Chapter 17. Probabilistic Commodity-Flow-Based Focusing of Monitoring Activities to Facilitate Early Detection of *Phytophthora ramorum* Outbreaks

(Project SO-R-08-01)

Steven C. McKelvey[1]
William D. Smith[2]
Frank Koch[3]

Introduction

This project summary describes a probabilistic model developed with funding support from the Forest Health Monitoring Program of the Forest Service, U.S. Department of Agriculture (BaseEM Project SO-R-08-01). The model has been implemented in SODBuster, a stand-alone software package developed using the Java software development kit from Sun Microsystems.

The goal of the probabilistic model and implementing software is to give the 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. Through the use of trace-forward information regarding the shipment of *P. ramorum* infected nursery stock provided by the Animal and Plant Health Inspection Service (APHIS), U.S. Department of Agriculture, supplemented by commodity flow data from the U.S. Departments of Commerce and Transportation, the analytical techniques and software identify areas with the greatest likelihood of new *P. ramorum* infestation, thus increasing the likelihood of successful intervention before the pathogen crosses the urban-forest interface. Briefly, this is accomplished by using partial survey results and 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.

Methods

The process of creating this ordered list consists of several stages. In the first stage a subset of sites vulnerable to infection by *P. ramorum* is surveyed. These sites will typically be areas east of the Rocky Mountains. As detailed below, the physical boundaries of these sites are defined by the U.S. Department of Transportation and can range in size from dozens of square miles to entire States. The surveyed sites are categorized as being recently infected, very likely to be uninfected (clean), or as sites for which current infection status is uncertain. Sites with an uncertain infection status are subsequently treated as though they were not surveyed. The combination of newly infected sites and recently certified clean sites is called an infection pattern.

[1] Associate Professor of Mathematics and Computer Science, Saint Olaf College, Northfield, MN 55057. mckelvey@stolaf.edu.

[2] Research Scientist, U.S. Department of Agriculture, Forest Service, Southern Research Station, 3041 Cornwallis Rd., Research Triangle Park, NC 27709. bdsmith@fs.fed.us.

[3] Research Ecologist, U.S. Department of Agriculture, Forest Service, Southern Research Station, Research Triangle Park, NC 27709. fhkoch@fs.fed.us. (Formerly Research Assistant Professor, North Carolina State University, Department of Forestry and Environmental Resources).

Once newly infected and known clean sites are identified, potential sources of infectious nursery stock are assigned probabilities of being active sources of infectious nursery stock. In the terminology of probability theory this is a Bayesian process in which the *a priori* 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 originated somewhere and the sources sending the most materials to these destinations are obvious suspects. Similarly, sources that send large amounts of nursery stock to sites known to be clean 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 (hereafter called destinations), a probability is computed that this site has become recently infected. This probability is based on two characteristics of the destination: the sources from which the destination's nursery stock was 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 relatively 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 any sites that are, in fact, infected and taking action to eliminate the threat of introducing *P. ramorum* into forests currently free of Sudden Oak Death.

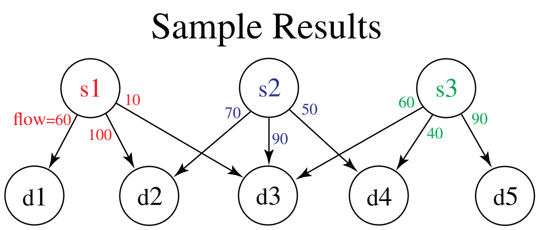
The data representing nursery stock flows between source and destination sites are adapted from the Freight Analysis Framework (FAF) Commodity Origin-Destination database (version 2.2). This relational database was created by the U.S. Federal Highway Administration to quantify the movement of commercial 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. More than 100 geographic regions across the United States (i.e., metropolitan areas or, in some cases, partial or entire States) serve as the source or destination sites for the nursery stock flow data.

