**Using Participatory Testing Networks and Machine Learning Algorithms to Predict Hybrid Reliability Under Untested Organic Environments**

**CPSC 444: Introduction to Spatial Analytics - Final Project Report**

**Student: Christopher Mujjabi**

**INTRODUCTION**

The demand for organic agricultural products has rapidly increased in the last decades due to the increasing consumer awareness of the impacts of conventional production systems on food quality, animal welfare, the environment, and human health (Oroian et al., 2017). The organic agricultural industry has transformed from a mere niche industry into a well-developed alternative to conventional systems and a secure source for non-GMO and chemical-free food with high nutritional value produced using environmentally friendly methods (Annunziata and Pascale, 2009). According to the 2021 Organic Trade Association (OTA) organic industry survey, U.S. organic sales reached a record high of 62 billion dollars in 2020, reflecting a 12 percent growth rate and more than twice the 2019 rate of 5 percent (OTA, 2019). In addition, there is a growing expansion in the certified organic acreage and the number of certified organic farms since more farmers are transitioning from conventional to organic operations. The 2019 organic survey by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) reported a 17 percent increase in certified organic farms from 14200 in 2016 to over 16500 farms. In addition, the total certified organic acreage increased by 9 percent since 2016 to over 5.5 million acres in 2019. The leading U.S. states with the highest number of certified organic farms in 2019 included California, Wisconsin, and New York, with over 3000, 1360, and 1320 farms, respectively (NSAC, 2020).

Despite the perpetual increase in certified organic farms and the steady expansion of the U.S. organic acreage, the current domestic organic food supply does not satisfy the increasing product demand (Brock et al., 2019). Organic agricultural systems rely on natural soil ecosystems as sustainable alternative sources for nutrients and the active mineralization of organic matter and manure instead of synthetic sources used in conventional systems. However, modern crop varieties developed under conventional breeding programs require intensive management and inputs to outcompete weeds, pests, and diseases (Andersen et al., 2015). The Organic Foods Production Act (OFPA), enacted under Title 21 of the 1990 Farm Bill to regulate the organic production standards (SARE, 2003) prohibits the use of chemical inputs such as pesticides and synthetic fertilizers (Osman et al., 2016). Consequently, the productivity of modern maize varieties in organic systems is often lower than their performance applying conventional management practices, creating a yield gap between the two systems (Andersen et al., 2015; de Ponti et al., 2012; Delbridge and King, 2014). Various studies and survey data from the USDA-NASS shows that the average organic maize yield is about 35 percent below an average conventional yield (Kniss et al., 2016; Langemeier and O’Donnell, 2020). The yield gap between conventional and organic systems is due to the lack of crop varieties that are well adapted to the extreme biotic and abiotic stresses like nutrient deficiency, weed pressure and pest competition prevalent in organic systems (Kniss et al., 2016). Over 95 percent of crop varieties commercially available to organic growers were bred under conventional high-input conditions. They, therefore, lack the resilience to produce high yields when grown under complex and diverse organic conditions with low soil nutrients and high weed and pest pressure (Lammerts van Bueren and Myers, 2011; Chozin et al., 2017).

Therefore, there is a need to establish decentralized breeding programs that utilize direct selection in target environments instead of indirect selection for performance in organic fields in favorable homogeneous on-station conditions (Bänziger and Cooper, 2001). Participatory plant breeding (PPB) is one of the used approaches in breeding for marginal heterogeneous environments. The PPB approach is an on-farm breeding method that aims to establish close collaborations between farmers and breeders to develop crop varieties that are well adapted to the farmer's specific regions and agronomic management practices (Zystro et al., 2012). However, the identification and selection of suitable varieties for the on-farm participatory variety testing rely on adequate genetic diversity in the available breeding populations (Al Bari and Horsley, 2014). The expired Plant Variety Protection (exPVP) maize germplasm is a publicly available source of genetically diverse plant materials for public and private research. The germplasm contains a set of elite inbred parents that have been used in the development of numerous commercial hybrids in the U.S. since the 1980s (White et al., 2020). Over 460 exPVP inbred maize lines are currently available for public breeders and researchers to use, and 750 additional lines will be added by 2028 (White et al., 2020). Incorporating the exPVP germplasm into organic breeding programs could be a crucial step in exploring the potential of these inbreds in new hybrid combinations with desirable agronomic performance and quality traits (Bari and Horsley, 2014).

