ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Suprunenko YF, Gilligan CA. 2025 Where to refine spatial data to improve accuracy in crop disease modelling: an analytical approach with examples for cassava. *R. Soc. Open Sci.* **12**: 250012. https://doi.org/10.1098/rsos.250012

Received: 6 January 2025
Accepted: 14 April 2025

Subject Category:
Ecology, conservation and global change biology

Subject Areas:
health and disease and epidemiology, mathematical modelling, ecology

Keywords:
crop landscape, epidemic invasion, epidemiological model, spatially explicit individual-based model, analytical approximation, infection rate

Author for correspondence:
Yevhen F. Suprunenko
e-mails: ys526@cam.ac.uk; yevhen.suprunenko@gmail.com

THE ROYAL SOCIETY
PUBLISHING

Where to refine spatial data to improve accuracy in crop disease modelling: an analytical approach with examples for cassava

Yevhen F. Suprunenko and Christopher A. Gilligan

Department of Plant Sciences, University of Cambridge, Cambridge, UK

YFS, 0000-0001-5927-7571

Epidemiological modelling plays an important role in global food security by informing strategies for the control and management of invasion and spread of crop diseases. However, the underlying data on spatial locations of host crops that are susceptible to a pathogen are often incomplete and inaccurate, thus reducing the accuracy of model predictions. Obtaining and refining datasets that fully represent a host landscape across territories can be a major challenge when predicting disease outbreaks. Therefore, it would be an advantage to prioritize areas in which data refinement efforts should be directed to improve the accuracy of epidemic prediction. In this paper, we present an analytical method to identify areas where potential errors in mapped host data would have the largest impact on modelled pathogen invasion and short-term spread. The method is based on an analytical approximation for the rate at which susceptible host crops become infected at the start of an epidemic. We show how implementing spatial prioritization for data refinement in a cassava-growing region in sub-Saharan Africa could be an effective means for improving accuracy when modelling the dispersal and spread of the crop pathogen cassava brown streak virus.

1. Introduction

A major challenge when modelling epidemics of crop disease is the lack of complete and accurate maps of the distribution of susceptible crops across a landscape [1]. The challenge is especially relevant when addressing epidemic threats to food production in climate-vulnerable regions. For example, in Ethiopia, the absence of high-quality crop type maps is a challenge for modelling the unfolding wheat rust epidemic [2,3] (but see the recently provided Ethiopian Crop Type 2020 dataset [4] for a single 2020/21 cropping season). In sub-Saharan Africa, the modelling of the expanding epidemic of cassava brown streak disease [5] may be sensitive to local spatial uncertainty in crop distribution [6], and therefore the quality of model predictions relies on the quality of the available data on the spatial distribution of cassava [7]. The problem of using incomplete or inaccurate crop maps is that they can amplify uncertainties in modelling the extent and rate of disease spread by (i) either under- or over-estimating the crop area as well as (ii) failing to account for the spatial structure of crops (e.g. clustering of crops often at multiple scales with concomitant crop-free gaps). In cases when it is impossible to rectify incomplete and inaccurate data, models and methods that use partial data have been proposed (e.g. in modelling the spread of livestock diseases when exact spatial locations and clustering of livestock holdings are unknown; see [8–12]). However, additional data collection is sometimes possible (e.g. by ground-based methods or via remote sensing [13–15]), and this could be used to refine incompletely mapped data with the purpose of improving the accuracy of epidemic modelling. The challenge then becomes how to use additional resources for data refinement most efficiently.

In this paper, we address the problem of finding geographic regions where new data on host crop distribution would lead to the strongest improvement of the accuracy in modelling the spread of crop diseases. We aim to develop a rapid analytical assessment method to estimate the impact of potential inaccuracies in crop datasets on the resulting predictions—derived from computer simulations of an individual-based model (IBM)—on epidemic spread. Analytical approximations have been used to address the problems of incomplete host maps in studies of humans and livestock by mapping local changes in the basic reproduction number R_0 to assess regions at risk [16] and to design and implement vaccination strategies [17]. Sellman *et al.* [12] also used local estimates of R_0 to assess the effects of clustered disaggregation of county-scale livestock premises on epidemic predictions. Suprunenko *et al.* [18] developed an alternative approach based upon the infection rate, r, to predict the impact of the spatial structure of a crop landscape on epidemic dynamics of crop disease. The method involved the calculation of the infection rate r, which has the advantage of encompassing susceptible–infected (SI) epidemics for which R_0 is undefined, while also applying to SIR epidemics, where $R_0 = r/\mu$ and μ is the removal (or recovery) rate of infected fields.

We now use the approach of Suprunenko *et al.* [18] to develop a method of spatial prioritization of areas for new data. Our focus is on crop disease, but the approach has wider applicability. For example, whereas the total number of livestock premises was known in the analyses of Sellman *et al.* [12], our method allows for uncertainty in the numbers of susceptible crops (analogous to livestock premises) across a landscape through which a pathogen is spreading.

We illustrate the application of the method to the analysis of the spread of cassava brown streak virus (CBSV) in an arbitrarily selected cassava production region in sub-Saharan Africa. Cassava is one of the most important staple food crops. An estimated 800 million people in Africa rely on cassava for their primary calorific intake [19]. Cassava production in sub-Saharan Africa has come under increasing pressure due to the rapidly expanding range of CBSV [20–23]. Recent work has provided improved maps of cassava production in sub-Saharan Africa [7] that were then used in combination with a parametrized epidemic model to predict spread and arrival times of CBSV throughout the region [5]. In this paper, we show how to identify local areas within the host crop (cassava) landscape where potential errors in maps of cassava production would impact most strongly on predictions of CBSV spread. The analyses show that surveying and refining data in certain areas substantially improve accuracy in predictions of the epidemic spread model. Hence, correcting for insufficient or inaccurate datasets could provide significant improvement to the preparedness of regions for the ongoing expansion of CBSV.

