A hidden Markov model approach for determining vessel activity from vessel monitoring system data

David Peel and Norman M. Good

Abstract: Many fisheries worldwide have adopted vessel monitoring systems (VMS) for compliance purposes. An added benefit of these systems is that they collect a large amount of data on vessel locations at very fine spatial and temporal scales. This data can provide a wealth of information for stock assessment, research, and management. However, since most VMS implementations record vessel location at set time intervals with no regard to vessel activity, some methodology is required to determine which data records correspond to fishing activity. This paper describes a probabilistic approach, based on hidden Markov models (HMMs), to determine vessel activity. A HMM provides a natural framework for the problem and, by definition, models the intrinsic temporal correlation of the data. The paper describes the general approach that was developed and presents an example of this approach applied to the Queensland trawl fishery off the coast of eastern Australia. Finally, a simulation experiment is presented that compares the misallocation rates of the HMM approach with other approaches.

Résumé : Plusieurs pêches commerciales dans le monde ont adopté des systèmes de surveillance des navires (VMS) pour vérifier leur conformité. Un avantage additionnel de ces systèmes est la récolte d'une quantité importante de renseignements sur la position des navires à des échelles spatiales et temporelles très fines. Ces données peuvent constituer une mine d'information pour l'évaluation des stocks, la recherche et la gestion. Cependant, comme la plupart des versions de VMS enregistrent la position du navire à intervalles fixes sans égard à l'activité de ce navire, il faudrait une méthode pour déterminer quels enregistrements de données correspondent à des activités de pêche. Notre travail décrit une approche probabiliste, basée sur des modèles de Markov cachés (HMM), pour déterminer l'activité des navires. Un HMM fournit un cadre naturel pour ce problème et, par définition, il modélise la corrélation temporelle intrinsèque des données. Nous décrivons l'approche générale que nous avons mise au point et présentons un exemple de cette approche utilisée avec des données de la pêche commerciale au chalut du Queensland au large de la côte de l'Australie orientale. Enfin, nous présentons une expérience de simulation qui compare les taux d'attributions erronées de la méthodologie des HMM à ceux d'autres méthodes.

[Traduit par la Rédaction]

Introduction

Vessel monitoring systems (VMS) are being adopted as compliance tools in many fisheries, for example, to detect vessels encroaching into closures or protected areas or registering number of days at sea to enforce fishing effort quotas. A secondary outcome of this technology is large databases of fine-scale spatial and temporal information on vessel positions. This data can be invaluable for stock assessment, research, and management in a number of ways. Some examples of the uses of fine-scale spatial fisheries data include mapping of fishing intensity at fine temporal and spatial scales (Larcombe et al. 2001; Marrs et al. 2002; Harrington et al. 2007), measuring depletion (Deng et al. 2005), assessing trawling impacts (Rijnsdorp et al. 1998), and estimating spatial abundance (Vignaux et al. 1998; Bertrand et al. 2005; Good and Peel 2007).

A VMS collects positional data of vessels at regular–irregular time intervals (i.e., a poll), which is generally in the range of every 1 or 2 hours. This polling occurs irrespective of the vessel's activity (i.e., trawling, at anchor, or steaming). To use the VMS data beyond compliance, we require an indication of the vessel's activity. In general, we are mainly interested in trawling or fishing activity. However, it is foreseeable that in some instances the spatial distribution of steaming or the location of anchorages may be of interest. In some VMS implementations the vessel's activity may be known, such as when information from net winch sensors is

Received 30 July 2010. Accepted 26 March 2011. Published at www.nrcresearchpress.com/cjfas on 6 July 2011.
J21947

Paper handled by Associate Editor Terrance Quinn II.

D. Peel. Wealth from Oceans National Research Flagship and CSIRO Mathematics and Informatics and Statistics, Castray Esplanade, Hobart TAS 7001, Australia.
N.M. Good.[*] Department of Employment, Economic Development and Innovation (DEEDI), Agri-Science Queensland, Sustainable Fisheries, Northern Fisheries Centre, P.O. Box 5396 Cairns, 4870, Australia.

Corresponding author: D. Peel (e-mail: David.Peel@csiro.au).

[*]Present address: CSIRO Mathematics and Informatics and Statistics – Australian e-Health Research Centre, Level 7, UQ CCR Building 71/918 Royal Brisbane and Women's Hospital, Herston QLD 4029, Australia.

doi:10.1139/F2011-055

available to determine trawling activity (e.g., Mejias 1999). Generally this is not the case; however, such data would provide an excellent test data set to quantify the error of a classification method.

As there is a physical limitation on how fast a vessel can draw a trawl net through the water (which is considerably slower than a vessel's normal steaming speed), an obvious indicator of a vessel's activity is the vessel's speed. So, in theory, for a given vessel, there would be a trawling speed and a steaming speed. The reality is not as straightforward, as the trawling and steaming speeds at any given time are determined by a number of factors, for example, vessel engine power, currents and tides, weather, size of net, type of gear, target species, captain–crew preference, and frequency of tows. Also the VMS data does not directly provide average vessel speed over the poll period, but rather vessel speed is calculated from the direct distance from the previous poll divided by the polling time interval. Since vessels do not always travel in straight paths, this measure of speed will generally be biased downwards. However, we would generally expect a steaming vessel to take a straighter, more direct route than when trawling. Therfore, the bias will reduce trawling speeds and increase steaming speeds and so may actually help discriminate between the two activities. The calculated speed may also be effected by non-trawling activity within the polling period (e.g., bringing in – letting out – emptying nets, turning, sorting catch, etc.), during which time the vessel may slow. These activities are ignored and considered as unmodelled noise in our approach.

Another indicator of a vessel's activity is the time of day. In many fisheries, trawling generally occurs only during certain periods of the 24 h day, beause of regulation or target species behaviour (e.g., fishing occurs only during the night). At the simplest, all points occurring during the non-trawling period could be removed. However, after examining the data in our example and consulting with fishermen, it seems that there is some variation on when fishing occurs (particularly at the boundary between trawl periods).

The existing literature does not fully address the problem of determining vessel activity from VMS data. One reference that examined the problem specifically was Mills et al. (2007), which presents a speed and direction-based approach. So it must be assumed that generally, in practice, some rule based solely on speed is used to filter out vessels travelling too fast to be trawling. However, although a simple speed cutoff approach may often work in practice, this approach has several limitations. Firstly, it ignores the temporal correlation of the data, which is the probability that a vessel trawling at the next poll is not independent of the activity at the current poll. Secondly, an outright classification of the data into vessel activity does not provide any measure of the uncertainty of the classification (e.g., polls that correspond to speeds close to the boundary of plausible fishing speeds are less certain to be trawling). Also in some applications some further adjustments or extras steps are required for the method to perform well. For example, in practice often vessels slowing to enter or exit anchorages appear to be trawling, based on speed alone. Finally, it is difficult to incorporate other covariates into the "model" (i.e., we may wish to include other covariates, such as habitat or bottom type, to either inform the classification or to possibly examine the relationship between vessel activity and the covariate).

Initially, to determine vessel activity, we investigated an approach using finite mixture models fitted to individual vessel's historical speed and time of day data (Peel et al. 2007). While this approach did solve many of the issues, the temporal correlation of the data was not addressed. To model the temporal correlation, a natural avenue to pursue is hidden Markov models (HMMs), which can be thought of as a mixture model with autocorrelated hidden membership variables. The work in this paper extends and refines HMM for VMS, initially described in Peel and Good (2007). HMM are being used in many applications, including speech recognition, econometrics, biology, and image processing (see MacDonald and Zucchini 1997 or McLachlan and Peel 2000 for a more complete list of references). The most relevant to the application described in this paper is modelling animal movement from tagging data (e.g., Patterson et al. 2009).

This paper describes the HMM method and how it can be applied to VMS data to probabilistically classify vessel activity. The method is evaluated in a simulation experiment, and a real world application to a reasonably sized data set is presented.

Materials and methods

Description of model

A HMM is a statistical model of a system that switches over time between a finite number of states, or nodes, along predefined links. A HMM assumes that the system is a Markov process. That is, the conditional probability of the future states, given the present, is independent of past states. In the context of this paper, the states, or nodes, correspond to the vessel's activity. The process involves the activity of the vessel discretely stepping around this network, with each step corresponding to a polling event. Some transition links in the network loop back to the same state, allowing a possible step outcome to be that the vessel does not change state or activity.

The true state is hidden, as we do not observe it directly; rather a quantity whose distribution is dependent on the current node–state is observed. In this application, as discussed in the previous section, an obvious observable indicator of vessel activity or state is the vessel's calculated speed. Other possible indicators are general location, time of day, and direction.