This study aimed to use a participatory research approach to characterize the performance of maize hybrids derived from exPVP germplasm under real-world organic management practices. The specific objectives of this study were to (1) identify the environmental factors and agronomic management practices that were most influential to the performance of the experimental hybrids, (2) estimate the probability of each hybrid to outperform the commercial check under untested organic fields using machine learning methods.

**MATERIALS AND METHODS**

Strip trial locations

A participatory variety testing approach was used to evaluate experimental maize hybrids using a vateriety testing network (VTN) of organic farms in Illinois and Indiana. The Participatory VTN was specifically developed for this project. A total of eight farms in Illinois and four farms in Indiana contributed to the participatory VTN. The on-farm strip trials were conducted with six farms in Illinois and one farm in Indiana in 2018, on three farms in Illinois and three farms in Indiana in 2019, and with four farms in Illinois and three farms in Indiana in 2020. A total of 20 locations were tested in the 3 years of the study.

Germplasm sources

A core set of hybrids from four different breeding programs (PROG) with varying breeding objectives was evaluated in on-farm trials. The Mandaamin Institute, an organic breeding program focused on developing maize varieties with enhanced protein quality and soil-nutrient uptake and efficiency, provided four hybrids (ORG1, ORG2, ORG4, and ORG5). Montgomery Consulting, a conventional breeding program focusing on native insect resistance, native herbicide tolerance, seed quality, and grain chemical composition, provided three hybrids (KEV1, KEV2, and KEV3). Five hybrids (UIUC1, UIUC2, UIUC3, UIUC4, and UIUC7) were derived from exPVP materials and developed at the University of Illinois under the Bohn-Lab.

Experimental design and management

Seven experimental hybrids and one check were planted in each location using a completely randomized design (CRD) with no replications. Each farm location was considered an individual replication for the corresponding year of testing. The hybrids were planted in four-row 30.5 meters long strip plots. The planting density and row spacing varied from farm to farm and ranged from 69,000 plants ha−1 to 87,500 plants ha−1 and 0.76 m to 1 m, respectively. The seed for all the UIUC experimental hybrids was produced in our organic nursery at the University of Illinois and was not treated before planting. Although management practices such as weed control and manure sources varied from farm to farm, all farms used methods and inputs that followed the organic standards and regulations established by the USDA's National Organic Program (NOP) under the OFPA Act.

Phenotypic data collection

The hybrids were evaluated for grain yield (YLD, t ha-1), plant height (PHT, cm), ear height (EHT, cm), test weight (TWT, kg/hL), as well as protein (PROT, %), oil (OIL, %), and starch (STR, %) content in the grain. The trait data were collected during the post-flowering farm visits. PHT and EHT were recorded as the mean of five average plants and measured following the Genome-to-Field standard operating procedure (G2F Initiative, 2018). At the R6 stage, two subplots of 0.0004 ha (1000th of an acre) size within each strip were staked out, and all ears within each subplot were manually harvested. Subsequently, the ears were dried and shelled using an Agriculex SCS-2 Single Corn Sheller. Grain yield was calculated in metric tones per hectare at 15.5% grain moisture. In addition, farm history and agronomic management practices such as cover crop type used before the trial (CCT), rotation system of the field (RTN) , manure source used (NSRC), planting density (PDTY) and weed pressure rate (WRT) for each farm were collected.

Weather Data

Weather information for each participating farm each year was downloaded using the R package daymetr. Weather data for the months May to September was used to calculate total precipitation (prcp), total solar radiation (srad), mean temperature (temp), minimum temperature (tmin) and maximum temperature (tmax) for the growing season in each location.

