NOVEL METHODS FOR SPECIES DISTRIBUTION MAPPING INCLUDING SPATIAL MODELS IN COMPLEX REGIONS

Lindesay Alexandra Sarah Scott-Hayward

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

2013

NOVEL METHODS FOR SPECIES DISTRIBUTION MAPPING INCLUDING SPATIAL MODELS IN COMPLEX REGIONS.

Lindesay Alexandra Sarah Scott-Hayward

Thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

in the Schools of Biology and Mathematics & Statistics

UNIVERSITY OF ST ANDREWS

ST ANDREWS

AUGUST 2013

Abstract

Species Distribution Modelling (SDM) plays a key role in a number of biological applications: assessment of temporal trends in distribution, environmental impact assessment and spatial conservation planning. From a statistical perspective, this thesis develops two methods for increasing the accuracy and reliability of maps of density surfaces and provides a solution to the problem of how to collate multiple density maps of the same region, obtained from differing sources. From a biological perspective, these statistical methods are used to analyse two marine mammal datasets to produce accurate maps for use in spatial conservation planning and temporal trend assessment.

The first new method, Complex Region Spatial Smoother [CReSS; Scott-Hayward et al., 2013], improves smoothing in areas where the real distance an animal must travel ('as the animal swims') between two points may be greater than the straight line distance between them, a problem that occurs in complex domains with coastline or islands. CReSS uses estimates of the geodesic distance between points, model averaging and local radial smoothing. Simulation is used to compare its performance with other traditional and recently-developed smoothing techniques: Thin Plate Splines (TPS, Harder and Desmarais [1972]), Geodesic Low rank TPS (GLTPS; Wang and Ranalli [2007]) and the Soap film smoother (SOAP; Wood et al. [2008]). GLTPS cannot be used in areas with islands and SOAP can be very hard to parametrise. CReSS outperforms all of the other methods on a range of simulations, based on their fit to the underlying function as measured by mean squared error, particularly for sparse data sets.

Smoothing functions need to be flexible when they are used to model density surfaces that are highly heterogeneous, in order to avoid biases due to under- or over-fitting. This issue was addressed using an adaptation of a Spatially Adaptive Local Smoothing Algorithm (SALSA, Walker et al. [2010]) in combination with the CReSS method (CReSS-SALSA2D). Unlike traditional methods, such as Generalised Additive Modelling, the adaptive knot selection approach used in SALSA2D naturally accommodates local changes in the smoothness of the density surface that is being modelled. At the time of writing, there are no other methods available to deal with this issue in topographically complex regions. Simulation results show that CReSS-SALSA2D performs better than CReSS (based on MSE scores), except at very high noise levels where there is an issue with over-fitting.

There is an increasing need for a facility to combine multiple density surface maps of individual species in order to make best use of meta-databases, to maintain existing maps, and to extend their geographical coverage. This thesis develops a framework and methods for combining species distribution maps as new information becomes available. The methods use Bayes Theorem to combine density surfaces, taking account of the levels of precision associated with the different sets of estimates, and kernel smoothing to alleviate artefacts that may be created where pairs of surfaces join. The methods were used as part of an algorithm (the Dynamic Cetacean Abundance Predictor) designed for BAE Systems to aid in risk mitigation for naval exercises.

Two case studies show the capabilities of CReSS and CReSS-SALSA2D when applied to real ecological data. In the first case study, CReSS was used in a Generalised Estimating Equation framework to identify a candidate Marine Protected Area for the Southern Resident Killer Whale population to the south of San Juan Island, off the Pacific coast of the United States.

In the second case study, changes in the spatial and temporal distribution of harbour porpoise and minke whale in north-western European waters over a period of 17 years (1994-2010) were modelled. CReSS and CReSS-SALSA2D performed well in a large, topographically complex study area. Based on simulation results, maps produced using these methods are more accurate than if a traditional GAM-based method is used. The resulting maps identified particularly high densities of both harbour porpoise and minke whale in an area off the west coast of Scotland in 2010, that might be a candidate for inclusion into the Scottish network of Nature Conservation Marine Protected Areas.

Declarations

1. Candidate's declarations:

I, Lindesay Scott-Hayward, hereby certify that this thesis, which is approximately 55,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in October 2006 and as a candidate for the degree of Doctor of Philosophy in Biology and Statistics in October 2006; the higher study for which this is a record was carried out in the University of St Andrews between 2006 and 2013.

Date:__________ Signature of Candidate:__________________________

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in Biology and Statistics in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date:__________ Signature of Supervisor:____________________________

Date:__________ Signature of Supervisor:____________________________

Date:__________ Signature of Supervisor:____________________________

3. Permission for electronic publication:

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

Embargo on both all of printed copy and electronic copy for the same fixed period of 1 year on the following grounds:

- publication would be commercially damaging to the researcher, or to the supervisor, or the University;

Date:___________ Signature of Candidate:_____________________________

Signature of Supervisor:_____________________________

Signature of Supervisor:_____________________________

Signature of Supervisor:_____________________________

Acknowledgements

- I owe a huge debt of gratitude to my supervisors, Carl Donovan and Monique Mackenzie for the opportunity they gave me, their hard work on my behalf, advice and friendship, and for the trips to New Zealand for both work and play. Thanks to my supervisor, John Harwood for his wise advice and thoughtful comments, and to Cameron Walker at the University of Auckland for all his help, the time he spared me in Auckland and for his unofficial supervision.

- A special thank you to Laura Marshall, Joanne Potts, Danielle Harris and Faye Donnelly for their support when it mattered and, especially their friendship during difficult times. Thanks also to my oldest friends, Ellena and Jennifer Growcott for remaining so, and for the example of their own dedication in completing their theses.

- Thanks to all at CREEM who through the use of cake and tea have kept me going. In particular Theoni Photopoulou, Angelika Studeny Glenna Evans, Rhona Rodger, Phil Le Feuvre and Lorenzo Milazzo.

- Thanks to my husband-to-be, Andrew Marshall, who made the last few years bearable with his kindness, warmth, encouragement, practical help and so much more.

- Last but not least, to my long-suffering family for their unquestioning support and continued belief that one day I would finally finish!

Data and Funding Acknowledgements

- Thank you to Erin Ashe and Rob Williams for the Killer Whale (*Orcinus orca*) data used in Chapter 4.
- Thanks to the Joint Nature Conservation Committee for use of the Joint Cetacean Protocol data used in Chapter 6.
- Thanks to the Natural Environment Research Council for funding my Ph.D, without which it would not have been possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

General Introduction

The most common definition of ecology is the study of the distribution and abundance of organisms [Andrewartha and Birch, 1954]. For hundreds of years, biologists have conducted field studies to determine the distribution of plants and animals, yet little is known about this for many species. This is, in part, because the fieldwork required to obtain this information takes time, costs money, tends only to occur in accessible areas and rarely covers the entire range of a species. Even when appropriate data have been collected, summarising these data and providing a clear insight into the factors which may determine the distribution of a species remains a challenging task.

Species Distribution Modelling (SDM) has provided a set of analytical tools that can be used to create models of a species distribution using the environmental characteristics of the locations where it is known to be present (and, sometimes, where it is absent). These models, which are usually statistical in nature, can be used to extrapolate species distributions to unsurveyed areas and to document changes in distribution over time. If the fit between the model predictions and the species' distribution is good, the model can also provide insight into the species' environmental tolerances or habitat preferences. SDM also allows the opportunity for prediction to locations or time-scales not surveyed.

The term SDM has been applied to both niche modelling [e.g. Rotenberry et al., 2006] and habitat suitability or distribution modelling [e.g. Hirzel and Le Lay, 2008, Guisan and Zimmerman, 2000]. An ecological niche is the ecological role and space that an organism fills in an ecosystem, and niche modelling attempts to identify the characteristics of this niche. Niche models may estimate:

- a species fundamental (or potential) niche, which is the full range of environmental conditions and resources it can occupy and use;
- its realised (or actual) niche, which is the part of the fundamental niche that an organism currently occupies as a result of limiting factors, such as competition;
- or its climatic niche, the area in which the climate is suitable for the species to succeed.

Habitat suitability modelling is based on the concept of a resource selection function, which describes the factors which determine the probability that a species will occur in a particular habitat [Manly et al., 2002]. In practice, these functions can relate the probability of occurrence to one or more environmental covariates.

Franklin [1995] describes SDM as “geographical modelling of biospatial patterns in relation to environmental gradients”, and it is the definition I have adopted for this thesis. It encompasses both *species distribution modelling* and *predictive* [*distribution*] *mapping*.

SDM is generally applied to presence data because much of the historical data from observational studies or taxonomic records was of this type. However, many studies are now designed to record both species absences and presences at a given location, and, in some cases, the number of individuals of each species. SDM uses models to link, usually sparse, data on species occurrence and abundance with data on environmental covariates, which is often plentiful. There are three main modelling approaches:

- Profile methods (which use presence data): these use the ranges of the environmental covariates that limit the occurrence of a species to define distribution. Values of covariates at a given location are compared with the values of those covariates observed at locations where the species is known to occur, and this comparison is used to determine the suitability of a habitat. One example of this approach is climate envelope modelling, which is based on the climatic niche of a species [Elith et al., 2006]. Geographic models use the location of occurrence points, but not the value of environmental covariates at these locations. The probability of occurrence at a particular point is assumed to be related to the distance of that point from the closest known presence point. Geographic models are not commonly used in SDM because they only describe the survey data itself and have limited predictive ability.

- Regression methods (which use presence/absence or count data) include Generalised Linear Models [GLMs; e.g. Guisan et al., 2002, McCullagh and Nelder, 1989] and Generalised Additive Models [GAMs; e.g. Embling et al., 2010, Hastie and Tibshirani, 1990]. Regression methods assume a continuous relationship between the response data and a set of environmental covariates.

- Machine learning methods use presence/absence or presence/pseudoabsence data, and are also known as data mining methods. The distinction in the literature between machine learning methods and regression methods is not at all clear. Both are distinct from profile methods in that they can be used with data other than presence-only. Regression models typically assume that the data are generated by a pre-specified stochastic model, whereas machine learning methods use algorithmic models and assume that the mechanism which generated the data is unknown. GAMs are very flexible regression models that can be used to fit complex and essentially arbitrary functions to the data. This makes them similar to machine learning methods, indicating that the distinction between regression and machine learning methods is blurred. The most common machine learning method used in SDM is MaxEnt [Maximum Entropy; Phillips et al., 2004, 2006, Elith et al., 2011], which is used to model presence-only data. Further information on machine learning methods can be found in Hastie et al. [2009] and Breiman [2001].

1.1 Why map species distribution?

Distribution maps generated by SDM can be used to:

- improve understanding of the ecology of a species,
- predict species occurrence at locations where survey data are lacking,
- assist in conservation planning and reserve design,
- assess a species' status by comparing past and current maps,
- predict the effects of climate change,
- evaluate the potential impacts of invasive species, to develop ecological restoration programmes,
- and to carry out environmental impact and risk analyses [Franklin and Miller, 2009].

One of the earliest examples of the application of SDM dates back to 1924, when Johnston (cited in Guisan and Thuiller [2005]) used it to predict the spread of the invasive cactus (*Opuntia sp.*) in Australia using correlations between the species' distribution and climate-related covariates. According to Guisan and Zimmerman [2000] and Zimmermann et al. [2010] the use of computers in SDM began in the 1970s when, for example, Nix (1977) made niche-based predictions of the spatial distribution of crop species in Australia [Nix et al., 1977].

Guisan and Zimmerman [2000] provides a review of 'predictive habitat distribution modelling' using a variety of regression based methods. They discuss the idea that the choice of model should not depend solely on statistical considerations but should include some thought about the nature of the species' response. As with most modelling, there is a trade-off between optimising accuracy and optimising generality (i.e. fitting a model perfectly to the data vs finding the underlying function that generated the data). They propose a framework for building an SDM [Figure 1 in Guisan and Zimmerman, 2000] that involves identifying a conceptual model based on the literature or laboratory experiments, this model is then used to inform both sampling design and statistical formulation. A formal version of the model is then fitted to the data, diagnostics are checked, predictions made and model performance evaluated. The same general framework was used in this thesis to analyse the datasets in Chapters 4 and 6.

An international workshop on SDM in 2000 resulted in two special issues of journals, one on the technical aspects of predictive habitat modelling using GLMs and GAMs [Guisan et al., 2002], and the other on the applications of SDM to a variety of terrestrial species data [Lehmann et al., 2002]. The first of these includes papers that discuss the use of GLMs and GAMs for building resource selection functions, whilst others describe the usefulness of these models for zero-inflated datasets that include a higher proportion of zeros than expected, presence-only and remote-sensed data, and problems associated with the incorporation of prediction uncertainties. Some of the applications of SDM in the second special issue include modelling historic distributions, modelling the response of species to climate change, and spatial conservation planning. These are common applications of SDM that are described further in the following sections.

In 2005 Guisan and Thuiller [2005] wrote a detailed review of the ecological principles and assumptions underpinning SDM and highlighted some critical limitations of the approach. They suggest that more care be taken to include information about competition and dispersion, because the boundary of a species' distribution may be determined by competition as well as environmental covariates. Furthermore, there may be issues because of a mismatch between the scales at which species distribution and environmental covariate data have been collected. Similar issues arise when predicting the effects of climate change on species distribution at a local scale. The predictions for climate change are likely to be on a coarser scale than those relating to competition or dispersion. In conclusion, Guisan and Thuiller [2005] believe that SDMs should be developed out of a collaboration between different aspects of biology, ecology, statistics and geography to ensure that they are 'better rooted in ecological theory, more dynamic and multispecific' .

Araújo and Guisan [2006] identified the additional issues of model parameterisation and model selection. A multitude of modelling techniques can be applied to species distribution data and all will give different results. Additionally, even different implementations of the same technique can give different results. This leads to the issue of model selection, both in terms of which implementation and which covariates to select. Model selection should be based both on biogeographical and ecological theory and how much each covariate explains the distribution of the data. One must also accept that there are likely to be strong drivers of distribution that are unavailable or unknown for use as covariates. The issues of model and parameter selection are discussed further throughout this thesis.

Recently, Hawkins [2012] identified a number of assumptions that are made when analysing spatial data. One of particular interest to this thesis is that spatial (and temporal) autocorrelation is widespread in the data used in SDM, but is often ignored. The main issue is a lack of independence in residuals (a common assumption of regression models), which leads to underestimation of standard errors if correlation is positive, and hence an overestimation of the significance of covariates. As Hawkins points out, this is only an issue if selection is done by significance tests. However, we must also consider that any estimates of the uncertainty associated with results, such as plots of confidence intervals, may be misleading. For example, these will be too narrow if correlation is positive. There is further discussion of autocorrelation and appropriate methods in these cases in Chapters 4 and 6.

This thesis focuses on the development of regression methods for modelling cetacean distributions. The aim is to create predictive distribution maps that are as accurate as possible. Clearly, highly accurate maps of any species can be produced if the location of every individual is known. However, this is an almost impossible task, for highly mobile and/or rare species, such as cetaceans. In practice, the information we have on the distribution of such species is patchy. The distribution of such species can still be estimated using modern statistical techniques, although a lack of understanding of their potential flaws may introduce unanticipated biases [Araújo and Guisan, 2006, Potts and Elith, 2006, Hawkins, 2012]. For example, many methods require unlikely assumptions about linearity or the independence of residuals. Other common problems with data collected on marine species that may lead to unrealistic measures of precision include autocorrelation and overdispersion. Both these issues arise in datasets analysed in this thesis and will be discussed in later chapters.

In the following sections, I describe how SDM can be used for three primary applications: analysing temporal trends, environmental impact assessment, and spatial conservation planning. In each application, statistical models that relate species presence/absence or abundance to environmental variables are derived from biological survey data, and these models are then used to fill in gaps in a species' distribution.

1.2 Temporal Distribution Trends

SDM is often used to create an atlas of species' distributions [e.g. Reid et al., 2003] or to map potential future distributions, given a change in environmental conditions [e.g. Teixeira and Arntzen, 2002]. A modelling process, which uses dynamic variables as the basis for mapping, enables prediction of trends and is therefore more flexible than simply mapping the occurrence of species.

Range maps showing the presence/absence of species have long been in use by organisations such as the British Trust for Ornithology or the Royal Society for the Protection of Birds. These have been used to assess historical changes in distribution and as the basis for predictions of future distributions.

Climate change is considered the single greatest long-term threat to birds and other wildlife, with mid-range climate warming scenarios predicting between 15% and 37% of species world-wide will be 'committed to extinction' by 2050 [Thomas et al., 2004]. These predictions rely on data about the relationship between species distribution and temperature. For example, Teixeira and Arntzen [2002] simulated the potential impact of climate warming on the range of the Iberian endemic Golden-striped salamander, *Chioglossa lusitanica*, using distribution models created using GLMs. They produced maps of species distribution extrapolated to the years 2050 and 2080 (equivalent to a rise in temperature of 2^oC and 3^oC), which predicted a substantial range reduction. Similarly, Ferrier et al. [2002] used GLMs to study the effect of climate change on biodiversity in northeast New South Wales, Australia and Pearson et al. [2002, 2004] coupled artificial neural networks with a climate-hydrological process model to identify bioclimatic envelopes for plant species in Great Britain. Araújo et al. [2005] used data on the distribution of 116 species of British breeding birds collected over the last 20 years to compare the performance of different methods for predicting shifts in range. They concluded that artificial neural networks and GAMs provided more accurate predictions than GLMs or classification tree analysis.

Lastly, SDMs have become an established tool for identifying locations where invasive species are likely to become established [Andersen et al., 2004] and for predicting the spread of pest and disease organisms [e.g. Kelly and Meentemeyer, 2002].

1.3 Impact Assessment

Environmental Impact Assessments (EIA) are mainly used to predict or document the potential effect on wildlife of the construction and operation of developments such as oil rigs, wind farms (both marine and land-based), bridges and roads. Ideally, these studies involve a designed survey that is carried out before any construction begins, and is repeated during and after construction. However, many studies only have access to 'before' (or 'after') data for assessing what species occur in an area that may be impacted by the development.

SDM can play an important role in providing spatially explicit predictions of animal presence before or after the development, and for comparing these distributions.

EIAs have traditionally used a Before-After-Control-Impact design [BACI; Green, 1979, McDonald et al., 2000, Smith, 2002, Fox et al., 2006] in order to determine whether a development has resulted in a significant change in the abundance or distribution of the species likely to be affected. However, it is often difficult to define the area over which a development may have an effect and to find a suitable control area that replicates this. Furthermore, BACI designs have little or no power to detect a re-distribution or displacement of animals within an impact area [Underwood, 1992]. It is more realistic to use a Before-After-Gradient (BAG) design [Mainstream, 2009, Barton et al., 2011] which assumes that the effect of a development will decline with increasing distance from the source [Ellis and Schneider, 1997, Morrison et al., 2008], and thus adds some element of spatial structure to the analysis. Displacement and/or habitat loss effects can then be detected [Guillemette and Larsen, 2002]. BACI designed analyses rarely use mapping as an output, but BAG designed analyses use before and after maps to indicate if there has been a re-distribution of the affected species [Petersen et al., 2006, 2011, Barton et al., 2011, Fox et al., 2006], even if the absolute abundance of the species remains the same. Differences in the density estimates can be used to calculate the magnitude of any avoidance effect, not just within the immediate vicinity of the development but also around the edges of the development area. This means there is no need to define a specific ‘impact’ area, as is required for a BACI.

Camphuysen et al. [2004] describe how high resolution large scale mapping of bird densities in marine waters is required to assess the potential impact of offshore wind turbine installations. They suggested that spatial and temporal modelling are the most appropriate methods for assessing changes in seabird distribution and abundance, weather effects, foraging areas and habitat disturbance and loss.

Petersen et al. [2011] highlighted the importance of spatial mapping in the assessment of the environmental effects of an offshore wind farm in the Nysted area of Denmark. They analysed the distribution of long tailed ducks (*Clangula hyemalis*) before and after construction using a GAM based model with spatially adaptive smoothing. There was no change in the absolute numbers of ducks in the study area after construction, but the SDM showed a marked decrease in the number of birds within the footprint of the wind farm. They also detected an increase in numbers in deeper waters. Long tailed ducks are known to prefer shallow waters, where they dive for their food [Nilsson, 1972]. Birds that are displaced to deeper water will probably use more energy in diving and may gain less energy from their prey. The re-distribution and subsequent energy budget issue would not have been identified without the help of SDM.

1.4 Spatial Conservation Planning in the Marine Environment

Results from SDM are often used to develop management frameworks for individual species. These frameworks may include the identification of areas which require protection because the species occurs at high density, or because they are of particular importance to some life history stages [Hoyt, 2012]. Mapping of species distribution plays a particularly important role in the decision making process for the designation of such protected areas [e.g. Embling et al., 2010, Ashe et al., 2010].

Halpern et al. [2008] concluded that "no area [of our oceans] is unaffected by human influence and that a large fraction (41%) is strongly affected by multiple drivers". The United Nations Environment Programme's Global Synthesis report suggests that there has been progress in the establishment of Marine Protected Areas (MPAs) in all parts of the world. However, only 1.17% of global ocean surface and 4.32% of continental shelf areas are currently designated MPAs, which falls well short of the 10% target set at the 7th Conference of Parties to the Convention on Biological Diversity in 2004 [UNEP, 2010]. Thus there is still an urgent need to designate more MPAs.

The UK and Scottish governments have recently begun the process of identifying candidate MPAs (now known as marine conservation zones in England and Wales) to comply with the Marine Scotland Act (2010) and the Marine and Coastal Access Act (2009; England and Wales). Prior to this, they had designated Special Areas of Conservation (SACs) for three of the four marine mammal species listed on Annex II of the EU Habitats and Species Directive (92/43/EEC). The UK government uses the International Union for the Conservation of Nature's (IUCN) definition of an marine conservation zone: "any area of intertidal or subtidal terrain, together with its overlying water and associated flora, fauna, historical and cultural features, which has been reserved by law or other effective means to protect part or all of the enclosed environment". A variety of other types of protected area, such as marine nature reserves and no-take zones, also fall within this definition. A key aspect of MPAs is that they need to be large enough to be biologically relevant but small enough to be managed in a cost effective way.

Designation of MPAs is best achieved through a multidisciplinary approach [Meffe, 1999]. Hoyt [2012] provided a checklist of twenty points for consideration in identifying good MPAs. Many of these points, such as assessing distribution and abundance, commissioning field studies, and determining critical habitat and prey preferences, are related to a species' distribution and can be addressed using SDM. However, local laws and policies, stakeholder involvement and human interactions must also be considered.

