Received: 28 December 2020 | Revised: 8 March 2021 | Accepted: 22 March 2021

DOI: 10.1111/nrm.12307

RESEARCH ARTICLE

Natural Resource Modeling WILEY

A Bayesian model identifying locations at risk from human-transported exotic pathogens

Steven C. McKelvey[1] | Frank H. Koch[2] | William D. Smith[2] | Kelly R. Hawley[3,4]

[1]Department of Mathematics, Statistics and Computer Science, St. Olaf College, Northfield, Minnesota, USA

[2]Eastern Forest Environmental Threat Assessment Center, USDA Forest Service, Research Triangle Park, North Carolina, USA

[3]Department of Mathematics, Statistics and Computer Science, St. Olaf College, Richfield, Minnesota, USA

[4]Patterson Companies Inc., St. Paul, Minnesota, USA

Correspondence

Steven C. McKelvey, Department of Mathematics, Statistics and Computer Science, St. Olaf College, 1500 St. Olaf Ave. Northfield, MN 55057, USA.
Email: mckelvey@stolaf.edu and mckelveys@gmail.com

Funding information

U.S. Forest Service, Grant/Award Number: 08-DG-11083150-010

Abstract

A two-phase Bayesian model is presented for updating risk assessments for locations susceptible to infection by exotic pathogens. Human transportation from previously infected regions to uninfected regions is the main dispersal mechanism. Information embedded in patterns within the transportation flow are exploited in the update process. We explore the sensitivity of the model's outputs to changes in inputs. A sample application of the model to sudden oak death, using fictitious infection data, is performed.

KEYWORDS

Bayesian analysis, human-mediated pathways, infection detection and prevention, pathogen dispersal, *Phytophthora ramorum*, probabilistic network model

[Article updated on June 18, 2021 after first online publication: a citation to “Susan” needs to be changed to “Frankel” and “Species” name of tanoak has been changed from “Lithocarpus densiflorus” to “Notholithocarpus densiflorus”; Citations have been added for “APHIS Bulletin (2019)” and “Grünwald, et. al. (2019)”]

https://doi.org/10.1111/nrm.12307

1 | INTRODUCTION

With the increase in global trade the importation of nonnative insects, pathogens, and other organisms, has become an increasingly significant threat to ecosystems throughout the world (see Brasier et al., 2004; Costello et al., 2007; Garbelotto et al., 2001; Levine & D'Antonio, 2003; Mack et al., 2000; Roy et al., 2014). Geographic barriers often exist that limit the natural spread of these organisms but human activity frequently provides mechanisms by which they can circumvent these barriers (see Hulme, 2009; Hulme et al., 2008; Mack et al., 2000).

The probabilistic model described in this paper provides a structure for identifying regions that are currently uninfected by a pathogen of interest but which, due to human activity, are at high risk of importing the pathogen and potentially becoming infected. The model implements a generalized approach that can be applied to many organisms. The human activity considered here is commodity flow, a common transport mechanism for exotic organisms (see Davidson et al., 2002; Levine & D'Antonio, 2003; McCullough et al., 2006; Office of Technology Assessment U.S. Congress, 1993). Our methodology applies to a wide variety of commodity types. Risks to a set of presently uninfected regions are based upon the amount of relevant commodities shipped to these regions from areas known to harbor the pathogen. Partial survey results, indicating which uninfected regions from a surveyed subset of all uninfected regions are receiving infectious commodities, are also utilized in calculating risk rankings. The approach is Bayesian, meaning a priori probabilities are updated based on new information. In the case of our model, the new information is the set of recent partial survey results.

The goal of this model is to identify presently uninfected regions that are at high risk of becoming infected in the near future. The hope is that this information can focus resources and policies to prevent or mitigate future infection.

The contribution of this paper is to provide a formal and rigorous analysis to the task of allocating scarce inspection resources to regions at risk of becoming infected by a dangerous infectious agent. The mathematics is standard, the specific application is new and of broad importance.

As an example of this procedure, we consider the case of sudden oak death (SOD), a disease caused by the pathogen *Phytophthora ramorum* (Garbelotto & Rizzo, 2005). In North America, this infection is currently restricted to areas along the central and northern coasts of California and southwestern Oregon (Hansen et al., 2019).

The primary impact of the pathogen is significant tree mortality in several oak (*Quercus*) species as well as tanoak (*Notholithocarpus densiflorus*). Although *P. ramorum* actually has more than 100 host species, most are rarely killed. Instead, a variety of tree, shrub, and herbaceous species support sporulation and spore dispersal, thus playing a potentially large role in the epidemiology of the pathogen (Rizzo et al., 2003, 2005; Tooley et al., 2004). Furthermore, many of these hosts are routinely used in commercial and residential landscapes across the country, exemplified by rhododendrons, azaleas, and camellias. There is a substantial transcontinental wholesale trade in this nursery stock which may provide the mechanism for spreading the pathogen beyond its currently limited range (Davidson & Shaw, 2003). Indeed, during the last two decades, nursery outlets in many US states have received infected plants from West Coast operations (APHIS, 2019; Frankel, 2008; Grünwald et al., 2019; Stokstad, 2004; Warfield et al., 2008). Nevertheless, while suitable hosts and environmental conditions are found in many parts of the country (Kelly et al., 2007; Koch & Smith, 2008; Venette & Cohen, 2006), *P. ramorum* is not believed to be established in any wildland areas outside of California, Washington and Oregon.

Early detection of newly infected sites is critical for the control and irradication of pathogens like *P. ramorum* (Anderson et al., 2017; Chavez et al., 2016; Goss et al., 2009) providing the motivation for our work.

The organization of the paper is as follows:

Section2 lays out the general structure of the model without regard to any particular application.

Section 3 identifies issues common to models relying on Bayesian analysis with attention to how those issues arise in the model at hand. This section can be skipped by readers familiar with Bayesian models.

Section 4 details the general model presented in this paper. For those more interested in an application of the model, this section can be skipped on the first reading.

Section 5 presents an example of the model applied to a fictitious situation motivated by the threat of *Phytophthora ramorum* and Sudden Oak Death (SOD).

Section 6 presents analytical characteristics of the model, specifically sensitivity to initial parameter values and the memory and running time requirements of computer implementations of the model as functions of problem size.

2 | GENERAL MODEL METHODOLOGY

The goal of our probabilistic model is to focus scarce inspection resources toward the early detection of pathogen outbreaks in those regions where the pathogen of interest is not yet established. This is accomplished using a standard discrete random variable Bayesian model (Lee, 2012; Ross, 1997) incorporating partial survey results (Turner et al., 2020) and commodity flow information, to create an ordered list of those regions not presently known to be infected. The list is ordered by likelihood that each region has, in fact, been recently infected through importation of infectious material (e.g., nursery stock).