Soil Data

Soil samples were taken (by Binod Ghimire, NRES) from each location at the beginning and during the growing season. These samples were analyzed and used to estimate different soil characteristics and elements such as cation exchange capacity (CEC), soil pH, soil organic matter (OM), Sulfur (S), phosphorus (p), Calcium (Ca), Magnesium (Mg), Potasium (K), Sodium (Na), available nitrogen (N03N), total organic carbon (TOC), total organic nitrogen (TON) and soil texture (percentage silt, sand and clay).

**STATISTICAL ANALYSIS**

The statistical analysis was conducted using R software version 4.0.5 (R Core Team, 2021) in the RStudio environment (RStudio Team, 2016). The parameter “reliability” (R) was calculated as the probability of a given experimental hybrid outperforming the commercial check variety. Traditionally, reliability is estimated using a using a parametric approach (R) below, described by Eskridge and Mumm (1992) assuming hybrid performance can be modeled using a normal distribution.

R = P (Z > - /)

where Z is a standard normal variable, and is the yield difference between the ith experimental hybrid and the commercial check on a given farm.

However, for this project, reliability for each experimental hybrid was treated as a “YES” or “NO” binary response when the hybrid outperfomed or failed to outperfom the check variety, respoectively. Different machine learning methods such as classification and regression tree CART, random forest and support vector machines (SVM) were used to predict the reliability of each hybrid given the management, soil and weather data of a particular farm. The accuracy of each method was estimated as the ratio of total true positives and true negatives out of the total observations. The best model with the highest accuracy will be selected as the best prediction model.

1. Using CART

A logistic regression model was conducted using the classification and regression tree (CART) analysis to classify the experimental hybrids into the 2 relibability responses. The variables used to fit the generalized linear model included the name of the hybrid, breeding program it was developed, the management practice, the soil conditions and the weather condition of a given location, as shown below.

Reliability ~ Hybrid + PROG + CCT + RTN + WRT + PDTY + NSRC + CEC + pH + OM + S + P + Ca + Mg + K + Na + NO3N + TOC + TON + Sand + Silt + Clay + prcp + srad + tmax + tmin + temp

The library cartool in R was used to split the data into 80% training and 20% testing sets. The libraries rpart and rpart.plot were used to contruct the decision tree for the reliability of the hybrids under different farms as presented in Fig1. The area under the curve was plotted and estimated using the library ROCR as presented in Fig. 2

1. Using Random Forest

The caret package was used to fit different random forest models with varying resampling methods. the models were fit using the similar variables used in the CART analysis and the same training and testing sets. The resampling methods used incuded bootstrapping, k-fold cross-validation, bootstrapping 0.632, leave-group-out cross validation (LGOCV), repeatedcv, optimism bootstrapping, adaptive cross validation and adaptive bootstrapping. The resampling for all models was done 10 fold. In addtion, a similar model was fit using support vector machines (SVM) to compare the accuracy between random forest and SVM. The accuracy of both models was plotted on a boxplot and presented in Fig 4.

**RESULTS AND DISCUSSION**

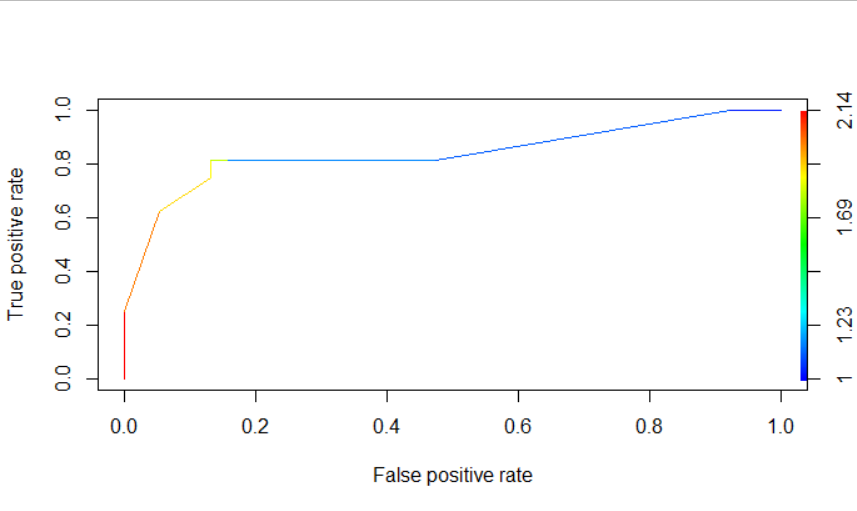
The objective of the study was to determine which environmental and management variables contributed to the performance and the reliability of the experimental hybrids. The most important variables used for in the random forest model were total organic nitrogen (TOC), phosphorus (P), temperature, percentage sand, sodium (Na), organic matter, maximum temperature, hybrid, cover crop and solar radiation. The second objective was to develop a prediction model that can accurately estimate the probability of an experimental hybrid to outperform the check at a given farm. The CART analysis resulted in a prediction accuracy of 0.7 with an area under the curve of 0.84 as shown in Fig 2.

