Hidden Markov models for geolocation of fish

Martin Wæver Pedersen

Kongens Lyngby 2007

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Abstract

The present thesis strives to estimate the geographical location (geolocation) and movement of demersal fish based on tidal data extracted from electronic data storage tags (DSTs).

The theory of the underlying diffusion model is presented with emphasis on the connection between the partial differential equation governing its time evolution and a homogeneous random walk. The paradigm of a hidden Markov model is applied to the DST data considering the global coordinates as the hidden states furnishing the observable tidal output. A Bayesian filter offers a straightforward framework for maximum likelihood estimation of model parameters. The most probable sequence of hidden states, i.e. the Most Probable Track, is found by employment of the Viberti algorithm.

A simulation study is conducted to examine the method performance in terms of computation time and parameter estimation. Furthermore it is sought to elucidate the filtering step in greater detail and evaluate the influence of spatial variation in environmental variables such as depth. Conclusively, the maximum likelihood estimator is tested for bias and precision followed by an analysis of the optimal track representation.

The dataset considered in the project consists primarily of depth and temperature records from Atlantic cod (*Gadus morhua*) tagged in the southern North Sea and eastern English Channel. The initial data preprocessing extracts the pertinent tidal information and depth to be transferred to the filtering algorithm. The variance structure of the observed time series is assessed by means of stationary tags at known geographical positions.

The geolocation method is implemented in the MATLAB v. 7.0 computing environment that offers a flexible presentation of the geolocation. Animating the time evolution of the marginal posterior distributions in an avi-file gives a detailed visualisation of the uncertainty in each discrete time step. The Most Probable Track images the mode of the joint posterior distribution and is a representation that can be contained in a single figure thereby easing interpretation of the results. Explicit estimation of the joint posterior distribution is unique for the method and opens for a wide range of applications.

The presented results concurred with the general pattern of previous studies of the data but excelled in terms of detail and computation time. The method showed flexibility and was prone to extensions of which some were implemented in simplified forms for illustrative purposes.

The estimated fish behaviour is based on statistical rigor and can serve as substantial argumentation in future decisions related to stock assessment and fisheries management.

KEY WORDS: Geolocation, diffusion process, Atlantic cod, data storage tags, hidden Markov model, maximum likelihood estimation, Most Probable Track

Resumé

Dette eksamensprojekt tilstræber at estimere den geografiske position (geolokalisering) og bevægelse af demersale fisk på baggrund af tidevandsdata fra elektroniske dataopsamlingsmærker (DSTs).

Teorien for den underliggende diffusionsmodel præsenteres med vægt på forbindelsen mellem den partielle differentialligning, der beskriver dens tidsudvikling, og en homogen random walk. Antagelserne i en hidden Markov model anvendes på DST-data, ved at opfatte den globale position som den skjulte tilstand, der giver anledning til det observerbare tidevandssignal. Et Bayesiansk filter er et værktøj, der er velegnet til efterfølgende maximum likelihood estimation af modelparametre. Den mest sandsynlige sekvens af skjulte tilstande, dvs. det Mest Sandsynlige Spor, findes ved anvendelse af Viterbi-algoritmen.

Et simulationsstudie udføres for at undersøge metodens ydeevne mht. beregningstid og parameterestimation. Ydermere tilstræbes det at belyse selve filtreringen og at evaluere indflydelsen af den rumlige variation i omgivelsernes karakteristika, såsom dybden.

Det, i projektet anvendte, datasæt består hovedsageligt af dybde- og temperaturmålinger fra eksemplarer af den Atlantiske torsk (*Gadus morhua*), mærket med DSTs i den sydlige del af Nordsøen og i den østlige del af Den Engelske Kanal. Den initielle datapræprocessering udtrækker den relevante tidevandsinformation og dybde, som skal overføres til filteralgoritmen. Variansstrukturen af den observerede tidsrække bestemmes ved analyse af stationære mærker på kendte geografiske positioner.

Geolokaliseringsmetoden implementeres i beregningsværktøjet Matlab v. 7.0, hvor en fleksibel præsentation af geolokaliseringen er mulig. Ved at animere tidsudviklingen af den marginale posteriorfordeling i en avi-fil opnås en detaljeret visualisering af usikkerheden i hvert diskret tidsskridt. Det Mest Sandsynlige Spor viser modus i den simultane posteriorfordeling og er en repræsentation, som kan være indeholdt i en enkeltstående figur og dermed letter resultatfortolkningen. Eksplicit estimation af den simultane posteriorfordeling er enestående for metoden og muliggør en lang række applikationer.

De præsenterede resultater var i overensstemmelse med de generelle tendenser set i tidligere studier af samme data, men excellerede mht. detaljegrad og beregningstid. Metoden viste sig fleksibel og nem at udvide, hvilket blev illustreret gennem simple implementationer.

Den estimerede fiskeadfærd bygger på statistisk stringens og kan anvendes som tungtvejende argumentation i fremtidige beslutninger angående bestandsvurdering og fiskeristyring.

Preface

This master thesis was prepared at the institute for Informatics and Mathematical Modelling (IMM) at the Technical University of Denmark (DTU) in partial fulfillment of the requirements for acquiring the Master degree in engineering. The extent of the thesis is equivalent to 40 ETCS points.

The thesis deals with application of statistical methods to data extracted from data storage tags mounted on North Sea fish with the purpose to obtain estimates of their geographical location.

My supervisors Uffe Høgsbro Thygesen and Henrik Madsen and colleagues Ken Haste Andersen and David Righton deserve acknowledgement for their contribution to the thesis in form of ideas and discussions. I am indebted to the kind people at the CEFAS Laboratory for sharing their wisdom and for providing the DST and environmental data.

This work was supported by Oticon Fonden.

Martin Wæver Pedersen, March 2007

Contents

Abbreviations

Abbreviation

AMPD	Animated Marginal Posterior Distributions.
acf	autocorrelation function.
BM	Brownian Motion.
CEFAS	Center for Environment, Fisheries and Aquaculture Science.
cdf	cumulated density function.
DAG	Directed Acyclic Graph.
DIFRES	Danish Institute for Fisheries RESearch.
DST	Data Storage Tags.
DTU	Technical University of Denmark.
FAO	Food and Agriculture Organization.
FEM	Finite Element Method.
ICES	International Council for Exploration of the Sea.

IMM	institute for Informatics and Mathematical Modelling.
IUCN	International Union for Conservation of Nature and natural resources.
MLE	Maximum Likelihood Estimate.
MPT	Most Probable Track.
ODE	Ordinary Differential Equation.
PDE	Partial Differential Equation.
pdf	probability density function.
POL	Proudman Oceanographic Laboratory.
POM	Princeton Oceanographic Model.
PSAT	Pop-up Satellite Archival Tags.
TLM	Tidal Location Method (Hunter et al., 2003).

Chapter 1

Introduction

This introductory chapter deals with the background and motivation for the thesis. It describes the previous studies related to estimation of geographical location (geolocation) of marine animals using various available technology. This work mainly comprises conventional tags and data storage tags (DSTs). Traditional methods employed to estimation of the position and movement of the fish are examined briefly. Conclusively, the aims for the present study are outlined along with an overview of the structure of the thesis.

1.1 Motives of geolocation

As technology became available the efficiency of fishing improved through the twentieth century. Along came a need to control the fishing effort in order to retain the depleting stocks of particular species that were of significant commercial interest. Recent examples of this endangerment of species are the Oceanic Whitetip Shark *(Carcharhinus longimanus)* and the Angel Shark *(Squatina squatina)* that are mostly caught as bycatch by pelagic fisheries and bottom trawl. This inconvenient situation has put the species on the "Red List of Threatened Species" published by the organisation "International Union for Conservation of Nature and natural resources", (IUCN, 2006).

Another example is the Haddock *(Melanogrammus aeglefinus)* that suffered from overfishing in the 1960s and up until recent years, but has now due to a series of regulations recovered its stock to some extent (FAO, 2004). The Haddock draws a lot of similarities to the Atlantic cod *(Gadus morhua)*, see Figure 1.1 (Bloch, 1785), both in taste and looks and unfortunately fate as well. The northwest Atlantic cod was during the early 1990s severely overfished which caused the stock to collapse leaving only relatively few specimens (FAO, 2004). This unnatural low stock resulted in other species taking the role as top predator now feeding on the Atlantic cod hence making it even harder for the species to recover.

The stock of the northeast Atlantic cod has recently diminished in size forcing experts of the "International Council for Exploration of the Sea" (ICES) to recommended a full stop of cod fishing in the North Sea.

Figure 1.1: *The Atlantic cod (Gadus morhua).*

The way to avoid scenarios as the ones mentioned goes through regulation of fishing efforts and an intelligent use of marine protected areas. In order to do so, informations on location of biomass, spawning grounds and fish behaviour must be assessed. Geolocation can supplement this assessment and hopefully be an aid to replenish the reduced stocks and extend our knowledge of fish behaviour.

1.2 History of tag based geolocation

Tagging of fish is a wide spread technique to gain information of behaviour and to obtain global positional estimates of the tagged individual.

1.2.1 Conventional tags

A tagging experiment consists of mounting simple markers on fish in a way that has the least possible effect on the behaviour and growth (Righton et al., 2006). A batch of fish is released into the sea with the intention that some percentage is recaptured and their tag recovered. This type of mark/recapture experiments, or "conventional tagging", were initially a mean to asses the mortality of fish by evaluating the return rate of the tags. As a side product, the experiments also supplied information of the recapture positions that gave rise to tag based geolocation.

Conventional tagging methods yield only a sparse dataset per returned tag, and therefore requires extensive tagging for major conclusions on the distribution of individuals to be made. Fortunately the procedure is associated with low costs and has been carried out since the mid sixties up until the present day, hence a substantial amount of data is available (Daan, 1978; Righton et al., 2007). However, the number of returns from a given geographical area is largely influenced by the fishing effort, thus diminishing the statistical power of the data.

The present thesis focuses on the Atlantic cod - henceforth referred to as cod - and the habitats of the North Sea and the English Channel. Figure 1.2 shows a map of the ICES areas that are contained in the considered domain. Previous work has shown that cod released in the southern North Sea tend to either stay in a limited area close to the release position or migrate north (Righton et al., 2007). Migration is often performed in an annual cycle bringing the cod to the central part of the North Sea (ICES IVb) in the summer, before returning south during the winter (Righton et al., 2007). This behaviour is confirmed by research based on DTSs (Righton et al., 2000). No annual migration cycle has so far been proven by conventional tagging for cod released in the English Channel. In fact, not much can be said about cod released in VIId besides that the majority was recaptured close to the release location regardless of its time at liberty (Righton et al., 2000; Righton et al., 2007).

The obvious drawback of conventional tagging is the scarce amount of data returned from one tag, rendering it difficult to deduce the behaviour whilst at

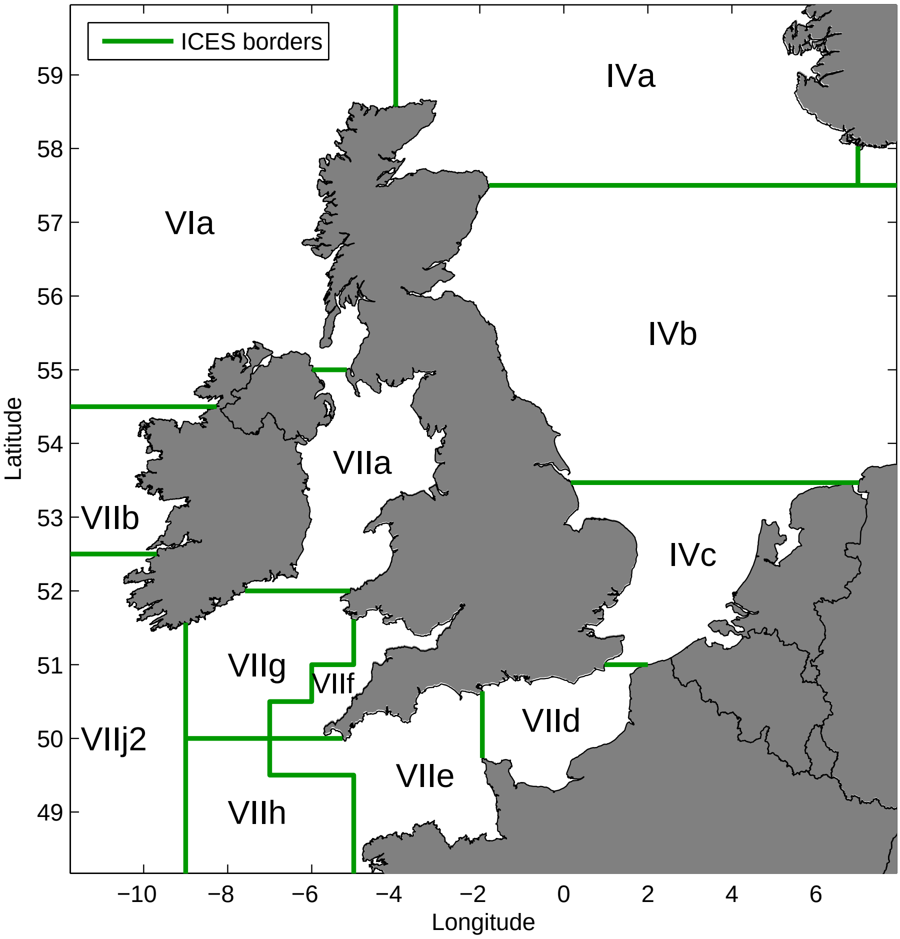

Figure 1.2: *Map showing the ICES areas.*

liberty. A cod recaptured close to its release position could possibly have made large excursions in the intervening period. It was therefore a great advance for the field of geolocation when DSTs where introduced as data collectors.

1.2.2 Data Storage Tags

DSTs come in a variety of types and sizes (see Section 5.2) and have in their short history been used for geolocation of many kinds of marine animals. For cod, the tagging procedure itself has developed as well, to cover both external and internal tagging of the fish (Righton et al., 2006). Compared to conventional mark/recapture tagging, the DST experiments have substantial added costs. It is therefore of great interest to extract maximal information from a successfully returned tag.

In the tagging procedure emphasis is put on minimising the traumatisation of the individual. The cod is either caught by line or by trawl and brought to the surface slowly to avoid swimbladder rupture. Here they are anaesthetised before the tag is mounted, either externally next to the first dorsal fin, or internally in the peritoneal cavity along with an external marker (Righton et al., 2006).

When the fish is released into the sea the DST logs information of the environment such as depth, light, temperature or salinity. The choice of measure depends in general on the species and its immediate environment. For example in the Baltic Sea, tagging experiments have been performed mostly with DSTs measuring depth, temperature and salinity exploiting the, in some areas, large gradients of these quantities (Neuenfeldt et al., 2006).

In the Pacific Ocean for tracking bigeye tuna, DSTs measuring ambient light have been used. The uncertainty of the light based geolocation is very seasonal dependent and increases especially around the equinox (Musyl et al., 2001).

Another type of DST used for geolocation is a pop-up satellite archival tag (PSAT). The tag self-releases from the animal at a preprogrammed time and transmits the data via satellite when reaching the surface. Due to the transmission process the PSAT has a large battery requirement compared to a DST and the amount of retrievable data is in general limited.

PSATs are normally used for animals that are not targeted by commercial fishermen, and therefore satellite transmission is the only way of retrieving the data. Among the applications are investigations of the dive behaviour and post-release mortality following interactions with longline fishing gear of olive ridley sea turtles (*Lepidochelys olivacea*) (Swimmer et al., 2006), and geolocation of Greenland sharks (*Somniosus microcephalus*) (Stokesbury et al., 2005).

Pressure measurements from demersal species have a great potential for geolocation. When the fish dwells at the sea bed for a longer period of time, the pressure recorded by the DST is constant except for variations following the tide. This tidal signal is compared to a numeric tidal forecast system and the possible positions can be found. A greater study using tidal patterns for geolocation was conducted successfully on plaice (*Pleuronectes platessa L.*) in the North Sea (Hunter et al., 2004). Tidal location work in progress focus also on other demersal species such as sole (*Solea solea*) and ray (*Raja clavata*), aiming to clarify migration routes and seasonal behaviour etc. Likewise, the Atlantic cod has been subject to ongoing DST research of which some results are presented in Turner et al. (2002); Righton et al. (2007).

1.3 Methods

The geolocation work based on electronic tagging experiments is extensive and covers a wide range of methodology and approaches. The first heuristic methods such as the Tidal Location Method (Metcalfe and Arnold, 1997), assesses the position of the fish by direct comparison of environmental variables with observations. The data analysis are to some extent influenced by subjectivity and the manual workload of data comparison can be very time consuming.

Later a state-space approach was presented in Sibert et al. (2003), that used the Kalman filter for tracking of bigeye tuna. This statistically well-founded method lead to straightforward estimation of model parameters by a maximum likelihood approach. Position estimates were given by the conditional mean and its error. The Gaussianity assumption of the method will in general be violated for fish swimming close to dry land, which is the case for many marine animals of commercial interest.

Nielsen (2004) suggested applying an extended Kalman filter as a solution to this, but a more flexible approach is the particle filter that does not rely on distribution assumptions or linearisations. Applications of the particle filter include a simulation study of light based geolocation (Nielsen, 2004), geolocation of cod in the Baltic (Andersen et al., 2007), tracking bluefin tuna in the Atlantic (Royer et al., 2005) to name a few. Major drawbacks of the method are the substantial computational efforts required by the filter and the numerical issues that arise in the smoothing step.

1.4 Aims of the present study

This thesis aims to build upon the above mentioned experiences and contribute to the field of geolocation by developing a method with emphasis on practical applications.

A hidden Markov model with a homogeneous diffusion process describing the movement, is assumed. The hidden positions are estimated by application of a

Bayesian filter to the DST observations. The filter handles arbitrary distributions, but avoids some of the numerical issues of the particle filter by considering the time evolution of the distribution itself, instead of representing it as particles. The reconstructions obtained from the Bayesian filter are smoothed in a backwards sweep yielding estimates of position conditioned on the whole set of observations.

The posterior distribution for the position, explicitly expresses its uncertainty and enables the method to output many interesting summary statistics. The thesis explores the concept of the “Most Probable Track”, that is a valuable representation of the joint posterior distribution. Also, an illustrative representation of the results is given in the form of an “Animated Marginal Posterior Distribution” that sequentially displays the estimated distribution in an avi-file.

The presented methods are evaluated both in a simulation model and in a study of data from tagged fish in the North Sea (cod and ray).

1.5 Thesis outline

The thesis is partitioned into three parts and should be read in sequence.

Part I: Fundamentals and theory of geolocation. The basic model assumptions and their supporting theory is introduced along with the filtering method, where especially the smoothing step is described in detail. Also the basic methodology with regards to likelihood estimation of the parameter(s) and determination of the Most Probable Track of the joint posterior distribution. This part ends with a simple simulated experiment that aims to verify the assumptions made in the modelling process via statistical hypothesis testing.

Part II: Geolocation of North Sea fish. A stochastic geolocation model based on depth measurements and their inherent tidal pattern is constructed in analogy with the simulation model of Part I. The model is tested on stationary DSTs from minipods for precision and validation before applying the method to data from North Sea fish. When possible the results are compared to previous research. Finally, model extensions are proposed based on the experiences made with the method, and their relevance is evaluated via simplified implementations.

Part III: Outlook and conclusion. Here the main results are discussed with a view into relevant topics for improvement and expansion of the presented method. The thesis is rounded off in a conclusion of the work.

Enclosed CD-ROM. The enclosed CD-ROM contains pdf-files and MATLAB fig-files containing plots of the real data sets used in the thesis. Also, is included animations of the results. The files are also found on the website www.student.dtu.dk/∼s002087.

Important notice: The files on the CD-ROM are not to be distributed without permission from CEFAS.

1.6 Symbol overview

Data from a tag contains time series of depth and temperature. The time is presented as a column vector

$$\boldsymbol{t} = [t_0, t_1, \dots, t_i, \dots, t_n]^T.$$

As default the sample rate of the tag, $t_{i+1} - t_i$, is 0.00694 day (10 minutes) if nothing else is stated.

The time series of depth is written

$$\boldsymbol{z} = [z_0, z_1, \dots, z_i, \dots, z_n],$$

where $z_i = -a$ denotes a water column height of a m.

Another time scale that will be useful later, contains the days inherent in the data i.e.

$$\boldsymbol{\tau} = [\tau_0, \tau_1, \dots, \tau_j, \dots, \tau_N]^T,$$

where τ_0 is the day of release and τ_N is the day of recapture. Note that j is used as index for this time scale. The time step of the τ scale is

$$\tau_{j+1} - \tau_j = k = 24\,\text{hours}.$$

The reason for this 24 hour interval is given later in the thesis.

Temperature is measured four times a day and written

$$\boldsymbol{q}_j = [q_1, q_2, q_3, q_4].$$

Subsamples of the depth time series from day j are subscripted also by the index of the earliest point in the sample

$$\boldsymbol{z}_{j,i} = [z_i, z_{i+1}, \ldots, z_i, \ldots, z_{i+m}].$$

A geolocated track is given in an array containing the global positions at the beginning of each day

$$\boldsymbol{x}_j = [x_{j,1}, x_{j,2}]^T,$$

where $x_{j,1}$ is the longitudinal coordinate (abscissae) and $x_{j,2}$ is the latitudinal coordinate (ordinate) for day j. With this terminology the geolocated release and recapture positions are written as $\boldsymbol{x}_0$ and $\boldsymbol{x}_N$ respectively and their observed (occasionally called reported) counterparts $\boldsymbol{x}_\dagger$ and $\boldsymbol{x}_\ddagger$.

Stochastic variables are written in capital letters hence the stochastic variables of the position is denoted $\boldsymbol{X}_j$.

An entire track is written in a matrix

$$\boldsymbol{\xi} = [\boldsymbol{X}_0 = \boldsymbol{x}_0, \ldots, \boldsymbol{X}_j = \boldsymbol{x}_j]^T.$$

To ease notation an observation matrix is defined as

$$\boldsymbol{\mathcal{Y}}_j = [\boldsymbol{Y}_0 = \boldsymbol{y}_0, \ldots, \boldsymbol{Y}_j = \boldsymbol{y}_j]^T,$$

that contains the observations from τ_0 up until time τ_j. This is not a matrix in a strict mathematical sense, as the number of elements in $\boldsymbol{y}_j$ varies depending on j

$$\boldsymbol{y}_j = \begin{cases} [\boldsymbol{x}_\dagger] & \text{for } j = 0 \\ [\boldsymbol{z}_{j,\widehat{i}}]^T & \text{for } j \in [1, \ldots, N-1] \\ [\boldsymbol{x}_\ddagger^T, \boldsymbol{z}_{j,\widehat{i}}]^T & \text{for } j = N \end{cases} .$$

The same vector including temperature observations is denoted

$$\boldsymbol{\mathcal{V}}_j = [\boldsymbol{v}_0, \ldots, \boldsymbol{v}_j]^T,$$

where

$$\boldsymbol{v}_j = \begin{cases} [\boldsymbol{y}_j] & \text{for } j = 0 \\ [\boldsymbol{y}_j^T, \boldsymbol{q}_j]^T & \text{for } j \in [1, \ldots, N] \end{cases} .$$

1.6.1 Additional notation

D	Diffusivity of the fish.
E	White noise error.
e	Tidal error.
ε	Error in auto regressive model.
λ	Weight in auto regressive model.
η	Bathymetry roughness error.
ψ_j	Normalisation constant for day j.
$\mathbb{E}(X)$	Expectation of the random variable X.
$\mathbb{V}(X)$	Variance of the random variable X.
$\mathbb{P}(X = x)$	Probability of the event $X = x$.
$f(\cdot)$	A function of not explicitly stated variables.
$\mathcal{F}(\phi) = \widehat{\phi}$	Fourier transform of ϕ.
$\mathcal{F}^{-1}\left(\widehat{\phi}\right) = \phi$	Inverse Fourier transform of $\widehat{\phi}$.
$X \sim \mathcal{N}(\mu, \sigma^2)$	X is Gaussian distributed with mean μ and variance σ^2.
$\mathcal{L}(A)$	Likelihood of A.
ℓ	Log likelihood function.

Part I

Fundamentals and theory of the geolocation method

Chapter 2

Diffusion

In the model building process, reasonable assumptions are made to simplify reality to an extent that makes the descriptive and implementational task feasible. The assumptions will always violate the true dynamics of the system and must therefore be borne in mind when evaluating the results.

The concept of Brownian Motion (BM) has traditionally been used for describing movement of particles that perform an erratic random behaviour through space. It was first observed by the botanist Robert Brown in 1828 and later formalised in the famous paper, Einstein (1905), that introduced the connection between BM and diffusion. A more mathematical oriented approach is found in Grimmett and Stirzaker (2001), whereas Okubo and Levin (2002) and Berg (1993) emphasise biological aspects of the topic.

BM may not seem appropriate as a model for the movement of fish as they are neither erratic nor are their actions (entirely) random. When the movement process is observed on a short time scale, this assertion is true. However, over a longer time period BM has proven to be a good descriptor of fish movement (Sibert and Fournier, 2001; Jonsen et al., 2003; Nielsen, 2004; Andersen et al., 2007). The concept of BM has different interpretations depending on field of research and it is therefore stressed that this thesis relies on the mathematical understanding, i.e. a homogeneous random movement in space.

For a particle performing a BM in d dimensions, the partial differential equation (PDE) governing the time evolution of the *probability density function* (*pdf*) associated to the position of the particle is given by the diffusion equation

$$\frac{\partial \phi}{\partial t} = D \sum_{i=1}^{d} \frac{\partial^2 \phi}{\partial x_i^2}, \tag{2.1}$$

where $-\infty < x_i < \infty$, $i \in [1, \ldots, d]$ and $t > 0$. D is the diffusivity parameter and $\phi = \phi(x_1, \ldots, x_d, t)$ is the *pdf* of the position of the particle.

The key assumption of this thesis is, that the movement of a fish causes the probability density of its position to evolve in time according to (2.1). It is a deliberate choice to omit an advection (drift) term, to maintain a simple model with a minimal parameter space. Also, it is rarely the case that the bias in fish movement remain constant over time and therefore it cannot be described by a simple advection model.

The present chapter shows three interpretations of the diffusion equation. This involves an analytical solution by Fourier transform and a discretised solution. A finite difference solution to the diffusion equation is shown to be analogous to a discrete random walk that in turn converges to the diffusion process when temporal and spatial steps shrink towards zero.

2.1 Analytical solution of diffusion

The general solution to the d-dimensional diffusion equation (2.1) is found by a vectorised combination of the separated one-dimensional solutions.

The solution in the one dimensional case of (2.1), $\frac{\partial \phi}{\partial t} = D\frac{\partial^2 \phi}{\partial x^2}$, can be obtained through a Fourier transform (denoted by $\mathcal{F}$) of the PDE with respect to x (Asmar, 2004). This yields

$$\begin{aligned} & \mathcal{F}\left(\frac{\partial \phi}{\partial t}\right) = \mathcal{F}\left(D\frac{\partial^2 \phi}{\partial x^2}\right) \\ \Leftrightarrow \quad & \frac{d}{dt}\widehat{\phi}(\omega, t) = -D\omega^2 \widehat{\phi}(\omega, t), \end{aligned} \tag{2.2}$$

where $\widehat{\phi}$ denotes the Fourier transformed version of ϕ. Equation (2.2) is an ordinary differential equation (ODE) in t when ω is fixed, with initial condition $\phi(x, 0) = f(x)$. For the solution to be a probability distribution it must hold for the initial condition that $\int f(x)dx = 1$ and that $f(x) \geq 0$ for all x.

