Enhanced risk-based surveillance of *Xylella fastidiosa*

Location

INRAE Avignon, Biostatistics and Spatial Processes unit
https://informatique-mia.inrae.fr/biosp/accueil

Supervisors

Edith Gabriel (BioSP, INRAE Avignon), Dino Ienco (UMR TETIS, Montpellier), Eric Verdin (PV, INRAE Avignon)

Summary

By devasting ecosystems and causing economic injuries, plant diseases require effective surveillance to detect epidemics at very early stage. The project aims at developing an enhanced risk-based surveillance coupling risk factors analysis and species distribution models. Methods and analyses will be driven by the epidemiology of *Xylella fastidiosa*. New predictors of its presence will be introduced, based on soil and leaf ionome composition variables. Then, an original machine learning approach will be developed to take into account both the heterogeneity of the data (inspection surveys, satellite images, sensor data… observed at different resolutions and times) and their spatio-temporal dimension and dependencies. Our predictive model will produce risk maps and help to design surveillance strategies. Artificial intelligence paradigms and tools will be used for, in a loop, (i) learning the risk (and its determinants) from data and (ii) planning data collection from the current risk evaluation.

Research questions

This project aims at finding new indicators and sampling strategies for epidemiological surveillance (ES) across several horizons of space and time by (i) leveraging the heterogeneity and the mass of data, (ii) revealing links between potential predictors and observations of pathogens from heterogeneous observation strategies, and (iii) extrapolating and predicting the presence of a pathogen or the risk of its presence.

- We want to develop a generic supervised learning approach to conceive an ES informed by massive and heterogeneous data.

- Because designing efficient ES of invasive pathogens, as *Xylella fastidiosa*, relies on the knowledge of risk factors and their interactions, we will focus on finding new indicators that will be generated by the above-mentioned supervised learning approach.

INRAE
science for people, life & earth

Research program

Invasive plant pathogens cause serious environmental and ecological damages with major economic impacts. *Xylella fastidiosa* is one of them, causing serious diseases in a wide range of plants such as grapevine and olive trees. Eradication management of *Xylella fastidiosa* currently involves surveying for early field symptoms, removing the infected plants once the bacterium is detected then establishing demarcated areas. Eliminating pesticides from plant protection involves a multitude of decisions for building and maintaining agro-food systems based on ecological and physical processes that favour plant health. The decisions required by farmers to deploy effectively prophylaxis are based on the availability of expert knowledge combined with numerous environmental variables whose states can be continuously updated via environmental monitoring – as part of systems referred to as epidemiological surveillance (ES). In the case of olive tree, the Leccino cultivar has demonstrated tolerance to the disease (D'Attoma et al., 2019) – that seems to be associated with a specific micronutrient content of the plant and a specific microbial composition of its endophytes. Future surveillance could thus focus on endophytic composition of trees in regions with climates and soils most favourable for Olive Quick Decline symptom expression. This would give prophylaxis the needed time to alter plant nutritional states and endophytes to mimic those of the Leccino cultivar.

Surveys cover large and complex landscapes. The ES therefore integrates multiple sources of data that potentially inform the presence (or risk of presence) of the pathogen. Environmental variables monitored in ES typically include abiotic conditions that are mechanistically relevant to disease epidemiology (i.e., temperature, relative humidity, leaf wetness…) and traces of the pathogen (symptoms on local crops; changes in leaf ionome composition…). Conceiving such an ES, incorporating a multitude of factors, across several horizons of space and time is thus challenging.

Martinetti and Soubeyrand (2018) proposed a risk-based surveillance strategy, based on a combination of machine learning techniques and network analysis, for understanding the main abiotic drivers of infections caused by *Xylella fastidiosa* and producing risk maps. Their predictors for explaining the presence of *Xylella fastidiosa* are mainly related to climate variables (precipitation seasonality, temperature in winter, solar radiation in summer…). Recent works, see e.g. Del Coco et al. (2020), reveal significant differences in soil and leaf ionome composition between safe areas and in presence of *Xylella fastidiosa*. Predictive models will be proposed on the basis of such new variables. Predictions will then be compared to the ones of Martinetti and Soubeyrand (2018). We will focus on proxies (eventually made of a combination of covariables) that allow large spatial coverage and cost reduction (in comparison with human inspections and specific biotic sensors).

