IMSI Intern Final Report

Stark Ledbetter
University of Washington
sledbett@uw.edu
United States Department of Agriculture
Dr. Justin McGrath, jmcgrath@illinois.edu
August 5, 2021

1. **Project Summary:** The USDA has developed a software called BioCro, which models crop growth based on weather conditions in a given season. The version of the model that we worked with takes 81 parameters. Parameters are constants in equations, like pi in the equation for the area of a circle. Some of these 81 parameters can be obtained from known facts about a given crop (for example, the coldest and hottest temperatures at which the crop can survive). These work like pi. They have been measured by scientists, and we have a good idea of their value. However, there are 11 parameters that are prohibitively difficult to measure directly. Instead, processes they affect are measured, and their values are inferred by fitting them to a model for each crop. To fit these parameters, we consider real data by measuring crops throughout a season (this part was done by others before the project started). Then we run the model with many hypothetical sets of values for the parameters, optimizing according to model fit criteria, until we find the set that gets the model predicted values closest to the measured values. We considered several statistical criteria to measure "closeness," and one of the main goals of this project was to find an effective definition of "closeness" for this type of modeling. Projects like this have been done before with other crop models. The most common way to do this is to fit the model to end-of-year yield (the weight of usable crop that gets harvested). BioCro predicts a lot of values other than end-of-year yield. In fact, it predicts the expected weight and size of various parts of the crop at every hour of the growing season. Therefore, it made sense to collect more measurements than just end-of-year yield, giving a more accurate set of fitted parameters. One big way to tell that the parameters are more accurate is that after fitting, the model still performs well when run with an entirely different (test) data set.

2. **Host company/Lab approval:**
   Permissions for future use of report: The report summary only may be used for reporting to the National Science Foundation and on the IMSI website. IMSI reserves the right to use nontechnical excerpts from the rest of the report in public facing documents without identifying the intern or host. If for any reason IMSI wishes to use excerpts of the other parts of the report with attribution, permission would first need to be obtained the from the author and the author's sponsor organization.

   I have reviewed this report and approve of its release for IMSI internal use, and I approve of the project summary for public release.

   Supervisor Name Supervisor Signature

3. **Introduction and problem statement:** BioCro is a model that predicts crop growth over time given crop-specific parameters and climate data as input. It is a system of C++ modules, each containing one or more differential or non-differential equations, based on models of key physiological and biophysical processes underlying plant growth. There are several different

working versions (comprised of lists of modules to use) of the model. Only the modules selected will affect the outcome of the model.

Goals of the project:
-In the Partitioning Logistic model, fit the 11 parameters not obtained from the literature.
-Identify suitable training and test data sets, given energy sorghum data measured in 2016 through 2019.
-Identify a suitable loss function to minimize using a numeral optimization method, like the Hooke-Jeeves method implemented in the dfoptim package for R.
-Define a statistic based on the loss function used to fit the parameters that measures model quality compared to observed data.
-If possible, find data for crops other than sorghum and fit parameters for those crops as well.

4. **Methods:** The data for energy sorghum was most complete for the 2017 and 2018 seasons. We had started trying to fit the parameters to the observed data in the 2016 season, but there were not enough data points for the fit to be meaningful. We also considered filtering to specific genotypes of energy sorghum, but that either overly reduces the number of data points so that the model fit is not effective, or the parameters we are fitting do not depend much on genotype. So we fit the parameters to the complete set of 2017 data, and used 2018 as a test set, and vice versa.

