Assessing the Variety of a Two-Parent Miscanthus Mapping Population

Author: Amy Richards (150004367)
Supervisor: Dr Kim Kenobi
Joint supervisors: Dr Kerrie Farrar, Dr Wayne Aubrey

Department of Mathematics
Aberystwyth University

27 September 2019

This dissertation is submitted in part completion of the degree of
MSc in Data Science (G490)

“The decisions we make
today are critical in ensuring
a safe and sustainable world for everyone,
both now and in the future”
Intergovernmental Panel on
Climate Change (IPCC) 2019

Summary

Attitudes towards the use of fossil fuels are quickly changing for worse as people learn and understand the effect that burning fossil fuels has on our world. One proposed alternative to their use is development of biofuel. This paper analyses a dataset spanning three years of measurements on Miscanthus plants grown in Aberystwyth, Wales. It analyses the behaviour of 105 genotypes of Miscanthus and develops a linear model which will predict the crop of each genotype by using the plants’ phenotypical properties. The predictions of the linear model are used to determine which genotypes should be further developed in order to improve crop quality and yield. The results of the assessment of variation in the Miscanthus plants in this paper suggest that the best genotype of Miscanthus for growth as biofuel in a climate similar to Aberystwyth will flower late in the year with a high transect density, tall stems and leaves located high up the stem.

Acknowledgements

Heartfelt thanks to my supervisor Dr Kim Kenobi for his patience and kindness through all stages of this project and for sharing his expertise in *R* to help me improve my proposed solution to the problem.

Thanks to Dr Kerrie Farrar for sharing the data with me and being available to discuss specific biological details of the data. Thank you for believing in me and giving me the confidence to complete the project with a positive and hopeful attitude to further academic development in future.

Thanks to Dr Wayne Aubrey for suggesting the use of LATEX for writing this dissertation.

Finally, thank you to my partner Dale for being there for me every step of the way. I would not have finished this without you.

Contents

List of Figures

1 Introduction

1.1 Background

The decline in fossil fuel reserves together with a recognition that greenhouse gas emissions such as carbon dioxide responsible for a deterioration in climate change have focused the attention of the world on reducing our dependence on the finite fossil fuel reserve. Leaders across the world have declared a climate emergency [1] and the general consensus is that burning fossil fuels is responsible for around half of the increased temperature worldwide and almost a third of the rise in sea level [2].

One proposed alternative to the use of fossil fuels is the development of bioenergy crops, i.e. energy derived from material that was once alive for example trees, crops or animal waste. Bioenergy crops is a research area that is growing considerably as attitudes towards traditional energy forms such as fossil fuels change for worse. Bioenergy can take the form of solid matter (biomass) for combustion, or liquid product (biofuel) and can be used to power vehicles e.g. cars and buses. Both biomass and biofuel can be derived from bioenergy crops: its development will be extremely important in trying to reduce our use of fossil fuels which will reduce our greenhouse gas emissions, and therefore slow climate change and its destructive effects.

The concept of using plants as an energy source has been in existence for years; until oil was discovered in 1859 forestry and agricultural crops were the most used energy sources in the world [3]. In 2000, the percentage of electricity production from renewable sources within the United Kingdom was only 2.6%, rising to 29.2% by 2017 and the percentage has reached 33.0% by 2019 [4]. This shows the huge increase in renewable energy production in only two decades. Reasons for this change include international and national incentives such as the European Union’s Renewable Energy Directive [5] setting the aim of countries within the European Union achieving 32% of their energy production from renewable sources by 2030.

Bioenergy crops are favourable since it is a high product with high energy potential and the ability to grow in low quality soil. Planting bioenergy crops on arable land has proved that they are able to improve the quality of the soil in the long term by increasing the organic carbon content of the soil [6]. The plant draws carbon from the atmosphere during growth and stores part of it as organic matter within the soil.

Another important property is that it has low requirements for biological, chemical and physical pre-treatment before being able to be used as bioenergy. A field of bioenergy crops can act as a filter system for pollution management; they draw surplus fertiliser and pesticides from the surface water before it pollutes groundwater or rivers/streams.

In order for the bioenergy supply chain to succeed it needs to be energy favourable, maintain or increase soil carbon levels and be cost-effective and sustainable for the environment without interfering with essential production [6]. One of the main areas of focus for researchers in this area is how to improve the biochemical constitution and structure of bioenergy crops to enable an increase in energy tonnage production.

Obviously the development of bioenergy crops will not solve all the problems arising from the use of fossil fuels and bioenergy cannot completely replace fossil fuels due to the huge cropping area required; but in examining options such as mixing biomass with coal [7] and improving the quality and yield of each harvest from the crops it is a step in the right direction.

1.2 Miscanthus

Ideal biomass uses the resources available such as water effectively, is a perennial crop i.e. does not require annual replanting, is not invasive and does not have high fertiliser or pesticide requirements. One bioenergy crop being considered that meets these requirements is *Miscanthus* [8], an African, Eurasian and Pacific Island plant genus in the grass family. The common name of the plant is *Elephant Grass*. Miscanthus is a high yield crop which grows to heights of over three metres in some climates, and looks similar to bamboo. Miscanthus is a second-generation bioenergy crop which means it is more efficient than first-generation bioenergy crops (e.g. corn, sugarcane, rapeseed and palm oil). Second generation crops can produce fuel from biomass that cannot be used as food e.g. woody crops, agricultural waste and bioenergy crops grown on arable land unsuitable for growing. Based on 25 miles per gallon, one ton of Miscanthus can produce enough fuel to drive a car 750 miles [9].

Commercial development of Miscanthus has been underway since 1998 led by Renewable Energy Crops Ltd [9]. The company offers a range of services which include planting and establishing Miscanthus and agronomic and harvest advice and consultancy.

The company also offers agreements to long-term growers to try to influence more farmers to grow Miscanthus on a wide scale.

The *Optimising Miscanthus Biomass Production* (OPTIMISC) project [10] funded by the European Union has the aim of optimising the Miscanthus supply chain by trialling a number of Miscanthus hybrids across Europe and working to try to destroy or reduce the properties limiting the potential use of Miscanthus as a bioenergy.

A relatively recent success in the development of Miscanthus as a bioenergy is the establishment of the Miscanthus Breeding Scheme [11] at the Institute of Biological, Environmental & Rural Sciences (IBERS) at Aberystwyth University in 2004. The aim of the scheme is to produce a biomass crop on land not in use for growing food. The scheme focuses on pairing genotypes and hybrids with growing conditions across Europe and the USA.