The process of creating this list consists of several stages. In the first stage some subset of vulnerable regions is surveyed. The surveyed regions are categorized as recently infected, uninfected, or regions for which infection status is uncertain. Regions with an uncertain infection status are subsequently treated as though they were not surveyed. The combination of newly infected regions and recently certified clean regions is called an *infection pattern*.

Once newly infected and known clean regions are identified, potential *sources* of infectious material are assigned probabilities of being the actual sources of infectious material. This is a Bayesian process in which the probability of infectious exports assigned to each potential source is updated from some previous value based on the observed infection pattern.

After the probabilities of exporting infectious materials have been updated, attention moves to the unsurveyed recipients of materials. For each unsurveyed recipient, hereafter called a *destination*, a probability is computed that the region has become recently infected. This probability is based on two characteristics of the destination; the sources of the destination's material and how much material comes from each source.

Once risks have been assigned to all unsurveyed destinations, inspection resources can be mobilized to high-risk destinations with the aim of identifying those regions that are, in fact, infected and actions can be taken to eliminate the threat of the imported pathogen before it becomes established.

3 | IMPORTANT ASSUMPTIONS AND CAVEATS

Bayesian models update probabilities as more data become available. In our case, the probabilities being updated by the model are the probabilities that some subset of sources are exporting infectious materials. As a starting point for the updating process, we must supply an initial estimated probability, or a priori probability, that each source is one that exports infectious material. Assuming independence, these source probabilities are combined to create probabilities for entire subsets of sources.

Typically, there is very little information upon which to base these a priori probabilities. One might choose to assign all sources the same a priori probability of being a source of infectious material. If there is reason to believe a particular group of sources is more likely to be exporting infectious materials than some other group, a user of this model could consider giving members of the riskier group a higher a priori probability of shipping infectious material.

Another probability that must be provided as an input to this model is a parameter that quantifies how the amount of material flowing from an exporter of infectious material to a destination affects the probability that the destination will become infected as a result of receiving that material. The *unit flow probability of infection* is the probability that a destination will become infected upon receiving a single unit of infectious material.

Precise values for the a priori probabilities of exporting infectious materials and the unit flow probability of infection are difficult to obtain. Fortunately, precision is not necessary. Because the goal of this model is to rank destination regions according to risk, what is important is the relative risk (i.e., which regions face greater risk than others) rather than precise risk values. This ranking turns out not be sensitive to the exact choices of a priori probabilities and the unit flow probability of infection, and thus reasonable estimates are sufficient for the model to serve its purpose.

In the case of *P. ramorum* we have confirmed the insensitivity of the model results to the initial probabilities of infection assigned to the sources. The risk-based rankings of destinations, as compared using the Spearman rank correlation coefficient (see Spearman, 1904), are detailed in Section 6.2. A similar sensitivity analysis for the unit flow probability of infection can be found in Section 6.3.

4 | THE MODEL

In this section, we provide the mathematical details of our Bayesian probabilistic model. The notation used in this section is summarized in Table 1.

4.1 | Model inputs

We begin by classifying physical regions as sources or destinations. A given region must be placed into exactly one of these categories. Sources are locations from which infectious materials may be exported. Destinations are locations that receive material from sources. The interaction between sources and destinations are represented by a bipartite network with links connecting every source with every destination. Flows on these links represent commodity transport of potentially infectious materials.

For each source/destination pair (s, d) we must determine the flow f_{sd} of material from s to d. Typical units measuring this flow are tons/year, or kilotons/year, although other units, such as monetary value/year, could also be used. For each source s, a value must be given for the a priori probability that s is, in fact, exporting infectious material. These a priori probabilities are denoted $P(J_s)$.

TABLE 1 Model notation

S	Set of sources
S'	A subset, proper or improper, of the set S of sources
s	A single source from the set S
D	Set of destinations
d	A single destination from the set D
J_s	The event that source s is exporting infectious material
$J_{S'}^{'}$	The event that the sources in S' are precisely the sources exporting infectious material
I_d	The event that destination d is newly infected
C_d	The event that destination d is known to be clean
I	The set of all newly infected destinations
C	The set of all destinations known to be clean through a survey
U	The set of all destinations for which infection status has not been recently determined by survey
PI	(I, C, U), an ordered triple of sets of destinations, known collectively as an infection pattern
N_{sd}	The event that destination d was recently infected by material from source s. The probability $P(N_{sd})$ is equal to the value of p_{sd} (see below)
f_{sd}	The amount of material, in units of mass per year, sent from source s to destination d
p_{sd}	The conditional probability that material from source s will cause infection at destination d given that source s is exporting infectious material. Generally this parameter is a function of the amount of material flowing from s and d
p	The unit flow probability of infection

An infection pattern PI must be specified as input. This represents a classification of destinations into three categories: recently infected (I), known clean (C), and status unknown (U). It is this infection pattern that drives both the updates of source infection probabilities and the assignment of risk to destinations with unknown infection status.

The last parameter that must be provided to the model is the unit flow probability of infection parameter, denoted p, which is described in the introduction.

4.2 | Phase 1: Updating probabilities of infectious exports

In Phase 1 of the model we use information gained by knowing the infection pattern PI to update the probabilities $P(J_s|PI)$ that each source s is actually exporting infectious material.

Rather than solve this problem directly, we initially tackle a related problem. Instead of computing the probability that a particular source s is a source of infectious material, we will consider every possible subset of sources and compute the probability that the subset of sources under consideration is precisely the collection of sources responsible for the observed infection pattern PI.

Standard notations for the power set of S, the set of all subsets of S, include 2^S and $\mathcal{P}(S)$. To emphasize the cardinality of the power set, we will use the notation 2^S in what follows.

Letting S' be a given subset of sources, we compute $P(J'_{S'}|PI)$ for every subset of S using Baye's Rule.

$$P(J'_{S'}|PI) = \frac{P(PI|J'_{S'})P(J'_{S'})}{\sum_{\hat{S}\in 2^S} P(PI|J'_{\hat{S}})P(J'_{\hat{S}})}. \tag{1}$$

Assuming independence, the second term in both the numerator and denominator of this fraction can be computed by multiplying the relevant probabilities. For any subset S' of sources,

$$P(J'_{S'}) = \left[\prod_{s\in S'} P(J_s)\right]\left[\prod_{s\notin S'} (1 - P(J_s))\right], \tag{2}$$

where the values $P(J_s)$ are precisely the individual a priori source probabilities (see Section 4.1).

