Prague University of Economics and Business
Faculty of Informatics and Statistics

Automatic detection of life events in animal tracking data

MASTER THESIS

Study program:	Knowledge and Web Technologies
Specialization:	Quantitative Analysis

Author: Karel Douda
Supervisor: RNDr. Ing. Petr Máša, Ph.D.

Prague, December 2022

Contents

Introduction

Traditionally, the study of animal movements and life cycles has been a domain of great uncertainty, owing to varying levels of difficulty in observing the full range of animal behaviour in the wild. This has been especially difficult for ornithologists studying migratory birds. With the advent of small electronic circuitry, global telecommunications and positioning systems, it has become easier to acquire the required primary data for study using lightweight animal tracking devices, known as loggers or tags (Kays et al., 2015). These loggers collect various metrics, most importantly position at a certain time, from the animal carrying it.

During the infancy of the animal telemetry field, it was usually possible to analyse the received data manually due to its sparsity and precisely scheduled frequency, but required near-direct observation of the animal and had a great degree of inaccuracy (Burger et al., 2008). With the increasing sophisticatedness of GPS loggers, energy efficient solar panels, batteries and faster communication networks (Kays et al., 2015), another common problem has arisen, especially for larger studies. Modern animal movement loggers can collect hundreds, if not thousands, of positions per day at very fine temporal scales (Gupte et al., 2022), and send them to the end user multiple times per day, in any order. It has become increasingly difficult for field experts to make sense of their data manually, while comprehensive data analysis providing useful results is not a trivial task. Furthermore, with the increased data frequencies and volume, it has also been more difficult to detech and classify errorneous positions as such (Gupte et al., 2022).

Speedy detection of animal life events, such as mortality or nesting, is crucial for actionable instructions to animal conservation fieldwork experts. Speedy detection of the tracked animal's death is especially important for helping to establish the cause of the mortality event, since specific causes will be more difficult to establish with the passage of time.

The goal of this diploma thesis is to provide animal conservation experts, namely ornithologists, with reliable methods of filtering unreliable data, simplifying or clustering too complex datasets for interpretation, interpolating or subsetting data to satisfy model constraints, and finally, detecting important life events from these filtered and clustered data. The structure of this diploma thesis will follow the CRISP-DM methodology, and the resulting models will be released on GitHub and integrated into the Anitra platform for animal conservation experts. Users of this platform, notably members of the LIFE Eurokite project will be consulted for evaluation of these final models. Methods likely used for this project will include basic rule-based filtering, soft (fuzzy) clustering, hierarchical clustering, long-short term memory networks and possibly hidden Markov models. Various options will be evaluated based on test and train datasets and all final outputs will be evaluated by the previously mentioned field experts.

1. Problem context

1.1 Animal telemetry

Animal telemetry is the study of animal lifecycles utilizing remotely sensing equipment. In practice, this involves attaching an electronic device capable of recording desired metrics, primarily location. Other metrics may include bodily temperature, accelerometers, magnetometers and many more. Frequencies of these measurement can be either fixed or varying, depending on the device or its programming. Usually, these metrics are stored on the device and either retrieved manually by recapturing the animal, or the device has a capability to send the data to its intended recipient, usually with an associated time delay.

Complexity of these telemetry devices varies greatly, from simple radio tags used to simply locate an animal, to complex GPS loggers with various sensors and the capability to quickly transmit data to a remote server for nearly immedieate display of the animal's location to the end user. However, it was not always as such, and as any field it has gone through its evolution.

For most of the history of modern zoology, animal tracking was on the sidelines of interest due to technological limitations of its time. Early radio-based tracking devices (tags) suffered from many issues, such as restrictive weight, complicated data acquisition requiring near-direct observation of the animal or lack of accuracy (Kays et al., 2015).

