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
Field of study: studijní obor

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

Prague, December 2022

Acknowledgements

-

Abstrakt

-

Klíčová slova

geospatial timeseries

JEL klasifikace

JEL1, JEL2, JEL3

Abstract

-

Keywords

geospatial timeseries, animal tracking

JEL classification

JEL1, JEL2, JEL3

Contents

List of Figures

List of Tables

List of abbreviations

BCC Blind Carbon Copy
CC Carbon Copy
CERT Computer Emergency Response Team
CSS Cascading Styleheets
DOI Digital Object Identifier
HTML Hypertext Markup Language
REST Representational State Transfer
SOAP Simple Object Access Protocol
URI Uniform Resource Identifier
URL Uniform Resource Locator
XML eXtended Markup Language

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. With the increasing sophisticatedness of GPS loggers, energy efficient solar panels, batteries and faster communication networks, another common problem has arisen, especially for larger studies. Modern animal movement loggers can hundreds, if not thousands, of positions per day, multiple times. 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.

Speedy detection of animal life events, such as mortality or nesting, is crucial for actionable instructions for animal conservation fieldwork experts. For example, speedy detection of animal death is very important for helping to establish the cause of the mortality event, since specific causes will be more difficult to establish later. Other life events

The goal of this diploma thesis is to provide animal conservation experts with reliable methods of filtering unreliable data, simplifying or clustering too complex datasets for interpretation and finally, detecting important life events from these filtered and clustered data. The structure of this diploma thesis roughly follows the CRISP-DM methodology, and the resulting models will be released on GitHub and integrated into the Anitra platform for animal conservation 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 Timeseries data

define timeseries, problems with existing methods for this problem (non-continuous sampling), possible solutions

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.

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.

2. Data understanding

In this chapter, data sourcing, description and availability will be provided.

2.0.1 Provenance

Data provenance is the description of origin of data. Data for this diploma thesis will come from two independent sources, but they share many common characteristics.

The first data source is Movebank, a project of the Max Planck institute in Germany, which collects geolocated animal tracking data from various device manufacturers and operators. Out of the box, the platform supports data visualisation, device and animal metadata management, project collaboration and data archival. The Movebank ecosystem has many community-built tools for data analysis, classification, visualisation and error detection. A drawback of the platform is the mandatory data publishing requirement, which may not be acceptable for some use cases. The Movebank platform's website can be found at `https://www.movebank.org`. The platform provides a public API, which allows data exports for analysis. The API will be one of the data sources used in the modeling and analysis. The Movebank data format is widely used and supported by various tools, and the naming conventions for metrics in this thesis will be taken from the Movebank data model.

The second data source is the Anitra platform, which shares many features of the Movebank ecosystem, but expands on the data management and does not require data sharing to the wider public. Anitra is also the first destination for the data from the Anitra devices and provides some additional metrics. The platform supports data exporting in the same format used by the Movebank platform. Data licensing for this platform will be elaborated upon later, but unlike Movebank, the data is not released under a permissive license. The author of this thesis has access to the full Anitra platform dataset, but only data with explicit permission of the owner will be used.

Satellite-collected datasets such as snow cover, cloud cover, satellite imagery, land use and elevation data can be used to enrich data and create more advanced ecological analyses. Local temperature, pressure and precipitation can be used to provide some other data to explain animal behaviour.

2.0.2 Data description

gnss datetime, latitude, longitude, temperature

2.0.3 Data enrichment

elevation, temperatures, pressure

3. Data preparation

Data preparation is the phase concerned with constructing the final dataset, including record and attribute selection, data cleaning, construction of new attributes and transformation of data for modeling tools (Wirth et al., 2000). In the context of this work, several data preparation tasks were performed.

3.1 Adding new attributes

For further analysis requirements, several new attributes had to be added. None of the attributes are complex and are relatively fast to add in one pass through of the data.

Inter-position distance, the distance between two points in metres, is calculated using the Geodesic formula. inter-position distance (Geodesic formula)

Time difference is calculated by subtracting the two timestamps of the measurement time.

Average speed is calculated as the inter-position distance divided by time difference.

Bearing angle is calculated using the (which?) formula.

3.2 Outlier detection

Outlier is an observation that appears to deviate markedly from other members of the sample in which it occurs (Grubbs, 1969). Outliers can be of many types, but in the context of this work, measurement errors are the most concering type, since not all methods are resilient towards them.

