graph we are not able to see that much since there is too much information that's been plotted and we are unable to see how the temperature varies throughout each day.However what we do see is that in the seconds of 3000 and 4000, there is a spike suggesting that some days do not follow the normal diurnal patterns.. In order to better observe the variation, we must plot for an individual day.

5.1 Temperature Variation throughout the day (Visual analysis)

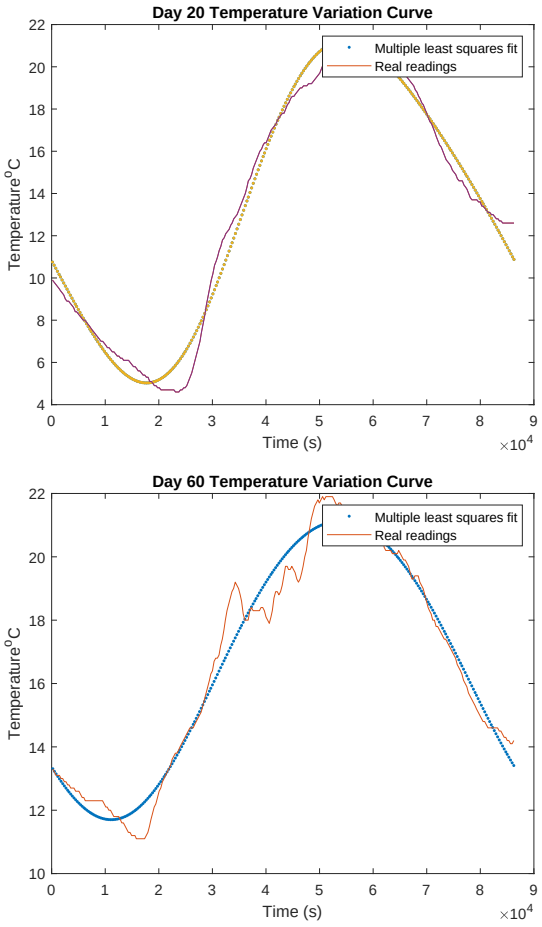

Figure 5.2: Temperature Variation Day 20 and 60

By plotting day 20 and 60 with the multiple least squares fitting for both, we can firstly see that the temperature for both days follows a general trend where the temperature typically begins to drop from 0 seconds (Midnight) continuously until it begins to increase exponentially until it drops again. This suggests that through the night, temperature will decrease slowly, but

as soon as the day begins, temperature will rise throughout the day where it typically will reach its maximum temperature throughout the middle of the day, in the case of day 60, the maximum temperature will be reached at 53594 seconds (14:53:14PM) and for the case of day 20, it maximum temperature is at 54548 seconds (15:09:08PM). What this shows is that for these days, their maximum temperature will be reached in the afternoon, from this we can make the assumption that these days are in the summer season. We see that for both days the temperature remains at its highest for a couple hours before it exponentially decreases as the sun begins to set. From the following plots we can overall see that the results obtained a accurate with slight noise with day 60, but overall constant and regular. Overall we can see that the data follows the normal diurnal temperature patterns. Throughout the year, most days follow a similar trend to the days in figure 5.2, However there are some days that have produced some irregular data.

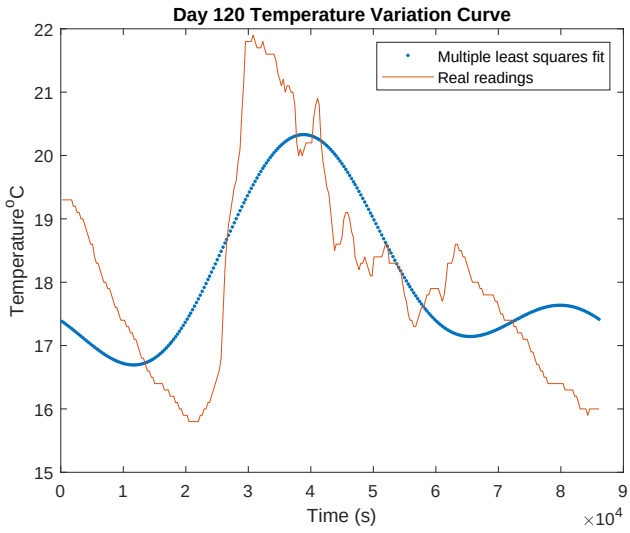

Figure 5.3: Day 120 temperature variation

Compared to Day 20 and 60, day 120 is irregular with its trend. Firstly we are able to see that there is a lot of noise in the data , we see that with the real readings, the temperature is not as constant as it should be. We can see with the multiple least squares plot, there is a lot of anomalies and noise in the system. Secondly, we see with the multiple squares fitting that a lot

of the data doesn't align with the predicted curve suggesting that the data collected for this day is inconsistent.This inconsistent and irregular pattern may be potentially die to external factors such as clouds, wind or rain.

