20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Generalised linear model and analysis of cereal plant biomass

M. Cespedes [a], **J. Cai** [b]

[a] *School of Mathematics and Statistics, University of South Australia, Mawson Lakes, SA 5095.*
[b] *Phenomics and Bioinformatics Research Centre, University of South Australia, Mawson Lakes, SA 5095.*
Email: cesmi001@mymail.unisa.edu.au

Abstract: Numerous literature discuss methods, models and results achieved by applied image analysis techniques in order to deduce plant information. Although all methods required data from destructive testing their recommendations suggest the expansion of their methods onto an automated system is within reach. This new system will provide a quicker, cost efficient non-intrusive way to gather plant information as an alternative to destructive testing. The estimation of plant biomass is important to many applications such as plant breeding and agriculture. The generic biomass model presented here is based on the separation of plant components which opens up the potential to account for the structural and density differences of plant components.

Various statistical techniques have been employed to assess the improvement of previous biomass models and challenges which arise when regression models do not conform to assumptions. This paper offers an alternative model which preserves the original data and linear model template, yet accounts for increasing systematic variability present in the data. Both informal and formal statistical methods were used to asses the goodness of fit and comparison between the commonly used method of least squares regression and the generalised model proposed here.

These models are an extension to those found in literature of plants under experimental conditions and the intention is to apply this model coupled with image analysis in order to deduce plant information in a non-intrusive way. The proposed biomass models presented here aim to combine image processing techniques with mathematical modeling in an effort to replace destructive testing.

Keywords: *Cereal biomass estimation, image processing, ordinary least squares, generalized lineal model, analysis of deviance, Akaike Information Criterion*

1 INTRODUCTION

Cereal plants provide a range of staple food in our society and on a world wide scale. Tester and Langridge (2012) highlight the need to improve current cereal strains, in order to maximise productivity and reach the increasing demands of a growing population. Plant scientists work closely with members from other disciplines to improve current methods in conducting experiments, in order to better study the behavior of plants subjected to genetic alterations. A common method to collect the required plant information is through destructive experiments. This time consuming method requires handling of every individual plant and destroys the subjects under study, in order to attain the required information. The non-intrusive method of photo imaging analysis will preserve plants intact for future study while enable us to collect the required information from test subjects.

The generalised mathematical models presented here were derived from several cereal genotypes, treatments and across a wide lifespan. The data collected are the results from destructive testing as well measurements taken from RGB images which were computed manually. The models are an extension to current estimation of dry plant biomass from literature available for grass species. Further discussion and extension of their methods found in literature are discussed in Section 2.

The long term intention is to combine the information obtained from image analysis, mainly leaf surface area and stem height as outlined in Section 3, with mathematical models presented here. Section 4 provides a brief background of the experimental data used as well as the formulation of our biomass models based on separated plant features. This paper proposes the preliminary process of a novel method to derive plant information utilising image analysis methods for an automated system in the near future.

2 CURRENT BIOMASS METHODS FOUND IN LITERATURE

Studies from Arvidsson et al. (2011), Berger et al. (2010), Leister et al. (1999), Munns et al. (2010), Paruelo et al. (2000) and Tackenberg (2007), highlight the benefits of using image analysis, both in RGB and infrared, to capture plant information such as plant surface area, fresh weight, vertical biomass, dry matter content, stomatal conductance, etc. These benefits include the efficiency of processing large amounts of plants quickly, under experimental conditions, as well as continued study of the same subjects into further development, as they remain undisturbed throughout the data collection stage.

Various studies suggest different methods to ascertain certain aspects of plant characteristics. An early use of imaging techniques on plants, found in Leister et al. (1999), studied two ecotypes of *Arabidopsis thaliana* which in general has a flat two dimensional structure. The aim of this study was to develop growth rate models from leaf area measured by imaging techniques. Tackenberg (2007) was another article of interest, as it too developed growth rates models but this time on grass species which have a complex three dimensional structure. Their methods required segmentation of the plant silhouette to estimate leaf surface area and so compute vertical biomass.