2. Material and methods

2.1. Cassava data

The map with currently the highest spatial resolution for cassava production in sub-Saharan Africa is known as CassavaMap, derived by Szyniszewska [7]. We used the cassava production layer from CassavaMap [7] and converted it to discrete fields of cassava as detailed in [5] (i.e. one raster 1 km-by-1 km cell can have a maximum of 1000 identical cassava fields) to create a rasterized map that describes

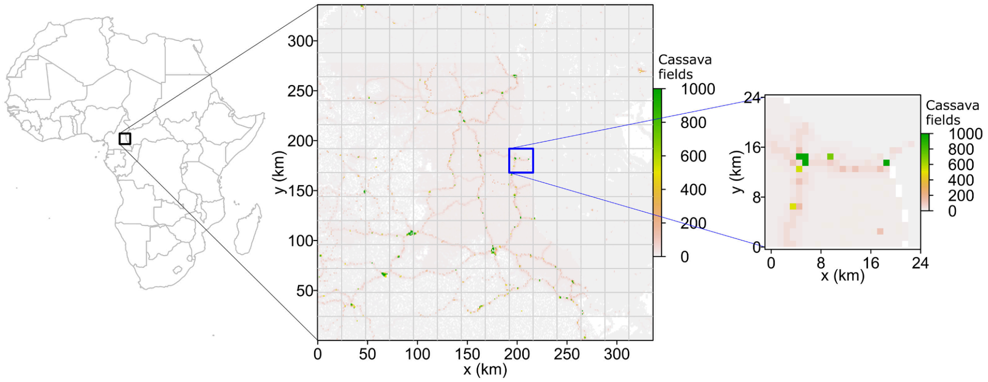


Figure 1. Host landscape. Data (as raster with spatial resolution 1 km by 1 km) on spatial distribution of cassava production extracted from CassavaMap [7] and converted to fields [5] (see §2); here, we show the entire 336 km-by-336 km landscape H (available from Figshare [24]). Raster cells with zero values and those with no data are both displayed in white: grey values (from the raster) indicate low values. We resolved the 336 km-by-336 km landscape into a square lattice with a mesh size of 24 km. For convenience, the 24 km-by-24 km area outlined by the blue boundary is used in figure 2 to illustrate the analyses that were conducted on the entire 336 km-by-336 km landscape.

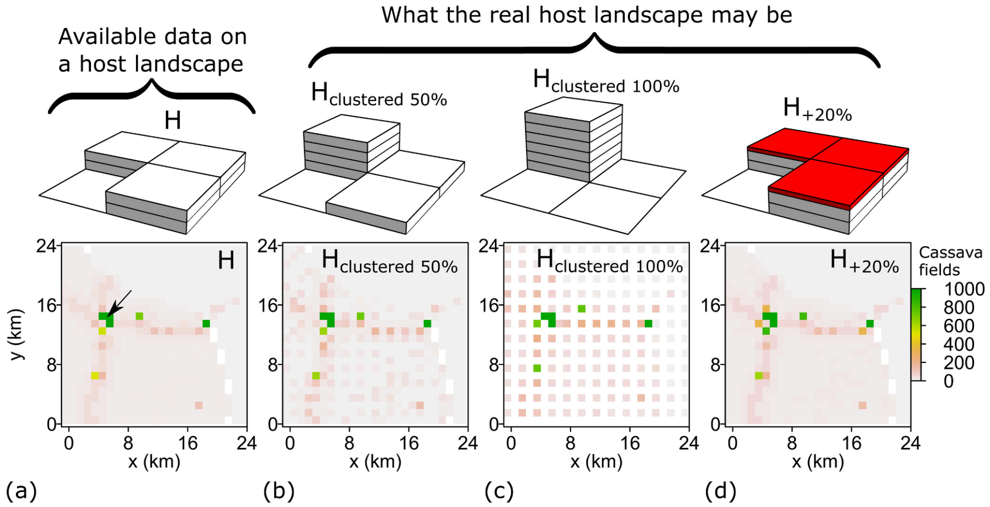


Figure 2. Host data and potential errors. (a) The landscape H represents the available data extracted from CassavaMap (figure 1). The small sample area shown here is the same area outlined by the blue boundary in figure 1. A black arrow in H denotes the location of a single initially infected cassava field (the same for all landscapes). (b–d) Different scenarios from H to account for potential errors in the mapped cassava landscape, illustrated on a 2 km-by-2 km area shown above each landscape (see also §2). Each tile represents a fixed number of fields, and the red tiles in (d) represent +20% in addition to two tiles underneath. In all panels, zero values are displayed in white.

the spatial distribution of cassava fields in sub-Saharan Africa. For brevity, hereafter we refer to this map as cassava (or, equivalently, as host or crop) data. A 336 × 336 km^2 area was arbitrarily selected on the border between Cameroon and the Central African Republic. The data describing this area were denoted as a host landscape H [24] (figure 1). As shown in figure 1, the cassava landscape is heterogeneous, with the spatial distribution of cassava fields driven by the spatial distribution of the human population.

2.2. Potential inaccuracies in cassava data

Here, we briefly summarize likely inaccuracies in CassavaMap that stem from inaccuracies in the underlying data, as reported by Szyniszewska [7]. The map was derived using administrative unit-level data on cassava production and harvested area from 32 countries [7], and disaggregating the data according to a rasterized human population density model, LandScan 2014 [25]. CassavaMap is available at a resolution of approximately 1 km by 1 km. According to Szyniszewska [7], errors in the spatial distribution in [25] or in production statistics would result in errors in CassavaMap [7]. Moreover, the size of the human population, as well as cassava production in the current year, may be very different from the size in 2014 [25], thus affecting the estimate of the area and the number of cassava fields. Lastly, due to the coarse administrative granularity of input data on cassava production (e.g. the data from some countries were available only on a national scale [7]), some areas in CassavaMap are likely to have cassava production distributed in a flatter manner than in reality because the spatial variation of cassava production *per capita* within administrative regions is absent in the input data. Recent work by Hassall *et al.* [6] has raised some additional challenges in mapping cassava at landscape and regional scales: for example, Hassall *et al*. [6] showed that estimated data (CassavaMap) were 'able to capture large-scale regional trends in cassava production but fail to capture the local variation' [6]. Here, in the absence of further improvements or detailed mapping of extensive regions, we treat CassavaMap as the best available starting point to illustrate the application of our method below.

For the purpose of this paper, we examine two out of several possible scenarios related to the accuracy of crop maps. First, where the actual spatial distribution of cassava fields is more clustered than presented in CassavaMap. To illustrate inaccuracies of this type, we used a previously established method [18] as one of the simplest examples of aggregating hosts locally: we considered the cassava landscape on a square lattice with a mesh size of 2 km and assumed that cassava fields can be aggregated within each 2 km-by-2 km cell.