For a particular infection pattern PI to be realized, it must be the case that each newly infected destination was infected by one or more sources and each known clean destination escaped infection. This observation leads to the equation

$$P(PI|J'_{S'}) = P\left(\left(\bigcap_{d\in I}\left(\bigcup_{s\in S'} N_{sd}\right)\right) \cap \left(\bigcap_{d\in C}\left(\bigcap_{s\in S'} N^c_{sd}\right)\right)\right), \tag{3}$$

where I is the set of infected destinations in PI, C is the set of known clean destinations in PI and N^c_{sd} is the complement of N_{sd}, the event that destination d was NOT infected by source s.

If we make the reasonable assumption that events N_{sd} are independent, then equation (3) becomes

$$P(PI|J'_{S'}) = \left[\prod_{d\in I} P\left(\bigcup_{s\in S'} N_{sd}\right)\right]\left[\prod_{d\in C}\prod_{s\in S'} (1 - P(N_{sd}))\right]. \tag{4}$$

Unfortunately, the events N_{sd} are not mutually exclusive so computing the probability of the union in (4) is laborious.

Lastly we compute the values of $P(N_{sd})$ which are also denoted p_{sd}. These values are based on the commodity flow data. For every source-destination pair (s, d), there is a flow f_{sd} of material from s to d. The value of $p_{sd} = P(N_{sd})$, given source s is infectious, is the probability that this flow of material will result in an infection of the previously uninfected destination d. To model this effect, we introduce p, the unit flow probability of infection. (See Section 4.1).

Under reasonable independence assumptions the resulting probability is

$$P(N_{sd}) = p_{sd} = 1 - (1 - p)^{f_{sd}}. \tag{5}$$

We can now use (1) to compute the updated probability that any subset S' of sources is the precise subset of sources that exports infectious materials. From here, it is straightforward to update the probability that any given source s is an exporter of infectious materials. This can be done by adding together the probabilities of infectiousness for every subset of sources that includes source s, yielding the values of $P(J_s|PI)$ for every source s.

TABLE 2 Summary of algorithm

Inputs
Set S of sources
The prior probabilities $P(J_s)$ for every $s \in S$
Set D of destinations
A partition of D into I (infected), C (clean) and U (uncertain) subsets
Commodity flows f_{sd} for every $s \in S, d \in D$
p, the unit flow probability of infection
Computations
Compute $P(N_{sd}) = p_{sd}$ for all $s \in S, d \in D$. (In SOD application $p_{sd} = 1 - (1 - p)^{f_{sd}}$)
Compute $P(PI\|J'_{S'})$ for every subset S' of S using Equation (4)
Compute $P(J'_{S'}\|PI)$ using equation (1), Baye's Formula
Compute $P(I_d\|PI)$ for every $d \in U$ using Equation (6)
Output
The quantities $P(I_d\|PI)$, the risks of undetected recent infection at destination d given the observed infection pattern

4.3 | Phase 2: Assigning infection probabilities to unsurveyed destinations

The goal of Phase 2 is to use the updated source infection probabilities to assign probabilities of infection $P(I_d|PI)$ to every destination d with an unknown infection status.

We can compute this value by conditioning on the specific source infection pattern.

$$\begin{aligned} P(I_d|PI) &= \sum_{S' \in 2^S} P((I_d|J'_{S'})|PI) P(J'_{S'}|PI) \\ &= \sum_{S' \in 2^S} P(I_d|(J'_{S'} \cap PI)) P(J'_{S'}|PI) \\ &= \sum_{S' \in 2^S} P\left(\bigcup_{s \in S'} N_{sd}\right) P(J'_{S'}|PI). \end{aligned} \tag{6}$$

After determining the probability of infection for each of the destinations for which infection status is currently unknown, we can order these destinations by risk of infection, identifying regions toward which inspection efforts might be productively focused (Table 2).

5 | APPLICATION TO SUDDEN OAK DEATH (*P. RAMORUM*)

Our model is applicable to any pathogen with a currently restricted distribution that could potentially be spread more widely by human activity. Its initial development was motivated by concern over the possible eastward spread of *P. ramorum*.

The relevant material being moved by human activity is wholesale nursery stock comprised, in part, of possibly infected *P. ramorum* host plants. While *P. ramorum* is only established in forest tracts in northern California and southern Oregon, the pathogen has been found in numerous commercial nurseries in both states, as well as in Washington (Tubajika et al., 2006). Infected plants have also been found at nurseries in more than 20 other US states, in many cases definitely traced to shipments from West Coast nurseries (California Oak Mortality Task Force, 2008; Grünwald et al., 2019). Thus, in the application of our model to sudden oak death, the source nodes are taken to be regions in California, Oregon, and Washington. All regions outside these three states are considered destination nodes.

The data representing the flows between source and destination nodes in the network are adapted from the Freight Analysis Framework[1] (FAF) Commodity Origin-Destination database (version 2.2). This relational database was created by the U.S. Federal Highway Administration to quantify the movement of commercial freight between major geographic regions in the United States. It is built upon publicly available data, most prominently the 2002 Commodity Flow Survey issued by the U.S. Bureau of Transportation Statistics, but it also incorporates specific data from other sources related to the movement of freight by water, air, and rail. In particular, the FAF database was developed using a modeling approach that estimated freight flows along certain shipment pathways not well represented by the Commodity Flow Survey. The FAF database consists of three four-dimensional matrices (for tons, ton-miles, and monetary value of shipments) in which the four dimensions are origin, destination, commodity, and transportation mode. More than 100 US geographic regions (i.e., metropolitan areas or, in some cases, partial or entire states) serve as origins and/or destinations of freight shipments. Shipments are reported for approximately 40 broad commodity categories.

For the nursery stock flow network, we used the FAF matrix describing tonnage flows. We extracted only those records associated with origin regions in CA (five regions), OR (two regions), and WA (two regions). We summed across all relevant modes of transport (truck, rail, water, air, and some mixed modes) the total tonnages of all commodities moving from each of these origin regions of interest to each destination region in the United States. We modified the total tonnage values with a scalar representing the proportion of the total tonnage that consists of nursery stock; we calculated this scalar using a relevant subset of data from the 2002 Commodity Flow Survey. The resulting network of nursery stock flow information consists of nine sources and 97 destinations. The network is shown in Figure 1.

With support from the USDA Forest Service, a software application was developed to implement the Bayesian model described in Section 2. The software was written using the Java 5 programming language developed by Sun Microsystems. The resulting software runs on a variety of platforms, including various versions of the Microsoft Windows operating system, the OS X operating system from Apple and many UNIX-like operating systems, including Linux. This application is called SODBuster. The program, its source code, and a user's guide are available from the author's web site.[2]

6 | MODEL CHARACTERISTICS

In this section we assess our model's characteristics, including running time, memory requirements, and the sensitivity of the model to unknown or poorly known model parameters. We explore these issues using an application of the model to sudden oak death.