A graphical representation of the model we used is provided (Fig. 1). The model includes the nodes corresponding to trawling, steaming, and at anchor (stationary). Also included are special "Entry" and "Exit" nodes. The purpose of the Entry and Exit nodes is to address an important issue in the data that we call false trawling. Because of the discrete nature of polling, when vessels approach or leave an anchorage at steaming speed a vessel will often be misclassified as trawling. To illustrate why this occurs, suppose a vessel is polled while returning to port at steaming speed, then arrives in port and is stationary until it is polled again. The calculated speed for this interval will be greater than zero as the vessel was steaming for the first portion of the poll interval, but will generally be less than the steaming speed (i.e., closer to trawling speed) because the vessel was stationary during

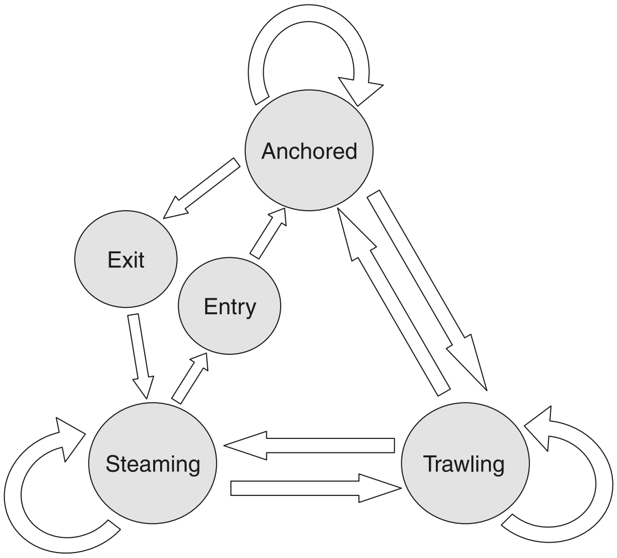


Fig. 1. Representation of a hidden Markov model (HMM) with the states–nodes (circles) and transition links (arrows) indicated. The states are the obvious anchored, trawling, and steaming, plus states exit and entry to avoid vessels entering port being falsely classified as trawling.

the second portion of the polling interval. The Entry–Exit nodes model this transition between anchorage and steaming. Similar nodes were not required between anchoring and trawling, because if trawling speed is biased downward upon entering a anchorage (trawling can occur when entering and exiting an anchorage, when the anchorage is very close to the trawl ground), it would most likely still be classified as trawling. However, the inclusion of such nodes may alleviate the observed non-normality of the trawling speeds that we discuss later in this section.

To describe the HMM more formally, consider a discrete Markov chain with s states and n time steps. The state of the process at time t is denoted by an indicator variable $z_{t,j}$, where $z_{t,j} = 1$ if the process is in state j at time t and 0 otherwise. The transition probability matrix $\mathbf{P}$ is defined such that

$$(1) \qquad P_{i,j} = \Pr\left[z_{t+1,j} = 1 | z_{t,i} = 1\right]$$

for all times $t = 1, \ldots, n - 1$. The current state at any given time t is not observed directly, so z is a hidden variable. However the measurement of calculated speed (y_t) is observed and it is assumed to be distributed as

$$(2) \qquad f(y_t | z_t; \phi) = \prod_{j=1}^{s} f_j(y_t, \theta_j)^{z_{t,j}}$$

where for node j ($j = 1, \ldots, s$), f_j is a suitably chosen probability density function, θ_j denotes the parameters of the density, and $\phi = \{\theta_1, \theta_2, \ldots, \theta_s\}$.

The choice of f_j ($j = 1, \ldots, s$) in eq. 2 should be driven by the data (i.e., the choice of state distributions could be based on examination of the empirical distributions of VMS calculated speeds; for example, see Fig. 2).

The anchorage distribution was taken as a half normal (Johnson et al. 1994), with fixed mean of zero and a small variance, to encompass speeds very close to zero:

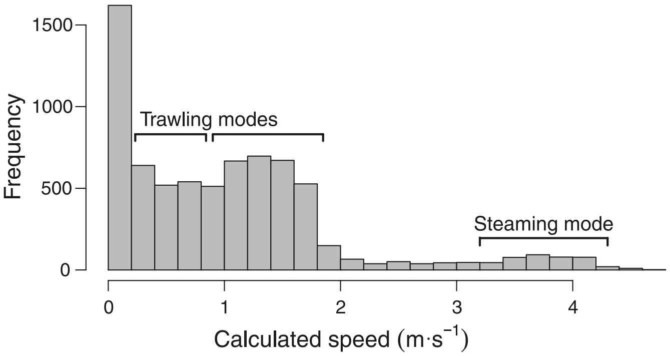


Fig. 2. Histogram of vessel monitoring systems (VMS) calculated speed (m·s^{-1}) for a selected vessel between January 2000 and March 2003. Obvious trawling and steaming modes are indicated. Notice also the almost bimodal nature of the trawling speeds. Polls with zero speed have been removed from the histograms for clarity.

$$(3) \qquad f_1(y; \theta_1) \sim \mathrm{HalfN}(y; 0, 0.01)$$

where HalfN(y; μ, σ^2) denotes the half-normal distribution with means μ and variances σ^2.

It was found that the distribution of trawling speeds was almost bimodal for some fisheries (as can be seen in Fig. 2). Upon further investigation, this seemed less prevalent when the polling frequency was increased. So the bimodal distribution is possibly due to non-trawling activities occurring during the polling period (e.g., sorting catch or turning around). To handle the non-normality of the data, a finite mixture model was used with unequal variances

$$(4) \qquad f_2(y; \theta_2) \sim \sum_{k=1}^{K} \pi_k N(y; \mu_{2k}, \sigma_{2k}^2)$$

where $N(y; \mu, \sigma^2)$ denotes the normal distribution with means μ and variances σ^2, and the mixing proportions are given by π_k.

The steaming distribution was taken to be a normal distribution

$$(5) \qquad f_5(y; \theta_5) \sim N(y; \mu_5, \sigma_5^2)$$

with mean μ_5 and variance σ_5.

The entry and exit distributions were assumed to be a uniform distribution with limits taken as the means of the anchorage and steaming distributions, i.e., for the entry state

$$(6) \qquad f_3(y; \theta_3) \sim U(y; 0, \mu_5)$$

and for the exit state

$$(7) \qquad f_4(y; \theta_4) \sim U(y; 0, \mu_5)$$

where $U(y;\ a,\ b)$ denotes the uniform distribution with bounds a and b.

The trawling state parameters π_k, μ_{2k}, σ_{2k} ($k = 1, \ldots, K$) and the steaming state parameters μ_5, σ_5 are unknown parameters to be estimated.

To model the effect of the time of day, the transition matrix, $\mathbf{P}$ is taken to depend on time of day d, with each hour of the day having its own independent transition matrix. For example, in a particular fishery the probability of moving to the trawling state during the day may be very low, and moving to anchorage–steaming transition probabilities may be high, and vice versa during nighttime. It is assumed that the vessel's trawl and steaming speeds will be unrelated to time of day (i.e., the state distribution parameters ϕ do not depend on time of day).

The model was also extended to model the change in behaviour of vessels corresponding to the fisheries they are targeting at any given time, with each fishery treated independently with certain parameters chosen to be fishery-specific. Trawling behaviour can be quite different between fisheries (Fig. 3). To incorporate a fishery component in the model, each fishery c was allowed to have its own independent transition probabilities. For example, in some fisheries where very long trawls are common, the transition probability of remaining in the trawling state would be higher than a fishery where short, quick periods of trawling are the norm. In addition, since vessels may possibly trawl at different speeds when in different fisheries, the state distribution was taken to depend on fishery: $\phi_c = \{\theta_{1c}, \theta_{2c}, \ldots, \theta_{sc}\}$.

So the transition matrix $\mathbf{P}$ (from eq. 1) is now the probability of moving form state i to state j given time of day (d) and fishery (c) (i.e., for $t = 1, \ldots, n - 1$):

$$(8) \qquad P_{i,j}(d,c) = \Pr\left[z_{t+1,j} = 1 | \{z_{t,i} = 1\}, d, c\right]$$

In our application, the fishery c can be determined by the spatial location, the time of day, and caught species. It should be pointed out that even though we use the term "fishery", the aim is not necessarily to accurately identify true targeted species. Rather, we simply wish to provide a mechanism in the model to handle heterogeneity caused by differing trawl behaviour. So it is not critical to perfectly identify target species as long as fishing behaviour within our defined fishery is reasonably homogeneous for the vessel in question. However, it does seem that target species is a reasonable criterion on which to base the grouping strategy.

In our example application, information on catch species is available, and there exists a reasonable spatial separation of fisheries. Therefore, it is possible to use a simple function of most caught species to determine the fishery. In some fisheries, using the most caught species to determine the fishery would definitely not be appropriate, and a more complex approach would be required (for example, the model could also estimate the target species–fishery).

Fitting of model

The HMM is fitted to the data for each vessel independently to ascertain the parameters $\Omega = \{\phi, \mathbf{P}\}$ of the HMM and the transition matrix of the conditional probabilities

$$(9) \qquad w_{i,j,t} = \Pr\left[z_{t,i} = 1 \quad \text{and} \quad z_{t+1,j} = 1 | y_t, d_t, c_t\right]$$

Taking the row sums of $w_{i,j,t}$ provides the posterior probabilities of the vessel belonging to each state at each time point (i.e., $\Pr[z_{t,j} = 1 | y_t, d_t, c_t]$). These posterior probabilities then provide a probabilistic measure of being in any given state at time t.

