Application of random regression models to model growth curve in Maize using phenotypes derived from multi-spectral images

Mahlet Anche[1], Kelly R Robbins[1], Michael A Gore[1], and Nicolas Morales[1]

[1]Plant Breeding and Genetics

November 22, 2022

Abstract

Vegetation indices (VIs) derived from multi-spectral imaging (MSI) can be used to collect non-destructive phenotypes that could be used to better understand development curves and interactions with environmental factors throughout the growing season. To investigate the amount of variation present in VIs derived from MSI and their relationship with important end-of-season traits, genetic and residual (co)variances for the VIs and their genetic and residual correlations with grain yield and grain moisture were estimated using maize data collected as part of the Genomes to Fields (G2F) initiative. One of the VIs considered in this study was normalized difference vegetation index (NDVI). In addition to NDVI, cumulative NDVI (cNDVI) was used as a phenotype to explore methods to simultaneously fit multiple phenotypes from MSI collected throughout the growing season. The potential of random regression models were investigated using either linear Splines or Legendre polynomial functions. Low to moderately high heritability estimates (0.10 – 0.35) was observed for NDVI values at each of the time points within years, indicating that there exists a reasonable amount of genetic variation. Moreover, strong genetic and residual correlations were found between grain yield and NDVI. Finally, it was found that using random regression with either of the functions converged using all time points and show a potential to be used as an alternative to multi-trait models.

Hosted file

`essoar.10508834.1.docx` available at https://authorea.com/users/540548/articles/600248-application-of-random-regression-models-to-model-growth-curve-in-maize-using-phenotypes-derived-from-multi-spectral-images

Application of random regression models to model growth curve in Maize using phenotypes derived from multi-spectral images

Mahlet T. Anche[1], Nicolas. Morales[1], Michael. A. Gore[1] and Kelly. R. Robbins[1]

[1]Cornell University, Plant Breeding and Genetics Section, School of Integrative Plant Science, Ithaca, NY 14853

ABSTRACT

Vegetation indices (VIs) derived from multi-spectral imaging (MSI) can be used to collect non-destructive phenotypes that could be used to better understand development curves and interactions with environmental factors throughout the growing season. To investigate the amount of variation present in VIs derived from MSI and their relationship with important end-of-season traits, genetic and residual (co)variances for the VIs and their genetic and residual correlations with grain yield and grain moisture were estimated using maize data collected as part of the Genomes to Fields (G2F) initiative. One of the VIs considered in this study was normalized difference vegetation index (NDVI). In addition to NDVI, cumulative NDVI (cNDVI) was used as a phenotype to explore methods to simultaneously fit multiple phenotypes from MSI collected throughout the growing season. The potential of random regression models were investigated using either linear Splines or Legendre polynomial functions. Low to moderately high heritability estimates ($0.10 - 0.35$) was observed for NDVI values at each of the time points within years, indicating that there exists a reasonable amount of genetic variation. Moreover, strong genetic and residual correlations were found between grain yield and NDVI. Finally, it was found that using random regression with either of the functions converged using all time points and show a potential to be used as an alternative to multi-trait models.

Keywords: multispectral images, NDVI, genetic variation, Random regression, Legendre polynomial

1. INTRODUCTION

Advancement in multi-spectral imaging (MSI) and image processing technologies has allowed collection of high-throughput phenotypes repeatedly during the growing season. These phenotypes, when recorded within the growing season of the plant can be used to provide insight about growth and development and plants response to the environmental stress. Normalized difference vegetation index (NDVI), one of the indices that is produced from combination of reflectance from the MSI, has been used to measure biomass and leaf area index (Babar *et al.* 2006). In addition to that, recent studies have shown when heritable variation in NDVI measurements exists, significant gain in genomic prediction accuracy can be obtain for grain yield when NDVI is used as a secondary trait (Anche *et al.* 2020).

Random regression models are common tools to model repeated phenotypes that are recorded in a continuous scale, such as time or age (Kirkpatrick *et al.*

1990; Meyer 2005). With increasing number of time points and with the large number of parameters that must be estimated, random regression models are proven to be very robust (Kirkpatrick *et al.* 1990; Lopes *et al.* 2012; Brito *et al.* 2017). Random regression models commonly use Legendre polynomials to model the variance and covariance of parameters at and among the time points (Meyer 2005). However, it has been noted that such high order polynomials are problematic at extreme values when data are sparse, resulting in poor estimation of variance and covariance (Misztal 2006). As an alternative to such high order polynomials, splines, which are piecewise functions consisting of segments that are connected by so-called knots, have gained popularity for analyzing repeated measurements in random regression models when data may be sparse or highly clustered (Robbins *et al.* 2005).