5.2 Plotting coefficients

In order to more accurately explain the data we must calculate rewrite the formula to calculate other coefficients such as the mean temperature, the amplitude and the phase. We model these coefficient using the following equation.

$$T_d(t) = \bar{T}+\Delta\tilde{T}_1\cos(\frac{2\pi}{\text{Day}}t-\phi_1)+\Delta\tilde{T}_1\sin(\frac{2\pi}{\text{Day}}t-\phi_1)+\Delta\tilde{T}_2\cos(\frac{2\pi}{\text{Day}/2}t-\phi_2)+\Delta\tilde{T}_2\sin(\frac{2\pi}{\text{Day}/2}t-\phi_2)$$

Where $\bar{T}$ is the mean temperature, $\Delta\tilde{T}_1$ is the daily harmonic amplitude, ϕ_1 is the daily harmonic phase, $\Delta\tilde{T}_2$ is the semi-daily harmonic amplitude and ϕ_1 is the semi-daily harmonic phase. [7] By using multiple least squares and converting the Cartesian equation to polynomial, we are able to calculate these coefficients. These coefficients will represent different data such as the mean temperature, the amplitude and the phase.

5.3 Mean temperature variation

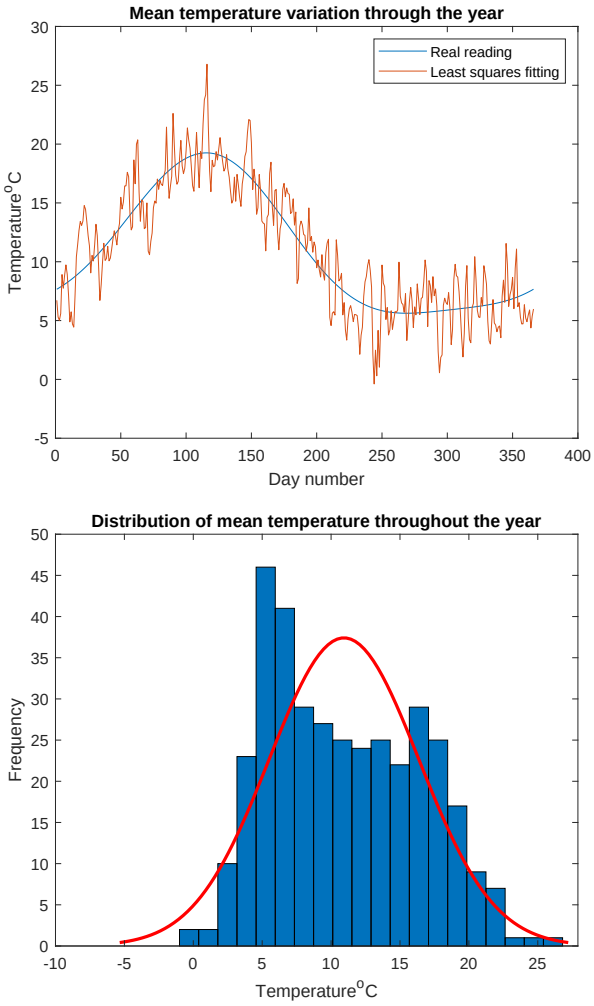

Figure 5.4: Yearly mean temperature variation and distribution

Skewness	Kurtosis	Standard deviation	Variance
0.2979	2.1457	5.5248	29.4284

Table 5.1: Calculated information for mean temperature

After plotting the mean temperature throughout the year, visually we are able to see that there is a constant trend, where the mean temperature maintains a follows a constant pattern throughout the year with some noise as shown by the original fit. In order to gain further information about how the mean temperature data is distributed we must plot a histogram.

From the distribution graph at the bottom of figure 5.4 we're able to see that the distribution of the mean temperature is fairly constant from the bell shape. We can see the data is fits the Gaussian distribution curve quite well with the exception of two of the bands which do not fit the curve. By calculating the skewness of the sample we get a value of 0.2979 suggesting that the mean temperature throughout the year is nearly symmetrical. As well as this, the kurtosis calculated is equal to 2.1457. This shows a platykurtic kurtosis which means most data points are present with high proximity to the mean [6]. The standard deviation calculated for the mean temperature is equal to 5.4248 which tells us that there is some spread within this dataset but not a large one. We also get a variance of 29.4284 again suggesting there is some spread in the model. Overall we can say from the data gathered that the mean temperature throughout the year roughly follows a symmetrical distribution with some spread.

5.4 Daily harmonic amplitude temperature

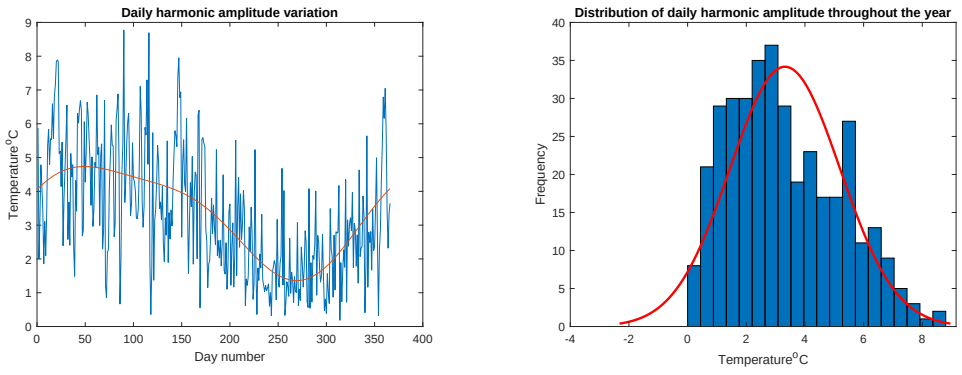

Figure 5.5: Daily harmonic amplitude variation and distribution

Skewness	Kurtosis	Standard deviation	Variance
0.4835	2.4381	1.8796	3.5329

Table 5.2: Calculated information for the daily harmonic amplitude variation

With the daily harmonic amplitude, we can see visually there is a uniform trend however, there is a lot of noise in the system, since a lot of the points are quite far away from the least squares fitting.