The second scenario represents inaccuracy where the distribution of cassava fields was underestimated in CassavaMap due to changing human populations over time. We compared the human population distribution model [25], used in CassavaMap and its most recent version, LandScan 2022 [26]. The comparison showed [27] that the total human population in the entire territory of sub-Saharan Africa used in CassavaMap [7] increased by approximately 35% from 2014 to 2022; while the human population in a smaller territory located, for example, between 13 and 15 degrees of longitude and between 4 and 6 degrees of latitude (roughly the area considered in this paper), increased by approximately 17% from 2014 to 2022. Hence, rounding up and considering a linear relationship between human population and the area of cassava fields, we assumed that the real area of cassava fields is +20% higher than recorded in CassavaMap.

To illustrate what the actual host landscape might be, we introduced additional scenarios into the original landscape H (figure 2a) and created the following three alternative landscapes as surrogates for the actual landscape:

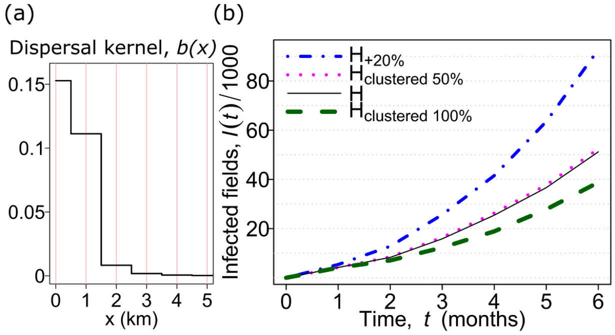
(1) $H_{\text{clustered 50\%}}$—To obtain the landscape denoted as '$H_{\text{clustered 50\%}}$', we considered the original landscape H on a 2 km lattice, and in each 2 km-by-2 km cell, 50% of all cassava fields were first removed and then placed onto a northwestern 1 km-by-1 km original raster cell. For an illustration of this first step, see the 2 km-by-2 km area shown above a landscape in figure 2a, where tiles denote the amount of cassava fields; in particular, removing 50% of cassava fields from the 2 km-by-2 km area (figure 2a) and then placing the removed fields onto a northwestern 1 km-by-1 km original raster cell produces a 2 km-by-2 km area shown above the landscape in figure 2b. If the maximal capacity of a northwestern 1 km-by-1 km original raster cell was reached, then the excess fields were placed onto a northeastern cell until its capacity was reached, followed by southwestern and finally southeastern cells (see figure 2b). All fully occupied cells in the original unmodified landscape H were assumed to remain fully occupied.

(2) $H_{\text{clustered 100\%}}$—To obtain the landscape denoted as '$H_{\text{clustered 100\%}}$', we followed the same procedure as for $H_{\text{clustered 50\%}}$ but using all cassava fields (i.e. 100% instead of 50%) in each 2 km-by-2 km cell (see figure 2c).

(3) $H_{+20\%}$—We obtained the landscape denoted as '$H_{+20\%}$' by adding 20% to the number of cassava fields present in each original 1 km-by-1 km raster cell. If the maximum number of fields in any single 1 km-by-1 km cell was reached because of adding new fields, then the excess of fields was distributed among eight nearest neighbouring cells by populating first the most occupied cell up to the maximum level, then the remaining most occupied cell, and so on (see figure 2d).

2.3. The model of pathogen invasion and spread

The raster-based compartmental epidemiological SI model of CBSV invasion and spread in sub-Saharan Africa was formulated in [5] and fitted to the data for CBSV spread in Uganda [5]. The model [5] used a raster-based power-law dispersal kernel that incorporated pathogen short-distance dispersal



(c)

Host landscape	Infected fields $I(6)/1000$ mean; median [5%; 95% percentiles]	$rN/10^5$
$\mathrm{H}_{+20\%}$	92.7; 93.5 [58.7; 126.0]	9.15
$\mathrm{H}_{\text{clustered }50\%}$	52.3; 48.6 [35.8; 78.0]	6.98
H	51.2; 48.0 [34.1; 76.0]	6.86
$\mathrm{H}_{\text{clustered }100\%}$	38.8; 35.8 [25.7; 61.2]	6.70

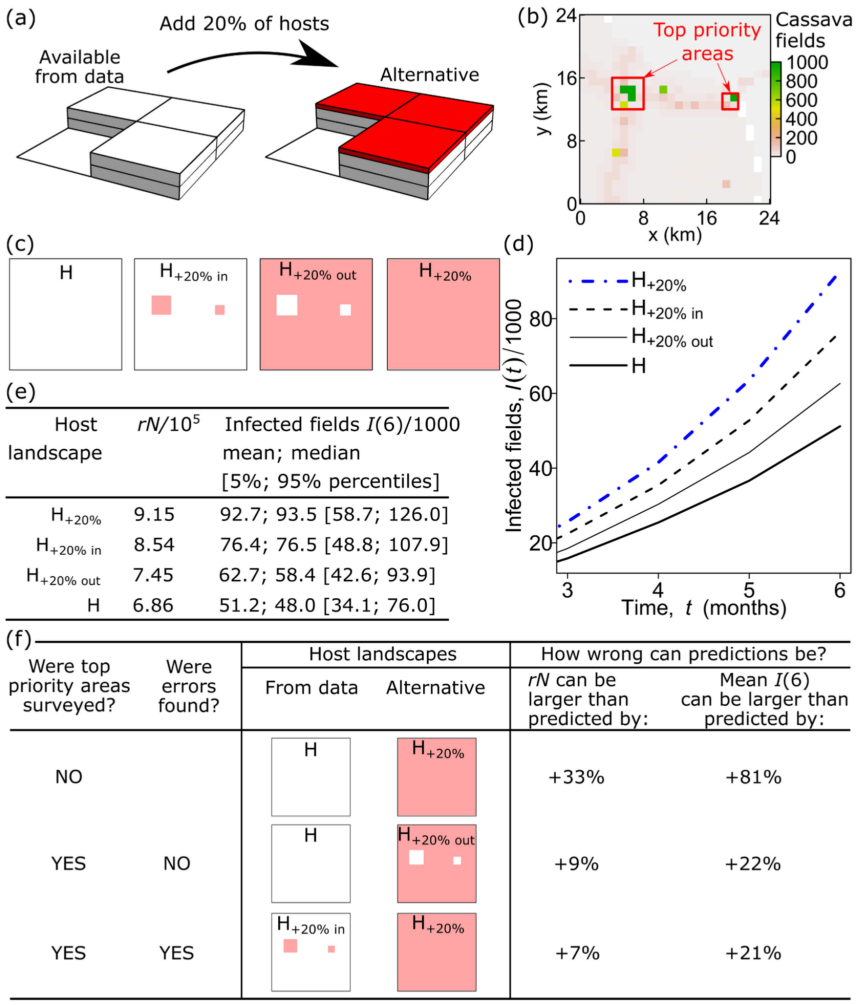
Figure 3. Impact of potential errors in host data on results of epidemic modelling. (a) Dispersal kernel $b(x)$ of a pathogen (CBSV), sampled from results of parameter estimation by Godding *et al.* [5] (see §2). (b) Mean number $I(t)$ of infected fields of cassava at $t = 6$ months after the start of an epidemic obtained from 1000 computer simulations for each landscape. (c) Mean, median and percentiles of results of computer simulations shown together with estimates of infection rate r from equation (2.1) multiplied by the total number N of cassava fields estimated for corresponding 24 km-by-24 km areas from figure 1.

