

Assessing the multi-pathway threat from an invasive agricultural pest: Tuta absoluta in Asia

Joseph Mcnitt, Young Yun Chungbaek, Henning Mortveit, Madhav Marathe, Mateus R Campos, Nicolas Desneux, Thierry Brévault, Rangaswamy Muniappan, Abhijin Adiga

▶ **To cite this version:**

Joseph Mcnitt, Young Yun Chungbaek, Henning Mortveit, Madhav Marathe, Mateus R Campos, et al.. Assessing the multi-pathway threat from an invasive agricultural pest: Tuta absoluta in Asia. Proceedings of the Royal Society B: Biological Sciences, 2019, 286 (1913), pp.1-9. 10.1098/rspb.2019.1159 . hal-03114727

HAL Id: hal-03114727
https://hal.inrae.fr/hal-03114727

Submitted on 19 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: McNitt J, Chungbaek YY, Mortveit H, Marathe M, Campos MR, Desneux N, Brévault T, Muniappan R, Adiga A. 2019 Assessing the multi-pathway threat from an invasive agricultural pest: *Tuta absoluta* in Asia. *Proc. R. Soc. B* **286**: 20191159. http://dx.doi.org/10.1098/rspb.2019.1159

Received: 24 May 2019
Accepted: 20 September 2019

Subject Category:
Ecology

Subject Areas:
computational biology, health and disease and epidemiology, ecology

Keywords:
biological invasion, insect pests, human-mediated spread, spread model, epidemic network models, agent-based modelling

Author for correspondence:
Abhijin Adiga
e-mail: abhijin@gmail.com

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4681763.

THE ROYAL SOCIETY PUBLISHING

Assessing the multi-pathway threat from an invasive agricultural pest: *Tuta absoluta* in Asia

Joseph McNitt[1], Young Yun Chungbaek[2], Henning Mortveit[2], Madhav Marathe[2], Mateus R. Campos[3], Nicolas Desneux[4], Thierry Brévault[5,6,7], Rangaswamy Muniappan[8] and Abhijin Adiga[2]

[1]Epic Systems Corporation, Verona, WI, USA
[2]Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, VA, USA
[3]French National Institute for Agricultural Research, Avignon, Provence-Alpes-Côte d'Azu, France
[4]Université Côte d'Azur, INRA, CNRS, UMR ISA, 06000, Nice, France
[5]BIOPASS, CIRAD-IRD-ISRA-UCAD, Dakar, Senegal
[6]CIRAD, UPR AIDA, Centre de Recherche ISRA-IRD, Dakar, Senegal
[7]AIDA, Université de Montpellier, CIRAD, Montpellier, France
[8]Feed the Future Integrated Pest Management Innovation Laboratory, Virginia Tech, Blacksburg VA, USA

HM, 0000-0003-3363-2947; MM, 0000-0003-1653-0658; MRC, 0000-0001-9317-3215; TB, 0000-0003-0159-3509; RM, 0000-0002-2992-2792; AA, 0000-0002-9770-034X

Modern food systems facilitate rapid dispersal of pests and pathogens through multiple pathways. The complexity of spread dynamics and data inadequacy make it challenging to model the phenomenon and also to prepare for emerging invasions. We present a generic framework to study the spatio-temporal spread of invasive species as a multi-scale propagation process over a time-varying network accounting for climate, biology, seasonal production, trade and demographic information. Machine learning techniques are used in a novel manner to capture model variability and analyse parameter sensitivity. We applied the framework to understand the spread of a devastating pest of tomato, *Tuta absoluta*, in South and Southeast Asia, a region at the frontier of its current range. Analysis with respect to historical invasion records suggests that even with modest self-mediated spread capabilities, the pest can quickly expand its range through domestic city-to-city vegetable trade. Our models forecast that within 5–7 years, *Tuta absoluta* will invade all major vegetable growing areas of mainland Southeast Asia assuming unmitigated spread. Monitoring high-consumption areas can help in early detection, and targeted interventions at major production areas can effectively reduce the rate of spread.

1. Introduction

As the intensity of trade and human mobility increases, so does the rate of exotic species invasions [1]. Climate change and the detrimental impact of intensive agriculture on natural resources further aggravate this problem [2]. Understanding the dynamics of invasive species spread is imperative for achieving zero hunger, no poverty, good health and well being, which are among the sustainable development goals drafted by the United Nations [3]. Models play an important role in predicting the spatio-temporal spread, identifying roles of different pathways, assessing efficacy of control strategies and exposing gaps in the understanding of the phenomenon [4,5]. However, impending invasions of agricultural pests present difficult challenges. Accounting for multiple drivers of dispersal invariably makes the model complex. At the same time, data inadequacy makes it nearly impossible to calibrate and validate these models. Despite these limitations, a natural goal for a modeller is to provide useful insights into the mechanisms of spread, and thus help design effective policies for its prevention and mitigation.

© 2019 The Author(s) Published by the Royal Society. All rights reserved.

Network propagation models have been widely used to study phenomena as diverse as infectious disease and invasive species spread, online social networks and cascading failures in infrastructure networks [6]. Douma *et al.* [7] survey the invasive species literature categorizing various efforts into flow-based pathway models and agent-based models. Network representations and analysis are being increasingly applied to capture human-mediated pathways of spread [8,9], multi-scale spread [10], monitoring [11], and mitigation [12]. Unlike pest risk maps generated by species distribution models [13], the resulting dynamics of such a validated model yields a causal description of the underlying complex system.

