Identifying Lookouts for Epidemio-Surveillance: Application to the Emergence of *Xylella fastidiosa* in France

Davide Martinetti[†] and Samuel Soubeyrand[†]

First and second authors: BioSP, INRA, 84914, Avignon, France.
Accepted for publication 15 November 2018.

ABSTRACT

Recent detections of *Xylella fastidiosa* in Corsica Island, France, has raised concerns on its possible spread to mainland France and the rest of the Mediterranean Basin. Early detection of infected plants is paramount to prevent the spread of the bacteria, but little is known about this pathosystem in European environments, hence standard surveillance strategies may be ineffective. We present a new methodological approach for the design of risk-based surveillance strategies, adapted to the emerging risk caused by *X. fastidiosa*. Our proposal is based on a combination of machine learning techniques and network analysis that aims at understanding the main abiotic drivers of the infection, produce risk maps and identify lookouts for the design of future surveillance plans.

The identified drivers coincide with known results in laboratory studies about the correlation between environmental variables, such as water stress and temperature, and the presence of the bacterium in plants. Furthermore, the produced risk maps overlap nicely with detected foci of infection, while they also highlight other susceptible regions where *X. fastidiosa* has not been found yet. We conclude the paper presenting a list of recommended regions for a risk-based surveillance campaign based on the predicted spread and probability of early detection of the disease.

Additional keyword: epidemiology.

Designing efficient surveillance of invasive pathogens first relies on the knowledge of risk factors and their interactions. Such knowledge allows the conception of risk-based surveillance strategies that are expected to lead to an efficient allocation of sampling effort. Examples of risk-based surveillance approaches are given by Parnell et al. (2014) for Huanglongbing disease, Ferrer et al. (2014) for avian influenza, and Whist et al. (2014) for Johnes disease. In these studies, the risk assessment component is grounded on a limited number of risk factors, and Parnell et al. (2014) established that effective targeted surveying plans can be obtained with relatively minimal information. However, more accurate risk maps can be obtained by combining numerous potential risk factors and, eventually, diverse sources of information (e.g., host characteristics, climatic data, and land cover). This is typically the case in niche modeling and species distribution modeling founded on geo-referenced species records, environmental databases, and regression tools such as generalized linear models (Elith and Leathwick 2009; Franklin 2010; Purse and Golding 2015). Such approaches have recently been implemented for *Xylella fastidiosa*, the pathogen under focus in this study, by Bosso et al. (2016) and Godefroid et al. (2018).

Epidemio-surveillance can be further improved by going beyond the static vision of risk factors (Hyatt-Twynam et al. 2017; Parnell et al. 2017). The idea here is to couple (i) risk factor analyses with (ii) risk evaluation based on current sanitary situations and information on disease dynamics. Evaluating risk using information on disease dynamics (point (ii)) and deducing adequate strategies for surveillance and control is nowadays relatively common in plant epidemiology (Cunniffe et al. 2016; Rimbaud et al. 2018). To

further this approach, Parnell et al. (2014) proposed to combine points (i) and (ii) by expressing the overall risk at any location as the product between the potential distribution of the disease (obtained from a sort of risk factor analysis) and its likely dispersal from known infection areas (obtained from surveillance data and a disease-spread model fitted to data). Then, a survey can be designed by drawing sampling locations proportionally to the overall risk.

In this article, we present an approach in line with the one proposed by Parnell et al. (2014) that couples a risk factor analysis and an epidemiological model to build an overall risk taking into account not only the suitability of the environment but also the potential spread of the pathogen. The uniqueness of our approach concerns how the risk factor analysis is carried out and how the epidemiological model is built. The risk factor analysis is performed via a machine learning approach, by using (i) a large number of high-resolution, climatic, and land-use explanatory variables; (ii) a battery of regression models allowing to handle model uncertainty and to overcome the constraints induced by the assumptions of a single model; and (iii) a spatially structured rebalancing technique applied to the dataset to cope with the eventual prominent influence of uninfected areas in the regression and to avoid spatial autocorrelation bias. This stage of the analysis allows the analyst to gain insight into significant risk factors governing the presence of the disease, to draw basic risk maps for the areas where data were collected (i.e., training dataset), and to project basic risk maps in other areas (the term basic indicates that the spread of the pathogen is not taken into account to assess the risk). For the second stage of the analysis, we built an epidemiological model taking the form of a discrete-time stochastic cellular automaton (Bian 2013) and defined conditionally on the basic risk map for weighting the links between the nodes of the network. We adopted this generic framework to tackle situations where little information is known on the spread of the disease, but information has been gained on the spatial heterogeneity of the risk of infection and susceptible host distribution. Obviously, this framework could be improved to take into account knowledge about long-distance dispersal, for instance. However, in our approach, the

[†]Corresponding authors: Davide Martinetti; E-mail: davide.martinetti@inra.fr, and Samuel Soubeyrand; E-mail: samuel.soubeyrand@inra.fr

Funding: This research was funded by the Horizon 2020 Framework Programme grant HORIZON 2020 Xf-ACTORS project SFS-09-2016.

epidemiological model is not intended to portray the ongoing outbreak, but to identify a set of locations that are relevant for surveillance, such as, for example, those locations where the epidemic is likely to be observed earlier than on a randomly chosen set of locations.

Thus, in this article, we develop an approach that exploits the knowledge of risk factors to build a generic dynamical model of outbreaks and, as a corollary, to identify lookouts that should allow the detection of pathogen emergence and disease range expansion in a limited time span with high probability. Here, lookouts are viewed as locations in space where one has a higher chance to detect an outbreak that would have started at a (some) more or less arbitrary location(s). In the medical literature, a similar surveillance strategy is referred to as sentinel surveillance and it is used when high-quality data are needed about a particular disease that cannot be obtained through a passive system. A sentinel system deliberately involves only a limited network of carefully selected reporting sites (usually hospitals), where in-depth analysis for the detection of the disease can be conducted. Lookouts can be considered as a generalization of hubs in epidemic networks. Indeed, in such networks, the hubs (i.e., the nodes that are linked with a high number of nodes) have to be under particular focus for surveying and controlling an infectious disease because of both their exposition to the disease and the ability to spread the disease. However, not all hubs are equally important and other types of nodes could be of interest, such as, for example, nodes in the center of the network (k-shell centrality) (Kitsak et al. 2010), nodes that are preferentially chosen by their neighbors (VoteRank) (Zhang et al. 2016), nodes

that connect efficiently with the highest number of neighbors (accessibility) (De Arruda et al. 2014), or nodes where an epidemic will show up with relative high frequency. Hence, the approach proposed in this article aims at identifying lookouts for epidemiological surveillance by considering different measures of nodes' importance.