Fourier transform of the initial condition gives

$$\begin{aligned} & \mathcal{F}\big(\phi(x,0)\big) = \mathcal{F}\big(f(x)\big) \\ \Leftrightarrow \qquad & \widehat{\phi}(\omega,0) = \widehat{f}(\omega). \end{aligned}$$

The general solution to the first order ODE, is

$$\widehat{\phi}(\omega,t) = A(\omega)e^{-D\omega^2 t}.$$

The transformed initial condition is used to find $A(\omega)$

$$\widehat{\phi}(\omega,0) = \widehat{f}(\omega) = A(\omega),$$

so that the specific solution to (2.2) becomes

$$\widehat{\phi}(\omega,t) = \widehat{f}(\omega)e^{-D\omega^2 t}.$$

Applying the inverse Fourier transform, $\mathcal{F}^{-1}$, the solution to the diffusion equation (2.1) is obtained

$$\phi(x,t) = f(x) * \mathcal{F}^{-1}\left(e^{-D\omega^2 t}\right), \tag{2.3}$$

where $*$ is the convolution operator. The second term in (2.3) can be evaluated and gives

$$H(x,t) = \mathcal{F}^{-1}\left(e^{-D\omega^2 t}\right) = \frac{1}{2\sqrt{\pi D t}} e^{-\frac{x^2}{4Dt}}. \tag{2.4}$$

This is recognised as the *pdf* of a Gaussian distribution with mean $\mu = 0$ and variance $\sigma^2 = 2Dt$. In PDE terminology it is known as a Gauss kernel.

The convolution operator is defined as the integral

$$f(x) * g(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x-y)g(y)\,dy.$$

With this definition (2.3) is rewritten as

$$\phi(x,t) = H(x,t) * f(x) = \int_{-\infty}^{\infty} H(x-y,t)f(y)\,dy.$$

In general this can be written

$$\phi(x,t) = H(x,t-s) * \phi(x,s) = \int_{-\infty}^{\infty} H(x-y,t-s)\phi(y,s)\,dy, \tag{2.5}$$

where $\phi(x,s)$ is the density at time s and $H(x,t-s)$ is the kernel for the time step $t-s$.

The conclusion is that the solution to the diffusion equation (2.1) is obtained by a convolution of the initial condition, $f(x)$, with a Gauss kernel, $H(x,t)$.

The closing section of this chapter shows that (2.5) is the continuous analogue to the solution of the discrete diffusion equation.

2.2 Discrete solution of diffusion

For a PDE to be solved numerically it must be discretised in some way to allow for implementation. There exists many ways to perform this discretisation that all have their pros and cons. PDE problems with geometric complex boundary conditions, as the one considered here (islands, bays etc.), is preferably solved with the Finite Element Method (FEM), which is a complex but powerful approach (Cook et al., 2001). The finite difference method (Asmar, 2004), is numerically simpler than FEM and will suffice as an approximation. Issues with complex boundaries are to some extent overcome implicitly by the nature of the problem in that the recorded depth of a DST is always below the sea surface. This restricts the possible positions of the fish to the sea, and serves as pseudo boundary conditions of the problem.

To obtain the finite difference scheme, the one-dimensional case of the diffusion equation is discretised by replacing differential quotients by difference quotients

$$\frac{\phi(x,t+k)-\phi(x,t)}{k} = D\left(\frac{\phi(x+h,t)-2\phi(x,t)+\phi(x-h,t)}{h^2}\right),$$

which is rearranged to yield the recursive equation

$$\phi(x,t+k) = r\phi(x-h,t) - (1-2r)\phi(x,t) + r\phi(x+h,t), \tag{2.6}$$

with

$$r = Dk/h^2, \tag{2.7}$$

where k is the time step and h denotes the spatial step.

The solution to the diffusion equation is a probability distribution which in its nature is bounded on an infinite domain as its integral is one. For this condition to hold for the discretised equation bounds are imposed on r. The future value $\phi(x,t+k)$ receives a contribution from the present, $\phi(x,t)$, and the two neighbouring cells, $\phi(x-h,t)$ and $\phi(x+h,t)$. The term “cell” refers to a position in the discrete temporal and spatial domain grid. The proportion carried on from each cell is limited by $(1-2r)$ which imply that $0 < r < 0.5$ and further

$$\frac{2Dk}{h^2} < 1, \tag{2.8}$$

when $D > 0$, $h > 0$ and $k > 0$. This is also known as the stability criteria for the finite difference solution (2.6).

The time updating equation (2.6) can be written as a vector multiplication

$$\phi(x, t+k) = [\phi(x-h, t), \phi(x, t), \phi(x+h, t)] \times H,$$

that gives the solution $\phi(x, t+k)$ for all x, where

$$H = [r, -(1-2r), r]^T, \tag{2.9}$$

is a 3×1 one-dimensional convolution kernel, the discrete analogue of (2.4).

A real data model may lead to values of h, k and D that cause the finite difference scheme to become unstable due to (2.8). The solution to this is to perform several time updates within one time step, effectively reducing the value of k. This corresponds to a convolution of the distribution at t with an extended kernel of size $(2m+1) \times 1$, where m is the number time updates performed within one time step of length k. For the case $m = 1$ this is equal to (2.9).

The next section views the finite difference scheme from an angle of stochastic processes and shows the direct link to a homogeneous random walk.

2.3 Random walk approximation to diffusion

The continuous time stochastic process that describes a particle exercising Brownian motion is the Wiener process. It is characterised by having independent and Gaussian distributed increments which Section 2.1 showed to be a property of $\phi(x, t)$.

Results in Chandrasekhar (1943); Okubo and Levin (2002) show that the Wiener process can be approximated by a simple random walk process

$$X_j = \sum_{i=1}^{j} U_i,$$

where U_i is the movement in one time increment, k, and has the distribution

$$\mathbb{P}(U_i = u) = \begin{cases} r & \text{for} \quad u = -h \\ 1-2r & \text{for} \quad u = 0 \\ r & \text{for} \quad u = +h \end{cases} . \tag{2.10}$$

The process, X_j, is a Markov process in that it has independent increments and $\mathbb{P}(X_{j+1} = x_{j+1}|X_j, \dots, X_0) = \mathbb{P}(X_{j+1} = x_{j+1}|X_j)$. A popular description says that for a Markov process it holds that “given the present, the future is independent of the past”, referred to as the *Markov property*.

The Central Limit Theorem says that when $k \downarrow 0$ and $h \downarrow 0$, the distribution of X_j will be Gaussian with mean zero and variance $j\mathbb{V}(U_i) = j(2rh^2) = j(2Dk)$ implying

$$X_j \sim \mathcal{N}(0, 2Dt),$$

which is equal to the Gauss kernel of (2.4), where $t = jk$ is the elapsed time interval.

The prediction or *time-update* of the process can be found by constructing the probability transition matrix of the process. The state space of the process is in principle infinite but can be written on index form where the probability to be in state x_1 at time j is denoted $\mathbb{P}(X_j = x_1)$. This probability is determined by using the standard rule of average conditional probability

$$\mathbb{P}(X_{j+1} = x_1) = \sum_x \underbrace{\mathbb{P}(X_{j+1} = x_1|X_j = x)}_{\text{transition}} \underbrace{\mathbb{P}(X_j = x)}_{\text{distribution}}. \tag{2.11}$$

Equation (2.11) is equal to the finite difference scheme in (2.6) when the transition probability is given by (2.10) and $x_1 - h \leq x \leq x_1 + h$ is fulfilled. The term $\mathbb{P}(X_j = x)$ is the distribution at the present and is equal to $\phi(x, t)$. Finally it is noted that (2.11) is a convolution sum and the discrete counterpart of (2.5).

2.4 Conclusion

It can be concluded that the three interpretations of diffusion presented in this chapter lead to identical calculations and results in continuous and discrete space, respectively. It is a powerful observation to have in mind that enables tools from a wide spectrum of mathematical fields to work in synergy.

Chapter 3

Filtering and estimation

Given the assumed behaviour model the movements of the tagged individual can be predicted. The estimated geolocations are obtained by numerical filtering that is described in this chapter. The method to be presented is closely related to already known filtering techniques such as the Kalman filter and state space modelling in general (Madsen, 2001), but relaxes their requirement of Gaussianity. The filtering problem is put into the framework of a hidden Markov model (Cappé et al., 2005). The principle is sketched in Figure 3.1.

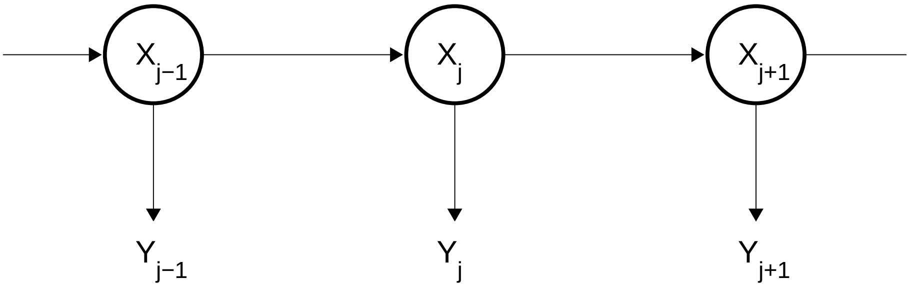

Figure 3.1: *Sketch of the hidden Markov model. X - hidden states (geolocations), Y - observable outputs (depths).*

The geolocation of the fish is considered as the hidden state, written X_j. The observed depth record from the DST is the output from the model, marked with Y in Figure 3.1. The objective is to process (filter) the observed output to gain estimates of the hidden states and their distribution.

3.1 Estimated positions

The filter works as a recursive process that relies on successive predictions and reconstructions of the position, $\boldsymbol{X}_j$. A short-hand notation for the observations up to time τ_j is

$$\boldsymbol{\mathcal{Y}}_j = [\boldsymbol{Y}_0 = \boldsymbol{y}_0, \dots, \boldsymbol{Y}_j = \boldsymbol{y}_j]^T,$$

where $\boldsymbol{y}_0$ is the observations related to day j (see also Section 1.6). The filter is initialised with the equation

$$\mathbb{P}(\boldsymbol{X}_0 = \boldsymbol{x} | \boldsymbol{\mathcal{Y}}_0) = \begin{cases} 1 & \text{for} \quad \boldsymbol{x} = \boldsymbol{x}_\dagger \\ 0 & \text{for} \quad \text{otherwise} \end{cases},$$

where $\boldsymbol{x}_\dagger$ is the release position that here is assumed to be known without uncertainty.

3.1.1 Prediction

The prediction step attempts to find $\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1} | \boldsymbol{\mathcal{Y}}_j)$, i.e. the probability of the position at the next time point given all preceding observations.

Using (2.11) and applying the *Markov property*, it is found that

$$\begin{aligned} \mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1} | \boldsymbol{\mathcal{Y}}_j) &= \sum_{\boldsymbol{x}} \mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1} | \boldsymbol{X}_j = \boldsymbol{x}, \boldsymbol{\mathcal{Y}}_j) \mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x} | \boldsymbol{\mathcal{Y}}_j) \\ &= \sum_{\boldsymbol{x}} \mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1} | \boldsymbol{X}_j = \boldsymbol{x}) \mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x} | \boldsymbol{\mathcal{Y}}_j). \qquad (3.1) \end{aligned}$$

This step is also called the *time-update* of the states or the *one-step prediction.*

3.1.2 Reconstruction

Whenever a new observation, $\boldsymbol{y}_{j+1}$, is introduced a reconstruction is performed of the position using Bayes' rule

$$\begin{aligned}&\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{Y}}_{j+1})\\ &= \mathbb{P}(\boldsymbol{Y}_{j+1} = \boldsymbol{y}_{j+1}|\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}, \boldsymbol{\mathcal{Y}}_j)\frac{\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{Y}}_j)}{\mathbb{P}(\boldsymbol{Y}_{j+1} = \boldsymbol{y}_{j+1}|\boldsymbol{\mathcal{Y}}_j)}.\end{aligned} \tag{3.2}$$

This step is also referred to as the *data-update*. After the reconstruction, the geolocation is conditioned on all preceding observations and the present one, $\boldsymbol{y}_{j+1}$.

In practice, the term $\mathbb{P}(\boldsymbol{Y}_{j+1} = \boldsymbol{y}_{j+1}|\boldsymbol{\mathcal{Y}}_j)$ can be considered a normalisation constant and (3.2) is reformulated as

$$\begin{aligned}&\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{Y}}_{j+1})\\ &= \psi_{j+1} \cdot \mathcal{L}(\boldsymbol{Y}_{j+1} = \boldsymbol{y}_{j+1}|\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1})\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{Y}}_j),\end{aligned} \tag{3.3}$$

where $\mathcal{L}(\boldsymbol{Y}_{j+1} = \boldsymbol{y}_{j+1}|\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1})$ is the unnormalised conditional probability of the observation given the position, henceforth (for convenience) known as the likelihood for the observation given the position or "observational likelihood". The *one-step prediction error*, $\psi_{j+1} = \mathbb{P}(\boldsymbol{Y}_{j+1} = \boldsymbol{y}_{j+1}|\boldsymbol{\mathcal{Y}}_j)^{-1}$, is equal to the normalisation constant that ensures that the probability of the whole outcome space sums to one.

3.2 Smoothed positions

A thorough presentation of the smoothing step is given as it is rarely considered on this form in the literature. The aim is to find the distribution of $\boldsymbol{X}_j$ conditioned on all observations, i.e. $\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_N)$.

First consider the random variables A, B and C. Their dependence relations are sketched in a Directed Acyclic Graph (DAG), see Figure 3.2.

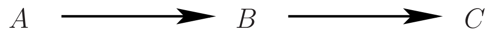

Figure 3.2: *Directed Acyclic Graph for the independence relations between A, B and C. A and C is seen to be conditional independent given B, this is a consequence of the Markov property.*

For clarity is introduced the notation $\mathcal{P}(\cdot)$ meaning the probability of "$\cdot$". According to Wasserman (2005) the Markov chain sketched in Figure 3.2 implies the following independence relations

$$\mathcal{P}(A = a|B, C) = \mathcal{P}(A = a|B),$$

i.e. A and C are conditionally independent given B. Now the smoothing equation can be found

$$\begin{aligned}
\mathcal{P}(A = a|C) &= \sum_b \mathcal{P}(A = a, B = b|C) \\
&= \sum_b \mathcal{P}(A = a|B = b, C)\mathcal{P}(B = b|C) \\
&= \sum_b \mathcal{P}(A = a|B = b)\mathcal{P}(B = b|C) && \text{(cond. independence)} \\
&= \sum_b \mathcal{P}(B = b|A = a)\frac{\mathcal{P}(A = a)}{\mathcal{P}(B = b)}\mathcal{P}(B = b|C) && \text{(Bayes' rule)} \\
&= \mathcal{P}(A = a)\sum_b \mathcal{P}(B = b|A = a)\frac{\mathcal{P}(B = b|C)}{\mathcal{P}(B = b)}. && (3.4)
\end{aligned}$$

The sketch in Figure 3.3 seeks to give a more intuitive interpretation of (3.4) as an update from $\mathcal{P}(A = a, B = b)$ to $\mathcal{P}(A = a, B = b|C)$ by multiplication with the ratio between the marginal of B, with and without the new information C. This is only valid due to the conditional independence of A and C given B. Summing over B in $\mathcal{P}(A = a, B = b|C)$ then yields the desired $\mathcal{P}(A = a|C)$.

To accomplish the aim of this section define

$$\begin{aligned}
A &= \boldsymbol{X}_j. \\
B &= \boldsymbol{X}_{j+1}. \\
C &= [\boldsymbol{Y}_{j+1} = \boldsymbol{y}_{j+1}, \dots, \boldsymbol{Y}_N = \boldsymbol{y}_N]^T. \\
\mathcal{P}(\cdot) &= \mathbb{P}(\cdot|\boldsymbol{\mathcal{Y}}_j),
\end{aligned}$$

where "$\cdot$" means not explicitly stated variables. It is noted that $\mathcal{P}(A = a|C) = \mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_N)$, which is the objective of this filtering step.

Applying the new definitions to (3.4) gives the smoothed estimate

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_N) \\
&= \mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_j)\sum_{\boldsymbol{x}_{j+1}} \mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{X}_j = \boldsymbol{x}_j)\frac{\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{Y}}_N)}{\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{Y}}_j)}. \quad (3.5)
\end{aligned}$$

The result is interpreted as the reconstruction, $\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_j)$, at time τ_j multiplied by a time-update *backwards in time* of the ratio between the smoothed

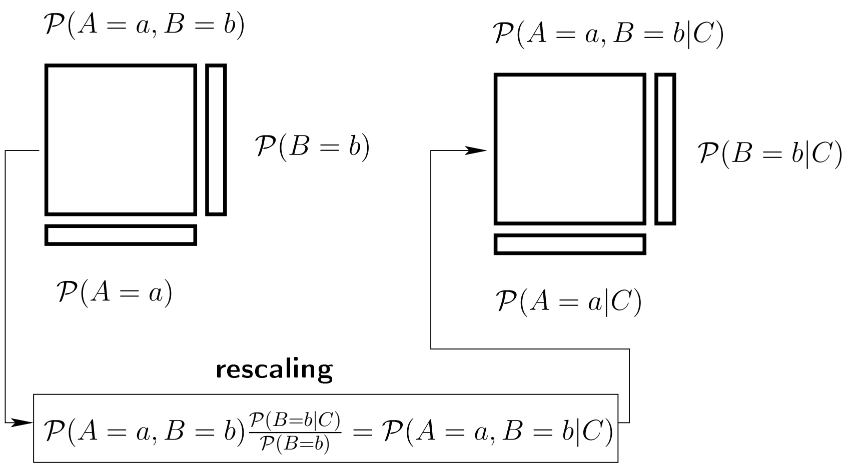
$\mathcal{P}(A=a, B=b)$
$\mathcal{P}(B=b)$
$\mathcal{P}(A=a)$
$\mathcal{P}(A=a, B=b|C)$
$\mathcal{P}(B=b|C)$
$\mathcal{P}(A=a|C)$
rescaling
$\mathcal{P}(A=a, B=b)\frac{\mathcal{P}(B=b|C)}{\mathcal{P}(B=b)} = \mathcal{P}(A=a, B=b|C)$

Figure 3.3: *A sketch of how the distribution of A given C is obtained. The joint distribution of A and B conditioned on C is given by a* ***rescaling*** *of the joint distribution of A and B with the new information, C, via the marginal distribution of B as indicated by the arrows. Summing over B in the conditional joint distribution gives the marginal of A given C as wished.*

position, $\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{Y}}_N)$, and the predicted position, $\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{Y}}_j)$ at the time τ_{j+1}. The symmetric transition matrix, according to (2.10), implies that forward and backward updates are identical calculations. The result of (3.5) is often referred to as the marginal posterior distribution of $\boldsymbol{X}_j = \boldsymbol{x}_j$ given $\boldsymbol{\mathcal{Y}}_N$.

The recursive scheme is initialised with the final reconstruction estimate, that is also a smoothed estimate, in that it is conditioned on all observations

$$\mathbb{P}(\boldsymbol{X}_N = \boldsymbol{x}_N|\boldsymbol{\mathcal{Y}}_N).$$

The smoothing step is very important for weeding out geolocated dead ends from the reconstruction step and generally makes the position estimates much more precise.

3.3 Likelihood estimation

The model may contain several parameters relevant for estimation e.g. the diffusivity, D, related to the swimming speed of the fish. Others may be variance parameters in the determination of the observational likelihood, $\mathcal{L}(\boldsymbol{Y}_j =$

$\boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j)$.

The parameters, subject to estimation, are denoted by $\boldsymbol{\theta}$ and are assumed to remain constant in time. Hence, the likelihood is given by the joint *pdf* of the observations, $\boldsymbol{\mathcal{Y}}_N$ (Brockwell and Davis, 1987; Shumway, 1988). This is found by recursive use of the standard formula $\mathbb{P}(A, B) = \mathbb{P}(A|B)\mathbb{P}(B)$ for events A and B. The likelihood for $\boldsymbol{\theta}$ is a function of the observations, $\boldsymbol{\mathcal{Y}}_N$, and becomes

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}_N) = \mathbb{P}(\boldsymbol{Y}_N = \boldsymbol{y}_N | \boldsymbol{\mathcal{Y}}_{N-1}; \boldsymbol{\theta}) \cdot \ldots \cdot \mathbb{P}(\boldsymbol{Y}_0 = \boldsymbol{y}_0; \boldsymbol{\theta}). \tag{3.6}$$

The terms of (3.6) is recognised as the denominator of (3.2) and is therefore regarded as the reciprocal of the normalisation constant, ψ_j, in (3.3). Hence it is concluded that maximising the likelihood for $\boldsymbol{\theta}$ is equal to minimising the *one-step prediction errors*. The likelihood value for $\boldsymbol{\theta}$ is therefore given by

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}_N) = \prod_{j=1}^{N} \frac{1}{\psi_j}.$$

This result is very convenient, in that the likelihood for $\boldsymbol{\theta}$ is implicitly calculated in the filtering process and does not require further computation. For parameter estimation in practice it is convenient to work with the logarithm of the likelihood function (the log likelihood function) defined as

$$\ell(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}_N) = \log \mathcal{L}(\,\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}_N).$$

This avoids the numerical problems associated to the very small numbers in the computation of (3.6).

The maximum likelihood estimate (MLE), $\widehat{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$, is a function of the observations and is defined as

$$\widehat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}_N).$$

Asymptotically, $\widehat{\boldsymbol{\theta}}$ is unbiased, efficient (smallest variance) and Gaussian distributed. Further description of the ML estimation technique and its properties is found in e.g. Rao (1965).

3.4 Sampling a random track

Evaluating the geolocation result solely based on the marginal posterior distributions, given by (3.5), does not suffice for a complete description. In this context, sampling a track from the joint posterior distribution of all positions, is a relevant supplement that will aid in assessing the possible routes of the fish. A track from the joint posterior distribution may reveal information that is not immediately evident from the marginal posterior distributions.

Formally the joint posterior distribution for all positions (often referred to merely as the joint posterior distribution) is defined as

$$\mathbb{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{Y}}_N), \tag{3.7}$$

where $\boldsymbol{\xi} = [\boldsymbol{X}_0 = \boldsymbol{x}_0, \ldots, \boldsymbol{X}_j = \boldsymbol{x}_j]^T$ denotes a track given by the positions at all time steps.

Sampling from the joint posterior distribution is done recursively by applying Bayes' rule. The sampling scheme runs backwards in time, initialised by sampling the terminal position at τ_N from the distribution

$$\mathbb{P}(\boldsymbol{X}_N = \boldsymbol{x}_N|\boldsymbol{\mathcal{Y}}_N),$$

and thereby obtaining $\boldsymbol{x}_N^s$, the sampled terminal position.

The position preceding τ_{j+1} is sampled from

$$\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_N, \boldsymbol{X}_N = \boldsymbol{x}_N^s, \ldots, \boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}^s).$$

This can be rewritten by applying the *Markov property* and Bayes' rule to obtain

$$\begin{aligned} &\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_N, \boldsymbol{X}_N = \boldsymbol{x}_N^s, \ldots, \boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}^s) \\ &= \mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_j, \boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}^s) \\ &= \mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}^s|\boldsymbol{X}_j = \boldsymbol{x}_j)\frac{\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_j)}{\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}^s|\boldsymbol{\mathcal{Y}}_j)}. \end{aligned} \tag{3.8}$$

The formula (3.8) uses the reconstruction, $\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j|\boldsymbol{\mathcal{Y}}_j)$, and updates it with the information of the previous (in an iterative not temporal sense) sample point $\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}^s|\boldsymbol{X}_j = \boldsymbol{x}_j)$. The term, $\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}^s|\boldsymbol{\mathcal{Y}}_j)$, is considered a normalisation constant in the implementation that makes the distribution sum to one. The Markov assumption is essential to this sampling method that would otherwise require a more complex simultaneous sampling from the joint distribution.

3.5 Finding the Most Probable Track

Another perhaps more interesting representation of the joint posterior distribution is the Most Probable Track (MPT). Previous studies employing the Kalman

Filter (Sibert et al., 2003) or particle filter technique (Nielsen, 2004) have suggested using a track that connects the conditional mean of all time steps. In the linear framework of the Kalman filter, the choice is rational. However, the potential multi modal distributions of a particle filtering can produce erroneous tracks in a nonlinear environment, possibly locating the most probable position on dry land. An environment is termed “nonlinear” if it contains islands or shores that cause the Gaussianity assumption to be violated.

The joint posterior distribution is given by (3.7). The mode in this distribution is de facto the most probable of all possible tracks in the outcome space and entitled the Most Probable Track. The non-trivial task of finding this track is solved by application of the Viterbi algorithm (Viterbi, 1967; Viterbi, 2006). The algorithm was developed for information theory and deep space communication and have found wide applications most prominently in speech recognition. It was later shown to be a computationally efficient technique for determining the most probable sequence in a hidden Markov model (Forney, 1973).

As it is a novel approach to track estimation in a geolocation context, the technique is here presented in some detail.

A track ending at a given position $\widetilde{\boldsymbol{x}}$ at time τ_j, is written

$$\boldsymbol{\xi}(\widetilde{\boldsymbol{x}}_j) = [\boldsymbol{X}_0 = \boldsymbol{x}_0, \dots, \boldsymbol{X}_j = \widetilde{\boldsymbol{x}}_j].$$

Furthermore the *branch metric* is defined

$$\mathcal{B}(\boldsymbol{x}_{j-1}, \boldsymbol{x}_j; \boldsymbol{y}_j) = \underbrace{\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j | \boldsymbol{X}_{j-1} = \boldsymbol{x}_{j-1})}_{\text{transition probability}} \underbrace{\mathcal{L}(\boldsymbol{Y}_j = \boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j)}_{\text{observational likelihood}},$$

as a product of the likelihood for the observation $\boldsymbol{y}_j$, given the new position $\boldsymbol{x}_j$ and the transition probability for jumping from $\boldsymbol{x}_{j-1}$ to $\boldsymbol{x}_j$.

A likelihood measure for a track is defined as

$$\mathcal{L}[\boldsymbol{\xi}(\widetilde{\boldsymbol{x}}_j)] = \mathcal{B}(\boldsymbol{x}_{j-1}, \widetilde{\boldsymbol{x}}_j; \boldsymbol{y}_j) \prod_{k=1}^{j-1} \mathcal{B}(\boldsymbol{x}_{k-1}, \boldsymbol{x}_k; \boldsymbol{y}_k),$$

which is proportional to the probability of $\boldsymbol{\xi}(\widetilde{\boldsymbol{x}}_j)$.

The *state metric* at a position, $\widetilde{\boldsymbol{x}}$ at time τ_j, is given by

$$\mathcal{S}(\widetilde{\boldsymbol{x}}_j) = \max_{\widetilde{\boldsymbol{x}}_j} \mathcal{L}[\boldsymbol{\xi}(\widetilde{\boldsymbol{x}}_j)],$$

meaning the likelihood of the most probable track leading to $\widetilde{\boldsymbol{x}}_j$.

As a consequence of the *Markov property* the maximisation can be done recursively

$$\mathcal{S}(\widetilde{\boldsymbol{x}}_j) = \max_{\widetilde{\boldsymbol{x}}_{j-1}} \{\mathcal{S}(\widetilde{\boldsymbol{x}}_{j-1})\mathcal{B}(\widetilde{\boldsymbol{x}}_{j-1}, \widetilde{\boldsymbol{x}}_j; \boldsymbol{y}_j)\}.$$

The algorithm sequentially finds the current state metric by maximising the product of the previous state metric and the attached branch metric. For all $\widetilde{\boldsymbol{x}}_j$, $\mathcal{S}(\widetilde{\boldsymbol{x}}_j)$ contains the likelihood of the most probable track leading to $\widetilde{\boldsymbol{x}}_j$. Logging the most probable track, in each recursion, for each $\widetilde{\boldsymbol{x}}_j$ is a simple way to obtain the Most Probable Track, $\widehat{\boldsymbol{\xi}}$. The track leading to $\widehat{\boldsymbol{x}}_N = \arg\max_{\boldsymbol{x}_N} \mathcal{S}(\boldsymbol{x}_N)$ is $\widehat{\boldsymbol{\xi}}$.

Chapter 4

Geolocation of simulated fish

Before applying the filter methods described in Chapter 3 to a real dataset, the performance of the method is investigated in a simulation study. Emphasis is put on clarification of bias and uncertainty on the parameter estimation of D. The track representation of a geolocation result is evaluated by determining a mean track, a mode track and the Most Probable Track according to Section 3.5.

The simulation model is not an attempt to make an entirely realistic model of reality, it is merely a mean to assess and illustrate the properties of the filtering technique.