We present a multi-pathway propagation model to study the spread of invasive agricultural pests. We applied it to study the spread of the South American tomato leafminer or *Tuta absoluta*, a pest of the tomato crop and representative of recent biological invasions that have significantly perturbed global food production. Indigenous to South America, *T. absoluta* was accidentally introduced to Spain in 2006, and since then has rapidly spread throughout Europe, Africa, Western and Central Asia, the Indian subcontinent and parts of Central America [14,15]. It is well accepted that trade played a critical role in *T. absoluta*'s rapid spread. On multiple occasions, it has been discovered in packaging stations and its spread pattern is correlated with prime trade routes [16]. Our study region is South and Southeast Asia—a region at the frontier of its current range—comprising of 10 countries: members of the Association of Southeast Asian Nations (ASEAN) and Bangladesh. In recent years, there has been a thrust to improve vegetable production in all the countries of this region. With the pest having already spread to major tomato producing areas in Bangladesh, there is a high chance that it will be introduced to the remaining countries in the near future. Such invasions can have devastating effect on the economy and livelihood of farmers. Moreover, the invasion in mainland Southeast Asia, in particular, is a serious threat to China [17], the largest producer of tomatoes, and Australasian neighbours. To our knowledge, this is the first study that explicitly considers multiple pathways of introduction and spread of *T. absoluta*. Earlier modelling efforts have only accounted for ecological aspects and self-mediated spread [14,18,19]. A precursor to this work [20] modelled the seasonal production and trade of tomatoes in Nepal to study the role of trade in the spread of *T. absoluta* using a gravity model and network dynamics.

Our model accounts for both self-mediated and human-mediated spread and encapsulates the spatial heterogeneity, temporal variations and multi-scale nature of the propagation mechanisms. To construct this model, we identified, analysed, and fused disparate datasets corresponding to biology, climate, production and agricultural commodity flow. With *T. absoluta* being an emerging invasion in the focus region, some of the pertinent questions are (i) what are the possible explanations for the observed spread; (ii) what are the possible patterns of future spread; and (iii) what steps could be taken to mitigate it. We develop a framework to parameterize and analyse the multi-pathway model with respect to ground truth by a novel application of popular supervised and unsupervised machine learning algorithms. Our approaches are motivated by recent research on machine learning surrogates for agent-based models [21] and interpretable artificial intelligence [22]. The analysis provides valuable insights into the dynamics of *T. absoluta* spread and its control, particularly from the perspective of human-mediated spread.

2. Methods

(a) Data

The global datasets used in the model and for analysis are described in the electronic supplementary material, table S1. Country specific data on seasonal production, consumption, processing and trade was obtained from websites of agriculture ministries, research articles and technical reports (electronic supplementary material, table S2). Almost all the datasets used are openly available. The details of *T. absoluta* biology can be found in the electronic supplementary material.

(b) Multi-pathway spread model

We developed a stochastic multi-scale propagation model to simulate the multi-pathway spread of *T. absoluta*. Key concepts are illustrated in figure 1*a*. The model parameters and their values are summarized in table 1. A discussion on the choice of model structure and assumptions is presented in the electronic supplementary material, S3.1. The study region is divided into cells, the smallest spatial units, by overlaying a grid ($0.25° \times 0.25°$). Each cell is in one of the three states: susceptible (S) denoting pest-free state, exposed (E) denoting that the pest has been introduced but the population has not yet built up to influence other cells, and infectious (I) denoting that the pest has established and the cell can influence its neighbours. The cell states are updated in discrete time steps, each corresponding to one month. The probability that a cell v transitions from state S to E is determined by (i) suitability to establish at that time step $\epsilon(v, t)$ and (ii) influence of neighbours in state I depending on the pathway. An exposed cell transitions to state I after a latency period of ℓ time steps. This is the time required for the excess population to build up to infect other cells. Once the pest has established in a cell, the cell remains infected forever, a fair assumption considering that, historically, eradication of *T. absoluta* has not been successful (the only exception being the UK). The infectiousness of a cell $\rho(v, t)$ is modelled as a linear function of host density at time t, for which we use the weighted sum of production volume of tomato, eggplant and potato in that cell at time t. The details are in the electronic supplementary material.

There are three pathways by which a cell can become infected: short-distance dispersal, local human-mediated dispersal and long-distance dispersal. Short-distance dispersal captures the spread through natural means; from an infested cell to cells in its Moore neighbourhood of range r_M. The probability that a susceptible cell gets exposed (E) at time step t through short-distance spread is as follows:

$$p_s(v, t) = \epsilon(v, t)\left(1 - \exp\left(-\alpha_s \sum_{v' \in M_v(r_M)} \rho(v', t)\right)\right). \tag{2.1}$$

The probability depends on the suitability of the cell $\epsilon(v, t)$, infestation level of each neighbouring cell in the Moore neighbourhood with range r_M, $\rho(v', t)$, and the scaling factor, α_s, which is the transmission rate for this pathway. The function form is explained in the electronic supplementary material, S3.1.

For human-assisted spread, we identified large urban areas in the region which we refer to as *localities* (figure 1*a*) and considered interactions within and between localities. These areas have significant trade flows owing to high consumption or production. Each *locality* consists of all grid cells which are within a certain distance (determined by *locality radius*) from its corresponding centre. Local human-mediated dispersal is modelled