We applied our approach to one of the most prominent plant pathogens presently emerging in Europe: *X. fastidiosa*. *X. fastidiosa* has been in situ detected and identified in 2013 in Italy, 2015 in France, and 2016 in Spain and Germany (Appendix). This bacterium has a large range of wild and cultivated host plants, it can be found in the xylem of the infected plants, as well as in the foregut of the insect vector, and may cause a rapid decline of its host (Purcell 2013). It is spread by insect vectors that feed on plant xylem (e.g., the meadow spittlebug, *Philaeus spumarius*) (Saponari et al. 2014) and by transport of infected plants. Since *X. fastidiosa* was detected, a surveillance and control protocol has been implemented and applied by local governmental agencies and other stakeholders. A large number of positive cases have been found in Corsica, especially in the South, while fewer positive cases have been found in Provence-Alpes-Côtes d'Azur (PACA hereafter), the South-East region of mainland France (Fig. 1). A closer look at data shows a particularly heterogeneous spatial distribution of *X. fastidiosa* in the contaminated regions, which are marked by relatively important variations in land use, altitude, distance to the shore and climatic conditions. The application of the approach briefly described above allowed us (i) to identify risk factors among a long list of potential explanatory variables related with climate and land use, and (ii) to

Fig. 1. Each dot represents the geographical location of a sampled plant: dark gray/red spots correspond to absence of *Xylella fastidiosa*, while light gray/green ones to presence. Left panel: Provence-Alpes-Côtes d'Azur (PACA) region. Right panel: Corsica Island. The zoomed portion of PACA contains all positive cases found in the region.

draw maps of basic and overall risk. For this application, we used observations made in Corsica as training data and we projected the risk in PACA. Such a projection is expected to guide epidemiological surveillance in this region, which is still today slightly impacted by *X. fastidiosa*.

MATERIALS AND METHODS

The methodology presented in this paper begins by coupling existing geolocalized observations of presence/absence of *X. fastidiosa* with other spatial multiscale environmental (a.k.a. abiotic) predictors, such as climate and land use. The produced fine-scale data are then used to learn relevant factors that can explain the presence of *X. fastidiosa* in France, to predict the risk of appearance of *X. fastidiosa* in the unsampled part of PACA and Corsica, and to estimate the epidemic dynamics of *X. fastidiosa* in order to design future surveillance campaigns.

Data. Sampling *X. fastidiosa* in France. A surveillance protocol has been implemented in France since July 2015 to face the emerging risk of diffusion of *X. fastidiosa* (Appendix). The protocol prioritizes those French regions (Corsica Island, PACA, and Languedoc-Roussillon) that are considered more susceptible to *X. fastidiosa* due to their geographical and climatic similarity to other non-French regions that already experienced *X. fastidiosa* infections. Corsica is precisely where the first infected plant was officially detected in 2015. The choice of sampling locations was random or evidence-based (Soubeyrand et al. 2018), while the surveillance protocol establishes a routine in case of positive detection: all infected, symptomatic and susceptible plants (reference to the list of known host plant is presented in the Appendix) in a 100-m diameter buffer zone around the positive-detected plant are uprooted, insecticide is applied to prevent the spread of the insect vectors, and reinforced surveillance is performed during the months following the first detection to test the actual eradication of the pathogen. Reinforced surveillance based on both clinical tests and visual inspections extends to a larger 10-km buffer in order to delimit the infected zone and detect new potential foci (references to the official surveillance and control strategies are presented in the Appendix).

Since 2015, 22,439 plants have been tested for the presence of *X. fastidiosa*: 21,332 negatives, 1,045 positives, and 62 inconclusive (last updated March 2018). For each sample the following features are recorded: date and GPS localization of the sample, species of sampled plants (more than 200 sampled species), and the result of the clinical analysis. Occasionally, additional information is also recorded, such as the circumstances of the sampling (prospection, reinforced surveillance in a buffer zone, event-driven observation, etc.), the subspecies of *X. fastidiosa*, or the presence and severity of symptoms on the sampled plant.

Bioclimatic predictors. Climatic conditions have long been observed to play a primary role in limiting species distributions and patterns. Bioclimatic envelopes (BIOCLIM) (Busby 1991) are well-established methods for species distribution modeling and are based on the assumption that a combination of different bioclimatic predictors can shape a species niche (Section 5.2.1 of Franklin 2010). In this work we considered a set of climatic variables from WorldClim Version 2 (Fick and Hijmans 2017) including monthly minimum and maximum temperatures, precipitations, solar radiation, water vapor pressure, as well as 19 other variables including seasonal means, variabilities and extremes of precipitation, and temperature on a spatial resolution of 1 km.

Land use and land aspect predictors. In recent years we observed an increased availability of detailed spatial (geographic information systems) databases concerning land use/cover and landscape aspects such as altitude and slope, that already found successful applications in species distribution modeling for invasive plants (Ibáñez et al. 2009) and plant pathogens (Meentemeyer et al. 2008). In this work we bring together a very recent land use atlas

at 20-m resolution with 17 classes (Inglada et al. 2017) and a 20-m resolution digital elevation model for altitudes.

Study area and data preparation. The study area is composed of the two French regions of Corsica and PACA. Since the sampling campaign accumulated more observations from Corsica Island (15,778 observations since 2015: 991 positives and 14,787 negatives), we will use Corsica for model training and PACA for predictions. Furthermore, due to the heterogeneity of the spatial resolution of the databases (*X. fastidiosa* presence/absence data are spatial points, bioclimatic variables are on 1-km grid, while land use and altitude are on a 20-m grid), we rescaled all data in order to share the same spatial resolution using the DFCI geographic coverage 1-km-grid (upper left corner of Figure 2 provides a sample of the DFCI grid used in the study and the Appendix provides more information on DFCI grids). For each cell of the resulting grid, we computed a set of 100 variables: 24 monthly average maximum and minimum temperatures, 12 monthly average precipitations, 12 monthly average solar radiations, 12 monthly average water vapor pressures, 19 bioclimatic variables (Fick and Hijmans 2017), 4 variables for minimum, median, mean and maximum altitude, and finally, the proportions of the 17 types of land use. Notice that the presence/absence of *X. fastidiosa* in the grid cells is computed as follows: 1 if there was at least one positive plant sample, 0 if all sampled plants in the cell were negative to *X. fastidiosa*, and undetermined if no plant was sampled from the cell. In Corsica, out of 11,102 grid cells, 1,976 (17.8%) of them had at least one sample, of which 309 had at least one positive sample (15.6% of sampled cells). One advantage of this construction is that it drastically reduces the selection bias that may be caused by the sampling strategy: as explained in the previous section, once an infected plant is detected, a reinforced surveillance is applied in a 10-km-diameter buffer around the first detection, hence causing a nonrandomly, highly spatially autocorrelated distribution of samples.

