|  |  |
| --- | --- |
|  |  |
| Predictors of evaporation in Melbourne |  |
|  | Kristen Osborne |
|  |  |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Executive Summary Understanding and predicting evaporation is key to the effective management of reservoirs and to ensuring Melbourne’s water supply. This report explores a range of temporal and metrological factors to determine their influence on evaporation and their effectiveness in predicting it.  An iterative approach was used to determine which of these difference factors were significant and to construct the simplest model containing all significant predictors. This model has then been used to predict evaporation in Melbourne under a range of possible scenarios.  Key findings:   * Daily evaporation levels appear to have relationships with several key factors, including minimum and maximum daily temperature, the month of the year, and the relative humidity at 9am. * The simplest model to predict evaporation uses month, minimum daily temperature, 9am relative humidity, and the interaction between month and humidity. * However, this model is less accurate at higher levels of predicted evaporation. | |  |

### Methods

**Initial analysis**

The first step in modeling the effects of certain factors on evaporation is to examine the individual relationships present. Code for this section is presented in segment 3 of the appendix. Figure 1 shows the average monthly evaporation for the 2018-19 period.

**A screenshot of a cell phone

Description automatically generated**

This figure shows there is a clear relationship between the month and the average evaporation, with the summer months having a much higher evaporation than winter months. The lowest evaporation was during June (less than 2.5mm) and the highest was during January (over 7.5mm).

The relationship between the day of the week and evaporation is unclear, as seen in figure 2.

A screenshot of a cell phone

Description automatically generated

On average, more evaporation occurred on Saturdays (5.94) and less occurred on Thursdays (4.68). However, the difference is only 1.26mm and there is no clear rationale as to why this may occur.

Figure 3 compares the maximum daily temperature with evaporation and shows that evaporation is generally higher when the max temperature is greater on any given day. Evaporation is also more variable at higher max temperatures.

**A close up of a map

Description automatically generated**

Figure 4 shows that evaporation has a similar relationship with minimum temperature as it does with the maximum.

A close up of a map

Description automatically generated

Similarly, more evaporation generally occurs on days when the minimum temperature is higher..

Finally, figure 5 compares daily evaporation with the relative humidity at 9am and shows a negative relationship between the two.

A screenshot of a computer

Description automatically generated

This means that daily evaporation is generally lower when humidity levels at 9am are higher.

**Model selection**

A linear model was first created using all the predictors listed above, as well as a variable representing the interaction between the month and the humidity. This new factor was included to determine if humidity’s effect is different depending on the month. Summary statistics relating to this initial model are presented in table 1. Code for all model selection is present in section 4 within the appendix.

**Table 1. Coefficients and significance values in model 1**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Coefficient** | **P-value** |
| Intercept | 6.03 | <0.001\* |
| Maximum temperature | 0.02 | 0.56 |
| Minimum temperature | 0.36 | <0.001\* |
| 9am relative humidity | -0.08 | <0.001\* |
| Month | various | <0.001\* |
| Day of the week | various | 0.10 |
| Interaction between month and 9am relative humidity | various | <0.001\* |

Note: significance was tested using the linear model for quantitative variables and using an ANOVA for qualitative variables (month, day of the week and month-relative humidity interaction) \*p values below 0.05 were considered significant.

Next, the predictor with the highest p-value (maximum temperature) was removed and the model was re-computed. This did not result in the significance of any predictors changing meaningfully, as seen in table 2.

**Table 2. Coefficients and significance values in model 2**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Coefficient** | **P-value** |
| Intercept | 6.03 | <0.001\* |
| Minimum temperature | 0.37 | <0.001\* |
| 9am relative humidity | -0.09 | <0.001\* |
| Month | various | <0.001\* |
| Day of the week | various | 0.10 |
| Interaction between month and 9am relative humidity | various | <0.001\* |

Note: significance was tested using the linear model for quantitative variables and using an ANOVA for qualitative variables (month, day of the week and month-relative humidity interaction) \*p values below 0.05 were considered significant.

