Estimating the spatial distribution of a plant disease epidemic from a sample

Stephen Parnell, Tim Gottwald, Mike Irey and Frank van den Bosch

First and forth authors: Centre for Mathematical and Computational Biology, Rothamsted Research, Harpenden, AL5 2JQ, United Kingdom; second author: US Horticultural Research Laboratory, US Department of Agriculture, 2001 South Rock Road, Ft. Pierce, Florida, 34945, USA; third author: US Sugar Corporation, Clewiston, Florida, 33440, USA.

Sampling is of central importance in plant pathology. It facilitates our understanding of how epidemics develop in space and time and can also be used to inform disease management decisions. Making inferences from a sample is necessary because we rarely have the resources to conduct a complete census of the population we are interested in i.e. we cannot assess the disease status of every plant in a field or of every field in a region. In plant pathology much progress has been made in estimating mean disease incidence from a sample (4). Important methods have also been developed to characterize the level of disease aggregation, by analyzing extra-binomial variation for example (4). However, often what is needed is a spatially explicit estimate of disease distribution in a population, i.e. a disease map, rather than an estimation of the mean and variation of a population (2). Surprisingly, this has received relatively little attention in plant pathology. However, a disease map is an essential tool in crop protection and allows for the spatially targeted deployment of control measures. This can minimize the costs involved in disease control via the optimal use of resources (1, 2).

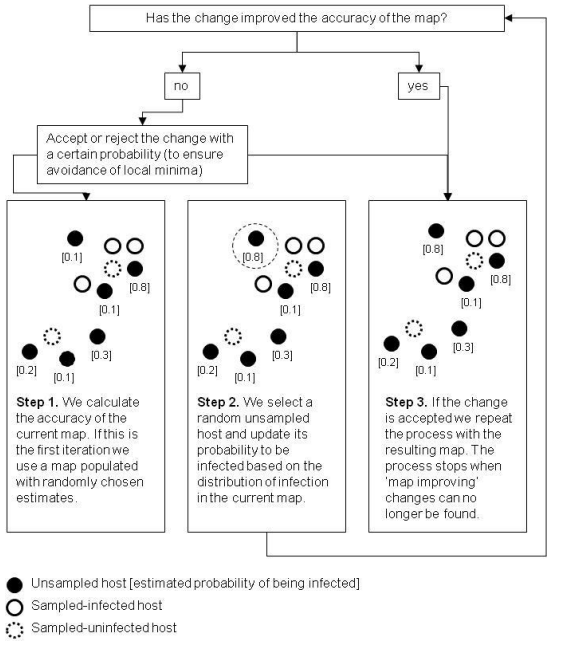

Fig. 1. Schematic of the estimation method.

In plant disease epidemiology existing approaches to disease mapping have come from the field of geostatistics (5). Geostatistical methods originate in mineral exploration and mining and involve estimating a surface map from a set of point samples. Although effective these methods assume a continuous variable and thus do not account for the spatial discontinuities that occur in the host distribution of a plant pathogen. Such spatial discontinuities occur when for example, there are missing trees in an orchard or significant distances between fields in a landscape and can have a significant effect on the temporal and spatial development of an epidemic. In this paper we describe a method to estimate the spatial distribution of a plant disease which accounts for the spatial structure of the host distribution. Additionally, we also account for the distance-dependent mechanisms by which real epidemics develop. We first describe the method (a schematic of which is also presented (Fig. 1)) and then apply the method to the distribution of an economically significant disease of citrus, Huanglongbing (Fig. 2). This is used as an example only and the method can be applied to any pathogen which exhibits distance-dependent patterns of spread.

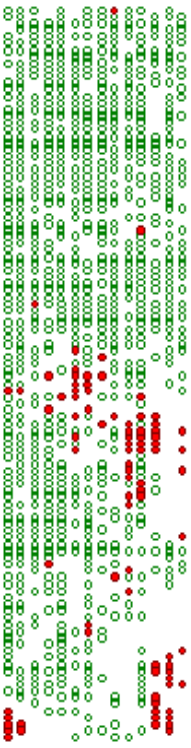

Fig. 2. A planting of commercial citrus trees in south Florida. Solid circles denote trees infected with Huanglongbing disease, clear circles denote uninfected trees. Data provided by M. Irey, US Sugar Corporation.

The method uses information on disease status at sampled host locations to estimate disease status at unsampled locations (where a 'host location' could be an individual plant, tree, field or host patch etc depending on the pathogen

and area being sampled). The final map consists of probabilities of infection [0,1] for all unsampled host locations. Central to the method is the need for a quantifiable indicator which describes the accuracy of the estimated map based on what is known at sampled locations. The derivation of such an indicator is given in box 1. Unless otherwise stated, when we refer to map accuracy we mean the accuracy to estimate disease status at sampled locations. The true accuracy of the map (i.e. the accuracy of the estimates of infection probability at unsampled locations) will obviously not be known in practice and so cannot be used to guide the method. However, we test the method using a hypothetical sample from a host distribution where disease status is fully known and can therefore be used to assess the accuracy of the method.