We looked into a few numerical optimization methods, including the auglag function in the nloptr package for R. Hooke-Jeeves seemed the most effective. We started by running it on a few different basic loss functions: χ^2 and Mean Absolute Percentage Error (MAPE).

$$\chi^2 = \sum_{observations} \frac{(model - observation)^2}{model}$$

$$MAPE = \frac{1}{\#\ of\ observations} \sum_{observations} \frac{|model - observation|}{observation} \cdot 100\%$$

MAPE sometimes fit better to the training sets, but performed inconsistently on the test sets. χ^2 performed more consistently. However, χ^2 is a statistic that is valid when expected values follow a Poisson distribution. The BioCro model does not predict a distribution of crop growth possibilities, but rather returns only the expected value of each measure of the crop for each time point, so we cannot assume anything about the expected distribution of the measured values. In particular, using χ^2 as a loss function seemed overly punishing to days with a lot of observations measured, resulting in high variance. The model would then be skewed towards these days, in order to try to minimize the χ^2 values from these days. In general, it may be valid to give more weight to days with more observations. However, all of these days occur toward the end of the growing season, so the observed values in the beginning of the growing season were essentially ignored. In particular, the resulting fitted parameters would often be impossible, such as the flowering stage starting before the vegetative stage. Another slight issue was that χ^2 tended to result in the model predicting higher values, and MAPE tended to result in the model predicting lower values, due to χ^2 having the model value in the denominator and MAPE having the observation value in the denominator.

Since the rationale for dividing by the model value in χ^2 is that the model value is equal to variance in a Poisson distribution, we decided to try a modified version where we divide by the variance (calculated from observed values on a given day; a different variance was used for each day). This corresponds closely with the Mahalanobis statistic, in the case where all the variables are independent. We also tried to calculate covariance with the observed values to use the full Mahalanobis statistic, but this did not result in a good fit, we believe because using observed values to calculate covariance was not a reliable method. Since there was no reliable method to calculate covariance, we reverted to assuming that all variables are independent.

$$Mahalanobis(day) = \sqrt{\sum_{observations\ that\ day} \frac{\left(observation - model(day)\right)^2}{\sigma^2(day)}}$$

We summed the Mahalanobis values over all days and all observed measurements. There were still some issues with the fit. One was that, like χ^2, Mahalanobis is unbounded, and increases as the number of data points increase. So, like χ^2, the model fit was skewed to days with more data points, which, as above, tended to result in ignoring the early growing season. We tried dividing by number of data points, both before and after taking the square root, and found that the following normed Mahalanobis value was generally independent of both scale of the observation value and number of data points:

$$\text{Normed } Mahalanobis(day) = \frac{1}{\#\ observations\ that\ day} \sqrt{\sum_{observations\ that\ day} \frac{(observation - model(day))^2}{\sigma^2(day)}}$$

The normed Mahalanobis statistic was also used in *Winter 2010*.