[1] For more information, see the Freight Analysis Framework (FAF) Version 2.2 User Guide, published by the Federal Highway Administration in 2006.

[2] http://www.stolaf.edu/people/mckelvey/#SOD.

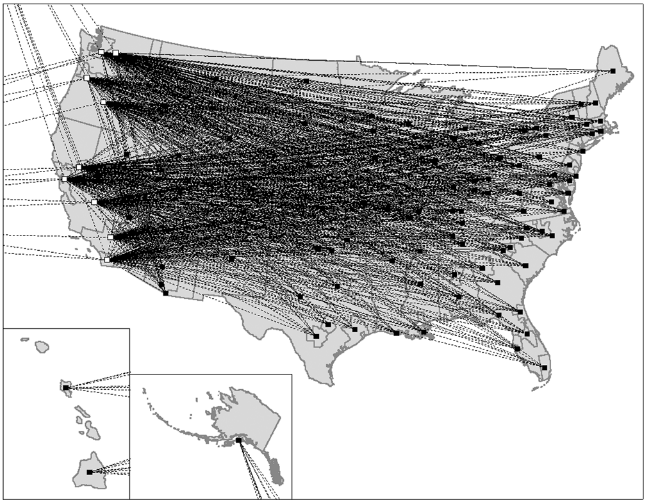

FIGURE 1 Nursery stock shipment network. Sources are denoted with white squares and destinations with black

6.1 | Sample application: Sudden oak death

As the basis for our investigation into the characteristics of our model we use the network commodity flow pattern from the sudden oak death application mentioned earlier. The nine regions that together make up the states of California, Oregon, and Washington serve as sources. All other regions in the United States (97 regions) are considered destinations. Commodity flows are calculated as described in Section 5.

A priori probabilities of infection for the nine sources were given initial values of 0.2. The unit flow probability of infection was given the value of 0.1. These values were chosen to demonstrate the characteristics of the model. They were not the result of any scientific study of nursery stock infectiousness at commercial nurseries. Once such studies are undertaken, these parameter values can be adjusted to reflect new scientific understanding.

The last data needed for the model are the results of surveying some subset of the destination regions. As of this writing, no survey resulting in a fully characterized inspection pattern has been undertaken. We created fictitious survey results for the purpose of testing our model. In the case of the baseline application, our survey consisted of results from 11 destination regions, four of which were found to be clear of infection while the remaining seven were classified as recently infected.

It is unfortunate that genuine field data is unavailable because the necessary studies have not been undertaken. One of the main benefits of mathematical modeling in general is the identification of important “missing” data to motivate future research. The articulation of this model provides rationale for undertaking the field work required to provide actual parameter values for the model. This is true not only for the case of Sudden Oak Death, but for other human-transported invasive organisms as well.

TABLE 3 Effect of Bayesian probability update

Node ID	Node name	A priori	Posterior
8	CA Los A	0.2000	0.9567
9	CA San D	0.2000	0.2223
10	CA Sacra	0.2000	0.2480
11	CA San J	0.2000	0.3060
12	CA rem	0.2000	0.5533
84	OR Portl	0.2000	0.2922
85	OR rem	0.2000	0.7424
109	WA Seatt	0.2000	0.3756
110	WA rem	0.2000	0.4085

Table 3 shows the updated probabilities that each source node is exporting infectious material. The magnitude of the update reflects the substantial effect of the partial survey results.

The top 10 riskiest unsurveyed regions, ranked according to their likelihood of infection, are shown in Table 4.

6.2 | Sensitivity to a priori probabilities of source infectiousness

The sensitivity of the model to changes in the a priori probabilities of infectiousness associated with each source is a key indicator of the model's effectiveness. The model's usefulness declines with the need for precise a priori probability values.

TABLE 4 The most at-risk unsurveyed destinations

Node ID	Node name	Prob. of infection
4	AZ Phoen	0.4702
59	NV LasV	0.4318
6	AZ rem	0.3824
28	ID	0.3307
60	NV rem	0.2503
53	MS	0.1930
98	TX Dalla	0.1846
102	UT Salt	0.1686
13	CO Denve	0.1389
41	LA rem	0.1375

The important results of our model are the reported relative likelihoods of infection at the unsurveyed destinations. To assess the robustness of our model with regard to the a priori probabilities of source infectiousness, we analyzed the ranked listings of destinations, sorted by likelihood of infection, that resulted from different a priori probabilities.

To perform this assessment, we used the same network and survey results as were discussed in section 6.1. Instead of assigning an a priori infection probability infectiousness value of 0.2 to each source, we created 20 instances of the problem. For each instance, we randomly selected nine values from a uniform distribution over the range from 0.01 to 0.99. These nine values were assigned to the nine sources as the a priori probabilities of infectiousness. In all cases the unit flow probability of infection was held fixed at $p = 0.1$. We then ran the model for each instance and noted the ordered listing of unsurveyed destinations, ranked by decreasing likelihood of infection.

Next we performed a pair-wise comparison of the orderings, computing the Spearman rank correlation coefficient for all 190 pairs. Figure 2 shows a histogram of the 190 rank correlation coefficients that resulted from this experiment. The correlation coefficient values ranged from 0.9552 to 0.9993, with a mean value of 0.9864 and a standard deviation of 0.0078.

The destination rankings appear quite stable in the face of widely differing values for the destination a priori probabilities. This should give managers confidence that the high-risk destinations are being reliably identified even if precision is not available for the a priori source infectiousness estimates.

6.3 | Sensitivity to unit flow probability of infection

The other model parameter value that is difficult to ascertain experimentally is the unit flow probability of infection, denoted p in our model. Recall that this is the probability that one unit of infectious material received by a previously uninfected destination will result in that destination becoming infected.

