

Probabilistic Techniques Used in SODBuster

Steven C. McKelvey, St. Olaf College

Submitted in partial fulfillment of the BaseEM project:

Probabilistic Commodity-Flow-Based Focusing of Monitoring Activities to Facilitate Early Detection of Phytophthora ramorum Outbreaks.

July 14, 2009

“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

1 Introduction

This technical paper describes the probabilistic model developed with funding support from the USDA Forest Health Monitoring Management Team and implemented in the SODBuster software. The software was developed using the Java software development kit from Sun Microsystems.

The goal of the probabilistic model and implementing software is to give the USDA Forest Service an analytical tool to help focus scarce inspection resources on the early detection of *Phytophthora ramorum* outbreaks in those parts of North America where *P. ramorum*, the organism that causes Sudden Oak Death (SOD), is not yet endemic. This is accomplished by using partial survey results, along with commodity flow information, to create an ordered list of those sites presently not known to be infected. The list is ordered by likelihood of each site having recently become infected through the importation of infectious nursery stock.

The process of creating this list consists of several stages. In the first stage some subset of vulnerable sites, in the case of *P. ramorum* these sites will typically be areas east of the Rocky Mountains, are surveyed. The surveyed sites are categorized as being recently infected, being very likely to be uninfected, or being a site for which infection status is uncertain. Sites with an uncertain infection status will be treated as though they were not surveyed. The combination of newly infected sites and recently certified clean sites is called an *infection pattern*.

Once newly infected and known clean sites are identified, known potential sources of infectious nursery stock are assigned probabilities of being sources of infectious nursery stock. In the terminology of probability theory 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 infection pattern observed. For example, those sources which happen to send a large amount of nursery stock to newly infected destinations will be assigned a high probability of exporting infectious materials because the new infections must have come from somewhere and the source sending materials to these destinations are good suspects. Similarly, sources that send large amounts of nursery stock to sites classified as known clean sites will be given a low probability of sending infectious exports because receiving these exports has not resulted in infection.

After the probabilities of exporting infectious materials have been updated attention moves to the unsurveyed recipients of nursery stock. For each unsurveyed recipient of nursery stock, called **destinations**, a probability is computed that this site has become recently infected. This probability is based on two characteristics of the destination, from which sources the destination’s nursery stock is sent and how much nursery stock comes from each source. If a given destination receives a significant amount of its stock from a high risk source, that destination will be assigned

a relatively high probability of infection. Conversely, if a destination receives very little stock from high risk sources, it will be assigned a low risk of infection.

Once risks have been assigned to the unsurveyed destinations, inspection resources can be mobilized to high risk destinations with the aim of identifying those sites that are, in fact, infected and actions can be taken to eliminate the threat of introducing *P. ramorum* into forests currently free of Sudden Oak Death (SOD).

The probability model described in the rest of this paper provides a careful and mathematically rigorous method of assigning these probabilities.

2 Important Assumptions and Caveats

The model described below is an example of a *Bayesian* model. Bayesian models update probabilities as more data become available. In our case the probabilities being updated by the model are the probabilities of a given source being one that is exporting infectious nursery stock. The very nature of updating something requires a starting point. In our case a model input is an initial probability, also called an *a priori* probability, of each source being one that exports infectious material.

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

Precise values for the *a priori* probabilities of exporting infectious materials and the unit flow probability of infection parameter are difficult to obtain. It is fortunate that precision is not necessary. Keeping in mind that the goal of this model is to rank destination (not yet infected) sites according to risk, what is important are the relative risks, which sites have greater risks than others, rather than the precise value of the risks. This ranking is not particularly sensitive to the exact choices of *a priori* probabilities. Choosing reasonable values is all that is required for this model to correctly perform its task.

The worst case running time and memory requirements of this model are exponential in the number of sources. This arises from the model's dependence on quantities associated with every subset of the sources. If a set contains n items, the set has 2^n subsets. Equations (1) and (5) show examples of this dependence.

Users can expect the running time and memory requirements of the model to approximately double for each additional source. The nationwide test data included with the SODBuster software uses nine sources and results in a model that requires just a few seconds to run. The exponential characteristic of this model's running time means careful consideration must be given before increasing the resolution of the model by reducing the size of source regions. Such a change would increase the number of sources, drastically increasing the running time and memory requirements of the model.

The output produced by the model includes updated (posterior) probabilities of sources exporting infectious nursery stock, the probabilities of infection for unsurveyed destinations and a map representing the relative risks of infection at each unsurveyed destination site.

The PDF version for this technical report can be found at the following URL:

3 The Model

Here we describe the mathematical details of the SODBuster probabilistic model. The notation used in this section is summarized in Figure 1.

3.1 Model Inputs

The first step in using the SODBuster probabilistic model is to classify physical areas as sources or destinations. A given region must be placed into exactly one of these categories. Sources are locations from which *P. ramorum* may be exported on outbound shipments of commercial nursery stock. Destinations are uninfected locations that receive nursery stock from sources.

For each source/destination pair (s, d) we must determine the flow of nursery stock from s to d . Typical units measuring this flow are tons/year, or kilotons/year. It is fine for such a flow to be zero, but no flow should take on a negative value. In the model these flow values are denoted by the symbol f_{sd} .

For each source s a value must be given for the a priori probability that s is exporting infectious nursery stock. In the model these a priori probabilities are denoted $P(J_s)$. For the reasons discussed in the introduction, the precise values of these probabilities are not especially important, but some care should be taken to make sure relative risks among sources are reflected in the a priori probability values.

An infection pattern PI must be specified. 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 probability of infection parameter, denoted p in the model, which is described in the introduction. As with the a priori source infection probabilities, a highly precise value for this parameter is not necessary as the final results are not strongly sensitive to the value.

