Phytopathogenic Mollicutes
Vol. 15 (1), June 2025, 111-112

doi: 10.5958/2249-4677.2025.00059.X

Epidemiology and control

Evaluating the threat of phytoplasma disease emergence in agroecosystems and natural habitats

Valeria Trivellone, Roxana Jafari Haddadian and Christopher H. Dietrich

Illinois Natural History Survey, Prairie Research Institute, University of Illinois at Urbana Champaign, Champaign, Illinois, United States of America

Abstract

Outbreaks of phytoplasma diseases annually cause billions of dollars in crop losses worldwide. A few efforts have been made to predict disease outbreaks and management continues to focus primarily on reducing pathogen spread following an outbreak. This study leverages machine learning to assess the global risk of emerging phytoplasma diseases using data from the literature on previous phytoplasma outbreaks in agroecosystems, combined with newly documented occurrences of phytoplasma-positive insects (potential vectors) in natural areas worldwide. By applying supervised machine learning on these datasets, key predictors of vector-host-phytoplasma interactions were identified and their importance in facilitating disease outbreaks was evaluated. The model highlights critical differences between two types of ecosystems and establishes a foundation for predicting new phytoplasma-host associations. These findings pave the way for targeted interventions to mitigate the risk of future outbreaks.

Keywords: vector-phytoplasma associations, machine learning, pathogen biodiversity, emerging plant diseases

Introduction

Outbreaks of plant diseases affecting agriculture have, until very recently, been considered rare and unpredictable. Thus, disease management continues to focus on reducing pathogen spread following an outbreak. Proactive strategies, such as landscape-level prediction of disease emergence incorporating data on natural vegetation and its interfaces with crops have rarely been attempted; however they should be feasible given sufficient information on vegetation cover, climate, and previously documented associations between pathogens and their hosts (including insect vectors). Here it was outlined a new approach for predictive modeling of disease emergence in phytoplasmas, one of the most widespread and important groups of vector-borne plant pathogens, which annually cause billions of dollars in yield losses to agriculture worldwide (Kumari *et al*., 2019). This approach leverages available landscape, socio-economic and climatic data, combined with data on associations between phytoplasmas, their insect vectors and plant hosts from agroecosystems and natural habitats to identify areas at risk of disease outbreaks. By using two databases created by the team at the University of Illinois, this study provides preliminary results of a modeling approach to predict emerging phytoplasma diseases.

Materials and Methods

Relevant data (coordinates, country, crop type, infected insects, rate of infection, phytoplasma strain, year) were compiled from 838 publications, yielding a database of 317 unique occurrences of phytoplasma-insect associations. About 42% specimens tested (135) were phytoplasma positive.

About 2,000 insect specimens representing distinct species and collected in various natural areas worldwide (722 sites in 50 countries) were analyzed using qPCR to quantify phytoplasma cells. Phytoplasma-positive samples were further characterized using a multi-locus approach (Trivellone *et al*., 2023). About 13% specimens (265) were positive.

Environmental variables known or hypothesized to affect phytoplasma outbreaks, such as temperature, precipitation, population density, altitude, cropland proportion, land cover type, and environmental performance index were selected. Bioclimatic (average temperature -˚C and precipitation -mm) and elevation (m above sea level) data‘ at 30 arc seconds (approximately 1 km^2) resolution were downloaded from Worldclim 2.1 (Fick and Hijmans, 2017). Global raster data were downloaded from the Socioeconomic Data and Applications Center (SEDAC) website. There were considered 6 predictors: UN WPP-Adjusted Population Density, 2000 and 2020 (CIESIN, 2018),

Corresponding author e-mail: Valeria Trivellone (valeria3@illinois.edu) human footprint (Venter *et al*., 2016), Terrestrial Biome Protection (global weights) /Biodiversity/Ecosystem Vitality (Wolf *et al*., 2022), Food Insecurity Hotspots Data Set, v1 (2009 – 2019) (CIESIN, 2020), Global Agricultural Lands in 2000 (Ramankutty *et al*., 2012), Landcover 2000 and 2050 (Millennium Ecosystem Assessment, 2005).

Data were analyzed using five supervised machine learning algorithms (Gradient Boosting Machine, Random Forest, Neural Network, General Linear Model, Recursive Partitioning) to build a final model that calculates the probability of infection of a potential insect vector (phytoplasma positive/negative) given the values of the most important selected variables. The models were trained using 20% of the data and the remainder of the data was used to test model accuracy. Classification accuracy and result reliability was evaluated using the Kappa statistic (comparison of observed to expected accuracy), with values ranging from -1 to 1 (<0 indicated that the classification is no better than a random classification). The models were fitted using the R-package caret (Kuhn, 2008).