By looking at the distribution curve alone we can see that most points fit the curve. By calculating the skewness we get a value of 0.4835, the value is very small which suggests that the distribution of the daily harmonic amplitude is very close to being symmetrical. Similarly to table 5.1, the kurtosis is calculated to be 2.4381, what this shows is that the data is close to being a normal distribution but since the kurtoisis isn't 3, it is a platykurtic (short tailed distribution) meaning most of the data points are in close proximity with the mean hut not all like with a normal distribution. [6]. The standard deviation is equal to 1.8796 which tells us there is very little spread in the data. What we can see from this is that the distribution of the daily harmonic amplitude is very symmetrical, more so than in table 4.1 where the variance for that system is equal to 29.4284 compared to the daily amplitudes variance of 3.5329.

5.5 Daily harmonic phase temperature

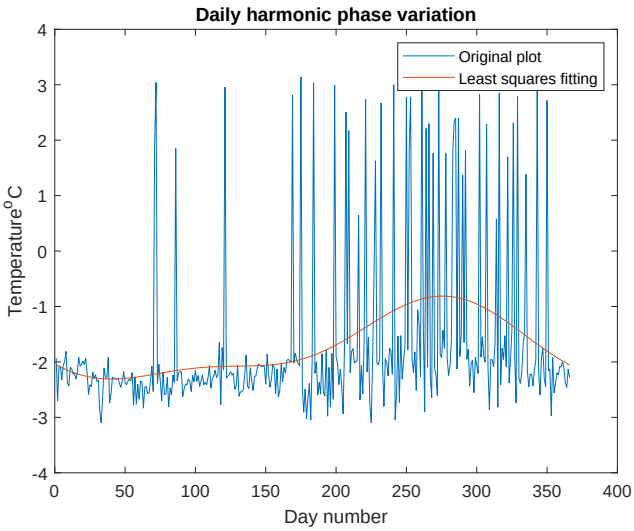

Figure 5.6: Distribution of daily phase amplitude

When plotting the daily harmonic phase variation, we can straight away see that there is a lot of noise in the system due to a lot of points being very far away from the least squares fitting. What this means is that a lot of the data points are meaningless and potentially insignificant. We do see a trend, however there are many spikes where the temperature will increase drastically. To understand the spread of the data and how its distributed we will plot a histogram with a Gaussian distribution curve once again.

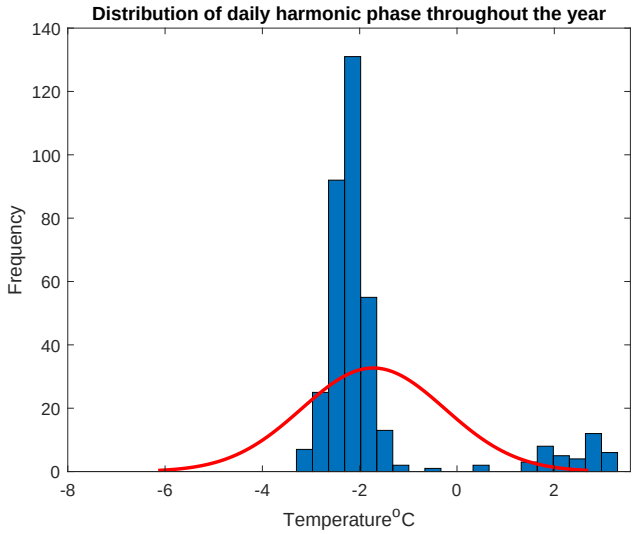

Figure 5.7: Distribution of daily phase amplitude

Skewness	Kurtosis	Standard deviation	Variance
2.3635	7.1966	1.4734	2.1710

Table 5.3: Calculated information for the daily harmonic phase temperature

After plotting the distribution curve and histogram, we see unlike the previous curves there is a much smaller one with a significantly smaller peak. Firstly, by calculating the the skewness of the data, we get a value of 2.3635. What this tells us about the data is that it is extremely positively skewed indicating that it is a type of distribution where the mean mode and median of the distribution are not negative or zero but rather positive. This is not desirable since it can lead to misleading results making the data unreliable. The kurtoisis of the data is equal to 7.1966, which indicates a Leptokurtic distribution (heavy tailed distribution). This type of distribution implies that there is a greater chance of having outliers in the data. Because of the large kurtoisis, we understand why there was a lot of noise in figure 3.8. Although the kurtosis and skweness is significantly large, the variance and standard deviation remain relatively low with values of 2.1710 and 1.4734. This coveys that there is a small spread between the mean.