We performed a series of tests in which we left the a priori probabilities of source infectiousness fixed at 0.2 for all sources while varying the value of p. Twenty runs were undertaken with each run having a p value randomly chosen from a uniform distribution on the

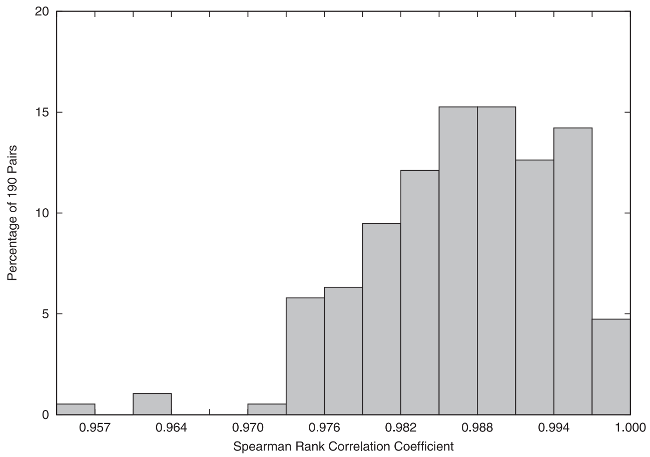

FIGURE 2 Distribution of rank correlation coefficients (p fixed)

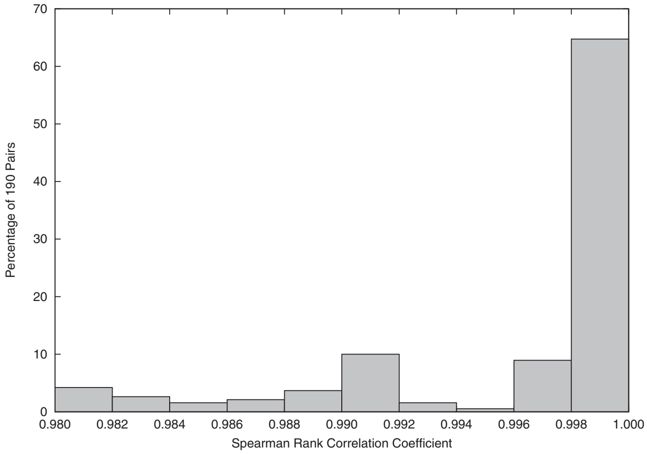

FIGURE 3 Distribution of rank correlation coefficients (varying p)

interval from zero to one. Pairwise Spearman rank correlation coefficients were computed for each of the resulting 190 pairs of rankings. A histogram depicting the distribution of these correlations is shown in Figure 3. These correlations fell in the range of 0.9814 to 1.0000, have a mean value of 0.9961 and a standard deviation of 0.0055.

6.4 | Computational issues: Running time and memory requirements

Memory usage and running time are two important characteristics of any algorithm. In the case of our model, both are linear in the number of destination nodes in the network, and exponential in the number of sources.

6.4.1 | Consequences of exponential memory and running time

The exponential characteristic of the model's resource consumption comes from the fact that in equation (1) a summation is taken over every subset of sources. The number of subsets of a given set is exponential in the number of elements in the set, specifically 2^n subsets exist for a set with n elements. As a consequence, for a fixed number of destinations we expect the running time and memory requirements of our model to double with each additional source.

The sudden oak death scenario consists of nine sources and 97 destinations. A contemporary desktop computer with common processor speeds and memory amounts can perform all the computations in a few seconds. However, running time becomes a serious issue as the number of sources increases.

Figure 4 shows the relationship between the number of sources and running time on networks very similar to the original sudden oak death example. (Destinations were reclassified as sources to increase the number of source regions in the trials). These times were attained using a quad-core (2.66 GHz) desktop PC with 8 gigabytes of RAM running the Linux operating system. An exponential curve is fitted to the data.

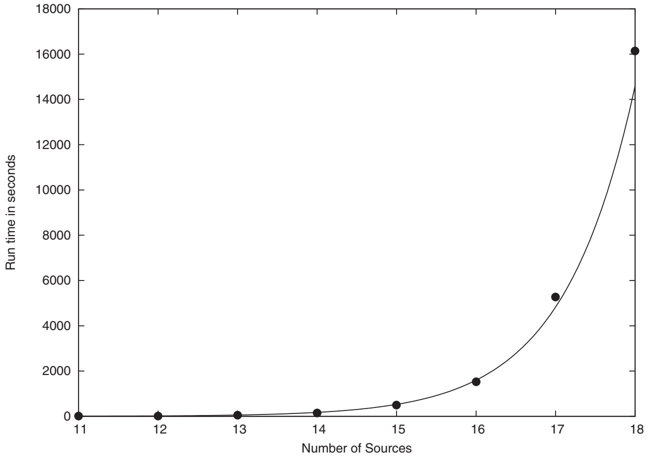

FIGURE 4 Running time versus source count

While the exponential character of the running time and memory requirements do not present a serious impediment to the use of our model in the case of sudden oak death, an application to a scenario with many sources would be problematic, even with access to a very powerful computing cluster or other advanced hardware.

6.4.2 | Polynomial memory and time through truncation approximation

Exponential use of time and memory resources is a serious drawback for any algorithm that aspires to address a wide range of scenarios. We are able to achieve polynomial run times and memory usage at the cost of approximating the sums given in Equations (1) and (6).

Adjusting the summation appearing in the denominator of (1) to include only smaller subsets of S, we derive the approximation

$$P(J'_{S'}|PI) \approx \frac{P(PI|J'_{S'})P(J'_{S'})}{\sum_{\hat{S}\in\mathbb{S}_c} P(PI|J'_{\hat{S}})P(J'_{\hat{S}})}, \tag{7}$$

where $\mathbb{S}_c$ is the subset of 2^S consisting of those elements of 2^S with cardinality c or less. The cardinality of $\mathbb{S}_c$ itself is

$$1 + n + \binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{c} = O(n^c).$$

Consequently, the number of terms in the summation, and the memory required to hold associated numerical values, is now polynomial in n, the number of sources.

Similarly, we can avoid exponential memory and running times associated with equation (6) by limiting consideration to subsets of size c or less in the summation, yielding the approximation

$$P(I_d|PI) \approx \sum_{S'\in\mathbb{S}_c} P\left(\bigcup_{s\in S'} N_{sd}\right) P(J'_{S'}|PI). \tag{8}$$

The rest of the Bayesian analysis remains unchanged.

The individual terms in the summations in (1) and (6) are small. Truncating the summations is unlikely to change the resulting relative risks in any important way. The terms being removed from the summations are probabilities that large subsets of S are precisely the sources of infection. One expects that, given ongoing careful monitoring, the number of sources emitting infectious materials will be small. Restricting our attention to the smaller subsets is a reasonable strategy.

Two important questions arise from the implementation of this approximation. First, does it actually result in improved running time, making the methodology more suitable for scenarios with more sources than in the sudden oak death example? Second, what effect does this truncation have on the accuracy of the results; are the highest-risk destinations still reliably identified when the truncated approximation is used?