Statistical analysis. The statistical analysis presented here aims at three fundamental objectives: (i) to understand which predictors can explain better the presence of *X. fastidiosa*, (ii) to make realistic predictions of the risk zones in Corsica and PACA, based on the predictors identified in the first step, and (iii) to simulate the epidemic dynamic of *X. fastidiosa* in order to inform future surveillance strategies. The first two steps rely on the combination of different machine learning techniques, while the latter is based on network analysis.

Feature selection. In order to identify a reduced set of predictors that better correlates with the presence/absence of *X. fastidiosa* and to avoid multicollinearity issues, we performed several machine learning techniques that assign a rank and a relative importance to each of the 100 variables of the data set presented in the previous section. The 21 classification learners used in this step include both regression and decision-tree algorithms, in both their standard formulation or in advanced reformulations that involves different levels of boosting (adaptive boosting, gradient boosting, etc.) or bagging (e.g., random forest), as well as other learners such as multivariate adaptive regression splines (MARS), support vector machines (SVM), weighted k-nearest neighbors, Bayesian classification with Gaussian processes, linear discriminant analysis, and naive Bayes classifiers. Furthermore, spatial sampling with repeated cross-validations has been used in order to avoid spatial autocorrelation issues. The R libraries *iml* and *mlr* propose a straightforward implementation of the listed algorithms (Bischl et al. 2016; Molnar 2018 and references therein). The feature importance assigned to each of the 100 variables by the 21 learners are ranked in decreasing order and the subset of the 40 highest ranked features is finally retained.

Risk maps. Producing *X. fastidiosa* risk maps for both Corsica and PACA region requires first the calibration of a statistical model capable of mapping the probability of appearance of *X. fastidiosa*, based on the value of the predictors chosen in the previous step for the grid cells where at least one plant sample have been collected

(training data set). Such technique usually goes under the name of species distribution modeling (Franklin 2010 and citations therein), or less frequently, ecological niche models, bioclimatic envelopes, habitat models, and resource selection functions. Since the training data set is unbalanced (presence of *X. fastidiosa* is detected on less than 16% of the cells), we rebalanced the dataset in order to avoid bias toward negative values. In problems of binary classification, strongly imbalanced classes often lead to unsatisfactory results regarding the prediction of new observations, especially for the minority class, in this case the presence of *X. fastidiosa*. Most classification methods work best when the number of observations per class are roughly equal. In order to lay more weight on the cases of the minority class, we deliberately remove randomly chosen instances from the majority class, while all cases of the minority class are kept. This arrangement attains at multiple purposes: firstly, the fitted models will train on a relatively higher proportion of observed presences of *X. fastidiosa*, allowing a deeper understanding of the actual relationship between the predictors and the presence of *X. fastidiosa*. Secondly, if on the one hand the detected presences can be explained by a favorable environment for *X. fastidiosa*, on the other hand the observed absences could be the consequence of exogenous factors that are not directly related to the suitability of the conditions for the propagation of the bacteria as follows: (i) epidemic dynamic, since the epidemic of *X. fastidiosa* in Corsica Island is still ongoing, we may have sampled in places where the epidemic has not arrived yet, even if conditions are favorable; (ii) validity of the diagnostic tests, the reliability of the diagnostic tests is still uncertain, hence we may have labeled as negative some actually positive samples; (iii) unobserved positives, the infection

from *X. fastidiosa* presents moderate to no symptoms, especially in the early stages of the infection, and combined with the fact that it has a broad host range and not all parts of an infected plant contain the bacteria, it may have led to missing actual infected plants in the surveyed cells.

For the choice of the predictive model, we followed a similar approach that in the previous section, comparing the performance of different models, such as generalized linear models (GLM), GLM with spatial autocorrelation (SAR-GLM) (Franklin 2010; Martinetti and Geniaux 2017, chapter 6.8), MARS (Friedman 1991), Random Forest (Breiman 2001), SVM, decision trees, and two boosted glm: extreme gradient boosting (XGBOOST) (Friedman 2002) and GLMBOOST (Bühlmann and Yu 2003). The performance of the different models has been measured with respect to the prediction error using the area under the curve (AUC, i.e., the proportion of times that a random selection from the positive group scores higher than a random selection from the negative group) as in Parnell et al. (2014) and the Kappa statistic (percentage of times that the predicted value agrees with that observed, corrected by the amount of agreement that could be expected to occur by chance; Cohen 1960). All models have been trained on different levels of rebalancing (all integer values between 17 and 50%) and 100 repetitions. The model that showed the best performances in terms of both AUC and Kappa statistics across the whole experiment (10-fold spatial cross-validation across all models and levels of rebalancing) was a Random Forest trained on a rebalanced dataset with 35% of positive observations. This is likely due to the ability of Random Forest of adapting to heterogeneous training data via feature bagging, that combined with spatial cross-validation

Fig. 2. Basic risk maps in Corsica and Provence-Alpes-Côtes d'Azur (PACA) inferred from *Xylella fastidiosa* surveillance data collected only in Corsica. Darker shades correspond to higher risk of presence of *X. fastidiosa* according to the fitted model. The zoomed portion of PACA also displays positive samples, that are entirely localized in high-risk regions, whereas they have not been included in the analysis. In upper left corner of the left panel is depicted a sample of the DFCI grid used in the data construction. The range of shading is different for the two panels.

allowed for a flexible predictive model accounting for spatial autocorrelation due to residual sampling bias. The retained model is finally applied to the test data set (unsampled cells of Corsica and all PACA) in order to produce basic risk maps (Fig. 2).