The SODBuster software package and underlying analytical methodology are best highlighted with a hypothetical example. Consider a fictitious network with three sources (S1, S2, and S3) and five destinations (D1, D2, D3, D4, D5). Suppose the three sources have been assigned *a priori* probabilities of exporting infectious nursery stock of 0.30 (S1), 0.40 (S2) and 0.25 (S3). These *a priori* probabilities might be based on a simple criterion such as the sites' relative numbers of APHIS-regulated wholesale nurseries (i.e., nurseries that ship plant stock associated with *P. ramorum*). Nonzero annual flows, by weight, of nursery stock are given by S1→D1=60, S1→D2=100, S1→D3=10, S2→D2=70, S2→D3=90, S2→D4=50, S3→D3=60, S3→D4=40, S3→D5=90. Furthermore, it is assumed that one unit of infectious nursery stock has a probability of 0.01 of causing an infection at the receiving site.

In this scenario, suppose officials conduct a survey of three of the five destinations. The survey shows that destinations D1 and D5 are shown to be clear of *P. ramorum* infection while destination D3 is infected.

The hypothetical data laid out so far are inputs to the SODBuster model. The next step is to run the model and examine the results. The model creates, for each of the three sources, a new (*a posteriori*) likelihood that the source is exporting material capable of infecting receiving sites. These updated likelihoods are 0.226, 0.854, and 0.240 for sources S1, S2, and S3, respectively. These updated values are mathematically rigorous and intuitively pleasing. D1, known to be uninfected, receives all its nursery stock from S1, suggesting the material being sent from S1 is not infectious. The cleanliness of D5 suggests that S3 is not a source of infectious material. D3, the sole known infected site, receives a lot of material from S2 and S3. We have already argued that the survey results suggest S3 is not a source of infectious materials, but the results offer no analogous shelter for S2. It is not surprising that S2 earns the highest likelihood of being a source of infectious material.

Armed with the updated source likelihoods, we move back to the unsurveyed destinations and see that destination D2 has a 0.521 likelihood of being infected, while destination D4 has a 0.400 likelihood of being infected. These values reflect the fact that D2 receives more material from a likely source of infection (S2) than does D4 (fig. 17.1).