Cereal crops and general grass species have a non-linear relationship between the vegetation weight and the number of green pixels captured in its image. The generic biomass models presented here are based on the separation of plant elements so that it may be used to account for the structural and density differences in plant components. This will allow us to conduct phenotyping analysis of each plant component as well as the plant as a whole and study their relationships. We aim to improve the estimation of biomass by considering the plants physical make up in our model as opposed to relying on the number of pixels captured in an image.

A relationship between this measurement of green vegetation, mainly the estimated leaf surface area, is the common method employed by many such as Tackenberg (2007), Leister et al. (1999), Munns et al. (2010) and Paruelo et al. (2000), for an estimation of plant biomass. We wish to expand on this notion and employ image analysis to distinguish plant features, such as stems and leaves in order to improve the estimation of plant biomass.

Ferryman et al. (2000) state object segmentation in images has been an important research topic in image processing and computer vision for several decades. As a result Cremers et al. (2007) and Ning et al. (2010) are a few of the developers which embrace different approaches and algorithms for object segmentation. After comparing several popular algorithms, in this instance we adopted the method developed by Cai et al. (2011) for plant segmentation. For the estimation of the plant stem height, we employ a popular shape detection method, Hough's Transform as described in Duda and Hart (1972).

Despite the significant technological advancements made, Frubank and Tester (2011) conclude plant phe-

nomics is still in its infancy. Currently there are no automated methods available to separate plant components and measure their corresponding growth patterns under stress. The data used to derive the biomass models presented here were from early to late growth stages, which have undergone varying treatments and incorporate characteristics from both destructive testing and RGB images.

3 IMAGE DATA ACQUISITION

The fully automated imaging station at the Australian Plant Phenomics Facility (APPF) Plant Accelerator can take thousands of plant images from different cameras and angles over a single day. Due to the amount of data collected, it is desirable to extract plant silhouettes and stem height data from these images automatically. The method by Cai et al. (2011) was used to segment the plant vegetation from the background. This method incorporates the difference between two images to obtain an initial estimation of the background which was refined by the Expectation-Maximization (EM) algorithm. Hough's transform by Duda and Hart (1972) was utelised to locate the stems for a given plant, from which the sum of stem height in millimeters was estimated. Figure 1 illustrates the use of both of these methods, preliminary testing on various plant images has yielded promising results.

Figure 1. Left - Automated plant segmentation. Right - Stem detection example.

Further investigation and fine tuning of these two image analysis methods remains to be conducted. However the authors feel confident the implementation of these algorithms will allow for an automated process to derive the necessary plant information required, in order to provide the numerical data for the above ground dry biomass models presented here.

4 OLS AND GLM BIOMASS MODELS

The data for the following models are the results from two experiments (24 and 56) conducted at the APPF. Gladius and Drysdale were the two cereal varieties under study which consisted of 43 and 40 plants respectively. The biomass models considered apply only for the early growth stages to 43 days, as a significant increase in weight was observed once the grain emerged. The plants underwent the following treatments; mild drought (D1), sporadic drought (D2) and well watered (WW).

Equation 1 was our initial model, parameters values were obtained by Ordinary Least Squares (OLS) whose values are listed in Table 1. Our objective is to provide a generalised model for cereal plants which incorporate physical plant features such as Leaf Surface Area (LSA) in mm^2 and the sum of stem height ($\sum L_i$) in mm which in future will be derived by imagine analysis techniques as described in Section 3. This generalised model for varying growth stages and treatments relies initially on LSA and its squared term at the beginning of the plants life-cycle. As the plant matures parameters $\sum L_i$ and sum of stem height squared ($\sum L_i^2$) begin to play a significant role in the biomass estimation of the plant.