5.6 Semi-daily harmonic amplitude temperature

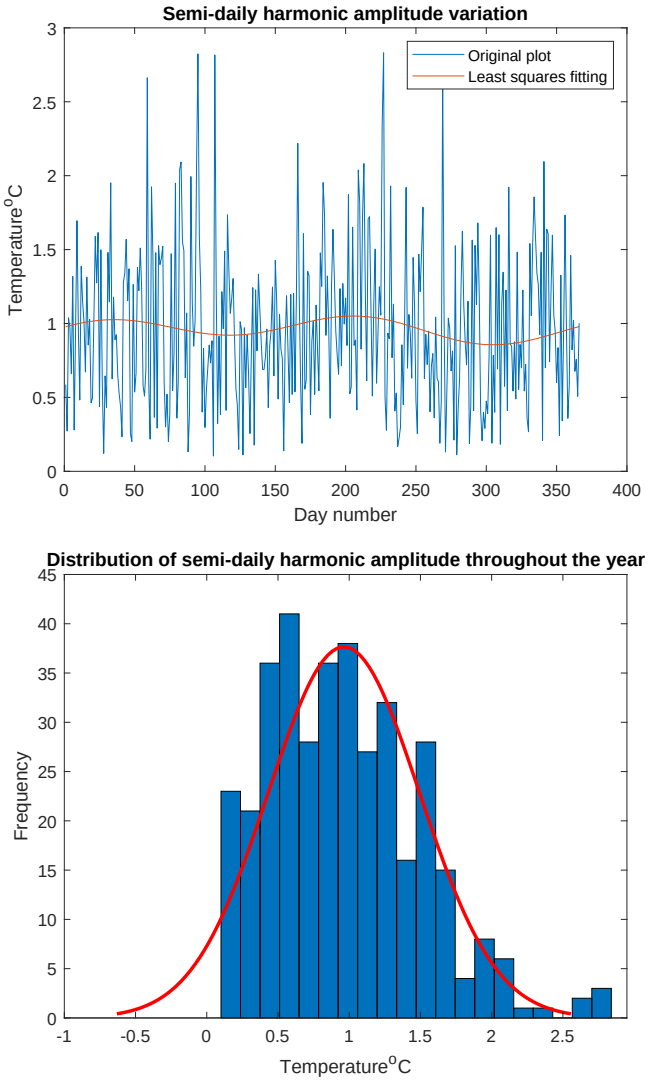

Figure 5.8: Semi-daily amplitude

Skewness	Kurtosis	Standard deviation	Variance
0.7245	3.5317	0.5312	0.2822

Table 5.4: Calculated information for the semi-daily harmonic amplitude temperature

By looking at the graphs, we visually see a lot of noise within the variation throughout the year. We see this as there are a lot of data points that are not within the least squares fitting which could mean there are outliers in the data. However, when we look at the histogram/distribution of the data points we see the data points fit the Gaussian distribution well. The skewness of the data point is equal to 0.7245, this displays a nearly symmetrical distribution that is slightly positively skewed. This suggests the results are ideal and reliable with a small number of misleading results. Calculating the kurtosis of the data gives us a value of 3.5317. A normal distribution has a kurtosis of 3, from this, we are able to see that the data follows the normal distribution with a slight excess of 0.5317. The standard deviation of the data is 0.5312 and the variation is 0.2822. This means there is little spread within this data.