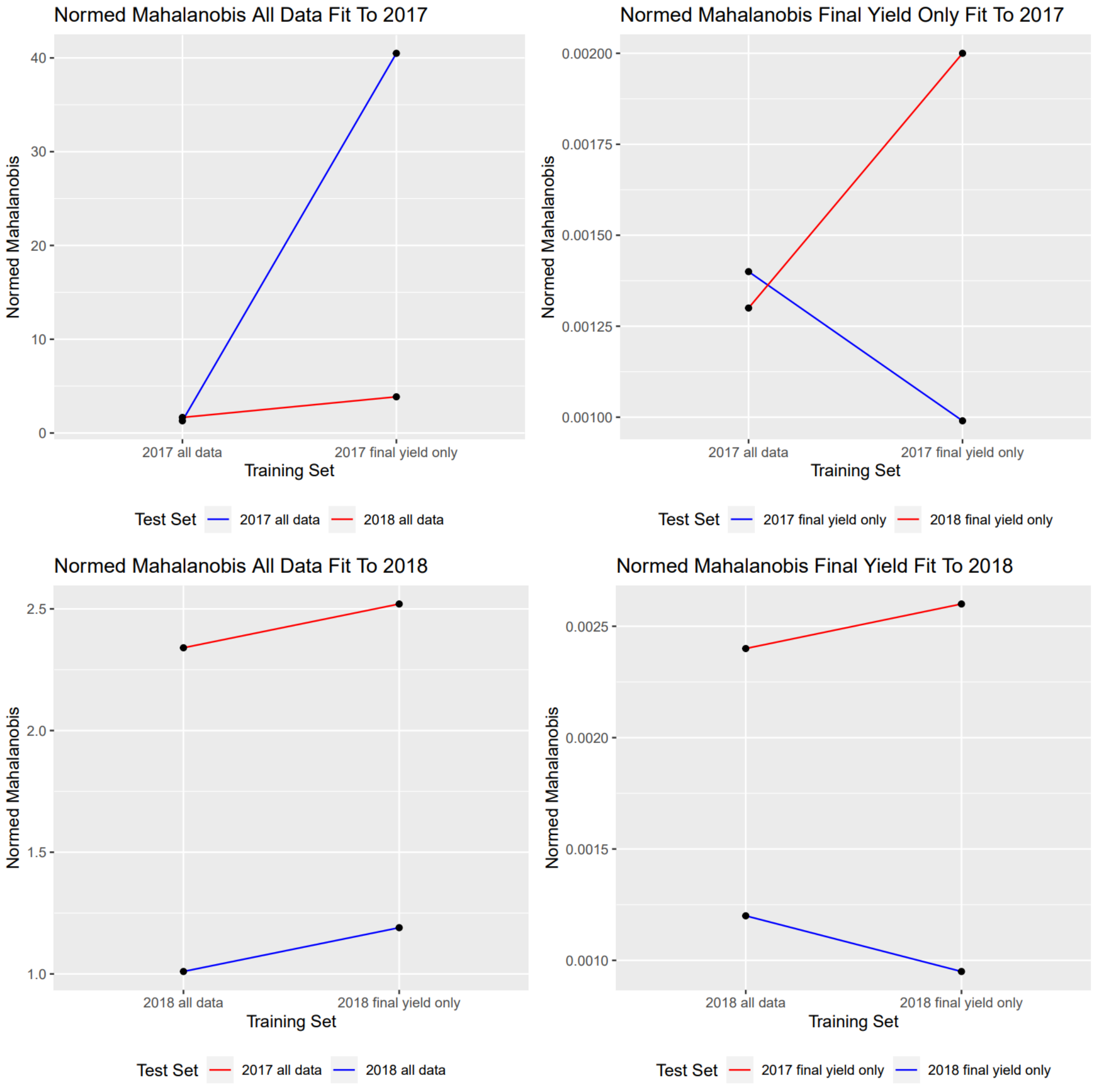
Summing the normed Mahalanobis statistic over all days and all observed measurements would result in 0 if the observations were all exactly equal to the model value, and

$$factor = \sum_{days,measurements} \frac{1}{\sqrt{number\ of\ observations\ that\ day}}$$

if the model value for each day was equal to the mean of the observed values that day. Thus, after summing, dividing by *factor* would result in the statistic being equal to 1 if the model value for each day was equal to the mean of the observed values that day. This was the final version of the loss function we used.

