Report for the 4th Mathematics in the Plant Sciences Study Group
Centre for Plant Integrative Biology
University of Nottingham, January 2011
Project: Defining ideotypes in the biomass crop Miscanthus

Kerrie Ferrar
IBERS, University of Aberystwyth

Kim Kenobi, Simon Preston, Theodore Kypraios and Simon Pearce
University of Nottingham

Simon White
University of Cambridge

Sophy Thomas
University of Chester

April 4, 2011

1 Introduction

Miscanthus is a tall grass which has efficient C4 photosynthesis and so is capable of rapid biomass accumulation; it acts as a combined solar panel and battery, converting the suns energy into storable biomass by fixing atmospheric CO_2 into complex molecules in its cell walls. Unlike its close relatives sugarcane, sorghum and maize, it is capable of survival and rapid growth in temperate climates and is therefore a leading candidate for bioenergy in the UK and beyond. Miscanthus is an undomesticated genus native to a wide geographical range across South East Asia; extensive genotypic collections have been made from the wild and are being evaluated in the UK.

Very broadly, Miscanthus can be considered to consist of two species: M. *sinensis* and M. *sacchariflorus*, which show contrasting morphologies. M. *sinensis* is clump forming with numerous fine stems while M. *sacchariflorus* has fewer, thicker stems and a spreading habit. M. x *giganteus* is a naturally occurring triploid hybrid between M. *sinensis* and M. *sacchariflorus* and is currently the single leading genotype of Miscanthus which is grown commercially for bioenergy. It combines many good traits from the two parental species but cannot be further improved via breeding as it is sterile. Not only do we need to broaden the genetic base of the crop but there are clear opportunities for huge improvements in terms of yield and quality in comparison to M. x *giganteus*.

2 The Dataset

In 2005 a replicated trial of 244 genotypes, including both M. *sinensis* and M.*sacchariflorus* types, was planted in Aberystwyth to assess the diversity within the genotypes, which had been brought to Europe previously, mainly by taxonomists and the horticultural industry. Although the majority of these plants do not represent the full potential of Miscanthus for biomass, they provide a unique resource in which to study diversity within the different genotypes and to link plant morphology traits to yield to enable development of ideotypes i.e. idealised plant forms towards which breeders can aim. The trial was extensively phenotyped in 2008 and 2009 for continuous biomass traits such as spring emergence (SE) and flowering time (FT), which can be combined to give a value for the active growing period for each plant, and canopy height (CH). In Miscanthus the harvested biomass consists mainly of stem material and so at the end of the growing season further phenotyping was undertaken to assess the stem morphology (height, diameter and number). Final yield was calculated following the spring harvest, which is delayed until after winter to improve the quality of the biomass in terms of moisture content (MC) and nutrient remobilisation/leeching.

2.1 Measurements taken

A range of measurements were taken on the Miscanthus plants. Some of these measurements were easy to take and could therefore be taken on all of the plants (approximately 1000 in total including the four replicates of 244 genotypes). On a subset of the genotypes (38 genotypes with four replicates) a series of more labour intensive measurements were taken. In Table 1 below we list the measurements taken, together with an indication of whether they were taken on all 244 genotypes or only the subset of 38 and a description of the meaning of the measurement.

3 Pairwise Analyses

Essentially our aim was to find good predictors of yield. We defined yield as the dry matter content of the plant. However, there are also other variables that need to be considered as well. If the plants are going to be harvested for bioethanol, it is desirable to have stems that are high in cellulose, low in hemicellulose and low in lignin. (See for example the presentation at http://www.agrireseau.qc.ca/energie/documents/Biofuel_crops.pdf by Donald L. Smith of McGill University.)

Measurement	Number of plants	Description of data
Species	1000	*sacchariflorus*, *sinensis* or hybrid
Year	1000	2007, 2008, 2009
Base Diameter *	1000	The diameter in mm of the base of the plant
Transect Count *	1000	The number of stems close to a stick that was passed through the plant
Tallest Stem * #	1000 only 2007/08	The height in cm of the tallest stem (Longest stem selected by eye) at time of harvest
Max Canopy Height * #	1000	Maximum height of canopy at time of harvest
Stem Diameter *	1000 only 2007/08	Average stem diameter at time of harvest
Day of year of first flowering *	1000	Day of year as number between 1 and 365/366
Moisture content * #	1000	The percentage of moisture in the whole plant when harvested
Dry Matter * #	1000	The mass in kg of dry matter in the yield
Cellulose #	1000	% cellulose in yield
Hemicellulose #	1000	% hemicellulose in yield
Lignin #	1000	% lignin in yield note these three percentages do not necessarily add to 100% since there is other material as well
Maximum stem height	38 only 2007/08	Length in cm of longest stem (All stems measured hence (cf Tallest Stem above) - only available for 38 genotypes)
Stem Count	38 only 2007/08	Number of stems in plant
Flowering Stem Count	38 only 2007/08	Number of stems in previous row that are flowering at time of harvest