5.7 Semi-daily harmonic phase temperature

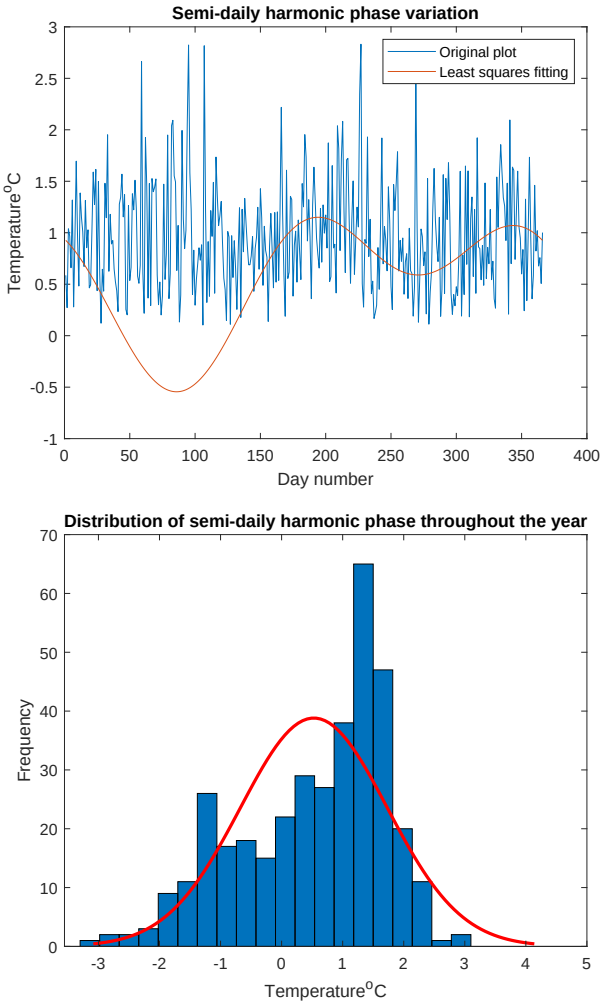

Figure 5.9: Semi-daily phase

Skewness	Kurtosis	Standard deviation	Variance
-0.6289	2.5348	1.2037	1.4488

Table 5.5: Calculated information for the semi-daily harmonic phase temperature

Calculating the skewness of the data gives a value of -0.6289. From this we see that the distribution is nearly symmetrical with it being slightly negatively skewed.The kurtosis is equal to 2.5348, this means that there is a short tailed distribution within the data, however since the kurtosis of normal distribution is equal to 3, we can see that the data is nearly a normal distribution because the difference is 0.4652.The standard deviation is 1.2037 and the variance is 1.4488,these values mean that there is little spread within the data and a little dispersion relative to the mean.

Chapter 6

Relationship between temperature and humidity

6.1 Mean temperature and humidity

To determine the type of relationship between temperature and humidity, we must use calculate the correlation coefficient in order to find the correlation between the two.

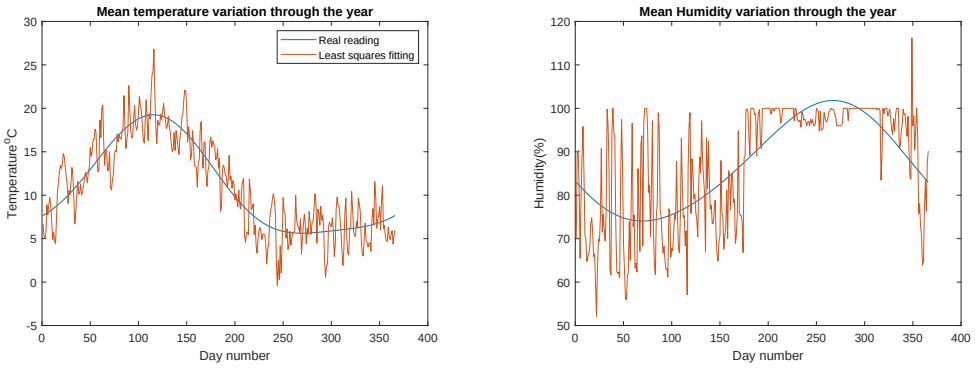

Figure 6.1: Mean temperature and humidity through the year

Mean temperature	Mean humidity
1.0000	-0.5773
-0.5773	1.0000

Table 6.1: Correlation coefficients for mean temperature and humidity

By looking at figure 6.1, we see that the mean temperature and the mean humidity follow a different trajectory. We see temperature increase throughout the summer months in particular with day 100 and day 110, where as with humidity, it remains relatively low. This is can be proven with the correlation matrix in table 6.1 where a correlation of -0.5773 is found. This shows a negative correlation meaning that as one increases the other will decrease, in this case as temperature increases, humidity decreases. However since the correlation coefficient is equal to -0.5773 its not a strong negative correlation. This may be due to the amount of noise we see as humidity varies throughout the year.

6.2 Daily harmonic amplitude

Next we will be comparing the daily harmonic amplitude of temperature and humidity.

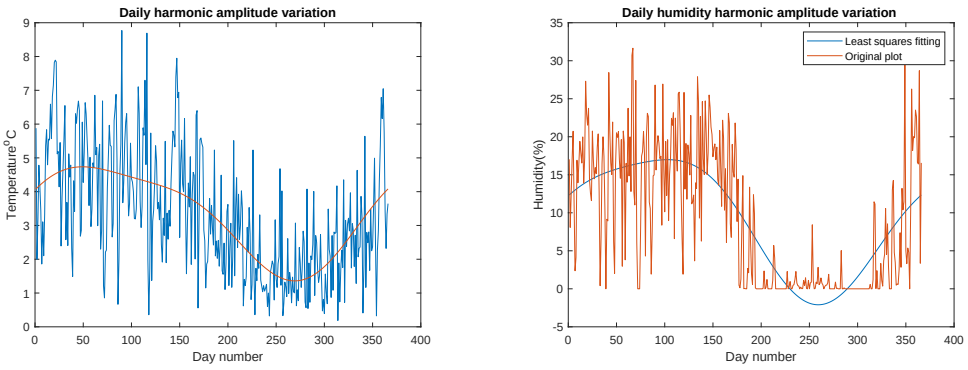

Figure 6.2: Daily harmonic amplitude variation

Temperature daily amplitude	Humidity daily amplitude
1.0000	0.7076
0.7076	1.0000

Table 6.2: Correlation coefficients for daily amplitude variation for temperature and humidity

Visually, we can see from figure 6.2 that both the temperature and humidity daily harmonic amplitude change similarly throughout the year. This is further backed up by the correlation coefficient which is 0.7076. This shows a strong correlation, meaning as one value increases, the other also increases as well. What we can deduce from this is when temperature hits its peak temperature humidity will hit its lowest humidity since the amplitude increases for both.

