Deep-Sea Research II 88–89 (2013) 34–46

Contents lists available at SciVerse ScienceDirect

Deep-Sea Research II

journal homepage: www.elsevier.com/locate/dsr2

State-space models for bio-loggers: A methodological road map

I.D. Jonsen [a],*, M. Basson [b], S. Bestley [a,e], M.V. Bravington [c], T.A. Patterson [b], M.W. Pedersen [d], R. Thomson [b], U.H. Thygesen [d], S.J. Wotherspoon [e,f]

[a] *Ocean Tracking Network, Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada*
[b] *CSIRO Marine and Atmospheric Research, Wealth from Ocean Flagship, Castray Esplanade, Hobart, Tasmania, Australia*
[c] *CSIRO Mathematics, Informatics and Statistics, Castray Esplanade, Hobart, Tasmania, Australia*
[d] *National Institute of Aquatic Resources, Technical University of Denmark, Charlottenlund, Denmark*
[e] *Institute for Marine and Antarctic Studies, University of Tasmania, Sandy Bay, Hobart, Tasmania, Australia*
[f] *Australian Antarctic Division, Kingston, Tasmania, Australia*

ARTICLE INFO

Available online 17 July 2012

Keywords:
Animal movement
Bayesian statistics
Foraging behaviour
Frequentist statistics
Hidden Markov model
Migration
Telemetry
Time series analysis

ABSTRACT

Ecologists have an unprecedented array of bio-logging technologies available to conduct *in situ* studies of horizontal and vertical movement patterns of marine animals. These tracking data provide key information about foraging, migratory, and other behaviours that can be linked with bio-physical datasets to understand physiological and ecological influences on habitat selection. In most cases, however, the behavioural context is not directly observable and therefore, must be inferred. Animal movement data are complex in structure, entailing a need for stochastic analysis methods. The recent development of state-space modelling approaches for animal movement data provides statistical rigor for inferring hidden behavioural states, relating these states to bio-physical data, and ultimately for predicting the potential impacts of climate change. Despite the widespread utility, and current popularity, of state-space models for analysis of animal tracking data, these tools are not simple and require considerable care in their use. Here we develop a methodological "road map" for ecologists by reviewing currently available state-space implementations. We discuss appropriate use of state-space methods for location and/or behavioural state estimation from different tracking data types. Finally, we outline key areas where the methodology is advancing, and where it needs further development.

1. Introduction

Within marine ecology, a very general set of ecological questions has emerged that researchers address using bio-logging techniques. For example, where and when do animals move? What are the physiological costs of these movements? How do these movements relate to environmental variability and the ever changing distributions of heterogeneous resources? What are the implications of individual movements at the population level? How might animal movements and population distribution be affected by future environmental change in the ocean? Answering these important ecological questions becomes a challenging statistical problem when using most bio-logging data.

Telemetry-based studies of animal movement ecology, physiology, and environmental interactions generally rely on a compartmentalized approach to analysis of tracking data. This approach typically has three stages: (1) error correction; (2) calculation of summary movement metrics from corrected tracks; and (3) biological inference through statistical or non-statistical analysis (Patterson et al., 2008). There are three key drawbacks to this approach. First, implicit assumptions, which may or may not be valid, regarding an animal's movements are required to remove spurious location errors (e.g., Austin et al., 2003). Second, measurement error effects are not separated from the underlying movement processes that are of interest, which can bias analyses (Bradshaw et al., 2007). Third, the analysis tools associated with this approach tend to be correlative, comparative (using simple hypothesis tests), or pattern-based; all of which are limited in their scope for direct examination of the ecological and physiological mechanisms that underpin animal movement. Adopting a stochastic, model-based approach that allows mechanistic models of the movement process to be fit directly to telemetry data, while accounting for measurement error when appropriate, is more rigorous and powerful but also requires more complicated analysis tools.

Recent reviews (Patterson et al., 2008; Schick et al., 2008) have publicized the notion of state-space model (SSM[1]) approaches for studying animal movement and these tools are now being more widely applied. We see signs within the bio-logging community that SSM approaches are making their way into analyses of many

---

* Corresponding author. Tel.: +1 902 494 3910.
E-mail address: ijonsen@dal.ca (I.D. Jonsen).

[1] See the Glossary for a definition of this and other technical terms.

0967-0645/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.dsr2.2012.07.008

kinds of telemetry data. A total of 21 oral and poster presentations at the Fourth International Symposium on Bio-logging (March 2011), held in Hobart, Tasmania, Australia, included some form of SSM analysis of telemetry data or further developed their implementation.[2] In many cases, SSMs are used to filter error-prone Argos (2011) or light-based (Hill, 1994) location data and/or to estimate unobserved behavioural states of animals. Despite these promising signs, we feel there is a gap to be bridged in the general understanding of SSMs, both in the technical aspects of their application and in the interpretation of their results. It is, therefore, timely to build a methodological road map for bio-loggers outlining currently available methods and applications appropriate to particular problems in the study of animal movement and behaviour.

Here we explain SSMs in the context of marine animal tracking data by providing the necessary background, technical details, and examples for field ecologists to appreciate the flexibility and power of these statistical tools. We outline general considerations for fitting SSMs, for querying their fit to data, and for model selection. We highlight several previously published SSM approaches for tracking data and the situations in which one may be preferred over others. Finally, we suggest areas where the methods are advancing and where further work is required.