Model 2 still contains a non-significant predictor – day of the week. For the next version, this was removed. The significant terms in the final model were:

* minimum daily temperature (in degrees Celsius),
* 9am relative humidity (%),
* month of the year, and
* interaction between month and 9am relative humidity.

The two terms excluded from the model are day of the week and maximum daily temperature. The former is unsurprising – there are no readily available reasons the specific day of the week would predict the weather in general or evaporation in particular. The latter is more surprising – many people would assume the temperature is a good predictor of how much evaporation will occur. Temperature is a good predictor, but the fact that maximum temperature is not significant when the model also includes the minimum temperature suggests that any predictive value the maximum temperature has is entirely accounted for by the minimum temperature. Put another way, the close connect between the maximum and minimum temperature is likely creating a situation where all the predictive values of daily temperature are accounted for by the minimum temperature.

**Model diagnostics**

There are four assumptions core to linear models. These are linearity, constant variance (homoscedascity), normality, and independence. Each of these was tested for in the final model. Code for all model diagnostics is present in section 5 within the appendix.

Linearity is an assumption that the relationship between the predicted variable and the predictor variables are linear and therefore well represented by a linear model. To assess this, figure 6 shows the correspondence between the fitted or ‘predicted’ values, and the residual – the actual value minus the predicted value. The curved nature of the line suggests a linear model may not be the best approach, but that this is more of an issue for higher predicted values of evaporation.

*A close up of a map

Description automatically generated*

Constant variance (homoscedascity) is the assumption that errors (or 'noise') present in the model are not larger for certain types of value. This is tested in figure 7 by plotting the fitted values again the standard residuals, a measure of the strength of difference between the observed and fitted values.

*A close up of a map

Description automatically generated*

Figure 7 shows that there appears to be more noise present in the model for higher fitted values. This is a concern because it suggests the quality of the model will be lower for these values.

Normality is the assumption that errors are normally distributed. This is tested by plotting the same standardised residuals by the theoretical quartiles in a normal Q-Q plot (figure 8).

*A close up of a map

Description automatically generated*

Figure 8 shows the model generally adheres to the assumption of normality – it is only around the 2nd theoretical quartile a divergence is seen. The assumption of normality is less important for those models using more than 200 observations however, and the 365 observations in this data are a protective factor.

Independence is the assumption that the data points the model is based on are independent. As the evaporation is heavily reliant on the weather, we can expect a strong correlation between the weather on one day and the next. For this reason, there is likely to be a violation of independence in this model. There are options to correct for this autocorrelation that could be explored in future work.

### Results

Table 3 presents the coefficients for each predictor in the final model.[[1]](#footnote-1) The intercept represents the average evaporation when the minimum temperature and relative humidity are 0, averaging across all months and all interactions between month and relative humidity.

**Table 3. Estimates and significance values in model 3**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Coefficient** | **P-values** |
| Intercept | 6.03 | <0.001\* |
| Minimum temperature | 0.37 | <0.001\* |
| 9am relative humidity | -0.09 | <0.001\* |
| Month | various | <0.001\* |
| Interaction between month and 9am relative humidity | various | <0.001\* |

Note: significance was tested using the linear model for quantitative variables and using an ANOVA for qualitative variables (month and month-relative humidity interaction) \*p values below 0.05 were considered significant.

The minimum temperature coefficient (0.37) indicates a positive relationship between the minimum temperature and evaporation. That is, higher minimum temperatures predict higher amounts of evaporation. Conversely, the 9am relative humidity coefficient (-0.09) shows a negative relationship between relative humidity and evaporation. When the proportion of relative humidity is higher, this predicts less evaporation.

Understanding the coefficients of the categorical predictors is more complicated as each category will have its own coefficient. Table 4 shows the coefficients for each month and the p-values.

