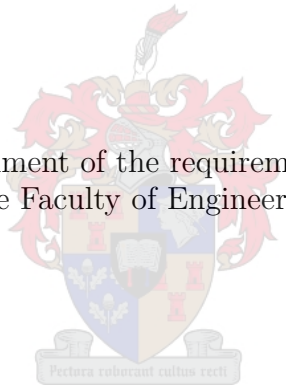
# Estimation Methods for Date Palm Yield – a Feasibility Study

by

Karlien Heyns




Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering (Industrial) in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof JF Bekker

March 2021

# Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 1 March 2021

# Abstract

With a growing population and a need for food security, crop yield prediction is vital; not only is it used by exporters and importers, but also by the farmer who needs to plan marketing strategies and determine prices. Methods on crop yield prediction are more abundant for annual plants than for perennials. Very few reliable crop yield prediction models have been developed on the date palm, which is grown in arid regions with plentiful water available.

Date fruit is a nutritious food which is produced in many countries and consumed widely around the world. Farming with date palms is a complex process with a large variety of factors affecting the annual yield. This study investigated the feasibility of predicting date yield using data collected by a research partner producing date fruit. Data on some farming practices as well as weather conditions was collected from 2010 onwards, at different levels of detail.

Machine learning techniques were considered for prediction of yield; however, four applicable linear regression techniques were identified and could be used with the available data for feature selection. The dataset has many features, but dominant features were extracted from the data. Some of the feature selection methods used were a correlation technique, stepwise regression and regularisation. These features were further used to develop regression models. It was found that some weather features were important, as well as features describing the date bunch mass. The latter were observed by sampling bunches from trees in different orchards.

Linear regression models were developed on orchard level and on farm level, *i.e.*, for the farm as a whole, and the best-performing linear regression models were selected (while avoiding overfitting). The yield predictions following from these models were compared to the actual annual yield recorded, as well as the estimated yield determined by a rather pragmatic yield prediction method devised by the research partner. The selected models produced a 4% prediction error while the farm method gives a 7% error. The proposed models reduced the prediction error and eliminate the need for laborious sampling work done to support the farm prediction model. The study found that certain data that is collected is not needed by the proposed linear regression models.

The study was done from an industrial engineering perspective, and a systematic process was followed to critically assess the data available. This was done to keep complexity of the models at a level suitable for reasonable and accurate yield prediction, and to eliminate some unnecessary data collection labour on the farm.

# Opsomming

Met ‘n groeiende wêreldopulasie en ‘n behoefte aan voedselsekuriteit is oesvoorspelling van gewasse noodsaaklik – nie net vir in- en uitvoerders nie, maar ook vir die produsent wat bemarkingstrategieë beplan en prysbepaling doen. Oesvoorspelling is meer gevorder vir eenjarige gewasse as vir meerjariges. Op die dadelpalm, wat groei in droë gebiede waar water volop beskikbaar is, is weinig betroubare oesvoorspellings ontwikkel. Die dadel is ‘n voedsame vrug wat verbou word in baie lande en regoor die wêreld geniet word. Die verbouing van dadelpalms is ‘n ingewikkelde proses, waar ‘n groot verskeidenheid faktore die oes beïnvloed. Hierdie studie het die lewensvatbaarheid ondersoek van die oesvoorspelling van dadels met data wat deur ‘n navorsingsvennoot, ‘n dadelprodusent, ingesamel is. Hierdie data, gedokumenteer sedert 2010, bevat bestuurspraktyke op die plaas, sowel as weerstoestande, met wisselende vlakke van detail.

Verskeie masjienleertegnieke is oorweeg vir die oesvoorspelling. Uiteindelik is vier lineêre regressietegnieke uitgesonder en kon hierdie vier gebruik word op die beskikbare data om veranderlikes te kies. Die datastel bestaan uit baie veranderlikes, maar dominante veranderlikes is geïdentifiseer met hierdie seleksiemetodes, wat ‘n korrelasietegniek, stapsgewyse regressie en regularisering insluit, en die veranderlikes is verder gebruik om regressiemodelle te ontwikkel. Sommige veranderlikes wat die weerstoestande in sekere tye bevat, en veranderlikes wat die vrugtrosmassas beskryf, is van groter belang. Die trosmassas is verkry deur ‘n steekproef van ‘n enkele boom in elke boord se trosse.

Lineêre regressiemodelle is ontwikkel op boord- en op plaasvlak, d.w.s. vir die plaas as ‘n geheel, en die modelle met die beste resultate is gekies. Die oeste wat deur hier-die modelle voorspel is, is vergelyk met die werklike jaarlikse oes en die geskatte oes bepaal met ‘n pragmatiese oesvoorspellingsmetode deur die navorsingsvennoot. Die gekose modelle verbeter die voorspellingsfout van die huidige pragmatiese metode se 7% na 4%. Die voorgestelde modelle verminder dus die voorspellingsfout en terselfdertyd sorg dit vir die weglating van tydsame en arbeidsintensiewe steekproefwerk wat die huidige skattingsmetode van die plaas benodig. Die studie het gevind dat sekere data wat ingesamel word, nie benodig word vir die voorgestelde modelle nie.

Die studie is vanuit ‘n bedryfsingenieursperspektief benader, en ‘n sistematiese proses is gevolg om die data krities te assesseer, om die kompleksiteit van die modelle op ‘n gepaste vlak vir sinvolle en akkurate oesvoorspelling te hou, en om onnodige arbeid vir datainsamelings op die plaas uit te skakel.

# Acknowledgements

I would like to express my sincere gratitude to the following people and organisations for their contribution to this thesis:

- My supervisor, Prof James Bekker, for mentoring me and teaching me much more than just industrial engineering, for allowing the work to be my own, and for wisdom and humour when I needed it.
- Jolene Wium, for providing the topic and access to the data, and for hope in a desperate time.
- The research partner, for the data provision and the opportunity to collaborate.
- Anne Erikson, for copy editing the document and in the process inspiring me to continue studying (the English language) after this endeavour.
- All the subject matter experts who engaged with me and opened new worlds to me, with special mentions for Prof Martin Kidd, Prof Karen Theron, Manie van Zyl and Marieta van der Rijst.
- My teammates – my friends, flatmates and family – for their unconditional support, interest and input, and for believing in me when I did not believe in myself.
- My mom. For always having a plan, always being proactive and positive, and always being there.
- My Creator. For showing me I do not always need to see the light at the end of the tunnel, only the Light before my steps.

# Contents

# List of Figures

# List of Tables

# Nomenclature

**Acronyms**

| | |
|---|---|
| $k$NN | $k$-nearest neighbour |
| AIC | Akaike information criteria |
| AICc | Corrected Akaike information criteria |
| ANN | Artificial neural network |
| ANOVA | Analysis of variance |
| BIC | Bayesian information criteria |
| CSV | Comma-separated values |
| CV | Cross-validation |
| DPP | Days post-pollination |
| DTR | Decision tree regressor |
| GDD | Growing degree days |
| KSA | Kingdom of Saudi Arabia |
| LASSO | Least Absolute Shrinkage And Selection Operator |
| LOOCV | Leave-one-out cross-validation |
| MAE | Mean absolute error |
| MLR | Multiple linear regression |
| MRA | Multiresolution analysis |
| OLS | Ordinary least squares |
| PAR | Photosynthetically active radiation |
| PCA | Principal component analysis |
| PCR | Principal component regression |
| PLS | Partial least squares |
| PPFD | Photosynthetic photon flux density |
| RMSE | Root mean squared error |
| RNN | Recurrent neural network |
| RSS | Residual sum of squares |
| SAWS | South African Weather Service |
| SIS | Sure Independence Screening |
| SME | Subject-matter expert |
| SVM | Support vector machine |

| | |
|---|---|
| SVR | Support vector regressor |
| UAE | United Arab Emirates |
| USA | United States of America |
| VIP | Variable importance in projection |

**Symbols**

| | |
|---|---|
| $i$ | index of the month of the year |
| $k$ | Number of features/parameters |
| $n$ | number of observations |
| $r$ | Pearson correlation coefficient |
| $ET_0$ | Reference evapotransipration |
| $HU_i$ | Sum of the calculated heat units of month $i$ |
| $Hum_i$ | Mean humidity in month $i$ |
| $MaxT_i$ | Maximum temperature in month $i$ |
| $MeanT_i$ | Mean temperature in month $i$ |
| $MinT_i$ | Minimum temperature in month $i$ |
| prev | Prefix for feature containing weather from the previous season |
| $Rain_i$ | Sum of the rainfall in month $i$ |
| $Wind_i$ | Mean of the daily maximum daily wind speed in month $i$ |

# Chapter 1

# Introduction

Food security is one of the major challenges in the world today due to an ever-increasing population constantly exploiting its resources. Considering limited natural and human resources, more economic food cultivation and production are constantly attempted. This is enabled by increasingly employing scientific measures, such as yield estimation, to produce adequate profitable products in demand.

Date palm production, a crop grown in arid regions, is the topic of this research. Where it is cultivated, the palm has an integral relationship with the life of the rural inhabitants and contributes to their social life and culture as well as the local economy. A variety of primary and secondary products of the date palm provide economic and social security to the people (Rajmohan, 2011).

Having been included on UNESCO's list of Intangible Cultural Heritage of Humanity in 2019 (Zacharias, 2019), the date palm is recognised not only for its fruit, a staple food of the Middle East (Sabir, 2019), but also its uses in rituals, furniture and even woven baskets, mats and hats. The global date palm market was valued at around USD 13 billion in 2018 and is expected to increase to USD 18 billion by 2023 (Shahbandeh, 2020). It is considered to be one of the main sources of income for many of the countries of cultivation, which are otherwise reliant on oil trade. In Egypt the date industry supports over a million people and it is estimated that a commercial date plantation of 40 ha requires an annual total of 8 000 working days. This can help to curb urbanisation (FAO, 2016). Annual fluctuations in date palm yield create uncertainty for the millions of people reliant on the industry, and in this regard yield forecasting will assist in economic certainty and food security over longer periods. Accurate yield estimation aids in determining marketing strategies and gaining optimal yield and also assists planning to mitigate the annual 1.3 billion tons of global food waste (FAO, 2021).

Yield estimation is gaining importance due to the global population growth and global warming that threaten water resources, especially as the date palm is a very water intensive crop. The estimation is of particular interest to the agricultural practitioner, requiring yield estimates for budgeting and export planning, but also for timely procurement of the right quantity and size of packaging. When predicting the yield of dates, the characteristics, specifically the size of the fruit as well as the total harvest mass are important outcomes. The suitability of these characteristics varies across the world as particular products are attractive to certain consumers. In some regions, smaller fruit is preferred while in others, the market wants larger, more succulent fruit.

This research project focuses on identifying the factors that influence the date palm yield at a South African producer. Data was obtained from the research partner and ways to learn from this data are investigated.

This chapter introduces the research presented in this thesis and background information on the subject is given. Next follows the research assignment, scope and objectives and a short description of the methodology.

## 1.1 Research background

This section sketches the background to the date palm industry and presents a few brushstrokes on the concept of yield estimation in general.

### 1.1.1 Date industry

In terms of its cultivation history, the date fruit is one of the oldest cultivated fruits in the world. The date industry originated in and is still dominated by the Middle East (Kayal, 2015). Date production is centred in the Northern Hemisphere, and the industry thrives in North Africa and the Arab States. The largest global producers are Egypt, Saudi Arabia, Iran, the United Arab Emirates, Pakistan, and Algeria. Israel and the USA are also smaller exporters. In the United States two varieties, namely the drier Deglet Nour and the fat, maple-hued Medjool are primarily produced (Kayal, 2015). The Medjool is also grown in South Africa, which is geographically far from the well-established date palm production regions of the world. An estimated 60% of the fruit produced in South Africa is exported and the rest is sold on the local market (McCubbin, 2007).

Dates exported from South Africa are very high in value ($8 per kg) while dates from Iran and the United Arab Emirates (UAE) have lower import value – around $1 per kg (Reilly et al., 2010). The Khapra Beetle, an important quarantine pest classified as one of the 100 worst invasive species worldwide (Athanassiou et al., 2019), is widespread in major date-growing regions including Algeria, Egypt, Iran, Iraq, Israel, Jordan, Lebanon, Morocco, Pakistan, Saudi Arabia, the UAE, and Yemen. Only South Africa, the United States and Mexico are not listed as Khapra Beetle countries. Countries such as Australia, with strict quarantine fruit entry regulations, are not permitted to import *khalaal*[1] dates from countries listed to have Khapra Beetle.

The northern parts of South Africa and southern Namibia also have an advantage by being located in the Southern Hemisphere, while the greatest date producing countries are all in the Northern Hemisphere. For the next few years, the northern parts of South Africa will be in season during Ramadan, according to Clive Garrett from the South African fruit producer ZZ2 (McGregor, 2019). This is crucial, as the date fruit plays a significant role in many religions, but especially in Islam. There is an old Arab saying: 'The uses of date palm are as many as the number of days in the year' (Khan and Khan, 2014). Traditionally, a date breaks the fast after the sun goes down at the end of each day of Ramadan.

More than 98% of the world's dates come from the Northern Hemisphere, with more than 90% of this production being consumed in the country of origin. When these countries are out of season, South Africa and its greatest competitors, Namibia and Australia, are able to supply dates. These countries still have a young developing market, with the production of dates only having been introduced in the last century. However, it is a large global industry, selective of its suitable climate, with dates cultivated in more than 40 countries, with 800 000 ha producing around eight million tons of dates a year (Reilly et al., 2010; Transparency Market Research, 2018).

According to agricultural experts, there are more than 3 000 date varieties. Dates, becoming increasingly popular in environmentally and health-conscious recipes, provide a long list of health benefits, ranging from their quality as a source of vitamins and minerals, to being a natural sugar. When consumed, date fruit is beneficial in many ways *i.e.*, aiding carbohydrate, protein, and fat metabolism, promoting colon and bowel function, balancing blood sugar, enhancing brain and eye health, increasing blood production, lowering blood pressure, and preventing osteoporosis. In an agricultural context, the date palm is also beneficial to farmers for intercropping – cultivating more crops simultaneously on the same field – as it provides shade (Johnstone, 2020).

### 1.1.2 Yield estimation and forecasting

Yield estimation models can aid in the realistic anticipation of the production mass of a season. This is vital for planning purposes for the agricultural practitioner and other role players in the value chain. Supportive models such as yield estimation models can aid in realistic anticipation for harvests and preparation through marketing and logistics.

[1]The term *khalaal* will be described in Section 2.1.3.

After investigating the expansive literature on yield estimation, specifically focusing on date fruit, only 57 closely related articles were isolated. Of these, 52 originate from Egypt, Saudi Arabia, Iran, and the UAE, confirming these countries' interest and market share in date cultivation. The articles were published between 2002 and 2019 (the literature review was conducted in 2019). The low number of relevant articles indicates clearly that there is research potential in this field.

For yield estimation of dates as well as other fruits like apples, a few articles suggest the counting of trees and fruits per tree from images instead of manual *in situ* counting. The literature on the date palm (such as Iqbal et al. (2014), Al-Saikhan (2008)) places great focus on pollination and other management practices and their effect, but this project will consider many features, including weather and fruit growth over time, to determine a yield estimation. Yield estimation models have been developed and forecasting methods assessed for many crops, including rice (Horie et al., 1992), apples (Wang et al., 2013), citrus (Malik et al., 2016) and grapes (Sabbatini et al., 2012). The focus of this study, however, is on perennials such as fruit trees and specifically palms, as opposed to annual crops *e.g.* wheat or maize.

## 1.2 Research assignment

There is a need for the further development of the South African date industry, as discovered during the research of literature and a glance at the industry. The study will investigate the feasibility of predicting date yield from factors influencing date production. There is a lack of knowledge regarding the factors influencing date palm yield, especially in the South African context. The research assignment is formulated as follows:

> Investigate the feasibility of developing one or more date yield prediction models at a specific production location.

## 1.3 Research scope

The scope indicates the research boundaries and what it encompasses. The data required for the development of the model will be obtained from the anonymous research partner, an agricultural company operating in South Africa, supplemented by an internet source, Prediction of Worldwide Energy Resources (POWER). The study will focus on a single date farm having many date orchards at one location. It is worth noting that the data was not captured for the purpose of this study, as the research partner collected data as deemed necessary for other purposes. The available data will thus be the main driver of this study.

## 1.4 Research objectives

The research project has the following objectives:

1. Gather information on, and grasp an understanding of, date cultivation through a literature study.
2. Identify key factors regarding yield estimation of date production.
3. Identify prediction models applicable to the specific type and structure of the data.
4. Investigate the possibility of developing a yield estimation model for date production in South Africa given historical agricultural practice data, yield data and various input features.
5. Improve accuracy of yield estimation for the research partner, considering critical features.

## 1.5 Proposed research methodology

The following research approach is proposed in order to achieve the objectives mentioned above:

1. Conduct a literature review to become familiar with the following topics:
   (a) The global date industry
   (b) The South African date industry (seeking gaps or opportunities)
   (c) Yield estimation done on various crops
   (d) Yield estimation on dates
      i. Effects of different factors (later referred to as 'features') such as weather and management practices
      ii. Existing yield prediction methods
2. Apply for ethics clearance for the project in order to gain permission for acquiring data.
3. Obtain data from the agricultural company about the following elements:
   (a) Meteorological history
   (b) Irrigation and soil fertilisation (growth conditions)
   (c) Growth measurements
   (d) Management practices such as pollination, thinning, soil coverage
   (e) Current prediction method
4. Critically assess the data for consistency, completeness, and possible patterns.
5. Investigate possible modelling techniques that are suitable for the data.
6. Employ acquired data to determine the correlation between the input features and output (yield), identifying more significant and less relevant features (factors).
7. Use identified features in the development of models to predict yield.
8. Compare models with the date yield prediction model currently used by the research partner.
9. Draw conclusions and make recommendations for best management practices based on findings. Recommendations will also include opportunities for further research.

These steps are merely a proposition based on investigation of the research assignment, aimed at addressing the objectives, but will evolve as the literature analysis is conducted.

## 1.6 Deliverables envisaged

The research project aims to provide the date producer with insight into the most vital factors influencing harvest yield and models to use for more accurate yield estimation. An estimator tool for more accurate yield predictions is also envisaged as a deliverable of this research.

## 1.7 Structure of the document

In Chapter 2 the literature of the date palm is investigated. The chapter examines the cultivation of dates with a focus on factors influencing yield. The global date market is also explored. Chapter 3 reports on existing crop and yield prediction models used in practice and in research. These models are presented in two categories: process, or mechanistic, models and statistical models. The focus is shifted to the field of date palm research, particularly date palm yield. Chapter 4 discusses the use of data in modelling, with an emphasis on data analysis and understanding. The topic of a required sample size is introduced and detailed. Finally, the factors that influence yield, identified in Chapter

2, are transformed into important data features, and a discussion on possible required datasets is presented. Chapter 5 introduces the available real-world datasets used in this project and begins exploring their usefulness, discussing those with which will be proceeded and those which do not provide enough promising relationships. Chapter 6 discusses predictive modelling theory and reviews methods for feature selection which are used in Chapter 7 to incorporate the data from Chapter 4. Possible date yield prediction models are presented in Chapter 7. Secondary findings are also reported after investigating the data. Finally, the conclusion in Chapter 8 summarises the project. Research findings and results are discussed and recommendations for further study are made.

## 1.8 Summary of the introduction

The present chapter introduced the research project in brief, describing the problem and the rationale of the study along with the research plan. Lastly, the structure of the thesis is outlined.

In the following chapter, the literature study on date palm and the global date market is presented. Date palm cultivation is discussed, including details on factors influencing yield, which is relevant to the outcome of this project.

# Chapter 2

# Background on the date palm

The previous chapter introduced the research project, presenting the problem and background to the context and defining the plan and objectives. In this chapter, references to date palms from the literature are presented. Investigation is done on the cultivation of the date palm, agricultural practices and, although this is not a natural science study, some environmental factors are considered. It explores the global date industry, the market for dates and also takes a look at post-harvest handling. Figure 2.1 gives an outline of this chapter and the next, starting with research on dates and following with an exploration into the date industry as well as yield estimation. The date industry and cultivation are presented in this chapter, while yield estimation models are investigated in Chapter 3.

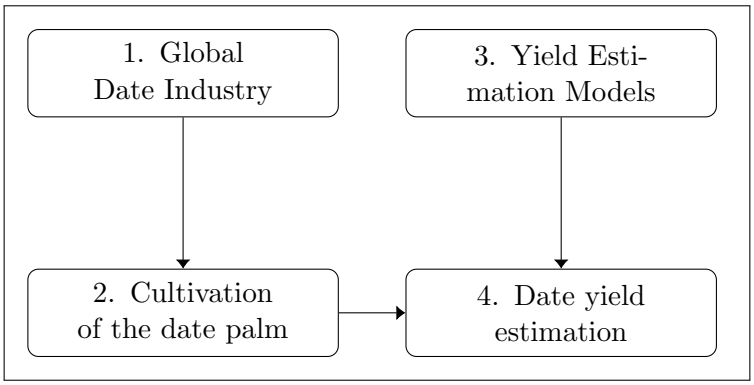



Figure 2.1: Literature Layout of Chapters 2 and 3

## 2.1 Date cultivation

*Phoenix dactylifera*, part of the family Arecaceae, is cultivated for its sweet fruit, commonly referred to as 'date'. The species name *dactylifera* is derived from the Ancient Greek word for 'date' or 'finger', *dáktulos* and the stem of the Latin verb *ferō* meaning 'I bear' (de Ferrara, 2012). Figure 2.2 shows an example of this tree, generally known as the date palm, bearing its fruit bunches. The date palm will be referred to as a tree, although there is controversy concerning the definition of a tree and whether palms qualify. Botanists define trees as woody plants with secondary growth – a definition that does not fit date palms – while the ecological definition is much broader: a tree is a plant providing habitat and shade, that produces leaves and flowers, stabilises soil and maintains biodiversity. The common definition, however, is that a tree is a large, woody plant with a single stem or trunk. Date palms fit the ecological and this common definition. Date production today is more than the exploitation of a source of nature, it creates an investment opportunity and is a source of revenue for many. It currently represents a significant global agricultural industry with production of more than eight million tons[1] of fruit in 2017. It is projected that by 2026 the global revenue of the date palm market will increase to US$ 10 353 million (Transparency Market Research, 2018).

### 2.1.1 Date palm trees

Date palm trees are large trees with a typical shape and can grow as tall as 30 m, with 12 to 250 feather-shaped leaves, up to 6 m long, produced each year. These leaves live for several years before

[1] In the United States, a ton (the short ton) is a unit of mass in the avoirdupois system, equivalent to 2 000 pounds or 907.18 kg. In Britain and everywhere outside of the US, a ton (the long ton) refers to 1 016.05 kg. The metric ton is 1 000 kg, and also written as just *ton*. A tonne also refers to the metric ton.

Figure 2.2: A typical date palm tree (Steinberger, 2013)

turning brown and being dropped. The trunk is covered with the leaf bases which prevents loss of water and also protects the tree against damage by animals.

The date palm is a dioecious species, meaning that the trees are either male or female. When trees are between four and eight years old they produce spathes which distinguish between male and female trees. Male spathes, torpedo-shaped and with a swollen appearance, are thicker and shorter than the female spathes, that have an elongated shape with a flat blunt tip (Intha and Chaiprasart, 2018). When the spathe splits open, floral strands, also called spikelets, appear. In male florescence, these are short, while the strands in female inflorescences are long and slender. Flowers borne on the strands, can also be examined in order to distinguish between male and female palm trees. Male staminate flowers are waxy and cream-coloured with visible petals, stamens (the male productive organs) and pollen. The female pistillate flowers are white, spherical florets that only contain pistils (three stigma), carpels and no petals.


Figure 2.3: Male and female date palm inflorescences (Intha and Chaiprasart, 2018)

Clusters of the female inflorescences are between 30 and 75 cm in length while male flower clusters are usually less than 23 cm long (McMullen, 2018). Pollen from the male flower is carried to the female flower by the wind, insects or manual (artificial) pollination in cultivation. A female flower

cluster then produces a date bunch. Large female trees in their peak production phase can bear eight to 12 bunches at a time, producing around 90 kg of fruit in a year with high levels of management. Fruit production peaks at around the age of 30 years and declines after the age of 60 years, until the end of the tree's reproductive life at around 80 years. These date bunches weigh around four to 18 kg.

Individual pollination and fruit harvesting are traditionally done by hand, during which the harvester must use ropes or ladders to manoeuvre up the tree, sometimes carrying a machete for trimming. In modern cultivation, this dangerous method is replaced by harvesting with cherry pickers and mechanical buckets. More on harvesting will follow in Section 2.5.2.

For a tree to produce dates, the temperatures must exceed 18 °C (Zaid and de Wet, 2002*a*). Therefore, dates are primarily cultivated in the very warm, dry parts of North Africa and the Middle East while some are grown in Namibia, Europe, Asia and the United States. This is because date palms can adapt to severe environmental conditions which include high temperature, salinity and drought.

The date fruit itself is an oblong soft to dry berry with a single stony seed that is surrounded by fleshy pulp (Sardar A. Farooq, 2012). A membranous endocarp separates the seed from the flesh. Figure 2.4 displays the anatomy of the date fruit in the final stage.

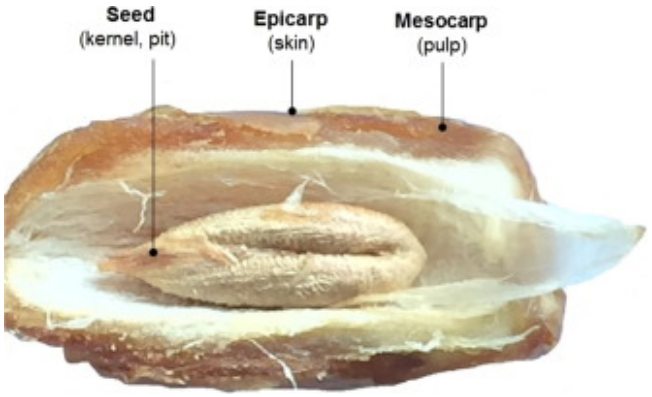



Figure 2.4: Anatomy of the date fruit (Ghnimi et al., 2017)

Depending on the cultivar, the colour of the date fruit can fluctuate from yellow to black and vary in length from 25 – 75 mm when suitable for harvesting (Lobo et al., 2013). These fruits are the traditional staple diet of the Bedouin who could survive on dates and water alone for months, while the pits are eaten by camels. Dates, having been cited for their medicinal benefits, can also be used in cooking or mashed to a pulp and then strained to produce honey-like date syrup. They are a good source of fibre and contain notable levels of iron and potassium, and fair amounts of calcium, copper, magnesium, manganese and sulphur (Rygg, 1975). It is a source of 16 amino acids as well as vitamins A, B1 and B2 (Chao and Krueger, 2007). The date is used medicinally for intestinal problems, oedema, liver and abdominal problems, treatment for colds and more.

### 2.1.2 Date cultivars

The terms *variety* and *cultivar* are often confused. 'Cultivar' is short for cultivated variety, meaning that the plants are selectively bred for traits maintained over generations. A 'variety' is the taxonomic rank. In the scientific name of an organism, the variety is written after the species name, and often preceded by its abbreviation, 'var'. Dates come in various cultivars with different characteristics. For instance, the *Medjool* date cultivar (also written *Majdool* and interestingly meaning 'unknown' in Arabic) grown and enjoyed in South Africa is large, dark brown in colour, has a wrinkled skin in edible stage and tastes sweet. It is known as the 'king of dates' and also called the 'crown jewel of dates' (de Ferrara, 2012). Another well-known cultivar, the crispy *Barhi*, has a round shape and a caramel taste.

### 2.1.3 Development stages

Dates also come in different styles and can be classified according to the growth stage or ripeness. Understanding these stages is vital for an investigation into date growth and development, as well as understanding the market. The cultivar of dates influences and could determine the business model, from the plantation management to the post-harvest processing and distribution processes.

The ripening of dates is divided in stages or styles with Arabic names, as seen in Figure 2.5, which displays the stages and the respective number of days post-pollination (DPP).

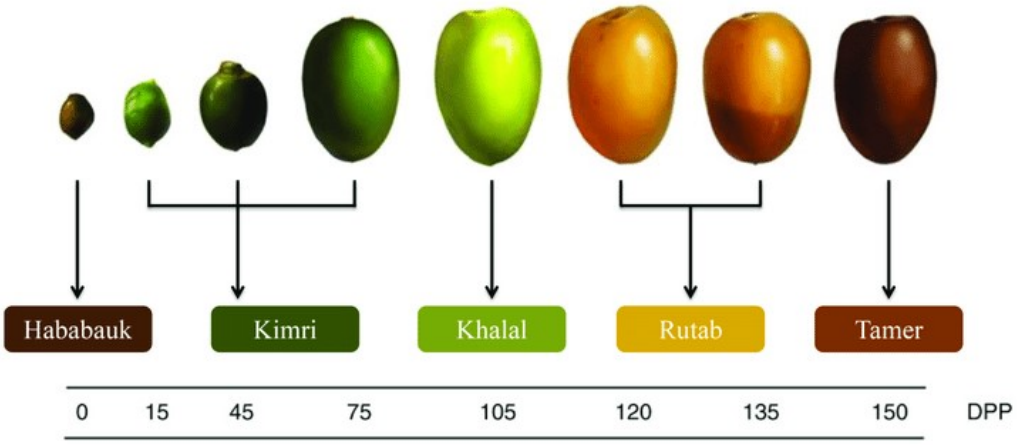



Figure 2.5: Stages of date fruit ripening (Al-Mssallem et al., 2013)

- The first stage after pollination is called *hababouk*, also written *hababauk*. This stage consists of the small spherical, cream-coloured female date flowering just after pollination and the small immature fruit (Nath, 2000).
- The next stage of date growth is *kimri*. *Kimri* dates are green, developing and inedible. During this stage, size and mass increase are rapid (Chao and Krueger, 2007).
- Next follow *khalaal*, also written *khalal*, or fresh dates, which are perishable and need to be refrigerated. The fruit colour changes from green to the colour characteristic of the cultivar. They contain 50% moisture and are firm and crunchy. During this phase in which the fruit reaches full size, the rate of gain in size and mass decreases slightly. The *Barhee* cultivar is usually enjoyed as a khalaal date. *Khalaal* production has a short harvest period and production peak. They require a defined transportation temperature, complicating their distribution.
- Next, the dates ripen from *khalaal* to *rutab* (ripe) dates, which have 30 to 35% moisture and high sugar levels. Darkening of skin, to amber or dark brown, occurs in rutab. More insoluble tannins develop in the fruit.
- *Tamar* (cured) dates are durable and have a long shelf life. These, containing 10 to 15% moisture, are the highest in sugar. The high sugar to water ratio prevents fermentation (Chao and Krueger, 2007). *Tamar* dates, also written *tamer* or *tamr*, are also known as bread dates, or the bread of the desert, since they are a main source of nutrition in many arid regions. For tamar dates, the main requirement is a suitable climate, characterised by long, hot, dry summers (Reilly et al., 2010).

## 2.2 Propagation

Zaid and de Wet (2002*b*) published three propagation techniques for the date palm. Dates can be

propagated by seed, offshoot or the more modern, recommended technique of tissue culture propagation.

Seed propagation is not a suggested method of vegetative propagation of the date palm. Also referred to as sexual propagation, seed propagation is not as time- and space-efficient or cost-effective as the other methods, although it is the easiest and quickest, due to the following:

1. Since these palms are heterozygous[1], great variation occurs within the progeny. Therefore, some preferable parent palm characteristics will not be transferred.
2. Because date palms are dioecious, half the progeny will be males. There is, however, no guarantee of sex of the progeny at the early developmental stages. This is only visible when the plant comes into bloom.
3. When compared to clonal plants, female date palms that originate from seedlings usually bring forth fruits of variable and inferior quality which also mature late. In a plantation of seedlings, good quality fruit is produced by less than 10% of the palms.

Offshoot propagation has advantages over seed propagation. Offshoot (also known as asexual or vegetative) propagation consists of a number of steps: offshoot selection, offshoot rooting, offshoot pruning, offshoot removal, and planting of offshoots. This process has the following benefits:

1. Offshoots grow from leaf axils (referred to as axillary buds) on the mother plant. Hence, the fruit quality will be similar to the mother date palm. Plants from offshoot are true to type to the parent plant, unlike in the case of seed propagation.
2. These plants bear fruit more than two years earlier than seedlings. Two developmental periods occur during the lifespan of the date palm. The vegetative period is when axillary buds develop into offshoots and the generative phase represents the period when buds form inflorescences and offshoots cease. This takes almost three years and another three or four years are required to reach the desired size to be separated and planted.

The technique of offshoot propagation does have some disadvantages as well. Only a few offshoots are produced, causing slow propagation. This production is also restricted to a specific phase in the life span of the date palm, making propagation from a mature plant specimen impossible.

Tissue culture or *in vitro* propagation, in essence cloning, is the best alternative, recommended by Zaid and de Wet (2002*b*) because of these factors:

1. Large-scale multiplication is possible.
2. Seasonality does not play a role since the plants are multiplied under laboratory (controlled) conditions.
3. The technique is economically reliable and suitable for large-scale production.
4. The result is the production of genetically homogenous plants.
5. Propagation of metaxenia characteristics is done quickly and relatively easily.

Tissue culture and micropropagation are often confused, as both are forms of asexual reproduction, categorised as vegetative propagation and used to produce identical plants – thousands in a small period of time. Micropropagation is the process of initially propagating the plants by growing plantlets in tissue culture followed by planting them out. Tissue culture is a micropropagation technique by which plant cells, tissues and organs are maintained and grown on an artificial medium under controlled environmental conditions to produce new plantlets from an explant.

[1] According to Vocabulary.com heterozygous is defined as having dissimilar alleles at corresponding chromosomal loci

An explant is defined as a cell or piece of tissue transferred from plants or animals to a nutrient medium. Researchers and agricultural practitioners have used different kinds of explants subsequent to the first implementation of date palm tissue culture. Research has shown that date palm micropropagation is most successful with explants of meristematic origin, including apical shoot tips, lateral buds as well as leaf primordial isolated from shoot tip (Al-Khayri and Naik, 2017). Meristematic tissue is very responsive to the culture medium. However, in the case of the multiplication of the Dhakki cultivar, shoot tips are the most successful explant with a high capacity for the regeneration of direct shoots. The method of using the inflorescence as an explant has recently become more prevalent, in which case offshoot harvesting is not necessary (Sardar A. Farooq, 2012). This study was done on mature palm trees in full fruit-bearing stage.

## 2.3 Layout of the orchards

The layout of date palm orchards differs across the world, because the reasoning behind it varies. The variation in spacing varies most importantly according to the cultivar along with climatic conditions. Unlike other management practices such as irrigation regimes, the decision on spacing is done once for an orchard and cannot usually be modified in the ensuing years and it must therefore be considered carefully. Spacing is most commonly 10 m × 8 m, being 10 m between rows and 8 m spacing between trees in the rows. A number of farmers apply an 8 m × 8 m spacing, although narrower spacing is generally not practical (Liebenberg and Zaid, 2002). The research partner's subject matter experts (SMEs) suggest a spacing of 8 m × 8 m and 10 m × 8 m and also note that a spacious arrangement could be advantageous because of reduced competition for resources.

## 2.4 Yield factors

Agricultural practices and environmental factors have an effect on the cultivation of the date palm, some of which may happen at the offset of the growing season but the effects of which may only be noticed much later. Globally, dates are cultivated in areas having very long and hot summers with little precipitation and very low humidity, where groundwater is in abundance or where irrigation is possible (Janick and Paull, 2008). Apart from the type of soil and the quality and quantity of water, various other factors play a minor to essential role in the life processes and bearing of fruit of the date palm. These include climate, irrigation, evapotranspiration, mineral nutrition and farming practices such as manual pollination and bunch thinning as well as the application of growth regulators. Subsequently, the climatic requirements of the date palm, irrigation, soil fertilisation, pollination, thinning, growth regulators and pests are discussed.

### 2.4.1 Climatic requirements

The combination of climate factors plays an important role in date palm production, namely temperature, rain, humidity, wind and light. From conversations with SMEs it is believed that the influence of the meteorological conditions of 18 months or even up to two years prior to a harvest has an effect on perennials, specifically palms. Zaid (2002) describes the climatic requirements, irrigation and water as well as fertilisation needs of the date palm. According to Zaid (2002) and SMEs, the following factors have a vital influence on fruit production, development and yield.

#### 2.4.1.1 Temperature

Temperature generally affects plant growth greatly, as is the case of the date palm. According to Zaid and Klein (2002) the date palm can endure large temperature variations (-5 to 50 °C). The fibrillium

and the leaf bases serve as protection for the terminal bud. The temperature of the terminal bud and trunk adjusts slowly to the temperature of the environment and is therefore quite stable. The difference in internal temperature is about 14 °C lower in summer and 12 °C higher in winter (Zaid and Klein, 2002). This enables the palm to resist frost in wintertime and great heat in summer for a number of days, on the condition of adequate water supply. Although normal temperature does not influence normal growth, cool weather may inhibit the ripening of fruit. The date palm requires an average optimal daily temperature for development for early-ripening, mid-season and late-ripening varieties, which from blossoming to fruit ripening is around 21 °C, 24 °C, and 27 °C respectively.

High temperatures over a long period in spring may also influence date development negatively in Deglet Nour which will only manifest at ripening (Rygg, 1975). Temperature also has a substantial influence on fruit quality. In a season where the maximum daily temperature in spring is higher than 37 °C, the Deglet Nour bears dull-coloured, low quality fruit with a dry texture, high sucrose percentage and a high acid content (Rygg, 1975). All dates affected by high spring temperatures are more acid than normal when ripening and are more difficult to hydrate. The suddenness of the onset of high temperatures during the heat-sensitive period may influence the degree of injury to the fruit. Soft varieties are not damaged by extreme heat to the same extent as Deglet Nour, but skin separation in ripe dates may be enhanced by high temperatures or high humidity or both at the beginning of the growing season.

Sub-zero temperatures can be somewhat endured, but the palm only grows in temperatures above 7 °C, it definitely thrives around 32 °C, and growing decreases at 40 °C and higher. When temperatures prevail below zero for a substantial period, metabolic disorders will lead to partial or even total leaf damage. When protoplasma of the palm leaves the cells, it freezes. During defrost, water invades intercellular spaces and causes leaves to become brown, wither and dry out. The impact of this damage depends on how intense the frost is and how long it prevails:

- At -6 °C, the ends of the leaflets become yellow-brown and wither.
- At -12 °C, the leaves on the outside of the crown wither and dry out.
- From -15 °C, the leaves of the middle crown freeze; if extreme temperatures prevail for some time, the central crown itself is affected and dries out.

Extensive leaf damage results in limited flowering the following year, influences the development and ripening of fruit unfavourably and this results in poor fruit quality. Because the growth centre of the date palm tree is largely protected against frost, a damaged palm can start growing normally again the following season. Inflorescences, the clusters of flowers, are also damaged by frost. This can be reduced by covering the inflorescences with paper bags after pollination. (Zaid and de Wet, 2002*a*). Bagging of bunches also decreases the risk of consumption by insects and birds (Al-Yahyai, 2018). Some date palm cultivars are susceptible to cold, among others these are the Amri, Barhee, Beid Hmam, Dayri, Deglet Nour, Khadrawy, Maktoum, Medjool, Menakher and Saidy (Zaid and Klein, 2002).

According to Battel (2017), heat units, a heuristic tool in phenology, are a more reliable method of predicting crop development than calendar days. Crops have a specific threshold temperature and beginning accumulation date. The base (threshold) temperature is defined as the minimum temperature for remarkable development in the crop to be expected. Crops also have an upper-end cutoff temperature, above which development is not greater than below it. There exist three methods for Growing Degree Days (GDD) calculation, namely Temperature Averaging, Baskerville-Emin (BE) Method and Electronic Weather Data Collection. The second, the BE method, involves fitting a curve to the temperature points greater than the threshold temperature and calculating GDDs as the area under the curve. For the third method, electronic devices record measurements every few minutes, in which case the base temperature can be subtracted from each and then the readings can be accumulated (Nugent, 2020). The most generally used method in agriculture is the Temperature Averaging method, making use of the average daily temperature calculated as $\frac{max+min}{2}$ and subtracting the base temperature.

The date producer for whom this study was done also recommended the Heat units equation

$$\text{Heatunits} = \frac{\max(\text{temperature}) + \min(\text{temperature})}{2} - \text{base}(\text{temperature}). \tag{2.1}$$

If the answer is negative, zero is assumed. The threshold temperature of the date palm is 18 °C and the palm requires a minimum of 2 000 heat units from flowering to fruit maturity, where Bhat and Kenna (2013) calculates heat units as the accumulation of maximum daily temperatures minus 18 °C. Heat units, or degree days, required to ripen the fruit vary with cultivars and range between 2 100 for early-ripening and 4 700 for late-ripening cultivars. Heat units stored in springtime impact the time of harvest (Yahia and Kader, 2011).

#### 2.4.1.2 Rainfall

Although there is research done on gene modification and editing to develop more drought-tolerant crops (Spendlove, 2016), precipitation or irrigation plays an essential role in the development of terrestrial plants. The date palm grows in dry to arid regions with hot summers, very little rainfall and extremely low humidity throughout the ripening period. Still, the palm needs plentiful water on a regular basis to thrive. Since these areas naturally receive very little rain, the water supply is supplemented by irrigation, often with water of high salinity.

The main production regions in the Northern Hemisphere receive winter rainfall, allowing desalination which is necessary due to the salt deposited on the soil during irrigation. Since the dates ripen in the summertime, this rain benefits the soils and is not harmful to the fruits. The main date-growing regions in the Northern Hemisphere receive almost no rain until November. Since harvest is normally from the middle of August until the beginning of November any precipitation in the flowering and harvest season may be adverse and may cause damage to pollination and to the fruits.

According to Bhat and Kenna (2013), low humidity and no summer rain contribute to the production of high quality dates. It is likely that rain immediately following pollination washes away some of the pollen applied (Zaid and de Wet, 2002*a*). Rain can also be unfavourable to fruit set[1] when it is accompanied by low temperatures or if colder temperatures occur after that rain. Furthermore, the receptivity of the flower is lower when there is contact with water. Rain damage occurs either when the rain is early or when the fruit ripens late. Rain may cause adverse cutting and cracking particularly in the *kimri* and late *khalaal* stages, so fruit bunches are usually covered by kraft paper. The most sensitive stages are *rutab* and *tamar* since rain and related humidity can cause harm, which includes rotting and fruit drop.

#### 2.4.1.3 Humidity

When discussing humidity, a distinction needs to be made between absolute humidity and relative humidity. Absolute humidity is defined as the amount of water vapour present in the air, expressed in moisture per cubic meter of air ($g/m^3$). Relative humidity is a percentage of the amount of moisture the air could potentially hold, expressed as a percentage, or the ratio of vapour pressure to saturated vapour pressure. If the temperature increases, relative humidity decreases (Khillar, 2019).

Air humidity has a significant influence on the cultivation and production of the palm. Very humid air is beneficial for infections like Graphiola leaf spot (*Graphiola phoenicis*), while other pests, such as the date mite (Bou-Faroua), diminish. In low air humidity fungal diseases are not present, while pest and mite outbreaks still occur (Zaid and de Wet, 2002*a*). High humidity at flowering and pollination has a negative effect by sustaining infection in closed spathes, causing rot. Air humidity has an

[1] Fruit set is the process in which inflorescences develop to become fruit and potential fruit size can be determined. It occurs after fertilisation – male pollen landing on receptive female flower parts (stigmas) and fertilising the eggs contained in them (Growers, 2020).

influence on the quality of the date fruit during maturation. In low humidity, especially when hot winds occur, fruits get physiological disorders, such as drying out, while in high humidity it becomes soft and sticky. In the case of high air humidity cuts or breaks are brought about in the skin during maturation. The fruits fall to the ground and are wasted. This only happens when high humidity occurs right before the *khalaal* stage. In the *rutab* stage the flesh softens, after which the skin is less likely to break in high humidity. However, the fruit absorbs damp and becomes tacky which makes handling more difficult. As seen in the Maktoom cultivar, high humidity shortly before the dates ripen may cause shrivel (Rygg, 1975). When the *tamar* stage is reached, air humidity does not have a substantial effect on the fruit.

#### 2.4.1.4 Wind

The date palm tree is well adapted for wind and shows no damage under windy conditions. It can endure a very warm, strong and dusty wind and the date palm may be utilised to protect other vulnerable crops against wind (Dowson, 1982). The sand and dust in the wind stick to the dates in the rutab and tamar stages.

In most areas in which dates are cultivated, harsh hot and dry winds are characteristic of the final part of the pollination season. This dries out the stigmas and consequently the styles of the female flowers and results in pollen reaching the ovule faster (Reuveni et al., 1986). Wind speed impacts the pollination effectiveness, light wind improving pollination, while strong winds blow away the pollen. Cold winds decrease the pollen germination. Dry, hot winds quicken the maturation process causing the fruit to wither and a white or yellow ring forms at the fruit base. In rare cases, very strong wind could break the rachis of the inflorescence (the fruit stalk), so that water and nutrients cannot reach the fruit and the bunch will die. There also is a possibility that wind carries mites from one palm to the other (Zaid and de Wet, 2002*a*).

#### 2.4.1.5 Light and photosynthetic performance

The growth of the date palm is hindered by violet and yellow light rays but improved by light at the red end of the spectrum. Mason (1925) states that red light stimulates photosynthesis the most. Although clouds decrease light intensity, the sky is unclouded in the date growing areas during the ripening period (February to May in the Southern Hemisphere and July to October in the Northern Hemisphere).

Photosynthetically active radiation (PAR) constitutes the fraction of sunlight with a wavelength range between 400 and 700 nm (blue to red spectrum of light), that can be used as the source of energy (photons) for photosynthesis by green plants usually expressed in $\mu$mol (photons) $m^{-2}s^{-1}$. The light reactions of photosynthesis are improved by increases in PAR. The photosynthetic response to different levels of PAR varies with plant species and leaf position. Compared to sunlit leaves, shaded leaves harvest PAR more efficiently at low light levels but less efficiently at high light levels. PAR changes with the seasons and varies depending on the time of day and the latitude. Photosynthetic photon flux density (PPFD) is the quantity used to measure the amount of incident light received by the plant and is the number of photosynthetically active photons in the 400 — 700 nm waveband incident per unit time on a unit surface, $\mu mol/m^2.s$ where $1mol = 6.023.10^{23}$ photons.

The quality of light has a significant influence on plants, particularly on their photosynthetic performance, resulting in vegetative and reproductive growth. Photosynthesis is negatively affected in parts of plants which are shaded by the rest of the canopy or by adjacent plants and the radiation that reaches the surface of the shaded leaves is solar diffuse radiation – diffuse photosynthetically active radiation (DPAR). According to Gurrea-Ysasi et al. (2018) some of the parameters used to determine the effect of solar diffuse radiation on photosynthesis are photosynthetic photon flux density (PPFD), the relationship between the blue region and the red region of the spectrum (B/R ratio) and the amount of infrared radiation compared to the red part of the spectrum (R/IR ratio). The ratio between B

and R varies between 0.5 for direct radiation and 0.95 for diffuse radiation. The filtering effect on incident solar radiation caused by the canopy or the presence of adjacent trees has been determined by calculating the PPFD Ratio between the shade and the sun (ER) from $ER = PPFD_{shadow}/PPFD_{sun}$. The trend of this ratio remains constant over the red to blue spectrum and the maximum filtering value of the incident direct radiation (which corresponds to a lower value of ER) occurs in a grove. It was also found that on a clear day parts of plants that are in the shade have even less PPFD to perform photosynthesis than on a cloudy day. It is therefore conclusive that trees growing in the centre of a grove have lower photosynthetic performance, growth and production than those on the edge, which receive more direct sunlight.

Gurrea-Ysasi et al. (2018) found that the solar zenith can reach around 70° in the afternoon, so that the rays of the sun strike very obliquely on the surface of the earth. For the date palm, radiation in the afternoon is similar to the radiation amount in the case of full sun. The ER of the date palm is higher (with factor 4.3) at 16:00 than any other time of the day. Trees planted on the edge of a grove or orchard will receive more PPFD and with higher productivity than in the centre of the orchard. This will be much more evident on the western edge in the Southern Hemisphere, because of the extended direct PAR as well as the lower filtering effect.

### 2.4.2 Irrigation

The date palm needs plentiful, normal quality water to produce a proper yield and water requirements differ depending on the location and cultivar. Growth, fruit production and yield are substantially influenced by the quantity and quality of available water and the water content of the soil. These water variations could result in a yield decrease or even failure in fruit production. According to Zaid and Klein (2002) the salinity level of the water must not exceed 5 – 6%. Adult date palms can however survive higher levels of around 9 – 10%. This fair tolerance to high salt concentration in the soil is ascribed to the ability of the date palm roots to exclude the salt (Rygg, 1975).

Normal growth of date palms depends on a well-drained soil. Although flooding can be resisted, irrigation must always be complemented with drainage. Conversely, in the case of inadequate drainage, and if the leaching of soluble salts is incomplete, high rates of evaporation are inclined to raise the salt concentration in the soil and surface water. Even though the palm is fairly tolerant to salinity, it impairs the absorption of water by creating adverse osmotic potential conditions in the soil solution around the root zone.

Date palms do tolerate different soil types, such as clay, loam and sandy soil. However, ideally the soil types should include free-draining sands or sandy loam with low levels of salt and good capacity for moisture-holding. In areas with minimal rainfall, high temperatures and evaporation, flood water or irrigation evaporates quickly, leaving the salts behind on the surface of the soil.

Transpiration is when liquid water evaporates from plant tissue and that vapour is lost to the atmosphere, mainly through the stomata, the miniscule openings on the leaf. Transpiration rate is dependent on many factors such as vapour pressure, energy supply, soil water content, waterlogging, soil water salinity and wind. Different crops have different rates of transpiration, as this rate also depends on crop characteristics and cultivation practices. Evapotranspiration (ET) is the effect of evaporation from the soil in combination with transpiration from the plant. For effective planning of irrigation, determining the evapotranspiration is essential when designing the irrigation system and deciding on an irrigation schedule.

The rate of evaporation is measured in millimetres (mm) per unit of time by the Penman-Monteith equation. Growing in different ranges of harsh climatic conditions, the date palm, having high evapotranspiration, needs large volumes water.

The Penman-Monteith equation is given by

$$\lambda \mathrm{ET} = \frac{\Delta(R_n - G) + \rho_a c_p \frac{(e_s - e_a)}{r_a}}{\Delta + \gamma(1 + \frac{r_s}{r_a})} \tag{2.2}$$

where
ET is evapotranspiration $[mm.day^{-1}]$,
$R_n$ is net radiation $[MJm^{-2}day^{-1}]$,
G is soil heat flux $[MJm^{-2}.day^{-1}]$,
$(e_s - e_a)$ represents vapour pressure deficit of the air $[kPa]$,
$\rho_a$ is mean air density at constant pressure $[kg.m^{-3}]$,
$c_p$ is the specific heat of the air $[MJkg^{-1}{}^{\circ}C^{-1}]$,
$\Delta$ is the slope of the saturation vapour pressure temperature relationship $[kPa^{\circ}C^{-1}]$,
$\gamma$ is the psychrometric constant $[kPa^{\circ}C^{-1}]$ and
$r_s$ and $r_a$ are the bulk surface and aerodynamic resistances $[s.m^{-1}]$.

Reference evapotranspiration ($\mathrm{ET}_0$) is a representation of the rate of the evapotranspiration from a reference surface, which could be a water surface, but is generally a uniform grass field. The reference crop is well watered and grows under optimal agronomic conditions, covering the soil completely. The $\mathrm{ET}_0$ can be calculated using the Penman-Monteith equation incorporating meteorological data. This equation estimates grass $\mathrm{ET}_0$ at the evaluated location, considering physiological as well as aerodynamic parameters.

During a study on water usage by young date palm trees, it was found that evapotranspiration of a tree in one year was approximately 1 857 mm, with a daily average of ETc = 5.08 mm per day (Abdelhadi et al., 2020).

The mature palm tree at the age of 10 years requires between 1 500 mm and 2 800 mm water annually, implying an irrigation requirement of up to 27 000 $m^3$/ha (123 trees/ha) (Bhat and Kenna, 2013).

In order to conservatively use a valuable resource, it is imperative to determine the right water quantity necessary for a region, if possible for a specific plantation.

The efficiency of applied water may be calculated as follows:

$$\mathrm{WP} = \frac{\mathrm{Yield}}{\mathrm{AW}} \tag{2.3}$$

where
WP is water productivity,
Yield is measured in t/ha
and AW is total applied water (rain and irrigation) in mm (Wichelns, 2014).

Global yearly date palm irrigation is shown in Table 2.1:

Water requirements are influenced by various factors. In order to determine the required water volume, Liebenberg and Zaid (2002) suggest that the following be taken into consideration:

1. Soil Quality: Adaptation could range from extremely sandy to clay soil. The quality of the soil corresponds to the soil's drainage capacity. For optimum soil conditions, water must penetrate to 2 m or deeper under the ground.
2. Soil salinity: In saline conditions, more water is necessary to enable a leaching process which removes all the salt from the soil.
3. Temperature: Higher temperatures are directly related to a higher rate of evaporation. In higher temperatures the plant will then require more water.

Table 2.1: Global date palm irrigation (Liebenberg and Zaid, 2002)

| Location | Quantity ($m^3$/ha/year) |
|---|---|
| Algeria | 15 000 – 35 000 |
| California, USA | 27 000 – 36 000 |
| Egypt | 22 300 |
| India | 22 000 – 25 000 |
| Iraq | 15 000 – 20 000 |
| Jordan Valley, Israel | 25 000 – 32 000 |
| Morocco | 13 000 – 20 000 |
| South Africa | 25 000 |
| Tunisia | 23 600 |

4. Humidity: The quantity of water required by the plant depends on the humidity of the soil – the lower the humidity levels, the more water is required by the plant.
5. Wind: Evaporation is increased by air movement. Strong constant wind causes a higher evaporation rate and this will result in increasing water demands.
6. Cloud cover: Since the heat of the sun increases evaporation, cloud cover plays a major role since more water is required when cloud cover decreases.

#### 2.4.2.1 Root drenching

Irrigation must be delivered where it is accessible to the roots of the plant. Four soil layers of similar depth are found. In the top layer 40% roots flourish, while 30% are found in the second and 20% in the third layer, only 10% roots grow in the bottom layer. Generally the roots of mature date palms are approximately 5 m deep, growing within a 3 m radius around the trunk of the tree. In terms of water extraction, 40% of the plant's total water need is accessed up to a depth of 50 cm, 70% is extracted from 100 cm below surface, and 90% from up to 150 cm. Only 10% of the water required by the plant is extracted from the last layer which is 150 cm and deeper. The depth of the roots of young date plantlets can fluctuate between 25 and 50 cm and the radius of the roots from 10 to 30 cm (Liebenberg and Zaid, 2002). Due to the shallow root depth a regular irrigation schedule is necessary. The required irrigation frequency depends on the type of soil. During the first summer of a young plant in the case of very sandy soil, daily irrigation is necessary.

Applying the water at the right time plays a vital role. The soil type has a certain characteristic water retention, resulting in the volume available to the palm. These two factors as well as the usage of water by the palm on a daily basis, assists in establishing when the next irrigation cycle is due. From light, medium to heavy soils, the mean values of available water are 100 mm/m, 140 mm/m and 180 mm/m respectively. Taking daily water usage of the date palm, available water in the soil and rooting depth of the palm into account, the cycle period for irrigation can be calculated.

Irrigation methods vary from traditional to modern. Basin, micro and drip methods are commonly suggested. Water should be applied to reach the top soil level only in order to enhance proper date palm root development.

#### 2.4.2.2 Irrigation methods

According to Liebenberg and Zaid (2002) the following different irrigation methods are applicable to irrigate crops.

1. Flood irrigation: Flood irrigation is the oldest and the most common irrigation method. Flood irrigation is easily carried out and is inexpensive. Unfortunately, this method is labour intensive and wastes much water, especially in sandy soil.

2. Furrow and basin irrigation: For this method, a smaller contained area is flood irrigated. This is more efficient than traditional flood irrigation. Drawbacks are that the method requires labour and might interfere with other mechanical operations.
3. Sprinkler irrigation: This method is more economical; it is the first modern irrigation method, increasing efficiency and enabling automation. It saves water and scheduling is easily managed. Disadvantages are expensive installation and maintenance and the method is not suitable in windy conditions.
4. Micro irrigation: Micro irrigation, also known as localised irrigation, was more recently introduced. Although it is expensive to install and requires good quality clean water, running costs are lower than sprinkler systems. Water usage is much more efficient as water is applied to the roots only and evaporation is reduced. It is less labour intensive, easy to automate and manage.
5. Drip irrigation: This modern irrigation technique was created in Israel because of water shortage. It has all the advantages of micro irrigation but has even more effective use of water and a lower running cost. Drip irrigation cannot be influenced by wind and requires very little labour. Installation is costly and the water has to be very clean. The challenge is to establish the optimum amount of water to be applied by the system.

Micro irrigation is recommended for dates because they are usually grown in sandy soils and because of the efficiency of this irrigation type. In the case of the immature date palm, water should not be sprayed on its crown. Micro irrigation with a 300° to 320° spray pattern is effective for small plants. The statute of micros can be optimised to ensure 100% coverage of the rooting region of immature date palms. Generally, from the planting of the palm to its fourth year, the covered area is approximately 12 $m^2$ and the water flow rate 96 $l$/h/palm. Approximately 18 $m^2$ area should be covered during the fifth to tenth year and the flow rate must increase to 104 $l$/h/palm. An area of 28 $m^2$ is covered from the tenth year and the flow rate should be 156 $l$/h/palm. Although more water enables leaching which the palm finds beneficial, to cover larger areas during the first years will waste water. More frequent irrigation during the initial period is beneficial to the shallow roots.

### 2.4.3 Soil fertilisation

Adding nutrients to the soil of date orchards is essential for the following reasons:

- To overcome nutrient deficiencies of the soil.
- To enhance appropriate forming, growth and development of the palms.
- To enhance the date palm yield capacity.

#### 2.4.3.1 Functions of nutrients

The nutrient elements in the soil have various functions. The fertiliser needs of the date palm are comparable to other cultivated crops (Klein and Zaid, 2002). Required in differing quantities, elements vital for plant growth and production include nitrogen, iron, magnesium, potassium, zinc, boron, calcium, sodium, chlorine, cobalt, copper, sulphur, manganese and phosphorus.

Plant life processes, in particular vegetative growth and photosynthesis, are important factors in producing a high yield. Nitrogen is an essential nutrient in this regard. Processes such as respiration, reproduction and the maintenance of the genetic identity, cannot occur without phosphorus. Linked to cell division, phosphorus also assists root development and flowering. Being in cell sap, potassium assists in the promotion of photosynthesis and transport of nitrogen in the plant. Potassium facilitates fibre strengthening and has an influence on the opening and closing of the stomata. It plays a role in the plant's resistance to drought and cold and increases fruit quality. For the improvement of the quality and yield of fruit, mineral nutrients such as potassium sulphate are applied to the date palm.

Most of the necessary micronutrients are found in the irrigation water. Klein and Zaid (2002) states that, in order to identify deficiencies, a simulation study based on leaf and soil analysis can be conducted.

Two micronutrients, boron and manganese, are essential. Boron plays an important role in pollination and reproduction processes such as the development of flowers and fruits. Boron also regulates the uptake of calcium, magnesium and potassium and plays a role in protein synthesis as well as cell division. Boron is responsible for an increase in cell membrane penetrability and assists with carbohydrate transport. Boron also plays a role in the synthesis of lignin. Spraying the bunches with potassium and boron will also improve fruit set, as well as the yield and quality.

Manganese initiates many chemical reactions of enzymes as well as certain physiological reactions. Since manganese plays a role in cell respiration, it also catalyses enzymes involved in the metabolism of nitrogen and chlorophyll synthesis.

#### 2.4.3.2 Application of fertiliser

The time of fertiliser application and application method are also determined meticulously. Two growth phases typically characterise the date season, *i.e.* vegetative and reproductive phases. The reproductive phase consists of two stages, the flower formation stage, followed by the fruit development stage. When fertiliser application is done according to the timing of these phases, the number of well-developed flowers and potential yield increase. Applying fertiliser directly after flower and fruit formation have commenced causes the best results. Klein and Zaid (2002) therefore recommend that the applications be done during June (flower formation) and November (fruit development) for the Southern Hemisphere.

It is important not to apply nutrition elements as foliar spray, because this can cause burn. In plantations where fertilisers cannot be supplied through the irrigation system, manual application is used. Small quantities of fertilisers are then applied by hand to each tree. In manual application, a uniform fertiliser allocation within the palm drip area but not near the palm base, must be ensured. This is time-consuming, requires labour and has the hazard of root burn if not evenly distributed.

The irrigation system may be utilised for fertiliser application and this method is called fertigation. Only soluble fertilisers are directed through the system while non-soluble nourishment still has to be done manually. The benefit of fertigation is that a suitable amount of fertiliser gets evenly applied within the irrigated area.

### 2.4.4 Artificial pollination and techniques

Pollination, when male pollen grains set down on female stigma, is one of the fundamental processes determining the fruit bearing and reproduction of plants. Since the date palm is dioecious, trees bear unisexual flowers, either male, producing pollen or female, producing fruit. The inflorescence consists of a spathe of floral spikelets, short in male, while long and slender in female inflorescences. A full-grown female palm tree produces between 15 and 25 spathes that contain 150 – 200 spikelets each. The pale white male inflorescences are borne singly, while the yellow female inflorescences are borne in clusters of three (Zaid and de Wet, 2002*c*).

In date palms natural pollination occurs by wind and insects. In order to ensure effective pollination that ensures successful fertilisation in a commercial orchard, the pollination process is done artificially by hand or mechanically. Pollen is taken from the male tree and applied to spathes on the female trees. Selecting good quality pollen is important as the type of pollen parent will determine the fruit size, time of ripening and the chemical composition (metaxenia) of the fruit. Seedling males differ to a large extent in their growth vigour as well as their spathe features and pollen quality. The amount of pollen grains in a spathe differs from 0.02 g to 82.29 g per spathe and it also contains different sizes of grains with mean diameters from 16 to 30 microns. Effective pollination can be implemented

two or four days after the female spathe has opened. In the Northern Hemisphere the ideal time for pollination is during March and April, whereas July and August are normally ideal for pollination in the Southern Hemisphere. The cultivar of the date and characteristics of the season could advance or delay the opening of the flowers.