The HMM can be fitted, obtaining estimates of $w_{i,j,t}$, using a brute force maximization of the likelihood

$$(10) \qquad L(y, \Omega) = \prod_{t=1}^{n} f(y_t | z_t; \phi) = \prod_{t=1}^{n} \prod_{j=1}^{s} f_j(y_t, \theta_j)^{z_{t,j}}$$

However it is more efficient to maximize the likelihood via the Baum–Welch algorithm (Baum et al. 1970; see also Rabiner 1989), which can be described as a special case of the EM algorithm (Dempster et al. 1977). In essence the EM algorithm involves two steps: the E and M steps.

On the (k + 1)th iteration, given estimates of parameters $\Omega^{(k)}$, the E step estimates the transition matrix of posterior probabilities, $w_{i,j,t}^{(k+1)}$, of moving from node i to node j at time point t. This is done via a forward–backward algorithm (see Appendix A). Since the trawling state's speed distribution is a mixture model, we must also estimate the posterior probabilities of each of the normal mixture components $\tau_{t,k}^{(k+1)}$, at time $t = 1, \ldots, n$, and $k = 1, \ldots, g$ (see Appendix A).

The M step estimates the state parameters Ω, via maximum likelihood, given estimates of the transition probabilities $w_{i,j,t}^{(k+1)}$. So for example, in our application the parameters π_k, μ_{2k}, σ_{2k}, ($k = 1, \ldots, K$), μ_5, and σ_5 are the unknown parameters to be estimated (see Appendix A for further details).

This circular algorithm is repeated, refining the solution, until some form of convergence is achieved (e.g., when the log-likelihood changes by less than some predefined tolerance). In practice, it is possible that there may be a lack of historical data for some vessels (e.g., if a new vessel joins the fleet). In this case default model parameters Ω based on similar boats could be used. Another issue of data size may occur when a vessel switches to a different fishery for a very short time only. The easiest solution to this problem is to pool the relevant data together with another fishery for that vessel.

Simulation study

A simulation study was conducted to examine the performance and validate the HMM method, when time correlation assumptions are true. The only related work that could be found in the literature was part of the work to establish if VMS data could be used to estimate depletion by Deng et al. (2005). In Deng et al. (2005), a simulation study was presented to examine the effect of polling frequency on the statistical power of trawling classification.

A simulated data set was formed by mimicking vessel behaviour. Firstly, on any given day a realistic time was generated for when the vessel would begin and end fishing. The polls within this fishing time were then allocated a trawling speed. This trawling speed was generated by sampling random time segments from fishermen's actual on-board high frequency GPS data of trawling activity and by calculating a speed from the distance between the polls. Polls either side of the trawling period were then set to steaming and given a steaming speed (N(4,0.3)), and polls corresponding to entering an anchorage were given a biased steaming speed (i.e., false trawling, U(0,4)). The remaining polls were set to speed zero to indicate anchoring. A thousand sample data sets were generated, each consisting of a single vessel polling each hour over a single year. For comparison, a number of methods ranging in complexity were applied to the simulated data.

Firstly, two filters based on a speed rule with speed cutoffs of 2 m·s^{-1} as a naive "guess", based on general fishery information, and a second more informed classification using a

Fig. 3. Plots showing the difference in trawling behaviour between two of the fisheries considered in this paper; eastern king prawn (which generally consists of faster straighter trawling) and scallop (where slower, shorter trawling is the norm). Histograms of the calculated distance travelled between polls for (*a*) eastern king prawn (*Penaeus plebejus*) and (*b*) scallop (*Amusium* spp.). Corresponding plots of trawling positions for (*c*) eastern king prawn and (*d*) scallop (data corresponding to speeds greater than 2 m·s^{-1} have been removed to give a rough indication of trawling). Note that longitude and latitude are not shown because of confidentiality.

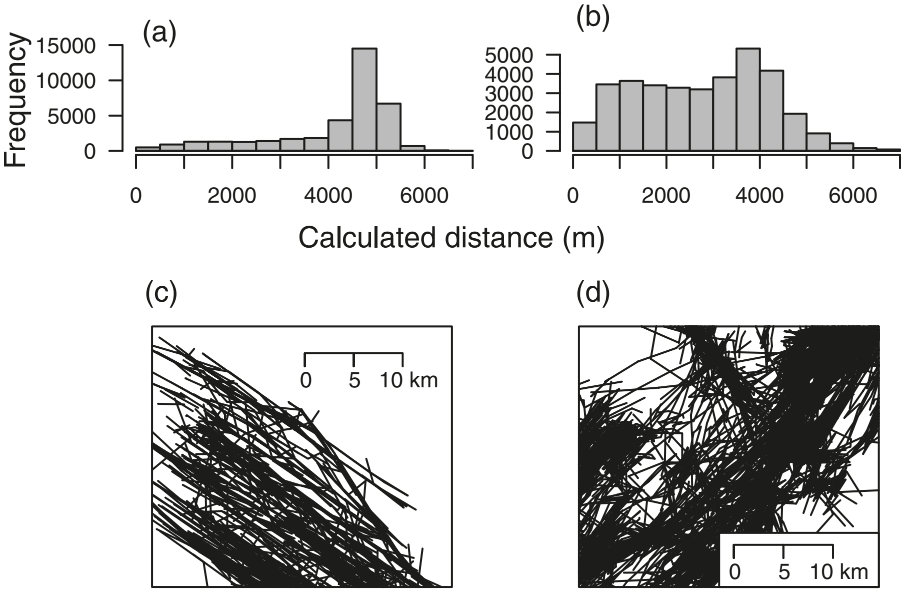


Table 1. Comparison of the misclassification rates for various methods, over a series of 1000 simulated data sets.

	Misclassification rate (%)					
Type of error	Filter: <2 m·s^{-1}	Filter: <1.7 m·s^{-1}	Filter+: <1.7 m·s^{-1}	2 group mixture	3 group mixture	HMM
Trawl as other	2.9	3.4	3.4	3.5	3.2	1.5
(90% CI)	(2.4, 3.3)	(2.9, 3.9)	(2.9, 4.0)	(2.8, 4.0)	(2.7, 3.7)	(0.8, 2.5)
Other as trawl	7.8	6.4	3.5	6.2	6.8	5.1
(90% CI)	(7.3, 8.2)	(6.0, 6.8)	(3.2, 3.9)	(5.6, 6.8)	(6.2, 7.5)	(4.0, 6.0)

Note: “Filter” corresponds to using a speed cutoff rule with the maximum trawling speed indicated; “Filter+” denotes a speed-based rule with a secondary postfixing of false trawling polls; the “2 and 3 group mixture” refer to the fitting of a standard univariate normal mixture to the data to classify between trawling and steaming; and “HMM” corresponds to the hidden Markov model approach. With regard to the types of error, “trawl as other” corresponds to the percentage of trawl polls that were misclassified as other activities (steaming, anchorage, or false trawling). Similarly, “other as trawl” corresponds to the percentage of predicted trawling activity that is actually other activities.

speed cutoff of 1.7 m·s^{-1} based on an empirical examination of the simulated data. For fairness, we also included an ad hoc classification to address the false trawling problem (i.e., we simply applied the filter speed cutoff rule and then forced any polls that were sandwiched between anchoring and steaming to be steaming).

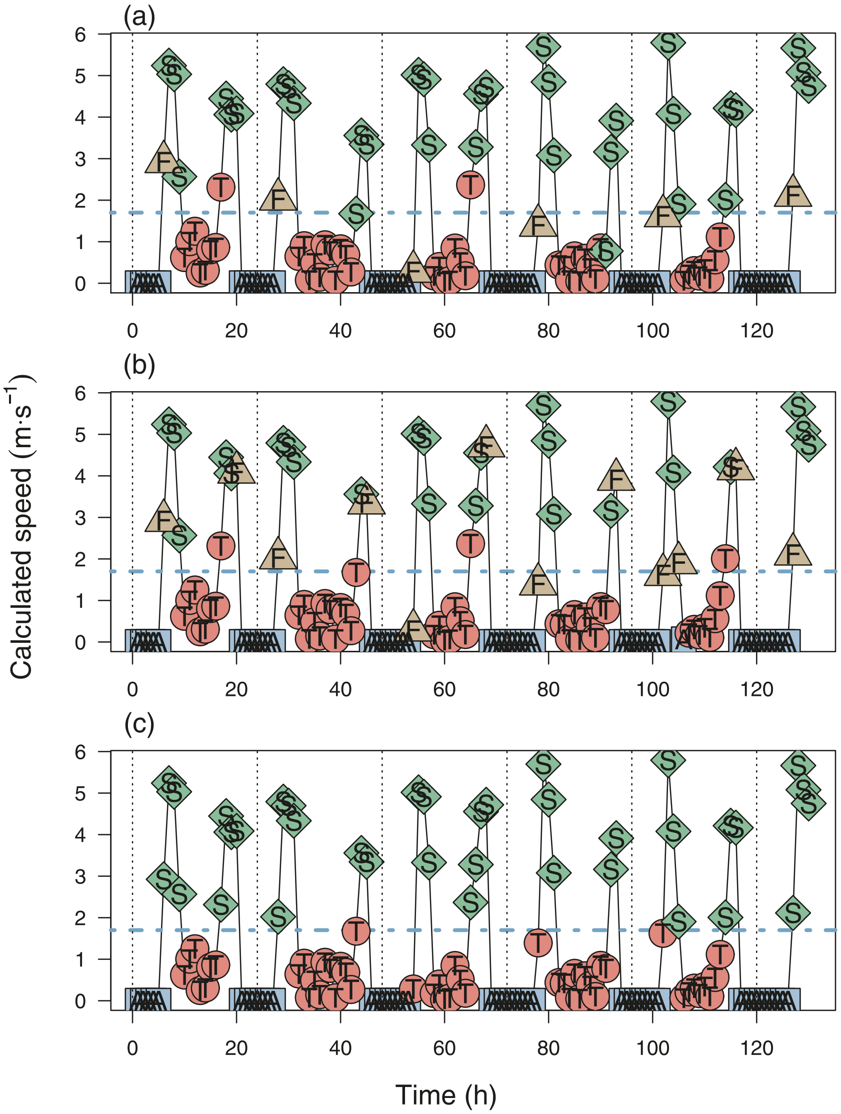
Secondly, two univariate normal mixture models to classify trawling were examined, one with a two-component mixture (trawling and steaming) and one with a three-component mixture (two components for trawling and one for steaming).