Box 1. Determining the accuracy of an estimated map based on what is known at sampled locations. Assuming we have already estimated disease status at unsampled locations we can then calculate the probability that each sampled site is infected given this estimated map. There are N locations in total, for each sampled location the distance dependent probability that it is diseased is determined using the estimate of disease status (if sampled) or infection probability (if unsampled) at all other locations, j.

$$P(\text{sample } i \text{ is diseased}) = \boldsymbol{\theta} \sum_{j \in N} \boldsymbol{exp}\left(-\boldsymbol{\mu} \boldsymbol{d_{ij}}\right)$$

Where d_{ij} is the distance between sampled location i and location j and θ and μ are epidemiological parameters. Next we use this to calculate the *deviance* which is our quantifiable accuracy indicator. The deviance is a measure of the difference between the estimate of disease status at each sampled site, P_i, and the actual disease at each sampled site (where 1 is diseased, 0 is non-diseased). An equally sized difference further from the true disease status is weighted more than one closer to the true disease status. Once summed over all sampled locations we have a single descriptor of map accuracy. The closer the estimated probability, P_i, to the actual disease status the smaller the contribution to the deviance. Therefore, the smaller the deviance the better the overall estimated map as measured against the information we have from the sampled locations.

The method begins by assigning random probabilities of infection to all unsampled host locations. Next, we randomly select an unsampled host location and update its probability of being infected. This is a function of the distance to all other host locations and their estimated infection probability (if unsampled) or disease status (if sampled; 1 if infected, 0 otherwise). This change is accepted or rejected depending on its effect on map accuracy (see box 1). If map accuracy is improved then the change is accepted. If map accuracy is not improved then the change is stochastically accepted or rejected with a certain probability in order to avoid the problem of local minima. This latter step utilizes a simulated annealing optimization algorithm. This process continues until no more changes can be found which significantly improve the accuracy of the map. This is defined by a stopping criterion which utilizes the gradient of map accuracy over a sequence of iterations. The method is repeated for multiple sets of the epidemiological parameters and the set which results in the map with the highest accuracy is used to generate the final map. The method therefore requires no prerequisite knowledge of the epidemiology of the disease and estimates all parameters internally. The only inputs are the locations of the hosts and the disease status of sampled hosts.

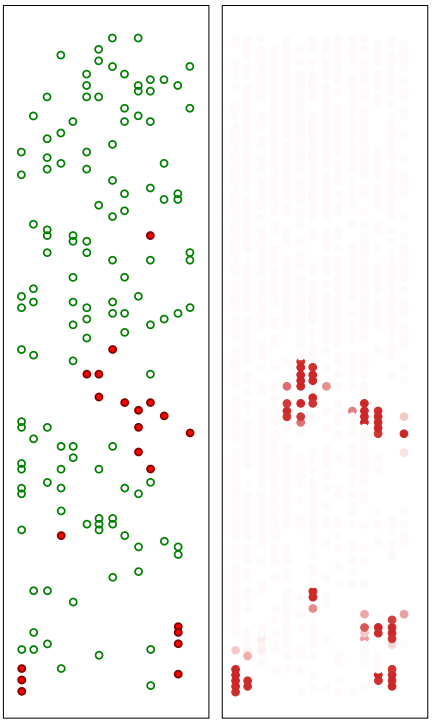

Fig. 3. The above plots relate to figure 2. The left plot is a simulated random sample of 15%. The right plot is the estimated map produced on the basis of this sample. The probability of infection increases with the shading of the circles.

We test the method against a dataset describing the distribution of Huanglongbing disease in a commercial planting of citrus in south Florida (Fig. 2). Huanglongbing is an economically significant bacterial disease of citrus spread by a psyllid vector (3). The planting contains approximately 1000 trees (Fig. 2). Discontinuities or voids in the host distribution can be seen which represent trees that have been previously removed from the planting (Fig. 2). We simulate a random sample of 15% from this host population and use this to generate our estimated map (Fig. 3). The accuracy of the estimated map can be assessed by comparison of the actual map (Fig. 2). The method captures the main aggregates of disease distribution well (Fig. 3). We are currently working on statistical descriptors to quantify this. One method is to extend the use of the deviance to unsampled locations as well as sampled locations. However, to reiterate, in practice the actual map will not be known. We are also exploring how sample size and sample placement effect map accuracy. Important questions under investigation include: how many samples are required in order to generate a map of sufficient accuracy for disease management decisions? Should samples be randomly positioned in space or clustered in order to maximize map accuracy? Additionally the method is being extended to a grid-based approach which allows for the situation whereby the location of all host individuals is not precisely known but there is some information on the density of hosts at a given spatial resolution.

Literature Cited

1. Dybiec, B., Kleczkowski, A., and Gilligan, C. A. 2004. Controlling disease spread on networks with

incomplete knowledge. Physical Review E 70:art. no.-066145.
2. Fleischer, S. J., Blom, P. E., and Weisz, R. 1999. Sampling in precision IPM: When the objective is a map. Phytopathology 89:1112-1118.
3. Gottwald, T. R., da Graca, J. V., and Bassanezi, R. B. 2007. Citrus Huanglongbing: the pathogen and its impact. Published online in Plant Health Progress doi: 10.1094/PHP-2007-0906-01-RV.
4. Madden, L. V., and Hughes, G. 1999. Sampling for plant disease incidence. Phytopathology 89:1088-1103.
5. Nelson, M. R., Orum, T. V., Jaime-Garcia, R., and Nadeem, A. 1999. Applications of geographic information systems and geostatistics in plant disease epidemiology and management. Plant Disease 83:308-319.