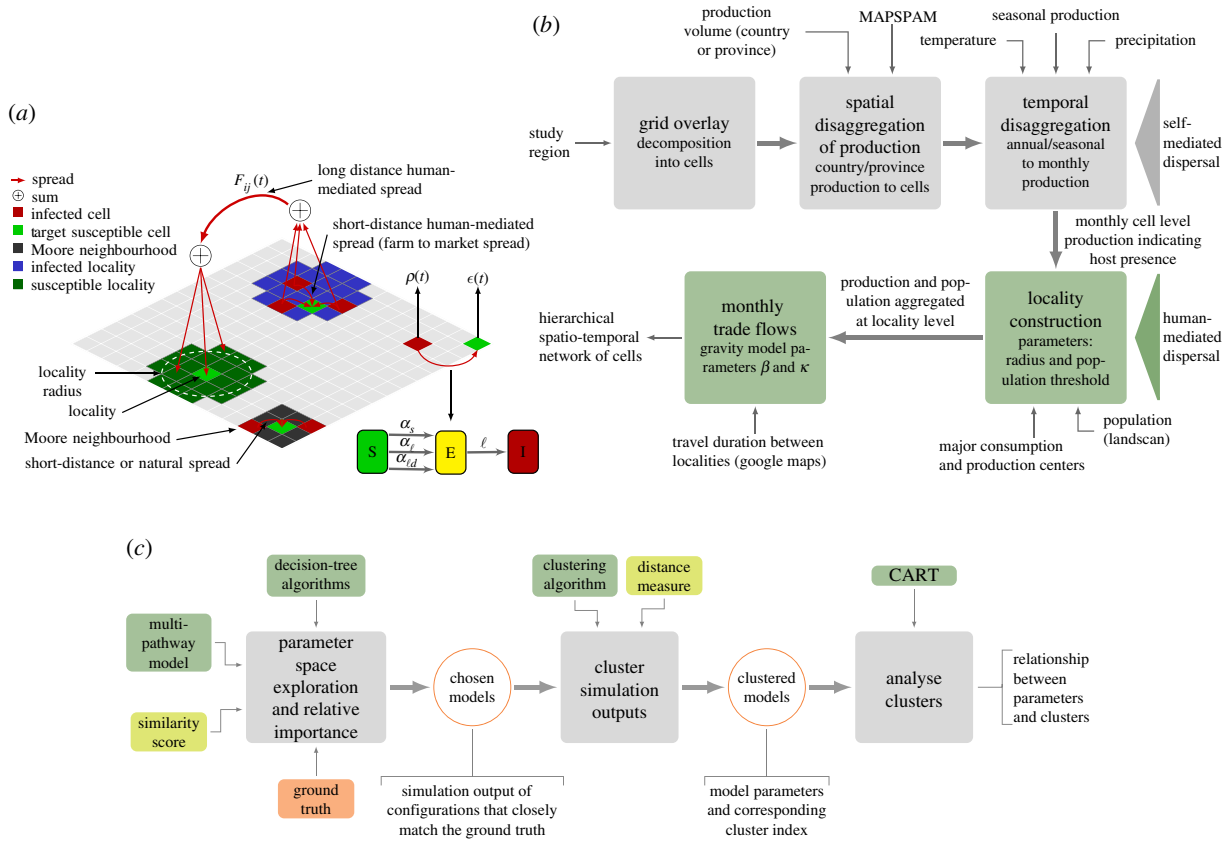


Figure 1. Multi-pathway model concept, construction and analysis. (*a*) Multi-pathway model. (*b*) Model construction pipeline. (*c*) Outline of the process used for analysing the multi-pathway spread. (Online version in colour.)

as the spread between cells belonging to a locality. Every cell v is influenced by cells in its locality L based on their infectiousness. The expression is similar to that in equation (2.1), but with cells in the locality instead of the Moore neighbourhood:

$$p_\ell(v,t) = \epsilon(v,t)\left(1 - \exp\left(-\alpha_\ell \sum_{v' \in \mathrm{L}} \rho(v',t)\right)\right), \tag{2.2}$$

where α_ℓ is the scaling factor. The details of locality construction are provided in the electronic supplementary material, S3.2.

Long-distance human-mediated dispersal corresponds to spread through trade between localities. For this purpose, we considered only tomato trade as there is not much evidence of *T. absoluta* spreading through trade of other hosts. We modelled domestic trade using a gravity model approach accounting for tomato production, processing, imports and exports in each locality, and the travel time between localities. The probability of spread is directly proportional to the trade flow F_{ij} from one locality (i) to another (j). Suppose cell v belongs to locality i. Then, the probability of cell v transitioning from S to E due to long-distance human-mediated dispersal is given by

$$p_{\ell d}(v,t) = \epsilon(v,t)\left(1 - \exp\left(-\alpha_{\ell d} \sum_{j \neq i} \sum_{v' \in \mathrm{L}(j)} F_{ji}\rho(v',t)\right)\right), \tag{2.3}$$

where $\alpha_{\ell d}$ is the pathway scaling factor.

(c) Network construction

Figure 1*b* provides a schematic of the network construction. The first step was to estimate monthly production volume of tomato, eggplant and potato for each cell. We estimated annual production in each cell followed by disaggregation to monthly production. The annual production was estimated using the vegetable production available at the highest resolution for each country (at the level of province to just one value for the entire country) and a synthetic dataset called the spatial production allocation model [23]. For monthly production, we used linear regression to model the production rate as a function of precipitation, temperature and elevation. Seasonal tomato and eggplant production data for different regions of the Philippines was used. For most of the other countries, only qualitative information on seasonal production is available (electronic supplementary material, table S2). The regression function was applied to locations of these countries where this information is available and visually compared to available data. The details are in the electronic supplementary material, S3.3.

To model locality-to-locality trade, we applied the approach of Venkatramanan *et al.* [20] with some modifications. We modelled the flow of fresh tomato crop between markets based on the following assumptions: (i) the total outflow from a city depends on the amount of produce in its surrounding regions and imports from countries outside the focus region at time t, and (ii) the total inflow depends on total consumption, processing demand, and exports from the city to countries outside the focus region. The details are in the electronic supplementary material, S3.4. Trade between countries of the focus region was not modelled as there is no adequate information on ports of entry or monthly flow volumes. But, it was accounted for while analysing the possible routes of introduction.

(d) Parameterization and experiment design

The goodness of fit of a parameter instance was determined by comparing the simulation output with *T. absoluta* incidence reports (figure 2*a*; electronic supplementary material, table S3 for Bangladesh). The spread was simulated with infestation starting from the location of first report. For each cell, v and model configuration C, the empirical probability $p(C, v, t)$ that it is in state I at time t was computed (averaged over 100 repetitions). The output was compared to ground truth using a similarity function adapted from [9]. Let v be a reporting cell and t_v denote the month of the actual report of pest presence. To account for uncertainty in reporting, we consider a time