**Table 3. Estimates and significance values for levels of the month variable in model 3**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Coefficient** | **P-values** |
| Jan | -1.97 | 0.58 |
| Feb | -1.15 | 0.74 |
| Mar | 3.29 | 0.23 |
| Apr | 0 (reference category) | N/A |
| May | -6.35 | 0.06 |
| Jun | -10.35 | 0.01\*\* |
| Jul | -7.33 | 0.04\* |
| Aug | -9.07 | 0.01\*\* |
| Sep | -3.22 | 0.32 |
| Oct | -8.13 | 0.01\*\* |
| Nov | -3.01 | 0.29 |
| Dec | -1.05 | 0.72 |

Note: \*Significant at the .05 level \*\* significant at the 0.01 level

April has been used as a reference point for table 3, meaning the coefficients of other months refer to if they predict an increase or decrease in evaporation as compared with April. Only four months had a significant effect in the final model: June, July, August and October. During any of these months, we can predict a reduction in evaporation as compared with April. Of these months, July had the largest coefficient (-10.35).

The same interpretation was applied to the interaction between month and humidity. This interaction was only significant for July, June, and October. In these three cases the coefficient was between 0.10 and 0.14, indicating that the interaction between these months and humidity predicts an increase in evaporation.

### Discussion

The model can be used to predict evaporation in future scenarios as well as to understand the influence of different factors on evaporation. Table 4 shows the predicted evaporation for different scenarios. Note that although the maximum temperature, the specific day and the year are listed, only the minimum temperature, the month, the 9am humidity and the interaction between month and humidity are used in the model (as in previous sections).

**Table 4. Predicted evaporation for various scenarios**

|  |  |  |  |
| --- | --- | --- | --- |
| Scenario | Predicted evaporation (mm) | Predicted evaporation (mm)  Upper confidence interval\* | Predicted evaporation (mm)  Lower confidence interval\* |
| 1. February 29, 2020  Minimum temp 13.8  Maximum temp 23.2  9am humidity 74% | 5.5 | 6.6 | 4.4 |
| 2. December 25, 2020  Minimum temp 16.4  Maximum temp 31.9  9am humidity 57% | 8.6 | 7.6 | 9.6 |
| 3. January 13, 2020  Minimum temp 26.5  Maximum temp 44.3  9am humidity 35% | 14.9 | 12.8 | 17.0 |
| 4. July 6, 2020  Minimum temp 6.8  Maximum temp 10.6  9am humidity 76% | 2.3 | 1.3 | 3.2 |

Note: \*95% confidence interval used

The upper and lower confidence interval predictions shown in table 4 indicate the range within which evaporation should lie in 95% of cases, as according to the model. For example, in scenario 1, the model predicts an evaporation of 5.5mm, and predicts that in 95% of identical situations the evaporation will be between 4.4mm and 6.6mm. The 95% confidence interval for scenario 2 is between 7.6mm and 9.6mm, indicating that in some of those 95% of cases scenarios 1 and 2 will have the same level of evaporation.

The scenario with the highest evaporation is number 3, with a predicted evaporation of 14.9mm. This is because it has both the higher minimum temperature and the lowest humidity. This scenario has almost double the variability, with a 4.2mm difference between the upper and lower confidence interval. This fits with the previous findings that it is more difficult to predict evaporation at higher values. The scenario with the lowest predicted evaporation (#4) has the lowest range between the upper (1.3mm) and lower (3.2mm) confidence intervals – 1.9mm.

During days when more than 10mm of evaporation takes place at the Cardinia Reservoir, measures should be taken to ensure continuous water supply. Based on table 4, we can be confident that in 95% of cases the 10mm threshold will not be exceeded in scenarios 1,2, and 4. As the confidence intervals for scenario 3 span between 12.8mm and 17.0mm, we can predict that in 95% of cases more than 10mm of evaporation will take place, and therefore measures of water remediation should occur.

### Conclusion

The analysis here has determined that, within the context of the model used, the month, minimum daily temperature, 9am relative humidity and the interaction between month and humidity all have a significant impact on the amount of evaporation in a day. The analysis did not find any significant impact of the day of the week, or the maximum temperature when considered within the model with the other factors. As noted, it is likely that if a measure of minimum temperature were not included in the model, the maximum temperature would be a significant predictor.