Table 1: The measurements taken on the Miscanthus plants. The asterisk and the hash symbols indicate variables that are used to generate figures 1 and 2 respectively

In our analysis, we first considered the pairwise relationships between the variables given in the table above. Scatterplots of pairs of variables are a useful way to visualise obvious correlations between the variables. Figure 1 below shows scatterplots of the variables marked with an asterisk in the table of variables above. All of these variables are measured on the full set of 1000 plants.

If we only consider the column of plots of dry matter against each of the other seven variables, the two variables that show the highest correlation with dry matter are maximum canopy height and tallest stem. These two variables are themselves highly correlated, which is not surprising given that they are measuring very similar features of the Miscanthus plants. It is also not surprising that taller plants give higher yields.

In Figure 2 we show the pairwise scatterplots for the variables marked with a # symbol in Table 1. (Note that the variables maximum canopy height, tallest stem, moisture content and dry matter feature in Figures 1 and 2.) From Figure 2 it is easy to see that maximum canopy height (and to a lesser extent tallest stem) is strongly negatively correlated with cellulose levels. Cellulose levels are negatively correlated with dry matter levels. Thus it may not be possible to define an ideotype of Miscanthus in which cellulose levels and yield (dry matter) are both high.

4 Linear modelling - all 1000 plants

A linear model is a model of the form

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon_i,$$

where Y_i is the value of the response variable for the i^{th} observation, β_0 is an intercept term, the β_js are the regression parameters with corresponding explanatory variables X_j and ϵ_i is the i^{th} error term.

In our case, we take the response variable as the dry matter (yield), and explore which explanatory variables account for a good proportion of the observed variability.

4.1 Model 1 A naive model

Here we consider the following explanatory variables: Species, Year, Base Diameter, Transect Count, Tallest Stem, Maximum Canopy Height, Stem Diameter and Day of Year of First Flowering.

Fitting this linear model gives an adjusted R-squared value of 0.7041, indicating that the model explains approximately 70% of the observed variability. Most of the explanatory variables are significant at the 0.1% level. The exceptions are Year ($p = 0.865$), Base Diameter ($p = 0.0135$) and Tallest Stem ($p = 0.784$).