To address the first question, we made slight modifications to the underlying network to allow us to consider situations with the number of sources ranging from 11 to 21. We noted the run time required for each of these situations, lastly fitting polynomials of order c to the running time results obtained when using c as the subset cutoff. Figure 5 shows the results for two cases, a cutoff value of $c = 5$ and $c = 7$. (Note the differing vertical scales). Tests involving

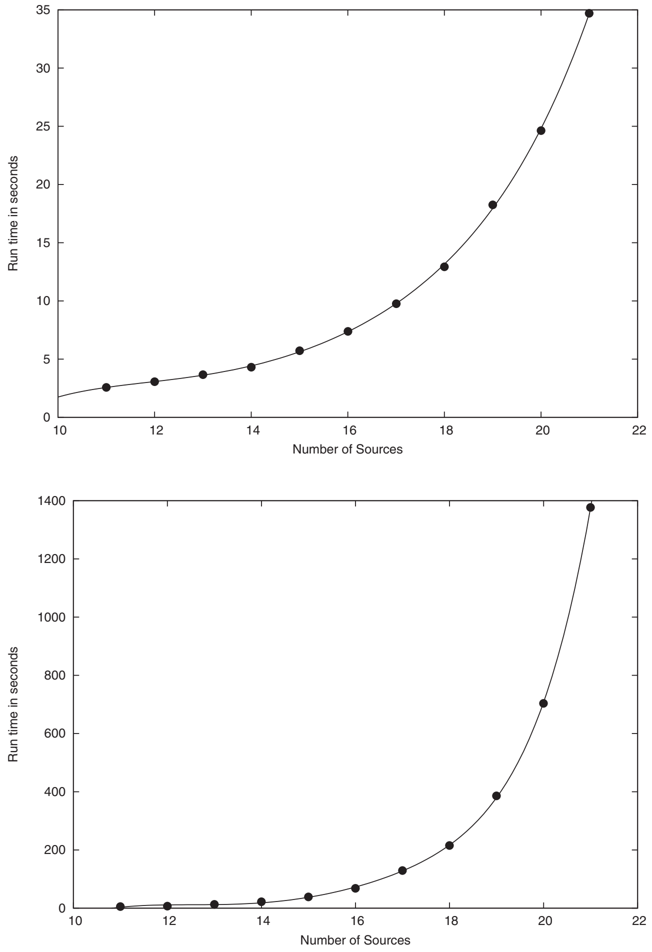

FIGURE 5 Polynomial fits to run time: Truncated model ($c = 5$ top, $c = 7$ bottom)

other values of c produced similar graphs. The empirical results are consistent with our assertion that a cutoff value of c allows us to achieve a run time that is $O(n^c)$. It is important to note that higher values of the cutoff c quickly result in substantial increases in the time needed to execute the computer application.

To assess the stability of the destinations' relative risks of infection with respect to various cutoffs c, we used the same networks as above with the number of sources ranging from 11 to 16. For these networks we created the true ordering of destination nodes, using the full model without series truncation. Next, we computed rankings using the truncated approximation with cutoff values of c ranging from 2 to 8. Lastly, we used the Spearman rank correlation coefficient to compare the destination rankings resulting from the truncation to the rankings achieved without truncation. The results are shown in Table 5.

Generally speaking, high cutoff levels yield rankings that are more consistent with the true rankings than the same network using lower cutoffs. This is consistent with the intuitive idea that using more of the available information should yield approximations closer to the true result. The table also indicates that a fixed cutoff level produces less consistency with true rankings as the number of sources increases.

The Spearman rank coefficient measures the consistency of ranking over the entire range, from lowest to highest. While this is important to understanding the mathematical characteristics of the various approximations, from a management point of a view a slightly different question is paramount. The manager wants to know if the approximations have the effect of moving high-risk destinations down the resulting risk ranking to an apparent level of moderate or low risk. In other words, the issue is whether the high-risk destinations remain near the top of the risk ranking produced by the approximate risk computation.

Consider the scatter plots shown in Figures 6 and 7.

Each of the 74 points in each plot corresponds to an unsurveyed destination. The axes represent the risk-based ordering of these destinations. For these plots the rankings are reversed, meaning a rank close to zero is a low-risk destination while a rank near 74 indicates a high-risk destination. The horizontal coordinate of each point represents the destination's risk as computed using the correct, untruncated version of the algorithm. The vertical coordinate represents the destination's risk as calculated using the truncated approximation. The dashed horizontal and vertical lines demarcate destinations in the riskiest 20%; destinations placed in

TABLE 5 Spearman rank correlation coefficients for truncated rankings

Source count	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$	$c = 7$	$c = 8$	No cut
11	0.974	0.992	0.997	0.999	0.999	0.999	1.000	1.000
12	0.904	0.976	0.980	0.982	0.982	0.982	0.982	1.000
13	0.905	0.975	0.979	0.981	0.999	0.980	0.980	1.000
14	0.903	0.909	0.911	0.912	0.956	0.910	0.910	1.000
15	0.868	0.840	0.843	0.995	0.887	0.999	0.843	1.000
16	0.841	0.811	0.815	0.955	0.866	0.962	0.825	1.000
17	0.908	0.692	0.699	0.807	0.742	0.821	0.712	1.000
18	0.904	0.708	0.713	0.812	0.756	0.824	0.728	1.000

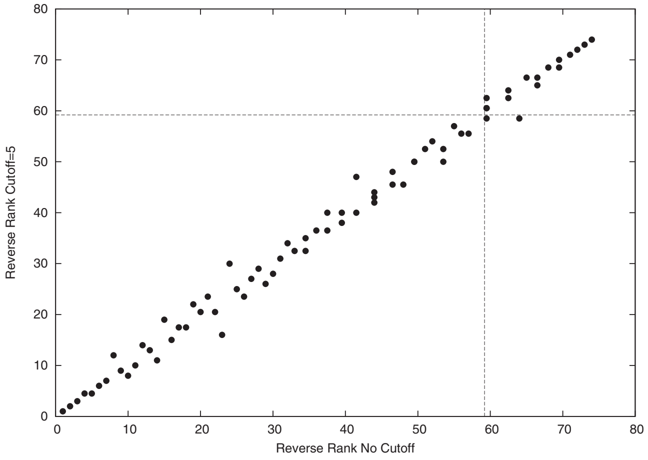

FIGURE 6 Risk scatter plot with 15 sources and cutoff $c = 5$

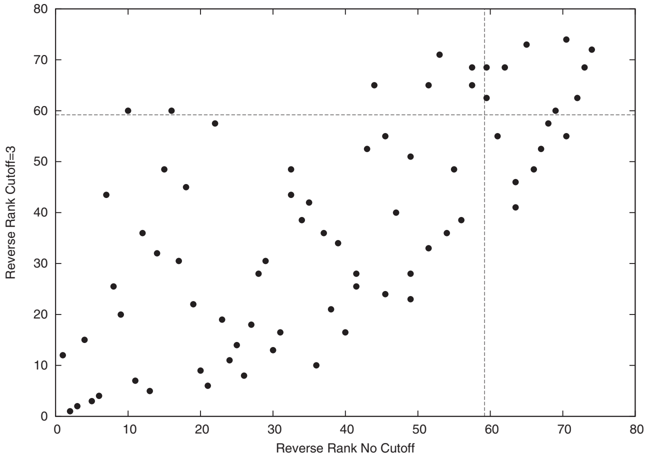

FIGURE 7 Risk scatter plot with 17 sources and cutoff $c = 3$

the top 20% by the truncated algorithm appear above the horizontal line, while destinations placed in the riskiest 20% by the nontruncated algorithm appear to the right of the vertical line.