3.2 Phase 1: Updating Probabilities of Infectious Exports

In the first step of the model we use information gained by knowing the infection pattern IP to update the probabilities of each source s being a source that is exporting infectious nursery stock. In the probabilistic notation of the model we seek to compute for each source s the value $P(J_s|PI)$, the conditional probability of J_s given the observed infection pattern PI .

Instead of tackling this problem directly, we tackle a related problem that will, eventually, give us the answer to the problem above. Instead of computing the probability that a particular source s is a source of infectious material, we will consider every possible subset of all the sources and compute the probability that the subset of sources is precisely the collection of sources that are exporting infectious materials.

If we let S' be a subset of sources, we are seeking to compute $P(J'_{S'}|PI)$ for every subset of S . The notation for all subsets of S is 2^S .

Applying Bayes Rule to this conditional probability gives us a relatively straight forward computation.

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

$P(A)$	The probability of event A .
$P(A B)$	The probability of event A given event B .
$P(A \cup B)$	The probability of event A or event B or both.
$P(A \cap B)$	The probability of both event A and event B .
S	Set 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 nursery stock.
$J'_{S'}$	The event that the sources in $S' \subseteq S$ are precisely the sources exporting infectious materials.
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, using in units of biomass 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.

Figure 1: Notation Used In The Probabilistic Model

The second term in both the numerator and denominator of this fraction is based on the a priori probability that is an input parameter to the model. Assuming that the infectious status of the sources are independent, these terms can be computed by multiplying together the relevant probabilities. In particular, for any subset S' of sources,

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

where the values $P(J_s)$ are precisely the a priori source probabilities that are input parameters to the model.

Now we move to the first term, the conditional probability. For a particular infection pattern PI to be realized it must be the case that each newly infected destination was infected by some source and each known clean destination escaped infection. This is reflected in probabilistic notation by the equation

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

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

Equation (3) can be interpreted as saying infection pattern PI arises from a set S' of infectious sources precisely when every destination in I is infected by *at least* one infectious source and every destination in C is infected by no infectious sources.

If we make the reasonable assumption that the infection of a destination d by some source s is independent of that same destination being infected by some other infectious source, then equation (3) can be simplified as follows

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

The probability of the union in the first term of equation (4) can be computed, laboriously, using the elementary rule for the probability of unions, namely

$$\begin{aligned} P(E_1 \cup E_2 \cup \dots \cup E_n) &= \sum_i^n P(E_i) \\ &\quad - \sum_i^n \sum_{j:i < j} P(E_i \cap E_j) \\ &\quad + \sum_i^n \sum_{j:i < j} \sum_{k:j < k} P(E_i \cap E_j \cap E_k) \\ &\quad - \dots + (-1)^{n+1} P(E_1 \cap E_2 \cap \dots \cap E_n) \end{aligned} \quad (5)$$

The independence of the events N_{sd} means the probabilities of the intersections in the above can be computed as products of the probabilities $P(N_{sd})$.

The last quantities to compute for Phase 1 are the values of $P(N_{sd})$ which are also denoted p_{sd} in our model. This is where the nursery stock flow data enters the computation.

For every source-destination pair (s, d) , there is a flow f_{sd} of material from s to d . This flow is never negative, but can be zero. 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 . The greater the flow, the more likely it is that infection will take place.

To model this effect we introduce the unit flow probability of infection, denoted p . We assume the value of p is strictly between zero and one.

Under reasonable independence assumptions, the resulting probability when f_{sd} units of infectious material are flowing from source s to destination d is

$$P(N_{sd}) = p_{sd} = 1 - (1 - p)^{f_{sd}}. \quad (6)$$

It is easy to see that when $f_{sd} = 0$, the probability of infection is zero. When $f_{sd} = 1$, the probability of infection is exactly the value p . As the flow f_{sd} grows without bound, the probability of infection approaches 1, as it should.

We have now described all the computations needed to fill in the values on the right hand side of equation (1). Equation (2) gives us the second term in the numerator and denominator while equation (4) gives us the first term. This means we can update the probabilities of each subset of sources as being the precise subset of sources that export infectious nursery stock.

From here, it is easy to update the probability that any given source s is an exporter of infectious nursery stock. This can be done by adding together the probabilities of infectiousness for every subset of sources that includes source s .

3.3 Phase 2: Assigning Infection Probabilities to Unsurveyed Destinations

At the end of Phase 1 we have computed the probability $P(J'_{S'}|PI)$ for every subset S' of S . Recall that S is the set of all sources. The goal of Phase 2 is to use these updated source infection probabilities to assign probabilities of infection to every destination d with an unknown infection status. In mathematical notation, we seek to compute values for $P(I_d|PI)$ for every destination d in U .

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

$$\begin{aligned} P(I_d|PI) &= \sum_{S' \in 2^S} P((I_d|J'_{S'})|PI) P(J'_{S'}|PI) \\ &= \sum_{S' \in 2^S} P(I_d|(J'_{S'} \cap PI)) P(J'_{S'}|PI) \\ &= \sum_{S' \in 2^S} P\left(\bigcup_{s \in S'} N_{sd}\right) P(J'_{S'}|PI). \end{aligned} \quad (7)$$

The first term of equation (7) can be computed using the identities previously shown in equations (5) and (6). The second term is an instance of the probabilities we calculated in Phase 1, starting with the application of Bayes Rule in equation (1).

Now that we have a probability of infection value for each of the destinations for which we have no survey information, we can order these destinations by risk of infection, identifying sites toward which inspection efforts might be productively focused.