$$Y = \alpha_0 + \alpha_1(LSA) + \alpha_2(LSA)^2 + \beta_1(\sum L_i) + \beta_2(\sum L_i^2). \quad (1)$$

An underlying assumption for the OLS regression model $Y = XA + \epsilon$, is dependent variable Y to be normally distributed. Hence the expectation $E[Y] = E[XA + \epsilon] = XA$ implies errors $\epsilon \sim N(0, \sigma^2)$. Analysis of this

model confirmed the variability of our data increases as both LSA and $\sum \mathrm{L}_i$ increase, despite the models high R^2 value of 0.94. Informal diagnostic checks of the residual plot showed the values increased symmetrically about zero and the corresponding histogram of the errors were positively skewed.

Parameter units	α_0 grams	α_1 gram/mm^2	α_2 gram/mm^4	β_1 gram/mm	β_2 gram/mm^2
OLS	0.0727	5.18×10^{-5}	6.59×10^{-11}	6.21×10^{-4}	2.27×10^{-6}
GLM_1	-0.027	4.72×10^{-5}	4.97×10^{-11}	1.75×10^{-3}	4.88×10^{-8}
GLM_2	-0.026	4.71×10^{-5}	5.14×10^{-11}	1.77×10^{-3}	-

Table 1. Coefficients for OLS and GLM models.

Two options to consider were to either collect additional data (if possible), re-apply the model and see if we then conform to normality assumption or consider the data of this nature is simply not normally distributed and the OLS multivariate linear regression is an incorrect model. The Kolmogorov-Smirnov test described in Miller (1956) was used to asses the normality of the residuals. Our null hypothesis was as follows: The residuals from our regression model come from a normal distribution. Versus the alternative hypothesis: The residuals from our regression model do not come from a normal distribution. The significance level for our two-sided hypothesis test was set at $\alpha = 0.05$. The test statistic was computed to be $\mathrm{D}_{83} = 0.8386$ and this was tested against $\mathrm{D}_{83,0.05} = 0.1326$. The null hypothesis was rejected as $\mathrm{P}(\mathrm{D}_n > \mathrm{D}_{n,\alpha}) = \alpha$. The result of this test reflects the diagnostic plots observed earlier, these results greatly increase our confidence the residuals do not come from a normal distribution.

Several transformations of the data were attempted in an effort to improve the residuals fluctuation in variability, but unfortunately they remained hereroskedastic. An alternatively we also considered non-linear models as suggested in Archontoulis and Miguez (2013). As our initial model has strong biological arguments and interpretation for each parameter, our approach deviates from Archontoulis and Miguez (2013). We simply need to account for the increase of variation as the response variables increase. This lead us to the application of Generalized Linear Models (GLM) as described in Fox (2008), Nelder and McCullagh (1989). GLM provides an extension to the usual linear regression models, normality and constancy of variance are no longer required. We generalize Y to have a distribution not necessarily of the normal distribution, but of the exponential family. Here we made the assumption for Y to have a Gamma distribution, as our response variable $Y_i \in (0, \infty)$ and we suspect the model variance increases approximately with the mean. Parameter estimates for the GLM were computed iteratively using numerical methods of Maximum Likelihood Estimates as there is no closed form. The canonical link function was chosen to be the identity as this preserves our initial model structure in equation 1. An underlying assumption with this new model is the conditional variance of the response grows with its mean but the coefficient of variation of the response remains constant.