Points appearing in the upper right hand quadrant of Figures 6 and 7 represent destinations that are in the riskiest 20% as ranked by both the truncated and nontruncated algorithms. As a manager's goal is to identify high-risk destinations, having many observations in this box indicates a manager can have some confidence that the truncated algorithm is effectively identifying truly high-risk destinations. The upper left-hand quadrant contains those destinations that are determined to be high risk by the truncated model, but are actually not in the riskiest 20% as determined by the nontruncated model. These destinations can be considered false positives in the sense that they are actually moderate- and low-risk destinations that the truncated algorithm falsely classifies as high risk. The lower right-hand quadrant contains destinations that are actually in the riskiest 20% but are classified by the truncated model as being in the lower 80% of the risk spectrum. These can be thought of as false negatives in the sense that they are truly high-risk destinations that are missed by the truncated algorithm. Lastly, the lower left hand quadrant contains destinations that are rated as low- or moderate-risk by both the truncated and nontruncated algorithms.

Figure 6 is typical of the situation where the Spearman rank correlation coefficient between the truncated and nontruncated results is near one. (In this example the coefficient is 0.9953). The figure shows a very small number of false positives and false negative rankings, which suggests that using the truncated version of the algorithm does not result in serious misclassifications. As a result, a manager could choose to take advantage of the computational quickness of the truncated algorithm with confidence that the resulting risk rankings are reasonably accurate classifications of the destinations' relative riskiness.

Figure 7 is typical of the situation where the Spearman rank correlation coefficient between the truncated and nontruncated results is far from one. (In this example the coefficient is 0.6992). The figure shows a substantially larger number of false positives and false negative rankings than Figure 6. Using the truncated version of the algorithm in such situations may result in serious misclassification of the risk associated with destinations.

7 | CONCLUSION

In this paper we have presented a model based on Bayesian probability to help focus inspection efforts in the face of human transported exotic pathogens. The model addresses situations where a pathogen exists in a limited number of well-known locations but is capable of moving to new locations via human activities. As survey results accumulate from previously uninfected areas, indicating that these areas have or have not become recently infected, our model produces lists of unsurveyed but vulnerable regions according to the risk each such region faces in light of the transportation flows of infectious material. These lists are ranked according to the likelihood of undetected new infection.

The numerical characteristics of our model are investigated through an application of the model to *P. ramorum*, the causal agent of sudden oak death. Commercial nursery stock can carry the pathogen to new areas. The transport of this nursery stock is the human activity of interest in this example.

The running time and memory requirements of the model are linear in the number of previously uninfected regions, but are exponential in the number of known sources of infectious material. In our application to the sudden oak death problem, the number of known sources is nine, and the model runs quickly on any modern desktop computer. However, the exponential time and memory requirements pose a serious limitation on the applicability of our technique to problems with a greater number of infectious sources.

In an attempt to allow application of our model to larger problems, we investigated the effect of approximating, through truncation, certain computations to reduce the memory and run time requirements. The resulting run times and memory requirements of the truncated model are polynomial in the number of sources, but there is some loss of accuracy in the model output. Several scenarios are presented to help understand the degree to which accuracy declines as running time and memory requirements are reduced through greater truncation.

The probabilistic framework presented in this paper gives managers a robust technique for identifying high-risk regions in the face of pathogens spread by human activity. In the case of sudden oak death, the model performs quickly and accurately. Our model can be applied to other pathogens as well, giving managers an important tool for slowing or eliminating the spread of harmful exotic pathogens.

Possible extensions of this study include updating the underlying network structure to include interactions more complicated than those representable through the bipartite network of supplies and demands used here. Using a more general network would allow for modeling of mixing during transport using transshipment nodes and common links in shipping patterns. Another interesting avenue for further research is the more explicit consideration of the geometry of the transportation network, including distance between nodes, proximity of nodes, etc. This introduces new spatial aspects to our model.

ACKNOWLEDGMENTS

The work upon which this publication is based was funded in whole or in part through a grant awarded by the Southern Region, State and Private Forestry, U.S. Forest Service. Award No: 08-DG-11083150-010.

AUTHOR CONTRIBUTIONS

S. C. McKelvey, lead researcher, defined the problem and developed probabilistic model; wrote SODBuster software and supporting documentation; and is the primary preparer of manuscript. F. H. Koch provided expertise on Sudden Oak Death, literature review, nursery stock transportation data and editorial advice on the manuscript and retrieved and cleaned all shipping data used in the SOD model example. W. D. Smith identified the need for inspection resource allocation modeling; assisted with the development of the probabilistic model; provided expertise on Sudden Oak Death and inspection processes; and gave significant editorial advice on the manuscript. K. R. Hawley performed significant literature review; assisted in the development of the paper's probabilistic model; tested the SODBuster software as it was being developed; designed and implemented the suite of statistical comparison tests used in discussing the robustness of the model in the face of computational simplifications; prepared all figures and tables; and proofread manuscript drafts.

ORCID

Steven C. McKelvey https://orcid.org/0000-0001-8971-2167

REFERENCES

APHIS. (2019). APHIS confirms detection of *Phytophthora ramorum-infected* plants in commerce. U.S. Department of Agriculture.

Anderson, D. P., Gormley, A. M., Ramsey, D. S. L., Nugent, G., Martin, P. A. J., Bosson, M., Livingstone, P., & Byrom, A. E. (2017). Bio-economic optimisation of surveillance to confirm broadscale eradications of invasive pests and diseases. *Biological invasions*, *19*(10), 2869–2884.

Brasier, C., Denman, S., Brown, A., & Webber, J. (2004). Sudden oak death (*Phytophthora ramorum*) discovered on trees in Europe. *Mycological Research News*, *108*(10), 1166–1175.

California Oak Mortality Task Force. (2008). *USA Phytophthora ramorum nursery chronology*. http://www.suddenoakdeath.org/html/chronology.html

Chavez, V. A., Parnell, S., & Van den Bosch, F. (2016). Monitoring invasive pathogens in plant nurseries for early-detection and to minimise the probability of escape. *Journal of Theoretical Biology*, *407*, 290–302.

Costello, C. M., Springborn, M., McAusland, C., & Solow, A. (2007). Unintended biological invasions: Does risk vary by trading partner? *Journal of Environmental Economics and Management*, *54*, 262–276.

