Risk Analysis, Vol. 33, No. 12, 2013

DOI: 10.1111/risa.12058

Time-Series Models for Border Inspection Data

Geoffrey Decrouez[1,*] and Andrew Robinson[1,2]

We propose a new modeling approach for inspection data that provides a more useful interpretation of the patterns of detections of invasive pests, using cargo inspection as a motivating example. Methods that are currently in use generally classify shipments according to their likelihood of carrying biosecurity risk material, given available historical and contextual data. Ideally, decisions regarding which cargo containers to inspect should be made in real time, and the models used should be able to focus efforts when the risk is higher. In this study, we propose a dynamic approach that treats the data as a time series in order to detect periods of high risk. A regulatory organization will respond differently to evidence of systematic problems than evidence of random problems, so testing for serial correlation is of major interest. We compare three models that account for various degrees of serial dependence within the data. First is the independence model where the prediction of the arrival of a risky shipment is made solely on the basis of contextual information. We also consider a Markov chain that allows dependence between successive observations, and a hidden Markov model that allows further dependence on past data. The predictive performance of the models is then evaluated using ROC and leakage curves. We illustrate this methodology on two sets of real inspection data.

KEY WORDS: Border inspection; hidden Markov model; independent model; leakage curves; Markov chain; ROC curve; serial dependence

1. INTRODUCTION

A primary challenge in the border inspection of imported goods is to determine, ideally in real time, how much inspection effort to apply to each incoming item. Contextual data are usually available about imported items before their arrival, for example, the contents, the country of origin, and the identities of the supplier or the importer. When combined with inspection history, the contextual data provide a potentially rich source of information that can be used to prioritize items for inspection effort, a process referred to as *profiling*. The priority for inspection can be represented in many ways, and here we choose to try to estimate the probability that each item is contaminated.

We begin with some definitions. A shipment arriving at the border contains a certain number of *consignments*. Each consignment can contain a different number of individual products or items of a different nature, also called *lines*. A fraction of these items are inspected. We define a *pathway* as the aggregation of like items. A pathway is analogous to a population or a process in statistical modeling, and its specification is usually guided by operational concerns as much as statistical rigor. For example, all air passengers entering a country could be considered a pathway, as could all returning nationals, or all returning nationals arriving at a certain airport, or even on a given flight number.

A container or a consignment is said to be "contaminated" if it contains biosecurity risk material, such as an invasive pest or disease. We will call

[1] Department of Mathematics and Statistics, The University of Melbourne, Parkville 3010, Australia.

[2] Australian Centre of Excellence for Risk Analysis (ACERA), Australia.

[*] Address correspondence to Geoffrey Decrouez, Department of Mathematics and Statistics, The University of Melbourne, Parkville 3010, Australia; dgg@unimelb.edu.au.

consignments that do not have pests *clean*. The list of pests and diseases identified as a potential threat depends on the pathway considered. For example, for the wheat pathway, quarantine services try to intercept consignments that could contain the smut fungus *Tilletia indica*, which causes karnal bunt, a disease of wheat. Karnal bunt on wheat has never been identified in Australia, and an incursion could have huge environmental and economical consequences.[1] The status of an incoming product may depend on various factors, as mentioned above; we refer to these factors as *covariates*. Based on historical data, one may estimate which covariate is more likely to differentiate between clean and contaminated consignments.

In this study, we consider the status of incoming products to be a binary time series, that is, the item is either clean or contaminated. Our rationale is twofold. First, it is of practical interest to have an estimate of the probability of compliance of incoming consignments. This probability should capture any instantaneous change in the status of incoming consignments: periods of low/high risks of presence of invasive species should be reflected in the evolution of this probability. Second, one wants to predict the status of a consignment based on past data to develop a strategy for surveillance of incoming products and efficiently prevent invasive species from entering the country. The novelty of this study is to include a dynamic component to the surveillance procedure by viewing the status of consignments as a time series. Serial dependence, when present, can then be used as a way to improve the prediction of arrival of risky containers. This study incorporates existing static methods for classifying shipments, and then uses this classification to predict the arrival of risky shipments in real time.

Table I. Status of Items Arriving at the Border Sorted by Time

Arrival Date	Quarantine Entry	Country	Supplier code	Importer Code	Fail
4/01/09	A	C	CCC7538	08357	FALSE
4/01/09	B	N	CCC7503	34750	FALSE
5/01/09	C	A	CCC2432	03409	TRUE
6/01/09	D	B	CCC6884	28234	TRUE
6/01/09	D	B	CCC6884	28234	TRUE
6/01/09	D	B	CCC6884	28234	TRUE
6/01/09	E	C	CCC9084	34890	FALSE
7/01/09	F	F	CCC1263	84546	FALSE

Note: Items arrive with information including the country of origin, the supplier and importer names/codes. The column labeled quarantine entry reports the identification number of a container. A container can contain several items or lines of different type. For example, quarantine entry A contains stock feedproducts of three different kinds. The status of an item is binary and equals 1 (=TRUE) if it fails the quarantine test. The three time-series models we consider in this study, corresponding to this data, are provided in Fig. 1.

To simplify our development, we assume perfect detection, that is, if a contaminated consignment is inspected, then the contamination is detected. We return to this point in the discussion in Section 4. We consider three discrete time-series models that account for different levels of serial dependence. First, the independence model assumes no correlation structure at all. This approach is equivalent to a static approach, since all decisions are based on information about covariates. Second, we model the observation sequence as a two-state first-order Markov chain, which allows a first-order correlation on past data. It can be used together with the information about the covariates to predict the status of the next shipment. Finally, we consider a binomial hidden Markov model (HMM). In an HMM, the status of an item is directly observed by quarantine services, and depends on a hidden state, which corresponds to the status of the pathway, for instance, to the status of the importer. An importer can be referred to as noncompliant if its consignments demonstrate a pattern of contamination. Table I presents artificial inspection data for the purpose of demonstrating the models, and the associated three models are depicted in Fig. 1. The HMM enables us to address the two objectives given before. First, the time series is not Markovian, which means that the status of an incoming consignment does not only depend on the current status, but also on the past observations. Second, once an HMM is fit to the data, prediction based on likelihood can be performed in a fast and efficient way using well-known algorithms.