It is important to note that this model did not test an exhaustive list of all possible factors that may affect evaporation. As with any linear model, there is the possibility that any untested factors could have a significant impact on predictions. To mitigate this, subject matter experts in the field of evaporation should be consulted on possible factors for inclusion in future versions of the model. Additionally, a number of validity assumptions for this model are questionable. For future predictions, the model should be improved with techniques to address the issues raised.

Finally, this model should be used with more caution when the predictions of evaporation are at higher values. The confidence range of evaporation is larger when the total predicted evaporation is greater, meaning it is harder to be confident about if the condition results in the exact predicted evaporation.

### Appendix – R code

### Segment 1 - loading libraries and loading data

##load libraries

library(tidyverse)

library(dplyr)

library(pander)

library(ggthemr)

library(gridExtra)

library(ggpubr)

library(lubridate)

library(ggiraph)

library(ggiraphExtra)

library(ggfortify)

#nicer look for table outputs

panderOptions('table.alignment.default', function(df)

ifelse(sapply(df, is.numeric), 'right', 'left'))

panderOptions('table.split.table', Inf)

panderOptions('big.mark', ",")

panderOptions('keep.trailing.zeros', TRUE)

#nicer theme for plots

ggthemr('flat')

##import data

raw\_data <- read\_csv("C:\\Users\\theja\\Downloads\\melbourne.csv")

summary(raw\_data)

###Segement 2 - preprocessing data

##create variables for month and day of the week

clean\_data <- raw\_data %>%

#use the weekdays and as.Date() function to get the day of the week and the month as new columns

mutate(day\_of\_week = (weekdays(as.Date(Date, "%Y-%m-%d"))), month = month(as.Date(Date, "%Y-%m-%d"), label=TRUE)) %>%

mutate(month = as.character(month))

##Segment 3 - inital statistical analysis

##examine the relationship between month and evaporation

#create data that's average evaporation by month

fig\_1\_data <- clean\_data %>%

select(month,`Evaporation (mm)`) %>%

group\_by(month) %>%

summarise\_at(vars(`Evaporation (mm)`),

list(`Average evaporation (mm)` = mean), na.rm = TRUE)

#plot average evaporation by month

fig\_1 <- ggplot(fig\_1\_data,aes(x = `Average evaporation (mm)`, y = month)) +

geom\_bar(stat = "sum") +

labs( y = '', title = 'Fig 1. Average evaporation by month, 2018-19', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

##examine the relationship between day of the week and evaporation

#create data that's average evaporation by day of the week

fig\_2\_data <- clean\_data %>%

select(day\_of\_week,`Evaporation (mm)`) %>%

group\_by(day\_of\_week) %>%

summarise\_at(vars(`Evaporation (mm)`),

list(`Average evaporation (mm)` = mean), na.rm = TRUE)

#plot average evaporation by day of the week

fig\_2 <- ggplot(fig\_2\_data,aes(x = `Average evaporation (mm)`, y = day\_of\_week)) +

geom\_bar(stat = "sum") +

labs( y = '', title = 'Fig 2. Average evaporation by day of the week, 2018-19', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

##examine the relationship between the maximum temperature and evaporation

#plot evaporation by max temp

fig\_3 <- ggplot(clean\_data, aes(x = `Evaporation (mm)`,y = `Maximum Temperature (Deg C)`)) +

geom\_point() +

labs( y = 'Max temperature (deg c)', title = 'Fig 3. Evaporation by max daily temperature, 2018-19', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

##examine the relationship between the maximum temperature and evaporation

#plot evaporation by min temp

fig\_4 <- ggplot(clean\_data, aes(x = `Evaporation (mm)`,y = `Minimum temperature (Deg C)`)) +

geom\_point() +

labs( y = 'Min temperature (deg c)', title = 'Fig 4. Evaporation by min daily temperature, 2018-19', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

##examine the relationship between the relative humidity and evaporation

#plot evaporation by min temp

fig\_5 <- ggplot(clean\_data, aes(x = `Evaporation (mm)`,y = `9am relative humidity (%)`)) +

geom\_point() +

labs(title = 'Fig 5. Evaporation by relative humidity, 2018-19', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

