8.3

USING HIDDEN MARKOV MODELS OF VARIABLE RELEASE SCHEDULES IN THE ESTIMATION OF UNKNOWN SOURCE PARAMETERS

John M. Shuford, Kevin Denelsbeck & Stephanie L. Seely
ENSCO Inc., Melbourne, FL

1. SYNOPSIS

ENSCO, Inc.[1] has been developing systems for imputing source locations using transport and dispersion modeling in combination with sample records of pollutant concentrations.

Goodness-of-agreement measures between modeled pollutant concentrations at fixed receptor locations and sample records at those locations provide the bases for likelihood calculations conditioned on hypothesized source location and release schedule.

Abstracting release schedules as Markov processes is reasonably justified from a physical/engineering perspective and permits exhaustive exploration of the enormous space of possible (unknown) release schedules using dynamic programming algorithms within a framework of hidden Markov models (HMM). The so-called Forward and Backward algorithms facilitate likelihood calculations marginalized over all possible state-paths (release schedules) of the hidden Markov process. Meanwhile, the Viterbi algorithm can be used to estimate the most-probable (*a posteriori*) state-path.

We propose an estimation framework using hidden Markov models and demonstrate the approach using synthetic source/receptor data generated using transport and dispersion models ingesting historical wind fields over a several-month period in 2005.

2. BACKGROUND

Suppose we are given the meteorological record of observed transport winds in a given region and the time-sequenced sample record at a given receptor location, and suppose that a material of interest (from a presumed unique source) has been detected in multiple samples over that time period. How then might we calculate the (posterior) probability that the assignable origin of the subject material lies in a given cell of the spatial grid?

It seems logical to answer this in a Bayesian framework. To do this, we require some simplifying notation. We let F (for "footprints") denote the meteorological record of analyzed transport winds during the period of record, and let $O = \{O_t \mid t = 1,\ldots,T\}$ be the record of observed sample analysis results over some sampling campaign that consists of T samples. Let M_x denote the model/hypothesis that location (grid-cell) **x** is the source. We also let λ denote the parameters of a stochastic process model for the source release schedule and amounts.

Our general goal is to use the posterior probability of hypothesis M_x given the observed data sets *O* and F, as a statistic measuring the strength of association between the sample record and the meteorology. Applying Bayes' theorem we can formulate the desired posterior probability in terms of the likelihood of the sample record. If we assume that F and M_x and λ are independent we can write

$$P(M_x \mid O, F, \lambda) = \frac{P(O \mid F, M_x, \lambda) P(M_x)}{P(O \mid F, \lambda)}.$$

This means that the posterior probability of hypothesis M_x given the data sets *O* and F and source stochastic model parameters λ is equal to the product of the likelihood, that is, $P(O \mid F, M_x, \lambda)$, times the prior probability of M_x divided by the conditional probability of the observed sample record *O* given the footprint dataset F and model parameters λ.

Assuming one has a way to calculate the likelihood $P(O \mid F, M_x, \lambda)$ function for every grid cell ***x***, and assuming that some reasonable prior can be specified (albeit possibly an uninformative one), then our goal of finding the most probable location of the source is served by identifying the grid cells associated with the largest values of $P(M_x \mid O, F, \lambda)$.

3. MATERIALS AND METHODS

3.1. A Stochastic Model for Source Schedule

A basic supposition is that the source of an airborne pollutant may not be operating in a steady-continuous manner – it may be intermittently turned-off or turned-on, and while on, the release amounts in successive time periods can vary. Without knowledge of the schedule, it is reasonable to treat the release schedule as a stochastic process. Doing so allows us to parameterize the process such that the space of allowable release schedules can be explored in some tractable manner.

As a first attempt at a reasonable abstraction of the stochastic behavior of unknown release schedules, it

[1] Corresponding author address: 4849 N Wickham Rd., Melbourne, FL 32940. email: shuford.john@ensco.com.

is convenient for us to consider first-order Markov processes. A Markov process is a sequence of discrete events or "states" characterizing the progression of a system or phenomenon in which the probability that a certain state is occupied at a given time t in the sequence depends only upon which state was occupied at the prior position, t-1 (Rabiner 1989). In other words, suppose a system can be understood as being in one of N states, say S_1, S_2, …, S_N, at any given occasion and that this system makes discrete transitions from one such state to another [Note: We allow that the system might stay in the same state or change to a different state at the next occasion for observing the system.] The progression of states occupied by that system is properly modeled as a Markov process if the current state of the system is the only condition affecting the respective probabilities for possible states to be occupied by the system at the next observing occasion. If we let q_t denote the state occupied by a Markov process at time t in the state-progression sequence (or "path"), then we can say that $a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$ is the transition probability for the Markov process to progress from state S_i to state S_j (Rabiner 1989). The set of transition probabilities comprise a square matrix $A = \{a_{ij}\}$.