We advocate fitting the three models to the entire data set, and then to subsets of the data corresponding to a specific importer or supplier. By doing this, we first test for the presence of correlation and patterns of contamination for the whole pathway. If the value of the covariates is more informative than the correlation structure for predicting the status of incoming shipments, which occurs when no systematic problems are detected, then the independence model will perform as well as other candidate models. We also investigate the presence of serial dependence for a specific importer or supplier, which occurs, for example, when a supplier suffers from seasonal contamination of his crops, or when an importer does not closely monitor the quality

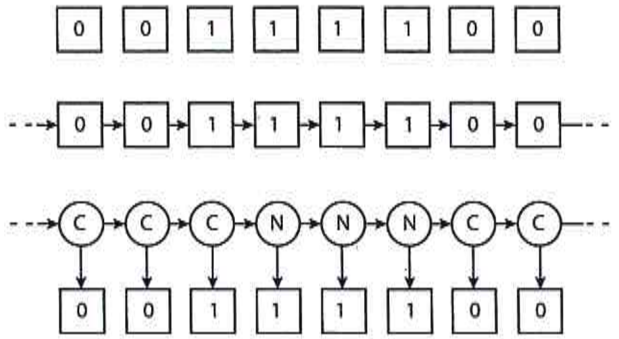

Fig. 1. The binary observation sequence corresponds to the data given in Table I. The value 1 stands for a noncompliant consignment, and a 0 for when no contamination is detected. The serial dependence is depicted using arrows. The top plot presents the independence model; see Section 2.3.1. Successive observations are assumed to be independent from each other, given the values of the covariates. In the middle plot, the binary status follows a two-state Markov chain: the status of the next arriving container depends only on the current status; see Section 2.3.2. The bottom plot shows a HMM. Circles are hidden states and correspond, for example, to the status of an importer or supplier (C for clean and N for noncompliant). The true status is unknown in real life, and is given here only for illustration. Squares are observations, corresponding to the status of an item in the shipment, from the data given in Table I. Observations are assumed to follow a Bernoulli distribution, where the probability of an item being contaminated depends on the hidden state C or N. Hidden variables are assumed to follow a Markov chain. See Section 2.3.3 for further details.

of the goods. We demonstrate this approach in our examples.

The article is organized as follows. Section 2 details the proposed methodology. We review and discuss different existing approaches used to classify covariates according to their level of risk in Section 2.2. We then describe in Section 2.3 the three time-series models considered: the independent model, the Markov chain, and the HMM. Finally, Section 2.4 addresses model selection and explains how the predictive performance of the selected model can be assessed using ROC and leakage curves. Section 3 illustrates the methodology on two real data sets taken from the mangosteen and the stockfeed pathways. Section 4 concludes with some additional remarks and a discussion.

2. METHODOLOGY

2.1. Quarantine Data

Quarantine data are usually accompanied by contextual data, such as the identity of the supplier or the importer, the description of the goods, or the country of origin. A first approach is to fit a time series to the full data set, possibly split over different time periods, and to use the contextual information as predictors or covariates. The aim here is to detect periods of high risk, where most of the goods belonging to a same pathway would suffer from contamination, for example, due to pests that respond to seasonal influences.

We also look for patterns of contamination at the importer and supplier level. In this case, the time-series models capture periods of high or low risk specific to a given supplier/importer. Quarantine inspection services are particularly interested in information about importers since they have a direct control over them and can take appropriate decisions during periods of high risk. In Section 3, we consider both approaches.

2.2. Classification of Covariates

The large amount of contextual data (covariates) requires methods to decide which information can be used to predict the high/low risk associated with subsequent shipments. A brief review of quarantine intervention for cargo at the border in Australia follows. Broadly speaking, pathway risk is managed by inspection at the border if the pathway risk is considered high enough. Otherwise, the risk is managed at the border by sighting documentation, for example, demonstrating relevant off-shore treatments. Two approaches are typically taken to the choice of intervention strategies: policy and modeling. In the former, the inspectorate forms a position as to the biosecurity risk posed by an import based on an import risk analysis, which is a summary of the state-of-the-art scientific knowledge of the biology of the likely pests and the projected impact upon social, economic, and ecological goods. The second approach is based on a statistical analysis of patterns of risk, using various analytical tools. There is an increasing acceptance that the outcomes of border inspections can and should be used to inform the inspectorate about the risk presented by the pathway.

In the United States, the Government Accountability Office (GAO) has published reports on cargo and agricultural inspections; see, for example, GAO-04-557T and GAO-08-96T. According to the latter report, the U.S. Department of Agriculture estimates the cost associated with the introduction of pests and diseases into the mainland to tens of billions of dollars annually. Moreover, since September 11, 2011, the United States is concerned with the attempt to smuggle weapons into the mainland, one possible method being by cargo containers. Since 2003, shipment and passenger inspections are carried out by the U.S. Customs and Border Protection (CBP). The large volume of imports prevents CBP inspectors from fully inspecting all containers and passengers. Instead, CBP uses a targeting strategy called Automated Targeting System (ATS), which assigns a risk level to shipments based on the shipment information, to help identify high-risk shipments and passengers for inspection, and prioritize their use of resources. The mathematical models used by CBP are not publicly available.

In this study, we focus on the contextual data giving the identity of the importer, the supplier, and the country of origin. The classification is performed on training data, where a time-series model is also fit; see next section. For each covariate, we compute the proportion of items found contaminated during this time period, also referred to as the consignment-level failure rate. If the failure rate is above a threshold, say R, then the corresponding level of the covariate is labeled "risky." The value of R may vary for each covariate. When more information is taken into account, one may use a ridge regression to estimate the risk level of a shipment. Explanatory variables correspond to the covariates, and the response variable to the risk level of a shipment. Then one identifies covariates that have the most influence from the size of the standardized, estimated coefficients; see, for example, Ref. 2 where the authors applied this method to the detection of nuclear materials in shipments.

Contextual data are available for each shipment arriving at the border. Costs associated with data collection for the proposed methodology are not higher than the costs associated with classification methods already implemented.