4.1 Construction of the model

The simulated geolocation will rely on depth measurements in an artificial domain. Environmental variables such as light, temperature and tidal information add a complexity to the model that is unwanted in this simulation and are therefore not included.

4.1.1 Bathymetry

The domain is constructed in close resemblance to the bathymetry of a real life situation e.g. with islands and varying depth gradient in order to achieve a non-trivial simulation.

The artificial lake that is used for most simulations is constructed in MATLAB based on the surface created by the command `peaks`. This is modified by means of simple arithmetic operations to become a lake as shown in Figure 4.1. The lake is discretised with 101 × 101 grid points.

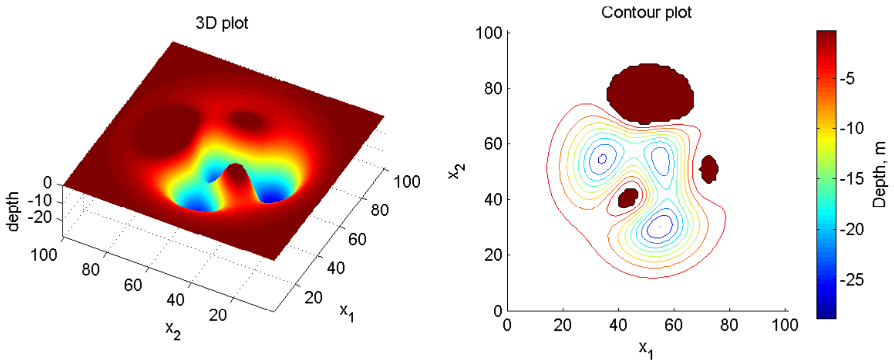

Figure 4.1: *Bathymetry for simulation.*

The domain contains three small islands that serves as a test for the handling of nonlinearities. The lake has very shallow areas close to the border of the domain and deeper areas near the middle. These gradients in depth and their effect on the uncertainty of the geolocation will be revealed from this bathymetry as well.

4.1.2 Simulation of random walk in the domain

The movement model for the fish is a two dimensional homogeneous random walk with transition probabilities according to (2.10) for each coordinate direction. The value of r is assumed to be constant in time. With this scheme each coordinate can maximally increase with h over a time step of k. For this simulation, the values of the increment parameters for time and space are for simplicity defined as

$$\begin{aligned} k &= 1, \\ h &= 1. \end{aligned}$$

The nonlinearities of the domain such as islands must be accounted for in the random walk simulation as it is not allowed for the fish to go ashore. This is handled by rejection sampling of the position.

4.1.3 Model for depth measurements

The fish is assumed to be demersal (at the sea bed) at all times resulting in an observed depth

$$Z_j = \mathcal{D}_j + E_j,$$

where $\mathcal{D}_j$ is the true depth extracted directly from the bathymetry at the simulated position, and E_j is the measurement error that is uniformly distributed with zero mean and range $[-\delta; +\delta]$.

Figure 4.2 shows an example of a simulated time series of depth measurements with $\delta = 2$.

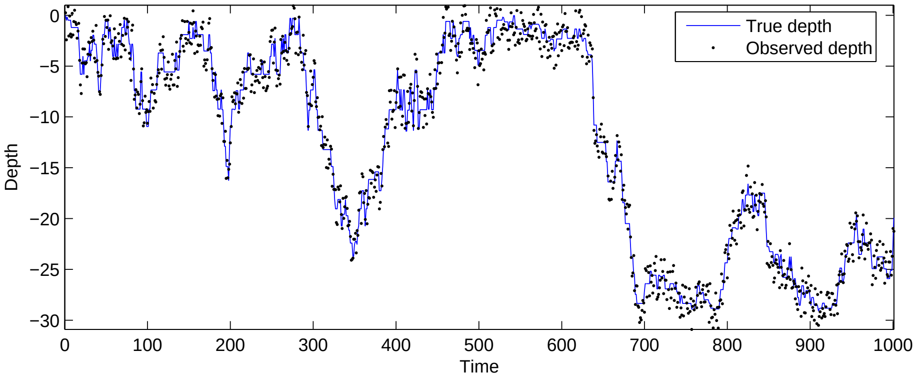

Figure 4.2: *Example of a simulated time series of depth measurements and the true depth. Note that the axes have no unit as they are measured in the standard space and time units h and k respectively.*

4.2 Likelihood estimation of D

An influential parameter of the simulation model is the diffusivity, D. It is related to the maximal swimming speed of the fish and generally adds to the understanding of the behaviour of the species. This "biomarker" may enable direct comparison of individuals.

The performance of the ML estimator is evaluated with respect to the following subjects.

- Validity of the likelihood ratio confidence interval for D, i.e. is the likelihood function of D behaving according to theory?
- Bias on the ML estimator via a t-test of empirical mean.
- Empirical standard deviation and the standard deviation of a single estimate computed from the Fisher Information via an F-test.

From the simulation model, 100 estimates of D were generated and used as dataset for the tests. The estimates were generated based on 500 step simulations with an observation uncertainty of $\delta = 4$.

In the remainder of this section the short notation $\ell(D)$ is used instead of $\ell(\boldsymbol{\theta};\, \boldsymbol{\mathcal{Y}}_N)$, where $\boldsymbol{\theta} = D$.

4.2.1 Likelihood Ratio tests

For each of the 100 estimates, a 95% confidence interval is constructed based on a Likelihood Ratio Test (Wasserman, 2005). The test is defined with the two hypotheses

$$H_0\colon\ D_0 = \widehat{D}, \qquad \text{versus} \qquad H_1\colon\ D_0 \neq \widehat{D},$$

where D_0 is the hypothesised (true) value of D, and $\widehat{D}$ is its ML estimate. The likelihood ratio test statistic is computed in the following way

$$Z_{LR} = 2\ell(\widehat{D}) - 2\ell(D_0). \tag{4.1}$$

Under H_0, Z_{LR} is asymptotically χ^2-distributed with one degree of freedom (one parameter). Based on (4.1) it is possible to create a 95% confidence interval for the parameter D

$$\begin{aligned} && \chi^2_{0.95}(1) &= 2\ell(\widehat{D}) - 2\ell(D_0) \\ \Leftrightarrow && \ell(D_0) &= \ell(\widehat{D}) - 0.5\chi^2_{0.95}(1). \end{aligned} \tag{4.2}$$

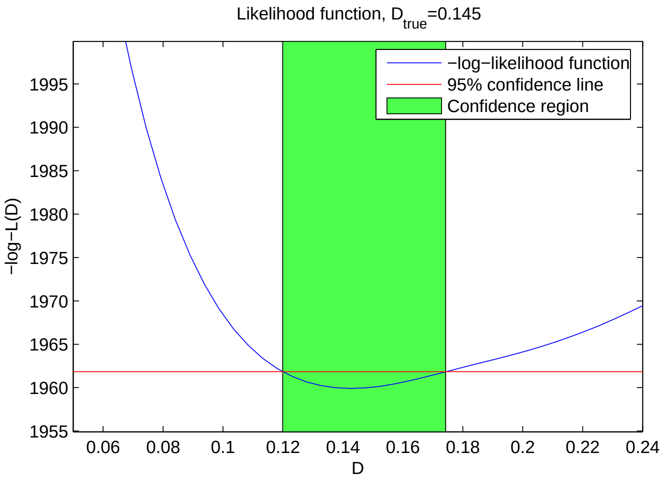

Figure 4.3: *Example of a negative log-likelihood function for the diffusivity parameter D.*

Equation (4.2) has two solutions for D_0 which can be seen graphically in the example in Figure 4.3, where a line is drawn at the likelihood value of (4.2).

For this example the true value, $D = 0.145$, is inside the confidence limits and consequently H_0 cannot be rejected.

4.2.1.1 Conclusion to Likelihood Ratio tests

The test was conducted by simulating 100 confidence intervals for D. Analysis of the results showed that 6 did not contain the true value of D. According to the significance level, $\alpha = 0.05$, it was expected that 5 of the 100 tests rejected H_0. The deviation from the expected number is small and acceptable for application purposes. No strong evidence of bias in the ML estimator could be found.

4.2.2 Test of empirical mean

The 100 estimates were evenly distributed around the empirical mean of 0.14533 which is shown in a histogram in Figure 4.4. The asymptotic Gaussianity of the ML estimate calls for a t-test (Madsen and Holst, 2000) to assess whether it can be rejected that the empirical mean of $\overline{D} = 0.14533$ is equal to the true value

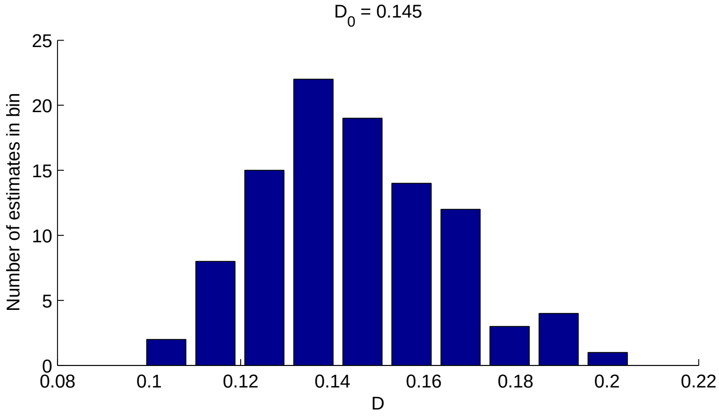

Figure 4.4: *Histogram of 100 simulated estimates of D.*

$D_0 = 0.145$. The problem is formulated as hypotheses

$$H_0\colon\ \overline{D} = D_0, \qquad \text{versus} \qquad H_1\colon\ \overline{D} \neq D_0,$$

The test statistic is

$$Z_t = \frac{\overline{D} - D_0}{\overline{s}/\sqrt{n}} = 0.1609,$$

where $\overline{s}$ is the empirical standard deviation and $n = 100$ is the number of estimates. The critical region at the $\alpha = 0.05$ level of significance is $\{z_t < -1.984 \vee z_t > 1.984\}$ implying that H_0 can not be rejected.

This test result shows that there is no provable indication of bias on the ML estimate of D.

4.2.3 Test of empirical variance

The variance of $\widehat{D}$ is determined in two ways. By the inverse of the observed Fisher information of $\widehat{D}$, and by computing an empirical variance of the results from numerous repetitions of the estimation procedure. The latter method is the definition of the variance concept and therefore converges to the true variance as the number of repetitions approach infinity. For the parameter D, the observed Fisher information corresponds to the second derivative (curvature) of the likelihood function. Comparison of this variance estimate with the empirical is done in an F-test (Madsen and Holst, 2000) with the hypotheses

$$H_0\colon\ \overline{s}^2 = \widetilde{s}^2, \qquad \text{versus} \qquad H_1\colon\ \overline{s}^2 \neq \widetilde{s}^2,$$

where $\overline{s}^2$ is the empirical variance of the 100 estimates of D and $\widetilde{s}^2$ is the mean of the individual variance estimates, s^2, of D.

The test statistic for the F-test is given by

$$Z_F = \frac{\overline{s}^2}{\widetilde{s}^2},$$

that under H_0 has the distribution $Z_F \sim F(99, \infty)$.

With the estimated values $\overline{s}^2 = 0.02026^2$ and $\widetilde{s}^2 = 0.01873^2$ the test statistic becomes

$$Z_F = 1.1701.$$

H_0 is rejected at a $\alpha = 0.05$ level of significance if Z_F is in the critical region: $\{z_F < 0.769 \vee z_F > 1.30\}$. The test statistic proves to be insignificant hence it can not be rejected that the two variances are equal.

This result is taken as argument for using the Fisher information to estimate the variance of $\widehat{D}$ in cases where it is needed.

4.3 Experimenting with the model

The filter is implemented in the MATLAB v. 7.0 software package with emphasis on functionality and ease of implementation rather than speed.

4.3.1 Brownian bridge

Validation of the simulation is done by considering a simple situation without observations of depth

$$\boldsymbol{\mathcal{Y}}_N = [\boldsymbol{x}_\dagger, \boldsymbol{x}_\ddagger]^T. \tag{4.3}$$

For this situation the resulting joint posterior distribution can be computed analytically and is known as a Brownian bridge.

It is known that the position, X_j, of a fish performing Brownian motion in one dimension given the initial position x_0, has a Gaussian distribution, $\mathcal{N}(x_0, 2Djk)$, according to (2.4). The recapture position, X_N, must then have the distribution $\mathcal{N}(x_0, 2DNk)$. The conditional distribution of X_j given X_N is obtained by conditioning in the joint distribution of the two, which is

$$\begin{bmatrix} X_j \\ X_N \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x_0 \\ x_0 \end{bmatrix}, \begin{bmatrix} 2Djk & 2Djk \\ 2Djk & 2DNk \end{bmatrix} \right),$$

where $\mathsf{Cov}(X_j, X_N) = \mathsf{Cov}(X_j, X_j + (X_N - X_j)) = \mathsf{Cov}(X_j, X_j) = 2Djk$ because X_j is disjoint from $X_j - X_N$ and therefore $\mathsf{Cov}(X_j, X_N - X_j) = 0$. The parameters in the conditional distribution becomes

$$\mathbb{E}(X_j | X_N) = x_0 + \frac{j}{N}(X_N - x_0), \tag{4.4}$$

$$\mathbb{V}(X_j | X_N) = 2Djk \left(1 - \frac{j}{N}\right), \tag{4.5}$$

according to standard formula for conditioning in a Gaussian distribution. The conditional of a Gaussian distribution is also Gaussian so $X_j | X_N$ is Gaussian distributed with mean given by (4.4) and variance given by (4.5), when $0 \leq j \leq N$.

A simulation of the brownian bridge with $N = 1000$ yielded the result shown in Figure 4.5.

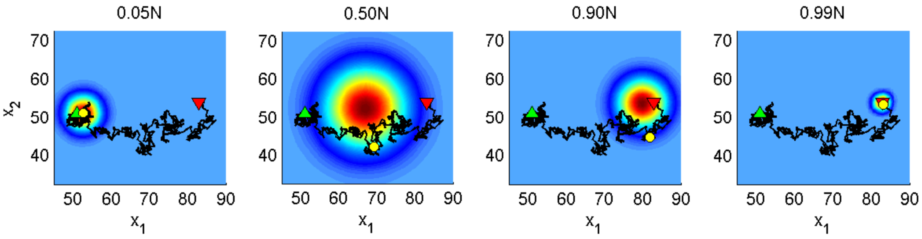

Figure 4.5: *The simulation behaves as a Brownian bridge when the depth is equal over the domain. The color map denote the probability of the position. Blue is least probable, red is most probable. Green triangle: release position. Red triangle: Recapture position. Yellow circle: The simulated position at the current time point.*

The simulation results follow the analytical, with the mean moving linearly from the release position to the recapture position. The variance increases until $j = 0.5N$ where it tops and afterwards reduces to zero at $j = N$.

This example illustrates the importance of the smoothing step. In general, a solution consisting of predictions relying purely on past observations is not constrained by the recapture position and becomes significantly more uncertain in time.

4.3.2 Tracks

This subsection presents and evaluates the various tracks that may be used to illustrate the result of a geolocation. The tracks considered are

- A track connecting the mean of the marginal posterior distributions at each time instant, termed a mean track.
- A track connecting the mode of the marginal posterior distributions at each time instant, termed a mode track.
- The Most Probable Track.

These are compared to the true simulated track. A high diffusivity was chosen, to simulate an active fish.

4.3.2.1 Brownian bridge

A simulation of 25 steps on a flat bathymetry (equal depth over domain) was performed, along with an estimation of the mean track and the MPT. The observation vector reduces to (4.3) as depth measurements hold no useful information (no depth gradient in bathymetry). The estimated posterior distribution behaves as a Brownian bridge as in Subsection 4.3.1.

The mean track, shown in Figure 4.6, follows for each coordinate the theoretical expression in (4.4), simply a straight line from the initial position to the terminal position.

In this example there exists many tracks that have the highest obtainable probability. Figure 4.6 shows one of the possible MPTs arbitrarily chosen by the algorithm. In this case rounding the mean track to closest integer coordinates also gives a MPT. All tracks, having 4 positive jumps in the x_1-direction, 2 negative in the x_2-direction and 19 zero jumps, have equal probability and are MPTs.

This test confirms that the Viterbi algorithm finds a track that, due to the simplicity of the problem, is known to be a MPT.

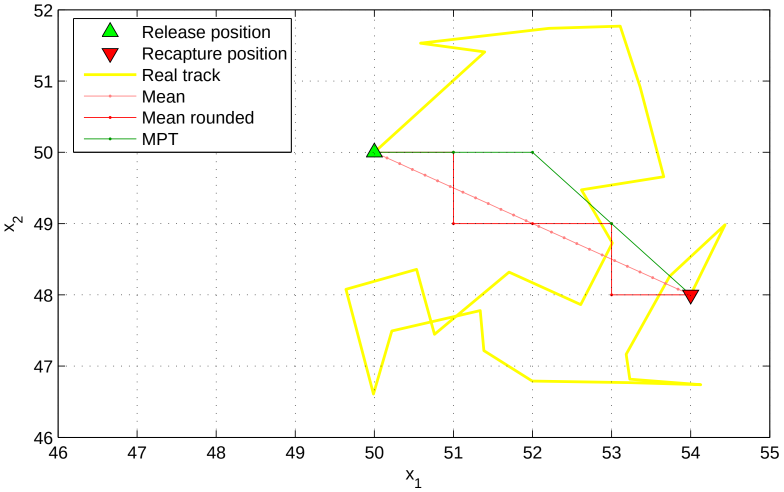

Figure 4.6: *Simulation result of a random 25 step track on a flat bathymetry along with estimated mean track and MPT.*

4.3.2.2 Linear environment

A track of 200 steps was simulated on the `peaks` bathymetry and plotted in Figure 4.7 along with the estimated mean track, mode track and MPT. The fish movement is only moderately influenced by the islands resulting in a track that is well estimated by all three estimators. It is noted however that the mode track occasionally shows an excessive erratic behaviour in contrast to the mean track that mostly has small jumps.

4.3.2.3 Nonlinear environment

The second simulation generated a 250 step track of a fish swimming in a nonlinear environment, see Figure 4.8. Very conspicuous is the behaviour of the mean and mode track that yield erroneous estimates when the fish swims close to the island. At the website, www.student.dtu.dk/~s002087 and on the enclosed CD-ROM, is shown the “Animated Marginal Posterior Distribution” for this simulation. When inspecting an AMPD it should be borne in mind that the color scale is not constant in time. The bimodal distributions of the marginals, result in the mode track jumping between two competing suprema repeatedly, causing the estimated track to move across the island. The mean track esti-

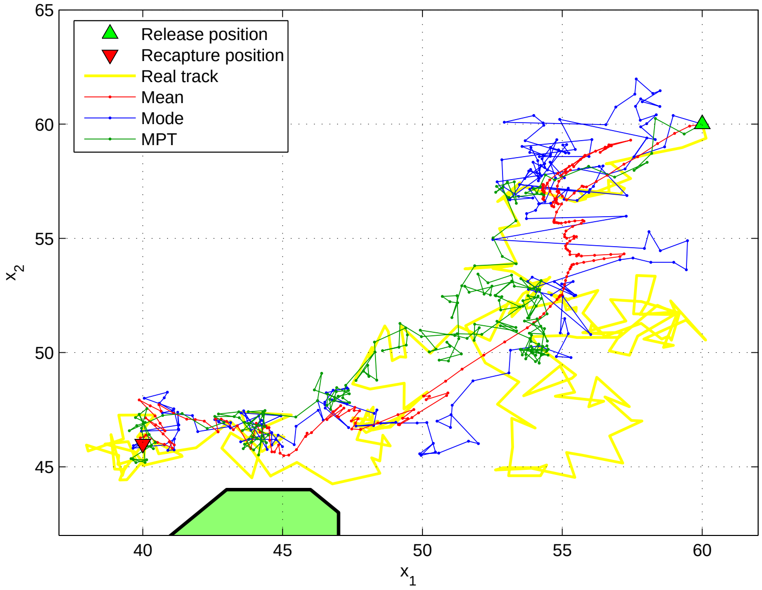

Figure 4.7: *Estimated tracks for a simulated fish (200 steps) with little influence from islands. All track estimates are quite accurate and follows the general trend of the simulated track.*

mates the fish to be located on the island and proves to be very misleading in a nonlinear environment. The MPT follows the general trend of the simulated track.

Sampling of 1000 random tracks from the joint posterior distribution gave the estimate that the fish moved east of the island with 64% probability. It is questions of this type that a sample of multiple random track can clarify.

4.3.3 Influence of δ

The uncertainty of the observations is one of the main influences on the uncertainty of the geolocation. This is illustrated in Figure 4.9.

The variance of the distribution clearly diminishes as δ decreases. At $\delta = 0.1$ the position of the fish is known without uncertainty except for the resolution of the discretisation. The effect is especially evident in the top row of Figure 4.9,

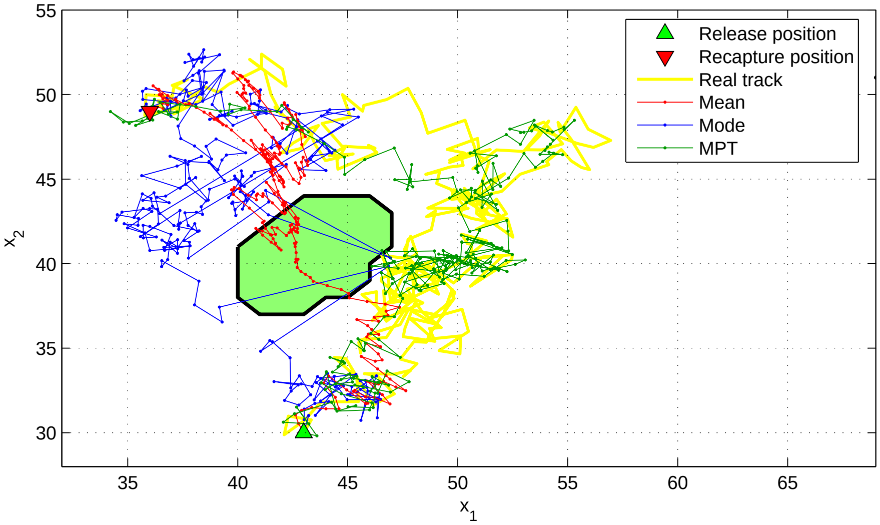

Figure 4.8: *Estimated tracks for a simulated fish (250 steps) swimming near an island. The mean track estimates positions on dry land, the mode track indicates crossing dry land, whereas the MPT shows a likely general trend.*

where the fish is swimming in a shallow area with little variation in the sea floor depth. This is contrary to the bottom row where quite precise geolocations are obtained even for large δ.

The results confirm what is fairly intuitive and stress that the power of the geolocator (depth observations) depends on its spatial gradient in the domain. When the data collection is planned this is an important note to keep in mind, especially for choice of DTS type. Areas such as the Baltic Sea contains large gradients of salinity but almost no tidal variation, in contrast to the North Sea that has the opposite properties.

4.4 Conclusion summary of simulation study

The simulation study illustrated several important aspects of the filter and estimation procedure.

The Brownian bridge example showed the importance of the smoothing step and how it restrains the uncertainty of the geolocation considerably. The maximum likelihood estimation of D could not be proved to be biased. This was

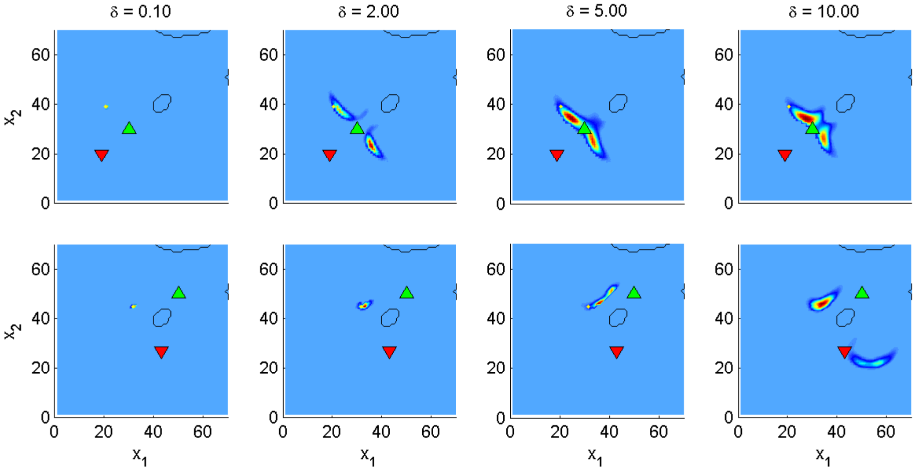

Figure 4.9: *Simulation of 500 steps here shown at $j = 250$ with various values of $\delta = [0.1, 2, 5, 10]$. Explanation of markers: Green: Release position, Yellow: Position at time of geolocation, Red: Recapture position. Top row shows the geolocation in a shallow area (little depth variation) near the border of the domain. The bottom row shows the geolocation near a larger depth gradient.*

concluded based on 100 likelihood ratio tests of $\widehat{D}$ and in a t-test of the empirical mean. The empirical variance of the 100 parameter estimates and the variance on single estimates computed from the observed Fisher information showed concurrence in an F-test. Finally the track comparison indicated that the MPT is the most rational representation of the joint posterior distribution compared to a track of the mean or mode of the marginal posterior distributions.

Part II

Geolocation of North Sea fish

CHAPTER 5

Introduction to tidal based geolocation

The method of tidal based geolocation relies on the spatial variation in tide. The complex amphidromic system of the North Sea make up an environment well-suited for this method. The North Sea habitat is described in Section 5.1.

The other main requirement for tidal based geolocation is a depth record from a DST. It is essential that the fish of interest is a demersal species that habitually visits the sea bed for a longer period of time (several hours). Being stationary causes the mounted DST to record the oscillating depth following the tidal variations of the sea thereby identifying its position. Description of the DSTs used for this study is given in Section 5.2

In short the technique compares the observed tidal signal from the DST with a prediction at a given position computed by a numerical model. The quality of the fit between the two signals decides the likelihood for the observation given this position. This allows for a rather unique determination of the position. The method is elaborated in Chapter 6.

The DST datasets and environmental databases for this study were provided by CEFAS. The time series considered in the thesis are plotted in separate MATLAB `fig`-files located on the enclosed CD-ROM.

Important notice: The files on the CD-ROM are not to be distributed without permission from CEFAS.

5.1 Habitat of the North Sea

Geolocation relies on gradients in environmental variables measured by the DST. The amphidromic system of the North Sea has proven to be a very powerful geolocator because of the great variations in tide especially near the English channel.

The North Sea is somewhat shallow, with depths in the range 30-70 m in most places. Slightly deeper areas are located to the north with depths in the range 150-200 m, see Figure 5.1 top left. Though the overall depth range is moderate some areas contain significant local variations with holes of e.g. 75 m in otherwise shallow areas. These areas make up great environments for fish to reside and therefore obvious places to do commercial fishing.

In contrast to the Baltic Sea, tidal variations are very pronounced in the North Sea and especially in the English Channel. The tidal wave originates in the Atlantic Ocean by the gravitational drag of the moon and sun and propagates through the English Channel and north of the British Isles into the North Sea. The tide has the largest amplitude near the shores and in bays and gulfs, indicated in Figure 5.1 bottom left. The bathymetry of the North Sea (including coast lines) results in a tidal system with two amphidromic points that are located in the south and eastern parts of the sea, see Figure 5.1 bottom right or left. These are areas where almost no tidal variation occurs.

Another important environmental variable of the North Sea is the temperature that shows much temporal and spatial variation. The shallow areas are subject to the largest temperature range over a year whereas the deep areas have only minor variations at the sea bed level. In the winter period the water column is mixed with a constant temperature in the range 6-9°C. In the summer months a vertical gradient exist with 6-9°C at the sea bed and up till 25°C at the sea surface near the shores.