There are some situations, however, in which we might postulate the existence of a Markov process whose progression of states is hidden from view – that is, a system or phenomenon that can be understood as a Markov process but for which we cannot directly observe the sequence of states that the system occupies. Instead, suppose we can only observe directly a sequence of "outputs" caused by the system where the outputs are values drawn from some probability distribution at each step. We further suppose that the underlying (hidden) state q_t of the Markov process at time t in the state-progression path determines the probabilities for the possible outputs at time t of the path. Such a process is described as a Hidden Markov Model or HMM (Rabiner & Juang 1986; Rabiner 1989). This general notion can be imagined as a Markov process lying in the background behind a "veil" and a sequence of outputs is all that we can observe in the foreground.

The probability of observing output O_t at time t depends on q_t – that is, the state occupied by the process at time t. So in addition to the transition probabilities for the underlying Markov model, one must also specify $P(O_t \mid q_t = S_i)$ for all i = 1,..,N. Utilizing these notions heuristically for our source-location purposes, we think of the release schedules of the (unknown) source as Markov processes whose states represent different levels of effluent release. Time, meanwhile, is discretized into disjoint intervals corresponding to sample exposure times at the receptor. So, we will treat the hypothesized source as occupying a single state (i.e. operating at some fixed output level) during each of the sampling intervals and then possibly transitioning to some other state (with its corresponding fixed release amount) for the subsequent sampling interval.

To illustrate how this might be applied, one might define a set of possible states corresponding to release levels that are discretized in logarithmic progression. For example, a 6- state Markov model for source releases might include a state with release level 0 (source "turned-off"), and five other "turned-on" states with release rates of, say, 0.01, 0.1, 1, 10, and 100 g/hr, respectively. A strawman state-transition probabilities matrix for this set of states can be concocted based on a general notion of "inertia" in the process – that is, that when the source is turned-off, it tends to stay turned-off for several sampling periods; but once turned on, it will tend to stay "turned-on" for the duration of several sampling periods before returning to the zero-output state. For example, we might imagine that when the process is "turned-off" at time step t, it will stay "turned-off" at time t+1 with probability of, say, 80% but would have a 4% probability of transitioning to any given one of the "turned-on" states. Meanwhile, when the process is at one of the positive release-rate states, it will tend to stay in that state with 40% probability, transition to another given "turned-on" state with 10% probability, and transition back to the zero state with 20% probability. Such a matrix would look like the following:

$$A = \begin{bmatrix} .80 & .04 & .04 & .04 & .04 & .04 \\ .20 & .40 & .10 & .10 & .10 & .10 \\ .20 & .10 & .40 & .10 & .10 & .10 \\ .20 & .10 & .10 & .40 & .10 & .10 \\ .20 & .10 & .10 & .10 & .40 & .10 \\ .20 & .10 & .10 & .10 & .10 & .40 \end{bmatrix}$$

When available, engineering judgment as to the behavior of actual industrial processes (associated with the pollutants of interest) should certainly be substituted for naïve guesses as to the behavior of those processes. However, we may not have to get this quite right – for practical purposes, it would probably suffice merely that physically "more plausible" state paths (release schedules) are rewarded with generally higher probability scores than physically "less plausible" state paths.

Time steps for the process, as we have said, are successive sample periods. The data set serving as

the observable "output sequence" in this HMM framework is the sample record (quantitative analysis results) at the single receptor location. But this raises the issue of how to assign a conditional probability for the output observed at time t given that the process is occupying a given state at that time. In other words, we need to specify $P(O_t \mid q_t = S_i)$ for all i = 1,..,N and, for that matter, for each time step t. To do this, we will need the results of transport and dispersion (T&D) modeling to measure how hypothesized releases would relate to observed collections of target material in the sample.

3.2. Likelihood of Sample Record

First, we use a transport and dispersion model to estimate predicted catch U_t for each sample period t under the hypothesis assuming steady-continuous unit releases from a hypothetical source at location x. Then, we calculate $C_{t,i}$ – the predicted catch at time t that is conditioned on the assumption that the source process is in state S_i. This is obtained by multiplying the U_t value by the release rate associated with state S_i.