MPAs can be a valuable tool for cetacean conservation, but they cannot guarantee a positive conservation result. Rather, they should be considered as part of marine spatial planning process in a broad ecosystem-based management approach. Gormley et al. [2012] describe how the establishment of an MPA for Hector's dolphin (*Cephalorhynchus hectori*) in the Banks Peninsula Marine Mammal Sanctuary in New Zealand has resulted in a 90% probability that survival of Hectors Dolphins has improved between the pre-sanctuary and post-sanctuary periods, with survival increasing by approximately 5%. However, this was the result of a long term study (21 years), which suggests that MPAs should be established with a commitment to long term monitoring. It is also important to manage other potential threats to cetaceans such as overfishing, by-catch, pollution and noise.

At least two marine mammal species, the Yangtze River Dolphin (*Lipotes vexillifer*), in China, and the Vaquita (*Phocoena sinus*) in the northern Gulf of California, have declined dramatically despite some protection, via the creation of MPAs [Hoyt, 2012]. In both cases, the MPA did not adequately cover the species' range and bycatch was not sufficiently controlled.

Examples where SDM has been used to select candidate MPAs for marine mammals include the proposal of a site for the endangered southern resident killer whales (*Orcinus orca*) [Reynolds III et al., 2009] on the west coast of North America [Ashe et al., 2010] and proposed MPAs for harbour porpoise off the west coast of Scotland [Embling et al., 2010]. Ashe et al. [2010] used observations of feeding behaviour to delineate the proposed area, whereas Embling et al. [2010] used animal density. Specifically, Ashe et al. [2010] used a GAM to model the distribution of observations of feeding killer whales in the inshore waters around San Juan Island, Washington State (USA) and adjacent Canadian waters. This model was combined with information on the levels of boat traffic, which may affect feeding behaviour.

Harbour porpoises are the only marine mammal species on Annex II of the 1992 EU Habitats Directive (92/43/EEC) for which the UK government has not designated an SAC. Embling et al. [2010] fitted a GAM to data collected off the west coast of Scotland over a three year period and used this model to identify areas of persistently high relative density of porpoise groups across years.

SDM has also been used to identify regions needed to protect a variety of other marine species. Sanchez et al. [2008] used GAMs combined with environmental variables to model the larval distribution of squid (*Loligo vulgaris*) in the northwest Mediterranean Sea. Louzao et al. [2009] and Rayner et al. [2007] used GLMs to establish habitat associations for Cory shearwater (*Calonectris diomedea*) and the endangered Cook's petrel (*Pterodroma cookii*) respectively. Rayner et al. [2007] showed that predictive habitat models offer an improvement on the more traditional population census methodologies for birds.

SDM can provide information on the distribution of species over time through the inclusion of temporal components. However, as with all statistical models, the accuracy of the resulting models must be taken into consideration. Standard mapping techniques may involve over-simplistic smoothing or violation of standard assumptions, thus introducing biases in the predictions. Furthermore, many studies lack, or have inadequate, estimates of uncertainty.

The chapters that follow focus on the use of SDM for conservation planning and temporal trend assessment. However, the same techniques can also be used for EIA, as described in the Conclusions (Chapter 8).

The distribution of a species may change over time, so there is a need to update the maps used for management decisions at regular intervals. To assist this process, a number of projects have attempted to archive all cetacean survey data in single, large databases, where they can be accessed by all interested parties. For example, OBIS SEAMAP is a web-based database that contains raw survey data for many different cetacean species [Read et al., 2011]. Users of OBIS SEAMAP can refine searches using regions and/or species and download raw data. Hoyt [2005] suggests that the first step in cetacean management is to use this database to identify the available information in an area of interest. There may be several surveys in the same area and results from these surveys have probably been modelled separately to give species distributions. However, ignoring the fact that the surveys overlap is both wasteful of effort and useful information, and limits the coverage of data in the area of interest. It would be useful to combine all the overlapping surveys in some way to provide the user with the distribution needed to begin MPA designation. This issue was also highlighted at a recent meeting of scientists studying turtle distributions off the east coast of the USA (Borchers *pers. comm.*), where multiple, overlapping surveys have been carried out by separate groups. That meeting concluded that a dedicated method for combining survey outputs, such as that developed in Chapter 7, would be a useful tool in assessment of turtle distribution off the east coast of the USA.

This thesis focuses on the development of statistical methods, particularly regression based, for mapping the distribution of marine species. However, the methods developed could equally be applied in many other contexts, including terrestrial ecology [e.g. Maggini et al., 2002, Ferrier et al., 2002, Teixeira and Arntzen, 2002] and demographic studies for example, income data from the 1996 Canadian census [Ramsay, 2002] or foreign resident distribution in Italy [Marra et al., 2011].

1.5 Statistical Issues

A problem frequently encountered when producing distribution maps for marine species is the presence of coastlines or islands with complex topography. This complex topography may exclude animals from certain areas, which are referred to here as 'exclusion zones'. The two case study analyses presented in this thesis (Chapters 4 and 6) are both in regions with complex topography.

Thin Plate Splines (TPS; Harder and Desmarais, 1972), which are currently the method of choice for constructing the density surfaces that form the basis of generalised additive SDM use the Euclidean distance between points to represent similarity. These methods can struggle to produce reliable distribution maps if animal densities are highly variable across exclusion zones [Ramsay, 2002, Wang and Ranalli, 2007, Wood et al., 2008], because the Euclidean distance is not always a realistic representation of the true distance an animal must travel between two points (Figure 1.1). This can result in 'leakage' in the model predictions, where high or low densities in one body of water can unduly influence the density estimates in another body of water from which it is separated by a land mass. The resulting prediction bias is an artefact of the distance measure. An example of this can be

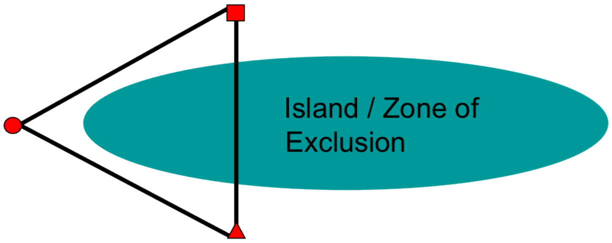


Figure 1.1: An example of 3 equidistant points, where the Euclidean distance between two of the points (triangle and square) crosses an exclusion zone. Realistically the similarity in these two points is the distance between the two without crossing the exclusion zone.

seen in Figure 1.2. The top plot is a simulated 'horseshoe' shaped region [Ramsay, 2002] with a zone between the two arms from which animals are excluded by topography. The bottom plot shows a TPS fit to a sample of 500 observations with low observation error. It is clear that the high values in the upper arm have been under-estimated (yellow at the lower edge of the arm) and the low values in the lower arm have been over-estimated (pale blue at the upper edge of the arm).

Recent alternatives to the TPS method are designed to respect complex boundaries. For example, Finite Element L-Splines (FELS, Ramsay, 2002) utilises a mesh that is constrained to the domain and the observed points within it. The FELS approach has been shown to be a marked improvement over TPS [Ramsay, 2002]. Further details of this method can be found in Chapter 2.

The Geodesic Low rank Thin Plate Spline method (GLTPS; Wang and Ranalli, 2007), involves a mixed model representation of the TPS basis and uses local neighbourhoods around points to estimate geodesic inter-point distances (see Section 2.4.4 for more details). The amount of leakage that is permitted by GLTPS can be small if the size of the chosen neighbourhood is also small, but there is no inherent constraint to prevent leakage across boundaries. This method also requires that a grid is chosen prior to modelling, and the final solution may be sensitive to this choice. Additionally, GLTPS uses a global basis function, meaning individual points can be influential over the entire surface. If the influence of

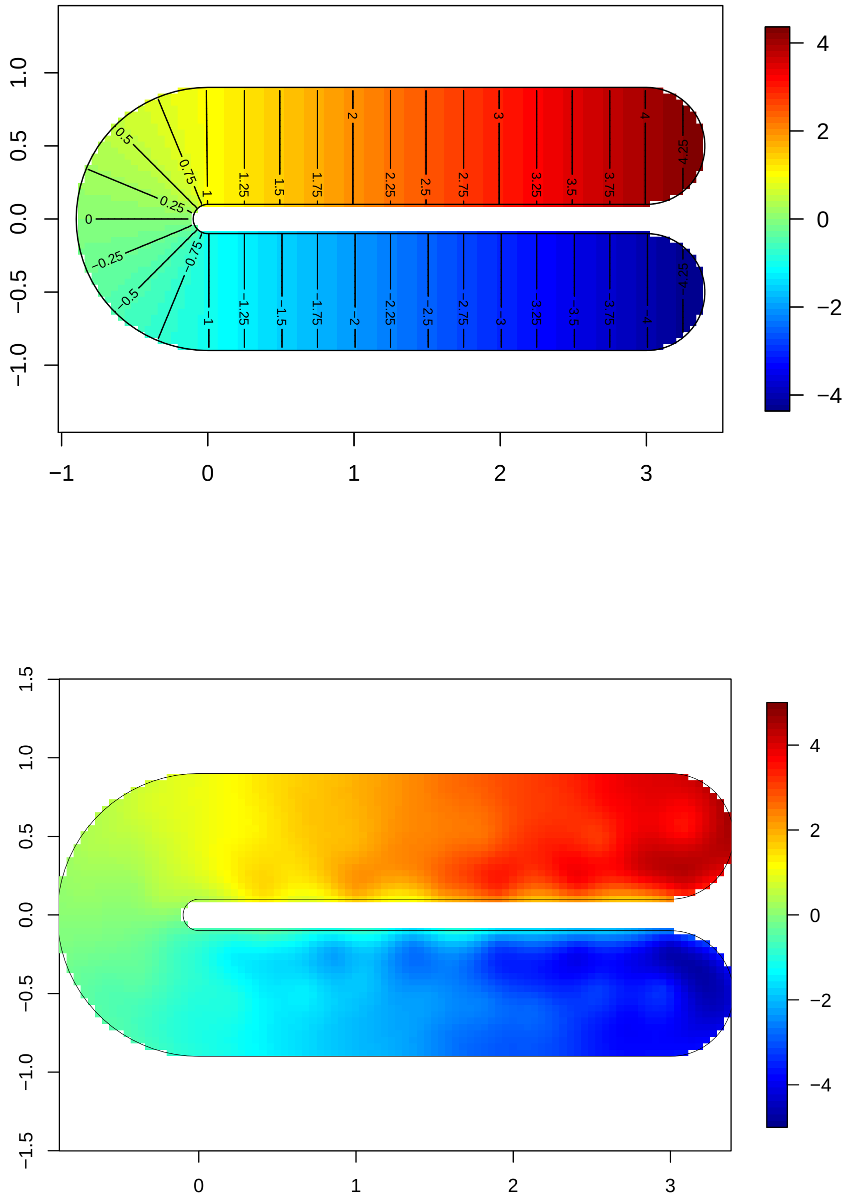


Figure 1.2: A horseshoe shaped simulated region from Ramsay [2002] containing an exclusion zone between the two arms (a). (b) shows a TPS fit to a sample of 500 low noise observations.

individual points is reduced to an area less than the area of the entire surface around each point, the basis function is termed a local basis function.

SOAP film smoothing (SOAP; Wood et al., 2008) uses a specific basis function to model the domain interior, alongside a cyclic penalised cubic regression spline to model each boundary. This method respects boundaries and employs at least two tuning parameters: one global parameter for the interior, and one for each boundary. It has been shown to perform well compared with TPS and FELS, but it has not yet been compared with the GLTPS method. Like TPS and GLTPS, this method is globally acting.

1.6 Thesis Outline

This thesis aims to address some of the problems involved in collating and analysing data on the abundance and distribution of individual marine species using regression based SDM. I review current methods for smoothing in complex areas and introduce a new, and relatively simple method, the Complex Region Spatial Smoother (CReSS), which respects boundaries. I also develop a process by which two overlapping density surfaces can be merged to produce a single, composite density surface.

This thesis relies heavily on smoothing methods and much of the following Chapter (2) is devoted to introducing this topic. Chapter 3 introduces the new CReSS method and compares its performance to that of three other methods using the two dimensional benchmark surface shown in Figure 1.2(a). An additional simulation surface which contains an island is introduced in Chapter 4. I also investigate the effects of data sparsity in a simulation setting. The end of the chapter comprises a case study for designation of an MPA based upon an analysis of feeding behaviour of killer whales off the west coast of North America. The methods and simulations from Chapters 3 and 4 are now published as Scott-Hayward et al. [2013].

In Chapter 5 I investigate improvements to the CReSS model that allow a spatially adaptive surface to be used. In Chapter 6, the methods of Chapters 3 and 5 are used to model the distribution of harbour porpoise and minke whale (*Balaenoptera acutorostrata*) in northern European waters using the Joint Cetacean Protocol (JCP) data resource. Finally, Chapter 7 develops a process for combining two density surfaces using data from the Relative Environment Suitability (RES) database [Donovan et al., 2011] as a template. An index of acronyms and statistical terminology can be found in Appendix A, whilst, appendices E and F, associated with Chapter 6, can be found on the accompanying CD.

Chapter 2

Background Methodology

Smoothing is a dominant theme throughout this thesis and so most of this chapter is devoted to a general review of smoothing techniques. Most of the methods described are extensions of the linear model, however, kernel smoothing is also discussed. Smoothing methods are useful in cases where a line or a surface is sought that is a smooth representation of the data. The goal of smoothing is to produce a graphical approximation of the underlying relationship that is less variable (or smoother) than the data themselves. This allows us to see past the random noise in the data and makes it easier to understand the relationship between response and predictor.

The methods described here begin with linear models and their generalised form (Generalised Linear Models (GLMs); McCullagh and Nelder, 1989) and then move on to Generalised Additive Models (GAMs; Hastie and Tibshirani, 1990), which allow non-linear relationships between predictor variables and the response (Section 2.2). Both GLMs and GAMs are very useful methods and are commonly used to model species distributions [Guisan et al., 2002, Guisan and Thuiller, 2005, Elith et al., 2006]. Different types of splines for use in GAMs are discussed, followed by a section detailing one and two-dimensional kernel smoothing (Section 2.3). I then discuss bivariate smoothing using TPS (Section 2.4) and focus on complex smoothing methods that are able to deal with the issue of smoothing in topographically complicated regions without producing the leakage artefacts mentioned in Chapter 1. The last section of this chapter deals with methods for selecting among competing models (Section 2.5).

2.1 Generalised Linear Models

This section begins with a brief recap of multiple regression linear models, which are a special case of a Generalised Linear Model (GLM, McCullagh and Nelder, 1989) with Gaussian errors and identity link function. Let's assume we have n observations consisting of a response variable, y, and covariates $x_1, ..., x_p$, then the linear regression formula is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \tag{2.1}$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. This can be written in matrix form as $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is a response vector of length n, $\mathbf{X}$ is an n x $(p+1)$ covariate matrix, $\boldsymbol{\beta}$ is a coefficient vector of length $p+1$ and $\boldsymbol{\epsilon}$ is a vector of unobserved errors with length n.

The least squares estimator of $\boldsymbol{\beta}$ is also the maximum likelihood estimator (for normally distributed errors), which is the basis for generalising the linear model. The least squares solution can be obtained by [McCullagh and Nelder, 1989]:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.2}$$

and therefore

$$\hat{\mathbf{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y} \tag{2.3}$$

$\mathbf{H}$ is known as the hat matrix and the least squares residuals are found using this matrix, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{1} - \mathbf{H})\boldsymbol{\epsilon}$.

GLMs [McCullagh and Nelder, 1989] are useful when the errors are known to have a distribution other than the Normal. A GLM consists of a random component specifying the distribution of the errors, a linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$ and a link function $g(\cdot)$. The distribution specified for the errors is from the exponential family, for example, normal (linear model), exponential, gamma, Bernoulli, binomial, Poisson, multinomial and negative binomial. The link function connects the mean response to the linear predictor.

Since the data to be used in Chapters 4 and 6 are count data, the Poisson distribution is shown as an example.

$$y \sim Pois(\lambda)$$

The Poisson distribution allows the variance to increase with the mean and if a log link function is used, for example, the predictions are required to be non-negative. A Poisson model with log link function can be written in terms of the response, y, and p covariates as:

$$y_i = e^{\eta_i} = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}} + \epsilon_i \tag{2.4}$$

or on the scale of the log link function:

$$g(\lambda_i) = \log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \tag{2.5}$$

This shows that the response on the link scale is linearly related to the covariates and, therefore, still a model which is linear in its parameters. The mean and the variance are assumed to be the same and equal to λ ($E[y] = V(y) = \lambda$). However, a common problem in practice when applying a Poisson regression to count data is that the variance increases at a faster rate than the mean ($\lambda < V(y)$). This means that the variance of the response is greater than the variance assumed by the model, a situation known as overdispersion. If this occurs the Poisson model will underestimate the uncertainty in the regression coefficients, leading to potentially misleading inferences. For example, a covariate may appear to be a significant predictor when it is not. The `glm` function in `R` [R Development Core Team, 2009] allows a quasi-Poisson family which is able to adjust the uncertainty depending on the amount of over/under dispersion seen in the data. For quasi-Poisson models, the variance is assumed to be proportional (rather than equal) to the mean:

$$V(y) = \phi\lambda$$

where ϕ is the dispersion parameter, which is estimated during the modelling process. For overdispersion ($\phi > 1$) the standard errors are rescaled, leading to wider confidence intervals and larger p-values for the intercept and slope parameters, relative to the Poisson model.

2.2 Generalised Additive Models

Generalised Additive Models (GAMs; Stone, 1985) are an extension of GLMs that allow the relationship between the response and a covariate to be non-linear. This is achieved through the use of parameter anonymous smooth functions. Smoothing splines typically carry out the smoothing in GAMs, frequently constructed from basis functions (Section 2.2.1). The generalised part of a GAM is the same as for a GLM and refers to the allowance of non-normal errors to be specified.

2.2.1 Basis Functions

Basis functions are a series of functions that collectively span the predictor data range, and are combined linearly to give an appropriate curve for the data. A simple example of a set of bases is a polynomial basis:

$$b(x) = \sum_{m=1}^{M} x^m$$

where M is the degree of the polynomial. For example, a degree 3 set of polynomial basis functions is:

$$b(x) = x + x^2 + x^3$$

Each of these three bases span the entire x-range and are known as ‘global’ bases. These can perform well but are restricted by their global nature; i.e. when fitting these as an ordinary multiple regression, the fitted curve at any point in the x-range is affected by the fitted curve elsewhere. This makes polynomials rather inflexible, but this can be alleviated to some extent by increasing the order of the polynomials. However, high degree polynomials show some oscillatory behaviour [Silverman, 1985]. To alleviate this, polynomial regression can be extended to a series of piecewise polynomials that join smoothly at break points, known as knots [Eilers and Marx, 1996]. For now, the knot locations are assumed given (equidistant, and increasing in x) but in reality their number and location is quite important (see Sections 2.3.1, 2.3.3 and 2.5.1).

There are two main types of knot based bases: truncated power series and b-splines and I will describe each of these in turn.

Let’s assume we have a single covariate, x, and want polynomials in x up to degree 3. A truncated power series basis [Hastie et al., 2009] can be expressed as:

$$b(x) = x + x^2 + \sum_{t=1}^{T} (x - \kappa_t)^3$$

where $\kappa_1 < ... < \kappa_T$ are fixed knots. This gives a third degree polynomial on each interval between two consecutive knots and with two continuous derivatives everywhere. These cubic bases are bounded below (by knot location) but not explicitly above, which can result in very large basis values for large x-ranges. B-splines tend to be preferred since they are positive over only a small subset of the data and between 0 and 1. They are calculated using a recursive relationship [de Boor, 2001].

$$b_{t,j}(x) = \frac{x - \kappa_t}{\kappa_{t+j-1} - \kappa_t} b_{t,j-1}(x) + \frac{\kappa_{t+j} - x}{\kappa_{t+j} - \kappa_{t+1}} b_{t+1,j-1}(x)$$

where

$$b_{t,1}(x) = \begin{cases} 1 & \kappa_t \leqslant \kappa_{t+1} \\ 0 & \text{otherwise} \end{cases}$$

This means that the t^{th} 1^{st} order ($j = 1$) B-spline will have a value of 1 between the two knots κ_t and κ_{t+1}. and zero elsewhere. First order bases (degree 0) are piecewise constants, second order (degree 1) are piecewise linear and give triangular bases between κ_t and κ_{t+2} [Eilers and Marx, 1996]. Figure 2.1 shows an example of degree 1, 2 and 3 B-spline bases with 5 knots each. The bases are shown alongside fit to some simulated motorcycle accident data [Silverman, 1985].

Now that basis functions have been discussed, we can see how they fit into a GAM framework. Let's suppose we have response, y and a single predictor x, we can write the formula for a GAM as follows:

$$y_i = \beta_0 + \sum_{t=1}^{T} \beta_t b_t(x_i) + \epsilon_i \tag{2.6}$$

where b_t are basis functions such as the ones previously described and the βs are the model coefficients. The model residual term, ϵ, may follow any of the exponential distributions mentioned in the GLM section. The additive part is the addition of function terms.