A mature male palm tree produces between 500 g and 1 kg of pollen (an average of 700 g), while 15 to 20 g of pollen is needed per female palm (around 2 kg per hectare). One male tree supplies enough pollen for pollinating 47 female palms. A fruit set is determined by the viability of the pollen as well as the temperature while pollinating. A daily temperature of 23.9 °C – 26.2 °C is optimal during pollination. A fruit set of 50 – 80% is adequate to ensure a good crop. If the temperature is too low during pollination, the flower cluster can be covered with paper bags (Lobo et al., 2013).

The time of flowering of the male and female palms should be coordinated so that ample pollen is available when the female spathes open. In the event of the male spathe opening two to four days before the female spathe, pollination can take place effectively.

The pollination technique depends on the pollen type available, but one of these is used: fresh male strands, pollen suspension or dried pollen (Zaid and de Wet, 2002*c*).

1. Fresh male strands: The most widely implemented pollination practice is cutting the strands of male flowers from a newly opened male spathe and placing two or three of them inverted and lengthwise between the strands of the female flowers. Different date varieties need different amounts of pollen. The number of male strands necessary for pollinating a female spathe varies from 1 to 10 depending on the date cultivar. The female flowers of some varieties are larger than others, which result in the requirement for more male strands.
2. Pollen suspension: The male spathes are harvested between one and two days after opening and placed in a dry, shaded area for drying. Before it is stored, moist pollen must be dehydrated fast and efficiently. Strands are detached and stored in paper bags until needed for the pollination of female flowers. When pollen is subjected to sunlight or heat, it will deteriorate rapidly and lose its vitality, the ability of the pollen grain to germinate and develop. A pollen grain suspension, consisting of 10% sucrose and 20 ppm GA3-hormone could be used for pollination (Zaid and de Wet, 2002*c*). Regarding fruit setting, pollination sprays were found to be as effective as pollination by hand.
3. Dried pollen: This method of pollination is more efficient, allowing proper use of the pollen and effective control of the timing of pollination. The dried pollen could have its origin from the previous season, from early ripening males of the same season, or from new male flowers. Many techniques exist for the application of dry pollen. The most familiar technique is dusting dry pollen on small cotton pieces and placing one or two pieces between the strands of female inflorescences. The puffer, a small manual insecticide duster, can be used for the application. Another technique is mechanical pollination. A special machine and pollinator are used to pollinate newly opened female spathes or spray the whole female leaf canopy right above the open spathes from the ground. In order to reach the inflorescences in high palm trees, the machine has an elevated platform. For this technique around two to three times more pollen is needed than in manual pollination. Fillers are characterised as follows: the particle size of the filler must be the same as the pollen grain and it must have no influence on the viability of the pollen or the germination on the female stigmates. Typical fillers are industrial talc, wheat flour and walnut-hull dust. Fillers are mixed with the pollen in a pollen/filler ratio suitable for the palm cultivar. The Medjool cultivar can be pollinated with a small quantity of pollen (10% pollen/talc ratio) (Zaid and Klein, 2002). Fruit set following mechanical pollination is generally not as successful as after manual pollination, but fruit quality and yields are comparable since the thinning of the inflorescences determines the final outcome of production.

According to Zaid and de Wet (2002*c*), some advantages of mechanical pollination are:

- Less labour with little training is necessary and it is a relatively quick process, making this a relatively low-cost pollination technique.
- A palm can be pollinated a number of times in a short period.
- A pollen mixture from different sources can be used, ensuring viable pollen and efficient fertilisation.
- Since labourers do not have to climb the high palm trees, accidents are less likely.

Chao and Krueger (2007) state that when growing naturally, some fruit falls from the tree, termed fruit drop, 25 – 35 days after the splitting of the spathe. A second fruit drop can occur in some varieties 100 days after the opening of the spathe.

### 2.4.5 Bunch removal and thinning

The quantity and quality of fruit bearing per date palm tree can be regulated by controlling the number of fruit bunches per tree as well as by controlling the number of individual fruits per bunch. The quality of fruit is usually a crucial element in marketing, determining the demand for and the price of the product. The practice of regulating fruit quantity can take place at pollination time but it may be necessary to repeat it six to eight weeks after pollination (Zaid and Klein, 2002). Regulating the number of fruits has different consequences on the date palm yield bearing ability to a different degree in various cultivars. Thinning is beneficial in many ways and although laborious, it is common practice in the cultivation of perennials for the positive effects it has. These are:

- Thinning improves the quality of the date fruit and it leads to an increase in fruit size, mass and flesh percentage, resulting in better marketability.
- Because of better quality, less fruit is spoiled, especially in humid areas where thinning ensures better ventilation in fruit bunches.
- Thinning minimises wear and deformation of fruit, increases the ripening process, improves colouration, fruit chemical properties and nutritional content.
- Thinning is also vital in reducing incidence of alternate bearing and regulates annual fruit production by ensuring adequate flowering in the following year.

Two types of thinning are under discussion:

1. Bunch removal: The date palm shows a correlation between the number of leaves and fruit bunches for ideal yield. Increasing leaf/bunch ratio results in increasing yield with an ideal ratio of 10 leaves per bunch. By removing fruit bunches the total fruit bunches per tree is limited to the accepted standard depending on the age and health of the palm. Bunch removal is uncomplicated and has a significant effect on the total yield. Remaining bunches are arranged according to layers to control fruit bearing and harvesting.
2. Bunch thinning: Several thinning practices are performed in date cultivation of which the manually removal of individual fruits from each strand has a significant effect but is an expensive method. Another thinning practice is the removing of some complete strands from the fruit bunch by cutting them out 5 – 6 cm from the bases. The number differs according to market demand or agricultural custom but in general one-third of the strands in the centre of the bunch are removed. Bunch pruning can also be done by shortening the length of the strands by a third of the end of the bunch during pollination or in Kimri phase.

### 2.4.6 Growth regulator treatment

Endogenous plant hormones called *gibberellins* (GAs) regulate basic aspects of growth and development such as seed germination, shoot growth, flowering, and fruit development. Artificially manufactured

chemical growth regulators are directed to plants in order to change the working of natural hormones or even to substitute the natural hormones. Growth regulators require repeated application to be effective and results are not always predictable, but the advantages are significant in fruit quality and yield (Bhattacharya et al., 2012).

Various growth regulators are used in date palm cultivation, of which Auxins and certain Gibberellins are well known. Treatment of the date palm with Indole acetic acid (IAA 10 ppm), Gibberellic acid (GA3 10 ppm), 6-Furfurylamino purine plus Gibberellic acid (BA+GA3 10 ppm) undergoing normal pollination highly increases the mass and length of the fruit and results in better fruit set. The growth regulator response varies between the varieties (El Hodoairi et al., 1992). During the early fruit development stage, cell division in the embryo happens at a fast rate. The fruit grows bigger because of the cells that enlarge rapidly. By spraying the fruit bunches with Auxins and Gibberellins the fruits size increases and ripening is delayed. Preharvest drop of fruit can be prevented by applying the growth regulator 2,4D and GA3 around 40 to 70 days after pollination (Lobo et al., 2013).

Treatment with growth hormones requires labour and time and adds to production cost. There is also an environmental risk involved.

### 2.4.7 Pests and diseases

Plant growth and yield are affected by biotic factors such as pests and diseases, in some regions more than in others. Southern Africa is barely affected by these biological hazards although in some areas of the world the yield can be reduced by up to 90% under severe infestation (Ahmed, 2007). The severity of infestation by these pests are influenced by the locality, cultivar, environmental conditions and management practices. According to Howard (1999) pests inhabiting the palm may infest the leaves, causing degeneration, and include larvae of certain insects, beetles, grasshoppers, true bugs such as the well-known Royal palm bug and mealybugs, aphids, white flies, greenhouse thrips and sale insects. The date palm hosts at least 12 types of armed scale insects of which the Green pit scale insect is the most common. The African Palm Weevil and Red date palm weevil cause significant damage to yield in almost all the date-growing areas of the world except for the southern part of Africa. Larvae of several species Rhynchophorus bore into the stems, palm buds and petioles. They cause damage to the tissue and sometimes assist a nematode in causing red ring disease. Some caterpillars inhabit the date palm flowers and fruit, causing wilting, yellowing, malformation of fruit and this leads to fruit drop and losses in production.

Different ways to control pests should be integrated in a management programme, such as:

- physical removal of the severely infested leaves;
- raising the soil around the palm to facilitate irrigation;
- pre-watering and frequent irrigation;
- application of chemical insecticides; and
- chemical insecticides may be hazardous by leaving chemical residues in water and the environment and by killing the natural enemies of pests. Botanical pesticides are the safer alternative but require proper pest population monitoring, and the use of resistant cultivars and biological control agents (El-Shafie et al., 2015).

According to Zaid et al. (2002) various diseases occur on the date palm worldwide which may be divided into major groups:

- Fungal diseases such as Bayoud disease, Black scorch disease and Brown spot, all cause whitening, browning or a scorched, charcoal-like appearance on leaves, stems, petioles or buds. Decay is most severe when it attacks the terminal bud, causing the death of the palm. Fungal diseases spread rapidly to epidemic proportions and are a major cause of reduction in yield. They also

have an epidemic aspect. Bayoud, for instance, has destroyed more than 12 million palms in Morocco and three million in Algeria in one century.

- Phytoplasmic diseases such as Lethal Yellowing, Al Wijam and Brittle Leaves Disease. Chlorotic striping and drying of the tips of the leaves are the first symptom of Brittle Leaves Disease. Yields drop significantly as the retardation in the growth of the terminal buds becomes evident. Lethal Yellowing causes flowers to drop, followed by a rapid yellowing of leaves and the death of the palm. Retardation in terminal bud growth after the occurrence of the disease have symptoms of rosette disease.
- Other diseases of unknown origin but with widespread occurrence include Bending Head, Dry Bone, Faroun Disease, Black Nose and many more. All these diseases have a detrimental effect on plant production and yield.

## 2.5 The date market

This section covers the date market. Topics discussed are of secondary importance but are included for the sake of completeness. The reader may skip this part without loss of continuity.

According to Transparency Market Research (2018), the global date market is segmented according to the supply region; namely Africa, Asia Pacific, Europe, Latin America, the Middle East and North America. The main market share in the global market consists of the Middle East and Africa in terms of value as well as volume. Tunisia holds the largest international market share in dates, around 20% every year, while Israel has just more than 10% share. Algeria and Egypt hold between 3.6 and 3.8% of the market. Consumption per capita in Oman exceeds that of the rest of the world with 68 kg per capita annually. Saudi Arabia has a yearly consumption of 34 kg per capita, while the per capita use for the rest of the world is currently much lower (Dhehibi et al., 2018).

The date market is divided according to the nature of the fruit, type or cultivar, end use, the form of the fruit and the region of production.

Consumable dates are divided into organic and conventional types. The conventional dates currently hold the biggest market share in that larger volume is sold and the total value sold is higher. The organic date market is expanding rapidly, especially in the USA and Europe, because of the growth in demand for organic food, *i.e.*, food free from chemicals, fertilisers and insecticides. The organic date also has a higher nutritional value and tastes better. Consumers also tend to prefer certified organic products (Transparency Market Research, 2018).

The leading high-value date varieties in the import markets are *Deglet Nour* and *Medjool* (or *Medjoul*), followed by *Barhi*, *Khalas*, *Fard* and *Zahidi*. Since the global market demands higher-quality dates such as *Medjool* or *Deglet Nour*, the competition for a share of the market is high (Mbaga, 2012*a*).

The Transparency Market Research (2018) report states that the global market for date fruit is segmented in terms of form into raw and processed dates. Dates are processed into purees, paste and syrups and may also be dried.

The global market for date palm products is catalogued into food service, health supplements, personal care, cosmetics and households. Transparency Market Research (2018) projects growth of 3.9% in the dietary supplements segment. There is a growing demand for healthy snacks and confectionery. Nutrition bars which contain dates as their main ingredient get more popular as the date contains dietary fibre, carbohydrates and many essential nutrients. Date-based purees, syrups as well as dried dates are used as one of the main ingredients in health products. There also is an increasing demand for date confectionery especially in the Asia Pacific and European markets. The marketing has also shifted focus from the Middle East theme to target health-conscious consumers. This is clearly seen in the United States, Bard Valley, where the marketing of more than a million dollars a year focused on creating the exciting new brand name, namely 'Nature's Delight' and marketing dates as the 'power

fruit' from nature. It influenced the country's date sales, which increased by 50% from 2010 to 2015 (Kayal, 2015).

### 2.5.1 Global yield

The Middle East is the world's biggest date producer with nearly 75% of the Arab Region planted with date palm in 2016, supplying approximately 77% of the global date production (Dhehibi et al., 2018). While the Middle East and North Africa boast the greatest production of dates (exceeding six million tons in the Arab countries), people in countries around the world enjoy this sweet fruit. In America, about 33 000 tons of dates are produced annually (Kayal, 2015).

According to Nagini et al. (2016), yield in agriculture is expressed as a degree of the yield per unit area of cultivated land and the seed production from that crop. The highest average yields (from 2000 to 2016) are recorded in three Arab states of the Persian Gulf, namely Kuwait (22.03 tons/ha), Qatar (11.13 tons/ha) and Oman (10.34 tons/ha). The yield of the date palm tree differs within countries, and many factors play a role including the cultivar, agro-ecological systems, and farming practice. The average global yield is 6 tons/ha (Dhehibi et al., 2018).

### 2.5.2 Date harvesting

Harvesting is seen as a critical function in the value chain because of its high impact on the date quality and eventually the final price. Picking is done manually or with mechanisation, especially for big farms. Mounted ladders or lifts including squirrel or self-propelled elevating platforms are used to lift harvesters up to the fruit. The harvesting process must be clean and faultless.

Date fruits are harvested and marketed at three of the development stages, namely during the sweet *khalaal*, the *rutab*, and the *tamr*. The decision to harvest at a specific stage is dependent on the different characteristics as well as the weather conditions and the market demand (Glasner et al., 2002).

- Fruits at the *khalaal* stage are physiologically mature, crisp and hard, bright yellow or red and perishable with moisture a content of 50 – 85%. They are consumed as fresh fruit.
- In the *rutab* phase dates are perishable and partially browned with a reduced moisture content of 30 – 45% and soft fibres.
- Dates harvested in *tamr* are amber to dark brown in colour, with an even lower moisture content of 25% and less. Their texture varies from pliable and soft to firm and hard. They are relatively safe from pest infestation and can be stored over long periods without any special conditions.

The degree of perishability increases from *tamr* to *rutab* to *khalaal*. Therefore, the level of processing necessary to prolong the storage ability is highest for *khalaal* and lowest for *tamr*. In the Northern Hemisphere harvesting is executed in August for *khalaal* and September to December for *tamr* and *rutab* stages. In the Southern Hemisphere this takes place in February and March to April respectively.

The *khalaal* usually find a ready market, being the first date fruits arriving at the market at the beginning of the harvesting season. Because of their perishability, the markets for sweet *khalaal* dates are normally local, especially in places with insufficient transport and logistics (Mbaga, 2012*b*). Some dates can be ripened to *rutab* stage after they have been harvested by quick freezing or by putting them under very low temperature of -18 °C or lower for at least 24 hours. They can also be drenched with acetaldehyde or ethanol gas.

When fully ripe, dates have high levels of sugar and low tannin content. In dry and semi-dry varieties sucrose is the main sugar, while fructose and glucose are the most dominant in soft varieties. They should contain less than 30% moisture at harvesting. Proper timing of harvesting can improve

date quality by preventing cracking of the skin, excessive dehydration of the fruit and insect infestation and damage caused by microorganisms (Kader and Hussein, 2009).

### 2.5.3 Post-harvest handling

This subsection discusses post-harvest handling concepts, including sorting, cooling and storage.

#### 2.5.3.1 Sorting and grading

After harvesting, dates are prepared for efficient marketing and processing by sorting and grading. This improves the market quality, reduces post-harvest losses and enables the product to meet the standards required by the customers. The level of effort going into the harvested product is to a large extent directly correlated to the quality of farming practices during the development and ripening of the fruit, such as the covering of fruit bunches with paper bags to avoid pests, rain and dust. Practices such as bunch thinning also play a role by reducing compactness of bunches and increasing size as well as quality of the fruit. The quality depends on many attributes. The most important are the fruit size, colour, texture (chewiness), cleanliness, skin defects, insect damage, decay or skin separation. The harvested fruit is initially sorted to remove foreign materials and defective or unpollinated dates. It may be cleaned by removing dirt, dust, and other foreign material by means of air pressure, washing with water and air drying to remove surface moisture (Kader and Hussein, 2009). Sorting and grading is subsequently performed according to quality and size.

Lobo et al. (2013) states that sorting of the fruit by hand is done on different criteria such as:

- physical size of fruit;
- physical or physiological disorders of the fruit, like colour darkening, skin puffiness and sugar spotting (sugar crystallisation below the skin)
- pathological disorders like decay (yeast infection), fungi, souring by bacteria; and
- damage done by insects and nature.

Several international Grade standards exist of which the CODEX-standard is generally applied to dates:

1. Size: Dates may be sorted according to size. According to the CODEX Standard a single date should have a minimum fruit size of 4.75 g (unpitted). According to Table 2.2, size categories are determined by the number of dates per 500 g.

Table 2.2: CODEX size grading system with numbers given per 500 g (Kader and Hussein, 2009)

| Size | Unpitted number | Pitted number |
|---|---|---|
| Small | More than 100 | More than 110 |
| Medium | 80 to 100 | 90 to 110 |
| Large | Less than 80 | Less than 90 |

   In the United States of America Medjool dates are graded according to the four grades based on the size of the fruit and fruit defects, as seen in Table 2.3.

2. Quality: Dates may be sorted according to the quality of the dates. According to CODEX standards, the definition of defects is as follows:
   - Blemishes – discolouration, dark spots, scars, surface appearance affecting an area with a diameter larger than 7 mm.

Table 2.3: Medjool grading system in USA (Kader and Hussein, 2009)

| Grade | Dates per kg | Description |
|---|---|---|
| Jumbo | 35 – 42 | No skin marks, skin separation or dryness |
| Large | 44 – 51 | No skin marks, skin separation or dryness |
| Extra Fancy | 44 – 53 | Few skin marks, all sizes together |
| Fancy | 44 – 57 | Some dryness and skin separation, all sizes together |

- Damaged – mashed or torn flesh exposing the pit detracting from the visual appearance.
- Unripe – light in mass or colour or shrivelled flesh or flesh of rubbery texture.
- Unpollinated – unpollinated dates are identified by thin flesh, immature characteristics or no pit.
- Dirt — sand, organic or inorganic material present with a diameter greater than 3 mm.
- Insects and mites – insect or mite damage or contamination by mites, dead insects or insect excretions.
- Scouring – breakdown of sugar into acetic acid and alcohol by yeast and bacteria.
- Mould – visible presence of mould.
- Decay – some decomposed flesh.

#### 2.5.3.2 Hydration

When dates have been harvested at the correct dryness they do not require hydration, but it might be necessary to hydrate them in order to soften the texture. They are submerged in very warm water or steamed at 60 - 65 °C and 100% humidity for four to eight hours. This changes dry dates into soft, glossy fruits. Air circulation can also be used to improve the process.

#### 2.5.3.3 Pasteurisation

It is possible to pasteurise date fruit by heating it at 72 °C and 100% humidity air until it cools down to 66 °C, and keeping it at this temperature for one hour. This process may, however, change the colour of the dates (Kader and Hussein, 2009).

#### 2.5.3.4 Cooling and storage

Dates may be cooled before or after packing to below 10 °C (preferably 0 °C) at a relative humidity of 65 – 75% before transportation or storage.

There are many factors to take into consideration for storage (Kader and Hussein, 2009). The most prominent are:

- Moisture content of dates must be controlled by artificial drying. The storage potential doubles for every 5% reduction in moisture content.
- Storage relative humidity is very important and should be controlled to be in equilibrium with the moisture content of the dates.
- Oxygen concentration in storage is very important. Storage in low oxygen (lower than 0.5%) atmospheres prevents insect infestation (Kader and Hussein, 2009).
- Storage temperature must be controlled — the storage potential of dates doubles for every 5 °C reduction in temperature.

- Effective insect control by disinfestation with ionising radiation at 0.75 to 1.0 kGy.
- Fumigation to eliminate insect pests is done with methyl bromide, although due to environmental concerns, alternative chemicals and physical control methods are being implemented (Chao and Krueger, 2007).

At relative humidity of 75% semi-soft dates (Deglet Nour, Halawy, Zahidi) can be stored for six months at 4 °C, for one year at 0 °C and more than a year at -18°C. Soft dates (Medjool, Barhee, Maktoom, Sayer and Dayri) can be stored for six months at 0°C and for more than six months at -18 °C.

#### 2.5.3.5 Ripening

As discussed above, dates are harvested when the moisture and sugar content are suitable for the specific market. This is featured in the texture and colour of the fruit. Should circumstances call for dates to be harvested before they are fully ripe, they can be artificially ripened. This is done by exposing the dates to ethylene for seven to 10 days and the ripening process is better at temperatures of 30 - 35 °C. Dates that are ripened in this way are of poor quality and the colour might be defective. *Rutab* and *tamar* dates varieties cannot be ripened artificially (Lobo et al., 2013).

### 2.5.4 Date marketing

Processed dates are globally marketed both for direct consumption or further processing into high-value confectionery products. Although there is growth in the demand, much more should be done in developing the potential utilisation of dates and date products in confectioneries, the pharmaceutical industry, handcrafts and furniture. Date palm leaf weaving is a craft known as *Safafa*. Developing new value-added date products will expand the date market. Products such as date sugar and date palm by-products such as sweet sap, date palm leaves and wood for furniture should be produced in modern, effective processing facilities equipped for creating different industrial grade and retail products. This can be sold on local and international markets (Dhehibi et al., 2018). The Kingdom of Saudi Arabia (KSA) and the United Arab Emirates (UAE) are the main exporters of dates and they produce the highest quality export date products (Dhehibi et al., 2018).

Date marketing channels differ from country to country, but Tunisia has the most professional and organised system which involves agents having different functions. The dates are initially harvested and processed at farm level. The fruit is then either transported to the closest market, especially the perishable *khalaal*, or to a packing plant.

It is imperative to choose the most economic but also effective transport for bringing the product to the client without losing quality (Mbaga, 2012*b*). Appropriate packaging can enhance date products and increase date sales in the global market. Besides packaging materials having to meet international standards (Zaid, 2002), the fruit must be properly protected in order to preserve its condition. It is either done for local long-distance and export transportation or packed for the final consumers at the supermarkets. According to Kader and Hussein (2009) the latest trends in packages are to use resealable bags or clamshell containers and to make use of recyclable materials and modified atmosphere packaging.

### 2.5.5 Farming practices and obstacles to marketing

The feasibility and profitability of date palm production depend on effective pollination (automated pollination mechanisms), soil management, successful harvesting (industrial maturation equipment and safe mechanical harvesting techniques), proper produce handling and sorting techniques. Production can also be improved by cultivating the new higher yielding cultivars.

In the date palm value chain, the weakest node is post-harvest handling, according to Dhehibi et al. (2018), as great losses occur after harvesting. Ineffective harvesting, sorting and grading are usually a result of poor farm management. According to statistics collected by the Food and Agriculture Organisation (FAO) almost a third of the date production in the UAE was used as animal feed during 2013 (Dhehibi et al., 2018). Management of preventive and acute treatment programmes for pests and diseases are not implemented properly and yield losses occur as a result. Yield losses can be mitigated with the management of pests and disease control, *i.e.*, by using red palm weevil detection devices (Al-Yahyai, 2018).

Date production is a growing industry but its contribution to the global economy can be much larger still. The final market price and total value of date production are determined by the way in which the dates are produced, harvested, sorted, graded and processed. Packaging and transport also play a role. The efficiency, safety and quality management of date marketing channels are ultimately the main determining factors to the final date market value (Dhehibi et al., 2018).

## 2.6 Synthesis: Literature study

The literature study briefly discussed the global date context, relating it to the South African market, and investigated yield prediction and estimation. The following chapter focuses on crop models, specifically yield simulation or prediction models, and the literature on date palm modelling.

# Chapter 3

# Literature review of existing yield models

This chapter takes a closer look at the literature relevant to the objective of this research, first by investigating different generic crop models and techniques existing in literature and used in practice, and secondly by forming an overview of research done on date palm yield prediction. The review presents an introduction to yield estimation models in general, used on various crop types, and different factors considered and investigated in the literature. The two main types of crop models are distinguished; namely process and statistical models, after which the focus is shifted to date palm models and research.

Date palm research and specifically yield prediction on this perennial were investigated. In order to explore the state of knowledge, an overview of relevant articles was obtained by following a semi-systematic review approach described by Snyder (2019). This was done on Scopus, the major synopsis and reference database of peer-reviewed literature. In order to gather the available literature, the search string shown in Table 3.1 was used.

Table 3.1: Scopus search string

| | | |
|---|---|---|
| **date** | **yield** | **predict*** |
| AND | OR | OR |
| **palm** | **harvest** | **estimat*** |
| | | OR |
| | | **forecast*** |

The search was refined by

- allowing only articles, books, and conference papers in the English language;
- including only documents published between 2010 and 2021; and
- excluding the following subject areas: Energy, Chemical Engineering, Chemistry, Medicine, Pharmacology and Toxicology, Material Science, Arts and Humanities, Business, Management and Accounting, Neuroscience, Health Professions, Psychology, Dentistry, Nursing, Immunology, Pharmacology, Veterinary, Social Sciences, Physics and Astronomy;

which produced 1 931 document results.

To get an overview of the literature on crop yield prediction (of all crops), the search string is adapted to 'yield AND (forecast* OR predict* OR estimat*)' with the same criteria, and 227 089 results are produced. It is clear that the field of crop yield prediction is expansive, while the date palm yield literature is not plentiful.

The documents from the date palm search, described in Table 3.1, were sorted and the abstracts were read to include only relevant articles. This implies including documents on date fruit yield (many related to oil palm instead of date palm) and fruit in general. Less than 60 documents were identified as particularly relevant to this study. Categorising according to country where research is done, the greatest majority of research is from Egypt, Saudi Arabia, Iran and the UAE. The impact of experiments, such as with irrigation or fertilisation, has been studied to a great extent. 46 documents relate management practices such as thinning, or treatment such as special fertilisation, to yield and some develop prediction models, while only two documents (Djerriri et al. (2018) and Al-Ruzouq et al. (2018)) from this search focus exclusively on yield estimation or prediction for the date palm, and research and analysis have also been done on the growth curve (Al-Khayri (2012)). In the literature on date palm yield estimation *per se*, a gap clearly exists.

Performing a literature overview provided insights into the field of crop yield estimation. It is evident that there is a lack of well-developed date palm yield prediction models in the literature. The literature is further investigated to study methods used to estimate crop yield in general.

Agriculture is a giant global industry providing food for an increasing world population, in the midst of climate change and with critical natural resources such as water, soil and atmosphere being under stress. Both short-term and long-term management procedures to sustain essential crops require information about crop-water and crop-nutrient productivity, the relative abundance of the specific crop under existing and possible future circumstances that could affect it, as well as the likelihood of geographical spreading. Knowledge of the impact of soil erosion, air pollution, water salinity, fertigation and agricultural management is key to ensuring the highest yield of best quality (Gornall et al., 2010).

Methods to estimate the yield of various crops differ from simple and straightforward to complicated and elaborate. A simple method is manual counting – measuring a representative area, counting the pods (in the case of beans) or fruit of the area and calculating an overall average, taking into account the mass to obtain a yield assessment (*Estimating Crop Yields: A Brief Guide*, 2013). This can be converted to a unit such as tons per hectare.

In the quest for better sustainable crop management and more economic cultivation of crops, decision-makers, together with scientists and researchers, have made numerous attempts to accurately calculate crop yield outcomes under different scenarios. Various methods of calculating relationships between initial inputs from different sources, the effect of significant variables and the complexity of these factors, necessitate the assistance of fast, accurate computations.

## 3.1 Basic outline of computerised yield models

In recent years, the computing power of machines has advanced tremendously, launching a predictive modelling revolution. A computer model, as defined by Kowalski (2015), is a program, run on a computer, that develops a model or simulation of a real-world feature, event or phenomenon. Agricultural models were developed to address numerous significant elements such as water or nutrient allocation, growing conditions, comparing potential and actual yield and much more.

According to Cai et al. (2017) yield models can be divided into two categories, namely process models and statistical models.

- Process yield models are based on mechanistic model principles while they are designed to be causal. Biophysical detail is employed, and assumptions are made on prominent underlying mechanisms by which real-life events are mimicked. Process models apply specific biological features of plants such as information on rate of photosynthesis and various ecological factors such as soil characteristics. These models are based on known factors that determine crop production and can define the influence of these variables on crop yield. Process models are dominant in the yield prediction field and include most of the well-known yield models.
- Statistical models are based on the characteristics and correlations of historical data, producing trends and patterns, and applying these to the future. These models simulate future crop yield although they are not structured on the biophysical processes that determine vegetative growth and do not rely on specific parameter values of individual crops.

In this research, statistical models are much more relevant as they focus on the data instead of relying on the agricultural and hortological principles. Both types of yield models require extensive parameter sets and input variables that are often difficult to obtain. Limitations for the application of these models are largely based on the accessibility of the input data, not necessarily on the model construction (Huth et al., 2014). Various computerised yield models have been developed for predicting

outcomes of different agricultural requirements on annual crops. These programs are also to a lesser extent applied on perennial vegetation such as shrubs and trees, requiring more complex modelling.

Although both annual and perennial crops produce measurable yield, the difference in the period of phenological development is significant. Annual crops are plants used in agriculture with a life cycle of only one year, producing seeds to continue the species as a new generation.

Perennial plants on the other hand have a life cycle of more than two years; they grow for many years until maturity, called gestation, and long-lived perennials such as shrubs and trees become woody with solid root systems. Climate and water supply have a substantial influence on the life cycle of perennials since they may grow continuously in a favourable climate while in a seasonal climate their growth is limited to the growing season. Perennial crops are harvested many times before replanting but production is complicated by various factors in the long life cycle, resulting in a changing productivity pattern over the crop's lifespan (Tregeagle, 2017). Productivity of trees varies to a great extent from the annual crop yield because a tree matures from a small plant to production and yield over an extended period, typically a few years. The onset of fruit production is followed by a long period of continuous annual production which declines gradually after many years and when trees reach a yield of non-profit quantity they are removed (Devadoss and Luckstead, 2010).

Most computerised yield models are mechanistic models (process models) based on mathematical expressions describing the physical and biological processes in the plants as well as the relationship between genetic, environmental and management factors, simulating predictions as output. The process of the model can usually be segmented into three parts: the input data, model construction and model output.

Data used as *inputs* for yield models consists of measurable parameters in the form of comprehensive datasets, which can be collected manually, nowadays mostly done by scientific instrumentation or remote sensing resources such as weather stations, measurement probes and satellite geographic cameras. This measured data includes environmental information such as climate factors (temperature, humidity, rain, wind), light intensity or radiance and soil data (soil type, soil water, nitrogen, phosphorus and potassium content, drainage, and erosion). According to Justice et al. (1998) various remote sensing data measurements may enhance the predicted yield. Measurements of the Moderate Resolution Imaging Spectroradiometer (MODIS) are particularly trustworthy with superior radiometric and spatial precision and the normalised Difference Water Index, a soil water index indicating plant status and production. Project management programs or other data capturing administrative tools are employed for capturing management information required by most data yield programs. This data includes irrigation and fertilisation details, root and other manual growth measurements, dates of crop phenological stages and significant management decisions. Plant genetic information of the specific species cultivar is found on biology databases. Raw data has to be manipulated and aggregated to function as suitable input for the model. Data imputation can be performed on temporal or time-varying data to compensate for absent data stretches and interpolation and extrapolation can be applied to fill in missing data in time. This also prevents potential overfitting of the model.

For the model construction, data analysis is done by means of statistical or data mining techniques and relevant features are identified and enhanced. Some of the data is employed to construct and calibrate the model and a quantity of data is utilised to validate the results and refine and improve the model (Tiwari and Shukla, 2018). The core of the crop growth and yield model is the employment of a set of equations to predict biomass production from absorbed carbon dioxide, solar radiation, and water. Simplified mathematical formulations of prominent causal mechanisms such as soil variables (*e.g.* the cation exchange capacity, quantity of organic matter, soil evaporation, rooting condition index and other fertility estimates), plant transpiration indexes and reference evapotranspiration, are encoded. Applying meteorological data together with remote sensing-based vegetation indices is an approach resulting in improved crop yield prediction. Graphical indicators such as the normalised difference vegetation index and its transformations and aggregations together with other pixel-intelligent computations are applied to satellite pictures to determine leaf area index, or the Fraction of Absorbed Photosynthetically Active Radiation. These are indicators of plant productivity, photosynthetic rate

and crop-specific masks from which growth rate and biomass production are simulated.

The computer program generates an *output* determined by the data provided as input, and the demand specified as a result. Possible results may constitute, among many others, scenarios of optimum water requirements referring to water quality and quantity, predictions of climate change effects on yield quality and quantity or on future distribution patterns of the crop and determining ideal irrigation schedules for crops. These results may be displayed by means of graphs, or data summarised in tables and digital images supported by text data.

More complex yield modelling systems are created by integrating different modules for specific outcomes, but some computer models simply serve as a framework to integrate significant yield programs.

Applying process yield models involves the scientific and precise collection of the required input data. Physical vegetative data and the monitoring of growth indicators are nowadays mainly done by applying pixel intelligence mapping from satellite images. Accurate plant physiology is derived from genotype information and environmental factors are measured with scientific instrumentation. These requirements are rarely met. The lack of specific information or even incomplete input data or input data of poor quality employed in yield forecasting models often lead to substandard outcomes. This especially proves to be the situation with data on date palm production.

## 3.2 Prominent process yield models

Numerous process, or mechanistic, models for yield prediction have been developed and successfully applied on annual crops, predicting yield or other outcomes and occasionally applied on perennials, of which the 12 models most prevalent in the investigated literature are discussed:

1. One such model which is technologically advanced by the Land and Water Division of the Food and Agriculture Organization (FAO), is the crop-water productivity model called *AquaCrop*. This crop growth model addresses food security by simulating the influence of water variation on herbaceous crop yield. The FAO published a paper (Steduto et al., 2012) outlining the science behind AquaCrop. The interaction between plant characteristics, climate factors and water is employed to separately calculate evaporation from the soil and transpiration by plants. AquaCrop distinguishes itself from other crop growth models in that it delivers substantial outcomes while only needing basic input data, achieving this by utilising a normalised water productivity to determine biomass.

   Many herbaceous crops including wheat, tomatoes and quinoa are discussed, followed by yield response to water of vines and fruit trees such as orange, peach, apple, almond and pear. The transpiration reductions of olive trees under deficit irrigation were also simulated with the AquaCrop model. However, dates are not among these fruit trees.

2. Another support tool to assist in decision-making, developed by the Land and Water Development Division is *CROPWAT*, which employs a modified Penman-Monteith method to determine the quantity of water required for irrigation of a crop field by taking crop, climate, and soil data into consideration. This program [1] includes features such as calculation of reference evapotranspiration, daily soil water balance output tables and adjustable irrigation schedules (Vozhehova et al., 2018).

3. The decision support system for agrotechnology transfer *(DSSAT)* has been used in the past three decades by researchers worldwide to aid agricultural stakeholders in decision making (Jones et al., 2003). The DSSAT cropping system model (CSM) is a complex tool which is modularly structured, and components are separated along scientific lines of discipline. The modules are Soil, Crop Template (simulating different crops by the definition of species input files), and

[1] The computing sense of "program" is spelt the same way in both American and British English while in British English "programme" is used for all other meanings of the word.

Weather. Another module is also available where competition for water and light among plants, soil characteristics and meteorological elements are taken into consideration. The primary scientific components on which this model is built are soil, crop, weather, and management (Jones et al., 2003).

4. Locally, *FruitLook*, a tool for planning, is a portal funded by the Western Cape Department of Agriculture, providing satellite images of farms, assisting in planning of water budgets, monitoring in terms of probe placement and allowing post-seasonal analyses of implemented changes with an aim at the future (Black, 2017). It has been particularly successful in increasing water use efficiency, the production per unit of water, production monitoring and crop water usage tracking. It has proved useful, as water scarcity is a major concern, specifically in the southern regions of South Africa.
5. The *Cropscan* tool has been employed to predict water use efficiency over wider areas with the interaction of satellite images and weather information with Fruitlook dataset information on crop growth and water use as well as leaf nitrogen content (Black, 2017).
6. *eLEAF*, based in the Netherlands, offers crop monitoring, irrigation planning, yield forecasting as well as water auditing services and subsequently assistance to farmers in water utilisation efficiency (Writers, 2014). For yield forecasting, it makes use of satellite-based crop production information based on pixel intelligence mapping and crop-specific yield models.
7. Another software modelling package for agricultural practitioners, *CLIMEX*, is applied to estimate the influence of climate change on species dissemination. Developed by Hearne Software, CLIMEX is an eco-physiological growth model, making use of simulation and modelling techniques. It mimics the biological mechanisms resulting in favourable and unfavourable growth seasons that determine the geographical spreading and relative abundance of species. Climate is the key factor determining the character of a growth season, while also taking into consideration parameters influencing a species. The existence of a species during unfavourable conditions depends on cold and heat (temperature index) and dry and wet (moisture index). Shabani et al. (2012) made use of CLIMEX, based on available distribution data, to develop a model of the potential distribution of the date palm under current and future climate scenarios. CLIMEX lacks biotic relationships and distribution in its modelling process, but software like the Geographic Index System may be integrated.
8. Environmental Policy Integrated Climate, *EPIC*, a forecasting tool developed to determine the influence of soil erosion on soil productivity, uses crop yield as indicator of the extent of the outcome rather than having yield prediction as a goal (Williams et al., 1989).
9. The agricultural production systems simulator, *APSIM*, utilises soil data such as water, nitrogen and phosphorus content and soil pH as well as erosion and management data to simulate biophysical processes. APSIM predicts crop production and soil-plant nitrogen interaction and is used in different climate change economic predictions and management practices. The APSIM framework has been used by Huth et al. (2014) in developing a production systems model for the oil palm, using data from different environments within Papua New Guinea.
10. Apart from remote sensing, genetic characteristics, climate, soil data and management information as input data, some forecasting programs such as *CERES*, also need root measurements compared to above-ground biomass growth. The way and tempo of root growth is an indication of the growth and yield of crops in water-deprived circumstances (Robertson et al., 1993).
11. The World Food Studies have developed a carbon driven tool, *WOFOST*, which uses information such as climate, solar radiation, soil and phenological data, above-ground biomass and photosynthetically active radiation (PAR) as input data and produces the constraints and stimuli to total biomass growth, the leaf area index and the final yield as the outputs (van Diepen et al., 1989).
12. The forecasting model *CropSyst* has been developed to simulate the canopy biomass increase related to the nitrogen content, soil water and climate. It estimates the transpiration coefficient

and the basal crop coefficient for perennial crops and has been successfully employed on the irrigation management of pears (Marsal and Stockle, 2011).

The abovementioned models are used for simulating crops, mostly annual crops. A discussion of other crop models specifically applied on perennial crops and date palms as well as models presented only in research will follow next.

## 3.3 Process yield models applied on perennials and date palm

This section discusses the application of process yield models, first on perennials and secondly on the date palm in particular. As mentioned, yield models on the date palm and perennials in general, are far less common.

### 3.3.1 Yield predictions on perennials

A relatively simple method to predict the yield of perennials is based on the Bavendorf Model. Three inputs are necessary: yield capacity, fruit set density and fruit mass at harvest (Köhne, 1985). Yield capacity per hectare ($c$) is the average tree canopy area in square metres ($m^2$) determined by tree size and number. For a tree

$$c = d \times h \tag{3.1}$$

where $d$ is the average canopy diameter and $h$ is the average canopy height. Yield capacity $c$ produces a sigmoid curve as the fruit-bearing canopy of the trees grows bigger. Fruit production is influenced by Fruit Set Density (FSD) which depends on flowering, pollination, fruit set and fruit drop. FSD is calculated by counting the fruit on representative trees. The FSD together with the total yield capacity determines the total yield. Fruit growth curves are employed to calculate the expected average size or mass of the fruit at harvesting.

Research on yield predictions of perennials has been done to some extent and the implementation of current well-known process yield models to obtain more accurate yield outcomes has been done with varying success. A few projects on perennials are mentioned below.

- The AquaCrop Yield model has been employed to investigate and determine the optimal fertigation of oranges under water stress (Qin et al., 2016). Soil water and nutrient dynamics have been processed by the three-dimensional FUSSIM program while the orange crop growth has been simulated by the AquaCrop model. The crop yield has been simulated as a function of the utilisation of water under predefined soil productive levels.
- In an attempt to optimise irrigation in a pear grove the CropSyst model was implemented by Marsal and Stockle (2011) under different irrigation conditions. The water-plant potential compared to transpiration (Ohm's Law) of the tree has been used while an agreement between predicted and confirmed tree transpiration, light interference and stem water potential was calculated. This project confirmed that CropSyst can provide valuable assistance for the management of stress irrigation circumstances of pear trees while predicted for periods of less than 40 days.
- The APSIM yield model framework has also been adapted into a semi-mechanistic computer model with the purpose of simulating development and yield of the oil palm, called OPSIM. This program then was adapted into another submodel called the APSIM-Oil Palm. This model takes into account oil palm physiology and physical processes as well as the causal relationship between the environment and the crop (Sung and Siang, 2018). The model successfully predicts for crop production in conditions where oil palm growth is limited by weather conditions and water.

- The performance of yield models is invariably verified when employed during research. One typical project was to benchmark the WOFOST model while reproducing the growth and yield of the Jujube under different deficient irrigation treatments (Bai et al., 2019). The objective was the implementation of more precision agricultural irrigation on Jujube crop and it was determined that 90% of full irrigation results in a balance between yield and water savings.

### 3.3.2 Process yield model application in date palm research

Currently known process yield models have rarely been employed in date palm research because of the lack of applicable, accurate or adequate input data. Date palm production is highly susceptible to climate factors from pollination to harvesting, but significantly and particularly so at certain phenological development stages with severe consequences. Farm management procedures are of the utmost importance. Diligence in executing effective pollination and performing precise scientific cultivation of the fruit, *e.g.* thinning, has a critical impact on the yield. The importance of these factors is often underestimated or absent in the algorithms of existing crop yield models. This leads to inaccuracy and inadequate calculated results, causing them to be of small merit.

Where known process yield models have been applied for research on the date palm, they were mainly employed to the extent of qualifying phenological or cultivation factors. Extensive research has been done on almost every possible phenological, environmental or managerial component of date palm cultivation in pursuit of testing hypotheses of experiments of which the following projects were applicable to this study.

- Research on the possible geographic spreading of the date palm under future climate change was done by implementing the CLIMEX model using the A2 Special Report on Emissions Scenarios (SRES), postulating rapid population and low economic growth under extensive technological development (Shabani et al., 2012). This was employed in association with two global climate system models CSIRO-Mk3.0 (CS) and MIROC-H (MR). The results disclosed that future spreading of the date palm tree will primarily be determined by undesirable cold and dry conditions. This outcome assists in strategic planning by indicating potential global settings in which to cultivate dates. The research revealed that the future climate conditions of areas in North and South America, including North Venezuela, will become more acceptable for the date palm tree. It is however predicted that Saudi Arabia, western Iran and Iraq will be less fitting for date palm cultivation by 2070.
- The CROPWAT yield model was implemented to establish irrigation requirements during the productive period of the Deglet Nour date palm (Mihoub et al., 2015). The study put emphasis on water quality, specifically salinity, since this is a primary factor in water management. To determine irrigation requirement CROPWAT was employed to determine reference evapotranspiration, crop coefficient, water holding capacity, leaching conditions and daily water requirements of the date palm. The study revealed that using localised irrigation methods on the date palm can reduce water usage by half, compared to methods such as border irrigation.

## 3.4 Statistical crop yield models

Statistical models have been extensively applied to determine crop yield by utilising historical data on climate and crop yields to calibrate uncomplicated regression equations. These endeavours had encouraging outcomes and proved to be an applicable alternative to process-based models under certain conditions (Shi et al., 2013).

According to Lobell and Burke (2010) the three major types of statistical models are:

1. time-series models utilising time series data from a single point or region;

2. cross-section models modelled on spatial variations data; and
3. panel models employing spatiotemporal variation data.

Research and technical development on perennials, in particular fruit trees, tend to be more feasible while depleting historical data. Some examples follow of statistical crop yield models implemented on perennials.

- Statistical models of crop growth and yield using regression models allow for a more elementary alternative for spectroradiometric studies. These models may include the effect of weather-generated pathogens, air pollution and various factors that are normally ignored by process-based models.
- Yield prediction of oranges was done in the southern United States and prediction of wine grape yield and quality in prominent vine regions (Lobell et al., 2006). Outputs from multiple climate models were employed to quantify the sensitivity of these crops to temperature and precipitation changes and to evaluate the relative contribution of climate and crop uncertainties to total uncertainties. Historical data on the influence of temperature and rain on yield was processed by statistical modelling. Yield functions used for crop and adjusted $R^2$ by linear regression was agreed upon based on historic data. An imperfect relationship between yield and monthly climate was deducted.
- Date palm crop yield forecasting has been done or suggested by implementing image processing and machine learning techniques (such as self-organising maps (SOMs), decision trees, and artificial neural networks (ANNs)) (Husain and Khan, 2020). Image classification was then used to perform the forecasting of the yield. Acquired images are stored in a rectangular grid; the colour or intensity at each point is converted into a numeric value. Data was reduced to a lower dimensional space using the ANN network and the weight matrix of each image stored. Labelled vectors are input to the ANN one at a time and used to train the model. Features such as certain fruit texture, variations, angles or measures are identified. Image classification is employed to label images into a predefined category. The system can then report and identify from a database. The total crop yield was estimated and forecasted by calculating identified fruit.
- In India the yield of Gala Red Lum Apples has been done by developing different models followed by an elimination process (Mushtaq et al., 2018). In these models the dependent variable or target to be predicted, is the yield. The independent variables are
  - crop density,
  - flower density per trunk cross-sectional area,
  - flower density per land area,
  - fruit clusters per trunk cross-sectional area,
  - fruit clusters per land area,
  - average fruit number per cluster and
  - average yield per fruit cluster.

  The relationships between the variables were calculated with linear, log-linear and polynomial regression models implementing a linear regression and a backward stepwise elimination option. The models were ranked on the principle of the coefficient of determination ($R^2$), adjusted $R^2$ and Akaike information criterion (AIC). The AIC, formally defined in Subsection 6.7.1, determines the out-of-sample prediction error and therefore relative quality of statistical models for a given set of data. A discussion on $R^2$ will be presented in Section 6.3.
- In Hungary statistical modelling was employed to estimate crop yield by utilising weather information and remote sensing vegetation indices (Kern et al., 2018). The intention of this study was to construct an applicable crop yield model by using multiple linear regression applying a few predictors. Another outcome was to calculate the effect of climate and environment features

on crop yield at different scales. This method also improved the statistical modelling with remote sensing data by taking vegetation pointers into consideration. Meteorological data and soil water content from meteorological reanalysis in monthly resolution were applied as forecasters of the models. In order to determine the importance of similar data on the predictive power of statistical models, vegetation index was included. Using stepwise regression, the most suitable models were identified with statistical evaluation. This delivered basic computations and equations with applicable coefficients, estimating crop yield of winter wheat, sunflower and maize effectively and with high accuracy.

- A model to determine date palm water requirements was developed by Sperling et al. (2014). A palm tree requires more than 2 000 mm water per year and water quality is a common concern. During these findings, a model was established to calculate palm tree evapotranspiration by determining the influence of environmental elements on canopy resistance and water loss and integrating water salinity into the model, while considering the quality of the irrigation water as an additional factor.

  These results were utilised by an adjusted 'Jarvis–PM' canopy conductance model employing meteorological and water quality data alone. The adapted procedure generated weekly irrigation proposals based on field water salinity (2.8 $dS.m^{-1}$) and climatic forecasts that resulted in a 20% decrease in irrigation water use compared to regular irrigation arrangements.

Many statistical methods were encountered in this literature study of yield models, of which regression is most common for application on small or simple datasets.

## 3.5 Synthesis on yield and dates in literature

In this research, the focus is yield and its most prominent influential factors. Referring to many articles written on date palm experiments, very few relate to yield predictions based on input data similar to the available data relevant in this project. Of the articles reviewed on date palm yield, all of the models employ data that is (at least partially) unavailable for this study, for instance satellite imagery, soil profile data, soil temperature and moisture content, irrigation and fertigation. Conversations with SMEs also led to the conclusion that it is a small field of research and practice in South Africa and therefor knowledge on the cultivation of dates is not abundant in the country. Therefore, two major needs are identified. First, there is a need for the determination of relevant factors. Secondly, there is a demand for the development of a date yield prediction model requiring less, more readily available, input factors. From the literature it follows that there is a shortage in the application of estimation techniques in date fruit yield. This study aims to address this need.

# Chapter 4

# Theory of learning from data

This chapter handles data discovery and exploration in general. It also presents a description of the necessary data for developing a date yield model by integrating the date palm literature, yield model literature and predictive modelling theory.

Data, as defined in the Cambridge Dictionary (*Data meaning in the Cambridge English Dictionary*, 2020), is "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer". According to the Oxford Dictionary (*Data meaning in the Oxford English Dictionary*, 2004), data can be defined as the "quantities, characters or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical or mechanical recording media". Raw (unprocessed) data is a collection of numbers and characters to be cleaned and prepared in order to be provided as input to analysis tools.

## 4.1 Knowledge discovery from data and exploratory data analysis

As described by Patel (2018), knowledge discovery from data and exploratory data analysis are done before the conceptualisation of a model. The general steps of the exploratory data analysis process are:

1. Variable identification: Define the type of each variable (numerical or categorical) and its function in the dataset.
2. Univariate analysis: Histograms and boxplots for every particular continuous variable.
3. Bivariate or multivariate analysis: Understand the relationships among different attributes in the data, typically with tools like heatmaps.
4. Detection and treatment of missing or erroneous values. Erroneous or aberrant values occur as a result of faulty inputs or calculation errors. Missing values occur during data collection or extraction.
5. Detection and treatment of outliers: Detect outliers, observations that deviate further away from other observations in the dataset. Visualisation tools such as boxplots (univariate) and Scatterplots (bivariate) are often employed.
6. Feature engineering: Create features and perform feature transformations. This is commonly done on date variables where they are transformed to a particular type, and to years and months of the year.

In order to perform knowledge discovery from data, there must be an understanding of the data in context. This is therefore the first step in analysing the data. Data understanding will be discussed next, as well as mentions of types of data and data handling methods.

### 4.1.1 Data understanding

The data on date production acquired from the research partner is of a semi-structured (CSV) and stationary type as the analysis is done on pre-existing historical data consisting of numbers or strings. Data is one of two types, either categorical or numerical. These can be subdivided further. Categorical data can be either nominal or ordinal, and numerical data either interval, absolute or ratio. Nominal data is discrete-valued, with no order relation. Ordinal values have an order relation *e.g.* small, medium, or large. For numerical data, interval data is measured on a scale with units of equal size,

such as temperature, year, or time. Ratio data is interval data with a natural zero point *e.g.* time. Absolute values are simply numerical values.

An Analytics Base Table is usually constructed for a dataset, with input features and a target feature, one target per entry. However, in some cases the raw datasets do not permit that because of a more complex structure than simply input-target entries.

To handle outliers in data, the following steps must be executed to firstly identify the outliers:

1. Examine minimum and maximum values for continuous features and make use of domain knowledge to determine if these are plausible values.
2. Examine cardinality of categorical features.
3. Compare disparity between the minimum, median, maximum, first quartile and third quartile values.
4. The maximum value is unexpected in case of the disparity between the third quartile and the maximum value being much larger than the disparity between the median and third quartile.
5. The minimum value is unexpected when the disparity between the first quartile and the minimum value is much larger than the disparity between the median and first quartile.

After the identification of outliers, the following must be considered:

- If the predictive model is robust to outliers, outliers remain in the data.
- Robust estimators are implemented if outliers need to be kept.
- Outliers can be removed.

Categorical data can be visualised with bar plots and continuous data with histograms. For each value, the central tendency and variation should be examined to understand the types of values a feature can take. For each continuous feature, the following should be examined:

- The mean and standard deviation of each feature, to get a notion of the main tendency and variation of the values within the dataset.
- Examine the minimum and maximum values to understand the possible range.

The most common issues related to data quality include missing values, irregular cardinality (a data quality issue that arises when the number of elements for a feature does not match the expectation), outliers (invalid values, included in a sample through error, and valid outliers, correct values that differ from the other values for a feature), noise, as well as skew or imbalanced data.

Missing values, where a value for a feature is not entered, may be a result of error in the data integration process or an artefact of the data collection process, for instance failure of a sensor or a typist error. It can also be meaningful, as in the case of a categorical feature not applicable to a specific entry. Approaches to take when handling these missing values, include dropping features with missing values, applying case analysis and removing instances with one or more missing values, or deriving a missing indicator feature from features with the missing value. Imputation is used to replace the missing values with a probable calculated value based on other values for that feature. The standard approach to imputation is replacing the missing values for a feature with a portion of the central tendency of the feature. Computation should generally not be applied on features missing more than 30% of the values.

### 4.1.2 Data exploration

There are two main goals when exploring the data: to fully understand the characteristics of the data, and to determine whether the data suffers from any data quality issues. Characteristics of the data are:

- the types of values a feature can take
- the ranges into which the values of a feature fall
- how the values for a feature are distributed across the range in which they fall

Examples of data quality issues are instances with missing values, outliers, noise, or different units of measure.

Outlying points can greatly affect the fit of a model, such as a linear regression model, because of the quadratic weight of residuals. The Z-score is the number of standard deviations a given data point lies from the mean (Brase and Brase, 2019). Generally, a threshold of 3 (or -3) is used to identify outliers if the data can be assumed to be normally distributed.

When reporting on the quality of the data, the descriptions of characteristics utilise the standard statistical measures of central tendency (mean, mode and median) as well as the standard measures of variation (the standard deviation and percentiles).

For a first step in understanding the relationships in the data, the correlation can be considered. A statistical scope of the strength of the relationship between the relative movements of two variables, is the correlation coefficient. It is also described as the degree to which one variable moves in relation to the other. The value of this coefficient can range from -1 to 1. A perfect negative correlation, a relation in which one variable linearly increases as the other decreases and *vice versa*, is indicated by a coefficient of -1. A perfect positive correlation, where the increase in one variable leads to a linear increase in the other, is represented by +1. A high positive coefficient indicates a definite positive relationship and a high negative coefficient close to -1 points to a strong inverse relationship. The Pearson correlation coefficient, also known as the Pearson product-moment correlation coefficient, can be used to find this relationship. Finding a correlation between two variables can be illustrated by placing the variables on a scatter plot. For the coefficient to be calculated, there must be some linear relationship. A scatter plot not portraying any resemblance to a linear relationship can be disregarded. A similarity of the scatter plot to a straight line, indicates a higher strength of association. The sample Pearson correlation coefficient of two variables $x$ and $y$ is calculated by

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{4.1}$$

where
$x_i$ is the $i^{th}$ observation point of variable $x$,
$y_i$ is the $i^{th}$ observation point of variable $y$ and
$n$ is the number of observations.

The use of the Pearson correlation coefficient in this study is an initial exploration of a dataset to discover relationships and determine the usefulness of further use of the data obtained from the research partner.

## 4.2 Measurable factors affecting yield of the date palm

Determining all the possible factors playing a role in the growth and production of the date palm leads to an indefinite study. Consequently, only major components can be taken into consideration. Accurate and consequent measuring of these factors for an applicable period as well as collecting and recording of the data establish a basis for employing the data.

This section discusses the constant parameters and the variable factors to consider in a date yield model, with reference to the literature study in Chapter 2 as well as the knowledge of the present chapter on data.

Factors to be included when predicting the yield consist of a combination of constant parameters together with variable elements.

### 4.2.1 Constant parameters

Constant components or parameters are factors that are fixed or controlled but have an influence on yield. This data is recorded once and included according to its applicability. Parameters include the following:

- The propagation type of the plant and date of planting both have an influence on the vegetative and productive state of the plant and are major factors in determining the size of yield. This information must be recorded manually during cultivation and was taken into consideration during this study.
- Spacing of the trees and the layout of the orchard definitely play a role in production and can be abstracted from the farm orchard layout.
- The topography of the area has an influence on water drainage and photosynthesis, playing a dominant role in production and yield. The effect of this aspect can rarely be determined and will be regarded as a constant. The orchard layout and geographical orientation of specific orchards can be obtained from a satellite photograph. Spacing and location of trees in a specific orchard as well as tentative topography could be derived from this.
- Pollination is a vital development stage and elements such as the quality of the male pollen, receptiveness of the female spathe and the meteorological circumstances determine the percentage fruit set. Although pollination is done manually, these mentioned elements cannot be manipulated or measured. The date of pollination must be recorded manually because this fact combined with various other factors play a distinct role in the quality and quantity of the yield. The outcome may be deducted by determining the fruit set percentage.

### 4.2.2 Variables

The following variables (factors that can change) have a definite influence on yield:

- The soil profile is of vital importance evaluating the water and nutrient content and condition. Soil condition should be measured on a regular basis by means of probes or soil samples. Data is applied to assist in determining the irrigation and fertilisation measures, strategies and programs.
  Irrigation: Sufficient water supply from rainfall combined with effectively applied irrigation plays a vital role in optimum date fruit yield.
  Fertilisation: Plant roots take up essential nutrients from the soil in order to perform normal life processes such as photosynthesis and other metabolic processes for vegetative growth. During cultivation, the condition of the soil can be enhanced by the accurate application of water and nutrients in the form of fertilisers. Required in differing quantities, nutrient elements vital for plant growth and production include nitrogen, iron, magnesium, potassium, zinc, boron, calcium, sodium, chlorine, cobalt, copper, sulphur, manganese, and phosphorus. Various tests have been done on several cultivars to determine the effect of different nutrient applications, mainly of nitrogen (N), phosphorous (P), potassium (K) and sulphur (S), on the yield.
- An essential measurable variable in producing a yield of high quality and quantity is the number and mass of the fruit bunches. In cultivation the number of bunches as well as the number of fruit per bunch are manually controlled.

  As discussed in Chapter 2, thinning greatly influences the yield. In fact, of the management practices implemented, thinning has the largest effect. The process is as scientific as its execution and this is affected by the training and skill of the labourer, or thinner. Moreover, there are

also factors like the state of mind of the thinner to consider. Date palms are not easy to treat manually because of their height and thorny leaves.

- Meteorological data is the most important information in predicting the yield and this is normally easy to measure and collect with a local weather station on the farm. The alternative is to obtain this data from an online weather source. As discussed in Chapter 2, weather elements such as heat, humidity, rainfall, wind and radiation play an enormous role during different vegetative stages of the date palm. Temperature, in particular minimum heat units required before pollination and fruit set and necessary for normal production proves to be the most influential factor in determining the yield. The effect of heatwaves during ripening also tends to have a high impact on the yield. Wind and high humidity may be detrimental during pollination and precipitation or high humidity causes fungal damage to ripe fruit. Radiation has an influence on photosynthesis, growth and production.

## 4.3 Summary of learning from data

This chapter presented data exploration methods. Finally, specific data useful for date palm yield prediction is discussed by considering the factors influencing yield as learnt in the literature study. Chapter 5 will present the real-world data made available for this research and begin exploring its adequacy.

# Chapter 5

# Exploring the real-world datasets used in the project

The previous chapter discussed the term "data" and presented ways of exploring acquired datasets. This chapter presents the real-world data available for exploration obtained from the research partner. The farm under study is located next to a major river in an arid region of South Africa, with the majority of orchards planted on the riverbank. The data is visualised, and some correlations calculated to discover useful relationships, or the lack thereof.

To discover if the available data will be suitable and sufficient for predictive modelling, the raw data needs to be transformed and investigated. The type, shape and size of the data will also determine the type of model used, should the data prove useful.

Meteorological data from the research partner was obtained for the years 2014 to 2020 and harvest data from 2010 to 2020. Additional acquired data to compensate for the deficient meteorological data, years 2008 to 2013 and 2020 for which harvest data is available, was obtained from an online source, the NASA Langley Research Center Prediction of Worldwide Energy Resources (POWER) Project, to complement the original data. The reconciliation of the second set of weather data from POWER was done by adapting maximum values, minimum values and mean values in accordance with the data obtained from the research partner for a period of five years.

The data from the investigated South African date farm and the POWER site include the datasets presented in Table 5.1, all converted to comma-separated values (CSV) format.

Table 5.1: Datasets available for the study

| **Name** | **Description** | **Timeframe** |
|---|---|---|
| Orchard description | Description of orchards incl. the year in which trees were planted, type of irrigation, area of the orchard (ha) | 1975 – 2017 |
| Fruit growth | Weekly growth measurements of 15 orchards | 2013 – 2019 |
| Harvests | Daily harvests off 61 orchards, complete from 2010 for 33 orchards | 2010 – 2020 |
| Growth stage monitoring | Dates of pollination, thinning of orchards | 2015 – 2019 |
| Bunches | Number of bunches on each tree and estimated average mass of bunches of all orchards | 2010 – 2020 |
| Farm weather | Hourly measurements for five years (temperature, humidity, precipitation, wind) | 2014 – 2019 |
| POWER weather | Daily measurements (temperature, humidity, precipitation, wind) | 2008 – 2013 |

All the obtained datasets, displayed in Table 5.1, are briefly described in this section, followed by a deeper exploration of the relationships in them.

## 5.1 Orchard description

The orchard description data contains relevant information about the characteristics of the orchards. Applicable attributes are age, type of irrigation, size of orchard in hectares, and type of propagation.

Propagation was discussed in Chapter 2 under Section 2.2, together with other literature on date cultivation. The orchard descriptions contain numerical entries in the age and size columns and categorical nominal entries in the irrigation type feature.

A layout of the orchards is displayed in Figure A.1 in Appendix A.

## 5.2 Fruit growth measurements

As the research partner focuses on exporting fruit of particular dimension, mostly of a large size and mass, care is taken to ensure fruits of adequate size. Selected fruits are individually measured to ensure the desired size and mass are recorded and achieved. The annual process starts just after fruit set [1], and measuring occurs from the end of October until early January. A single fruit on a bunch at the top, near the middle and at the bottom on a marked tree in 15 of the 61 orchards is measured weekly. Length and diameter measurements are taken of these three fruits and the weighted average mass is calculated. The assumption is made that the three fruits on this tree are representative of the orchard in which it grows. Where a measured fruit falls from the tree during the growth period, a different but similar fruit is measured instead. This sometimes leads to large dips in the growth curve. These and other inaccuracies are corrected during data processing.

From seasons 2013 to 2018, the measurements on all orchards were taken for 12 weeks or less, where the weeks are numbered 1 to 12. In the 2019 season, nine weekly measurements were taken for each orchard. Note that orchards are referenced by numbers, which are not necessarily chronological.