Measurement errors are inherent to the problem area; small IoT devices are heavily energy-constrainted and animal movements are very time sensitive, therefore there might not be enough time for the device to detect the error or take another data sample.. Global positioning systems usually come with several meters of inprecision from the actual position of the receiver, but these inprecisions may become exaggerated in areas with poor reception or radio-reflective surfaces.

For example, earlier GNSS-based tracking systems used roughly over 40% of their energy on getting GPS positions and had an average precision of 15 meters in open-air conditions (Jain et al., 2008)

The resulting deviations may be in tens of metres to several kilometers between two relatively close positions in cities or dense forests, for example.

There are many types of errors with varying levels of difficulty for automatic detection; while all of the errors are obvious to an expert, making an algorithm for all of them proved difficult.

Speed-based position error detection

In the context of animal movements, a valid upper bound for movement speed can be specified with reasonable level of certainness. Alternatively, according to Gupte et al., setting a dataset-dependent upper bound (such as the 90th percentile) for positions with incoming and outgoing extreme speeds can be recommended.

Angle-based position error detection

3.3 Attribute selection

data cleansing - normalisation, outlier detection (rules), interpolation, ordering, simplification?

Data standardization

(definition)

relativizace dat - pokud bude potřeba

zmínit potřebu fixních intervalů mezi časy pro HMM a podobné modely

3.4 Data labeling

Due to the intended results of the analysis process, precise and quickly interpretable result requirements, labeling of the data is a necessity. Therefore, a data labeling module was added to the Anitra animal tracking application for expert classification. Datasets will be enriched with data labels for each specific model. However, classic unsupervised learning algorithms such as k-means are not

3.5 Data splitting

For further sub-analysis, each dataset can be split into transfer and static states. Transfer state is assigned to the time the animal is moving between two specific locations, static state describes any other position. Transfer states can be further analysed to discover foraging or

nest-building activity, and static states can be further analysed to discover mortality states, nesting states or night roosts.

K-means activity classification

In previous research, Schwager et al. demonstrated that the well-known K-means unsupervised classification algorithm can be used to classify tracking data into various categories, in their study they demonstrated a capability of classifying data into two clusters based on activity (or lack thereof). Using speed, horizontal head angle and vertical head angle, they were able to produce animal-independent uniform clusters for determining animal activity and inactivity states. Unfortunately for more general applications, head angles may be impossible to determine. Loggers may be placed on other parts of the animal that may make it impossible to determine these angles (such as on the back or on the leg), or may rotated in various different ways. This may create inconsistent clustering not only between different animals, but between different loggers.

(try this with other attributes)

4. Modeling

created models

reinforcement learning?

4.0.1 Stationary position clustering

kernel density estimation, cluster statistics

4.0.2 Cluster classification

4.0.3 Mortality detection

General mortality detection

Device specific mortality detection

5. Evaluation

???

6. Deployment

???

Závěr

Závěr je povinnou částí bakalářské/diplomové práce. Obsahuje shrnutí práce a vyjadřuje se k míře splnění cíle, který byl v práci stanoven, případně shrnuje odpovědi na otázky, které byly položeny v úvodu práce.

Závěr k diplomové práci musí být propracovanější – podrobněji to je uvedeno v Náležitostech diplomové práce v rámci Intranetu pro studenty FIS.

Závěr je vnímán jako kapitola (chapter), která začíná na samostatné stránce a která má název Závěr. Název Závěr se nečísluje. Samotný text závěru je členěn do odstavců.

Bibliography

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

GRUBBS, Frank E., 1969. Procedures for Detecting Outlying Observations in Samples. *Technometrics*. Vol. 11, no. 1, pp. 1–21. Available from DOI: 10.1080/00401706.1969.10490657.

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

JAIN, Vishwas Raj; BAGREE, Ravi; KUMAR, Aman; RANJAN, Prabhat, 2008. wildCENSE: GPS based animal tracking system. In: *2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp. 617–622.

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.

SCHWAGER, Mac; ANDERSON, Dean M; BUTLER, Zack; RUS, Daniela, 2007. Robust classification of animal tracking data. *Computers and Electronics in Agriculture*. Vol. 56, no. 1, pp. 46–59.

WIRTH, Rudiger; HIPP, Jochen, 2000. CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1, pp. 29–40.

Appendices

A. Formulář v plném znění

B. Zdrojové kódy výpočetních procedur