Finally, a HMM was applied to the data, which, as we have discussed, can be thought of as a time-dependent mixture model. The results of the simulation are provided (Table 1), and it can be seen that the misallocation rate for the HMM was substantially lower than a traditional speed rule. However, it should be noted that modifying the simple speed cutoff using an ad hoc fixing of false trawling improves performance drastically. The HMM handles the simulated data very well and correctly identifies most of the polls (Fig. 4). Occasional discrepancies occur at the boundary between trawling and false trawling, as would be expected.

Example

To demonstrate the practical application of the HMM approach, we present an example from the coastal trawl fisheries of Queensland, Australia. VMS has been in operation in the fishery since 2000, covering a fleet of up to approximately 500 trawl vessels. The fishery covers a range of 2000 km of coastline from the New South Wales border to the tip of Cape York (Fig. 5). The fleet mainly targets eastern king prawn (*Penaeus plebejus*), tiger–endeavour prawn (*Penaus esculentus*, *Penaeus semisulcatus*, and *Penaeus mono-*

Fig. 4. Examples of a typical fit to a section of time for a simulated vessel. (*a*) True states are shown: pink circles with T denote trawling, green diamonds with S correspond to steaming, blue squares with A show the anchored state, and brown triangles with F indicate the false trawling; (*b*) the corresponding HMM fit; and (*c*) the two-component mixture model fit. The polls in the fit plots have been allocated in an outright classification to states (based on maximum posterior probability) for clarity. The vertical dotted lines denote midnight each day, and the horizontal green line shows the 1.7 m·s^{-1} speed filter.



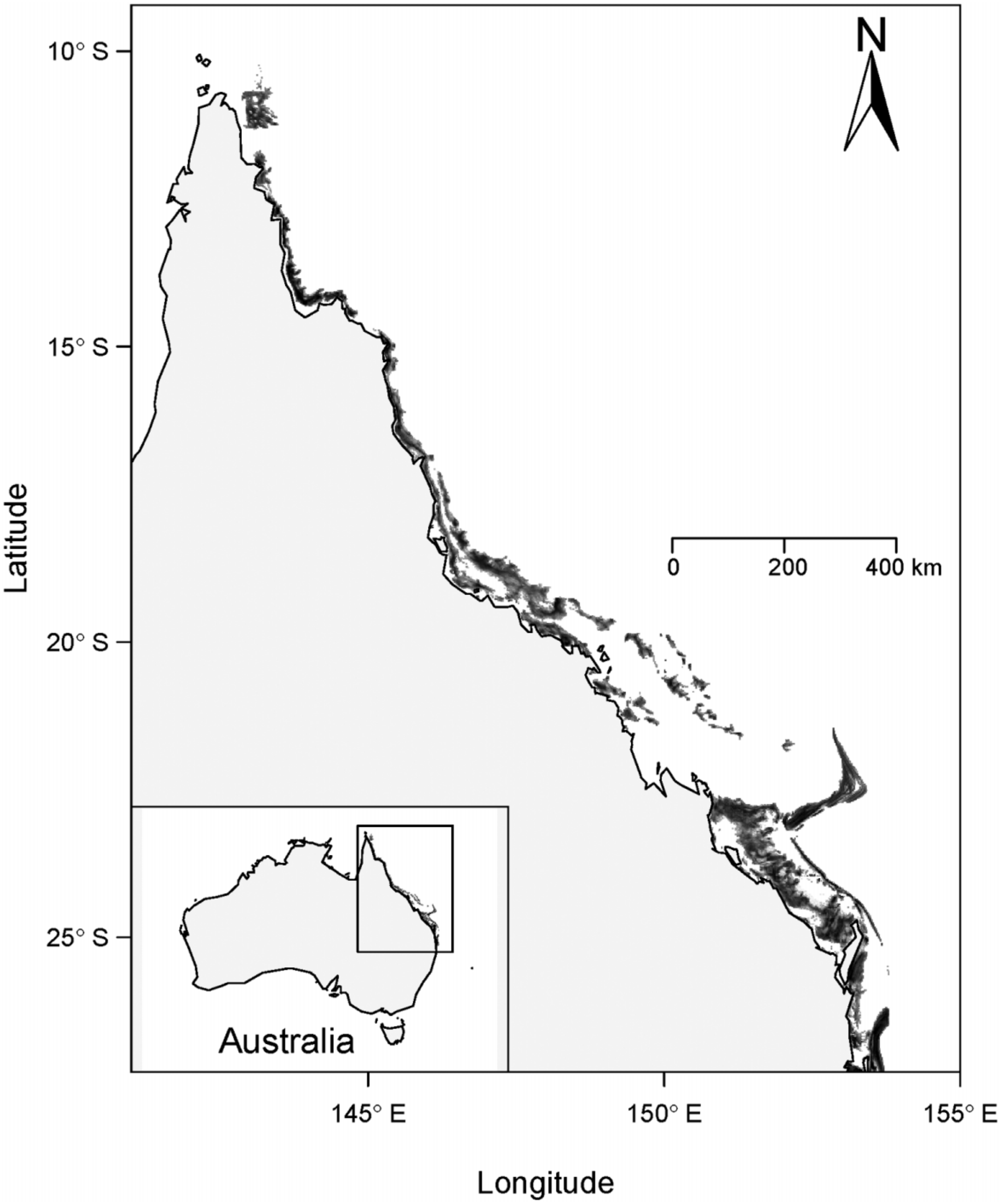
don – Metapenaeus endeavouri and *Metapenaeus ensis*), banana prawn (*Fenneropenaeus merguiensis*), and scallop (*Amusium japonicum* and *Amusium pleuronectes*).

This paper examines 3 years of VMS data recorded between January 2000 and March 2003 from 578 vessels, consisting of 2 791 901 polls, or positions. Extending the analysis to a longer time series of data should, if anything, improve the results, assuming numerical difficulties can be avoided and there is no unmodelled temporal change in vessel behaviour (for example, if the vessel's trawling speed changes over time because of net design changes). The fishery uses a fishermen logbook system to record catch amounts and species per day; this data is matched to the VMS data to provide a catch associated with each VMS data point. Further investigation has been done into using maximum entropy (Vignaux et al. 1998) to better combine logbook catches and VMS data (see Good and Peel 2007; Peel and Good 2007). Once a VMS record has a catch composition, the spatial separation of the fisheries allowed allocation to fisheries to be based on a simple maximum catch rule.

Because of the confidential nature of VMS data, we are unable to disclose the polling frequency, exact locations of individual vessels, or display fishing effort for areas where fewer than five boats were active. To work around this condition, in some of the figures spatial noise has been added, a transformation applied, no longitude–latitude indicated, and some data in sensitive areas hasve been removed.

A model consisting of five states was used (Fig. 1), and the choice of state distributions was based on examination of the empirical distributions of VMS calculated speeds. The

Fig. 5. Location and extent of Queensland trawl fishery. Light grey denotes land, and dark grey indicates general fishing extent.



raw VMS data was matched to logbook records of catch to establish vessel fishery.