Now, we find the difference between (log-transformed) predicted catch $C_{t,i}$ and the analyte measurement O_t. Standardize this difference by dividing by an estimate of the standard deviation of that difference:

$$Z_{t,i} = \frac{\log[C_{t,i}+1] - \log[O_t+1]}{\sqrt{\sigma^2_{\log[C_{t,i}+1]} + \sigma^2_{\log[O+1]}}}$$

Use the normal probability model to assign a "tail area probability" to the event of observing a standard-normal variate at least as great as $| Z_{t,i} |$; that is, we will let

$$P(O_t \mid q_t = S_i) = 2(1 - \Phi(| Z_{t,i} |)),$$

where $\Phi(x)$ is the probability distribution function for a standard normal random variable.

In using a HMM framework, the task at hand is to evaluate the conditional probability of the output sequence for every possible state path. Doing so allows one to 1) obtain the likelihood of the observed sequence marginalized (summed) over all state paths; and 2) infer the a posteriori most probable state path given the observed sequence.

Brute-force exhaustion over all N^T possible state paths is prohibitive for values for N and T that are typically encountered in application (indeed, in this report we illustrate use of a model using N=6 and T=428). Such a calculation is feasible only by use of efficient dynamic programming algorithms known as the Forward algorithm (and the closely related Backward algorithm) as well as the Viterbi algorithm that exploit the assumed Markovian structure of the underlying model (Rabiner & Juang 1986; Rabiner 1989).

The Forward algorithm is an efficient approach to calculating the likelihood of the observed sequence (i.e. sum of probabilities over all possible state paths) given the model parameters (Rabiner 1989).

The Viterbi algorithm is a dynamic programming algorithm that finds the state path for which the probability of the observed sequence is maximized. Like the Forward and Backward algorithms, it proceeds inductively (including a trace-back step) and exploits the Markovian structure of the model (Viterbi 1967; see also Forney 1973, Hayes 2002).

Our purpose in invoking the mathematical machinery of hidden Markov models is, of course, to serve the purpose of calculating the likelihood function $P(O \mid F, M_x, \lambda)$ of our Bayesian HMM for every grid cell x of a regular spatial grid surrounding our sampler location.

If we consider λ to represent the parameters of our HMM framework – that is, the number N of different states, the particular release rate associated with each of the N states, the state transition probability matrix A, and the vector π of initial state probabilities – then the likelihood function $P(O \mid F, M_x, \lambda)$ can be obtained via the Forward algorithm (described in the previous section) by summing $\alpha(T,k)$ over all k. We emphasize that this likelihood is marginalized over all possible state paths.

A maximum *a posteriori* (MAP) estimate of the location of the true source is sought by evaluating the likelihood function at every grid cell x and multiplying this likelihood by the prior probability of that grid cell. If one has prior belief as to the probable location of the source, it is certainly useful to incorporate that information in the prior probability. Lacking such prior belief, as is usually the case, we can assume an uninformative prior – that is, a uniform probability distribution over the grid. In that case, the MAP estimate corresponds to the grid cell with maximum likelihood.

Since our concern is for the relative likelihood of different candidate grid cells, it is convenient to ratio the posterior probability obtained for each grid cell to

the posterior probability of the "winning" cell and to express the "score" associated with each grid cell as the logarithm of an odds ratio.

$$L_x = \log\left[\frac{P(M_x \mid O, F, \lambda)}{P(M_{\max} \mid O, F, \lambda)}\right]$$

3.3. Implementation

We have implemented this Bayesian/HMM technique in Java. One of the input datasets contains the sample record – that is, a listing of all the samples to be included in the analysis along with the measured quantity (or concentration) of subject effluent in the sample. Another input dataset contains "footprints" – that is, the predicted catch (or concentration) in each sample from each hypothetical grid-point source, assuming a steady-continuous release schedule. The number of records in the second dataset is therefore the product of the number of samples and the number of cells comprising the spatial grid. A third input file is a model-specification file. This file contains specifications of the number of states N of the underlying Markov model, the release rate associated with each of these states, the initial probabilities for each state, and the N × N state transition matrix.

The outputs of the application include a level-plot showing the logarithm of the posterior odds for each grid cell relative to the highest-scoring grid cell (Figure 1). The level-plot display is interactive: the user can mouse-click on a cell of this level-plot to bring-up a window that shows the highest-scoring state path (as determined by the Viterbi algorithm) for the selected grid cell. All output data are exportable to ASCII flat files for analysis or graphing by other applications.

Figure 1. Screen-shot of the interactive levelplot of log posterior odds scores for cells in a spatial grid.

4. Trial Application

To test whether the basic ideas that are embodied in this approach "make sense" or at least show promise, we have performed some trial applications of the Bayesian/HMM model to simulated source-receptor scenarios.