by an insect vector and long-distance dispersal by movement of virus-infected materials used for planting [5]. The power-law dispersal kernel was characterized [5] by the exponent α, the proportion p of dispersed inoculum that remains within the source (1 km-by-1 km) cell, and the infection rate per contact density β. The contact density refers to the spatial density of susceptible cassava fields that are contacted by an initial infected field via dispersal kernel $b(x)$. Here, using the approach of [18,28], we reformulated the raster-based model from Godding *et al.* [5] as an IBM with a staircase-like radially symmetric dispersal kernel (i.e. a radially symmetric version of the raster-based $b(x)$ from Godding *et al.* [5]). The IBM describes the spread of infection among discrete fields of cassava (treated as individual hosts located at the centroid of corresponding raster cells, as in [18,28]). An infected field infects susceptible fields at a distance x with a rate given by the product of the parameter β and a rotationally symmetric dispersal kernel, $b(x)$. We applied the power-law dispersal kernel and selected the following parameter values from the posterior distribution from Godding *et al.* [5]: $\alpha = 3.75$, $p = 0.12$ and $\beta = 10^{-3} \times \exp(6)$ (see figure 3a); note that the parameter β in this paper is 1000 times smaller than β in [5] because here it is measured in units (number of fields per area)$^{-1}$ × time^{-1} instead of units (number of raster cells per area)$^{-1}$ × time^{-1} used in [5]. We considered CBSV invasion and spread over the four landscapes described above, i.e. H, $\mathrm{H}_{\text{clustered }50\%}$, $\mathrm{H}_{\text{clustered }100\%}$ and $\mathrm{H}_{+20\%}$. Epidemics were simulated using the `ModelSimulator` software presented by Cornell *et al.* [29].

2.4. Estimating the impact of potential inaccuracies in crop data on the invasion of a pathogen

To determine the impact of potential inaccuracies in the cassava landscape on pathogen invasion using computer simulations, we considered the difference between the mean number of infected fields, $I(t)$, at $t = 6$ months after the start of an epidemic in the alternative landscape, H_{Alt}, and the (360 km-by-360 km) landscape H derived from CassavaMap, i.e. $I_{\mathrm{H}_{Alt}}(6) - I_{\mathrm{H}}(6)$, where the alternative landscape is one of the following: $\mathrm{H}_{\text{clustered }50\%}$, $\mathrm{H}_{\text{clustered }100\%}$ or $\mathrm{H}_{+20\%}$ (figure 3b,c). For simplicity, in all cases, the same initial location of primary infection is used (figure 2a).

To determine the impact of potential inaccuracies on CBSV invasion analytically, we used an approximation [18] for the infection rate at which susceptible fields (i.e. individual hosts in an IBM) become infected at the start of an epidemic:



Host landscape	$rN/10^5$	Infected fields $I(6)/1000$ mean; median [5%; 95% percentiles]
$H_{+20\%}$	9.15	92.7; 93.5 [58.7; 126.0]
$H_{+20\%\text{ in}}$	8.54	76.4; 76.5 [48.8; 107.9]
$H_{+20\%\text{ out}}$	7.45	62.7; 58.4 [42.6; 93.9]
H	6.86	51.2; 48.0 [34.1; 76.0]

Were top priority areas surveyed?	Were errors found?	Host landscapes		How wrong can predictions be?	
		From data	Alternative	rN can be larger than predicted by:	Mean $I(6)$ can be larger than predicted by:
NO		H	$H_{+20\%}$	+33%	+81%
YES	NO	H	$H_{+20\%\text{ out}}$	+9%	+22%
YES	YES	$H_{+20\%\text{ in}}$	$H_{+20\%}$	+7%	+21%

Figure 4. Finding areas where spatial host data should be refined to improve accuracy in crop disease modelling. (a) Presume that the alternative landscape, $H_{+20\%}$, is a realistic surrogate for the actual landscape (cf. figure 1d), i.e. $H_{+20\%}$ has +20% more cassava fields than landscape H extracted from available data (see §2 for details). (b) Spatial prioritization for data refinement: in each 24 km-by-24 km cell within the entire landscape, we identified five 2 km-by-2 km cells with the highest impact on epidemic spread and denoted them as 'top priority areas', outlined by the red boundary. A small sample area shown here is the same area outlined by the blue boundary in figure 1. (c) The four landscapes in which no fields are added (H); fields are added inside ($H_{+20\%\text{ in}}$) or outside top priority areas ($H_{+20\%\text{ out}}$) or across the domain ($H_{+20\%}$), subject to a maximum increase of 20% over the default map. (d) Mean number of infected fields obtained from 1000 computer simulations. (e) Numerical values for mean, median and percentiles of infected fields, together with estimates of quantities rN for each landscape. (f) The effect of spatial prioritization and subsequent data refinement within identified top priority areas on the accuracy of epidemic model predictions (see §3 for details). The computer code and data, including entire landscapes and the map of top priority areas used in this work, are available from Figshare [24,27].

$$r = \beta n_S \times \left(1 + \frac{1}{n^2} \int_0^{\sqrt{A}/2} 2\pi x b(x) g(x)\, \mathrm{d}x\right). \tag{2.1}$$

Here, n_S (or n) represents the spatial density of susceptible (or susceptible and infected) fields; β is the infection rate per contact density and $b(x)$ is a pathogen dispersal kernel; $g(x)$ is the spatial autocovariance of fields containing the host crop across the landscape of interest, A, and takes values on the interval $x \in (0, \sqrt{A}/2)$; $g(x)$ is also known as a second-order spatial cumulant (e.g. [29]). Equation (2.1) provides a spatially local estimate of the infection rate, r, on the local square area A. Denoting the total number of fields within the area A as N, we have $n = N/A$. We assumed that a pathogen was introduced by a single infected field randomly selected from all susceptible fields within area A; therefore, $n_S = (N-1)/A$. Note that equation (2.1) can be used for estimates of the basic reproduction number R_0 in the corresponding SIR compartmental epidemiological model (e.g. [18,30]): R_0 can be approximated as the ratio of r and μ, i.e. $R_0 = r/\mu$, where μ is the removal or recovery rate of infected fields.