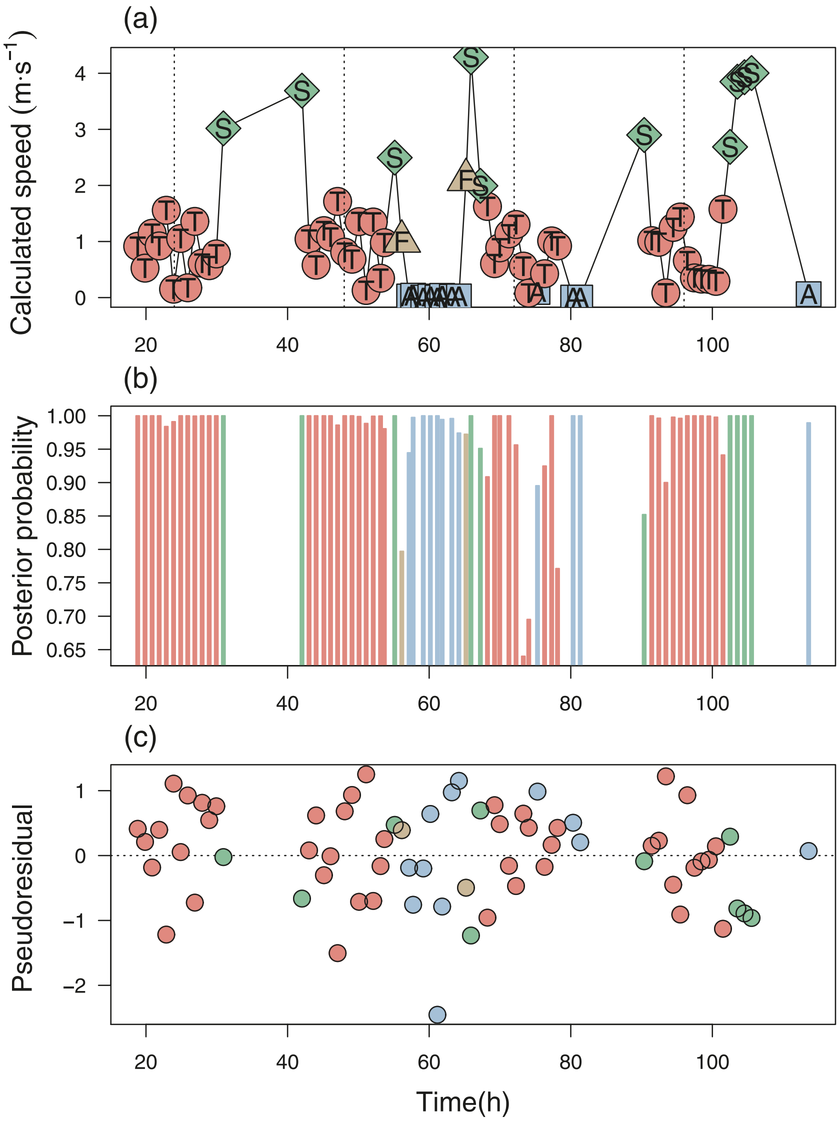
Results

The HMM was applied to the full fleet for the 3 years of data, with each vessel modelled independently by its own HMM. As expected, the model successfully determines the high speed polls to generally be steaming, the near stationary polls as anchorage, and the polls in the band between the two generally as trawling (for example, Fig. 6*a*). Also, the HMM has reasonably indicated several polls as false trawling (i.e., entry–exit from anchorage). The magnitude of the corresponding posterior probabilities (Fig. 6*b*) can also give an indication of the certainty of the allocation. For model diagnostics and outlier detection, pseudoresiduals (Patterson et al. 2009; Zucchini and MacDonald 2009) can also be examined (see Fig. 6*c*).

For some vessels there is evidence of a difference in trawl behaviour depending on targeted species (i.e., fishery), as can be seen in the histograms of calculated speed (Fig. 7). In the scallop fishery, generally there was a much greater number of lower trawling speeds because of shorter, less straight trawling behaviour (Fig. 7*a*). Whereas, for the eastern king prawn fishery there was a much more distinct mode corresponding to trawling because of longer, straight trawling behaviour (Fig. 7*b*). Interestingly, including a fishery component did not substantially change the final result for many of vessels examined. This is generally due to the vessels not changing between drastically different fisheries (e.g., the two extremes shown in Fig. 3).

The probabilistic allocation to the trawling state can be used to produce spatial and temporal maps of fishing effort for the whole fleet. At the simplest level, the fishery is divided into grid cells, and the total time trawled in each grid is calculated and plotted (Fig. 8). If spatial effort is examined

Fig. 6. Examination of a typical fit to a section of time for a single vessel. The polls have been allocated in an outright classification to states for clarity, based on maximum posterior probability (*a*). The pink circles with T denote trawling, green diamonds with S correspond to steaming, blue squares with A show the anchored state, and brown triangles with F indicate the false trawling. The *x* axis corresponds to time and the *y* axis shows the calculated speed (m·s^{-1}). The vertical dotted lines denote midnight each day. The corresponding maximum posterior probability are shown (*b*) to indicate the uncertainty of the allocations. The corresponding pseudoresiduals can be used for model diagnostics or outlier detection (*c*).

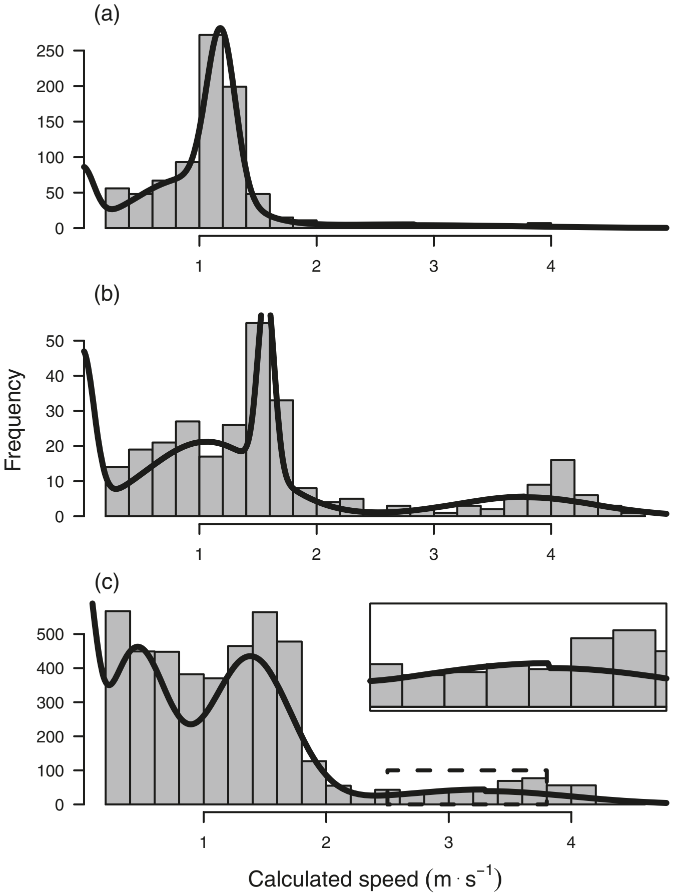


over long time periods, some form of standardization of effort would possibly be required because of changes in fishing technology and practices (O'Neill et al. 2003; Bishop et al. 2004; Maunder and Punt 2004). During this time period of the data in this paper, the Queensland fisheries logbook system recorded position at the degree level (and more recently to one tenth of a degree). Hence, the use of VMS data for effort mapping provides a substantial increase in spatial resolution over fishermen logbooks alone. It should also be noted that the VMS data provide a substantial increase in temporal resolution. Logbook entries correspond to a single point representing a full 24 h period of fishing, whereas the VMS system provides location information at a much higher frequency (for confidentiality we cannot disclose the actual polling frequency).

We can also plot the outright classifications of tracks on maps (Fig. 9). Closer examination of the allocations in this way does show some misclassified effort on the entrance and exit of an anchorage (Fig. 10*b*). However, the actual percentage of data misclassified around the anchorage is very small, especially when compared with a direct application of a speed rule (Fig. 10*a*).

The computation time to fit the model was quite reasonable; for example, a single vessel (10 548 polls over 3 years) took about 3 min to fit on a standard Intel i5 powered laptop (speed rule and mixture are both in the order of seconds). In practice, the fitting procedure should only have to be completed once and then the posterior probabilities stored to be used from then on to indicate vessel activity. As new data are collected, either a new fit to the complete time series

Fig. 7. Example of a fit for a single vessel of calculated speed corresponding to when the vessel is deemed to be in the scallop (*a*), tiger–endeavour prawn (*b*), and eastern king prawn (*c*) fisheries. The state distributions are denoted as a mixture and are overlaid on a histogram of calculated speed corresponding to the respective fishery. Polls with near zero speed have been removed from the histograms for clarity. It should be emphasized that the HMM used is not based on calculated speed alone, but also includes time of day and the underlying temporal correlated aspects of the data. Hence these plots of raw calculated speed and state distributions do not convey these important factors, but they are still useful to visualize the resulting fits. A larger section of plot corresponding to the dotted rectangle is inset in the final plot (*c*) to highlight the discontinuity due to uniform state–space distributions.



could be done, or if various conditions hold, the new data can be quickly assigned in a discriminant analysis context based on the existing HMM model parameters (i.e., simply implement a single E step).

Discussion

Hidden Markov models provide a flexible, automated method to determine vessel activity. The HMMs by their very nature model the inherent temporal correlation of the data. We have previously developed more complicated approaches combining several methods, or steps, plus ad hoc adjustments. However, the advantage of the HMM approach is that most of the issues with the data are addressed within a single, simple, statistical framework.

The flexibility of the HMM approach is a strength; however, this freedom also raises the question of model selection, for example, the choice of state distributions. The fact that the models are described within a statistical framework does provide some guidance through, for example, goodness-of-fit tests, criterion such as AIC–BIC, likelihood ratio hypothesis testing (in the case when models being considered are nested), and model diagnostics (such as pseudoresiduals). With regard to the assumed state distributions in our example, there is possibly a better choice for the entry–exit state distribution than the uniform distributions. The entry–exit state corresponds to a vessel that spends a proportion of the polling period stationary and the remaining portion at steaming speed. Hence, possibly a distribution consisting of the sum of randomly weighted steaming and anchorage speed distributions would be more appropriate. The effect of the blunt use of a uniform distribution can sometimes be seen by the non-smooth nature of the overall distribution at the upper limit of the entry–exit uniform.