The hybrid most typical of Miscanthus is Miscanthus x giganteus collected by Aksel Olsen from Yokohama, Japan in 1935. The plant was taken to Denmark where it was treated and spread through Europe and North America for planting for horticultural use [8]. Due to its high yield over a range of conditions, Miscanthus x giganteus has been grown successfully across the world - from a Mediterranean climate in Spain to the northern climates of Scandinavia [12]. In Europe Miscanthus x giganteus has been widely studied since 1983 for combustion to produce heat and electricity [13]. Miscanthus x giganteus is a hybrid of *Miscanthus Sinensis* and *Miscanthus Sacchiflorus*; it is unique within the species as it retains levels of photographic-synthetic activity at low temperatures [14]. A large number of Miscanthus hybrids have been trialled across Europe over the last 30 years and they have high yield potential together with low inputs which make them very useful in increasing biomass production across Europe in future [15].

Figure 1: Miscanthus x giganteus in Pembrokeshire

While Miscanthus x giganteus has been the most popular hybrid across Europe for several years, some concerns have appeared. There are high costs in growing the hybrid on a wide scale and there have been several cases of the hybrid failing to become established in colder climates at high latitudes [14].

Miscanthus x giganteus is a triploid and this means there will be difficulty in breeding this hybrid [16]. One hybrid cannot meet all the requirements of all uses of Miscanthus and growing large areas of one clone increases the risk of disease and pests. We need a wide genetic base and development of different genotypes of Miscanthus to overcome the limitations involved in this [17]. The current focus of scientists and biologists is looking at combinations of different hybrids that can be developed in the Miscanthus genus so that it can be grown in different climates and latitudes across Europe and the world.

Miscanthus has very high water use efficiency and one concern in overplanting Miscanthus is the increasing demand for water; the location of Miscanthus plantations must be planned carefully [6] to ensure that there is no lack of water in the area. Another concern is that increasing Miscanthus production can have a negative effect on the wider natural environment [6].

Most cost concerns connected with biofuels involve the development of biofuels from annual food crops with high fertiliser costs [6]; the use of Miscanthus reduces this concern as Miscanthus is not used for food and does not require high levels of fertiliser to grow.

When growing Miscanthus, nitrogen fertiliser is unnecessary except on low fertility soil, pesticides are completely unnecessary but herbicides are necessary during the first years of establishment and afterwards natural weed prevention by shade is sufficient [6].

A number of trials have been carried out across Europe with the aim of researching an improvement in Miscanthus yield and finding an ideal genotype for the climate and environment in which it is to be grown. One trial of interest is a trial carried out in 2000 in Germany [18]. The researchers found that while much interest had been carried out into the Miscanthus x giganteus hybrid, the rhizomes (an underground plant stem that includes nodes from which new sprouts and stems grow) of this hybrid are inadequate during winter months. Miscanthus x giganteus is not well adapted to all climates and poor survival of this hybrid was observed in northern European areas over the winter of the first year after planting.

The team of researchers identified an opening in the research area to expand the genetic base of Miscanthus by finding different hybrids which are more suitable for different climates. Another important reason for expanding the genetic base is that using only one Miscanthus hybrid to develop bioenergy carries considerable risk of pest and disease attack [18]. The Miscanthus genus has a wide geographical range and therefore there is a considerable genetic variation that can be examined in order to breed new genotypes [18].

Fifteen Miscanthus genotypes were collected and planted in southern Germany for the trial which measured phenotypic growth properties such as height, bud density, stem diameter, time of flowering and rate of senescence over three years. The main conclusions from this research were that at least two years data was required in order to compare the growth of different genotypes, the yield and quality of the biomass cannot be predicted using data from the first year of plant growth. The reason for this is that Miscanthus is a rhizomatous annual plant i.e. Miscanthus has an underground stem which creates roots and shoots from its nodes, and so the properties of growth change as the plant matures and so a number of years of data are required to truly analyse the potential of a genotype [18].

The most important properties when considering the prediction of the behaviour of different genotypes was the time of flowering and rate of senescence over the autumn. It was discovered that lowest biomass quality was found in hybrids with senescence late in the year and that the highest biomass quality occurred in hybrids with early senescence and thin stems [18].

Twenty-one genotypes were studied in a trial in northern France the focus of which was to analyse differences between flowering date and growth rate of each genotype. It was discovered that canopy height had a strong positive correlation with biomass yield [19]. Genotype behaviour can be estimated by measuring its growth rate and canopy height since the biomass potential of a crop is a product of growth rate and growth length [19].

Another study of 15 Miscanthus genotypes was held on five sites across Europe in Sweden, Denmark, England, Germany and Portugal [17]. The main conclusion of this study was that genotypes with the greatest yield in Sweden and Denmark were the genotypes with the smallest yield in Portugal and Germany [17]; the genotypes therefore have different results depending on where in the world and what kind of environment they grow in. Several phenotypic properties were measured over the trial including stem height, stem diameter, flowering time, rate of senescence and yield. Results from this study were used to contribute to knowledge of where in Europe is most suitable for growing the different genotypes. Like the study in Germany, the conclusion was that plant height, which is influenced by time of flowering, was the most important property in selecting the best genotype [17].

In 2001, another study was held with a focus this time on optimising harvest time for Miscanthus on six sites across Europe [20] as part of the OPTIMISC project [10]. The results of this study show that harvesting time delay varies between genotypes due to the differences in stem diameter and time of flowering i.e. senescence [20].

Another study under OPTIMISC is a study of how drought stress effects the growth and quality of Miscanthus crops for bioenergy. Fifty Miscanthus genotypes were studied under drought conditions and control conditions in a glasshouse experiment [21]. Drought is one of the most common abiotic stresses and drought events will only increase as climate change worsens [22]. A great attraction of Miscanthus is that its perennial growth and wide-ranging root system mean that it is better at finding and using water stores under the soil than annual plants [23]. It was discovered that there are wide differences between the weights of plants showing a great variation in how different genotypes survive drought [21]. There is therefore great potential for research and observation on Miscanthus genotypes better able to cope with drought conditions than other genotypes.

A Miscanthus breeding scheme was started in the biology department of IBERS at Aberystwyth University back in 2004. The objective was to breed a range of Miscanthus hybrids that will survive in different growing conditions for use as biomass fuel in future [15]. A large number of crosses between species of Miscanthus were attempted and these trials are being used to try to improve mathematical models used to predict hybrid behaviour. Data from trials in Asia and Europe has been combined, together with observations from glasshouse experiments at Aberystwyth to derive a mathematical model to predict the environmental conditions required to synchronise flowering [24]. The observations taken from the plants include plant height, photosynthesis rate, water usage, response to drying out and data on flowering, emergence and senescence,

Further development of the Miscanthus crop depends on several factors, one of which is competition with other forms of energy and the market value of Miscanthus will depend on the cost of coal, gas, oil and production costs of nuclear energy and other renewable energies such as wind, solar, wave, tide and hydro [2]. Use of Miscanthus will increase if hybrids are discovered that can tolerate biotic and abiotic stress and be planted on a large scale and then managed at low cost [15].