Network analysis. The final step of the statistical analysis is intended to model the epidemic dynamics of *X. fastidiosa*. We adopt a weighted complex network approach where the nodes of the network are constituted by the cells of the grid and the edges (binary links that represents the probability of contagion between pairs of nodes) are computed as follows: for every pair of nodes (cells) i and j ,

$$W(i, j) = \mathbb{I}_{(i, j) \in \mathbb{E}} r(i) r(j)$$

where $r(k) \in [0, 1]$ is the risk predicted for node k by the model presented in the previous section, $\mathbb{E}$ is the set of neighboring cells of the grid (the symmetric squared matrix W is known as adjacency matrix) and $\mathbb{I}_A$ is the indicator function that take value 1 if the condition A is verified, and 0 otherwise. The neighboring relationships between cells of the 1-km-grid (upper left panel of Figure 2) is based on the first six nearest neighbors and it has been chosen since it can realistically represent the short-distance spread ability of the insect vectors, estimated to range between a few meters to up to 1 km (Cornara et al. 2018; Saponari et al. 2014; Strona et al. 2017, and references therein). Furthermore, we chose to compute the link weight for neighboring nodes i and j using the product of the corresponding risks, i.e., $r(i)r(j)$, in such a way that the probability of contagion is proportional to the combined risk of both nodes (other binary aggregation operators could have been used, such as triangular norms and binary residuals operators [Klement et al. 2000] or dispersal kernels [Lindström et al. 2011], but it is beyond the scope of this paper).

To simulate the epidemic dynamic on the network, we set up a susceptible-infected-susceptible (SIS) model in which each node i , at a given time t , can be either in the infected or susceptible state ($\mathbb{I}_t(i)$ equals 1 if the node is infected and 0 if it is susceptible at time t). We assume that all N nodes are equally susceptible in such a way that the epidemic is ruled solely by the network topology, an approach that goes under the name of individual-based cellular automata in the characterization made by Bian (2013). Each SIS simulation starts at time t_0 with a relatively small focus of 10 infected neighboring nodes (<0.1% of the total number of nodes). At each discrete-time step t , the SIS model iterates over all N nodes in such a way that, for every node i , its state is defined by the following equation:

$$I_t(i) = \mathbb{I}_{\{I_{t-1}(i)=0\}} \left(\sum_{j \in \mathbb{E}(i)} (W(i, j) I_{t-1}(j)) > \beta \right) + \mathbb{I}_{\{I_{t-1}(i)=1\}} (1 - \mathbb{I}_{\{U_{it} > \lambda\}}) \quad (1)$$

where $\mathbb{E}(i)$ is the subset of $\mathbb{E}$ containing the neighbors of node i and U_{it} is a random number sampled from a uniform distribution in the unit interval. The two parameters $\beta \in [0, 1]$ and $\lambda \in [0, 1]$ drives the way in which the epidemic propagates: from the first part of Equation 1, a noninfected node at time $t - 1$ can become infected at time t if and only if the sum of the edge weights of its infected neighbors is above a certain threshold β . From the second part of the same equation, infected nodes at time $t - 1$ return to the susceptible state at time t with probability λ . In other words, the infection can only propagate between nodes that have sufficiently strong links. It is important to clarify that the proposed SIS model is not intended to portray a realistic representation of the ongoing epidemic of *X. fastidiosa*, but rather to identify a set of influential nodes that

Fig. 3. Map of the frequency of passage of epidemics simulated with the susceptible-infected-susceptible model. Darker shades correspond to higher frequency of passage. The range of shading is different for the two panels.

are relevant in terms of surveillance and potential spread of the epidemic. For these reasons, the values of the parameters $\beta = 0.15$ and $\lambda = 0.1$, as well as the stopping time of the simulated epidemic $T = 150$, have been chosen in such a way that, depending on the location of the initial focus, the epidemic either dies out or becomes endemic, usually on a clustered subset of network's nodes. Computationally, we ran the SIS model from all N possible starting nodes of the network and, for each node i , we recorded (i) the epidemic size

$$S(i) = \sum_{j=1}^N I_{T(j)}$$

i.e., the number of infected nodes at the end of the simulated epidemic with focus in i , and (ii) $F(i, j) = \mathbb{I}_{\{I_t(j)=1 \wedge I_t(i)=1, t \in \{0, \dots, T\}\}}$, a vector of size N where ones correspond to the index of the nodes j that became infected at least once when the simulated epidemic started from node i . While epidemic size can be used to estimate the epidemic risk associated to a node, the value

$$F(j) = \sum_{i=1}^N F(j, i) r(i)$$

can be interpreted as the frequency of passages of an epidemic through node j , weighted according to the risk of the initial focus $r(i)$.

Finally, we also benefit from the constructed network structure to compute other metrics in order to evaluate the relative importance of nodes as lookout points. From a recent review on vital nodes

identification methods for weighted networks by Lü et al. (2016, Chapter 7), we chose to compute the following indicators: (i) k-shell index, which specifies the degree of network centrality of nodes (Garas et al. 2012; Kitsak et al. 2010), (ii) VoteRank, which identifies a set of decentralized nodes with the highest spreading ability (Zhang et al. 2016), and (iii) generalized random walk accessibility (De Arruda et al. 2014), which quantifies the potential of each node to access other nodes via a random walk of arbitrary length, penalizing walks with longer lengths. In the case of spatial networks, DeArruda et al. (2014) proved that accessibility (Travençolo and Costa 2008) has the overall best performance in determining the importance of a node with respect to an epidemic process.

Surveillance strategies on networks. A surveillance strategy consists of choosing a subset of nodes of the network that is more efficient in detecting the presence of an ongoing epidemic on the network (Herrera et al. 2016). A trivial strategy would be to select all nodes of the network, but that would be unfeasible for obvious reasons. We chose to select a subset of 1% of nodes (which roughly corresponds to the number of nodes sampled every year since the beginning of the campaign). We restrict this exercise to the case of PACA region, since Corsica already has a high prevalence of infection and early detection is no longer a priority. In order to choose 407 nodes in PACA region (i.e., 1% of the total 40,674 nodes), we consider the following seven strategies:

- rand: random: select nodes at random.
- racq: random acquaintance: select a random acquaintance of random nodes (this should bias toward more connected nodes).

Fig. 4. Map of the k-shell index and VoteRank. Darker shades correspond to higher k-shell index. Furthermore, the white/yellow dots highlight the 200 points (100 for Provence-Alpes-Côtes d’Azur and 100 for Corsica) that have higher spreading ability according to the VoteRank algorithm. The range of shading is different for the two panels.

