05-Aug-2019

Dear Dr Adiga:

Your manuscript has now been peer reviewed and the reviews have been assessed by an Associate Editor. The reviewers’ comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. As you will see, the reviewers and the Editors have raised some concerns with your manuscript and we would like to invite you to revise your manuscript to address them.

We do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions”, click on "Create a Revision”. Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" - in the "File Upload" section. This should document, point by point, how you have responded to the reviewers’ and Editors’ comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as ‘tracked changes’ to be included in the ‘response to referees’ document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:

If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:

If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:

It is a condition of publication that you make available the data and research materials supporting the results in the article. Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:

All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,

Dr Sasha Dall

mailto: proceedingsb@royalsociety.org

Associate Editor

Board Member: 1

Comments to Author:

Two experts have now reviewed your paper “Assessing the Multi-pathway Threat from an Invasive Agricultural Pest: Tuta absoluta in Asia” (RSPB-2019-1159). Reviewer 1 is quite positive on the paper and indicates the need of only formal changes in the text, mostly aimed at clarifying methods and making a more complete review of the literature. Reviewer 2 is more critical, particularly on the models and related decisions. In my opinion all his/her comments are highly relevant, but you should pay particular attention those regarding the predictive nature of the study, and those related to CA models detailed in the third paragraph of the review.

**Response**: Thank you for considering this manuscript for revision. We found the reviewers’ comments to be constructive. We believe that we have addressed all the comments. Here are the main changes:

* We have provided more information on the analysis of the model, which uses machine learning algorithms in a novel manner. The figure in the supplementary material providing the outline of this process has been moved to the main part. Recent publications on interpretability of machine learning algorithms have been added to show that similar methods have been considered in other domains.
* In our original submission, we had dedicated two sections on predictions, and this has been made clear. A figure from the supplementary material has been moved to the main document to highlight this part.
* We have explained our viewpoint of CA models.

We thank the reviewers for their constructive comments. Our responses are given below.

# Referee: 1

Comments to the Author(s)

I found the manuscript very interesting and well-written. Data are presented in great detail (especially in the supplementary material), as well as the methodology and the results. I appreciate the huge effort in collecting and sharing such a complex body of data on international trades, climate, crop production and other relevant variables. The proposed methodology is very innovative and allows to discover new insights into the spread of Tuta absoluta and on possible ways of controlling the pest.

Invasive-species dynamics are hard to predict and model, especially when dispersal can be mediated by human transport over large regions. The authors consider a multi-pathway dynamic and actually find two possible spreading pathways: with and without long-distance human-mediated transport of the insect pest. The fitted models of short and long-distance transport (based on past detections in Bangladesh) are then applied to the rest of the study region in order to predict the future spread of Tuta absoluta. Finally, a simple control strategy is proposed (quarantine in high-outflow regions) and its effect is estimated by means of the constructed model.

The methodology used in the paper is rather new and involves a novel way of exploring the parameter space by means of surrogate machine learning techniques. My only concern is that such novel practices are rather unknown to the general public and may deserve more explanation. In my opinion Figure S5 in the Supplementary Materials could easily be included in the main text since it helps understanding the methodology.

**Response**: We agree. Due to space constraints, we had to make this decision. Figure S5 has been moved to Fig. 1c. More details have been added to it. Our survey of recent literature revealed similar use of CART in the emerging field of interpretable machine learning [Guidotti et al. 2019]. We have added this reference too. Due to these changes however, we exceeded the word limit. We had to shorten a part of the discussion.

Other few remarks are listed in the following:

1. In Section ‘Parametrization and experiment design’, right after Eq.4, the sentence about the use of CART model is rather misleading since it’s only after 1) selecting outputs with similarity over 75% and 2) clustering of outputs (as explained in S4 of SM) that the classification via CART is used. I suggest to remove it from there or explain the entire process before CART.

**Response**: We apply CART in two places: (a) for parameter exploration, motivated by the use of machine learning surrogates (e.g. [Lamperti et al. 2017]), and (b) cluster analysis as indicated in your comment. We have made this clear in Fig. 1c. A couple of sentences have been added in the methods (space was a constraint). The reference on interpretable AI [Guidotti et al. 2019] which talks about interpreting black box deep learning models using proxy decision-tree models. Thanks for this suggestion. Hopefully, this is much more clear now.

1. The advantage of the proposed methodology is that it allows to determine the influence of the different parameters of the model on the final output (seen as a machine learning problem where parameters are features or independent variables and the outcomes are the dependent one). Nonetheless, most reader will find this technique overly complicated w.r.t. a classical optimization (gradient descent, genetic algorithm or any other heuristic and meta-heuristic technique). In my opinion the authors should stress more the relevance of the proposed technique.