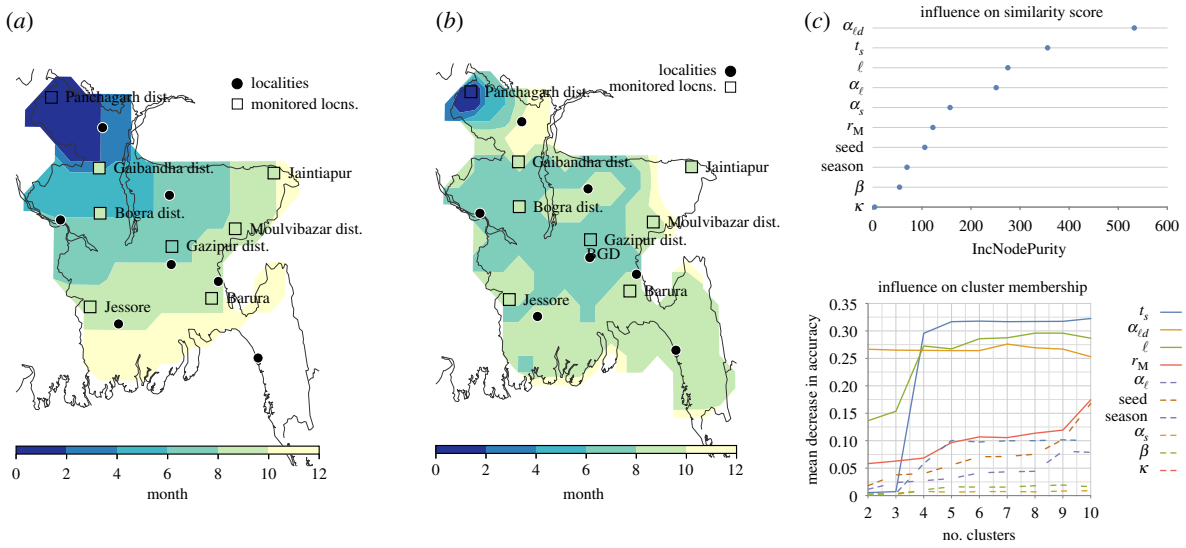


Figure 2. Explaining observed spread pattern in Bangladesh. The contour plots show the spread starting from the location of first report in Panchagarh district for a simulation time of 12 months. Here, the time of infection for a cell is the minimum time step t such that the empirical probability that the cell is infected by time t is greater than or equal to 0.8. Also highlighted are the eight monitored locations and the localities applied in the model. The colours of the monitored locations correspond to the time of infection relative to the first report (Panchagarh). Two distinct spread patterns emerged from the cluster analysis. (*a*,*b*) Representative spreads observed for each class. The similarity in each case was $\mathcal{S} > 0.8$. Importance of model parameters with respect to (i) similarity score $\mathcal{S}$ and (ii) cluster membership based on the random forest method. The latter plot shows how the results vary with an increase in the number of clusters for hierarchical clustering algorithm. More results are presented in the electronic supplementary material, figure S9. Videos depicting the spatio-temporal spread for each class are provided in the supplement. (*a*) Class A, (*b*) class B and (*c*) parameter importance. (Online version in colour.)

Table 1. Model parameters, their values and notes on parameter choices and ranges.

parameter	description	value/range
r_M	range of Moore neighbourhood	{1, 2, 3} corresponding to spread per month of approximately 25 km, 50 km and 75 km, respectively
ℓ	latency period to transition from E to I	{1, 2, 3} months based on the time for the pest to complete life cycle (*T. absoluta* biology in the electronic supplementary material, S2)
season	disaggregation of annual production to monthly values	*uniform* throughout the year or *seasonal* based on regression analysis (Methods)
β	gravity model distance function exponent	{0, 1, 2}
κ	gravity model distance function cut-off	between 4 and 16 h of travel time
seed	location and time of initial infestation	scenarios based on countries (see the electronic supplementary material, table S5)
locality radius	determines cells assigned to a locality	100 km (see the electronic supplementary material, S3.2 for locality construction and analysis)
t_s	time of initial infestation during parameterization	{3, 4, 5} corresponding to March, April and May, respectively, based on first report in Bangladesh (electronic supplementary material, table S3)
α_s, α_ℓ, $\alpha_{\ell d}$	pathway scaling factors	in the interval [0, 500]

window $U_\tau = [t_v - \tau, t_v + \tau]$ during comparison, where τ is the uncertainty parameter. We set $\tau = 2$, that is, error within ±2 months is tolerated. Supposing $\mathcal{C}_R$ is the set of cells corresponding to ground truth, then the similarity $\mathcal{S}$ is given by

$$\mathcal{S}(C) = \frac{1}{|\mathcal{C}_R|} \sum_{v \in \mathcal{C}_R} \left(\sum_{t \in U_\tau} p(C, v, t) + \sum_{t \notin U_\tau} (1 - p(C, v, t)) \right). \tag{2.4}$$

For parameter space exploration, we were motivated by a recent approach of using machine learning surrogates [21]. In our iterative 'go with the winners' process [24], the subspace under consideration is sampled uniformly (the first part of figure 1*c*). Then, with model parameters as features and the similarity score as the dependent variable, we use classification and regression trees (CART) approach to identify parts of the subspace for which the similarity score is high and reject the remaining. In the following iteration, these subspaces are sampled uniformly, and the process continues. The approach is very similar to the *reverse engineering approach* used to build an interpretable learner for a black box model ([22], figure 10). In the interpretable machine learning framework, the black box is a machine learning algorithm like a neural network or tree ensemble, while in our case, the black box is an agent-based

model. Simulations were performed for more than 500 000 parameter combinations using a high performance computing cluster. Configurations with similarity score $\mathcal{S}(C) \geq 0.75$ were chosen for further analysis.

(e) Analysis of spread pattern

The objective here is to analyse the variability in the simulation outcomes within the set of best-fit configurations. We leverage well-known machine learning techniques in a novel way to address this question. The methodology is outlined in figure 1*c*. First, we cluster the simulation outputs (time and cell-indexed empirical probabilities) of selected configurations from the parameterization phase. This step captures the variability in outcomes; simulation outputs belonging to different clusters can be considered to be significantly different from one another. In the second step, we attempt to infer relationships between model parameters and cluster membership. To this end, our approach is to cast this as a classification problem using CART with model parameters as the features and cluster index as the label. The relationships are inferred from the decision tree that resulted from the algorithm. To avoid any bias introduced by the clustering algorithm, we also apply more than one method—hierarchical agglomerative clustering and the *k*-means algorithm. In both cases, we use the Euclidean distance as the distance measure to compare two simulation outputs. The analysis is repeated for different values of *k*, the number of clusters. More details are provided in the electronic supplementary material, S5.