According to Faraway [2006] there are three ways of fitting GAM models in the statistical computing environment, `R` [R Development Core Team, 2009]. The `gam` package is based upon the work of Hastie and Tibshirani [1990]. The `mgcv` package is part of the basic packages that comes with the default installation of `R` and is based upon work by Wood

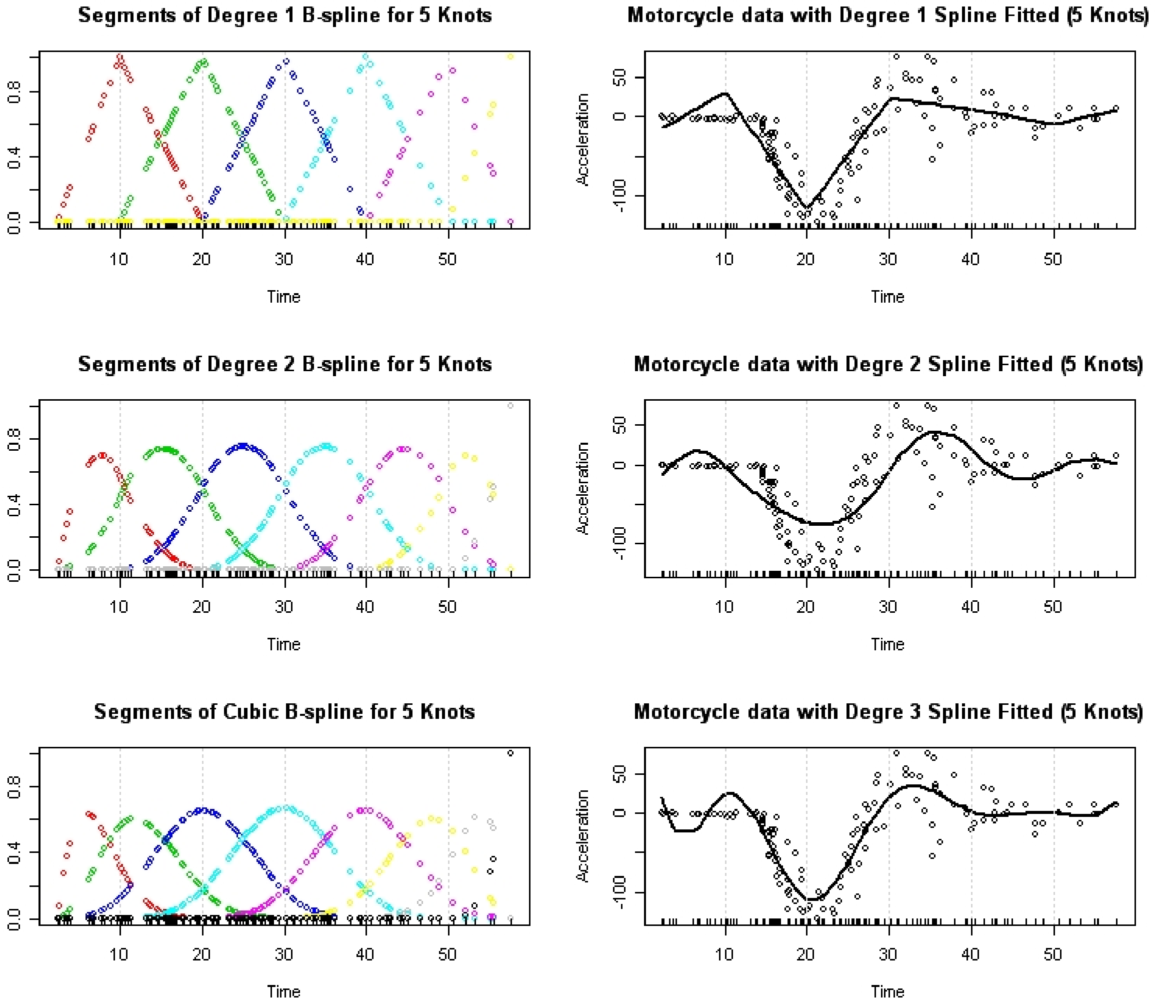


Figure 2.1: An example of 5 knot B-spline bases of degree 1, 2 and 3 (left). The figures on the right are splines fitted to simulated motorcycle accident data from Silverman [1985], depicting acceleration vs time to an impact event.

[2000]. The gam package allows more choice of smoothers (e.g. moving average, local regression, smoothing spline) and a back-fitting algorithm, while the mgcv package has wider functionality, uses a penalised smoothing spline approach for fitting and penalised least squares for estimation. The third package is gss [Gu, 2002], which takes a smoothing spline based approach. These types of spline smoother are covered in Section 2.3.

I have focused on the methodology of mgcv as this package is used throughout the thesis. A brief description of gam, the other commonly used package, can be found in Faraway [2006]. In mgcv, splines are the only choice of smoother and the amount of smoothing is typically chosen internally. Whilst automatic selection avoids the work and subjectivity of making the selection by hand, it can also fail and human intervention may sometimes be necessary. The user must also choose which spline basis to use for each covariate. Some of the commonly used splines included in this package are the cubic regression spline, cyclic cubic regression spline and the thin plate regression spline.

The implementation of GAMs in statistical software packages such as R has simplified the application of the models and led to their increased use in applied fields. Consequently GAMs are frequently presented in the environmental and ecological literature as the final model, often without critical assessment. Despite this, however, GAMs can easily be mis-specified, for example, through an inappropriate choice of smoothness parameter.

The next few sections will describe some alternative smoothing approaches including, regression splines, smoothing splines and penalised regression splines, all of which can be used in a GAM framework, and lastly non-parametric kernel smoothing.

2.3 Smoothing Methods

We consider here a regression problem, where a functional relationship is sought between a single response variable and a set of covariates: the observed data are n observations consisting of the covariate matrix $\mathbf{x}$ (e.g. for 2 covariates, $\mathbf{x}_i = [x_{1,i}, x_{2,i}], i = 1, ..., n$) and the corresponding scalar response y_i. The general model assumed is $y_i = s(\mathbf{x}_i) + e_i$, and the problem consists of approximating the underlying function s from the data in the presence of noise (e_i). The surface approximation, $\hat{s}$, can then be used for prediction or to explain the systematic process generating the observations.

2.3.1 Regression Splines

The parametric approach is to assume that $s(x)$ belong to a parametric family of basis functions. For example, $s(x) = \beta_0 + \beta_1 x$ gives a simple linear regression. The parametric approach is quite flexible because we are not constrained to just linear terms, like in this example. We can add many different types of terms, such as polynomials and other functions of the variable, x, to achieve flexible fits, whilst still in a linear framework. Furthermore, this approach has the advantage that parameters may have intuitive interpretations. Non-parametric methods do not have an easily interpretable equation and therefore the relationship between predictors and the response may have to be described graphically. Since, we can write down the parametric formula, the information required for prediction is greatly reduced from the observed data, to the estimated model parameters. This means that extrapolation and interpolation are both easier with this kind of model.

Using the truncated power cubic basis function discussed in Section 2.2.1 we can construct a cubic regression spline as follows:

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{t=1}^{T} \beta_{2+t}(x - \kappa_t)^3$$

where the β's are regression coefficients and $\kappa_1 < ... < \kappa_T$ are fixed knots. This gives a third degree polynomial on each interval between two consecutive knots and have two continuous derivatives everywhere. The flexibility of the cubic regression spline, for example, relies entirely upon the number and location of the knots, therefore knot specification is of great importance. Too many knots and the model will be over-fitted, resulting in a surface that fits too closely to the data and is therefore too ‘wiggly’ (Figure 2.2, 50 knots). Too few and the result will be a model that is too smooth (Figure 2.2, 3 knots) and leaves pattern in the noise. Ideally, we would like to find, and fit closely to, the underlying function driving the process that generated the noisy data. Furthermore, the fit of the model tends to depend strongly on the locations chosen for the knots. Typically, knots are either spaced evenly throughout the range of x or at quantiles of the distribution of unique x values as in Figure 2.2 [Wood, 2006, Faraway, 2006], but other automated and data-driven methods are discussed in Section 2.5. The degree of smoothing for regression splines is determined by the number and placement of knots but, as is discussed in the next section, we could also use some penalty to determine overall smoothness.

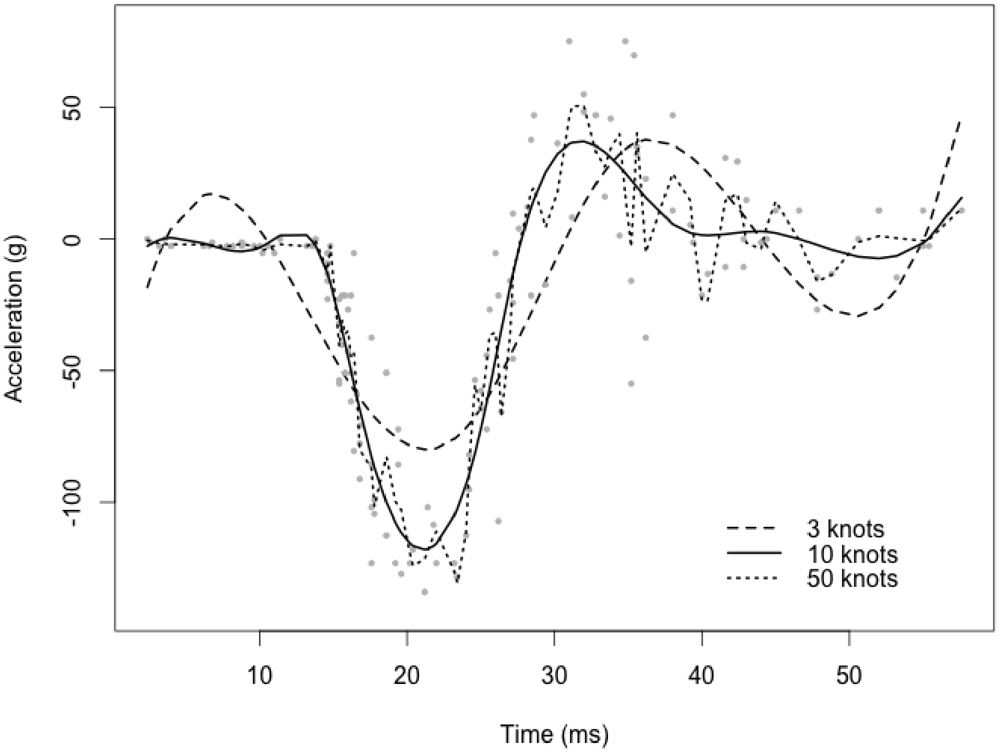


Figure 2.2: An example of cubic regression spline smoothing using three different numbers of knots (3, 10, 50). The data are simulated motorcycle accident data from Silverman [1985], depicting acceleration vs time to an impact event.

2.3.2 Smoothing Splines

Smoothing splines are very similar to regression splines but avoid the knot selection issues encountered with regression splines by having a knot at each unique x-value and adding a penalty term to prevent over-fitting to the data. $s(x)$ can be approximated by minimising sums of squares and a penalty function [Reinsch, 1967]:

$$min\Big(\sum_{i=1}^{n}\{y_i - s(x_i)\}^2 - \lambda\int\{s^{''}(x)\}^2\mathrm{dx}\Big)$$

where $\lambda > 0$ is the smoothing parameter and $\int[s''(x)]^2dx$ is a roughness penalty. A large value of λ means the roughness measure dominates the function to be minimised, resulting in a very smooth curve. Alternatively, a small value of λ results in a very ‘wiggly’ curve, which in the extreme will be an interpolating spline passing through each observation (assuming unique x-values; Green and Silverman, 1994). Therefore, we can balance fit against smoothness. The solution for $\hat{s}$ using this roughness penalty is a cubic spline, so $\hat{s}$ is a piecewise cubic polynomial in each interval (x_i, x_{i+1}). Knots need not be chosen but λ must be specified or estimated. This choice can be made using Cross Validation (CV), which is a popular general-purpose selection method [Faraway, 2006]. Ideally we would use the minimum Mean Squared Error (MSE) to determine fit but this requires the unknown true function $s(x)$. Leave-one-out CV has been shown to be a good approximation of MSE [Hastie et al., 2009]. However, this method of CV is computationally expensive and the more efficient Generalised Cross Validation (GCV; Craven and Wahba, 1979) is commonly used instead [Hastie et al., 2009]. For large sample sizes this has been shown to minimise the MSE but does have a tendency to overfit [Ruppert, 2002]. Methods for selecting the smoothing parameter, including CV and GCV, are discussed further in Section 2.5. It is also worth a note at this point that λ is global and therefore one value dictates model smoothness for the whole x-range. This is not good for approximating a function with locally varying smoothness.

2.3.3 Penalised Regression Splines

Penalised Regression Splines (PRS; Parker and Rice [1985], Wahba [1990], Eilers and Marx [1996]) are a compromise between regression splines and smoothing splines. Therefore, the flexibility of PRS is determined by both knots and λ. They use less knots than one at every unique x-value but still use a penalty to help avoid over-fitting. Generally knots are placed at equally spaced quantiles (see Section 2.5.1 for more on knot selection) and λ is typically chosen using GCV [Eilers and Marx, 1996, Ruppert, 2002]. All chosen knots (number of knots, $T <$ number of data points, n) are included but the influence of each knot is constrained. In matrix form these splines can be written

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda^2 \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S} \boldsymbol{\beta} \tag{2.7}$$

where the first term assesses model fit and the second term penalises models that are too ‘wiggly’. The trade-off between the two is controlled by the smoothing parameter λ. As for smoothing splines, $\lambda = 0$ gives us an unpenalised regression spline, whilst $\lambda \rightarrow \infty$ leads to a straight line estimate for s. λ is squared here to allow the properties of λ to remain the same if a transformation occurs in the x variable [Ruppert et al., 2003]. Matrix $\mathbf{S}$ is the penalty matrix $diag(0, 1_T)$ of size $(T + 2, T + 2)$ where T is the total number of knots.

For a given λ the penalised least squares estimator of β, is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X} + \lambda \mathbf{S})^{-1} \mathbf{X^T y} \tag{2.8}$$

and therefore, similar to a GLM [Ruppert et al., 2003],

$$\hat{\mathbf{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X^T X} + \lambda \mathbf{S})^{-1} \mathbf{X^T y} = \mathbf{Hy} \tag{2.9}$$

where $\mathbf{H}$ is known as the hat or influence matrix. So long as there are enough knots to make the basis more flexible than we think we need, it is now the choice of λ that determines the flexibility of the model. As for smoothing splines, this choice is typically made using CV or GCV.

PRS are a cross between regression splines, where knots are manually chosen and smoothing splines, where every datum is a knot point and over-fitting is addressed using a roughness penalty. Penalisation shrinks all basis coefficients toward zero, whereas knot selection shrinks some coefficients to zero and leaves others unshrunk. Ruppert et al. [2003] showed that, provided the knots in a penalised spline cover the range of x values, the number and position of the knots makes little difference to the results, assuming however, that we are not trying to approximate a function with locally varying smoothness. Once the knots and λ are selected, one must also decide what spline basis to use, for example a truncated power function or B-spline. Often in practice, when n is large, software that uses smoothing splines may actually use PRS. For example, in the `R` function `gam` (in the `mgcv` library) this limit is set to 200, above which the smoothing spline becomes a PRS, and this would seem to have little effect on the end result [Wood, 2006].

2.3.4 Kernel Smoothing

To choose s from some smooth family of functions, we make some assumptions about s, so that it has some degree of smoothness and continuity. With no formulaic output, the relationship between predictors and the response is usually described graphically or as a set of predictions. An advantage of the non-parametric approach over the parametric approach is that less is assumed about the model so we reduce the bias from, for example, the wrong choice of model form.

This section begins with kernel smoothing in one-dimension (one covariate), for simplicity, and then extends the discussion to two-dimensional smoothing (two covariates). Two-dimensional spline smoothing is discussed in the next section. Kernel smoothing fits a different, simple model, separately at each observation point using only those observations closest to the target point. A simple kernel approach is to construct the local mean estimator using observed data denoted by $\{x_i, y_i; i = 1, ..., n\}$,

$$\hat{s}(x) = \frac{\sum_{i=1}^{n} w(x_i - x; h) y_i}{\sum_{i=1}^{n} w(x_i - x; h)}$$

first proposed by [Nadaraya, 1964] and [Watson, 1964]. The kernel function $w((x_i - x); h)$ is usually a smooth positive function which peaks at $(x_i - x) = 0$ and decreases monotonically as $(x_i - x)$ increases in size. The basic idea is to give the most weight to the observations whose covariate values x_i lie close to the point of interest x and less to those that are remote. More is discussed about types of kernel later in this section. The smoothing parameter, h, controls the width of the kernel function and hence the degree of smoothing applied to the data. As the smoothing parameter increases, the resulting estimator may smooth over local features of the data, but if the smoothing parameter is too small, the function will simply interpolate between the observed points, resulting in a very wiggly surface.

The local mean estimator shows some artificial flattening at the boundaries (edges of covariate space) which leads to large bias in this region [Fan, 1993]. An alternative approach, with minimal boundary bias, is to fit a local linear regression and the issue now becomes a least squares problem.

$$min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; h) \tag{2.10}$$

The solution to which, is the local linear estimator $s(\hat{x})$ [Cleveland, 1979]:

$$s(\hat{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{f_2(x; h) - f_1(x; h)(x_i - x)\} w(x_i - x; h) y_i}{f_2(x; h) f_0(x; h) - f_1(x; h)^2}$$

where $f_r(x; h) = \{\sum (x_i - x)^r w(x_i - x; h)\}/n$. A useful property of this local linear estimator is that as the smoothing parameter, h, becomes very large, the curve estimate approaches the fitted least squares regression line. The local mean estimator converges to a straight line parallel to the x axis, with intercept $\bar{y}$, when h is large.

This thesis is mainly concerned with two-dimensional smoothing of space and the local linear approach can easily be extended to non-parametric regression in two dimensions. For observed data denoted by $\{x_{1i}, x_{2i}, y_i; i = 1, ..., n\}$ the weighted least squares formulation is an extension of Equation 2.10.

$$min_{\alpha,\beta,\gamma} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_{1i} - x_1) - \gamma(x_{2i} - x_2)\}^2 w(x_{1i} - x_1; h_1) w(x_{2i} - x_2; h_2) \qquad (2.11)$$

where h_1 and h_2 are smoothing parameters associated with each weight function (one for x_1 and one for x_2). More on smoothing parameter selection is discussed later in this section. A more general two-dimensional kernel function could be used but Bowman and Azzalini [2003] suggest the product of two separate weight functions for each covariate is sufficient. It is often simpler and more compact to define the estimator in matrix notation. Let $\mathbf{X}$ denote an n x 3 matrix whose i^{th} row consists of the elements $\{1, (x_{1,i} - x_1), (x_{2,i} - x_2)\}$ and $\mathbf{W}$ an n x n matrix of zeros with a product of two separate weight functions for each covariate, $w(x_{1,i} - x_{1,1}; h_1)w(x_{2,i} - x_{2,1}; h_2)$, for each of the n observations down the lead diagonal.

Thus the local linear estimator can be written as the first element of the least-squares solution $(\mathbf{X^T W X})^{\mathbf{-1}} \mathbf{X^T W y}$, where $\mathbf{y}$ denotes a vector of responses of length n.

I have outlined the general framework for a local linear approach to kernel smoothing, that will be used in Chapter 7. However there are two parameters which must be chosen *a priori*; the weight function, w, and the smoothing parameter, h. Ideally, we would like a weight function that meets two conditions. Firstly, a smooth weight function results in a smooth estimate and secondly, a weight function that is non-zero only on a bounded interval is preferred to one, for example, approaching zero as $(x_i - x)$ gets large. This means that observations with near zero weight can be ignored, significantly reducing computational speed. The uniform kernel, for example, is compact in its support (Figure 2.3) but can produce a stepped fit, which is, therefore, not smooth (Figure 2.4). For convenience, a Gaussian density function is commonly used as the kernel. K(u) represents a kernel function and for all examples here $u = (x_i - x)/h$.

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - x_1)^2}{2h^2}} \tag{2.12}$$

where h denotes the smoothing parameter and controls the width of the kernel function (Figure 2.4). This results in a smoother looking fit compared with the uniform kernel (Figure 2.4), but does not appear to meet the second of our desired properties: compact support (Figure 2.3). In theory, the contribution of every point must be calculated. However, for the Gaussian kernel, h is the standard deviation of the normal density function and therefore we can show that observations within an effective range of $3h$ in the covariate axis will contribute to the estimate. Observations out-with this range are deemed to have weight near zero and need not be computed.

Another common choice for kernel is the Epanechinikov kernel:

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & |u| < 1 \\ 0 & \text{Otherwise} \end{cases}$$

This kernel shows some smoothness, is non zero on a bounded interval (Figure 2.3) and is computationally rapid. Many kernels will produce similar results and so the choice of kernel is not crucially important.

However, the choice of h is critical to the performance of the estimator [Bowman and Azzalini, 2003]. The aim is to produce an estimate that is as smooth as possible whilst maintaining the 'wiggliness' of the underlying function. This becomes an issue of a bias-variance trade off. As h increases the bias increases due to the inclusion of points far from the point of interest, and the variance decreases due to the effects of averaging. The opposite occurs as h decreases. Ideally, we would like to choose h such that we minimise the Mean

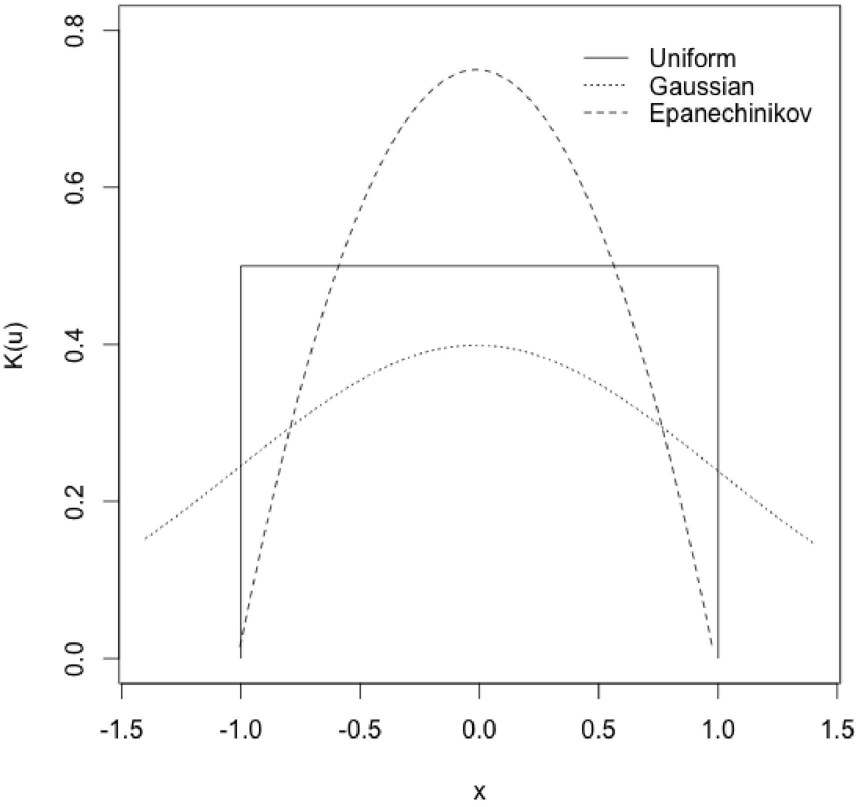


Figure 2.3: Figure showing three types of kernels; Uniform, Gaussian and Epanechinikov

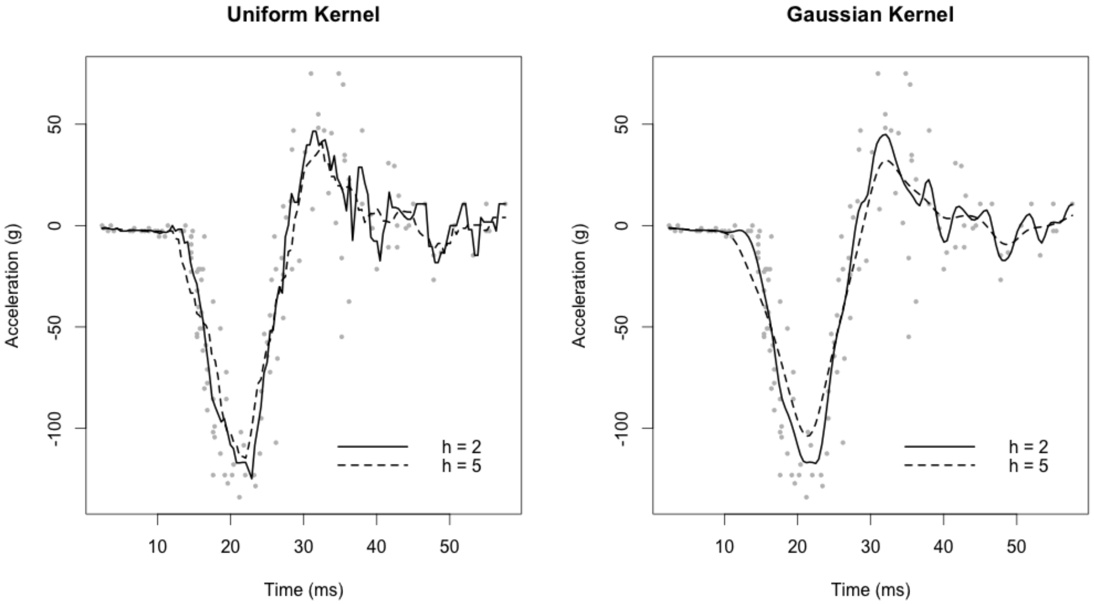


Figure 2.4: An example of kernel smoothing using a uniform kernel (left) and a Gaussian kernel (right). Each type of kernel has been fitted with two different smoothing parameters, h. The data is simulated motorcycle accident data from Silverman [1985], depicting acceleration vs time to an impact event.