In considering the research seeking to improve or breed ideal genotypes of bioenergy crops, the phenotypes usually targeted focus on resistance to disease, yield optimisation, improvement in crop quality, waste reduction, or the introduction of tolerance to challenging environmental conditions such as drought [25]. The central aim of breeding bioenergy crops must be to intensify product sustainability – the biomass produced must increase for each unit of land without environmental harm or more agronomic inputs such as supplementary watering, fertilisation and use of pesticides. The performance of the crop also has to be consistent; bioenergy crops have to provide dependable high yield annually even when climactic conditions change.

2 Materials and Methods

2.1 Project Objective

The IBERS department at Aberystwyth is one of the foremost researchers in the development of Miscanthus plants for bio-fuel. The objective of this project is to analyse data collected by IBERS on the annual cycle of Miscanthus plants over a period of three years to determine which Miscanthus genotypes are most suited to further development. There are 302 plants in the data spread over 105 different genotypes.

2.2. The Data

The raw data for this project was provided by Dr Kerrie Farrar in the form of five CSV files; three containing data on the annual cycle of Miscanthus plants containing the harvest results for two years. Dr Farrar is based in the Institute of Biological, Environmental & Rural Sciences (IBERS) at Aberystwyth University and Dr Farrar's research focuses on how plants develop in response to their genome, environment and biotic interactions. Dr Farrar runs the IBERS Energy Crop Biology research group and the aim of the research is to improve bioenergy crop yields in order to provide sustainable biomass as an alternative to energy based on petroleum, liquid fuels for transport and bulk chemicals.

There are two parts to the Miscanthus growth annual data; single data and continuous data. The single data consists of individual measurements and observations on the plants and the continuous data consists of weekly measurements over several months on the plants. There are 105 different genotypes of Miscanthus observed over the three years; 2011, 2012 and 2013. Harvest data is available for 2012 and 2103 only. The harvest data consists of columns for fresh weight, moisture content percentage and final dry weight of each plant. The column involved in this project is the dry weight of each plant at the end of the harvest since this is the biomass created by the plant i.e. how much can be burned as biofuel. See *appendix A* for a snapshot of the raw data and field plan of the plants.

Figure 2: Miscanthus mapping population parents for this project

Figure 3: Miscanthus harvest at Bluestone Park in Pembrokeshire

Property	**Year**	**Description**
Stem diameter (mm)	2011, 2012, 2013	Three stems are selected at random from each plant, one measurement of diameter is taken from each stem halfway up the stem.
Clump diameter (cm)	2011, 2012, 2013	Measurement of the diameter of the base of the plant. One plant has several stems, this is the diameter of all stems together as a clump.
Transect density	2011, 2012, 2013	Measurement of plant density. The measurement is taken by placing a stick through the plant, raising the stick halfway up the plant and counting how many stems touch the stick on one side.
Tallest stem ligule (cm)	2011, 2012, 2013	Measurement from the base of the plant to the ligule (thin growth at the junction of leaf and stem) on the tallest stem.
Tallest stem leaf (cm)	2012, 2013	Measurement from the base of the plant to the true leaf on the tallest stem.
Tallest stem flower (cm)	2011, 2012, 2013	Measurement from the base of the plant to the flower on the tallest stem.
Flowering density	2011, 2012, 2013	Percentage of plant in flower. Sets out when the plant is less than 50% in flower, more than 50% in flower and more than 80% in flower.
Senescence density	2011, 2012, 2013	Percentage of the plant that has senesced i.e. withered, declined with age. Sets out when over 80% of the plant has senesced.
Canopy height (cm)	2011, 2012, 2013	Weekly measurements of the canopy height of each plant are taken over a period of 7 months.
Observations on flowering	2011, 2012, 2013	Sets out when each plant flowers. There are two stages; A and F. Stage A is the first sign of flowering and stage F denotes that the plant has flowered. These measurements are taken weekly over a period of 5 months.
Emergence	2012, 2013	Data for 2012 includes weekly measurements on the emergence of each plant over a period of five months. Data for 2013 includes measurements for three months.

Table 1: Phenotypic properties of Miscanthus data

2.2.1 Data Preparation

The data required a great deal of preparation before the data could begin to be used and analysed. The raw data had a number of problems such as complicated or unclear column names, missing data and an overcomplicated layout which complicated the code written for data analysis.

Three scripts are written to achieve this; `Data_cleanup_2011.R`, `Data_cleanup_2012.R` and `Data_cleanup_2013.R`. The three Excel files on Miscanthus annual life cycle are similar to each other but there are minor changes such as column names and a number of measurements which are different from year to year and so I dealt with each year of data separately.

Genotype ID

Each plant is given a unique ID in the column `UID`, for example `28287` and an ID to distinguish between genotypes in the column `MxNumber`, for example `-Mx1553#114`. Each `UID` is unique but the `MxNumber` is not unique since there are several plants for each genotype.

A number of records in the `MxNumber` column were not numbers but a specific genotype of Miscanthus `-Goliath` or `-Giganteus`. The first stage of data preparation was to change the MX numbers to integers for use as the ID of the genotypes. The prefixed characters – `Mx1553#` were stripped from the ID to leave a three digit integer for each row of data. The `Goliath` records were changed to `Gol` and the `Giganteus` records were changed to `Gig`. This ID was stored in a new column `Genotype`. The final stage was to arrange the data in increasing order of genotype ID numbers over the single and continuous data for each year. It was observed that there were usually three entries for each genotype, two for Giganteus and twelve for Goliath.

Dead Plants

There were 22 records of dead plants in the raw data; each measurement for these plants was *0* or *NA* and this data is of no use to the investigation. I elected to remove these entries from the data in order to completely remove them from further analysis.

Stem Diameter

Over each year of data there are three measurements for stem diameter. I summarised this in a column meanStemDiameter in order to compare more easily between genotypes and plot data more visually efficiently. I defined another column, sdStemDiam which contains the standard divergence of the three measurements of stem diameter for each plant.

Clump diameter

Over each year of data there are two measurements for clump diameter. I summarised this in a column meanClumpDiam which is an average measurement of the clump diameter of each plant. As with the stem diameter, I defined a column sdClumpDiam to capture the standard diversion of the stem diameter measurements for each plant.

DoYFirst3Emergence

A score of 0-3 is given to each plant based on its emergence using these parameters –

- 0: no emergence at all
- 1: 1-10cm
- 2: up to 15cm
- 3: up to 20cm

A column DoYFirstEmergence was defined containing the day of the year that each plant achieved a score of 3 for emergence. Once the plant has achieved a score of 3, canopy height is measured.

Arbitrary Records

There were several invalid entries in several columns in the data. One example of this are entries of -9999 in the columns for flowering density and senescence density. These entries were amended to NA. Other entries in these columns were 318 and these were therefore amended to y. The pair NA and y are a great deal easier to analyse than arbitrary numbers such as -9999 and 318.