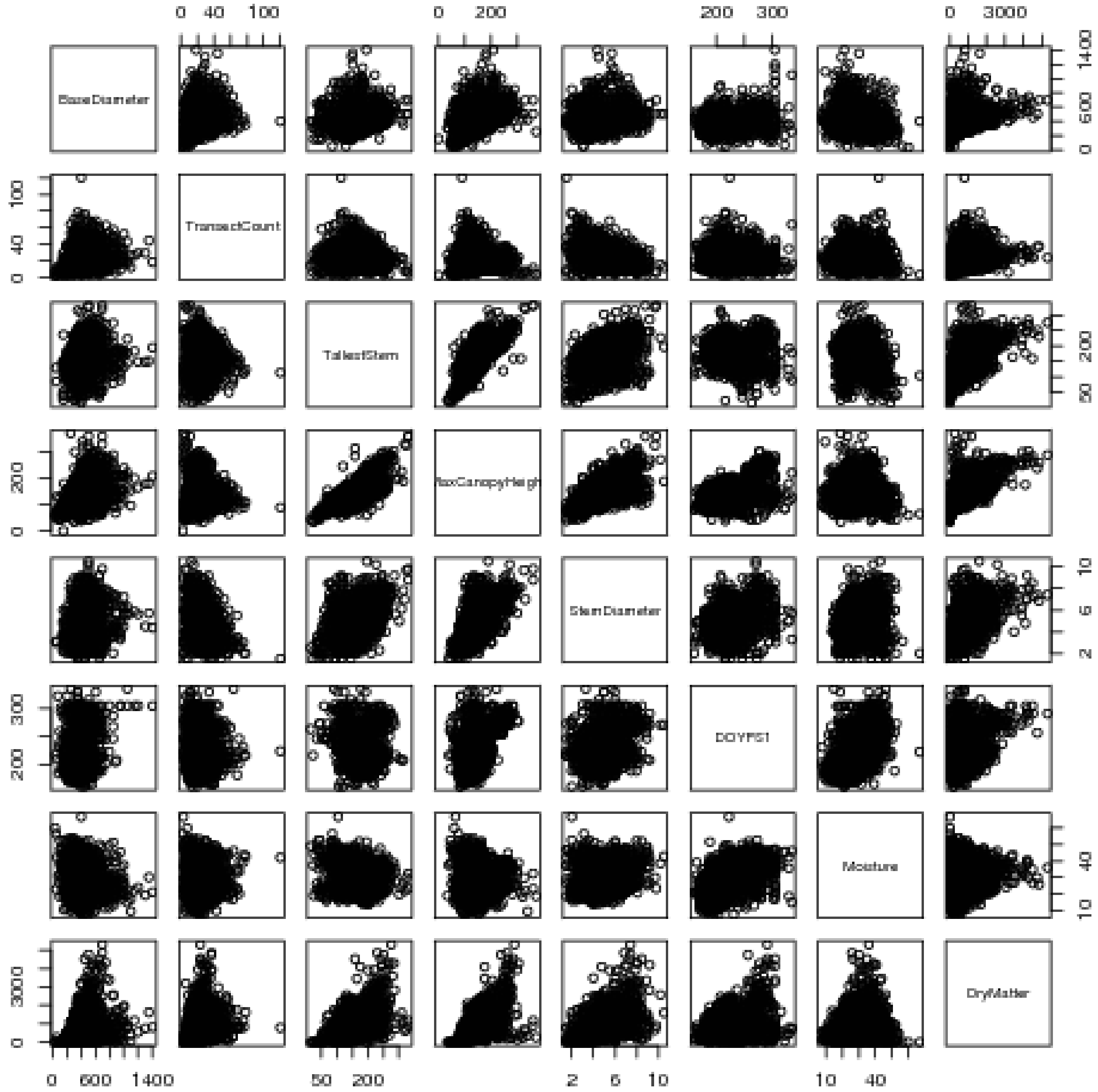


Figure 1: Scatterplot of the variables marked with a * in Table 1

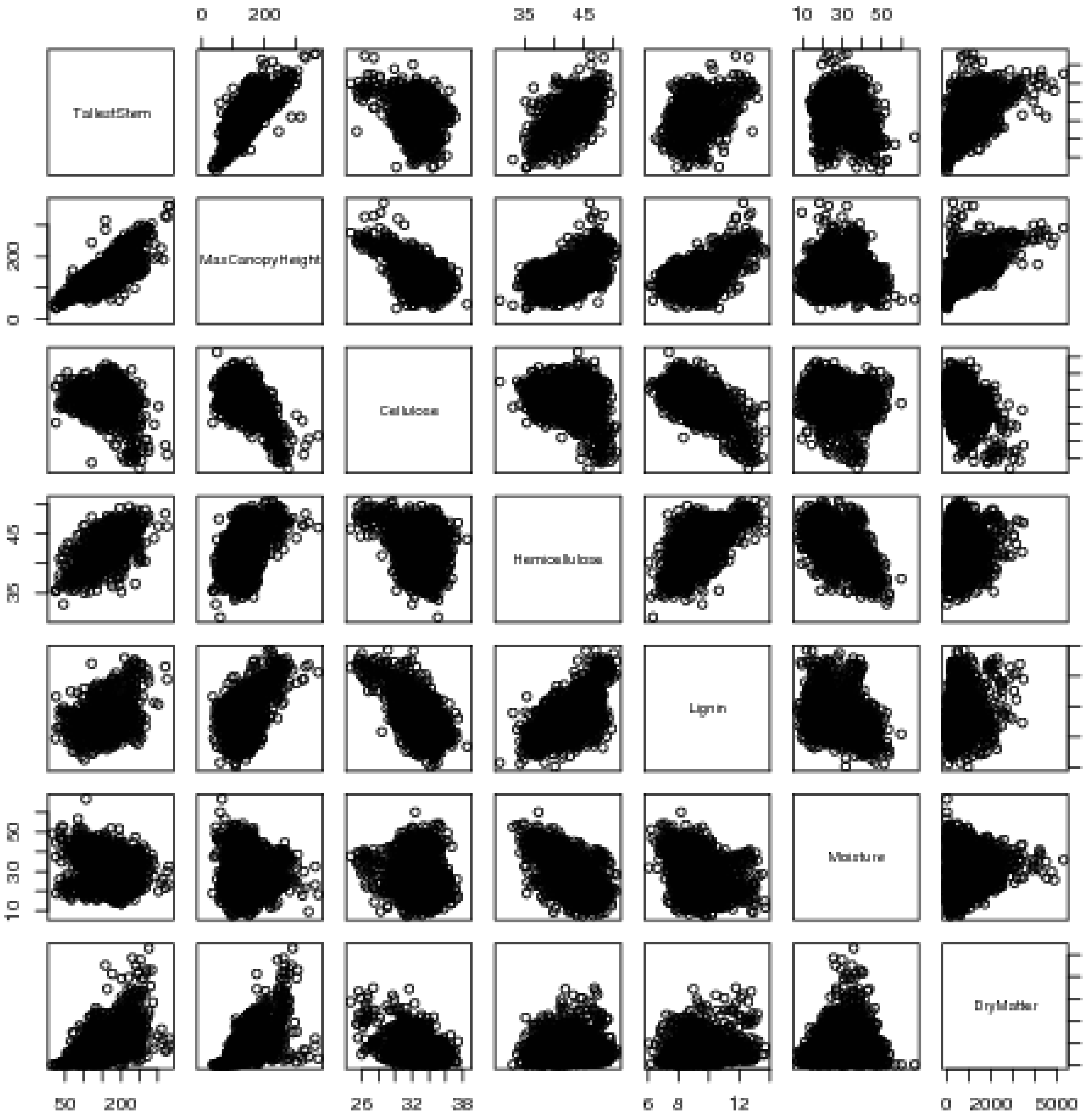


Figure 2: Scatterplot of the variables marked with a # in Table 1

Variable	Estimate for β_j	Std. Error	t value	Pr($>|t|$)
(Intercept)	-2240	118	-19	1.17e-72
sacchariflorus	-180	65.8	-2.73	0.00631
sinensis	-101	57.2	-1.78	0.076
Year (2009)	34.1	19.2	1.77	0.0764
Area	-0.00646	0.000565	-11.4	4.2e-29
Volume	1.61e-05	1.93e-06	8.3	2.27e-16
TransectCount	14.5	0.878	16.6	1e-56
StemDiameter	22.1	8.96	2.47	0.0135
DOYFS	2.17	0.318	6.84	1.15e-11
MaxCanopyHeight	8.99	0.602	14.9	4.01e-47
BaseDiameter	4.32	0.352	12.3	3.75e-33

Table 2: The details of model 3

4.2 Model 2 - Estimating plant volume

In this model, we create a new variable - Volume, by calculating

$$Volume = BaseDiameter^2 \times CanopyHeight.$$

Since Tallest Stem did not emerge as at all significant in Model 1, we drop this from Model 2. The explanatory variables we use for Model 2 are:

Species, Year, BaseDiameter, Volume, Transect Count, Stem Diameter, Day of Year of First Flowering, Maximum Canopy Height.

In this case the adjusted R-squared is 0.7068 a tiny improvement on Model 1. In this case all of the variables are significant at the 0.1% level except for Year.

4.3 Model 3 - Including Area as well as Volume

As a final experiment, we consider including the variable ‘Area’, given by

$$Area = BaseDiameter^2,$$

as well as the volume variable defined in Model 2.

This improves the adjusted R-squared value to 0.7299 and many of the variables are highly significant (i.e. $p < 0.001$). The details of this model are shown in Table 2.

From Table 2 we can see that many of the estimates are positive. In particular, an increase in the volume variable corresponds to an increase in yield, as does an increase in Transect Count, an increase in Stem Diameter, later first flowering, an increase in Maximum Canopy Height and an increase in Base Diameter.

Note that the coefficient for Area in this model is negative. Although on its own this is somewhat counterintuitive, it arises as a result of competition between the Area and Volume variables.