We compared the impact of alternative landscapes on epidemic invasion using the difference between rN in an alternative landscape ($(rN)_{\mathrm{H}_{Alt}}$) and in landscape H ($(rN)_{\mathrm{H}}$). Positive (negative) values of this difference, $(rN)_{\mathrm{H}_{Alt}} - (rN)_{\mathrm{H}}$, mean that an epidemic on the alternative (i.e. presumed real) landscape on average is larger (smaller) than an epidemic modelled on the landscape H.

As the quantity rN is measured locally, to select the size of the local area to estimate rN, we used the results of Suprunenko *et al.* [18] where for the dispersal kernel $b(x)$ (cf. figure 3a), it was shown that the calculation of r on an area $A = 8 \times 8$ km^2 (an area determined by the spatial scale of the pathogen dispersal kernel; see [18]) provides a better agreement with estimates from computer simulations than on a larger area, $A = 24 \times 24$ km^2 (an arbitrarily selected larger area considered in [18]). Therefore, we considered a host landscape on a square lattice with a mesh size of 8 km for which we estimated values of n_S, n, $g(x)$, r and N in each 8 km-by-8 km cell.

2.5. Finding areas where inaccuracies in host data have the strongest impact on the modelled epidemic invasion

Here, we aim to construct a spatially resolved map of the degree of impact of potential errors in the host data on modelled epidemics. To achieve the aim using computer simulations would involve time-consuming efforts requiring a large number of computer simulations using different initial conditions and different implementations of potential errors. Instead, we adapt the analytical approach (based on equation (2.1)) to identify spatial reconfigurations of a host landscape that provide the strongest deceleration of an invading pathogen [18]. First, we assumed that $\mathrm{H}_{+20\%}$ (figure 4a) is a more realistic surrogate of the actual landscape than H (i.e. it is a more accurate landscape than H due to the increase in human population since 2014, when the data for CassavaMap were collated [7]), and that the effect of changing the host landscape could be shown on a square lattice with a mesh size of 2 km. We divided each 8 km-by-8 km cell (on the lattice used to calculate local r) into 2 km-by-2 km cells. Calculating the product of the local infection rate r (calculated on 8 km-by-8 km cells as described above) and the number $N_{2\mathrm{km}}$ of hosts in each 2 km-by-2 km cell on the $\mathrm{H}_{+20\%}$ and H landscapes, we obtained the rasterized map of values $(rN_{2\mathrm{km}})_{\mathrm{H}_{+20\%}} - (rN_{2\mathrm{km}})_{\mathrm{H}}$. The areas with the largest values of the difference $(rN_{2\mathrm{km}})_{\mathrm{H}_{+20\%}} - (rN_{2\mathrm{km}})_{\mathrm{H}}$ are areas where the selected inaccuracy has the strongest impact on an epidemic invasion.

2.6. Comparing the analytical solution with computer simulations

We used computer simulations to check the effect on epidemic dynamics from a change in the landscape due to spatial prioritization provided by the analytical solution. We considered the situation when the landscape was divided into relatively small regions (analogous to local administrative regions) where spatial prioritization was based on the data within that region independently from other local regions. Therefore, we considered the entire landscape on a square lattice with a mesh size of 24 km, within the overall domain of 336 km by 336 km, and within each 24 km-by-24 km cell, we used the algorithm presented above to select the top five 2 km-by-2 km cells with the largest values of the difference $(rN_{2\mathrm{km}})_{\mathrm{H}_{+20\%}} - (rN_{2\mathrm{km}})_{\mathrm{H}}$. We refer to selected cells as 'top priority areas' for data refinement (figure 4b shows top priority areas within a single 24 km-by-24 km cell as an example; top priority areas identified simultaneously in all 24 km-by-24 km cells within the entire landscape are available from Figshare [24]). Next, we constructed two new landscapes in addition to landscapes H and $\mathrm{H}_{+20\%}$: (i) the landscape denoted '$\mathrm{H}_{+20\%,\ \mathrm{in}}$' was obtained from H by introducing +20% of hosts *only inside* the selected five 2 km-by-2 km cells *in each* 24 km-by-24 km area; (ii) the landscape denoted '$\mathrm{H}_{+20\%,\ \mathrm{out}}$' was obtained from H by introducing +20% of hosts in all cells *only outside* the selected five cells *in each* 24 km-by-24 km area; the difference between the four landscapes on a sample 24 km-by-24 km area is illustrated in figure 4c. If the area with the highest impact is identified correctly, then it is expected that the epidemic on landscape $\mathrm{H}_{+20\%,\ \mathrm{in}}$ should be closer (in terms of the mean number of infected fields) to the epidemic on landscape $\mathrm{H}_{+20\%}$ than on landscape H; similarly, the epidemic on $\mathrm{H}_{+20\%,\ \mathrm{out}}$ should be closer to the epidemic on H than on $\mathrm{H}_{+20\%}$. Using computer simulations of the IBM together with analytical estimates described above, we compared the characteristics of an epidemic invasion on the four landscapes described in this section (figure 4d,e).

3. Results

Computer simulations of the IBM of a potential CBSV epidemic using the selected dispersal kernel (figure 3a) demonstrated that potential inaccuracies in the host data (figure 2) can influence epidemic invasion strongly. In particular, we found that the mean number of infected cassava fields could be 24% fewer on a more clustered landscape compared with the H landscape calculated from CassavaMap (i.e. −24% in $H_{\text{clustered }100\%}$ relative to H; figure 3b,c). By contrast, up to 81% more fields were infected compared with H when allowance was made for up to 20% missing fields on CassavaMap (i.e. +81% in $H_{+20\%}$ relative to H; figure 3b,c). Recall that the alternative landscapes are designed to allow for inaccuracies in H: $H_{\text{clustered }100\%}$ allows for failure to account for clustering if fields were over-dispersed among CassavaMap cells from coarse production data; the $H_{+20\%}$ landscape allows for growth in human population density in the decade since the raw data for CassavaMap were collated.