These simple radio tags were prevalent until the advent of the satellite-based Argos system in 1980 (Douglas et al., 2012), which finally allowed acquiring location data from animals at greater distances. Position data was extrapolated from the signal emitted by the Argos sattelites using a doppler-shift based method, which resulted in its main disadvantage, position inaccuracy, which has to be compensated for by sophisticated algorithms, which reduce data volume (Douglas et al., 2012). Early Argos tags were also unsuitable for small animals. Argos remains today as the only viable option for marine life or animals living in very remote areas, such as the Arctic.

Spread of the worldwide GSM network, along with further advances in electronics and microprocessors, more precise GPS positioning, lightweight batteries and compact solar panels brought on the reneissance of animal telemetry, allowing it to be used for medium-sized vertebreates (Kays et al., 2015). Wildlife conservation experts may now enjoy nearly instantaneous knowledge of the animal' location or life state, allowing them to, for example, quickly respond to possible mortality cases or to locate nests, animal aggregation places, wintering locations for migratory animals...

Today, animal tracking suffers from the opposite problem of having too much data to work with. Studies may deal with tens of millions of positions and their associated metrics, which makes human analysis of these very impractical. Due to this, various methods and analysis tools were developed or adopted for animal tracking. Some of these methods are elaborated upon in the following chapters. and sections.

1.2 Time series data

Time series is defined as *a set of observations arranged chronologically* (Parzen, 1961). Usually the field of time series analysis is concerned with consistently samped time series data with one variable, such as economic statistics collected annually, however there are methods in data preprocessing that are capable of normalizing the data or methods using the data directly.

Sequence analysis methods on multivariate time series are a possible approach for data analysis. Sequence analysis methods include Hidden Markov models, LSTM neural networks, cSPADE algorithm and more. All these methods listed work best with evenly spaced data in time series, so data interpolation and subsetting will have to be utilized.

Hidden Markov models have already been sucessfully applied to the animal telemetry analysis context, which is described in section 1.4.

1.3 Geospatial data

Geospatial data is a subtype of data where each data point is assigned a georeferenced position, usually with latitude and longitude. Sometimes, these data can be also enriched with time information, making the data sequential. For the purpose of this thesis, all geospatial data are associated with time, since animal tracks are sequential. Methods of multivariate time series analysis are thus suitable for animal tracking problems.

Substantial research has been done on various machine learning algorithms for analysis of spatial data. According to Kanevski et al., typical geospatial data problems include: spatial predictions and interpolation, modelling with uncertainities, multivariate joint predictions of several variables, risk mapping, modelling of spatial variability and uncertainty, optimisation of monitoring networks, space filtering models, machine learning, data mining in high dimensional geo-feature space. It should be noted that most of these methods listed in the previous enumeration are unsupervised, possibly yielding results with difficult interpretation.

Geospatial data, especially in the context of IoT, can be met with further challenges, such as measurement errors, varying data frequencies and data delays.

Methods used for geospatial data analysis and classification include kernel-based (soft) clustering, DBSCAN or Optics clustering. Geospatial data interpretations also provides hints for errorneous positions, such as uncharacteristicly shaped angles between two locations, impossible speeds, etc.

1.4 Previous research

In the context of animal tracking, there have been various efforts for processing data. These methods may involve identifying errors in the dataset, classifying animal behaviour, interpolating data.

Gupte et al. provided a guide (or more specifically, a methodology and software) for preprocessing high frequency animal telemetry data with an emphasis on lowering position errors. Fleming et al. used strong statistical background for error analysis, creating several recommended models and testing them on various telemetry devices. Douglas et al. provide methods for handling very imprecise Argos data and extracting the most likely positions out of them using various methods, including relatively simple rules.

Langrock et al. explain and use extensions of hidden Markov models in the context of animal telemetry processing of bisons, classifying data into stationary and moving states. Leos-Barajas et al. explore hierarchical hidden Markov models in the context of animal tracking, incorporating various HMMs at different time scales. Michelot et al. created the R package moveHMM for the wider animal telemetry community. Wijeyakulasuriya et al. propose and test various methods (including random forests and long-short term memory neural networks) for prediction of animal positions. Jeantet et al. use convolutional neural networks for 1D classification of behaviour of juvenile turtles.