5 Linear modelling - using more detailed measurements on 38 genotypes

One aspect of the problem presented related to the actual measurement process for the covariates of interest. Several of these required a great deal of effort and time to collect, for example, the total number of stems of each plant is a time intensive counting activity.

Thus as part of the trial, a subset of 38 genotypes were chosen from the 244 to be measured in further detail over the years 2008 and 2009, while still being measured as the remaining plants. This dataset consists of 300 observed plants ($38 \times 2 \text{ years} \times 4 \text{ plots}$, less incomplete sets).

Our aim here is to model the total dry matter yield in terms of both the detailed and simple measurements for these 300 plants, and thereby express the detailed measurements in terms of the simple ones.

5.1 Initial exploration

Figure 3 plots the dry matter yield across the years and replicates for all plants. The measurements of 2007 are to be discounted as a refinement of the protocol and since many of the plants appear to not have matured by this time. In 2008 and 2009 there seems to be no obvious difference in the majority of the samples, however there is a indication that the outliers with a greater yield are different. This is unfortunate, since these are the ideally the plants of interest, and yet many standard statistical methods struggle with what may be non-normal distributions.

We begin by modelling the total yield using the detailed, time intensive and more accurate measurements on the 300 plants. After some experimentation, the details of which are not shown, the resulting model is:

$$\begin{aligned}&\log(\text{DryMatter}) = \\ &\qquad -5.152 + 0.901(\text{StemDiameter}^2 \times \text{MaxCanopyHeight} \times \text{StemCount}),\end{aligned}$$

and Figure 4 shows the fitted and actual yields under this model. The model was fitted to offer a balance of good predictive power and parsimony. Recall, we are typically interested in the higher yield plants and so the unsatisfactory fit for smaller yields is not such a concern.

5.2 Using the simpler alternatives

Unfortunately, the measurement of StemCount, the total number of stems for each plant, is time consuming and so we would like to instead use simpler measurements.

We consider the TransectCount, a quick method to count the number of stems across the largest transect of the plant, BaseDiameter and StemDiameter.

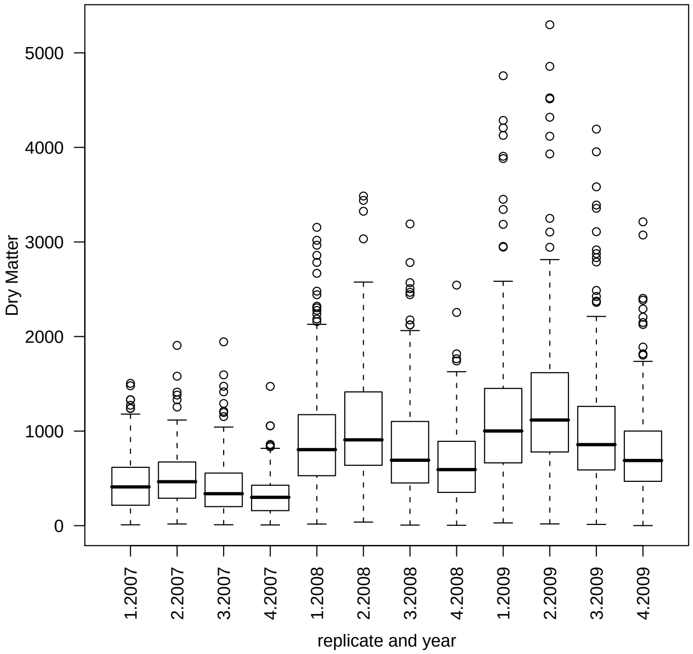


Figure 3: Comparison of Dry Matter yield across years (2007, 2008, 2009) and replicates in the trial (1, 2, 3, 4). Excluding 2007, there seems a difference in the outliers, but not in the median of the samples.