The development and growth of the fruit are measured at the end of the *Hababauk* phase until harvesting. These length and diameter measurements are used to calculate an average mass of a fruit. The data for the 2012 season had incomplete entries and was removed. The data for Orchards 42 and 90 was also removed as each only had entries for one season. After the removal of duplicates found in the data, the dataset of 1 124 observations, is summarised in Table 5.2, where the diameter and length are in mm and the mass is in g.

Table 5.2: Short statistical description of growth measurements dataset

| | Season | Week | Average Diameter (mm) | Average Length (mm) | Processed average mass (g) |
|---|---|---|---|---|---|
| mean | 2015.92 | 39.29 | 27.21 | 40.52 | 12.51 |
| std | 1.95 | 17.90 | 6.55 | 14.49 | 7.99 |
| min | 2013 | 1.00 | 7.50 | 9.00 | 0.20 |
| median | 2016 | 47.00 | 28.88 | 43.00 | 12.40 |
| max | 2019 | 52.00 | 41.50 | 68.75 | 36.30 |

Figure 5.1 displays the weekly mass measurements for four arbitrarily chosen orchards; namely 9, 10, 46, and 70, over the years 2013 to 2019. The weeks are numbered instead of being displayed by date to associate annual data according to week numbers of the year.

The stacked histogram in Figure 5.2 displays the distribution of the mass of the fruit measured in week 12 for seasons 2013 to 2019. The average mass in this week ranges from 14.5 to 29.5 g and it is clear that some years (2013 and 2018) produce bigger fruit while 2017 has smaller fruit.

Week 12 is the last week in which measurements were taken for all seasons from 2013 to 2019. Figure 5.2 shows the trend of smaller fruit in 2017 clearly with the outliers around 15 g.The variation found in fruit mass among the various years led to the assumption that increases or decreases in fruit mass could give an indication of expected date yield.

[1] The term *fruit set*, discussed in Chapter 2, refers to the fruit after pollination.

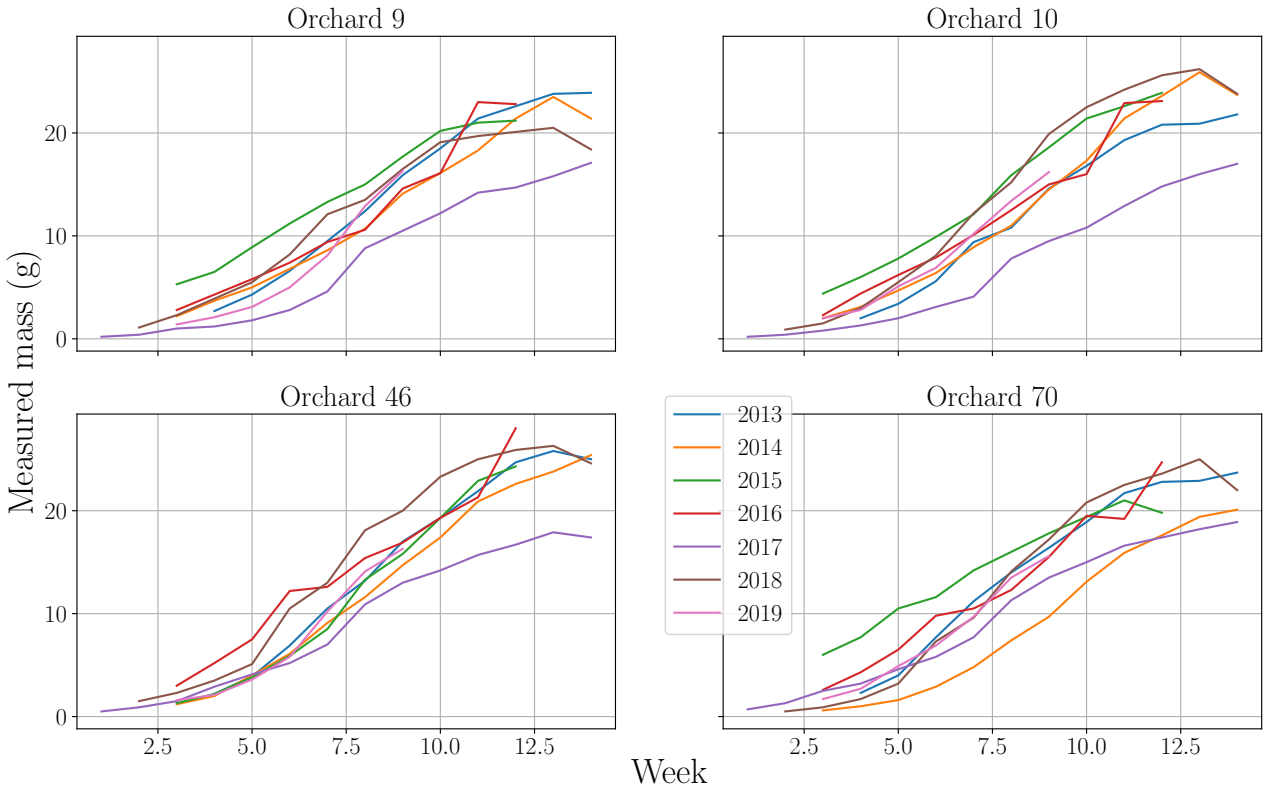



Figure 5.1: Weekly measurements for four orchards

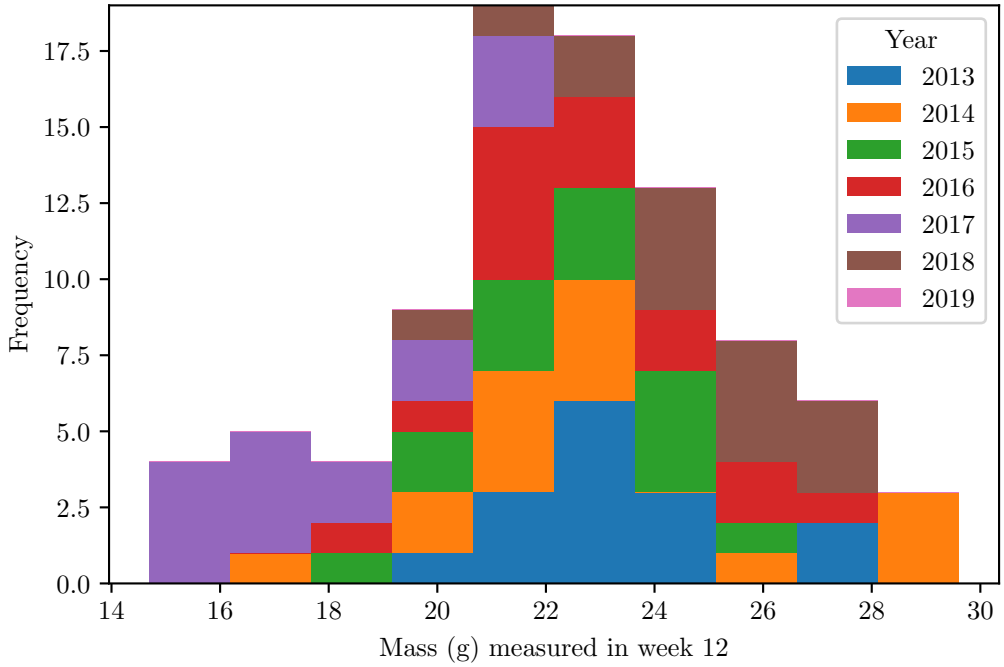



Figure 5.2: Distribution of fruit mass measured in week 12 of years 2013 - 2019

## 5.3 Growth stage monitoring

For each orchard, from 2015 to 2019, the following values were recorded: orchard number, cultivar, flowering date, pollination date, date of first thinning and the date of beginning of harvest. Most useful is the pollination date of each orchard, which could be used to analyse information on the

meteorological circumstances of the pollination. Although the thinning date is recorded, the extent or severity of the thinning is omitted.

## 5.4 Bunch data

For each orchard over the years 2010 to 2019, the number of bunches and bunch mass are recorded from sample trees. After thinning, usually in December, the bunches on each tree in each orchard are counted. These numbers are summed and divided by the number of trees in the orchard, to obtain a mean number of bunches per tree for the particular orchard in that season. A mean bunch mass representing the kilograms per bunch of the orchard is also found by weighing a few bunches on one tree and finding the mean. On the farm, the current practice is a pragmatic method using the product of the number of bunches and estimated bunch mass to make predictions of the expected yield. The bunch dataset is a visual representation of each orchard displaying the number of bunches on each tree for all the years. These "maps" also indicate the orientation of the orchards with respect to natural landmarks such as the river.

## 5.5 Harvest data

The harvest dataset contains 26 024 entries of all harvests, the mass of the fruit production harvested from full-bearing date palms and made on every harvest date. Most dataset entries specify the harvest date, mass and the orchard from which it was harvested, although some entries contain only the harvest date and the harvest mass. Harvesting starts in early to middle February and continues until April or sometimes early in May. Fruit is usually harvested when ripe, although from 2017 harvests included unripe fruit which was artificially ripened without reducing moisture content.

Table 5.3: Yearly date harvest mass

| Year | Mass (kg) |
|---|---|
| 2010 | 1 010 588 |
| 2011 | 1 098 823 |
| 2012 | 802 648 |
| 2013 | 870 598 |
| 2014 | 791 985 |
| 2015 | 943 248 |
| 2016 | 898 332 |
| 2017 | 1 049 651 |
| 2018 | 993 454 |
| 2019 | 862 256 |
| 2020 | 1 267 714 |

Table 5.3 displays the seasonal harvests. This study aimed at finding the factors influencing these values and the fluctuations among them. The large standard deviation of 132 720.15, motivates a proper method for predicting the harvest at an early stage.

For a more accurate depiction of yield, the area of the orchard must be considered, as well as the year the palm trees in the particular orchard were planted.

For the calculation of mean yield per area (in kg/ha), all harvest entries from 2010 to 2020 where the orchard was specified, were considered. The hectare per orchard was obtained for the Orchard description dataset containing entries with values for calculated hectare, year planted, irrigation method and growth stage for some orchards. The yield per hectare calculation is only possible for orchards

present in both the Harvest and the Orchard description dataset, *i.e.*, if data on the harvest quantity of the orchard, as well as information of the orchard size, are available.

Table 5.4: Mean yield per area in kg/ha for all available orchards

| Orchard | Mean Yield (kg/ha) | Orchard | Mean Yield (kg/ha) |
|---|---|---|---|
| 3 | 57 494 | 4 | 31 170 |
| 5 | 12 539 | 8 | 16 598 |
| 9 | 10 592 | 10 | 15 377 |
| 11 | 13 368 | 12 | 18 194 |
| 13 | 13 246 | 14 | 12 294 |
| 15 | 16 422 | 16 | 5 467 |
| 17 | 14 951 | 19 | 4 394 |
| 75 | 16 865 | 70 | 18 182 |
| 30 | 4 209 | 33 | 16 093 |
| 31 | 3 313 | 35 | 18 483 |
| 37 | 1 977 | 40 | 6 770 |
| 43 | 17 576 | 44 | 16 951 |
| 45 | 17 048 | 61 | 16 781 |
| 51 | 19 350 | 38 | 17 951 |
| 39 | 20 451 | 50 | 19 765 |
| 42 | 18 606 | 48 | 7 617 |
| 47 | 20 471 | 56 | 15 636 |
| 57 | 17 311 | 58 | 16 450 |
| 46 | 18 202 | 49 | 18 190 |
| 41 | 6 764 | 67 | 1 584 |
| 71 | 16 433 | 74 | 14 573 |
| 90 | 13 838 | | |

For this study, only complete orchard harvests for orchards with entries from 2010 to 2020 were considered. Thus, younger orchards harvested for the first time after 2010 were not included in the rest of the study.

After noticing an abnormally large yield for Orchard 3, it was found that all harvests in 2016 and most harvests in 2017 from another orchard, Orchard '3,4', were entered in Orchard 3 and Orchard 4. The mean was recalculated with 2016 and 2017 excluded. Considering harvests from 2010 to 2019 with 2016 and 2017 removed, the picture changes and the previous outliers are eradicated. The highest yield per area is obtained in Orchard 47 and the orchard with the smallest yield is Orchard 67. Orchard 47 was planted in 1991 and is irrigated with flood irrigation while Orchard 67 was planted in 2011 with micro irrigation.

Invalid entries in the harvest data for years 2016 and 2017 were corrected. Rearranging according to the year in which the trees of the orchard were planted and correcting the erroneous entries in Orchard 3 and 4 with consideration of production unit numbers for clarification, leads to the data displayed in Figure 5.3. The orchards are displayed from oldest to youngest, as the legend displaying the year in which the trees of the orchard were planted, indicates.

The bars in Figure 5.3 display yield per area in hectares for all harvested orchards. Here the influence of the age of the date palms is evident. As discussed in the literature, the trees start bearing

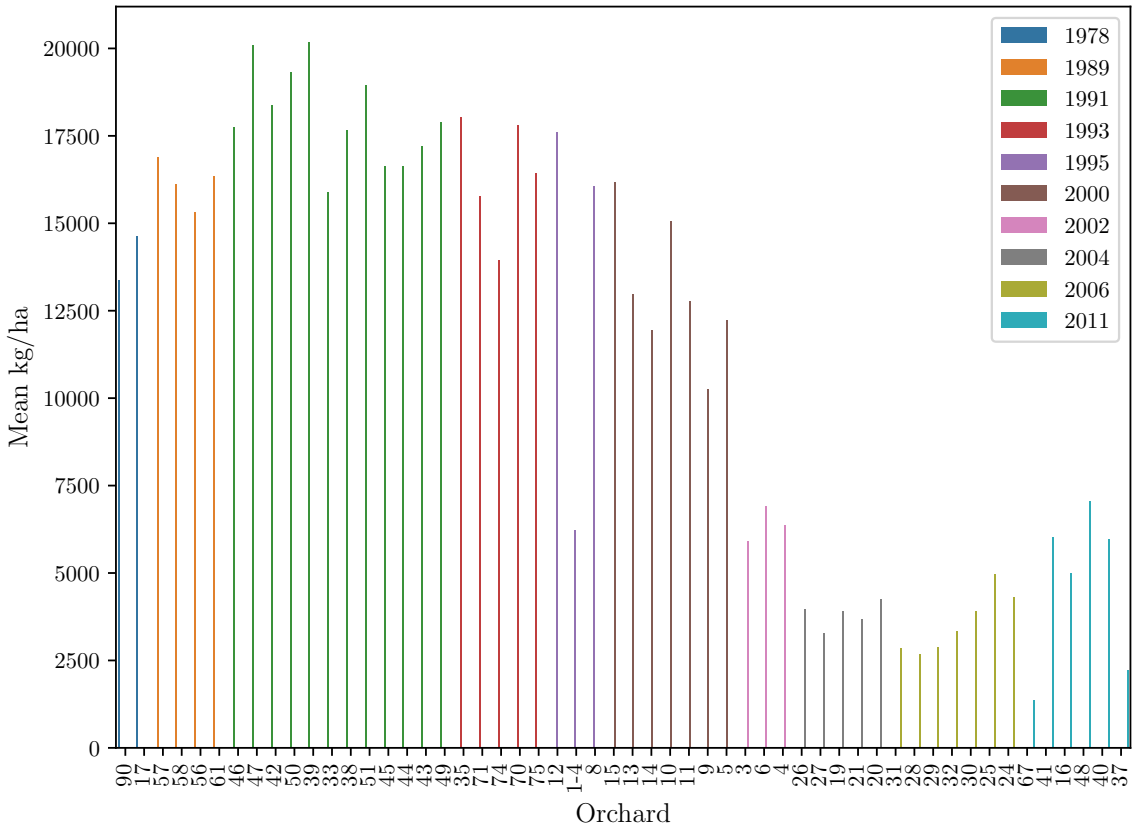



Figure 5.3: Mean yield per ha of orchards

fruit after seven years, depending on propagation type, reach maturity and peak production at 30 to 35 years of age, and may live over 100 years.

Although it is not possible to characterise the distribution of the data (from the visualisations) with certainty, normality is assumed for most calculations where a specified distribution is required. Insufficient data may be the cause for a normal distribution to appear skewed, or of the difficulty of characterising the distribution.

## 5.6 Meteorological data

Weather measurements (temperature, humidity, rainfall, wind speed, wind direction) were obtained from the research partner and the online source, POWER (NASA, 2021). These datasets, of which all the columns are classified as numerical interval data, are described in Table 5.5. The hourly and daily datasets are concatenated. Because the data on the Penman Eto (evapotranspiration) is only available for four years and is inadequate, it is not explored further. In the column names, 'Temperature' is abbreviated as 'Temp'.

Data exploration is done on the entire concatenated weather dataset, from 2008 to 2019 to account for weather up to two years before harvest. For the histogram plots, hourly entries were aggregated to daily entries.

From Figure 5.4, displaying rolling means over 10-day periods of adapted versions of the four weather measurements, namely temperature, wind speed, rainfall and humidity, it is noticed that the weather station stopped functioning and did not record maximum wind speed for a period in 2016. Consulting the original weather dataset, this period is identified as 2016-11-24 11:00 to 2016-12-28 10:00 where the value remained '0.38' for the hourly maximum wind speed measurement.

Table 5.5: Short statistical description of combination of weather datasets

| | Min Temp (°C) | Mean Temp (°C) | Max Temp (°C) | Sum of Rainfall (mm) | Humidity (%) | Wind Speed (m/s) |
|---|---|---|---|---|---|---|
| mean | 16.54 | 23.18 | 30.07 | 0.25 | 33.61 | 5.54 |
| std | 6.22 | 6.51 | 7.07 | 1.70 | 13.96 | 1.94 |
| min | 0.40 | 7.88 | 10.10 | 0.00 | 5.86 | 0.38 |
| median | 16.50 | 23.54 | 30.84 | 0.00 | 31.29 | 5.43 |
| max | 34.70 | 37.52 | 45.10 | 45.80 | 86.31 | 16.02 |

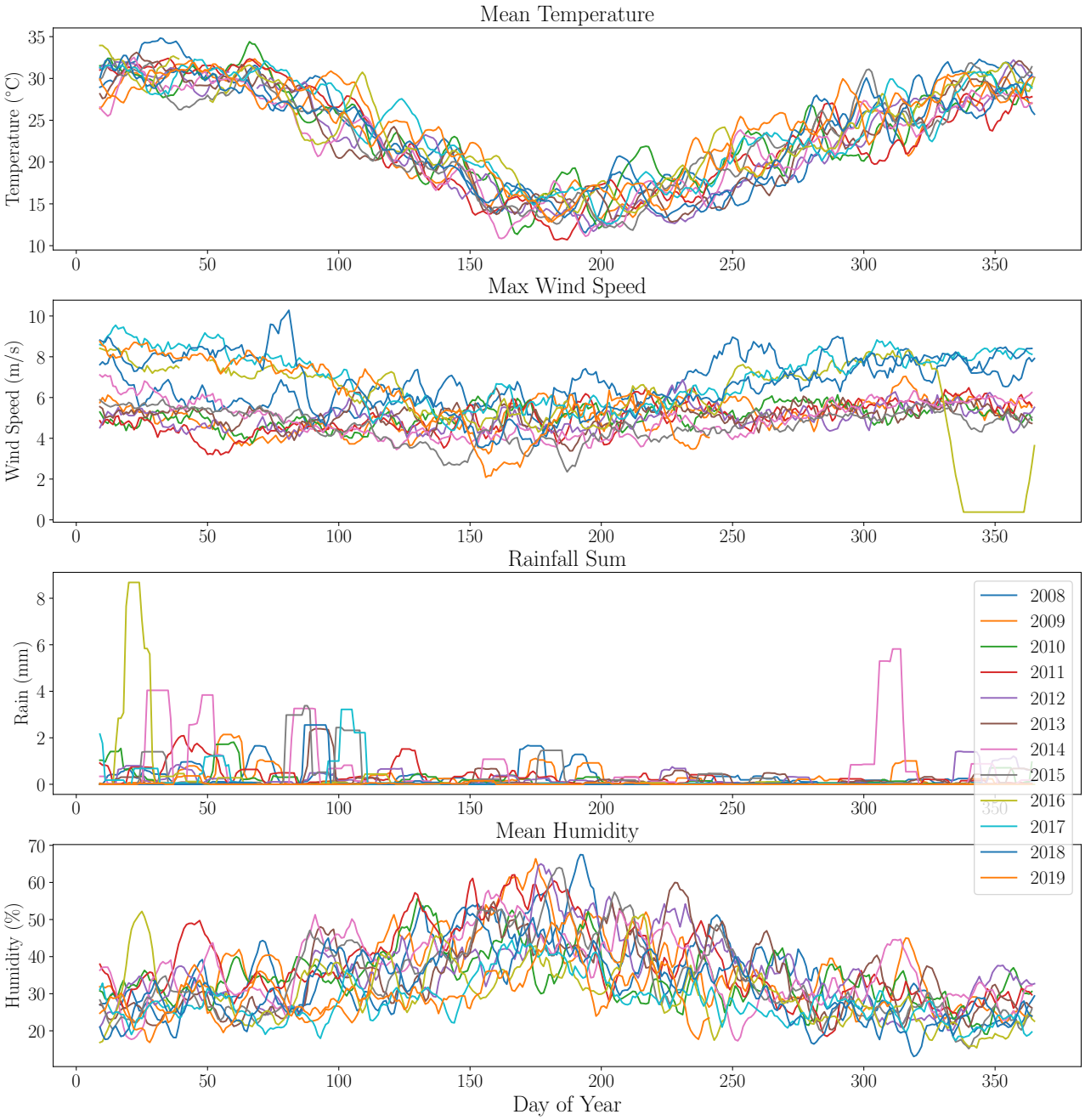



Figure 5.4: Mean temperature, maximum wind speed, sum of rainfall and mean humidity, averaged over dekads (10-day periods)

### 5.6.1 Air temperature

The literature study revealed that the optimum air temperature for the Medjool date palm is 38 °C. The boxplots in Figure 5.5 display statistical measures of the processed meteorological data, with the means shown by the triangles.

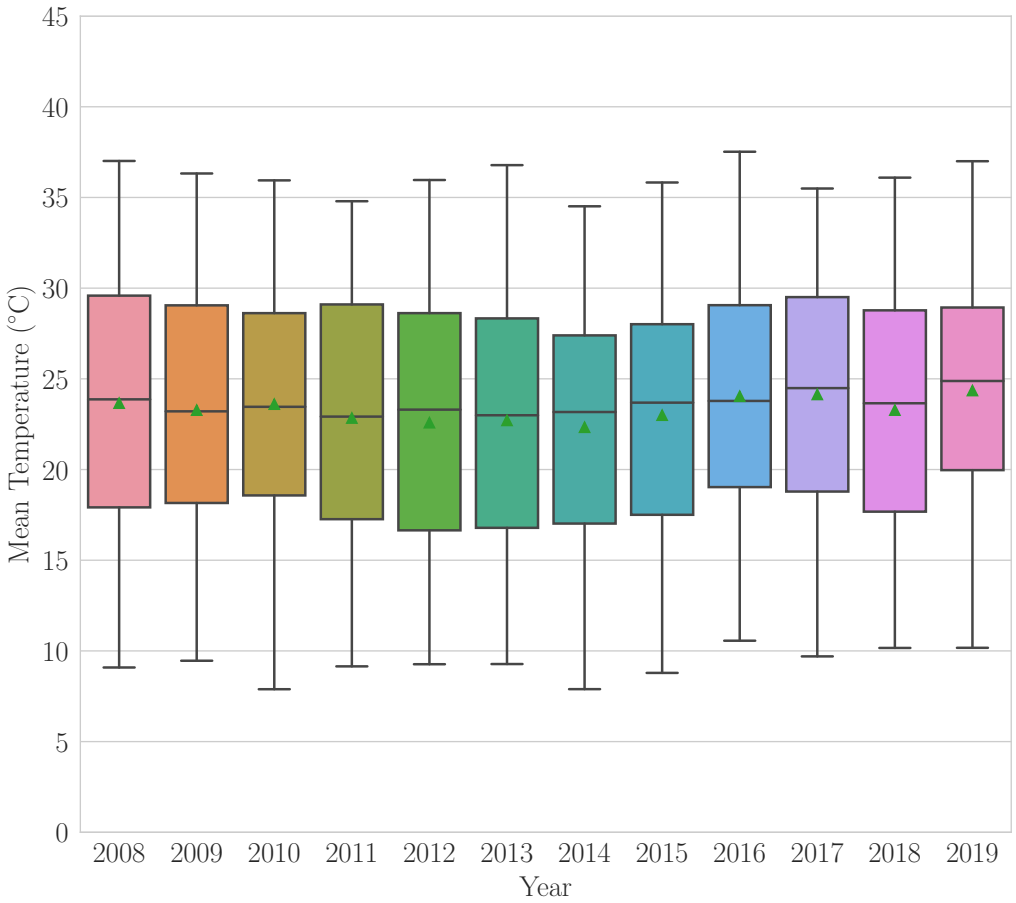



Figure 5.5: Boxplots of daily air temperatures per year

High temperatures are also visible in Figure 5.6, where 2016 has the highest average temperatures.

### 5.6.2 Humidity

According to a subject matter expert (SME), the Medjool cultivar is particularly sensitive to high temperatures above 40 °C in combination with low humidity (below 5%) just before harvesting, *i.e.* in middle January and early February, before the fruit is ripe. Transpiration leads to dehydration of the fruit, causing it to wrinkle before it is completely dry and possibly results in fruit drop.

Considering the weather from 26 December 2013 to 15 August 2019, Table 5.6 shows that a combination of low humidity and high temperatures occurred in the month of January, before fruit maturation and harvesting which usually commence early in February.

### 5.6.3 Rainfall

Although irrigation is the main source of water for palms, rainfall is also investigated, as Chapter 2 describes the adverse effects of rain at certain times. Rainfall is summed daily, as an average of hourly rainfall measurements would not be a useful indication and the total rainfall is taken into account to

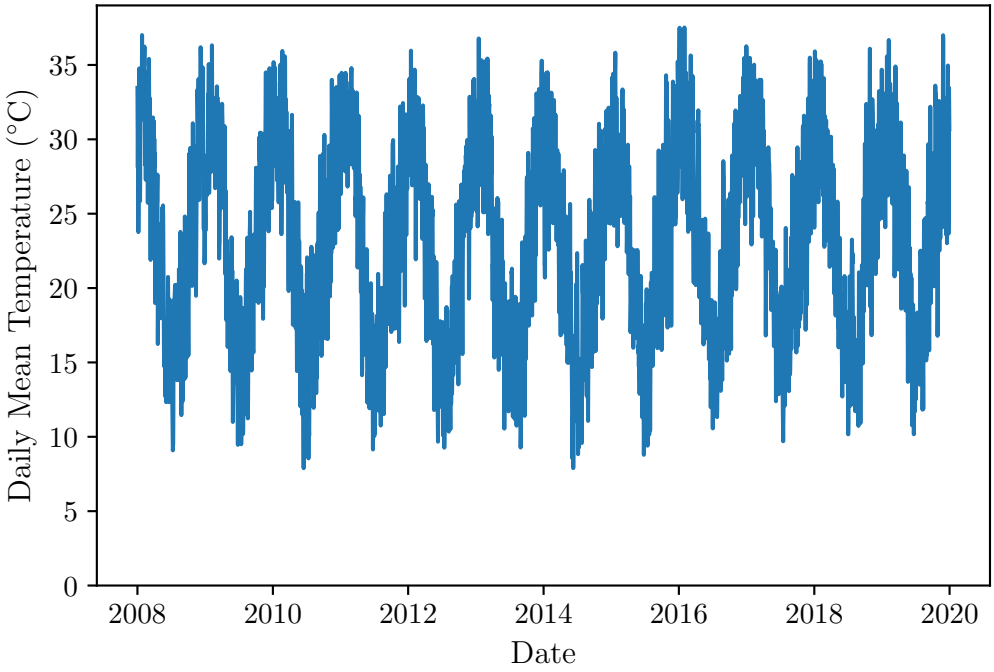



Figure 5.6: Daily mean air temperature

Table 5.6: Entries in weather data with high temperature and low humidity in January

| Date and Time | Temperature (°C) | Humidity (%) |
|---|---|---|
| 2018-01-04 16:00 | 43.22 | 4.9 |
| 2019-01-21 15:00 | 41.00 | 4.0 |
| 2019-01-21 16:00 | 41.68 | 4.0 |
| 2019-01-21 17:00 | 41.67 | 4.1 |
| 2019-01-21 18:00 | 41.54 | 4.4 |

determine possible detrimental effects. It is noticeable that the rainfall is particularly low, with the rainfall of this arid region between 50 and 100 mm per annum.

## 5.7 Exploration of raw datasets

The raw datasets obtained from the research partner are processed and investigated to discover meaningful relationships and to evaluate the use of the dataset in the rest of the study.

### 5.7.1 Growth measurements

The slopes or gradients of weekly growth measurements, using the calculated average mass, are determined with linear regression for each orchard and each season. This was done to distinguish fast-growing from slower-growing orchards. Furthermore, the date of pollination has been recorded only from 2015. Being different every year the orchards were grouped according to the pollination date and the differences were correlated with the weather measurements to find relationships. The weather of the past week is only one of the many factors determining the growth rate of the particular week. Other factors include the current development stage of the fruit, the soil modal profile, as well

as the age of the palm tree.

The slopes (gradients) of weekly growth measurements are determined with linear regression for each orchard and each season. The means of all the orchards are taken for each season, to find slower and faster fruit-growing years. This showed that 2017 had the slowest fruit growth rate and considering each orchard per season, Orchards 58 (2014 and 2017), 70 (in 2015), 8 (in 2013), 6 (in 2016), 9 (in 2018) and 4 (in 2019) grew the slowest. Maximum growth rates per year are measured for Orchard:

- 13 in 2013
- 8 in 2014
- 10 in 2015
- 46 in 2016
- 3 in 2017
- 5 in 2018
- 50 in 2019 (which also has the fastest growth rate of all the available seasons and orchards)

### 5.7.2 Pollination date, growth rates and harvest mass

The pollination date of each orchard, combined with the differences in fruit mass and the corresponding weeks, are tabulated. The differences are simply calculated as the differences in calculated fruit mass, from weekly measurements, for each orchard and season. Table 5.7 shows the second entry for a few sample orchards (for which pollination dates as well as growth measurements are available). The first entry has no previous entry with which to compare the mass of the current week.

Table 5.7: Differences in weekly fruit mass

| Pollination Date | Orchard | Difference (g) | Season | Week |
|---|---|---|---|---|
| 2015-08-07 | 3 | 1.3 | 2015 | 4 |
| 2015-08-03 | 4 | 1.6 | 2015 | 4 |
| 2015-08-07 | 5 | 1.8 | 2015 | 4 |
| 2015-08-07 | 6 | 1.4 | 2015 | 4 |
| 2015-08-12 | 7 | 2.1 | 2015 | 4 |
| 2015-08-13 | 8 | 1.5 | 2015 | 4 |
| 2015-08-12 | 9 | 1.2 | 2015 | 4 |
| 2015-08-13 | 10 | 1.6 | 2015 | 4 |
| 2015-08-12 | 13 | 1.8 | 2015 | 4 |

The means of differences are grouped by the measured week and pollination date, aggregating orchards. An excerpt of four entries is displayed in Table 5.8.

Table 5.8: Means of differences in fruit mass

| Pollination Date | Week | Difference (g) |
|---|---|---|
| 2015-08-03 | 4 | 1.6 |
| 2015-08-03 | 5 | 1.8 |
| 2015-08-03 | 6 | 2.5 |
| 2015-08-03 | 7 | 2.9 |

A dataset is constructed considering the date of pollination of an orchard and calculating the statistics for temperature and humidity on that day. Furthermore, the statistics for temperature and

humidity are calculated for the week of all the growth stages. These features include characteristics such as the orchard number, the pollination date, the yield of the orchard, and the growth rate of the orchard for the season; and the following meteorological measures on the pollination date as well as in each of the development stages *Hababouk, Kimri, Khalal, Rutab* and *Tamar*: Humidity, Maximum temperature, Mean temperature, Sum of rainfall. A dataset is also constructed for the accumulation of heat units and the fruit mass at weeks 1 to 12.

To find possible relationships between the features, correlation analysis was done on this dataset. The correlation coefficients between the heat units and the mass at the various weeks are weak, all below 0.1. The Pearson correlation coefficient between the mean difference in fruit mass, *i.e.*, the growth rate, and the harvest of the orchard at the end of the season, is -0.172. All coefficients between the harvest and the other features (weather conditions of the various growth stages) as well as coefficients between the mean rate and the weather at the growth stages are all low, *i.e.*, positive coefficients are below 0.4 and negative coefficients are above -0.4. To account for the possibility that the relationship between features and output may not be linear, a new method was briefly attempted with the investigation of the data by means of support vector regression (SVR). Support vector machines and SVR as such, are a class of algorithms characterised by the usage of kernels, absence of local minima, sparseness of the solution and capacity control obtained by acting on the number of support vectors. The following features were set up for use in predicting weekly differences in growth:

1. Mean temperature on the day of pollination.
2. Sum of radiation for the week preceding the measurement.
3. Maximum humidity for the week.
4. Mean temperature for the week.
5. Minimum temperature for the week.
6. Number of the week (starting the first measurement of the season as Week 1).
7. Month of pollination.
8. Day of pollination.
9. Sum of rain for the week.
10. Year of pollination.
11. Maximum temperature for the week.

After considering and correcting faulty entries of the growth measurements, the SVR produced an $R^2$ score of 0.43. $R^2$ and other evaluation metrics are formally discussed in Section 6.3, with the discussion of models used in the rest of this study. The weak relationships with the weather on the pollination dates shown with the correlation coefficients and the low score from the SVR led to a termination of the investigation of the usefulness of pollination dates for this study.

A dataset was constructed for all the orchards from 2010 to 2020 with the following columns: Orchard, Season, Fruit mass in week 12, Harvest, Number of palm trees, Harvest per tree and Year planted (age of orchard).

Exploratory investigations were done on this dataset as well. Taking the 12th week of growth measurements and the age of the orchard into account, linear regression was performed to predict the harvest mass per tree at the end of the season. The resulting model has an Adjusted $R^2$ value of 0.539 and a root mean squared error (RMSE) of 21.241 kg. This is relatively high considering the mean harvest per tree, for this model training data is 83.86 kg. However, the coefficients give an indication of the relationships between the variables. The unscaled coefficient of the orchard age is 4.07 and of the fruit size in week 12 is -1.16. The positive influence of the orchard age is evident and biologically rational. The negative coefficient of the fruit size, predictable from the negative correlation with the harvest, points to the inverse relationship and to the fact that more fruits indicate increased total harvest but decreased fruit size.

### 5.7.3 Clarifying differences in number of date bunches

A short investigation into the data on the number of bunches and bunch mass is presented here. The main purpose is first to find if there is a difference between the number of bunches produced by trees on the inside and on the outside of an orchard. If there is, in fact, no statistically significant difference between the production of trees on the inside and outside, labour costs can be reduced by only counting, for instance, a single tree instead of all the bunches in the orchard. Secondly, the investigation aims to explain the difference in bunch counts and evaluate if the solar exposure accounts for the number of bunches on a tree. For all the orchards, the correlation coefficient values between the harvest per tree and the number of bunches per tree are calculated and vary from 0.3 to 0.7, Generally, in younger trees there is a stronger relationship between number of bunches and harvest mass. The yearly growth difference is much more visible in young trees than in older, full-grown trees.

The research partner's farm is located in an arid region in South Africa, with the majority of the date palm orchards under investigation lying in an elongated block beside a permanent river. The map on Google Earth was consulted throughout, to gather information on the layout and position of the orchards. For reference purposes, a map is displayed in Figure A.1 in Appendix A. The soil type and moisture content of the various blocks differ vastly between the soil near the river and that closer to the desert further from the riverbank. Particularly, orchards growing on the riverbank have more loamy soil and are not as exposed to the afternoon sun. In contrast, the orchards on the opposite side of the block are on the west side, more exposed to the afternoon sun and with more sandy soil.

The map was used specifically to distinguish between the orchards bordering the river on the eastern side of the block of orchards with silt soil, and orchards on the western side, with drier sandy soil. The rows of trees on the outer edges of the orchard, without adjacent orchards next to them, were identified, as well as the trees in the centre of the orchard that were surrounded by at least four trees. The numbers of bunches of the inside trees and the outside trees of the orchards for the years 2010 to 2020, or the years in which the trees bore fruit bunches, were used to calculate the average number of inside and outside bunches for an orchard. The orchards were grouped according to their positions in the block of orchards. In layman's terms the groups were categorised as follows:

1. Orchards on the riverside, referred to as "river orchards".
2. Orchards in the middle, referred to as "middle orchards".
3. Orchards on the dry western side, referred to as "dry orchards".

To synthesise from these measurements, the hypotheses state the following:

- The null hypothesis ($H_0$): The means of the number of bunches on trees on the inside and outside of the orchards are equal.
- The alternative hypothesis ($H_1$): The means of the number of bunches on trees on the inside and outside of the orchards are unequal.

The t-test is used to determine how significant the differences between groups are and if the differences, measured in means, can be confirmed or could happen by chance.

Student's t-tests are used to compare averages. The $t$-score is a ratio between the difference between two groups and the difference within the groups. Every $t$-value, or $t$-score, is accompanied by a $p$-value. This $p$-value represents the probability that the results from the sample data are by chance. In the two-tailed t-test for equal means, $t$ can be very large or very small, indicating a difference in the means. The two-sample t-test assuming unequal variances was used, because of the nature of the bunch data.

For the orchards on the western side, the mean number of inside bunches and the mean number of outside bunches were compared by means of the t-test. The $t$-value and $p$-value were calculated with

a two-tailed t-test for a chosen alpha level of 0.05. The calculated $t$-value is larger than the table value corresponding to an $\alpha = 0.05$ and the $p$-value is smaller than the $\alpha$. The hypothesis that there is no statistically significant difference between the means, can be rejected. The results of the t-test for the dry orchards are shown in Table 5.9. The small $p$-value indicates that the hypothesis can be rejected, and that there is indeed a statistically significant difference between the number of inside and outside bunches.

Table 5.9: Results of two-sample t-test assuming unequal variances for comparing means of bunches on outside and inside trees in dry orchards

| | **Outside bunches** | **Inside bunches** |
|---|---|---|
| Mean | 17.562 | 14.298 |
| Variance | 2.275 | 3.393 |
| Observations | 44 | 44 |
| Hypothesised Mean Difference | 0 | |
| df | 83 | |
| t Stat | 9.093 | |
| P(T<=t) one-tail | 2.13E-14 | |
| t Critical one-tail | 1.66342 | |
| P(T<=t) two-tail | 4.26E-14 | |
| t Critical two-tail | 1.989 | |

This difference in the means can be explained by the position of the orchards and the solar exposure they receive. Photosynthetically active radiation (PAR), photosynthetic performance and light quality of the date palm are discussed in Subsection 2.4.1.5. This serves as substantiation for the explanation of the difference in means. In terms of agricultural practices, it cannot be assumed that the number of bunches on a particular tree in the orchard is representative of the entire orchard, and so the inside and outside bunches must be counted. It can be assumed that the trees more exposed to sunlight will produce greater harvests, if the bunch count is correlated with the harvest mass.

### 5.7.4 Exploration of the harvest data

Finding a relationship between the weekly growth measurements of the single fruit and the harvest mass of that orchard at the end of the season could lead to predictions of yield with many benefits such as assisting in budgeting and early planning. For this analysis, the Pearson coefficient is calculated. Numerically, the Pearson coefficient is represented in the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1.

As seen in the correlation matrix in Table 5.10 the correlation between the weekly measured fruit mass and the size of the harvest at the end of the season is very weak. The Pearson correlation coefficient between the harvest per tree and the measured fruit mass in week 14 (the last week for any measurements) is -0.2. The negative correlation with week 12 is a bit stronger, with a coefficient of -0.22. This value, close to zero, indicates a weak negative correlation between the two variables. This indicates that an increase in the measured mass in week 12 leans towards a decrease in harvested yield at the end of that season. It can be gathered that the number of fruits per palm is indicative of the yield, and fruit size decreases as the number of fruits increases. However, one would expect larger fruit to indicate a larger yield.

The scatter plot in Figure 5.7 shows the fruit mass as measured in week 12 (showing the strongest correlation in Table 5.10) and the corresponding harvests per orchard. There is a strong correlation between the size of the harvest and the age of the palm trees, which is expected as the literature shows the full production of the date palm peaking at around 30 years of age. Taking the number of fruits per tree into account would be important if this number varies greatly among the trees. However, the

Table 5.10: Pearson correlation coefficient matrix

| | Mass week 3 | Mass week 4 | Mass week 9 | Mass week 10 | Mass week 12 | Mass week 14 | Harvest | #palmtrees | Harvest/tree | Year Planted |
|---|---|---|---|---|---|---|---|---|---|---|
| Mass week 3 | 1.00 | 0.97 | 0.47 | 0.49 | 0.24 | 0.45 | 0.00 | 0.12 | -0.07 | -0.03 |
| Mass week 4 | 0.97 | 1.00 | 0.55 | 0.57 | 0.30 | 0.38 | -0.06 | 0.07 | -0.09 | -0.08 |
| Mass week 9 | 0.47 | 0.55 | 1.00 | 0.91 | 0.82 | 0.66 | -0.02 | 0.12 | -0.06 | -0.12 |
| Mass week 10 | 0.49 | 0.57 | 0.91 | 1.00 | 0.90 | 0.76 | -0.07 | 0.05 | -0.09 | -0.14 |
| Mass week 12 | 0.24 | 0.30 | 0.82 | 0.90 | 1.00 | 0.90 | -0.20 | -0.03 | -0.22 | -0.15 |
| Mass week 14 | 0.45 | 0.38 | 0.66 | 0.76 | 0.90 | 1.00 | -0.17 | 0.00 | -0.20 | -0.21 |
| Harvest | 0.00 | -0.06 | -0.02 | -0.07 | -0.20 | -0.17 | 1.00 | 0.74 | 0.85 | -0.61 |
| #trees | 0.12 | 0.07 | 0.12 | 0.05 | -0.03 | 0.00 | 0.74 | 1.00 | 0.31 | -0.05 |
| harvest/tree | -0.07 | -0.09 | -0.06 | -0.09 | -0.22 | -0.20 | 0.85 | 0.31 | 1.00 | -0.65 |
| Year Planted | -0.03 | -0.08 | -0.12 | -0.14 | -0.15 | -0.21 | -0.61 | -0.05 | -0.65 | 1.00 |

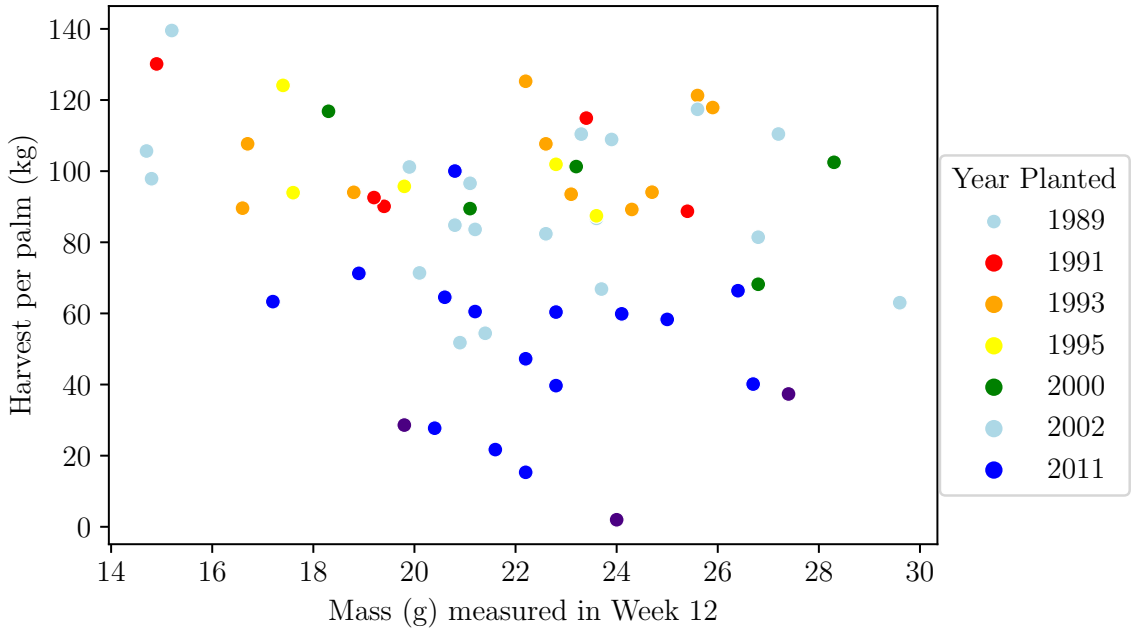



Figure 5.7: Mass measured in week 12 vs harvest per palm tree

bunch removal and thinning would usually lead to a constant number of fruits. As learned from the literature and site visit, the thinning process is meticulous and plays a vital role in the eventual size of the fruit. Depending on the intended market, the size of the fruit could be more important than the total production.

Due to the impact of human interventions – especially the effects of thinning – predicting the harvest from the fruit measurements is not a simple task and nor is it conclusive.

### 5.7.5 Temperature to heatwaves and heat units

The research partner expressed a supposition that heatwaves affect the yield, which is supported by the literature on temperature requirements and tolerances, and the knowledge that climatic conditions have an effect on the production. To determine if there is indeed a relationship between the heatwaves and yield, the data is further explored. According to the South African Weather Service (SAWS), heatwave criteria in South Africa are defined as follows: "If the maximum temperature at a particular town is expected to meet or exceed 5 °C above the average maximum temperature of 'the hottest month' for that particular place, as well as persisting in that mode for three days or more, then a heatwave may be declared." The SAWS has outlined threshold values to be met or exceeded at different regions for a heatwave to occur.

Considering a threshold value of 44 °C, the temperatures of only five entries are equal to or exceed this threshold. These entries, all from January 2016, are shown in Table 5.11.

Table 5.11: Entries in weather data with temperature above 44 °C

| Date and Time | Temperature (°C) |
|---|---|
| 2016-01-04 18:00:00 | 44.0 |
| 2016-01-05 16:00:00 | 44.1 |
| 2016-01-05 17:00:00 | 44.1 |
| 2016-01-05 18:00:00 | 45.1 |
| 2016-01-05 19:00:00 | 44.8 |

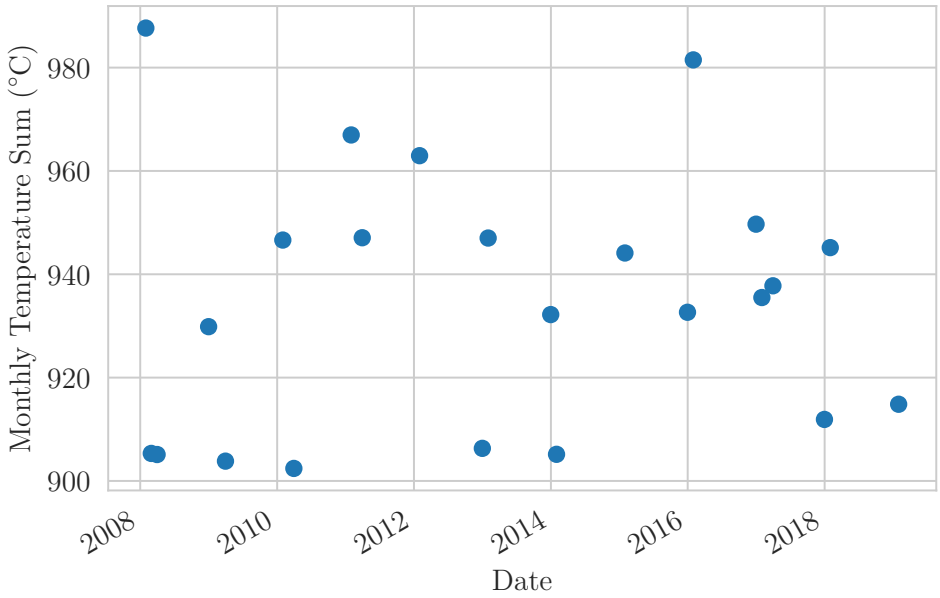



Figure 5.8: Monthly sum of mean air temperatures where the sum > 900 °C

The weather statistics of a site close to the farm were used as reference. In the month of January, the average temperature on this site as reported by Climate-Data.org (2020), is 28.4 °C. The annual rainfall is 79 mm, and its average annual temperature is 21.5 °C.

Rather than only detecting maximum temperatures per day the investigation is continued by summing the daily mean temperatures of the entire month. To clearly identify especially warm periods, Figure 5.8 displays the monthly *sums* of daily mean temperatures above 900 °C. For each year, at least one month had a temperature sum of 900 °C or higher. This was done to visually compare the hottest months. From this figure it is clearly visible that January of 2008 and 2016 are the hottest months, as displayed in Figure 5.6 as well. The highest monthly sums are 987.65 °C in January 2008 and 981.49 °C in January 2016, equating to an average daily temperature of around 31 °C, including night temperatures.

Following the heatwave criteria from the SAWS, heatwaves were detected in the original dataset for the purpose of data analysis. The research partner revealed the possibility that extremely hot temperatures at the end of October 2018 (more than 5 degrees above October's maximum of 28.1 °C) for three consecutive days led to a decrease in the 2019 harvest. To investigate heatwaves, as well as heat units, the original weather dataset was used.

Considering all weather measurements, Table 5.12 shows the days which were the beginning of five-day heatwaves, where the maximum temperature on a day is 5 degrees warmer than the particular month's average maximum.

The definition of a heatwave of three consecutive days with maximum temperatures higher than the month's average maximum temperature of the region, leads to the number of heatwaves shown in Table 5.13.

Suggestions to investigate the influence of the heatwaves were followed, however no strong indication of a correlation was found, and the heatwaves were not furthered considered.

As described in Chapter 2, growing degree days (GDD) are often used in phenology. Calculating GDD can determine if the crop grows in adequately high temperatures in order for blooming and other phases to occur properly. The heat unit requirement of date palm cultivars varies. The Medjool requires temperatures in excess of 1 500 heat units, or degree days, a concept introduced in Subsection 2.4.1.1, above 18 °C and are generally cultivated in arid conditions.

Table 5.12: Starting days of five-day heatwaves

| Date | Maximum Temperature (°C) | Date | Maximum Temperature (°C) |
|---|---|---|---|
| 2014-07-30 | 26.9 | 2016-08-27 | 36.0 |
| 2014-09-04 | 33.3 | 2017-05-01 | 32.8 |
| 2014-09-05 | 34.3 | 2017-10-29 | 37.5 |
| 2015-10-26 | 37.6 | 2018-07-22 | 28.2 |
| 2016-04-11 | 37.7 | 2018-10-06 | 38.1 |
| 2016-04-12 | 37.1 | 2018-10-25 | 37.4 |
| 2016-04-13 | 38.1 | 2019-05-11 | 32.9 |
| 2016-04-14 | 38.1 | 2019-05-12 | 34.9 |
| 2016-08-25 | 32.9 | 2019-08-21 | 33.3 |
| 2016-08-26 | 34.7 | 2019-08-22 | 33.5 |

Table 5.13: Number of three-day heatwaves in each month

| Year | # 3-day Heat-waves | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2014 | 7 | | | | | 1 | 1 | | | 4 | 1 | | |
| 2015 | 6 | | | | | | | | | 1 | 5 | | |
| 2016 | 15 | 1 | | | 6 | | | | 5 | | 1 | 2 | |
| 2017 | 12 | | | | 2 | 5 | 1 | 3 | | | 1 | | |
| 2018 | 12 | 1 | | | | 1 | | 3 | | | 6 | 1 | |
| 2019 | 14 | | | 1 | | 6 | | 1 | 6 | | | | |

The Year Heat units column in Table 5.14 displays units for the previous year, as harvesting takes place at the beginning of the year. The Season Heat units column contains the heat units calculated from the daily temperatures between October and November, the period during which the fastest growth occurs.

Table 5.14: Sum of heat units of previous year with its harvest mass

| Year | Season Heat units | Year Heat units | Dates Total mass (kg) |
|---|---|---|---|
| 2014 | 685.90 | 115.10 | 812 083 |
| 2015 | 2 379.30 | 4 246.10 | 943 248 |
| 2016 | 2 489.40 | 4 496.90 | 898 332 |
| 2017 | 2 427.27 | 4 753.04 | 1 049 651 |
| 2018 | 2 427.03 | 4 799.07 | 993 454 |
| 2019 | 2 361.57 | 4 395.97 | 862 251 |

As the heat and exceptionally warm periods are explored, extremely cold measurements should also be identified. As discovered from the literature study in Chapter 2, extended periods of extremely cold temperatures can adversely affect the date yield. However, no periods longer than a single day of minimum temperature 5 degrees below the minimum average temperature for the month were found in the weather data. The entries with temperatures 5 degrees lower than the minimum average are displayed in Table 5.15. Because low temperatures are not very prevalent in this area, and only five

Table 5.15: Entries in the original weather data with temperatures 5 degrees lower than the minimum average of the region

| Date | Temperature (°C) |
|---|---|
| 2014-01-06 | 13.90 |
| 2014-03-26 | 12.40 |
| 2014-06-11 | 0.40 |
| 2016-03-28 | 12.82 |
| 2018-11-06 | 8.42 |

entries were identified in the original data, it was not further investigated.

### 5.7.6 Converting meteorological measurements into features

The Pearson correlation is calculated between total yearly harvest and weather features to investigate the existence of relationships between the yield and weather. The features were resampled to a single value per year, which effectively disregards valuable information and many intricacies in the data. These coefficients are displayed in Table 5.16. Correlation coefficient values above 0.5 or below -0.5

Table 5.16: Correlation coefficient between total yearly harvest mass and yearly resampled values for weather measurements

| Weather measurement | $r$ |
|---|---|
| Penman Evapotranspiration | 0.48 |
| Max Humidity | **-0.74** |
| Mean Humidity | -0.47 |
| Min Humidity | -0.14 |
| Radiation Sum | 0.33 |
| Max Temp | 0.26 |
| Mean Temp | **0.59** |
| Min Temp | 0.39 |
| Max Wind Speed | -0.26 |

indicating stronger relationships are printed in bold face. It seems that the maximum humidity is negatively correlated while the mean temperature has a positive effect. For this reason, it might be useful to include mean temperature and humidity when constructing a dataset for use in the study.

After processing, the weather data is summarised to be used as predictors, or explanatory variables. The type of algorithm to be used for prediction of the yield now becomes relevant. It is already evident that the weather data may be useful as it contains a range of measurements. However, since the sample size of the harvests equates to only 11 observations (from the 11 years), the model is restricted in terms of complexity. Keeping the hourly measurements would require a model able to translate these hourly entries to an annual yield. Condensing these entries to annual (one per year) could disregard too much valuable information. Quarterly summaries may also be forfeiting too much detail. The effect of two years' weather conditions on the palm tree production is considered. Thus, the meteorological data of the current and previous (**prev**) season is summed into monthly predictors (explanatory variables) of the most useful factors. These are factors which are gathered from the literature and are present in the data. The predictors in the newly constructed meteorological dataset are:

1. Accumulated heat units of the month, named **Heatunits**.

2. Mean humidity of the month, named **Humid**.
3. Maximum temperature of the month, named **Max Temp**.
4. Mean temperature of the month, named **Mean Temp**.
5. Minimum temperature of the month, called **Min Temp**.
6. Sum of the rainfall of the month, simply called **Rain**.
7. Mean wind speed of the month, named **Wind**.

Only temperatures and humidity for the previous season are included implying that wind and rain for the previous season are not considered for the outcome of the current season. This is because the literature study has provided enough evidence that the wind and rain for the previous season will not affect the production of the palm in the following season.

The processed dataset to be used as input features is created from the data and presented here. For illustrative purposes, only the features created from data recorded in the month of January (for both the previous and current season) are statistically summarised in Tables 5.17 and 5.18. Similar features for the months February to December are included in the complete dataset to be used as input to the predictive models.

Table 5.17: Statistical description of weather dataset features, containing 11 observations, constructed from January data

| | **Jan Wind** | **Jan Heatunits** | **Jan Humid** | **Jan Min Temp** | **Jan Mean Temp** | **Jan Max Temp** | **Jan Rain** |
|---|---|---|---|---|---|---|---|
| mean | 6.46 | 385.14 | 28.04 | 17.06 | 30.33 | 42.91 | 20.26 |
| std | 1.57 | 25.40 | 3.18 | 1.84 | 0.86 | 1.12 | 25.82 |
| min | 4.69 | 339.48 | 23.39 | 13.90 | 28.81 | 41.21 | 0.00 |
| median | 5.74 | 388.59 | 27.06 | 16.68 | 30.48 | 42.96 | 14.00 |
| max | 8.69 | 429.75 | 32.98 | 19.80 | 31.66 | 45.10 | 86.80 |

Table 5.18: Statistical description of weather dataset features, containing 11 observations, constructed from previous year January data

| | **prev Jan Humid** | **prev Jan Heatunits** | **prev Jan Mean Temp** | **prev Jan Min Temp** | **prev Jan Max Temp** |
|---|---|---|---|---|---|
| mean | 28.40 | 390.24 | 30.54 | 17.13 | 42.82 |
| std | 2.80 | 26.60 | 0.93 | 1.79 | 1.09 |
| min | 24.62 | 339.48 | 28.81 | 13.90 | 41.21 |
| median | 27.37 | 391.90 | 30.54 | 16.68 | 42.90 |
| max | 32.98 | 429.75 | 31.86 | 19.80 | 45.10 |

### 5.7.7 Converting bunch data into features

The bunch data, which was introduced in Subsection 5.4, consists of a number of bunches per tree in each orchard and for each season as well as an average bunch mass, estimated by measuring an upper, middle and lower bunch on a representative tree in an orchard. These two numbers, bunch mass and bunch count, for each orchard, can be added to the constructed dataset in the preceding section as features named bunch mass and bunch count. They are added for each orchard with a prefix containing the orchard number, *e.g.* '8'bunch mass, and will be handled separately as the individual orchards are considered. The newly constructed dataset then contains the weather features prev Jan Max Temp, *etc.* as well as the bunch features, '12'bunch count, *etc.* for each orchard.

## 5.8 Consideration of yield-influencing factors

Chapter 4 discussed the factors influencing yield and indicating what data is required for a complete statistical model predicting yield. These factors are further investigated in this section. The challenges created by the data available for this study include the small sample size and the restriction in the number of yield-influencing factors. Information on a number of these factors is available and used in the form of processed datasets while records of others could not be supplied and are only discussed for the sake of completeness. These factors are summarised in Table 5.19, indicating which data is available for all years under study and will be used in the rest of the study, and others which can only be discussed from literature for use in future studies.

Table 5.19: Yield influencing factors which are and are not available and used in this study

| Factor | Data available | Data used |
|---|---|---|
| Type of propagation | Yes | No |
| Planting date (age of orchard) | Yes | Yes |
| Layout of orchard | Yes | Yes |
| Topography | No | No |
| Date of pollination | Yes | No |
| Soil data (temperature, moisture content) | No | No |
| Type of irrigation | Yes | No |
| Soil nutrient content | No | No |
| Thinning intensity | No | No |
| Number of bunches | Yes | Yes |
| Bunch mass per orchard | Yes | Yes |
| Meteorological data | Yes | Yes |
| Growth regulators | No | No |
| Pest and disease control | No | No |

Among those measured factors that will not be utilised are the Type of propagation and Date of pollination. Propagation, discussed in Chapter 2, refers to the method of cultivation *e.g.* seed propagation, offshoots or tissue culture. All the orchards concerning this study were propagated by tissue culture. Because of this constant value and no alternative against which to compare it, its influence cannot be determined. The date of pollination was preliminarily used to gain information about the weather conditions during pollination time, but the correlations with these conditions did not provide reasonable substantiation for further investigation. Since the quantities and schedule of the irrigation are not available, the type of irrigation is also not used. Data on other management practices influencing the yield, *i.e.*, growth regulators and pest and disease control, is also not available.

As discussed in this chapter, the meteorological data, orchard descriptions, growth measurements, thinning dates, number of bunches after thinning and the pollination dates are available for this study.

The additional factors for which data is not available can be included in future studies. Recommendations follow on how this data can be manipulated. For the farming practices irrigation, fertilisation and thinning, approximate values from literature are also provided.

Since neither data on the soil, either the modal profile or water and nutrient content were collected for an extensive period, these factors could not be considered during this study. The relevant data that should be recorded in this regard is the water and nutrient content. Irrigation applied to the orchards will influence the moisture content of the soil, which is measurable by soil probes. The irrigation quantities can also be recorded. The water required by the date palm depends on the daily evapotranspiration (ET) rates which are influenced by the atmospheric conditions (solar radiation, temperature, wind and humidity) and the physical and physiological characteristics of the crop (Bhat et al., 2012). According to studies directed by Djerbi (1995) on the date palm it is shown that during a season, for growth, development and producing 1 kg of date fruit the date palm tree requires approximately 2 400 litres of water. Therefore, when estimated to produce a yield of 100 kg per tree, annual irrigation and rainwater should total up to 240 $m^3$ per tree or 28 800 $m^3$ per hectare (120 trees).

Soil samples can be tested for the nutrient content. The quantities of applied fertiliser can also be recorded. The annual application may vary considerably to the extent of 500 – 900 g N, 250 – 800 g P, 300 – 1 300 g K per tree, depending on the cultivar. Djerbi (1995) estimated that in order to produce 100 kg dates, the basic fertilisation needs of the date palm tree are approximately 740 g N, 220 g P and 830 g K applied through irrigation water. Ezz et al. (2010) has determined in order to produce 111.5 kg dates, the Zaghloul and Hallway cultivars require the following annual fertilisation: 700 g of N, 500 g of P and 1 300 g of K per tree. Not much research has been done on the Medjool cultivar but Alhejjaj et al. (2020) found that foliar application of 800 mg/litre potassium improved yield of the Medjool date palm by 31%, fruit size by 10.3% and fresh mass by 25.1%, compared to no application. It is further recommended by Oosthuyse (2018) to maximise nitrogen application by applying potassium as $KNO_3$ or to apply nitric acid.

For the purpose of this study no information could be obtained on the measure or method of annual thinning, but data on the number of bunches after thinning was made available. The number of bunches per tree was employed in an attempt to find a relationship between the number and the total yield obtained.