2.3. Mathematical Models

In this section, we describe the models. Particular focus is given to how the covariates are included. Expression of the likelihood and technical details about parameter estimation are postponed to the Appendix. We denote observations by $y_1, \dots, y_n$, where n is the total number of observations, and $y_i \in \{0, 1\}$, for $i = 1, \dots, n$, corresponding to the status of a consignment. Throughout this article, capital letters are used to denote random variables, and the corresponding small letters to observations. Let $\mathbf{Y}_{1:i} = (Y_1, \dots, Y_i)$, for $i = 1, \dots, n$. Covariates at time i are denoted by $\mathbf{z}_i = (1, z_{i,1}, \dots, z_{i,K})^t$, $i = 1, \dots, n$, where t is the transpose operator and $K \geq 0$ represents the total number of covariates. Let $\mathbf{z}_{1:i} = (\mathbf{z}_1, \dots, \mathbf{z}_i)$. The full observation sequence is split in two parts. The first part is referred to as the training period, where the models are fit and where model selection is addressed. The second part is used for prediction (see next section) and is referred to as the test period.

2.3.1. Independent Model

Random variables $Y_1, \dots, Y_n$ are assumed to be mutually independent, with $\mathbf{P}(Y_i = 1 \mid \mathbf{Y}_{1:i-1}, \mathbf{z}_{1:i}) = \mathbf{P}(Y_i = 1 \mid \mathbf{z}_i) = p_i$, where $\text{logit}(p_i) = \alpha^t \mathbf{z}_i$, $i = 1, \dots, n$, $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_K)^t$ are the covariate coefficients.

2.3.2. Markov Chain

The observation sequence $Y_1, \dots, Y_n$ is assumed to follow the dynamics of a Markov chain. It satisfies $\mathbf{P}(Y_i \mid \mathbf{Y}_{1:i-1}, \mathbf{z}_{1:i-1}) = \mathbf{P}(Y_i \mid Y_{i-1}, z_{i-1})$. Future observations are independent of the history of the process given the present value. The dynamics of the chain at time i are completely characterized by its transition matrix

$$\mathbf{A}(i) = \begin{pmatrix} a_{00}(i) & 1 - a_{00}(i) \\ a_{10}(i) & 1 - a_{10}(i) \end{pmatrix}$$

at time i, where $a_{00}(i) = \mathbf{P}(Y_{i+1} = 0 \mid Y_i = 0, \mathbf{z}_i)$ and $a_{10}(i) = \mathbf{P}(Y_{i+1} = 0 \mid Y_i = 1, \mathbf{z}_i)$, with the initial distribution $\mathbf{P}(Y_1 = y_1 \mid \mathbf{z}_1)$. The transition probabilities $a_{00}(i)$ and $a_{10}(i)$ depend on the covariates via the logistic link function, $\text{logit}(a_{00}(i)) = \beta^t \mathbf{z}_i$ and $\text{logit}(a_{00}(i)) = \gamma^t \mathbf{z}_i$, where $\beta = (\beta_0, \beta_1, \dots, \beta_K)^t$ and $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_K)^t$, for $i = 1, \dots, n$.

2.3.3. HMMs

HMMs are a very popular class of models due to their simplicity, and have been applied to many applied problems, for example, in speech recognition,[3] in biomedical modeling,[4,5] and in molecular biology,[6] but not yet to our knowledge in the area of border inspections. Covariates can be added to a HMM; see, for instance, applications in cognitive science[7] and in experimental psychology.[8] For other applications of HMM, with and without covariates, we refer the reader to Ref. 9.

In HMM, the observation distribution depends on the state of an unobserved underlying two-state Markov chain $\{X_i\}$. We use the same notation as in Section 2.3.3 for the dynamics of $\{X_i\}$. Given the chain is in state j, where $j = 0$ and 1 corresponds to the states C and N in Fig. 1, the probability of noncompliance is $p_j(i) = \mathbf{P}(Y_i = 1 \mid X_i = j, \mathbf{z}_i)$, so that the conditional distribution of $Y_i|X_i$ is Bernoulli. We refer to $p_j(i)$ as the state-dependent probabilities, and we model their dependence on covariates with the logistic link function $\text{logit}(p_0(i)) = \delta'\mathbf{z}_i$, and $\text{logit}(p_1(i)) = \zeta'\mathbf{z}_i$, for $i = 1, \dots, n$, with $\delta = (\delta_0, \delta_1, \dots, \delta_K)'$ and $\zeta = (\zeta_0, \zeta_1, \dots, \zeta_K)'$. By construction, the random variables $\{Y_i\}$ are mutually independent conditionally on $\{X_i\}$.

2.4. Model Selection and Prediction

Model selection is performed using the Akaike information criterion (AIC), defined by $\text{AIC} = -2\log L + 2k$, where L is the likelihood of the fitted model, k the number of parameters to estimate, n the sample size, and the Bayesian information criterion (BIC), defined by $\text{BIC} = -2\log L + k\log n$. The selected model is the one with the smallest AIC or BIC. There is no clear reason to prefer one criterion over the other. On the one hand, if the family of models considered contains the true model, then the BIC is asymptotically consistent, which means it selects the true model as the sample sizes go to infinity. In our setting, there is not a true model that contains a clear subset of all possible covariates. On the other hand, the AIC does not assume the existence of a true model. However, as the sample size increases, the AIC tends to choose models that contain too many parameters, which is not the case for the BIC because of the heavy penalty imposed on models with many parameters. As such, we report both the AIC and BIC.

The AIC and BIC are used to detect the presence or absence of serial dependence in the time series, and to indicate when a Markov chain or HMM provides a better fit than the independence model. When this is the case, the serial dependence suggests the presence of contamination patterns, and these models are used for prediction. When there is no evidence of patterns of contamination, that is, when the AIC/BIC selects the independence model, prediction should be made from existing methods of classification presented in Section 2.2.

When a Markov model or HMM is selected from the training data, we use it to predict the probability of noncompliance of subsequent shipments over a new period of time called the test period. The covariates classified as "risky" during the training period are still classified as "risky" in the test period. The one-step-ahead probability $\mathbf{P}(Y_{i+1} = 0 \mid \mathbf{y}_{1:i}, \mathbf{z}_{1:i})$ is computed using the parameters estimated during the training period. For the HMM, it can be expressed as a ratio of likelihoods, and can be efficiently computed using the so-called forward variable.[14] The entire probability distribution for the forecast can be computed. It provides a way to obtain not only point forecasts, but also interval forecasts. Since our process is binary, confidence intervals for the forecasts are not informative. Higher-order joint forecast distributions can be computed similarly, see Chapter 5 in Ref. 9.