Numerous factors potentially inform *Xylella fastidiosa* presence and dissemination (more than 100 factors, e.g. related to plant-health sanitary alerts, meteorological predictions, land use, satellite environmental data, abiotic sensor data…) and these factors are observed at various space-time scales and degrees of specificity. We thus need predictive models that deal with target variables and covariables that are potentially spatially and temporally autocorrelated and that can be noisy, more or less reliable, and collected at diverse spatial resolution x coverage x density. Machine learning, more specifically random forest approaches, has been successful in geoscience and spatial predictions, as well as in the study of systems that are dynamics. See e.g. Reichstein et al. (2019) and Salcedo-Sanz et al. (2020) for reviews on machine learning methods and information fusion in geoscience. However, machine learning methods rarely account for spatial and/or temporal dependencies, and hence high risk to lead to wrong extrapolation, interpretation, causal relation, and ignorance of confounding effects.

Several attempts to account for spatial or temporal dependencies have been made, but the mainstream approaches are based on hand-crafted features. In this project we want to go further, by jointly considering both the spatial and temporal dimensions to deal with spatio-temporal autocorrelation. Spatio-temporal dependencies will be overcome by deploying machine learning approaches (Meyer et al., 2019; Szatmári and Pásztor, 2019), including species distribution models (Norberg et al., 2019) and tailor them to the heterogeneity of the data and to our ES objectives: risk prediction and risk factor identification.

Coupling risk maps provided by our predictive models and optimization or allocation algorithms will then allow us to identify optimal risk-based surveillance (Mastin et al., 2020). However, instead of considering static situations, we will propose sequential sampling strategies conditional on the risk but also on previous samples, by mutually feeding risk mapping and data collection in a loop. Developing such a dynamic vision of sampling is especially paramount for *Xylella fastidiosa*: we know that a mild winter can favour bacteria expansion and symptom expression in regions beyond the usual suitability envelope, and surveys should be adapted accordingly.

Main steps

- Literature survey on machine learning approaches and epidemiology of *Xylella fastidiosa.*
- Identify new proxies – based on soil and leaf ionome composition – for *Xylella fastidiosa* and compare predictions with the literature.
- Develop machine learning methods accounting for heterogeneous data with spatio-temporal dependencies and compare them with competing strategies.
- Optimize the ES from our models and algorithms adequate for space-time sampling processes and/or (if samples can be made by our partners) diagnostic test for the new proxies of *Xylella fastidiosa* presence.

Profile

Strong understanding of machine learning and/or spatial statistics concepts;
Good coding skill in R (and/or Python, C++);
Interest for reading research papers;
Interest for applied mathematics and epidemiology.

Application

Please send your CV, cover letter, 1 or 2 reference letter and a copy of your M1 and M2 transcripts to edith.gabrie@inrae.fr, dino.ienco@inrae.fr and eric.verdin@inrae.fr **before 30 April 2021**.

References

D'Attoma G et al. (2019) Ionomic Differences between Susceptible and Resistant Olive Cultivars Infected by Xylella fastidiosa in the Outbreak Area of Salento, Italy. *Pathogens*, 8(4):272

Del Coco L et al. (2020) Soil and Leaf Ionome Heterogeneity in Xylella fastidiosa Subsp. Pauca-Infected, Non-Infected and Treated Olive Groves in Apulia, Italy. *Plants*, 9(6), 760.

Martinetti D and Soubeyrand S (2018) Identifying Lookouts for Epidemio-Surveillance: Application to the Emergence of Xylella fastidiosa in France. *Phytopathology*, 109(2):265–76.

Mastin J et al. (2020) Optimising risk-based surveillance for early detection of invasive plant pathogens. *PLOS Biology*, 18(10):e3000863.

Meyer H et al. (2019) Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecological Modelling*. 411:108815.

Norberg A et al. (2019) A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3):e01370.

Reichstein M et al. (2019) Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195—204.

Salcedo-Sanz S et al. (2020) Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, 63, 256—272.

Szatmári G and Pásztor L. (2019) Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. *Geoderma.* 337, 1329–1340.