- risk: risk-based: chose the 1% of nodes with the higher value of the risk index (Fig. 2).
- pass: frequency-based: chose the 1% of nodes with the higher value of the frequency of passages of the epidemic (Fig. 3).
- kshel: k-shell: chose the 1% of nodes with the higher value of the k-shell index (Fig. 4).
- vote: VoteRank: chose the first 407 nodes found by the algorithm VoteRank (Fig. 4).
- grwa: generalized random work accessibility: chose the 1% of nodes with the higher value of the index (Fig. 5).

The chosen points for surveillance are depicted in Figure 6. The performance of the proposed surveillance strategies is quantified by means of the following four measures (Fig. 7):

- 1%-persistence early detection: the temporal lag between the surveillance subset reaching 1% of persistence of infection and the entire network reaching the same threshold.
- 2%-persistence early detection: same as before, but with 2% threshold.
- Peak timing: the temporal lag between the surveillance subset reaching its epidemic peak and the entire network reaching its peak.
- Peak ratio: the difference between the epidemic persistence reached by the surveillance subset and the entire network.

RESULTS

Relevant variables for explaining the presence of *X. fastidiosa*.

Repeated runs of different feature-selection algorithms revealed a robust subset of relevant predictors that stand out among the

100 candidate explanatory variables. We list them in order of importance.

- Precipitation seasonality, i.e., the measure of the variation in monthly precipitation totals over the course of the year (Fick and Hijmans 2017; O'Donnell and Ignizio 2012). The higher the index, the bigger the variability of precipitations during the year. Since the study area is located in a temperate climatic region, the index basically translates into the variability in mean precipitations between the humid season (from late fall to early spring) and the dry one (from late spring to early fall). The correlation between this predictor and the presence of *X. fastidiosa* is positive (the more variability in precipitations, the higher the probability of finding *X. fastidiosa*).
- Minimum temperature in winter: average minimum temperatures in December, January, February, and March. The correlation between these predictors and the presence of *X. fastidiosa* is also positive (the higher the minimal temperature, the more likely the presence of *X. fastidiosa*).
- Precipitation during the dry season: mean precipitations in the summer months of July and August. The correlation between these predictors and the presence of *X. fastidiosa* is negative (the lower the precipitation, the more likely the presence of *X. fastidiosa*).
- Solar radiation (MJ/m²) in late spring and summer: solar radiation measures the radiant energy emitted by the Sun and absorbed by the Earth. It depends on the height of the Sun over the horizon, atmospheric conditions, and the tilt of the absorbing surface. The correlation between these predictors and the presence of *X. fastidiosa* is positive (the higher the solar radiation, the more likely the presence of *X. fastidiosa*).

Fig. 5. Map of the generalized random walk accessibility index. Darker shades correspond to higher values of the accessibility. The range of shading is different for the two panels.

Other predictors also appeared to have some relevance, such as maximum temperatures in summer (positive correlation with *X. fastidiosa*'s presence) and precipitation in January and February (negatively correlated).

Risk maps. Basic risk maps obtained with the Random Forest model for both Corsica and PACA are shown in Figure 2. Recall that the model has been trained only on those grid cells of Corsica where at least one plant has been sampled (1,976 cells out of 11,102), while the rest of unsampled grid cells of Corsica and the entire region of PACA constitute the test data set. In order to check the relevancy of the obtained results, we superimposed the location of sampled positive plants over the risk maps of PACA (zoomed box in left panel of Figure 2). It appears that cells with high risk coincide with actual detected presence of the bacteria, hence validating, at least qualitatively, the predicted risk maps. Alternatively, another method for assessing the reliability of the model is to compute the Gower distance (Gower 1971) between train and test datasets: the higher the distance of an observation in the test set from the closest observation in the train set, the higher the uncertainty on the predictions (the map of the Gower distance is not reported in the manuscript, but it is available upon request).

Improving surveillance strategy. Risk maps are useful for portraying the susceptibility of the two regions to the disease, but they may be ineffective in predicting the epidemic spread. For this reason, we modeled the epidemic dynamic through a network approach. Figures 3, 4, and 5 depict the relative importance of each node in terms of different measures of their spreading ability. In particular, Figure 3 shows the weighted frequency of passages of the epidemic for each cell according to the simulated SIS model, Figure 4 shows the k-shell index of each cell and, in white or yellow dots, the 100 most influential spreaders according to the VoteRank

algorithm, while Figure 5 shows the generalized random walk accessibility index. In order to compare the different surveillance strategies proposed in Figure 6, we reran an SIS simulation from each node of the network and record the four measures described

Fig. 7. Performance of surveillance strategies. The gray/green and black/blue curves represent the epidemic persistence for a simulated epidemic on the surveillance subset and on the entire network, respectively. The 1%-persistence early detection (zoomed panel) measures the time lag between the gray/green and black/blue curves reaching the 1% threshold. The peak timing measures the time lag between the gray/green and black/blue curves reaching their respective peaks. Finally, the peak ratio measures the difference in epidemic persistence between the peaks of the two curves.

Fig. 6. Example of surveillance subsets with 400 nodes according to seven strategies: random and random acquaintance (upper left), risk (upper center), frequency of passages (upper right), k-shell (bottom left), VoteRank (bottom center), and generalized random walk accessibility (bottom right).

above for each simulation. The results of the different surveillance strategies are shown in Table 1 and Figure 8. It appears that three strategies, VoteRank, generalized random walk accessibility and risk-based, outperform k-shell, frequency of passages, and the random-based surveillance strategies in terms of early detection and peak ratio. In fact, the three best-performing strategies can achieve an average 1%-persistence early detection close to 50 time steps (100 time steps for the 2%-persistence early detection), while the persistence of the outbreak at its epidemic peak is roughly 20 percentage points higher on these surveillance subsets than on the entire network. On the other hand, the epidemic peak is reached earlier on random, k-shell, and frequency of passages strategies (at least in mean), but the differences do not look to be very significant (bottom left panel of Figure 8).