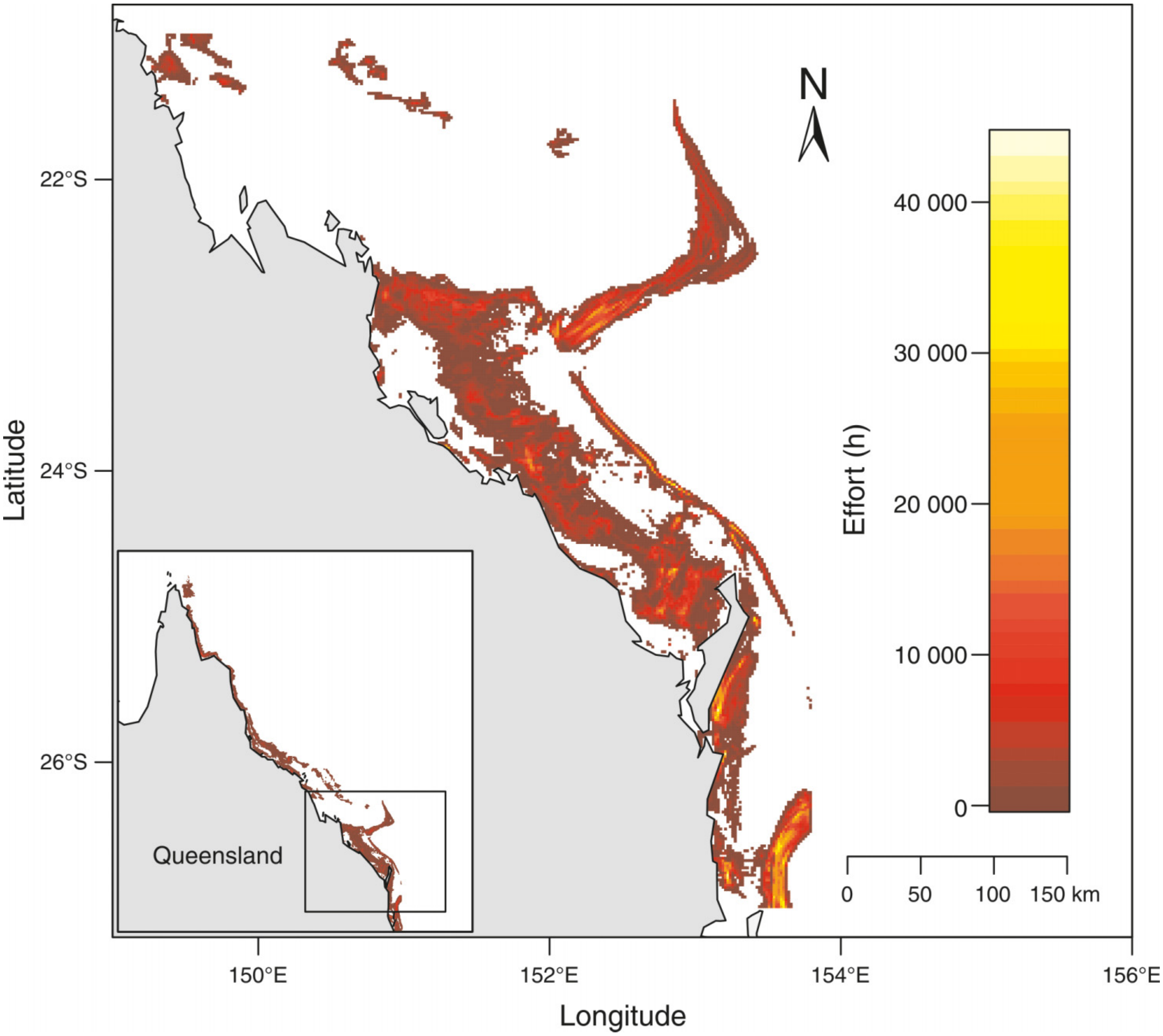
Since the fitting of the HMM is reasonably automated, as seen by our example, it is feasible to use the approach on fleets consisting of a large number of vessels. However, compared with using a simpler approach, more care must be taken with the fitting procedure. In particular, it is important to avoid so-called spurious model fits, where the model fit converges to fit a random colinearity in the data (i.e., a mixture component fits small linear clumps randomly found in the data; see McLachlan and Peel (2000) for further details). To avoid spurious solutions, the use of a range of different starting values may be required. Alternatively, if spurious solutions are occurring, then constraining the covariance matrices of the normal mixture to be equal should avoid the issue. Therefore, it would be always be advisable to assess the individual vessel results to confirm that reasonable fits are being obtained.

We presented a real world example of the use of a HMM to demonstrate the practicality of the method. Specifically, the model was able to efficiently handle a large amount of VMS data (in the order of minutes per vessel) and provide sensible results. One possible drawback is that the HMM approach is much more complex to implement than a simple cutoff. Compared with the finite normal mixture approach, there is a small increase in complexity, with the added forward–backward steps required within the Baum–Welch algorithm and the computation of transition rates.

In our example, the need for a fishery component in the model, generally, did not seem to be as important as first thought. This was because the vessels seemed to generally stay in similar fisheries. This meant that much of the fishery heterogeneity in the data was modelled adequately by each vessel having its own independent HMM.

The use of VMS data provides a large increase in spatial precision over traditional logbook data. However, care should be taken not to overestimate the spatial accuracy of the resulting effort distributions. Since the VMS position is accurate (say, for example, to 50 m), it is tempting to examine the data at an extremely fine spatial scale. However, the spatial uncertainty of vessel positions between polls needs to also be considered. To better portray this uncertainty, a natural progression from the assumption that vessels travel in a straight line is to develop more realistic models (Horne et al.

Fig. 8. An example of mapping fishing effort intensity (hours) for a section of the Queensland fishery between January 2000 and March 2003 (for all fisheries). White indicates no effort, dark colours denote low intensity, and yellow shows high intensity (areas where a low numbers of vessels are active have been deleted for privacy). Gray colouring indicates land.



2007; Hintzen et al. 2010). In our example, we have high-frequency GPS logger data, that could be used to produce empirical distributions of models of vessel location between polls depending on spatial distance between polls.

Further work is required to develop greater ground truthing and validation. Collection of true activity for a sample of data would be one obvious way to accomplish this. Without further data on true vessel activity, it is impossible to truly validate the result but it would seem that the classifications are generally reasonable, with distinct steaming and trawling spatial locations visible. Furthermore, pseudoresiduals on the whole seem to be well behaved.

A simulation experiment compared the HMM with other approaches ranging from a simple speed cutoff rule to a mixture model approach. As expected, when temporally correlated data were analyzed, the HMM considerably outperformed a simple speed cutoff approach. Interestingly, with a postcorrection to force non-trawling on to entry–exit polls, the speed cut-off rule improved drastically. This improvement may be slightly artificial, since in the simulation all polls entering–exiting anchorage were simulated as false trawling, so obviously the ad hoc fix works extremely well and reduces the rules error. In reality, the proportion of such cases may be different and less clear cut. We could also argue that even if the modified non-HMM approach could be further refined, so that for a given situation it produced similar results to the HMM, the method would be less defendable and less robust to data changes.

The HMM approach lends itself easily to a more complex model if required. For example, the model could be adapted to estimate targeting as well as vessel activity. In the example presented in this paper, it was assumed that the most caught species and the spatial location on any given trawl indicated the target species. This is true to a large extent, although not perfect, and obviously the targeting behaviour of fishermen will be temporally correlated. The HMM method could also

Fig. 9. Plots of selected regions of VMS data for the scallop (*a*) and the eastern king prawn (*b*) fisheries over a single year. The HMM was fitted and transects with trawl probability < 0.5 are denoted by grey lines; trawl probability ≥ 0.5 are denoted by black lines.