Renaming Columns

In the continuous data, the columns are numbered as numbers from the days of the year. In the entries ShootCounts2011, the numerical columns e.g. 87 are changed to ShootCountsDoY87, where DoY stands for 'Day of Year'. In renaming such columns they are easier to read in the code. Other examples are changing X178.1 to CanopyHeightDoY178 and changing X213.2 to FloweringStageDoY213.

Outliers

At the end of each Data-cleanup file I dealt with outliers. Some were completely invalid and these were changed to NA as there was no way of knowing whether the value had been written down wrongly or was a mistake. Some were changed to the correct value, for example CanopyHeightDoY274 and CanopyHeightDoY281 were both 130, we could deduce that the entry in the column CanopyHeightDoY289 of 1130 was typed wrongly and should have been 130.

After preparing the data, the data was saved as six new CSV files; Single_2011_clean.csv, Single_2012_clean.csv, Single_2013_clean.csv, Continuous_2011_clean.csv, Continuous_2012_clean.csv, Continuous_2013_clean.csv. The script read.in.clean.data.R is used in order to read into the CSV files for investigation. See *appendix B* for a snapshot of the data after preparation.

2.3 Methodology

Predictive modelling is the process of developing a model which produces a correct prediction of some variable of interest. We are interested in predicting the variable Y using a set of related variables, X_1, X_2, ...X_i and then using the information gleaned from the model to predict Y using the observations from X_1, X_2, ...X_i. For this project, we want to develop a predictive model in order to predict the yield (i.e. dry weight) of 105 different Miscanthus genotypes. This model will help investigators in the field to select which genotypes to develop further. A good model will be simple (low number of parameters), intelligible and give correct predictions.

A model should only be as complicated as is entirely necessary to describe the data, i.e. a model with 10 variables is only better than a model with one variable if it is a much better fit for the data which is not always true.

2.3.1 R and RStudio

The statistical coding language R [26] is used within the integrated development environment RStudio [27]. R is chosen because it provides an intensive environment for the analysis, processing, transformation and imaging of data. There are a number of packages in R that are useful and relevant to this project. See *Appendix C* for the packages used during this project.

2.3.2 Linear Model

The multiple linear model is a type of predictive model. It uses linear regression to predict the value of the variable Y based on one or more predictive variables X. The aim is to model the continuous variable Y as a mathematical function of one or more variables X so that we can use the model to predict Y when only X is known.

The equation can be generalised as:

$$Y_i = \beta_0 + \beta_1 X_i + ... + \varepsilon_i$$

Where Y_i is the ith response, β_0 is the interception on the y-axis, β_1 is the gradient, X_i is the value i of the variable X and ε_i is the error i, which is the part of Y the model cannot explain.

Linear Regression Assumptions

1. **Linearity**: A linear relationship must exist between Y and the predictors X_1, X_2, ...X_i. It is important to examine the data for outliers since linear regression is very sensitive to the effects of outliers in the data. Outliers can be seen in the data through distribution plots of each variable.

2. **Normality**: Multiple regression assumes that the residuals are distributed normally. This can be checked by *Plot Q-Q*. if the data is not normally distributed, a non-linear transformation such as a *log* transformation can deal with this.

3. **Multi-linearity**: Multiple regression assumes that the independent variables do not have a high correlation i.e. too high a correlation exists between the independent variables and each other.

4. **Homoscedasticity**: The variance of error terms is similar across the values of the independent variables. This can be seen by plotting residual terms versus the fitted, the points should be distributed equally across all values of the independent variables.

R^2 – The Coefficient of Determination

R^2 is used as an indicator of linear model performance, it shows the predictive ability of the model and the closer the value to 1, the better the model. R^2 does not measure the accuracy of a model but it does measure its usefulness. Mathematically, R^2 is the percentage variance of the dependent variable explained by the linear model.

The variables considered in developing the linear model for this project are –

- **TransectCount**
  Transect intensity of the plant

- **MaxCanopyHeight**
  Maximum canopy height of the plant

- **DoYmaxCanopyHeight**
  The day of the year the plant reaches its maximum canopy height.

- **ClumpDiameter**
  The diameter of the clump of the plant

- **TallestStemLigule**
  The height of the ligule of the tallest stem of the plant

- **TallestStemTrueLeaf**
  The height of the leaf on the tallest stem of the plant

- **TallestStemFlowerBase**
  The height of the flower on the tallest stem of the plant

- **FloweringStageDoYFirstA**
  The day of the year the plant reaches the first stage of flowering (flagleaf)

- **FloweringStageDoYFirstF**
  The day of the year the plant flowers

- **DoYFirst3Emergence**
  The day of the year the plant attains a score of 3 for emergence i.e. reaches 20cm in height

2.3.3 StepAIC

StepAIC seeks to select the best predictors for a multiple linear regression model by comparing models with different predictors in a sequential way.

StepAIC is a combination of a *step*wise model selection i.e. selecting a model by sequential means; and *AIC* which stands for *Akaike Information Criteria*. The information criterion objectively determines the best model by selecting the model with the lowest AIC value was the aim is to find a balance between the suitability of the model and its complexity.

You could look at all possible combinations of linear models that can be created with the predictor variables individually but the *StepAIC* function makes this task quick and effective.

StepAIC does not necessarily improve the model but it does simplify the model without much effect on performance. The function removes the variable with the lowest AIC value at each iteration as this is the variable which causes the least loss of information when removed from the model. The function also adds variables back into the model in subsequent iterations in order to see if the addition increases the AIC value.

When the function has been through all possible iterations of the variables, the result is a model with the best possible set of variables. When a model has been selected using StepAIC, it can be used to predict the values of the variable Y.

3 Results

3.1 Investigative Analysis

The three scripts Clean_Data_Exploration_2011.R, Clean_Data_Exploration_2012.R and Clean_Data_Exploration_2013.R are used to carry out the investigative analysis.

3.1.1 Annual Distribution of Plants

I defined eight groups using all variables FloweringIntensityLessThan50, FloweringIntensityGreaterThan50, FloweringIntensityGreaterThan80 and SenescedGreaterThan80.

- Group 0: No flowering density recorded
- Group 1: Flowering density less than 50%
- Group 2: Flowering density greater than 50%
- Group 3: Flowering density greater than 80%
- Group 4: Senescence greater than 80%
- Group 5: Flowering density less than 50% and senescence greater than 80%
- Group 6: Flowering density greater than 50% and senescence greater than 80%
- Group 7: Flowering density greater than 80% and senescence greater than 80%

Figure 4: 2011 Group Distribution

We can see that 29% of plants have not yet begun to flower and another 29% have flowering density of less than 50% in 2011. This suggests that the great majority of plants have not yet matured and not yet reached the flowering stage. Only 10% of plants have reached the stage of senescence.