We used the product, *rN*, of the local infection rate multiplied by the local number of hosts as an analytical approximation to map the impact from inaccuracies in the host data on the modelled epidemic dynamics. Geographic areas with the largest impact were indicated as priorities for data refinement efforts (i.e. where additional data on the spatial distribution of fields of susceptible crops would be valuable).

To verify the analytical method of spatial prioritization, we tested the effect of data refinement in top-priority areas. To compare the impact of inaccuracies in the identified top priority areas and in the rest of the landscape, we constructed two additional landscapes where potential inaccuracies were adjusted for by changing the number of host fields either only inside or only outside those top priority areas (figure 4c). Calculated values of *rN* and computer simulations of CBSV spread confirmed that the impact of inaccuracies in the selected top priority area was indeed stronger than the impact from the rest of the landscape (figure 4d,e).

Finally, we estimated how the accuracy of epidemic model predictions would change because of spatial prioritization and subsequent data refinement within the top priority areas. Using the example studied in figure 4a–e, we focused on the following three cases (figure 4f):

(i) In the first case, we assumed no additional data were collected. Since we lacked information about the actual landscape and had no additional data in this case, we assumed the maximum deviation. Consequently, we assumed that the actual landscape had +20% more cassava fields than landscape H. Thus, $H_{+20\%}$ was considered the most realistic surrogate for the actual landscape. In this case, the potential inaccuracy of the model predictions was +81% in the alternative landscape $H_{+20\%}$ as compared with the landscape H extracted from the data (CassavaMap), i.e. $100\% \times (I_{H_{+20\%}}(t) - I_H(t))/I_H(t) \approx 81\%$ at $t = 6$ months, where $I(t)$ is the mean number of infected fields as inferred from computer simulations of the IBM.

(ii) In the second case, we assumed that the top priority areas were identified, and the subsequent survey of the top priority areas showed no errors in the landscape H in those areas. Assuming the maximum deviation, given the unknown actual landscape outside the surveyed areas, $H_{+20\%\text{ out}}$ was considered the most realistic surrogate for the actual landscape, i.e. inaccuracies could be present only outside the identified areas. In that case, the mean number of infected fields, $I(6)$, was +22% larger than model predictions based on the landscape H.

(iii) In the third case, we assumed that the survey of the identified top priority areas confirmed the presumed inaccuracies, i.e. the number of hosts in those areas was larger by the presumed +20% than in the landscape H. In that case, the refined data described the landscape $H_{+20\%\text{ in}}$ (i.e. inaccuracies were present only in the identified areas), and $H_{+20\%}$ was considered the most realistic surrogate for the actual landscape. As a result, an epidemic on $H_{+20\%}$ was larger by +21% than predictions based on the refined data ($H_{+20\%\text{ in}}$).

Note that in both cases (ii) and (iii), the identification and subsequent survey of the top priority areas reduced the potential deviation in model predictions for the variable $I(6)$ by approximately four times (from 81% down to 22% or 21%; figure 3f). A qualitatively similar reduction was also observed in the analytical estimate of *rN*: there was approximately fourfold reduction of deviations (from 33% down to 9%) when no errors were found in the data and approximately fivefold reduction of deviations (from 33% down to 7%) when the presumed errors in the data were confirmed. Thus, these findings suggest that the method based on the estimate of the quantity *rN* could be used to identify top priority areas for additional survey that would improve accuracy in epidemic modelling.

4. Discussion

We have presented a solution to the problem of identifying local geographic priorities for host data refinement efforts that would improve the accuracy of the number of infected fields of crops predicted by an epidemic spread model during the early stage of an epidemic (figure 4). As an example of an epidemic of a crop disease, we considered the spread of CBSV through an arbitrarily selected 336 km-by-336 km cassava landscape (derived from CassavaMap [7]) in sub-Saharan Africa where the potential errors in host data (figure 2) can have a substantial effect on predictions of the number of infected fields (figure 3).

The use of an analytical approach enabled the identification of priority areas for host refinement in relation to potential impacts on epidemic spread. It would be practically unfeasible to derive an equivalent solution by computer simulation due to the large number of alternative configurations that would have to be considered; for example, in addition to multiple simulations for each initial condition in each landscape, one would need to consider many alternative landscapes where the selected type of inaccuracy is placed in all possible locations. Within the analytical approach, we used the quantity rN because it has a number of characteristics that are important for the solution derived in this paper. The variable rN is measured locally, i.e. within a selected local area; rN captures changes in both r (that is influenced by the pathogen dispersal kernel and the local spatial structure of the host landscape) and N, the number of hosts in the local area; rN is an additive characteristic in the sense that rN characterizing a larger area is simply a sum of rN estimated from smaller areas within the larger one; therefore, rN allows a direct comparison of contributions from different local areas. An additional advantage of the analytical approach over computer simulations is that it can be used to derive general insights into the dependence of the spatial prioritization for data refinement on the interplay of several factors. These include different types of potential inaccuracies in host data as well as the spatial pattern, especially intercrop clustering, relative to the dispersal kernel of an invading pathogen.

For practical applications, it is important to stress that we assumed that all host crops within a landscape have the same probability of becoming infected at the start of the epidemic. Based on this assumption, areas with the strongest impact from the selected inaccuracy were identified as the top priority areas for data refinement. However, where there are local differences in the probability of crops becoming infected in certain areas, spatial prioritization for data refinement would need to account for that heterogeneity. Differences could arise due to local use of pesticides, cultivation of partially resistant varieties or where environmental conditions differ. Extension of the analytical method to deal with these types of heterogeneities could be considered in future work.

In addition to the analytical approximation equation (2.1) used in this paper, there are other equivalent approximations for the infection rate as well as the closely related quantity, the basic reproduction number R_0. Other approximations that are applicable to non-random host distribution (and therefore relevant to this work) have been derived, for example, by Bolker [31], North & Godfray [32] and more recently by van den Bosch *et al.* [33]. As shown in [18], the estimation of the infection rate according to equation (2.1) differs from earlier estimates [31–33] only when calculated in a local area of a host landscape with a low density of hosts, i.e. when the difference between the density of susceptible hosts, n_S, and the total density, n, matters. In the case of cassava in sub-Saharan Africa, rural areas are often characterized (in CassavaMap) by a low number of cassava fields per unit area; therefore, the selected approximation equation (2.1) has a slightly higher accuracy as compared with alternative approximations [31–33]. Other approximations for infection rate and the basic reproduction number, R_0, would be useful in some specific cases of spatial distribution of hosts. For example, van den Bosch *et al.* [33] derived analytical expressions for R_0 in regular host distributions generated by a Strauss process and in spatially clustered distributions generated by a Neyman–Scott process. For random spatial host distributions, Suprunenko *et al.* [30] derived a spatially local approximation for r, and Wadkin *et al.* [34] improved the accuracy of the approximation for r and R_0 by accounting for host depletion. Mikaberidze *et al.* [35] considered a single rectangular crop field for which they derived R_0 from a system of integro-differential equations. In addition, R_0 can be calculated by using all pairwise probabilities of a host infecting any other hosts (e.g. [16] and references therein). Further work is needed to review the merits of the different approaches, ideally for real systems.