DISCUSSION

The subset of retained predictors agrees substantially with previous results found in the literature. In particular, the correlation between water stress (severe droughts in summer and high precipitation seasonality) and the presence of *X. fastidiosa* is in line with experimental studies in planta (Choat et al. 2009; McElrone et al. 2003; Thorne et al. 2006). Similarly, high temperatures have been found to favor the survival, multiplication, and proliferation of both the insect vector (Cornara et al. 2018) and *X. fastidiosa*

(Sicard et al. 2018, and references therein). An unexpected result is the complete absence of land-use predictors among the set of relevant features that correlates with the presence of *X. fastidiosa*: a possible explanation is the insufficient level of detail of the land cover characterization, since it contains just a few categories (17), combined with the fact that *X. fastidiosa* can proliferate on a great variety of (wild, agricultural, and ornamental) plants that can be found indifferently in several land-cover categories. In general, it is encouraging to find that our large-scale approach is capable of reproducing known results from small-scale studies, like in planta experiments or long-term observational studies. In view of its

TABLE 1. Average value of the four performance measures for the seven surveillance strategies

Strategy	Early detection		Peak timing	Peak ratio
	1%	2%		
rand	4.28	-44.56	-39.37	-0.0019
racq	1.72	-40.06	-38.56	-0.0008
risk	-48.83	-106.0	-15.60	0.2288
pass	-30.05	-60.29	-38.80	0.0610
kshell	-10.93	-98.64	-27.36	0.0508
vote	-49.61	-111.1	-10.03	0.2297
grwa	-46.26	-101.8	-26.10	0.2573

Fig. 8. Empirical distributions of the performance measures for the seven surveillance strategies based on: random, random acquaintance, risk, frequency of passages, k-shell, VoteRank, and generalized random walk accessibility. The top two panels show the 1%- and 2%-persistence early detection, the bottom left panel shows the peak timing, and the bottom right panel shows the peak ratio.

applications in the early phases of an emerging epidemic, our method is particularly attractive since it has reduced costs and computation time, while its ability to recover the relevant features that characterize the pathosystem can guide scientists and other experts formulating a reduced number of hypotheses about the disease and its spread.

By comparing the risk maps produced for Corsica and PACA (Fig. 2) with the detected cases of *X. fastidiosa*, it is clear that the fitted model is rather efficient at identifying susceptible regions. The case of PACA is particularly illustrative: despite the model being trained only on Corsica's data, the zones in PACA with the highest predicted risk correspond to those places where *X. fastidiosa* has been actually identified. Furthermore, other parts of the regions appear to be under moderate to high risk: the lower part of Durance River Basin and the Rhône Valley. From an economic perspective, both regions are of particular concern, since they host historical high-valued crops that are also potentially susceptible to *X. fastidiosa*, such as vineyards, olive, and fruit orchards. For Corsica, the high-risk zones coincide with those regions where *X. fastidiosa* has been repeatedly found, particularly in the southern and western part of the island.

In the perspective of designing a surveillance strategy, different metrics for the importance of a cell in terms of its spreading potential can be used. The most straightforward is the risk of presence of the disease from abiotic factors, as depicted in Figure 2. Nonetheless, this type of risk only depends on the local factors that determine environmental suitability for the installation of the bacteria, ignoring the connections with neighboring sites and the dynamic of a potential epidemic. For these reasons, we first simulated an epidemic SIS model on a spatial network issued from the risk map of Figure 2, and on the same network, we computed different indexes for identifying potential lookouts for the disease spread. Figure 3 represents a map highlighting those regions of PACA and Corsica where the epidemic simulated via the SIS model from every possible starting cell has passed more often. Figure 4 depicts a similar map where darker shades indicate higher values of the k-shell index. Nodes with low values of the k-shell index are located at the periphery of the network, while those nodes with high degree of connection that are located at the core of the network should have the higher values of the index. Therefore, only hubs at the core of the network present the highest values of k-shell index. In Figure 4, we also depicted the most influential spreaders according to the VoteRank algorithm (white or yellow dots, 100 for PACA and 100 for Corsica). Those nodes should represent a set of decentralized spreaders with the best spreading ability. Finally, in Figure 5 we report a map highlighting the index of generalized random walk accessibility that measures how many nodes can be effectively accessed during the epidemic spreading. By comparing Figures 3, 4, and 5, we can appreciate the different granularities achieved with the four methods. From finer to coarser, we find the following: (i) the generalized random walk accessibility index (Fig. 5) has the same refinement as the risk map in Figure 2, (ii) the k-shell index (Fig. 4) is a spatial smoothing of the basic risk index in Figure 2, (iii) the SIS model (Fig. 3) defines sharp high-risk regions, and finally, (iv) the VoteRank algorithm (Fig. 4) identifies a discrete set of relevant spreaders. Overall, the reported methods coincide in pointing toward already infected regions (especially southern Corsica and the zoomed area in Figure 2 in PACA) as well as the western part of PACA, that has been moderately sampled up to now (Fig. 1). The computed indices have also been used to define different surveillance strategies, based on choosing a subset of nodes with the higher value for each of the different measures of node importance from the entire network. This exercise has been performed only in the PACA region, since the epidemic in Corsica Island seems to be already well established, and surveillance for early detection is no longer a priority. The surveillance strategies consisted of choosing 1% of nodes from the entire network having the higher values of the indices of risk, frequency of passages,

k-shell, VoteRank, generalized random walk accessibility and, for the matter of comparison, a random and a random acquaintance strategy (the chosen nodes are shown in Figure 6). The choice of using only 1% of nodes meets the requirement of an optimized strategy relying on a reduced number of sampling sites, and furthermore, it corresponds to roughly 400 sites in PACA, in line with the sampling effort of previous campaigns (further simulations have been performed for surveillance subsets of sizes comprised between 0.1 and 5% of the total number of nodes, i.e., between 41 and 2,034 nodes, without any significant change in the results, hence they are not reported). The performance of the different strategies on simulated epidemics has been compared by means of four measures of surveillance effectiveness: 1%- and 2%-persistence early detection, peak timing, and peak ratio. The comparison of the six strategies (Fig. 8; Table 1) clearly favors the ones based on risk, VoteRank, and generalized random walk accessibility, which show better performances in terms of early detection and peak ratio. The k-shell-based and the frequency-of-passages-based strategy performs worse than the previous ones, but still largely outperforms random and random acquaintance strategies, which are rather poor. In general, all strategies based on some type of index computed as a function of risk outperform random strategies, since they perform better educated choices. On the other hand, to explain the differences between risk-based strategies, we may look at the spatial distribution of the suggested surveillance locations in Figure 6. It is clear that the level of spatial clustering of the sampling nodes compromises the efficiency of the surveillance strategy: too much clustering (k-shell and frequency of passages) results in inefficient strategies, while VoteRank, risk-based, and generalized random walk accessibility ensure a good spatial cover as well as sufficient separation between sampling points.