Figure 5: 2012 Group Distribution

By 2012, 63% have reached group 7 compared with only 4% in 2011 suggesting that the plants have matured by 2012. All plants have reached the senescence stage by 2012 with the majority of plants having flowering density over 80% and senescence over 80%

Figure 6: 2013 Group Distribution

2013 shows the same distribution pattern as 2012 with the great majority of plants in group 7. This suggests that plants mature quickly from their first year to the second year and then slow down and follow the same pattern annually which is expected of perennial plants such as Miscanthus.

3.1.2 Comparison of Variable Average

There are 302 plants in the data, spread over 105 genotypes. The average of a number of variables for each genotype was calculated in order to plot them visually efficiently. Since a large number of genotypes are being examined, I divided the data into three groups so that it could be more easily seen on a graph. The following graphs are arranged in increasing order of genotypes in 2013, i.e. arranged by the year when the plants were at their most mature.

Figure 7: Average Stem Diameter Measurements over 3 Years

The great majority of stem diameters is our less in years subsequent to 2011. The general pattern shows that stem diameter was at its lowest in 2012 and higher in 2013. The mean stem diameter value in 2011 was 5.27, 4.30 in 2012 and 4.63 in 2013.

Stem diameter is one of the variables greatly influenced by environment. Stems are cut at the end of each year for use as biofuel and so it is reasonable that there will not be an annual increasing pattern of stem diameter. Stems are also selected at random from the plant so it is not guaranteed that the biggest stems are chosen for measurement.

Figure 8: Average Clump Diameter Measurements over 3 Years

The general pattern is that the clamp diameter increases annually; this is reasonable since the plant matures annually. There should be more stems each year and therefore clump diameter will increase. There is a great growth from 2011 to 2012 and then this slows from 2012 to 1013. There are some genotypes that barely see an increase in clump diameter from 2012 to 1013.

Figure 9: Average Transect Density Measurements over 3 Years

Transect density also follows the general pattern of annual increase. There is again less difference between 2012 and 2013 suggesting that the number of stems reaching a height halfway up the plant is less than the number of stems at the base of the plant. This explains the greater difference between clump diameter in 2012 and 2013.

Figure 10: Average Tallest Stem Ligule Measurements over 3 Years

The distribution of the tallest stem ligule also follows the pattern of annual increase. We can see that the genotypes with the highest tallest stem ligule in 2011 still have a high tallest stem ligule measurement in 2013 i.e. if the genotype has a high ligule measurement in 2011, the pattern shows that the genotype always has a high ligule measurement in following years.

Figure 11: Average Canopy Height Measurements over 3 Years

This variable shows an obvious annual increase. All genotypes have a maximum canopy height which is higher in following years which again suggests that if a genotype has a high canopy height in one year, it is also likely to have a high canopy height in subsequent years.

Figure 12: Average Measurements for Average Day of Year Maximum Canopy Height is Reached for each Genotype

These measurements are not distributed increasingly as MaxCanopyHeight. No clear pattern is apparent but we can see that several genotypes in 2011 had a day of year maximum canopy is reached substantially lower than subsequent years. This variable is very strongly influenced by the environment i.e. weather and growing conditions and it is therefore reasonable not to see a clear annual pattern. The ideal genotype will reach its maximum canopy height later in the year i.e. it grows throughout the growing season.

TallestStemFlowerBase, TallestStemTrueLeaf, FloweringStageDoYFirstA, FloweringStageDoYFirstF and DoYFirst3Emergence were not plotted in this way since too much data was missing between years to allow comparison.

3.1.3 Paired Plots

Paired plots is a distribution plot matrix showing the relationship between variables. It is a useful method for seeing patterns and inconsistencies in data in order to further analyse that data. Below are paired plots for the variables over the three years; by analysing these plots, we can see whether there is a linear relationship between the variables over the years.

Correlation is a statistical measurement which gives the linear dependence between two variables. A measure is taken of a correlation valued between -1 and +1. A value close to 0 suggests a very weak relationship between the variables and a low correlation value ($-0.2 < x < 0.2$) suggests that the predictor X does not explain the variation of Y and that we should look for other explanatory variables.

Figure 13: Stem Diameter Paired Plot

We can see a linear relationship between the three years with a stronger relationship between 2012 and 2013 than between 2011 and 2013 once again. This is expected given the annual distribution of plants as seen in chapter 3.1.1. While it seems that there are outlier values in the data after looking at the raw data I decided it is not an outlier value – it is not too different to the other values. There is a linear relationship between the years since the stems are selected at random from the plants to be measured, this variable is not as reliable for use to predict yield when compared with other variables.

Figure 14: Clump Diameter Paired Plot

We see a comparatively strong linear relationship between the three years. However the linear relationship is stronger between 2012 and 2013 than 2011 and 2013 which is expected given the annual distribution of the plants. This shows that the high clump diameter in one year means there will also be a high clump diameter in the next year i.e. there is a strong linear relationship between clump diameters of the three years.

Figure 15: Transect Density Paired Plot

We can see that transect density measurements in 2011 are lower than measurements in subsequent years. There were many fewer stems in 2011 compared with following years and so it is reasonable that the transect density is lower. However, there is a comparatively strong linear relationship between the three years suggesting that the genotypes with high transect density in 2011 still have high transect density in 2012 and 2013.

Figure 16: Tallest Stem Ligule Paired Plot

There is a very strong linear relationship between the three years which again suggests that plants with greater ligule height on the tallest stem still have a high measurement in subsequent years.

Figure 17: Tallest Stem Flower Paired Plot

There is a linear relationship between the three years with the strongest relationship between 2012 and 2013. This suggests that tallest stem flower in 2012 influenced the height of the tallest stem flower in 2013. It is expected that there is no strong relationship between 2011 and the two other years – given the distribution of plants in 2011, the majority had not reached flowering. There is too much data missing between the years to calculate correlation values for this plot.

Figure 18: Tallest Stem True Leaf Paired Plot

There is a very strong linear relationship between the height of the tallest stem true leaf from 2012 and 2013 which once again suggests that the leaf height of a plant in 2012 influences the height of the true leaf on the same plant in 2013.

Figure 19: Maximum Canopy Height Paired Plot

There is a strong linear relationship between each year of maximum canopy measurements. This once again suggests that plants with high canopy height continue to be amongst the plants with high canopy height in subsequent years.

Figure 20: Day of Year Maximum Canopy Height Paired Plot

There is no strong linear relationship between the three years which suggests that the day the plant reaches maximum canopy height does not influence the day in subsequent years. This may be for several reasons including environmental effects such as different weather between the years.