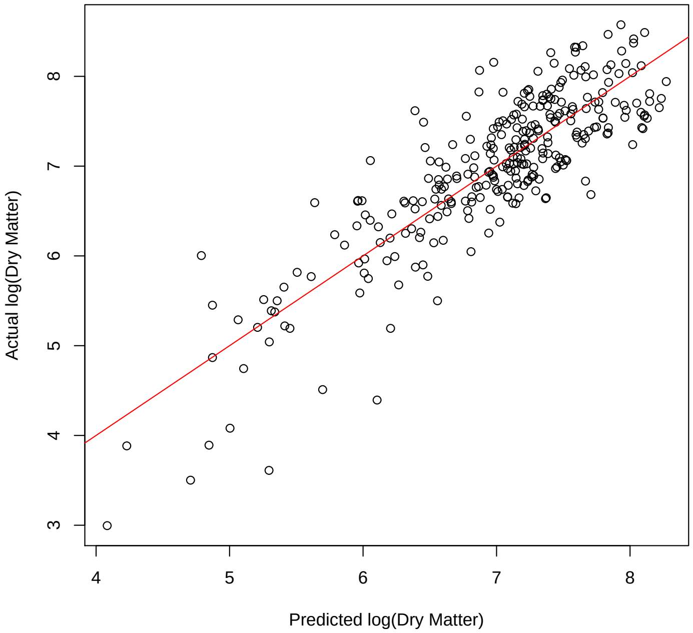


Figure 4: Predicted Dry Matter yield using the full detailed measurements on the 300 plants.

We fit a polynomial linear regression of the logarithm of TransectCount to model the StemCount and obtain the fit:

$$\begin{aligned}\log(\mathrm{StemCount}) = \\ &- 0.045 + 2.377 \log(\mathrm{TransectCount}) - 0.233 \log(\mathrm{TransectCount})^2.\end{aligned}$$

Thus StemCount can be modelled as a non-linear function of TransectCount, namely a quadratic in its logarithm. The fitted line is shown in Figure 5, though these is still significant variability about the predicted values. Including additional categorical terms of the year and replicate of each plant greatly increase the model fit, explaining more of the variation. However, if year is a factor due to plant maturity this should be evident in other direct measurements on the plant and no environmental data were available. Nor can we reasonably explain a within replicate difference, and Figure 3 does not suggest one for the complete set of plants. (Note that Figure 3 is the data for all 244 genotypes across 3 years.)

If we attempt to include further covariates to model the StemCount, again using poly-

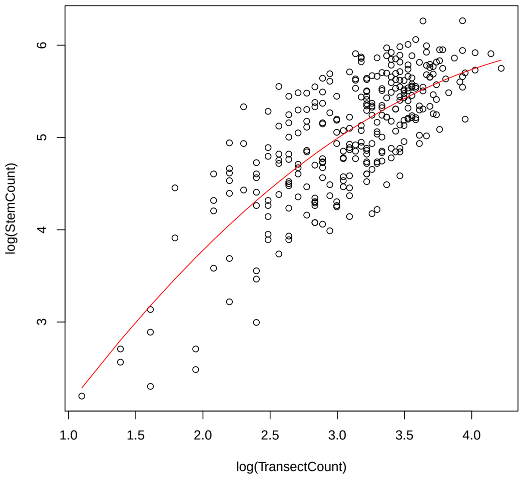


Figure 5: Fitted polynomial of TransectCount to StemCount.

nomial fitting, we can obtain the following model,

$$\log(\mathrm{StemCount}) = 0.93 + 2.28\log(\mathrm{TransectCount})$$
$$- 0.23(\log(\mathrm{TransectCount}))^2 - 7.75 \times 10^{-7}(\mathrm{BaseDiameter}^2) - 0.09\mathrm{StemDiameter},$$

where the additional terms are significant, and improve the model fit. However, their effect size is minimal (despite the BaseDiameter being measure in millimetres).Thus we do not consider this model further except to plot the fitted verses residuals for the model in Figure 6.

5.3 Modelling yield using predicted StemCount

Finally, we are able to model the dry matter yield, substituting the StemCount with our predicted StemCount from the previous section. It must be noted that we fail to properly propagate the variance of the predicted StemCount through our modelling. To achieve this would require a true multistage modelling approach, instead of the independent steps used here.

Thus, the model of interest is, having replaced StemCount by its predicted value and adapting the StemDiameter into polynomial terms in order to adjust for dropping MaxStemHeight, we obtain

$$\log(\mathrm{DryMatter}) = -2.07 + 4.77\log(\mathrm{StemDiameter})$$
$$- 0.81\log(\mathrm{StemDiameter})^2 + 0.68\mathrm{Predicted}\{\log(\mathrm{StemCount})\}.$$