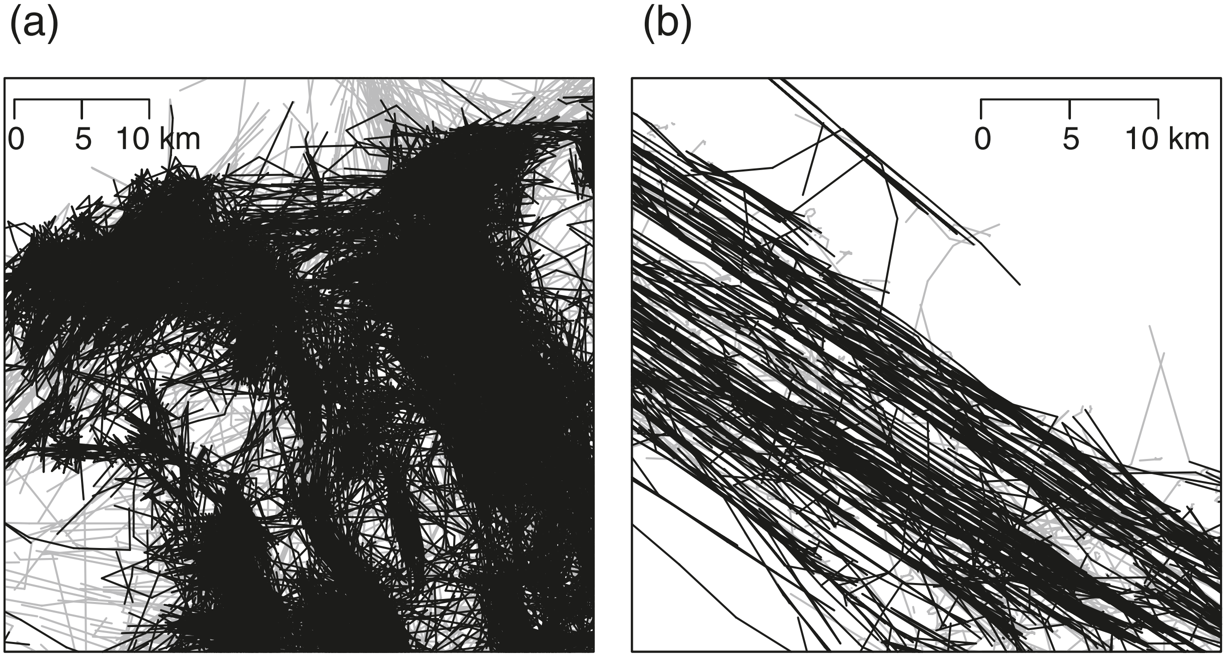
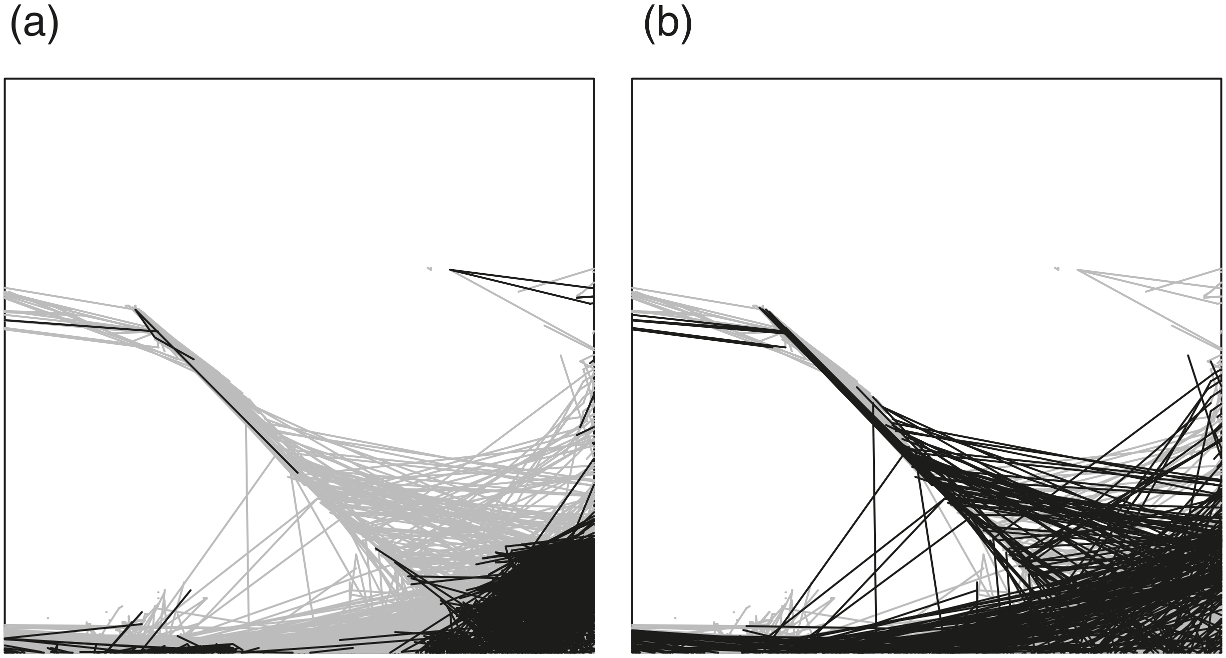


Fig. 10. Plot demonstrating some misclassification related to navigation into a port. Tracks of VMS data from a particular region: (*a*) trawling determined via a HMM and (*b*) trawling determined via speed cutoff rule (based on a mixture model). Transects with trawl probability < 0.5 are shown in grey, and trawl probability ≥ 0.5 are denoted by black lines.



allow additional covariates (e.g., net type or size) to be included that help predict targeting behaviour and incorporate the temporal correlation. Assuming that the change over of fishing gear required to switch between fisheries only occurs when a vessel is anchored, then the model can be extended so that each fishery has its own corresponding branch of the Markov model, with the interchange node being a common anchorage node. Then the caught species could be included as an observation dependant on node–state. However, a reasonable amount of data would be required to allow the model to be extended in such a manner.

Another reasonably straightforward extension to the model would be to include other information or data that help predict vessel activity (e.g., weather, habitat–bottom type, or vessel direction) into the HMM model. This could be incorporated as a covariate on any of the parameters, as we saw with our inclusion of time of day in our example. Furthermore, in some cases the inclusion of covariates could actually be the motivation in itself, where we are trying to elicit relationships between the activity and the covariate (e.g., relating trawling to habitat type). Alternatively, extra data could be incorporated as another state-observed variable, for example, direction could be included by making the state distribution bivariate and, assuming speed and direction are independent, with the second variable corresponding to direction (i.e., modelled as an Von Mises distribution). Including this extra information may increase the acuracy of the method.

In some VMS implementations, polling frequency is not fixed and may vary. For example, when a vessel nears a closure or protected area, some systems increase polling frequency. In this case the model would have to be modified to accommodate different time steps; one approach could be to reframe the model as a continuous time Markov model.

In summary, we have found that using a HMM approach successfully addresses many of the issues in VMS data, such as temporal correlation and misclassification of entry–exit to anchorage as trawling. The HMM was shown to work within a practical implementation and generally was validated via simulation experiment. Overall HMMs provide a powerful, eloquent tool to extract vessel activity information from VMS data.

Acknowledgements

We thank Simon Hoyle and Neil Gribble for initiating the project. The project was funded by the Australian Fisheries Research and Development Corporation (FRDC). Thanks also go to Toby Patterson for the seminar that inspired the use of HMM in this work, Mark Bravington for his insightful statistical suggestions, and Tony Courtney for his advice on trawl fisheries and enthusiasm for the results VMS can provide. Thanks to the large amount of VMS data provided by Cameron Baker and the Queensland DPI&F VMS unit. Also thanks to the fishermen who provided GPS logger data and the anonymous referees for their valuable comments.

References

Baum, L.E., Petrie, T., Soules, G., and Weiss, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. **41**(1): 164–171. doi:10.1214/aoms/1177697196.

Bertrand, S., Burgos, J.M., Gerlotto, F., and Atiquipa, J. 2005. Lévy trajectories of Peruvian purse-seiners as an indicator of the spatial distribution of anchovy (*Engraulis ringens*). ICES J. Mar. Sci. **62**(3): 477–482. doi:10.1016/j.icesjms.2004.12.002.

Bishop, J., Venables, W.N., and Wang, Y.G. 2004. Analysing commercial catch and effort data from a Penaeid trawl fishery: a comparison of linear models, mixed models, and generalised estimating equations approaches. Fish. Res. **70**(2–3): 179–193. doi:10.1016/j.fishres.2004.08.003.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B (Method.), **39**: 1–38.

Deng, R., Dichmont, C., Milton, D., Haywood, M., Vance, D., Hall, N., and Die, D. 2005. Can vessel monitoring system data also be used to study trawling intensity and population depletion? The example of Australia's northern prawn fishery. Can. J. Fish. Aquat. Sci. **62**(3): 611–622. doi:10.1139/f04-219.

Good, N., and Peel, D. 2007. An index of abundance for prawn stocks in Queensland using maximum entropy methods and VMS data. *In* GIS/spatial analyses in fishery and aquatic sciences. Vol. 3. *Edited by* T. Nishida, P.J. Kailola, and A. Caton. Fishery–Aquatic GIS Research Group, Saitama, Japan. pp. 161–180.

Harrington, J.J., Semmens, J.M., and Haddon, M. 2007. Spatial distribution of commercial dredge fishing effort: application to survey design and the spatial management of a patchily distributed benthic bivalve species. Mar. Freshw. Res. **58**(8): 756–764. doi:10.1071/MF06101.

Hintzen, N.T., Piet, G.J., and Brunel, T. 2010. Improved estimation of trawling tracks using cubic Hermite spline interpolation of position registration data. Fish. Res. **101**(1–2): 108–115. doi:10.1016/j.fishres.2009.09.014.

Horne, J.S., Garton, E.O., Krone, S.M., and Lewis, J.S. 2007. Analyzing animal movements using Brownian bridges. Ecology, **88**(9): 2354–2363. doi:10.1890/06-0957.1. PMID:17918412.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. Continuous univariate distributions, Vol. 1. John Wiley and Sons, New York.