Figure 21: Flowering Stage A Paired Plot

There is a comparatively weak linear relationship between the measurements for Flowering Stage A, which is the first indicator of flowering in the plant. This suggests that the day the plant reaches flowering stage A does not influence the day it reaches this stage in subsequent years. Too much data is missing between the years to calculate correlation values for this variable.

Figure 22: Flowering Stage F Paired Plot

There is a linear relationship between the three years but it seems to be a random pattern on earlier days in the year. Fewer plants reach flowering stage F than reach flowering stage A. Too much data is missing between the years to calculate correlation values for this variable.

Figure 23: Day of Year First Emergence Paired Plot

There is a linear relationship between the measurements for day of emergence between 2012 and 2103. A better pattern would be seen if there was more data available for this variable. There are not enough measurements to match between years to calculate a correlation value.

Figure 24: Single Data Paired Plot 2011

We see that the strongest correlation is between tallestStemLigule and tallestStemFlowerBase which is reasonable since the flowers are above the ligule on the plant, i.e. a ligule must exist in order for a flower to exist. The relationship between them is so strong it is reasonable to suppose that it is almost autocorrelated.

We see a subtle linear relationship in the meanStemDiam plots which suggest that this variable does not have much influence on the other variables. There is some kind of linear relationship between meanStemDiam and meanClumpDiam.

There seems to be a negative linear relationship between meanStemDiam and TransectCount which is reasonable to suppose that the more stems a plant has the smaller the diameter of each individual stem.

Figure 25: Continuous Data Paired Plot 2011

There is a strong linear relationship between FloweringStageDoYFirstA and FloweringStageDoYFirstF which is expected since flowering stage F follows flowering stage A. There is no obvious relationship between maxCanopyHeight and the two flowering stages which suggests that canopy does not influence when a plant flowers.

Figure 26: Single Data Paired Plot 2012

There is some kind of linear relationship evident between each variable. We see strong correlations between the three variables tallestStemLigule, tallestStemFlowerBase and tallestStemTrueLeaf. This is expected since a leaf follows the development of a ligule and a flower follows the development of a leaf. There is a concern that these three variables have too high a correlation to be used as individual variables in a linear model.

The value Yield has a comparatively strong linear correlation with the transect density which is expected since if the number of stems is higher on the plant, there should be more plant as dry weight at the end of the harvest to use as biofuel.

The value Yield has a linear correlation with the three variables ligule, leaf and tallest stem flower. This suggests that more dry weight comes from taller plants, which is a reasonable assumption.

There is a weaker correlation between Yield and meanStemDiam suggesting that this variable is not very important when considering high yield.

Figure 27: Continuous Data Paired Plot 2012

DoYFirst3Emergence, the score for emergence, has a negative linear correlation with Yield. This shows that early emergence leads to greater yield. The earlier a plant reaches a score of 3 for emergence, i.e. 20 cm in height it is fair to assume that the plant will reach its maximum height and therefore flower earlier. This suggests that plants that flower early produce less biofuel mass after harvest which is reasonable. This was supported by the negative linear relationship between DoYFirst3Emergence and maxCanopyHeight. The plants that emerge early are likely to attain a greater height.

There is a strong linear relationship between Yield and maximum canopy height, suggesting that the taller the plant the more biofuel mass is produced.

There is a subtle correlation between Yield and the two flowering stages, FloweringStageDoYFirstA, FloweringStageDoYFirstF which suggests that they are not important variables to be considered in discussing Yield. More data points in these variables are required in order to reach a definite conclusion on their influence.

Figure 28: Single Data Paired Plot 2013

Once again, Yield shows some kind of linear relationship with all variables. The three tallestStem variables again have a very high correlation with each other. The values for meanStemDiam seem to be more intermittent compared with 2012 and no clear linear relationship is apparent. The variables on the whole interact with Yield similarly to 2012.

Figure 29: Continuous Data Paired Plot 2013

There is a clear negative linear correlation between MaxCanopyHeight and DoYFirst3Emergence which shows that the earlier a plant reaches a score of 3 for emergence, the taller its canopy height will be.

The data for FloweringStageDoYFirst3 seems more intermittent than 2012 and has less of a correlation with Yield.

We see a negative correlation between Yield and DoYFirst3Emergence which shows that an earlier emergence means greater yield.

3.2 2012 Linear Model

Figure 30: 2012 Correlation Plot

We can see that there are strong positive relations between Yield and several variables – TransectCount, TallestStemLigule, TallestStemTrueLeaf, TallestStemFlowerBase and MaxCanopyHeight suggesting that these variables will be important in developing a linear model i.e. these are the variables with the greatest influence on biofuel production.

There is almost no correlation at all between StemDiameter and Yield (-0.09), suggesting that the diameter of the stem has no influence on the amount of biofuel harvested.

Using the paired plots and correlation plot, it is predicted that the best variables to predict Miscanthus yield are MaxCanopyHeight and TransectCount.

Let us create simple (one variable) linear models to observe R^2 for each variable. The model in this case will be –

1m (Yield ~ Predictor)

Predictor	R^2
TransectCount	0.4476
MaxCanopyHeight	0.4423
TallestStemTrueLeaf	0.3449
TallestStemLigule	0.2 21
DoYFirst3Emergence	0.2074
TallestStemFlowerBase	0.2005
DoYMaxCanopyheight	0.1131
ClumpDiamter	0.0412
FloweringStageDoYFirstA	0.0122
FloweringStageDoYFirstF	0.0032
StemDiameter	-0.0046

We can see from the table that one of these simple linear models cannot explain Yield fully i.e. one variable on its own cannot predict Yield and therefore we use a multiple linear model in order better to predict Yield. Looking at the predictor StemDiameter, the negative R^2 value means that the linear regression using StemDiameter alone is worse than using the average value i.e. the model is a worse fit than using a horizontal line.

After discussion with a biologist, MaxCanopyHeight is removed from the linear model development since it is a composite property i.e. a combination of several variables and is not a very good predictor for use in building a linear model. Usually selecting one of the variables TallestStemLigule, TallestStemFlowerBase and TallestStemTrueLeaf for use instead of three is sensible as they have a high correlation with each other. In this case, we will let StepAIc select which variables to retain in the model.

Let us look at a linear model using *all* predictors –

The R^2 value from using all predictor variables is **0.6629** which is much better than the R^2 values achieved by using only one predictor. The modified R^2 value is **0.6452**. The coefficients of the variables appear to be sensible. The four negative variables are connected with flowering and early flowering means less biofuel production.

Looking at the output of the `summary` function, we see that most of the variables *are not significant* in looking at the `Signif.codes` row. This shows that there are too many variables in the model which do not really add any useful information to the model.

Let us use StepAIC to select the most significant properties of the model:

The R^2 value is now **0.6583** and the modified R^2 is **0.6486** which is special – the value has hardly changed but the model has been reduced from 9 predictive variables to only 5. We can see that the two most significant variables in this model are TransectCount and TallestStemTrueLeaf.