Squared Error (MSE).

$$MSE(x) = E(s(x) - \hat{s}_h(x))^2 \tag{2.13}$$

However, this cannot be found in practice since it involves the unknown function $s(x)$, which represents the true underlying function that produced our noisy data.

There are three designs of smoothing parameter; fixed, nearest-neighbour and variable. Fixed selection means that h is constant across the surface but often leads to large changes in variance of $\hat{y}$ due to large changes in the density of the data in the covariate axis, particularly at the limits. Fixed neighbourhoods also tend to contain less points on the boundaries whereas nearest-neighbourhoods get wider to encompass more points. This means the bias for nearest-neighbours is much reduced, particularly in areas of data sparsity. A variable smoothing parameter is one that can vary with x and allow some parts of the curve to be smoother (large h) than others. This is particularly useful for locally adaptive smoothing, when the true s varies a lot, but comes at a cost. Finding an optimal h for every x_i becomes a very computer intensive problem.

Fixed and nearest-neighbour smoothing parameters can be chosen automatically using CV, which is discussed, along with the less computationally expensive GCV method, later in the chapter (Section 2.5). Parameter h is chosen such that this criteria is minimised. Unfortunately, in practice the minimum CV does not always correspond with the minimum MSE and visual assessment is also recommended [Bowman and Azzalini, 2003]. It is important to note that in bivariate smoothing, h must be found for x_1 and x_2, resulting in h_1 and h_2, thus further increasing the computational burden.

Having discussed one and two-dimensional kernel smoothing, we now return to spline based smoothing to discuss both simple and complex bivariate methods.

2.4 Bivariate Smoothing Splines

A spline basis can be used that allows for interaction terms between covariates and so the $b_t(x)$ in Equation 2.6 can be replaced or supplemented with a bivariate basis, $b_t(x_1, x_2)$. Since, one of the aims of this thesis is to look at mapping techniques, we consider a bivariate smooth that allows, for example, an interaction between Latitude and Longitude.

Firstly, the most common bivariate smooth, a thin plate spline, is discussed. There after, the discussion focuses on bivariate smoothing in topographically complex regions.

2.4.1 Thin Plate Regression Splines

Thin Plate Splines (TPS) are a well studied generalisation of a smoothing spline, providing a flexible smooth function in multiple dimensions [Harder and Desmarais, 1972, Green and Silverman, 1994]. Only two-dimensional penalised low rank thin plate regression splines are considered here, where the number of underlying basis functions is less than the set of n observations. As with one-dimensional regression splines, a low rank TPS requires some decision as to the number and location of basis functions - referenced spatially by points called knots, $\boldsymbol{\kappa}_t$ $(t = 1, .., T)$. Since each basis is defined to be symmetric about its knot, $\boldsymbol{\kappa}_t$, they are a type of radial basis function. TPS can be used to estimate the smooth surface, s, by finding the function $\hat{s}(x)$ that minimises Equation 2.7, page 31.

Figure 2.5 shows a graphical example of a TPS basis and the structure can be written:

$$\mathbf{b}(d_{i,t}) = d_{i,t}^2 \log d_{i,t} = (\|\boldsymbol{\kappa}_t - \mathbf{x}_i\|)^2 \log(\|\boldsymbol{\kappa}_t - \mathbf{x}_i\|) \tag{2.14}$$

where $\mathbf{x}_i = [x_{1,i}, x_{2,i}]^T$ and $\boldsymbol{\kappa}_t = [\kappa_{1,t}, \kappa_{2,t}]^T$ are coordinates in $\mathbb{R}^2$. Variable $d_{i,t}$ represents the distance, in this case Euclidean, between the t^{th} knot ($\boldsymbol{\kappa}_t$) and i^{th} datum ($\mathbf{x}_i$) [Harder and Desmarais, 1972, Green and Silverman, 1994].

Given $\boldsymbol{\kappa}$, the regression spline equation in a GAM framework for the smooth surface $\hat{s}$ at a point $\mathbf{x}_i$ using this low rank radial basis is

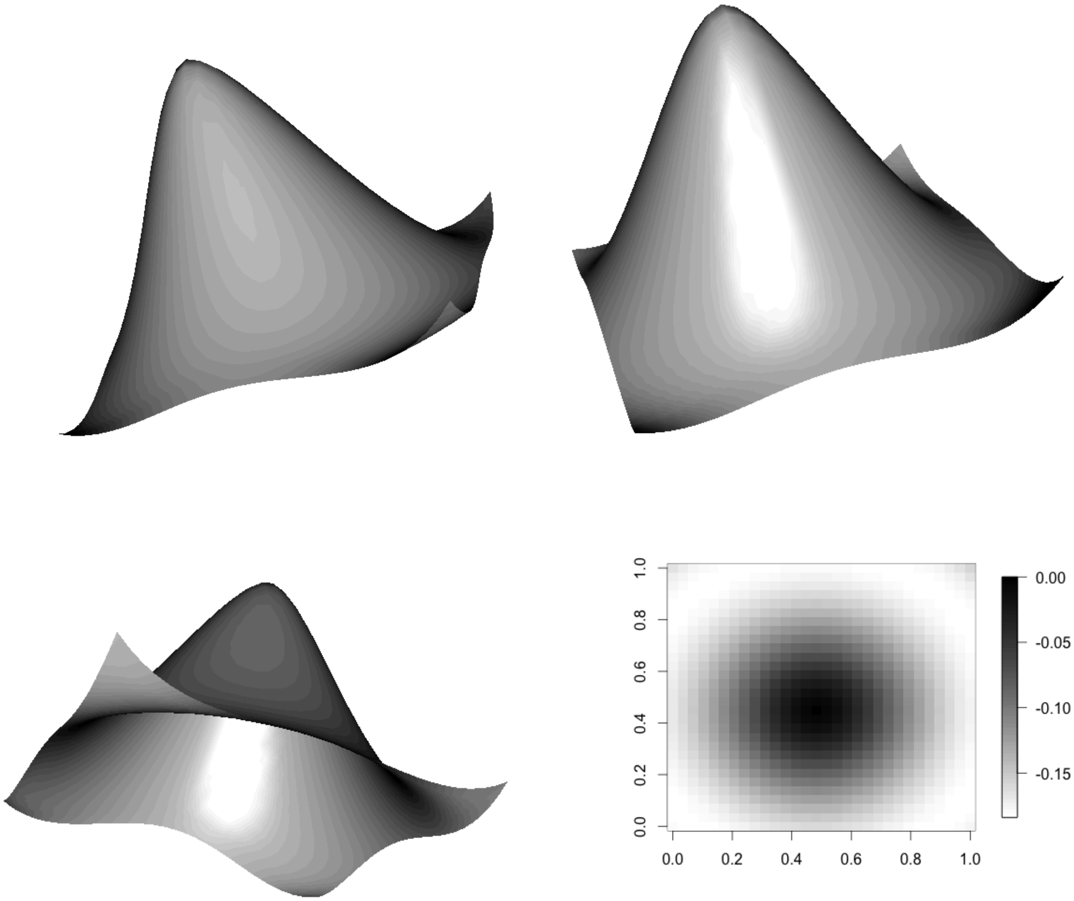


Figure 2.5: A graphical representation of a single thin plate spline basis function

$$\hat{s}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \sum_{t=1}^{T} \hat{\delta}_t b_t(d_{i,t}) \tag{2.15}$$

where the $\hat{\beta}$'s and $\hat{\delta}$'s are estimated coefficients. Knots, $\boldsymbol{\kappa}$, must be chosen *a priori* and selection procedures are discussed in Section 2.5.1. This method does not address the problem of leakage, as seen in Figure 1.2, Chapter 1. Several new methods exist for dealing with this and they are described next.

2.4.2 Review of Complex Bivariate Smoothing Methods

The methods described so far do not address the problem of 'leakage' in the model predictions. As we saw in Figure 1.2 (Chapter 1), 'leakage' occurs when high or low densities in one area can have undue influence across a boundary, such as a coastline, into another area. This section describes three methods that are designed to model these complex topographical areas and are therefore referred to as 'complex' methods. The performance of these methods, alongside TPS, is assessed in Chapters 3 and 4 .

Finite Element L-Splines (FELS, Section 2.4.3) allow for complex topographies by using a mesh that is constrained to the domain and the observed points within it [Ramsay, 2002]. The FELS approach has been shown to work very well [Ramsay, 2002] but it requires the estimated function to meet the boundary at right angles. More on this condition is discussed in the next section.

Another recent alternative, the Geodesic Low rank Thin Plate Spline method (GLTPS; Wang and Ranalli, 2007), involves a modification to the TPS basis and uses the neighbourhood around each point to estimate geodesic distance between points when constructing the TPS basis (see Section 2.4.4 for more details). The amount of leakage that is permitted by GLTPS can be small if the size of the neighbourhood chosen is also small, but there is nothing to explicitly prevent leakage across boundaries. This method also requires that a grid is chosen prior to modelling and the resolution of this grid partly determines the extent of leakage. Another drawback of their implementation of this method is that it uses a basis function which is global in nature and model coefficients are estimated using observations both near to and far from the specified knot locations.

The most recent alternative to conventional TPS is SOAP film smoothing (SOAP; Wood et al., 2008) which uses a soap basis to model the interior alongside a cyclic penalised cubic regression spline to model an unknown boundary (Section 2.4.5). This method respects boundaries and has been shown to perform well compared with TPS and FELS but has not yet been compared with the GLTPS method. Additionally, it respects boundaries but employs a global smoother which may struggle to approximate surfaces with spatially varying complexity. We will now discuss each method in turn in more detail.

2.4.3 Finite Element L-Splines (FELS)

FELS [Ramsay, 2002] utilise a mesh that is constrained to the domain and the observed points within it. The FELS approach has been shown to be a marked improvement over TPS [Ramsay, 2002]. The FELS method uses a bivariate L-spline and then finite element analysis is used to find a solution to the resulting partial differential equations. The domain, A, is covered by a system of triangles, and the basis functions are piecewise quadratics that have a value of one at a vertex and decrease to zero on each of the distal edges [Ramsay and Silverman, 2005]. A description of L-splines can be found in Wahba [1990].

The L-spline smoothing function is a tool for estimating smooth univariate curves from data of the form $\{(\mathbf{x}_i, y_i), ..., (\mathbf{x}_n, y_n)\}$ and is contained within the roughness penalty. For two dimensions the bivariate L-spline approximation to s is the function $\hat{s}$ which minimises the functional

$$\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda \int_A (L_p s) dA \tag{2.16}$$

L_p is a Laplacian operator, Δ, with possibly non-constant coefficients [Heckman and Ramsay, 2000] and must be chosen carefully so that the minimiser to Equation 2.16 does not depend upon the choice of coordinate system (Ramsay 1999). Δ is defined by $\Delta s = s_{x_1x_1} + s_{x_2x_2}$ for all s [Ramsay, 1999]. Finite element analysis is used to find the simplest bivariate L-spline function, that will minimise Equation 2.16. A more detailed description of this method can be found in Ramsay [2002].

The major disadvantage of the FELS method is the strong boundary condition. The normal derivative of s must be zero on the boundary of A and therefore the contours of the estimated function must meet the boundary at right angles. Whilst FELS outperforms TPS in complex regions, the boundary condition limits its performance compared to other complex region methods.

I included, here, only a brief description of FELS since a comparison is not included in this thesis because the conclusions drawn are very similar to those of Wang and Ranalli [2007] & Wood et al. [2008]. Specifically, GLTPS (Section 2.4.4) and the SOAP (Section 2.4.5) both clearly outperform the FELS method.

2.4.4 Geodesic Low-Rank Thin Plate Splines (GLTPS)

The geodesic distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ in a region A, in $\mathbb{R}^2$, is the length of the shortest path between $\mathbf{x}_i$ and $\mathbf{x}_j$ that lies within A. If A is convex then the geodesic distance equals the Euclidean distance. Wang and Ranalli [2007] describe GLTPS within a mixed model framework using a modified version of low rank thin plate splines (regression splines), where an estimated geodesic distance is used to determine the similarity between all observations and knot locations.

The calculation of an accurate estimate of geodesic distance can be complicated and time consuming. Wang and Ranalli [2007] estimate the geodesic distance by viewing the data set of n points as a set of vertices in a graph. Edges are included between every data point and its w closest data points (using Euclidean distance to measure closeness). This permits calculation of a matrix of distances between the $(i, j)^{th}$ pair of points, restricted to paths involving this set of edges. The resulting restricted inter-point distances are equal to the Euclidean distance if there is an edge between them, and infinity otherwise.

Floyds algorithm [Floyd, 1962] is then used to establish the shortest path between points based upon this restricted distance matrix. Floyds algorithm is described in detail in Appendix B. Wang and Ranalli [2007] recommend using the smallest w for which there are no infinite values in the shortest path distance matrix (all points can be reached from every other point).

The mixed model representation of low rank TPS with geodesic distances is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}^{*}\mathbf{u} + \epsilon \tag{2.17}$$

where matrix $\mathbf{Z}^{*}$ is defined:

$$\mathbf{Z}^{*} = [C(|x_i, \kappa_t|_G)][C(|\kappa_t, \kappa_{t'}|_G)]^{-1/2} = [g_{i,t}^2 \log g_{i,t}][g_{t,t'}^2 \log g_{t,t'}]^{-1/2} \tag{2.18}$$

κ_t are the knot locations $| \cdot |_G$ denotes geodesic distance, $i = 1, .., n$ and $t = 1, ..., T$. The function C is the same as the TPS basis function in Equation 2.14, but with geodesic distance ($g_{i,t}$) between the t^{th} knot ($\kappa_{\mathbf{t}}$) and i^{th} datum ($\mathbf{x}_i$) or between two knots ($\kappa_{\mathbf{t}}$, $\kappa_{\mathbf{t}'}$), replacing $d_{i,t}$. For a given number of knots, knot placement is chosen using a space-filling design by John et al. [1995].

While the GLTPS technique has been shown to perform better than TPS and FELS it does not preclude the shortest distance between two points crossing a boundary. The choice of w is fixed for the entire surface and represents a trade-off between accuracy and computational feasibility. Ideally, w is small so that in areas where the exclusion area between boundaries is small, the possibility and extent of leakage is also small. However, if w is too small, the points in the network may be poorly connected and result in distances larger than they should be. Furthermore, if w is too big then points are connected directly by Euclidean distance and boundaries will be breached.

To alleviate problems associated with relatively small w, Wang and Ranalli [2007] use a pre-defined grid over the region. The finer the grid, the greater the likelihood of the distances between points converging on the true geodesic distance. While this provides a lower likelihood of leakage, grid resolution is, in practice, constrained by computational resources. Notably, Wang and Ranalli [2007] do not explicitly include the boundary points in the grid.

Owing to leakage and plotting artefacts created using distances calculated by Wang and Ranalli [2007], the GLTPS method in this thesis uses an alternative method of calculation of geodesic distance (Section 3.2). Thus, the modelling framework of GLTPS is assessed without being compromised by geodesic distance calculations.

2.4.5 Soap Film Smoother (SOAP)

SOAP uses the same GAM framework for fitting as the TPS but specifies a soap basis rather than a TPS basis [Wood et al., 2008, Wood, 2010]. SOAP smoothing is constructed using two sets of basis functions; one for the interior region of interest and one for finding values on each boundary. These are then summed to form

$$s(x_1, x_2) = \sum_{j=1}^{J} \alpha_j a_j(x_1, x_2) + \sum_{t=1}^{T} \gamma_t g_t(x_1, x_2) \tag{2.19}$$

where the γ_k and α_j are the parameters to be estimated. The boundary basis is the first part in Equation 2.19, where a_j are known cyclic cubic spline basis functions for J knots.

For the internal part of the smooth, a set of functions $\rho(x_1, x_2)$ are found such that they are each solutions to the Laplace's equation in two dimensions

$$\frac{\delta^2 \rho}{\delta {x_1}^2} + \frac{\delta^2 \rho}{\delta {x_2}^2} = 0$$

except at each one of the knots. Then Poisson's equation is solved in 2-dimensions

$$\frac{\delta^2 g_t}{\delta {x_1}^2} + \frac{\delta^2 g_t}{\delta {x_2}^2} = \rho_t(x_1, x_2)$$

for T knots. When the boundary condition $\rho_t(x_1, x_2) = 0$ is applied, the set of basis functions for the soap film smoother, $g_t(x_1, x_2)$ is found.

Therefore, knots must be chosen for the internal basis and for every boundary basis constructed. For further details of this method refer to Wood et al. [2008].

2.5 Model and Parameter Selection Methods

This chapter has, so far, been concerned with descriptions of several different methods of smoothing. However, we must be able to determine the performance of each of our models to find the best combination of parameters that gives the best trade-off between fit to the data and the underlying function. This is a form of model selection and attempts to achieve a balance between goodness of fit and parsimony. Better fits to the data can be achieved by adding more parameters but parsimony gives us simpler and easier to interpret models. We could exactly work out model fit if we knew the underlying function that gave rise to the data we have. For most data this is unknown and we must make use of an empirical (data-based) information-theoretic approach. When we simulate data as in Chapters 3 and 4 we know the underlying function and therefore we can make use of the MSE (Equation 2.20). This considers differences in predictions from the underlying function and can be calculated for data points and out-of-set prediction locations. Out-of-set refers to prediction points that are not also data points. MSE is calculated at each location and a mean taken to get an average fit for the whole surface. For simulations this is calculated for a given set of model parameters, θ (e.g. knot number, h, λ):

$$\widehat{MSE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{y}(x_i;\theta) - y_i^*\}^2. \tag{2.20}$$

where y_i^* is the true function value at x_i. However, we don't usually know the true function and must assess our model in another way. We can calculate the fidelity of the model to the data, using Residual Sums of Squares (RSS, Equation 2.21). This is a measure of predictive ability at the data points, is relatively simple and does not require knowledge of truth.

$$\widehat{RSS}(\theta) = \sum_{i=1}^{n} \{\hat{y}(x_i; \theta) - y_i\}^2. \tag{2.21}$$

where y_i are the observed response values i.e. y^* + error. Unfortunately, RSS is not very suitable for model selection because it measures fit to the data and not to the underlying function or data unseen by the model. This is because the minimiser for RSS is at the interpolant ($\hat{y}_i = y_i$), which leads to the smooth that is closest to interpolation (smoothing parameter = zero). CV achieves a solution to this problem by splitting the data into two sets. The model is fitted to the first set (training set) and predictions are made to the second set (validation set). The process is repeated for multiple training and validation sets. The predictions can then be compared to the actual observations in the validation set. The model of choice is the one that minimises some summary of the error. Leave-one-out CV has a training set of $n - 1$ data points and a single validation point. It is defined as:

$$\mathrm{CV}(\theta)\mathrm{score} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{y}_{-i}(x_i; \theta) - y_i\}^2 \tag{2.22}$$

where y_{-i} is the estimate calculated with the current values of the control parameters, θ (number of knots, h or λ), from all of the data points except the ith. This formula requires n models to be fitted and is therefore a computationally expensive process. Another more efficient type of CV is k-fold CV where, for example, in 10-fold CV 10% of the data is removed for fitting and then used for prediction. This is repeated 10 times, rather than n, where each validation set is 10% of the data, sampled without replacement. Thus the data set is split into 10 unique validation sets each containing 10% of the data. The formula for

10-fold CV is:

$$\text{10-fold CV}(\theta) = \frac{1}{10}\sum_{f=1}^{10}\sum_{q}\{\hat{y}_q(x_i;\theta) - y_q\}^2 \tag{2.23}$$

with q being an index providing a random sample of 10% of the data, without replacement. Maggini et al. [2006] showed k-fold CV to be the best compromise between model stability and performance. However, efficiencies can be made [Hastie et al., 2009] using an approximate CV called Generalised Cross Validation (GCV) first developed by Craven and Wahba [1979] for smoothing splines.

$$GCV(\theta) = \frac{n\sum_{i=1}^{n}\{\hat{y}_i(x_i;\theta) - y_i\}^2}{[tr(\mathbf{I} - \mathbf{H})]^2} \tag{2.24}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{H}$ is the hat matrix (see Section 2.3.3) and the trace of $(\mathbf{I} - \mathbf{H})$ represents the effective number of parameters.

There are several other selection criterion that trade off fit (RSS) against smoothness in various ways. Two common-used criteria used are Akaike's Information Criterion (AIC; Akaike [1973]) and Bayesian Information Criterion (BIC; Schwarz [1978]). These are useful for selecting the best model, but if all models are poor the one model picked to be best will still be poor in a general sense because they are relative measures. The AIC estimates the expected, relative distance between the fitted model and the unknown true function that generated the observed data. It is defined as follows:

$$\text{AIC}(\theta) = -2\text{log}(\mathcal{L}(\hat{\theta}|y) + 2K \tag{2.25}$$

where $\mathcal{L}(\hat{\theta}|y)$ is the likelihood of the estimated model parameters $(\hat{\theta})$ given the data (y) and K is the number of estimable parameters (number of covariates + intercept + σ for a simple regression). The first part of the equation measures fidelity to the data and the second penalises for the number of parameters estimated in the model. AIC has a tendency to over-fit and therefore may select a model that is more complex than necessary [Rust et al., 1995]. The BIC has a stronger penalty based upon sample size, n, and thus the second part of the equation above is now $K\log(n)$. This penalty eliminates the over-fitting seen for AIC but has a tendency to under fit [Burnham and Anderson, 2010, Hastie et al., 2009].

2.5.1 Knot Selection

Knot selection is particularly important for regression splines and is typically done by selecting quantiles of the data, where the maximum number of knots is found using $\min(\frac{1}{4}n, 35)$ [Ruppert et al., 2003]. This method does not use any information in the data other than the sample size, n.