To assess the relative importance of model parameters, we adopted the approach of Lamperti *et al.* [21]. We use the random forest algorithm [25] to assess the importance with respect to (i) similarity score ($\mathcal{S}$) and (ii) spread pattern, which in our case, is akin to cluster membership. The set-up is similar to the parameterization and cluster analysis case, with CART replaced by the random forest algorithm. Details are in the electronic supplementary material, S5. To evaluate parameter importance with respect to the similarity score, we used *mean increase in node purity* as the criterion (as it is a regression problem), and for cluster membership, we used *decrease in accuracy* (classification problem).

3. Results

(a) Variability in spread pattern

The clustering analysis of the configurations selected during the parameterization phase (approx. 8000 of them) reveals two distinct spread patterns primarily determined by the pathway parameters. The first class of models (figure 2*a*), referred to as class A, is characterized by the absence of long-distance human-mediated spread ($\alpha_{\ell d}$ negligible) and brisk spread between geographically adjacent cells, driven by the latency period ℓ, the Moore range r_M and the short-distance scaling factor α_s. By contrast, for class B models (figure 2*b*), the long-distance pathway ($\alpha_{\ell d}$) plays a significant role and there is relatively slow spread between geographically adjacent neighbours. Both hierarchical clustering and *k*-means clustering (electronic supplementary material, figures S5(b) and S6(a)) are consistent in this regard.

The class A spread pattern does not capture the gap between the time of first report (Panchagarh) and the report in Gaibandha district (figure 2*a*). Even though the distance between the two locations is only 185 km, the latter reported the presence only after 10 months of first report, suggesting that self-mediated spread might have been much slower. In the model output on the other hand, the corresponding cell gets infected between the second and fourth months. In class B, this location is infected much later in comparison. However, the eastward spread towards the location Jaintiapur is slower than what was observed (figure 2*b*). Even though Panchagarh is quite far from this location, pest presence was reported by February 2017, just nine months after the first report. As a baseline, we also simulated the spread using the cellular automata model developed by Guimapi [19] for Bangladesh. The spread pattern is similar to class A as the model does not account for long-distance hops. However, the predicted rate of range expansion is much higher than our models (see the electronic supplementary material, S6.3 for model details and results).

In the case of spread pattern, the importance was derived for each *k* (number of clusters) and clustering algorithm. Some results are presented in figure 2*c*. We note that the long-distance scaling factor ($\alpha_{\ell d}$) is among the top three important parameters. The start month (t_s) is also important for two reasons. Firstly, the distance between two time-shifted simulation outputs can be large. Secondly, outputs are sensitive to seasonal variations or temporality of the network. Latency period (ℓ) and Moore range (r_M) together control the extent of radial spread in a time step. Typically, for class A models, r_M is high and ℓ is low and the other way round in the case of class B models. Analysis of trade flows and seasonality is presented in the electronic supplementary material, S6.2.

(b) Scenarios of pest introduction to countries in Southeast Asia

To identify routes of introduction to other countries in the region, we applied both class A and class B models. The starting point of the spread corresponds to the Panchagarh district (figure 2*a*). Both model classes strongly indicate that *T. absoluta* is already present in parts of Myanmar (curves corresponding to time step 24, or 2 years from first report). Also, the pest is likely to enter Thailand from Myanmar, and subsequently move to Laos, Cambodia and Vietnam as it spreads eastwards, and to China when it spreads northwards. From Thailand, spreading southward, it will enter Malaysia and subsequently enter Indonesia (figure 3*a*,*b*).

We also analysed the international tomato trade network (electronic supplementary material, S6.1) to assess the risk owing to imports from *T. absoluta* infested countries outside this region. Malaysia and Singapore are important hubs with tomato imports from *T. absoluta* infested regions. There is a possibility that *T. absoluta* is directly introduced to these regions. However, in both cases, the import volume is very low. Also, the introduction risk depends on the preventive measures taken by the exporting countries. With respect to both trade and natural pathways, there is a low chance that the pest will be introduced into Philippines from neighbouring countries, as there are no shared borders with any country in the region nor evidence of tomato trade. However, human mobility is a possible pathway. For example, the Middle East is the top destination for Filipino workers.

(c) Predicted spread is model and region dependent

In the case of class A models, the eastward spread is faster than southward spread (figure 3*a*). This is mainly because the Moore neighbourhood is smaller at the narrow region in the south of Myanmar and Thailand bordering Malaysia.

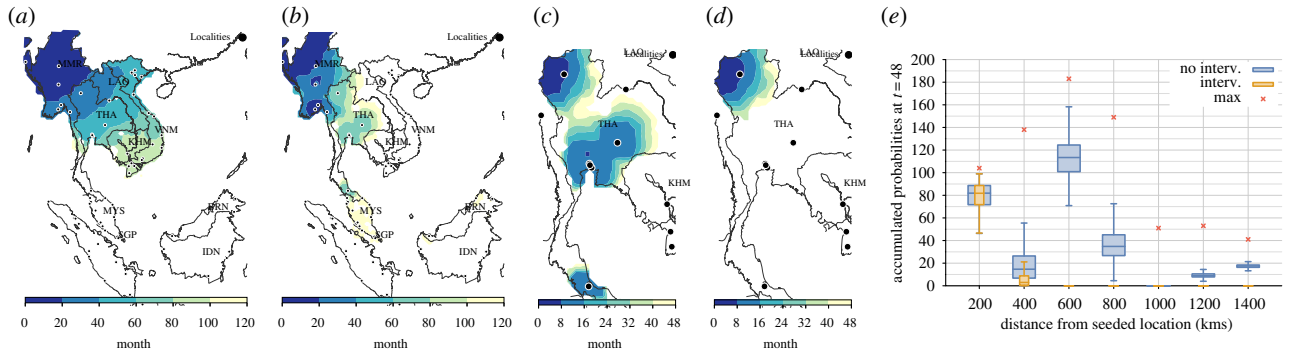


Figure 3. Predicting rate and pattern of spread. The contour plots show the spatio-temporal spread. The colours indicate the time interval at which there is at least a 50% chance that a location will be infected. *Spread in Southeast Asia*. Figures (*a*,*b*) correspond to the spread starting from northern Myanmar for 120 time steps or 10 years; (*a*,*b*) correspond to representative simulation output for class A and class B models respectively. *Domestic spread with and without intervention.* In (*c*,*d*), representative spread dynamics of class B models ($r_M = 1$, $\ell = 3$) are shown for the country of Thailand. More plots are in the electronic supplementary material, figure S12. In each case, a cell close to a high production region was seeded and simulation run for 48 time steps; (*c*) corresponds to unmitigated spread while (*d*) corresponds to spread after cutting off flows from chosen localities. (*e*) Average spread with respect to origin of infection for all class B models. The cells are binned based on their distance from the origin of infection. Given time step t (48), let $\Pr_{\leq t}(v)$ be the probability that cell v is in state I by time t. For each configuration, we computed the 'total infection' for every bin at time t by aggregating $\Pr_{\leq t}(v)$ for each v in the bin. The cross points referred to as 'max' correspond to the total number of cells in each bin, which is also the maximum possible accumulated probability for that bin. (Online version in colour.)