Not much research has been done on thinning of the Medjool cultivar but for all date palms the bunches per tree should be controlled relating to the age and size of the tree. A prominent, highly cited study by Nixon (1956) for optimum thinning of Medjool dates suggests the bunches per tree should be optimised by removing bunches resulting in a bunch/leaf ratio of 10 leaves per bunch and aiming for 17 bunches per tree. The purpose of thinning is mainly to improve fruit characteristics, causing an increase in fruit size, mass and the flesh/pit ratio. Removing 15% of the strands of a bunch will result in increased bunch mass but removing 30% whole strands will result in decreasing bunch mass and total yield. Fruit dropping of 10 – 20% for the Medjool cultivar should be taken into consideration when the thinning measure is finalised. The ideal will be to have 30 strands per bunch and 10 fruits per strand. In practice it should be an objective to manipulate the fruit quantity per bunch around 300 and aiming for an average fruit mass of 20 g per date (as a semi-dry fruit from 18 to 28 g), in which case the bunch mass will be approximately 6 kg. Research on Medjool dates by Nixon (1956) shows that reducing date fruit to 20 fruits per strand results in almost ten percent gain in total yield, and five percent loss in fresh fruit weight, compared to 16 fruits per strand. A lower number of fruit produces an optimal bunch mass and total yield of a suitable size with high quality fruit.

## 5.9 Summary of real-world data description

This chapter presented the raw datasets obtained from the research partner. Exploration on these datasets was done to identify useful relationships. The effect of solar exposure on the photosynthesis and bunch quantity of the trees was also investigated. Finally, a section is devoted to explain which data will be used in the study and which not.

The following chapter presents theory on predictive modelling and sketches the background of modelling used in the study. The objectives of the study and the data types identified in this chapter point to the predictive modelling techniques described in the next chapter.

# Chapter 6

# Predictive Modelling

Chapter 5 presented the real-world data available for use in this study. It concluded with a summary of which data will be used. This chapter introduces various statistical methods suitable for use on the datasets. First, predictive modelling and linear models are discussed in the context of crop yield modelling. The focus on this specific model type is justified by the characteristics of the data from Chapter 5. Various evaluation metrics are presented, with a focus on those used in this study. A discussion follows on minimum sample size requirements for cases where datasets comprise a small number of observations. Then, different types of feature selection methods are investigated, again with the datasets introduced in the previous chapter being taken into consideration.

## 6.1 Predictive model theory

Relating to the discussion on yield models in Chapter 3, a statistical yield model is characterised by the realisation of a trend function and an algorithm for the model (or multiple algorithms if using a meta model) (Cai et al., 2017).

A subfield of data analytics, namely predictive data analytics, as defined by Kelleher et al. (2015), is the ability to build and apply models that make projections based on relationships and recurring configurations found in built-in data. Predictive data analytics has a large variety of applications. It is often utilised in the following scenarios:

- When predicting price in businesses such as online retailers, predictive analytics models can make predictions based on historical sales.
- Risk assessment, considered in decision-making, benefits from predictive analytics models predicting the decision-associated risk.
- Propensity modelling, in which the likelihood of customers taking various actions is predicted, makes use of historical data on customer behaviour.
- Document classification can be done automatically with the use of predictive data analytics. Diagnoses are made by doctors and scientists who are supported by predictive analytics models to gain insight from past examples.

The word 'predictive' in the term *predictive data analytics* refers to more than the temporal aspect of determining what will happen in the future. With regard to data analytics, 'prediction' is the assignment of a value to an unknown variable. Predicting could refer to determining a document type, making a medical diagnosis or predicting prices for houses in an area.

Supervised machine learning is used to instruct and guide these predictive models. Machine learning is the process of obtaining knowledge from data (Müller and Guido, 2016). The research field combines statistics, artificial intelligence and computer science. The term *machine learning* is believed to have been coined by Arthur Samuel in 1959 with the well-known definition 'Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed'. Although no published documentation exists to substantiate this, the description does provide insight into the topic. Machine learning has various applications, including prediction, natural language processing and statistical pattern recognition.

Machine learning types are usually categorised by the way in which the algorithms learn. According to these categories, the main types of machine learning are:

1. Supervised learning – predictions are made based on a dataset or set of examples with input and output, or labels, made visible, and includes classification, regression (also known as function approximation) and forecasting.
2. Semi-supervised learning – less expensive and time-consuming unlabelled examples are used in conjunction with some labelled data to enhance supervised learning.
3. Unsupervised learning – inferences are drawn from completely unlabelled data; meaningful patterns and groupings inherent in data are found, such as clustering or dimension reduction.
4. Reinforcement learning – this type is used for analysis and optimisation of agent behaviour based on feedback from the environment with the use of trial-and-error and delayed reward.

The focus of this study is specifically on supervised regression problems, owing to the type of data available. Since the constructed datasets are able to provide an output in the form of yield, unsupervised learning is not applicable. A regression problem involves the prediction of continuous values, where classification is the prediction of a categorical variable. When deciding on an algorithm, the following aspects must be considered in particular: training time, accuracy, and ease of use.

The following algorithms could be employed for a regression problem:

- Linear regression (correlation between a continuous interrelated variable and a number of predictors).
- Logistic regression (classification algorithm used for categorisation).
- Linear support vector machine (SVM) and Kernel SVM.
- Trees and ensemble trees.
  Decision trees, random forest, gradient boosting (all based on decision trees) to further partition the feature space into regions. The latter two generally achieve good accuracy and overcome the over-fitting problem of decision trees.
- Hierarchical clustering.
- Spatiotemporal downscaling – particularly for use with climate data.
- Principal component analysis (PCA), singular value decomposition (SVD) and latent Dirichlet allocation for dimension reduction.
- Neural networks and deep learning – neural networks (NN) consisting of an input layer, hidden layers (defining the model complexity and modelling capacity) and an output layer to loosely model the working of the human brain. A variety of NN algorithms are used, including recurrent neural networks (RNN) for sequence generation and multilayer perceptrons (MLP), a class of feedforward artificial neural network (ANN).

Predictive modelling may be utilised to build a best probability model to determine possible crop estimates. It consists of four stages known as descriptive analysis, data treatment (outlier fixing, replacement of missing values), data modelling and estimation of performance (Nagini et al., 2016). A popular predictive modelling technique for continuous data, regression analysis, determines the relationship between a predictor (independent) and target (dependent) variable. Regression analysis has an array of applications, among others, the use of time series data in forecasting, and finding the hidden relationship among the variables. This technique is used to analyse the data and fit a line or curve using the data points by minimising the offset between the data points and the line of the graph (Nagini et al., 2016). Li (2017) suggests the use of principal component analysis for dimension reduction.

After the most prominent features have been identified and selected, the model is fitted with those features and developed for evaluation. The case of a small sample size $n$ was thoroughly investigated and discussed in Section 6.4. This is one of the main challenges of this study, together with understanding and navigating the various influences on the target, the crop yield.

A traditional crop yield model approach is not in the scope of this study, specifically because data on many more factors is required. Also, implementing an existing model would imply a black box model, which does not aid in gaining new insight into the data. An advanced machine learning approach such as neural networks require many more observations (a much larger $n$ of hundreds to thousands of observations). In fact, various algorithms and approaches (*e.g.* AdaBoost, XGBoost, neural nets and support vector machines) were attempted with the data split into a train and test set. These failed not only because of testing on so few data points but also because of the training set being too small. Algorithms like support vector machines also do not assume linear relationships, which were assumed for this study due to the small sample size.

A small sample size is a relative concept. Whether a sample can be considered small depends on the quantity estimated. When estimating the mean of a one-dimensional Gaussian distribution, $n = 10$ may not be too small, though when it comes to estimating the probability of rare events, $n = 10^6$ could be considered sparse. For regression, sample sizes of less than 100, test observations typically have too large test errors.

Bootstrapping, a Monte Carlo method, was considered as a method to improve statistical results with a small sample size. However, bootstrapping does not produce a better point estimate. A sample of $n = 10$ observations contains information from only 10 observations, so bootstrapping cannot give more information or improve the reliability of the confidence intervals as it samples from those 10 observations.

For this study, only linear relationships between features are considered. The reason is two-fold. The model is fundamentally limited by the small sample size. There is also the consideration of parsimony, and a need to choose the simplest possible model, as the variables eventually included in the model should tell a story of their influence, rather than just predict an outcome. These linear relationships are found with linear models. Linear regression is useful as it is a basic yet powerful model for predicting numerical values.

## 6.2 Linear models

This section briefly discusses linear models. When exploring linear models, or any model development, it is imperative to understand the characteristics of a well-designed model. The following characteristics are fundamental for a desirable model:

- Parsimonious – the model must be as simple as possible, in terms of the number of included variables. Parsimony is based on the principle that instead of complex models with numerous variables, simple models with fewer variables are preferred. More variables in the model increase the dependence of the model on the data (Hosmer et al., 2013).
- Identifiability – the values of estimated parameters should be unique, one estimate per parameter.
- Goodness of fit – as much variation in the dependent variable as possible should be explained.
- Theoretical consistency – coefficients must have the expected positive or negative signs.
- Predictive power – the model must be capable of being used to make reliable forecasts.

The most basic model structure is the linear model with equation

$$\mathbf{y} = \beta\mathbf{X} + \epsilon \tag{6.1}$$

where $\mathbf{y}$ is the variable to be explained, known as the response or dependent variable and also, in terms of the machine learning model, called the target or output. $\mathbf{y}$ is a one-dimensional vector of length $n$, where $n$ is the number of observations. $\mathbf{X}$ is the matrix containing the explanatory or independent variables, which are also called the regressors, predictors or inputs, depending on the context. $\mathbf{X}$ has

length $n$ and is $m$-dimensional[1], with $m$ the number of candidate features. $\beta$, the one-dimensional vector of length $m$ hence holds the coefficients, the parameters of interest. The error term, denoted $\epsilon$, is also called the residuals.

Linear regression is one of the simplest and most common modelling techniques. It assumes a linear association between the predictor variables and the response. The most commonly used estimator is the ordinary least squares (OLS), also known as simply least squares. The estimates, or coefficients of the predictor variables, are chosen so as to minimise the square of the distance between predicted ($\hat{y} = \sum_{j=1}^{m} \hat{\beta}_j x_{ij}$) and actual ($y$) values, as a result minimising the loss function

$$\sum_{i=0}^{n} (y_i - \hat{y}_i)^2. \tag{6.2}$$

A lower minimum square error leads to better explanatory power of the regression model. When performing OLS regression, five assumptions are made. These are:

1. Linearity: Every independent variable, multiplied by a coefficient, is summed to predict the output value.
2. No endogeneity: The covariance of the error and the independent variables are zero for any error or variable.
3. Normality and homoscedasticity: The error term is normally distributed. As no errors are expected, on average, the expected value of the error is zero. Homoscedasticity relates to constant variance. The errors should have equal variance. $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = \sigma^2_{\epsilon_m}$. In a heteroscedastic dataset, the points start close to the regression line and move further away. In this example, the smaller values of the independent and dependent variables have a better prediction than the bigger (spread out) values. This is often solved by the removal of outliers or with a log transformation.
4. No autocorrelation: This assumption is also known as no serial correlation. Errors are assumed to be uncorrelated. Serial correlation between errors is common in time series data.
5. No multicollinearity: Two or more independent variables are not highly correlated.

Related to the second assumption, is a problem called Omitted Variable Bias. This is introduced to the model when an independent variable that is relevant, is not included. As each independent variable explains the dependent variable $y$, they are correlated, or have a relationship, to some degree. Similarly, $y$ is also influenced by and correlated with the omitted but relevant variable. The omitted variable is also possibly correlated with at least one independent variable. However, it was mistakenly excluded from the regressors. Since everything that cannot be explained by the model forms part of the error, the error becomes correlated with all the variables. Incorrectly excluding a variable, such as the case of omitted variable bias, leads to biased and counterintuitive estimates with adverse effects on the regression analysis. An incorrect inclusion of a variable leads to inefficient estimates without biasing the regression. Including irrelevant variables is hence preferred above excluding relevant ones.

For multiple linear regression with more than one predictor variable, standardisation and scaling are done on the values. This ensures that the features are normally distributed as it removes the mean of the data (mean values will become zero) and it scales to unit variance. The mean of the training sample's values is subtracted from each of them before they are divided by the standard deviation of the training samples. The standard score of a sample $x$ is calculated as: $z = \frac{x-u}{s}$ where $u$ is the mean of the training sample and $s$ is the standard deviation of the training sample.

When developing the models, it is necessary to distinguish between better and worse models. The following section presents how the linear models can be evaluated.

[1] $p$ is most commonly used in academic writing to denote the number of features or dimensions. However, to avoid confusion with the $p$-value used in statistics, $m$ will be used in this document.

## 6.3 Interpreting linear regression results and reporting evaluation metrics

In regression analysis, an equation describes the statistical correlation between the independent (predictor) variables and the dependent (response) variable. The fit of that equation to the data is verified by the residuals and the interpretation of the results from the analysis. These results include the $p$-values and coefficients of the independent variables. For each independent variable in the equation, the $p$-value tests the null hypothesis that the coefficient of that variable is equal to zero and hence has no effect on the dependent variable. When a $p$-value is small, considered to be when below 0.05, it indicates that the null hypothesis can be ignored and shows the variable to be statistically significant. A larger $p$-value would suggest that the predictor does not influence the response. Regression coefficients are also known as slope coefficients, as they characterise the mean change in the response for a unit of change in the predictor, assuming all other predictors are kept constant. This is more easily interpreted with simple linear regression than with multiple regression. The constant term is also called the y-intercept. The sum of the components of $\mathbf{X}$ multiplied by their coefficients gives a zero term, leading to the intercept as the remainder.

All developed models have to be evaluated and compared. To determine the fit of the model, certain measures are calculated to evaluate models and compare them with the measures of other models (Hawkins et al., 2003). These include the mean of the squares of the residuals called Mean Squared Error (MSE) and R-squared ($R^2$) and are subsequently discussed. Cross-validation as a means of validating models and computing the evaluation metrics is then presented.

### 6.3.1 R-squared measure

R-squared or $R^2$, is the coefficient of determination, a useful measure for evaluating the goodness of fit of the model. The $R^2$ value represents the proportion of variance explained by the model. It is a relative metric and applicable for comparing models trained on the same data.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}. \tag{6.3}$$

The sum of squared errors (SSE) is the total of the squared differences between the actual values and predicted values, and total sum of squares (SST), also written TSS, is the sum of the squared differences between the actual values and their mean. The ratio $\frac{SSE}{SST}$ is the proportion of total variation that cannot be explained by the model. The $R^2$ is thus interpreted as the proportion of variance of the response variable explained by the model. No explained variance would result in an $R^2$ value of 0 while a model in which all the variance is explained would have $R^2 = 1$. It is a positively oriented score; higher values are better. However, a value of 1 on the test set indicates a high probability that information is being leaked. It is also an indication of an overfitted model. A negative $R^2$ is possible and occurs in situations where the predictions made by the model fit the data worse than simply predicting the mean of the output variable. For a more accurate evaluation, the $R^2$ is adjusted to account for the addition of more features (predictor variables). The adjusted $R^2$ only increases when a newly added predictor variable improves the model performance more than would be expected by chance. This score is useful when the focus is on the most parsimonious model. It is more commonly used in statistical inference than in machine learning. Using the number of features $m$ and the number of observations $n$, the Adjusted $R^2$ can be calculated as

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - m - 1} \tag{6.4}$$

where $n - 1$ is the degrees of freedom. For the sake of predictive ability alone, $R^2$ is favoured, while Adjusted $R^2$ is useful for simpler, more parsimonious models in which causality is important.

### 6.3.2 Mean errors

The mean absolute error (MAE) and root mean squared error (RMSE) are two of the most often used metrics to determine accuracy for continuous variables, which are the type of variables in the obtained datasets. They differ in the way they are calculated, although both are absolute measures which share units with the dependent variable. The MAE is a measure of the average magnitude of errors, direction not considered, in a set of predictions. The MAE can be defined as the average over the test sample of the absolute differences between prediction and measured observation where all individual differences are weighted equally, calculated by

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|. \tag{6.5}$$

Without the absolute value taken, the signs of the errors are taken into consideration and the equation is known as the mean bias error. The MAE is conceptually easy for regression problems. It shows how far off the model predictions are, on average. MAE is slower to compute as an optimisation metric when used in training loops.

The RMSE however, is a quadratic scoring rule, measuring the average magnitude of the error. As the name implies, it is the square root of the average of the squared differences between prediction and actual observation, calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}. \tag{6.6}$$

While MAE and RMSE both express the average model prediction error in the units of the variable of interest, the dependent or target variable, and both are negatively oriented scores, they differ considerably. As RMSE squares the errors before averaging, RMSE weights the large errors highly. Consequently, RMSE should be used as a metric when huge errors are definitely unwanted. RMSE does not particularly increase with the variance of the errors but is enlarged with the variance of the frequency distribution of error magnitudes. There may be cases where the variance of the frequency distribution of error magnitudes needs attention but generally the variance of the errors is of more interest.

MAE is always equal to or smaller than RMSE, and RMSE tends to be larger than MAE as the test sample size increases. It is beneficial to consider RMSE when large errors must be penalised, while MAE is more useful for interpretation purposes as RMSE does not describe the average error in isolation.

Expressing the error in percentage is most commonly done with the mean absolute percentage error (MAPE). The calculation for this statistical measure is given by

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%. \tag{6.7}$$

It is useful when there are no extreme outliers and no zeros in the data.

The performance of a developed model fitted to the data in question, is relative. For that reason, it is useful to have a baseline model with baseline evaluation metrics against which to measure a developed model. The capability of such a model represents the lowest acceptable performance on the specific dataset. For regression, a central tendency measure can be used as the result for all predictions. An example of a baseline or dummy regression model predicts the mean value of the target variable. The RMSE can be used, and the $R^2$ of this model is zero. Any $R^2$ above zero is hence better than the baseline, or the ultimate lowest acceptable performance.

### 6.3.3 Cross-validation

Cross-validation is a useful statistical method commonly employed in applied machine learning to compare and select a predictive model. It is relatively simple and results in a less optimistic estimate of the model skill that generally has a lower bias than methods such as a train/test split. Cross-validation is a re-sampling procedure used to estimate the performance of a predictive model when predicting on unseen data. So, cross-validation uses a limited sample in order to estimate the expected general performance of the model when applied to make predictions on data not used during the model training.

In this section, the methods for calculating measures such as $R^2$ and errors are presented with the focus on cross-validation specifically, since it is useful in cases where the dataset is too small for a split into a training and test set.

The MSE and $R^2$ estimates can be calculated in three ways:

1. Resubstitution estimates: Using exactly the same data that was used for fitting the model to calculate the estimates of the target variable, leads to the resubstitution estimate of $R^2$. However, it has been known for decades that this method is overoptimistic about the ability of the fitted model to generalise to future observations.
2. Train/test split: The data is split into a training set and a test set, the first for fitting the model and the latter an entirely new set used to estimate $R^2$. The splitting of data in this manner is generally more useful in the case of a very large dataset.
3. Cross-validation (CV), particularly leave-one-out cross-validation (LOOCV). All the data is used for both fitting the model and assessing it. The estimate of $R^2$ which is obtained by LOOCV can also be denoted as $q^2$.

The first method calculating the resubstitution estimate is not considered because of its false optimism. The second, also called holdout CV, is not useful as the dataset used in this research is too small. When considering cross-validation, it must be taken into account that the procedure is computationally expensive to perform. The advantages are reliability and an unbiased estimate of the performance of the model. However, LOOCV is not appropriate when the dataset is very large, or the model is too computationally expensive to evaluate. With its high accuracy of model performance estimation, it is the right choice for evaluating models when a small dataset is used.

A procedure known as $k$-fold cross-validation (CV) is a specific type of cross-validation used for estimating the skill of a machine learning algorithm when it has to predict on data not used during the model training. The procedure has a hyperparameter, $k$, used to control the number of subsets into which the dataset is split. Each subset is used as a test set while the others are used for the training. Consequently, $k$-fold CV is done by fitting and evaluating $k$ models, where $k$ estimates of the model's performance on the particular dataset are provided. This performance can be reported with summary statistics including the mean and standard deviation.

LOOCV is a special case of CV, a configuration where $k$ is set to equal the number of examples or observations in the dataset. This version of $k$-fold CV where it is taken to its logical extreme, has the highest computational cost, as it requires a model to be fitted and evaluated for every single observation in the dataset. In the case of a small dataset with only a few observations, this is of course not as computationally expensive. This configuration has the advantage of providing a robust estimate of the performance of the model, as each row in the dataset is used to test the fitted model. The final score of a model is then calculated as the average of all $k$ estimates.

### 6.3.4 Akaike information criterion

The Akaike information criterion (AIC) is used for the evaluation and particularly, the comparison of models. The AIC is an estimator of relative quality of statistical models for a given set of data. This

means AIC scores are useful exclusively in comparison with other AIC scores for the same dataset. Burnham et al. (2011) define AIC as

$$\text{AIC} = 2m - 2\log(\hat{L}). \tag{6.8}$$

The term $-2\log(\hat{L})$ is known as the deviance.

$$\log(\hat{L}) = -\frac{n}{2}\log(\frac{\text{RSS}}{n}) \tag{6.9}$$

where RSS denotes the residual sum of squares from the fitted model (Burnham et al., 2011). AICc, the corrected AIC, which is adjusted for small sample sizes and is calculated as

$$\text{AICc} = \text{AIC} - \frac{2m(m+1)}{(n-m-1)} \tag{6.10}$$

with $m$ the number of parameters, $n$ the number of observations and $\hat{L}$ the maximum value of the likelihood function for the model.

## 6.4 Minimum required sample size

When a linear model is to be fitted to data, this data must be adequate and satisfactory. A sufficient sample size is also relevant, even when considering that linear models may require less data than, for instance, other machine learning models such as neural networks that require very large datasets. In an era of ever-increasing data quantities, to find or aggregate data for certain disciplines is still problematic. Big Data has been a focus in technology companies, with experts gaining experience in handling these datasets. However, small datasets also present challenges. They are sometimes more difficult to handle and require a different skill set. When dealing with small datasets, many challenges arise. Outliers are more influential, and the effect of noise is significant. According to Deeb (2015), small datasets can be handled best by using mathematical tools such as statistical tests, having a limited set of hypotheses, cleaning the data, performing feature selection and making use of regularisation.

Therefore, identifying a minimum sample size required for a study is a relevant problem in fields such as medicine, biology and engineering. In the case of this study, in the field of agriculture, the problem of insufficient data persists. In farming practices, detailed data capturing is often not as high a priority as in other fields. Although this has started to change, it was not predominant a decade ago. This generally leads to sample sizes being smaller in agricultural disciplines than in economics.

With the focus on regression, a problem arising with small samples is the higher likelihood of inconclusive or contradictory results. This is especially true in cases of considerable variation. A larger sample size is often suggested by researchers; however, a quantitative minimum sample size, $n$, is seldom advised. Studies have advised on the required quantity of observations relating to the number of predictors ($m$), for instance $n > 50 + m$ (Harris, 1985) or $n \sim 50m$ (Elazar J. Pedhazur, 1991) or $n > 50 + 8m$ (Green, 1991).

If a value is provided at all, the recommended minimum $n$ evidently varies. For decades, the effort to obtain clear and reliable guidelines for minimum $n$ has relied on inferential statistical calculation, where two hypotheses are compared, *i.e.* the possibility that a null hypothesis can be rejected, and a Type II error can be avoided. Power analyses are traditionally used in inferential statistics (such as t-tests and ANOVA) to determine what sample size will ensure a high probability that the null hypothesis will be correctly rejected.

Power describes the probability that the statistical test will find a statistically significant difference if such a difference does exist. Power of 0.8 or greater is widely taken as acceptable, implying that the sample size is sufficiently large and there is at least an 80% chance of finding a difference that is

statistically significant, in the case such a difference exists. In general, a larger sample size magnifies the power of the test. The size of the sample relates to the amount of information collected, in which case it is easier to reject the null hypothesis, consequently avoiding a Type II error. For a power calculation, with which a power value between 0 and 1 is calculated, the following elements are required:

- The type of test that is proposed (*e.g.* independent or paired t-test, ANOVA or regression)
- The significance level ($\alpha$) used (usually 0.01 or 0.05)
- The expected effect size
- The proposed sample size

If a difference is statistically significant, it does not imply that it can assist in decision-making. The effect size, which is a standardised measure, helps to determine if the observed difference that is statistically significant is also meaningful. In the case of an experiment, it is most commonly calculated as the difference between the two groups (for instance, the mean of the treatment group minus the mean of the control group) and dividing it by the standard deviation of one of the groups (Cohen, 1977).

Based on fundamental and operational reasons, the use of power analysis is not reliable (Jenkins and Quintana-Ascencio, 2020). Power analysis carries forward basic problems with null hypothesis inference. This has been the conventional foundation for statistical analyses, but it has also aptly received widespread criticism.

From an operational point of view, taking into consideration the four interdependent concepts: power, effect size, sample size and the level of significance $\alpha$ (Cohen, 1977), statistical power analysis presents challenges to estimate minimum $n$.

Estimating minimum sample size is solved using the following: a desired power level, effect size (which is slope in linear regressions or flexibility in economics) and significance level. Preliminary data can complement these assumptions, but the challenge is its availability. Also, the data is not always predictive. As a result, a challenge arises because a projected effect size develops into a goal of the research. The basis of model selection should rather be information theory metrics and parsimony. This is based on the principle of Occam's razor, when only the bare essentials are taken into consideration. The use of adjusted $R^2$ is then to critically evaluate the suitability of a certain model and applying the requirement as stated by Whitehead to aim for simplicity but to distrust or critically judge it.

As a solution to the above problems, statistical advances that use information theory allow for a different approach. Subsequently, an experiment by Jenkins and Quintana-Ascencio (2020) implementing this will be discussed to further examine the required minimum $n$. Jenkins and Quintana-Ascencio (2020) aimed to find a minimum sample size $n$ – the number of observations – required for accurate inference, to determine the shape of data made with null (random), simple linear, and quadratic regressions. They also evaluated the effect of variance on the minimum required $n$. For the purposes of the present study, however, only the linear case needs to be considered, as it is the only shape for which the data is tested. The expectation is that it is not a null model, and a quadratic shape would lead to even more possible features in a situation already of the type $m \gg n$, where $m$ is the number of parameters or features and $n$ the number of observations. Jenkins and Quintana-Ascencio (2020) simulated data on a selection of variances and effect sizes and solved regression models at a range of $n$ to determine a minimum $n$ where the data matches the regression model.

Among the range of a perfectly fitted model (with all the points on a line and $R^2 = 1$) and random scatter (where $R^2 = 0$) virtually unlimited options of combinations for the factors affecting power of regressions – variance, effect size and $n$ – are found. Approximate margins at low and high combinations of effect size and variance for a data shape (such as a straight-line pattern) are determined, after which regressions are recurrently evaluated with different $n$. The work of Jenkins and

Quintana-Ascencio (2020) is limited to first- and second-order polynomial linear models, two types of linear models which include additive combinations of coefficients of a predictor variable and constants. The simple linear model is also simply known as the linear model. In order to avoid confusion between the class and the models of the class, for this section the model will be referred to as the 'straight-line' model in the linear class. A second-order polynomial, also a linear model, is known as the quadratic equation

$$y = \alpha + \beta X + \gamma X^2 + \epsilon. \tag{6.11}$$

The results from these experiments should be applicable to multiple regressions (*i.e.* where covariates are included). The Akaike information criterion (AIC) is used to determine the most plausible model among the analysed set. AICc is the AIC value corrected for smaller sample sizes. AICc weight ($w_i$) is the proportion of the total predictive power of the full set of models being assessed (Bevans, 2020). A $w_i$ value is a criterion that scales from 0 to 1 indicating the probability that a model is the most likely. The AICc should not be used to suggest the model that fits the data best but rather suggests a model as the best regarding the trade-off between bias and variance of the fitted model parameters, for a particular $n$ (Burnham et al., 2011). Evaluating AICc values themselves, smaller values indicate a better model fit. AICc is used in cases with smaller $n$, and AICc values approach uncorrected AIC values at $n \sim 40$.

Null (random), straight-line and quadratic data were created with 50 observations per class, by prescribing a model and adding variance. Datasets were representative of scatter plots with either little or much added variation. For the analyses a range of samples of $n = 4$ (dictated by the minimum degrees of freedom for a quadratic model) up to $n = 50$ was sampled from a full dataset with $n = 50$. Each sample was evaluated for each of the three types of models, and they were compared by the weights ($w_i$) for AICc values.

Results indicate that $n \leq 7$ is not sufficient to compare quadratic to null and straight-line models, even if the standard deviation is very low. For a model with high $\sigma$, it can be concluded that a straight-line pattern would be most accurately detected with $n \geq 25$ (Jenkins and Quintana-Ascencio, 2020).

Attention to sample size has mostly focused on power analysis. However, Jenkins and Quintana-Ascencio (2020) approached the issue of sample size in a different manner by addressing the question of minimum required $n$ to accurately match a model to a data shape. The question is handled by model selection, where models are representative of alternative hypotheses. The answer is dependent on variance, but not on effect size or whether it is a straight-line or quadratic model. Jenkins and Quintana-Ascencio (2020) recommend a minimum $n = 8$ for cases with very low variance (*i.e.* a tight pattern). However, with high variance, this minimum is increased to $n \approx 25$ to clearly match a model to the data pattern. This recommendation is prudent, for observational studies that rely on regressions. Effect size was expected to affect the answer, a concept originating from power analysis, where the focus is on statistical significance of a slope coefficient. That is, however, not relevant to model selection based on AIC, where the answer is not reliant on statistical significance. Bolker (2008) recommends that models are first compared using AIC (or BIC), after which the goodness-of-fit for the selected model is determined with $R^2$ or adjusted $R^2$. The key recommendations from Jenkins and Quintana-Ascencio (2020) ($n \geq 8$ for cases with very little variance, but $n \geq 25$ for any more variance) are made under the assumption that samples are evenly spread in the dataset, and that regression assumptions are not defied.

Other independent variables (covariates) were not included in the models, other than the key independent variable. The definition of covariate is most precise in its use in Analysis of Covariance (ANCOVA). In ANCOVA, the independent variables of interest are categorical. Adjustment for the effect of an observed, continuous variable, the covariate, must be made. In this definition, the covariate is continuous, always observed, and is never the key independent variable. The term covariate is sometimes used as a synonym for any continuous predictor variable in the model. The ANCOVA definition also states that the covariate is a control variable, giving a third meaning as a categorical

control variable (Grace-Martin, 2020). Because of this confusion, it is mainly avoided throughout this document. Covariates can aid in describing variation in empirical data and produce more accurate coefficient estimates.

When only small $n$ is available, relatively weak (scattered) evidence might be salvaged with influential covariates, especially if variables are scaled so that varying units are standardised. Thoughtful planning and foreknowledge of the system under study could ensure that data of applicable and relevant covariates is acquired. These analyses by Jenkins and Quintana-Ascencio (2020) used fixed effects, but they can be informative for mixed-effects regressions often implemented in the natural sciences and medical research. A mixed-effects model is a statistical model consisting of a combination of fixed and random effects (Baltagi, 2008). A mixed-effects model defines the relationship between a response variable and other explanatory variables obtained where at least one of the explanatory variables is a categorical grouping variable representing an experimental unit (Magezi, 2015). Mixed-effects regressions with random effects would require $n \gg 25$ to clearly represent data patterns.

Research based on regressions or meta-regressions using $n \geq 25$ may improve reproducibility. Considering the study by (Jenkins and Quintana-Ascencio, 2020) it is clear that the data discussed in Chapter 5, when aggregated as yearly entries, is not sufficient for conclusive results. However, it may be worthy as input to models, aiding in gaining insight of significant role-playing factors, through feature selection.

## 6.5 Feature selection methods and dimensionality reduction

When developing a model to make predictions, it is necessary to identify the predictors, also known as input variables or features, to include, in a process known as feature selection. Simply put, feature selection chooses the best predictors for the target variable. Especially when the available features are many, such as in this study, dimensionality reduction can also be considered.

Both feature selection and dimensionality reduction are used for reducing the number of features in a dataset. However, the important difference is that feature selection is the process of selecting and excluding given features without modifying them, while dimensionality reduction is the transformation of features into a lower dimension and the creation of a projection of the data resulting in new input features. Feature selection is also similar to dimensionality reduction techniques in that both methods seek fewer input features to a predictive model. Therefore, feature selection can be seen as a type of dimensionality reduction or an alternative to it. Feature selection can be categorised in two ways. The first way is seen in the machine learning realm. Supervised feature selection methods can be categorised into either wrapper, filter or intrinsic methods. Feature selection methods are also commonly called variable selection methods in statistics. The second way of categorising them is as either test-based, penalty-based or screening based.

Feature selection can be categorised according to the first set of classes as follows:

1. Unsupervised: The target variable is not used. Redundant variables or variables with many missing values are removed. This method is not considered for this study as there are no features with missing values and more information can be gained from supervised methods also considering the target variable.
2. Supervised: The target variable is considered. Irrelevant variables are eliminated. The supervised feature selection methods are typically presented in three classes based on the combination of the selection algorithm and the model building.
    - Wrapper: Search for well-performing subsets of features, *e.g.* stepwise methods.
    - Filter: Select subsets of features by evaluating their relationship with the target. Implementation of these methods is much faster than wrappers, *e.g.* correlation.
    - Intrinsic or embedded: Algorithms perform automatic feature selection during model fitting, *e.g.* lasso, ridge, elastic net.

## 6.6 Problem with many input features

Too many input variables can impair the performance of machine learning algorithms. The number of features in the data can be considered to represent dimensions on a $m$-dimensional feature space and the observations of the points in that space. A large number of dimensions, especially when the points or observations represent a small sample, can dramatically impact the performance of machine learning algorithms.

It is therefore desirable to reduce the number of input features or reduce the dimensions. These two options are specifically distinguished from each other as feature selection is only one of the methods contained in dimensionality reduction.

Here it is important to note that linear regression is chosen for use on the datasets in this study. The least squares estimates will have low bias provided that the true relationship between the response and the predictors is approximately linear. If $n$ denotes the number of observations and $m$ is the number of independent variables (features), also referred to as predictors, the following three cases can occur (James et al., 2013):

- If $n \gg m$, the number of observations is much greater than the number of features, the least squares estimates tend to have low variance, and performs well on test observations.
- $n \geq m$, the number of observations is approximately equal to the number of features: a lot of variability can be present in the least squares fit, which results in overfitting and consequently poor predictions on future observations not used in the training of the model.
- If $m > n$, the number of observations is smaller than the number of variables, there is no longer a unique least squares coefficient estimate. The method cannot be used, as the variance is infinite.

The variance can often be substantially reduced at the cost of a negligible increase in bias through the constraining or shrinkage of the estimated coefficients. This can result in substantial improvements in the accuracy with which the response can be predicted for unseen observations (James et al., 2013).

Reducing the number of features improves accuracy. Less misleading data means modelling accuracy improves and it also reduces training time, as the algorithms train faster on less data. The problems resulting from high-dimensional data include overfitting. Overfitting occurs when the model has too many features or terms for the number of observations. Generally, at least 10 to 15 observations for each term are recommended in a linear model. Following this convention, in this study, there are consequently enough observations for a single feature.

Problems occurring when fitting a model, namely overfitting and its opposite, underfitting, will be briefly discussed.

Overfitting is a modelling error that occurs when the function is too closely fit to the limited set of training data. It usually takes the form of developing an overly complex model to explain idiosyncrasies found in the data under investigation. Consequences of overfitting are:

- Coefficients of the independent variables are unbiased and consistent.
- Coefficients are inefficient – coefficient variances and hence standard errors are estimated too large, therefore there is a risk to wrongly accept that coefficients are not significant when they are.

To avoid overfitting, the theory behind the relationship between the dependent and independent variables should be considered. Not all variables should be accepted as valid and included in the model. Therefore t- and F-tests should be done to compare different versions of the model.

The opposite of overfitting is underfitting. It occurs when a statistical model is unable to capture the underlying structure of the data. In an underfitted, oversimplified model, some parameters or

terms that would appear in a correctly specified model are missing. Underfitting leads to the following problems:

- Biased coefficients – in repeated applications the estimated coefficients will not coincide with the true values.
- Error variance is biased.
- Inconsistent coefficients – bias does not disappear for larger samples.
- Variance of estimated coefficients and hence standard errors have positive bias.
- t-tests are inaccurate or invalid, the null hypothesis is not rejected too easily.

To avoid underfitting, variables that are relevant according to theory should be included. The $R^2$ value, adjusted $R^2$, t-tests and signs of the coefficients should be considered, as well as the residuals.

Overfitting can generally be solved by:

- using a less complex model;
- using more training observations; and
- regularisation.

When overfitting a model, the regression coefficients become representative of the noise rather than the genuine relationships between variables (Frost, 2020).

An overfit model occurs when the regression line captures every single point in the graph.

The bias-variance trade-off is also of concern here. Bias is an expression of the difference between what is captured by the model and what the available data shows. A model with high bias does not closely match the dataset. On the other hand, a model with low bias matches the dataset to a high degree. Variance is either because of sensitivity or the result of small fluctuations in the data. Typically models with high bias have low variance, and models with high variance have low bias. In a model with high variance random noise in the data are captured rather than the intended outputs. The ideal is finding a line with low bias and low variance.

A line capturing every data point is, in fact, not ideal. It captures the abnormal nuances of the small sample of data well; however, it may not necessarily perform equally well on unseen, out-of-sample data. These nuances of the sample data are the outliers and distinctive characteristics of the sample data, not necessarily possessed by unseen data. One of the options to discourage overfitting or simply reduce it, is regularisation. Section 6.4 touched on this where the concept of small datasets and required sample size were discussed. Methods commonly used to manipulate datasets with too many variables or features to avoid overfitting are regularisation (shrinkage), dimension reduction and subset selection.

## 6.7 Regression methods used in the study

This section introduces the various feature selection methods considered. In some of these, the development of the model is also done while the features are selected. In these cases, a linear regression model is used, as crop yield prediction is a regression problem, and a linear model is most prudent considering a small sample size. This section is thus written in the context of linear regression.

The question of feature subset selection is concerned with finding a subset of the original features of a dataset, in order to run an induction algorithm on data containing only the chosen features that will ultimately generate a predictive model with the highest possible accuracy. It is vital to choose a subset of the most relevant features (Hand et al., 2001).

Literature on the matter as to which features to employ in a statistical model initially paid attention to stepwise regression in 1967 (Breaux, 1967) and autometrics (Hendry and Richard, 1987). From there on more advanced procedures were developed, such as the non-negative garrote (Breiman, 1995), the LASSO developed by Tibshirani (1996) and the sure independence screening (Fan and Lv, 2008). These methods, together with other statistical methods put these algorithms into three classifications (Desboulets, 2018), namely:

1. Test-based: This group is categorised on statistical tests to choose between candidate features.
2. Penalty-based: Applying a limit on parameters inside assessments causing sparsity among them.
3. Screening-based: Ranking features by importance.

The following sections present the four methods used in this study, namely forward stepwise regression, elastic net regression, a correlation filter and partial least squares regression, and discuss in which categories they are classified as well as their characteristics.

### 6.7.1 Stepwise regression

Stepwise regression is the first method in the category of test-based algorithms and it falls under the category of wrappers, adding significant variables (forward stepwise) or retrieving insignificant variables (in the backward approach) according to a defined statistical criterion. Implementation is straightforward, although in some situations consistent selection is not ensured.

Stepwise regression (Breaux, 1967) is one of the oldest methods used for model selection. Among $m$ variables, $2^m$ models can possibly be constructed, as a result they should possibly all be considered. To overcome this computational challenge of testing all possibilities (which is done in another algorithm, best subset selection), stepwise regression saves computational power by investigating only a subset of all the possible regressions in search of the true model. The forward approach entails the process starting with a null model containing only the intercept and gradually adding variables (step by step). The backward approach starts with a full model containing all the variables and removing them gradually. Both approaches can be considered. The selection within a single step is based upon some criteria. All one-variable increments are considered, and decisions are made according to these conditions. The criterion is usually, among other options of measures, the lowest AIC, AICc, BIC, Mallow's $C_p$, highest $R^2$ or adjusted $R^2$, lowest prediction error, lowest $p$-value or leave-one-out cross-validation. There are some criticisms of stepwise regression, mainly about the lack of search. Biased estimation and inconsistent selection are some of the concerns, because this method proceeds along a single path without backtesting. The exception is the forward-backward stepwise, but only a single previous step is considered.

In the case where many independent variables may play a role in the behaviour of the response variable, stepwise regression is used to select important variables in order to obtain a simpler model.

Forward stepwise regression consists of the following steps:

1. Begin with the Null Model $M_0$ that contains no variables, only the intercept, if chosen to be included, $y = \beta_0$.
2. Start adding the most significant variables one after another.
3. A pre-specified stopping rule is reached, or all the variables have been included in the model.

The most significant variable to be added next to the model is determined in a few possible ways. One is to determine the independent variable that leads to the largest increase in $R^2$. Another is the variable that provides the smallest Residual Sum of Squares compared to other predictors considered at that point. A stopping rule often implemented is to stop when the number of included features reaches $n - 1$, as the aim of the algorithm is not to create a model with more features than samples

included. Another stop rule is satisfied when all remaining variables to consider have a $p$-value larger than a threshold if they were to be added to the model. This threshold is usually fixed, for instance 0.05 or 0.2, but can also be determined by the AIC or Bayesian Information Criterion (BIC). AIC determines the threshold by considering the degrees of freedom of the variable, while BIC finds the threshold according to the effective sample size. BIC is recommended for large sample sizes, usually exceeding 100 observations per independent variable.

The created models (with varying numbers of included features) are compared by the following most popular criteria:

1. Mallow's $C_p$
2. AIC
3. BIC
4. Adjusted $R^2$

Mallow's $C_p$ is defined as

$$C_p = \frac{1}{n}(RSS + 2m\hat{\sigma}^2) \tag{6.12}$$

where $\hat{\sigma}^2$ is an estimate of the variance of the error associated with each response observation. $\hat{\sigma}^2$ is typically estimated using the full model with all predictors. The $C_p$ statistic adds a penalty of $2m\hat{\sigma}^2$ to the training RSS to adjust for the fact that the training error tends to be overly optimistic. The penalty increases as the number of predictors included in the model increases, which adjusts for the decrease in training RSS. The model with the smallest Mallow's $C_p$ is considered most desirable.

AIC is defined for a large class of models fit by maximum likelihood. AIC is defined in Subsection 6.3.4, but in this context can also be calculated as

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2m\hat{\sigma}^2) \tag{6.13}$$

while for a linear model BIC is defined as

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + log(n)m\hat{\sigma}^2) \tag{6.14}$$

where $m$ represents the number of predictors.

The adjusted $R^2$ is, as its name suggests, adjusted to account for the fact that the $R^2$ always increases with more variables. The inclusion of unnecessary features in the model is penalised with the adjusted $R^2$:

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-m-1)}{TSS/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-m-1} \tag{6.15}$$

where $n$ is the total sample size, and $m$ refers to the number of included predictors. The first three criteria have rigorous theoretical justification relying on asymptotic arguments, *i.e.* when the sample size grows very large, whereas the adjusted $R^2$, although intuitive, is not as often used in statistical theory.

Stepwise regression models have a few parameters that need consideration. One of the most important of these is the $p$-value of each included variable. The $p$-value should be 0.05 or below to be significant at 95%. Other types of stepwise regression also exist, such as backward selection. Forward stepwise regression, however, has advantages especially when the number of variables to consider is greater than the sample size. It does not need to consider all the possible predictors, as it only considers models with a number of variables less than the sample size. Backward stepwise regression is advantageous when the number of candidate variables is smaller than the sample size, because it

considers the effects of all the variables. Limitations of stepwise regression include that the tests are biased. The fit may appear better than it really is. The models may also be oversimplifications of the real models of the data. These limitations can be addressed by verifying the resulting model, such as with cross-validation. Stepwise model selection has been controversial. The main problems with stepwise methods are summarised by Harrel (2001) as:

1. $R^2$ values are biased on the high side.
2. The F-statistics do not have the claimed distribution.
3. The standard errors of the parameter estimates are too small.
4. Because of this, the confidence intervals around the parameter estimates are too small.
5. $p$-values are too small, because of multiple comparisons.
6. Parameter estimates are biased away from zero.
7. Collinearity problems are magnified.

In the case of few observations, as in this study, it remains a method to evaluate, while considering domain knowledge.

The application of stepwise regression as a screening method leads to progressed results. In stepwise regression and other testing methods such as autometrics (Hendry and Richard, 1987), selection and estimation are implemented consecutively, while penalty-based algorithms perform both simultaneously. Penalty-based procedures are then when tests are directly implemented inside inference.

### 6.7.2 Elastic net regression

The overfitting problem presented by high-dimensional datasets can also be stemmed by regularisation, specifically elastic net, which falls under the category of embedded feature selection methods. Elastic net is the combination of lasso and ridge regression and all three are penalty-based methods.

Penalty-based methods involve applying a penalty on the estimated parameters, or implementing an altered loss function, encouraging sparsity (so that the values of some parameters shrink to zero). In high-dimensional cases, namely where $m > n$, sparsity is vital. This can be handled, and inference made possible, with the application of penalties on parameters. Generated sparse models can be employed to integrate a test inside the estimation. Inference of such models is based on the hypothesis that certain variables are not relevant. Penalty-based methods assist in coefficient estimation. By merging both testing and inference procedures into an integrated structure a distinctive invention of statistical modelling is brought about, hence it is an embedded method. The predominant penalty-based algorithms in this category are the ridge (Marquardt and Snee, 1975) and lasso or LASSO (short for Least Absolute Shrinkage And Selection Operator) of Tibshirani (1996), the two most common types of regularisation.

Regularisation refers to the process of introducing additional information in order to prevent overfitting. This information is generally in the form of a penalty for complexity. Regularisation favours simpler models over more complex models. Regularised linear regression addresses concerns such as variance-bias trade-off, multicollinearity, sparse data handling (*i.e.* in cases of $m > n$), and feature selection. Shrinkage methods, also considered penalties of complexity, are modern techniques in which shrinkage is applied to the regression coefficients. The model is fitted with all $m$ predictor variables, while regularising the coefficient estimates towards zero relative to the least squares estimates. It uses a penalty to penalise for large coefficients, or a large number of coefficients and will typically shrink the coefficient values towards zero. Two categories exist under penalty-based methods: penalties on the norm and concave penalties. Although norm penalties are typical and commonly used, concave penalties also have certain applications and include the non-negative garrote of Breiman (1995), SCAD (Smoothly Clipped Absolute Deviation) and MCP (Minimax Concave Penalty).

The two norm penalties applicable to both lasso and ridge are subsequently discussed. For norm penalties the general goal is to resolve, in penalised form, the following equations, each of the penalty-based methods applied to $L_\gamma$ norms, $L_1$ and $L_2$ respectively:

$$\min_{\beta \in \mathbb{R}^m} \frac{1}{2}||\mathbf{y} - \beta\mathbf{X}||_2^2 + \lambda||\beta||_1 \text{ (lasso regression)} \tag{6.16}$$

$$\min_{\beta \in \mathbb{R}^m} \frac{1}{2}||\mathbf{y} - \beta\mathbf{X}||_2^2 + \lambda||\beta||_2^2 \text{ (ridge regression)} \tag{6.17}$$

where $\lambda \geq 0$ is the tuning parameter and $m$ is the dimension or number of parameters.

In the case of one input variable the linear relationship is a line but for higher dimensions this association can be applied as a hyperplane connecting the predictor variables to the target variable. An optimisation process with the aim of minimising of the total of the squared error between the predictions ($\hat{y} = \sum_{j=1}^{m} \hat{\beta}_j X_{ij}$) and the target values ($y$) is employed to find the coefficients of the model, using

$$\text{loss} = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \tag{6.18}$$

A disadvantage of linear regression is that estimated values of the coefficients of the model grow very large, increasing the sensitivity of the model to inputs. This is mainly relevant to scenarios with limited observations, or $m \gg n$ problems. The $L_1$ penalty, which can be included in the cost function for linear regression to be called lasso regularisation, minimises the size of all coefficients. During this penalty a model is penalised based on the total absolute coefficient values. The lasso is the solution to the optimisation problem

$$\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|. \tag{6.19}$$

Applying 6.19, lasso assigns a penalty to the coefficients in the linear model. The hyperparameter $\lambda$ is used to control the amount of weight allocated to the penalty. A default $\lambda$ value of 1.0 will grant full potential of the penalty while the penalty is excluded when $\lambda$ is set to zero. Insignificant values of $\lambda$, such as $1e-3$ or smaller, are commonly used. A result of penalising the absolute values is that for some value of $\lambda$, some parameters are set to zero. With the value of a number of coefficients set to zero, feature elimination is effectively performed and consequently, lasso produces models that use regularisation in order to improve the model as well as to perform feature selection (Max Kuhn, 2013). The solution of ridge regression (Hoerl and Kennard, 1970) is analytical, but it encourages a dense model as it does not shrink coefficient values ($\beta$s) to exactly zero and consequently coefficients are not eliminated in the model. The $L_1$ norm of lasso (Tibshirani, 1996), however, is singular at the origin, resulting in a sparse solution.

When sample sizes are relatively small, ridge regression can be implemented to improve predictions made from new data and reduce variance by decreasing sensitivity to training data. Ridge allocates a penalty, the squared magnitude of the coefficients multiplied by $\lambda$, to the loss function. Ridge, just as lasso, puts a penalty on coefficients that are overemphasised by the model. The value of the lambda parameter plays an important role in the amount of weight assigned to the penalty for the coefficients. The ridge formula incorporates the sum of the error with the sum of the squares of coefficients. Ridge regression is the solution to the optimisation problem

$$\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{m} (\hat{\beta}_j)^2. \tag{6.20}$$

Elastic net, the combination of the characteristics of both lasso and ridge regression, decreases the influence of some features but it does not remove all the features. The formula combines the lasso and ridge formulas, implementing both penalties to minimise the objective function

$$\frac{1}{(2n)}||y - X\beta||_2^2 + \alpha \times l1\_ratio \times ||\beta||_1 + \frac{1}{2}\alpha \times (1 - l1\_ratio) \times ||\beta||_2^2 \tag{6.21}$$

The L1 and L2 penalty are equivalent to $a \times L1 + b \times L2$ where $\alpha = a + b$ and $\text{l1\_ratio} = \frac{a}{(a+b)}$.

The regularisation methods lasso, ridge and elastic net improve the performance of the linear model. Lasso will remove a number of features and minimise overfitting in the linear model. Ridge lowers the effect of the features that are not vital in predicting the target variable. By merging feature elimination from the lasso model and feature coefficient reduction from ridge, elastic net improves the predictions of the linear model and lowers the incidence of overfitting.

### 6.7.3 Correlation feature selection

Filter or screening methods, often relying on correlation, comprise another category of feature selection methods. Screening relies on a relationship between the dependent variable and independent variables, or regressors. These methods handle excessive dimensional features and have computational complexity ($log(m) = O(n^{\alpha})$ with $0 \leq \alpha \leq 1$, dimension $m$ and sample size $n$). Screening is the most computationally efficient compared with other selection methods. As a rule, screening does not perform model selection, but instead it simply ranks the variables according to significance. It has to be combined with other procedures, which are theoretically mainly penalty-based, to be employed for model selection. An essential step in variable selection is simply to decrease the number of features in the original set of data and screening methods are highly effective tools for this. Screening implements a ranking assessment, which may or may not be linear, allowing it to be employed in both frameworks. Some are model-free while others are determined by specific models (such as a linear model). The techniques in this group mainly differ according to the ranking measure they apply. They generally use correlation coefficients. The first method, mentioned above, and from which nearly all of the others are derived, is Sure Independence Screening (SIS) (Fan and Lv, 2008). SIS makes use of simple correlation on standardised variables, $\hat{\omega}(x_j, y) = \tilde{\mathbf{x}}_\mathbf{j}\tilde{\mathbf{y}}$, producing a ranking of the $\mathbf{x_j}$. A threshold is then used to choose the top-ranked features.

Similar to the screening-based method SIS, other correlation-based methods exist. Weston et al. (2003) illustrates the use of correlation criteria in micro-array data analysis. Guyon and Elisseeff (2003) discuss variable and feature selection, distinguishing between variables and features by referring to variables as the raw input variables and features as the constructed variables. They advise on feature selection methods, specifically that in case features need to be assessed individually to gather an understanding of their influence or whether they are too many – both of which are applicable to this study – a variable ranking method should be used. Variable ranking is often used by variable selection algorithms as a principal selection mechanism thanks to its simplicity and empirical success.

Correlation-based methods are classified as filter methods. Correlation-based filters are often implemented for classification tasks but are also useful for ranking continuous variables in the case of a continuous outcome. Two types of filter methods, namely univariate and multivariate, exist. Univariate filter methods are used to evaluate and rank a single feature according to certain criteria, treating each feature individually as well as independently of the feature space. Features are ranked according to certain criteria, according to which the highest-ranking features are selected. A possible drawback of univariate methods is that a redundant variable may be selected, as the relationships between features are not considered. Multivariate filter methods evaluate the entire feature space. They consider features in relation to others in the dataset. These methods can handle redundant, duplicated and correlated features, but there is added complexity. Generally, correlation-based feature selection eliminates features highly correlated with others and chooses features highly correlated with the output or target. For the purpose of this study, the greater interest is in the correlation of a

feature with the target, therefore its relevance to the output rather than its redundancy. In order to find a simple method with low computational complexity as an alternative feature selection method, univariate filter methods are investigated. Correlation filter methods are univariate. Correlation, as discussed in Chapter 4, is a measure of how two variables change together. A high correlation is often useful for determining if a variable can be used to predict another. Therefore, correlation filter methods generally look for features that are highly correlated with the target, especially for linear machine learning models. Several methods can be used to measure the correlation between variables. Max Kuhn (2013) states that the classic approach to quantifying the relationship of numeric predictors with the outcome is the utilisation of the sample correlation statistic. In the case of continuous features and a continuous target, the Pearson correlation coefficient $\rho$ is most often used. For samples, the correlation coefficient is denoted by $r$ while the correlation coefficient when applied to populations is represented by $\rho$. The Pearson correlation coefficient $r$ was shown in Equation 4.1. The Pearson correlation coefficient assumes a Gaussian distribution to each variable; that there exists a straight-line relationship between the two variables; and the data is equally distributed around the regression line. The fact that the Pearson's correlation only detects linear dependencies between the variable and the target is, however, not a drawback in this study as only linear relationships are considered. For the purpose of this study, the correlation coefficients will simply be used with a user-defined threshold to determine features to include in the linear regression model, using the correlation ranking method as a means to reduce the high dimensional problem and solve the 'short, fat data problem' (Verma et al., 2018). The Scikit-learn Python module (Pedregosa et al., 2011) has an implementation of this to be used for validation. It implements the correlation statistic in the f_regression() function. This function can be used for feature selection, for instance by choosing the $K$ most relevant features via the SelectKBest class. The SelectKBest class using the f_regression function consists of two steps: First, the correlation between each regressor and the target is computed. The correlation coefficient is then converted to an $F$ score and a corresponding $p$-value. The features with the top $K$ scores are chosen, where $K$ is a hyperparameter that can be tuned by the user.

### 6.7.4 Dimensionality reduction

Dimensionality reduction methods, which include partial least squares (PLS) and principal component analysis (PCA), cannot be grouped with the other categories of feature selection methods because of the transformation of the data onto new axes rather than a selection of a subset of features. PCA is an unsupervised method and therefore may disregard important variables and it is too restrictive as most implementations require the number of features $m$ to be less than the number of observations $n$, or only use the first $n$ features.

PLS is similar to PCA in the way it finds a linear transformation and gets rid of multicollinearity. However, PLS takes the variability of the target or dependent variable into consideration, which is its advantage over PCA. For this reason, PLS is implemented in this study. The PLS method, also known as 'projection to latent structure', is used for constructing predictive models in the case of many and highly co-linear factors. PLS, developed by Herman Wold in the 1960s for econometrics, is mainly used by chemical engineers and chemometricians. PLS is particularly suited when the matrix of predictors $\mathbf{X}$ is of the type $m > n$ with more variables than observations, and when there is multicollinearity among the features. In these cases, standard multiple linear regression would fail unless it is regularised. The PLS regression procedure combines aspects of PCA and multiple regression, by first extracting a set of latent factors that explains as much as possible of the covariance between the predictors and targets, as a result projecting into a lower-dimensional subspace. The second step is a regression to predict values of the targets with the use of the decomposition of the predictors (Turek et al., 2020). In the regression form of PCA, principal component regression (PCR), the set of independent variables $\mathbf{X}$ is transformed to $\mathbf{X}' = W\mathbf{X}$. $W$ is a linear transformation, resulting in a new set that is linearly independent. $\mathbf{X}'$ is the factor scores. But as mentioned, in this method only the independent variables are taken into consideration, not the dependent variable. PLS regression seeks transformations of the original data into a new set of uncorrelated variables and constructs a set of linear combinations of

the inputs for regression, using $y$ in addition to $\mathbf{X}$ for this construction. The basic objective of PLS is to project the data in a latent variable space in such a way that it maximises the covariance between feature space $X'$ and response $y$. PLS projects both $\mathbf{X}$ and $y$ into a lower-dimensional subspace such that the covariance between $\mathbf{X}'$ and $y'$ is maximal.

In the context of the three categories filter, wrapper and embedded methods, partial least squares regression is described as follows (Mehmood et al., 2012):

- Filter methods: The purpose of these methods is variable identification. Filter methods make use of the output from the PLS regression (or PLSR) algorithm, identifying a subset of variables of importance.
- Wrapper methods are based on iterating between model fitting and variable selection. Filter methods are used to identify the variables which are piped back into refitting the PLSR, reducing the models.
- Embedded methods: In these methods the variable selection occurs at component level. They are a combination of variable selection and modelling in a single step. On each component of the PLS, the method searches for an optimal subset of variables. The variable selection is therefore nested, or embedded, within the PLS algorithm. These methods are based on a single iterative procedure, while wrapper methods are based on a double iterative procedure. Therefore, the embedded methods are generally faster to implement in comparison with wrapper methods.

The first type, filter methods, are used to implement PLS in this research. The filter methods carry out feature selection in two steps. The PLS regression model is fitted to the data, after which the features are selected by introducing a threshold on some measure of relevancy obtained from the fitted PLS model. These procedures are usually effective, but they give no indication with regard to the prediction relevancy of the selected variables. A threshold on the filter measure is necessary to classify variables as selected or not, hence the selection is particularly influenced by the selected threshold. Examples of filter measures in PLS are PLS regression coefficients $\beta$, loading weight vectors $w_a$ and variable importance in projection. Variable importance in projection is a filter measure used to select variables is the variable importance in PLS projections, which was introduced as 'Variable influence on projection', termed VIP by Eriksson et al. (2006). Examples of implementations of this method are in wavelength selection (Gosselin et al., 2010) and the search of biologically relevant QSAR descriptors (Olah et al., 2004).

The $v_j$ weights are measures of the contribution of each variable according to the variance explained by each PLS component. Now $v_j$ can be expressed as

$$v_j = \sqrt{m \sum_{a=1}^{A} [(\mathbf{q_a^2 t'_a t_a})(w_{aj}/||\mathbf{w_a}||)^2] / \sum_{a=1}^{A} (\mathbf{q_a^2 t'_a t_a})} \tag{6.22}$$

where $t_a$ is the $a^{th}$ column vector of score matrix $\mathbf{T}$ and
$q_a$ is the $a^{th}$ element of regression coefficient vector $q$ of $\mathbf{T}$.
$w_a$ is the $a^{th}$ column vector of weighting matrix $\mathbf{W}$, which gives the weighted variability of $j^{th}$ variable in the retained dimensions.
$m$ is the number of variables in regressor matrix $\mathbf{X}$.
The VIP score calculates the contribution of each variable according to variance explained by each PLS component. The expression $w_{ja}/||wa||$ represents the importance of the $j^{th}$ variable in the $a^{th}$ PLS component. The $q_a^2 t'_a t_a$ is the variance of y explained by the $a^{th}$ PLS component. And the summation of $q_a^2 t'_a t_a$, the denominator term, is the total variance explained by the PLS model with $A$ components.

Variable $j$ can be eliminated if $v_j < u$ for some user-defined threshold $u \in [0, \infty)$. Generally it is accepted that a variable should be included in the selection if $v_j > 1$ (Gosselin et al., 2010).

## 6.8 Synthesis of theory on feature selection methods

This chapter detailed the theory of linear models and described prediction in the context of regression. The rationale for linear regression for the data in this study was also discussed, namely, to ensure simplicity of interpretation and considering what the size of the dataset allows. Evaluation metrics for assessing model performance and comparing models were presented, as well as the preferred metrics for use in this study. Following general predictive modelling, specific algorithms for selecting features were discussed. The methods were chosen considering one of the main objectives of the study, namely identification of important yield-influencing factors. For their abilities to handle regression problems (with continuous data), $m > n$ datasets, and especially small datasets, the following four methods were identified for implementation in this study:

1. A screening method making use of the correlation coefficient which is validated with SelectKBest from Scikit-learn.
2. A subset selection method namely forward stepwise regression with the AICc criterion.
3. A regularisation method combining
   (a) lasso and
   (b) ridge regression

   named elastic net regression is chosen for its regularisation properties.
4. Finally, partial least squares regression is chosen for its dimension reduction characteristics, and ability to transform the entire dataset.

These four methods represent various ideologies of feature selection and regression methods, but share common properties enabling them to perform feature selection and prediction on the data discussed in Chapter 5.

In the following chapter, the results of the combination of the weather data with the harvest data and the methods will be presented. The focus will shift to the bunches data, for corroborated findings on what influences the number of bunches on a palm. Finally, the bunches data will be concatenated to the weather data to form a new set of possible features, predicting the yield of both individual orchards and the farm as a whole.

# Chapter 7

# Implementation of feature selection methods

The previous chapter discussed the theory of linear model development and considered methods useful for choosing features and developing regression models. In this chapter the feature selection and model development are presented. Each of the chosen feature selection methods is implemented, first on orchard level and then on farm level for the entire yield. The prediction model is then developed and evaluated. The implementation of all feature selection methods as well as the results of the models developed by fitting these features will be discussed in this chapter. The features evaluated for selection comprise the weather dataset constructed from the data in Chapter 5. Following the weather data, the bunches data for each orchard is added for the respective evaluations per orchard. The investigation of the bunch data is continued by consulting the map to consider orchard and tree positions as an explanation for bunch quantity.

## 7.1 Feature selection on constructed weather data to predict yield

The regression models investigated in Chapter 6 are implemented to identify the most significant features. They are first applied to each of the 33 individual orchards which have data available from 2010. In these models the independent variable $\mathbf{X}$ refers to the features in the dataset and $y$ is the yield per tree of the particular orchard. Subsequently, the feature selection methods are applied to the entire farm yield, which refers to all the orchards on the farm, not only to the 33 orchards.

All the possible features in the dataset, manipulated and derived from the weather data, are named and abbreviated as set out below. $\mathbf{X}$ comprises the following weather factors: heat units, mean humidity, mean temperature, maximum temperature and rainfall.