The challenge associated with spatial prioritization for data collection or control and management efforts has attracted attention in recent studies of disease spread in agricultural systems. For example, when studying livestock disease with limited host data, Dawson *et al.* [10] modelled disease spread on the UK cattle trade network and showed that nodes with the highest number of livestock movements should be prioritized for data collection to get more accurate model predictions of an epidemic. Spatial prioritization for data collection aiming to improve the accuracy of spatial modelling of crop diseases is a relatively new topic that supplements a larger field of research on spatial prioritization for surveillance for plant pests and pathogens [36–38] as well as spatial prioritization of management efforts for greater crop yield [39].

Future work could develop the analytical approach presented here to address some open problems. While our analyses have assumed rotationally symmetrical dispersal kernels, crop pathogens often disperse anisotropically. Wheat stem rust [3], for example, is dispersed anisotropically by wind; therefore, incorporating anisotropic dispersal of pathogens into methods of this paper could potentially help in refining wheat field data at various scales [4] to improve accuracy in epidemic modelling. As another example, the lack of spatial information within raster cells in rasterized host data with coarse spatial resolution can potentially be a source of large inaccuracy in raster-based epidemic models. Tildesley *et al.* [8] addressed this problem in optimal control of foot and mouth disease of cattle and showed that data on exact farm locations were not required and that using randomized aggregate county-scale data was sufficient when the model parameters could be re-fitted to the outbreak data on randomized locations. In addition, it was shown that missing spatial information within raster cells in rasterized host data can also be imputed by using land-cover maps [9] or can be predicted by using computational methods such as the Farm Location and Agricultural Production Simulator [11]. However, the investigation of the problem of the coarse spatial resolution in rasterized host data in a broader range of applied research questions would be valuable.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Data [24] and code [27] are available from Figshare.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. Y.F.S.: conceptualization, investigation, methodology, software, visualization, writing—original draft, writing—review and editing; C.A.G.: conceptualization, funding acquisition, resources, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. Y.F.S. and C.A.G. acknowledge support from the Epidemiology and Modelling Group, Cambridge, UK. C.A.G. also acknowledges support from the Bill and Melinda Gates Foundation (INV-010472).

Acknowledgements. The authors are grateful to Dr Alison Scott-Brown, Dr Renata Retkute and Dr Ruairi Donnelly for their helpful comments on the manuscript.

References

1. Cunniffe NJ, Koskella B, E. Metcalf CJ, Parnell S, Gottwald TR, Gilligan CA. 2015 Thirteen challenges in modelling plant diseases. *Epidemics* **10**, 6–10. (doi:10.1016/j.epidem.2014.06.002)
2. Meyer M *et al*. 2021 Wheat rust epidemics damage Ethiopian wheat production: a decade of field disease surveillance reveals national-scale trends in past outbreaks. *PLoS ONE* **16**, e0245697. (doi:10.1371/journal.pone.0245697)
3. Bradshaw CD *et al*. 2022 Irrigation can create new green bridges that promote rapid intercontinental spread of the wheat stem rust pathogen. *Environ. Res. Lett.* **17**, 114025. (doi:10.1088/1748-9326/ac9ac7)
4. Blasch G *et al*. 2024 Ethiopian Crop Type 2020 (EthCT2020) dataset: crop type data for environmental and agricultural remote sensing applications in complex Ethiopian smallholder wheat-based farming systems (Meher season 2020/21). *Data Brief* **54**, 110427. (doi:10.1016/j.dib.2024.110427)
5. Godding D, Stutt ROJH, Alicai T, Abidrabo P, Okao-Okuja G, Gilligan CA. 2023 Developing a predictive model for an emerging epidemic on cassava in sub-Saharan Africa. *Sci. Rep.* **13**, 12603. (doi:10.1038/s41598-023-38819-x)
6. Hassall KL *et al*. 2024 Validating a cassava production spatial disaggregation model in sub-Saharan Africa. *PLoS ONE* **19**, e0312734. (doi:10.1371/journal.pone.0312734)
7. Szyniszewska AM. 2020 CassavaMap, a fine-resolution disaggregation of cassava production and harvested area in Africa in 2014. *Sci. Data* **7**, 159. (doi:10.1038/s41597-020-0501-z)
8. Tildesley MJ, House TA, Bruhn MC, Curry RJ, O'Neil M, Allpress JLE, Smith G, Keeling MJ. 2010 Impact of spatial clustering on disease transmission and optimal control. *Proc. Natl Acad. Sci. USA* **107**, 1041–1046. (doi:10.1073/pnas.0909047107)
9. Tildesley MJ, Ryan SJ. 2012 Disease prevention versus data privacy: using landcover maps to inform spatial epidemic models. *PLoS Comput. Biol.* **8**, e1002723. (doi:10.1371/journal.pcbi.1002723)

10. Dawson PM, Werkman M, Brooks-Pollock E, Tildesley MJ. 2015 Epidemic predictions in an imperfect world: modelling disease spread with partial data. *Proc. R. Soc. B* **282**, 20150205. (doi:10.1098/rspb.2015.0205)

11. Burdett CL, Kraus BR, Garza SJ, Miller RS, Bjork KE. 2015 Simulating the distribution of individual livestock farms and their populations in the United States: an example using domestic swine (*Sus scrofa domesticus*) farms. *PLoS ONE* **10**, e0140338. (doi:10.1371/journal.pone.0140338)

12. Sellman S, Tildesley MJ, Burdett CL, Miller RS, Hallman C, Webb CT, Wennergren U, Portacci K, Lindström T. 2020 Realistic assumptions about spatial locations and clustering of premises matter for models of foot-and-mouth disease spread in the United States. *PLoS Comput. Biol.* **16**, e1007641. (doi:10.1371/journal.pcbi.1007641)