However, in the case of class B (figure 3*b*), the spread is much faster in the same region aided by domestic trade flows from northern and central Thailand to the southern region. The class A spread pattern predicts that within the next 4–5 years, much of the northern part of mainland Southeast Asia will be invaded. The class B spread pattern predicts that in the same period, *T. absoluta* will spread all over Malaysia and Singapore. However, the rate of spread observed is slower than that observed in Bangladesh for both classes. Also, even though the models exhibited a similar rate of spread for Bangladesh, we observed high variance in intensity of infestation as well as range expansion for the rest of the region. The results are in the electronic supplementary material, figure S11. The reason for slow spread is as follows. Bangladesh has the highest tomato volume per country surface area ($\approx$2.5 tonnes km^{-2}). The next country is Vietnam ($\approx$1.5 tonnes km^{-2}). Therefore, in the case of Bangladesh, not only is the extent of infestation in a cell $\rho(\cdot)$ typically high, but also because it is a densely populated country, most cells have vegetable production. Hence, the rate of spread is much higher for relatively lower values of pathway parameters and Moore range. Also, we observed a strong dependence on Moore range (electronic supplementary material, figure S11*b*). In geographically larger countries, the production is scattered. Therefore, the lower the Moore range, the slower the spread.

(d) Influence of domestic trade on spread pattern and rate

Here, we focus on long-distance dispersal and therefore restrict our discussion to class B models. For the country-specific studies, the starting location was decided based on our analysis of possible entry points through different pathways (electronic supplementary material, S4.1). We observed the following common spread pattern. When the invasive species is introduced to a country, dispersal is slow until the invasion front reaches a production source. Once it establishes at a source, the spread is very fast. Depending on the country, within 12–24 time steps (or 1–2 years), it spreads to almost all major localities of the country (see figure 3*c* for example). Production areas which are very close to high-consumption localities (large urban areas) are particularly vulnerable. Because local production typically does not satisfy demands of such localities, they have high inflows from other production areas and possibly from other countries. As a result, these localities are quickly infected. Once introduced to such localities, farmer–market interactions (local human-mediated dispersal) can facilitate the introduction of the pest to nearby production regions where it can establish and proliferate. Similar observations have been made in a number of works [26,27].

Given that monitoring and quarantining are both resource-intensive and potentially disruptive, developing strategies that involve few locations yet provide near-optimal control is a goal for modellers. Market-level phytosanitary measures in terms of import restrictions have been undertaken by countries [28]. Here, we evaluated a simple strategy for containing the spread through the trade pathway. Localities associated with high annual outflows were identified (at most four in each country). As discussed earlier, pest establishment in these areas can potentially lead to rapid range expansion. The outflow from the targeted localities was cut off to mimic control at the trade/market level. In the strictest sense, this can be implemented by restricting trade of host crops. But, it is possible that phytosanitary measures have the same effect. Figure 3*d* shows the spread after intervention. More results are present in the electronic supplementary material, figure S9. Consistently, across countries, we observed a significant reduction in range expansion as well as intensity of spread. Besides, as seen in figure 3*d*, stifling these flows localizes the spread that resembles those of class A models, but with much less intensity.

4. Discussion

The variability in the spread patterns that explain the incidence reports exposes the lack of understanding of the pathways of spread. Nevertheless, the analysis does strongly indicate the role of human-assisted spread of *T. absoluta*. The pest was reported in May 2016 in the northwestern part of

Bangladesh bordering India. The region is among the top three tomato producers in the country. By the beginning of the next production season, *T. absoluta* was found in almost every major urban region. Similar correlation between the tomato trade and *T. absoluta* spread was observed in Nepal [20]. Studies on self-mediated spread (flying capability or by wind) can definitely help estimate more accurately the rate of self-mediated spread. It is also important to consider alternate scenarios of introduction. We recall that the far eastern part of Bangladesh (locality Jaintiapur in figure 2*b*) reported pest presence nine months after the first report. This place happens to be close to an important trade route connecting northeastern Bangladesh to Meghalaya in India, where *T. absoluta* was officially reported in January 2017. Therefore, it is possible that multiple incursions took place.

Historically, international trade has played a strong role in the spread of *T. absoluta* between countries. For example, the pest was first reported by India in 2014. By early 2016, it was discovered in the Kathmandu area of Nepal and in the northern part of Bangladesh in May 2016. Both countries import significant volumes of tomatoes from India. However, there has been no report from Pakistan, another neighbour which does not import tomatoes from India. There are similar examples outside the region such as its slow advance from South America to Central America, or the fact that it is not reported in China despite being present in neighbouring Central Asian countries since 2015. We recall the discussion on slow predicted rate of spread in mainland Southeast Asia compared to the observed rate in Bangladesh. One reason for this could be the unaccounted trade flows between countries. International trade within this region is not documented well. It is critical to address the data gaps concerning international trade, particularly considering that production and trade between countries in this region have been increasing over the years (details are in the electronic supplementary material, S6.1).

While several integrated pest management strategies have been suggested for managing *T. absoluta*, hardly any work has been done in designing effective interventions at the trade level. Some countries have already taken measures in this regard. In the USA, the Animal and Plant Health Inspection Service of the Department of Agriculture (USDA-APHIS) has instituted quarantine regulations for imports from regions where the pest is present [28]. Identifying important locations to mitigate an epidemic ([26,27,29]) or monitor are problems being increasingly studied with the lens of network analysis. While good algorithms exist for undirected networks, solutions for these problems on directed weighted networks are few and far in between.