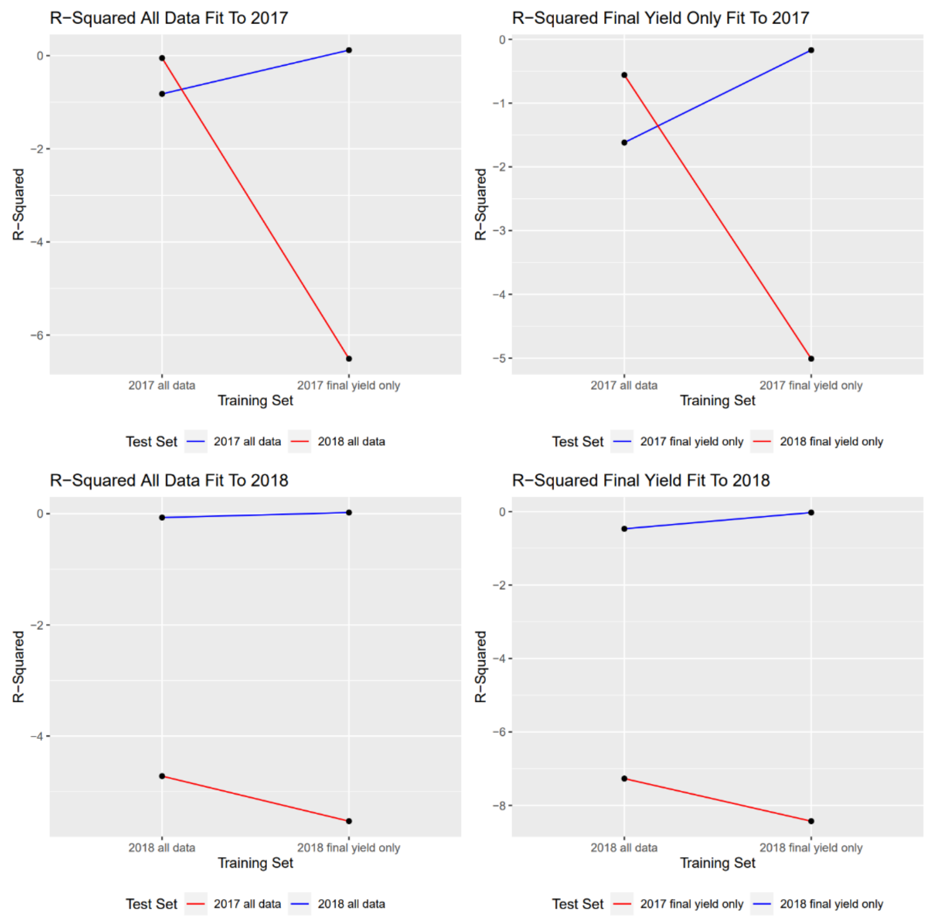
We also found some miscanthus data from *Nunn et. al. 2017*. Since this was just one data point per day (2 days per location), there was no variance to be calculated, so we used χ^2 as the loss function to minimize for fitting parameters for miscanthus. This worked relatively well and was simple.

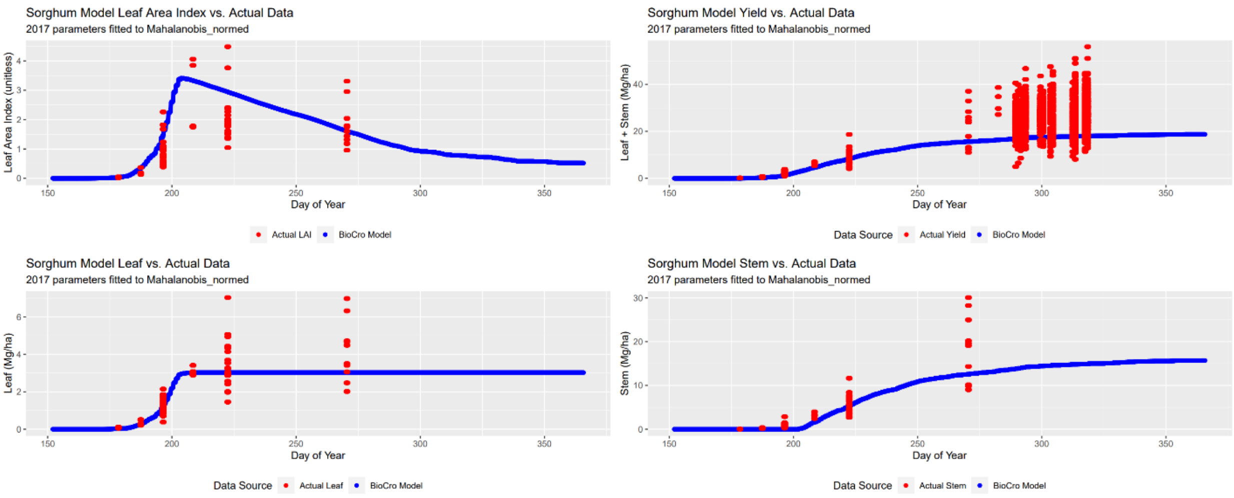
5. **Results:**



The graphs shown are for energy sorghum data, fit to all the data for the training set year, and then fit to only end-of-year yield data. Occasionally, the training set does better when fit to end-of-year yield data, especially if we only consider how well the end-of-year data is fit. However, the test set (in red on all graphs) always does worse when only fit to end-of-year yield data, even when only end-of-year yield data in the test set is considered. Thus, the advantage of BioCro is the modeling of crop growth throughout the course of the season, which can lead to a much more accurate model.