Source	A priori	A posteriori
s1	30.00%	22.62%
s2	40.00%	85.42%
s3	25.00%	24.03%

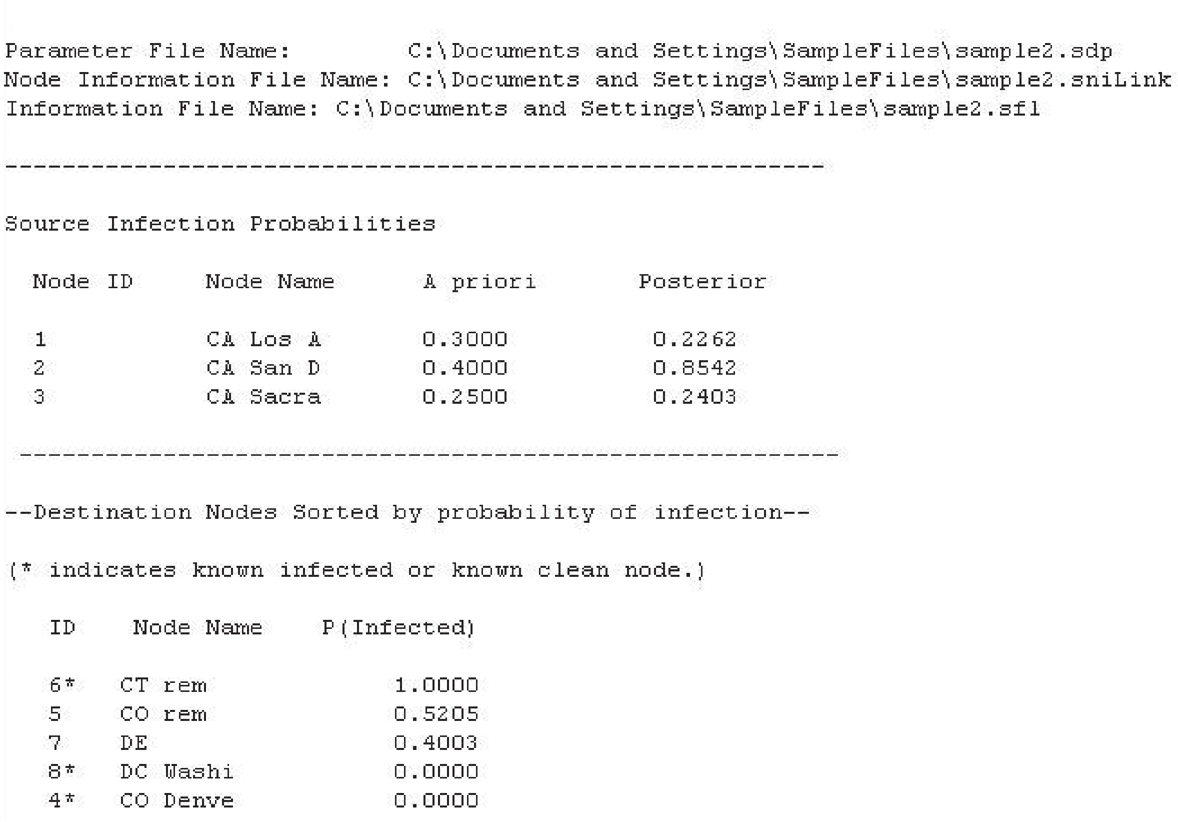
Sink	Status	Pr(Inf)
d1	Clean	0.00%
d2	?	52.05%
d3	Inf.	100%
d4	?	40.03%
d5	Clean	0.00%

Figure 17.1—Sample results for the SODBuster example described in the Methods section.

Results

This model has not yet been applied to genuine *P. ramorum* survey information, so we are not able to provide the results of a true application of the model. The software produces results that fall into two broad categories. The first, textual results, are displayed within a graphical window. Values for all input parameters are presented so the user can easily determine which output goes with each collection of parameter values. The output also presents the *a posteriori* likelihoods of each source being, in fact, a source of infectious materials. Lastly, the output presents the risk of each destination site being infected and sorts this list in various ways (fig. 17.2). The listings can be saved for subsequent printing to a hard copy by any text editor. If the program's user so chooses, these data can be organized into a comma-delimited text file designed to be easily imported into a standard spreadsheet program for further processing and analysis.

The non-textual output consists of a series of maps of the continental United States. The first of these maps show the underlying commodity transportation network superimposed upon the various FAF regions (fig. 17.3). The second (fig. 17.4) uses solid red discs of various sizes to indicate the relative risks of destinations. Both of these maps can be saved as PNG images for further processing, subsequent inclusion in documents, etc.



```
Parameter File Name:        C:\Documents and Settings\SampleFiles\sample2.sdp
Node Information File Name: C:\Documents and Settings\SampleFiles\sample2.sniLink
Information File Name: C:\Documents and Settings\SampleFiles\sample2.sfl

---------------------------------------------------------

Source Infection Probabilities

  Node ID     Node Name      A priori       Posterior

  1           CA Los A       0.3000          0.2262
  2           CA San D       0.4000          0.8542
  3           CA Sacra       0.2500          0.2403

 ---------------------------------------------------------

--Destination Nodes Sorted by probability of infection--

(* indicates known infected or known clean node.)

   ID    Node Name    P(Infected)

   6*   CT rem              1.0000
   5    CO rem              0.5205
   7    DE                  0.4003
   8*   DC Washi            0.0000
   4*   CO Denve            0.0000
```

*Figure 17.2—Sample output information from the SODBuster computer program. Note values of input parameters (*a priori*) are presented as well as the likelihoods of each source being a source of infection as determined by the model (*a posteriori*).*