In the feature names:

- Month names are abbreviated, *e.g.* January is written as Jan.
- The mean humidity is abbreviated simply as Hum.
- Maximum, minimum are written as Max, Min.
- The weather data for two years prior to harvest is considered – weather of the season before the current season is denoted by prev.
- The heat units for the month are abbreviated HU.

In the equations:

- $i$ represents the index of the month of the year, $i \in 1, 12$
- Maximum Temperature in month $i$: $\mathbf{MaxT}_i$
- Mean Temperature in month $i$: $\mathbf{MeanT}_i$
- Minimum Temperature in month $i$: $\mathbf{MinT}_i$
- Sum of the calculated heat units of month $i$: $\mathbf{HU}_i$
- Sum of the rainfall in month $i$: $\mathbf{Rain}_i$
- Mean humidity in month $i$: $\mathbf{Hum}_i$
- Mean of the daily maximum daily wind speed in month $i$: $\mathbf{Wind}_i$

- Weather from the previous season has a prefix: **prev**

Rainfall figures for the previous season were not taken into consideration as irrigation would outweigh their effect completely and the negative effects of rain (washing away pollen) would not be applicable.

Referring to Section 6.4, it is evident that the size of the dataset for use in this study does not justify splitting it into a training and test set as both would be too small for model training and testing purposes. For the feature selection methods implemented in this chapter, all the observations were used to train the models.

The tables display the equations of the fitted regression models as well as the assessment from LOOCV, for this purpose the LOOCV RMSE (written RMSE) is chosen. The RMSE, in the same units as the target, is the square root of the MSE. Lower values are preferable, indicating a smaller error in prediction.

### 7.1.1 Correlation-based method and SelectKBest on orchard-level yield

In this section, regression on the individual harvests of the orchards and weather data will be discussed. As discussed in Section 6.5, correlation by itself is not typically used as a feature selection method, but it provides valuable insight, as suggested by the section on screening methods. For the purpose of investigating the relationships between the constructed weather features and the yield, correlation is a purposeful starting point. Correlation coefficient calculations and linear regression were used on the weather features and the yield of the farm as a whole, not considering specific orchards as individual harvests. Presently, the Pearson correlation coefficient is calculated between the harvests per palm tree of all 33 orchards for which data is available for all the years investigated, and the weather features, derived from processing the daily weather measurements to monthly features. The coefficients were used to determine positive and negative effects of the weather factors; rain, temperature, sum of heat units and humidity in the different months of the year. In order to discover these role-playing features from the data, a correlation coefficient threshold is arbitrarily chosen. This threshold is specified in a range, with 0.7 and -0.7 distinguished from the others for the results they produce. A lower threshold (0.6 and -0.6) results in too many features, some of which may be irrelevant while a higher threshold is too restrictive, possibly disregarding valuable factors.

For the sake of simplicity and clarity, the Pearson correlation coefficient is used outside an existing implementation such as in the machine learning library Scikit-learn, mentioned in Section 6.5. A scatter plot of the identified features and corresponding harvests of the relevant orchard is then used to plot trend lines with the use of linear regression. Since simple linear regression is used for each explanatory variable, scaling and normalising are not required. The goodness of fit of the line is expressed with an $R^2$ value. The $R^2$ (equivalent to $r^2$ in the case of simple linear regression) is the coefficient of determination, the proportion of the variance in the dependent variable that is predictable from the independent variable.

For the regression lines, an $R^2$ value above 0.5 indicates a good fit, suggesting a strong influence on the harvest by the given weather feature. Therefore, for each orchard individually, the equation of the regression line is calculated together with its $R^2$ value and the mean squared error loss. By plugging the weather values into the equation of these regression lines, this equation can be used to calculate the harvest per tree of the corresponding orchard for the harvest of the following year. A relationship found by regression does not imply causation and a change in the independent variable does not necessarily lead to or cause the change in the dependent variable. However, the influence of some of these features is, as discussed in the literature analysis, supported by agricultural research and science. The most prominent feature, with high positive correlation coefficient values between the feature values and the orchard harvests, and high $R^2$ values when plotted against each other, is the maximum temperature in August. The most prominent feature with a negative influence on yield, with high negative correlation coefficient values and a high $R^2$-value when plotted against the relevant

harvests, is the maximum temperature in May.

As discussed in the literature analysis, heatwaves, especially at specific times of the year, have a detrimental effect on fruit drop, or abscission. In agricultural practice, provision is made for this by thinning or pruning less.

In the tables displaying selected features, found in the appendix, the adjusted $R^2$ value is also displayed in parentheses with the $R^2$, as a higher number of independent variables inaccurately increases the $R^2$ value. The adjusted value is a more modest depiction of the coefficient of determination. The adjusted $R^2$ addresses two problems that arise with the use of $R^2$. The first is that the $R^2$ value increases at random with every added predictor; consequently, a model with more terms appears to fit better although it is simply caused by an increase in the number of predictor variables. The second challenge is that a model with too many predictors may model random noise in the data, causing overfitting, which decreases the accuracy of the model.

The Python project Scikit-learn (Pedregosa et al., 2011) provides a function SelectKBest for feature selection that selects features according to the $K$ highest scores. Using its f_regression as the score function, the individual effect of many regressors is tested. The procedure involves finding the correlation between each feature $i$ and the target variable as

$$\frac{(X_i - \overline{X_i} \times (y - \bar{y}))}{(\sigma(X_i) \times \sigma(y))}. \tag{7.1}$$

The correlation is converted to an F-score and its associated $p$-value. This method is very similar to computing the correlation and setting a threshold. However, the number of chosen features is not determined by the threshold value of the correlation coefficient, but by a user-specified $K$. This method is implemented for two main reasons. First, it is used as validation for the correlation-based method. Secondly, specifying to choose only one feature per orchard provides the opportunity to find the most relevant features. The $p$-values resulting from this method are exactly the same as those obtained from finding the Pearson correlations and their associated $p$-values.

To avoid overfitting, since the dataset is so small, regression lines with a resubstitution $R^2$-value of 1.00, for which the adjusted $R^2$ is not calculable since too many independent variables were included, were discarded. In these cases where the number of features $m$ is greater than the number of observations $n$, $m$ was reduced by considering features with the highest correlation coefficient with the yield. These features are used as regressors to form a new regression model.

The $p$-values of the regression coefficients indicate the probability of getting the same results if there was in fact no relationship between the predictor and response. A $p$-value larger than 0.05 is not statistically significant and the null hypothesis of no association between the response and predictor, after adjusting for the other explanatory variables in the model, cannot be rejected. If zero is included in the 95% confidence interval for the parameter estimate, the possibility cannot be ruled out with 95% confidence that the association between the two variables is zero.

Correlations for features with correlation coefficients $r > 0.6$ with four or more orchards planted before 2000 are displayed in Table 7.1. In this table as well as the rest of the chapter, the **Oc** column shows the orchard numbers.

Table 7.1: Correlation coefficient values for highly correlated weather features in older orchards

| Oc | prev May Humid | prev Jun Humid | Aug HU | Apr Mean Temp | Aug Mean Temp | Aug Max Temp | Jul Rain |
|---|---|---|---|---|---|---|---|
| 17 | 0.450 | 0.568 | 0.566 | 0.511 | 0.573 | **0.728** | 0.437 |
| 90 | 0.177 | -0.094 | 0.345 | **0.762** | 0.323 | 0.363 | 0.435 |
| 56 | **0.868** | **0.657** | 0.486 | 0.079 | 0.459 | **0.604** | 0.577 |
| 57 | 0.479 | 0.487 | **0.714** | 0.551 | **0.700** | **0.779** | 0.490 |
| 58 | 0.445 | 0.281 | **0.741** | **0.634** | **0.715** | **0.809** | 0.484 |
| 61 | 0.429 | 0.394 | **0.663** | **0.629** | **0.651** | **0.743** | 0.526 |
| 33 | 0.177 | 0.255 | **0.735** | 0.530 | **0.706** | **0.620** | 0.470 |
| 38 | 0.327 | 0.107 | 0.414 | 0.370 | 0.368 | 0.402 | 0.438 |
| 39 | **0.739** | 0.363 | 0.477 | 0.538 | 0.451 | **0.630** | **0.732** |
| 42 | **0.758** | 0.292 | 0.442 | 0.453 | 0.409 | **0.629** | **0.711** |
| 43 | 0.432 | 0.384 | **0.654** | 0.219 | **0.621** | **0.718** | 0.359 |
| 44 | 0.220 | 0.131 | **0.790** | **0.643** | **0.759** | **0.778** | 0.259 |
| 45 | 0.374 | 0.236 | **0.702** | 0.592 | **0.679** | **0.737** | **0.620** |
| 46 | 0.495 | 0.099 | 0.575 | 0.510 | 0.539 | **0.652** | 0.585 |
| 47 | **0.666** | 0.367 | 0.428 | 0.462 | 0.389 | 0.496 | **0.753** |
| 49 | **0.744** | 0.170 | 0.468 | 0.412 | 0.450 | 0.491 | **0.684** |
| 50 | **0.660** | 0.055 | 0.383 | 0.411 | 0.353 | 0.440 | **0.619** |
| 51 | **0.759** | 0.267 | 0.504 | 0.411 | 0.468 | 0.544 | 0.531 |
| 35 | 0.456 | 0.321 | 0.393 | 0.338 | 0.352 | 0.464 | 0.563 |
| 70 | 0.498 | **0.656** | **0.705** | 0.343 | **0.678** | **0.692** | 0.366 |
| 71 | 0.366 | **0.739** | 0.578 | 0.282 | 0.581 | **0.613** | 0.241 |
| 74 | 0.590 | **0.633** | **0.644** | 0.299 | **0.610** | **0.686** | 0.358 |
| 75 | 0.449 | **0.601** | 0.397 | 0.125 | 0.362 | 0.366 | 0.555 |
| 8 | 0.362 | 0.526 | 0.520 | 0.274 | 0.489 | 0.593 | 0.493 |
| 12 | 0.559 | 0.389 | 0.218 | 0.206 | 0.176 | 0.248 | **0.735** |
| 5 | 0.237 | 0.373 | 0.226 | 0.103 | 0.206 | 0.129 | -0.300 |
| 9 | 0.022 | 0.119 | 0.346 | 0.164 | 0.322 | 0.281 | -0.372 |
| 10 | -0.036 | 0.132 | 0.371 | 0.180 | 0.337 | 0.154 | -0.114 |
| 11 | -0.380 | -0.194 | 0.426 | 0.414 | 0.427 | 0.041 | -0.375 |
| 13 | 0.225 | 0.458 | 0.344 | 0.276 | 0.320 | 0.451 | -0.091 |
| 14 | 0.090 | 0.301 | 0.417 | 0.318 | 0.408 | 0.391 | -0.335 |
| 15 | 0.203 | -0.016 | 0.344 | 0.238 | 0.306 | 0.132 | -0.063 |
| 4 | -0.461 | -0.231 | 0.110 | -0.050 | 0.128 | -0.172 | -0.626 |

The correlation coefficient $r$ values greater than 0.6 are printed in **bold face**.

Table 7.2 displays correlation coefficient values for weather features with four or more $r > 0.6$ with orchards planted in the year 2000 or later.

Table 7.2: Correlation coefficient values for highly correlated weather features in younger orchards

| Oc | prev Oct HU | prev Oct Mean Temp | May HU |
|---|---|---|---|
| 17 | 0.370 | 0.357 | -0.357 |
| 90 | 0.081 | 0.027 | -0.129 |
| 56 | 0.278 | 0.283 | -0.281 |
| 57 | 0.447 | 0.418 | -0.213 |
| 58 | 0.482 | 0.447 | -0.042 |

Continued on next page

Table 7.2 – continued from previous page

| Orchard | prev Oct HU | prev Oct Mean Temp | May HU |
|---|---|---|---|
| 61 | 0.434 | 0.401 | -0.279 |
| 33 | **0.657** | **0.612** | 0.012 |
| 38 | 0.316 | 0.279 | 0.046 |
| 39 | 0.271 | 0.255 | -0.272 |
| 42 | 0.271 | 0.257 | -0.281 |
| 43 | 0.492 | 0.470 | -0.039 |
| 44 | **0.713** | **0.680** | 0.329 |
| 45 | 0.559 | 0.525 | -0.155 |
| 46 | 0.375 | 0.346 | -0.045 |
| 47 | 0.232 | 0.197 | -0.308 |
| 49 | 0.341 | 0.346 | -0.241 |
| 50 | 0.151 | 0.137 | -0.245 |
| 51 | 0.433 | 0.437 | 0.046 |
| 35 | 0.217 | 0.184 | -0.072 |
| 70 | 0.497 | 0.469 | -0.089 |
| 71 | 0.393 | 0.373 | -0.302 |
| 74 | 0.459 | 0.430 | -0.157 |
| 75 | 0.154 | 0.111 | -0.341 |
| 8 | 0.358 | 0.317 | -0.231 |
| 12 | -0.064 | -0.107 | -0.342 |
| 5 | 0.522 | 0.537 | 0.445 |
| 9 | **0.784** | **0.801** | **0.711** |
| 10 | 0.463 | 0.437 | 0.597 |
| 11 | **0.720** | **0.715** | **0.613** |
| 13 | **0.690** | **0.697** | 0.417 |
| 14 | **0.793** | **0.814** | **0.615** |
| 15 | **0.629** | **0.637** | **0.708** |
| 4 | 0.503 | 0.534 | **0.622** |

Table 7.3: Multiple linear regression on features with correlation threshold of **0.8** and greater

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 1989 | 56 | prev May Hum | $3.987 prevHum_5 - 52.043$ | 19.482 |
| 1989 | 58 | Aug Max Temp | $9.943 MaxT_8 - 210.391$ | 20.282 |
| 1991 | 44 | Aug Min Temp | $18.898 MinT_8 + 13.678$ | 17.509 |
| 1995 | 12 | Oct Hum | $8.145 Hum_{10} - 116.078$ | 19.028 |
| 2000 | 9 | prev Oct Mean Temp | $14.262 prevMeanT_{10} - 265.805$ | 20.087 |
| 2000 | 11 | Jan Wind | $10.086 Wind_1 + 27.385$ | 13.749 |
| 2000 | 14 | prev Oct Mean Temp, Nov Heat units | $8.286 prevMeanT_{10} + 0.321 HU_{11} - 196.888$ | 13.600 |

Table 7.4: Multiple linear regression on features with correlation threshold of **-0.8** and less

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 1978 | 17 | May Max Temp | $-7.491MaxT_5 + 343.229$ | 9.982 |
| 1989 | 56 | prev May Mean Temp, prev Jun Max Temp | $-9.058prevMeanT_5 - 12.641prevMaxT_6 + 654.536$ | 19.062 |
| 1989 | 57 | May Max Temp | $-18.679MaxT_5 + 735.897$ | 23.296 |
| 1989 | 61 | May Max Temp | $-17.779MaxT_5 + 692.726$ | 22.653 |
| 1991 | 51 | prev Jun Max Temp | $-17.234prevMaxT_6 + 617.098$ | 17.604 |
| 2000 | 11 | Aug Rain | $-4.130Rain_8 + 106.650$ | 13.205 |

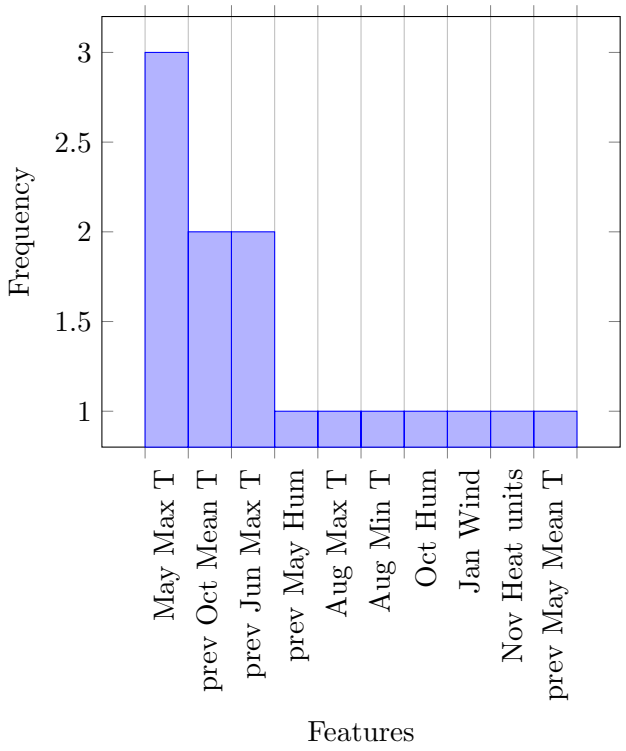



Figure 7.1: Number of features with $r < -0.8$ and $r > 0.8$ with individual orchard harvests per tree

Tables 7.3 and 7.4 display linear regression equations on features with a correlation threshold of 0.8 and greater, and -0.8 and less, respectively, with the yields on orchard-level. Section B.1 in Appendix B contains tables with equations for less strict thresholds. For a correlation coefficient of 0.8 and greater, or -0.8 and smaller, with the orchard harvests per tree, a bar graph is drawn of the applicable features in Figure 7.1.

As seen from the correlations, and confirmed by the coefficients, the influence on the yield of maximum temperature in May is always negative, previous October mean temperature positive, while previous June maximum temperature is negative.

By considering only the weather features included in the dataset as possible predictors of yield, it can be assumed that a lower maximum temperature in May, higher mean temperature in October of the previous year and lower maximum temperature of June also of the previous year, influence the yield of the season positively and may result in a higher harvest mass.

Table 7.5 displays the selected features for verifying the correlation threshold method, with their associated $p$-values indicating statistical significance of the correlation coefficient, and RMSE when a simple Least Squares regression is fitted. The $p$-values all below 0.05 suggest statistical significance.

Table 7.5: SelectKBest method, $K = 1$

| Year | Oc | Weather features | $p$ | RMSE |
|---|---|---|---|---|
| 1978 | 17 | May Max Temp | 0.00220 | 9.982 |
| 1989 | 56 | prev May Humid | 0.00054 | 19.482 |
| 1989 | 57 | May Max Temp | 0.00175 | 23.296 |
| 1989 | 58 | Aug Max Temp | 0.00258 | 20.282 |
| 1989 | 61 | May Max Temp | 0.00191 | 22.653 |
| 1978 | 90 | Apr Mean Temp | 0.00642 | 15.987 |
| 1993 | 35 | Oct Humid | 0.05930 | 28.525 |
| 1991 | 38 | Jul Heat units | 0.06295 | 26.201 |
| 1991 | 33 | Aug Heat units | 0.01003 | 12.534 |
| 1991 | 39 | prev May Humid | 0.00934 | 25.911 |
| 1991 | 42 | prev May Humid | 0.00688 | 24.138 |
| 1991 | 47 | Jul Rain | 0.00745 | 23.745 |
| 1991 | 43 | Aug Max Temp | 0.01290 | 27.030 |
| 1991 | 44 | Aug Min Temp | 0.00122 | 17.509 |
| 1991 | 45 | May Max Temp | 0.00895 | 25.544 |
| 1991 | 46 | prev Jun Max Temp | 0.02331 | 23.980 |
| 1991 | 49 | prev May Humid | 0.00862 | 31.595 |
| 1991 | 50 | prev Jun Max Temp | 0.02610 | 27.249 |
| 1991 | 51 | prev Jun Max Temp | 0.00094 | 17.604 |
| 1993 | 70 | Aug Heat units | 0.01541 | 18.058 |
| 1993 | 71 | Dec Wind | 0.00734 | 12.306 |
| 1993 | 74 | Aug Max Temp | 0.01987 | 18.611 |
| 1993 | 75 | prev Oct Min Temp | 0.01491 | 23.086 |
| 1995 | 8 | May Max Temp | 0.02918 | 23.413 |
| 1995 | 12 | Oct Humid | 0.00195 | 19.028 |
| 2000 | 5 | Mar Min Temp | 0.00642 | 14.699 |
| 2000 | 9 | prev Oct Mean Temp | 0.00305 | 20.087 |
| 2000 | 10 | prev Jul Max Temp | 0.00382 | 13.845 |
| 2000 | 11 | Aug Rain | 0.00230 | 13.205 |
| 2000 | 13 | Mar Min Temp | 0.00841 | 23.164 |
| 2000 | 14 | Nov Heat units | 0.00226 | 15.649 |
| 2000 | 15 | Jun Max Temp | 0.00787 | 15.277 |
| 2002 | 4 | Jan Wind | 0.00579 | 17.629 |

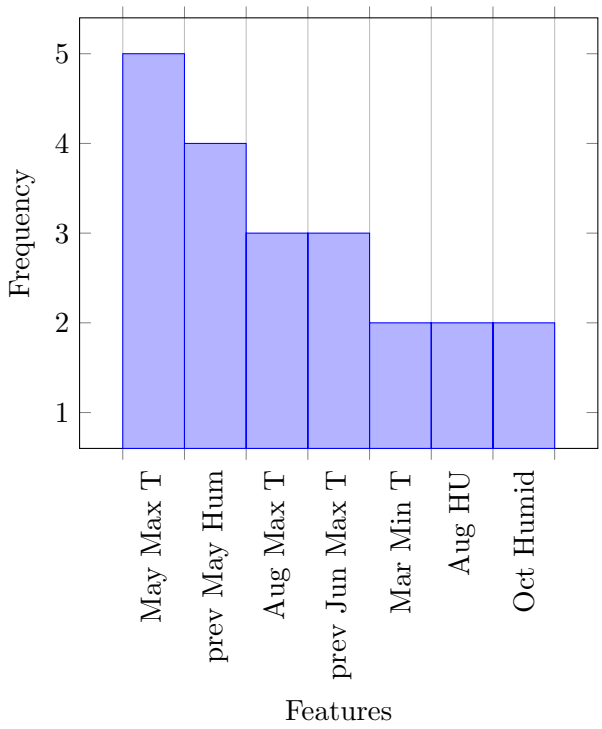



Figure 7.2: SelectKBest weather features chosen for more than one orchard

With SelectKBest, the frequency of chosen features is displayed in Figure 7.2 for features chosen for more than one orchard. From this, it can be concluded that the feature May Max Temp is dominant in the feature set, although all the features displayed in the figure have significance over features for other months and measurements which are not presented.

### 7.1.2 Correlation approach on the entire yield

Next the focus is shifted to the entire yield of all the orchards. For a better understanding of the effect of temperature on the yield, the temperatures of the year prior to the harvest are also processed into monthly features.

Applying the same approach as with the individual orchards, considering correlation coefficient thresholds for the total yield of the season, the following features result. A correlation $r > 0.8$ is found with the features prev Oct Heat units and prev Oct Mean Temp, which evidently have the strongest positive relationships with the total yield. Considering a lower threshold $0.7 \leq r < 0.8$, the features prev Oct Max Temp, Sep Heat units and Sep Mean Temp emerge. Lowering the threshold to 0.6 produces the following features as well: prev Jan Max Temp, May Heat units, Aug Heat units, Apr Min Temp, Aug Min Temp, Sep Min Temp, May Mean Temp, Aug Mean Temp, and Dec Max Temp. Negative relationships are found for Aug Humidity with an $r < -0.7$, and a slightly less strict threshold of $-0.7 < r \leq -0.6$ produces the features prev Oct Humidity, Apr Humidity, and Aug Rain.

Between the entire yield and each of the weather features, no correlation coefficient of -0.8 or smaller exists.

A correlation threshold of 0.7 and greater or -0.7 and smaller, generates the following selected variables:

- prev Oct Heat units
- prev Oct Mean Temp
- prev Oct Max Temp

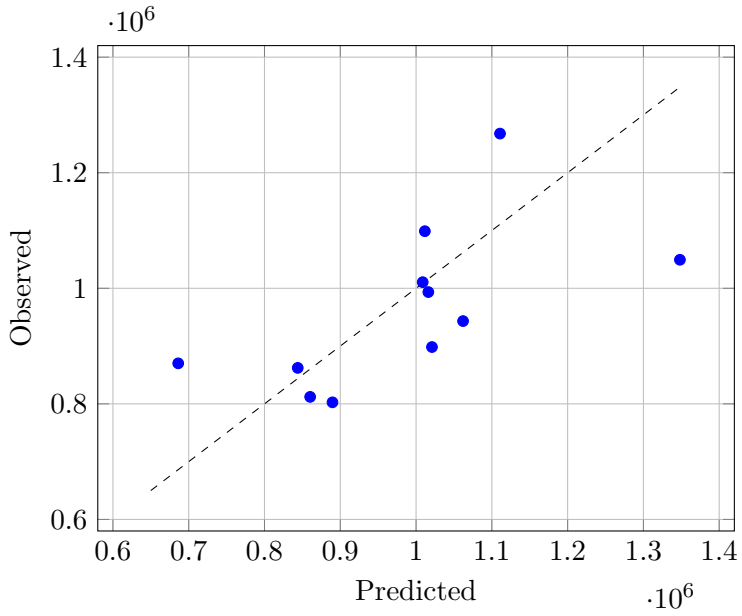



Figure 7.3: Observed vs predicted for $r > 0.7$ features

- Sep Heat units
- Aug Humidity
- Sep Mean Temp

The implementation of these features in a linear regression model leads to an adjusted $R^2$ of 0.7380 and a RMSE of 42 963.58 kg.

The OLS assumptions are checked for this to ensure that the linear model is, in fact, trustworthy. With respect to the error term population mean of zero, the mean of the residuals (the mean error) is -6.561E-10, which is sufficiently close to zero. The model is accordingly unbiased as the average value of the error term equals zero. Checking the exogeneity assumption, the correlations between the residuals and each of the predictors are:

- prev Oct Heat units: -1.16039E-15
- prev Oct Mean Temp: -8.4526E-16
- prev Oct Max Temp: -1.27011E-15
- Sep Heat units: -1.43583E-15
- Aug Humidity: 3.60143E-15
- Sep Mean Temp: -1.36026E-15

It can therefore be concluded that there is no endogeneity. Checking heteroscedasticity is challenging with so few data points, although the variance of the errors seems to be consistent. Finally, no independent variable is a perfect linear function of other explanatory variables.

The predicted values are calculated with LOOCV and displayed against the measured values in Figure 7.3. In the case of ideal predictions, the scatter plot lies on the dotted line, where the predicted values are similar to the observed values.

Evaluating model predictions is a critical step in the development of a model, and visualisation is a useful aid in the evaluation process. Scatter plots are often drawn to show the correlation between actual (or observed) and predicted values. Piñeiro et al. (2008) investigated the correct way to plot these variables on the axes, as even commonly used software like Statistica and Excel differ in the way

variables are placed on the axes. Piñeiro et al. (2008) showed empirically and demonstrated analytically that model evaluation based on linear regressions should be done with the observed values placed on the y-axis and the predicted values on the x-axis (Observed vs Predicted). In this graphical evaluation, the $r^2$ can be used again, as a measure of the proportion of the variance in observed values that is explained by the predicted values.

Figure 7.3 displays the predicted values versus the target values when using the correlation threshold of -0.7 and +0.7 to select features for linear regression on the total seasonal yield. The dotted line is the position of ideal points, where the prediction is the target value.

The correlation threshold method is verified with the function SelectKBest from the Scikit-learn project (Pedregosa et al., 2011). With $K$ set to one feature, the top feature is heat units during October of the previous season, prev Oct HU with a $p$-value of 0.00026. This feature in a simple linear regression produces an RMSE of 85 332.72 kg when used to fit a least squares regression. Setting $m$ to a maximum of 9, as dictated by restriction of the sample size, results in the features and their corresponding F-values and $p$-values as shown in Table 7.6. The $p$-values which are all below 0.05 indicate the statistical significance of the correlations.

Table 7.6: SelectKBest on entire yield with number of features set to 9

| **Feature** | **F** | ***p*-value** |
|---|---|---|
| prev Oct Humid | 8.07565 | 0.01935 |
| prev Oct heat units | 33.73145 | 0.00026 |
| prev Oct Mean Temp | 28.20877 | 0.00049 |
| prev Oct Max Temp | 9.17049 | 0.01429 |
| Sep Heat units | 12.68347 | 0.00611 |
| Aug Humid | 9.92790 | 0.01172 |
| Aug Min Temp | 7.19909 | 0.02508 |
| Sep Min Temp | 7.55339 | 0.02254 |
| Sep Mean Temp | 11.76389 | 0.00751 |

### 7.1.3 Forward stepwise regression on the individual orchard yield and weather features

In the case where many independent variables may play a role in the behaviour of the response variable, stepwise regression is used to select important variables in order to obtain a simpler model. It is implemented to predict the yield of the individual orchards from the weather features.

By using the AICc as selection criterion, the features were identified for each orchard, and the relevant regression was fitted to obtain the leave-one-out cross-validated root mean squared error (RMSE) and the AICc used for the choice of variables. The AICc alone has no value for interpretation. The table showing chosen features for each orchard can be found in Appendix B Section B.2. As seen from the RMSE values and Figure 7.5 forward stepwise regression with $m$ set to the highest possible value produces overfitted models. This is, of course, because of the nature of the small dataset, with a $m > n$ shape and only 11 observations. Although an investigation into the models reveals very small $p$-values ($< 0.005$) of the t-values of the estimated coefficients, a low F-score of the model and a very high adjusted $R^2$ close to 1, the purpose of the implementation of these methods is identifying significant features. For this purpose, all the features are taken into consideration even if the model is overfitted.

As seen in Figure 7.4, the most selected features, chosen eight or more times for the 33 orchards, are the maximum temperature in May and the minimum temperature of the previous October.

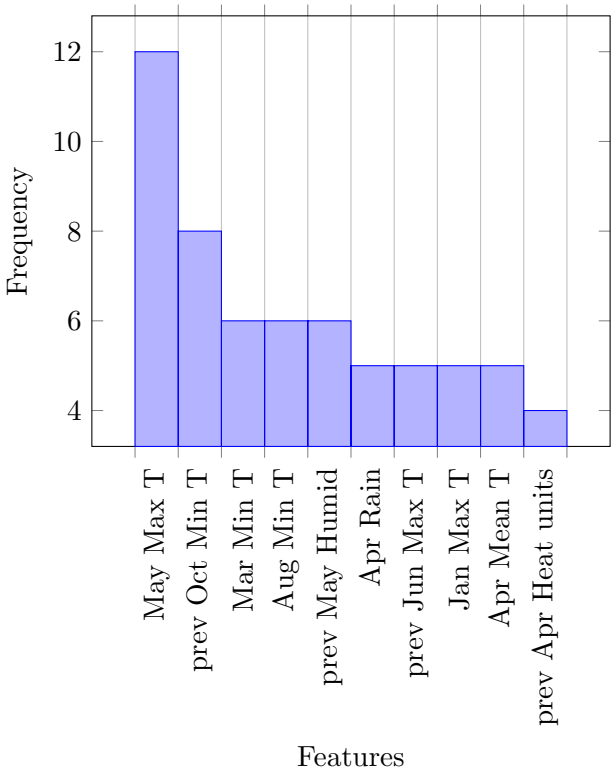



Figure 7.4: Features most chosen with stepwise regression on orchard-level

### 7.1.4 Forward stepwise regression on entire yield

Considering the farm as a whole, stepwise regression is done with the independent variables kept the same as above (144 possible features to choose from) with the total yield taken as the dependent variable. The following features are selected by this method:

- Previous Oct Heat units
- Previous Feb Mean Temp
- Jul Max Temp
- Dec Rain
- Previous Sep Max Temp
- Previous Nov Mean Temp
- Previous Nov Min Temp
- Previous Jan Min Temp
- Previous Apr Heat units

By using the AICc as selection criterion, the features were identified for each orchard, and the relevant regression was fitted to obtain the leave-one-out cross-validated root mean squared error (RMSE) and the AICc used for the choice of variables. The AICc alone has no value for interpretation. Although the low RMSE values indicate well-fitted models, they may be indicative of overfitting, however they are useful for determining relevant role-playing features.

The table showing the resulting RMSE of each regression for each orchard can be found in Appendix B. The most selected features, chosen eight or more times for the 33 orchards, are the maximum temperature in May and the minimum temperature of the previous October.

This model has an adjusted $R^2$ value of $\sim 1$ and a leave-one-out cross-validated RMSE of 231 kg, on the total yearly harvests ranging from 812 083 to 1 267 714 kg, which clearly indicates how

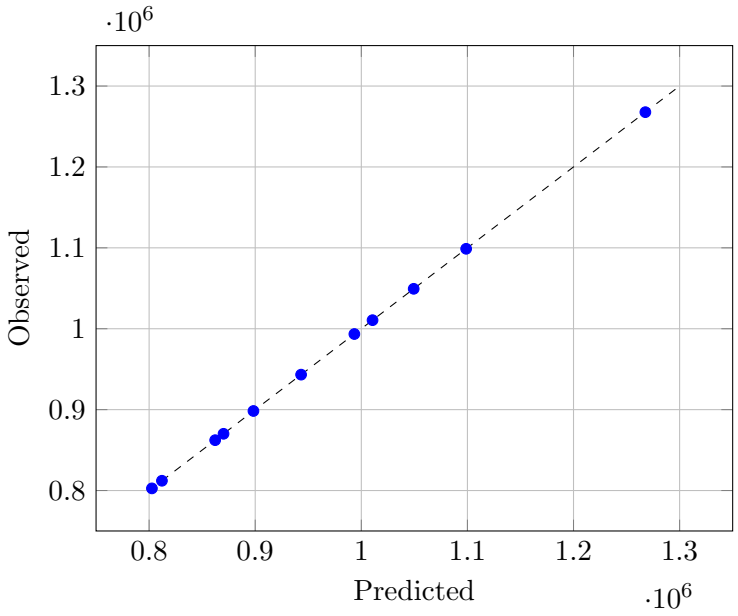



Figure 7.5: Observed vs predicted values for forward stepwise regression on entire yield

overfitted the model is. This is because of the large number of chosen features on the small dataset, so that $n \sim m$. The coefficients for the variables are shown to indicate the sign (and thus effect) of the feature.

- prev Oct Heat units: 3 654.88
- prev Feb Mean Temp: 32 823.43
- Jul Max Temp: 21 190.30
- Dec Rain: 3 482.12
- prev Sep Max Temp: -6 711.40
- prev Nov Mean Temp: 7 337.21
- prev Nov Min Temp: -1 204.09
- prev Jan Min Temp: 339.09
- prev Apr Heat units: 1.62

### 7.1.5 Elastic net regression on weather and yield data of individual orchards

This section presents the implementation of the elastic net algorithm, a combination of the lasso and ridge penalties, which outputs a default of one less feature than the number of observations. James et al. (2013) states that if $n$ is not much larger than $m$, a high variability may be present in the least squares fit. Because of the small sample size and the high number of features chosen by the elastic net algorithm (for which elastic net has received criticism), this may cause overfitting and consequently poor predictions on out-of-sample observations not used in model training. To avoid this overfitting, all the features from elastic net are iterated and a least squares model fitted each time, applying AICc to find the lowest AICc. The model with the smallest AICc value is chosen, resulting in the final necessary features.

The hyperparameters obtained from cross-validated grid search and used for the elastic net regression on the individual orchards, are $\mathsf{alpha} = 0.01$ and $l1\_ratio = 0.97$.

To illustrate, an example of Orchard 17 is presented. The iteration over number of features and best corresponding AICc is as follows:

1. One feature: 22.38
2. Two features: 19.81
3. Three features: 23.07
4. Four features: 8.07
5. Five features: 2.89
6. Six features: 4.49
7. Seven features: -2.39
8. Eight features: 18.10
9. Nine features: 83.48

Consequently, the chosen features will be from an iteration of seven features producing the best AICc value when evaluating the resulting model. As a result, the lowest AICc for orchard 17 has seven features (the seven most prominent regressors as identified with elastic net), producing a model with an adjusted $R^2$ of 0.9988. The values for the adjusted $R^2$ resulting from the elastic net regression for each orchard are included in Appendix B Section B.3.

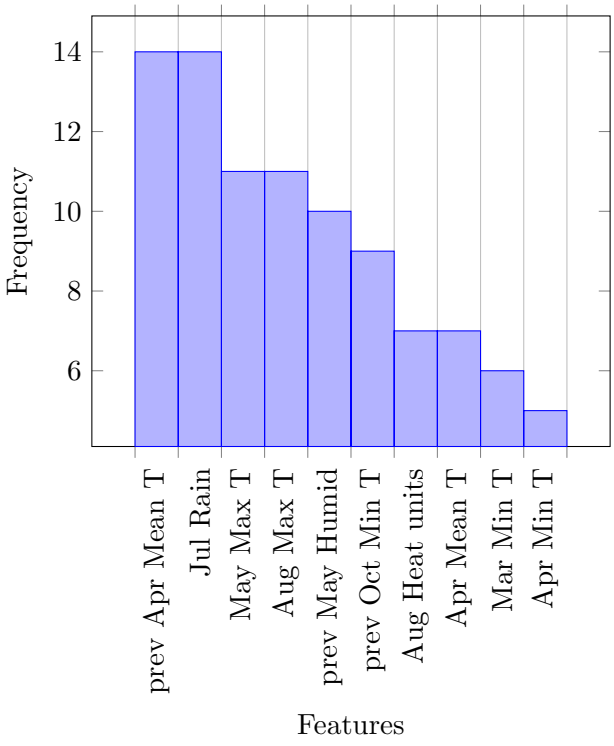



Figure 7.6: Top 10 features chosen with elastic net and AICc for all orchards

The table showing all the features chosen for each orchard and the corresponding AICc value can also be found in Appendix B. The top 10 features, with the number of orchards for which they are chosen, are displayed in Figure 7.6.

Taking into consideration the signs of the coefficients of the features, the effect of rain in July, the mean temperature of the previous April and the maximum temperature in August are positive, while the maximum temperature in May is negative.

For interpretability, coefficients are not scaled. Scaling is done on the initial dataset used to select features with elastic net. However, once the features are selected, the values are used for fitting the linear regression and finding the model with the smallest AICc value.

The chosen features for each orchard, following from elastic net regression analysis, the equation for the harvest per tree and the respective RMSE and Adjusted $R^2$ values, can be found in Appendix B Section B.3.

### 7.1.6 Elastic net regression applied on data of the entire farm

Performing elastic net on the scaled values of the weather features and the total yield per season (the farm production as a whole) with the alpha value and L1 ratio found from cross-validation, alpha set to 0.01, $\mathsf{l1\_ratio} = 0.79$, results in a model with the following features:

- prev Oct Heat units
- prev Oct Min Temp
- Sep Mean Temp
- Sep Heat units
- Apr Min Temp
- prev Jun Min Temp
- prev Oct Mean Temp
- Mar Rain
- Jun Humidity

Using these as the independent variables of a linear regression model, the model has an adjusted $R^2$ of 0.99987 and an LOOCV RMSE of 20 129.74 kg which could be deemed acceptable given the scale of the target values.

Applying stepwise feature selection on these features, and evaluating with AICc, the lowest AICc is obtained with a model including three features:

- prev Oct Heat units
- prev Oct Min Temp
- Sep Mean Temp

For this model the adjusted $R^2$ is 0.8695, AICc is 16.32 and LOOCV RMSE is 65 726.78 kg.

Figure 7.7 displays the observed yield of the farm vs the farm yield predicted by elastic net regression. Elastic net, with the use of AICc stepwise selection, emerges as an improvement on forward stepwise regression for three reasons. It is an easily implementable algorithm, with the use of packages such as Python's Scikit-learn, and it is not as computationally expensive. Less features are selected by this method, which is beneficial for instances with a small number of observations as is the case in this study, although for the purposes of identifying all significant features, choosing more features would be an advantage.

### 7.1.7 Partial least squares regression on weather features and yield data on orchard-level

This section presents the application of partial least squares regression (PLSR) to predict the yield of the individual orchards from the different weather features. Implementing PLSR in Python as well as with XLStat in Excel, with the optimal number of extracted factors or $\mathsf{number\ of\ components} = 5$, provides the VIP scores used to determine the most significant features.

The plot in Figure 7.8 displays how the optimal number of components to be used as parameters in the PLSR model was determined.

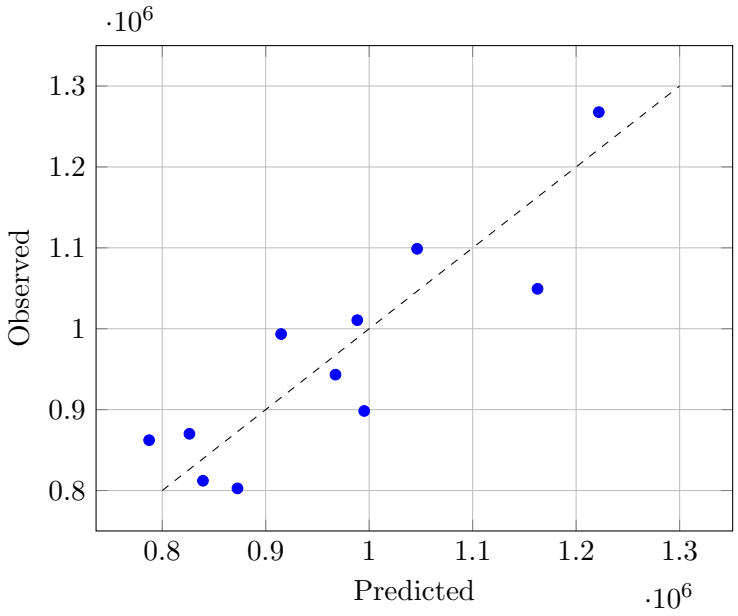



Figure 7.7: Observed yield vs predicted yield from elastic net regression with three features: prev Oct HU, prev Oct Min Temp, Sep Mean Temp on entire yield

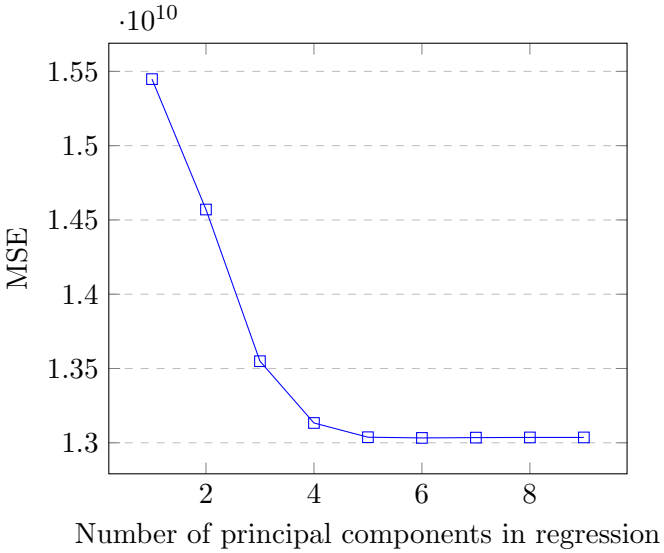



Figure 7.8: Cross-validation to determine optimal number of components

For the identification of important features, VIP scores are used as discussed in Chapter 6. Variable $j$ can be eliminated from the selection if $v_j < u$ for some user-defined threshold $u \in [0, \infty)$. It is generally accepted that a variable should be included in the selection if $v_j > 1$ (Gosselin et al., 2010). For this study, using a total number of 144 candidates of weather variables as $X$ and the total farm harvest as dependent variable $Y$, the threshold $u$ is set to 1.5, a slightly higher threshold than 1, to filter only the most relevant candidates and for ease of comparison with other implemented methods.

Considering the orchards separately, Table B.7 in Appendix B Section B.4 shows the features with a VIP score greater than 1.8, as 1.5 results in too many features and a threshold of 2 results in cases where no features qualify, to clearly distinguish the differences between orchards in the same area.

The 10 most chosen features are displayed in Figure 7.9, where the frequency displays the number of orchards for which a feature is chosen.

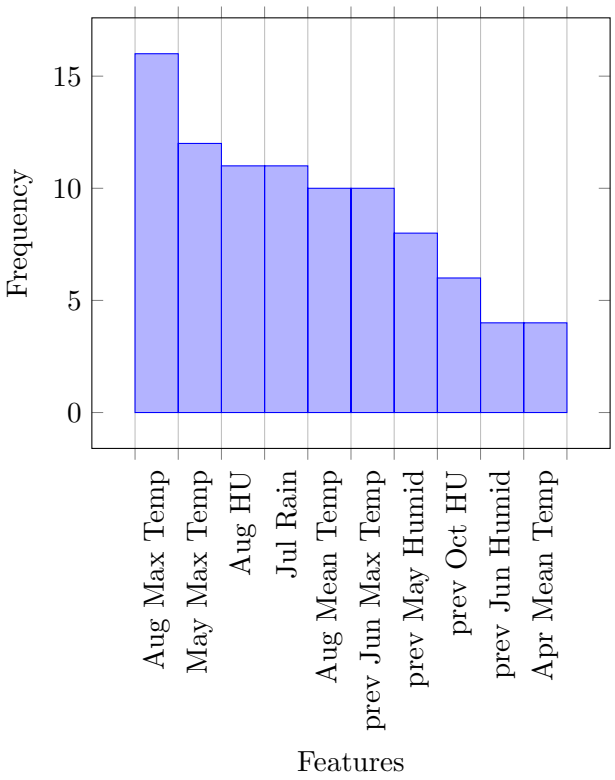



Figure 7.9: Top 10 features of all the orchards with VIP score $> 1.8$ chosen with PLS VIP

### 7.1.8 Partial least squares regression on entire yield

For the yield of the entire farm, the selected features are the ones for which the VIP score is greater than 1.5. The resulting model has a resubstitution $R^2$ of $\sim 1$ and a LOOCV $R^2$ of 0.26. The LOOCV RMSE of 114 181.87 kg, for a target sample with a mean value of 964 421.45 kg and standard deviation of 132 720.15, indicates that this model produces more accurate results than using the mean of the target as the prediction.

Choosing features with VIP scores greater than 1.8, yields the following features:

- prev Oct Humid
- prev Oct Heat units
- prev Oct Mean Temp
- prev Oct Max Temp
- Aug HU
- Sep HU
- Aug Humid
- Apr Min Temp
- Aug Min Temp
- Sep Min Temp
- Sep Mean Temp

The dotted line in Figure 7.10 indicates where the dots should ideally lie, if the PLS predicted the targets correctly.

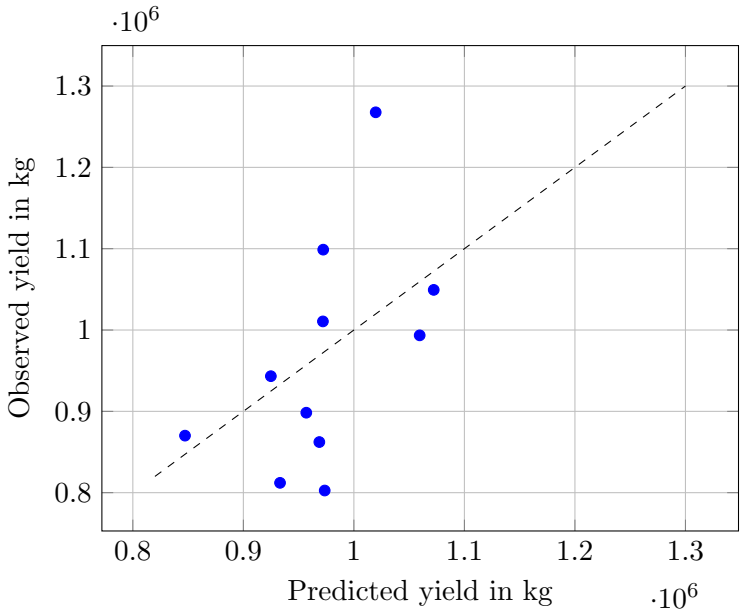



Figure 7.10: Observed vs predicted yield for PLS regression with five components on entire yield

### 7.1.9 Summary of regression results on orchard level

From the implemented feature selection methods above, the selected features common to the most orchards from Figures 7.1, 7.2, 7.4, 7.6, and 7.9 are displayed in Table 7.7. Features printed in purple are shared by three or more of the methods.

Table 7.7: Performance of regression methods summarised on all orchards

| **SelectKBest** | **Corr** (>0.8, <-0.8) | **Forward stepwise** | **Elastic Net** & AICc | **PLS** (VIP >1.8) |
|---|---|---|---|---|
| May Max T | May Max T | May Max T | prev Apr Mean T | Aug Max T |
| prev May Humid | prev Oct Mean T | prev Oct Min T | Jul Rain | May Max T |
| Aug Max T | prev Jun Max T | Mar Min T | May Max T | Aug HU |
| prev Jun Max T | prev May Humid | Aug Min T | Aug Max T | Jul Rain |
| Mar Min T | Aug Max T | prev May Humid | prev May Humid | Aug Mean Temp |
| Aug HU | Aug Min T | Apr Rain | prev Oct Min T | prev Jun Max T |
| Oct Humid | Oct Humid | prev Jun Max T | Aug HU | prev May Humid |
| | Jan Wind | Jan Max T | Apr Mean T | prev Oct HU |
| | Nov HU | Apr Mean T | Mar Min T | prev Jun Humid |
| | prev May Mean T | prev Apr HU | Apr Min T | Apr Mean T |

Features selected by 3 or more methods are: maximum temperature in May (May Max Temp), August (Aug Max Temp) and June of the previous year (prev Jun Max Temp), mean humidity of May of the previous year (prev May Humid), minimum temperature in March (Mar Min Temp), and heat units in August (Aug HU).

Thus far, a few remarks can be made. Although initially only used for exploration, the correlation methods select the features accurately. They indicate the strength of the relationships between predictor and target. All the features selected by three or more methods have strong correlations with the target. In conditions of low available computational power or time limitation, performing the Pearson correlation calculations between the constructed features and the targets may be sufficient. Using the correlation coefficient with a threshold of positive and negative 0.8 has, in this particular situation with a small sample size and relatively high dimensionality, four out of 10 features in common with

the forward stepwise regression. Because of the many features chosen by this implementation of forward stepwise regression, it often produces overfitted models and may be regarded as too expensive computationally.

The current implementation of elastic net regression requires choosing the appropriate hyperparameter with cross-validation and performing stepwise selection to choose the optimal model using AICc. This increases the complexity of the model although it is still less complex than forward stepwise regression.

Judged by the execution time of the implementations, forward stepwise regression is the slowest (its fastest execution takes 1 min 38 s per orchard), followed by the elastic net implementation incorporating stepwise selection (taking at least 4.67 s per orchard). This is executed after hyperparameter tuning with cross-validation, taking 5.58 s. PLS regression takes 125 ms and evaluating correlation thresholds 78 ms. These times are displayed in Table 7.8.

Table 7.8: Running times of implementations of methods

| **Method** | **Running time** |
|---|---|
| Forward stepwise | 1 min 38 s |
| Elastic net and stepwise | 10.25 s |
| PLS | 125 ms |
| Correlation | 78 ms |

All the thresholds for the correlation-based method and the VIP scores are user-specified and not chosen by the algorithms. Therefore, a hyperparameter tuning time such as for elastic net regression is not included for the other methods. Table 7.8 shows the superior complexity of the stepwise method over the other methods. However, the execution times do not provide any further insight and are not considered hereafter.

### 7.1.10 Summary of regression results on the entire yield

The selected features from the five selected regression methods are displayed in Table 7.9.

Table 7.9: Features selected from the various methods predicting total yield

| **SelectKBest** | **Corr(>0.7,<-0.7)** | **Forward Stepwise** | **Elastic Net** | **PLS (VIP >1.5)** |
|---|---|---|---|---|
| prev Oct Humid | prev Oct HU | prev Oct HU | prev Oct HU | prev Oct HU |
| prev Oct HU | prev Oct Mean T | Prev Feb Mean T | prev Oct Min T | prev Oct Mean T |
| prev Oct Mean T | prev Oct Max T | Jul Max T | Sep Mean T | Sep HU |
| prev Oct Max T | Sep HU | Dec Rain | Sep HU | Sep Mean T |
| Sep HU | Aug Humid | prev Sep Max T | Apr Min T | Aug Humid |
| Aug Humid | Sep Mean T | prev Nov Mean T | prev Jun Min T | prev Oct Max T |
| Aug Min T | | prev Nov Min T | prev Oct Mean T | prev Oct Humid |
| Sep Min T | | prev Jan Min T | Rain Mar | Aug HU |
| Sep Mean T | | prev Apr HU | Jun Humid | Sep Min Temp |
| | | | Apr Min T | |
| | | | Aug Min T | |

Features selected by three or more methods displayed in purple are:

- prev Oct Heat units

- prev Oct Mean Temp
- Sep Mean Temp
- Sep Heat units

In essence, these methods were all used to identify the most important features on which to fit an OLS regression model. For the PLS method, two approaches are followed. First, the features from the VIP selection method are used in an OLS model. Secondly, the PLS output itself with the new components is evaluated.

## 7.2 Justifying selected weather features

In an attempt to account for the features selected and identified during this study by employing the different regression methods and discussed in Chapters 5 to 7, the features are interpreted and analysed with reference to previous scientific studies and research on factors influencing crop growth, development and production. Validation for these features can be interpreted differently because factors are interrelated resulting in diverse outcomes, but some selected features can be justified with some precision. Substantiation for the selection of these features, if it exists in literature, is presented. Since the selected features have a significant influence on the yield and are related to a specific phenological stage of the date palm tree, justifications are reviewed in this regard.

### 7.2.1 Features encouraging production and yield

These features for which justification is sought in the literature, have a positive effect on the yield, as judged by the coefficient sign in the developed models.

**High heat units during August, High heat units during September, High mean temperature in September**
Heat units are calculated by means of different methods, all using a preferred base temperature and a principle of calculating the mean temperature above the base temperature. For this study, Equation 2.1, where the heat units of the month are the sum of daily averaged temperatures minus the base temperature of 18 °C, is used. By any acceptable formula, the number of heat units is directly proportional to the mean temperature. This correlation implies that higher heat units are the result of higher mean temperatures. High heat units and high mean temperature are therefore both considered the result of high temperature. During feature selection, choosing one of the two features Heat units and Mean temperature for the same month, can for that reason be explained by the same theory. After fruit set, seeded fruit has a rapid growth rate and the increase in size and mass follows a sigmoid growth curve. For the date fruit the *Kimri* stage is characterised by an intense increase in size and mass and the colour changes from green to yellow gold (Chao and Krueger, 2007). For this to happen, extensive photosynthesis is necessary. Depending on environmental circumstances, pollination and fruit set occur around July and early August, and fruit development gets under way in August. Photosynthesis is essential for healthy physiological development, including fruit growth. For photosynthesis to take place, radiant energy from the sun is required, among other factors. High heat units during the August and September period promote vigorous photosynthesis, resulting in dynamic fruit development and gain in mass.

**High mean temperature during October of the previous year, high heat units during October of the previous year, high maximum temperature in October of the previous year**
It is evident from the data that a higher yield was realised where the farm experienced temperatures above the average during October of the preceding year. It is a well-established phenomenon that the physiology of perennial plants and in particular of fruit trees is dependent on factors stretching over more than one season. During photosynthesis energy from sunlight converts water, carbon dioxide,

and minerals into oxygen and energy-rich organic compounds. Carbohydrates such as starch, accumulate in the leaves. At night, respiration transforms the starch to produce sucrose, a source of energy for metabolic processes such as growth of shoots, stems, roots and reproductive structures as well as defence against fungi and prevention of plant mortality (Kozlowski, 1992). An increase of photosynthetic production units causes plants to accumulate carbohydrates in different organs, called sinks, during periods of excess production (White et al., 2015). Proteins may also be stored for growth and production in subsequent years. This is typically applicable at times of high temperatures where the sun's radiation is more effective with a high RAD. The plant maximises its lifetime by choosing the best growth schedule within each season and also by allocating resources between vegetative growth, protection and reproduction for the year and storage for the next year. This confirms why the selected features referring to high heat units, high mean temperature and high maximum temperature during October of the previous year had an encouraging result on production and yield.

### 7.2.2 Features having a negative effect on production and yield

The features presented here generally have a negative coefficient in the developed models and are presumed to negatively affect the yield.

**Maximum Temperature in May**

When investigating the influence of high temperatures during May, the period preceding the flowering of the date palm tree, it is necessary to acknowledge the specific weather conditions required for fruit tree flowering to take place. One definite requirement is the need for trees to have previously been exposed to a certain period of low temperature for bud formation to be promoted. Studies by Shabana and Sunbol (2007) have shown that normal bud development of the date palm may be inhibited by unfavourable ecological factors such as high temperatures and that it requires a definite phase of cold hours during the period preceding the flowering season to stimulate the physiological processes needed for normal stimulation of flower bud formation and development. The feature of maximum temperature in May selected during this study, indicates an insufficient period of the required low temperature necessary for bud formation, consequently resulting in lower fruit production and yield.

**High rainfall in August and high humidity during August**

Pollination of the palm date occurs in the period from June to August. Precipitation during pollination time is likely to cause some reduction in the fruit setting of which the effect ultimately may be visible in the yield quantity (Zaid and de Wet, 2002*a*). Previous researchers have emphasised the damage of rain and its influence on date palm fruits. Some plausible explanations for this were stated. Rain or water might wash away some applied pollen from the stigmas of the female flowers before fertilisation can occur. Rain causes a reduction in the receptivity of female flowers by enclosing them with water. The relative air humidity around the flowers is also increased by water, promoting the rotting of the inflorescences as they are more prone to attacks from cryptogamic diseases. Rainfall also causes a decrease in air temperature. Al-Musawi (2019) found that pollen tubes of the date palm grow faster at higher temperatures but at 15 °C the pollen tubes did not reach the base of the style even after eight hours. Low temperatures between 8 and 20 °C causes an increase of parthenocarpic fruits and normal fruit development declines.

**High humidity during April**

When investigating the selected feature referring to high relative humidity during April, various negative impacts of humidity are considered. The palm tree undergoes a period of dormancy from the completion of the harvest until flowering. Depending on the harvesting date it is possible for the male spathes with flowers to develop and appear from April onwards. High humidity at flowering and pollination has a negative effect by sustaining fungal diseases in closed spathes, causing rot and loss of potential fruit.

Except for the weather data, additional supplementary data obtained from the research partner include the growth measurements and the bunch counts. This was investigated separately. The usefulness of the growth measurements was evaluated in Chapter 5 after which it was concluded that

the relationships were not strong enough for further investigation. Subsequently, the data on bunch count and mass was utilised.

## 7.3 Consideration of yield prediction with bunches data

Yield prediction models in the literature were discussed in Chapter 3. Next follows an investigation into the procedure implemented for yield prediction on the farm under study in this research and the documented data used for that purpose. Yield predictions as well as fruit size predictions are currently made as far as possible ahead, for various planning and logistics reasons including to find seasonal workers, order packaging material and to negotiate prices with buyers. The fruit sizes are categorised, as discussed in the literature analysis, as small, medium, large, jumbo and super jumbo. The yield prediction process currently implemented on the farm is labour-intensive. It is executed around December – harvest time starts in February – by counting the number of bunches on every tree in every orchard, and weighing an upper, middle and lower bunch on a tree in every orchard representing orchards of its age. The current procedure used by the research partner to calculate the expected harvest mass of the orchard is to multiply the number of bunches per orchard by the mass per bunch. This calculation is done for all the orchards, resulting in an overestimated harvest prediction for some orchards and an underestimation for others. However, with these compensations occurring, the overall outcome is a prediction with acceptable accuracy above 90% compared to actual yield obtained.

Table 7.10 displays the predictions and their accuracies of the arbitrarily selected Orchards 8, with 107 trees; and 12, containing 117 trees. The date palms in both of these orchards were planted in 1995. The number of bunches per tree is found by dividing the total number of bunches of the orchard by the number of trees in the orchard and the bunch mass is in kg per bunch.

The harvest predictions in Table 7.10 are an overestimation in some years and an underestimation in others. This is due to an incorrect estimation of bunch mass, which is usually an average for orchards with similar characteristics but should be done more granularly. The estimated bunch mass is determined by averaging the mass of one upper, one middle and one lower bunch in a representative orchard. Refining this process and improving results of bunch mass estimation advocates manually reducing the produce.

The bunch data (estimated bunch mass and bunch count of 33 orchards) is available for use in this study. This can be used as possible features in the feature selection methods implemented above for better comparison with, and possible improvement of the current method implemented by the research partner.

## 7.4 Incorporating bunch mass and number of bunches as additional features

In this section, it is discussed how the data on the counted number of bunches and the estimated mass per bunch is included in the list of possible features of the dataset. This is only applied on the individual orchards and not the entire yield where the farm is considered as a whole. The reason for this approach is that the bunch data differs between the orchards, whereas the weather data is constant across orchards and accordingly for the entire farm. Adding the bunch data is done separately for illustrative purposes as it is much more labour-intensive for an agricultural practitioner in South Africa to count and weigh the bunches than to extract the weather measurements from the weather station. The weather measurements are common for all orchards, but the bunches data is unique to each orchard. The four methods used to identify role-playing weather features will also be implemented in this section to determine if the bunch data is regarded as influential on the yield and if adding the bunch data has a significant effect on the evaluation metrics of the regression models.

Table 7.10: Number of bunches, harvests and predictions for orchards 8 and 12

| **Orchard** | **8** | | | | |
|---|---|---|---|---|---|
| Season | Yield/ tree | bunch count | Predicted Harvest/tree | bunch mass overestimated | %harvest overestimated |
| 2010 | 138.85 | 12.96 | 136.11 | -0.212 | -1.98 |
| 2011 | 158.82 | 14.57 | 167.56 | 0.599 | 5.50 |
| 2012 | 107.04 | 12.82 | 108.99 | 0.152 | 1.82 |
| 2013 | 89.47 | 13.09 | 91.65 | 0.167 | 2.44 |
| 2014 | 102.49 | 13.59 | 95.12 | -0.542 | -7.19 |
| 2015 | 101.29 | 13.43 | 94.01 | -0.542 | -7.19 |
| 2016 | 88.44 | 13.13 | 98.48 | 0.765 | 11.35 |
| 2017 | 116.83 | 12.97 | 84.32 | -2.506 | -27.83 |
| 2018 | 68.20 | 9.90 | 64.33 | -0.390 | -5.67 |
| 2019 | 92.31 | 15.15 | 98.47 | 0.407 | 6.67 |
| 2020 | 116.16 | 15.21 | 98.90 | -1.134 | -14.86 |
| **Orchard** | **12** | | | | |
| Season | Yield/ tree | bunch count | Predicted Harvest/tree | bunch mass overestimated | %harvest overestimated |
| 2010 | 155.97 | 15.35 | 161.18 | 0.340 | 3.34 |
| 2011 | 167.90 | 15.26 | 175.45 | 0.495 | 4.50 |
| 2012 | 115.27 | 13.56 | 115.29 | 0.002 | 0.02 |
| 2013 | 116.97 | 14.94 | 104.58 | -0.830 | -10.59 |
| 2014 | 108.09 | 14.03 | 98.24 | -0.702 | -9.11 |
| 2015 | 107.83 | 12.64 | 88.49 | -1.530 | -17.94 |
| 2016 | 90.36 | 13.26 | 92.79 | 0.184 | 2.70 |
| 2017 | 79.55 | 12.38 | 80.44 | 0.073 | 1.13 |
| 2018 | 103.62 | 15.59 | 109.13 | 0.353 | 5.31 |
| 2019 | 74.08 | 14.60 | 102.19 | 1.925 | 37.94 |
| 2020 | 99.90 | 15.14 | 105.96 | 0.400 | 6.07 |

The features bunch mass and bunch count represent the bunch mass and number of bunches of the respective orchards. The orchard harvest is only compared with the bunch mass of the particular orchard, not the bunch data of other orchards.