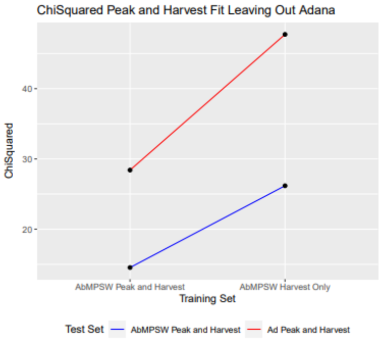
The R-Squared values are often negative, despite a reasonable fit for the model. This is likely due to the combination of a large variance on some measurements on some days, and the fact that the BioCro model is not a linear regression (rather, it is a large system of differential and non-differential equations). Thus, a “null model” that is constant and equal to the mean of the measured values (used as comparison in R-Squared) is not actually attainable, as crops do not

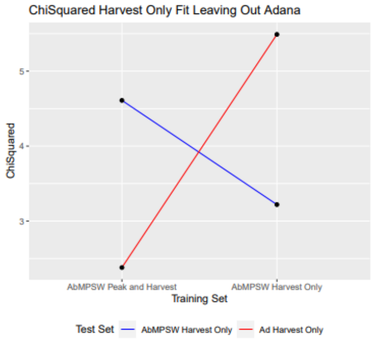


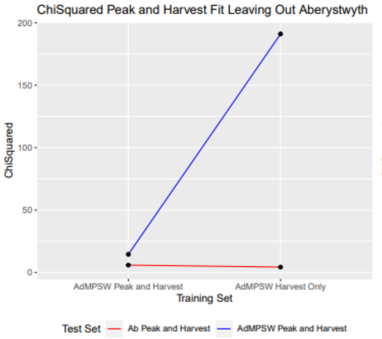


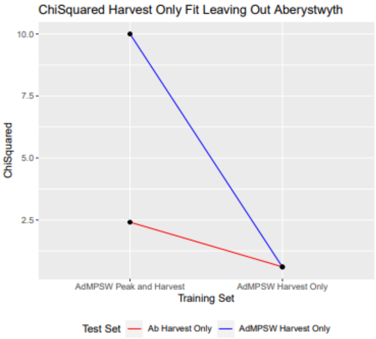
grow this way. As an example, the graphs below show the fitted model for 2017, giving an R-Squared value of -0.82. In general, the trend across all the fitted models was that leaf area index was overestimated and/or Leaf mass was underestimated, so it is possible that the BioCro model should be adjusted to fix this.

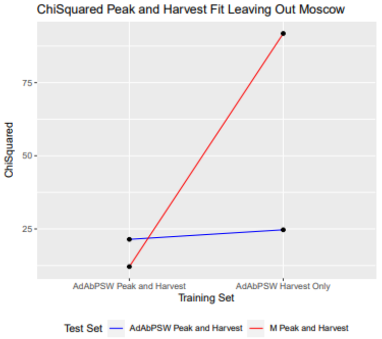
The graphs on the next page show the results for χ^2 for miscanthus giganteus growth in 6 different European cities in 2014. *Nunn et. al. 2017* gave yield measurements for partway through the season (near peak growth time) and at the end of the season. The model was fitted 6 different times, leaving one city's data out of the test set each time. Thus, the training sets consist of all possible combinations of 5 cities, and the test set (again in red) was the sixth city. This fitting was less reliable due to lack of data (only 10 data points in each test set, 2 in each training set, each already an average calculated by *Nunn et. al. 2017*, and half that many when fitting to harvest only), and the need to use χ^2 due to the lack of data. However, χ^2 was generally higher