**Response**: We thank the reviewer for pointing this out. We had not explained this process in Methods. We have provided it in the revision. In principle, any surrogate that can explain the behaviour of the ABM is a candidate for such an analysis. Even gradient descent could be used if its results can be interpreted. It is our opinion that decision-tree based algorithms are not only powerful, but also provide good explanations. Further, we have chosen very popular and well-used models for this purpose. Such approaches are being used to analyse complex models such as ABMs [Lamperti et al. 2018] and artificial neural networks [Guidotti et al. 2019].

1. In Section ‘Influence of domestic trade on spread pattern and rate’, you find a rather well-known result in network theory, i.e. that hubs of the networks are facilitators of epidemic spread. There exists an extensive literature on the subject that should be cited.

**Response**: We thank the reviewer for bringing this up. We have added sentences with references in Introduction, Results and Discussion.

4. Towards the end of the first paragraph of the Discussion (page 13, before Section ‘Literature Survey’), you mention Nopsa et al. as one of the few works ‘[…] Identifying the optimal set of nodes in a network to reduce infectious disease’. This is not true. Lot of works in plant epidemiology based on a network approach have used ‘important’ nodes for developing surveillance and control strategies: just a couple of examples on another major plant pest (Xylella fastidiosa) are Strona et al. 2017 ‘Network analysis reveals why Xylella fastidiosa will persist in Europe’ and Martinetti at al. 2018 ‘Identifying Lookouts for Epidemio-Surveillance: Application to the Emergence of Xylella fastidiosa in France’.

**Response**: We agree with the reviewer. That sentence has been modified. Not only did we add the above-mentioned references, we also found few references that comment on surveillance and mitigation in directed networks. Thanks.

## Referee: 2

We thank the reviewer for the useful comments.

Comments to the Author(s)

This paper studies the relationship between exotic species invasions and trade and human mobility. This is a data-driven modelling approach to human-mediated spread which poses two interesting modelling challenges: (1) model complexity due to multiple drivers of epidemic dispersal and (2) model validation issues derived from existing data limitations. The paper is interesting and potentially relevant to the community of agricultural pest management. In particular, this seems like a useful means to visualise and recapitulate previous epidemics (a kind of “post-mortem” analysis) but it is somehow lacking in terms of predictions.

**Response:** Please see our response below about the concern on predictions.

An interesting novelty of the paper is the application of a machine learning framework to calibrate the model best model parameters (see ‘Analysis of spread pattern’, pp. 7). In effect, this applies a method originally develop in the community of agent-based models to the modelling of epidemic spread. My main concern is not the parameter fitting (but one should recall that the workings of clustering algorithms are difficult to understand) but the novelty of the modelling approach, which is unclear.

**Response**: To better explain the clustering/CART approach, we have moved the relevant figure from the supplementary material to the main paper. In addition, we have added more details to the figure. Hopefully, this helps the reader to understand the framework better. Also, due to lack of space, we could not provide definitions or descriptions of the clustering algorithms. However, these algorithms are very popular and implemented for most popular software such as python and R.

I found the modelling terminology a bit confusing. The model is called “stochastic multi-scale propagation” or “multi-pathway spread” but a more apt classification seems “old-school cellular automata” (CA) model of epidemic spreading. The authors should review similar CA models and tell us what is really new in this case. Perhaps merging the “literature review” section with the introduction can help in this regard.

**Response**: The terminologies “multi-pathway spread” and “multi-scale propagation” have been used in the invasive species literature before. For example, Douma et al. (2016) surveys different “pathway” models among which network propagation based methods are used in a significant number of works. Our model is “multi-scale” in nature like for e.g. Wildemeersch et al. (2019), who view the landscape as a network of habitats, and each habitat as a small-scale network. In our case, the propagation occurs at different spatial scales: (a) from one grid cell to another, (b) within a locality, and (c) between localities. Even though the functional forms for the probability of infection are similar, the pathways are distinguished by the structure of the neighbourhoods. All the above mentioned references are included in the paper. To summarize, (a) we have just adopted these terms from the existing literature on the invasive species, and (b) our models, even though similar in spirit to some of the existing works, are significantly different in structure.