The analysis of deviance table below summarises the comparison between the null and full model.

```
Response: VegetationW
Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         82    108.112
LSA         1  100.628        81      7.484 < 2.2e-16 ***
LSA.2       1    0.609        80      6.875 0.0007623 ***
SumStemH    1    3.203        79      3.672 1.153e-14 ***
SumStemH.2  1    0.000        78      3.672 0.9730389
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The GLM models and analysis were computed in R, parameters for both models are on Table 1. The analysis of deviance table reflects predictor variable SumStemH.2 failed to reduce the residual deviance when its term was added sequentially. The GLM_1 model was computed repeatedly with terms in various order and unfortunately SumStemH.2 consistently failed to reduce the deviation. Hence this term was removed and the GLM model was computed again with the remaining four terms, the final row on table 1 shows the correct final model. The overall residual deviance for the GLM_2 was not significant. As we had 79 degrees of freedom and residual

deviance of 3.67, we may compute $1 - \chi^2_{79,3.67} = 1 > 0.05$, which indicates the model has captured majority of the variance. The diagnostic plots for the GLM_2 model provided an informal way to assess the goodness of fit, the authors are happy to provide and further discuss these plots upon request.

The residual vs fitted plot in the OLS model displayed a systematic trend, this lead us to reject the initial model in pursuit of an improved version. The same plot for the GLM model showed the residuals scattered approximately symmetrically about zero and displayed no systematic trend, with the exception of a single residual which indicated an was outlier present in the data. The scale-location plots provide an informal check of the adequacy of the assumed variance function for the model. It displayed no systematic trend or unusual clustering. This plot was particularly important, as it indicated there were no major violations to our initial variance assumption as stated earlier.

The residual vs leverage plot provides an insight for any significant outliers present. Cook's distance as discussed in Cook (1977), provides a value between zero and one for a specific data value which corresponds to an estimated influence that data value has on the model. All the diagnostic plots indicated an outlier was present. This warrants further investigation, as this outlier may have had a large influence on the parameter values. As all plants are unique, plant biomass may be greatly influenced by various aspects, some of which are beyond the experimental scientist control. It comes of no surprise to have a few observations which deviate from the norm. The outlier present in the data corresponds to a two day old barley seedling under WW treatment from experiment 56. This particular plant had an unusual low biomass to LSA ratio which we believe was the reason the information for this plant to differs significantly from the rest. We had no immediate reason to assume the LSA, biomass and corresponding RGB image were incorrect, hence the entire data set remained in tact. Figure 2 demonstrates the application of the GLM_2 model and illustrates with its 95% confidence interval how the variance increases with the mean. The GLM_2 model will differ from OLS method in the estimation of biomass, as for a given RGB image for plant j, the values LSA_j and $[\sum \text{L}_i]_j$ will be estimated by image analysis. This model can then compute $\text{Y}_j \pm Z_{1-\alpha/2}\sqrt{V[Y_j]}$ which is the estimated biomass with corresponding $(1-\alpha)100\%$ confidence interval as a tolerance. For small values of LSA and $\sum \text{L}_i$ the standard error will be relatively small and conversely for large explanatory values the standard error will reflect the increased variability of the data.