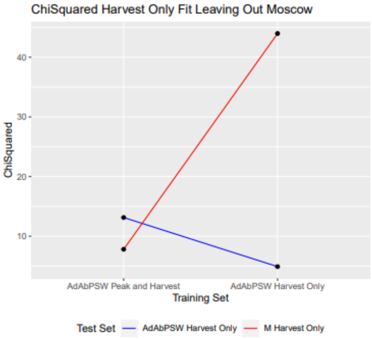


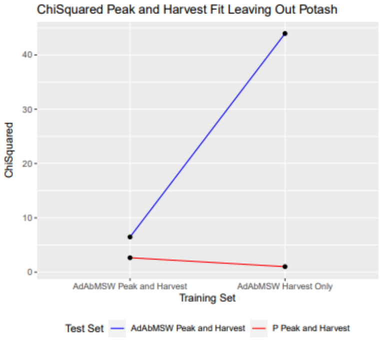


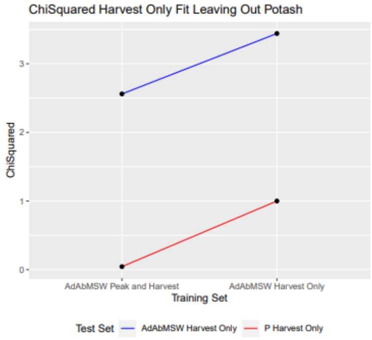


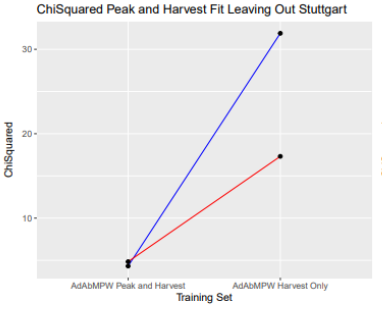


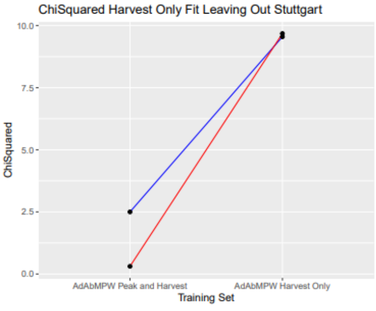












when fit to harvest only data, showing again that parameters fit to more than just end-of-year yield produces a more accurate model.

6. **Conclusions:** The R script used to parametrize the models using the Hooke-Jeeves algorithm and our modified normed Mahalanobis statistic will be useful for all future versions of the BioCro model, and all future data obtained by measuring other crops over a growing season. This should help BioCro accomplish the goal of predicting crop growth in a different or future climate, given its growth in a particular climate, which will be useful as the world faces issues such as population growth and global warming.

Inconsistency in the number of data points for each day, and lack of sufficient data to calculate covariance were some of the major difficulties faced in this project. Biomass data is very time-consuming to obtain from crops, and is usually done sparingly early in the season to preserve the crop for harvest. However, more data will always be useful to result in a more accurate model.

Acknowledgements: Thanks to Surya Eada for statistics expertise, Dr. Justin McGrath for devising an interesting project, Dr. Ed Lohocki for an excellent introduction to BioCro, and Patrick Ryan for building the perfect computer to run these models ad nauseum.

References:

Nunn, C., Hastings, A. F., Kalinina, O., Özgüven, M., Schüle, H., Tarakanov, I. G., Van Der Weijde, T., Anisimov, A. A., Iqbal, Y., Kiesel, A., Khokhlov, N. F., McCalmont,

J. P., Meyer, H., Mos, M., Schwarz, K.-U., Trindade, L. M., Lewandowski, I., & Clifton-Brown, J. C. (2017). Environmental influences on the growing Season duration and ripening of Diverse MISCANTHUS Germplasm grown in six countries. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.00907

Winter, C. L. (2010). Normalized Mahalanobis distance for comparing process-based stochastic models. *Stochastic Environmental Research and Risk Assessment*, *24*(6), 917–923. https://doi.org/10.1007/s00477-010-0386-z