The residuals should be equally distributed around the average value 0. The residuals are the difference between the values observed in the response variable (Yield in this case) and the values which the model predicted. We can see that the residuals in this example are not very proportionate. Let us consider this further using diagnostic plots.

By using the StepAIC function we have quantified the effect of the predictive variables on Yield which is the dry weight of Miscanthus at harvest. We could now conclude that TransectCount, TallestStemTrueLeaf and FloweringStageDoYFirstA have a substantial positive effect on Yield but TallestStemLigule and FloweringStageDoYFirstF have a negative substantial effect on Yield.

Figure 31: lm1.aic Diagnostic Plots

Under an assumption of linearity, there should be no pattern in the graph *Residuals vs Fitted*. There is no obvious pattern in this graph. The graph Q-Q shows the normality of the model – the points are close to the diagonal line and so it was concluded that the model was sufficiently normal and sufficiently linear for use although improvement was possible through further statistical research to ensure that the assumptions of linear models were true for this model.

Figure 32: lm1.aic Regression Term Plot

We can see from the regression term plots the contribution of each to the linear model. The greater the slope of the graph, the more it contributes to the model. From this, we can see that TransectCount, TallestStemTrueLeaf and FloweringStageDoYFirstA have positive effects on the model, and that FloweringStageDoYFirstF and TallestStemLigule have a negative effect on the model. We see that data for FloweringStageDoYFirstF is more intermittent than the other variables.

The conclusion from this model is that the five variables are the most important when developing a model to predict Yield and this is the formula for its prediction –

$$\mathit{Yield} = -\ 92.7537 + 2.9513\mathit{TransectCount} + 1.8650\mathit{TallestStemTrueLeaf} - 1.0239\mathit{TallestStemLigule} + 0.6288\mathit{FloweringStageDoYFirstA} - 0.4990\mathit{FloweringStageDoYFirstF}$$

3.3 2013 Linear Model

Figure 33: 2013 Correlation Plot

The correlation plot for 2013 follows a very similar pattern to 2012 which is expected.

It is surprising how irrelevant FloweringStageDoYFirstA is for Yield given that it is one of the variables which was in the 2012 final linear model. On the whole, the predictive variables follow a similar pattern with the same type of correlation with Yield.

Similar to the development of the 2012 linear model, we see the values of R^2 when using one predictor at a time using the model –

1m (Yield ~ Predictor)

Predictor	R^2
MaxCanopyHeight	0.3793
TransectCount	0.3335
TallestStemTrueLeaf	0.2215
TallestStemLigule	0.139
TallestStemFlowerBase	0.1387
DoYFirst3Emergence	0.06892
ClumpDiameter	0.05081
FloweringStageDoYFirstF	0.02456
DoYMaxCanopyheight	0.01183
StemDiameter	0.007831
FloweringStageDoYFirstA	0.001108

MaxCanopyHeight and TransectCount are still the best predictive variables for predicting Yield alone. It is surprising that FloweringStageDoYFirstA is the worst predictor on its own since this is one of the variables which was in the final linear model for 2012.

Let us look at the linear model using *all* predictors –

An R^2 value of **0.5705** which means that 57.05% of the variation in the model can be explained by using all predictors.

The next step is to try to simplify the model and select the most important properties by using StepAIC –

The R^2 value after running StepAIC on the model is **0.5612** and the modified R^2 is **0.5467** which is again very good given that the model has been reduced from 9 variables to 4.

Figure 34: lm1.2013.aic Diagnostic Plots

The diagnostic plots show that there is not much pattern in the residuals plot against the fitted values which confirms the assumption of linearity. The Q-Q plot shows the model has a Normal distribution. These plots can be improved, especially Normality on the far side of the line of best fit.

Figure 35: lm1.2013.aic Regression Terms

We can see from the regression term plots the contribution of each term to the linear model. TransectCount and TallestStemTrueLeaf and FloweringStageDoYFirstA have a positive effect on Yield and TallestStemLigule has a negative effect on Yield.

Once again, it appears that the values for FloweringStageDoYFirstF are more intermittent than the other variables.

The conclusion from developing this model is that TransectCount, TallestStemTrueLeaf, TallestStemLigule and FloweringStageDoYFirstF are the most important predictors for this year. The formula for the 2013 model is –

Yield = – 931.448 + 1-365 TransectCount + 8.29 TallestStemTrueLeaf – 4.834 TallestStemLigule + 1.888FloweringStageDoYFirstF

3.4 Final Linear Model

Let us use the 2012 linear model in order to try to predict 2013 Yield values.

Figure 36: lm1.aic Actual Values vs Fitted Values

This plot shows the actual 2013 Yield values against the values that the 2012 lm1.aic model predicted. It is seen that the model regularly overestimates. This can be seen better on the plot below which is the same plot on axes which are at the same scale.

Figure 37: lm1.aic Actual Values vs Fitted Values

For an entirely true prediction the points will lie perfectly across the line $x = y$ i.e. *Actual = Fitted*. It is clear that this model is not perfect since most of the points lie above the line. The points are scattered, especially towards higher values. The model always overestimates values above approximately 60.

Given the linear model assumptions; TallestStemTrueLeaf is removed from the model to see if this improves it. This variable is selected for removal from the model since TallestStemTrueLeaf and TallestStemLigule have such a close correlation, this can break an assumption of multi-linearity. FloweringStageDoYFirstA is also removed from the model as it has high multi-linearity with FloweringStageDoYFirstF.

We see an R^2 value of **0.6447** and a modified R^2 of **0.6366**. This new model is used to try to predict 2013 Yield values once again and plot the actual values against the predictive values –

Figure 38: lm2 Actual Values vs Fitted Values

Figure 39: lm2 Actual Values vs Fitted Values

We see a negligible improvement as a result of removing the two multilinearity variables.

The conclusion here is that it is likely that more sophisticated statistical methods are required to improve this model. The advantage of the model is that we know which variables will affect production and which properties to look for in genotypes for further development.

4 Conclusions

4.1 Key Findings

The key finding arising from this project is that the prediction of the behaviour and yield of Miscanthus plants is a complicated and difficult task with a very large number of possible variables and environmental factors to be considered.

However, we have developed an insight into how the phenotypic properties of Miscanthus interact with each other, with the yield at harvest and also how they change and develop annually.

Yield is a very complicated property and no single property can explain any more than 45% of the dry weight which varied from 0.6kg to 182.1kg. Using all properties in a linear model increased the explanation of the yield to 66% and remained at 66% when using a simplified model after running the StepAIC function.

The amount of biomass product produced by a plant is connected with the canopy phenology – the length of the canopy from flowering time to senescence and when considering the leaves on the plant. We see this in considering the properties chosen by StepAIC as the most important properties for predicting yield.

