Leveraging machine learning models and multi-environmental soybean trial data for environmental characterization and genomic prediction in Argentina.

Carlos Hernandez

Research question: Is it possible predict soybean genotypes across different environments in Argentina?

In many situations, and for various reasons, it is important to know the behavior of a genotype in an environment where it was not tested. On the other hand, it is also important to evaluate which environmental variables can position some genotypes better in some environments or in others.

a. Data description

Climate data can be taken from the TERRACLIMATE dataset, while soil data will be provided by the SoilGrid dataset. The crop dataset consists of soybean variety trials conducted in the central region of Argentina by the National Institute of Agriculture Technology (INTA). The crop data will include the trial site, season, yield, variety, maturity group, planting, emergency and harvest date, and in some of them the phenology for R1, R5, R7 and R8.

1. Methods

b. Envirotyping

Using climate and soil data, different soil and climate zones will be delimited and characterized in the region under study. The process will consist of using the first principal components of spatial component analysis, which will be used as input to a cluster model to delineate different clusters. An optimal number of clusters will be carried out to obtain the best number of cluster. In addition, together with crop data, the most important variables that describe the variability between genotypes will be evaluated.

c. Genomic prediction

Using climate, trial and soil data as input for a machine learning model to be defined, the genotype performance at a given site will be estimated. To evaluate the accuracy of the model, the dataset will be divided into two groups, one for training and the other for model validation. The metrics performance can be R2, RMSE, KGE and NGE.