Table II. Definition of a, b, c, and d

	Item Contaminated	Item Clean
Inspected	a = true positive	b = false positive
Not inspected	c = false negative	d = true negative

The decision rule that derives from these models is as follows. The predicted probability $\mathbf{P}(Y_{i+1} = 0 \mid \mathbf{y}_{1:i}, \mathbf{z}_{1:i})$ is compared with a threshold $p_T \in [0, 1]$. If $\mathbf{P}(Y_{i+1} = 0 \mid \mathbf{y}_{1:i}, \mathbf{z}_{1:i}) < p_T$, then inspect the consignment, otherwise do not inspect. The limit cases correspond to $p_T = 0$ where we inspect nothing, and $p_T = 1$ where we inspect everything. For a given p_T, we record the number of items inspected and not inspected that are actually clean and contaminated for the test period. We denote these numbers by $a(p_T)$, $b(p_T)$, $c(p_T)$, and $d(p_T)$; see Table II. For convenience, we drop the explicit dependence on p_T. ROC curves and leakage curves provide a good way of assessing the predictive power of a model. The ROC curve plots the true positives (given by the ratio $a/(a + c)$, also known as sensitivity) against the false positives (given by $b/(b + d)$, which is 1 minus the so-called specificity), and provides an idea of how many false positives to expect for a given effectiveness. The leakage curve plots the probability of an item being not inspected given that it is contaminated, $c/(a + c)$, against the total number of inspected items $a + b$, thus giving an idea of how much effort must be undertaken to reduce leakage below a given level.

2.5. Risk Analysis

The project reported in this article reflected the context of developing statistical tools to support risk-based approaches to the management of the biosecurity risk of imports. The reported innovation focuses on at-border intervention, although mitigatory

Table III. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for the Time-Series Models Fitted to the Mangosteen Pathway; the Best AIC/BIC Are Highlighted

	Independence		Markov		HMM	
Covariates	AIC	BIC	AIC	BIC	AIC	BIC
()	445.2	**449.1**	443.8	451.8	**440.3**	460.3
(I)	442.7	**450.7**	443.6	459.5	**441.1**	477.0
(S)	**420.4**	**428.4**	450.6	466.5	424.1	460.1
(I+S)	**420.3**	**432.3**	453.2	477.2	426.2	478.0

interventions are also in place offshore (preborder) and onshore (postborder). It is accepted that the pathways for which such tools might be deployed should be considered low risk, that is, there should be a broad acceptance that there will be some leakage of contaminated consignments, and the purpose of the intervention is to reduce the rate rather than eliminate the possibility. This policy approach is in line with the prescriptions of Ref. 10, which stated "zero risk is unattainable and undesirable"; see also Ref. 11.

3. ILLUSTRATION

3.1. Mangosteen Pathway

The mangosteen pathway comprises 644 observations from December 2005 to December 2010. During this period of time, 23.3% containers were found contaminated. All mangosteens are imported from a single country, and we consider only the importer and supplier codes as covariates. We fit the data using the models described in the previous section on the first 400 observations, then test the prediction on the remaining 244. The data set is too small to consider a subset corresponding to a single supplier or a single importer. The threshold R is taken to be 0.15 for suppliers (which yields the classification of 67.7% of them as "risky") and 0.2 for importers (70.2% as risky). The AIC and BIC are presented in Table III. Covariates are indicated in brackets, (I) for importer and (S) for supplier, and (I+S) for a combination of both. Empty brackets denote no covariates.

The independence model and the HMM fit the data better than the Markov chain. The AIC are similar for the independence model and the HMM, and the BIC favors the independence model since it contains less parameters. In fact, the good fit of the HMM is due to the instantaneous dependence of the observation to the present covariate, corresponding to the term $\mathbf{P}(Y_i = y_i \mid X_i = x_i, \mathbf{z}_i)$ in the expression of the likelihood (see the Appendix) and not to the dependence structure captured by the Markov chain. Including supplier identity as a covariate improves the fit for this time series.

This pathway provides an example where the procedure described in Section 2 indicates a lack of systematic contamination patterns. Prediction of the risk level of subsequent shipments should be based on their classification.

3.2. Stockfeed Pathway

We now focus on the inspection of stockfeed (e.g., food for cattle, chickens, wheat gluten, aquarium fish food) arriving at the Australian border. In the six years spanning October 2005–October 2011, an annual average of nearly 6,500 consignments related in some way to stockfeed were imported into Australia, amounting to nearly 18 per day. For stockfeed products, all consignments or all lines are inspected, which can be very expensive.

The training period is taken to be from 2005 to the end of 2010, and the test period is the year 2011. As for the mangosteen pathway, directly analyzing the full time series does not show the benefits of using a model that allows correlation, compared to the independence model. However, the data set is large enough to analyze the time series corresponding to a specific importer or supplier. For quarantine services, it is often more valuable to analyze the level of risk of an importer, since importers pay for inspections. We focus on one importer, whose time series shows the presence of noncompliant items, and for whom we have enough data for the six-year period. The data set chosen comprises 662 observations, 582 for the training period, and 80 for the test period. For this importer, we do not consider any covariates since it is mainly importing from two suppliers (representing 43% and 42% of the total imports), and from two countries (86% and 12%). Considering the models with no covariate has the advantage to target periods of high-level risk for this importer due to the serial dependence of the sequence.

The AIC and BIC are given in Table IV. The Markov chain and the HMM represent a better fit for this importer, compared to the independence model. The corresponding estimates of the transition matrix and the state dependent probabilities for the HMM, together with 95% confidence intervals, are:

$$\hat{\mathbf{A}} = \begin{pmatrix} 0.995\ (0.967, 0.999) & 0.005\ (0.001, 0.033) \\ 0.124\ (0.031, 0.799) & 0.876\ (0.201, 0.969) \end{pmatrix}, \tag{1}$$

Table IV. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for the Time-Series Models Fitted to the Stockfeed Pathway, for a Specific Importer; the Best AIC/BIC Are Highlighted