###Segment 4 - model selection

##model 1

#define the model using all predictors

all\_predictors\_mdl <- lm(`Evaporation (mm)` ~ month +

day\_of\_week +

`Maximum Temperature (Deg C)` +

`Minimum temperature (Deg C)` +

`9am relative humidity (%)` +

month:`9am relative humidity (%)`, data = clean\_data)

#summary information for model

summary(all\_predictors\_mdl)

#anova to understand significance for qualitative variables

anova(all\_predictors\_mdl)

##model 2

#the greatest non significant p value in model 1 was `Maximum Temperature (Deg C)`, so removing that from model 2

mdl\_v1 <- lm(`Evaporation (mm)` ~ month +

day\_of\_week +

`Minimum temperature (Deg C)` +

`9am relative humidity (%)` +

month:`9am relative humidity (%)`, data = clean\_data)

#summary information for model

summary(mdl\_v1)

#anova to understand significance for qualitative variables

anova(mdl\_v1)

##model 3

#the greatest non significant p value in model 2 was the day of the week, so removing that from model 3

model\_data <- clean\_data %>%

mutate(evaporation = `Evaporation (mm)`, min\_temp = `Minimum temperature (Deg C)`, rel\_humidity = `9am relative humidity (%)`)

mdl\_v2 <- lm(`Evaporation (mm)` ~ month +

`Minimum temperature (Deg C)` +

`9am relative humidity (%)` +

month:`9am relative humidity (%)`, data = clean\_data)

#summary information for model

summary(mdl\_v2)

#anova to understand significance for qualitative variables

anova(mdl\_v2)

###Section 5 - model diagnostics

##plot the residual against the fitted values

fig\_6 <- autoplot(mdl\_v2, which = 1)+

labs( title = 'Fig 6. Residual values vs fitted values', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

#plot the standard residual against the fitted values

fig\_7 <- autoplot(mdl\_v2, which = 3)+

labs( title = 'Fig 7. Standard residual vs fitted values', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

#plot the standard residual by theoretical quartiles

fig\_8 <- autoplot(mdl\_v2, which = 2)+

labs( title = 'Fig 8. Normal Q-Q', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

#plot the autocorrelation

fig\_8 <- autoplot(mdl\_v2, which = 4)+

labs( title = 'Fig 8. Normal Q-Q', legend ='') +

theme(plot.title = element\_text(size=12),legend.position = "none")

### Section 5 - predictions

#create tibble with values to prediction based on scenario 1

scenario\_1 <- tibble(month = "Feb"

, `Minimum temperature (Deg C)` = 13.8

,`9am relative humidity (%)` = 74

)

#run prediction

predict\_1 <- as\_tibble(predict(mdl\_v2

, newdata = scenario\_1

, interval = "confidence"

,level = 0.95))

#create tibble with values to prediction based on scenario 2

scenario\_2 <- tibble(month = "Dec"

, `Minimum temperature (Deg C)` = 16.4

,`9am relative humidity (%)` = 57

)

#run prediction

predict\_2 <- as\_tibble(predict(mdl\_v2

, newdata = scenario\_2

, interval = "confidence"

,level = 0.95))

#create tibble with values to prediction based on scenario 3

scenario\_3 <- tibble(month = "Jan"

, `Minimum temperature (Deg C)` = 26.5

,`9am relative humidity (%)` = 35

)

#run prediction

predict\_3 <- as\_tibble(predict(mdl\_v2

, newdata = scenario\_3

, interval = "confidence"

,level = 0.95))

#create tibble with values to prediction based on scenario 4

scenario\_4 <- tibble(month = "Jul"

, `Minimum temperature (Deg C)` = 6.8

,`9am relative humidity (%)` = 76

)

#run prediction

predict\_4 <- as\_tibble(predict(mdl\_v2

, newdata = scenario\_4

, interval = "confidence"

,level = 0.95))

1. Code for this section is present in section 4 of the appendix. [↑](#footnote-ref-1)