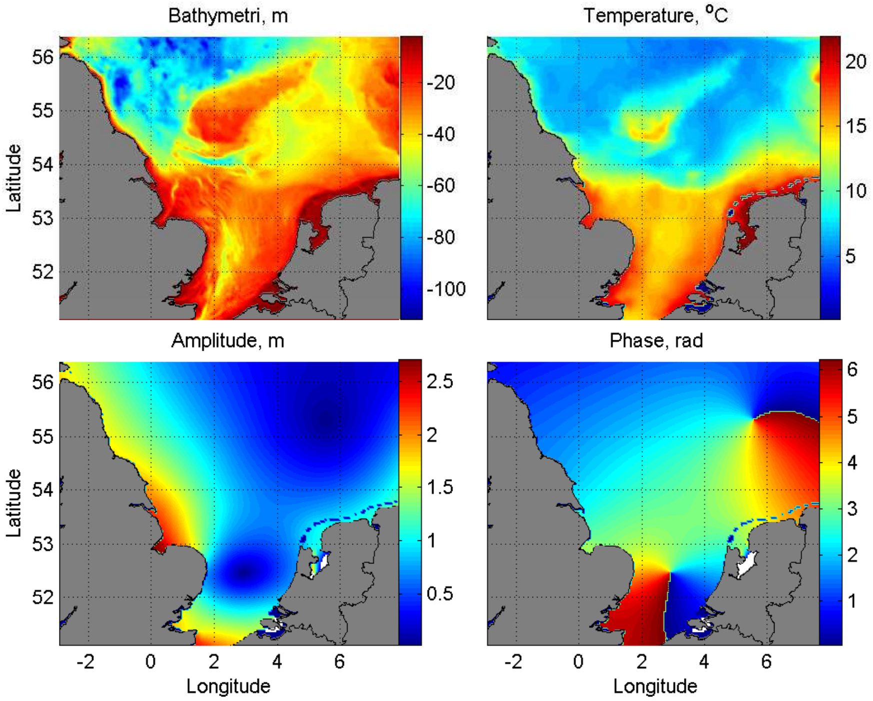

Figure 5.1: *Habitat of the North Sea. Top left: Bathymetry of the North Sea. Top right: Sea bed temperature the 18th of July 2001 in °C. Bottom left: Amplitude of the M2 tidal constituent i metres. Bottom right: Phase of the M2 tidal constituent in radians.*

5.2 Data Storage Tags

It has within the last 10-15 years become possible to construct DSTs in a size that can satisfactorily be applied to fish of length 50-70 cm, such as cod. There exists at this point various types of DSTs. The ones used by CEFAS to create the data analysed in this thesis are listed here in a short summary. In Figure 5.2 is shown examples of the tags.

5.2.1 Star-oddi centi

Manufactured by Star-Oddi. Due to its size is most suited for external tagging. Dimensions are 46 × 15 mm (length × diameter) and weighs 19 g in air and 12 g in water. The resolution of depth measurements on the tag depends on its full measuring range but lies at approximately 0.03-0.075 m. The accuracy of the measured depth is ±(0.4-1) m. The temperature range is -1°C to 40°C with a resolution of 0.032°C and an accuracy of ±0.1°C.

5.2.2 Star-oddi milli

Manufactured by Star-Oddi. Somewhat similar to Star-oddi centi but smaller. Can be used for both internal and external tagging. Dimensions are 38.4×12.5 mm (length × diameter) and weighs 9.2 g in air and 5 g in water. The resolution of depth measurements on the tag depends on its full measuring range but lies at approximately 0.03-0.09 m. The accuracy of the measured depth is ±(0.4-1.2) m. The temperature range is -1°C to 40°C with a resolution of 0.032°C and an accuracy of ±0.1°C.

5.2.3 LTD 1200 (Mk3C)

An older tag from 2001 at the time manufactured by LOTEK but now by CEFAS. Dimensions are 57×23 mm (length × diameter) and weighs 17 g in air and 1.8 g in water. The resolution of depth measurements on the tag is approximately 0.05 m. The accuracy of the measured depth is ±1 m. The temperature range is 0°C to 30°C with a resolution of 0.05°C and an accuracy of ±0.1°C.

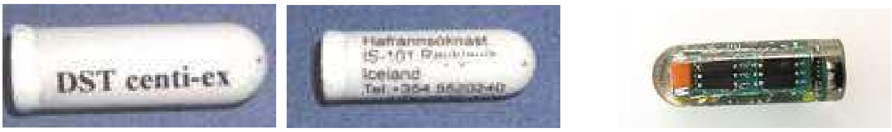

Figure 5.2: *Various types of DSTs used for geolocation. Left: Star-oddi centi, Center: Star-oddi milli, Right: LDT 1110 (similar to 1200).*

Experience with the various tags say that the accuracy of the tag should be interpreted as a bias on the measurements that is “defined” when the tag is manufactured. This bias is constant for the life time of the tag and the only uncertainty of the tag lies in its resolution.

Examples of depth measurements from a DST is shown in the following chapter in Figure 6.1.

Chapter 6

Statistical analysis of depth record

This chapter explains how the observational likelihood is obtained from the DST depth record. The subject has many aspects that are decided upon based on objective statistical analysis when possible.

Recall from Section 3.1 that the observational likelihood is written as

$$\mathcal{L}(\boldsymbol{Y}_j = \boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j),$$

that is the likelihood of observing $\boldsymbol{y}_j$ given the position $\boldsymbol{x}_j$. The evaluation of the likelihood varies depending on the type of information found in the depth record at the time point, τ_j. Either the fish rests at the sea bed and records a tidal signal or it performs a behaviour that does not record a tidal signal of sufficient quality. In Section 6.1 details are given as to how tidal information in the depth record is detected and extracted. This results in a classification algorithm for the entire set of depth observations.

When the depth record has been successfully classified into tidal/non-tidal intervals, the observational likelihood can be determined. For the case "tidal information available" the observations $\boldsymbol{Y}_j = \boldsymbol{y}_j$ is assumed to follow an m-dimensional Gaussian distribution

$$\boldsymbol{Y}_j \sim \mathcal{N}_m\big(\widehat{\boldsymbol{z}}_j(\boldsymbol{x}_j), \boldsymbol{\Sigma}(\boldsymbol{x}_j)\big), \tag{6.1}$$

where m is the number of observations, $\widehat{\boldsymbol{y}}_j(\boldsymbol{x}_j)$ is the database prediction and $\boldsymbol{\Sigma}(\boldsymbol{x}_j)$ is the covariance matrix at the position $\boldsymbol{x}_j$. Section 6.2 describes how the database prediction of the tide given the position is calculated. The structure of the covariance matrix is estimated by analysis of separate contributions in Sections 6.3, 6.4 and 6.5. More specifically, Section 6.3 deals with white noise, Section 6.4 assesses the database resolution error and Section 6.5 investigates the error that arise from small scale movements of the fish.

The results are summarised in Section 6.6 which explicitly states how the observational likelihood is determined for observations with or without tidal information.

6.1 Extraction of tidal information from data

Many aspects must be taken into consideration as to how tidal information can be extracted efficiently from the measured time series of pressure. The characteristic wave form caused by the tide is relatively easy detectable by eye, see Figure 6.1, but comes in many forms varying both across the life of a single individual and between individuals.

Figure 6.1.1 shows a smooth tidal signal where the fish is resting at the sea floor for a longer period without any disturbances. For cod, a tidal signal this clean is rarely seen for periods longer than 24 hours.

In Figure 6.1.2 the tidal signal is very evident but perturbed with noise possibly due to small scale foraging behaviour. This type of disturbances can also be due to environmental conditions such as storms or currents.

Figure 6.1.3 shows a stationary fish that occasionally makes small excursions up into the water column either for foraging or in some cases for relocation in the surroundings.

Figure 6.1.4 is a more extreme version of Figure 6.1.3, where the cod is active during the night time and rests in the daylight period. This kind of behaviour is also noted in Righton et al. (2000). The behavioural pattern could also be an indication of tidal stream transport where the fish swims along the tidal wave and obtains a swimming speed that would not otherwise be possible.

A time series of depth is written $\boldsymbol{z} = [z_0, \dots, z_n]$ (depth measurements are given by the negative water column height) corresponding to the time vector $\boldsymbol{t} = [t_0, \dots, t_n]^T$. The examined data were sampled at varying rates but all

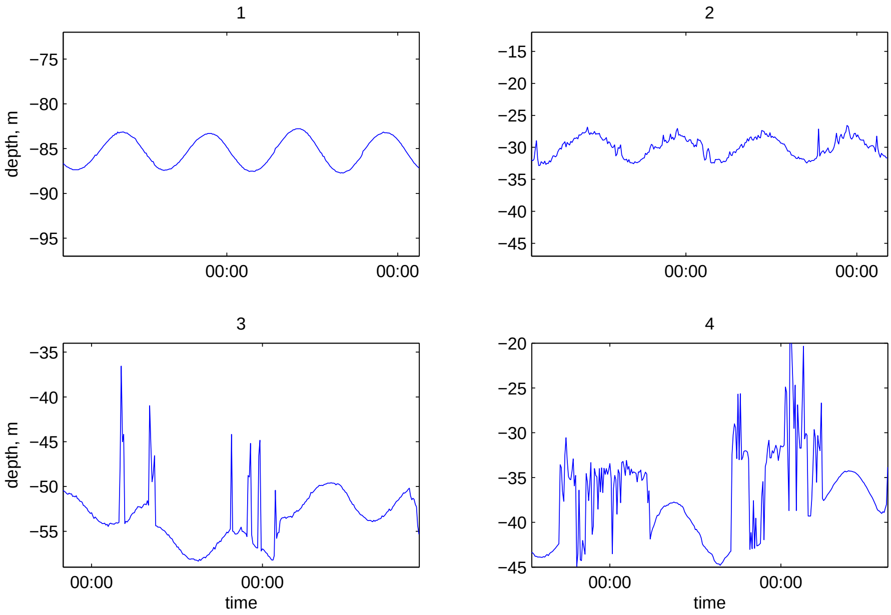

Figure 6.1: *Some types of tidal information all found in tag #2255. See text for description.*

converted to the standard sample rate of 10 minutes i.e.

$$t_{i+1} - t_i = 10 \text{ minutes},$$

which is used throughout the remainder of the thesis if not stated otherwise.

6.1.1 Definition of the applied linear model

Creating an efficient algorithm that can detect and extract all types of tidal information is an extensive signal processing and curve fitting task that exceeds the scope of this thesis. Instead a simple method with reasonable efficiency is chosen inspired by the one described in Hunter et al. (2003).

A set of observations, $\boldsymbol{z}_i = [z_i, \ldots, z_{i+m}]$, is extracted from $\boldsymbol{z}$. The observations are assumed to follow a linear model on the form

$$\boldsymbol{Z}_i = \boldsymbol{w}_i \boldsymbol{\beta}_i + \boldsymbol{E}_i, \tag{6.2}$$

where $\boldsymbol{E}_i$ is a Gaussian white noise error, $\boldsymbol{\beta}_i = [a_i \ b_i \ c_i]^T$ is the parameter vector and

$$\boldsymbol{w}_i = \begin{bmatrix} 1 & \cos(\omega t_i) & \sin(\omega t_i) \\ \vdots & \vdots & \vdots \\ 1 & \cos(\omega t_{i+m}) & \sin(\omega t_{i+m}) \end{bmatrix}, \tag{6.3}$$

is the design matrix where ω is the angular frequency. The value of ω is in principle unknown and time dependent because $\boldsymbol{z}_i$ is a superposition of all tidal constituents that have varying frequencies. It is here assumed that $\omega = 12.14$ rad/day, equal to the angular frequency of the dominating tidal constituent, M2. See Section 6.2 for explanation of tidal constituents.

The maximum likelihood estimate of $\boldsymbol{\beta}_i$ is found by solving the normal equations

$$\widehat{\boldsymbol{\beta}}_i = (\boldsymbol{w}_i^T \boldsymbol{w}_i)^{-1} \boldsymbol{w}_i^T \boldsymbol{z}_i.$$

The model fit yields various summary statistics that can be used to evaluate if the model matches the data and is suited for geolocation.

The extraction procedure iterates by sliding a window of m data points across the data and collecting for each ten minute interval the relevant summary statistics. With this information, appropriate criteria can be set up in order to determine intervals where the fish has dwelled at the sea bed. If the interval is accepted as a tidal signal, it is stored and used later for comparison with the database prediction. It is crucial that the tidal extraction algorithm does not falsely identify a tidal signal as this will lead to very wrong geolocations and maybe even terminate the process.

The value of m has great influence on the performance of the algorithm. Cod are rarely at the sea bed for a straight 24 hour period and even an m-value of 108 (18 hours) is probably too long for most cod and will certainly miss some intervals with tidal data. On the other hand a too short interval has a higher probability of misclassification and therefore large uncertainties will be attached to intervals of correctly classified tidal information.

After some experimentation it was found that $m = 60$ was a good choice of interval length, corresponding to 10 hours.

6.1.2 Classification of depth record

The following three summary statistics were used to classify the recorded signal in intervals that contained a tidal pattern and in ones that did not contain a tidal pattern.

6.1.2.1 The standard error of the fit

This is the root mean square of the residuals ($rmse$) given by

$$S = \sqrt{\frac{1}{m-p}\sum_{k=i}^{i+m}(z_k - \widehat{z}_k)^2},$$

where p is the number of estimated parameters, in this case $p = 3$ and $\boldsymbol{w}_i\widehat{\boldsymbol{\beta}} = [\widehat{z}_i, \ldots, \widehat{z}_{i+m}]$ is the prediction of $\boldsymbol{z}_i$. This parameter measures the deviation between the observed and fitted curve, and will be large if the observed data does not conform to a sine wave.

The unit of S is that of z_i and therefore S dependents on the magnitude of the tidal range in $\boldsymbol{z}_i$ that varies over $\boldsymbol{z}$ because of the phenomena high high tides and low high tides. These are results of the shifting positive and negative interference between the many tidal constituents. It is therefore difficult to define a limit value for S that in all cases effectively separates a tidal signal from a non-tidal signal. An example of a classification based only on S is shown in Figure 6.2.

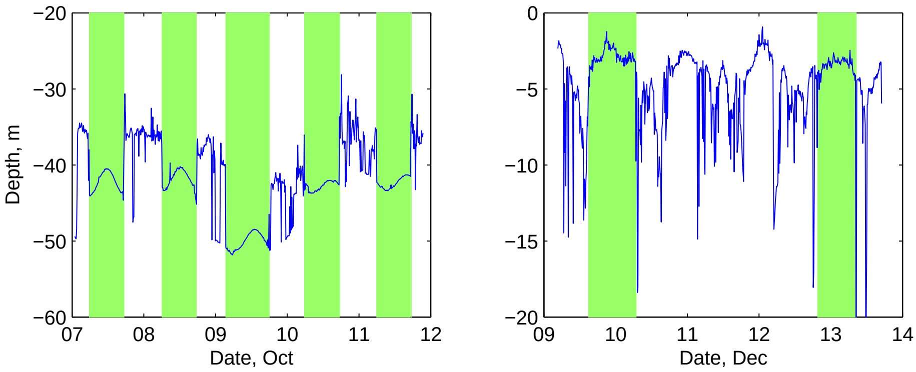

Figure 6.2: *Examples of tidal classification. Green intervals have a rmse below the limit 0.42 m. Left: Tidal information correctly classified n. Right: Tidal information falsely classified. Both from tag #2255.*

6.1.2.2 The R^2 of the fit

The intervals wrongly classified as tidal data by S can be constrained by a requirement on the R^2 as well. This, "coefficient of determination", denotes the proportion of the variance in the observations that is explained by the fitted curve. This should preferably be close to 1. In such a case the observed data has a smooth wave form. The R^2 statistic is independent of the magnitude of the tidal range.

6.1.2.3 The amplitude of the fit

A horizontal line is represented very well by the linear model in (6.2) i.e. with amplitudes close to zero and a in $\boldsymbol{\beta}$ equal to the value of the line. Such measurements arise either when the fish dwells close to an amphidromic point or when it swims at a constant depth. To avoid this plausible possibility for misclassification a constraint is put on the amplitude, $A = \sqrt{b^2 + c^2}$. For the tidal signal to be confidently used for geolocation A must be above some limit value.

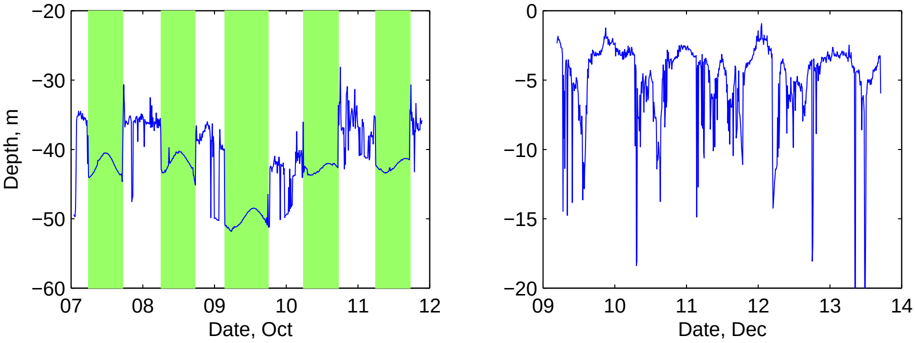

Figure 6.3: *Examples of tidal classification. Classified using the S, R^2 and the amplitude A. Compared to Figure 6.2 the right pane is now correctly classified.*

Figure 6.3 indicates the performance of the extraction algorithm. The limit values used was

$$\begin{aligned} S &< 0.42 \text{ m}, \\ R^2 &> 0.85, \\ A &> 0.6 \text{ m}. \end{aligned}$$

These values were applied to all DTS data used in this thesis.

6.1.3 Preprocessing and additional notation

For reasons of computational speed and numerical stability the filter runs in intervals of 24 hours. Based on this discretisation a new time vector is defined

$$\boldsymbol{\tau} = [\tau_0, \ldots, \tau_j, \ldots, \tau_N]^T.$$

See Figure 6.4 for a sketch of the time line. For $\boldsymbol{\tau}$ it holds that

$$\tau_0 = t_0, \qquad \text{and} \qquad \tau_N = t_n,$$

and that $\tau_j, j \in [1, \ldots, N-1]$ is the time one minute past midnight on the intervening Julian days. One temporal increment holds 144 observations and corresponds to

$$k = \tau_{j+1} - \tau_j = 24 \text{ hours}, \qquad j \in [1, \ldots, N-2].$$

When all useful tidal information in the time series is detected, an indicator array, $\boldsymbol{I} = [I_0, \ldots, I_j, \ldots, I_{N-1}]$, is created where

$$I_j = \begin{cases} 1 & \text{if tidal criteria are fulfilled for at least one of } [\boldsymbol{z}_{j,i}, \ldots, \boldsymbol{z}_{j,i+144}] \\ 0 & \text{otherwise} \end{cases},$$

where $j \in [0, \ldots, N-1]$. The best fit, defined as the one with the lowest standard error within each I_j, is found and chosen as representative for I_j. This is written more explicitly for the interval j as

$$\widehat{i}_j = \arg\min_i S_j(i), \quad \text{s.t.} \quad I_j = 1,$$

where $S_j(i)$ is the standard error of the i'th fit contained in the time interval $[\tau_j, \tau_{j+1}]$. The $\widehat{i}_j$'s are assembled in a vector

$$\widehat{\boldsymbol{i}} = [\widehat{i}_0, \ldots, \widehat{i}_{N-1}].$$

The depth record of the optimal fit is referred to as $\boldsymbol{z}_{j,\widehat{i}}$.

The reader is referred to Section 1.6 for a summary of the notation.

6.2 POL environment database

There exists tidal models that can predict the tide at any given time with some precision. Such models split the tide into a number of constituents that represent different modes with varying frequencies. A superposition of all modes

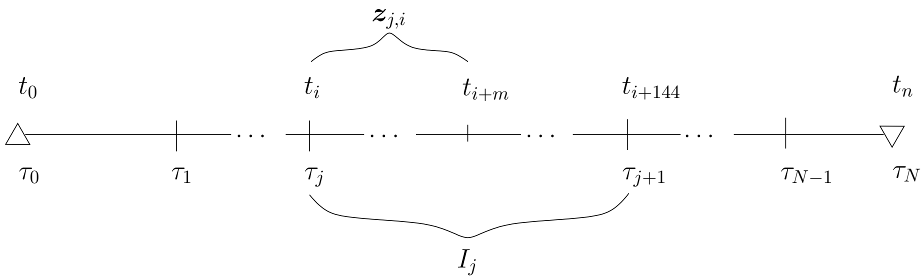

Figure 6.4: *Sketch of the time line for a DST time series. The fish is released at τ_0 and recaptured at τ_N.*

yields the resulting wave that approximates the one observed in practice.

The tidal behaviour is very dependent on the global position. The tide observed in the North Sea is a result mainly of a tidal wave propagating from the Atlantic Ocean both through the English Channel and north of the British Isles creating a complex tidal pattern with two amphidromic points as mentioned earlier, see Figure 5.1.

The tide is dominated by the M2 tidal component (where "M" stands for Moon and "2" stands for two periods a day), which has a period of 12.42 hours as a result of Earth's rotation and the orbit of the Moon.

6.2.1 Description

The database provided by CEFAS was originally acquired at the Proudman Oceanographic Laboratory (POL), which is a scientific research institution focusing on oceanography encompassing global sea-levels and geodesy, numerical modelling of continental shelf seas and coastal sediment processes[1].

The database has a grid of 1/9° latitude and 1/6° longitude which is a resolution of approximately 12×12 km. The area covered is from 48.17° to 59.95° latitude, and $-11.75°$ to 7.92° longitude, equal to a grid of 119×107 cells.

As a consequence of the spherical coordinates given in the database the requirement of a rectangular grid is no longer fulfilled i.e. the spatial step h varies across the domain. Effectively this causes the diffusivity to be dependent on the spatial position. The range of the value of h in the POL domain is from 9.33

[1] For further information see www.pol.ac.uk

km to 12.42 km, which is a substantial variation. It is however not considered too large to invalidate the method if D is kept at a constant value. Moreover, a single individual is expected to stay within a significantly smaller area than the one spanned by the entire domain thereby supporting the choice of a constant value. The value of h is set to 10.88 km, the average value of the extremes.

The database contains the bathymetry of the region as well as seven tidal constituents: M2, S2, N2, K2, O1, K1, M4. It was created in 1994 and is based on a storm surge model and meteorological data. Little is known of the uncertainties of the database other than qualified guesses from people experienced with the database.

It is certain that the maximal uncertainty is found at the shores where the tide is extreme and that, at open sea and at a distance from amphidromic points, the database should be reliable to within 10 cm on the tidal prediction.

6.2.2 Tidal prediction

The tidal variation for a single constituent at a given position is calculated in the simple way

$$z = fA(\boldsymbol{x})\cos[\omega t - \theta(\boldsymbol{x}) + G],$$

where $A(\boldsymbol{x})$ and $\theta(\boldsymbol{x})$ are amplitude and phase respectively that depends on position, $\boldsymbol{x}$. The constants f and G are calibration parameters that are calculated separately and depends on the tidal constituent and the $t = 0$ definition. The prediction of the seven constituents given by the database are summed to yield the complete tidal prediction at the given position.

6.3 Analysis of stationary tags

For the observational likelihood to be calculated, a model that describes the possible random variations in a tidal pattern, must be formulated. When a tidal pattern is observed in the depth record, the fish is assumed to be stationary at the sea floor. It is therefore relevant to study stationary tags to assess the uncertainties that are non fish related e.g. tidal prediction uncertainty, tag measurement uncertainty, influences of weather etc. Tags are kept stationary by so called minipods.

6.3.1 Analysis of tidal noise

A change in weather conditions will impose fluctuations on the depth record even at the sea bed level. This is confirmed in an examination of the depth measured by the stationary tag #1536, that shows a period of increased noise, Figure 6.5.

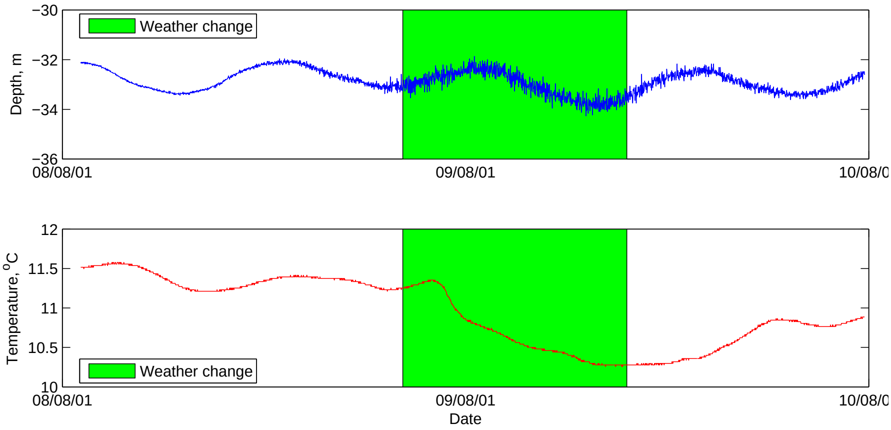

Figure 6.5: *Measurements of depth and temperature from Tag #1536 at the 9th of August, 2001. The depth has increased fluctuations and the temperature drops approximately a degree at the time.*

According to the Danish Meteorological Institute[2] low pressure, 995 hPa, was observed in western Jutland at the 9th of August 2001, leading to increased wind speeds of >17 m/s on the main land. The harsh weather conditions seem to affect the data recorded by the stationary tag, see Figure 6.5. The depth is measured with significantly increased noise due to waves and the temperature is seen to drop about 1 °C possibly because of mixing with surrounding colder water.

From Figure 6.5 there are obviously two types of variation in the observed depth. Variation following the tide which has the approximate period of 12.4 hours and the superposed white noise type variation from the waves that varies at a frequency higher than the tag sample rate of 1 minute. The observations are assumed to follow the stochastic process

$$Z_i = \mathcal{D}_i + E_i,$$

[2] www.dmi.dk

where E_i is white noise i.e. $E_i \sim \mathcal{N}(0, \sigma_E^2)$, and $\mathcal{D}_i$ is a slowly varying process that comprises the mean depth, the tidal variation and storm surge. A one-differencing of the process is performed to remove the slow process leaving the superposed noise

$$V_i = Z_{i+1} - Z_i = (\mathcal{D}_i + E_i) - (\mathcal{D}_{i-1} + E_{i-1}) \simeq E_i - E_{i-1}.$$

The white noise variance, σ_E^2, is estimated as half of the empirical variance of V_i

$$\sigma_E^2 = \frac{1}{2}\mathbb{V}(V_i).$$

The green intervals in Figure 6.5, consists of 800 data points and cover the period where the storm is at its highest. The standard deviation of the white noise is found to

$$\widehat{\sigma}_E = \sqrt{0.5 \cdot 0.072763} = 0.19074\,\mathrm{m}.$$

This type of variation in depth cannot be predicted by the tidal model and must therefore be incorporated in the error model.

The assumption of Gaussianity of V_i and thereby E_i, can be checked by a Q-Q plot of the quantiles of the Gaussian distribution to quantiles of the empirical distribution of data. A Q-Q plot for V_i is displayed in Figure 6.6 along with the autocorrelation function (acf) for V_i.

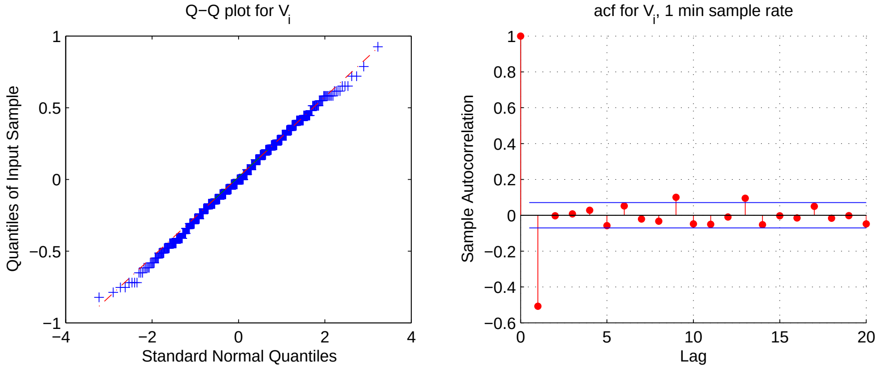

Figure 6.6: *Statistical analysis of V_i. Left pane: Q-Q plot for V_i. Right pane: acf for V_i. The process shows apparent Gaussianity as assumed.*

The Q-Q plot shows that the quantiles of V_i have strong agreement with the quantiles of a standard Gaussian distribution. The theoretical autocovariance function for a differenced white noise process is given by

$$\mathsf{Cov}(V_i, V_{i+\Delta}) = \mathbb{E}(V_i V_{i+\Delta}) - \mathbb{E}(V_i)\mathbb{E}(V_{i+\Delta}).$$

The process has zero expectancy hence the last term can be omitted and after some computation it is found that

$$\mathbb{E}(V_i V_{i+\Delta}) = \begin{cases} 2\sigma_E^2 & \text{for } \Delta = 0 \\ -\sigma_E^2 & \text{for } \Delta = 1 \\ 0 & \text{otherwise} \end{cases} .$$

Normalisation of this autocovariance function by $2\sigma_E^2$ gives an *acf* similar to the estimated shown in Figure 6.6 right pane.