Graphical user interface

Description automatically generated

Fig 1. Classification Tree for hybrid reliability using CART Analysis

The classification tree generated had multiple branches with different soil factors (P, OM, sand), management practices(manure source and cover crop type) and weather variables (precipitation) as shown in Fig 1. Experimental hybrids outperformed the check when the phosphorus was above 287 ppm. Under farms with no manure application, hybrids KEV2, KEV3, ORG1, ORG2, UIUC1, UIUC2, UIUC3 and UIUC4 outperformed the check while the rest did not. In addition, experimental hybrids outperformed the check in farms with organic matter above 44.5. Hybrids grown in farms that used grass cover crops, mixed cover crops or no cover crop did not outperform the check.



**AUC = 0.84**

Fig 2. Plot for area under the curve for the CART anlysis

The prediction accuracy increased when the random forest method was used. However, the different cross validation and bootstrapping resampling methods obtained different accuracies as presented in Table 1. The k-fold cross validation obtained the lowest accuracy of 0.87 while bootstrapping and bootstrapping0.632 obtained the highest accuracy of 0.93 with an interval of 0.82 – 0.97.

Table 1. Accuracies for the different resampling cross-validation and bootstrapping methods

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Accuracy | Kappa | 95% CI |
| CV | 0.8704 | 0.6947 | 0.751 - 0.9463 |
| Adaptive CV | 0.9074 | 0.7894 | 0.797 - 0.9692 |
| Repeated CV | 0.9074 | 0.7739 | 0.797 - 0.9692 |
| LGOCV | 0.9074 | 0.7739 | 0.797 - 0.9692 |
| Bootstrapping | 0.9259 | 0.8224 | 0.821 - 0.9794 |
| Bootstrapping0.632 | 0.9259 | 0.8157 | 0.821 - 0.9794 |
| Optimism Boot | 0.9074 | 0.7739 | 0.797 - 0.9692 |
| Boot\_All | 0.8704 | 0.6947 | 0.751 - 0.9463 |
| Adaptive Boot | 0.9074 | 0.7739 | 0.797 - 0.9692 |

Chart, box and whisker chart

Description automatically generatedComparing random forest method with SVM, random forest showed the highest mean accuracy of 0.92 while SVM obtained a mean accuracy of 0.88 as shown in Fig. 3 below.

Fig 3. Prediction accuracies comparison between random forest and SVM models

Therefore, the random forest model with either the bootstrapping0.632 or normal bootstrapping was chosen as the best prediction model to estimate the reliabiltiy of an experimental hybrid.

**CONCLUSIONS**

The study showed that different agronomic management practices such as manure sources used, and cover crop used significantly influenced the performance of the experimental hybrids. in addition, soil health variables such as phosphorus, sodium, organic matter, total organic nitrogen, and soil texture also influenced hybrid reliability. Weather variables such as average temperature and total precipitation during the growing season also impacted the reliability of the hybrids. Machine learning models such as CART, random forest and SVM successfully predicted the probability of an experimental hybrid to outperform the commercial check. In general, random forest models using bootstrapping resampling method showed the highest prediction accuracy of 0.92. These algorithms can be used to predict how experimental hybrids in the Bohn Lab breeding program may perform in different environments without testing them. Farmers with access to soil testing facilities and weather data can use this algorithm to select hybrids that will perform well under their distinctive management practices, hence avoid surprises at the end of the season. However, the model should be tested using a different dataset to eliminate possibilities of overfitting.