We are well aware of the CA literature. Some of the authors of this manuscript have worked on their theoretical aspects extensively. For example, one of the authors of this manuscript has authored a book on dynamical systems - “Mortveit, Henning, and Christian Reidys. *An introduction to sequential dynamical systems*. Springer Science & Business Media, 2007.”, where, CAs and their extensions are discussed at length. The following is our take on the reviewer’s comment: A “traditional” CA typically consists of some form of uniform lattice, where a cell’s neighbours are those cells which share their boundaries with it. Our model in some sense can be viewed as a combination of CAs. The CA structure of the short distance pathway is different from the structure of the short-distance human-mediated spread. At the same time, the concept of CA being fairly generic, we admit that some researchers might view the entire model as a CA.

We have moved literature review from Discussion to Introduction. This was a very useful suggestion from the reviewer as the background for the modelling part is set right in the beginning.

In addition, some of the modelling decisions are unclear or require additional empirical support/justification. Each cell of the CA can be in one of three states: susceptible(S) => exposed (E) => infectious (I). Transitions from one state to another are governed by simple stochastic rules. For example, Infectiousness is a simple linear function of density. However, I cannot judge if this choice is realistic enough or not. Many complex systems show non-linear effects with varying population sizes. An empirical curve correlating infectiousness and density will support the current model (or some explanation of why this has been chosen).

**Response**: We have added a new section on model parameters and assumptions to address the reviewer’s concerns. Due to space constraints, we could not add this to the main document. In particular, this includes modelling infectiousness as a linear function of density. We have used results from a recent paper to support this choice [Sylla et al. 2018].

Another aspect is the long-distance dispersal kernel (see equation 1). This is different from the standard CA model, which presupposes an unrealistically local Moore neighbourhood (see Table 1).

**Response**: It is not clear to us which pathway the reviewer is referring to here since equation 1 refers to short-distance and not long-distance dispersal. Therefore, we will comment on both the pathways. For the short-distance pathway, we are considering Moore neighbourhood. Guimapi et al. (2016), in their CA approach to model *T. absoluta* spread use Moore neighbourhood. We chose to follow their modelling procedure in this regard. However, we agree that other methods could be incorporated (such as von Neumann neighbourhood or a more radial spread), and those might be more realistic. We will definitely consider this for the future. The long-distance pathway is discussed in the response to the following comment.

Again, it is unclear why the exponential form was chosen instead of other possibilities. Such exponential kernels are typical of gravity models employed for example in socio-economics. I concede this a more realistic approach than the Moore definition, but still it does not seem an accurate model for global trade networks, e.g., a power-law distribution (see paragraph starting at pp 5, line 11).

**Response:** For the long-distance pathway, there are two components: (i) The structure of the network based on the gravity model, and (ii) exponential kernel for the probability function from one locality to another.

* **The trade network structure**: Typically, we have found power-law distribution being associated with unweighted version of a network. Here, without the directionality and weights the network does not represent the domestic flows. Also, generally, fitting a network’s degree distribution to a power-law function is possible when the network is large. Our networks have a few hundred edges each. However, we recall that in the gravity model that we apply, the flows are inversely proportional to some power of time taken to travel from source to destination. The resulting network might resemble Kleinberg’s *small world network* [Kleinberg, NIPS 2001].
* **Spread through trade network**: We surveyed the literature on modelling spread through trade network. Researchers have applied many different methods to model this aspect. We have included this discussion in the same section in the supplementary material.
* **Exponential form of function probability**: We had provided explanation for the choice of this function form in the first version of the manuscript in the supplementary material. That explanation has been moved to “Model structure, parameter ranges and assumptions” in the supplementary material.

The analysis reveals the role of human-assisted spread of the pest of tomato, which quickly expands its range through city-to-city trade routes (see Fig 3). This is hardly surprising, but the analysis shows a potential to discriminate between different factors driving the epidemic spreading (e.g., role of markets).

In this context, it will be interesting to move from detailed reconstructions (and visualisations) of previous epidemics (like the present manuscript) to accurate predictions.

**Response**: We do predict the possible spread of the pest in the entire region as well as in the individual countries under different introduction scenarios. Please see the following paragraphs in results.

* Scenarios of pest introduction to countries in Southeast Asia: Here, we discuss at length the possible spread pattern in Mainland Southeast Asia.
* Predicted spread is model and region dependent: Here, we have assessed various counterfactual scenarios of domestic spread for each country.

However, considering the reviewer’s concerns, we have tried to make this clear. Firstly, the figure on the spread in Mainland Southeast Asia has been moved from supplementary material to the main document (Fig. 3a and 3b). Secondly, Fig. 2 and 3 titles have been changed to make it explicit that the former is about explanation of observed spread and the latter is about prediction. Hope this helps.