	Independence	Markov	HMM
AIC	162.6	136.4	**135.6**
BIC	166.9	**145.1**	157.4

and $\hat{p}_0 = 0.01\,(10^{-6}, 0.019)$ and $\hat{p}_1 = 0.53\,(0.017, 0.999)$. We used the percentile bootstrap with 999 resamples to compute confidence intervals.[9, 12] A discussion about the confidence intervals is in order here. The first state of the Markov chain is clearly persistent; once the chain enters this state, there is a large probability that it remains in it, which is not as pronounced for the second state. The confidence interval for $\hat{p}_0$ is very narrow around the point estimate, so that the probability of noncompliance is very low when the chain is in the first state. However, the high persistence of the first state has a substantial effect on the confidence interval for $\hat{p}_1$, which is not informative. In fact, values of $\hat{p}_1$ close to 0 occur when the chain of the bootstrap resamples does not leave the first state at all. Alternatively, computing the Hessian to obtain confidence bounds is not reliable when some parameters are close to the boundaries of the parameter space see (Chapter 3 in Ref. 9), which is the case here for two parameters. However, we can still point out that the state 0 corresponds to a smaller probability of contamination than when the chain is in state 1. Expression of $\hat{\mathbf{A}}$ suggests the following interpretation for the two hidden states: one corresponding to a "clean" status with associated zero state dependent probability, and one "contaminated" status with a positive value for the probability of failing the quarantine test. The high value of the probability of remaining in the clean state is very high, which accounts for the small proportions of items failing the quarantine test. For ergodic Markov chains, the stationary distribution gives the proportion of time the chain spends in one state. It corresponds to 96.1% of the time for the first state, and 3.9% of the time in the second state. The goodness of fit of the HMM is discussed in the Appendix.

An advantage of the HMM over the Markov chain first lies in the interpretation of the hidden states. The Viterbi algorithm provides a fast procedure to estimate the most likely states of the Markov chain that have given rise to the observations, given the parameter estimates.[9] Denote the clean state by C, and the contaminated or noncompliant state by N. We present in Table V sections of the time series under study, together with the estimated hidden states using the Viterbi algorithm. Periods of high risk emerge more clearly from the hidden states: clusters of 1s in the observation sequence are put together in a unique N state, and we refer to this importer as noncompliant during this time interval. This suggests that quarantine services should monitor more closely this importer during this period. The hidden state associated with isolated 1s in the time series is classified as C since no high-risk period is detected, and no immediate action needs to be taken.

The corresponding ROC and leakage curves are presented in Fig. 2. The AUC for the Markov chain is 0.687, and 0.772 for the HMM. It can be seen from these curves that both the Markov chain and the HMM help reduce leakage for this importer. The HMM performs better compared to the Markov chain. From the model estimates, the probability of observing a 0 given a long sequence of 0s has been observed in the recent history is 0.9854. The ROC curve obtained from the HMM model suggests inspecting consignments when the probability of noncompliance drops slightly below the latter value, that is, $p_T = 0.985$, where p_T was defined in Section 2.4. The non-Markovian nature of the HMM implies that after a period of noncompliance, corresponding to the N states estimated using Viterbi, it will take some time for this probability to reach the threshold again. Practically, this means that inspection services should keep monitoring an importer for some time after a pattern of contamination is detected, until a period of low risk is reached again, corresponding to the C state. The HMM provides guidance to estimate how long this monitoring should take place. For the threshold $p_T = 0.985$, we infer from the ROC and leakage curves that for the year 2011, this will induce 29% of false positive, 86% of true positives, and 14% of leakage, corresponding to the inspection of 27 consignments out of 80, which represents a reduction of intervention to about a third of the pathway being inspected. The evolution of the predicted probability of compliance under the Markov chain and the HMM models are given in Table VI.

4. DISCUSSION

We have developed and fitted a collection of time series to border inspection data. When there

Table V. Observation Sequence and Most Likely Hidden States Estimated Using the Viterbi Algorithm; Symbols C and N Correspond to the Compliant and Noncompliant State, Respectively

Sequence	...0	0	0	1	0	1	0	0	0	1	1	1	1	0	0	0	0...
Hidden state	...C	C	C	N	N	N	N	N	N	N	N	N	N	C	C	C	C...
Sequence	...0	0	0	1	0	0	0	...	0	1	1	1	0	1	1	0	0...
Hidden state	...C	C	C	C	C	C	C	...	C	N	N	N	N	N	N	C	C...

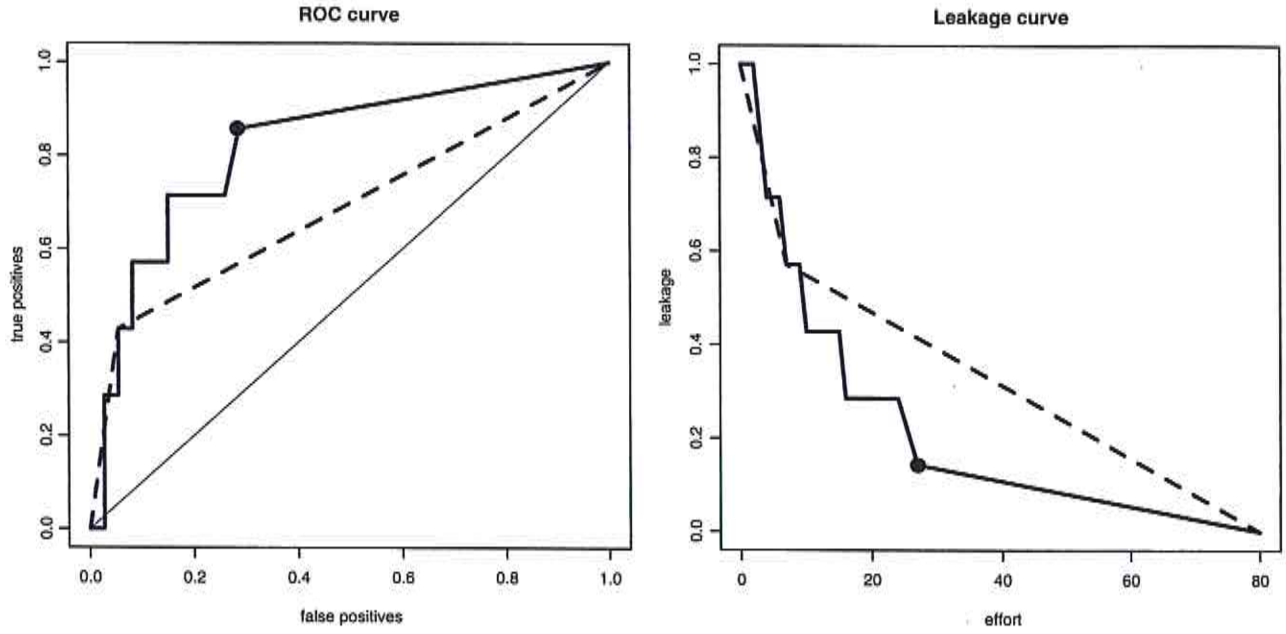

Fig. 2. ROC and leakage curves for the HMM (solid line) and the Markov chain (dashed line), with no covariates. Different sections of these curves correspond to the threshold p_T, not on the model complexity; see Section 2.4 for details. The black dots correspond to the adopted strategy for $p_T = 0.985$.