Larcombe, J.W.P., McLoughlin, K.J., and Tilzey, D.J. 2001. Trawl operations in the South East Fishery, Australia: spatial distribution and intensity. Mar. Freshw. Res. **52**(4): 419–430. doi:10.1071/MF99169.

MacDonald, I.L., and Zucchini, W. 1997. Hidden Markov and other models for discrete-valued time series. Chapman & Hall, London.

Marrs, S.J., Tuck, I.D., Atkinson, R.J.A., Stevenson, T.D.I., and Hall, C. 2002. Position data loggers and logbooks as tools in fisheries research: results of a pilot study and some recommendations. Fish. Res. **58**(1): 109–117. doi:10.1016/S0165-7836(01)00362-9.

Maunder, M.N., and Punt, A.E. 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. **70**(2-3): 141–159. doi:10.1016/j.fishres.2004.08.002.

McLachlan, G.J., and Peel, D. 2000. Finite mixture models. John Wiley and Sons, New York.

Mejias, A., Jr. 1999. Vessel monitoring system sensor applications in the Gulf of Mexico shrimp fishery. *In* Proceedings of the International Conference on Integrated Fisheries Monitoring, 1–5 February 1999, Sydney. *Edited by* C.P. Nolan. FAO, Rome. pp. 291–303.

Mills, C.M., Townsend, S.E., Jennings, S., Eastwood, P.D., and Houghton, C.A. 2007. Estimating high resolution trawl fishing effort from satellite-based vessel monitoring system data. ICES J. Mar. Sci. **64**(2): 248–255. doi:10.1093/icesjms/fsl026.

O'Neill, M.F., Courtney, A.J., Turnbull, C.T., Good, N.M., Yeomans, K.M., Staunton-Smith, J., and Shootingstar, C. 2003. Comparison of relative fishing power between different sectors of the Queensland trawl fishery, Australia. Fish. Res. **65**(1–3): 309–321. doi:10.1016/j.fishres.2003.09.022.

Patterson, T.A., Basson, M., Bravington, M.V., and Gunn, J.S. 2009. Classifying movement behaviour in relation to environmental conditions using hidden Markov models. J. Anim. Ecol. **78**(6): 1113–1123. doi:10.1111/j.1365-2656.2009.01583.x. PMID: 19563470.

Peel, D., and Good, N. 2007. Refined trawl signature and trawl track definitions using a hidden Markov model approach. *In* Innovative stock assessment and effort mapping using VMS and electronic logbooks. *Edited by* N.A. Gribble, N. Good, D. Peel, M. Tanimoto, and R. Officer. Department of Primary Industries: Fisheries Research and Development Corporation, Brisbane, Australia. pp. 55–77.

Peel, D., Good, N., and Tanimoto, M. 2007. Mapping the spatial intensity of fishing effort using a speed filter method. *In* Innovative stock assessment and effort mapping using VMS and electronic logbooks. *Edited by* N.A. Gribble, N. Good, D. Peel, M. Tanimoto, and R. Officer. Department of Primary Industries: Fisheries Research and Development Corporation, Brisbane, Australia, pp. 20–54.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, **77**(2): 257–286. doi:10.1109/5.18626.

Rijnsdorp, A.D., Buys, A.M., Storbeck, F., and Visser, E.G. 1998. Micro-scale distribution of beam trawl effort in the southern North Sea between 1993 and 1996 in relation to the trawling frequency of the sea bed and the impact on benthic organisms. ICES J. Mar. Sci. **55**(3): 403–419. doi:10.1006/jmsc.1997.0326.

Vignaux, M., Vignaux, G.A., Lizamore, S., and Gresham, D. 1998. Fine-scale mapping of fish distribution from commercial catch and effort data using maximum entropy tomography. Can. J. Fish. Aquat. Sci. **55**(5): 1220–1227. doi:10.1139/f97-297.

Zucchini, W., and MacDonald, I.L. 2009. Hidden Markov models for time series: an introduction using R. Chapman & Hall, New York.

Appendix A

This section contains a more detailed description of the fitting process using the Baum–Welch algorithm (Baum et al. 1970). Let s denote the number of states in the HMM, n the number of time intervals (polls). Then given the observed vessel speeds $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, time of day $\boldsymbol{d} = \{d_1, \ldots, d_n\}$, and fishery $\boldsymbol{c} = \{c_1, \ldots, c_n\}$, the algorithm on the $(k + 1)$th iteration is given by

E step

Forward recursion The values $a_{i,t}^{(k)}$ ($i = 1, \ldots, s$; $t = 1, \ldots, n - 1$) corresponding to the forward probabilities (i.e., the probability of being in a state at time t, given the previous $1, \ldots, t - 1$ observations) are calculated as

$$\textbf{Initialization} \quad a_{i,1}^{(k+1)} = P_{0,i}^{(k+1)}(d_1, c_1)f_i^{(k+1)}(y_1, \theta_{i,c_1})$$

Induction

$$a_{i,t+1}^{(k+1)} = \left[\sum_{h=1}^{s} a_{h,t}^{(k+1)} P_{h,i}^{(k+1)}(d_t, c_t)\right] f_i^{(k+1)}(y_{t+1}, \theta_{i,c_{t+1}})$$

where $P_{0,i}(d_1, c_1)$ are the initial probabilities of being in each state.

Backward recursion The values $b_{i,t}^{(k+1)}$ ($i = 1, 2, \ldots, s$; $t = n - 1, n - 2, \ldots, 1$) corresponding to the backward probabilities (i.e., the probability of being in a state at time t, given the future $t + 1, \ldots, n$ observations) are calculated as

$$\textbf{Initialization} \quad b_{i,n}^{(k+1)} = 1$$

Induction

$$b_{i,t}^{(k+1)} = \sum_{h=1}^{s} P_{h,i}^{(k+1)}(d_t, c_t) f_i^{(k+1)}(y_{t+1}, \theta_{i,c_{t+1}}) b_{h,t+1}^{(k+1)}$$

Also within the E step we estimate the posterior probabilities $w_{i,j,t}^{(k)}$ of moving from node i to node j at time point t

$$w_{i,j,t}^{(k)} = \frac{a_{i,t}^{(k+1)} P_{h,i}^{(k+1)}(d_t, c_t) f_i^{(k+1)}(y_{t+1}, \theta_{i,c_{t+1}}) b_{j,t+1}^{(k+1)}}{\sum_{h=1}^{s}\sum_{i=1}^{s} a_{h,j}^{(k+1)} P_{h,i}^{(k+1)}(d_t, c_t) f_i^{(k+1)}(y_{t+1}, \theta_{i,c_{t+1}}) b_{i,j+1}^{(k+1)}}$$

Since the trawling state distribution is a mixture model, we must also estimate the posterior probabilities of each of the K normal mixture components, i.e., for observation at time $t = 1, \ldots, n$, and $m = 1, \ldots, K$

$$\tau_{t,m}^{(k+1)} = \frac{\pi_{m,c_t}^{(k)} N(y_t; \mu_{2,m,c_t}, \sigma_{2,m,c_t})}{\sum_{h=1}^{K} \pi_{h,c_t}^{(k)} N(y_t; \mu_{2,h,c_t}, \sigma_{2,h,c_t})}$$

M step

For example, for the state distributions and corresponding Ω, let $w_{\cdot,j,t}^{(k+1)}$ denote the probability of the vessel being in state j at time t, i.e., $w_{\cdot,j,t}^{(k+1)} = \sum_{i=1}^{s} w_{i,j,t}^{(k+1)}$ and ψ be the set of all times that the vessel was allocated to fishery C

$$\mu_{5,C}^{(k+1)} = \frac{\sum_{t\in\psi} w_{\cdot,j,t}^{(k+1)} y_t}{\sum_{t\in\psi} w_{\cdot,j,t}^{(k+1)}}$$

$$\sigma_{5,C}^{(k+1)} = \frac{\sum_{t\in\psi} w_{\cdot,j,t}^{(k+1)} (y_t - \mu_{5,C}^{(k+1)})^2}{\sum_{t\in\psi} w_{\cdot,j,t}^{(k+1)}}$$

where also since the trawling component is a mixture distribution we also have the K component mixture parameters to estimate

$$\pi_{m,C}^{(k+1)} = \sum_{t\in\psi} \tau_{t,m}^{(k+1)}$$

$$\mu_{2,m,C}^{(k+1)} = \frac{\sum_{t\in\psi} \tau_{t,k}^{(k+1)} w_{\cdot,j,t}^{(k+1)} y_t}{\sum_{t\in\psi} \tau_{t,m}^{(k+1)} w_{\cdot,j,t}^{(k+1)}}$$

$$\sigma_{2,m,C}^{(k+1)} = \frac{\sum_{t\in\psi} \tau_{t,k}^{(k+1)} w_{\cdot,j,t}^{(k+1)} (y_t - \mu_{2,m,C}^{(k+1)})^2}{\sum_{t\in\psi} \tau_{t,m}^{(k+1)} w_{\cdot,j,t}^{(k+1)}}$$

Reference

Baum, L.E., Petrie, T., Soules, G., and Weiss, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. **41**(1): 164–171. doi:10.1214/aoms/1177697196.