The proposed surveillance framework relies heavily on the 22,439 tests performed on sampled plants collected during a 3-year campaign and a great deal of the results presented here would not be possible without such abundance and quality of data. Nonetheless, even in the perspective of an earlier response to the outbreak, preliminary results could have been produced with fewer observations (especially fewer detected presences of *X. fastidiosa*). One solution could have been to produce different surveillance strategies based on more uncertain risk maps and then adjust the ongoing sampling campaign according to the suggested sampling location. The strategy providing more positive identifications would have increased the volume of the data set and validated empirically the appropriateness of the proposed strategy. This process could have been repeated with on the fly adjustments in order to increase the number of positive identifications and improve the quality of the risk and prediction models.

We conclude our remarks with a note on the methodology used in the paper. A great deal of the obtained results issued from a combination of different machine-learning and data analysis techniques. These methods, despite not being specifically designed for epidemiological studies, can still prove to be very effective if used wisely. The key for their success is twofold: firstly, we need to compensate for the lack of specificity by testing a plethora of different approaches, while the comparison of their performance, for example through a cross validation, can help select the best performing ones. Secondly, although the abundance and particularly the redundancy of information can be an obstacle in classic modeling, due to an increase in the number of parameters, the used techniques can take the best of the provided information, without the need of in-depth a priori knowledge on the underlying physical phenomena. This last point makes them particularly attractive, especially in the early phases of an emerging epidemic, provided a certain number of positive cases have been detected, when time, resources and knowledge of the pathosystem are scarce. Thus, in the early study of the large-scale distribution of *X. fastidiosa* in Europe,

our approach gives new insights completing those obtained with alternative modeling approaches that have been implemented, namely species distribution modeling (Bosso et al. 2016; Godefroid et al. 2018), network analyses (Strona et al. 2017), and mechanistic modeling and inferences (Abboud et al. 2018; Soubeyrand et al. 2018; White et al. 2017).

APPENDIX

A comprehensive list of positively detected cases of *X. fastidiosa* is maintained by the European Commission, Ref. Ares (2017) 3773669 and can be found at https://ec.europa.eu/food/sites/food/files/plant/docs/ph_biosec_legis_list-demarcated-union-territory_en.pdf.

The list of known host species for *X. fastidiosa* is presented here: https://ec.europa.eu/food/plant/plant_health_biosecurity/legislation/emergency_measures/xylella-fastidiosa/susceptible_en.

Surveillance and control strategies for France are described in official plans DGAL/SDQSPV/2017-653 and DGAL/SDQSPV/2017-39; see <https://info.agriculture.gouv.fr/gedei/site/bo-agri/instruction-2017-653> and <https://info.agriculture.gouv.fr/gedei/site/bo-agri/instruction-2017-39>. The 2-km DFCI grid used in this work can be downloaded from <https://www.data.gouv.fr/fr/datasets/carroyage-dfci-2-km/>.

ACKNOWLEDGMENTS

We thank DGAL (French General Directorate for Food), Anses (French Agency for Food, Environmental and Occupational Health & Safety), SRAL (French Regional Directorate for Food), FREDON (French Regional Federation for Pest Control), LNR-LSV (French National Reference Laboratory for Plant Health), and certified laboratories for data collection, molecular analyses, and data availability. We also thank the Cesbio laboratory for making the land-use map available under 'Open Database' license (<https://opendatacommons.org/licenses/odbl/summary/>).