### 7.4.1 Correlation-based selection and regression on weather and bunches data to predict yield

In the same way in which the correlation coefficient values were used to select appropriate features from the weather dataset, the method is implemented with the bunches features added to the dataset. The user-defined threshold of the coefficients again determines the minimum strength of relationships between features tolerated when choosing features. The correlation coefficients are considered for the weather features and the mass per bunch and bunches per tree of the respective orchards. A low threshold (of 0.6 and -0.6) is chosen so that even a slight relationship is recognised. With the number of bunches and the estimated mass per bunch included in the list of features, it is clear that the estimated bunch mass of an orchard is related to the harvest of that orchard.

Table 7.11 shows correlation coefficient values for the bunches features with the yields of all the orchards.

Table 7.11: Correlation coefficient values for bunches features with orchard harvests

| Orchard | Bunch mass | Bunch count | Orchard | Bunch mass | Bunch count |
|---|---|---|---|---|---|
| 17 | 0.281 | 0.316 | 50 | **0.787** | 0.384 |
| 90 | **0.676** | -0.277 | 51 | **0.616** | 0.159 |
| 56 | **0.691** | -0.374 | 35 | **0.894** | **-0.626** |
| 57 | **0.802** | 0.084 | 70 | **0.707** | 0.150 |
| 58 | **0.808** | 0.101 | 71 | 0.407 | 0.508 |
| 61 | **0.775** | 0.352 | 74 | **0.629** | 0.247 |
| 33 | **0.630** | -0.035 | 75 | **0.794** | 0.346 |
| 38 | **0.858** | 0.418 | 8 | **0.815** | 0.488 |
| 39 | **0.780** | 0.441 | 12 | **0.913** | 0.489 |
| 42 | **0.765** | 0.424 | 5 | -0.122 | **0.606** |
| 43 | **0.734** | 0.167 | 9 | -0.271 | **0.672** |
| 44 | 0.562 | -0.055 | 10 | 0.267 | 0.446 |
| 45 | **0.697** | 0.337 | 11 | 0.058 | 0.484 |
| 46 | **0.821** | 0.218 | 13 | -0.045 | 0.359 |
| 47 | **0.846** | 0.365 | 14 | -0.135 | 0.593 |
| 49 | 0.531 | 0.314 | 15 | 0.220 | 0.389 |
| 4 | -0.090 | **0.939** | | | |

$r$ values over 0.6 and below -0.6 are printed in bold face. The Year column again displays the year the orchard was planted so that the orchards are presented in order of age. The data is presented in such a fashion to identify possible similarities based on the age of the trees. The correlation thresholds of 0.6 and -0.6 are utilised on the dataset with bunches data included. The features with a correlation coefficient above this threshold are displayed in Table B.8 in Section B.5. Features with a correlation coefficient of -0.6 and lower with the yield are displayed in Table B.9.

### 7.4.2 SelectKBest with bunches

The validation of the Correlation method, done with Scikit-learn's SelectKBest, is presented in this section. The bunches data included as features with the weather features clearly has relevance as a yield predictor for individual orchards.

The number of bunches per tree was only chosen once as the top predictor, for the youngest orchard. This is justified by the immaturity of the trees not having reached their full bearing stage representing their optimum number of bunches.

The features chosen for more than one orchard are displayed with their number of occurrences in Figure 7.11, including the two features containing bunches data. Bunch mass is chosen for 12 of the orchards. In comparison with the weather features, it has much more popularity with this correlation method.

### 7.4.3 Forward stepwise regression to predict yield with weather and bunches data

The bunch data (bunch mass and bunch count) are also added to the list of features to choose from for the forward stepwise regression algorithm. The results indicate that the bunch mass may aid harvest prediction more than the mean number of bunches per tree.

The features selected with forward stepwise regression where bunch data is included in the form of the two features bunch mass and bunch count as possible predictors, are displayed in Table B.10 in Section B.6 in Appendix B.

Table 7.12: SelectKBest method on weather and bunches features, $K = 1$

| Year | Oc | Weather features | $p$ | RMSE |
|---|---|---|---|---|
| 1978 | 17 | May Max Temp | 0.00219 | 9.982 |
| 1989 | 56 | prev May Humid | 0.00053 | 19.482 |
| 1989 | 57 | May Max Temp | 0.00175 | 23.296 |
| 1989 | 58 | Aug Max Temp | 0.00258 | 20.282 |
| 1989 | 61 | May Max Temp | 0.00191 | 22.653 |
| 1978 | 90 | Apr Mean Temp | 0.00642 | 15.987 |
| 1993 | 35 | bunch mass | 0.00020 | 15.054 |
| 1991 | 38 | bunch mass | 0.00072 | 14.403 |
| 1991 | 33 | Aug Heat units | 0.01003 | 12.534 |
| 1991 | 39 | bunch mass | 0.00463 | 26.952 |
| 1991 | 42 | bunch mass | 0.00604 | 25.504 |
| 1991 | 47 | bunch mass | 0.00103 | 24.547 |
| 1991 | 43 | bunch mass | 0.01017 | 23.248 |
| 1991 | 44 | Aug Min Temp | 0.00121 | 17.509 |
| 1991 | 45 | May Max Temp | 0.00894 | 25.544 |
| 1991 | 46 | bunch mass | 0.00193 | 23.218 |
| 1991 | 49 | prev May Humid | 0.00862 | 31.595 |
| 1991 | 50 | bunch mass | 0.00407 | 30.240 |
| 1991 | 51 | prev Jun Max Temp | 0.00094 | 17.604 |
| 1993 | 70 | bunch mass | 0.01495 | 19.249 |
| 1993 | 71 | Dec Wind | 0.00734 | 12.306 |
| 1993 | 74 | Aug Max Temp | 0.01986 | 18.611 |
| 1993 | 75 | bunch mass | 0.00350 | 20.918 |
| 1995 | 8 | bunch mass | 0.00222 | 18.035 |
| 1995 | 12 | bunch mass | 8.88945e-05 | 12.776 |
| 2000 | 5 | Mar Min Temp | 0.00641 | 14.699 |
| 2000 | 9 | prev Oct Mean Temp | 0.00304 | 20.087 |
| 2000 | 10 | prev Jul Max Temp | 0.00382 | 13.845 |
| 2000 | 11 | Aug Rain | 0.00229 | 13.205 |
| 2000 | 13 | Mar Min Temp | 0.00840 | 23.164 |
| 2000 | 14 | Nov Heat units | 0.00225 | 15.649 |
| 2000 | 15 | Jun Max Temp | 0.00786 | 15.277 |
| 2002 | 4 | bunch count | 1.83381e-05 | 9.406 |

The features chosen for four or more orchards are displayed in Figure 7.12 where the bunch mass and bunch count features, although unique to each orchard, are regarded as single features for interpretation and visualisation.

### 7.4.4 Elastic net regression to predict yield from weather and bunch features

The bunches data is subsequently added to the weather features and regularised with elastic net regression to predict the yield of the individual orchards. The results are summarised in Table B.11 in Appendix B. As suggested by the large number of orchards for which the bunch mass is chosen, it is indicative of the yield to be expected. This is a relatively intuitive finding, although bunch mass could be a misleading predictor of total mass per tree, or harvest density, if the bunch quantity is very low. However, it is evident from the frequency of chosen features in Figure 7.13 that the bunch quantity is not as relevant. With the elastic net regression algorithm, bunch count is only chosen for the harvest prediction of four orchards, of which one is Orchard 4, a young orchard with immature palms.

Figure 7.13 shows features chosen for more than three orchards by elastic net regression where

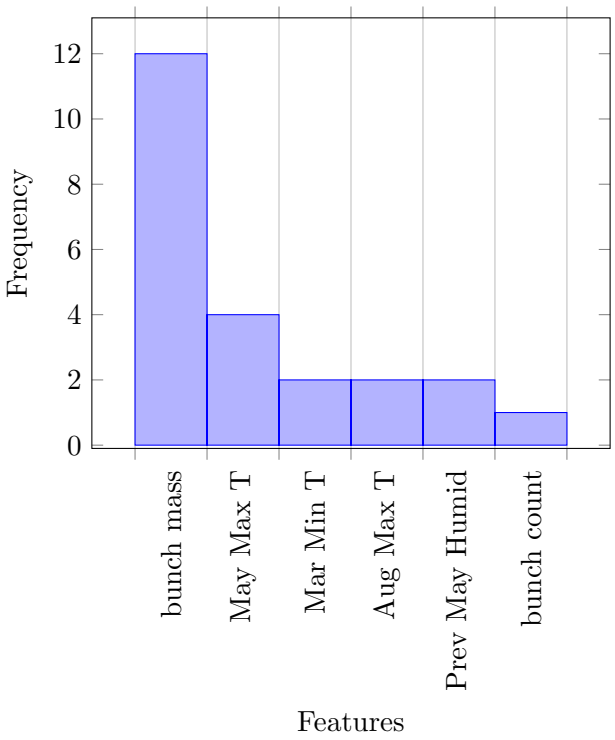



Figure 7.11: Most chosen features from weather and bunches data with SelectKBest method

bunches data is included.

### 7.4.5 Partial least squares regression to predict yield from the weather and bunch data

Because of the nature of the PLS algorithm and following convention, only LOOCV RMSE is reported for this method. The bunches data is applied here as two features, namely the bunch mass and bunch count are included in the list of features as input to PLSR.

The features with a VIP score higher than 2 are chosen as significant for the individual orchards. They are shown in Table B.12 in Appendix B. Orchards not listed do not have any features with a VIP score above 2. The LOOCV RMSE is also displayed.

Figure 7.14 displays features with VIP score greater than 2 for more than two orchards with bunches included. The features bunch mass, which is measured in kg/bunch, and bunch count which is the number of bunches per palm, have a VIP score $> 2$ for 17 and 2 of the orchards respectively.

It is clear from these orchards that the number of bunches does not affect the size of the harvest considerably, but the mass of the average bunch can be directly related to the harvest. This could support the fact that a tree has a certain capacity, and more bunches may only imply smaller bunches. This is confirmed by the correlation coefficient values between the number of bunches per tree and the mass of a bunch, which is very low or negative. For all the regression models in which the bunch data were included, the model improved. The addition of data naturally results in better estimates. The data should, however, be relevant and in the case of bunch data, especially the mean estimated bunch mass of the more mature orchards, it is a relatively accurate predictor of the yield of those orchards.

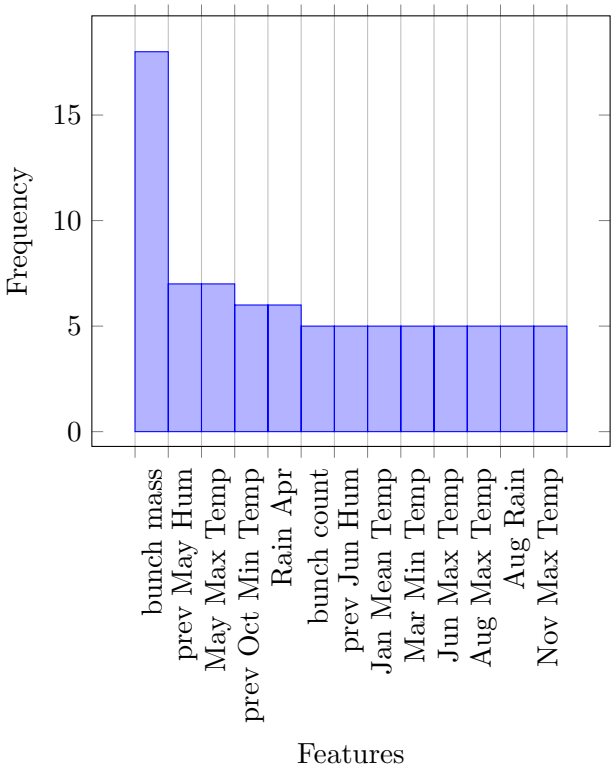



Figure 7.12: Forward stepwise regression features chosen from weather and bunches features for five or more orchards

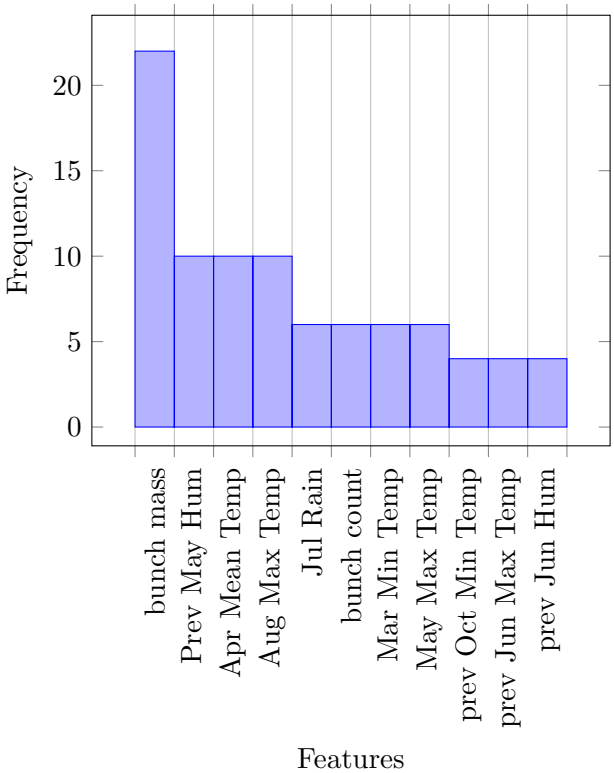



Figure 7.13: Elastic net regression features chosen from weather and bunches data for four or more orchards

## 7.5 Synthesis of the feature selection methods

The implementation of the feature selection methods on the weather and bunch data was done in order to identify most relevant predictors of yield, both at orchard level and farm level. The four methods,

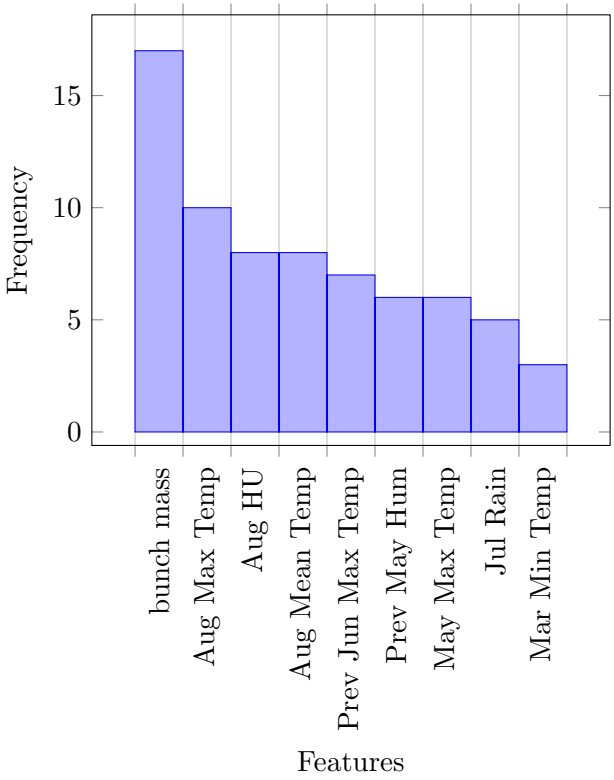



Figure 7.14: Features with VIP score greater than 2 for three or more orchards chosen from weather and bunches data

correlation-based, forward stepwise regression, elastic net regression and partial least squares, highlighted the most significant features from the weather conditions throughout the year influencing the yield. Following was a discussion on these features and the biological explanation from the literature. Next, the focus shifted to the bunch data. First, it introduced the current method of yield prediction on the farm, where the product of the bunch mass and quantity are used. Secondly, the bunch data was investigated as possible predictors of yield using the same four methods. The bunch mass and count were added to the constructed weather dataset as additional features for input to the feature selection methods. The results with these features added showed that by counting and weighing the bunches, the yield can be predicted more accurately. From the results of these methods implemented above, a formal comparison between the current method of the research partner and newly proposed prediction models is subsequently presented.

## 7.6 Comparison of current prediction with linear models from this study

For this section, the researcher distinguishes between the entire yield of the farm as mentioned in the sections above, hereafter called 'farm yield', and the sum of the 33 orchards for which detailed information is available, hereafter referred to as 'orchard total yield'. For the farm yield, a single set of weather features and the average bunch mass and bunch count of the farm are used as independent variables or regressors. For the orchard total yield, a set of weather features for each individual orchard as well as the bunch count and bunch mass for the particular orchard form the regressors. The sum of the individual orchard yield predictions represents the total yield prediction for these orchards.

The yield prediction method currently employed by the research partner, also hereafter referred to as the *current method*, is a simple calculation where for each orchard, the number of bunches and estimated bunch mass are multiplied. The sum of the product of each orchard concludes the orchard

total yield prediction. The root mean squared error of this method on the farm yield, calculated from the squared difference between the predictions and the actual farm yield values of 11 years, is 86.06 t. This represents a mean absolute prediction error of 8.61% on the yield shown in Table 5.3. The RMSE of this method on the orchard total yield, computes to 46.39 t and the MAPE to 7.74%. These predictions and associated errors will be used for comparison with the linear models developed in this section.

Linear regression models, as introduced in Chapter 6, are used as an alternative yield prediction method to the current model and incorporate meteorological conditions to discover significant features. The aim of these models is to improve current prediction by reducing estimation errors.

The prediction models are developed at orchard level, after which the predictions can be summed to obtain the orchard total yield prediction. For the farm yield, the farm averages of the bunch data are used. Prediction models at orchard level are investigated individually, as it could be useful for the research partner to distinguish between orchards with different characteristics such as age, and soil type, recognising how they are influenced differently.

The proposed alternative yield prediction models are developed according to the following steps:

1. Weather features for the linear models of each orchard are selected. All feature selection models, namely correlation, forward stepwise, elastic net and partial least squares, are considered. The features most selected by the methods in Section 7.1 for the particular orchard, are used.
2. The initial model is developed, consisting of all the weather features identified in Step 1 as $\mathbf{X}$ and the yield per tree of the particular orchard as $y$.
3. The $p$-values of the coefficients of these features are evaluated. Where $p$ is too high, generally considered so when above 0.05, the feature is eliminated and the model is redeveloped without it.
4. Step 3 is repeated if high $p$-values persist and $\mathbf{X}$ comprises more than one feature, else if $\mathbf{X}$ contains one feature, it is used.
5. The model is evaluated in terms of RMSE, MAE and MAPE.
6. The bunch features are added as predictors. The two bunch features are also used as $\mathbf{X}$, both separately and together, to evaluate all possible combinations of the weather and bunch features.

The $p$-value for each independent variable tests the null hypothesis that the feature has no correlation with the target. Since the objective is finding significant features, this serves as an alternative criterion to AICc used in forward stepwise and the adaptation of elastic net, and correlation used in the first correlation-based method. With the small dataset available, elastic net had to be adapted to avoid choosing too many features in an attempt to avoid overfitting. For this reason, $p$-values can be useful, especially after previous identification of possible role-playing features is done.

For simplicity, the linear prediction models to be compared with the current multiplication model are named and include the following features:

1. **Model A**: Weather features as identified in the above description.
2. **Model B**: Bunch count is used as the only feature.
3. **Model C**: Bunch mass is the only feature.
4. **Model D**: Both bunch count and bunch mass are used as features. This model has the same variables as the current model, but in a linear regression form.
5. **Model E**: Weather and both bunch features represent $\mathbf{X}$.
6. **Model F**: Weather and bunch mass are the predictors.

### 7.6.1 Comparison of current model and linear models on orchard level

These models A to F are developed for each of the 33 orchards. The rationale for including or eliminating features, as well as some remarks are described. For most relatively mature orchards, the bunch count coefficient $p$ is very high. Although removing the bunch count feature from the model does not always improve the RMSE, the $p$-values for the other coefficients are generally very low and the adjusted $R^2$ higher than with the bunch count included. This indicates that bunch count may not be beneficial for yield prediction.

For some orchards where the coefficients of the bunch features have unacceptably high $p$-values (ranging from 0.4 to 0.8), it is still an improvement to use these features in a linear regression model with a lower RMSE than the current model. It becomes evident that the bunch mass, together with the most prominent weather features from the selection methods for the respective orchard as features in a linear regression model are a promising improvement of the yield prediction model currently implemented by the research partner.

In the case of young trees planted in 2002, being between eight and 18 years of age, the bunch count is an exceedingly more accurate predictor than bunch mass. Since the trees are still growing, the bunch count poses a good indication of the expected yield. In Model F the $p$-value for the bunch mass is 0.61. Using the bunch count along with the two weather features produces a model with a RMSE of 6.94 kg per tree, as compared to Model E with a RMSE of 6.72 kg per tree with high $p$-values. For young trees, it might be worthwhile to count the bunches instead of weighing them.

Considering the orchard-level results in general, using the bunch mass in a linear regression model to predict the yield already produces a better, or at least a similar, result to multiplying the bunch mass and count. Applying bunch mass only would be more cost- and time-effective as bunch counting is very labour-intensive. Another advantage of employing a linear regression model is that merely one representative tree per orchard is required for the mass estimation.

Because of the few predictors included in the final models and the small number of observations, using resubstitution errors to see if the models are feasible is sensible. Resubstitution, as mentioned in Chapter 6, uses all the data for training the model and for calculating errors. Although the errors are more optimistic, they can be compared with the resubstitution errors of the current model and serve the purpose of comparison rather than purely evaluation. For this section, resubstitution errors are reported, mainly for comparison purposes. Low resubstitution errors motivate further model development with more features.

Table 7.13 shows the RMSE values of the models for each of the 33 orchards, where RMSE is not cross-validated for ease of interpretation and because the comparison is being done with an already existing (and implemented) method. The unit of the error value is kg per tree.

Table 7.13: Comparison of models

| **Oc** | **Current Model** | **Linear Models A to F with features:** | | | | | |
|---|---|---|---|---|---|---|---|
| | Bunch count × bunch mass | Weather | Bunch count | Bunch mass | Both bunches features | Weather and bunches | Weather and bunch mass |
| 17 | 13.7 | Max Temp in May and Aug 7.36 | 12.89 | 13.04 | 11.25 | 6.87 | 6.98 |
| 90 | 16.37 | Apr Mean Temp 13.64 | 20.24 | 15.52 | 15.18 | 8.07 | 8.22 |

**Table 7.13 continued from previous page**

| Oc | Current Model | Linear Models A to F with features: | | | | | |
|---|---|---|---|---|---|---|---|
| 8 | 12.46 | Max Temp in May and Aug 18.6 | 22.06 | 14.63 | 10.97 | 7.08 | 12.25 |
| 12 | 11.96 | Oct Hum, Min Temp May, Jul Rain 13.6 | 23.94 | 11.22 | 10.88 | 3.83 | 4.25 |
| 56 | 26.4 | prev May Hum, prev Jun Max Temp 11.41 | 26.46 | 20.64 | 20.58 | 7.37 | 7.46 |
| 57 | 23.81 | Max Temp in May 18.9 | 33.36 | 20.02 | 19.05 | 9.73 | 13.43 |
| 58 | 24.88 | Max Temp in May and Aug 13.95 | 26.18 | 15.5 | 15.35 | 7.04 | 7.29 |
| 61 | 29.42 | Max Temp in May 18.23 | 29.96 | 20.24 | 15.61 | 8.23 | 13.95 |
| 33 | 40.13 | Aug Heat units 10.61 | 15.63 | 12.15 | 12.06 | 8.83 | 8.99 |
| 38 | 13.58 | Mean Temp in Jul 20.43 | 22.57 | 12.76 | 11.78 | 10.24 | 10.3 |
| 39 | 18.16 | prev May Humid 19.68 | 26.23 | 18.29 | 17.59 | 12.3 | 13.03 |
| 42 | 16.2 | prev May Humid 18.19 | 25.24 | 17.94 | 15.5 | 11.13 | 12.34 |
| 43 | 19.12 | Aug Max Temp 20.02 | 28.34 | 19.53 | 18.51 | 13.54 | 14.43 |
| 44 | 19.59 | Aug Min Temp 13.32 | 24.5 | 20.3 | 19.58 | 9.7 | 10.3 |
| 45 | 22 | Max Temp in May 19.95 | 28.01 | 21.32 | 20.39 | 16.86 | 17.29 |
| 46 | 16.42 | prev Jun Max Temp 20.8 | 27.44 | 13.62 | 13.06 | 9.85 | 12.48 |
| 47 | 16.35 | Jul Rain 20.96 | 29.66 | 16.99 | 16.55 | 13.24 | 13.43 |
| 49 | 22.63 | prev Jun Max Temp 23.37 | 33.07 | 29.51 | 20.98 | 8.79 | 21.83 |
| 50 | 15.14 | prev Jun Max Temp 23.02 | 28.4 | 18.99 | 14.82 | 9.72 | 15.34 |
| 51 | 17.16 | prev Jun Max Temp 14.6 | 27.29 | 21.78 | 17.35 | 8.18 | 12.36 |
| 35 | 10.62 | prev Oct Min Temp, May Max Temp, prev Jun Max Temp 17.85 | 27.03 | 13.02 | 10.57 | 10.82 | 11.22 |
| 70 | 15.26 | prev Oct Min Temp, prev Jun Max Temp, May Max Temp 10 | 20.5 | 14.66 | 14.37 | 9.92 | 10 |
| 71 | | Mar Min Temp, Dec Wind | | | | | |

**Table 7.13 continued from previous page**

| Oc | Current Model | Linear Models A to F with features: | | | | | |
|---|---|---|---|---|---|---|---|
| | 55.4 | 7.83 | 14.29 | 15.15 | 11.89 | 3.95 | 4.77 |
| 74 | 16.82 | Aug Max Temp<br>14.72 | 19.59 | 15.71 | 11.97 | 10.02 | 12.26 |
| 75 | 15.07 | prev Oct Min Temp, May Max Temp<br>10.64 | 23.58 | 15.28 | 14.27 | 6.16 | 8.39 |
| 5 | 13.99 | Mar Min Temp<br>11.99 | 14.73 | 18.37 | 13.26 | 10.85 | 11.39 |
| 9 | 19.23 | prev Oct Humid, prev Aug Min Temp<br>10.72 | 18.74 | 24.35 | 18.1 | 10.37 | 10.63 |
| 10 | 14.5 | prev Jan and prev Jul Max Temp<br>10.7 | 16.84 | 18.13 | 11.47 | 9.84 | 10.67 |
| 11 | 15.98 | Rain in May and Aug<br>6.67 | 16.44 | 18.76 | 13.04 | 5.51 | 5.7 |
| 13 | 26.24 | prev Oct HU, Mar Min Temp, Jul Min Temp<br>5.18 | 25.35 | 27.13 | 24.59 | 5.96 | 5.16 |
| 14 | 19.75 | Nov Heat units<br>13.32 | 18.49 | 22.76 | 17.32 | 12.92 | 13.32 |
| 15 | 14.68 | May HU, Sep Mean Temp, May Hum<br>6.14 | 17.02 | 18.03 | 12.54 | 5.38 | 5.39 |
| 4 | 10.04 | Jan Wind, Oct Mean Temp<br>10.19 | 7.67 | 22.22 | 7.06 | 6.72 | 9.99 |

For Model E and F of all of the 33 orchards, the equations are in Tables B.14 and B.15 in Section B.9.

Table B.13 in Section B.9 displays the cross-validated errors for these orchards. Although the values are higher than the resubstitution errors, as expected from cross-validation, they are still reasonably low considering the small sample size. Cross-validation is not further discussed in this subsection, as the focus is on the comparison with the research partner's method, which is not an estimator and as a result cannot be cross-validated.

Considering the trade-off between prediction accuracy and saving labour cost and time, the linear model, using weather and bunch mass, produces acceptable results – without the effort of bunch counting and improving on the current model. The bunch count as a feature also has a high $p$-value, indicating that it is not a significant feature. The mean $p$-values for the linear models of all the orchards containing all the features (Model E) are 0.39 for the bunch count and 0.22 for the bunch mass. The RMS errors when including and excluding bunch count, are shown for the individual orchards in kg per tree in Table 7.13. Furthermore, the orchard total yield values and the farm yield are indicated with MAPE. Individual orchards do require a unique weather feature set as input to their prediction models.

The orchard total yield, found by summing the values of all 33 orchards, as well as the actual observations, predictions with the current model, and predictions with the new model **F** – using the

respective weather features of the orchard and the bunch mass – are presented in Table 7.14. Predictions resulting from including bunch count (model E) are also presented. The absolute percentage errors (APE) for each year's observation $i$ is calculated as $\frac{|y_i - \hat{y}_i|}{y_i} \times 100\%$.

Table 7.14: Summation of the orchard total yield in kg and respective absolute percentage errors

| Year | Actual orchard total | Current predict | Current APE (%) | Predict with Model F | Model F APE (%) | Predict with Model E | Model E APE (%) |
|---|---|---|---|---|---|---|---|
| 2010 | 672 738 | 615 724 | 8.47 | 650 170 | 3.35 | 657 265 | 2.30 |
| 2011 | 673 829 | 690 741 | 2.51 | 684 032 | 1.51 | 679 583 | 0.85 |
| 2012 | 535 033 | 528 695 | 1.18 | 534 921 | 0.0002 | 530 182 | 0.91 |
| 2013 | 498 088 | 474 360 | 4.76 | 488 211 | 1.98 | 483927 | 2.84 |
| 2014 | 438 418 | 467 205 | 6.57 | 469 481 | 7.09 | 453 807 | 3.51 |
| 2015 | 466 126 | 496 056 | 6.42 | 470 355 | 0.91 | 460 774 | 1.15 |
| 2016 | 478 968 | 535 200 | 11.74 | 505 955 | 5.63 | 507 594 | 5.98 |
| 2017 | 554 679 | 509 831 | 8.09 | 543 693 | 1.98 | 544 331 | 1.87 |
| 2018 | 484 885 | 505 373 | 4.23 | 493 189 | 1.71 | 484 996 | 0.0002 |
| 2019 | 407 279 | 513 545 | 26.09 | 417 688 | 2.56 | 437 321 | 7.38 |
| 2020 | 602 635 | 572 271 | 5.04 | 554 981 | 7.91 | 572 894 | 4.94 |

Calculating the RMSE and MAPE on the orchard totals, produces the error values and percentages presented in Table 7.15.

Table 7.15: Prediction errors on orchard totals for current model, Model E (weather and both bunch mass and count) and Model F (weather and bunch mass)

| Current model | | Model F | | Model E | |
|---|---|---|---|---|---|
| RMSE | MAPE | RMSE | MAPE | RMSE | MAPE |
| 46 389 kg | 7.74% | 21 301 kg | 3.15% | 17 775 kg | 2.89% |

Figure 7.15 displays the observed values, predictions with the current model and predictions with Models E (weather and both bunch features) and F (weather and bunch mass) for individual orchards and summed to obtain the orchard total yield. As seen, the difference in prediction between Models E and F is relatively small and Model F sometimes outperforms E, even though fewer features are included.

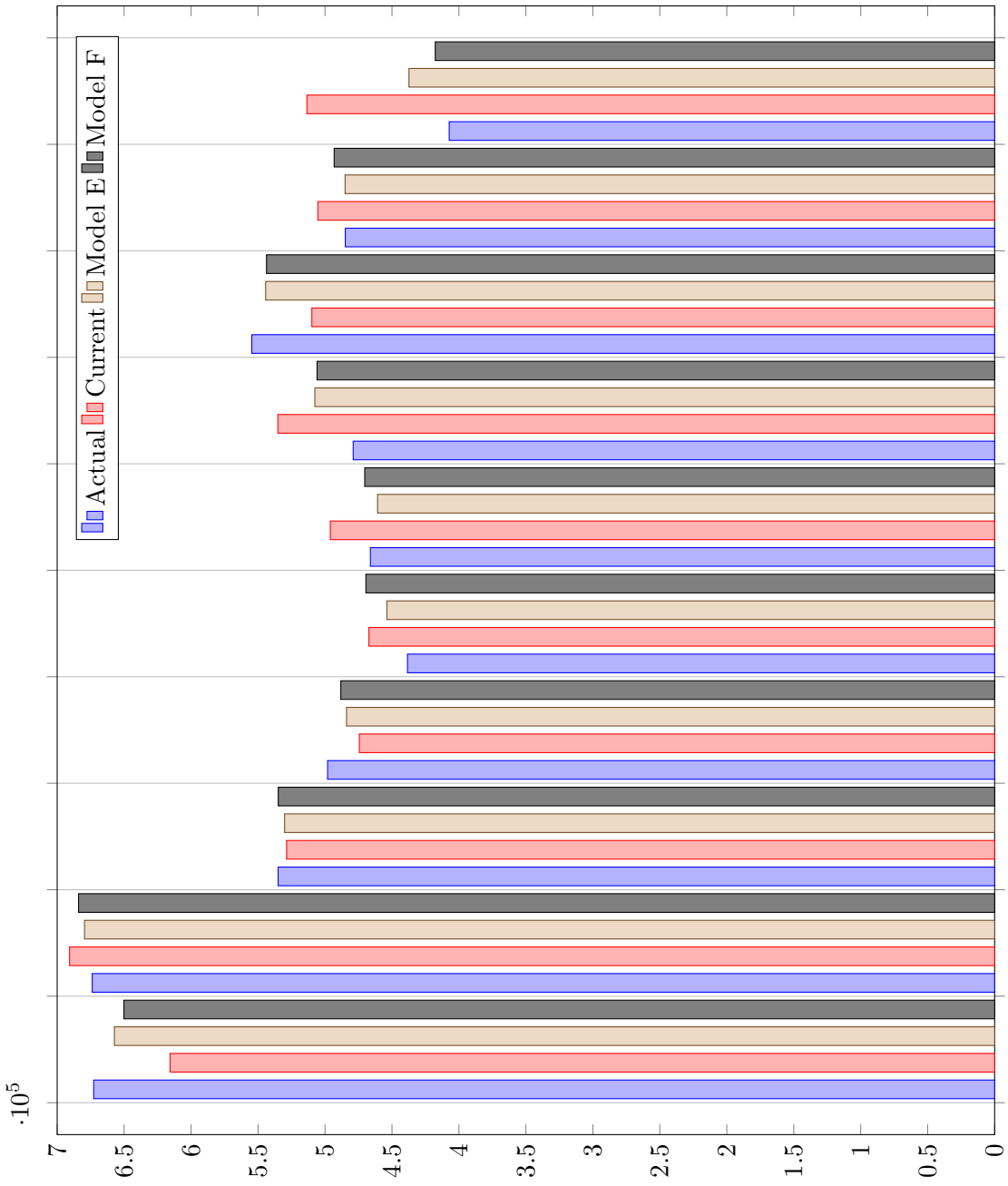



Figure 7.15: Actual and predicted values for the orchard total

### 7.6 Comparison of current prediction with linear models from this study

The mean absolute errors (MAE) in tonnes for the current and newly developed linear models are shown in Figure 7.16. Model D (both bunch mass and count) is displayed specifically for comparison with the current model. This indicates that using the same features of the current model in the linear model is already an improvement. Figure 7.17 shows the observed vs predicted values in kg and the

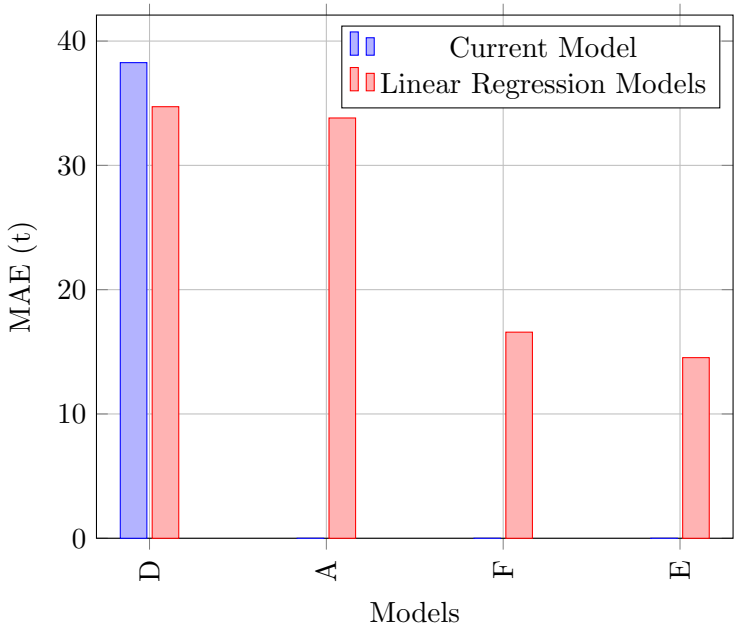



Figure 7.16: Comparison of absolute errors of models predicting orchard total yield

dotted line indicating ideal positions. As Model E uses more features, both weather features for the individual orchards and bunch features, the predictions are indeed more accurate. Comparable results are achieved by Model F, in which the bunch count feature is omitted. For this reason, Model F is preferred.

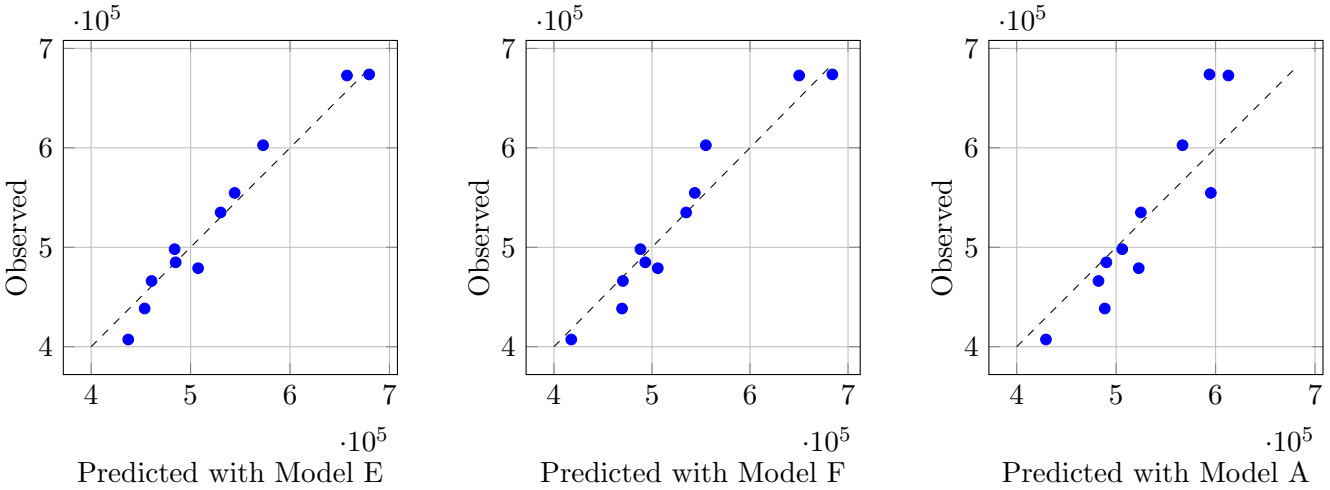



Figure 7.17: Observed values vs predicted values in kg for linear models E (predicted with weather and bunch features), F (predicted with weather and bunch mass) and A (predicted with weather features) predicting orchard total yield

This subsection discussed the prediction of the orchard total yield and calculated the RMSE, MAE and MAPE of the Models A to F, which incorporate weather and bunch features in various combinations. They were compared with the current model of the research partner and it was found that the models can predict yield better than the current method used.

### 7.6.2 Comparison of current method and linear models on farm level

The same models A to F are developed for the farm yield. The features prev Oct Mean Temp and Sep Mean Temp are identified as the most significant weather features via the same approach used for the individual orchards: the weather features chosen by most of the feature selection methods are evaluated on their coefficient $p$-values to eliminate less significant features. The bunch features are calculated as the mean bunch mass and mean bunch count of the farm.

Table 7.16: Comparison of RMSE (in kg) of current and linear models predicting the farm yield

| **Current model** | **Linear Models A to F with features:** | | | | | |
|---|---|---|---|---|---|---|
| Bunch count × bunch mass | A: Weather | B | C | D | E: All features | F |
| | prev Oct Mean Temp, Sep Mean Temp | Bunch count | Bunch mass | Both bunch features | Weather and bunches | Weather and bunch mass |
| 86 055.7 | 53 966.3 | 124 528.5 | 128 002.6 | 88 083.8 | 45 466.9 | 47 276.5 |

Table 7.16 displays the RMSE in kg of the models A to F for the farm yield.

The farm yield actual observations, predictions with the current model, and predictions with the new model **F** – using the respective weather features of the orchard and the bunch mass – are presented in Table 7.17. Predictions resulting from including bunch count (model E) are also presented. The absolute percentage errors (APE) for each year's observation $i$ is calculated as $\frac{|y_i - \hat{y}_i|}{y_i} \times 100\%$.

Table 7.17: Farm yield in kg and respective absolute percentage errors

| **Year** | **Actual farm yield** | **Current predict** | **Current APE (%)** | **Predict with Model F** | **Model F APE (%)** | **Predict with Model E** | **Model E APE (%)** |
|---|---|---|---|---|---|---|---|
| 2010 | 1 010 588 | 961 726 | 4.83 | 1 030 201 | 1.94 | 1 022 802 | 1.21 |
| 2011 | 1 098 823 | 1 130 048 | 2.84 | 1 057 739 | 3.74 | 1 068 633 | 2.75 |
| 2012 | 802 648 | 830 039 | 3.41 | 884 394 | 10.18 | 877 449 | 9.32 |
| 2013 | 870 131 | 900 539 | 3.49 | 812 762 | 6.59 | 809 041 | 7.02 |
| 2014 | 812 083 | 922 699 | 13.62 | 800 562 | 1.42 | 786 685 | 3.13 |
| 2015 | 943 248 | 917 761 | 2.70 | 930 583 | 1.34 | 923 770 | 2.07 |
| 2016 | 898 332 | 1 054 969 | 17.44 | 977 542 | 8.82 | 986 607 | 9.83 |
| 2017 | 1 049 359 | 1 024 379 | 2.38 | 1 086 682 | 3.56 | 1 080 637 | 2.98 |
| 2018 | 993 454 | 1 039 021 | 4.59 | 993 060 | 0.04 | 979 748 | 1.38 |
| 2019 | 862 256 | 942 986 | 9.36 | 821 145 | 4.77 | 853 410 | 1.03 |
| 2020 | 1 267 714 | 1 440 299 | 13.61 | 1 213 966 | 4.24 | 1 219 855 | 3.78 |

Figure 7.18 displays the observed farm yield, the prediction of the yield of the entire farm by the research partner and the Model E (all features) and F (prev Oct Mean Temp, Sep Mean Temp and average bunch mass). Models E and F both outperform the current model.

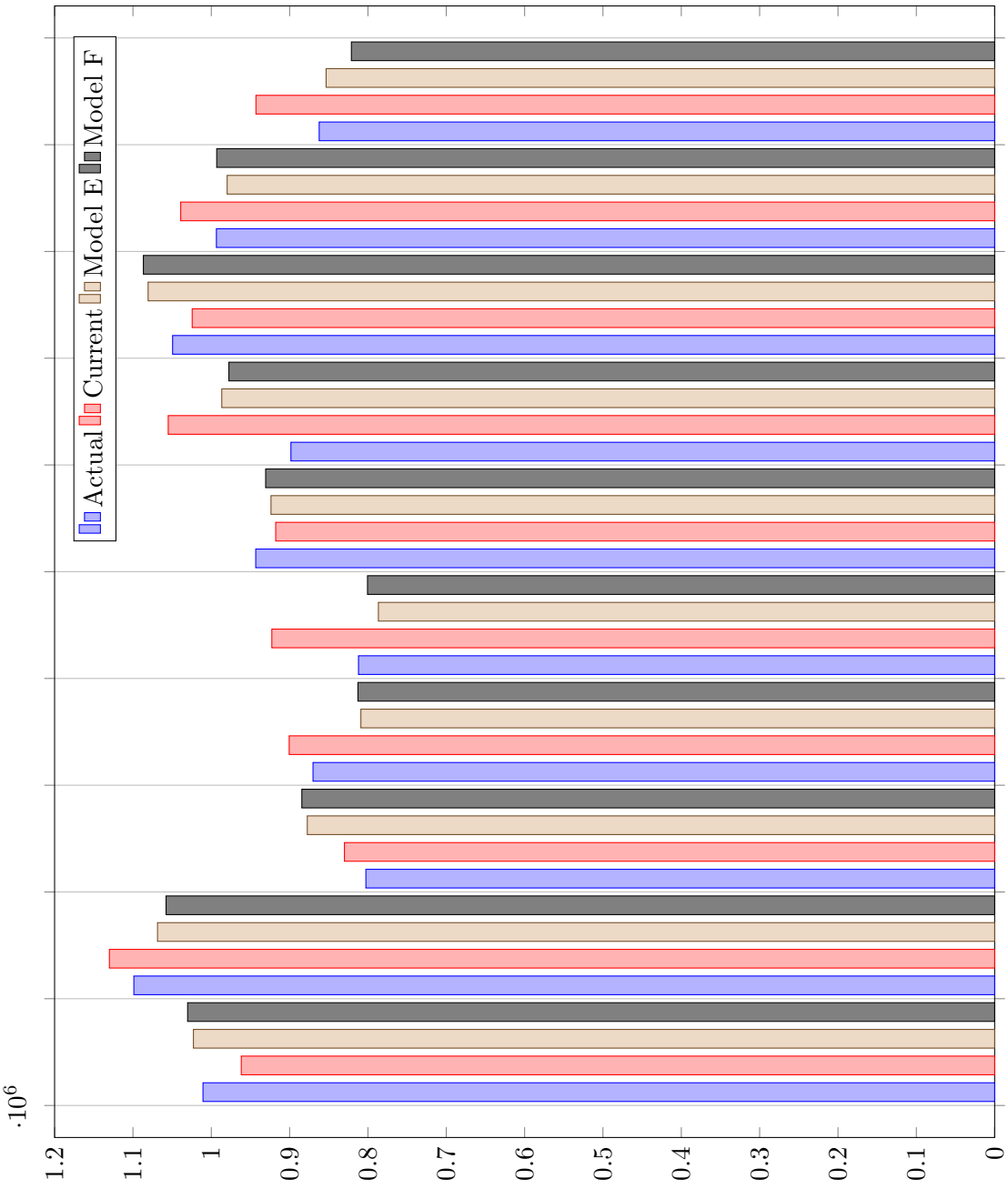



Figure 7.18: Actual and predicted values for the farm total

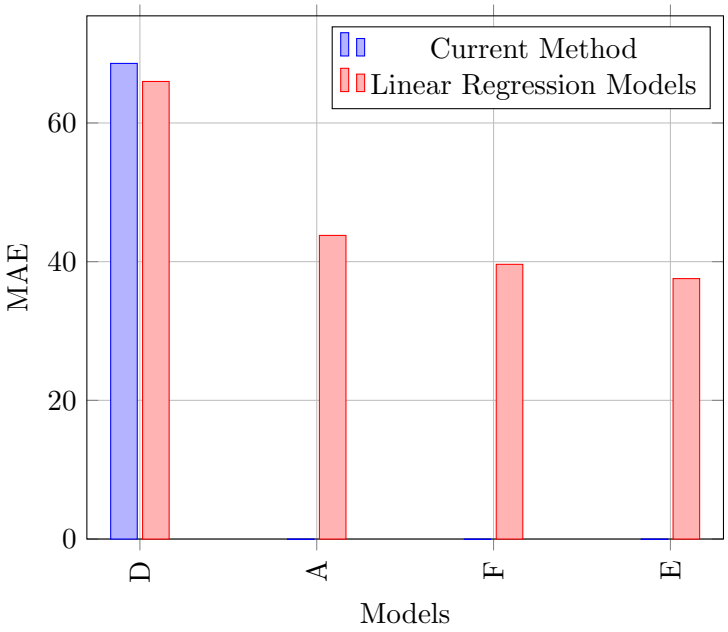



Figure 7.19: Comparison of absolute errors of models predicting farm yield

Figure 7.19 displays the MAE in tonnes of the farm yield predictions by the current method, Models D (both bunch features also used by the current model), A (Mean Temp in prev Oct and Mean Temp in Sep), F (weather and bunch mass) and E (all features, weather and both bunch count and bunch mass). The blue bar represents the error of the current model for comparison. Model D uses the same features, the average of all the bunch mass estimates and the average bunch count. Using these two features in a linear regression model produces a slightly smaller MAE than the current model, while prev Oct Mean Temp and Sep Mean Temp as features leads to a much-improved error. The accessible weather features can contribute fundamentally to the prediction accuracy. This can be further improved by adding the bunch features as predictors to the linear model. A MAPE of 4.5% is produced, an improvement on the original 7.12% of the current multiplication model. When cross-validating the models, the errors would be higher. The LOOCV RMSE for the linear model predictions with the two weather features is 82 272 kg. Although, in this section the focus is on the comparison of the combinations of features to use in a linear model with the original multiplication method implemented by the research partner.

The observed versus predicted graphs in Figure 7.20 show the proximity of the predicted values to the ideal line. A scatter plot with more accurate points closer to the dotted line would indicate overfitting for such a small sample size, which refers to the importance of parsimony of models in this study. The results of Model F are comparable with Model E, as is the case with the prediction of the orchard total.

Considering significant features in terms of $p$-values and with the focus on the yield of the entire farm, the mean bunch count has a $p$-value of 0.51 while the mean bunch mass has a $p$ of 0.21. While bunch mass is also not significant, bunch count clearly does not prove to be a significant feature. The decrease in error by including bunch count as a feature is smaller than the improvement brought about by using model F compared to the current model. It is therefore not recommended that bunch count is used, but bunch mass and the individual orchard features used for predictions on orchard level.

A major consideration is whether a single model for the entire farm can be used, *i.e.*, a single set of features: prev Oct Mean Temp, Sep Mean Temp, and mean bunch mass, or whether orchard-level models with individualised features should be used and the outputs summed. A small decrease in error occurs when using orchard-level predictions.

Equations of both Model E and F for the individual orchards can be found in Section B.9 in Appendix B to be used for the prediction of orchard yield per tree, which can be multiplied by the

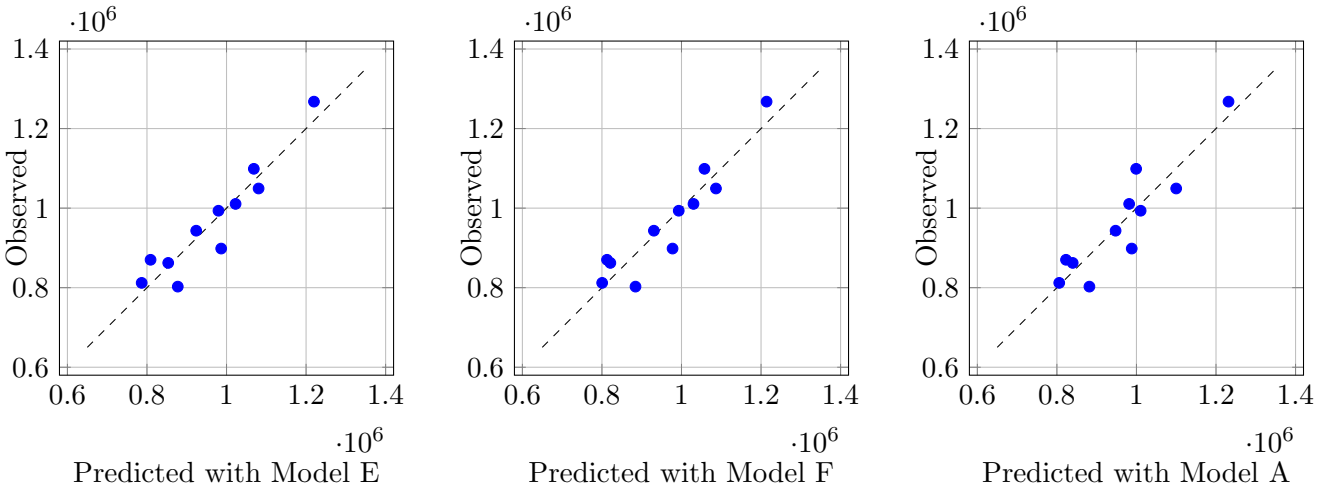



Figure 7.20: Observed values vs predicted values for linear models predicting farm yield with weather features prev Oct Mean Temp and Sep Mean Temp and bunch features mean bunch mass and mean bunch count. Model A makes predictions only using weather features. Model E predicts with both weather and bunch features. Model F uses weather features and the bunch mass feature.

number of trees of the particular orchard and summed for the total orchard yield.

For the farm yield, the equation for model E is

$$55\,199.1 prevMeanT_{10} + 20\,520.8 MeanT_9 + 36\,362.0 bmass + 22\,252.6 bcount - 1\,378\,041.3 \quad (7.2)$$

and the equation for model F is

$$64\,189.1 prevMeanT_{10} + 21\,063.5 MeanT_9 + 22\,200.1 bmass - 1\,177\,225.6. \quad (7.3)$$

Calculating the RMSE and MAPE on the farm yield, produces the error values and percentages presented in Table 7.18.

Table 7.18: Prediction errors on the farm yield for Models E (weather and both bunches) and F (weather and bunch mass)

| **Current model** | | **Model F** | | **Model E** | |
|---|---|---|---|---|---|
| RMSE | MAPE | RMSE | MAPE | RMSE | MAPE |
| 86 055.7 kg | 7.12% | 47 276.5 kg | 4.24% | 45 466.9 kg | 4.04% |

As seen from Table 7.18, as compared with Table 7.15, the errors of the summation of orchard-level predictions are smaller than the total farm yield prediction errors. Both methods improve on the current model. Instead of the summation of the predictions of Model F in Section B.9, a single calculation by Equation 7.3 can be used for the total farm yield prediction at the cost of 1.09% MAPE. This subsection presented the developed models A to F to predict total farm yield, calculated the RMSE, MAE and MAPE of these models A to F, and compared them with the current model of the research partner.

The current method does not calculate any estimates, nor does it consider any observations prior to the current observation. For this reason, cross-validation errors cannot be calculated. The resubstitution errors are therefore used and compared with the newly developed methods' resubstitution errors. The leave-one-out cross-validated errors are calculated for the linear models predicting the farm yield, and though they cannot be compared with the current method's resubstitution errors, they indicate more realistically the performance of the model.

Model A on the farm yield:

- Resubstitution RMSE: 53 966 kg
- LOOCV RMSE: 82 272 kg

Model E on the farm yield:

- RMSE: 45 467 kg
- LOOCV RMSE: 80 693 kg

Model F on the farm yield:

- RMSE: 47 276 kg
- LOOCV RMSE: 83 086 kg

High LOOCV errors and low resubstitution errors confirm that the sample size is too small for definitive conclusions, but the results are still useful. Considering Model E, a LOOCV MAE of 71 077 kg and LOOCV RMSE of 80 693 kg (or 80.7 t) could still be regarded as favourable, or at least promising, for a yield with a mean actual value of 964 421 kg, or 964.4 t.

## 7.7 Chapter summary

This chapter laid out the results from implementing the selected prediction methods and presented a conclusive comparison between the research partner's current method of prediction, and the developed models in this study. First, the constructed weather features were used to predict the harvest of individual orchards as well as of the farm as a whole. The feature selection methods were used to identify the most significant features from the weather conditions influencing the yield. From these features linear regression models were developed for the comparison with the yield prediction method employed by the research partner.

Comparing the predicted yield values with the observed yield values as well as the current predictions of the research partner showed that by either using the orchard predictions individually and summing these predictions for a total, or using mean values to predict the total farm yield, the errors measured as RMSE, MAE and MAPE, are lower than the current prediction model. The investigation also revealed that using weather features in addition to the bunch mass and bunch count produces more accurate predictions.

Using the summation of orchard-level predictions, where the best prediction of each orchard's yield is done with a unique subset of weather and bunch features, produces a MAPE of 2.89% (or 3.15% using only weather and bunch mass), improving on the research partner's method MAPE of 7.74%.

Considering the total farm yield and using a single set of weather features and the mean values of the bunch features, the MAPE is 4.04% (or 4.24% without using bunch count), in comparison with the research partner's MAPE of 7.12%.

Although the prediction errors are slightly larger, it is recommended to use the bunch mass and the weather features as these proved to be most significant. Proper validation could not be done with the small sample but is advised for future implementation.

The following and final chapter serves the purpose of summarising the project and integrating the real-world context with the statistical results, reflecting on the discoveries and considering suggestions for future research.

# Chapter 8

# Conclusion

This chapter concludes the project by presenting an overview of the findings and makes recommendations for the agricultural practitioner. Suggestions for further research are made and finally, the project is summarised.

Crop yield prediction plays a vital role in food production globally and agricultural practitioners rely on yield predictions to make informed management and financial decisions. Accurate yield prediction also supports food import and export decisions for increased food security. The accurate prediction of crop yield poses numerous challenges and is extremely complex. The genotype information of the crop, and its interaction with environmental factors as well as management practices on the farm, are some of the factors contributing to this complexity.

The field of crop yield prediction is well established for annual crops, but not as widely developed for perennials. This project researched a yield prediction approach on the date palm using specific mathematical tools and data collected by the research partner.

## 8.1 Research findings

The onset of this project was motivated by a belief by both the researcher and the research partner that it was possible to predict the yield of the date crop by growth data, consisting of measurements of the fruit dimension, combined with machine learning. Investigation, however, proved that the growth data was insufficient and non-representative. Therefore, the investigation was redirected towards applying regression on other available data types, and it was found that weather data and fruit bunch data are of significance in date yield prediction. Several modelling techniques to predict date yield were investigated and it became evident that regression was the most suited technique because of the small datasets available.

The researcher had limited yield data available consisting of records for 11 years. It was rational to consider the effects that features of date production have on the date yield, hence the linear regression approach. Several models were constructed using different combinations of features to finally isolate the most prominent features. The methods used in the study are forms of linear regression, as the small sample size would not allow more complexity, and all methods are applicable to high-dimensional datasets. The four regression approaches considered, were:

1. correlation-based method, validated with the SelectKBest method.
2. forward stepwise regression.
3. elastic net regression.
4. partial least squares regression.

First, the Pearson correlation coefficient was calculated for each feature to gather information on the strength of the relationships and their effect on the yield. It was validated with the SelectKBest method, also utilising correlation. For this first method the features were ranked according to the strength of the correlations. The researcher suggests a threshold of -0.7 and +0.7. The research partner can, however, consider different thresholds of correlation coefficient values. Forward stepwise regression normally is applied because of its ability to handle a large number of features. It is useful in comparing possible combinations of features, although it may be replaced with modern methods such as regularisation. Regularisation, specifically elastic net regression (a combination of ridge and lasso regression), was implemented. Its characteristic as a penalty-based linear regression grants it the ability to reduce the number of included features. Lastly, a dimension reduction method, namely

partial least squares, was employed to transform the input data and identify the features regarded most important for yield prediction.

The results are in favour of the newly proposed regression models using weather information as well as bunch mass, as these are significant features producing lower prediction errors. The approach can be applied to individual orchards or the entire farm, where unique sets of features are identified for each orchard. The farm prediction can be made with four features, namely the mean temperature in September and the previous October, the mean bunch count and bunch mass.

The farm consists of many date palm orchards of the same cultivar but differing in age and size. The current method employed by the research partner to predict yield on the farm is to calculate the sum of the products of the bunch features, namely bunch count and bunch mass, of each orchard. These predictions have a mean absolute percentage error of 7% when compared to the actual yield. By using the identified weather and bunch features from the four methods in a linear model, the predictions were improved to a mean absolute percentage error of 4%. Primarily, it was found that using the same bunch data as used by the current prediction method in a linear regression model, produces more accurate results, both on orchard level and for the entire yield. Secondly, it was shown that certain weather factors could further improve the prediction model. Finally, it is concluded that for most orchards, as well as the entire yield estimation, the prediction models require some weather data and the estimated bunch mass. The counting of the bunches is not a requirement for more accurate yield prediction, consequently all the intensive labour and costs associated with bunch counting can be saved.

## 8.2 Limitations of the study

In this project investigating an agricultural problem of date palm yield prediction, a primary focus and challenge was to explore possible methods able to handle a small sample size. A small sample is particularly relevant in agricultural problems, where the number of role-playing factors is normally large. Factors influencing date fruit development and eventually yield, vary from environmental to artificial, from the wind speed at the time of pollination to the amount of fertiliser applied. Using only the data available for this study, the secondary challenge was how to best aggregate the data, or present it in a summarised format, to construct a dataset as input for predictive modelling. With the newly constructed set, the problem persisted (high-dimensional), where the number of features exceeded the number of observations.

Various machine learning algorithms were attempted with a train and test set, but the small sample size did not warrant either splitting the data or implementing an approach more complex than multiple linear regression. The size of the test set usually presents around 25% of the dataset, which in the case of such a small dataset would be a meagre two or three data points. For that reason, it was more sensible to build a predictive model by training on all the data.

Certain important practices on the farm were not included in the study due to absence of data. These include the degree of thinning, time of pollination, amount, time and frequency of fertiliser application, and the amount of irrigation. Data on these practices could add more meaningful features to the proposed regression models and improve the prediction accuracy.

Because of the small sample size, this project took the form of a feasibility study to determine if it would be beneficial to predict date yield with the considered factors and whether an improved model is possible.

## 8.3 Contributions of the research

The date palm is not extensively cultivated in Southern Africa, and literature on the topic, especially on date yield prediction, is limited. For this reason, and the application of methods on a very small-sized but high-dimensional dataset, this study is relevant to agricultural practitioners, date consumers and research regarding small sample size and high-dimensional problems. It is specifically of value to the research partner, who can replace the current method of yield prediction by applying the research recommendations. The task of yield prediction is widely addressed in agricultural practice and research, as discussed in Chapter 3, with the use of mechanistic models and horticultural knowledge, considering an extensive variety of factors, ranging from management practices to soil profiles. This study considered the weather as the main factor, mainly due to unavailability of data on the other factors. An intention of the study was not to provide an absolute predictive model but to initiate research in the field of date yield prediction in South Africa. It is confirmed that traditional machine learning approaches require more data than one farm can typically measure during its years of production, in the case where it takes one year to gather enough data for one observation. However, it was verified that correlation is a useful statistic for gaining insight into the effects of the features on the yield. This study has shown that traditional approaches would require more data, but already indicated important factors in the determination of the yield of a season.