Ruppert et al. [2003] describe two types of automated knot selection; Myopic Algorithm and Full Search. The myopic algorithm, proposed by Ruppert [2000], takes a set of trial knot values, for example $T = (5, 10, 20, 40, 80, 120)$. The model is fitted for $T = 5$ and $T = 10$. Then, using GCV selection, if $\text{GCV}_{T=10} < 0.98\text{GCV}_{T=5}$ the model with the lowest GCV score is used. Otherwise, the GCV score for $T = 20$ is calculated and compared with $\text{GCV}_{T=10}$. The process is continued until the GCV scores are within 2% of each other, or the maximum knot number is reached ($T = 120$ in this case). The disadvantage of this algorithm is that it never looks beyond T, which means that it might stop too early.

The second method proposed by Ruppert [2002] is the full search algorithm. In this method, the GCV is computed for all T values. The value of T that minimises the GCV score is selected. This is computationally more expensive than the myopic algorithm but fitting regression splines is reasonably efficient so this is not really a problem. Furthermore, the advantage over the myopic algorithm is that it completes a thorough search.

These two methods will automatically select knot number but they do not provide means of placing the knots. The knots could be placed using quantiles, as mentioned earlier, or some kind of space-filling algorithm such as those proposed by Johnson et al.

[1990], John et al. [1995], Nychka and Saltzman [1998]. However, it would be useful to be able to choose both knot number and knot placement in one automated step. Recently, Walker et al. [2010] proposed a Spatially-Adaptive Local Smoothing Algorithm (SALSA) that automatically chooses the location and number of knots to be used in the spline model. The first approach to this problem was by Friedman and Silverman [1989] who developed a forward and backward knot selection algorithm (Turbo). The advantage of SALSA is that the forward/backward selection step is restricted, reducing the number of models to be evaluated. Whilst this reduces the computational burden these methods are still quite computationally expensive. SALSA is described in further detail in Chapter 5.

2.6 Summary

The best choice of smoother will depend on the characteristics of the data and knowledge about the true underlying relationship. The choice will also depend on whether the fit is to be made automatically or with manual input. Furthermore, the effectiveness of a smoother is often more related to the selection of the smoothness parameter than the selection of a particular form of smoother.

If pure regression splines are used the number of knots must be chosen for each covariate. This and their location determines the flexibility in the surface. With no penalty the GAM model structure can be written and fitted as an ordinary GLM, where each basis enters the model as an additional covariate. However, some care needs to be taken not to have too many bases and thus over parameterise the problem. This method is of particular importance for the method development in Chapter 3.

In Chapter 7 kernel smoothing is used as no interpolation, extrapolation or parameter interpretation is required. In order to develop a new method to deal with complex topographies regression splines are used and compared with TPS, GLTPS and SOAP methods. SALSA is not currently available as a knot selection method for a two-dimensional smoother so knot placement is done using a space-filling algorithm, for all methods, and the number of knots determined using a full search algorithm. In Chapter 5 we begin to develop the SALSA method for two dimensions.

A reference list for all parameters and acronyms may be found in Appendix A. Furthermore, all the coding work in this thesis is developed using the statistical computing environment `R` [R Development Core Team, 2009].

Chapter 3

Modelling Species Distribution in Complex Topographies

3.1 Introduction

This chapter describes a new smoothing method, the Complex Region Spatial Smoother (CReSS), and demonstrates its ability by comparison with existing and recently developed methods, Thin Plate Splines (TPS; Section 2.4.1), Geodesic Low Rank Thin Plate Splines (GLTPS; Section 2.4.4) and a SOAP film smoother (SOAP; Section 2.4.5), using a simulated benchmark surface first developed by Ramsay [2002]. Three of these four methods (GLTPS, SOAP and CReSS) are designed for use in areas where animals must travel around land/water (e.g. islands or lakes); areas referred to as exclusion zones. The reasons for exclusion from an area can vary widely. For example, these could include a particular depth contour or altitude, an isotherm, a river system or main roads.

TPS in a Generalised Additive Model (GAM) framework are commonly employed to construct density surfaces in the field of ecology [e.g. Guisan et al., 2002, Ashe et al., 2010]. However, this method has been shown to leak across boundaries [Ramsay, 2002, Wang and Ranalli, 2007, Wood et al., 2008] and are therefore used here to complete the review. Finite

Element L-Splines (FELS) have a major boundary issue (see Section 2.4.3) and SOAP has been shown to perform better than the FELS approach and so FELS are not considered here.

Why is a new method needed when we have the other complex methods available? The motivation for CReSS development was multi-faceted. While other recently developed methods model data for complex regions, SOAP is complicated, new and largely untested and GLTPS uses an estimated geodesic distance that could still give leakage (see Chapter 1 for details of leakage) and the smooth is globally acting. The CReSS method uses a more accurate estimate of the geodesic distance than Wang and Ranalli [2007] and allows the choice of a local or global radial basis function. Additionally, in contrast to the TPS and GLTPS methods, we use points on the domain boundary in the function construction. This explicitly constrains connections between points to be within the domain, allowing distances to be determined more accurately, even for sparse data sets.

More specifically, this chapter will describe the CReSS method, which involves estimation of geodesic distance, locally varying radial basis functions and model averaging (Section 3.2). The method is tested and compared with other methods using the horseshoe benchmark region designed by Ramsay [2002] (Section 3.3). The details of the CReSS method and much of the simulation results in Section 3.3 and 4.2.2 can also be found in Scott-Hayward et al. [2013].

3.2 Methodological Details

CReSS contains elements similar to the three methods described previously. Like TPS and SOAP, it uses the GAM framework. It uses a different basis to TPS, but the type is still a radial function. As with GLTPS, a geodesic distance metric is used. However, unique to CReSS, a model averaging approach is adopted, which has proven very successful in mapping animal densities in complex regions (the application that motivated its development).

Described here in more detail are the main components of the CReSS method, beginning with the method of estimating the geodesic distance.

3.2.1 Improved Geodesic Distance Estimation

Before fitting a model using the CReSS method we must first calculate geodesic distances between data locations and knot locations. As explained in section 2.4.4, to determine the geodesic distance between points, GLTPS constructs a network with vertices at the points. In CReSS we also build a network to estimate geodesic distances, but our vertex set includes the corners of the boundary polygons and the knot locations, as well as the data points. Polygons are used to identify the boundary of the exclusion zone, for example a coastline. This boundary is defined by one or more polygons, the vertices of which are included in the network to accommodate the calculation of the geodesic distance between the pairs of data points. Thus the edge set is made up of all line segments between distinct pairs of vertices, with the length of all edges that do no cross the exclusion zones being calculated using the Euclidean norm, and all other edges being assigned infinite length. In the case when the edge between two data points is infinite, the geodesic distance is calculated using non-infinite edges (see example in Figure 3.1). Floyds algorithm Floyd [1962] is used to determine the shortest distance through the network between all pairs of data points, knot points, and the boundary points (which GLTPS does not explicitly include). Floyds Algorithm is described in detail in Appendix B. The use of the polygon points in this process means the estimation of geodesic distance using this method is as accurate as the definition of the exclusion area polygons.

3.2.2 Basis Structure

A local radial basis can more easily accommodate spatially varying complexity than a globally acting TPS basis (Equation 2.14). Although the local basis is globally defined it is not globally acting, since the radial basis is effectively zero after a certain point. A

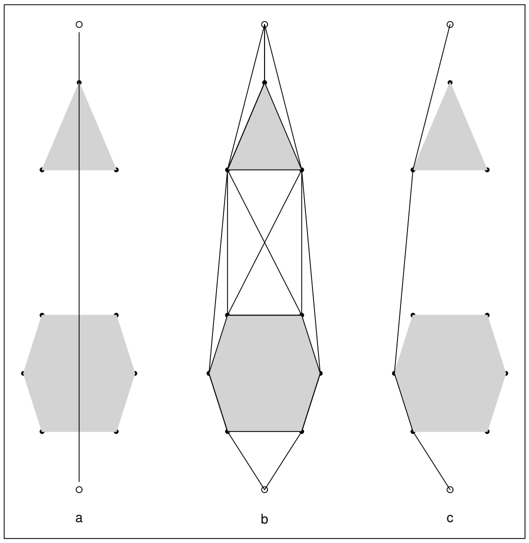


Figure 3.1: An example of graph construction using the CReSS method. The grey areas represent exclusion zones, filled circles represent the polygon vertices and lines represent edges. (a) the Euclidean distance between two points (open circles). (b) the distance network created by CReSS for these two points and (c) the geodesic distance between the two points using only the edges shown in (b).

test region (Figure 3.2), which includes a triangular exclusion zone, is used to show the global nature of TPS, using one of the usual TPS basis functions (Figure 3.3(a)). For instance, when choosing a new basis, the behaviour of many radial basis functions near the boundaries is cause for concern [Fornberg et al., 2002]. The values of the TPS basis increase with distance from each knot location. This can often lead to errors at the edges of the plot [Fornberg et al., 2002], and give rise to pronounced edge-effects. These effects are exaggerated when non-Euclidean distances are used, since the furthest distance from a knot point is no longer at the edge of the plot and the radial pattern may no longer be guaranteed if distances are modified to accommodate boundaries. In some cases, the basis is distorted, leading to areas of reinforcement (Figure 3.3(b)) where large distances compound. This

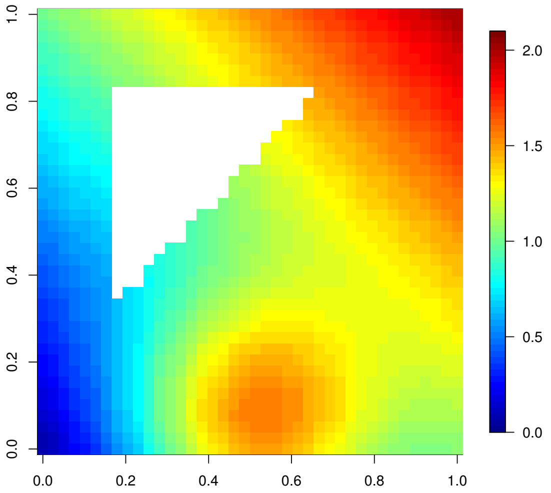


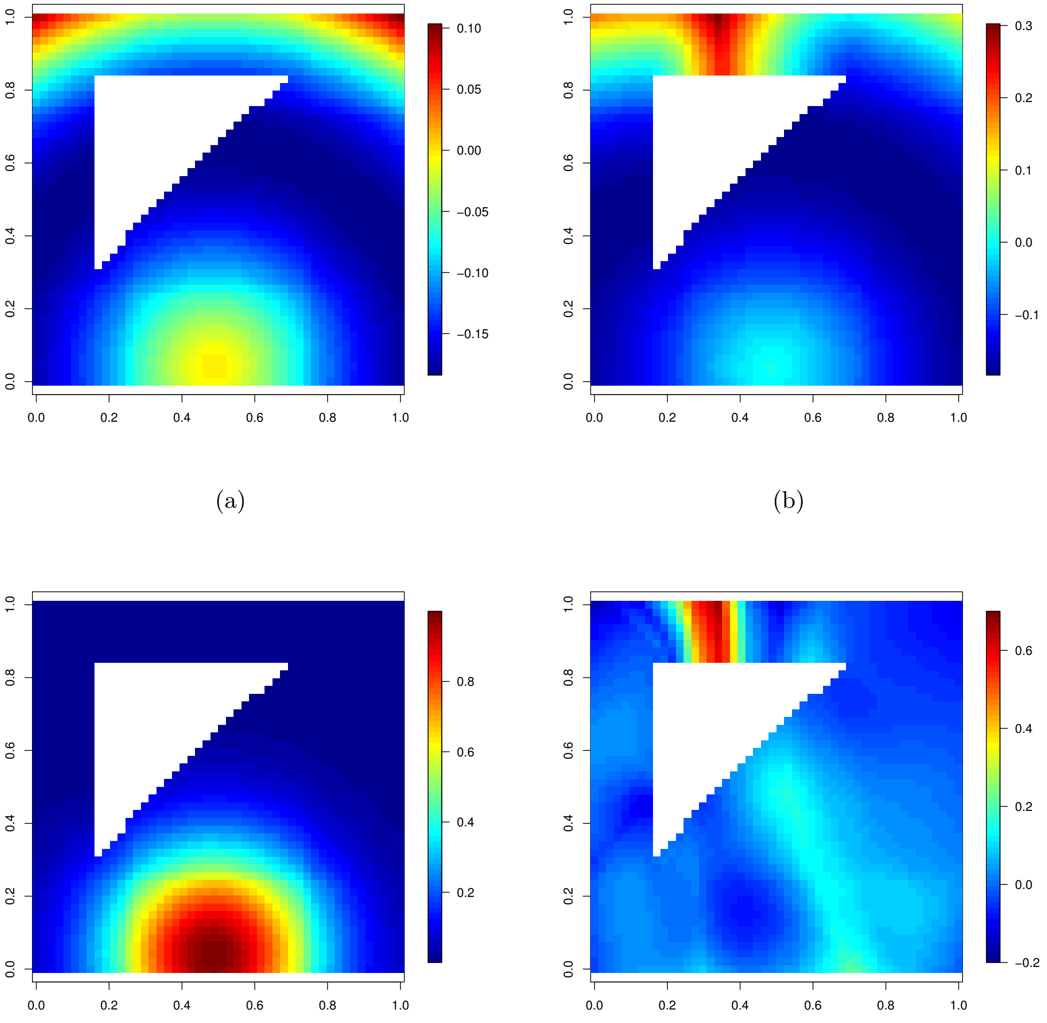
Figure 3.2: Underlying function used to show the problems of reinforcement

could lead to large prediction errors (mean squared error) and make some surfaces difficult to approximate (Figure 3.3(d)). A local basis restricts the distance from each knot over which the basis is effective and reduces the likelihood of reinforcement occurring (Figure 3.3(c)).

Thus, CReSS replaces the global radial basis function $b_t(\theta)$, (Equation 2.14, page 39) with

$$b_t(g, r) = \exp^{(-g/r^2)} \tag{3.1}$$

where r dictates the decay of this Exponential function with distance [Rathbun, 1998], and thus the extent of its local (or global) nature. Notably g indicates a geodesic distance which in practice will be between some t-th knot and i-th data location, indexed accordingly as g_{it}. Parameter r takes values such that if r is small that model will have a set of local basis functions and if r is large that model will have a set of global basis functions. However, the exact values of r are dependent upon the range and units of the spatial covariates. We have



(a)

(b)

(c)

(d)

Figure 3.3: Graphical representations of one basis function (out of a possible 30 knots) for (a) TPS, (b) global exponential basis (large r) and (c) local exponential basis (small r). The global basis shows reinforcement at the top of the triangle and (d) the area of greatest prediction error (shown as mean squared error) for the surface in Figure 3.2.

removed the planar parts of Equation 2.15 (page 41) since a linear trend in x_1 or x_2 could be based on unrealistic Euclidean distances. Thus the equation for the smooth surface, s, at point $\mathbf{x}_i$, parameter r and T knots using the CReSS method is

$$\hat{s}_r(\mathbf{x}_i) = \hat{\beta}_0 + \sum_{t=1}^{T} \hat{\delta}_t b_t(g_{it}, r).$$

3.2.3 Model Averaging

Rather than using predictions from a single 'best' model, we find the relative merits of a set of models and average the results. For each model in the set we change the number of knot locations and/or the size of parameter r.

For model selection we use frequentist model averaging [Claeskens and Hjort, 2009, Buckland et al., 1997], with AIC$_c$ [Hurvich and Tsai, 1989] for model weights. AIC$_c$ is a small sample AIC and Burnham and Anderson [2002] recommend it be used when the ratio $n/k < 40$, where n is the sample size and k is the total number of estimated regression parameters (including the intercept and σ^2). For values of this ratio > 40 AIC and AIC$_c$ converge. Other information criteria may be substituted for AIC$_c$.

The full set of models is limited to models with ΔAIC$_c$ < 10 since Burnham and Anderson [2002] suggest that a model with ΔAIC$_c$ > 10 shows no empirical support for that model. This reduced set of models is the candidate model set (M), and the relative merits of these models is found by calculating weights [Claeskens and Hjort, 2009, Buckland et al., 1997] using

$$w_m = \frac{\exp(-\frac{1}{2}\Delta_m)}{\sum\limits_{m=1}^{M} \exp(-\frac{1}{2}\Delta_m)} \tag{3.2}$$

where $m = 1, ..., M$ and Δ is the difference in AIC$_c$ between model m and the best model (lowest AIC$_c$). If the Bayesian information criterion is used, Equation 3.2 becomes the Schwarz [1978] approximation of the Bayes factor. Predictions are made for all models in model set M and their weights, w_m, are used to calculate a weighted sum of predictions to get an overall prediction. To calculate predictions, we calculate the geodesic distance between prediction locations and knot locations, using the method described in section 3.2.1, and use these to generate the bases.

A range of knot sets is considered where each set contains a different number of knots and a range of values for parameter r. For example, 10 knot sets ($\tau = 10$) and 5 r's ($R = 5$) results in 50 possible models. Of these models, the candidate knot set, M, is some number ≤ 50. Figure 3.4 gives an overview of the CReSS algorithm.

```
CReSS:
    Inputs
        spatial coordinates,
        geodesic distance matrix (data to knot locations),
        $\tau$ knot sets,
        and basis ‘range’ parameter $r$ ($R$ values in total considered)
    Model Fitting
        $\tau \times R$ candidate models calculated
        for $j$ in 1: $\tau$
            for $k$ in 1:R
                Calculate locally radial basis functions for given $r$ and knot set, $j$
                Fit models for given $k$ and $j$ using maximum-likelihood as per GAM
                Retrieve AIC$_c$ score (or other fit statistic)
    Model Selection
        Calculate model weights for models with $\Delta AIC_c < 10$ ($M$ models, $M \leq \tau \times R$)
    Model Prediction
        Calculate weighted sums of predictions (using AIC$_c$ weights) from $M$ models using
geodesic distances from prediction locations to knot locations.
```

Figure 3.4: Pseudocode outlining the structure of CReSS.

3.2.4 Choice of knots and r

Knot placement in this paper follows Wang and Ranalli [2007] by using a space filling design, such as that of John et al. [1995] from the `FIELDS` package [Furrer et al., 2010]. The knot sets therefore represent a range of different numbers of knots, whose locations in each case are determined by the space filling algorithm. The knots for all methods were generated from the observed data so any simulation run comparing methods has the same knot choices. Parameter r was chosen such that the smallest value can pick up local trends, but not so small as to cause discontinuities, and the largest value for r must be large enough to approximate a plane.

Notes on the Development of CReSS:
Several other methods were trialled during the development of CReSS, but they gave poor results. Ridge regression [Hastie et al., 2009] and mixed models [Ruppert et al., 2003] using geodesic distances were the main alternatives but both were unstable. In both cases we hoped to effectively shrink the coefficients for some of the bases but not others to allow the surface to be locally varying. There was also a tendency for methods to perform well on a simple simulation with low or medium noise, but to perform poorly on complicated and noisy surfaces. For each of the methods tried, including the final version of CReSS, we also assessed the performance of a variety of local basis functions (for example: Gaussian, Exponential and Wendland [Wendland, 2005]). The exponential function was generally found to give the best results; with ‘best’ determined using mean squared error.

3.3 Simulation

The performance of CReSS was evaluated using two simulation studies and compared with TPS, GLTPS and SOAP. The first simulation employs the horseshoe benchmark function [Ramsay, 2002, Wang and Ranalli, 2007] (Figure 3.5), which is commonly used in complex smoothing literature. The second simulation is inspired by a land reclamation project in the Persian Gulf near the coast of the United Arab Emirates (Figure 4.1) and is described in Chapter 4. Both are examples of areas with irregular shaped boundaries and sharp changes in the response across these boundaries, though the latter exhibits more complexity.

3.3.1 Horseshoe Simulation

The horseshoe (Figure 3.5) varies smoothly from approximately 4 to -4 from the right hand end of the top arm to the right hand end of the lower arm and was established by Ramsay [2002]. Three test scenarios were generated by adding a normal errors noise term with standard deviation 0.05, 1 and 5 to the function values (low, medium and high noise respectively) and randomly choosing $n = 600$ points from each noisy surface. Predictions were obtained on a grid of $N = 3584$ points (a regularly spaced grid, with points removed outwith the benchmark area). For the GLTPS method, the estimated geodesic distances calculated using code provided by Wang and Ranalli [2007] were poor and led to plotting artefacts. Therefore, the improved estimates of geodesic distance, calculated using the method in Section 3.2, were used for both GLTPS and CReSS. SOAP was constructed using a cyclic penalised cubic regression spline (40 knots) to estimate the unknown boundary values [Wood et al., 2008] and since there is no guidance for boundary knot allocation, the same knot numbers as used in Wood et al. [2008] were also used here. For CReSS, parameter r took 134 values between 2 and 10,000 for basis calculation ($r = 2$ to 20 by 1, 25 to 95, by 5 and 100 to 10,000 by 100), which gave a range of bases with local (small r) to global effects (large r). All methods employ between 10 and 100 knots (by 5) generated using the space-filling algorithm. As per authors' recommendations, model selection was performed using GCV for TPS and SOAP, AIC for GLTPS and AIC$_c$ for weights calculation for CReSS. Table 3.1 gives all parameter values for each simulation scenario.

Two measures were employed to determine the relative performance of each of the methods: estimation bias, $\hat{\mathbf{b}}$ and Mean Squared Error (MSE). MSE is described in Chapter 2. The estimation bias, $\hat{\mathbf{b}}$, is a vector of bias evaluations $\hat{b}_j$ at each of N points, $\mathbf{x}_j$ $(j = 1, ..., N)$:

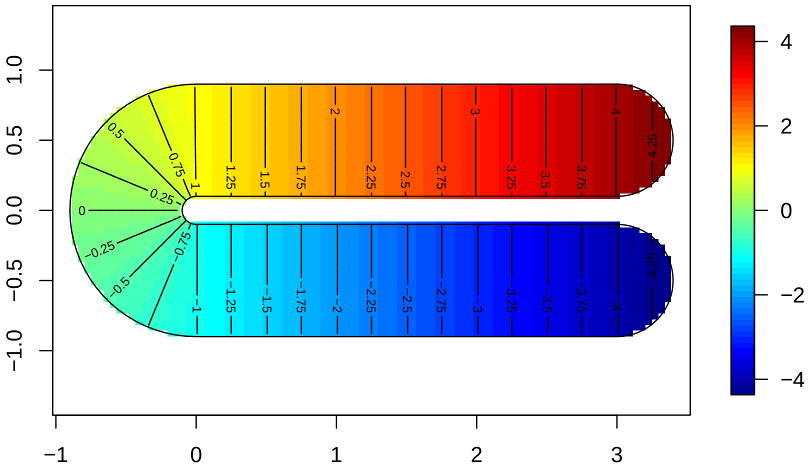


Figure 3.5: The underlying function on the horseshoe region, first seen in Ramsay [2002].

$$\hat{b}_j = 100^{-1} \sum_{p=1}^{100} \hat{z}_p(\mathbf{x}_j) - z^*(\mathbf{x}_j) \quad \text{for } j = 1, ..., N, \tag{3.3}$$

where $\hat{z}_p(\mathbf{x}_j)$ is the method's estimate of the true value, $z^*(\mathbf{x}_j)$, at replicate p (random data realisations from a surface with noise) for $p = 1, ..., 100$. MSE considers differences between predictions and the underlying function and is calculated for out-of-set prediction locations (locations unseen by the fitting process) for each replicate.