is evidence of serial dependence, the principal benefit of a HMM for this problem is that it provides a framework within which one can answer the question: If we inspect an item and it is contaminated, how does that change our best estimate of the probability that the next item will be contaminated? That is, the HMM provides a way of using recent history to flexibly allocate inspection resources among a suite of different pathways, depending on recent history. The second benefit of HMM lies in the interpretation of the model and can help inspection services to estimate how long a specific importer or supplier should be monitored after he was found noncompliant. For the two pathways considered here, there was no evidence of systematic contamination at the pathway level. For the stockfeed pathway, we provided an example with the presence of contamination patterns, and explained how the proposed methodology can be used to help inspection decisions.

As noted in the Section 1, for the reported example, we assumed perfect detection, that is, inspection of a contaminated item would certainly detect the contamination. This is not a realistic assumption for most inspection systems. Our model will extend to inspection systems in which imperfect detection is possible, and can do so explicitly so long as independent estimates of the detection rate can be made. We modeled the outcome of inspection using a conditional Bernoulli random variable. It is straightforward to model the outcome of inspection using two conditional Bernoulli random variables: the first that the item is contaminated, say with probability p, and the second that the contamination is detected, say with probability π. The random variable so modeled is identical to a single Bernoulli random variable with probability of success $p \times \pi$. Then the independent estimate of the failure to detect contamination can be integrated into the modeling framework.

Table VI. Evolution of the Predicted Probability of Compliance $\mathbf{P}(Y_{i+1} = 0 \mid \mathbf{y}_{1:i}, \mathbf{z}_{1:i})$ with Time, for a Fraction of the Year 2011

Status	HMM	Markov Chain
0	0.9854	0.9805
1	0.9854	0.9805
0	0.8399	0.3889
0	0.9149	0.9805
0	0.9541	0.9805
0	0.9720	0.9805
0	0.9798	0.9805
0	0.9831	0.9805
1	0.9846	0.9805
1	0.8210	0.3889
1	0.5474	0.3889
0	0.5319	0.3889
1	0.6384	0.9805
1	0.5345	0.3889
0	0.5316	0.3889
0	0.6380	0.9805
0	0.7638	0.9805

Note: The first column displays the status of the consignment. The second column is the predicted probability without covariates, for the HMM, and the third column for the Markov chain.

Alternatively, if the estimate is unknown, but assumed to be constant across the pathways, then it can be accounted for in the interpretation of the output, but would likely not affect decisions such as whether covariates are useful, and which covariates are best.

The current models are fit on the full observation sequence. In the case of a policy change, any deviation from 100% inspection will influence parameter estimation and prediction on future shipments. Maximum likelihood can still be performed using the EM algorithm, designed to estimate parameters for models with latent variables. Maximum likelihood estimation is unaffected by missingness in the data; however, the estimates may be more biased and will be less efficient relative to those fitted on the full data set. For HMM, this amounts to including the missing observations with the unobserved hidden states. An efficient way to compute the likelihood for HMM when data are missing can be found, for example, in Ref. 9; see Section 2.3.3. Since the one-step-ahead probability is expressed as the ratio of likelihood, it can still be quickly calculated when data are missing.

As pointed out by an anonymous reviewer, we have elected to fit our models using symmetric objective or loss functions. The disadvantage of symmetry in the loss function is that it implies that false positives are of equal importance as false negatives for the purposes of parameter estimation. This implication is operationally unrealistic in most settings, and very much so in biosecurity, where false positives result in the unnecessary inspection of uncontaminated consignments, whereas false negatives result in a potential increase in the biosecurity risk. One way to approach the problem is to fit models using asymmetric loss. An alternative is to reinterpret the results in the light of this realization, supported by devices such as ROC curves.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous referees for their many constructive comments, which greatly improved the structure and content of the article. The authors acknowledge financial support from the Australian Research Council Discovery Grant DP110103159 and the Australian Centre of Excellence for Risk Analysis. The authors are also grateful to Adam Dawes and Brendan Woolcott of the Australian Department of Agriculture, Fisheries and Forestry for providing inspection data and many useful conversations.

APPENDIX

A.1. Mathematical Models

We provide some more details about the models, in particular the expression of their likelihood, and the estimation of their parameters. We recall some notation. Observations are denoted $\mathbf{Y}_{1:i} = (Y_1, \dots, Y_i)$, for $i = 1, \dots, n$. Covariates at time i are denoted by $\mathbf{z}_i = (1, z_{i,1}, \dots, z_{i,K})^t$, $i = 1, \dots, n$, where $K \geq 0$ represents the total number of covariates. Therefore, $\mathbf{z}_i(1) = 1$ and $\mathbf{z}_i(j) = z_{i,j-1}$, for $j = 2, \dots, K+1$. Let $\mathbf{z}_{1:i} = (\mathbf{z}_1, \dots, \mathbf{z}_i)$.

A.1.1. Independent Model

The joint log likelihood under this model takes the product form:

$$\log \mathbf{P}(\mathbf{Y}_{1:n} = \mathbf{y}_{1:n} \mid \mathbf{z}_{1:n}) = \sum_{i=1}^{n} \log \mathbf{P}(Y_i = y_i \mid \mathbf{z}_i) .$$

The probability of contamination $p_i = \mathbf{P}(Y_i = 1 \mid \mathbf{z}_i)$ at time i depends only on the value of the covariates at time i through $\text{logit}(p_i) = \alpha^t \mathbf{z}_i$. The $(K+1)$ parameters are estimated using maximum likelihood, which requires solving a system of $(K+1)$ nonlinear equations, with $(K+1)$ unknown parameters,

$$\sum_{i=1}^{n} \mathbf{z}_i(j)(y_i - p_i) = 0, \quad j = 0, \dots, K+1 ,$$

where the dependence on the model parameters is via p_i. An alternative to the independence model would be the independence mixture model, where observations are independent but depend on a hidden state; see, for example, Ref. 12 for an application of mixture models to counts of movements by a fetal lamb. We chose to compare the two alternative models with the independence model instead, bearing in mind that the primary goal is to detect patterns of contamination.