LITERATURE CITED

- Abboud, C., Bonnefon, O., Parent, E., and Soubeyrand, S. 2018. Dating and localizing an invasion from post-introduction data and a coupled reaction-diffusion-absorption model. Research Report, INRA, BioSP, Avignon.
- Bian, L. 2013. Spatial approaches to modeling dispersion of communicable diseases: a review. *Trans. GIS* 17:1-17.
- Bischof, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M. 2016. mlr: Machine Learning in R. *J. Mach. Learn. Res.* 17:5938-5942.
- Bosso, L., Russo, D., Di Febbraro, M., Cristinzio, G., and Zoina, A. 2016. Potential distribution of *Xylella fastidiosa* in Italy: A maximum entropy model. *Phytopathol. Mediterr.* 55:62-72.
- Breiman, L. 2001. Random forests. *Mach. Learn.* 45:5-32.
- Bühlmann, P., and Yu, B. 2003. Boosting with the L2 loss: Regression and classification. *J. Am. Stat. Assoc.* 98:324-339.
- Busby, J. (1991). Bioclim: A Bioclimate Analysis and Prediction System. Plant Protection Quarterly, Australia.
- Choat, B., Gambetta, G. A., Wada, H., Shackel, K. A., and Matthews, M. A. 2009. The effects of Pierce's disease on leaf and petiole hydraulic conductance in *Vitis vinifera* cv. Chardonnay. *Physiol. Plant.* 136:384-394.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20:37-46.
- Cornara, D., Bosco, D., and Fereres, A. 2018. *Philaenus spumarius*: When an old acquaintance becomes a new threat to European agriculture. *J. Pest Sci.* 91:957-972.
- Cunniffe, N. J., Cobb, R. C., Meentemeyer, R. K., Rizzo, D. M., and Gilligan, C. A. 2016. Modeling when, where, and how to manage a forest epidemic, motivated by sudden oak death in California. *Proc. Natl. Acad. Sci.* 113: 5640-5645.
- De Arruda, G. F., Barbieri, A. L., Rodríguez, P. M., Rodrigues, F. A., Moreno, Y., and da Fontoura Costa, L. 2014. Role of centrality for the identification of influential spreaders in complex networks. *Phys. Rev. E* 90:032812.
- Elith, J., and Leathwick, J. R. 2009. Species distribution models: Ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* 40:677-697.
- Ferrer, E., Alfonso, P., Ippoliti, C., Abeledo, M., Calistri, P., Blanco, P., Conte, A., Sánchez, B., Fonseca, O., and Percedo, M. 2014. Development of an active risk-based surveillance strategy for avian influenza in Cuba. *Prev. Vet. Med.* 116:161-167.
- Fick, S. E., and Hijmans, R. J. 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37:4302-4315.
- Franklin, J. 2010. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press.
- Friedman, J. H. 1991. Multivariate adaptive regression splines. *Ann. Stat.* 19: 1-67.
- Friedman, J. H. 2002. Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38:367-378.
- Garas, A., Schweitzer, F., and Havlin, S. 2012. A k-shell decomposition method for weighted networks. *New J. Phys.* 14:083030.
- Godefroid, M., Cruaud, A., Streito, J.-C., Rasplus, J.-Y., and Rossi, J.-P. 2018. Climate change and the potential distribution of *Xylella fastidiosa* in Europe. *bioRxiv* 289876.
- Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27:857-871.
- Herrera, J., Srinivasan, R., Brownstein, J., Galvani, A., and Meyers, L. A. 2016. Disease surveillance on complex social networks. *PLOS Comput. Biol.* 12:e1004928.
- Hyatt-Twynam, S. R., Parnell, S., Stutt, R. O., Gottwald, T. R., Gilligan, C. A., and Cunniffe, N. J. 2017. Risk-based management of invading plant disease. *New Phytol.* 214:1317-1329.
- Ibáñez, I., Silander, J. A., Jr., Wilson, A. M., LaFleur, N., Tanaka, N., and Tsuyama, I. 2009. Multivariate forecasts of potential distributions of invasive plant species. *Ecol. Appl.* 19:359-375.
- Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., and Rodes, I. 2017. Operational high-resolution land cover map production at the country scale using satellite image time series. *Remote Sens.* 9:95.
- Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. 2010. Identification of influential spreaders in complex networks. *Nat. Phys.* 6:888-893.
- Klement, E. P., Mesiar, R., and Pap, E. 2000. Triangular Norms. Springer, Netherlands.
- Lindström, T., Hakansson, N., and Wennergren, U. 2011. The shape of the spatial kernel and its implications for biological invasions in patchy environments. *Proc. Roy. Soc. London B: Biol. Sci.* 278:1564-1571.
- Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., and Zhou, T. 2016. Vital nodes identification in complex networks. *Phys. Rep.* 650:1-63.
- Martinetti, D., and Geniaux, G. 2017. Approximate likelihood estimation of spatial probit models. *Reg. Sci. Urban Econ.* 64:30-45.
- McElrone, A. J., Sherald, J. L., and Forseth, I. N. 2003. Interactive effects of water stress and xylem-limited bacterial infection on the water relations of a host vine. *J. Exp. Bot.* 54:419-430.
- Meentemeyer, R. K., Rank, N. E., Anacker, B. L., Rizzo, D. M., and Cushman, J. H. 2008. Influence of land-cover change on the spread of an invasive forest pathogen. *Ecol. Appl.* 18:159-171.
- Molnar, C. 2018. iml: Interpretable Machine Learning. R package version 0.5.1. <https://CRAN.R-project.org/package=iml>
- O'Donnell, M. S., and Ignizio, D. A. 2012. Bioclimatic predictors for supporting ecological applications in the conterminous United States. U.S. Geological Survey Data Series 691.
- Parnell, S., Gottwald, T. R., Riley, T., and van den Bosch, F. 2014. A generic risk-based surveying method for invading plant pathogens. *Ecol. Appl.* 24: 779-790.
- Parnell, S., van den Bosch, F., Gottwald, T., and Gilligan, C. A. 2017. Surveillance to inform control of emerging plant diseases: An epidemiological perspective. *Annu. Rev. Phytopathol.* 55:591-610.
- Purcell, A. 2013. Paradigms: Examples from the bacterium *Xylella fastidiosa*. *Annu. Rev. Phytopathol.* 51:339-356.
- Purse, B. V., and Golding, N. 2015. Tracking the distribution and impacts of diseases with biological records and distribution modelling. *Biol. J. Linn. Soc.* 115:664-677.
- Rimbaud, L., Dallot, S., Bruchou, C., Thoyer, S., Jacquot, E., Soubeyrand, S., and Thébaud, G. 2018. Heuristic optimisation of the management strategy of a plant epidemic using sequential sensitivity analyses. *bioRxiv* 315747.
- Saponari, M., Loconsole, G., Cornara, D., Yokomi, R. K., De Stradis, A., Boscia, D., Bosco, D., Martelli, G. P., Krugner, R., and Porcelli, F. 2014. Infectivity and transmission of *Xylella fastidiosa* by *Philaenus spumarius* (Hemiptera: Aphrophoridae) in Apulia, Italy. *J. Econ. Entomol.* 107:1316-1319.
- Sicard, A., Zeilinger, A. R., Vanhove, M., Schartel, T. E., Beal, D. J., Daugherty, M. P., and Almeida, R. P. 2018. *Xylella fastidiosa*: Insights into an emerging plant pathogen. *Annu. Rev. Phytopathol.* 56:181-202.
- Soubeyrand, S., de Jerphanion, P., Martin, O., Saussac, M., Manceau, C., Hendrikx, P., and Lannou, C. 2018. Inferring pathogen dynamics from temporal count data: The emergence of *Xylella fastidiosa* in France is probably not recent. *New Phytol.* 219:824-836.
- Strona, G., Carstens, C. J., and Beck, P. S. 2017. Network analysis reveals why *Xylella fastidiosa* will persist in Europe. *Sci. Rep.* 7:71.

- Thorne, E. T., Stevenson, J. F., Rost, T. L., Labavitch, J. M., and Matthews, M. A. 2006. Pierce's disease symptoms: Comparison with symptoms of water deficit and the impact of water deficits. *Am. J. Enol. Vitic.* 57: 1-11.
- Travençolo, B. A. N., and Costa, L. F. 2008. Accessibility in complex networks. *Phys. Lett. A* 373:89-95.
- Whist, A., Liland, K., Jonsson, M., Sæbø, S., Sviland, S., Østeras, O., Norström, M., and Hopp, P. 2014. Designing a risk-based surveillance program for *Mycobacterium avium* ssp. *paratuberculosis* in Norwegian dairy herds using multivariate statistical process control analysis. *J. Dairy Sci.* 97:6835-6849.
- White, S. M., Bullock, J. M., Hoofman, D. A., and Chapman, D. S. 2017. Modelling the spread and control of *Xylella fastidiosa* in the early stages of invasion in Apulia, Italy. *Biol. Invasions* 19:1825-1837.
- Zhang, J.-X., Chen, D.-B., Dong, Q., and Zhao, Z.-D. 2016. Identifying a set of influential spreaders in complex networks. *Sci. Rep.* 6:27823.