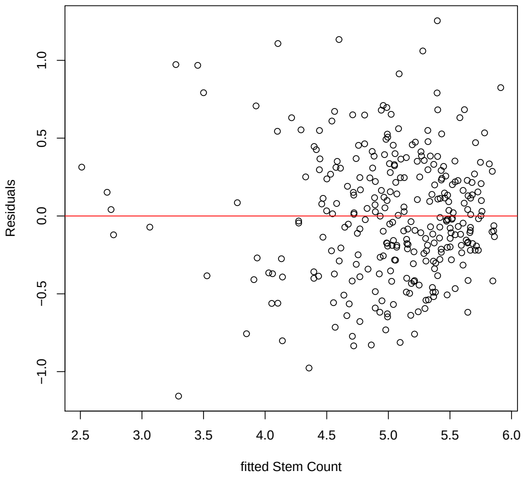


Figure 6: Fitted verses residuals for modelling StemCount as polynomial terms in TransectCount, BaseDiameter and StemDiameter

Note that the Predicted (log(StemCount)) is calculated using the equation in Section 5.2 above. As with all models considered, no interaction terms are significant, and the inclusion of cubic terms results in insignificant coefficients tested with nested likelihood ratio tests.

The residuals and qqplot are shown in Figure 7, from which we see the general fit is satisfactory. However, there are issues with the lower and upper tails of the model in terms of the normality assumption. Something that was of concern given Figure 3.

Finally, it is left to compare the predicted Dry Matter yield using the measured and predicted StemCount. Figure 8 compares the residuals of the measured and predicted StemCount models, showing a generally good fit for both models. Also, the comparison between the two model predicted dry matter yields is fairly consistent.

5.4 Detailed measurements summary

We have considered a sub-component of the larger Miscanthus trial, specifically considering the issue of determining predictive and efficient covariates for modelling the dry matter yield. There are many further issues to consider, such as the proper propagation of the uncertainty of the predicted StemCount into the predictions for the yield, and the bivariate nature of the dry matter yield and moisture content. Further, since the primary interest is in larger yielding plants, the issues of normality in the upper tail should be considered.