13. Vizzari M, Lesti G, Acharki S. 2024 Crop classification in Google Earth engine: leveraging Sentinel-1, Sentinel-2, European CAP data, and object-based machine-learning approaches. *Geo-Spat. Inf. Sci.* 1–16. (doi:10.1080/10095020.2024.2341748)

14. Alvarez-Vanhard E, Corpetti T, Houet T. 2021 UAV & satellite synergies for optical remote sensing applications: a literature review. *Sci. Remote Sens.* **3**, 100019. (doi:10.1016/j.srs.2021.100019)

15. Kaivosoja J *et al*. 2021 Reference measurements in developing UAV systems for detecting pests, weeds, and diseases. *Remote Sens.* **13**, 1238. (doi:10.3390/rs13071238)

16. Tildesley MJ, Keeling MJ. 2009 Corrigendum to 'Is R0 a good predictor of final epidemic size: foot-and-mouth disease in the UK'. *J. Theor. Biol.* **259**, 863. (doi:10.1016/j.jtbi.2009.05.015)

17. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the basic reproduction number (R0). *Emerg. Infect. Dis.* **25**, 1–4. (doi:10.3201/eid2501.171901)

18. Suprunenko YF, Cornell SJ, Gilligan CA. 2025 Predicting the effect of landscape structure on epidemic invasion using an analytical estimate for infection rate. *R. Soc. Open Sci.* **12**, 240763. (doi:10.1098/rsos.240763)

19. Nweke FI, Lynam JK, Spencer DSC. 2002 *The cassava transformation: Africa's best-kept secret*. East Lansing, MI: Michigan State University Press. See https://www.jstor.org/stable/10.14321/j.ctt7ztc0t.

20. Patil BL, Kanju E, Legg JP, Fauquet CM. 2015 Cassava brown streak disease: a threat to food security in Africa. *J. Gen. Virol.* **96**, 956–968. (doi:10.1099/vir.0.000014)

21. Legg JP *et al*. 2011 Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. *Virus Res.* **159**, 161–170. (doi:10.1016/j.virusres.2011.04.018)

22. Alicai T, Omongo CA, Maruthi MN, Hillocks RJ, Baguma Y, Kawuki R, Bua A, Otim-Nape GW, Colvin J. 2007 Re-emergence of cassava brown streak disease in Uganda. *Plant Dis.* **91**, 24–29. (doi:10.1094/pd-91-0024)

23. Muhindo H, Wembonyama F, Yengele O, Songbo M, Tata-Hangy W, Sikirou M. 2020 Optimum time for harvesting cassava tubers to reduce losses due to cassava brown streak disease in northeastern DRC. *J. Agric. Sci.* **12**, 70–81. (doi:10.5539/jas.v12n5p70)

24. Suprunenko YF, Gilligan CA. 2024 Data from: Where to refine spatial data to improve accuracy in crop disease modelling: an analytical approach with examples for cassava. Figshare. (doi:10.6084/m9.figshare.26236490)

25. Bright E, Rose A, Urban M. 2015 LandScan Global 2014 [Data set]. Oak Ridge National Laboratory. ()

26. Sims K *et al*. 2023 LandScan Global 2022 [Data set]. Oak Ridge National Laboratory. ()

27. Suprunenko YF, Gilligan CA. 2024 Computer code for: Where to refine spatial data to improve accuracy in crop disease modelling: an analytical approach with examples for cassava. Figshare. (doi:10.6084/m9.figshare.26236478)

28. Suprunenko YF, Cornell SJ, Gilligan CA. 2024 Computer code for: Predicting the effect of landscape structure on epidemic invasion using an analytical estimate for infection rate. Figshare. (doi:10.6084/m9.figshare.25804810)

29. Cornell SJ, Suprunenko YF, Finkelshtein D, Somervuo P, Ovaskainen O. 2019 A unified framework for analysis of individual-based models in ecology and beyond. *Nat. Commun.* **10**, 4716. (doi:10.1038/s41467-019-12172-y)

30. Suprunenko YF, Cornell SJ, Gilligan CA. 2021 Analytical approximation for invasion and endemic thresholds, and the optimal control of epidemics in spatially explicit individual-based models. *J. R. Soc. Interface* **18**, 20200966. (doi:10.1098/rsif.2020.0966)

31. Bolker B. 1999 Analytic models for the patchy spread of plant disease. *Bull. Math. Biol.* **61**, 849–874. (doi:10.1006/bulm.1999.0115)

32. North AR, Godfray HCJ. 2017 The dynamics of disease in a metapopulation: the role of dispersal range. *J. Theor. Biol.* **418**, 57–65. (doi:10.1016/j.jtbi.2017.01.037)

33. van den Bosch F, Helps J, Cunniffe NJ. 2024 The basic-reproduction number of infectious diseases in spatially structured host populations. *Oikos* **2024**, e10616. (doi:10.1111/oik.10616)

34. Wadkin LE, Holden J, Ettelaie R, Holmes MJ, Smith J, Golightly A. Estimating the reproduction number, R0, from individual-based models of tree disease spread. *Ecol. Modell.* **489**, 110630. (doi:10.1016/j.ecolmodel.2024.110630)

35. Mikaberidze A, Mundt CC, Bonhoeffer S. 2016 Invasiveness of plant pathogens depends on the spatial scale of host distribution. *Ecol. Appl.* **26**, 1238–1248. (doi:10.1890/15-0807)

36. Tuomola J, Yemshanov D, Huitu H, Hannunen S. 2018 Mapping risks of pest invasions based on the spatio-temporal distribution of hosts. *MBI* **9**, 115–126. (doi:10.3391/mbi.2018.9.2.05)

37. Mastin AJ, Gottwald TR, van den Bosch F, Cunniffe NJ, Parnell S. 2020 Optimising risk-based surveillance for early detection of invasive plant pathogens. *PLoS Biol.* **18**, e3000863. (doi:10.1371/journal.pbio.3000863)

38. Soubeyrand S *et al*. 2024 Building integrated plant health surveillance: a proactive research agenda for anticipating and mitigating disease and pest emergence. *CABI Agric. Biosci.* **5**, 72. (doi:10.1186/s43170-024-00273-8)

39. Buddenhagen CE *et al*. 2022 Where to invest project efforts for greater benefit: a framework for management performance mapping with examples for potato seed health. *Phytopathology* **112**, 1431–1443. (doi:10.1094/phyto-05-20-0202-r)