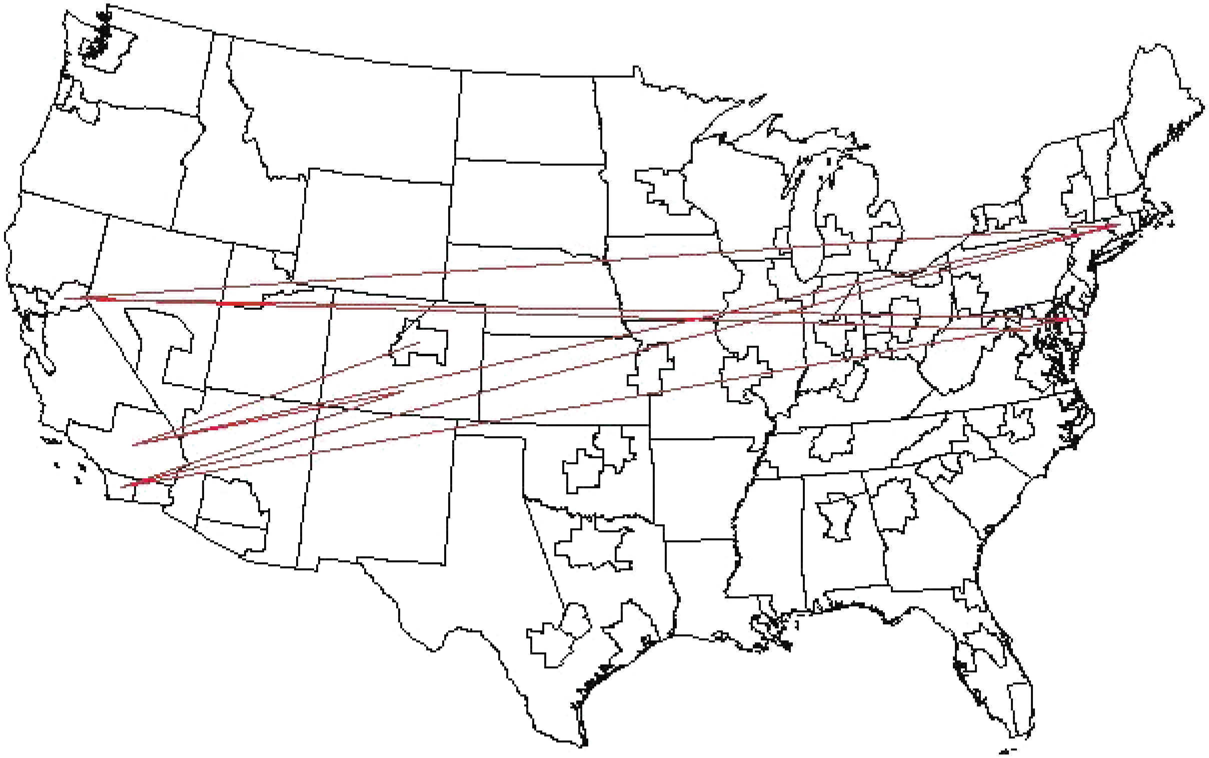


Figure 17.3—Example map output from the SODBuster computer program showing the underlying commodity transportation network (red lines) superimposed upon the various Freight Analysis Framework (FAF) regions. (Data source: U.S. Federal Highway Administration)

Discussion

There are a number of important assumptions and caveats with respect to the probabilistic model underlying the SODBuster software. Bayesian models update probabilities as more data become available. In our case, the probabilities 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 one 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. However, if there is reason to believe a particular group of sources is more likely to