In this study, we aim to investigate the application of random regression model using linear Spline (RRS) and Legendre polynomials (RRm) to model the growth curve for maize hybrids. For that purpose, NDVI was used as a repeated phenotype. In addition to that, we also aimed to the use of cumulative NDVI that was calculated the NDVI values as a phenotype to model the growth of the maize hybrids.

2. Materials and Methods

Agronomic phenotypic data: The phenotypic data were available from Genomes to Fields (G2F) initiative (http://datacommons.cyverse.org/browse/iplant/home/shared/commons_r
The trials were planted at research sites in Aurora, New York for 2016, 2017 and 2019 ((McFarland *et al.* 2020) using a randomized complete block design with two replicates in each years. For 2016, 2017 and 2019, plot-level phenotypic data was collected for important end-of-season traits for 195, 184 and 131 hybrids, respectively. For the year 2016 and 2017, there were 5 time points and for 2019 there were 6 time points that were extracted. In this study, results using data from 2019 will be reported. Table 1 shows the dates on which the image data was collected, the growing degree days and the growth stage of the plants for the 2019 data.

Table 1. MSI record dates, growing degree days (GDDs) and growth stages for the 2019 data

Record date	GDDs	Growth stages
07-16-2019	874	VT
07-24-2019	1064	R1-R2
07-29-2019	1167	R1-R2
08-05-2019	1306	R3-R4
08-15-2019	1485	R3-R4
09-10-2019	1872	R5-R6

A DJI Matrice 600 equipped with a DRTK-GPS guidance system and a Mi-

casense RedEdge 3 were used for the aerial surveys. Images of the Micasense calibration panel were taken before and after each flight in 2019. Orthomosaics from each aerial survey flight were constructed in Pix4Dmapper (https://www.pix4d.com), which were then used to calculate summary statistics, such as mean, median and variance for individual bands/reflectance for each plot.

Using these summary statistics, NDVI was calculated for each plot using the following equations as shown in Anche et.al. (2020). Among the different summary statistics that were available for the NDVI, mean values for each plot were used in this study. Cumulative NDVI (cNDVI) was used as an alternative phenotype and was calculated by taking the definite integral of NDVI values between two consecutive time points and taking the cumulative of each interval:

$cNDVI = \int_{\mathrm{GDD}_t}^{\mathrm{GDD}_{t+1}} \mathrm{NDVI}$ (1)

where GDD_t is the growing degree days at time t and GDD_{t+1} is growing degree days at $t+1$. (How to calculate growing degree days is presented in Anche *et.al.,* (2020)).

Genotypic data: Genome-by-sequencing (GBS) data were available for 600K Single Nucleotide Markers (SNP) for 1557 parental lines. After quality control for missing data and minor allele frequency, 122K SNP markers were used to calculate the additive genomic relationship matrix between parents, where the additive genomic relationship was calculated by taking half of the parental inbred genomic relationships (VanRaden 2008; Anche *et al.* 2020).

Single trait genomic best linear unbiased prediction (ST-BLUP) model as shown in Anche *et al.* (2020) was used to estimate the genetic and residual variances for plot level grain yield, and NDVIs and cNDVIs at each time points. Genetic and residual correlation between grain yield and NDVI and cNDVI values at different time points were estimated using multi-trait genomic best linear unbiased prediction (MT-GBLUP)

Random Regression Model: Random regression models fit with linear splines (RRS) was used to fit NDVI and cNDVI values at different time points. The RRS is the same as was used in Anche *et al.* (2020). Here RR model with m order Legendre polynomials (RRLm) was also used to fit NDVI and cNDVI phenotypes at different time points. The Legendre polynomial of m order can be denoted as $\varphi_m(w)$, where w is the standardized time to lie between -1 and 1 as follows (Schaeffer 2004);

$w = \frac{2(t_i - t_{\min})}{(t_{\max} - t_{\min})} - 1$ (2)

where $t_i = 1, \ldots n$; and $t_{\min}$ is the earliest time point and $t_{\max}$ is the latest time point. In this study, the growing degree days (GDD) are used as a measure of time points.