Finally, a test for distribution is performed. The one, commonly recognised as the most powerful, is the Anderson-Darling test (D'Agostino and Stephens, 1986). This is similar in methodology to the Kolmogorov-Smirnov test (Conover, 1971) but allows for the parameters of the test distribution to be estimated from the data. The hypotheses are

H_0: The data comes from a Gaussian distribution.
H_1: The data does not come from a Gaussian distribution.

The result was that the H_0 hypothesis is rejected at a significance level of $\alpha = 0.05$, but not at the $\alpha = 0.025$ level with a test statistic of 0.75731. It should be noted that the Anderson-Darling test assumptions of independent observations in V_i were violated because of the correlation. Even so, the test passed at an acceptable significance level to allow for practical implementation.

It is concluded based on the above analysis that changes in weather conditions can lead to a white noise effect on the depth measurements with a standard deviation of at least $\widehat{\sigma}_E = 0.19074$ m. This noise covers also the uncertainty inherent in the resolution of the tag and other unknown non-fish related white noise sources. Inspection of the observations tells that the variance of the white noise is time varying but is for simplicity modelled here with this constant value.

6.3.2 Tidal prediction uncertainty

The POL database has a resolution of approximately 12 × 12 km on the bathymetry as well as on the amplitude and phase of the tide. An arbitrarily fine grid of the tide can be obtained by interpolation but this is not appropriate for the bathymetry. Therefore it has no meaning to refine the resolution with the intention to get a more precise geolocation.

It may be worthwhile, though, to interpolate the phase and amplitude to check the tidal prediction at the exact location of a stationary tag to get an impression of how the optimal prediction looks. Tag #1536 of the above analysis is reused here. In Figure 6.7 is shown a plot of the measured depth at 55.24° latitude, 2.57° longitude, and its predicted depth with the mean subtracted.

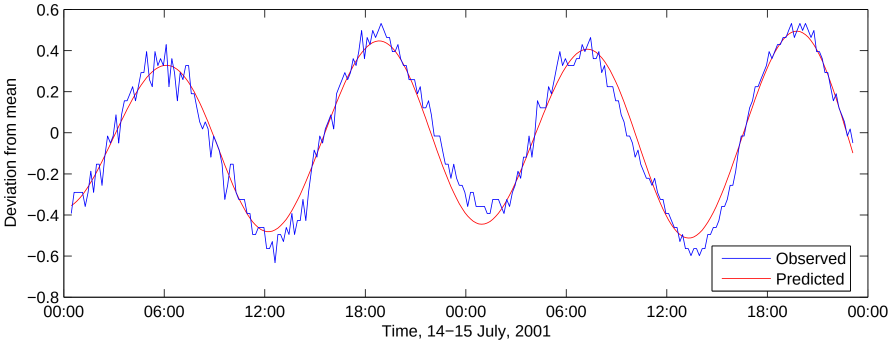

Figure 6.7: *Observed tide and predicted tide at exact location for tag #1536.*

The predicted pattern fits well to the observed. Deviation from the prediction is explained by the white noise of the previous section. Later in the time series, a change in conditions seems to have moved the minipod slightly and introduced a bias (approx. 20 cm) in the depth (not shown). An assessment of the tidal prediction uncertainty is given in the next section.

6.4 Uncertainty due to database resolution

As mentioned, the resolution of the POL database introduces errors in the predictions if the actual location is not perfectly on top of a grid cell. Comparison of an observed tide at one position with a predicted tide at another results in residuals with an oscillating structure, see Figure 6.8. This type of error must be assessed and accounted for when observations are compared with predictions.

6.4.1 Tidal error

The tidal variation with respect to tidal range and times of high and low tide varies as a function of the position and consequently the error of the database must be a function of the position as well. A new term is introduced called

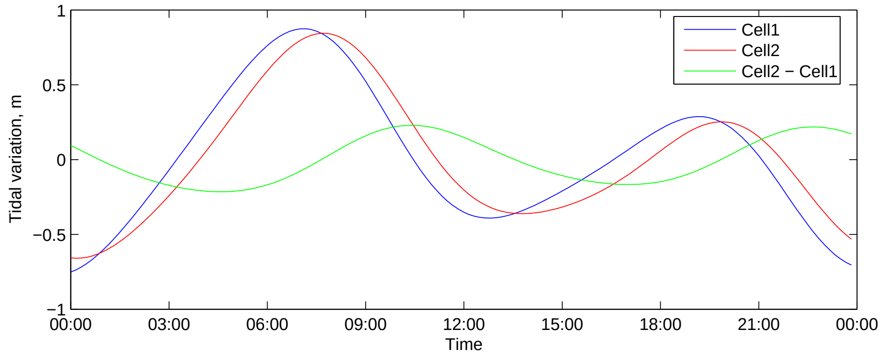

Figure 6.8: *Tidal prediction from two adjacent grid cells close at 52.5° latitude, 1.75° longitude. A position with relatively large tidal variation close to the shore.*

tidal roughness. It is a quadratic measure for the difference in tidal prediction, $\widehat{\boldsymbol{z}}(\boldsymbol{x})$ from a position $\boldsymbol{x}$ to the predictions at its adjacent positions in the grid, $\widehat{\boldsymbol{z}}(\boldsymbol{x}+\Delta\boldsymbol{x})$, where $\boldsymbol{x}+\Delta\boldsymbol{x}$ denotes one of the 8 adjacent positions. The tidal roughness is proportional to the variance, $\sigma_e^2(\boldsymbol{x})$, of the tidal prediction error for a position. The variance has a slight temporal variation but is here assumed to be constant given $\boldsymbol{x}$.

The tidal prediction for two adjacent positions in the southern North Sea are shown in 6.8. The difference between the two predictions has a wave form with a period time equal to that of the two predictions.

The tidal roughness for a cell is estimated as the maximal empirical variance of the differences with its adjacent positions in the database grid i.e.

$$\widehat{\sigma}_e^2(\boldsymbol{x}) = \max_{\Delta\boldsymbol{x}} \mathbb{V}[\widehat{\boldsymbol{z}}(\boldsymbol{x}) - \widehat{\boldsymbol{z}}(\boldsymbol{x}+\Delta\boldsymbol{x})].$$

For the example shown in Figure 6.8 it was found that $\widehat{\sigma}_e(\boldsymbol{x}) = 0.175$ m.

The map of $\widehat{\sigma}_e(\boldsymbol{x})$ is shown in Figure 6.9. The largest roughness is observed at shores and at narrow passages e.g. the English Channel whereas the mid North Sea has very little variation. Amphidromic points are local minima because the amplitude of the tide is diminished.

The error $\widehat{\boldsymbol{z}}(\boldsymbol{x}) - \widehat{\boldsymbol{z}}(\boldsymbol{x}+\Delta\boldsymbol{x})$ for fixed $\boldsymbol{x}$ and $\boldsymbol{x}+\Delta\boldsymbol{x}$, is a function of time. The error is assumed to follow a $\mathcal{N}(0,\sigma_e^2)$ distribution with an oscillating correlation structure in accordance to Figure 6.8. The estimated *acf* of this error, shown in Figure 6.10, which explicitly expresses the correlation structure of this

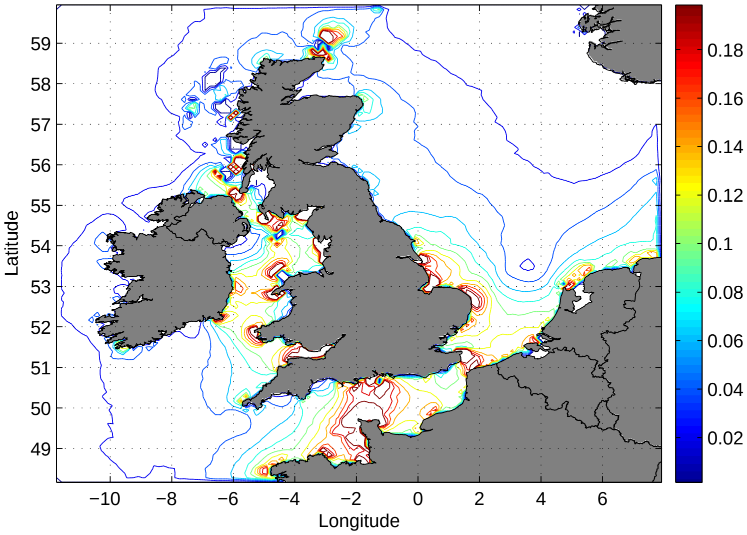

Figure 6.9: *Map of $\widehat{\sigma}_e(\boldsymbol{x})$ across the domain. Note that $\widehat{\sigma}_e(\boldsymbol{x})$ of 0.2 m and above is indicated by one contour. These high values occur at the shores whereas the open sea has little tidal variation particularly at the amphidromic points.*

semidiurnal variation. It consist of a pure sine wave that can be expressed as

$$\rho_{i,i+\delta} = \cos\left(\frac{2\pi}{p}\delta\right), \tag{6.4}$$

where $\rho_{i,i+\delta}$ is the autocorrelation at lag δ and p is the period time that is defined as the period of the dominating M2 tidal constituent which is 12.42 hours or $p = 74.52$ ten-minute intervals.

6.4.2 Bathymetry error

The bathymetry of the North Sea is stored in a discrete grid with resolution 12×12 km which is far too coarse to capture all variation of the sea bed. This imposes an uncertainty on the data in the database that influences the predicted depth of a given position.

The bathymetry uncertainty is position dependent and is particularly large in areas with a considerable variation in depth e.g. at shores or banks. A conservative method is used to assess the uncertainty for each position in the bathymetry.

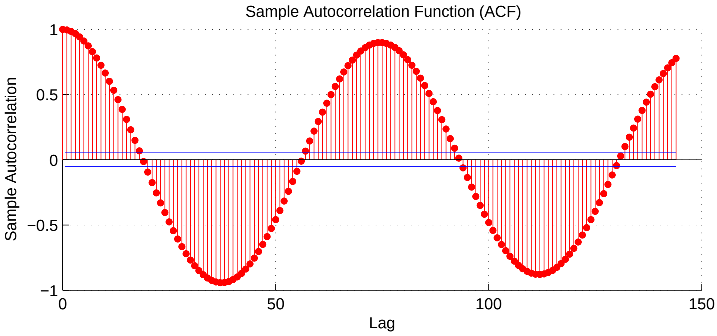

Figure 6.10: *Autocorrelation function for $\widehat{z}(\boldsymbol{x}) - \widehat{z}(\boldsymbol{x} + \Delta\boldsymbol{x})$ at for fixed x and $x + \Delta x$. The acf has a period of approximately 72 lags i.e. 12 hours (when the sample rate is 10 min).*

The depth given in the database at position $z(\boldsymbol{x})$ is assumed to be uniformly distributed in an interval of length Δz and therefore has the variance

$$\sigma_\eta^2(\boldsymbol{x}) = \mathbb{V}[z(\boldsymbol{x})] = \frac{\Delta z^2}{12},$$

where

$$\Delta z = \max_{\Delta\boldsymbol{x}} z(\boldsymbol{x} + \Delta\boldsymbol{x}) - \min_{\Delta\boldsymbol{x}} z(\boldsymbol{x} + \Delta\boldsymbol{x}).$$

Here $\boldsymbol{x} + \Delta\boldsymbol{x}$ refers to one of the 8 adjacent positions. Performing this calculation for the entire domain yields the result shown in Figure 6.11.

Figure 6.11 shows that the roughness is increased at the shores and around banks whereas the flat eastern North Sea has almost zero roughness.

6.5 Error from fish movement

Small scale movements of the fish will impose bias on the recorded depth signal and thereby causing it not to conform to the correlation structure in (6.4). For a fish the change in depth within a 10 minute interval can be fairly large especially in sloped or rocky areas that are often favorite habitats for fish to linger.

It is not directly possible to estimate the correlation structure that arise in the depth measurements when this kind of behaviour is present, hence an intuitive model is provided. For a depth record the future time step, Z_{i+1}, must be

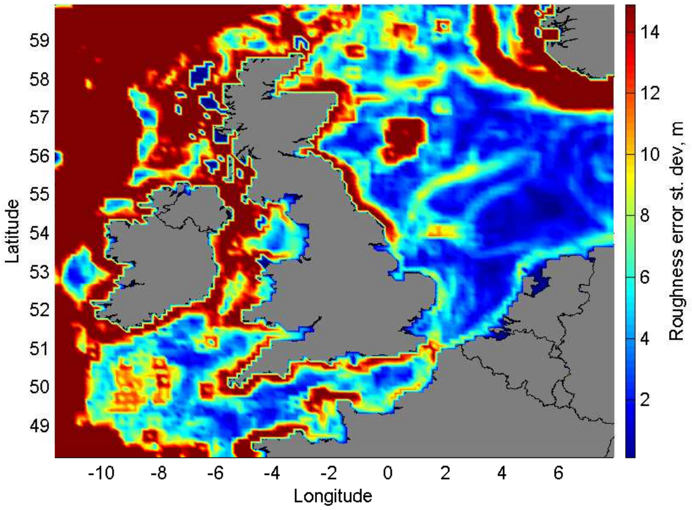

Figure 6.11: *Map of $\widehat{\sigma}_\eta(\boldsymbol{x})$ across the domain. Note that $\widehat{\sigma}_e(\boldsymbol{x})$ of 15 m and above are all shown as red.*

equal to the present, Z_i, times a weight plus a random error. When all other contributions are removed e.g. mean depth, tidal variation etc., the AR(1) model is written as

$$Z_i = \lambda Z_{i-1} + \varepsilon_i, \tag{6.5}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ by assumption and the weight $|\lambda| < 1$. The values of the parameters, σ_ε^2 and λ, require detailed knowledge of individual fish movement at the microscopic level to assess and will probably still have a large interindividual variation.

The formulation in (6.5) gives rise to a covariance structure of the measurements given by

$$\mathsf{Cov}(\varepsilon_i, \varepsilon_{i+\delta}) = \sigma_\varepsilon^2 \lambda^{|\delta|},$$

with heuristic estimates of the parameters set to $\sigma_\varepsilon^2 = 0.05$ m and $\lambda^{40} = 0.05$.

6.6 Likelihood for observation

This section describes how the observational likelihood given the position, i.e. the term $\mathcal{L}(\boldsymbol{Y}_j = \boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j)$, is determined.

6.6.1 Demersal behaviour

At time intervals where tidal information is present, a likelihood for the observation given the position, $\boldsymbol{x}_j$, is calculated based on how well the observation, $\boldsymbol{y}_j$, fits to the prediction, $\widehat{\boldsymbol{z}}_j(\boldsymbol{x})$. The assumed model (6.1) was

$$\boldsymbol{Y}_j \sim \mathcal{N}_m\big(\widehat{\boldsymbol{z}}_j(\boldsymbol{x}_j), \boldsymbol{\Sigma}(\boldsymbol{x}_j)\big).$$

Deviations in the database predictions from the observed depths will follow the error schemes outlined in Sections 6.3, 6.4 and 6.5. These are used to construct an estimate of $\boldsymbol{\Sigma}(\boldsymbol{x}_j)$. Listed here in summary they are

E_i – White noise term that describes the error caused by the sensor resolution of the tag, by noise from environmental influences such as storms and currents and other unknown sources that invoke white noise. The white noise variance becomes

$$\mathsf{Cov}(E_i, E_{i+\delta}) = \begin{cases} \sigma_E^2 & \text{for} \quad \delta = 0 \\ 0 & \text{for} \quad \delta \neq 0 \end{cases}.$$

e_i – This describes the periodic error that is caused by the resolution of the tidal database. It contributes to the covariance structure with

$$\mathsf{Cov}(e_i, e_{i+\delta}) = \sigma_e^2 \cos\left(\frac{2\pi}{p}\delta\right).$$

η_i – The error term that accounts for the fact that the bathymetry has a low resolution compared to the detail level of the sea floor. This results in a constant term affecting the whole covariance matrix

$$\mathsf{Cov}(\eta_i, \eta_{i+\delta}) = \sigma_\eta^2.$$

ε_i – Small scale movements of the fish may cause minor changes in the depth and thereby perturb the tidal signal. This can be modelled as an AR(1) process with covariance structure

$$\mathsf{Cov}(\varepsilon_i, \varepsilon_{i+\delta}) = \sigma_\varepsilon^2 \lambda^{|\delta|}.$$

An illustration of the contributions is shown in Figure 6.12.

The *pdf* for $\boldsymbol{Y}_j$ is written here explicitly for the sake of clarity

$$\begin{aligned} & f_{\boldsymbol{Y}_j}(\boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j) \\ = & \frac{1}{(2\pi)^{m/2}\sqrt{\det \boldsymbol{\Sigma}(\boldsymbol{x}_j)}} \exp\left(-\frac{1}{2}[\boldsymbol{y}_j - \widehat{\boldsymbol{z}}_j(\boldsymbol{x}_j)]^T \boldsymbol{\Sigma}(\boldsymbol{x}_j)^{-1} [\boldsymbol{y}_j - \widehat{\boldsymbol{z}}_j(\boldsymbol{x}_j)]\right), \end{aligned} \tag{6.6}$$

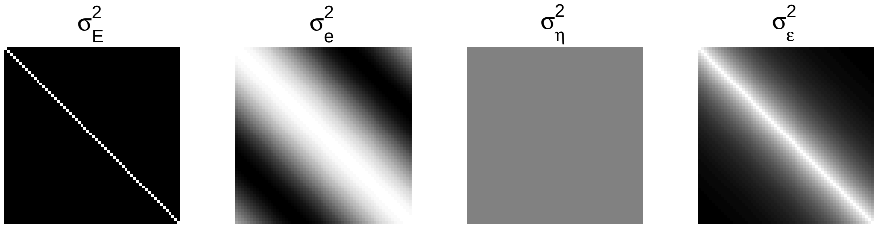

Figure 6.12: *Illustration of the correlation structure of the four contributions to* $\boldsymbol{\Sigma}(\boldsymbol{x}_j)$*. Each matrix is* $m \times m$ *(*60×60*). The color scales are:* σ^2_E *- white is 1, black is zero.* σ^2_e *- white is 1, black is* -1*.* σ^2_η *- gray is 1.* σ^2_ε *- white is 1, black is zero.*

for $j \in [1, \ldots, N-1]$

The observational likelihood is found by considering (6.6) as a function of the position $\boldsymbol{x}_j$

$$\mathcal{L}(\boldsymbol{Y}_j = \boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j) = \mathbb{P}\big[\boldsymbol{Y}_j = \boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j ; \widehat{\boldsymbol{z}}(\boldsymbol{x}_j), \boldsymbol{\Sigma}(\boldsymbol{x}_j)\big]. \qquad (6.7)$$

Calculating the observational likelihood for the entire domain yields a unnormalised probability distribution (hence likelihood) for the observation, parameterised by the position.

6.6.2 Pelagic behaviour

So far, only time intervals containing tidal information have been considered. When a tidal signal cannot be extracted from the time series the behaviour of the fish is unknown. Here it is conservatively assumed that the fish is pelagic, e.g. migrating or foraging in the water column, away from the sea bed.

In the absence of a tidal signal there is still some information in the time series that can be used for geolocation. A very strict model would say that the fish cannot be in shallow waters if a large depth is measured. This is true, but with the limited resolution of the database bathymetry, the possibility of the fish being in some position cannot be ruled out based solely on the depth. Instead, the bathymetry uncertainty (Subsection 6.4.2) is used to calculate a reasonable likelihood for the observation given the position.

An indicator variable, $I_j(\boldsymbol{x})$, is defined where

$$I_j(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad \mathcal{D}_j(\boldsymbol{x}) < \overline{z}_j \\ 0 & \text{if} \quad \text{otherwise} \end{cases}, \qquad \text{for} \quad z(\boldsymbol{x}) < 0,$$

where $\overline{z}_j$ is the maximal depth recorded in $\boldsymbol{z}_j$ and $\mathcal{D}_j(\boldsymbol{x})$ is a random variable that follows a truncated Gaussian distribution i.e. $\mathcal{D}_j(\boldsymbol{x}) \sim \mathcal{N}\big(z(\boldsymbol{x}), \widehat{\sigma}_\eta(\boldsymbol{x})\big)$ where $z(\boldsymbol{x}) < 0$. The value $z(\boldsymbol{x})$ is the depth at the position $\boldsymbol{x}$ given by the database. The likelihood of a position is assigned as the expectation of the indicator of the position

$$\begin{aligned} \mathcal{L}(\boldsymbol{Y}_j = \boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j) &= \mathbb{E}[I_j(\boldsymbol{x})] \\ &= \mathbb{P}[\mathcal{D}_j(\boldsymbol{x}) < \overline{z}_j], \end{aligned} \tag{6.8}$$

for $j \in [1, \ldots, N-1]$. Defining Φ as the *cumulated density function* (*cdf*) of a standardised Gaussian distribution with the constraint (truncation)

$$\overline{z}_j < 0, \quad \text{and} \quad z(\boldsymbol{x}) < 0.$$

Now (6.8) can be written as

$$\mathcal{L}(\boldsymbol{Y}_j = \boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j) = \Phi\left(\frac{\overline{z}_j - z(\boldsymbol{x})}{\widehat{\sigma}_\eta(\boldsymbol{x})}\right) \Phi\left(\frac{-z(\boldsymbol{x})}{\widehat{\sigma}_\eta(\boldsymbol{x})}\right)^{-1}, \tag{6.9}$$

which is the *cdf* evaluated at $\overline{z}_j$ normalised by the *cdf*-value at the zero-crossing as a consequence of the truncation. The likelihood will decline according to the *cdf* and approach zero as $\overline{z}_j - z(\boldsymbol{x}) \to -\infty$ (remember that depth measurements are negative). The sketch in Figure 6.13 illustrates the calculation performed in (6.9).

6.6.3 Recapture position

The equations (6.7) and (6.9) assess only the likelihood in the time interval $\tau_1, \ldots, \tau_{N-1}$. Experience from past tagging experiments say that the terminal position cannot be assumed to be known without uncertainty and must enter into the likelihood for the observation, $\boldsymbol{y}_N$. Though uncertain, the recapture position is of particular importance if no tidal information is present close to τ_N.

The terminal position, $\boldsymbol{X}_\ddagger$, has the assumed distribution

$$\boldsymbol{X}_\ddagger \sim \mathcal{N}(\boldsymbol{x}_\ddagger, \sigma_N^2 \boldsymbol{I}),$$

where $\boldsymbol{I}$ is the 2×2 identity matrix and $\sigma_N = 20$ km based on experience from past experiments. The observational likelihood for the final time step when tidal

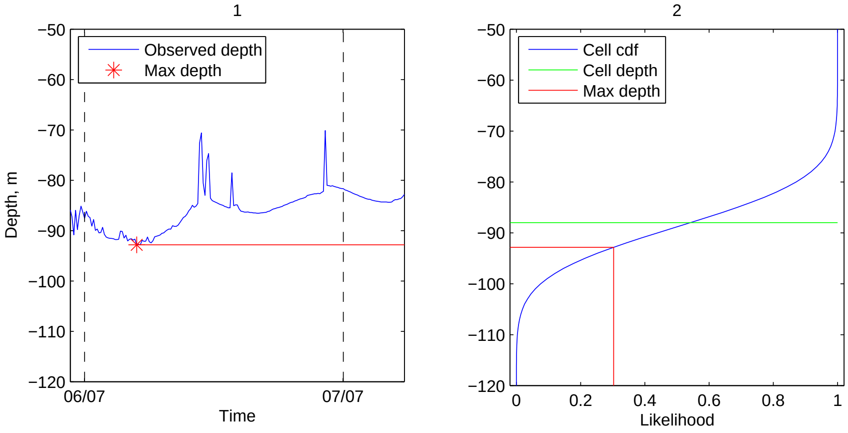

Figure 6.13: ***1**: Observed depth at 6th of July 2001 of tag #2255. **2**: Principle in calculation of the likelihood at a position* (55.8° *latitude,* −0.25° *longitude). The deepest observation in the record is* −92.8 *m. This is compared to the depth value of the grid cell,* −88 *m, by evaluation of the expression in (6.9). In this example the likelihood becomes 0.30.*

information is present is given by

$$\mathcal{L}(\boldsymbol{Y}_N = \boldsymbol{y}_N | \boldsymbol{X}_N = \boldsymbol{x}_N) = \mathbb{P}(\boldsymbol{X}_\ddagger = \boldsymbol{x}_N)\mathbb{P}(\boldsymbol{Z}_{N,\widehat{i}} = \boldsymbol{z}_{N,\widehat{i}} | \boldsymbol{X}_N = \boldsymbol{x}_N).$$

where $\boldsymbol{Z}_{N,\widehat{i}}$ follows the distribution given in (6.1). In the case no tidal information is extracted the observational likelihood becomes

$$\mathcal{L}(\boldsymbol{Y}_N = \boldsymbol{y}_N | \boldsymbol{X}_N = \boldsymbol{x}_N) = \mathbb{P}(\boldsymbol{X}_\ddagger = \boldsymbol{x}_N)\Phi\left(\frac{z(\boldsymbol{x}) - \overline{z}_N}{\widehat{\sigma}_\eta(\boldsymbol{x})}\right)\Phi\left(\frac{-\overline{z}_N}{\widehat{\sigma}_\eta(\boldsymbol{x})}\right)^{-1}.$$

These results are due to the conditional independence of $\boldsymbol{X}_\ddagger$ with the depth observations given $\boldsymbol{X}_N = \boldsymbol{x}_N$.

Finally it should be stressed that the position of the release of the fish is known without uncertainty and therefore needs no *data-update*.

Chapter 7

Results

This chapter presents the results obtained when the theory and methods described in the previous chapters are applied to data from DSTs mounted on fish.

The presented tags are chosen to emphasise important aspects of the method and serve as validation and evaluation with a view to improving the model further. Some of the tags have been subject to investigation by CEFAS in the recent years. Selected results have been published (Hunter et al., 2005; Righton and Mills, 2007) and are used for comparison in this study.

The analysis has focused on the following six tags

- #1209, stationary tag, Section 7.1.
- #2255, cod, Section 7.2, (Righton and Mills, 2007).
- #1186, cod, Section 7.3.
- #2324, thornback ray, Section 7.4, (Hunter et al., 2005).
- #1432, cod, Section 7.5.
- #6448, cod, Section 7.5.