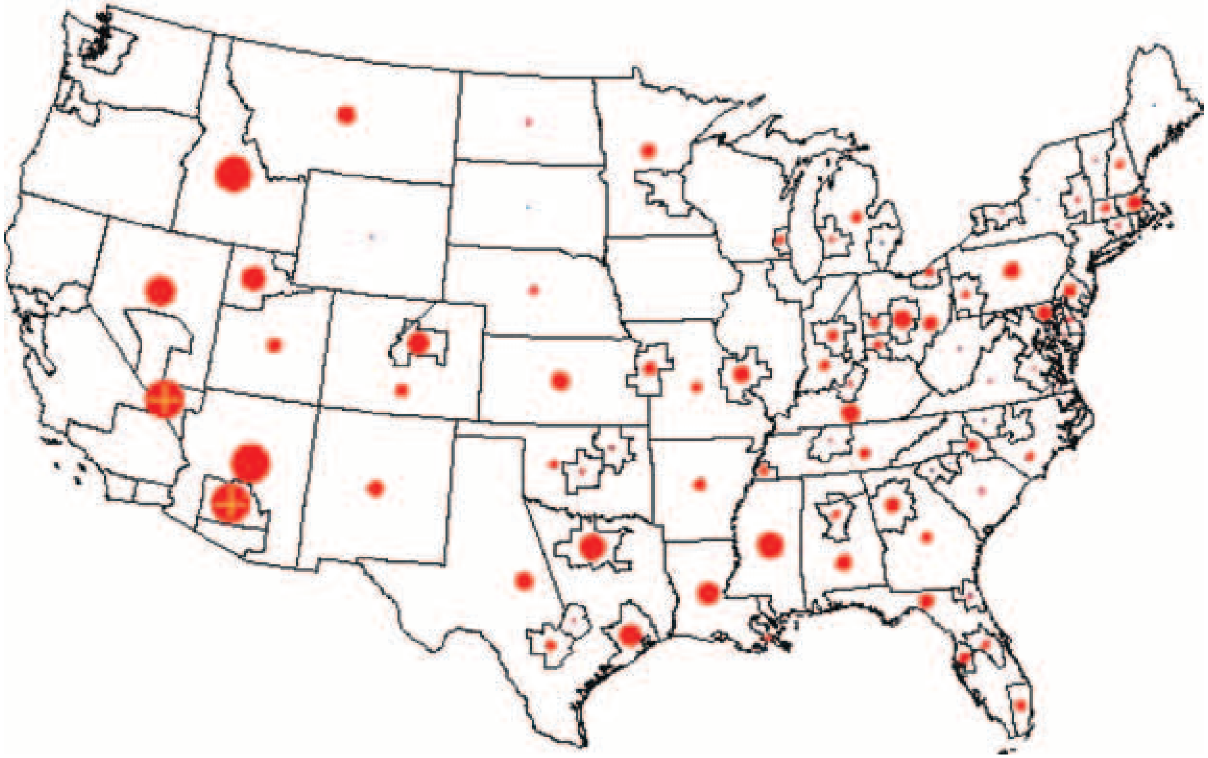


Figure 17.4—Example map output from the SODBuster computer program showing solid red discs to indicate the relative risks of destinations, superimposed upon the various Freight Analysis Framework (FAF) regions. (Data source: U.S. Federal Highway Administration)

be exporting infectious nursery stock than some other group, a model user could consider giving members of the riskier group higher *a priori* probabilities of sending infectious material. Generally, *a priori* probabilities should reflect known information regarding the relative risks of various sources.

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 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. Fortunately, precision is not necessary. Keeping in mind that the goal of this model is to rank not yet infected destinations 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. We expect that 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 related exponentially to the number of included source sites. 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. For example, the set {1,2,3} has 2^3 subsets, specifically {} (the empty set), {1}, {2}, {3}, {1,2}, {1,3}, {2,3}, and {1,2,3}. The model performs summations over all of these subsets. 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. 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.

A more detailed description of the probability model and a users' guide for the SODBuster software can be found on the following Web sites:

- Detailed Model Description: http://www.stolaf.edu/people/mckelvey/SOD.dir/BaseEMTech.pdf.
- Software Users' Guide: http://www.stolaf.edu/people/mckelvey/SOD.dir/UserGuide.pdf.