A Wilcoxon paired signed rank test [Wilcoxon, 1945] was used to see if MSE scores for TPS, GLTPS and CReSS were significantly different to SOAP (until CReSS, the most recently developed method).

3.3.2 Results

CReSS, SOAP and GLTPS all perform substantially better than TPS in this trial, at all noise levels (Table 3.2 and Figure 3.6). Beyond this distinction it is difficult to visually

Table 3.1: Horseshoe Simulation settings. The Gaussian noise (σ) is taken from a $\mathrm{N}(0, \sigma^2)$ distribution and added to P realisations of the Horseshoe function values. Each realisation is of size n. The selection criteria are specific to each method and chosen based on authors recommendations.

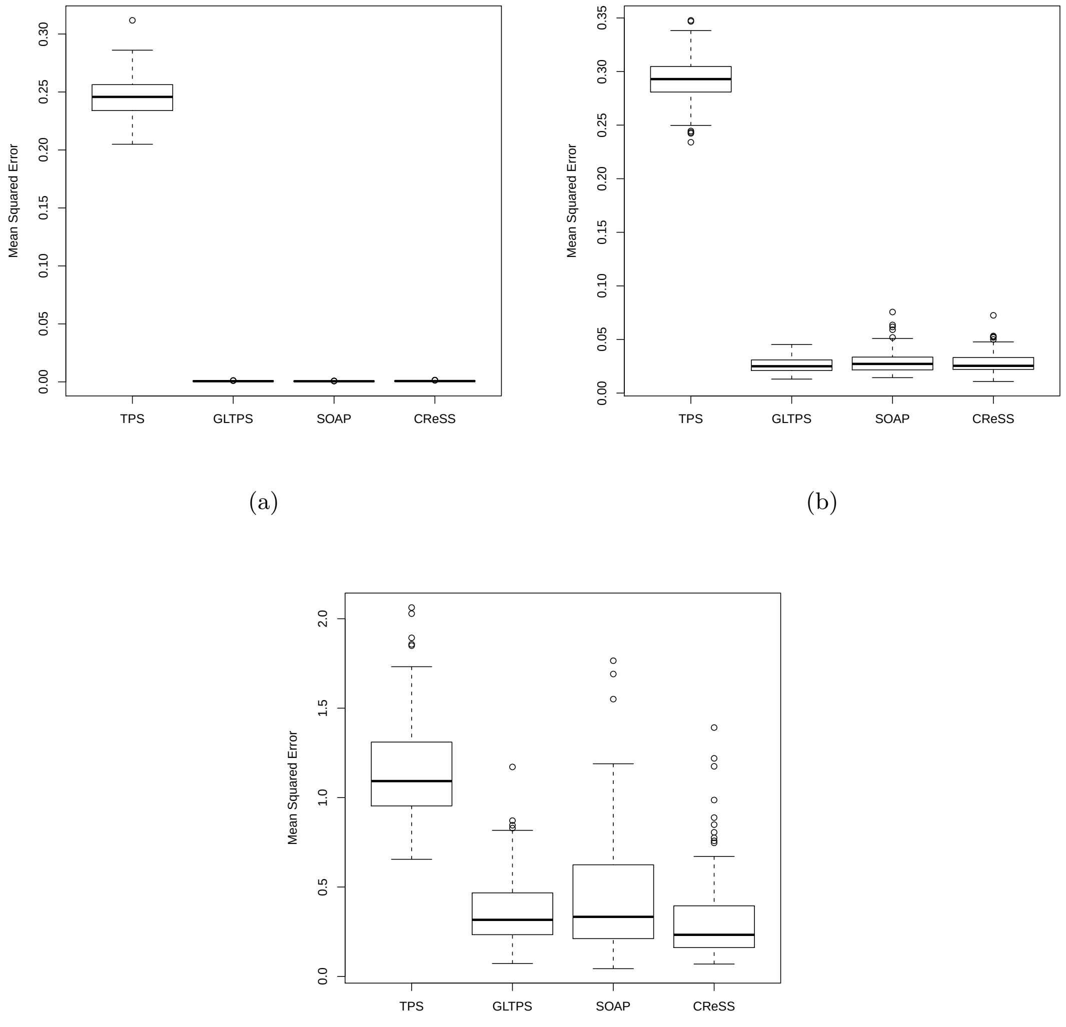
Parameters	**All methods**			
Gaussian Noise (σ)	0.05, 1, 5			
# realisations (P)	100			
Prediction grid size (N)	3584			
Sample size (n)	600			
# knots	(10, 15, ... , 95, 100)			
# knot sets (τ)	19			
	CReSS	**SOAP**	**GLTPS**	**TPS**
Selection Criterion	AIC_c	GCV	AIC	GCV
Extra Parameters	r = 2:10000 (134 values)	$\mathrm{k}_{\mathrm{outer}} = 40$	-	-

appreciate the extent of any differences between the methods (Figure 3.6), so a Wilcoxon paired signed rank test [Wilcoxon, 1945] tested for differences between all methods and SOAP. CReSS had the best mean MSE score (and smallest variance) at high noise and performed significantly better than SOAP (Table 3.2). At other noise levels the methods were very similar, with SOAP marginally but significantly better at low noise than all of the other methods. However, in real terms, the magnitude of any differences between methods, for the low noise simulated sets, were insignificant. At medium noise, GLTPS performed significantly better than SOAP and there was no significant difference between CReSS and SOAP.

Consistent with other analyses [Ramsay, 2002, Wang and Ranalli, 2007, Wood et al., 2008] the main error for TPS is along the inner edges of the two arms, while GLTPS, SOAP and CReSS show their greatest error in the elbow region (Figure 3.7-3.9). The range of the estimation bias, $\hat{\mathbf{b}}$, is comparable for all of the complex region methods, but slightly lower for CReSS at high noise (Figure 3.7-3.9). CReSS also described the increasing function along the arms better than SOAP or GLTPS (Figure 3.9).

Table 3.2: Mean MSE scores and standard deviation for all methods at all noise levels on the horseshoe simulation. A * indicates the MSE results of a method are significantly better than SOAP ($p < 0.05$; Wilcoxon paired signed rank test), a † indicates that the results for SOAP are significantly better. The bold scores represent the best average for each statistic at each noise level.

Method	Low		Medium		High	
	μ	σ	μ	σ	μ	σ
TPS	0.24608†	1.69x10^{-2}	0.2925†	0.0224	1.163†	0.285
GLTPS	0.00062†	1.54x10^{-4}	**0.0261***	**0.0064**	0.365	1.198
SOAP	**0.00055**	**7.68x**10^{-5}	0.0294	0.0114	0.458	0.358
CReSS	0.00073†	2.23x10^{-4}	0.0286	0.0100	**0.327***	**0.258**

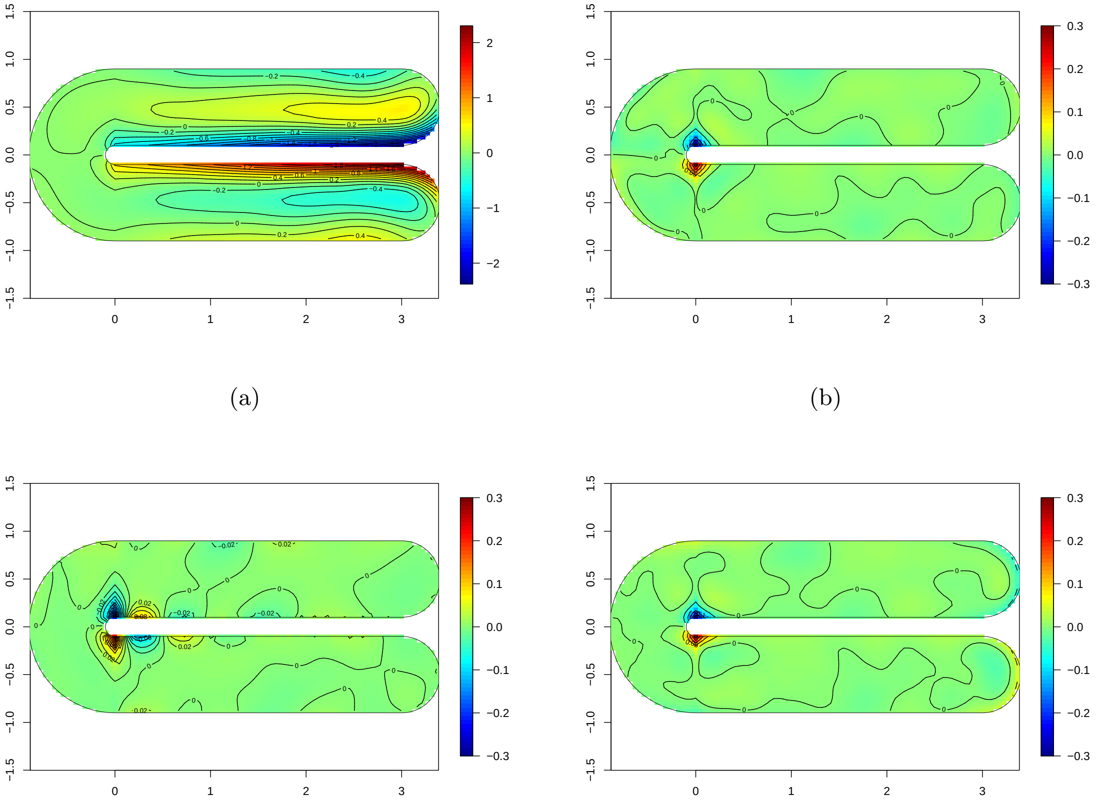


(a)

(b)

(c)

Figure 3.6: Boxplots of MSE scores for 100 simulations on the horseshoe (a) Low noise (σ = 0.5), (b) Medium noise (σ = 9) and (c) High noise (σ = 50).



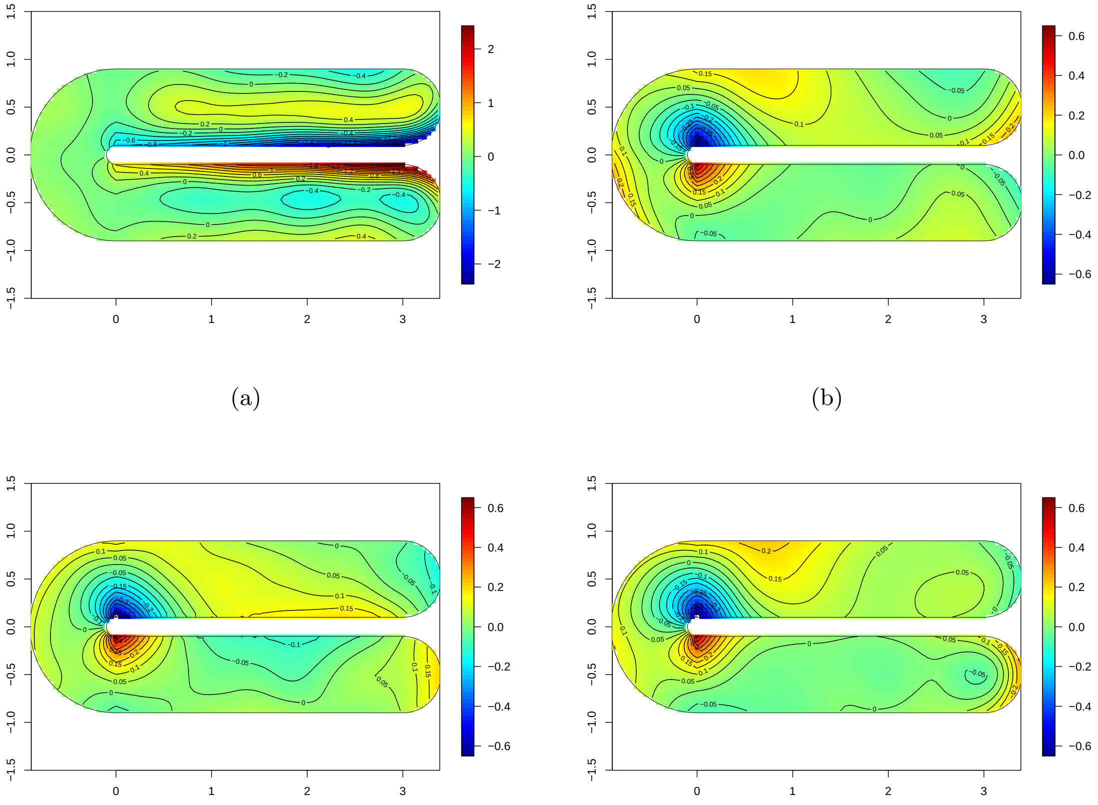
(a)

(b)

(c)

(d)

Figure 3.7: Bias for low noise, (a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS



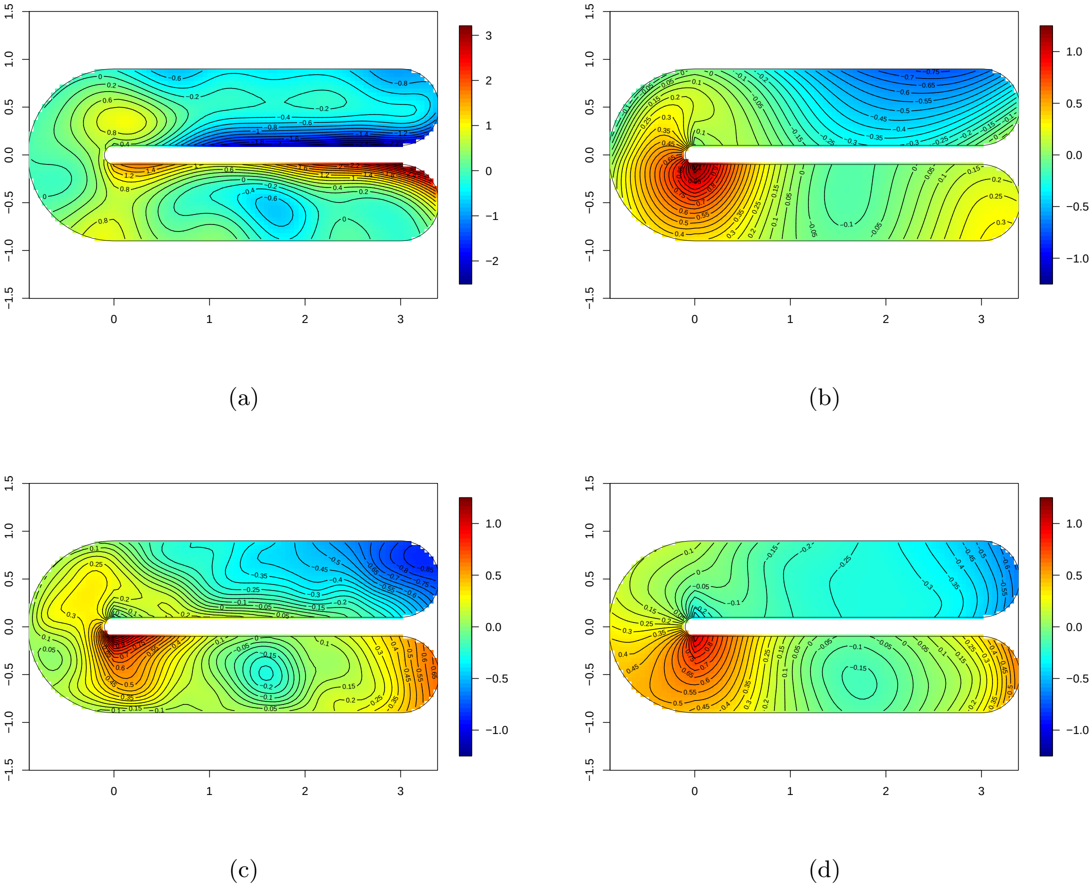
(a)

(b)

(c)

(d)

Figure 3.8: Bias for medium noise, (a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS



(a)

(b)

(c)

(d)

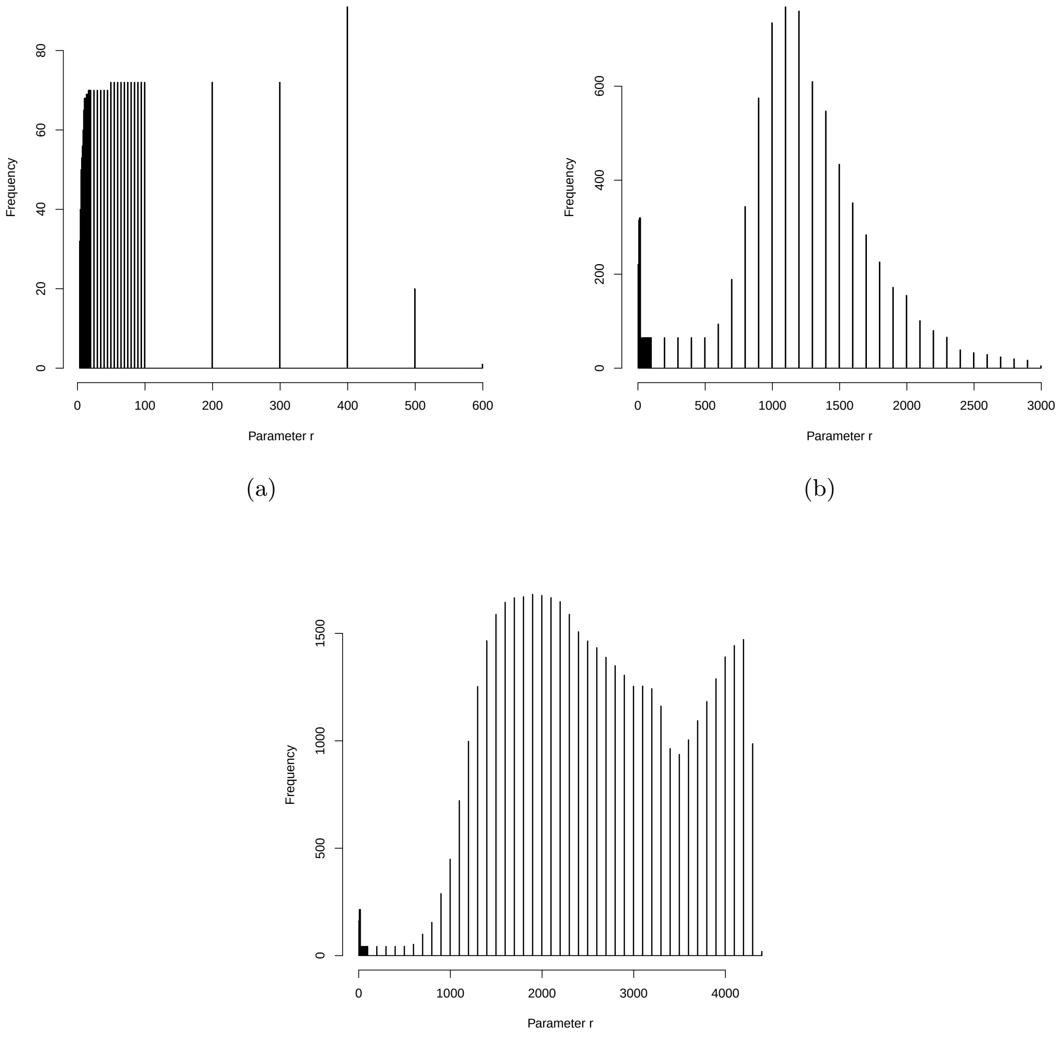
Figure 3.9: Bias for high noise, (a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS

The choice of knots, parameter r and number of models averaged over the different noise levels in CReSS are shown in Table 3.3. In general both the number of models averaged and the size of r increase as noise increases. More specifically, Figure 3.10 shows the range of values used for r and their frequency of being averaged. As the noise increases the range of r averaged increases with a distinct shift in distribution to larger values. The number of knots is similar across noise levels, although there appears to be a tendency for more knots to be chosen at low noise.

A single model tended to be chosen at low noise using CReSS, but for high noise many more were chosen and averaged (Table 3.3). The behaviour of the MSE score was examined as the number of models averaged increases. Within high noise between 58 and 653 models, of a possible 2546 (19 knot number choices and 134 possible r's), were averaged. Figure 3.11 shows the effect this has on MSE score: as the number of models averaged increases, the MSE score decreases.

Table 3.3: Parameter choices made by CReSS for each of the three noise levels averaged over 100 simulation realisations. The parameters include the mean number of models averaged per realisation, the mean for parameter r and the mean number of knots. Numbers in brackets show the minimum and maximum.

	Low	Medium	High
Mean Number of Models Averaged	24.45	91.34	471.40
	(1, 127)	(11, 250)	(58, 653)
Mean r	68.71	976.5	2513
	(3, 600)	(2, 3000)	(2, 4400)
Mean Number of Knots	75.57	45.03	54.48
	(30, 100)	(10, 100)	(10, 100)



(a)

(b)

(c)

Figure 3.10: Distribution of parameter r chosen using 100 simulation realisations for (a) low, (b) medium and (c) high noise levels. The allowed choice of r ranged from 2 to 10,000.

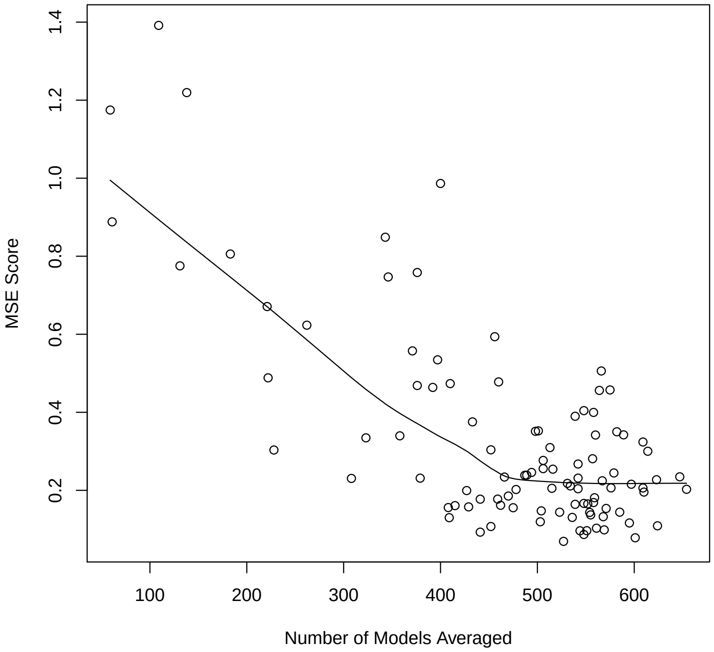


Figure 3.11: Variation in MSE with the number of models averaged. The line represents a locally weighted polynomial regression smooth of the data. The total number of models that could be averaged is 2546.