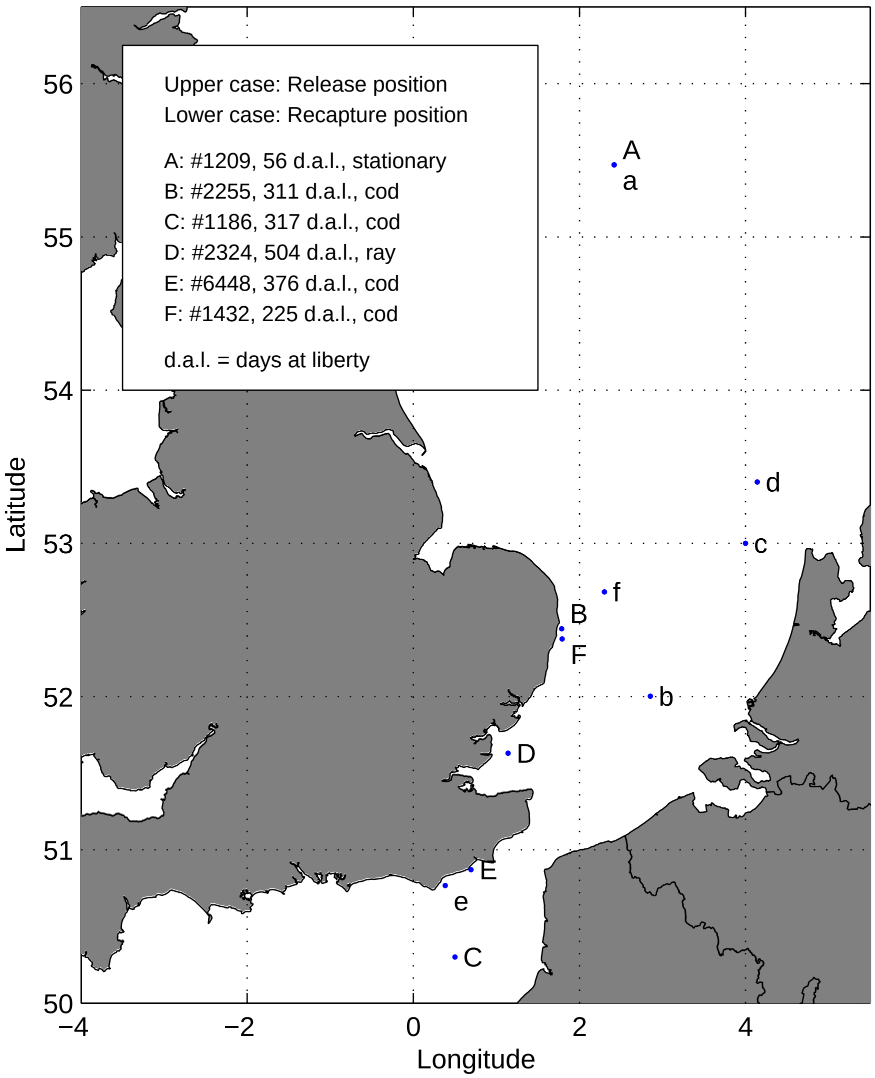
Upper case: Release position
Lower case: Recapture position
A: #1209, 56 d.a.l., stationary
B: #2255, 311 d.a.l., cod
C: #1186, 317 d.a.l., cod
D: #2324, 504 d.a.l., ray
E: #6448, 376 d.a.l., cod
F: #1432, 225 d.a.l., cod
d.a.l. = days at liberty
A
a
B
F
f
b
c
d
D
E
e
C
56
55
54
53
52
51
50
Latitude
−4
−2
0
2
4
Longitude

Figure 7.1: *Reported release and recapture positions for DSTs.*

In Figure 7.1 is shown the reported release and recapture positions of the tags.

For each tag an animation is generated, which is an avi-file, showing the evolution of the marginal posterior distributions in time. The abbreviation AMPD for "Animated Marginal Posterior Distributions" is used henceforth. The animations are found at the web site www.student.dtu.dk/∼s002087 and on the enclosed CD-ROM. The animations was created with MATLAB's `avifile` command and compressed with the Cinepak AVI codec to limit the file size. This compression results in some loss of detail especially in the plot of the depth record and the tidal intervals. It is therefore recommended to inspect these from the printed plots or from the MATLAB `fig`-files on the CD-ROM.

Important notice: The files on the CD-ROM are not to be distributed without permission from CEFAS.

7.1 Stationary tag, #1209

As a first check, a tag from a minipod is geolocated and compared to its actual known global position. This will reveal, to some extent, the uncertainty and bias of the method and give an impression of how well a stationary fish can be geolocated. The tag type was LTD 1200 (see Section 5.2).

7.1.1 Inspection of the data

The minipod was deployed at the coordinates 55.47° latitude and 2.42° longitude, see Figure 7.1. The depth time series of the stationary tag #1209 is shown in Figure 7.2.

Most of the data is marked in green colour indicating that tidal information could be extracted. It would be expected for a stationary tag that the entire data set showed a tidal pattern. Apparently, a change in the weather conditions at the 8th of August and again towards the end, imposed noise onto the observations and locally rendered the signal useless for tidal comparison.

7.1.2 Results

The AMPD show that the geolocation algorithm finds the tag to be positioned in a grid cell adjacent to the reported true position, indicating a minor bias.

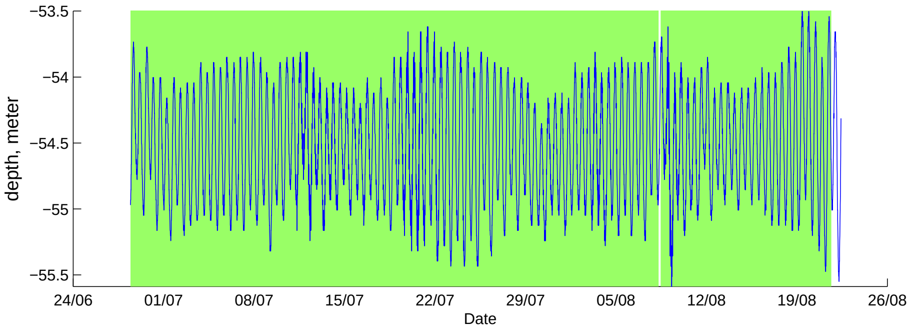

Figure 7.2: *Time series from tag #1209, released 28th of June 2001 and recaptured 22nd of August 2001. Tidal information intervals are marked in green.*

A number of other stationary tags have been analysed with a similar result although no consistency in bias could be detected. All available tags were deployed in the same geographical area and within a time period of a few days. It is not possible to make strong conclusions based on such a sparse data set that furthermore were influenced by changing weather.

The overall conclusion is that the stationary tag are geolocated satisfactory.

7.2 Cod #2255

This tag contained a very high quality data set, perfectly suited for tidal based geolocation. For this reason the tag has undergone a thorough study at CEFAS using the the Tidal Location Method (Hunter et al., 2003, explains the TLM). Hence, the overall behavioural pattern is known, along with geolocated positions at time instances with strong tidal information. These results can be used for cross validation with the method described in this thesis and help to uncover deviations and pinpoint potential errors in the present model.

7.2.1 Inspection of the data

The tag type was LTD 1200. The cod was released at the 3rd of April 2001 at 52.44° latitude, 1.78° longitude and recaptured the 6th of February 2002 at 52.00° latitude, 2.85° longitude (see Figure 7.1).

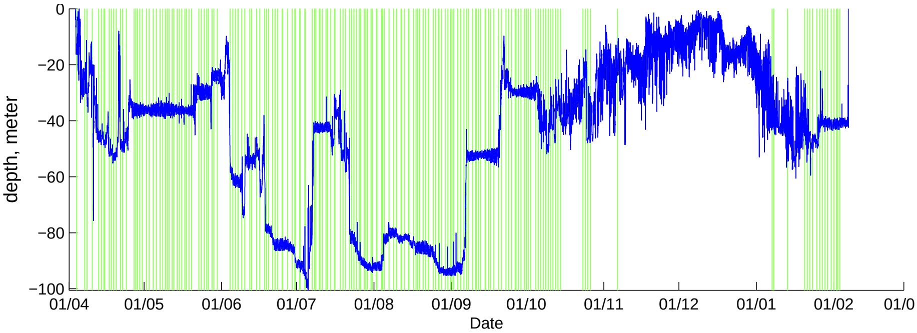

Figure 7.3: *Time series from tag #2255, released 3rd of April 2001 and recaptured 6th of February 2002. Tidal information intervals are marked in green.*

The entire time series of the tag is shown in Figure 7.3. The time series lasts for 311 days which is one of the longer data set obtained from cod in the North Sea. Moreover does it hold much tidal information and of impressing high quality, at times more smooth than the data measured by the stationary tags. Tidal information is present both at the day of release and day of recapture, which enables the reported positions to be cross validated with the geolocations and unveil possible uncertainties of both.

7.2.2 Results

The data is processed by the geolocation filter and the smoothed position estimates, $\mathbb{P}(\boldsymbol{X}_j = \boldsymbol{x}_j | \boldsymbol{\mathcal{Y}}_N)$ for $j \in [0, \ldots, 311]$, are obtained, where τ_0 is the 3rd of April 2001 and τ_{311} is the 6th of February 2002.

7.2.2.1 Estimation of D

The MLE of D for tag #2255 was found to

$$\widehat{D} = 22.4 \text{ km}^2/\text{day},$$

with a standard deviation of 2.7 km^2/day estimated from the observed Fisher information. The estimate of D represents the average diffusivity that fits the model best given the depth record. The diffusivity is a measure for how active the fish was during its time at liberty.

In the two dimensional space the diffusivity is related to the average swimming speed of the fish in the way

$$D = \frac{\rho v^2}{2}, \tag{7.1}$$

where v is the constant speed of the fish and ρ is the decorrelation time of this swimming speed, e.g. Visser and Thygesen (2003).

The value of ρ is not known and is not immediately possible to estimate based on the data. Instead a conservative value of $\rho = 12$ hours, is chosen. This means that the *acf* of the velocity has decreased to insignificant values after 12 hours.

The maximal average swimming speed for an interval of 24 hours is selected to be 0.5 body lengths per second. Breen et al. (2004) and its references provide a reasonable fundament for this decision. Among the cod considered in this study, lengths were in the range 50-70 cm, corresponding to maximal speeds in the range 22-30 km/day. The conservative value of 30 km/day is chosen and results in a maximal value for the diffusivity of

$$D_{max} = 225 \text{ km}^2/\text{day}.$$

Comparing this value to the MLE for #2255, $\widehat{D} = 22.4$ km^2/day, the activity level of this cod appears to be low on average.

7.2.2.2 Animated marginal posterior distributions

When inspecting an AMPD it should be borne in mind that the color scale is not constant in time. The AMPD reveal time intervals of low uncertainty where the fish is stationary. These are typical at times when tidal extraction was possible. The low activity intervals were mostly present in the summer months where the fish stayed near the eastern shore of England in the middle North Sea. This type of cod behaviour is also reported in Turner et al. (2002); Righton et al. (2007). The fish had a high level of activity in the initial and final part of its time at liberty, migrating north and south respectively. The behaviour displayed by this cod seems to conform very well to general trends shown in past tagging experiments.

7.2.2.3 Most Probable Track

To visualise the result of the geolocation, the MPT is calculated from the estimated joint posterior distribution, see Figure 7.4 right pane.

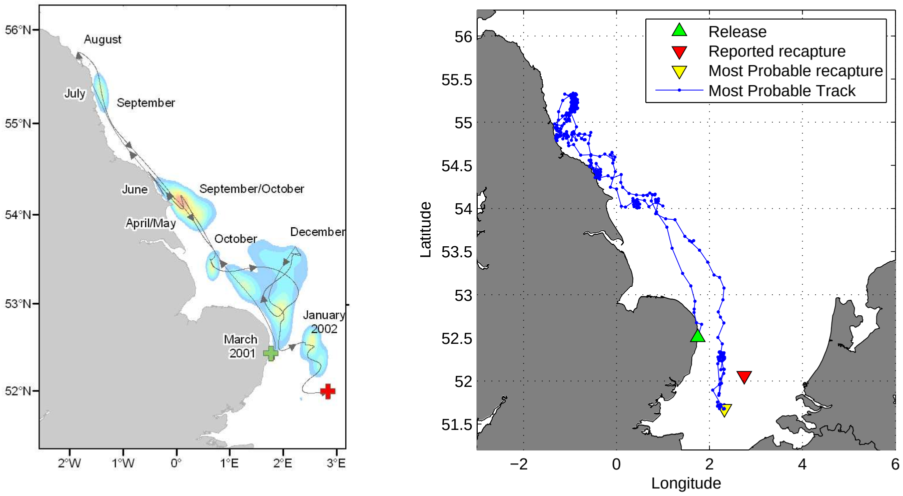

Figure 7.4: *Comparison of geolocation methods. Left pane: CEFAS's results based on a modified TLM. Right pane: MPT computed from the estimated joint posterior distribution of #2255.*

The cod was captured and released close to Lowestoft (eastern England) and, according to the MPT, immediately began a migration to the north, settling down a month later at approximately the 1st of May at 54.5° latitude, −0.5° longitude. Here it stayed for a month before relocating a bit further north to an area around 55° latitude, −1° longitude, where it stayed for a longer period until late September. Then activity level gradually increased (also evident from the Figure 7.3) and eventually a southwards migration brought the cod to a position at 51.75° latitude, 2.5° longitude, around the 9th of January and was recaptured a month later at approximately this position.

Figure 7.4 left pane shows the result obtained with CEFAS's TLM method supplemented by temperature measurements Righton and Mills (2007). The coloured areas are pseudo *pdfs* that are calculated based on a MCMC algorithm. The two plots show largely identical movement patterns of the fish. There are minor deviations due to the difference in method most evidently in the final southern migration.

It was found that the most probable recapture position is estimated to differ significantly ($p < 0.0001$) from the reported recapture position, see also Figure 7.4 and the AMPD. The deviation cannot be explained purely by a possible bias in geolocation and it is therefore concluded that the reported recapture position must be encumbered with uncertainty. In contrast, the release position is geolocated precisely based on the tidal pattern observed by the tag after only a short time at liberty (7 hours). This supports the geolocation method and decreases the faith in the correctness of the reported recapture position.

7.2.2.4 Bathymetry roughness

A close inspection of the data shows some curiosities that may be interesting to examine further. At the 10th of April, the fish visits a depth of −75 m and returns to around −20 m within a short time interval of approximately 10 hours, see Figure 7.5. First of all, this is interesting for a biologist as it requires a great effort from the fish to perform a depth change of this magnitude, that is unlikely to be carried out purely by regulation of its swimbladder (Harden Jones and Scholes, 1985).

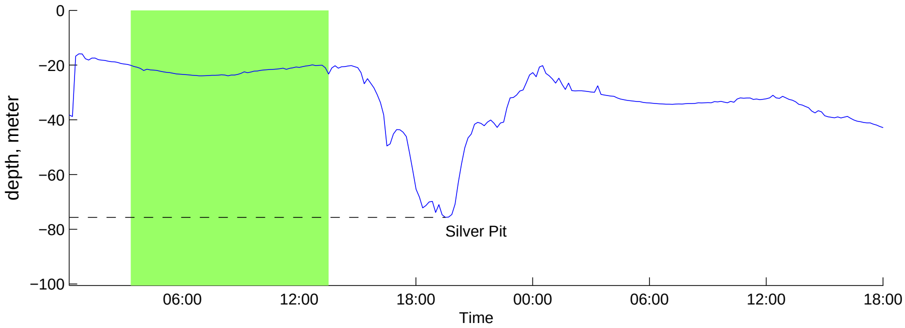

Figure 7.5: *Sample from the depth record of tag #2255, at the 10th of April. The fish stays at* −20 *m of depth until 15:00 and then swims to a depth of* −75 *m and returns around midnight to* −20 *m. The geolocation estimates its position to be in the Silverpit.*

In terms of the geolocation method, this occurrence is interesting because the closest location with a depth of −75 metres is at least 200 km away from the release position according to the bathymetry. The fish reaches this position within seven days, that is from the release at the 3rd of April to the 10th of April. Travelling a distance of this magnitude requires a very determined migration of the cod with a constant high activity for all days. The depth record, however, does not indicate a constant migration behaviour at a high speed.

Apparently it is a mystery but fortunately a part of the interval also holds tidal information that can be used for geolocation, see Figure 7.5. It turns out that the fish is crossing an area called the Silver Pit, which is a submerged valley located east of the English shore at Spurn Head. The geolocation method finds the position in the Silver Pit despite no depth of the observed magnitude is present here according to the bathymetry. This proves the importance of the bathymetry error that compensates for the coarse resolution of the database.

7.2.3 Discussion of results

The results found in this section opens for some topics that need to be discussed.

First of all, is a new geolocation method at all necessary when the Tidal Location Method yields a similar result? Yes, and for several reasons. It is by far preferable to assess the geolocation based on a rigorous statistical framework that excludes subjectivity and opens for an automated process that eventually can be formed into a Matlab applet for easy access. The method gives results both with and without a tidal signal and adjusts the uncertainty thereafter. The estimation of a "biomarker" (the diffusivity) and its uncertainty makes comparison of individuals a straightforward procedure.

The tag revealed that the resolution of the database may occasionally limit the applicability of the model. The omission of deep areas such as the Silver Pit reduces confidence in the bathymetry. Fortunately, the bathymetry error accounts for the effect but a more realistic uncertainty assessment should definitely be possible with an improved bathymetry resolution.

The behaviour of the cod was observed to shift in intervals between migration and a resting/foraging, i.e. a high and low activity level. Modelling the fish with a constant diffusivity over the time at liberty may yield uncertainty estimates that are unrealistically high or low. This is subject is addressed in Chapter 8.

7.3 Cod #1186

This cod was released in the eastern English channel in order to verify a long time claim of fishermen, that cod in this area remain largely stationary.

7.3.1 Inspection of the data

The tag type was Star-Oddi centi. The cod was released at the 11th of March 2005 at 50.3° latitude, 0.5° longitude and recaptured the 20th of January 2006 approximately at 53° latitude, 4° longitude. The depth record for the tag is presented in Figure 7.6 along with the estimated tidal information intervals. The reported release and recapture positions are shown in Figure 7.1.

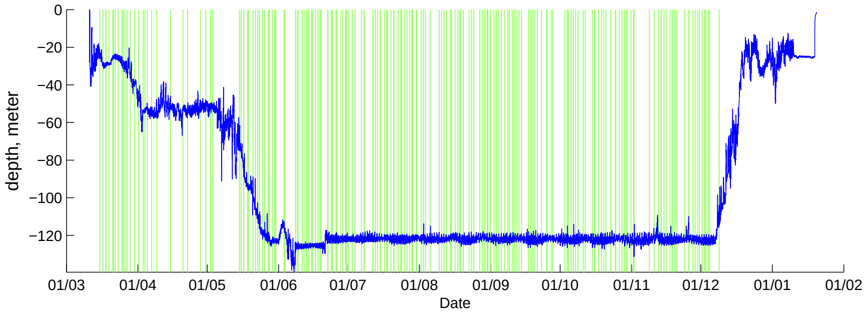

Figure 7.6: *Time series from tag #1186, released 11th of March 2005 and recaptured 20th of January 2006. Tidal information intervals are marked in green.*

The record lasts for 317 days and is very different from #2255, most notably because the cod came from a different population and another ICES management area (VIId), see Figure 1.2. From its release in March, the cod travels towards deeper waters and settles at a depth of 120 metres in mid June. It stays there until December where it ascends and eventually is caught at 25 metres of depth. The recapture position is quite close to an amphidromic point which is also evident from the depth record of the last ten days where only a very vague tidal signal is measured.

Much of the tidal information present in the record have perturbations that can be explained by small scale movements of the fish. This noise causes the extraction algorithm to fail frequently. The tidal wave can occasionally be spotted by the eye. However, the superposition of the movement noise requires a much more advanced fitting algorithm to extract the wave form in an automated process.

7.3.2 Results

The MLE for the diffusivity was 118.9 km^2/day with a standard deviation of 18.9 km^2/day. On average a much more active fish than #2255.

7.3.2.1 Animated marginal posterior distributions

The AMPD show again that the cod has a behaviour that is seasonal dependent. After its release it spends until late May travelling west to its favorite location at the mouth of the English Channel. This location is called the Hurd Deep. It resides in the area at a constant depth for six months before making a rapid migration through the Channel to its recapture position in the southern North Sea. There is no tidal information in the last month of the record and the geolocated final migration is therefore mainly a result of the reported recapture position.

In the six stationary months, the depth record contains plenty of tidal information but the marginal posterior distributions have considerable uncertainty. The fish moves at an iso depth contour where also the phase of the tide is constant thus the marginal posterior distribution is inflated. Another source for increased uncertainty is the amplitude and phase both showing little spatial variation at the position.

7.3.2.2 Most Probable Track

The MPT is shown in Figure 7.7 and is very interesting from a fish management point of view. The figure also shows the ICES areas that each are assigned individual fishing regulations with respect to the species they are inhabiting. Therefore it is interesting to observe that #1186 visits four different ICES areas (VIIh, VIIe, VIId and IVc) in its time at liberty (see Figure 1.2). This result contradicts the claim of the fishermen and shows that the regulation of individual ICES areas should be executed with this type of biomass movement in mind. It may be, that in spring months the cod are found in the English channel, but in the last part of the year they inhabit areas further to the west imploring a segmented regulation.

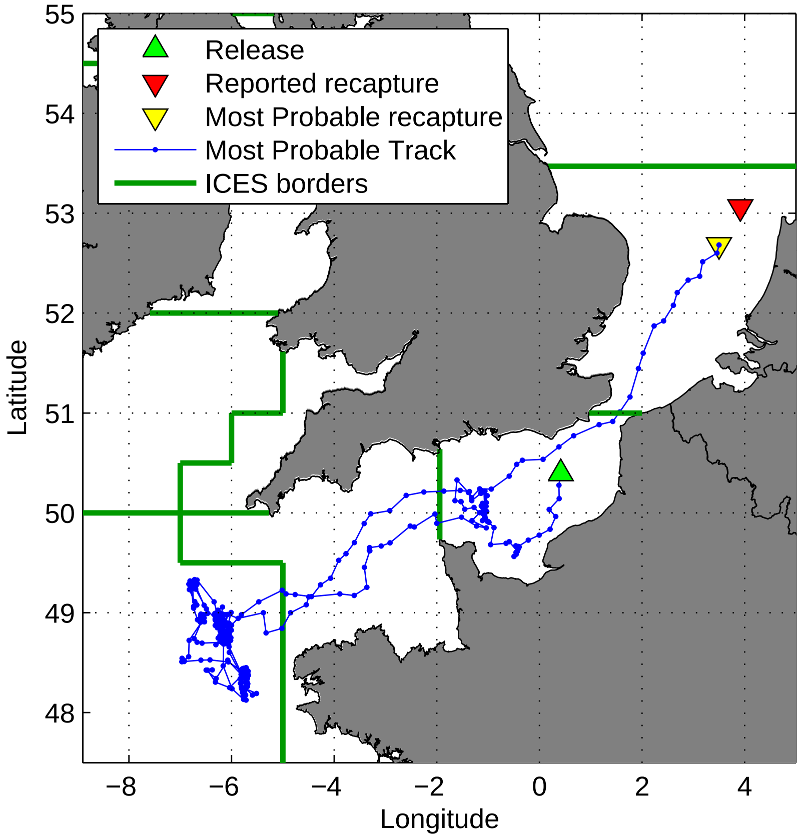

Figure 7.7: *Most Probable Track for #1186. The cod is seen to visit four different ICES management areas within its time at liberty.*

7.3.3 Discussion of results

The main input to the model evaluation given by these results, is the confirmation of the behavioural change also observed in the #2255 tag. The final migration to the recapture position forces the diffusivity estimate to increase and therefore the uncertainty estimates in the periods with low activity is artificially increased. Fortunately much tidal information is usually present in these time periods enforcing a narrow distribution.

The results emphasise the importance of gradients in the environmental variables. This cod spent time in an area with a small tidal range and synchronised tidal phase resulting in a less precise estimate of its position.

The large migrations shown by this cod bring much more detailed information to attention compared to what the same study with conventional tagging data would. The tag gives strong evidence of a biomass that moves between regulation areas which should definitely be accounted for when fish quotas are decided upon.

7.4 Thornback ray #2324

The previous two tags have given a strong affirmation of the validity of the method and the supporting theory. The present tag has been chosen to challenge the model with a highly active depth record. The tag was mounted on a thornback ray (*raja clavata*) which is, like the cod, a demersal species that often dwells at the bottom.

Until recent years, the study of elasmobranchii has been limited to conventional mark/recapture experiments that have only shown minor dispersal of the thornback ray in the southern North Sea. Research has focused on the area around the Thames Estuary that is known as a preferred spawning ground for the species. Results have shown that release and recapture were often in close proximity to each other (Walker et al., 1997). The electronic tagging investigation of Hunter et al. (2005) presents a different result revealing migratory behaviour that could not be assessed by conventional methods.

7.4.1 Inspection of the data

The tag type was LTD 1200. The ray was released at the 6th of October 1999 at 51.63° latitude, 1.14° longitude and recaptured after 504 days at liberty at a position reckoned to be 53.4° latitude, 4.14° longitude, see Figure 7.1. Battery depletion caused the recording to end after 425 days. The recorded depth for the ray is shown in Figure 7.8.

The vertical behaviour of the ray is more erratic than the two cod tags considered so far (#2255 and #1186), and it has remarkably fewer periods of tidal information. Only in August 2000 does it seem to settle at the sea bed for a longer period of time with only few vertical excursions.

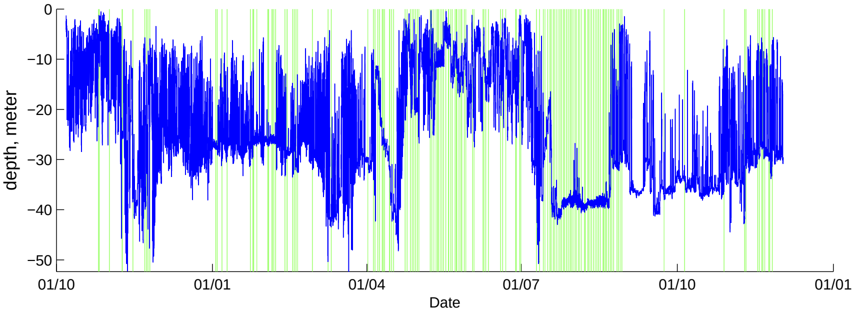

Figure 7.8: *Time series from tag #2324, released 6th of October 1999, battery depleted the 2nd of December 2000. Tidal information intervals are marked in green.*

A close inspection of the time series tells that it is essential to perform the tidal extraction in relatively short time intervals e.g 8-12 hours. The animal displays many periods of nocturnal behaviour and occasionally also indications of tidal stream transport. Very few tidal patterns last for a full 24 hour cycle in contrast to the #2255 tag.

7.4.2 Results

The MLE for the diffusivity was 155.3 km^2/day with a standard deviation of 23.5 km^2/day. The high value of the diffusivity indicates a high average level of activity in agreement with the depth record.

7.4.2.1 Animated marginal posterior distributions

The resulting geolocation has intervals containing multi modal distributions, see Figure 7.9, which is interesting from a modelling point of view. This emphasises the power of this direct solution of the PDE in comparison with the linear results obtained from a Kalman filtering.

In the right pane of Figure 7.9 the fish swims close to the eastern amphidromic point causing the geolocation to suffer an increase in uncertainty despite that tidal information was extracted. It illustrates the influence of the amphidromic point and the ambiguity that it imposes on the geolocation.

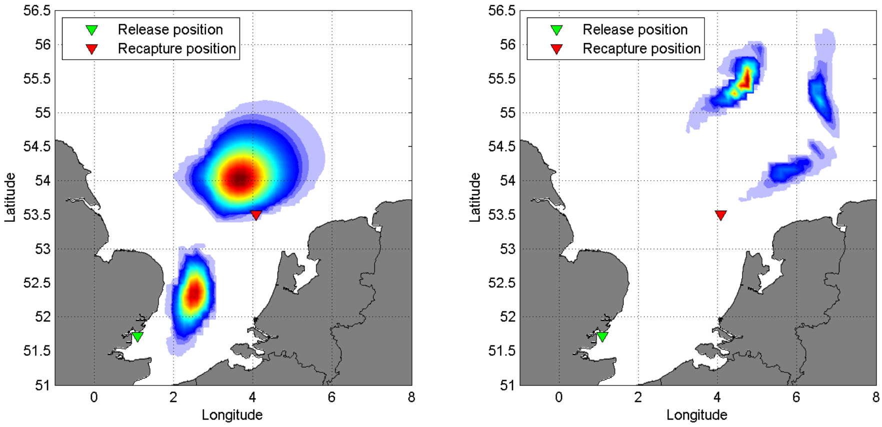

Figure 7.9: *Highly non-Gaussian marginal posterior distribution for #2324. Left pane: the 24th of December 1999. Right pane: 27th of September 2000. Red is most probable white is least probable.*

7.4.2.2 Most Probable Track

The MPT is shown in Figure 7.10 along with a track connecting the mean of the marginal posterior distributions.