A.1.2. Markov Chain

The joint log likelihood under this model is given by:

$$\log \mathbf{P}(\mathbf{Y}_{1:n} = \mathbf{y}_{1:n} \mid \mathbf{z}_{1:n}) = \log \mathbf{P}(Y_1 = y_1 \mid \mathbf{z}_1) + \sum_{i=2}^{n} \log \mathbf{P}(Y_i = y_i \mid Y_{i-1} = y_{i-1}, \mathbf{z}_{i-1}) .$$

Parameters are estimated using maximum likelihood, by solving a nonlinear system of $(2K+2)$ equations with $(2K+2)$ unknown parameters,

$$\sum_{i=1}^{n-1} \mathbf{z}_i(j)(v_{00}(i) - a_{00}(i)(v_{00}(i) + v_{01}(i))) = 0 ,$$

$$\sum_{i=1}^{n-1} \mathbf{z}_i(j)(v_{10}(i) - a_{10}(i)(v_{10}(i) + v_{11}(i))) = 0 ,$$

for $j = 0, \ldots, K+1$, where $v_{jk}(i) = 1$ if there is a transition from j to k at time $i+1$, that is, $y_i = j$, $y_{i+1} = k$, and $v_{jk}(i) = 0$ otherwise.

A.1.3. HMMs

The joint log likelihood is:

$$\log \mathbf{P}(\mathbf{Y}_{1:n} = \mathbf{y}_{1:n}, \mathbf{X}_{1:n} = \mathbf{x}_{1:n} \mid \mathbf{z}_{1:n}) = \log \mathbf{P}(X_1 = x_1 \mid \mathbf{z}_1) + \sum_{i=2}^{n} \log \mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}, \mathbf{z}_{i-1}) + \sum_{i=1}^{n} \mathbf{P}(Y_i = y_i \mid X_i = x_i, \mathbf{z}_i) ,$$

where $x_i, y_i \in \{0, 1\}$. We use the depmixS4 package in R to maximize the likelihood and estimate the parameters.[17] There are $(4K+5)$ parameters to estimate, which includes the estimate of the initial distribution. Estimation of the parameters is performed using the EM algorithm due to its scalability for large data sets when fitting an HMM with two hidden states.

A.2. Goodness of Fit

We discuss methods to assess the goodness of fit for HMM for the model with no covariates. Methods of goodness of fit for HMM generally assume that the Markov chain is stationary. When estimating the parameters of the HMM, one can either decide to maximize the log likelihood given in Section 2.3.3 by taking the term $\mathbf{P}(Y_1 = y_1 \mid \mathbf{z}_1)$ to be the stationary distribution of the Markov chain, or by estimating it directly. The depmixS4 package estimates the initial distribution, so that the chain is not assumed to be in its stationary state at the origin. We argue here that this is not an impediment to use goodness-of-fit methods, as long as we discard the first few observations of the sequence.

Convergence of a Markov chain with transition matrix $\mathbf{A}$ toward its stationary distribution occurs at an exponential rate, which rate depends on the size of the eigenvalues λ_i of $\mathbf{A}$. For a stochastic matrix $\mathbf{A}$, 1 is always an eigenvalue since $\mathbf{A1} = \mathbf{1}$, with the right eigenvector $\mathbf{1}$ with all entries equal to 1. If a stationary distribution π exists, then the left eigenvector is π, since $\pi\mathbf{A} = \pi$. Moreover, if $\mathbf{A}$ is an irreducible, aperiodic stochastic matrix, then $\lambda_1 = 1$ satisfies $\lambda_1 > |\lambda_i|, i \neq 1$, for any other eigenvalue. For a two-by-two matrix, then one can show that $\mathbf{A}^n = \mathbf{1}\pi + O(|\lambda_2|^n)$, so that convergence toward the stationary distribution is at the exponential rate $|\lambda_2|$.

The estimated matrix $\hat{\mathbf{A}}$ in 1 has eigenvalue $\lambda_2 = 0.871$. By discarding the first 40 samples, given that λ_2^{40} is of order 10^{-3}, one can assume that the chain has reached its stationary distribution. Checking for the goodness of fit can be done on the remaining observation sequence. The value of λ_2 provides guidance to know how many samples to remove, which varies from one pathway to another.

First, we check if the model captures the serial dependence structure present in the data. In Ref. 1, the author provides a graphical technique for assessing the goodness of fit of a stationary HMM by comparing the empirical m-dimensional distribution, say $\bar{F}_n^m$ where n is the sample size, with the estimated one, say $\hat{F}^m$. Here we focus on bivariate and trivariate distributions ($m = 2$ and 3). The bivariate empirical distribution is defined as:

$$\bar{F}_n^2(y_1, y_2) = \frac{1}{n-1} \sum_{t=1}^{n-1} \mathbf{1}(Y_t \leq y_1, Y_{t+1} \leq y_2) ,$$

for $y_i \in \{0, 1\}$, where $\mathbf{1}(\cdot)$ is the indicator function, and the estimated bivariate distribution $\hat{F}^2(y_1, y_2)$ is given by:

$$\sum_{k_1=0}^{1}\sum_{k_2=0}^{1}\pi_{k_1}a_{k_1k_2}\mathbf{P}(Y\leq y_1\mid X=k_1)\mathbf{P}(Y\leq y_2\mid X=k_2),$$

where the Markov chain has entries $\mathbf{A}=(a_{ij})$ and stationary distribution π. The trivariate distributions are defined in a similar way. Assuming observations come from a stationary HMM with m-dimentional distribution F^m, it is shown in Ref. 1 that both the estimated and empirical distributions converge to F^m as n tends to ∞. Therefore, plotting one against the other, the plot should line on a 45° line passing through the origin. We present in the left panel of Fig. A1 a comparison of the m-multivariate distributions for $m=2$ and $m=3$. The model captures the bivariate and trivariate distributions very well.