A major contribution to the study of date palm yield, is the confirmation of the definite influence of the weather conditions, particularly temperature and humidity, not only of the year leading up to a harvest but also of the previous year, 12 to 24 months prior to the harvest. SMEs of the research partner have speculated for some time that the harvest of a particular year is affected by factors of the past two years, but they do not know which and to what extent. Temperature affects the yield more than wind or rain, particularly in this region in South Africa.

The study confirmed that using weather and bunch data in a linear predictive model is feasible and the outcome warrants the effort of more data gathering. It was also confirmed that linear predictive modelling improved the current method of the research partner and that certain current measurements (*i.e.* bunch count) is deemed insignificant and can be omitted.

The researcher concludes that this study essentially provided a proof-of-concept that it is possible to improve date yield prediction, albeit with limited data.

## 8.4 Recommendations for future work

A number of recommendations for further work are deduced from the results, and these are presented next.

### 8.4.1 Recommendation for producers

From the research findings, it is recommended that the research partner use the mean temperature in the previous October and current September, as well as the bunch mass for the farm-level date yield prediction for the following year. For agricultural practitioners, meticulous documentation of business practices is recommended. Practical, reliable means of measuring and recording should be established and maintained. Although it was learnt that the individual fruit size estimated for an orchard is not useful in predicting the total harvest mass, it may be useful for the farmer to gather insight on the current stage of the fruit, and support harvesting and marketing decisions. The agricultural practitioner can build tolerance by incorporating cultivation of other crops – known as intercropping – for 'off-years', as the date palm is subject to alternate bearing (where the natural production of the palm fluctuates in alternate years, if not regulated by thinning practices).

### 8.4.2 Future work

There are various other factors, besides the meteorological conditions, that contribute to the yield. From the literature study on crop models, and specifically research on date palms, the three most influential are thinning, fertilisation and irrigation (Djerbi, 1995; Nixon, 1956). For future work, the study needs to be expanded to consider these human interventions and agricultural practices, especially the extensive application of thinning. Irrigation practices, the soil modal profiles of the various orchards and fertilisation regimes must also be taken into consideration. Data on these factors may be more difficult to obtain but will sketch a more accurate picture. This would require increased stakeholder engagement.

Meteorological data is used as the main predictor of harvest density in this study, although many other factors should be considered and the weather data itself should be expanded. More types of measurement, including radiation and evapotranspiration, that were not available for all the years under study, can be included. For more reliable predictions, more data is required, both in terms of the number of observations and the variety of measured factors. The way the target variable was set up where one observation represented the harvested mass of one season; more observations equate to more years. The models developed in this work could be further improved by validating the resultant predictions made with data from the 2021 and successive harvests. The results can be compared with an existing crop model such as those discussed in Chapter 3, other than the method employed by the research partner. The crop model must, however, be suitable for perennials and specifically date palms, and the research shows that these models are not abundant.

For future work, the final model development (after implementation of the feature selection methods) could be done with best subset selection instead of using $p$-values for determining feature significance. The motivation for using $p$-values in this study is because the features were already filtered and chosen by the four types of regression models serving the purpose of feature selection. However, best subset using $R^2$ might be more reliable.

## 8.5 Project summary

This project focused on the prediction of date palm yield from meteorological and other input data as obtained from the research partner, to identify role-playing factors from the data and to improve the current prediction method used on the research partner's farm. These goals were indeed achieved, and the newly developed regression models improve the yield predictions in the order of three percentage points, which, for a large date farm equates to a 40 ton error improvement on an annual 1 000 ton yield. The proposed models also suggest that similar or better yield predictions can be achieved while eliminating current labour-intensive data collection practices.

Chapter 1 introduced the problem and its background. It was further expanded in Chapter 2 which looked specifically at the date palm and its cultivation. The chapter examined the cultivation of dates with a focus on factors influencing yield. Finally, the global date market was explored. Chapter 3 reported on existing crop and yield prediction models used in practice and in research. These models were presented in two categories: process, or mechanistic, models and statistical models. The focus was shifted to the field of date palm research, particularly date palm yield. Chapter 4 discussed the use of data in modelling, with an emphasis on data analysis and understanding. Finally, the yield-influencing factors learnt in Chapter 2 were transformed into important data features, and a discussion of possible required datasets was presented. Chapter 5 introduced the available real-world datasets used in this project and began exploring their suitability, discussing those which would be proceeded with and those which did not provide enough promising relationships. In Chapter 6 the topic of a required sample size was introduced and detailed. The chapter also discussed predictive modelling theory and reviewed methods for feature selection which were used in Chapter 7 to incorporate the data from Chapter 4 and to propose yield prediction models. The resulting models were compared with the yield prediction method of the research partner, both on individual orchards and considering

the entire farm, and it was found that the linear models do improve the prediction accuracy. Finally, the conclusion in Chapter 8 summarised the project. Research findings and results were discussed and recommendations for further study were made.

# References

Abdelhadi, A. W., Salih, A., Sultan, K., Alsafi, A. and Tashtoush, F. (2020), 'Actual water use of young date palm trees as affected by aminolevulinic acid application and different irrigation water salinities', *Irrigation and Drainage* **69**. 16

Ahmed, M. (2007), 'The efficacy of four systemic insecticides using two methods of application against the green date palm pit scale insect (*ASTEROLECANIUM PHOENICIS* Rao) (*PALMAPSIS PHOENICIS*) (Homoptera: Asterolecaniidae) in Northern Sudan', *Acta Horticulturae* **736**(34), 369–389. 22

Al-Khayri, J. (2012), 'Determination of the date palm cell suspension growth curve, optimum plating efficiency, and influence of liquid medium on somatic embryogenesis', *Emirates Journal of Food and Agriculture* **24**, 444–455. 29

Al-Khayri, J. M. and Naik, P. M. (2017), 'Date palm micropropagation: Advances and applications', *Ciência e Agrotecnologia* **41**, 347 – 358. 11

Al-Mssallem, I. S., Hu, S., Zhang, X., Lin, Q., Liu, W., Tan, J., Yu, X., Liu, J., Pan, L., Zhang, T., Yin, Y., Xin, C., Wu, H., Zhang, G., Ba Abdullah, M. M., Huang, D., Fang, Y., Alnakhli, Y. O., Jia, S., Yin, A., Alhuzimi, E. M., Alsaihati, B. A., Al-Owayyed, S. A., Zhao, D., Zhang, S., Al-Otaibi, N. A., Sun, G., Majrashi, M. A., Li, F., Tala, Wang, J., Yun, Q., Alnassar, N. A., Wang, L., Yang, M., Al-Jelaify, R. F., Liu, K., Gao, S., Chen, K., Alkhaldi, S. R., Liu, G., Zhang, M., Guo, H. and Yu, J. (2013), 'Genome sequence of the date palm *Phoenix dactylifera* L.', *Nature Communications* **4**, 2274. 9

Al-Musawi, M. (2019), 'Effects of Water on Fruit Set and Weight of Date Palm (*Phoenix dactylifera* L.)', *Annals of Agri Bio Research* **24**, 221 – 226. 106

Al-Ruzouq, R., Shanableh, A., Gibril, M. and Al-Mansoori, S. (2018), 'Image segmentation parameter selection and ant colony optimization for date palm tree detection and mapping from very-high-spatial-resolution aerial imagery', *Remote Sensing* **10**(9). 29

Al-Saikhan, M. S. (2008), 'Effect of thinning practices on fruit yield and quality of Ruzeiz date palm cultivar (*Phoenix dactylifera* L.) in Al-Ahsa Saudi Arabia', *Asian Journal of Plant Sciences* **7**(1), 105–108. 3

Al-Yahyai, R. (2018), 'Strategies to improve date palm production and hence dates quality in the sultanate of oman', Workshop Lecture. 12, 28

Alhejjaj, H., Ayad, J., Othman, Y. and Abu-Rayyan, A. (2020), 'Foliar potassium application improves fruits yield and quality of 'Medjool' date palm', *Fresenius Environmental Bulletin* **29**, 1436–1442. 64

Athanassiou, C., Phillips, T. and Wakil, W. (2019), 'Biology and control of the khapra beetle, *Trogoderma granarium*, a major quarantine threat to global food security', *Annual Review of Entomology* **64**. 2

Bai, T., Zhang, N., Chen, Y. and Mercatoris, B. (2019), 'Assessing the performance of the WOFOST model in simulating jujube fruit tree growth under different irrigation regimes', *Sustainability* **11**, 1466. 35

Baltagi, B. H. (2008), *Econometric Analysis of Panel Data*, fourth ed. edn, Wiley, New York. 76

Battel, B. (2017), 'Understanding growing degree-days', https://www.canr.msu.edu/news/understanding_growing_degree_days. Accessed: 2020-03-28. 12

Bevans, R. (2020), 'An introduction to the akaike information criterion', https://www.scribbr.com/statistics/akaike-information-criterion/. Accessed: 2020-10-30. 75

Bhat, N., Lekha, V., Suleiman, M., Thomas, B., Ali, S., George, P. and Al-Mulla, L. (2012), 'Estimation of water requirements for young date palms under arid climatic conditions of Kuwait', *World Journal of Agricultural Sciences* **8**(5), 448–452. 64

Bhat, V. and Kenna, G. (2013), 'Date palm (*Phoenix dactylifera*) (arecaceae)', *Agnote* . 13, 16

Bhattacharya, A., Kourmpetli, S. and Davey, M. R. (2012), 'Practical applications of manipulating plant architecture by regulating gibberellin metabolism', *Journal of Plant Growth Regulation* pp. 249–256. 22

Black, D. (2017), 'Managing your farm with Fruitlook', https://bluenorth.co.za/managing-your-operation-with-fruitlook/. Accessed: 2020-04-30. 33

Bolker, B. M. (2008), *Ecological Models and Data in R*, Princeton University Press, Princeton. 75

Brase, C. H. and Brase, C. P. (2019), *Understanding Basic Statistics*, international metric edition edn, Cengage Learning. 40

Breaux, H. J. (1967), On stepwise multiple linear regression, Technical Report 1369, Army Ballistic Research Lab, Aberdeen Proving Ground, Maryland. 79

Breiman, L. (1995), 'Better subset regression using the nonnegative garrote', *Technometrics* **37**(4), 373–384. 79, 81

Burnham, K., Anderson, D. and Huyvaert, K. K. (2011), 'AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons', *Behavioral Ecology and Sociobiology* **65**, 23–35. 73, 75

Cai, Y., Moore, K., Pellegrini, A., Elhaddad, A., Lessel, J., Townsend, C., Solak, H. and Semret, N. (2017), Crop yield predictions – high resolution statistical model for intra-season forecasts applied to corn in the US, Technical report, Gro Intelligence Incorporated. 30, 66

Chao, C. T. and Krueger, R. R. (2007), 'The Date Palm (*Phoenix dactylifera* L.): Overview of Biology, Uses, and Cultivation', *HortScience* **42**(5), 1077 – 1082. 8, 9, 21, 27, 105

Climate-Data.org (2020), 'Klein pella climate', https://en.climate-data.org/africa/south-africa/northern-cape/klein-pella-565600/. Accessed: 2020-03-10. 58

Cohen, J. (1977), Chapter 1 – The concepts of power analysis, *in* J. Cohen, ed., 'Statistical Power Analysis for the Behavioral Sciences', Academic Press, pp. 1 – 17. 74

*Data meaning in the Cambridge English Dictionary* (2020), https://dictionary.cambridge.org/dictionary/english/data. Accessed: 2020-03-08. 38

*Data meaning in the Oxford English Dictionary* (2004), https://www.lexico.com/definition/data. Accessed: 2020-03-13. 38

de Ferrara, M. (2012), 'Stuffed dates', http://glorious.atenveldt.org/home/a-s-special-southwind-2012/table-of-contents/cooking/stuffed-dates. Accessed: 2019-09-01. 6, 8

Deeb, A. E. (2015), 'What to do with "small" data?', https://medium.com/rants-on-machine-learning/what-to-do-with-small-data-d253254d1a89. Accessed: 2020-03-17. 73

Desboulets, L. D. D. (2018), 'A review on variable selection in regression analysis', *Econometrics* **6**, 45. 79

Devadoss, S. and Luckstead, J. (2010), 'An analysis of apple supply response', *International Journal of Production Economics* **124**, 265–271. 31

Dhehibi, B., Ben Salah, M. and Frija, A. (2018), *Date Palm Value Chain Analysis and Marketing Opportunities for the Gulf Cooperation Council (GCC) Countries*, IntechOpen. 23, 24, 27, 28

Djerbi, M. (1995), *Précis de phoeniciculture*, FAO, Rome. 64, 129

Djerriri, K., Ghabi, M., Karoui, M. and Adjoudj, R. (2018), Palm trees counting in remote sensing imagery using regression convolutional neural network, *in* 'International Geoscience and Remote Sensing Symposium (IGARSS)', Vol. July, Institute of Electrical and Electronics Engineers Inc., pp. 2627–2630. Conference of 38th Annual IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018 ; Conference Date: 22 July 2018 Through 27 July 2018; Conference Code:141934. 29

Dowson, V. (1982), 'Date production and protection with special reference to North Africa and the Near East', *FAO Technical Bulletin* **35**, 294. 14

El Hodoairi, M. H., Bawa, O. and El Barkouli, A. A. (1992), 'The effects of some growth regulators on fruitset of date palms (Phoenix dactylifera L.) trees', *Acta Horticulturae* **321**(35), 334–342. 22

El-Shafie, H., Peña, J. and Khalaf, M. (2015), Major hemipteran pests, *in* W. Wakil, J. Romeno Faleiro and T. Miller, eds, 'Sustainable Pest Management in Date Palm: Current Status and Emerging Challenges', 1st edn, Sustainability in Plant and Crop Protection, Springer, Cham, pp. 169–204. 22

Elazar J. Pedhazur, L. P. S. (1991), *Measurement, Design, and Analysis: An Integrated Approach*, 1st edn, Psychology Press, New York. 73

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstr, C. and Wold, S. (2006), 'Multi- and megavariate data analysis: Part i basic principles and applications', *Ume Sweden: MKS Umetrics AB* pp. 1–103. 85

*Estimating Crop Yields: A Brief Guide* (2013), https://apo.org.au/sites/default/files/resource-files/2010-11/apo-nid56597_0.htm. Accessed: 2020-01-14. 30

Ezz, T., Kassem, H. and Marzouk, H. (2010), 'Response of date palm trees to different nitrogen and potassium application rates', *Acta Horticulturae* **882**, 761–768. 64

Fan, J. and Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911. 79, 83

FAO (2016), Workshop on "Irrigation of Date Palm and Associated Crops", Food and Agriculture Organization of the United Nations, Damascus, Syrian Arab Republic. 1

FAO (2021), 'Food loss and food waste', http://www.fao.org/food-loss-and-food-waste/flw-data)#:~:text=One%2Dthird%20of%20food%20produced,way%20to%20final%20household%20consumption. Accessed: 2021-01-30. 1

Frost, J. (2020), 'Overfitting regression models: Problems, detection, and avoidance', https://statisticsbyjim.com/regression/overfitting-regression-models/. Accessed: 2020-10-01. 78

Ghnimi, S., Umer, S., Karim, A. and Kamal-Eldin, A. (2017), 'Date fruit (*Phoenix dactylifera* L.): An underutilized food seeking industrial valorization', *NFS Journal* **6**, 1 – 10. 8

Glasner, B., Botes, A., Zaid, A. and Emmens, J. (2002), Chapter ix: Date harvesting, packinghouse management and marketing aspects, *in* A. Zaid, ed., 'Date Palm Cultivation', Food and Agricultural Organization of the United Nations, chapter IX. 24

Gornall, J., Betts, R., Burke, E., Clark, R., Camp, J., Willett, K. and Wiltshire, A. (2010), 'Implications of climate change for agricultural productivity in the early twenty-first century', *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**(1554), 2973–2989. 30

Gosselin, R., Rodrigue, D. and Duchesne, C. (2010), 'A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications', *Chemometrics and Intelligent Laboratory Systems* **100**(1), 12 – 21. 85, 101

Grace-Martin, K. (2020), 'Confusing statistical terms #1: The many names of independent variables', https://www.theanalysisfactor.com/the-many-names-of-independent-variables/. Accessed: 2020-08-01. 76

Green, S. (1991), 'How many subjects does it take to do a regression analysis', *Multivariate Behavioral Research* **26**(3), 499 – 510. 73

Growers, L. W. (2020), 'Fruit set factors', https://www.lodigrowers.com/improving-fruit-set/. Accessed: 2020-04-10. 13

Gurrea-Ysasi, G., Blanca-Gimenez, V., Fita, I., Fita, A., Prohens, J. and Rodriguez-Burruezo, A. (2018), 'Spectral comparison of diffuse par irradiance under different tree and shrub shading conditions and in cloudy days', *Journal of photochemistry and photobiology. B, Biology* **189**, 274—282. 14, 15

Guyon, I. and Elisseeff, A. (2003), 'An introduction of variable and feature selection', *J. Machine Learning Research Special Issue on Variable and Feature Selection* **3**, 1157 – 1182. 83

Hand, D., Mannila, H. and Smyth, P. (2001), *Principles of Data Mining*, MIT Press, Cambridge, Massachusetts. 78

Harrel, F. (2001), *Regression Modeling Strategies*, Springer, New York. 81

Harris, R. J. (1985), *A Primer of Multivariate Statistics*, 2nd edition edn, Academic Press, New York, New York. 73

Hawkins, D., Basak, S. and Mills, D. (2003), 'Assessing model fit by cross-validation', *Journal of Chemical Information and Computer Sciences* **43**, 579 – 586. 70

Hendry, D. and Richard, J.-F. (1987), Recent developments in the theory of encompassing, CORE Discussion Papers 1987022, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).
**URL:** *https://EconPapers.repec.org/RePEc:cor:louvco:1987022* 79, 81

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67. 82

Horie, T., Yajima, M. and Nakagawa, H. (1992), 'Yield forecasting', *Agricultural Systems* **40**(1-3), 211–236. 3

Hosmer, D., Lemeshow, S. and Sturdivant, R. (2013), *Applied logistic regression*, 1 edn, John Wiley & Sons, Incorporated, New York. 68

Howard, F. W. (1999), 'An introduction to insect pests of palms', *Acta Horticulturae* **486**, 133—140. 22

Husain, M. and Khan, R. A. (2020), 'Date palm crop yield estimation – A framework', *SSRN Electronic Journal* **7**(6). 36

Huth, N., Banabas, M., Nelson, P. and Webb, M. (2014), 'Development of an oil palm cropping systems model: Lessons learned and future directions', *Environmental Modelling and Software* **62**, 411–419. 30, 33

Intha, N. and Chaiprasart, P. (2018), 'Sex determination in date palm (*Phoenix dactylifera* L.) by PCR based marker analysis', *Scientia Horticulturae* **236**(December 2017), 251–255. 7

Iqbal, M., Jatoi, S. A., Niamatullah, M., Munir, M. and Khan, I. (2014), 'Effect of pollination time on yield and quality of date fruit', *Journal of Animal and Plant Sciences* **24**(3), 760–764. 3

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer Publishing Company, Incorporated. 77, 98

Janick, J. and Paull, R. E. (2008), *The Encyclopedia of Fruit & Nuts*, Cabi Publishing. 11

Jenkins, D. and Quintana-Ascencio, P. (2020), 'A solution to minimum sample size for regressions', *PLos ONE* **15**. 74, 75, 76

Johnstone, H. (2020), "The dates are drying': Profits shrivel for farmers as the heat rises in Tunisia', https://www.theguardian.com/global-development/2020/jan/14/the-dates-are-drying-profits-shrivel-for-farmers-as-the-heat-rises-in-tunisia. Accessed: 2021-01-03. 2

Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A. and Ritchie, J. (2003), 'The DSSAT cropping system model', *European Journal of Agronomy* **18**, 235 – 265. 32, 33

Justice, C., Vermote, E., Townshend, J., Defries, R., Roy, D., Hall, D., Salomonson, V., Privette, J., Riggs, G., Strahler, A., Lucht, W., Myneni, R., Knyazikhin, Y., Running, S., Nemani, R., Wan, Z., Huete, A., Van Leeuwen, W., Wolfe, R., Giglio, L., Muller, J., Lewis, P. and Barnsley, M. (1998), 'The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research', *IEEE Transactions on Geoscience and Remote Sensing* **36**(4), 1228–1249. 31

Kader, A. A. and Hussein, A. M. (2009), 'Harvesting and postharvest handling of dates', *ICARDA* pp. 1–15. 25, 26, 27

Kayal, M. (2015), 'Dates: The sticky history of a sweet fruit', https://www.nationalgeographic.com/people-and-culture/food/the-plate/2015/06/18/dates-the-sticky-history-of-a-sweet-fruit/. Accessed: 2019-08-18. 2, 24

Kelleher, J. D., Mac Namee, B. and D'Arcy, A. (2015), *Fundamentals of Machine Learning For Predictive Data Analytics*, Massachusetts Institute of Technology. 66

Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., Bognár, P. and Lichtenberger, J. (2018), 'Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices', *Agricultural and Forest Meteorology* **260-261**, 300 – 320. 36

Khan, H. and Khan, S. A. (2014), 'Dates for health', http://radianceweekly.in/portal/issue/rail-fare-hike-bitter-medicine-or-fatal-blow/article/dates-for-health/. Accessed: 2019-08-26. 2

Khillar, S. (2019), 'Difference between absolute and relative humidity', http://www.differencebetween.net/science/nature/difference-between-absolute-and-relative-humidity/. Accessed: 2020-04-20. 13

Klein, P. and Zaid, A. (2002), Chapter vi: Land preparation, planting operation and fertilisation requirements, *in* A. Zaid, ed., 'Date Palm Cultivation', Food and Agricultural Organization of the United Nations, Rome, chapter 6. 18, 19

Köhne, S. (1985), 'Yield estimation based on measureable parameters', *South African Avocado Grower's Association Yearbook* **8**(103), 3. 34

Kowalski, K. (2015), 'Explainer: What is a computer model?', https://www.sciencenewsforstudents.org/article/explainer-what-computer-model. Accessed: 2020-04-23. 30

Kozlowski, T. T. (1992), 'Carbohydrate sources and sinks in woody plants', *Botanical Review* **58**(2), 107–222. 106

Li, H. (2017), 'Which machine learning algorithm should i use?', https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use. Accessed: 2019-10-18. 67

Liebenberg, P. and Zaid, A. (2002), Chapter vii: Date palm irrigation, *in* A. Zaid, ed., 'Date Palm Cultivation', Food and Agricultural Organization of the United Nations, Rome, chapter 7. 11, 16, 17

Lobell, D. and Burke, M. (2010), 'On the use of statistical models to predict crop yield responses to climate change', *Agricultural and Forest Meteorology - AGR FOREST METEOROL* **150**, 1443–1452. 35

Lobell, D., Field, C., Nicholas, K. and Bonfils, C. (2006), 'Impacts of future climate change on California perennial crop yields: Model projections with climate and crop uncertainties', *Agricultural and Forest Meteorology* **141**, 208–218. 36

Lobo, G., Yahia, E. and Kader, A. (2013), Biology and postharvest physiology of date fruit, *in* M. Siddiq, S. M. Aleid and A. A. Kader, eds, 'Dates: Postharvest Science, Processing Technology and Health Benefits', Woodhead Publishing, England, chapter 3, pp. 57–80. 8, 20, 22, 25, 27

Magezi, D. A. (2015), 'Linear mixed-effects models for within-participant psychology experiments: an introductory tutorial and free, graphical user interface (LMMgui)', *Frontiers in Psychology* **6**, 2. 76

Malik, Z., Ziauddin, S., Ahmad, R. and Safi, A. (2016), 'Detection and Counting of On-Tree Citrus Fruit for Crop Yield Estimation', *International Journal of Advanced Computer Science and Applications* **7**(5), 519–523. 3

Marquardt, D. and Snee, R. (1975), 'Ridge regression in practice', *American Statistician – AMER STATIST* **29**, 3–20. 81

Marsal, J. and Stockle, C. (2011), 'Use of CropSyst as a decision support system for scheduling regulated deficit irrigation in a pear orchard', *Irrigation Science* **30**, 139–147. 34

Mason, S. E. (1925), 'Partial thermostasy of the growth center of the date palm', *Journal of Agricultural Research* **31**, 415 – 453. 14

Max Kuhn, K. J. (2013), *Applied Predictive Modeling*, Springer, New York. 82, 84

Mbaga, M. (2012*a*), *An Overview of Dates Marketing*, CRC Press. 23

Mbaga, M. D. (2012*b*), Date marketing, *in* E. S. A. Manickavasagan, M. Mohamed Essa, ed., 'Dates: Production, Processing, Food, and Medicinal Values', CRC Press, Boca Raton, chapter 11. 24, 27

McCubbin, M. (2007), 'The South African date palm industry – strengths and weaknesses', *Acta Horticulturae* **736**, 59–63. 2

McGregor, N. (2019), 'South africa: Major investment in dates production', https://www.freshplaza.com/article/157231/South-Africa-Major-investment-in-dates-production/. Accessed: 2019-08-15. 2

McMullen, S. (2018), 'How many dates will one palm tree produce?', https://homeguides.sfgate.com/many-dates-one-palm-tree-produce-101844.html. Accessed: 2019-08-29. 7

Mehmood, T., Liland, K., Snipen, L. and Sæbø, S. (2012), 'A review of variable selection methods in Partial Least Squares Regression', *Chemometrics and Intelligent Laboratory Systems* **118**, 62 – 69. 85

Mihoub, A., Samia, H., Sakher, M., Hafed, K., Koull, N., Kawther, L., Tidjani, B., Abdesselam, B., Mhamed Bensalah, L., Yamina, K. and Amor, H. (2015), 'Date Palm (*Phoenix dactylifera* L.) irrigation water requirements as affected by salinity in Oued Righ conditions, North Eastern Sahara, Algeria', *Asian Journal of Crop Science* **7**, 174–185. 35

Müller, A. and Guido, S. (2016), *Introduction to Machine Learning with Python: A Guide for Data Scientists*, O'Reilly Media.
**URL:** *https://books.google.co.za/books?id=vbQlDQAAQBAJ* 66

Mushtaq, T., Mir, S., Nazir, N., Raja, T., Pandith, A., Rasool, K. and Lone, M. (2018), 'Statistical model for yield estimation of 'Gala Red Lum' apples after bloom in Northern India', *Current Journal of Applied Science and Technology* **29**, 1–8. 36

Nagini, S., Kanth, T., Rajini, V. and Kiranmayee, B. V. (2016), 'Agriculture yield prediction using predictive analytic techniques', *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016* pp. 783 – 788. 24, 67

NASA (2021), 'Prediction of worldwide energy resources', https://power.larc.nasa.gov/. Accessed: 2021-01-02. 48

Nath, M. (2000), Study on the processing and preservation of locally grown dates, Master's thesis, Bangladesh Agricultural University, Mymensingh. 9

Nixon, R. (1956), 'How many fruits per strand should be left in thinning the Medjool date?', *Date Growers' Institute Report* **33**. 64, 129

Nugent, J. (2020), 'Calculating growing degree days', https://www.canr.msu.edu/uploads/files/Research_Center/NW_Mich_Hort/General/CalculatingGrowingDegreeDays.pdf. Accessed: 2020-02-28. 12

Olah, M., Bologa, C. and Oprea, T. (2004), 'An automated PLS search for biologically relevant QSAR descriptors', *Journal of computer-aided molecular design* **18**, 437 – 449. 85

Oosthuyse, S. (2018), 'Effect of potassium fertilizer, either $KNO_3$, $K_2SO_4$ or KCl, in a saline fertigation solution on the growth of Medjool date palm plants pot-grown in river sand or river sand/calcium carbonate', *Acta Horticulturae* **1217**, 357–364. 64

Patel, A. (2018), 'Chapter 4: Knowledge from the data and Data Exploration Analysis', https://medium.com/ml-research-lab/chapter-4-knowledge-from-the-data-and-data-exploration-analysis-99a734792733. Accessed: 2020-02-29. 38

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830. 84, 89, 96

Piñeiro, G., Perelman, S., Guerschman, J. and Paruelo, J. (2008), 'How to evaluate models: Observed vs. predicted or predicted vs. observed?', *Ecological Modelling* **216**, 316–322. 95, 96

Qin, W., Heinen, M., Assinck, F. B. and Oenema, O. (2016), 'Exploring optimal fertigation strategies for orange production, using soil-crop modelling', *Agriculture, Ecosystems and Environment* **223**, 31–40. 34

Rajmohan, K. (2011), Date palm tissue culture: A pathway to rural development, *in* S. M. Jain, J. M. Al-Khayri and D. V. Johnson, eds, 'Date Palm Biotechnology', Springer Science & Business Media, Netherlands. 1

Reilly, D., Reilly, A. and Lewis, I. (2010), *Towards an Australian date industry: an overview of the Australian domestic and international date industries*, Rural Industries Research and Development Corporation, Barton, A.C.T. 2, 9

Reuveni, O., Abu, S. and Golobovitz, S. (1986), 'Date palm pollen germination and tube elongation on pistillate flowers cultured at different temperatures', *Acta Horticulturae* **175**, 91 – 96. 14

Robertson, M., Fukai, S., Hammer, G. and Ludlow, M. (1993), 'Modelling root growth of grain sorghum using the CERES approach', *Field Crops Research* **33**(1), 113 – 130. 33

Rygg, G. (1975), 'Date development, handling, and packing in the united states', *USDA Agriculture Handbook* . 8, 12, 14, 15

Sabbatini, P., Dami, I. and Howell, G. S. (2012), 'Predicting harvest yield in juice and wine grape vineyards', https://www.canr.msu.edu/uploads/resources/pdfs/Predicting_Harvest_Yield_in_Juice_and_Wine_Grape_Vineyards_(E3186).pdf. 3

Sabir, I. (2019), 'Dates have been a staple food of the Middle East and the Indus Valley for thousands of years', https://www.pakistangulfeconomist.com/2019/07/08/dates-have-been-a-staple-food-of-the-middle-east-and-the-indus-valley-for-thousands-of-year Accessed: 2020-12-01. 1

Sardar A. Farooq, Roohi S. Khan, T. T. F. (2012), Tissue culture studies in date palm, *in* E. S. A. Manickavasagan, M. Mohamed Essa, ed., 'Dates: Production, Processing, Food, and Medicinal Values', CRC Press, Boca Raton, chapter 2. 8, 11

Shabana, H. and Sunbol, A. (2007), 'Date palm flowering and fruit setting as affected by low temperatures preceding the flowering season', *Acta Horticulturae* **736**, 193–198. 106

Shabani, F., Kumar, L. and Taylor, S. (2012), 'Climate Change Impacts on the Future Distribution of Date Palms: A Modeling Exercise Using CLIMEX', *PLoS ONE* **7**(10), 1–12. 33, 35

Shahbandeh, M. (2020), 'Global date palm industry value 2014–2023', https://www.statista.com/statistics/960213/date-palm-market-value-worldwide/. Accessed: 2020-01-02. 1

Shi, W., Tao, F. and Zhang, Z. (2013), 'A review on statistical models for identifying climate contributions to crop yields', *Journal of Geographical Sciences* **23**(3), 567 – 576. 35

Snyder, H. (2019), 'Literature review as a research methodology: An overview and guidelines', *Journal of Business Research* **104**, 333–339.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0148296319304564* 29

Spendlove, T. (2016), 'Can crops survive without water?', https://www.engineering.com/IOT/ArticleID/11389/Can-Crops-Survive-Without-Water.aspx. Accessed: 2019-10-30. 13

Sperling, O., Shapira, O., Tripler, E., Schwartz, A. and Lazarovitch, N. (2014), 'A model for computing date palm water requirements as affected by salinity', *Irrigation Science* **32**. 37

Steduto, P., Hsiao, T., Fereres, E. and Raes, D. (2012), *Crop yield response to water*, Food and Agriculture Organization of the United Nations, Rome. 32

Steinberger, S. (2013), https://pixabay.com/photos/date-palm-palm-dates-223247/. Accessed: 2019-09-07. 7

Sung, C. T. B. and Siang, C. S. (2018), *Modelling crop growth and yield in palm oil cultivation*, Burleigh Dodds Science Publishing. 34

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288. 79, 81, 82

Tiwari, P. and Shukla, P. (2018), 'Crop yield prediction by modified convolutional neural network and geographical indexes', *International Journal of Computer Sciences and Engineering* **6**, 503–513. 31

Transparency Market Research (2018), Date palm market (nature - organic, conventional; form - raw, processed; variety - deglet noor, medjool, barhi, zahidi, others; end use - household, foodservice, dietary supplements, personal care & cosmetics, food industry) - global industry analysis, size, share, growth, trends, and forecast 2018 - 2026, Technical report, Transparency Market Research. 2, 6, 23

Tregeagle, D. T. (2017), The Dynamics of Perennial Crop Production and Processing, PhD thesis, University of California, Berkeley. 31

Turek, C., Wróbel, S. and Piwowar, M. (2020), 'OmicsON – Integration of omics data with molecular networks and statistical procedures', *PloS one* **15**(7), e0235398–e0235398. 84

van Diepen, C., Wolf, J., van Keulen, H. and Rappoldt, C. (1989), 'WOFOST: a simulation model of crop production', *Soil Use and Management* **5**(1), 16–24. 33

Verma, S., Lucas, A., Zhang, X., Veturi, Y., Dudek, S., Li, B., Li, R., Urbanowicz, R., Moore, J., Kim, D. and Ritchie, M. (2018), 'Collective feature selection to identify crucial epistatic variants', *BioData Mining* **11**(5). 84

Vozhehova, R., Lavrynenko, Y., Kokovikhin, S., Lykhovyd, P., Biliaieva, I., Drobitko, A. and Nesterchuk, V. (2018), 'Assessment of the CROPWAT 8.0 software reliability for evapotranspiration and crop water requirements calculations', *Journal of Water and Land Development* **39**, 147–152. 32

Wang, Q., Nuske, S., Bergerman, M. and Singh, S. (2013), 'Automated Crop Yield Estimation for Apple Orchards', *Experimental Robotics* pp. 745–758. 3

Weston, J., Elisseeff, A., Schölkopf, B. and Tipping, M. (2003), 'Use of the zero-norm with linear models and kernel methods', *J. Machine Learning Research Special Issue on Variable and Feature Selection* **3**, 1439 – 1461. 83

White, A., Rogers, A., Rees, M. and Osborne, C. (2015), 'How can we make plants grow faster? A source-sink perspective on growth rate', *Journal of experimental botany* **67**, 31–45. 106

Wichelns, D. (2014), 'Do estimates of water productivity enhance understanding of farm-level water management?', *Water* **6**(4), 778–795. 16

Williams, J., Jones, C. A., Kiniry, J. and Spanel, D. (1989), 'EPIC crop growth model', *Transactions of the ASAE* **32**. 33

Writers, S. D. S. (2014), 'DMCii help Dutch company eLEAF provide much needed crop information to African farmers', https://www.spacedaily.com/reports/DMCii_help_Dutch_company_eLEAF_provide_much_needed_crop_information_to_African_farmers_999.html. Accessed: 2020-04-29. 33

Yahia, E. and Kader, A. (2011), Date (*Phoenix dactylifera* L.), *in* 'Postharvest Biology and Technology of Tropical and Subtropical Fruits: Cocona to Mango', Woodhead Publishing, England, pp. 41–79. 13

Zacharias, A. (2019), 'Date palm culture makes UNESCO Intangible Heritage list', https://www.thenationalnews.com/uae/heritage/date-palm-culture-makes-unesco-intangible-heritage-list-1.950386. Accessed: 2020-11-02. 1

Zaid, A. (2002), *Date Palm Cultivation*, Vol. 156, Food and Agricultural Organization of the United Nations. 11, 27

Zaid, A., de Wet M. Djerbi, P. and Oihabi, A. (2002), Chapter XII: Diseases and pests of date palm, *in* A. Zaid, ed., 'Date Palm Cultivation', Food and Agricultural Organization of the United Nations, Rome, chapter 12. 22

Zaid, A. and de Wet, P. (2002*a*), Chapter iv: Climatic requirements of date palm, *in* A. Zaid, ed., 'Date Palm Cultivation', Food and Agricultural Organization of the United Nations, Rome, chapter 4. 8, 12, 13, 14, 106

Zaid, A. and de Wet, P. (2002*b*), Chapter v: Date palm propagation, *in* A. Zaid, ed., 'Date Palm Cultivation', Food and Agricultural Organization of the United Nations, Rome, chapter 5. 9, 10

Zaid, A. and de Wet, P. (2002*c*), Pollination and bunch management, *in* A. Zaid, ed., 'Date Palm Cultivation', Food and Agricultural Organization of the United Nations, chapter VIII. 19, 20

Zaid, A. and Klein, P. (2002), Date palm technical calendar, *in* A. Zaid, ed., 'Date Palm Cultivation', Food and Agricultural Organization of the United Nations, chapter XI. 11, 12, 15, 20, 21

# Appendix A

# Orchard layout

For a general understanding of the position of the orchards, the map with the orchard numbers is displayed in Figure A.1, showing orchards shielded by the trees next to the river and orchards lying on the unshaded sunny side.

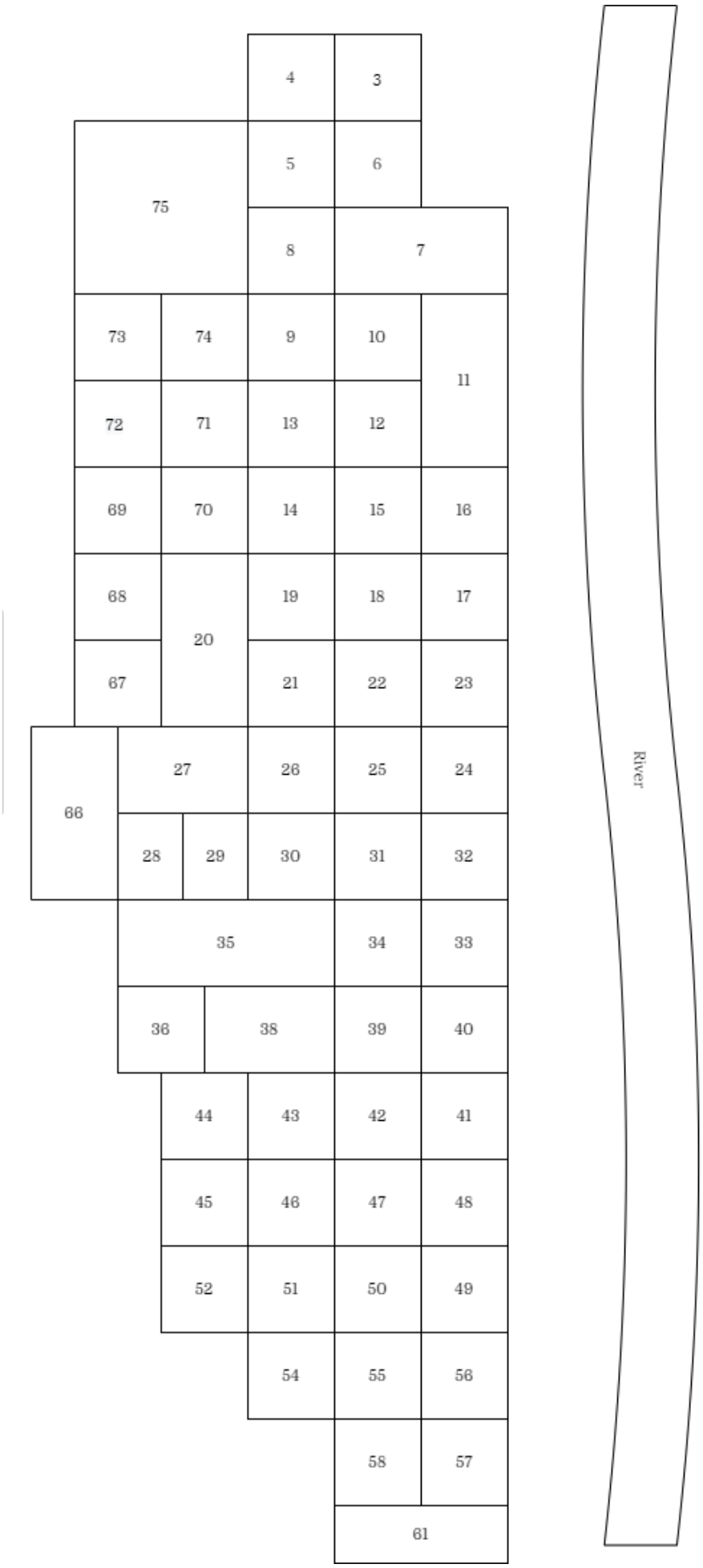



Figure A.1: Layout of orchards located next to the river

# Appendix B

# Results of regression models predicting yield with weather and bunch features

This appendix contains results of Chapter 7 in more detail. In the tables shown here, the same abbreviations apply as in Chapter 7 where prev denotes the weather of the previous season and months are numerically denoted by the $i^{th}$ month of the year a subscript $i$ in equations. Temperature is denoted Temp in weather feature names and T in the equations. Minimum and maximum are abbreviated Min and Max respectively. The mean Humidity of the month is written as Hum and the mean wind speed is written simply as Wind. The heat units of a month is written as HU or HU. The sum of the rainfall of the month is written as Rain.

## B.1 Correlation based method on weather

In Section 7.1.1 the features were identified using thresholds. Using these features in an OLS regression model as done in Chapter 7 yields the following equations displayed in Tables B.1, B.3, B.4 and B.2 for the corresponding threshold.

Table B.1: Multiple linear regression on features with correlation threshold of **0.7** and greater

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 1978 | 17 | Aug Max Temp | $4.62MaxT_8 - 54.565$ | 11.099 |
| 1978 | 90 | Apr Mean Temp | $11.224\ MeanT_4 - 134.710$ | 15.987 |
| 1989 | 56 | prev Mar Hum, prev May Hum | $0.487prevHum_3 + 3.767prevHum_5 - 57.832$ | 21.398 |
| 1989 | 57 | Aug HU, Aug Max Temp | $0.14HU_8 + 9.272MaxT_8 - 179.646$ | 26.341 |
| 1989 | 58 | Aug HU, Aug Mean Temp, Aug Max Temp | $2.240HU_8 - 67.208MeanT_8 + 6.479MaxT_8 + 1098.612$ | 18.960 |
| 1989 | 61 | Aug Max Temp | $11.118MaxT_8 - 256.329$ | 26.364 |
| 1991 | 33 | Aug HU, Aug Mean Temp | $1.819HU_8 - 51.667MeanT_8 + 1022.708$ | 11.219 |
| 1991 | 39 | prev May Hum, Jul Rain | $2.212prevHum_5 + 5.026Rain_7 + 17.738$ | 29.382 |
| 1991 | 42 | prev May Hum, Jul Rain | $2.327prevHum_5 + 4.265Rain_7 + 18.216$ | 25.020 |
| 1991 | 43 | Aug Max Temp | $9.638MaxT_8 - 201.089$ | 27.030 |
| 1991 | 44 | prev Oct HU, Aug HU, Aug Min Temp, Aug Mean Temp, Aug Max Temp | $-2.838e-02prevHU_{10} + 1.672HU_8 + 9.165MinT_8 - 4.828e+01MeanT_8 + 2.885MaxT_8 + 837.851$ | 13.636 |
| 1991 | 45 | Aug HU, Aug Max Temp | $0.161HU_8 + 6.89MaxT_8 - 100.354$ | 26.025 |
| 1991 | 47 | Jul Rain | $9.1Rain_7 + 103.779$ | 23.745 |
| 1991 | 49 | prev May Hum | $4.174prevHum_5 - 50.441$ | 31.595 |
| 1991 | 51 | prev May Hum | $3.377prevHum_5 - 20.168$ | 23.544 |
| 1993 | 70 | Aug HU | $0.270HU_8 + 116.271$ | 18.058 |

Continued on next page

**Table B.1 – continued from previous page**

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 1993 | 71 | prev Jun Hum | $1.981prevHum_6 + 1.322$ | 13.276 |
| 1993 | 75 | prev Oct Min Temp | $21.586prevMinT_{10} - 126.792$ | 23.086 |
| 1995 | 12 | Oct Hum, Jul Rain | $5.886Hum_{10} + 3.773Rain_7 - 63.298$ | 23.051 |
| 2000 | 9 | prev Oct HU, prev Oct Mean Temp, May HU, Nov HU, May Mean Temp, Mean Temp Nov | $-0.818prevHU_{10} + 33.064prevMeanT_{10} - 0.579HU_5 + 0.154HU_{11} + 20.025MeanT_5 + 2.004MeanT_{11} - 1012.880$ | 33.364 |
| 2000 | 10 | prev Jan Max Temp, Aug Min Temp | $8.153prevMaxT_1 + 6.671MinT_8 - 296.311$ | 15.843 |
| 2000 | 11 | Jan Wind, Feb Wind, prev Oct HU, prev Oct Mean Temp, Oct Max Temp, Dec Max Temp | $1.706Wind_1 + 4.528Wind_2 + 0.495prevHU_{10} - 5.554prevMeanT_{10} + 6.016MaxT_{10} - 13.949MaxT_{12} + 442.029$ | 16.267 |
| 2000 | 13 | prev Nov Max Temp | $20.756prevMaxT_{11} - 763.763$ | 22.002 |
| 2000 | 14 | prev Oct HU, prev Oct Mean Temp, prev Oct Max Temp, Nov HU, Apr Min Temp, Mean Temp Nov, Jan Max Temp | $-0.266prevHU_{10} + 10.216prevMeanT_{10} + 3.789prevMaxT_{10} + 0.386HU_{11} + 1.067MinT_4 - 3.999MeanT_{11} + 4.371MaxT_1 - 452.446$ | 51.563 |
| 2000 | 15 | May HU, Sep HU, Sep Mean Temp, Jun Max Temp | $0.088HU_5 - 0.937HU_9 + 32.236MeanT_9 + 4.579MaxT_6 - 643.033$ | 21.298 |
| 2002 | 4 | Jan Wind, Feb Wind, Mar Wind, Oct HU, Oct Mean Temp, Oct Max Temp | $14.174Wind_1 - 4.577Wind_2 - 1.024Wind_3 - 0.188HU_{10} + 12.164MeanT_{10} + 0.83MaxT_{10} - 293.844$ | 33.739 |

Table B.2: Multiple linear regression on features with correlation threshold of **-0.7** and smaller

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 1978 | 17 | Dec Wind, May Max Temp | $-2.832Wind_{12} - 5.631MaxT_5 + 297.615$ | 9.394 |
| 1989 | 56 | prev May HU, prev May Mean Temp, prev Jun Max Temp | $0.263prevHU_5 - 17.005prevMeanT_5 - 12.780prevMaxT_6 + 798.617$ | 21.266 |
| 1989 | 57 | May Max Temp | $-18.679MaxT_5 + 735.897$ | 23.296 |
| 1989 | 58 | May Max Temp | $-13.491MaxT_5 + 558.509$ | 20.593 |
| 1989 | 61 | May Max Temp | $-17.779MaxT_5 + 692.726$ | 22.653 |
| 1991 | 45 | May Max Temp | $-14.918MaxT_5 + 612.881$ | 25.544 |
| 1991 | 49 | prev Jun Max Temp | $-18.965prevMaxT_6 + 669.803$ | 27.328 |
| 1991 | 51 | prev Jun Max Temp | $-17.234prevMaxT_6 + 617.098$ | 17.604 |
| 1993 | 71 | Dec Wind, Mar Min Temp | $-4.978Wind_{12} - 4.696MinT_3 + 196.594$ | 10.168 |
| 2000 | 5 | Mar Min Temp | $-7.765MinT_3 + 204.502$ | 14.699 |

Continued on next page

**Table B.2 – continued from previous page**

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 2000 | 9 | prev Oct Hum | $-6.849prevHum_{10} + 268.060$ | 23.491 |
| 2000 | 10 | prev Jul Max Temp | $-11.828prevMaxT_7 + 425.512$ | 13.845 |
| 2000 | 11 | prev Oct Hum, Aug Hum, Aug Rain | $-2.902prevHum_{10} + 0.455Hum_8 - 3.490Rain_8 + 168.706$ | 13.406 |
| 2000 | 13 | Mar Min Temp | $-11.152MinT_3 + 261.098$ | 23.164 |
| 2000 | 14 | prev Oct Hum | $-6.045prevHum_{10} + 250.685$ | 20.256 |
| 2000 | 15 | Apr Hum, May Hum | $-1.947Hum_4 - 1.126Hum_5 + 200.514$ | 16.875 |
| 2002 | 4 | Feb Hum | $-3.820Hum_2 + 170.438$ | 18.355 |

Table B.3: Multiple linear regression on features with correlation threshold of **0.6** and greater

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 1978 | 17 | Jan Min Temp, Aug Max Temp | $2.287MinT_1 + 3.508MaxT_8 - 57.641$ | 11.600 |
| 1978 | 90 | prev Apr HU, prev Aug HU, prev Apr Mean Temp, prev Aug Mean Temp, prev Aug Max Temp, prev Dec Max Temp, Apr HU | $-0.639prevHU_4 + 1.919prevHU_8 + 22.878prevMeanT_4 - 63.452prevMeanT_8 + 3.674prevMaxT_8 - 5.878prevMaxT_{12} + 0.452HU_4 + 871.858$ | 22.545 |
| 1989 | 56 | prev Mar Hum, prev May Hum, prev Jun Hum, prev Oct Min Temp, Aug Max Temp | $0.521prevHum_3 + 2.233prevHum_5 - 0.495prevHum_6 + 16.013prevMinT_{10} + 6.145MaxT_8 - 340.383$ | 16.549 |
| 1989 | 57 | Aug HU, Aug Max Temp | $0.14HU_8 + 9.272MaxT_8 - 179.646$ | 26.341 |
| 1989 | 58 | Aug HU, Aug Min Temp, Aug Max Temp | $0.057HU_8 + 6.297MinT_8 + 6.975MaxT_8 - 145.702$ | 19.884 |
| 1989 | 61 | Aug HU, Apr Max Temp, Aug Max Temp | $0.031HU_8 + 5.086MaxT_4 + 8.089MaxT_8 - 343.108$ | 27.580 |
| 1991 | 33 | prev Oct HU, prev Oct Mean Temp, Aug HU, Nov Hum, Aug Max Temp | $1.396prevHU_{10} - 33.494prevMeanT_{10} - 0.038HU_8 + 1.837Hum_{11} + 1.857MaxT_8 + 532.412$ | 12.227 |
| 1991 | 39 | prev May Hum, Aug Max Temp, Jul Rain | $1.221prevHum_5 + 6.642MaxT_8 + 5.849Rain_7 - 158.248$ | 18.621 |
| 1991 | 42 | prev May Hum, Feb Hum, Aug Max Temp, Jul Rain | $1.382prevHum_5 + 0.306Hum_2 + 6MaxT_8 + 4.81Rain_7 - 147.473$ | 19.050 |
| 1991 | 43 | Aug HU, Min Temp Feb, Aug Max Temp | $0.203HU_8 + 7.307MinT_2 + 2.821MaxT_8 - 101.144$ | 25.104 |
| 1991 | 44 | prev Oct HU, prev Oct Mean Temp, Aug HU, Aug Min Temp, Aug Max Temp | $1.091prevHU_{10} - 29.652prevMeanT_{10} + 0.052HU_8 + 10.614MinT_8 + 3.876MaxT_8 + 439.397$ | 39.163 |
| 1991 | 45 | Aug HU, Aug Max Temp, Jul Rain | $0.212HU_8 + 5.013MaxT_8 + 6.418Rain_7 - 55.615$ | 16.789 |

Continued on next page

**Table B.3 – continued from previous page**

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 1991 | 46 | , Mar HU, Min Temp Feb, Aug Max Temp | $+0.265HU_3 + 4.353MinT_2 + 5.685MaxT_8 - 228.732$ | 27.973 |
| 1991 | 47 | prev May Hum, Oct Hum, Jul Rain | $1.486prevHum_5 + 2.144Hum_{10} + 5.61Rain_7 - 8.159$ | 43.452 |
| 1991 | 49 | prev May Hum, Jul Rain | $2.929prevHum_5 + 4.938Rain_7 - 12.152$ | 39.944 |
| 1991 | 50 | prev May Hum, Mar HU, Jul Rain | $1.597prevHum_5 + 0.337HU_3 + 3.749Rain_7 - 60.078$ | 26.929 |
| 1991 | 51 | prev May Hum | $3.377prevHum_5 - 20.168$ | 23.544 |
| 1993 | 70 | prev Jun Hum, Aug HU, Aug Max Temp | $1.682prevHum_6 + 0.153HU_8 + 2.21MaxT_8 - 38.232$ | 16.166 |
| 1993 | 71 | prev Jun Hum, prev Jul Hum, Jan Hum, Aug Max Temp | $1.381prevHum_6 + 0.227prevHum_7 + 1.619Hum_1 + 2.121MaxT_8 - 93.913$ | 14.602 |
| 1993 | 74 | prev Jun Hum, Aug HU, Mar Hum, Aug Max Temp | $0.986prevHum_6 + 0.093HU_8 + 2.236Hum_3 + 3.21MaxT_8 - 130.065$ | 20.091 |
| 1993 | 75 | prev Jun Hum, prev Oct Min Temp, Mar Hum | $0.678prevHum_6 + 11.055prevMinT_{10} + 2.879Hum_3 - 134.682$ | 22.824 |
| 1995 | 12 | Oct Hum, Nov Hum, Jul Rain | $6.198Hum_{10} - 0.461Hum_{11} + 4.043Rain_7 - 60.707$ | 27.097 |
| 2000 | 9 | Jan Wind, Apr Wind, prev Oct HU, prev Oct Mean Temp, prev Oct Max Temp, May HU, Nov HU | $2.813Wind_1 - 13.89Wind_4 - 0.416prevHU_{10} + 16.089prevMeanT_{10} + 4.357prevMaxT_{10} + 0.085HU_5 + 0.458HU_{11} - 465.776$ | 43.201 |
| 2000 | 10 | prev Jan Max Temp, Aug Min Temp, Nov Min Temp | $4.234prevMaxT_1 + 7.529MinT_8 + 3.131MinT_{11} - 170.367$ | 15.390 |
| 2000 | 11 | prev Oct HU, prev Oct Mean Temp, May HU, Oct HU, Oct Max Temp, Dec Max Temp | $1.954prevHU_{10} - 48.896prevMeanT_{10} + 0.18HU_5 + 0.399HU_{10} - 2.213MaxT_{10} - 5.007MaxT_{12} + 1109.299$ | 10.816 |
| 2000 | 13 | prev Oct HU, prev Oct Mean Temp, prev Nov Max Temp, Nov HU, Apr Min Temp, Jan Max Temp | $-1.27prevHU_{10} + 39.225prevMeanT_{10} + 16.724prevMaxT_{11} + 0.114HU_{11} + 6.178MinT_4 - 2.016MaxT_1 - 1311.080$ | 31.530 |
| 2000 | 14 | prev Oct HU, prev Oct Mean Temp, prev Oct Max Temp, Jun HU, Nov HU, Apr Min Temp, Jan Max Temp | $0.466prevHU_{10} - 14.339prevMeanT_{10} + 7.387prevMaxT_{10} - 0.229HU_6 + 0.433HU_{11} + 1.644MinT_4 + 5.067MaxT_1 - 311.306$ | 37.002 |
| 2000 | 15 | prev Oct HU, prev Oct Mean Temp, prev Jan Max Temp, May HU, Sep HU, Nov Min Temp, Jun Max Temp | $-1.337prevHU_{10} + 38.012prevMeanT_{10} - 1.621prevMaxT_1 + 0.256HU_5 + 0.176HU_9 + 5.282MinT_{11} - 4.071MaxT_6 - 479.127$ | 48.772 |

Continued on next page

**Table B.3 – continued from previous page**

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 2002 | 4 | Jan Wind, Feb Wind, Mar Wind, May HU, Oct HU, Oct Mean Temp, Jan Max Temp, Oct Max Temp | $7.072Wind_1 - 6.968Wind_2 - 4.881Wind_3 + 0.503HU_5 - 2.361HU_{10} + 79.922MeanT_{10} + 6.244MaxT_1 - 1.819MaxT_{10} - 1624.457$ | 17.584 |

Table B.4: Multiple linear regression on features with correlation threshold of **-0.6** and smaller

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 1978 | 17 | Dec Wind, May Max Temp | $-2.832Wind_{12} - 5.631MaxT_5 + 297.615$ | 9.394 |
| 1989 | 56 | prev May HU, prev May Mean Temp, prev May Min Temp, prev Jun Max Temp | $0.220prevHU_5 - 13.618prevMeanT_5 - 1.915prevMinT_5 - 12.750prevMaxT_6 + 748.770$ | 23.392 |
| 1989 | 57 | May Max Temp | $-18.679MaxT_5 + 735.897$ | 23.296 |
| 1989 | 58 | May Max Temp | $-13.491MaxT_5 + 558.509$ | 20.593 |
| 1989 | 61 | May Max Temp | $-17.779MaxT_5 + 692.726$ | 22.653 |
| 1991 | 39 | prev Jun Max Temp, May Max Temp | $-11.405prevMaxT_6 - 10.171MaxT_5 + 789.222$ | 22.955 |
| 1991 | 42 | prev Jun Max Temp | $-14.027prevJunMaxTemp + 530.928$ | 23.838 |
| 1991 | 43 | prev Mar Max Temp, prev Jun Max Temp | $-8.377prevMaxT_3 - 10.811prevMaxT_6 + 760.811$ | 27.269 |
| 1991 | 44 | Mar Rain | $-1.315Rain_3 + 130.226$ | 21.533 |
| 1991 | 45 | May Max Temp | $-14.918MaxT_5 + 612.881$ | 25.544 |
| 1991 | 46 | prev Jun Max Temp | $-13.891prevMaxT_6 + 517.732$ | 23.980 |
| 1991 | 49 | prev Jun Max Temp | $-18.965prevMaxT_6 + 669.803$ | 27.328 |
| 1991 | 50 | prev Jun Max Temp | $-14,979prevMaxT_6 + 558.540$ | 27.249 |
| 1991 | 51 | prev Jun Max Temp | $-17.234prevMaxT_6 + 617.098$ | 17.604 |
| 1993 | 70 | prev May Mean Temp, May Max Temp | $-9.653prevMeanT_5 - 7.224MaxT_5 + 536.839$ | 17.310 |
| 1993 | 71 | Dec Wind, Mar Min Temp, May Max Temp | $-1.373Wind_{12} - 6.19MinT_3 - 6.011MaxT_5 + 398.586$ | 6.018 |
| 1993 | 74 | prev May HU, prev May Mean Temp, Jun Min Temp | $0.435prevHU_5 - 21.196prevMeanT_5 - 4.445MinT_6 + 494.197$ | 28.603 |
| 1993 | 75 | Jun Min Temp | $-8.444MinT_6 + 139.327$ | 28.837 |
| 1995 | 8 | May Max Temp, Apr Rain | $-9.237MaxT_5 - 1.288Rain_4 + 425.548$ | 18.941 |
| 1995 | 12 | Jan Wind, Feb Wind, Mar Wind, May Min Temp | $-2.019Wind_1 + 12.29Wind_2 - 21.758Wind_3 - 2.948MinT_5 + 196.975$ | 32.873 |
| 2000 | 5 | Mar Min Temp | $-7.765MinT_3 + 204.502$ | 14.699 |
| 2000 | 9 | prev Oct Hum, prev Aug Min Temp, May Hum, Jun Hum | $-8.133prevHum_{10} - 15.954prevMinT_8 + 1.877Hum_5 - 0.883Hum_6 + 352.635$ | 16.484 |

Continued on next page

**Table B.4 – continued from previous page**

| Year | Oc | Weather features | Equation for harvest/tree | RMSE |
|---|---|---|---|---|
| 2000 | 10 | prev Sep Mean Temp, prev Jul Max Temp | $-5.062prevMeanT_9 - 10.106prevMaxT_7 + 477.146$ | 11.700 |
| 2000 | 11 | prev Feb Hum, prev Oct Hum, Aug Hum, Sep Hum, Dec Hum, Aug Rain | $1.817prevHum_2 - 2.361prevHum_{10} + 2.659Hum_8 - 0.215Hum_9 - 3.317Hum_{12} - 6.348Rain_8 + 118.232$ | 25.234 |
| 2000 | 13 | Mar Min Temp | $-11.152MinT_3 + 261.098$ | 23.164 |
| 2000 | 14 | prev Oct Hum, May Hum, Jun Hum, Mar Min Temp | $-3.516prevHum_{10} - 0.932Hum_5 - 0.142Hum_6 - 7.217MinT_3 + 335.539$ | 14.675 |
| 2000 | 15 | prev Oct Hum, Apr Hum, May Hum | $0.088prevHum_{10} - 1.969Hum_4 - 1.145Hum_5 + 199.612$ | 19.861 |
| 2002 | 4 | prev Oct Hum, prev Aug Max Temp, Feb Hum, May Hum, Oct Hum, May Rain, Jul Rain | $-2.091prevHum_{10} - 5.34prevMaxT_8 + 1.717Hum_2 - 0.632Hum_5 - 1.806Hum_{10} - 1.338Rain_5 - 2.738Rain_7 + 320.403$ | 33.964 |

## B.2 Detailed results of forward stepwise regression

From Chapter 7 Section 7.1.3 Table B.5 shows the features for each orchard chosen with forward stepwise regression. For these tables, the same naming convention of the features is used as described in Chapter 7, *e.g.* HU abbreviates Heat units.

The RMSE resulting from the forward stepwise regression of each orchard as discussed in Chapter 7 Section 7.1.4 is displayed in Table B.5.