Interestingly, the tracks differ significantly. The MPT is without the entire branch to the north of the recapture position which is seen in the mean track. This interval belongs to the final part of the record where tidal information was scarce thus making the geolocation very uncertain due to the large value of $\widehat{D}$.

The MPT sketches a route that, based on the AMPD, would not be expected to be the most probable. This emphasises the fact that the MPT gives the mode of the joint posterior distribution for all positions at all time steps, which can be very different from the track connecting the mean or mode of the marginals. The deviation in tracks happens mostly in time intervals with highly uncertain geolocations, i.e. no tidal information.

7.4.3 Discussion of results

This geolocation shows that the uncertainty is inversely proportional to the amount of tidal data found in depth record. Moreover was it confirmed that the

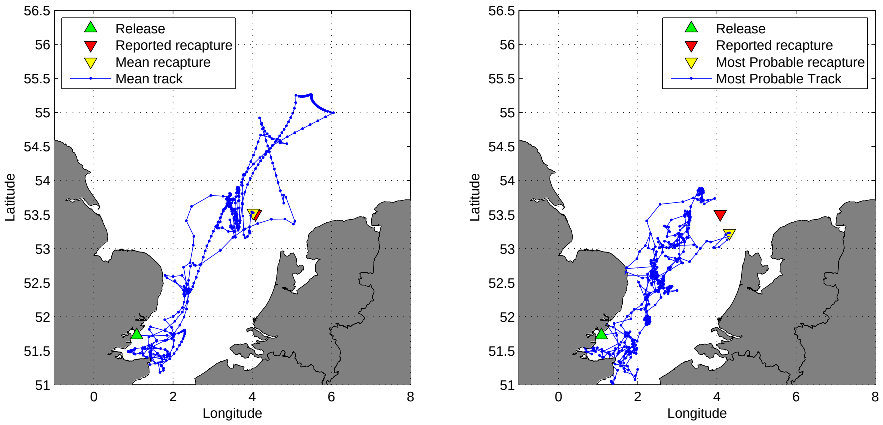

Figure 7.10: *Comparison of track representations for tag #2324. Left pane: Track connecting the mean of the marginal posterior distributions. Right pane: MPT.*

precision of the geolocation is largely influenced by amphidromic points, evident from the AMPD.

Analysis of the data by the TLM gave the same overall results as presented here (Hunter et al., 2005). The fish leaves the Thames Estuary in the winter and relocates further east to an area off the Dutch shore before it returns to spawn in a period from May to July 2000.

The presented MPT seems unlikely judging from the animation of the AMPD or the mean track displayed in Figure 7.10. The MPT was, in the simulation study, shown to be preferable as representative for the joint posterior distribution. It is, however, doubtful whether this conclusion is directly transferable to a real data model. It is a general problem in signal processing and time series analysis to assess the bias introduced in estimates when the data generating system (in this case the fish) differs in behaviour from the assumed model. Extensive research could be done on this subject but it is not within the scope of this thesis.

Conducting a field study to verify a MPT is not feasible with the currently available means. In theory it could be done by mounting an acoustic tag along with a a data storage tag on a fish and follow its movements with acoustic tracking devices. Comparison of the observed path with the estimated MPT from the DST observations yields a measure for the accuracy of the estimate. Such a study spanning possibly months has immense economical costs, which is why the use of powerful statistical methods should be applied to gain maximal knowledge from the DST data.

For a tag encumbered with this kind of substantial uncertainty, expressing the geolocation as a track may have little relevance. Inspection of the AMPD gives much more detailed information of the geolocation e.g. the varying uncertainty of the position dependent on the presence of tidal signal.

7.5 Cod #1432 and cod #6448

The tag type was LTD 1200 and Star-Oddi milli for #1432 and #6448 respectively. In Figure 7.11 is shown the depth record for tag #6448 and in Figure 8.4 for tag #1432.

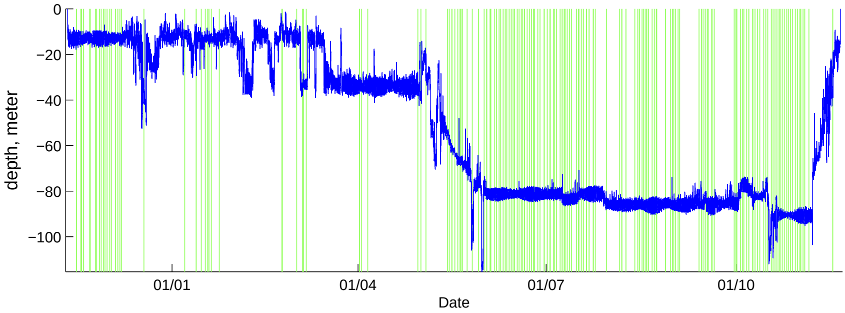

Figure 7.11: *Time series from tag #6448, released 11th of November 2004 and recaptured 20th of November 2005. Tidal information intervals are marked in green.*

The geolocations of these tags support the hypothesis that population conclusions can be made based on electronic tagging experiments. The two cod show a behaviour similar to that of #2255 and #1186. The MPT for both is shown in Figure 7.12.

7.5.1 Results #6448

The cod was released on the 11th of November 2004 at 50.87° latitude, 0.70° longitude, and recaptured the 20th of November 2005 at 50.77° latitude, 0.38° longitude, having a total of 376 days at liberty, see Figure 7.1. The MLE of D

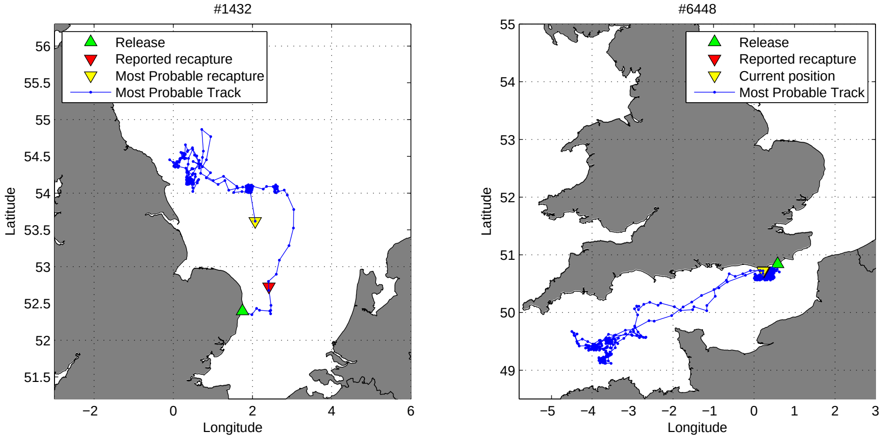

Figure 7.12: *MPT for, left pane: tag #1432. Right pane: #6448. Tag #1432 show a path very similar to #2255 and #6448 show a path similar to #1186.*

was 36.6 km^2/day with standard deviation 5.0 km^2/day.

It is interesting that the cod was recaptured in immediate vicinity of its release position, apparently indicating a stationary fish. The geolocated track however unveils an intervening migration towards the west similar to that of #1186. In the first period after the release the cod inhabits the eastern English Channel. Around early May 2005 it migrates west to the Hurd Deep and stays there during the summer months before returning in early November 2005.

7.5.2 Results #1432

The cod was released on the 30th of March 1999 52.38° latitude, 1.79° longitude, and recaptured the 8th of November 1999 52.68° latitude, 2.3° longitude, having a total of 225 days at liberty, see Figure 7.1. The MLE of D was 77.6 km^2/day with standard deviation 12.8 km^2/day.

The cod is highly active until mid May where it is geolocated just west of the Dogger Bank where it stays until mid October. Here the fish increases its activity level again and the tidal signal is lost. The recapture position was for this tag chosen to be without uncertainty, forcing the geolocation to reach the reported recapture in the final time step.

7.5.3 Discussion of results

It was previously hinted that the recapture position can be incorrectly reported and should not be considered a fixed point in the likelihood update. Objective assessment of the uncertainty of the recapture position is not possible and therefore its variance is manually defined. For #1432 the recapture position is very important due to the lack of tidal signal in the final period of the observations. This leads to the conclusion that the geolocated recapture is entirely dependent on the uncertainty of the recapture position.

The overall pattern of #1432 is similar to #2255 with respect to behaviour and choice of location. The track supports the hypothesis that cod from the southern North Sea tend to migrate north to a summer habitat where they stay until the winter.

The cod captured and released in the English Channel, #6448, agrees with the behaviour estimated from the #1186 tag. Apparently the Hurd Deep makes up an attractive environment for a cod to inhabit in the latter six months of the year.

The results presented in this section demonstrate that reproducible results are indeed obtainable by DST geolocation. Generalisation of behaviour to population level can be done with great confidence based on relatively few tag returns compared to a study based on conventional tagging data.

7.6 Summary of the main findings

The results presented in this chapter showed a stochastic geolocation with a precision that surpass that of previous geolocation methods (Hunter et al., 2003; Nielsen, 2004). Direct comparison is not immediately possible (or fair) due to difference in species and in the environmental variables used for geolocation. That being said, the presented method yields a much more detailed result quantified by an animation of the marginal posterior distributions and the Most Probable Track. The importance of an AMPD increases when tidal data is scarce in the recorded depth because the MPT does not express the time dependent uncertainty of the geolocations.

As the particle filter, the method does not rely on a Gaussianity assumption in contrast to most applications of the Kalman filter. The uncertainty of the geolocation is therefore best viewed by considering the estimated marginal posterior distributions. Different from the particle filter, the method estimates the marginals via direct solution of the diffusion equation thus avoiding the need for an excessive amount of particles. The price is a loss of flexibility if one wishes to extend the space of hidden states.

Estimation of the joint distribution for the positions at all discrete time instances, opens for sampling of random tracks or assessment of the MPT via the Viterbi algorithm. These tracks serve as illustration of possible routes and differ from the mean track when the marginal posterior distributions are highly non-Gaussian.

The geolocations presented here relied purely on depth measurements and release and recapture positions. The combination of demersal fish and an environment with a significant tidal variation makes a perfect setting for this type of geolocation. When the extraction algorithm fails to find a tidal pattern the uncertainty increases dramatically. To encompass an accurate geolocation even in these intervals, the method needs complementary observations of e.g. temperature, salinity of light. This subject is explored in Chapter 8 where temperature observations are experimentally included in the model.

The ML estimates of D are summed up in Table 7.1.

Tag	$\widehat{D}$	sd($\widehat{D}$)	95% C.I.
#1209	5.6	1.1	[3.4;7.8]
#2255	22.4	2.7	[17;28]
#1186	118.9	18.9	[81;157]
#2324	155.3	23.5	[108;202]
#1432	77.6	12.8	[52;103]
#6448	36.6	5.0	[47;57]

Table 7.1: *Maximum likelihood estimates and approximate 95% confidence intervals of the diffusivity parameter D for all considered tags. All units are km^2/day. sd$(\cdot)$ means standard deviation estimated from the observed Fisher information.*

The values of D are anticipated to lie within the interval $[0; 225]$ km^2/day to accommodate the maximal swimming speed of a cod. The estimates of D span the entire available spectrum and no general trends can be drawn.

The stationary tag, #1209, shows that the resolution of the database imposes a lower limit on D that is different from zero.

The estimate of D is a measure for the average activity over the entire period the fish was at liberty. This includes both migratory and resident behaviour in one variable. Judging from the AMPD, the behaviour of most of the considered fishes is split in intervals of high and low activity respectively. Therefore, the single parameter representation seems insufficient for a fair description of the activity level. Doubtless, the introduction of a dual parameter model will yield more realistic uncertainty estimates of the geolocations and give values of D that are easier interpretable. This subject is explored in Chapter 8.

Chapter 8

Model extensions

This chapter addresses some improvements and extensions of the tidal based geolocation model presented in the previous chapters. The extensions are implemented for illustrative purposes and are in some cases simplified to keep the problem tractable. The aim is to sneak peak at some straightforward extensions and to show the versatile structure of the model.

Creating a simple regime model for the activity of the fish is hoped to give more realistic uncertainty estimates where behaviour changes of the fish are accounted for. Another important subject is the use of temperature data to improve the precision of the geolocation. This implementation also corroborates that the filtering technique is applicable to other environmental variables.

The results of the extensions are presented in the final section to give a compact overview of the important features.

8.1 Regime model

As discussed, the behaviour of the geolocated fish tends to be divided into intervals of high and low activity. This observation has been subject to previous investigation in Turner et al. (2002) where DTSs were used to prove a seasonal dependent behaviour. Modelling this behaviour with a single constant diffusivity forces the model to, in some parts overestimate and in other parts underestimate the uncertainty of the geolocation. This issue is unwanted and can be remedied by introducing a new state that describes the activity level of the fish.

The activity state is a time dependent indicator function that on a daily basis describes the activity of the fish as high or low. The state is in principle hidden (not directly observable), which extends the estimation problem of the hidden Markov model immensely. It is therefore sought to classify the activity state before the subsequent filtering and estimation by a preprocessing of the depth record.

8.1.1 Classification of behaviour

The depth record alone has in previous studies proven to hold much information about the activity level of the fish (Righton et al., 2000). In the following analysis, emphasis is put on minimising the probability of misclassification. Two types of error can be committed: a) classifying high diffusivity as low, b) classifying low diffusivity as high. Error type a) is critical and could lead to a completely erroneous geolocation and maybe even render the algorithm unstable. The artificially increased variance of the geolocation that a type b) error would result in, is on the other hand more acceptable.

For each of the 24 hour intervals the classification can be formulated as a hypothesis test where

H_0: The fish has a high level of activity (large value of diffusivity).
H_1: The fish has a low level of activity (small value of diffusivity).

Only when H_0 is rejected at a sufficiently high level of significance can the small value of diffusivity be applied.

The test determining whether H_0 can be rejected is a more subtle subject. An algorithm considering the skewness and range of the depth within a 24 hour period was tried. A fish swimming in mid water with occasional excursions to the sea floor shows a negatively skewed distribution of depth and a fish showing a large depth range and a skew around zero is probably migrating as well. The algorithm was rejected for being too individual specific and for having a high probability of misclassification.

Instead focus was turned to an algorithm that considered the quality of a sine wave fit within the 24 hour interval. Very similar to the linear model in (6.2) used for tidal extraction but with the difference that the interval length, m, of the fit needs to exceed 10 hours. When a fish performs tidal stream transport it is possible that it rests at the sea bed for a longer period, waiting for the tide and then swims at a high speed for a period. Intervals of this type should be modelled with a high value of diffusivity.

Heuristic experimentation with the fitting algorithm resulted in the choice of $m = 96$, that corresponds to a 16 hour fit. This proved as a value that rejected apparent migratory behaviour but allowed for occasional small scale movements.

8.1.1.1 Pruning outliers

A fish resting at the sea bed may make small vertical excursions into the water column, creating “outliers” that deviate significantly from the pattern of the tidal signal. It takes only few of these outliers to make the extraction method of (6.2) fail to see an otherwise clear tidal pattern. Pruning of outliers was unwanted for the purpose of tidal extraction but is essential here in order to capture all relevant intervals with low activity.

Maverick observations can be spotted by considering influence statistics such as Cook’s D, DFBETAS, DFFITS and covariance ratios.
The call `influence.measures` in R[1] returns the potentially influential observations with regards to the fitted model. Further description of the function is found in the R reference manual and the references therein.

The observations classified as outliers are suppressed and a new model is fitted based on the updated dataset. It is assumed that outliers in the initial model can be regarded as outliers in the updated model. From the new fit the summary statistics S ($rmse$) and R^2 are extracted.

8.1.1.2 Classification

The hypothesis, H_0, is rejected if both of the following criteria are fulfilled

$$\begin{aligned} S &< 0.42 \text{ m}, \\ R^2 &> 0.85. \end{aligned}$$

[1] R is a free software environment for statistical computing and graphics, www.r-project.org.

If H_0 cannot be rejected the high diffusivity must be accepted.

The 16 hour window is slid across the time series in steps of 10 minutes analogous to the classification technique described in Section 6.1. This results in an indicator array, $\boldsymbol{BH}$, for the behaviour for which it holds

$$BH_j = \begin{cases} 1 & \text{if } H_0 \text{ is accepted for at least one of } [\boldsymbol{z}_{j,i}, \ldots, \boldsymbol{z}_{j,i+144}] \\ 0 & \text{otherwise} \end{cases} .$$

Remember a 24 hour period contains 144 observations, see Figure 6.4. The value of BH_j determines whether the the prediction $\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1} | \boldsymbol{\mathcal{Y}}_j)$ is computed from a high or low diffusivity behaviour model.

8.1.2 ML estimation of $\boldsymbol{D}$

The parameter space of the model is now extended to $\boldsymbol{D} = [D_0 \ D_1]$, where D_0 and D_1 are low and high diffusivity respectively. The ML estimation of $\boldsymbol{D}$ constitutes a two-dimensional minimisation problem of the negative log-likelihood function, $-\ell(\boldsymbol{D})$.

The problem is handled with the MATLAB function `fmincon`, that is found in the Optimization toolbox. The function finds a minimum of a constrained nonlinear multivariable function. The imposed constraints are the disallowance of negative values in $\boldsymbol{D}$ and the maximal swimming speed limit of 225 km^2/day.

For this medium-scale problem, `fmincon` applies an algorithm based on Sequential Quadratic Programming, Quasi-Newton and line-search. Thorough documentation of the `fmincon` function is found in the help-file for the Optimization toolbox which also includes references.

The results of the implementation is found in Section 8.3.

8.2 Temperature

Installed in some DSTs is a sensor that measures the ambient temperature, $\boldsymbol{q}_j$ at time τ_j. These observations, when added to the observation vector $\boldsymbol{y}_j$, contribute to the *data-update* of the filtering step.

To ease notation define

$$\boldsymbol{v}_j = \begin{cases} [\boldsymbol{y}_j] & \text{for } j = 0 \\ [\boldsymbol{y}_j^T, \boldsymbol{q}_j]^T & \text{for } j \in [1, \ldots, N] \end{cases},$$

and

$$\boldsymbol{\mathcal{V}}_j = [\boldsymbol{v}_0, \ldots, \boldsymbol{v}_j]^T,$$

which contains all observations before and including time τ_j.

8.2.1 POM database

The Princeton Ocean Model (POM) covers a smaller region but with an improved resolution compared to the POL model. More specifically, this is 1/30° latitude and 1/20° longitude which is a resolution of approximately 3.3×3.7 km. The covered area is from 51.02° to 56.48° latitude and −3.93° to 9.53° longitude on a 165×270 grid. Tidal information was obtained from an interpolation of the POL model. A detailed description of the POM database is found in Young (2002).

8.2.1.1 Temperatures

The database package used in this thesis contained temperature predictions for the years 1999-2003 for the North Sea. The temperature predictions are stratified in the water column at six so called sigma levels. A sigma level is a constant percentage of the total depth at the position. For all sigma levels, the temperature is predicted at four points in time on a daily basis, 0:00, 6:00, 12:00 and 18:00 hours. To reduce this very large data set (3.4 billion database entries), only the sigma level at the sea bed was used. According to CEFAS experts this will suffice for illustrative purposes under the assumption that cod stay near the sea bed at summer times, and further that the water is mixed in the winter period and therefore has approximately equal temperature at all sigma levels. A full scale implementation of temperature should naturally include all sigma levels.

8.2.2 Analysis of stationary tags

Analysis of temperature records from the stationary tags aid to the understanding of the errors that measurements and database predictions are encumbered

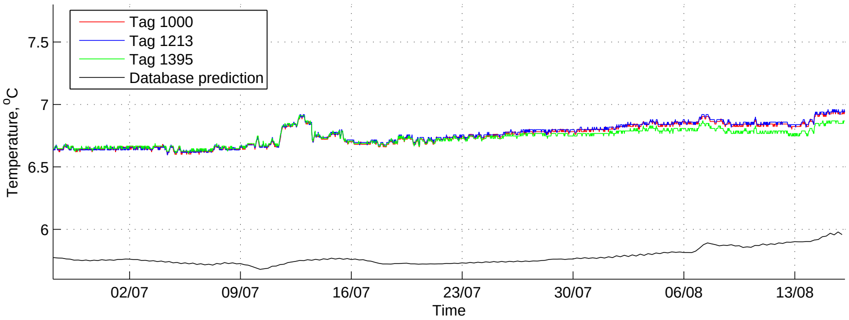

Figure 8.1: *Temperature prediction along with the temperature record from three stationary tags that were deployed at the same position. Time range is 1.5 months.*

with. In Figure 8.1 is shown the temperature record from three tags deployed at the same position, and the corresponding database prediction.

The tag measurements seem to have negligible error compared to the database prediction uncertainty. For this position the bias is approximately 1°C and the empirical standard deviation of the residuals is 0.06°C. Apart from the bias, the temperature is predicted quite precisely. However, the analysis of the remaining stationary tags revealed no consistency in the bias. The experiment was conducted in the summer period where the vertical temperature gradient peaks. This could possibly explain the observed bias at the sea bed level.

The bias was investigated further with the result that it appears consistent at a given position but varies between positions. A complete mapping of this spatial variability in bias is a task too comprehensive for this thesis and is departed here. Instead it is accounted for by increasing the error variance on the observations.

8.2.3 Likelihood for observation

With the addition of the temperature the reconstruction in (3.3) is now changed to

$$\begin{aligned}&\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{V}}_{j+1})\\ &= \psi_{j+1} \cdot \mathcal{L}(\boldsymbol{V}_{j+1} = \boldsymbol{v}_{j+1}|\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1})\mathbb{P}(\boldsymbol{X}_{j+1} = \boldsymbol{x}_{j+1}|\boldsymbol{\mathcal{V}}_{j}),\end{aligned} \tag{8.1}$$

where

$$\mathcal{L}(\boldsymbol{V}_j = \boldsymbol{v}_j | \boldsymbol{X}_j = \boldsymbol{x}_j) = \mathcal{L}(\boldsymbol{Q}_j = \boldsymbol{q}_j | \boldsymbol{X}_j = \boldsymbol{x}_j)\mathcal{L}(\boldsymbol{Y}_j = \boldsymbol{y}_j | \boldsymbol{X}_j = \boldsymbol{x}_j),$$

due to the assumed conditional independence of $\boldsymbol{Q}_j$ and $\boldsymbol{Y}_j$ given $\boldsymbol{X}_j = \boldsymbol{x}_j$.

The new term $\mathcal{L}(\boldsymbol{Q}_j = \boldsymbol{q}_j | \boldsymbol{X}_j = \boldsymbol{x}_j)$ is found in a way similar to the one described in Section 6.6. The temperature observations are divided into intervals of 24 hours and subsampled (by taking average) at the times 0:00, 6:00, 12:00 and 18:00 hours. The observation at time τ_j is assumed to be given by the linear model

$$\boldsymbol{q}_j(\boldsymbol{x}) = \widehat{\boldsymbol{q}}_j(\boldsymbol{x}) + \boldsymbol{E}_j,$$

where $\widehat{\boldsymbol{q}}_j(\boldsymbol{x})$ is the predicted temperature from the database at the position $\boldsymbol{x}$. To keep the parameter space at a minimum, the error $\boldsymbol{E}_j$ is assumed to be Gaussian white noise i.e. $\boldsymbol{E}_j \sim \mathcal{N}_4(0, \sigma_E^2)$, with standard deviation

$$\sigma_E = \begin{cases} 2.5°\text{C} & \text{if the fish is at the bottom} \\ 3°\text{C} & \text{otherwise} \end{cases} . \tag{8.2}$$

The choice of values is based on analysis of the stationary tags. The tidal extraction algorithm is used to determine if the fish is at the bottom.

This error structure should account for, only using sea bed temperatures, diurnal variation in temperature, bias on prediction, measurement noise. It is important to emphasise that the main geolocating variable is still the tidal information, the temperature merely serves to aid this, especially at times when the fish shows a high activity level.

In Figure 8.2 is given an example of $\mathcal{L}(\boldsymbol{Q}_j = \boldsymbol{q}_j | \boldsymbol{X}_j = \boldsymbol{x}_j)$ calculated from tag #2255 at the 18th December of May, 2001.

The results of the implementation is found in Section 8.3.

8.3 Results of model extensions

This section holds the results of the theory from Sections 8.1 and 8.2 applied to the data of cod #2255 and #1432 that both contained temperature sensing devices.

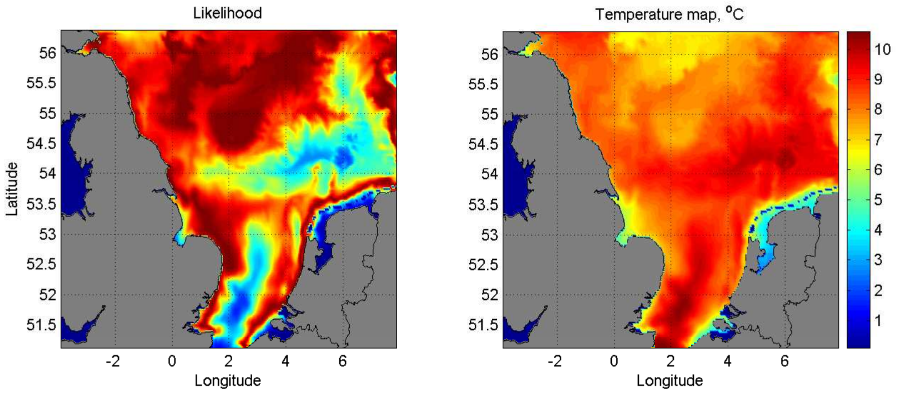

Figure 8.2: *Left pane: Likelihood for a temperature observation over all positions at the 18th of December 2001. Blue least likely, red most likely. Likelihood for the temperature of 7.24° C, measured by tag #2255.*

8.3.1 Cod #2255 extended

The finer discretisation and the second-order optimisation task increase the total computation time significantly. The AMPD have changed remarkably and shows high precision at times of low activity which is quite abundant for this fish. At times of high activity, particularly in the latter part of the record, the uncertainty of the geolocation is increased. The impact of this implementation is pointed out in Figure 8.3. The shown plots has not included temperature observations.

A view of the AMPD for the geolocation, including both temperature and activity regime, shows a jerky distribution at shifts in activity level. To reduce this effect it may be advantageous to change the prediction horizon from 24 hours to 6 hours and obtain a more smooth animation.

The ML parameter estimate was

$$\widehat{\boldsymbol{D}} = [1.17,\ 83.4]\ \mathrm{km}^2/\mathrm{day},$$

which is Gaussian distributed with the covariance matrix

$$\boldsymbol{j}(\widehat{\boldsymbol{D}})^{-1} = \begin{bmatrix} 0.18^2 & -0.29 \\ -0.29 & 14.1^2 \end{bmatrix},$$

where $\boldsymbol{j}(\widehat{\boldsymbol{D}})$ is the observed Fisher information determined from the estimated Hessian. The ML estimate is converted to average swimming speed via an

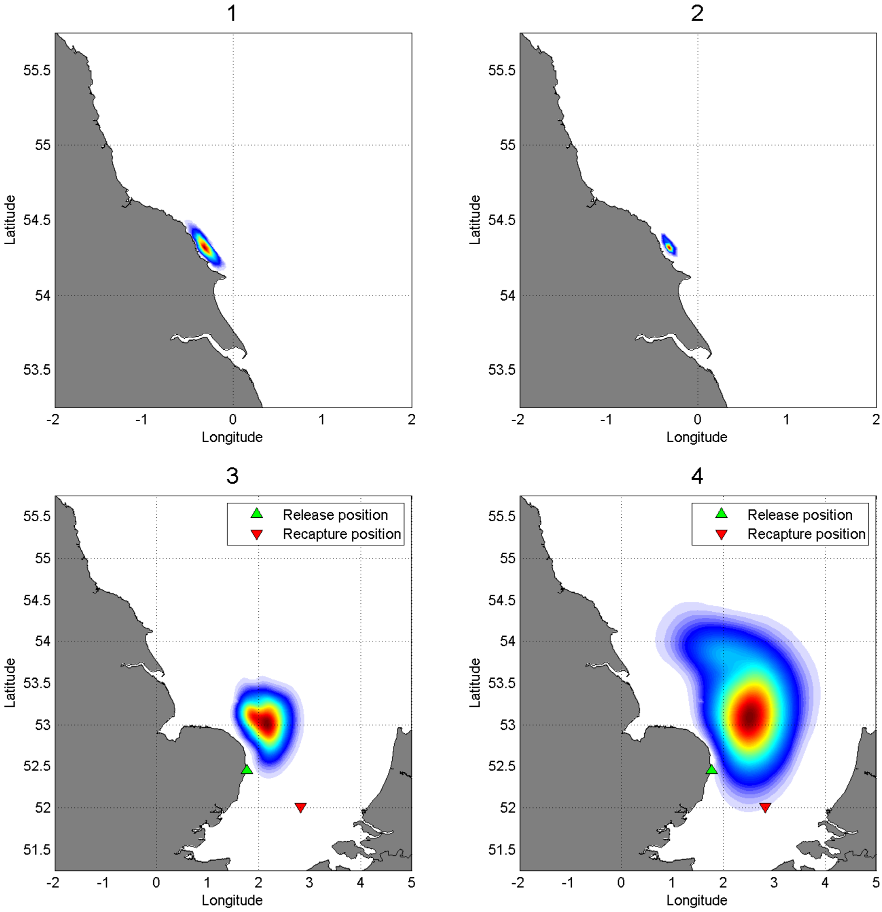

Figure 8.3: *Comparison of the basic and regime model for tag #2255. Top row: Marginal posterior distribution at the 12th of May 2001, low activity, **1**: old model, **2**: regime model. Bottom row: Marginal posterior distribution at the 18th of December 2001, high activity, **3**: old model, **4**: regime model. This calculation has not included temperatures.*

assumed decorrelation time of $\rho = 12$ hours and (7.1), yielding

$$\widehat{v} = [2.2,\ 18.3]\ \text{km/day}.$$

For comparison, the univariate parameter estimation of Section 7.2 resulted in $\widehat{v} = 9.5$ km/day.