6.3 Daily harmonic phase

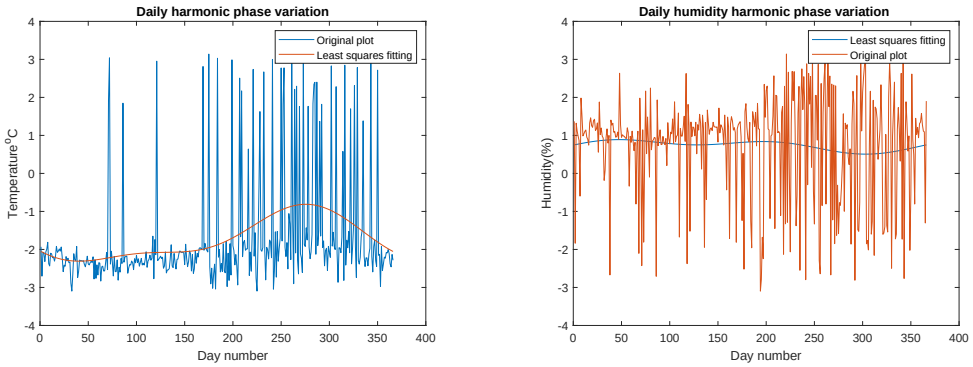

Figure 6.3: Daily harmonic phase variation

Temperature daily phase	Humidity daily phase
1.0000	-0.1507
-0.1507	1.0000

Table 6.3: Correlation coefficients for daily phase variation for temperature and humidity

By looking at the two graphs in figure 6.3, its clear that there is any significant correlation. From table 6.3, we see that the correlation coefficient is equal to -0.1507, which indicates there is a very slight and insignificant negative correlation. So from this we can deduce that there is no significant correlation between the daily harmonic phase of temperature and humidity.

6.4 Semi-daily harmonic amplitude

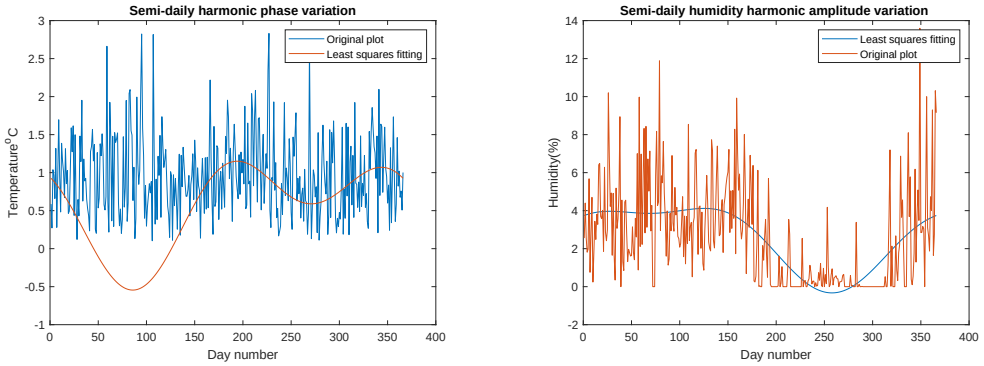

Figure 6.4: Semi-daily harmonic amplitude variation

Temperature Semi-daily amplitude	Humidity Semi-daily amplitude
1.0000	0.1287
0.1287	1.0000

Table 6.4: Correlation coefficients for Semi-daily amplitude variation for temperature and humidity

Looking at figure 6.4 like the daily harmonic phase, shows little to no correlation visually. Table 6.4 shows a correlation of 0.1287 suggesting there is an insignificant positive correlation, meaning there is so significant relationship between the semi-daily harmonic amplitude of temperature and humidity.

6.5 Semi-daily harmonic phase

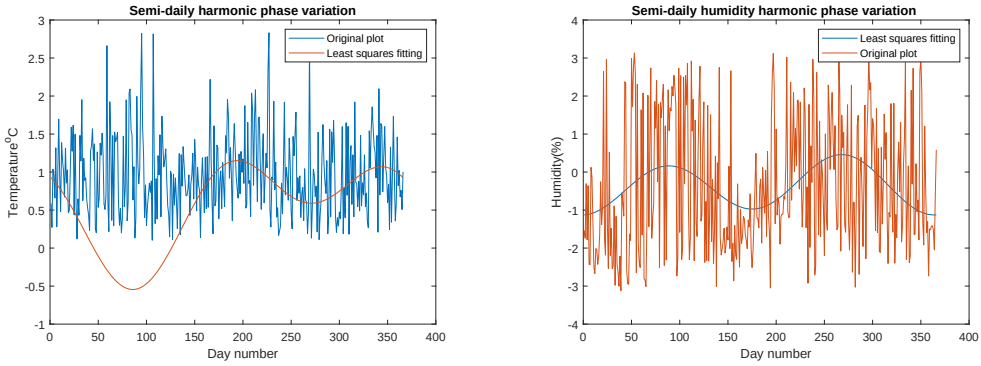

Figure 6.5: Semi-daily harmonic phase variation

Temperature Semi-daily phase	Humidity Semi-daily phase
1.0000	-0.1474
-0.1474	1.0000

Table 6.5: Correlation coefficients for Semi-daily phase variation for temperature and humidity

Looking at figure 6.5, we see that the semi-daily phase of temperature and humidity varies differently. By looking at table 6.5, the correlation coefficient is equal to -0.1474, this exhibits a very weak and insignificant correlation which suggest there is no significant correlation between the two.