Davidson, J., Rizzo, D., Garbelotto, M., Tjosvold, S., & Slaughter, G. (2002). *Phytophthora ramorum and sudden oak death in California: Ii. transmission and survival*. General technical report PSW-GTR-184, Forest Service, U.S. Department of Agriculture.

Davidson, J. M., & Shaw, C. G. (2003). Pathways of movement for *Phytophthora ramorum*, the causal agent of sudden oak death. In *Sudden Oak Death Online Symposium*. http://www.apsnet.org/online/proceedings/sod/pdf/shaw_davidson.pdf

Frankel, S. J. (2008). Sudden oak death and *Phytophthora ramorum* in the USA: A management challenge. *Australasian Plant Pathology*, *37*, 19–25.

Garbelotto, M., & Rizzo, D. (2005). A California-based chronological review (1995-2004) of research on *Phytophthora ramorum*, the causal agent of sudden oak death. *Phytopathologia Mediterranea*, *44*(2), 1–17.

Garbelotto, M., Svihra, P., & Rizzo, D. (2001). Sudden oak death syndrome fells three oak species. *California Agriculture*, *55*(1), 9–19.

Goss, E. M., Larsen, M., Chastagner, G. A., Givens, D. R., & Grünwald, N. J. (2009). Population genetic analysis infers migration pathways of *Phytophthora ramorum* in US nurseries. *PLOS Pathogens*, *5*, e1000583.

Grünwald, N. J., LeBoldus, J. M., & Hamelin, R. C. (2019). Ecology and evolution of the sudden oak death pathogen *Phytophthora ramorum*. *Annual Review of Phytopathology*, *57*, 301–321.

Hansen, E., Reeser, P., Sutton, W., Kanaskie, A., Navarro, S., & Goheen, E. M. (2019). Efficacy of local eradication treatments against the sudden oak death epidemic in oregon tanoak forests. *Forest Pathology*, *49*(4), e12530. https://doi.org/10.1111/efp.12530

Hulme, P. E. (2009). Trade, transport and trouble: Managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, *46*, 10–18.

Hulme, P. E., Bacher, S., Kenis, M., Klotz, S., Kuhn, I., Minchin, D., Nentwig, W., Olenin, S., Panov, V., Pergl, J., Pysek, P., Roques, A., Sol, D., Solarz, W., & Vila, M. (2008). Grasping at the routes of biological invasions: A framework for integrating pathways into policy. *Journal of Applied Ecology*, *45*, 403–414.

Kelly, M., Guo, Q., Liu, D., & Shaari, D. (2007). Modeling the risk for a new invasive forest disease in the united states: An evaluation of five environmental niche models. *Computers, Environment and Urban Systems*, *31*, 689–710.

Koch, F. H., & Smith, W. D. (2008). Mapping sudden oak death risk nationally using host, climate, and pathways data. In S. J. Frankel, & K. M. Palmieri (Eds.), *Proceedings of the sudden oak death third science symposium, number PSW-GTR-214 in General Technical Report* (pp. 279–287). US Department of Agriculture, Forest Service, Pacific Southwest Research Station.

Lee, P. M. (2012). *Bayesian statistics: An introduction*. John Wiley and Sons, Inc.

Levine, J. M., & D'Antonio, C. M. (2003). Forecasting biological invasions with increasing international trade. *Conservation Biology*, *17*, 322–326.

Mack, R. N., Simberloff, D., MarkLonsdale, W., Evans, H., Clout, M., & Bazzaz, F. A. (2000). Biotic invasions: Causes, epidemiology, global consequences, and control. *Ecological Applications*, *10*, 689–710.

McCullough, D. G., Work, T. T., Cavey, J. F., Liebhold, A. M., & Marshall, D. (2006). Interceptions of nonindigenous plant pests at US ports of entry and border crossings over a 17-year period. *Biological Invasions*, *8*, 611–630.

Office of Technology Assessment U.S. Congress. (1993). *Harmful non-indigenous species in the United States* (Number OTA-F-565) U.S. Government Printing Office.

Rizzo, D. M., Garbelotto, M., & Hanson, E. M. (2003). Sudden oak death: Endangering California and Oregon forest ecosystems. *Frontiers in Ecology and the Environment*, *1*, 197–204.

Rizzo, D. M., Garbelotto, M., & Hanson, E. M. (2005). *Phytophthora ramorum*: Integrative research and management of an emerging pathogen in California and Oregon forests. *Annual Review of Phytopathology*, *43*, 309–335.

Ross, S. M. (1997). *Introduction to probability models* (6th ed.). Academic Press.

Roy, B. A., Alexander, H. M., Davidson, J., Campbell, F. T., Burdon, J. J., Sniezko, R., & Brasier, C. C. (2014). Increasing forest loss worldwide from invasive pests requires new trade regulations. *Frontier in Ecology and the Environment*, *12*(1), 457–465. https://doi.org/10.1890/130240

Spearman, C. (1904). The proof and measurement of association between two things. *British Journal of Psychology*, *15*, 72–101.

Stokstad, E. (2004). Plant pathology-nurseries may have shipped sudden oak death pathogen nationwide. *Science*, *303*(5666), 1959.

Tooley, P. W., Kyde, K. L., & Englander, L. (2004). Susceptibility of selected ericaceous ornamental host species to *Phytophthora ramorum*. *Plant Disease*, *88*, 993–999.

Tubajika, K. M., Bulluck, R., Shiel, P. J., Scott, S. E., & Sawyer, A. J. (2006). The occurence of *Phytophthora ramorum* in nursery stock in California, Oregon and Washington states. *Plant Health Progress*. Advance online publication. *7*(1), 1–10. https://doi.org/10.1094/PHP-2006-0315-02-RS; http://www.plantmanagementnetwork.org/pub/php/research/2006/ramorum/

Turner, R. M., Plank, M. J., Brockerhoff, E. G., Pawson, S., Liebhold, A., & James, A. (2020). Considering unseen arrivals in predictions of establishment risk based on border biosecurity interceptions. *Ecological Applications*, *30*(8), e02194.

Venette, R. C., & Cohen, S. D. (2006). Potential climatic suitability for establishment of *Phytophthora ramorum* within the contiguous United States. *Forest Ecology and Management*, *231*, 18–16.

Warfield, C. Y., Hwang, J., & Benson, D. M. (2008). Phytophthora blight and dieback in North Carolina nurseries during a 2003 survey. *Plant Disease*, *92*(3), 474–481.

How to cite this article: McKelvey, S. C., Koch, F. H., Smith, W. D., & Hawley, K. R. (2021). A Bayesian model identifying locations at risk from human-transported exotic pathogens. *Natural Resource Modeling*, e12307. https://doi.org/10.1111/nrm.12307