**REFERENCES**

Andersen, M.M., X. Landes, W. Xiang, A. Anyshchenko, J. Falhof, et al. 2015. Feasibility of new breeding techniques for organic farming. Trends Plant Sci. 20(7): 426–434. doi: 10.1016/j.tplants.2015.04.011.

Annunziata, A., and P. Pascale. 2009. The 113th EAAE Seminar “A resilient European food industry and food chain in a challenging world. Consumers’ behaviours and attitudes toward healthy food products: The case of Organic and Functional foods. Crete

Bänziger, M., and M. Cooper. 2001. Breeding for low input conditions and consequences for participatory plant breeding: Examples from tropical maize and wheat.

Al Bari, M.A., and R. Horsley. 2014. Usefulness of Expired Proprietary (Ex-PVP) Maize (Zea mays l.) Germplasm for U.S. Northern Breeding Programs.

Brock, C., D. Jackson-Smith, S. Kumarappan, and J.-S. Douglas. 2019. A Profile of Organic Corn Production in the Midwest and Northeast PART I: Farm Characteristics, Production Practices, and Soil Management Strategies https://ayoqq.org.

Chozin, M., S. Sudjatmiko, N. Setyowati, F. Fahrurrozi, and Z. Muktamar. 2017. Analysis of traits association in sweet corn inbred lines as grown under organic crop management. Sabrao J. Breed. Genet.

Delbridge, T.A., and R.P. King. 2014. The Conventional-Organic Crop Yield Gap: Evidence from Farm-Level Data. Minneapolis.

Kniss, A.R., S.D. Savage, and R. Jabbour. 2016. Commercial Crop Yields Reveal Strengths and Weaknesses for Organic Agriculture in the United States. PLoS One 11(8). doi: 10.1371/journal.pone.0161673.

Langemeier, M., and M. O’Donnell. 2020. Conventional and Organic Enterprise Net Returns. Lafayette, Indiana.

NSAC. 2020. 2019 NASS SURVEY FINDS BIG GAINS FOR ORGANIC. Natl. Sustain. Agric. Coalit. https://sustainableagriculture.net/blog/2019-organic-survey-finds-big-gains-for-organic/ (accessed 1 July 2021).

Oroian, C., C. Safirescu, R. Harun, G. Chiciudean, F. Arion, et al. 2017. Consumers’ Attitudes towards Organic Products and Sustainable Development: A Case Study of Romania. Sustainability 9(9). doi: 10.3390/su9091559.

Osman, A.M., C.J.M. Almekinders, P.C. Struik, and E.T. Lammerts van Bueren. 2016. Adapting spring wheat breeding to the needs of the organic sector. NJAS - Wageningen J. Life Sci. doi: 10.1016/j.njas.2015.11.004.

OTA. 2019. U.S. organic sales break through $50 billion mark in 2018. Washington DC.

de Ponti, T., B. Rijk, and M.K. van Ittersum. 2012. The crop yield gap between organic and conventional agriculture. Agric. Syst. 108. doi: 10.1016/j.agsy.2011.12.004.

SARE. 2003. History of Organic Farming in the United States. Transitioning to Org. Prod.: 3–4.

White, M.R., M.A. Mikel, N. Leon, and S.M. Kaeppler. 2020. Diversity and heterotic patterns in North American proprietary dent maize germplasm. Crop Sci. 60(1): 100–114. doi: 10.1002/csc2.20050.

Zystro, J., A. Shelton, and S. Snapp. 2012. Participatory Plant Breeding Toolkit. Org. Seed Alliance. http://seedalliance.org/wp-content/uploads/2017/04/PPB\_Toolkit\_FINAL.pdf (accessed 12 July 2021).