2. Current progress

As of the time of the writing of this document, error detection based on simplistic rules is completely finished. Several outlines for the text part of the diploma thesis are also done, some are present in this document. Research into possible methods is also underway, with emphasis on HMMs and neural networks. Evaluating effectiveness of HMMs and CNN/LSTM NN for behaviour classification will be underway soon. A module for labeling sections of data for training and validation was added to the Anitra platform, and now a phase of data collecting is also underway. As the author is not an expert in animal movement and would introduce his layman biases to the data and thus bring down the possible end quality of the data, the author does not help in this phase.

3. Methods

This diploma thesis will utilize design research methods, as the final output will be an evaluated artifact usable on animal tracking data. The practical part of the thesis will follow the CRISP-DM methodology, and the assesment and evaluation will use qualitative research methods and approaches.

Bibliography

BURGER, Alan E; SHAFFER, Scott A, 2008. Perspectives in ornithology application of tracking and data-logging technology in research and conservation of seabirds. *The Auk*. Vol. 125, no. 2, pp. 253–264.

DOUGLAS, David C; WEINZIERL, Rolf; C. DAVIDSON, Sarah; KAYS, Roland; WIKELSKI, Martin; BOHRER, Gil, 2012. Moderating Argos location errors in animal tracking data. *Methods in Ecology and Evolution*. Vol. 3, no. 6, pp. 999–1007.

FLEMING, CH; DRESCHER-LEHMAN, J; NOONAN, MJ; AKRE, TSB; BROWN, DJ; COCHRANE, MM; DEJID, N; DENICOLA, V; DEPERNO, CS; DUNLOP, JN, et al., 2021. A comprehensive framework for handling location error in animal tracking data. *bioRxiv*, pp. 2020–06.

GUPTE, Pratik Rajan; BEARDSWORTH, Christine E; SPIEGEL, Orr; LOURIE, Emmanuel; TOLEDO, Sivan; NATHAN, Ran; BIJLEVELD, Allert I, 2022. A guide to pre-processing high-throughput animal tracking data. *Journal of Animal Ecology*. Vol. 91, no. 2, pp. 287–307.

JEANTET, Lorène; VIGON, Vincent; GEIGER, Sébastien; CHEVALLIER, Damien, 2021. Fully Convolutional Neural Network: A solution to infer animal behaviours from multi-sensor data. *Ecological Modelling*. Vol. 450, p. 109555.

KANEVSKI, M; POZDNUKHOV, A; TIMONIN, V, 2008. Machine learning algorithms for geospatial data. Applications and software tools.

KAYS, Roland; CROFOOT, Margaret C; JETZ, Walter; WIKELSKI, Martin, 2015. Terrestrial animal tracking as an eye on life and planet. *Science*. Vol. 348, no. 6240, aaa2478.

LANGROCK, Roland; KING, Ruth; MATTHIOPOULOS, Jason; THOMAS, Len; FORTIN, Daniel; MORALES, Juan M, 2012. Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology*. Vol. 93, no. 11, pp. 2336–2342.

LEOS-BARAJAS, Vianey; GANGLOFF, Eric J; ADAM, Timo; LANGROCK, Roland; VAN BEEST, Floris M; NABE-NIELSEN, Jacob; MORALES, Juan M, 2017. Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics*. Vol. 22, no. 3, pp. 232–248.

MICHELOT, Theo; LANGROCK, Roland; PATTERSON, Toby A, 2016. moveHMM: an R package for the statistical modelling of animal movement data using hidden Markov models. *Methods in Ecology and Evolution*. Vol. 7, no. 11, pp. 1308–1315.

PARZEN, Emanuel, 1961. An approach to time series analysis. *The Annals of Mathematical Statistics*. Vol. 32, no. 4, pp. 951–989.

WIJEYAKULASURIYA, Dhanushi A; EISENHAUER, Elizabeth W; SHABY, Benjamin A; HANKS, Ephraim M, 2020. Machine learning for modeling animal movement. *PloS one*. Vol. 15, no. 7, e0235750.