The simulation test-bed consisted of a dataset of observed surface weather data from Dugway Proving Ground (DPG) Utah for the period June through December 2005. That collection of data includes:

- Data Set A – a set of high-resolution observations from a DPG mesonet known as SAMS (Surface Atmospheric Measurement System). In this dataset, the SAMS stations recorded fifteen-minute averages of air temperature, wind speed, wind direction, and pressure.

- Data Set B – a dataset having lower resolution in space and time, comprised of surface, upper-air, and global gridded data, supplemented by hourly observations from a single SAMS location in order to mimic the presence of a standard surface reporting station.

The transport and dispersion simulations suppose a single receptor at a fixed location. We define a grid of hypothetical source locations (16x19), having a spatial resolution of 0.05-degree, centered on the hypothetical receptor.

Numerous different source-location/release schedule scenarios have been created to exercise the Bayesian/HMM source-locator technique and software application. In each of these scenarios, a simulated sample record was created supposing a hypothetical source location operating according to: a predetermined (intermittent) release schedule, modeled transport based on the high-resolution weather data (Data Set A), and a hypothetical collector at the center of the grid taking consecutive 12-hour samples. Over the seven-month period described by this data set, that amounted to a sample record of 428 simulated samples. The simulated sample results were left-censored according to a pre-set "detection limit" – that is, if the simulated collection fell below the detection limit, the "measured" target amount in the sample was set to zero. In some of the scenarios, random errors of 5% of the simulated sample catch were added to the sample record.

8.3

In addition to the sample record, the estimation technique requires as an input the predicted "catch" of effluent at the subject sampler due to hypothetical plumes originating from each point of a grid of putative sources assuming a steady-continuous release schedule. In our case, the predicted airborne concentrations are calculated by the transport and dispersion model SLAM, which is a Lagrangian Gaussian-puff and trajectory model that can ingest a wide variety of meteorological data and formats (ENSCO, Inc. 2007; Shuford, et al., 2006). In our simulations, the lower-resolution meteorological data set B was used to drive the T&D model "footprints".

For purposes of illustration, let us examine one simulation scenario in which a hypothetical source is located approximately 33 km SSE of the hypothetical sampler and in which the source emits a material of interest intermittently according to the schedule pictured in Figure 2. In this scenario, we do indeed add "white" noise and impose a left-censoring to the simulated sample record.

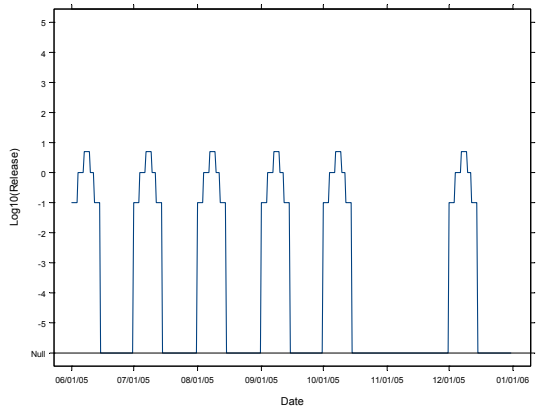

Figure 2. Simulated release schedule for illustration scenario.

Applying the Bayesian/HMM source-location technique with the 6-state Markov model described earlier, we obtain two outputs from the program. The first output, of primary interest, is the level-plot of the log posterior odds scores for each cell of the spatial grid (Figure 3). In this figure, the location of the hypothetical sampler is denoted by the solid white dot, and the correct location of the hypothetical source is shown by the open white circle.

The ranking of the ten best scoring grid cells are annotated in the corresponding cells. As we can see, the correct source location fell within the 4th-highest scoring cell and within the clearly highlighted region of the levelplot.

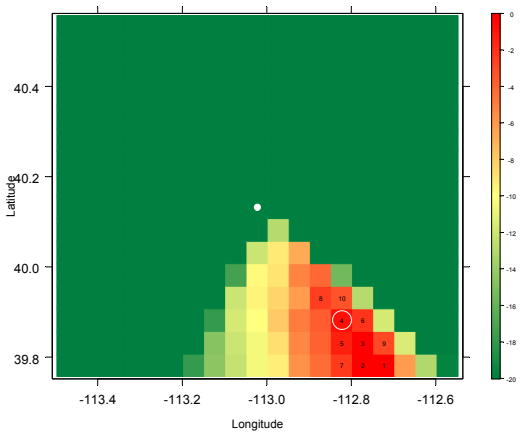

Figure 3. Levelplot of log posterior odds scores from the Bayesian/HMM program using a 6-state source stochastic model. The location of the hypothetical sampler is denoted by the solid white dot, and the correct location of the hypothetical source is shown by the open white circle

Another output of the program is the *a posteriori* estimated most probable state sequence (i.e. release schedule) obtained via the Viterbi algorithm. This is shown below in Figure 3.