It is tested in a Likelihood Ratio Test (Wasserman, 2005) whether the implementation of the regime model has a significant effect on the results. The hypotheses are formulated

$$H_0\colon\ D_0 = D_1, \qquad \text{versus} \qquad H_1\colon\ D_0 \neq D_1.$$

This essentially tests if a two-diffusivity model improves the likelihood of the MLE significantly compared to a one-diffusivity model. The test statistic is found to

$$Z_{LR} = 2[\ell(\widehat{\boldsymbol{D}}) - \ell(\widehat{\boldsymbol{D}}_0)] = 185,$$

where $\widehat{\boldsymbol{D}}_0$ is the MLE under H_0 and $\widehat{\boldsymbol{D}}$ is the MLE under H_1.

The test statistic, Z_{LR}, is χ^2 distributed with one degree of freedom resulting in a p-value for the test of $p < 10^{-41}$, which is highly significant at all reasonable levels. This result provides evidence that #2255 switches its activity level in a way that is well estimated by the classification algorithm of Subsection 8.1.1. It is concluded that the regime model is a considerable improvement with respect to the uncertainty of the geolocation.

8.3.2 Cod #1432 extended

The influence of temperature observations is illustrated clearly by tag #1432. The depth record lack tidal information in the initial and final part, see Figure 8.4, and therefore the basic geolocation model relied heavily on the reported release and recapture positions.

It is expected that temperature observations will reduce the influence of the recapture position. Figure 8.4 shows that the temperature observations initialises around 7 °C and rises steadily to 13.6 °C over a period of 2.5 months. Here it drops abruptly to 8 °C and then continue to rise now more erratic until mid October where a sudden rise of 2 °C occur. Thereafter the temperature slowly declines ending at 12.2 °C at recapture.

The ML estimate of $\boldsymbol{D}$ was

$$\widehat{\boldsymbol{D}} = [0.85,\ 82.0]\ \text{km}^2/\text{day},$$

with the estimated covariance matrix

$$\boldsymbol{j}(\widehat{\boldsymbol{D}})^{-1} = \begin{bmatrix} 0.13^2 & -0.014 \\ -0.014 & 10.4^2 \end{bmatrix}.$$

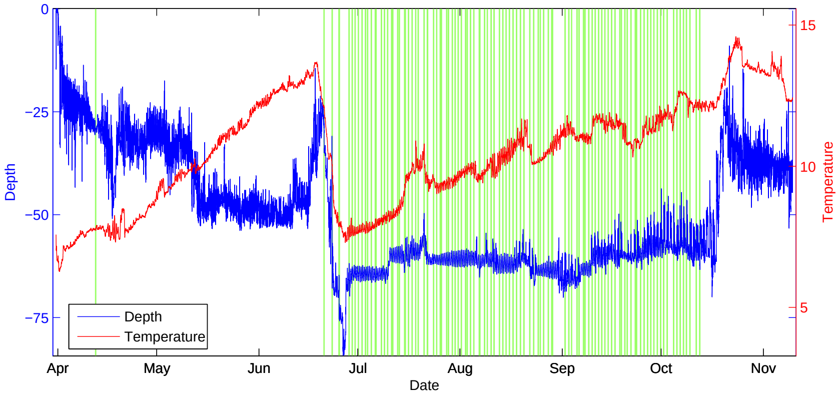

Figure 8.4: *Depth and temperature record for #1432. Intervals where a tidal signal was used for geolocation are marked in green. Time range is 30th of March to 8th of November 1999.*

The ML estimate is converted to average swimming speed

$$\widehat{v} = [1.8,\ 18.1]\ \mathrm{km/day}.$$

The estimate of $\boldsymbol{D}$ seems more realistic compared to the basic geolocation model that gave $\widehat{v} = 17.6$ km/day as average diffusivity.

Again a Likelihood Ratio Test is performed to assess if the two-diffusivity model has improved the uncertainty estimates significantly (for details see Subsection 8.3.1). A highly significant p-value of $p < 10^{-33}$ was found.

Now, apart from having a similar route, the two tags #2255 and #1432 also agree in parameter estimates. Even based on few data it seems reasonable to expect future estimates of $\boldsymbol{D}$ to be in the same order of magnitude.

The change in the geolocations following the inclusion of temperature observation is best displayed by the AMPD. However, also the MPT has changed significantly. For the basic model, the migrations were assumed to happen over a longer period of time due to the lack of tidal signal. The new estimated MPT, see Figure 8.5, shows that the fish stays in close proximity to its release position for two months before travelling north. The path chosen for this migration differs as well. The new MPT estimates a route crossing over the shallow area closer to the shores instead of swimming around, as the old MPT suggests. The return migrations are initialised at contemporary time steps but differ slightly in path.

8.4 Discussion of model extension results

The basic model framework presented in Chapter 6 proved to have much room for expansion of which two important issues were implemented here in a simplified version. However, the change in results was substantial and should not be overlooked.

Parameter estimates of the two component diffusion regime were realistic and resulted in a sensible improvement of the AMPD and their variances. The Likelihood Ratio Tests showed that the change in likelihood was statistically significant and proved that a two-mode regime model is reasonable description of cod behaviour. The parameter $\boldsymbol{D}$ is now independent of the amount of tidal information in the depth record and has the interpretation as the level of activity in each regime. This makes individuals more comparable and future tagging experiments may open for parallels to be drawn from the diffusivity to the physiology and biology of the fish.

The temperature proved particularly useful in time periods without tidal signal where it contributed with a coarse estimate of the position. In this way the migration route was determined more precisely which in the end means that the MPT is more reliable. The added price for a temperature sensing tag is minimal compared to the potential gain in accuracy. The temperature had only little influence on the computation time of a geolocation. One unclarified subject is the is error assessment of the database that seemed much more complex than modelled here. Perhaps inclusion of all sigma levels and further investigation of a larger dataset from stationary tags will improve the understanding of this.

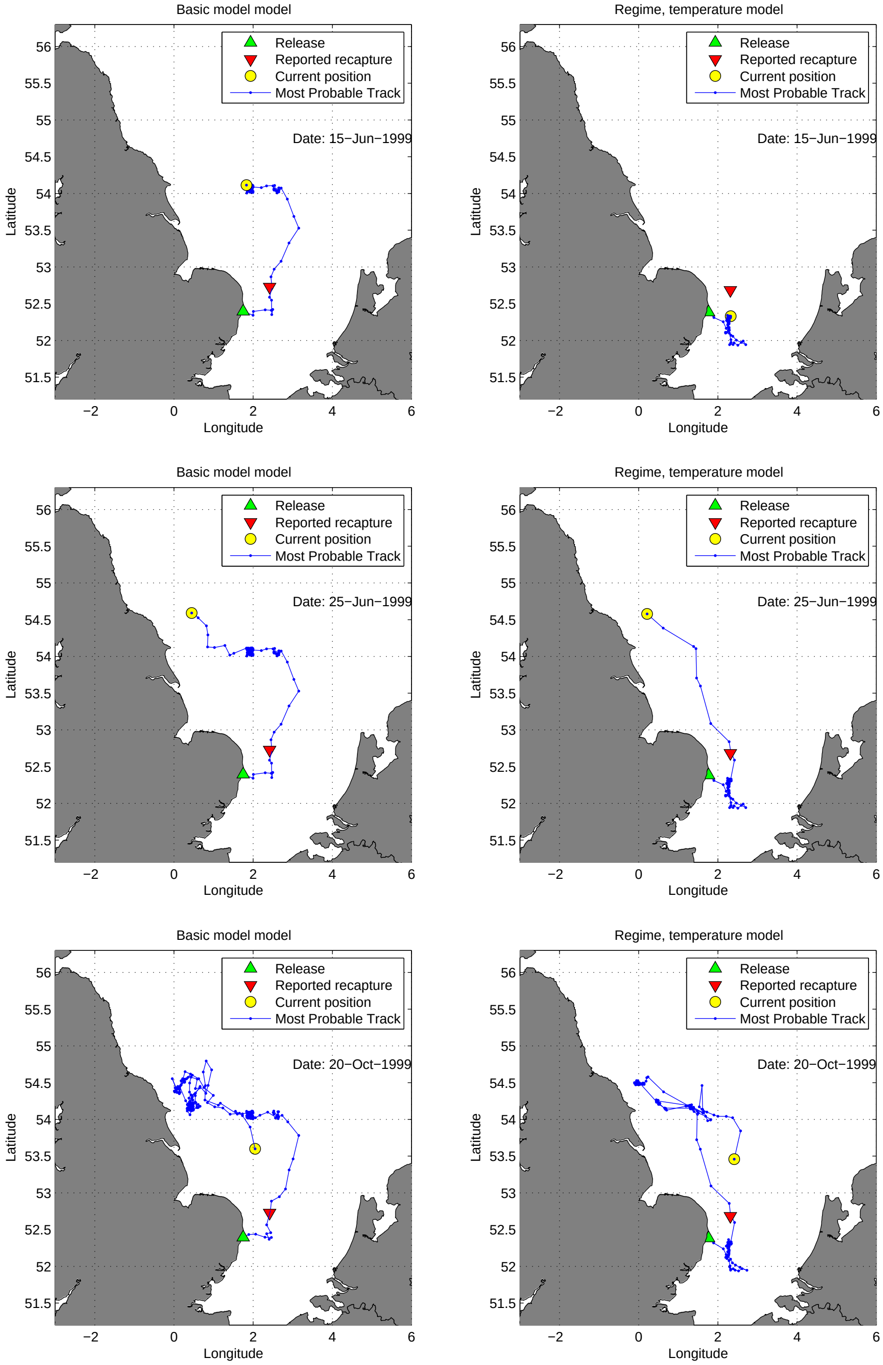

Figure 8.5: *Comparison of the MPT for tag #1432 calculated from the basic model (left column) and the updated model (right column).*

Part III

Outlook and conclusion

Chapter 9

Discussion and future work

This chapter discusses the major contributions of the dissertation and elaborates on the potential of the important matters and erudition that was brought to attention during the work.

Within the field of geolocation it is a classic assumption that the movements of the fish are random, possibly with a bias. The choice is conservative and simplifies the filtering step thus increasing the tractability of the geolocation problem. The validity of this assumption is widely discussed. Critics claim that it is fundamentally wrong to assume a behaviour model that predicts the fish not to move i.e. zero expectancy of the change in position. Furthermore the fish acts to survive and spawn in ways that depend on the environment and the internal biological state, and not randomly as modelled. The argumentation is valid, but at present the simplifying assumptions are a necessity for the geolocation algorithm due to a lack of sufficiently detailed data.

The use of an involved model requires great confidence in its validity and can lead to erroneous estimates if violated. There is no doubt that model complexity can be increased, but is it beneficial? In the present study the advection term was deliberately omitted from the behaviour model. Surely a fish is more advective than diffusive at times but direction and velocity are perturbed with temporal variation appealing to a model with time varying parameters. In future work this is an extension worth implementing. Advection could, experimentally, be included in a seasonal regime model obeying migration trends inferred from previous tagging research. One should be wary though, as this may inhibit the model's ability to discover new trends. The safe choice is therefore the basic diffusion model that comprises any behaviour of the fish, and only restricts its maximal swimming speed by adjustment of the diffusivity parameter.

A large scale implementation of the geolocation method should consider departing the, in some aspects limited, finite difference solution of the diffusion equation. A rectangular discretisation of the domain is required for the convolution operation which inhibits a shift to a continuous representation. Complex boundary geometry, as the one present in the North Sea, is not easily implemented in a finite difference scheme. The land areas were here not implemented in the finite difference scheme meaning the fish in principle could move freely in the domain. The *data-update* step was used to prevent geolocations on dry land. This can in some cases cause the distribution to be artificially repulsed from land areas and thereby introduce a bias in the geolocation. The correct boundary model is *reflecting*, which keeps the fish off dry land and conserves, without renormalisation, the probability mass in the domain.

The method encompassing the aims, not reachable by the finite difference solution, is the Finite Element Method. The method is readily applicable to an arbitrary shaped domain and delivers possibly continuous output result based on local interpolation functions in the elements. The discrete grid can have an arbitrary spatial structure that may be refined in regions of specific interest to obtain a more precise solution. FEM relies on heavy linear algebra operations that is likely to increase computation time, the main drawback of the method. Furthermore, the method rely on more advanced theory which increases complexity in the implementation phase.

For the basic model, computational requirements were not an issue of severe interest. However, with the added complexity of model extensions and estimation of an expanded parameter space, a move to a computational efficient programming language is on a longer term preferable. The need for fast linear algebra operations and a multi-dimensional minimum finding function leads to FORTRAN as the recommended language. It is widely used within scientific computing for demanding tasks and possesses much of the functionality of MATLAB along with modules for minimisation. Another advantage of FORTRAN is the possibility of parallelisation of the geolocation code that would further reduce computation time. Implementation in FORTRAN is a considerable task but will surely turn out beneficial with respect to computational performance.

An illustrative and intuitive presentation of the results is sought in order to communicate broadly the essential findings of the geolocation. Track representations comprising the mean track, the mode track and the Most Probable

Track were evaluated here. It was argued that the MPT is the rational choice for this filtering technique mainly because of its robustness. The computations leading to the MPT are somewhat tedious due to the immense magnitude of the optimisation problem. Variants of the method exists, such as the Lazy Viterbi algorithm, that intelligently reduces computation time but may in rare cases lead to an erroneous track. This may be applied to obtain a fast estimate of the MPT.

A track representation of the results does not describe the uncertainty of the geolocation and may in some cases be very misleading. Optimally, results are given by a MPT combined with an animation of the marginal posterior distributions, possibly supplemented by a sample of random tracks. Probabilities of specific fish behaviour can be directly estimated by such a sample, e.g. the probability of the fish entering a marine protected area or swimming east/west of an island. Immediate access to the estimated joint posterior distribution makes such assessments straightforward to determine for the presented geolocation method.

It was shown that simple inclusion of temperature measurements in the observational likelihood resulted in a significant change in the geolocations. A future full scale implementation of temperature should include all available sigma levels. Moreover is it advisable to conduct a thorough study of the error of the provided forecast model that proved to be perturbed with an inconsistent bias. Some DST dataset include observations of light intensity that could add extra precision to the geolocation. Especially precision of the latitudinal coordinate may benefit from light information and can enhance estimates of migration.

The current tidal extraction algorithm relies on the fit of a linear model to the observations. A high quality fit implies that a tidal pattern is present and that the fish is assumed to rest at the sea bed. The algorithm rarely misclassifies a non-tidal pattern as a tidal pattern but occasionally tidal patterns obvious for the eye are overlooked by the algorithm. Preprocessing of the time series, e.g. by low pass filtering to remove small scale movement noise, may improve results. However, one should act with prudence as chances of misclassification may increase. Certainly, advanced signal processing tools should be applied in further development of the algorithm to ensure that maximal information is extracted from data.

An approximation of the spherical coordinate system of the database was here made as a simplification. The relatively narrow latitude range covered by the North Sea keeps the committed error small. Application of the geolocation method to species in the Atlantic or Pacific oceans might benefit from a mapping of the spherical grid to a rectangular grid to keep D constant in space.

Results of the tidal based geolocation indicated that the recapture position is encumbered with some uncertainty. Probably, the fish considered here are caught from trawl fishing where determination of the exact recapture position is difficult. The results of tag #2255 showed in the final time step a deviation from the reported recapture position that was too large to be explained purely by the uncertainty of the geolocation ($p < 0.01$). This finding was supported by the tag #6448. The importance of an accurately reported recapture position depends on the amount of tidal data in the final part of the record. For some tags (#1432 and #1186) the recapture position and its uncertainty becomes decisive for the geolocation in the end period of the time at liberty. In such a situation the uncertainty of the recapture position has large influence on the ML estimate of D in particular for highly migratory fish.

The presented method has expanded the field of tidal based geolocation and has proven to give results of convincing quality for cod data. Future work should consider applying the method to other demersal species in the North Sea. Also, experimenting with data from other environments and species such as sea turtles, tuna or sharks, can reveal potential areas of application on the longer term.

Chapter 10

Conclusion

The aim of the project was to create a method capable of estimating the probability distribution of the position of a marine animal based on a log file from a data storage tag. For this to be possible the following requirements must be met

- The ambient environment of the marine animal must have sufficient spatial and possibly temporal variation to allow for differentiation between positions.
- Access to prediction models that for a given position can forecast the value of the environmental descriptor chosen as geolocator.
- Access to electronic data storage tags equipped with sensory devices for measuring the relevant quantities.

These characteristics were implemented in a simulation study to assess the performance of the geolocation method. No bias on the maximum likelihood estimate of the diffusivity, $\widehat{D}$, could be proved based on a t-test. An F-test showed that the variance of $\widehat{D}$ is well approximated by the inverse of the observed Fisher information of $\widehat{D}$. Several track representations were investigated. A track connecting the mean of the marginal posterior distributions, a track connecting the mode of the marginal posterior distributions and a track termed the Most Probable Track determined by the Viterbi algorithm. The mean and mode tracks are easy to compute but not suited for possibly multi modal distributions. The Most Probable Track is computationally demanding but, of the three, was shown to give the best representation of the track.

Depth records extracted from DSTs was used as basis for tidal based geolocation. The quality of a least squares fit of a linear model was used to locate tidal patterns in the observations of depth. The extracted tidal data was used as primary geolocator for the method by comparison with tidal predictions obtained from a numerical forecast model created by Proudman Oceanographic Laboratory. Formally, a spatial likelihood distribution for the observation as a function of time, was determined by assuming a linear model for the data. The variance structure of the model was estimated by inspection of stationary DSTs at known locations and by examination of the resolution of the forecast model.

The geolocation method was applied to dataset from four cod and one thornback ray tagged in the southern North Sea and eastern English channel. Estimated marginal posterior probability distributions of the position were presented in the form of an animation. Also, Most Probable Tracks were determined along with estimates of the diffusivity and their standard deviations. For two tags, the cod #2255 and the thornback ray #2324, results were compared with previous findings obtained from the Tidal Location Method. The conclusion was an overall concurrence but the present geolocation method showed improvements with respect to level of detail, e.g. by track representation and uncertainty assessment.

The work with the geolocation method spawned many new ideas for future extensions of which some where implemented with simplifications. During the work an alternative forecast model became available that included temperature predictions on a high resolution grid (approx. 3.5×3.5 km). A linear model for the temperature with Gaussian white noise error were assumed for calculation of the spatial likelihood distribution. The temperature proved influential but should in the simplified case only be considered a supplement to the more powerful tidal information.

The basic geolocation results lead to the conclusion that the behaviour of the fish has large temporal variation. A regime model using high and low values of diffusivity was chosen to comply with this finding. The results of the extended model gave more realistic uncertainty measures and resulted in diffusivity estimates that were independent of the amount of tidal information in the tag. A likelihood Ratio Test showed a significant increase in the model likelihood thus providing statistical proof of shifts in the behaviour.

The statistical basis of the method allows, potentially, for generalisation of multiple concurring geolocation results in a population model. The results presented here showed interindividual reproducibility that agreed with trends seen in conventional tagging experiments. On the longer term these results can aid in the determination of marine protected areas and seasonal fish stock assessment.

Overall, the work resulted in a functional geolocation method that can provide detailed information of the position of a fish based on its depth record. Analysis of DST data recorded in the North Sea proved the method's potential and relevance for application in future geolocation tasks.

List of Figures

Bibliography

Andersen KH, Nielsen A, Thygesen UH, Hinrichsen HH, Neuenfeldt S (2007) Using the particle filter to geolocate baltic cod with special emphasis on determining uncertainty. *In prep* .

Asmar NH (2004) *Partial Differential Equations with Fourier Series and Boundary Value Problems* Pearson Prentice Hall, 2 edition.

Berg HC (1993) *Random walks in biology* Princeton University Press.

Bloch ME (1785) *Ichthyologie ou histoire naturelle des poissons*, Vol. 2 of 6 NYPL, Rare Books Division.

Breen M, Dyson J, O'Neill FG, Jones E, Haigh M (2004) Swimming endurance of haddock (*Melanogrammus aeglefinus L.*) at prolonged and sustained swimming speeds, and its role in their capture by towed fishing gears. *ICES Journal of Marine Science* 61:1071–1079.

Brockwell P, Davis R (1987) *Time series: Data analysis and theory* New York.

Cappé O, Moulines E, Ryden T (2005) *Inference in Hidden Markov Models* Springer.

Chandrasekhar S (1943) Stochastic problems in physics and astronomy. *Rev. Mod. Phys.* 15:1–89.

Conover WJ (1971) *Practical nonparametric statistics* John Wiley & Sons.

Cook RD, Malkus DS, Plesha ME, Witt RJ (2001) *Concepts and Applications of Finite Element Analys* John Wiley & Sons Inc.

Daan N (1978) Changes in cod stocks and cod fisheries in the north sea. *Rapp. P.-v. Réun. Cons. int. Explor. Mer* 172:39–57.

D'Agostino R, Stephens M (1986) *Goodness-of-Fit Techniques* Marcel Dekker, New York.

Einstein A (1905) Uber die von der molekularkinetischen theorie der wärme gefordete bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der Physik* 17:549.

Forney GD (1973) The viterbi algorithm. *Proc. IEEE* 61:268–278.

Grimmett G, Stirzaker D (2001) *Probability and random processes* Oxford University Press, 3rd edition.

Harden Jones FR, Scholes P (1985) Gas secretion and resorpotion in the swimbladder of cod (*Gadus morhua*). *Journal of Comparative Physiology B* 155:319–331.

Hunter E, Aldridge JN, Metcalfe JD, Arnold GP (2003) Geolocation of free-ranging fish on the european continental shelf as determined from environmental variables. *Marine Biology* 142:601–609.

Hunter E, Buckley AA, Stewart C, Metcalfe JD (2005) Repeated seasonal migration by a thornback ray in the southern north sea. *Journal of the Marine Biological Associations, UK* 85:1199–1200.

Hunter E, Metcalfe JD, Holford BH, Arnold GP (2004) Geolocation of free-ranging fish on the european continental shelf as determined from environmental variables ii. reconstruction of plaice ground tracks. *Marine Biology* 144:787–798.

Jonsen ID, Myers RA, Flemming JM (2003) Meta-analysis of animal movement using state-space models. *Ecological Society of America* 84:3055–3065.

Madsen H (2001) *Time series analysis* IMM, 2nd edition.

Madsen H, Holst J (2000) *Modelling Non-Linear and Non-Stationary Time Series* IMM.

Metcalfe JD, Arnold GP (1997) Tracking fish with electronic tags. *Nature* 387:665–666.

Musyl MK, Brill RW, Curran DSG JSHJR, Hill RD, Welch DW, Eveson JPB CH, Brainard RE (2001) Ability of archival tags to provide estimates of geographical position based on light intensity. *Electronic tagging and tracking in marine fisheries.* pp. 343–367.

Neuenfeldt S, Hinrichsen HH, Nielsen A, Andersen KH (2006) Reconstructing migrations of individual cod (*Gadus morhua* l.) in the baltic sea by using electronic data storage tags. *Fisheries Oceanography, in press* .

Nielsen A (2004) Estimating fish movement Ph.D. diss., Royal Veterinary and Agricultural University.

Okubo A, Levin SA (2002) *Diffusion and ecological problems* Springer, 2nd edition.

Rao CR (1965) *Linear statistical inference and its applications* New York: Wiley.

Righton D, Kjesbu OS, Metcalfe JD (2006) A field and experimental evaluation of the effect of data storage tags on the growth of cod. *Journal of Fish Biology* 68:385–400.

Righton D, Quayle V, Hetherington S, Burt G (2007) Movements and distribution of cod (*Gadus Morhua L.*) in the southern north sea and english channel: results from conventional and electronic tagging experiments. *Journal of the Marine Biological Associations, UK* .

Righton D, Turner K, Metcalfe JD (2000) Behavioural switching in north sea cod: implications for foraging strategy? *ICES Annual science conference 2000* .

Righton D, Mills CM (2007) Reconstructing the movements of free-ranging demersal fish in the north sea: a data-matching and simulation method. *Marine Biology* in submission.

Royer F, Fromentin JM, Gaspar P (2005) A state–space model to derive bluefin tuna movement and habitat from archival tags. *Oikos* 109:473–484.

Shumway R (1988) *Applied Statistical Time Series Analysis* Prentice Hall, New Jersey.

Sibert J, Musyl MK, Brill RW (2003) Horizontal movements of bigeye tuna (*Thunnus obesus*) near hawaii determined by kalman filter analysis of archival tagging data. *Fisheries Oceanography* 12:141–151.

Sibert J, Fournier D (2001) Possible models for combining tracking data with conventional tagging data. *Kluwer Academic Publishers, Dordrecht* p. 443–456.

Stokesbury MJW, Harvey-Clark C, Gallant J, Block BA, A. MR (2005) Movement and environmental preferences of greenland sharks (*Somniosus microcephalus*) electronically tagged in the st. lawrence estuary, canada. *Marine Biology* 148:159–165.

Swimmer Y, Arauz R, McCracken M, McNaughton L, Ballestero J, Musyl M, Bigelow K, Brill R (2006) Diving behavior and delayed mortality of olive ridley seat turtles *Lepidochelys olivacea* after their release from longline fishing gear. *Marine Ecology-Progress Series* 323:254–261.

FAO (2004) Fisheries global information system Technical report, Food and Agriculture Organization, www.fao.org.

IUCN (2006) Red list of threatened species Technical report, IUCN, www.iucnredlist.org/info/gallery2006.

Turner K, Righton D, Metcalfe JD (2002) The dispersal patterns and behaviour of north sea cod (*Gadus Morhua*) studied using electronic data storage tags. *Hydrobiology* 483:201–208.

Visser AW, Thygesen UH (2003) Random motility of plankton: diffusive and aggregative contributions. *Journal of plankton research* 25:1157–1168.

Viterbi AJ (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform.Theory* IT-13:260–269.

Viterbi AJ (2006) A personal history of the viterbi algorithm. *IEEE Signal Processing Magazine* 23:120–142.

Walker P, Howlett G, Millner R (1997) Distribution, movement and stock structure of three ray species in the north sea and eastern english channel. *ICES Journal of Marine Science* 54:797–808.

Wasserman L (2005) *All of statistics* Springer-Verlag.

Young E (2002) Tidal validation of a three-dimensional, primitive equation model for the irish and celtic seas region. *CEFAS Publication* .

Index