The serial dependence can also be addressed via the autocorrelation function ρ. For a stationary binomial HMM with transition matrix $\mathbf{A}=(a_{ij})$, stationary distribution π, and state dependent probabilities $\mathbf{p}=(p_0,p_1)$, we get $\rho(k)=w^k(1+\alpha)^{-1}$, with $w=a_{00}-a_{10}$, $\alpha=(\pi\mathbf{p}^t-\pi\mathbf{P}\mathbf{p}^t)/(\pi\mathbf{P}\mathbf{p}^t-(\pi\mathbf{p}^t)^2)$, and $\mathbf{P}=\left(\begin{smallmatrix}p_0 & 0\\ 0 & p_1\end{smallmatrix}\right)$.(18) The right panel in Fig. A1 compares $\rho(k)$ for $k=0,\ldots,20$ with the sample autocorrelation function. The HMM captures the decay of the autocorrelation well, with a particularly good match at lags 1 and 2.

The empirical run length of contaminated items arriving at the border is compared to the model prediction. A run of s consecutive ones is defined to be the observation of a 01 sequence followed by another $s-1$ consecutive ones, and a zero, in that order. Denote by S the length of a run of ones. The event $\{S=s\}$ has probability $q_s=\mathbf{P}(S=s)$ given by:

$$\mathbf{P}(Y_{i+2}=\ldots=Y_{i+s}=1,\,Y_{i+s+1}=0\mid Y_i=0,\,Y_{i+1}=1),$$

for $s=1,2,\ldots$. An explicit expression for q_s for a stationary HMM can be found, for example, in Ref. 13, Section 2.6.3. The empirical estimate for q_s is f_s/f where f_k denotes the number of runs of s consecutive ones and $f=\sum_s f_s$. Table A1 compares the observed run-length distribution with the one predicted by the model for the HMM model fitted in Section 2.3.3. There is a good match for $s=1$ and 2, with an outlier at $s=4$. We expect a closer match as the amount of historical data increases.

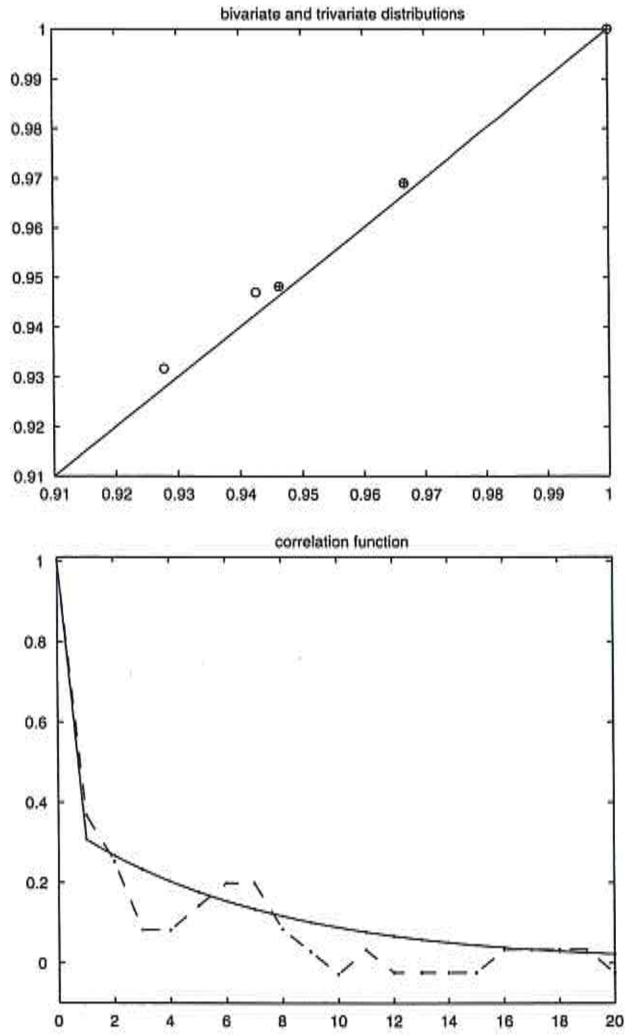

Fig. A1. The top panel presents the empirical (x-axis) versus predicted (y-axis) bivariate (crosses) and trivariate (circles) distributions. The bottom panel displays the estimated (dashed line) versus predicted (solid line) autocorrelation function. The observation sequence corresponds to a specific importer from the stockfeed pathway, presented in Section 3.2.'

Table A1. Observed Versus Predicted Run-Length Distribution of Ones, $\mathbf{P}(S=s)$

s	1	2	3	4	5	6
Predicted distribution q_s	0.738	0.141	0.064	0.030	0.014	0.007
Empirical distribution f_s/f	0.727	0.091	0	0.182	0	0

REFERENCES

1. Elliston L, Yainshet A, Hinde R. Karnal bunt: The regional economic effects of a potential incursion. ABARE, eReport 04.4, prepared for Plant Health Australia, Canberra, 2004.
2. Chen X, Cheng J, Xie M. A penalized regression approach in detection of nuclear materials in shipment to the United States. Joint Statistical Meetings, Vancouver, 2010.
3. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989, 77:257–286.
4. Albert PS. A two-state Markov mixture model for a time series of epileptic seizure counts. Biometrics, 1991; 47:1371–1381.
5. Leroux BG, Puterman ML. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics, 1992, 48:545–558.
6. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology: Applications to protein modeling. Journal of Molecular Biology, 1994, 235:1501–1531.
7. Dutilh G, Wagenmakers EJ, Visser I, van der Maas HLJ. A phase transition model for the speed accuracy trade-off in response time experiments. Cognitive Science, 2011, 35:211–250.
8. Visser I, Raijmakers MEJ, van der Maas HLJ. Hidden Markov models for individual time series. Pp. 269–289 in Valsiner J, Molenaar PCM, Lyra MCDP, Chaudhary N (eds). Dynamic Process Methodology in the Social and Developmental Sciences. New York, Springer, 2009.
9. Zucchini W, MacDonald IL. Hidden Markov Models for Time Series. An Introduction Using R. Boca Ratan, FL: Chapman and Hall, 2009.
10. Beale R, Fairbrother J, Inglis A, Trebeck D. One Biosecurity: A Working Partnership. Commonwealth of Australia, 2008.
11. DAFF Reform of Australia's biosecurity system. CC BY 3.0.
12. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall, 1993.
13. Visser I, Speekenbrink M. depmixS4: An R package for hidden Markov models. Journal of Statistical Software, 2010, 36:1–21.
14. Altman RM. Assessing the goodness-of-fit of hidden Markov models. Biometrics, 2004; 60:444–450.
15. MacDonald IL, Zucchini W. Hidden Markov and Other Models for Discrete-Valued Time Series. New York: Chapman and Hall, 1997.