3.4 Discussion

This chapter introduced a new method for dealing with ‘leakage’ problems in topographically complex regions such as those with complex coastlines. On this simple horseshoe shape, the performance of CReSS was comparable with or better than other complex region methods (GLTPS and SOAP). The CReSS method is a novel hybrid of three techniques used in spatial modelling; geodesic distances, local radial basis functions and model averaging. The estimation of geodesic distance was improved by making our estimation as accurate as the description of the exclusion zone, compared with that of Wang and Ranalli [2007], which does not preclude the shortest distance between two points crossing a boundary. In fact, for a fair comparison to the GLTPS method we used our improved geodesic distances.

The use of locally varying basis functions allowed the method to accommodate local smoothing requirements. As the noise level increased, the range of these basis functions tended towards being global, much like a thin plate spline, in order to smooth through the noisy data. At low noise much smaller r were chosen allowing the model to fit more closely to the data. These choices were all automated, based on AIC$_c$ score, from the same set of r's for each noise level, removing the need for decisions by the user.

The last technique to complete the CReSS method was model averaging. The improvement that CReSS provides over other models at high noise may be a result of model averaging; as noise level increased, more models were averaged. The results showed that there was a clear advantage in terms of MSE score to average many models. If some of the models in the set, which are averaged, under or over-fit to the data, their effect is averaged out in the final surface. This allows the final surface to better approximate the underlying function and not over or under-fit to the noise, especially at high noise levels.

The calculation of geodesic distance required for both CReSS and GLTPS is computationally expensive but need only be done once. GLTPS was the fastest method for model fitting, but SOAP was quickest if time for distance calculation is included. However, the authors of SOAP provide no guidance on how to select the number of boundary knots. Here the number used was the same as in Wood et al. [2008], but if this was chosen by trial and error, the process would be more time consuming. Whilst CReSS computes multiple models, each one is a simple, computationally efficient GLM (since a CReSS model is linear in its parameters) and, excluding distance calculation, takes roughly the same time as SOAP to complete one simulation run. Therefore, on this simple, planar simulation region, any one of the three complex methods could be used to get similar results, and the choice comes down to user-friendliness and convenience. SOAP is not particularly user friendly and is complicated to understand but is conveniently packaged for use in `R` [R Development Core Team, 2009] alongside the well-used `mgcv` package [Wood, 2000]. CReSS is a very simple method to understand and can be fully automated with few user inputs. It is currently being applied to analyses for spatial modelling of impact assessment data and will be packaged as part of a Marine Scotland funded project.

It was clear from the results of the TPS analyses that there is a need for complex region methods since the errors from 'leakage' can be large. The results of the horseshoe trial do not provide compelling evidence for the introduction of a new method, in part because the horseshoe is rather easy for the methods to approximate. The next chapter challenges these methods with a topographically more complicated area that contains an island, which means there are at least two ways to get to any point on the surface. This may cause problems with reinforcement, discussed in Section 3.2.2, for global smoothing methods such as TPS and GLTPS. Thus, with further simulations using the CReSS method, any future directions for this research are reserved for the following chapter.

Chapter 4

Modelling Species Distribution Using Complex Topography Methods Including Islands

4.1 Introduction

The simulation used to evaluate the performance of CReSS and other smoothing approaches in Chapter 3 has become a widely-accepted standard for a complex two dimensional problem. However, it is relatively simple; it is a plane bent around in a horseshoe shape. In this chapter, we evaluate the performance of the same methods using a more complex region inspired by the palm structures in the Persian Gulf off the coast of Dubai (Figure 4.1). The upper island and edge pieces represent the outer breakwater with two channels, whilst the inner segment represents a palm leaf with three fronds on each side. The manufactured surface varies smoothly from approximately -40 to 110 units and is constructed using the definitions in Table 4.1 and the zones in Figure 4.1. It was created to test the performance of the method when the function changes greatly across small exclusion areas and in particular where there is an island. Furthermore, it was designed to test for the reinforcement issue outlined in Section 3.2.2.

This chapter will use simulation to investigate two questions: how do the methods developed in this thesis perform in a complicated region, and how do they perform when data are sparse? The same questions are subsequently addressed in Section 4.3 using a sparse and topographically complex data set on killer whale feeding behaviour.

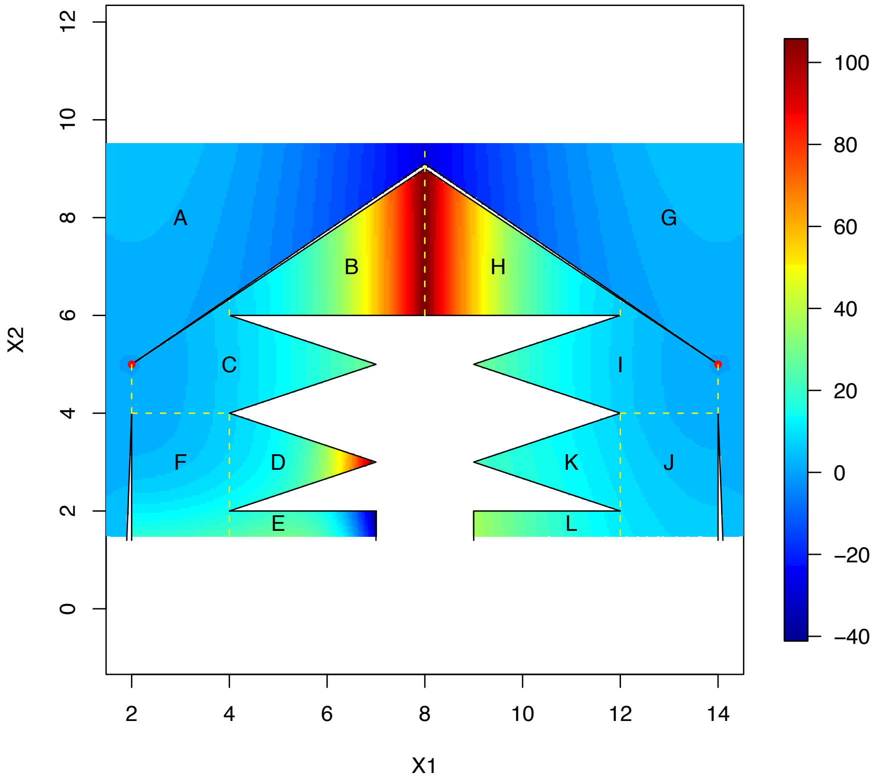


Figure 4.1: The underlying function on the simulated palm region. The letters refer to the regions in Table 4.1 used to construct the function.

Table 4.1: The benchmark surface, F, seen in Figure 4.1 is defined by functions for each region as shown. We denote the geodesic distance between two points x_1 and x_2 as $d(x_1, x_2)$. The leftmost red dot at coordinate (2,5) we denote by L. The rightmost dot at coordinate (14,5) we denote by R.

Region	$\mathbf{F}(\mathbf{X}) = \mathbf{F}(\mathbf{x_1}, \mathbf{x_2})$
A	$d(\mathbf{X}, L) - (x_1 - 2)^2$
B	$d(\mathbf{X}, L) + (x_1 - 2)^2 + (x_1 - 4)^3$
C	$d(\mathbf{X}, L) + (x_1 - 2)^2$
D	$d(\mathbf{X}, L) + (x_1 - 2)^2 + (4 - x_2)^3 + (x_1 - 4)^4$
E	$d(\mathbf{X}, L) + (x_1 - 2)^2 + (4 - x_2)^3 - (x_1 - 4)^4$
F	$d(\mathbf{X}, L) + (x_1 - 2)^2 + (4 - x_2)^3$
G	$d(\mathbf{X}, R) - (14 - x_1)^2$
H	$d(\mathbf{X}, R) - (14 - x_1)^2 - (12 - x_1)^3$
I	$d(\mathbf{X}, R) + (14 - x_1)^2$
J	$d(\mathbf{X}, R) + (14 - x_1)^2 + (4 - x_2)$
K	$d(\mathbf{X}, R) + (14 - x_1)^2 + (4 - x_2) - (12 - x_1)^2$
L	$d(\mathbf{X}, R) + (14 - x_1)^2 + (4 - x_2) - (12 - x_1)$

4.2 Simulation

The following section describes the two simulations (data rich and data sparse) on the palm shape. All complex region methods are applied along with TPS.

4.2.1 Methods

Six test cases were generated by randomly choosing $n = 500$ (data rich scenario) or $n = 100$ (data sparse scenario) points from the surface and adding a Normal error term with standard deviation 0.5, 9 and 50 to the function values. The noise was added such that the signal-to-noise ratio was similar to that of the horseshoe simulation used in Chapter 3. The signal-to-noise ratio was calculated using $var(y)/var(y - y_n)$, where y is the underlying function and y_n is the underlying function with noise added. Predictions were obtained on $N = 2518$ points.

CReSS, SOAP, GLTPS and TPS based models were all compared using these simulation data. SOAP was constructed with unknown boundary values but there is no published or available guidance for selecting the number of boundary knots. The default value (10 knots) is too small in many situations. Boundary knots were therefore selected using an extensive but non-exhaustive trial and error search to give the best results possible. For the data rich simulation ($n = 500$), we used 50 knots for the outer boundary and 40 for the island. However, for the data sparse simulation, there were not enough degrees of freedom available for these knot numbers. After a non-exhaustive search, 10 knots were chosen each for both the inner and outer boundaries. As for the simulation study in Chapter 3, parameter r, for the CReSS method, took 134 values between 2 and 10,000. For all methods a choice of 10 to 100 knots was allowed for the data rich trials and 10 to 75 for the data sparse trials. These varying model complexities were discriminated between using GCV (TPS and SOAP), AIC (GLTPS) and AIC$_c$ (CReSS). In the case of SOAP these knots were allocated to the interior soap basis. A summary of the simulation settings for all methods can be

Table 4.2: Palm Simulation settings. The Gaussian noise (σ) is taken from a $\mathrm{N}(0, \sigma^2)$ distribution and added to P realisations of the Palm function values. Each realisation is of size n. The selection criteria are specific to each method and chosen based on authors recommendations.

Parameters	**All methods**			
Gaussian Noise (σ)	0.5, 9, 50			
# realisations (P)	100			
Prediction grid (N)	2518			
Sample size (n)	100		500	
# knots	(10, 15, ... , 70, 75)		(10, 15, ... , 95, 100)	
# knot sets (τ)	14		19	
	CReSS	**SOAP**	**GLTPS**	**TPS**
Selection Criterion	AIC$_c$	GCV	AIC	GCV
Extra Parameters	r = 2:10000	n=100: (k$_{\text{outer}}$ = k$_{\text{island}}$ = 10)	-	-
	(134 values)	n=500: (k$_{\text{outer}}$ = 50, k$_{\text{island}}$ = 40)		

found in Table 4.2.

As with the horseshoe simulation in the previous chapter, model fit was assessed using estimation bias and Mean Squared Error (MSE). In a simulation setting we know truth so we can calculate MSE, however in reality we need a measure that mirrors the MSE score without knowing truth. This analysis uses 10-fold Cross Validation (CV) because it assesses fit to data unseen by the model (see Chapter 2 section 2.5 for details). A Wilcoxon paired signed rank test [Wilcoxon, 1945] was also used to see if MSE scores for TPS, GLTPS and CReSS were significantly different to SOAP (the most recent method).

The results for the data rich and data sparse simulations are presented separately in the following two sections.

4.2.2 Data Rich Results

CReSS exhibited the best performance and lowest MSE scores at low and medium noise across all model types and TPS gave the worst performance to both the data and underlying function across all noise levels (Table 4.3 and Figure 4.2). It is difficult to see, from Figure 4.2, any differences in MSE scores between methods, so a Wilcoxon signed rank test [Wilcoxon, 1945] was used to test for significant differences between all methods and SOAP (the most recent method). At both medium and high noise CReSS performed significantly better than SOAP ($p < 0.05$) and while CReSS performed significantly better than SOAP on average at low noise, it was statistically indistinct for that noise level. SOAP did not perform best at any noise level and at low and high noise had the highest variances for MSE scores. Given that SOAP was statistically no worse than CReSS or GLTPS at low noise, the high variance indicates that when SOAP performed badly, it performed very badly. In comparison with GLTPS, CReSS performed better at low and medium noise but not at high noise levels. However, there were some fitting artefacts that led to GLTPS being numerically good but graphically poor at high noise, which are mentioned further later.

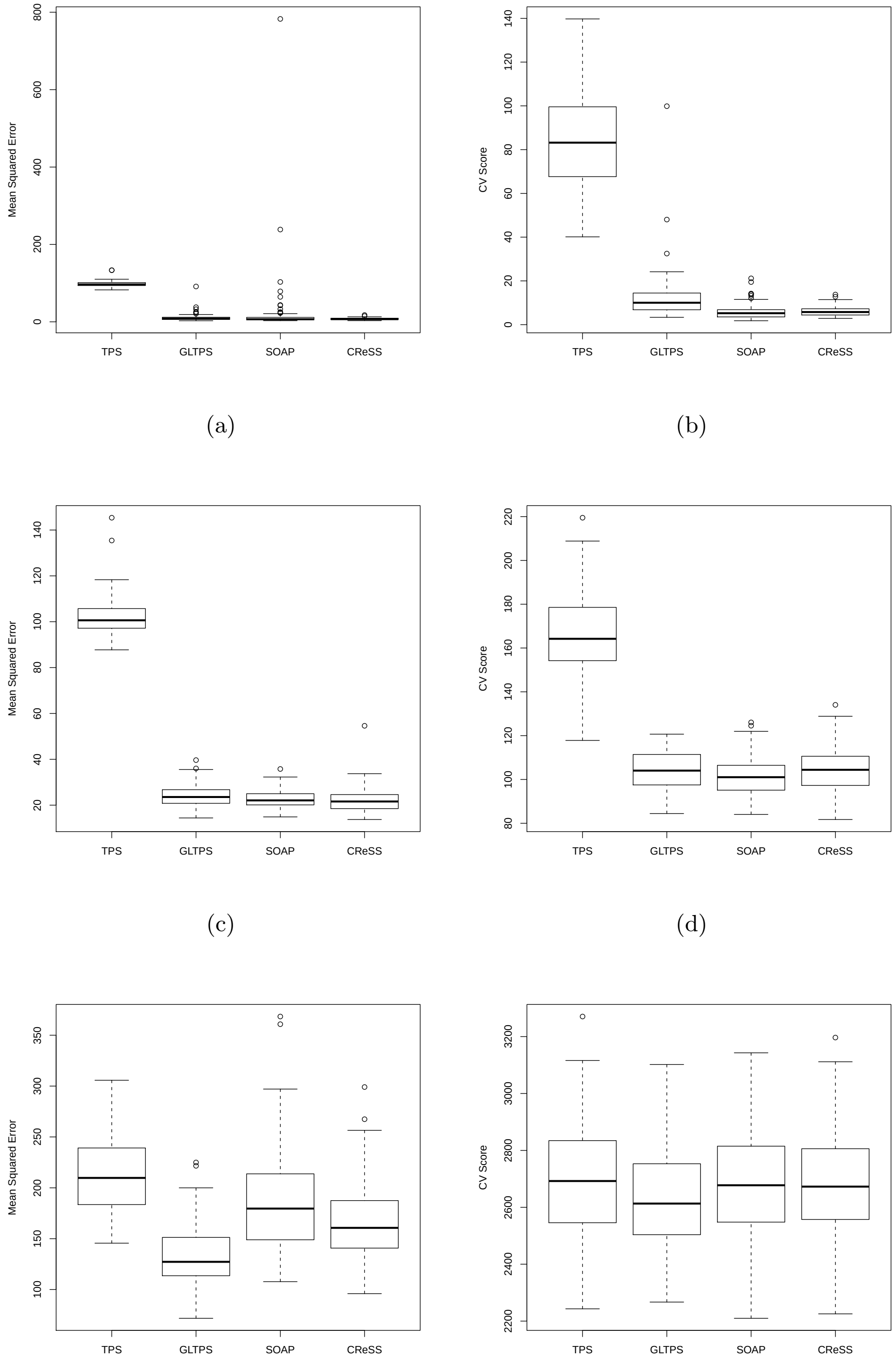
Table 4.3: Mean MSE scores and standard deviation (sd) for all methods at all noise levels for the palm simulation using 500 data points. A * indicates the MSE results of a method are significantly better than SOAP ($p < 0.05$, Wilcoxon signed rank test), a † indicates that the results for SOAP are significantly better. The bold scores indicate the best average for each statistic at each noise level.

Method	Low		Medium		High	
	mean	sd	mean	sd	mean	sd
TPS	97.48†	7.42	101.95†	7.92	213.47†	33.43
GLTPS	10.51	9.95	24.18†	4.78	**131.45***	**29.54**
SOAP	21.47	81.52	22.70	**3.73**	188.24	51.33
CReSS	**7.70**	**2.70**	**21.97***	5.49	167.26*	38.53

CV was able to distinguish TPS from the other methods at low and medium noise, but no clear distinction could be made at high noise levels or between complex methods at any noise level (Figure 4.2). A Wilcoxon signed rank test [Wilcoxon, 1945], comparing the CVs for SOAP to all other methods revealed the best methods at low noise were CReSS and SOAP, SOAP at medium noise and GLTPS at high noise. These results were similar to the MSE results except for medium noise, indicating that 10-fold CV is not necessarily a good measure for selecting between methods. As a mirror for the MSE score, CV also did not perform very well. An investigation into the difference in MSE scores for CReSS and SOAP and the difference in CV scores between the two methods revealed that CV correctly classified the rank of one method over the other 58% of the time for low noise, and 45% and 46% for medium and high noise levels respectively. These numbers are much lower than would be expected if CV was to be used as a ranking measure in practice.

The lack of fit of TPS became more pronounced as noise increased, mainly due to leakage across the island, where the difference in underlying function values is greatest (Figures 4.3(a), 4.4(a), 4.5(a)). At high noise there was also some evidence of leakage through the palm fronds from the hotspot at the top of the simulated surface. As expected, there was no evidence of leakage for GLTPS, SOAP or CReSS, however, all methods (including TPS) struggled to model the high and low function values to the left of the stem (Figures 4.3 - 4.5). These errors may be due to lack of coverage by the data points or an inflexibility in knot number and/or placement.

CReSS respected all the boundaries, keeping the high values below the breakwater and the low values above it without leakage. However, CReSS exhibits some negative bias just under the breakwater and just above the central palm shape (Figures 4.3(d) - 4.5(d)). These are two areas where perhaps the radial nature of CReSS struggles to approximate the striations of the underlying function. SOAP dealt well with the outer breakwater, but there was some evidence of errors on the ends of the upper fronds and, as noise increases, on the upper edge of the central palm shape (Figure 4.5(c)). GLTPS showed a good numerical fit, particularly at high noise, but exhibited some artefacts (striations) to the upper right and left of the island. These are particularly apparent on a prediction plot for a single realisation with medium noise (Figure 4.6(b)). We consider that this is due to reinforcement issues arising from the global basis function and that the local concentration of errors and associated artefacts made it a poor choice in practice.