Results

The most important variables driving positivity rate in known and potential insect vectors were precipitation, temperature, population density, proportion of cropland (for both habitat types), altitude (for agroecosystems) and environmental performance index (for natural habitats) (Figure 1). The top five predictors were used to train each classification model. The class imbalance ratio (proportion of negative samples) was calculated for agroecosystems and natural habitat datasets (0.75 and 6.5, respectively) and a correction was applied for the highly imbalanced natural habitat dataset. For both datasets, Random Forest performed better than the other algorithms (k= 0.30, for agroecosystems and k= 0.51, for natural habitats). For agroecosystems, sites in lowlands with higher population density were associated with higher probability of infections. For natural habitats, higher infection probability was associated with dry conditions and lower proportion of crop lands.

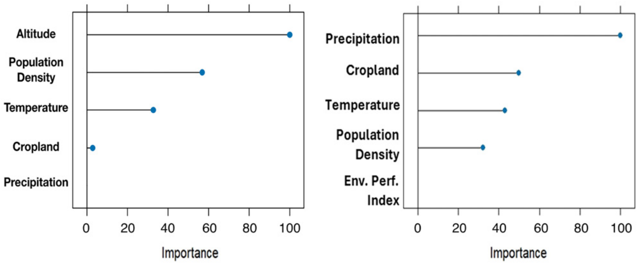

Figure 1. Top 10 variables by importance estimated using random forest to predict phytoplasma-infected insects in agroecosystems (left) and natural habitats (right).

Discussion

Data compiled from literature on phytoplasma outbreaks in agroecosystems were used combined with data from intensive screening of insect specimens collected from natural areas worldwide, to create a classification model predicting the probability of phytoplasma-positive insect vectors in different habitat types based on climatic, socio-economic and landscape characteristics. Although the models performed accurately on the test data, and data on recorded positive samples in natural habitats may be adequate, it was only possible to include 317 records from 47 sites in agroecosystems. Addition of data from agroecosystems, particularly in under sampled areas of the tropics, may increase the predictive power of the model. Next steps include adding more records from tropical regions, and from plant infections, using infection rate as the continuous variable to estimate the severity of outbreaks. Another step will be to evaluate a dataset explicitly incorporating interfaces between natural areas and agroecosystems to estimate the risk of spillover.

Acknowledgements

This work was supported by the U.S. National Science Foundation Grant DEB-2244871. The authors thank all the who contributed to setting up and testing various steps of the workflow for the project, and the Office of Undergraduate Research Apprenticeship Program at University of Illinois for granting to RJH the opportunity to complete part of this research.

References

CIESIN Center for International Earth Science Information Network - 2018. https://doi.org/10.7927/H49C6VHW [accessed 14.02. 2024].

CIESIN Center for International Earth Science Information Network - 2020. https://doi.org/10.7927/cx02-2587 [accessed 14.02. 2024].

Fick SE and Hijmans RJ 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37: 4302–4315.

Kuhn M 2008. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5): 1–26.

Kumari S, Nagendran K, Rai AB, Singh B, Rao GP and Bertaccini A. 2019. Global status of phytoplasma diseases in vegetable crops. *Frontiers in Microbiology*, 10: 1349.

Millennium Ecosystem Assessment 2005. http://dx.doi.org/ 10.7927/ H47P8WB9 [accessed 14.02. 2024].

Ramankutty N, Evan AT, Monfreda C and Foley JA 2010. Global agricultural lands: croplands, 2000. Data distributed by the Socioeconomic Data and Applications Center (SEDAC). http:// sedac.ciesin.columbia.edu/es/aglands.html.

Trivellone V, Cao Y and Dietrich CH 2023. Multilocus next-generation sequencing of leafhopper-associated phytoplasmas highlights gaps in knowledge for some phytoplasma lineages and genetic *loci*. *Phytopathogenic Mollicutes*, 13(1): 115-116.

Venter O, Sanderson EW, Magrach A, Allan JR, Beher J, Jones KR, Possingham HP, Laurance WF, Wood P, Fekete BM and Levy MA 2016. Global terrestrial human footprint maps for 1993 and 2009. *Scientific Data*, 3(1): 1-10.

Wolf MJ, Emerson JW, Esty DC, de Sherbinin A and Wendling ZA 2022. Environmental performance index. New Haven, CT: Yale Center for Environmental Law & Policy. epi.yale.edu.