Variable selection strategies applied to identify climatic variables impacting the detection of crop diseases

Suarez F [1], *Fiore J* [2], *Balzarini M* [3], *Gimenez-Pecci MP* [4], *Bruno C* [5]

[1] suarezfranco@agro.unc.edu.ar , Unidad de Fitopatología y Modelización Agrícola (UFyMA) jua– INTA – CONICET

[2] juanmfiore@mi.unc.edu.ar , UFyMA – INTA – CONICET

[3] gimenez.mariadelapaz@inta.gob.ar , Unidad de Fitopatología y Modelización Agrícola (UFyMA). INTA

[4] monica.balzarini@unc.edu.ar , Unidad de Fitopatología y Modelización Agrícola (UFyMA), Estadística y Biometría, Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias. Argentina

[5] cebruno@agro.unc.edu.ar , Unidad de Fitopatología y Modelización Agrícola (UFyMA), Estadística y Biometría, Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias. Argentina

The epidemiological development of infectious diseases results from the interaction of at least three main factors: a conductive environment (weather conditions), a susceptible host, and a virulent pathogen. Currently, it is possible to easily access large volumes of georeferenced climate data, which is why new models have been generated based on the relationship between environmental conditions and disease in a pathosystem. When it is necessary to model the relationship of the disease with the climate, the challenge of working with multiple climatic parameters appears, generally correlated and redundant. Variable selection is the process of selecting a subset of relevant variables to build robust models. The objective of this work was to compare the performance of three variable selection methods: Stepwise, Boruta (B), and Least Absolute Shrinkage and Selection Operator regression (LASSO) in modelling, context to predict disease risks in crops from high-dimensional climatic variables. Data from three pathosystems, with different *n* records and *p* climatic variables, were processed: one related to Aspergillus flavus, another to Corn stunt spiroplasma (CSS), and the third to Maize dwarf mosaic virus (MDMV), with georeferenced records from the monitoring of these diseases in maize (*Zea mays L.*) crops in Argentina. Each database had the presence/absence value of the pathogen and climatic variables such as relative humidity, temperature, accumulated rainfall, and wind speed, were downloaded, covering the period before sowing until harvest and summarized weekly or monthly. The databases were partitioned into 2 data subsets, one for training (80% of the data) and the other for validation (20% of the data). The selection of variables and the adjustment of the models were carried out with the training base and with repeated cross-validation of k=10 and 5 repetitions. The validation of each model obtained was carried out with the validation base and the selection methods were compared using the values of precision and area under the Receiver Operating Characteristics (ROC) curve obtained. The generated models presented good metrics and inspire the construction of alarm systems for these diseases, based on climatic variables. LASSO produces an intermediate parameterization between the selection methods Stepwise (minimum number of predictors) and B (maximum number of predictors) and achieved greater predictive capacity in the classification models to discern between the presence and absence of the pathogen of the corn. In the case of A. flavus LASSO present a precision of 84.09%, CSS 77.52%, and MDMV 70.69%.

Keywords: LASSO regression, Stepwise, Boruta.