Emulators—based on Gaussian processes, for example [30]—and machine learning surrogates [21] are emerging as solutions to overcome computational challenges, parameterization and sensitivity analysis of complex agent-based models. Our approaches were motivated by these works. We are not aware of any previous work that analyses the dynamics of simulation systems using unsupervised learning as presented in this paper. However, clustering has been considered in the context of multi-resolution simulation models as an interfacing component between simulators with different resolutions [31]. Our use of CART to explain the clustering is motivated by recent work in interpretable aritificial intelligence [22], where deep learning models are interpreted using decision-tree proxy models.

(a) Challenges and limitations

Modelling emerging invasions is particularly challenging. Limited data on incidence and understanding of the underlying dynamics makes it nearly impossible to calibrate and validate the models. We have had to simplify or ignore some of the processes that might significantly influence the spread. For example, our model uses monthly production as a surrogate for infectiousness of a cell. Complex phenology models can be used instead (as in Carrasco *et al.* [9]), but would add to the complexity of the model. Because our focus region spans multiple countries, identifying and collecting data for each country was a lengthy process. For many countries, data had to be collected (or even inferred) from several publications and reports (electronic supplementary material, table S2). Furthermore, these datasets were misaligned in time and spatial resolution. It is important to account for heterogeneity in production, consumption, awareness, cultural factors, etc. both within and between countries. Some countries are technologically more advanced than others, which manifests as differences in yield, crop loss, trade infrastructure, pest awareness and preparation for invasion [2].

In particular, it is hard to model human-assisted spread owing to lack of seasonal trade data. To determine outflows and inflows for each locality, we had to identify major ports for import(s) and export(s) as well as estimate the fraction of production which was used for processing and was available only for a few countries. The farm–market-consumer interactions (local human-mediated spread) involves various actors such as farmers, wholesalers, retailers, wet markets, supermarkets and so on. Modelling this is a challenge in itself. If data on actual flows of vegetables is provided, the gravity model can be improved or replaced by more sophisticated approaches. Also, the relationship between long-distance invasion risk and trade volume is hard to determine. While a direct relationship between volume and risk is plausible, whether the relation is linear (as assumed by our model) is unclear.

(b) Conclusion

Traditionally, in developing countries, crops such as the tomato are seasonal. However, over the past decade, owing to rising demand and opportunities to export, there has been a thrust towards year-round production using protected cultivation methods and resilient varieties. An increase in urban population, short shelf life of vegetables, and the advantages of short marketing chains have encouraged urban agriculture in developing countries [32]. Our results indicate that such urban and peri-urban agriculture is particularly vulnerable to invasive species attacks. In particular, in Southeast Asia, vegetable production and internal trade have steadily increased. In comparison, the export of tomatoes outside of the focus region has risen steeply in recent years (after 2011), while the imports generally indicate a downward trend. Therefore, invasions from pests such as *T. absoluta* can have a huge negative impact on the socio-economic fabric of this region. The modelling and analysis framework presented here is generic and applicable to other invasive species. The methodology is modular and leverages popular learning algorithms to analyse complex models under data scarcity. Other potential applications for this work include studies of natural or

human-initiated disasters, climate change and optimization of food flows.

Data accessibility. The authors declare that the data supporting the findings of this study are available within the paper and its electronic supplementary material, or from the authors upon reasonable request.

Authors' contributions. A.A. defined the scope of the research. A.A., J.M., T.B., M.R.C. and N.D. collected and interpreted data. A.A. and M.M. conceived and designed the experiments. J.M., A.A. and Y.Y.C. performed the analysis. H.M. and R.M. provided assistance in interpreting the results. A.A. and J.M. wrote the paper with significant inputs from M.R.C. and Y.Y.C. A.A. supervised the research. All authors discussed the results and commented on the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work was supported in part by the United States Agency for International Development under the Cooperative Agreement no. AID-OAA-L-15-00001, Feed the Future Innovation Laboratory for Integrated Pest Management, DTRA CNIMS Contract HDTRA1-11-D-0016-0001, NSF BIG DATA grant no. IIS-1633028, NSF DIBBS grant no. ACI-1443054, NIH grant no. 1R01GM109718 and NSF NRT-DESE grant no. DGE-154362.

Acknowledgements. We are grateful to Yousuf Mian, Nguyen Van Hoa and Kimhian Seng for their help with obtaining country-specific information on production, trade and pest incidence. We thank Richard Beckman, Irene Eckstrand, Srinivasan Venkatramanan, Stephen Eubank and Erin Raymond for useful discussions on model design and paper organization.