The RR model that is used to fit the NDVI and cumulative NDVI (cNDVI) at

time point t is;

$$y_{ijn:t} = f_j + \sum_{k=0}^{m} a_i w_{ijn:t} + \sum_{k=0}^{m} \text{pe}_l w_{ijn:t} + e_{ijn:t} \ , \ (3)$$

where $y_{ijn:t}$ is the nth observation on ith hybrid at time t in the jth fixed factor; f_j is a fixed effect that accounts for the mean growth curve; and in $\sum_{k=0}^{m} a_i w_{ijn:t}$, m is the order of Legendre polynomial, and a_i is the additive genetic effect for hybrid i and $w_{ijn:t}$ is the time covariables related to time t, and in $\sum_{k=0}^{m} \text{pe}_l w_{ijn:t}$, pe_l is the permanent environmental effect for plot l and $e_{ijn:t}$ is a random residual effect.

3. Results and Discussion

Figure 1 shows the heritability estimates for NDVI values at different growth stages and grain yield for the 2019 data. These estimates are obtained using ST-GBLUP model described in Anche *et al.* (2020). Heritability estimates for NDVI values range from 0.11 to 0.35, the lower estimate observed at the last time point and the highest at the first time point.

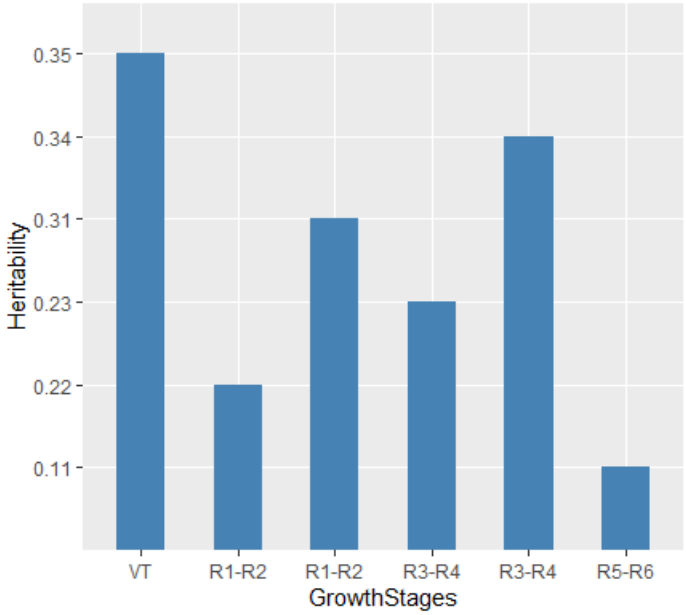

Figure 1. Heritable estimates for NDVIs at different growth stages for the 2019 data.

For the 2019 data, genetic correlation ranging from 0.65-0.97, was found between NDVI values at different growth stages and was found to be very high when compared to the correlation that was found for the 2016-2017 data in Anche *et al.* (2020). Due to high genetic correlation, multi-trait analysis between the first and the fifth time point was not able to converge. A similar convergence

issue was found in a multi-trait analysis between the second and the fifth time point.

The RRS model was fit to model NDVI; however, the model failed to converge. This is also true when cNDVI was used as phenotype instead of NDVI. In addition to fitting RRS, we also fit RR model using 2nd and 3rd order Legendre polynomials. For the 2019 data, RR with 2nd and 3rd order Legendre polynomial converged, when only cNDVI were used as a phenotype. In order to investigate how well the RR model performs, estimates from RR model was correlated with the estimates from ST-GBLUP model.

Figure 2 shows, the correlation between the genetic estimates from the RR model with 2nd order Legendre polynomial and the ST-GBLUP when cNDVI was used a phenotype. As shown in Fig. 2, correlation between the estimates ranges from 0.98 to 0.99 for the 2019 data (Figure 3 (a)). As mentioned above, 3rd order Legendre polynomial was fit to model the cNDVI values, and the model converged well for all the 2019 data. The correlation between the estimates from RR model with 3rd order Legendre polynomial and ST-GBLUP is presented in Figure 3 (b) and it ranges from 0.53 to 0.64. This result indicates that, with fewer number of time point, RR model with 2nd order Legendre polynomial is robust enough to model the data points. In addition to that, the results indicate that, with increasing time point that are recorded, RR model with Legendre polynomials could be used as an alternative to multi-trait models to model repeated phenotypes from MSI.

