New machine learning model offers simple solution to predicting crop yield

September 3 2024, by John Lovett

Sam Fernandes, left, assistant professor of agricultural statistics and quantitative genetics, and Igor Fernandes, statistics and analytics master's student, have worked to improve crop yield prediction using environmental and genetic data. Credit: U of A System Division of Agriculture / Paden Johnson

A new machine-learning model for predicting crop yield using

environmental data and genetic information can be used to develop new, higher-performing crop varieties.

Igor Fernandes, a statistics and analytics master's student at the University of Arkansas, entered agriculture studies with a data science background and some exposure to agronomy as an undergraduate assistant for Embrapa, the Brazilian Agricultural Research Corporation. With an outsider's perspective and a history of working with environmental data through one of his former advisers, he developed a novel approach to forecasting how crop varieties will perform in the field.

His interest in the subject led to a recently published study co-authored with his adviser, Sam Fernandes, an assistant professor of agricultural statistics and quantitative genetics with the Arkansas Agricultural Experiment Station, the research arm of the University of Arkansas System Division of Agriculture.

The study, [published](#) in the *Theoretical and Applied Genetics* journal, is titled "Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials."

"Igor came in from statistics with no genetics background," Sam Fernandes said. "So, he had this idea that was not at all what we would use in genetics, and it was just surprising that it worked well."

Igor Fernandes's model, which focused on environmental data, led him to a close second in this year's international Genome to Fields competition. Co-authors of the study that stemmed from the competition entry included Caio Vieira, an assistant professor of soybean breeding for the experiment station, and Kaio Dias, assistant professor in the department of general biology at the Federal University of Viçosa in Brazil.

Environment and genetics

While the competition entry showed environmental data alone worked better than expected at predicting crop yield, the researchers saw an opportunity to build a comprehensive study that compared the novel approach to established prediction models used in genomic breeding.

Genomic breeding, a process of screening thousands of candidates for field trials based on DNA alone, can save time and resources needed to develop a new plant variety, such as growing better in drought conditions. An important part of genomic breeding involves genomic prediction to estimate a plant's yield using its DNA.

"Let's say you have thousands of candidates, and you get the DNA from all of them," Sam Fernandes explains. "Based on the DNA along with information from previous field trials, you are able to tell which one will be the highest yielding without planting it in the field. So, you're saving resources that way. This is genomic prediction."

Adding information to a model on how that plant would interact with environmental conditions increases the accuracy of the genomic prediction and is becoming more common as more environmental data from testing centers becomes available. The practice is called "enviromics." Still, there is no consensus on the best machine learning approach to combine environmental and genetic data.

"One advantage of including the environmental information in the models is that you can address what we call genotype-by-environmental interaction," Sam Fernandes said. "Since the environment does not affect all of the individuals in the same way, we try to account for all of that, so we are able to select the best individual. And the best individual can be different depending on the place and season."

The study used the same data on corn plots from the Genomes to Fields Initiative that were used in the competition, but the researchers adjusted inputs as genetic, environmental, or a combination of both in "additive" and "multiplicative" manners. When including environmental and genetic data in a more straightforward "additive" manner, the prediction accuracy was better than the more complicated "multiplicative" manner.

The simpler model took less time for the computer to process, and the mean prediction accuracy improved 7% over the established model. The experiment was validated in three scenarios typically encountered in plant breeding.

"One of the unique things that Igor did is how he processed the environmental data," Sam Fernandes said. "There are fancier models [in which] people can throw in all sorts of information. But what Igor did is a simple, yet efficient way of combining the genetic and environmental data using feature engineering to process the information and get a summary of variables that is more informative."

Collectively, the researchers say the results are promising, especially with the increasing interest in combining environmental features and genetic data for prediction purposes. Their immediate goal is to apply it to increase the capability of screening genotypes for field trials.

More information: Igor K. Fernandes et al, Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials, *Theoretical and Applied Genetics* (2024). DOI: 10.1007/s00122-024-04687-w

Provided by University of Arkansas