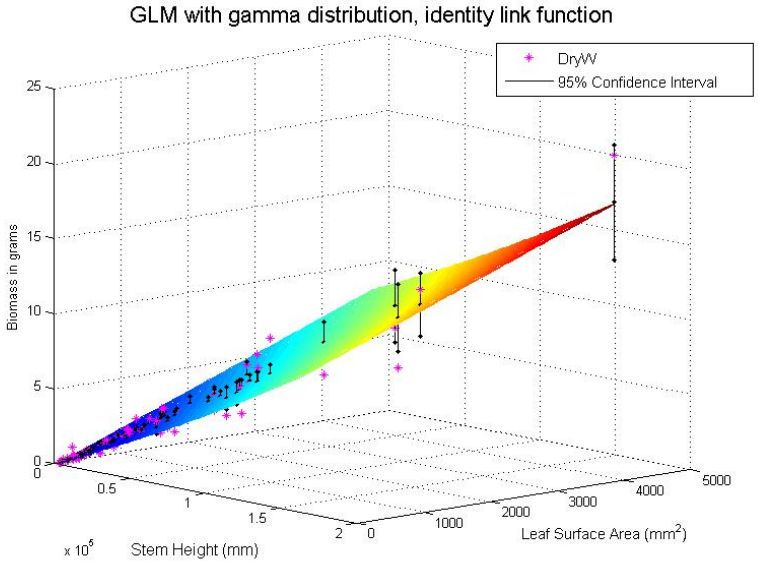

Figure 2. Biomass model with 95% CI intervals and DryW values

Various authors such as Archontoulis and Miguez (2013), Nelder and McCullagh (1989), Fox (2008) state an optimal model for the data of interest would consist of highly significant response variables limited to minimal number of terms to avoid over fitting. We wish to capture the general trend of the data and systematic effects present. Additionally the intention for this model will be to predict biomass estimation for future experimental

data, hence we seek a model which satisfies regression assumptions. For a brief comparison between the two methods presented here we refer to Akaike Information Criterion (AIC) plots, as described in Akaike (1974). In this instance we had a large number of observations relative to the number of parameters in the saturated model. This makes the AIC score a suitable means by which to compare the two methods. The AIC score takes into account the goodness of fit and the number of terms in a model, it is not limited to a specific type of model unlike the coefficient of determination, commonly referred to as R^2, which is limited to OLS models. Instead we may compute the AIC scores to compare two models which employ very different methods for parameter estimates. The OLS model, whose parameters are derived by reducing the sum of the errors squared can be compared against a GLM model which uses maximum likelihood estimates to approximate parameter values.

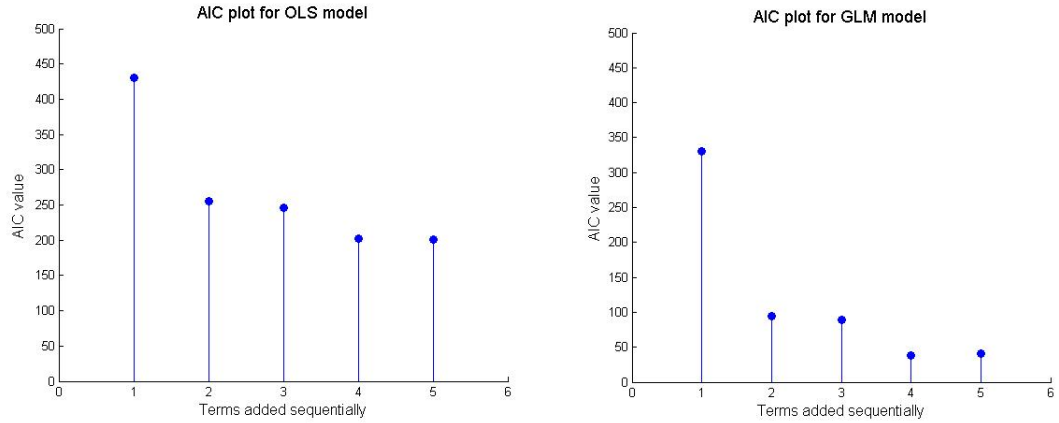

Figure 3. Akaike Information Criterion plots for comparison of OLS and GLM models.

Figure 3 show a large different between the AIC values for the OLS and GLM models. The parameter terms were added sequentially denoted by indices one to five, one being the null model and five being the full saturated model. The AIC score provides a relative estimate of the information lost when a given model is used to represent the data, in comparing models we choose the one which has the lowest AIC score. The best AIC score for the OLS model is 201 which corresponds to the full saturated model, in comparison with the GLM model which yielded an optimal AIC score of 39. The large difference in AIC values reflect not only which is the better model, but it also the effect different methods for parameter estimates have on the same model structure. The comparison of these plots, in addition to the derivation of the GLM_2 model and the statistical analysis discussed in this paper, illustrates GLM_2 to be a superior regression model for a generalised biomass estimate of cereal plants in comparison to the OLS model.

5 CONCLUSION

This paper provides the initial exploration of a novel biomass model for cereal plants which incorporate individual plant features. Current literature explore new methods to deduce plant information from image analysis and the models derived here are an extension from their ideas. By utelising image analysis in promising areas such as EM algorithm for deducing LSA and Hough's transforms to compute stem height, we conclude this paper paves the way for an automated detection system to deduce cereal plant biomass in a non-intrusive way. Once the data is collected from plant images, a robust generalised biomass model is needed to accurately estimate above ground dry plant biomass. This paper explores the process and investigated the mathematical background, required to improve the standard OLS model to the GLM_2 model which has statistically improved biomass estimate.