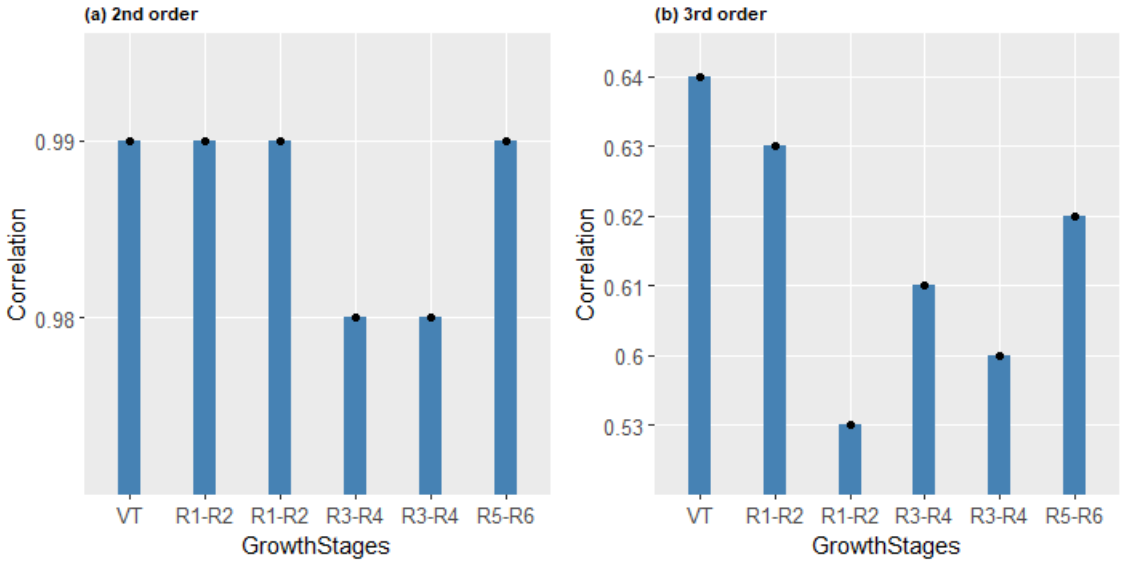

Figure 2. Correlation between genetic estimate from ST-GBLUP and RR model with 2nd order (a) and 3rd order (b) Legendre polynomial

DATA AVAILABILITY STATEMENT

All the phenotypic data is available on CyVerse Data Commons up on request. Phenotypic data from MSI is also available up on request on https://imagebred.org/.

ACKNOWLEDGMENTS

This material is based upon the work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, Hatch 100397 (M.A.G), 1010428 (M.A.G.), 1013637 (M.A.G.), 1013641 (M.A.G.), Iowa corn, and Cornell University startup funds (M.A.G. and K.R.R.). The information, the data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA_E), U. S. Department of Energy, under award number DE-AR0000661. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.

Anche, M. T., N. S. Kaczmar, N. Morales, J. W. Clohessy, D. C. Ilut *et al.*, 2020 Temporal covariance structure of multi-spectral phenotypes and their predictive ability for end-of-season traits in maize. Theor. Appl. Genet.Babar, M. A., M. P. Reynolds, M. Van Ginkel, A. R. Klatt, W. R. Raun *et al.*, 2006 Spectral reflectance to estimate genetic variation for in-season biomass, leaf chlorophyll, and canopy temperature in wheat. Crop Sci.Brito, L. F., F. Gomes da Silva, H. Rojas de Oliveira, N. Souza, G. Caetano *et al.*, 2017 Modelling lactation curves of dairy goats by fitting random regression models using Legendre polynomials or B-splines. Can. J. Anim. Sci.Kirkpatrick, M., D. Lofsvold, and M. Bulmer, 1990 Analysis of the inheritance, selection and evolution of growth trajectories. Genetics.Lopes, F. B., C. U. Magnabosco, F. Paulini, M. C. da Silva, E. S. Miyagi *et al.*, 2012 Analysis of longitudinal data of Nellore cattle from performance test at pasture using random regression model. Springerplus.McFarland, B. A., N. AlKhalifah, M. Bohn, J. Bubert, E. S. Buckler *et al.*, 2020 Maize genomes to fields (G2F): 2014-2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. BMC Res. Notes.Meyer, K., 2005 Random regression analyses using B-splines to model growth of Australian Angus cattle. Genet. Sel. Evol. 37: 473.Misztal, I., 2006 Properties of random

regression models using linear splines. J. Anim. Breed. Genet.Robbins, K. R., I. Misztal, and J. K. Bertrand, 2005 A practical longitudinal model for evaluating growth in Gelbvieh cattle. J. Anim. Sci.Schaeffer, L. R., 2004 Application of random regression models in animal breeding. Livest. Prod. Sci.VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414–4423.