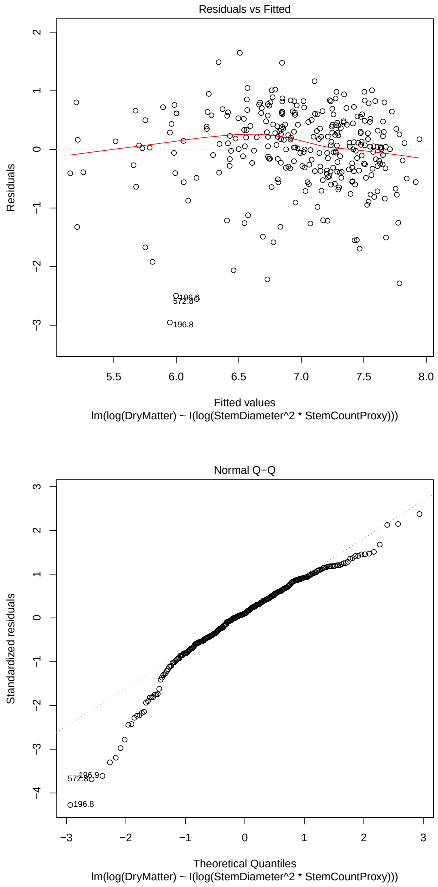


Figure 7: Residuals and qqplot for predicted yield using predicted StemCount

6 Discussion

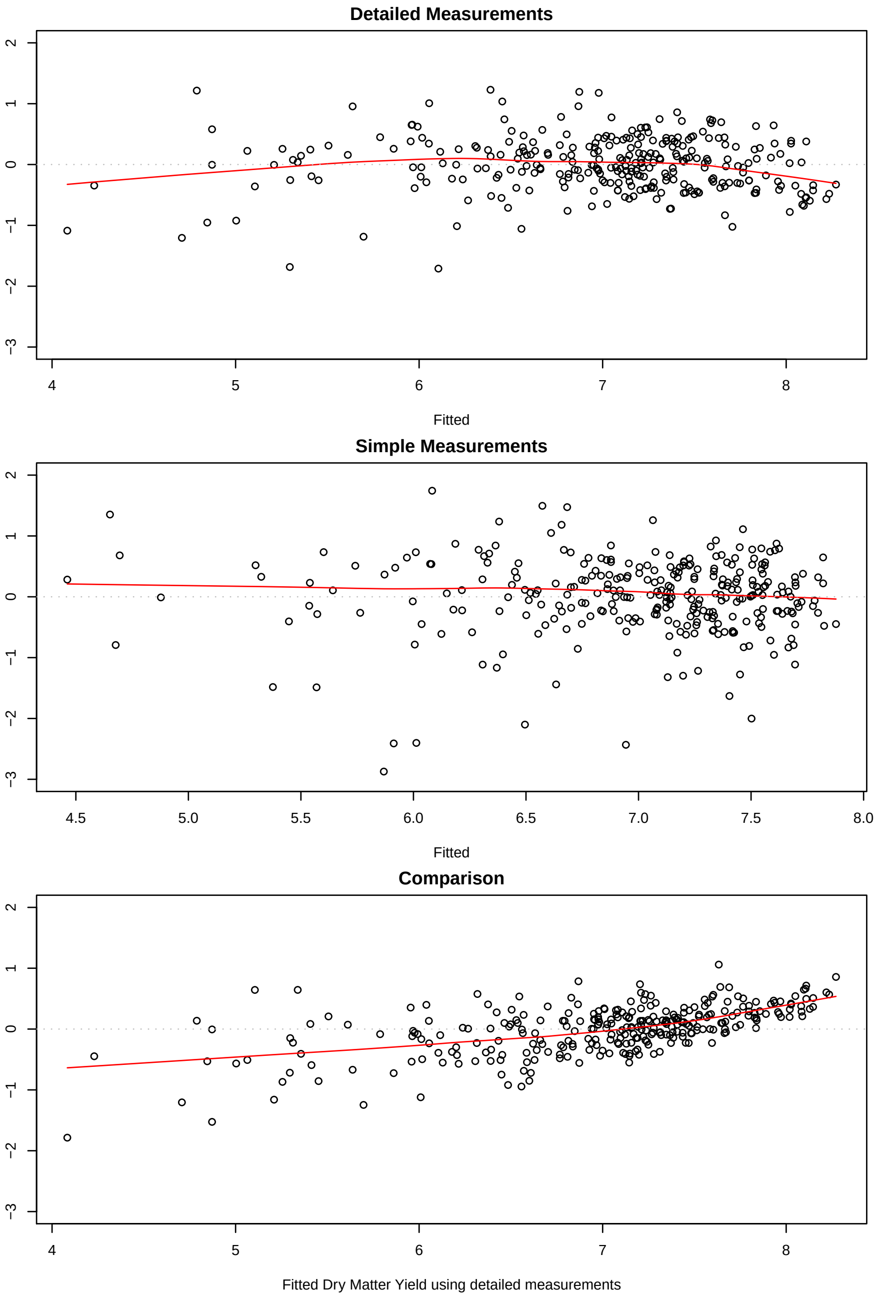
We have considered the application of linear modelling to a fairly large dataset from a field study of the plant Miscanthus, which has potential as a bioenergy crop. Some of the measurements that we used as potential explanatory variables for determining yield were relatively easy to take, and are therefore available on the entire set of approximately 1000 plants. Other measurements were more labour intensive, and are only available for a subset of approximately 38 of the 244 genotypes.

We started by looking at pairwise relationships between the measurements on all of the plants. The variable we used to quantify yield was the mass of dry matter produced by the plants. The variables that emerged as showing the strongest correlation with dry matter were the height of the tallest stem, and the closely related covariate maximum canopy height.

We built linear models, initially focussing on the traits measured on the entire set of plants. In a linear model that included ten explanatory variables with no interactions, we were able to explain over 70% of the variability in the data. This model included two derived explanatory variables, which were proxies for the volume of the plant and the cross-sectional area of the plant.

The other aspect of the linear modelling was to assess how we could express the more detailed measurements in terms of the less time-consuming measurements. This would enable us to extrapolate the findings from the subset of 38 plants and obtain simple models of yield for the full set of 1000 plants. We fitted a model to capture the predictive power of the time-consuming measurement StemCount using a function of the more easily measured TransectCount.

Overall, the linear modelling shows significant explanatory power and we need further investigation to more fully assess our ability to make predictions about which traits determine the yield of Miscanthus plants.



Detailed Measurements

Simple Measurements

Comparison

Figure 8: First two plots compare the fitted verses residuals for the models with the measured and predicted StemCount, and finally the difference between the two fitted dry matter yields. The red lines show the regression lines fitted using a form of local regression that uses tobust locally linear fits.