2. General explanation of state-space methods

State-space models encompass a range of time series methods that estimate the state of an unobservable process from an observed data set. The earliest example of state-space methods used for estimation purposes is the celebrated Kalman filter (KF) (Kalman, 1960, see also Section 5) which is now used in applications from aerospace to finance, as well as for geolocation of animals from tagging data (Sibert et al., 2003). The state-space paradigm is not limited to time series analysis: it applies also to pure analysis problems, one example from marine ecology being larval transport and growth (e.g., Christensen et al., 2008), and to dynamic optimization problems, e.g., arising in behavioural ecology (Houston and McNamara, 1999). In ecology, state-space methods are used to model single individuals, population dynamics (Brinch et al., 2011; de Valpine and Hastings, 2002), and marine ecosystem dynamics (Evensen, 2003).

The notion of the state is pivotal in SSMs, as the name suggests. In a bio-logging context, the most common state variables used (to date) are an animal's position (Anderson-Sprecher and Ledolter, 1991; Jonsen et al., 2003; Royer et al., 2005; Thygesen et al., 2009) and it's behaviour, the latter is usually represented as a discrete variable with two or more nominal categories such as "foraging" and "not foraging" or "migratory" and "resident" (Jonsen et al., 2005; Morales et al., 2004; Patterson et al., 2009; Pedersen et al., 2011b). Mathematically, a collection of variables constitute the state of a dynamical system, if they summarize the previous history of the system, so that predictions about the future can be made solely from the current state. Choosing the right number of state variables to describe a real system is a delicate balance between realism and feasibility, and choosing the most important state variables requires insight into not just the biology of the animal, but also the data quality and the nature of the statistical estimation problem.

The key ingredient in a stochastic SSM is the process model, an equation that describes how the state evolves randomly in time. A simple example of a process model is a random walk, which verbally can be formulated as follows: if the position of an animal at time t is known to be $\mathbf{x}_t$, then the position one day later is Gaussian with mean $\mathbf{x}_t$ and variance $\mathbf{V}$. Here $\mathbf{V}$ is a parameter, the variance of the daily move distance. A mathematical formulation of this is

$$\mathbf{x}_{t+1|t} \sim \mathrm{N}(\mathbf{x}_t,\mathbf{V}) \tag{2.1}$$

Note that the process model is written in terms of conditional probability distributions. If the state $\mathbf{x}_t$ at time t is known, how is the later state $\mathbf{x}_{t+1}$ distributed? Mathematically, the notion of a state is formalized by the Markov property: given the current value of the state, future state variables are statistically independent of past state variables. These conditional distributions, which are known as transition probabilities, are sufficient to describe all state dynamics. Note also that we distinguish between states and parameters: states like $\mathbf{x}_t$ evolve in time and describe the immediate situation, while parameters like $\mathbf{V}$ are typically constant in time and describe the underlying properties of the animal or mechanism of the system.

In this example, the state variable is continuous, i.e. may take any real value. In other cases, the state may be purely discrete, i.e. only a finite (or countable) number of different states are possible. In yet other examples, the state is composed of both continuous and discrete state variables. Similarly, the time variable may be continuous or discrete. Even if, for example, the position of an animal is defined at all times, we may choose to have the model represent only the daily position.

The process model is written entirely without reference to available data. Of course, choosing a suitable process model requires thought about available data, but the state of the system evolves regardless of how or if we observe the system.

For estimation purposes, the process model is complemented by one or more equations, the observation model, which describe the link between each observed data point and the state of the system at the time of observation, or at regular time intervals within which observations may or may not occur. The observation model describes what happens at the time of observation, so does not make any reference to the dynamics in the underlying process model. This structure is depicted in Fig. 1. Also these equations specify conditional probability distributions: if the state at time t were known to be $\mathbf{x}_t$, how is the measurement $\mathbf{y}_t$ distributed? It is important to appreciate the generality of this framework: the measurement may be a state variable subject to measurement error, for example, a position estimated with the Argos satellite system, but the measurement may also be any other quantity which holds some information about the state, for example, an animal's travel speed or sea surface temperature (SST). The observation model may specify several different types of measurements which may be taken simultaneously or at different points of time. Like the process model, the observation model

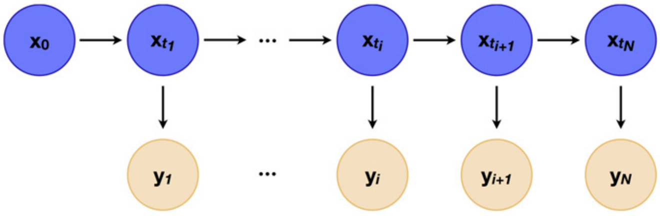

Fig. 1. Structure of a SSM for estimation. The state variable evolves randomly in time, starting at $\mathbf{x}_0$ and ending at $\mathbf{x}_{t_N}$. At times $t_1, t_2, \ldots, t_N$, measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are taken. In this graphical representation of the model, an arrow from one variable (say, $\mathbf{x}_{t_i}$) to another (say, $\mathbf{y}_i$) indicates that the model is written in terms of the conditional distribution of $\mathbf{y}_i$ given $\mathbf{x}_{t_i}$, and furthermore that conditional on $\mathbf{x}_{t_i}$, $\mathbf{y}_i$ is statistically independent of all other variables since they are not connected with arrows.

---

[2] http://www.cmar.csiro.au/biologging4/documents/AbstractsandProgram_final.pdf, last accessed on 25/05/2012.