We have found that –

- The date of emergence is an important property in seeking to predict the yield of a Miscanthus genotype. If a genotype emerges early it is more likely to achieve higher heights and flower late thus producing more dry weight at harvest time. The date of emergence is therefore an exceptionally important property to monitor when developing Miscanthus genotypes. The date of emergence is more dependable for use as a property compared with canopy height. Canopy height is a combination of several properties and it is very likely that if a plant has an early emergence date it will also grow to a great canopy height.

- Date of flowering has a great influence on Miscanthus biofuel yield. Flowering means the plant will not grow any more and therefore ideally we want genotypes that flower late.

4.1.1 Genotypes for Further Development

From the results we understand that a genotype that will produce a high Yield will have the following properties –

- High TransectCount – more stems mean there will be more biomass at harvest time

- Early DoYFirst3Emergence – plants that emerge early usually reach a greater canopy height which means more biomass is harvested

- Late FloweringStageDoYFirstF – the ideal genotype flowers late in the year. After a plant has flowered, it does not grow any more so the later it flowers the higher the plant grows.

Let us look at the genotypes with the 30 highest TransectCount values, the 30 lowest values for DoYFirst3Emergence and the 30 highest values for FloweringStageDoYFirstF over the two years in order to select the best genotypes for further development. The table below shows a summary of the genotypes within the group of 60 genotypes having more than one of the ideal properties.

Genotype	High Transect Density	Early Emergence Day	Late Flowering Stage F Day
10	✓	✓	
12	✓	✓	
22	✓	✓	
40		✓	✓
41	✓	✓	
50	✓	✓	✓
69	✓	✓	
72	✓	✓	
77	✓	✓	
82	✓	✓	
89	✓		✓
98	✓		✓
102	✓	✓	
110	✓	✓	
116	✓		✓

Unfortunately none of the genotypes had all three properties but there are 15 genotypes which have two. When everything in this project is combined, these are the genotypes that should be further developed in order to try to improve the yield and quality of the Miscanthus crop.

4.2 Further Research

In order to improve this project, it would be useful to study a greater range of years i.e. study five or six years rather than two. Stronger conclusions can be taken from more data. Only two years of harvest data is enough to build a simple model but predictors will be better if there was more data to analyse.

In order to extend this project, it would be useful to see how the plants respond to drought conditions. Drought will become a more common problem as climate change worsens and so research into genotypes able to withstand drought is of great importance in this research area. Using the results of a project to examine the effects of drought on Miscanthus genotypes together with this project would lead to more certain findings on which types of genotypes develop further.

It is a pity that so much data is missing. Having removed rows which included *NA* values or values missing from the data entirely in order to use these in a linear model, only 118 plants out of the original 302 remained in the dataset. The statistical analysis would have been better if there were more data points to use. Due to the nature of taking measurements on live things like plants, it is unlikely that a dataset without missing or incorrect data could exist.

4.3 Critical Review

On the whole I am very happy with the development of this project. While I am disheartened that the linear model does not predict yield as well as I wanted, I am still happy with the content, presentation and development of the project.

The addition of interactive terms to the linear model may have improved it but given the diagnostic plots from the models created, it is clear that there is a bigger problem and that interactive terms would not solve this.

Looking back at the timeline of this project, I would spend less time plotting graphs and more time on statistical analysis instead. I spent a great deal of time plotting the data for 2011 without realising that this was the least important and was negligible for the main purpose of the project. Basically the three year distribution plots, paired plots and correlation plots show almost the same information. It is still useful to see the same data in different ways in order to see patterns and correlations that may not be seen in another plot.

The project would read better if the analysis were carried out in reverse chronology i.e. 2013 first since that is the most important year given the aim of the project. The data for 2011 was not very important in considering the prediction of the yield of the Miscanthus plants.

Including time series plots from the continuous data would have been useful for the project. Seeing the development of continuous variables over time visually would have added to my understanding of how the variables develop.

For the distribution plots with three years on the same axes, this could be an oversimplification of the data. There is a great variety between the plants of each genotype and so using a group mean is not the most valid form for comparing genotypes. However, if the genotypes had not been moderated in this way, the three year distribution plots would have been unreadable and difficult to draw information from.

This is the kind of variation seen between genotypes –

It would be very difficult to determine which genotype exactly is best since there are only three clones of each genotype and the difference in measurements within genotype is very big.

The personal development from completing a project of this size has been amazing. I am now confident and proficient in using the statistical language *R* and I hope it is a skill I will continue to use and develop. The same is true of handling LATEX. I had not used LATEX before this project and it has been a steep learning curve but very beneficial to learn. It is an extremely useful skill I hope to continue to develop and use.

An additional difficulty for this project was my choice of writing in Welsh. This has been difficult, especially considering scientific and mathematical terms lacking a clear translation. However, this has been a useful and valuable process for me. I am now more confident writing in Welsh than I was when I completed my A-levels in a Welsh language school before beginning university. This is a useful and invaluable confidence for me in seeking jobs where the ability to use and write fluent Welsh is necessary.

A Raw Data

Figure 40: Planting Layout Block 1

This is block 1 of the field plan for the Miscanthus hybrid plants. Three 9x12 blocks were planted, giving a total of 324 plants. The plants highlighted are the plants that die. We can see that the plants are randomly distributed.

Figure 41: Snapshot of 2012 Single Data

Note: Not all rows are shown here.
We can see that the data has a comparatively unclear and overcomplicated layout.

Figure 42: Snapshot of 2012 Continuous Data

Note: Not all rows or columns are shown here.
We can see again that the data has an overcomplicated layout as well as missing data.

Figure 43: Snapshot of 2012 Harvest Data

Note: Not all rows are shown here.
We can see that the harvest data agrees with the continuous and single data, we can pair each plant to its harvest value using the column UID. The column we use as the Yield value is the column WholeDW which is the dry weight of the plant at harvest.

B Prepared Data

Figure 44: Snapshot of 2012 Single Data

Figure 45: Snapshot of 2012 Continuous Data

Note: Not all columns or rows are shown here.
We can see that the paired data has a similar and intelligible layout.

C R packages Used

Package	Function	Description
MASS	StepAIC	The StepAIC function is used to complete a step by step regression on a linear model. A result of using the function is a model that is simplified without loss of effectiveness.
ggplot2	qplot, ggtitle, geom_point	ggplot2 is used throughout the project to create plots of the data. ggplot2 plots can be widely adapted and this was a useful feature.
gridExtra	grid.arrange	This function is used to organise three plots in one. This was very useful as there are so many plants to show on one graph.
corrplot	cor	This function is used to create a correlation plot of all variables. This was useful since there was no way to see one variable on a paired plot but it was possible to combine variables from the single data and continuous data on a correlation plot.
waffle	waffle	This function is used to create a waffle plot of the annual distribution of plants. This is a more effective way of showing data than a pie chart.

References