Overall, we see that there is a significant negative correlation between the mean temperature and mean humidity as well as a very significant positive correlation between the daily harmonic amplitude for both. This suggests the data follows the rules of humidity and temperature where, as the temperature increases, the humidity decreases.

Conclusion

After extracting and analysing all the data from the weather station, we can conclude that overall the quality of the data was accurate. Firstly, when plotting the temperature variation throughout a day, we could see that on most days, the data follows normal day and night patterns with the temperature starting of low and continuing to decrease until increasing again as the sun comes up and reaching its peak around midday where it then decreases again. Since it follows these diurnal temperature patterns, we can presume that the quality of data is good and adequate.

However, when we plotted the temperature variation of every single day onto one plot (Figure 5.1 from chapter 5), we were able to see that some days didn't follow the normal day and night patterns as well as this, day 120 (Figure 5.3) had an irregular trend with a lot of noise in the data. From this we can deduce that the data although having most days following a normal day and night pattern had a small amount of outliers which suggests that the data is accurate for most days but not completely.

Secondly, when plotting the distribution of mean temperature, daily harmonic amplitude,daily harmonic phase,semi-daily harmonic amplitude and semi-daily harmonic phase throughout the year, we came up with the following results. After plotting the mean temperature variation curve, we saw that the temperature followed a smooth trend that was constant with a very small amount of noise. This suggests a good quality of data. As well as this, the distribution was also constant with a skewness of 0.2979 which meant the distribution was nearly symmetrical. This lead to the assumption that the quality of these results were accurate.However,with the case of daily harmonic phase temperature, we saw a skewness of 2.3635. This shows an extremely positive skew which often leads to undesirable results, because of

this we can deduce that the results of the daily phase amplitude may not be accurate. .

Finally by comparing the relationship between temperature and humidity through Pearson's correlation coefficient, we were able to determine that there was a negative correlation of -0.5773 between the two. This meant that as the temperature increased, the humidity decreased. This suggests the quality of data is good and accurate since when temperature is at its highest usually, the humidity will be at its lowest. For example, throughout the night, humidity remains high in contrast to temperature that remains low.

From this we can conclude that overall the data received from this weather station is accurate, mainly due to how the data shows normal day and night patterns as well as correctly showing the relationship between temperature and humidity.

Understanding the data and determining the quality of it is a very important aspect of data analysis. Having inaccurate data can lead to increased risk of developing incorrect assumptions and ideas, which in the real world can lead to disastrous outcomes. It is because of this that accurately sorting out the data effectively and determining the quality of data is essential when analysing data. For future research, comparing this data to other years would be ideal in order to see how the quality differs between the two as well as to see if there are similar/different trends that happen in different years. Analysing data from other years would give more validity to the data as well as make it more reliable. As well as this, using data from other weather stations would be a good idea to show how different locations are affected by temperature and humidity and how the location can impact the quality of data. For example, in an urban area the maximum daily temperature may be higher than in a rural area and if not there may be external factors that affect the accuracy of the data.

To conclude this project, We've learnt that the accurate extraction of data is important in order to gain valid information. From the data gathered we can see that the quality of the data from this weather station is overall accurate with a small number of anomalies that tarnish the accuracy slightly.

Appendix

Listing 6.1: Plotting coefficients of temperature (mean,amplitude and phase)

```
clc ;

which extract
AprMayJun2019 = readtable("Outdoor_AprMayJun_2019.xls")
    ;
JulAugSep2019 = readtable("Outdoor_JulAugSep_2019.xls")
    ;
OctNovDec2019 = readtable("Outdoor_OctNovDec_2019.xls")
    ;
JanFebMar2020 = readtable("Outdoor_JanFebMar_2020.xls")
    ;

yeardata = [AprMayJun2019;JulAugSep2019;OctNovDec2019;
    JanFebMar2020]

Times =table2array(yeardata(:,1));
Times = Times-Times(1)+17;
Temp =table2array(yeardata(:,3));
Humidity = table2array(yeardata(:,4));
TimeAndTemp = [Times,Temp];
TimeAndHumidity = [Times,Humidity]

for day = 0:365
daytemp = extractTemp(TimeAndTemp,day)
dayhumid = extractHumid(TimeAndHumidity,day)


```

```
Gamma = SFit(daytemp)
SGamma(day+1,1:5) = Gamma
end
plot(SGamma(:,1))
hold on
plot(SGamma(:,2))
plot(SGamma(:,3))
plot(SGamma(:,4))
plot(SGamma(:,5))

hold off
legend('Mean temperature','Daily harmonic amplitude','Daily harmonic phase','Semi daily harmonic amplitude','Semi daily harmonic phase')
figure(2)
histogram(SGamma(:,2))
figure(3)
histogram(SGamma(:,3))
figure(5)
histfit(SGamma(:,1))
```