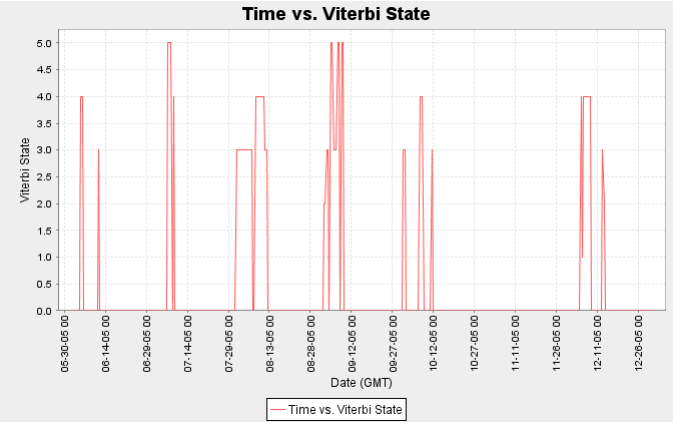

Figure 4. Estimated source release schedule obtained via Viterbi algorithm for the illustration scenario. Release schedule estimate shown here corresponds to the highest-scoring grid cell.

8.3

In this simulation scenario, the technique provided reasonably accurate estimates of the location of the source and the "unknown" intermittent release schedule at that source.

5. DISCUSSION

We have described here an approach to estimating the location of an assignable (and presumably unique) source of a monitored effluent given a collection record at a receptor location that includes multiple collections of the monitored pollutant over a sampling campaign. That approach uses a Bayesian framework within which hidden Markov models of release schedule serve to calculate the likelihood of the sample record.

The advantages of the HMM approach are that we can exhaustively search for a solution over a huge space of possible release schedules (versus sampling that space via Monte Carlo techniques) and that the computational cost of exploring that space is comparatively very low due to the availability of efficient dynamic programming algorithms.

It is admitted that the use of Markov process models for release schedules is somewhat heuristic: certainly the use of a discretized set of N "states" to represent the possible release rates from a source is at best a stair-step approximation to a more continuous time series with higher-frequency components, and there is nothing physically dictating that changes in release rate at a source should correlate with the start and stop times for sample exposures at our receptor. However, since the temporal resolution of information concerning releases is limited by those exposure durations, we really should not hope to succeed in exploring release schedules of finer resolution. So, the discretization of time may not be unreasonable.

To the extent that true release processes possess a Markovian character, we would expect that engineering judgment might help refine the specification of the state-transition matrix. The matrix used in our trial applications was a somewhat arbitrary and capricious attempt to capture an expected day-to-day "inertia" of the source operation. We note, however, that techniques for iterative re-estimation of the HMM parameters have been described extensively in the literature (e.g. Rabiner & Juang 1986; Rabiner 1989; Ephraim and Merhav 2002). Such capability was not implemented in this effort due to resource limitations.

6. REFERENCES

Ephraim, Y. and Merhav, N. (2002). Hidden Markov processes. IEEE Transactions on Information Theory, 48(6), 1518- 1569.

Forney, G. D. Jr. (1973). The Viterbi algorithm. Proceedings of the IEEE, 61(3), 268-278.

Hayes, J. F. (2002, May). The Viterbi algorithm applied to digital data transmission. IEEE Communications Magazine [50th Anniversary Commemorative Issue]. (Reprinted from IEEE Communications Magazine, 13(2) [March 1975].)

ENSCO Inc. (2007). Technical Documentation for the Short-range Layered Atmospheric Model (SLAM). ENSCO, Inc. (Melbourne, Florida) Technical Report.

Rabiner, L. R. and Juang, B. H. (1986, January). An introduction to hidden Markov models. IEEE ASSP Magazine, 3(1), 4-16, January 1986.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE, vol. 77, pp. 257-286.

Shuford, J., Eoll, J., Seely, S., Moore, W., and Dreher, J. (2006). Comparison of two transport descriptors within a hybrid receptor modeling system. In Proceedings of the 8th Conference on Atmospheric Chemistry and 14th Joint Conference on the Application of Air Pollution with the AWMA. [Jan 30 – Feb 2, 2006 Annual Meeting of the American Meteorological Society, Atlanta, Georgia].

Viterbi, A. J. (1967, April). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory, IT-13(2), 260-269.