References

1. Hulme PE. 2009 Trade, transport and trouble: managing invasive species pathways in an era of globalization. *J. Appl. Ecol.* **46**, 10–18. (doi:10.1111/jpe.2009.46.issue-1)
2. Early R *et al.* 2016 Global threats from invasive alien species in the twenty-first century and national response capacities. *Nat. Commun.* **7**, 12485. (doi:10.1038/ncomms12485)
3. United Nations. 2019 Sustainable development goals, April. See https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/.
4. Cunniffe NJ, Koskella B, Metcalf CJE, 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)
5. Epstein JM. 2008 Why model? *J. Artif. Soc. Soc. Simul.* **11**, 12.
6. Barrat A, Barthelemy M, Vespignani A. 2008 *Dynamical processes on complex networks*. Cambridge, UK: Cambridge University Press.
7. Douma J, Pautasso M, Venette R, Robinet C, Hemerik L, Mourits M, Schans J, van der Werf W. 2016 Pathway models for analysing and managing the introduction of alien plant pests an overview and categorization. *Ecol. Modell.* **339**, 58–67. (doi:10.1016/j.ecolmodel.2016.08.009)
8. Nopsa JFH, Daglish GJ, Hagstrum DW, Leslie JF, Phillips TW, Scoglio C, Thomas-Sharma S, Walter GH, Garrett KA. 2015 Ecological networks in stored grain: key postharvest nodes for emerging pests, pathogens, and mycotoxins. *BioScience* **65**, 985–1002. (doi:10.1093/biosci/biv122)
9. Carrasco L, Mumford J, MacLeod A, Harwood T, Grabenweger G, Leach A, Knight J, Baker R. 2010 Unveiling human-assisted dispersal mechanisms in invasive alien insects: integration of spatial stochastic simulation and phenology models. *Ecol. Modell.* **221**, 2068–2075. (doi:10.1016/j.ecolmodel.2010.05.012)
10. Wildemeersch M, Franklin O, Seidl R, Rogelj J, Moorthy I, Thurner S. 2019 Modelling the multi-scaled nature of pest outbreaks. *Ecol. Modell.* **409**, 108745. (doi:10.1016/j.ecolmodel.2019.108745)
11. Martinetti D, Soubeyrand S. 2019 Identifying lookouts for epidemio-surveillance: application to the emergence of *Xylella fastidiosa* in France. *Phytopathology* **109**, 265–276. (doi:10.1094/PHYTO-07-18-0237-FI)
12. Strona G, Carstens CJ, Beck PS. 2017 Network analysis reveals why *Xylella fastidiosa* will persist in Europe. *Sci. Rep.* **7**, 71. (doi:10.1038/s41598-017-00077-z)
13. Pearson RG. 2007 Species' distribution modeling for conservation educators and practitioners. *Synth. Am. Mus. Nat. Hist.* **50**, 54–89.
14. Desneux N *et al.* 2010 Biological invasion of European tomato crops by *Tuta absoluta*, ecology, geographic expansion and prospects for biological control. *J. Pest Sci.* **83**, 197–215. (doi:10.1007/s10340-010-0321-6)
15. Biondi A, Guedes R, Wan F, Desneux N. 2018 Ecology, worldwide spread, and management of the invasive South American tomato pinworm, *Tuta absoluta*: past, present, and future. *Annu. Rev. Entomol.* **63**, 239–258. (doi:10.1146/annurev-ento-031616-034933)
16. Karadjova O, Ilieva Z, Krumov V, Petrova E, Ventsislavov V. 2013 *Tuta absoluta* (Meyrick) (Lepidoptera: Gelechiidae): potential for entry, establishment and spread in Bulgaria. *Bulg. J. Agric. Sci.* **19**, 563–571.
17. Han P, Zhang YN, Lu ZZ, Wang S, Ma DY, Biondi A, Desneux N. 2018 Are we ready for the invasion of *Tuta absoluta*? unanswered key questions for elaborating an integrated pest management package in Xinjiang, China. *Entomol. Generalis* **38**, 113–125.
18. Tonnang HE, Mohamed SF, Khamis F, Ekesi S. 2015 Identification and risk assessment for worldwide invasion and spread of *Tuta absoluta* with a focus on Sub-Saharan Africa: implications for phytosanitary measures and management. *PLoS ONE* **10**, e0135283. (doi:10.1371/journal.pone.0135283)
19. Guimapi RY, Mohamed SA, Okeyo GO, Ndjomatchoua FT, Ekesi S, Tonnang HE. 2016 Modeling the risk of invasion and spread of *Tuta absoluta* in Africa. *Ecol. Complexity* **28**, 77–93. (doi:10.1016/j.ecocom.2016.08.001)
20. Venkatramanan S *et al.* In press. Modeling commodity flow in the context of invasive species spread: study of *Tuta absoluta* in Nepal. *Crop Prot.* (doi:10.1016/j.cropro.2019.02.012)
21. Lamperti F, Roventini A, Sani A. 2018 Agent-based model calibration using machine learning surrogates. *J. Econ. Dyn. Control* **90**, 366–389. (doi:10.1016/j.jedc.2018.03.011)
22. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. 2019 A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **51**, 1–42. (doi:10.1145/3236009)
23. You L, Wood-Sichra U, Fritz S, Guo Z, See L, Koo J 2017 Spatial Production Allocation Model (SPAM) 2005 v3.2. See http://mapspam.info.
24. Aldous D, Vazirani U. 1994 'Go with the winners' algorithms. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pp. 492–501. Santa Fe, NM: IEEE.
25. Breiman L. 2001 Random forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)
26. Kiss IZ, Green DM, Kao RR. 2006 The network of sheep movements within Great Britain: network properties and their implications for infectious disease spread. *J. R. Soc. Interface* **3**, 669–677. (doi:10.1098/rsif.2006.0129)
27. Pautasso M, Xu X, Jeger MJ, Harwood TD, Moslonka-Lefebvre M, Pellis L. 2010 Disease spread in small-size directed trade networks: the role of hierarchical categories. *J. Appl. Ecol.* **47**, 1300–1309. (doi:10.1111/jpe.2010.47.issue-6)
28. USDA. 2012 New pest response guidelines: tomato leafminer (*Tuta absoluta*). *Animal and Plant Health Inspection Service, Plant Protection and Quarantine*. See http://www.aphis.usda.gov/import_export/plants/manuals/emergency/downloads/Tuta-absoluta.pdf.
29. Banks NC, Paini DR, Bayliss KL, Hodda M. 2015 The role of global trade and transport network topology in the human-mediated dispersal of alien species. *Ecol. Lett.* **18**, 188–199. (doi:10.1111/ele.2015.18.issue-2)
30. Fadikar A, Higdon D, Chen J, Lewis B, Venkatramanan S, Marathe M. 2018 Calibrating a stochastic, agent-based model using quantile-based emulation. *SIAM/ASA J. Uncertain. Quantification* **6**, 1685–1706. (doi:10.1137/17M1161233)
31. Cassandras CG, Panayiotou CG, Diehl G, Gong W, Liu Z, Zou C. 2000 Clustering methods for multiresolution simulation modeling. In *Enabling Technology for*

Simulation Science IV, SPIE Proceedings, vol. 4026, Aerosense 2000, pp. 37–49. Orlando, FL: International Society for Optics and Photonics.

32. Moustier P, Renting H. 2015 Urban agriculture and short chain food marketing in developing countries. In *Cities and Agriculture–Developing Resilient Urban Food Systems* (eds H de Zeeuw, P Drechsel), pp. 121–138. Oxford, UK: Routledge.