Listing 6.2: Smoothing procedure for temperature coefficients

```
k1 =  2*pi./365;
k2 =  2*pi./182.5;
Xvalues = (1:366)';

XMeanTemp = [ones(size(Xvalues)) cos(Xvalues*(k1)) sin(Xvalues*(k1)) cos(Xvalues*(k2)) sin(Xvalues*(k2)) ];

BetaMeanTemp= inv(XMeanTemp'*XMeanTemp)*XMeanTemp'*SGamma(:,1);
Betadailyharmonicamplitude = inv(XMeanTemp'*XMeanTemp)*XMeanTemp'*SGamma(:,2);
betadailyharmonicphase = inv(XMeanTemp'*XMeanTemp)*XMeanTemp'*SGamma(:,3);
Betasemidailyharmonicamplitude = inv(XMeanTemp'*XMeanTemp)*XMeanTemp'*SGamma(:,4);
```

```
betasemidailyharmonicphase = inv(XMeanTemp'*XMeanTemp)*
    XMeanTemp'*SGamma(:,5);
% plot((Xvalues),(XMeanTemp*BetaMeanTemp))
%
% hold on
%
% plot(Xvalues,SGamma(:,1))
% hold off
% legend('Real reading','Least squares fitting' )
% xlabel('Day number')
% ylabel('Temperature^{o}C')
% title('Mean temperature variation through the year')
%

%
% var(SGamma(:,1))
% skewness(SGamma(:,1))
% kurtosis(SGamma(:,1))
% std(SGamma(:,1))

plot(Xvalues,SGamma(:,4))
hold on
plot((Xvalues),(XMeanTemp*betasemidailyharmonicphase))
hold off
legend('Original plot','Least squares fitting')
xlabel('Day number')
ylabel('Temperature^{o}C')
title('Semi-daily harmonic phase variation')
```

Listing 6.3: Smoothing procedure for humidity coefficients

```
k1 =  2*pi./365;
k2 =  2*pi./182.5;
Xvalues = (1:366)';

XDays = [ones(size(Xvalues)) cos(Xvalues*(k1)) sin(
    Xvalues*(k1)) cos(Xvalues*(k2)) sin(Xvalues*(k2)) ]
    ;

```

```
7  BetaMeanHumidity= inv(XDays'*XDays)*XDays'*SZeta(:,1);
8  betadailyharmonicamplitudeHumid = inv(XDays'*XDays)*
       XDays'*SZeta(:,2);
9  Betadailyharmonicphasehumid = inv(XDays'*XDays)*XDays'*
       SZeta(:,3);
10 BetaSemiDailyHarmonicAmplitudeHumid = inv(XDays'*XDays)
       *XDays'*SZeta(:,4);
11 BetaSemiDailyHarmonicPhaseHumid = inv(XDays'*XDays)*
       XDays'*SZeta(:,5);
12
13 plot((Xvalues),(XDays*BetaSemiDailyHarmonicPhaseHumid))
14
15 hold on
16
17 plot(Xvalues,SZeta(:,5))
18 hold off
19 legend('Least squares fitting','Original plot')
20 xlabel('Day number')
21 ylabel('Humidity(%)')
22 title('Mean Humidity variation through the year')
```

Listing 6.4: Functions to calculate information about data

```
1
2  skewness(SGamma(:,5))
3  kurtosis(SGamma(:,5))
4  std(SGamma(:,5))
5  var(SGamma(:,5))
6  corrcoef(SGamma(:,1),SZeta(:,1))
7  corrcoef(SZeta(:,2),SGamma(:,2))
8  corrcoef(SZeta(:,3),SGamma(:,3))
9  corrcoef(SZeta(:,4),SGamma(:,4))
10 corrcoef(SZeta(:,5),SGamma(:,5))
```

Bibliography

[1] Randal J. Barnes. Matrix differentiation - department of atmospheric sciences. (https://atmos.washington.edu/ dennis/MatrixCalculus.pdff).

[2] Code Emporium. Linear regression and multiple regression. (https://www.youtube.com/watch?v=$K_E H2abOp00t = 304sab_c hannel = CodeEmporium$), 2022.

[3] Mathisfun. Linearsquaresregression. (https://www.mathsisfun.com/data/least-squares-regression.html), November 2021.

[4] Tiffany Means. How the atmosphere heats and cools during a 24-hour period. (https://www.thoughtco.com/diurnal-temperature-range-3444244), Jul 2019.

[5] Kody Powell. Derivation of multiple least squares for fitting models with multiple inputs. (https://www.youtube.com/watch?v=oLwLFdy8sv4amp;t=444samp;ab$_c hannel = KodyPowell$), $Feb2017$.

[6] suvarna3. Skewness and kurtosis: Shape of data: Skewness and kurtosis. (https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/), May 2021.

[7] Kai Wang, Yuguo Li, Yi Wang, and Xinyan Yang. On the asymmetry of the urban daily air temperature cycle. *Journal of Geophysical Research: Atmospheres*, 122(11):5625–5635, 2017.