ACKNOWLEDGEMENT

The first author would like to thank the financial supports from the Grains Research and Development Corporation as well as the Phenomics and Bioinformatics Research Centre. Additionally the first author extends her sincere gratitude to Associate Professor John Van Der Hoek and Dr Chris Brien for statistical advice. Both

first and second authors would like to extend our gratitude to Brendan Fong from Oxford University for the initial idea of segmented biomass model and for Dr Boris Patent from the Plant Accelerator for providing the experimental data.

REFERENCES

Akaike, H. (1974, December). A new look at the statistical model identfication. *IEEE Transactions on Automatic Control 19*(6).

Archontoulis, S. V. and F. E. Miguez (2013, May). Non-linear regression models and applications in agricultural research. *Agronomy Journal*, 1–13.

Arvidsson, S., P. Pérez-Rodriguez, and B. Mueller-Roeber (2011). A growth phenotyping pipeline for Arabidopsis thaliana integrating image analysis and rosette area modeling for robust quantification of genotype effects. *New Phytologist 191*, 895–907.

Berger, B., B. Parent, and M. Tester (2010). High-throughput shoot imaging to study drought responses. *Journal of Experimental Botany*, 1–10.

Cai, J., M. Golzarian, and S. J. Miklavcic (2011, July). Novel image segmentation based on machine learning and its application to plant analysis. *International Journal of Information and Electronics Engineering 1*, 79–84.

Cook, D. (1977, February). Detection of influential observation in linear regression. *Technometrics 18*(19), 15 – 18.

Cremers, D., M. Rousson, and R. Deriche (2007). A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International Journal of Computer Vision 72*(2), 195 – 215.

Duda, R. O. and P. E. Hart (1972, January). Use of Hough transformation to detect lines and curves in pictures. *Communications of the ACM 15*, 11–15.

Ferryman, J., S. Maybank, and A. Worrall (2000). Visual surveillance for moving vehicles. *International Journal of Computer Vision 37*(2), 187 – 197.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models* (2nd ed.). Sage Publication Inc.

Frubank, T. and M. Tester (2011). Phenomics - technologies to relieve the phenotyping bottleneck. *Trends in Plant Science 16*(12), 635 – 644.

Leister, D., C. Varotto, P. Pesaresi, A. Niwergall, and S. F. (1999). Large scale evaluation of plant growth in Arabidopsis thaliana by non-invasive image analysis. *Plant Physiology and Biochemistry 37*(9), 671–678.

Miller, L. (1956). Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association 51*(273), 111–121.

Munns, R., R. James, X. Sirault, R. Furbank, and H. Jones (2010). New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. *Journal of Experimental Botany 61*(13), 3499–3507.

Nelder, J. and P. McCullagh (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

Ning, J., L. Zhang, D. Zhang, and C. Wu (2010). Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition 43*(2), 445 – 456.

Paruelo, J., W. Lauenroth, and P. Roset (2000). Technical note: Estimating aboveground plant biomass using photographic technique. *Journal of range management 53*(2), 190–193.

Tackenberg, O. (2007). A new method for non-destructive measurement of biomass, growth rates, vertical biomass distribution and dry matter content based on digital image analysis. *Annals of Botany 99*, 777–783.

Tester, M. and P. Langridge (2012). Breeding technologies to increase crop production in a changing world. *Science 327*, 818–822.