Table B.5: Forward stepwise regression with AICc model selection on features

| Oc | Weather features | RMSE |
|---|---|---|
| 17 | May Max Temp, Mar Min Temp, prev Dec Min Temp, Apr Rain, prev Feb Max Temp, prev Apr HU, Aug Min Temp, prev Jan HU, Oct Wind | 0.0025 |
| 56 | prev May Hum, Aug HU, prev Oct Min Temp, May Max Temp, prev Jul Hum, Jun Max Temp, prev Jan Max Temp, prev Aug Max Temp, Dec Max Temp | 0.0056 |
| 57 | May Max Temp, prev Jun Max Temp, Apr Min Temp, Jan Max Temp, Mar Hum, Jun Hum, Rain Feb, prev Nov Mean Temp, HU May | 0.5439 |
| 58 | Aug Max Temp, Jul Rain, Aug Min Temp, Prev Jan Mean Temp, prev Aug Min Temp, prev Nov Max Temp, Nov Min Temp, May Wind, prev Aug HU | 1.4552 |
| 61 | May Max Temp, Apr Min Temp, Feb Hum, prev Mar Max Temp, prev Nov Mean Temp, Jan Hum, prev Dec Max Temp, Apr Wind, Wind Jul | 4.2796 |
| 90 | Apr Mean Temp, Jan Max Temp, prev Dec Max Temp, HU Jun, Jul Rain, Dec Rain, Rain Feb, Oct Hum, prev Sep Min Temp | 0.0004 |
| 35 | Oct Hum, Apr Max Temp, Rain Sep, Oct HU, prev May Hum, prev Dec Max Temp, Mar Min Temp, prev Mar Hum, Mar HU | 0.0331 |
| 38 | Jul HU, prev Jun Max Temp, Nov Hum, prev Apr HU, prev Feb HU, Wind Sep, Oct Rain, Sep Hum, Mean Temp Nov | 2.4029 |

Continued on next page

**Table B.5 – continued from previous page**

| Oc | Weather features | RMSE |
|---|---|---|
| 33 | Aug HU, Jan HU, Apr Rain, Jun Wind, Oct Min Temp, prev Mar Min Temp, Feb Hum, Jun Max Temp, Dec Max Temp | 0.098 |
| 39 | prev May Hum, Apr Mean Temp, Mar Wind, Jul Min Temp, prev Jan Max Temp, Apr Rain, Mar HU, prev Dec Mean Temp, prev Jun HU | 0.7756 |
| 42 | prev May Hum, Apr Mean Temp, Min Temp Feb, Dec Hum, prev Oct Min Temp, Dec Max Temp, prev Jun Min Temp, prev Mar Max Temp, prev Sep Mean Temp | 0.0109 |
| 47 | Jul Rain, Sep Min Temp, Apr Rain, Jan Max Temp, May Rain, Apr Hum, Jul HU, Rain Sep, prev Jul Hum | 0.7696 |
| 43 | Aug Max Temp, prev Oct Min Temp, prev Jun Hum, May Max Temp, Apr Max Temp, prev Nov Max Temp, Oct Min Temp, Aug Rain, prev May HU | 1.0182 |
| 44 | Aug Min Temp, Jul Rain, Aug HU, prev Nov Hum, Oct HU, Oct Hum, May Max Temp, Jun Mean Temp, prev Jun Hum | 0.1571 |
| 45 | May Max Temp, Jan Mean Temp, Aug Max Temp, Sep Hum, Jul Max Temp, Apr Rain, Jan Min Temp, Nov Max Temp, prev Apr HU | 0.9073 |
| 46 | prev Jun Max Temp, prev Apr Hum, prev Apr Min Temp, Rain Feb, Apr Mean Temp, Mar Wind, prev Nov Hum, prev Jul HU, prev Jul Max Temp | 0.1776 |
| 49 | prev May Hum, prev Feb Hum, prev Nov Max Temp, prev Aug Max Temp, Jan Max Temp, May Min Temp, prev Aug Mean Temp, Aug Min Temp, prev Apr HU | 0.2402 |
| 50 | prev Jun Max Temp, prev Aug Min Temp, Jan HU, Mar Min Temp, Feb Max Temp, prev Jan Hum, prev Jun Min Temp, Feb Hum, Nov HU | 0.0005 |
| 51 | prev Jun Max Temp, Apr Mean Temp, Prev Feb Mean Temp, Mar Rain, Jun Min Temp, May Wind, prev May Min Temp, Oct Min Temp, Aug Rain | 0.0069 |
| 70 | Aug HU, prev Oct Min Temp, Aug Wind, May Min Temp, Dec Rain, Aug Min Temp, prev Jan Max Temp, Sep Max Temp, prev Feb Hum | 0.2828 |
| 71 | Dec Wind, Mar Max Temp, Jul HU, Jul Max Temp, Mar Hum, prev Jul Hum, Dec HU, Nov Rain, Nov Min Temp | 0.0604 |
| 74 | Aug Max Temp, prev Oct Min Temp, Mar Max Temp, prev Aug Hum, Jan Rain, prev Aug Min Temp, Jun Max Temp, Dec Wind, Sep Mean Temp | 0.5591 |
| 75 | prev Oct Min Temp, May Max Temp, prev Dec Hum, Apr Min Temp, Aug Hum, Dec Rain, prev Nov Max Temp, prev Sep Hum, Jan Wind | 0.0893 |
| 8 | May Max Temp, prev Oct Min Temp, Jul Hum, Dec Mean Temp, Jan Max Temp, prev Aug Hum, prev Jul Max Temp, Feb Mean Temp, Jan Rain | 0.1379 |
| 12 | Oct Hum, Apr Max Temp, May Min Temp, prev Jul Mean Temp, Jun Min Temp, Oct HU, Oct Wind, prev Feb Min Temp, May Max Temp | 0.02 |
| 5 | Mar Min Temp, prev Nov Min Temp, Sep Min Temp, prev Feb HU, May Rain, Jun Hum, Aug Min Temp, prev May Hum, prev Jun HU | 0.0135 |
| 9 | prev Oct Mean Temp, May Max Temp, prev Mar Min Temp, prev Nov Min Temp, Dec Hum, May Hum, prev Sep HU, prev Feb Max Temp, Apr Max Temp | 0.1166 |
| 10 | prev Jul Max Temp, prev Sep Mean Temp, May Mean Temp, prev Feb Min Temp, prev Dec Min Temp, prev Sep HU, Oct Max Temp, Dec Wind, Feb HU | 0.8654 |
| 11 | Aug Rain, May Rain, Nov Max Temp, Jul HU, Jun Hum, Jan Mean Temp, prev Feb Hum, Apr Wind, prev Oct HU | 0.0011 |
| 13 | Mar Min Temp, Jul Min Temp, prev Jan Max Temp, Jun Mean Temp, Jan HU, prev Nov Hum, Min Temp Dec, Feb Mean Temp, Oct Mean Temp | 0.0454 |
| 14 | Nov HU, Mar Min Temp, prev Oct Max Temp, prev Jun HU, Aug Rain, prev Jan Min Temp, Jun Rain, Jun Min Temp, prev Mar Hum | 0.0262 |
| 15 | Jun Max Temp, Mar Max Temp, prev Oct Min Temp, prev Oct Max Temp, Mar Mean Temp, prev Sep Max Temp, prev Jun HU, Sep Hum, prev Jun Hum | 0.0193 |
| 4 | Jan Wind, Feb Mean Temp, prev Aug HU, prev Oct Max Temp, prev Oct Mean Temp, Nov Max Temp, prev May HU, prev Aug Max Temp, May Max Temp | 0.0003 |

## B.3 Detailed results of elastic net regression

Chapter 7 Section 7.1.5 discusses the implementation of elastic net regression and Table B.6 displays these equations for each orchard. The 'Year' column denotes the year the trees in the orchard were planted.

Table B.6: Elastic net regression and AICc model selection on features

| Year | Oc | Weather features | Equation for harvest/tree | RMSE | AdjR2 |
|---|---|---|---|---|---|
| 1978 | 17 | May Max Temp, Mar Min Temp, prev May Hum | $-7.021MaxT_5 - 3.000MinT_3 + 0.540prevHum_5 + 351.788$ | 7.608 | 0.880 |
| 1978 | 90 | Apr Mean Temp, prev Apr Mean Temp, prev Nov Min Temp, Oct Hum | $7.847MeanT_4 + 6.683prevMeanT_4 + 2.917prevMinT_{11} + 3.055Hum_{10} - 334.183$ | 8.230 | 0.938 |
| 1989 | 56 | prev May Hum, prev Oct Min Temp, Aug Max Temp | $2.358prevHum_5 + 13.550prevMinT_{10} + 5.961MaxT_8 - 321.033$ | 10.852 | 0.947 |
| 1989 | 57 | May Max Temp, Aug Mean Temp, prev May Hum, Apr Min Temp, Jul Mean Temp | $-11.587MaxT_5 + 4.714prevMeanT_8 + 2.113MinT_4 + 2.508prevHum_5 + 8.950MeanT_7 + 157.605$ | 13.653 | 0.954 |
| 1989 | 58 | May Max Temp, Aug Min Temp, Jul Rain | $-7.831MaxT_5 + 14.886MinT_8 + 3.958Rain_7 + 282.608$ | 11.494 | 0.893 |
| 1989 | 61 | Apr Min Temp, May Max Temp, Aug Max Temp, Jul Rain, Mar Hum, Apr Max Temp, prev Apr Mean Temp | $5.996MinT_4 - 5.387MaxT_5 + 3.903MaxT_8 + 3.835Rain_7 + 2.513Hum_3 + 3.365MaxT_4 + 3.899prevMeanT_4 - 213.630$ | 7.925 | 0.991 |
| 1991 | 33 | Aug HU, prevJan Hum, Nov Hum, Jul Rain | $0.270Hum_{11} + 0.172HU_8 + MeanT_7HU_1 - 2.402prevHum_1 + 2.376Rain_7159.292$ | 6.276 | 0.860 |
| 1991 | 38 | Jul Rain, prev Apr Mean Temp, prev Oct Min Temp, Jul Mean Temp, prev Apr Min Temp | $4.723Rain_7 + 3.871prevMeanT_4 + 12.820prevMinT_{10} + 2.941MeanT_7 + 7.775prevMinT_4 - 246.297$ | 22.000 | 0.813 |
| 1991 | 39 | prev May Hum, Jul Rain, Apr Mean Temp, Aug Max Temp, prev Apr Mean Temp, May Max Temp, Aug Min Temp | $1.947prevHum_5 + 5.168Rain_7 + 4.567MeanT_4 + 2.797MaxT_8 + 3.123prevMeanT_4 - 0.649MaxT_5 + 3.599MinT_8 - 244.160$ | 2.650 | 0.999 |

Continued on next page

**Table B.6 – continued from previous page**

| Year | Oc | Weather features | Equation for harvest/tree | RMSE | AdjR2 |
|---|---|---|---|---|---|
| 1991 | 42 | Jul Rain, prev May Hum, Aug Max Temp, prev Jul HU, Apr Mean Temp, prev Apr Mean Temp, prev Oct Min Temp | $3.814Rain_7 + 1.663prevHum_5 + 4.201MaxT_8 + 0.178prevHU_7 + 3.507MeanT_4 + 2.269prevMeanT_4 + 3.915prevMinT_{10} - 254.627$ | 3.163 | 0.997 |
| 1991 | 43 | prev Oct Min Temp, Jul Mean Temp, Aug HU, Feb Min Temp, Aug Max Temp | $16.259prevMinT_{10} + 7.041MeanT_7 + 0.164HU_8 + 5.604MinT_2 + 3.636MaxT_8 - 378.551$ | 7.564 | 0.984 |
| 1991 | 44 | Aug Min Temp, Aug Max Temp, prev Sep Mean Temp, Jul Rain, Aug HU, prev Nov Mean Temp | $11.236MinT_8 + 4.502MaxT_8 - 3.212prevMeanT_9 + 1.703Rain_7 + 0.060HU_8 + 2.052prevMeanT_{11} - 84.827$ | 0.613 | 1.000 |
| 1991 | 45 | Aug Max Temp, Jul Rain, Jan HU, Jul Mean Temp, prev Jan Hum, Aug HU, May Max Temp | $8.055MaxT_8 + 4.321Rain_7 - 0.353HU_1 + 4.132MeanT_7 - 1.204prevHum_1 + 0.058HU_8 - 1.821MaxT_5 + 15.423$ | 1.796 | 0.999 |
| 1991 | 46 | Jul Rain, prev Jun Max Temp, prev Apr Min Temp, Apr Max Temp, Oct Hum, Apr Mean Temp | $4.655Rain_7 - 9.128prevMaxT_6 + 6.207prevMinT_4 + 4.293MaxT_4 + 1.804Hum_{10} + 3.193MeanT_4 + 17.140$ | 9.088 | 0.994 |
| 1991 | 47 | Jul Rain, Apr Mean Temp, Jun Min Temp, prev Apr Mean Temp, prev May Hum | $5.956Rain_7 + 7.955MeanT_4 - 6.918MinT_6 + 6.887prevMeanT_4 + 1.500prevHum_5 - 275.99616.270$ | 0.965 | |
| 1991 | 49 | prev Jun Max Temp, Apr Mean Temp, Feb Max Temp, Jul Rain, Oct Min Temp | $-15.934prevMaxT_6 + 12.262MeanT_4 - 8.120MaxT_2 - 0.367Rain_7 + 17.709MinT_{10} + 438.226$ | 11.663 | 0.966 |
| 1991 | 50 | Jul Rain, prev Jun Max Temp, Feb Max Temp, prev Apr Mean Temp, prev Apr HU, Dec HU | $4.043Rain_7 - 13.412prevMaxT_6 - 3.326MaxT_2 - 23.268prevMeanT_4 + 1.125prevHU_4 - 0.293HU_{12} + 1096.211$ | 13.600 | 0.970 |
| 1991 | 51 | prev Jun Max Temp, Apr Mean Temp, prev May Hum, prev Apr Mean Temp, Jan HU, Apr Hum | $-12.127prevMaxT_6 + 5.031MeanT_4 + 1.317prevHum_5 + 4.295prevMeanT_4 - 0.237HU_1 - 0.508Hum_4 + 300.566$ | 2.077 | 0.998 |
| 1993 | 35 | Jul Rain, prev Oct Min Temp, Aug Min Temp, Jul HU | $6.823Rain_7 + 11.163prevMinT_{10} + 11.237MinT_8 + 0.332HU_7 - 44.014$ | 24.249 | 0.850 |
| 1993 | 70 | prev Jun Hum, Aug HU, May Max Temp, prev Oct Min Temp, prev Apr Mean Temp | $0.868prevHum_6 + 0.182HU_8 - 3.865MaxT_5 + 10.685prevMinT_{10} + 4.716prevAprMeanT_4 - 25.962$ | 4.044 | 0.979 |

Continued on next page

**Table B.6 – continued from previous page**

| Year | Oc | Weather features | Equation for harvest/tree | RMSE | AdjR2 |
|---|---|---|---|---|---|
| 1993 | 71 | Mar Min Temp, May Max Temp, prev Jun Hum | $-5.488MinT_3 - 6.206MaxT_5 + 0.655prevHum_6 + 355.272$ | 4.125 | 0.970 |
| 1993 | 74 | Jun Min Temp, prev Apr Mean Temp, Apr Min Temp, prev May Hum, Aug Max Temp | $-6.654MinT_6 + 4.752prevMeanT_4 + 1.864MinT_4 + 0.880prevHum_5 + 4.040MaxT_8 - 186.429$ | 7.772 | 0.943 |
| 1993 | 75 | prev Oct Min Temp, May Max Temp, prev Apr Mean Temp, Jun Min Temp, prev Nov Min Temp | $18.196prevMinT_{10} - 6.966MaxT_5 + 6.510prevMeanT_4 - 4.285MinT_6 + 2.204prevMinT_{11} - 23.849$ | 5.560 | 0.981 |
| 1995 | 8 | May Max Temp, Jul Mean Temp, prev Oct Min Temp, Jun Min Temp, Jul Rain, Aug Max Temp | $-6.676MaxT_5 + 9.127MeanT_6 + 11.460prevMinT_{10} - 3.634MinT_6 + 1.955Rain_7 + 2.079MaxT_8 + 12.500$ | 3.799 | 0.994 |
| 1995 | 12 | Oct Hum, prev Apr Mean Temp, Jul Rain, prev Jul Max Temp | $4.585Hum_{10} + 7.608prevMeanT_4 + 5.641Rain_7 - 6.304prevMaxT_6 - 34.618$ | 14.937 | 0.943 |
| 2000 | 5 | Mar Min Temp, May Max Temp, Sep Min Temp, Nov HU, prev May Hum, prev Dec Mean Temp | $-6.380MinT_3 + 4.685MaxT_5 + 4.797MinT_9 + 0.111HU_{11} + 0.371prevHum_5 - 1.819prevMeanT_{12} + 0.085$ | 1.069 | 0.999 |
| 2000 | 9 | May Mean Temp, Jun Wind, prev Aug Min Temp, Mar Min Temp | $-11.781Wind_6 - 5.425prevMinT_8 + 11.973MeanT_5 - 4.788MinT_3 - 2.168$ | 14.558 | 0.917 |
| 2000 | 10 | May Hum, prev Jul Max Temp, prev Jan Max Temp, Jun Min Temp, Oct HU, Nov Rain | $-1.405Hum_5 - 3.542prevMaxT_7 + 4.720prevMaxT_1 - 1.726MinT_6 - 0.166HU_{10} + 0.372Rain_{11} + 78.837$ | 8.620 | 0.980 |
| 2000 | 11 | Aug Rain, May Rain, Sep Mean Temp, Oct Mean Temp, prev Oct Mean Temp | $-2.019Rain_8 - 1.356Rain_5 + 2.980MeanT_9 + 3.841MeanT_{10} + 1.591prevMeanT_{10} - 87.761$ | 2.729 | 0.994 |
| 2000 | 13 | Mar Min Temp, Nov Mean Temp, Jul Min Temp, Mar Rain, prev Nov Max Temp, prev Oct HU | $-9.468MinT_3 + 5.020MeanT_{11} + 5.571MinT_7 - 0.238Rain_3 + 2.608prevMaxT_{11} + 0.090prevHU_{10} - 40.286$ | 5.289 | 0.995 |
| 2000 | 14 | Nov HU, Mar Min Temp, prev Oct Max Temp, prev Oct Mean Temp, prev Nov Mean Temp | $0.336HU_{11} - 3.963MinT_3 + 2.501prevMaxT_{10} + 3.886prevMeanT_{10} - 3.284prevMeanT_{11} - 45.629$ | 1.746 | 0.998 |

Continued on next page

**Table B.6 – continued from previous page**

| Year | Oc | Weather features | Equation for harvest/tree | RMSE | AdjR2 |
|---|---|---|---|---|---|
| 2000 | 15 | May Hum, Sep Mean Temp, prev May Hum, prev Sep Mean Temp | $-1.624Hum_5 + 4.495MeanT_9 + 0.803prevHum_5 - 3.401prevMeanT_9 + 92.763$ | 7.223 | 0.955 |
| 2002 | 4 | Jan Wind, Oct Mean Temp, prev Apr Hum, prev Aug Max Temp | $9.648MeanWind_1 + 4.738MeanT_{10} + 2.014prevHum_4 - 2.538prevMaxT_8 - 112.366$ | 4.885 | 0.970 |

## B.4 Partial least squares regression on the weather features

PLSR is applied to every orchard by considering $X$ the constructed weather dataset. The results are presented in Chapter 7 Section 7.1.7.

Table B.7: Features with VIP score $> 1.8$ for PLS on all orchards

| Orchard | Weather features | CV RMSE |
|---|---|---|
| 17 | prev Jun Hum, prev Jul Hum, Aug HU, Jan Hum, Jan Min Temp, Aug Mean Temp, May Max Temp, Aug Max Temp, Wind Dec | 18.13 |
| 56 | prev Mar Hum, prev May Hum, prev May HU, prev May Mean Temp, prev Jun Max Temp | 27.17 |
| 57 | Aug HU, Aug Mean Temp, Apr Max Temp, May Max Temp, Aug Max Temp | 40.93 |
| 58 | Aug HU, Aug Min Temp, Apr Mean Temp, Aug Mean Temp, May Max Temp, Aug Max Temp, Mar Rain | 25.33 |
| 61 | Apr HU, Aug HU, Apr Min Temp, Apr Mean Temp, Aug Mean Temp, Apr Max Temp, May Max Temp, Aug Max Temp, Jul Rain | 41.23 |
| 90 | prev Apr HU, prev Aug HU, prev Apr Mean Temp, prev Aug Max Temp, prev Dec Max Temp, Apr HU, Apr Mean Temp | 24.23 |
| 35 | prev Jun Max Temp, Mar HU, Oct Hum, Jul Rain | 36.88 |
| 38 | Jul HU, Jul Mean Temp | 29.42 |
| 33 | prev Jan Hum, prev Oct HU, prev Oct Mean Temp, prev Sep Min Temp, Aug HU, Nov Hum, Aug Mean Temp, May Max Temp, Aug Max Temp | 15.54 |
| 39 | prev May Hum, prev Jun Max Temp, May Max Temp, Aug Max Temp, Jul Rain | 35.68 |
| 42 | prev May Hum, prev Jun Max Temp, Feb Humid, Aug Max Temp, Jul Rain | 31.39 |
| 47 | prev May Hum, Oct Hum, Jul Rain | 33.88 |
| 43 | prev May Min Temp, prev Mar Max Temp, prev Jun Max Temp, Aug HU, Feb Min Temp, Aug Mean Temp, May Max Temp, Aug Max Temp | 32.35 |
| 44 | prev Oct HU, Aug HU, Aug Min Temp, Aug Mean Temp, Aug Max Temp | 21.00 |
| 45 | prev Oct HU, Aug HU, Feb Min Temp, Apr Mean Temp, Aug Mean Temp, May Max Temp, Aug Max Temp, Jul Rain | 34.27 |
| 46 | prev Jun Max Temp, Mar HU, Aug HU, Feb Min Temp, Mar Mean Temp, May Max Temp, Aug Max Temp, Jul Rain | 28.96 |
| 49 | prev May Hum, prev Jun Max Temp, Feb Max Temp, Jul Rain | 47.27 |
| 50 | prev May Hum, prev Jul Mean Temp, prev Jun Max Temp, Mar HU, Mar Mean Temp, Jul Rain | 34.54 |
| 51 | prev May Hum, prev Jun Max Temp, Apr Hum, Aug Max Temp, Mar Rain, Jul Rain | 34.71 |

Continued on next page

Table B.7 – continued from previous page

| Orchard | Weather features | CV RMSE |
|---|---|---|
| 70 | prev Jun Hum, prev May HU, prev May Mean Temp, prev Jun Max Temp, Aug HU, Aug Mean Temp, May Max Temp, Aug Max Temp | 22.59 |
| 71 | prev Jun Hum, prev Jul Hum, Jan Hum, Mar Min Temp, May Max Temp, Aug Max Temp, Dec Wind | 18.01 |
| 74 | prev May Hum, prev Jun Hum, prev May HU, prev May Mean Temp, Aug HU, Mar Hum, Jun Min Temp, Aug Mean Temp, Aug Max Temp | 19.21 |
| 75 | prev Oct Min Temp, Mar Hum | 30.13 |
| 8 | prev Oct Min Temp, Mar Hum, May Max Temp, Aug Max Temp, Apr Rain | 33.02 |
| 12 | Oct Hum, Jul Rain | 30.48 |
| 5 | Mar Min Temp | 26.45 |
| 9 | prev Oct HU, prev Oct Mean Temp | 28.92 |
| 10 | prev Jan Max Temp, prev Jul Max Temp, Aug Min Temp | 24.55 |
| 11 | Aug Rain | 14.49 |
| 13 | prev Oct HU, prev Oct Mean Temp, prev Nov Max Temp, Nov HU, Mar Min Temp, Apr Min Temp, Nov Mean Temp | 34.25 |
| 14 | prev Oct HU, prev Oct Mean Temp, Nov HU, Mar Min Temp, Nov Mean Temp | 23.94 |
| 15 | May HU, Sep HU, Apr Hum, May Hum, Nov Min Temp, Sep Mean Temp, Jun Max Temp | 23.76 |
| 4 | Jan Wind | 17.33 |

## B.5 Correlation-based approach on weather and bunch data to predict yield

Tables B.8 and B.9 display weather and bunch features for which a correlation coefficient value higher than 0.6 and lower than -0.6 respectively. The 'Year' column is displayed again for an indication of the orchard's age.

Table B.8: Correlation threshold of 0.6 and higher with bunches data included in features

| Year | oc | Weather features | RMSE |
|---|---|---|---|
| 1978 | 17 | Jan Min Temp, Aug Max Temp | 11.600 |
| 1978 | 90 | **bunch mass**, prev Apr HU, prev Aug HU, prev Apr Mean Temp, prev Aug Mean Temp, prev Aug Max Temp, prev Dec Max Temp, Apr HU, Apr Mean Temp | 6.092 |
| 1989 | 56 | **bunch mass**, prev Mar Hum, prev May Hum, prev Jun Hum, prev Oct Min Temp, Aug Max Temp | 9.920 |
| 1989 | 57 | **bunch mass**, Aug HU, Aug Mean Temp, Aug Max Temp | 19.434 |
| 1989 | 58 | **bunch mass**, Aug HU, Aug Min Temp, Apr Mean Temp, Aug Mean Temp, Aug Max Temp | 15.530 |
| 1989 | 61 | **bunch mass**, Aug HU, Apr Mean Temp, Aug Mean Temp, Apr Max Temp, Aug Max Temp | 36.857 |
| 1991 | 33 | **bunch mass**, prev Oct HU, prev Oct Mean Temp, Aug HU, Nov Hum, Aug Mean Temp, Aug Max Temp | 3.312 |
| 1991 | 38 | **bunch mass** | 14.403 |

Continued on next page

Table B.8 – continued from previous page

| Year | Oc | Weather features | RMSE |
|---|---|---|---|
| 1991 | 39 | **bunch mass**, prev May Hum, Aug Max Temp, Jul Rain | 22.199 |
| 1991 | 42 | **bunch mass**, prev May Hum, Feb Hum, Aug Max Temp, Jul Rain | 22.777 |
| 1991 | 43 | **bunch mass**, Aug HU, Feb Min Temp, Aug Mean Temp, Aug Max Temp | 23.790 |
| 1991 | 44 | prev Oct HU, prev Oct Mean Temp, prev Jan Max Temp, prev Oct Max Temp, prev Dec Max Temp, Apr HU, Aug HU, Aug Min Temp, Sep Min Temp, Apr Mean Temp, Aug Mean Temp, Aug Max Temp | 42.243 |
| 1991 | 45 | **bunch mass**, Aug HU, Aug Mean Temp, Aug Max Temp, Jul Rain | 21.253 |
| 1991 | 46 | **bunch mass**, Mar HU, Feb Min Temp, Mar Mean Temp, Aug Max Temp | 21.619 |
| 1991 | 47 | **bunch mass**, prev May Hum, Oct Hum, Jul Rain | 30.637 |
| 1991 | 49 | prev May Hum, Jul Rain | 39.944 |
| 1991 | 50 | **bunch mass**, prev May Hum, Mar HU, Mar Mean Temp, Jul Rain | 59.421 |
| 1991 | 51 | **bunch mass**, prev May Hum | 23.670 |
| 1993 | 35 | **bunch mass** | 15.054 |
| 1993 | 70 | **bunch mass**, prev Jun Hum, Aug HU, Aug Mean Temp, Aug Max Temp | 10.633 |
| 1993 | 71 | prev Jun Hum, prev Jul Hum, Jan Hum, Aug Max Temp | 14.602 |
| 1993 | 74 | **bunch mass**, prev Jun Hum, Aug HU, Mar Hum, Aug Mean Temp, Aug Max Temp | 9.594 |
| 1993 | 75 | **bunch mass**, prev Jun Hum, prev Oct Min Temp, Mar Hum | 17.107 |
| 1995 | 8 | **bunch mass** | 18.035 |
| 1995 | 12 | **bunch mass**, Oct Hum, Nov Hum, Jul Rain | 9.797 |
| 2000 | 5 | **bunch count** | 17.226 |
| 2000 | 9 | Jan Wind, Apr Wind, prev Oct HU, prev Oct Mean Temp, prev Oct Max Temp, May HU, Nov HU, May Mean Temp, Nov Mean Temp, **bunch count** | 90.885 |
| 2000 | 10 | prev Jan Max Temp, Aug Min Temp, Nov Min Temp | 15.390 |
| 2000 | 11 | Jan Wind, Feb Wind, Mar Wind, prev Oct HU, prev Oct Mean Temp, May HU, Oct HU, Oct Mean Temp, Oct Max Temp, Dec Max Temp | 37.877 |
| 2000 | 13 | prev Oct HU, prev Oct Mean Temp, prev Nov Max Temp, Nov HU, Apr Min Temp, Nov Mean Temp, Jan Max Temp | 49.473 |
| 2000 | 14 | Apr Wind, prev Oct HU, prev Oct Mean Temp, prev Oct Max Temp, May HU, Jun HU, Nov HU, Apr Min Temp, May Mean Temp, Jun Mean Temp, Nov Mean Temp, Jan Max Temp, Dec Max Temp | 116.937 |
| 2000 | 15 | prev Oct HU, prev Oct Mean Temp, prev Jan Max Temp, May HU, Sep HU, Nov Min Temp, May Mean Temp, Sep Mean Temp, Jun Max Temp | 556.854 |
| 2002 | 4 | **bunch count**, Jan Wind, Feb Wind, Mar Wind, May HU, Oct HU, Oct Mean Temp, Jan Max Temp, Oct Max Temp | 13.922 |

Table B.9: Correlation threshold of -0.6 and lower, with bunch count and bunch mass included in features

| Year | oc | Weather features | RMSE |
|---|---|---|---|
| 1978 | 17 | Wind Dec, May Max Temp | 9.394 |
| 1989 | 56 | Jan Wind, Feb Wind, Mar Wind, Oct Wind, Nov Wind, prev May HU, prev May Mean Temp, prev May Min Temp, prev Jun Max Temp | 640.439 |

Continued on next page

Table B.9 – continued from previous page

| Year | Oc | Weather features | RMSE |
|---|---|---|---|
| 1989 | 57 | May Max Temp | 23.296 |
| 1989 | 58 | May Max Temp | 20.593 |
| 1989 | 61 | May Max Temp | 22.653 |
| 1991 | 39 | prev Jun Max Temp, May Max Temp | 22.955 |
| 1991 | 42 | prev Jun Max Temp | 23.838 |
| 1991 | 43 | prev Mar Max Temp, prev Jun Max Temp | 27.269 |
| 1991 | 44 | Mar Rain | 21.533 |
| 1991 | 45 | May Max Temp | 25.544 |
| 1991 | 46 | prev Jun Max Temp | 23.980 |
| 1991 | 49 | prev Jun Max Temp | 27.328 |
| 1991 | 50 | prev Jun Max Temp | 27.249 |
| 1991 | 51 | prev Jun Max Temp | 17.604 |
| 1993 | 35 | **bunch count** | 28.503 |
| 1993 | 70 | prev May Mean Temp, May Max Temp | 17.310 |
| 1993 | 71 | Dec Wind, Mar Min Temp, May Max Temp | 6.018 |
| 1993 | 74 | prev May HU, prev May Mean Temp, Jun Min Temp | 28.603 |
| 1993 | 75 | Jun Min Temp | 28.837 |
| 1995 | 8 | May Max Temp, Apr Rain | 18.941 |
| 1995 | 12 | Jan Wind, Feb Wind, Mar Wind, May Min Temp | 32.873 |
| 2000 | 5 | Mar Min Temp | 14.699 |
| 2000 | 9 | prev Oct Hum, prev Aug Min Temp, May Hum, Jun Hum | 16.484 |
| 2000 | 10 | prev Sep Mean Temp, prev Jul Max Temp | 11.700 |
| 2000 | 11 | prev Feb Hum, prev Oct Hum, Aug Hum, Sep Hum, Dec Hum, Aug Rain | 25.234 |
| 2000 | 13 | Mar Min Temp | 23.164 |
| 2000 | 14 | prev Oct Hum, May Hum, Jun Hum, Mar Min Temp | 14.675 |
| 2000 | 15 | prev Oct Hum, Apr Hum, May Hum | 19.861 |
| 2002 | 4 | prev Oct Hum, prev Aug Max Temp, Feb Hum, May Hum, Oct Hum, May Rain, Jul Rain | 33.964 |

## B.6 Forward stepwise regression on weather and bunch features

Table B.10 displays the results from including the 'bunch mass' and 'bunch count' features for each orchard in the predictor dataset and selecting the best subset of features with forward stepwise regression. The small RMSE and high $R^2$ values are attributed to the large number of features chosen for the small dataset.

Table B.10: Features chosen with forward stepwise regression for all the orchards with bunches data included

| Year | Oc | Features | RMSE | Adj$R^2$ |
|---|---|---|---|---|
| 1978 | 17 | May Max Temp, Apr Min Temp, prev Dec Min Temp, Apr Rain, prev Feb Max Temp, prev Apr HU, Aug Min Temp, prev Jan HU, Oct Wind | 0.0025 | 0.99982 |
| 1978 | 90 | **bunch count**, **bunch mass**, Apr Mean Temp, Jan Max Temp, Jul Wind, prev Jun Min Temp, prev Mar HU, prev May Mean Temp, Sep Min Temp | 0.0258 | 0.999989 |

Continued on next page

Table B.10 – continued from previous page

| Year | Oc | Features | RMSE | Adj$R^2$ |
|---|---|---|---|---|
| 1989 | 56 | prev May Hum, Aug HU, prev Oct Min Temp, May Max Temp, prev Jul Hum, Jun Max Temp, prev Jan Max Temp, prev Aug Max Temp, Dec Max Temp | 0.0056 | 1.0 |
| 1989 | 57 | **bunch mass**, May Max Temp, Apr Min Temp, prev Apr Hum, prev Jul Max Temp, Jan Min Temp, Apr Rain, prev Jun Max Temp, prev May Hum | 0.3231 | 0.99948 |
| 1989 | 58 | **bunch mass**, Aug Max Temp, prev Jan Mean Temp, prev Jul Max Temp, prev Jul Min Temp, Feb Min Temp, Nov Min Temp, prev Feb Hum, prev Nov Mean Temp | 2.5777 | 0.996308 |
| 1989 | 61 | **bunch mass**, May Max Temp, Apr Min Temp, Feb Hum, Nov Max Temp, prev Jan HU, Mar Min Temp, prev Dec Hum, Feb Rain | 0.0169 | 0.999998 |
| 1991 | 33 | Aug HU, Jan HU, Apr Rain, Jun Wind, Oct Min Temp, prev Mar Min Temp, Feb Hum, Jun Max Temp, Dec Max Temp | 0.098 | 0.99997 |
| 1991 | 38 | **bunch mass**, Nov Max Temp, prev Jun Min Temp, prev Jul Min Temp, Jan Rain, prev Aug Min Temp, prev May Max Temp, prev Oct Max Temp, Mar Hum | 0.0012 | 1.0 |
| 1991 | 39 | **bunch mass**, prev May Hum, Apr Mean Temp, prev Dec Hum, prev Mar HU, prev Sep Hum, Apr Rain, prev Oct Min Temp, May Wind | 0.1432 | 0.99988 |
| 1991 | 42 | **bunch mass**, prev May Hum, prev Feb Hum, Aug Wind, prev Sep Max Temp, prev Jun Hum, Jul HU, prev Nov Min Temp, May Min Temp | 0.2316 | 0.99994 |
| 1991 | 43 | **bunch mass**, Aug Max Temp, prev Aug HU, prev Sep Min Temp, prev Apr Hum, Dec Rain, Nov Hum, prev Nov HU, prev Sep HU | 0.0886 | 0.99993 |
| 1991 | 44 | Aug Min Temp, Jul Rain, Aug HU, prev Nov Hum, Oct HU, Oct Hum, May Max Temp, Jun Mean Temp, prev Jun Hum | 0.1571 | 0.998887 |
| 1991 | 45 | **bunch count**, May Max Temp, Jan Mean Temp, Aug Max Temp, Sep Hum, Jul Max Temp, Apr Rain, Jan Min Temp, Nov Max Temp | 0.9073 | 0.99963 |
| 1991 | 46 | **bunch mass**, Jun Max Temp, Feb Min Temp, Apr Mean Temp, prev May Hum, prev Sep Hum, prev Jun Min Temp, prev Jun Hum, Jul Min Temp | 0.5188 | 0.999482 |
| 1991 | 47 | **bunch mass**, **bunch count**, prev Nov Max Temp, prev May Hum, Jan Mean Temp, Oct Hum, prev Aug Min Temp, Aug Wind, Apr Min Temp | 1.1383 | 0.997496 |
| 1991 | 49 | prev May Hum, prev Feb Hum, prev Nov Max Temp, prev Aug Max Temp, Jan Max Temp, May Min Temp, prev Aug Mean Temp, Aug Min Temp, prev Apr HU | 0.2402 | 0.999973 |
| 1991 | 50 | **bunch mass**, prev Jul Mean Temp, prev Jun Max Temp, prev Aug Min Temp, Mar Max Temp, Jun Min Temp, Sep Mean Temp, Apr Max Temp, prev Jan Min Temp | 0.1081 | 0.999951 |
| 1991 | 51 | prev Jun Max Temp, Apr Mean Temp, prev Feb Mean Temp, Mar Rain, Jun Min Temp, May Wind, prev May Min Temp, Oct Min Temp, Aug Rain | 0.0069 | 0.999995 |
| 1993 | 35 | **bunch mass**, Jul Min Temp, Dec Min Temp, prev Jan HU, prev Feb Min Temp, Jul Hum, Sep Max Temp, prev Jan Min Temp, Sep Rain | 0.0067 | 1.0 |
| 1993 | 70 | **bunch mass**, prev Oct Mean Temp, prev Jun Hum, prev Apr Min Temp, prev Jul Hum, Jan Min Temp, prev Aug Hum, Dec HU, Aug HU | 0.8128 | 0.999091 |

Continued on next page

Table B.10 – continued from previous page

| Year | Oc | Features | RMSE | Adj$R^2$ |
|---|---|---|---|---|
| 1993 | 71 | Dec Wind, Mar Max Temp, Jul HU, Jul Max Temp, Mar Hum, prev Jul Hum, Dec HU, Nov Rain, Nov Min Temp | 0.0604 | 0.99999 |
| 1993 | 74 | **bunch mass**, Aug Max Temp, prev Oct Min Temp, Mar Max Temp, prev Aug Hum, Jan Rain, Aug Wind, Oct HU, Aug Mean Temp | 0.6542 | 0.999782 |
| 1993 | 75 | **bunch mass**, prev Oct Min Temp, Aug Rain, Oct Min Temp, Jun Max Temp, prev Jan Mean Temp, Apr Wind, Jan Mean Temp, Apr Max Temp | 0.0114 | 0.999998 |
| 1995 | 8 | **bunch mass**, Apr Rain, prev Oct HU, prev Jul Max Temp, prev May Min Temp, Jun Mean Temp, Dec Max Temp, prev Nov Mean Temp, prev Aug Hum | 0.0559 | 0.999993 |
| 1995 | 12 | **bunch mass**, Oct Hum, May Rain, prev Jan Hum, Aug Max Temp, Oct Min Temp, Dec Rain, Jan Mean Temp, prev Mar HU | 0.5422 | 0.999948 |
| 2000 | 5 | **bunch count**, Mar Min Temp, prev Nov Min Temp, Sep Min Temp, prev Feb HU, May Rain, Jun Hum, prev Oct Min Temp, Apr Max Temp | 0.0296 | 0.999999 |
| 2000 | 9 | prevOct Mean Temp, May Max Temp, prev Mar Min Temp, prev Nov Min Temp, Dec Hum, May Hum, prev Sep HU, prev Feb Max Temp, Apr Max Temp | 0.1166 | 0.999927 |
| 2000 | 10 | **bunch mass**, prev Jul Max Temp, prev Sep Mean Temp, May Mean Temp, prev Feb Min Temp, Feb Rain, May Min Temp, prev Jan Max Temp, prev Apr HU | 0.7289 | 0.999682 |
| 2000 | 11 | Aug Rain, May Rain, Nov Max Temp, Jul HU, Jun Hum, Jan Mean Temp, prev Feb Hum, Apr Wind, prev Oct HU | 0.0011 | 0.999997 |
| 2000 | 13 | Mar Min Temp, Jul Min Temp, prev Jan Max Temp, Jun Mean Temp, Jan HU, prev Nov Hum, Dec Min Temp, Feb Mean Temp, Oct Mean Temp | 0.0454 | 0.999991 |
| 2000 | 14 | Nov HU, Mar Min Temp, prev Oct Max Temp, prev Jun HU, Aug Rain, prev Jan Min Temp, Jun Rain, Jun Min Temp, prev Mar Hum | 0.0262 | 0.999999 |
| 2000 | 15 | Jun Max Temp, Mar Max Temp, prev Oct Min Temp, prev Oct Max Temp, Mar Mean Temp, prev Sep Max Temp, prev Jun HU, Sep Hum, prev Jun Hum | 0.0193 | 1.0 |
| 2002 | 4 | **bunch count**, prev Sep Min Temp, Nov Max Temp, Jun Rain, Aug Rain, Oct Wind, Jan Wind, Nov Hum, prev Sep HU | 0.0325 | 0.999683 |

## B.7 Elastic net regression on the weather and bunch features

Table B.11 displays the selected features after including the bunch data for each orchard in the input dataset for elastic net regression.

Table B.11: Features chosen with elastic net for sample orchards with bunches data included

| Oc | Features | RMSE | Adj$R^2$ |
|---|---|---|---|
| 17 | May Max Temp, Mar Min Temp, prev May Hum | 7.608 | 0.8805 |
| 90 | **bunch mass**, Apr Mean Temp, prev Aug Max Temp, Mar Mean Temp, prev Nov Min Temp, prev Feb Mean Temp, Jan HU | 2.868 | 0.9979 |

Continued on next page

Table B.11 – continued from previous page

| Oc | Features | RMSE | Adj$R^2$ |
|---|---|---|---|
| 5 | Mar Min Temp, May Max Temp, Sep Min Temp, Nov HU, prev May Hum, prev Dec Mean Temp | 1.069 | 0.9992 |
| 33 | **bunch mass**, prev Jan Hum, prev Oct HU, prev Sep Min Temp, Apr Mean Temp, May Max Temp | 3.865 | 0.9837 |
| 38 | **bunch mass**, Jul HU, Nov Min Temp, Nov Max Temp, prev Jul HU, Sep Max Temp, prev Oct Min Temp, Jul Min Temp | 0.230 | 1.0 |
| 39 | **bunch mass**, prev May Hum, Jul Rain, Apr Mean Temp, Mar Rain, Aug Max Temp | 1.955 | 0.9976 |
| 42 | **bunch mass**, prev May Hum, Jul Rain, Aug Max Temp, prev Jul HU, Apr Mean Temp, Jul Min Temp | 4.078 | 0.998 |
| 43 | **bunch mass**, prev Oct Min Temp, Aug Max Temp, Feb Min Temp | 14.431 | 0.8913 |
| 44 | **bunch mass**, Aug Min Temp, Aug Max Temp, prev Sep Hum, prev Oct HU, Nov Min Temp, Apr Mean Temp | 1.383 | 0.9993 |
| 45 | **bunch mass**, Aug Max Temp, Jul Rain, prev Jan Hum, Apr Mean Temp, Jan Mean Temp, Feb Min Temp | 3.735 | 0.9991 |
| 46 | **bunch mass**, Feb Min Temp, Apr Mean Temp, prev Jun Max Temp, Feb Mean Temp, Apr Hum, prev Sep Hum | 2.251 | 0.9986 |
| 47 | **bunch mass**, prev May Hum, Apr Mean Temp | 15.018 | 0.8501 |
| 49 | **bunch count**, prev Jun Max Temp, Apr Mean Temp, Jul Rain, Feb Max Temp | 28.235 | 0.9218 |
| 50 | **bunch mass**, **bunch count**, prev Jul Mean Temp, prev May Hum | 17.965 | 0.8866 |
| 51 | **bunch mass**, prev Jun Max Temp, Apr Mean Temp, prev May Hum, Apr Hum, prev Feb Mean Temp | 7.441 | 0.9891 |
| 56 | **bunch mass**, prev Jun Max Temp, prev May Hum, Aug Max Temp, prev Jun Hum, Apr Rain | 9.187 | 0.9848 |
| 71 | **bunch count**, Mar Min Temp, May Max Temp, prev Jun Hum, prev Aug Mean Temp | 1.685 | 0.9978 |
| 74 | **bunch mass**, Jun Min Temp, Aug Max Temp, prev Jun Hum, prev Apr Mean Temp, Sep Max Temp | 4.483 | 0.9847 |
| 75 | **bunch mass**, prev Oct Min Temp, Jun Min Temp, prev Jan HU, Apr Min Temp, prev Jul Max Temp, Jul Mean Temp | 0.908 | 0.9997 |
| 57 | **bunch mass**, May Max Temp, Apr Min Temp, Aug Max Temp, prev May Hum, prev Mar Max Temp | 5.836 | 0.9924 |
| 58 | **bunch mass**, Aug Max Temp, Aug Min Temp, Jul Rain | 4.905 | 0.9867 |
| 61 | **bunch mass**, **bunch count**, May Max Temp | 12.567 | 0.9056 |
| 35 | **bunch mass**, Nov Min Temp, Jul Min Temp | 8.594 | 0.9566 |
| 70 | **bunch mass**, Aug HU, prev Jun Hum, prev Oct Min Temp, Jul Mean Temp, prev Apr Mean Temp | 6.292 | 0.9898 |
| 8 | **bunch mass**, Apr Rain, Aug Max Temp, prev Jan Hum | 8.667 | 0.9529 |
| 12 | **bunch mass**, **bunch count**, Oct Hum, prev Jul Max Temp, Jul Rain, prev Feb Min Temp | 5.254 | 0.9950 |
| 9 | May Mean Temp, Jun Wind, prev Aug Min Temp, Mar Min Temp | 14.558 | 0.9175 |
| 10 | May Hum, prev Jul Max Temp, prev Jan Max Temp, Jun Min Temp, Oct HU, Nov Rain | 8.620 | 0.98 |
| 11 | Aug Rain, May Rain, Sep Mean Temp, Oct Mean Temp, prev Oct Mean Temp | 2.729 | 0.994 |
| 13 | Mar Min Temp, Nov Mean Temp, Jul Min Temp, Mar Rain, prev Nov Max Temp, prev Oct HU | 5.289 | 0.9954 |
| 14 | Nov HU, Mar Min Temp, prev Oct Max Temp, prev Oct Mean Temp, prev Nov Mean Temp | 1.746 | 0.9977 |
| 15 | May Hum, Sep Mean Temp, prev May Hum, prev Sep Mean Temp | 7.223 | 0.9546 |

Continued on next page

Table B.11 – continued from previous page

| Oc | Features | RMSE | Adj$R^2$ |
|---|---|---|---|
| 4 | **bunch count**, prev Oct Mean Temp, Jun Wind, Jan Min Temp, prev Aug Max Temp, May Hum | 1.099 | 0.9986 |

## B.8 Partial least squares regression on the weather and bunch data

The implementation of PLS on all the orchards with bunches data included in the list of possible features, to predict yield was done in Subsection 7.4.5. Table B.12 shows features with a VIP score greater than 2.

## B.9 Developed prediction models after comparing with current model

### B.9.1 Cross-validation of models predicting individual orchard yields

Table B.13 displays the leave-one-out cross-validated RMSE for each orchard in kg per tree. The (resubstituted) RMSE of the current method is also displayed purely for reference.

Table B.13: LOOCV errors in kg per tree of Models A, D, E and F for the individual orchards

| Oc | Current method RMSE | Model A | Model D | Model E | Model F |
|---|---|---|---|---|---|
| 17 | 13.70 | 10.42 | 16.46 | 15.27 | 12.49 |
| 56 | 26.40 | 18.09 | 27.50 | 13.88 | 12.01 |
| 57 | 23.81 | 23.30 | 24.76 | 16.46 | 22.23 |
| 58 | 24.88 | 20.67 | 21.51 | 16.52 | 11.72 |
| 61 | 29.42 | 22.65 | 23.74 | 12.57 | 20.75 |
| 90 | 16.37 | 15.99 | 29.00 | 12.68 | 12.16 |
| 35 | 66.21 | 30.00 | 16.58 | 22.05 | 17.54 |
| 38 | 13.58 | 26.92 | 14.08 | 16.35 | 14.76 |
| 33 | 40.13 | 12.53 | 15.53 | 14.54 | 11.64 |
| 39 | 18.16 | 25.91 | 32.89 | 28.93 | 20.52 |
| 42 | 16.20 | 24.14 | 21.93 | 20.01 | 19.42 |
| 47 | 16.35 | 23.75 | 29.60 | 22.32 | 17.95 |
| 43 | 19.12 | 27.03 | 26.27 | 20.87 | 20.08 |
| 44 | 19.59 | 17.51 | 26.75 | 17.62 | 16.58 |
| 45 | 22.00 | 25.54 | 32.60 | 30.85 | 25.92 |
| 46 | 16.42 | 23.98 | 26.63 | 20.47 | 18.96 |
| 49 | 22.63 | 27.33 | 38.90 | 15.98 | 39.98 |
| 50 | 15.14 | 27.25 | 25.72 | 16.80 | 26.54 |
| 51 | 17.16 | 17.60 | 27.95 | 21.60 | 18.86 |
| 70 | 15.26 | 16.48 | 20.37 | 24.90 | 19.43 |
| 71 | 16.70 | 10.17 | 15.07 | 8.47 | 8.25 |
| 74 | 16.82 | 18.61 | 16.11 | 14.85 | 16.82 |
| 75 | 15.07 | 15.82 | 22.02 | 12.08 | 13.32 |
| 8 | 12.20 | 27.02 | 13.90 | 12.91 | 21.45 |

Continued on next page

Table B.13 – continued from previous page

| Oc | Current method RMSE | Model A | Model D | Model E | Model F |
|---|---|---|---|---|---|
| 12 | 11.96 | 29.12 | 13.51 | 13.20 | 12.03 |
| 5 | 13.99 | 14.70 | 16.50 | 16.33 | 14.96 |
| 9 | 19.23 | 17.02 | 22.09 | 21.11 | 18.06 |
| 10 | 14.50 | 14.52 | 16.19 | 15.17 | 14.82 |
| 11 | 15.98 | 9.80 | 15.61 | 13.22 | 17.29 |
| 13 | 26.24 | 20.36 | 32.31 | 22.02 | 20.04 |
| 14 | 19.75 | 15.65 | 23.01 | 16.63 | 16.36 |
| 15 | 14.68 | 9.64 | 17.61 | 11.18 | 8.62 |
| 4 | 10.04 | 15.05 | 8.80 | 13.44 | 20.71 |

### B.9.2 Model E on individual orchards

Bunch mass (in kg per bunch) is denoted by 'bmass' and bunch count (in number of bunches per tree) is denoted by 'bcount'. Table B.14 displays the equations of the 33 orchards for Model E (containing individualised weather features, bunch count and bunch mass) predicting harvest per tree. The predicted outcome can simply be multiplied by the number of trees in the orchard for the orchard yield, and the total of the 33 orchards can be predicted by summing the orchard yield predictions.

Table B.15 displays the equations of the 33 orchards for Model F, which includes individualised weather features and bunch mass.

As discussed in Section 7.6, the trade-off between labour of counting the bunches and accuracy of prediction is for the research partner to consider. The results of the section also indicated that bunch count may not be a prudent feature to include as its $p$-value is too high, indicating that it is not a significant feature. For this reason and to save on labour cost and time, Model F is recommended by the researcher. The coefficient values of the weather features especially, generally remain of similar sign and size in Model E and F. This indicates that these features are significant, as they are not greatly affected by the inclusion of other variables.

Table B.12: Features with PLS VIP score greater than 2 for all orchards

| Orchard | Weather features | CV RMSE |
|---|---|---|
| 17 | Jan Min Temp, May Max Temp, Aug Max Temp, Dec Wind | 17.97 |
| 90 | Apr Mean Temp | 22.81 |
| 33 | prev Oct HU, Aug HU, Aug Mean Temp | 14.99 |
| 38 | **bunch mass** | 26.96 |
| 39 | **bunch mass**, prev May Hum, Jul Rain | 34.24 |
| 42 | **bunch mass**, prev May Hum, prev Jun Max Temp, Jul Rain | 30.24 |
| 43 | **bunch mass**, prev Jun Max Temp, Aug HU, Feb Min Temp, Aug Mean Temp, Aug Max Temp | 31.16 |
| 44 | Aug HU, Aug Min Temp, Aug Mean Temp, Aug Max Temp | 20.66 |
| 45 | **bunch mass**, Aug HU, Aug Mean Temp, May Max Temp, Aug Max Temp, Jul Rain | 33.44 |
| 46 | **bunch mass**, prev Jun Max Temp, Feb Min Temp, Aug Max Temp | 27.79 |
| 47 | **bunch mass**, Jul Rain | 31.85 |
| 49 | prev May Hum, prev Jun Max Temp, Jul Rain | 46.11 |
| 50 | **bunch mass**, prev May Hum, prev Jun Max Temp | 32.84 |
| 51 | **bunch mass**, prev May Hum, prev Jun Max Temp | 34.10 |
| 56 | prev May Hum, prev May HU, prev May Mean Temp, prev Jun Max Temp | 26.31 |
| 71 | prev Jun Hum, Mar Min Temp, Dec Wind | 17.51 |
| 74 | Aug Max Temp | 18.88 |
| 75 | **bunch mass** | 28.81 |
| 57 | **bunch mass**, Aug HU, Aug Mean Temp, May Max Temp, Aug Max Temp | 39.07 |
| 58 | **bunch mass**, Aug HU, Aug Mean Temp, May Max Temp, Aug Max Temp | 24.06 |
| 61 | **bunch mass**, Aug HU, Apr Mean Temp, Aug Mean Temp, Apr Max Temp, May Max Temp, Aug Max Temp | 39.47 |
| 35 | **bunch count**, **bunch mass** | 33.85 |
| 70 | **bunch mass**, prev Jun Hum, Aug HU, Aug Mean Temp, Aug Max Temp | 21.65 |
| 8 | **bunch mass**, May Max Temp, Apr Rain | 30.60 |
| 12 | **bunch mass**, Oct Hum | 27.15 |
| 5 | Mar Min Temp | 26.27 |
| 9 | | 28.29 |
| 10 | prev Jan Max Temp, prev Jul Max Temp, Aug Min Temp | 24.57 |
| 11 | | 14.63 |
| 13 | prev Nov Max Temp, Mar Min Temp | 34.31 |
| 14 | | 23.98 |
| 15 | | 23.80 |
| 4 | **bunch count** | 16.60 |

Table B.14: Model E equations for individual orchards

| Oc | Weather features | Equation |
|---|---|---|
| 17 | May and Aug Max Temp | $-6.294MaxT_5 + 1.525MaxT_8 - 2.015bmass + 1.029bcount + 255.084$ |
| 90 | Apr Mean Temp | $9.867MeanT_4 + 7.61bmass - 1.846bcount - 136.077$ |
| 56 | prev May Hum, prev Jun Max Temp | $1.927prevHum_5 - 8.208prevMaxT_6 + 6.586bmass + 1.199bcount + 198.698$ |
| 57 | May Max Temp | $-13.847MaxT_5 + 10.495bmass + 7.145bcount + 374.834$ |
| 58 | May and Aug Max Temp | $0.544MaxT_5 + 7.599MaxT_8 + 9.75bmass - 1.778bcount - 199.929$ |
| 61 | May Max Temp | $-11.064MaxT_5 + 10.557bmass + 7.208bcount + 269.901$ |
| 33 | Aug HU | $0.166HU_8 + 4.342bmass - 1.034bcount + 88.102$ |
| 38 | Jul Mean Temp | $7.572MeanT_7 + 11.961bmass - 1.909bcount - 55.495$ |
| 39 | prev May Hum | $2.263prevHum_5 + 9.53bmass + 5.319bcount - 130.189$ |
| 42 | prev May Hum | $2.034prevHum_5 + 9.58bmass + 8.674bcount - 169.692$ |
| 43 | Aug Max Temp | $6.492MaxT_8 + 10.069bmass + 3.523bcount - 230.204$ |
| 44 | Aug Min Temp | $16.318MinT_8 + 8.894bmass + 3.262bcount - 89.872$ |
| 45 | May Max Temp | $-9.623MaxT_5 + 7.955bmass + 3.497bcount + 320.088$ |
| 46 | prev Jun Max Temp | $-9.672prevMaxT_6 + 11.319bmass + 6.059bcount + 213.168$ |
| 47 | Jul Rain | $4.654Rain_7 + 12.84bmass + 2.509bcount - 25.634$ |
| 49 | prev Jun Max Temp | $-15.463prevMaxT_6 + 12.68bmass + 17.153bcount + 188.697$ |
| 50 | prev Jun Max Temp | $-9.071prevMaxT_6 + 12.707bmass + 12.101bcount + 98.921$ |
| 51 | prev Jun Max Temp | $-12.765prevMaxT_6 + 11.235bmass + 7.812bcount + 278.208$ |
| 35 | prev Oct Min Temp, prev Jun and May Max Temp | $3.587prevMinT_{10} - 1.58MaxT_5 - 3.763prevMaxT_6 + 13.317bmass - 0.722bcount + 151.73$ |
| 70 | May and prev Jun Max Temp, prev Oct Min Temp | $-8.662MaxT_5 + 12.655prevMinT_{10} - 2.886prevMaxT_6 - 0.011bmass + 1.578bcount + 321.922$ |
| 71 | Mar Min Temp, Dec Wind | $-4.2MinT_3 - 4.283Wind_{12} + 5.089bmass + 1.778bcount + 122.956$ |
| 74 | Aug Max Temp | $3.543MaxT_8 + 9.779bmass + 3.984bcount - 153.622$ |
| 75 | prev Oct Min Temp, May Max Temp | $18.343prevMinT_{10} - 4.245MaxT_5 + 7.544bmass + 5.209bcount - 83.338$ |
| 8 | May and Aug Max Temp | $-0.234MaxT_5 + 4.145MaxT_8 + 9.456bmass + 7.488bcount - 191.511$ |
| 12 | Oct Hum, Jul Rain, May Min Temp | $4.192Hum_{10} + 0.74Rain_7 + 0.224MinT_5 + 10.47bmass + 1.935bcount - 119.071$ |
| 5 | Mar Min Temp | $-6.028MinT_3 + 0.989bmass + 2.208bcount + 143.278$ |
| 9 | prev Aug Min Temp, prev Oct Hum | $-10.484prevMinT_8 - 5.598prevHum_{10} + 1.303bmass + 1.298bcount + 262.705$ |
| 10 | prev Jan and prev Jul Max Temp | $-3.963prevMaxT_7 + 5.415prevMaxT_1 + 9.262bmass + 3.306bcount - 139.188$ |
| 11 | May and Aug Rain | $-3.436Rain_8 - 0.839Rain_5 + 7.188bmass + 2.53bcount + 22.415$ |
| 13 | Mar Min Temp, prev Oct HU | $-8.147MinT_3 + 0.327prevHU_{10} - 4.009bmass + 0.539bcount + 175.046$ |
| 14 | Nov HU | $0.428HU_{11} + 4.061bmass + 1.829bcount - 76.252$ |
| 15 | May Hum, Sep Mean Temp, May HU | $-3.362Hum_5 + 6.642MeanT_9 - 0.326HU_5 + 5.163bmass + 0.313bcount + 63.58$ |
| 4 | Jan Wind, Oct Mean Temp | $1.618MeanWind_1 + 2.198MeanT_{10} + 2.248bmass + 3.644bcount - 56.78$ |

Table B.15: Model F equations for individual orchards

| Oc | Weather features | Equation |
|---|---|---|
| 17 | May and Aug Max Temp | $-6.301MaxT_5 + 2.02MaxT_8 - 3.664bmass + 265.546$ |
| 90 | Apr Mean Temp | $9.447MeanT_4 + 8.652bmass - 163.7397$ |
| 56 | prev May Hum, prev Jun Max Temp | $1.816prevHum_5 - 8.559prevMaxT_6 + 6.222bmass + 235.725$ |
| 57 | May Max Temp | $-12.302MaxT_5 + 10.594bmass + 441.472$ |
| 58 | May and Aug Max Temp | $-0.828MaxT_5 + 6.69MaxT_8 + 9.477bmass - 151.965$ |
| 61 | May Max Temp | $-12.15MaxT_5 + 9.466bmass + 431.511$ |
| 33 | Aug HU | $0.165HU_8 + 4.254bmass + 71.261$ |
| 38 | Jul Mean Temp | $6.491MeanT_7 + 11.947bmass - 69.305$ |
| 39 | prev May Hum | $2.305prevHum_5 + 10.521bmass - 55.795$ |
| 42 | prev May Hum | $2.339prevHum_5 + 9.528bmass - 46.149$ |
| 43 | Aug Max Temp | $6.737MaxT_8 + 9.715bmass - 184.099$ |
| 44 | Aug Min Temp | $16.684MinT_8 + 7.357bmass - 31.625$ |
| 45 | May Max Temp | $-10.257MaxT_5 + 8.098bmass + 394.705$ |
| 46 | prev Jun Max Temp | $-8.164prevMaxT_6 + 12.303bmass + 255.722$ |
| 47 | Jul Rain | $4.818Rain_7 + 13.156bmass + 11.862$ |
| 49 | prev Jun Max Temp | $-16.098prevMaxT_6 + 6.158bmass + 538.654$ |
| 50 | prev Jun Max Temp | $-9.074prevMaxT_6 + 12.683bmass + 288.44$ |
| 51 | prev Jun Max Temp | $-14.551prevMaxT_6 + 6.77bmass + 487.361$ |
| 35 | prev Oct Min Temp, prev Jun and May Max Temp | $4.129prevMinT_{10} - 1.988MaxT_5 - 3.901prevMaxT_6 + 14.378bmass + 147.035$ |
| 70 | prev Jun and May Max Temp, prev Oct Min Temp | $-9.279MaxT_5 + 13.184prevMinT_{10} - 2.682prevMaxT_6 - 0.607bmass + 358.959$ |
| 71 | Mar Min Temp, Dec Wind | $-5.095MinT_3 - 4.380Wind_{12} + 4.801bmass + 164.205$ |
| 74 | Aug Max Temp | $4.949MaxT_8 + 6.711bmass - 123.183$ |
| 75 | prev Oct Min Temp, May Max Temp | $18.239prevMinT_{10} - 4.426MaxT_5 + 8.161bmass - 4.394$ |
| 8 | May and Aug Max Temp | $2.258MaxT_5 + 4.781MaxT_8 + 11.44bmass - 210.08$ |
| 12 | Oct Hum, Jul Rain, May Min Temp | $4.343Hum_{10} + 0.659Rain_7 + 0.407MinT_5 + 11.086bmass - 101.732$ |
| 5 | Mar Min Temp | $-7.982MinT_3 - 5.143bmass + 243.726$ |
| 9 | prev Aug Min Temp, prev Oct Hum | $-12.388prevMinT_8 - 5.967prevHum_{10} - 2.113bmass + 321.957$ |
| 10 | prev Jan and prev Jul Max Temp | $-7.821prevMaxT_7 + 6.632prevMaxT_1 - 1.328bmass + 36.479$ |
| 11 | May and Aug Rain | $-3.769Rain_8 - 1.585Rain_5 - 0.4bmass + 113.712$ |
| 13 | Mar Min Temp, prev Oct HU | $-8.319MinT_3 + 0.335prevHU_{10} - 5.432bmass + 193.049$ |
| 14 | Nov HU | $0.508HU_{11} - 0.032bmass - 44.087$ |
| 15 | May Hum, Sep Mean Temp, May HU | $-3.413Hum_5 + 6.662MeanT_9 - 0.324HU_5 + 4.616bmass + 72.734$ |
| 4 | Jan Wind, Oct Mean Temp | $8.293MeanWind_1 + 7.712MeanT_{10} - 2.253bmass - 171.185$ |