(a)

(b)

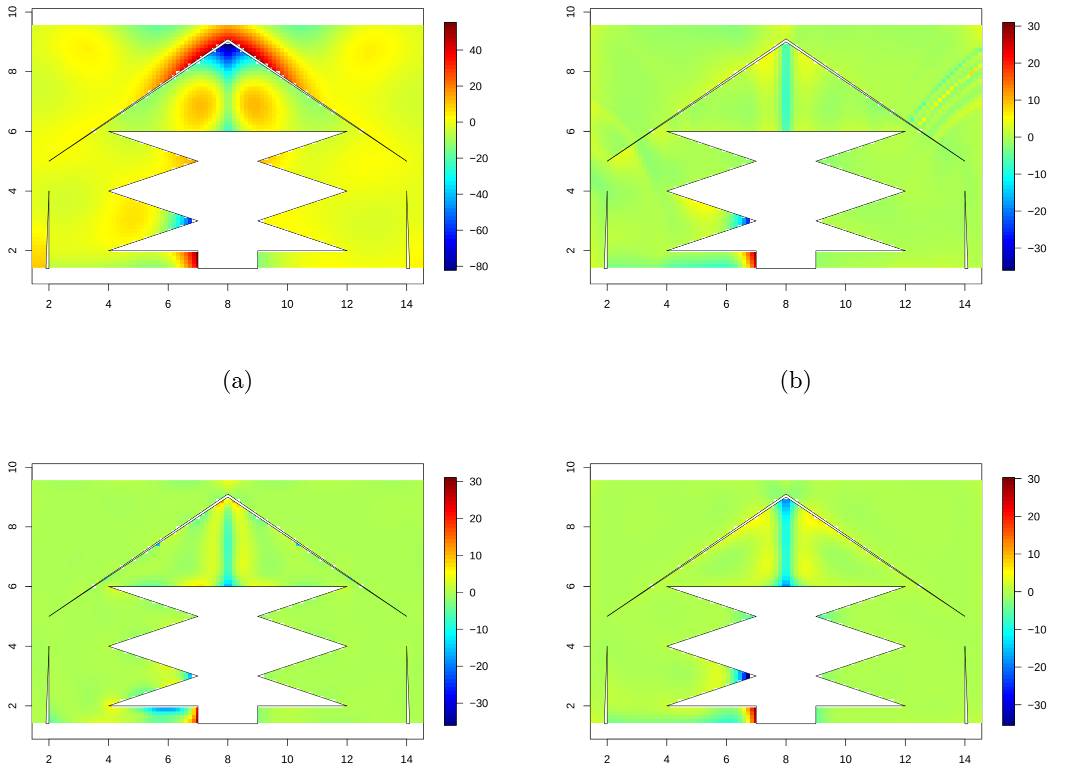
(c)

(d)

(e)

(f)

Figure 4.2: Boxplots of MSE (left) and CV scores (right) for 100 simulations on the palm function. (a, b) Low noise (σ = 0.5), (c, d) Medium noise (σ = 9) and (e, f) High noise (σ = 50)



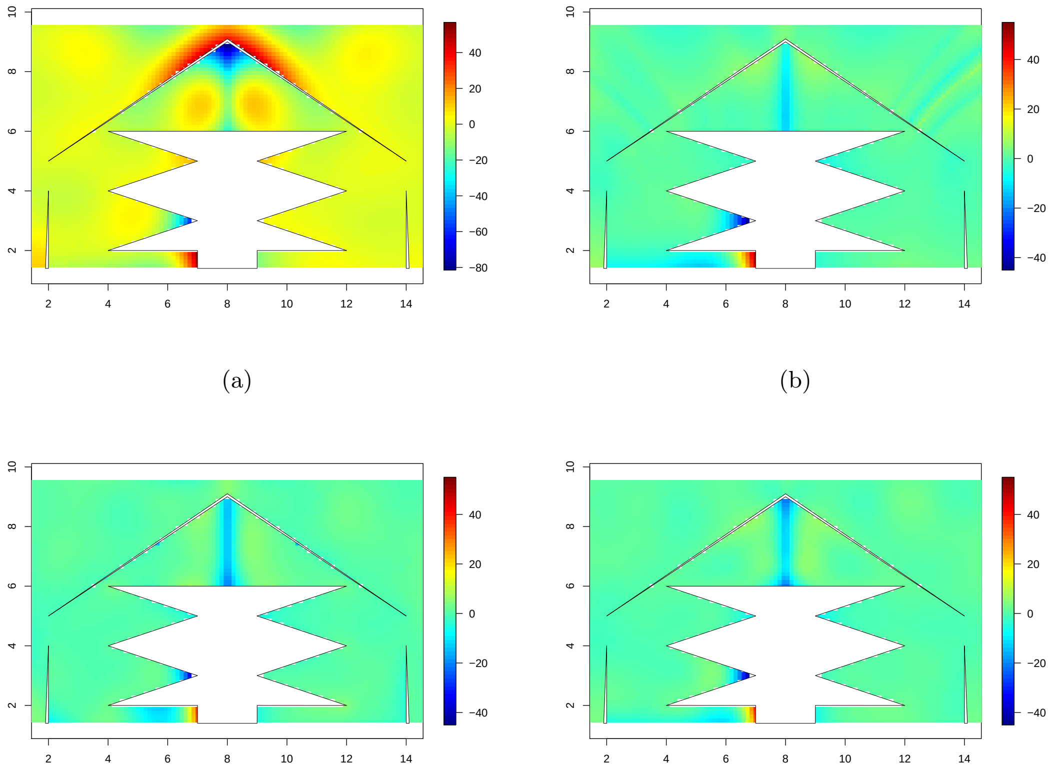
(a)

(b)

(c)

(d)

Figure 4.3: Bias for low noise a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS



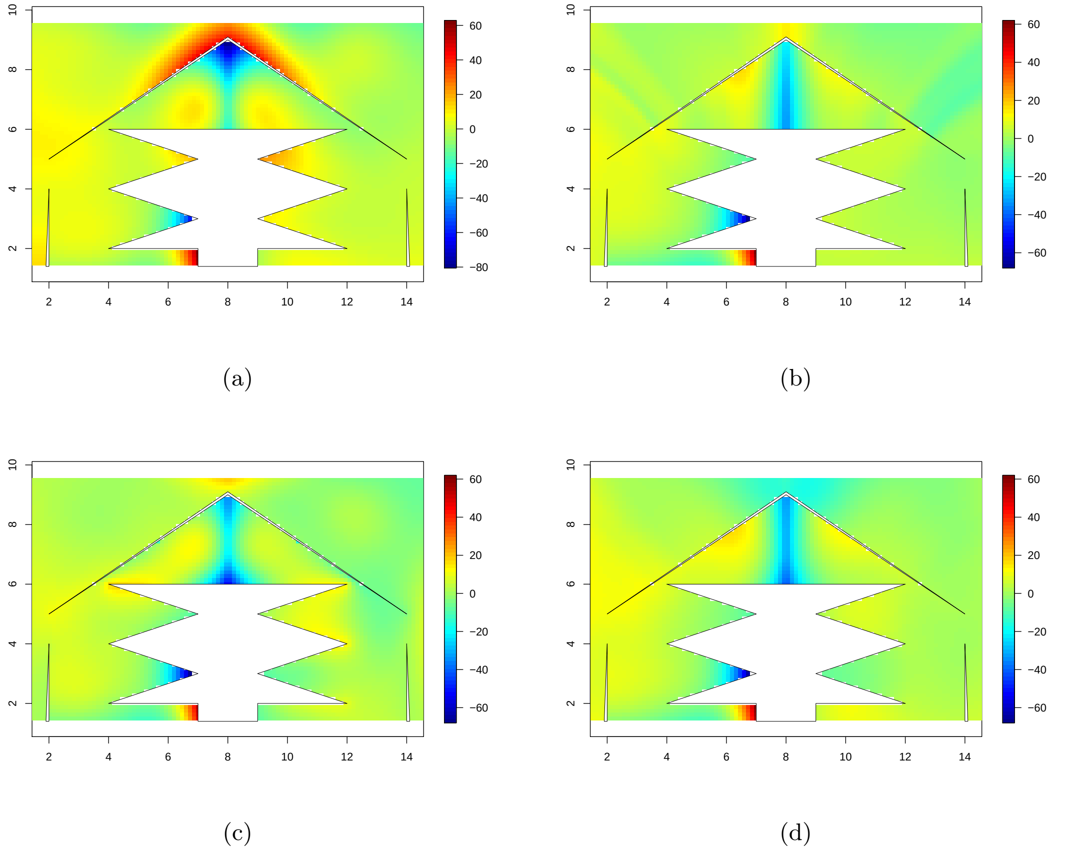
(a)

(b)

(c)

(d)

Figure 4.4: Bias for medium noise (a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS



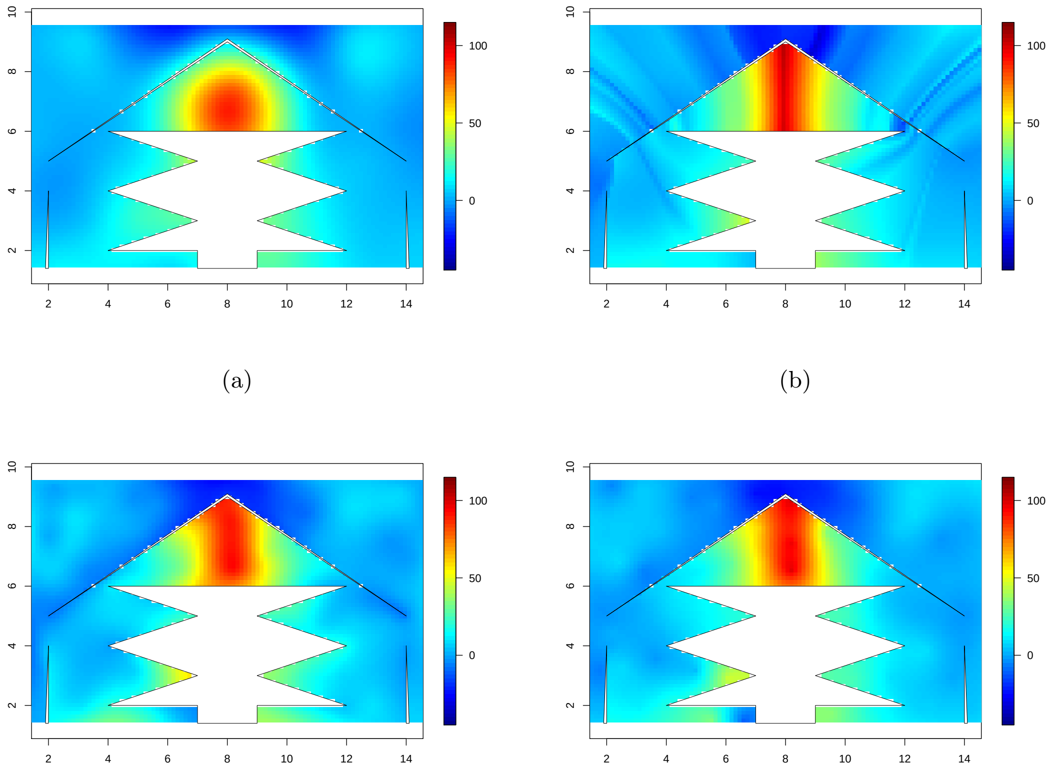
(a)

(b)

(c)

(d)

Figure 4.5: Bias for high noise, (a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS



(a)

(b)

(c)

(d)

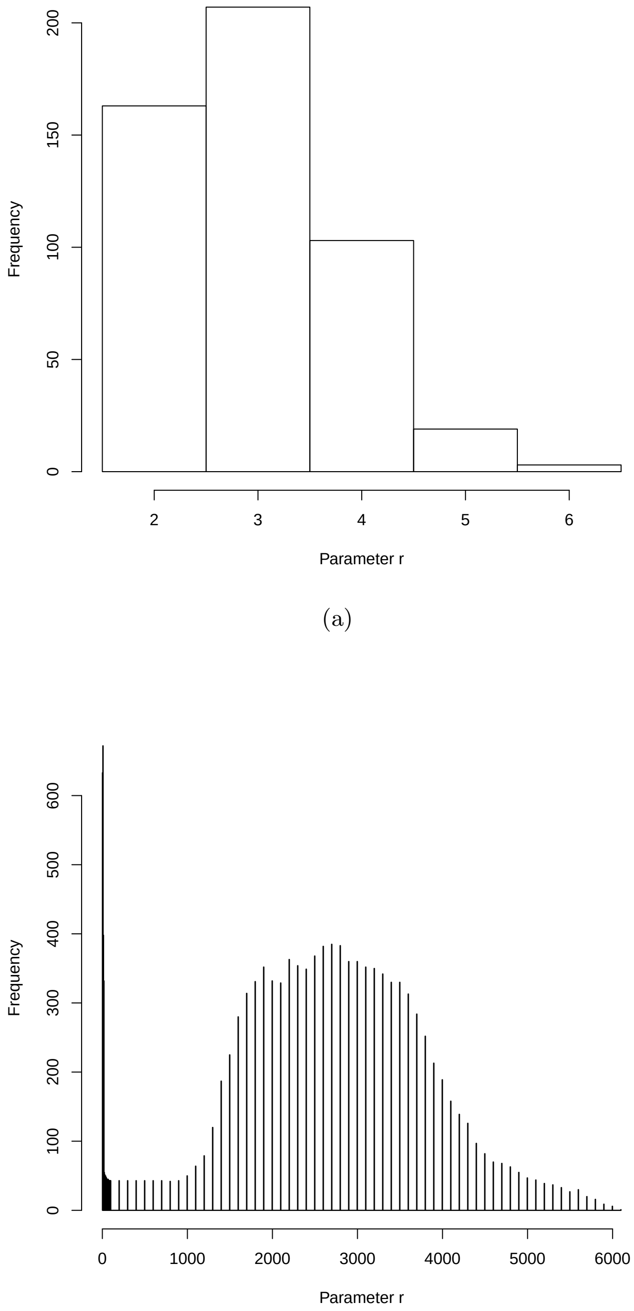
Figure 4.6: Example predictions for medium noise, (a) TPS, (b) GLTPS, (c) SOAP, (d) CReSS (iteration 80)

In general, as noise level increased, the models for averaging became more smooth; the number of models averaged by CReSS and the value of parameter r increased and the mean number of knots decreased (Table 4.4). Figure 4.7 shows the distribution of r for medium and high noise. For low noise, only two values of r were ever chosen and so there is no distribution to represent in a figure. Many more values of r were chosen for high noise and the distribution was shifted to higher values. However, a number of models using small r were still chosen.

In contrast to the horseshoe simulation (Chapter 3), the MSE score did not necessarily improve because more models were averaged (Figure 4.8). With high noise levels, between 3 and 497 models, of a possible 2546 (19 knot number choices and 134 possible r's), were averaged and Figure 4.8 shows the effect this has on MSE score. The effect is not as convincing as was the case for the horseshoe simulation, although there may be some advantage in increasing the number of models averaged to about 200. Above this level, the MSE score increased with the number of models averaged. Figures are not presented for low or medium noise as only a few models were averaged in each case (Table 4.4).

Table 4.4: Parameter choices made by CReSS for each of the three noise levels averaged over 100 simulation realisations. The parameters include the mean number of models averaged per realisation, the mean for parameter r and the mean number of knots. Numbers in brackets show the minimum and maximum.

	Low	Medium	High
Mean Number of Models Averaged	1.45	4.95	132.07
	(1, 4)	(1, 13)	(3, 497)
Mean r	2.75	2.97	2211
	(2, 3)	(2, 6)	(2, 6100)
Mean Number of Knots	92.5	60.2	45.4
	(65, 100)	(20, 100)	(10, 100)



(a)

(b)

Figure 4.7: Distribution of parameter r chosen using 100 realisations of the simulation for (a) medium and (b) high noise levels. Results from the low noise trials are not shown because only two different r's were ever chosen. The allowed choice of r ranged from 2 to 10,000.

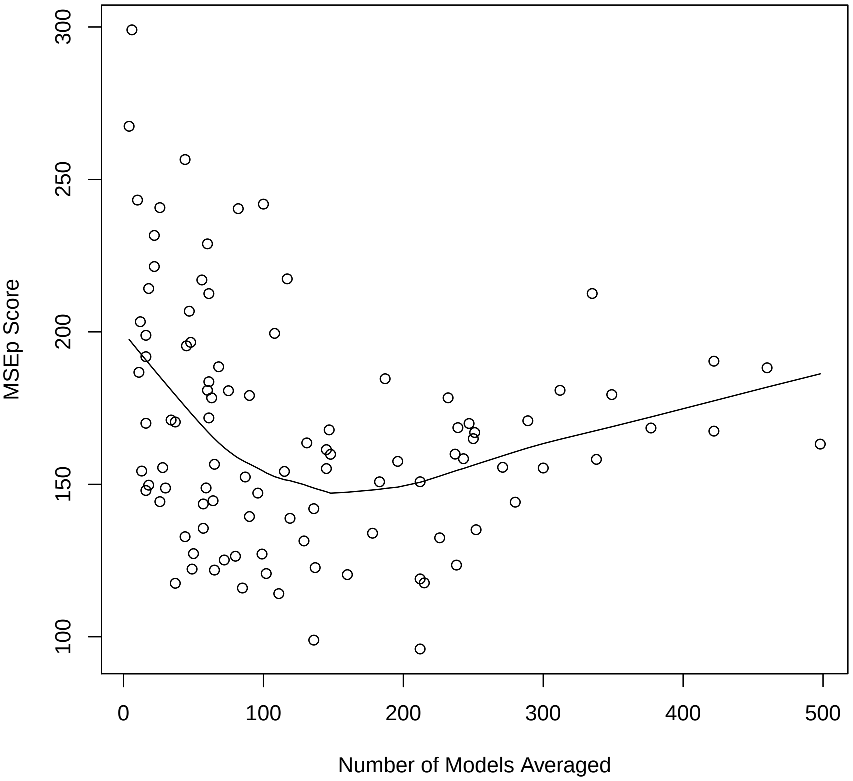


Figure 4.8: Variation in MSE with the number of models averaged for models fitted to high noise. The points represent how many models were averaged and the resulting MSE score for each of the 100 simulation realisations. The line represents a locally weighted polynomial regression smooth of the data. The total number of models that could be averaged is 2546.

4.2.3 Data Sparse Results

The results for $n = 100$ are much more conclusive. Numerically CReSS provided the best fits and was more stable (low standard deviation) at all noise levels than any of the other methods (Table 4.5 and Figure 4.9). CReSS had the lowest MSE scores of all methods and was significantly better than SOAP at all noise levels ($p << 0.05$; Table 4.5). Due to the skewed nature of the results, particularly for SOAP and GLTPS, the median MSE is also included. At low noise, the results for SOAP and GLTPS are surprisingly variable (Figure 4.9) and the MSE and CV plot has been limited on the y-axis to give a better comparison with other methods. At medium and high noise, CReSS and GLTPS are both significantly better than SOAP and a Wilcoxon signed rank test [Wilcoxon, 1945] between the two indicates that CReSS performs better than GLTPS at low ($p < 0.05$) and high noise ($p < 0.1$). TPS was consistently the worst performing method, however at high noise, it was significantly worse than SOAP at the 5% but not the 10% level of significance.

Table 4.5: Mean and median MSE scores and standard deviation (sd) for all methods at all noise levels for the palm simulation using 100 data points. A * indicates the MSE results of a method are significantly better than SOAP ($p < 0.05$, Wilcoxon signed rank test), a † indicates that the results for SOAP are significantly better. The bold scores indicate the best average for each statistic at each noise level.

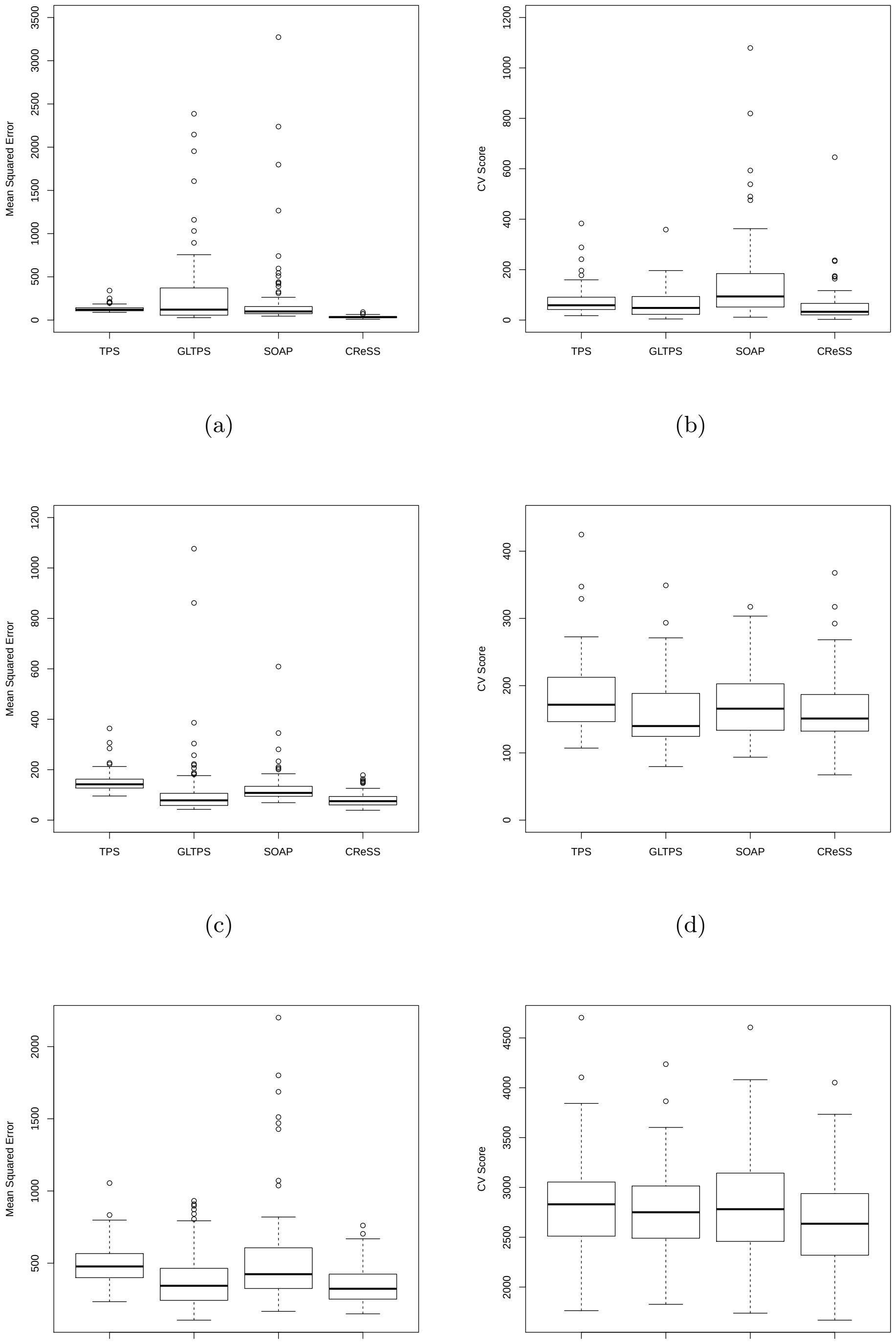
Method	Low		Medium		High	
	mean (median)	sd	mean (median)	sd	mean (median)	sd
TPS	130 (119)	36.6	152 (142)†	40.5	499 (152)†	477
GLTPS	1745 (121)†	5926	144 (78.4)*	363	383 (191)*	343
SOAP	1806 (99.3)	11207	141 (107)	181	521 (348)	423
CReSS	**35.7 (33.2)***	**13.3***	**81.6 (75.2)***	**27.8**	**344 (125)***	**323**

10-fold CV was unable to distinguish between the models (Figure 4.9) at any level of noise and the patterns from Wilcoxon signed rank tests bear little resemblance to those seen from the boxplots of MSE scores. As a mirror for the MSE score, CV also did not perform very well. An investigation into the difference in MSE scores for CReSS and SOAP and the difference in CV scores between the two methods revealed that CV correctly classified the rank of one method over the other 89% of the time for low noise, and 49% for both medium and high noise levels. Whilst the classification for low noise is good, the other noise levels show a much lower correct classification rate than would be expected if CV was to be used as a ranking measure in practice.

In general, the estimation biases were higher, for a given method and noise level, than those for the data rich results (Figures 4.10 to 4.12). TPS showed high levels of leakage across the breakwater and this increased with higher noise. As expected, there was no evidence of leakage for GLTPS, SOAP or CReSS, however, all methods (including TPS) struggled to model the high and low function values to the left of the stem (Figures 4.10 - 4.12). Like the data rich simulation it is thought that these errors may be due to lack of coverage by the data points, which was much poorer in this simulation, or an inflexibility in knot number and/or placement. Similar to the data rich simulation, the GLTPS method shows striations, which are particularly prevalent at low noise (Figure 4.13(b)) and are also apparent in the figure showing example predictions at medium noise (Figure 4.13). Therefore, the fit assessment for all methods should involve both numerical and visual assessment.

The SOAP method dealt well with the breakwater but was hard to parametrise; increasing the number of boundary knots from the default (10 knots) did not improve performance. With low noise, the use of 20 knots for both the inner and outer boundaries resulted in a mean (317132), median (200) and standard deviation (3x10^6) of MSE scores that were higher than those obtained with 10 boundary knots. Furthermore, increasing the number of boundary knots meant that fewer parameters were available for the number of knots within the domain. There seemed to be an increase in errors around the boundary in comparison with the data rich simulation (Figure 4.10(c) versus Figure 4.3(c)), particularly around the ends of the palm fronds. This is possibly due to a smaller number of boundary knots for this simulation.

CReSS respected all the boundaries, dealt well with the breakwater and the pattern of biases for was consistent with that seen in the data rich simulation. Specifically, CReSS exhibits some negative bias just under the breakwater and just above the central palm shape (Figures 4.10(d) - 4.12(d)).



(a)

(b)

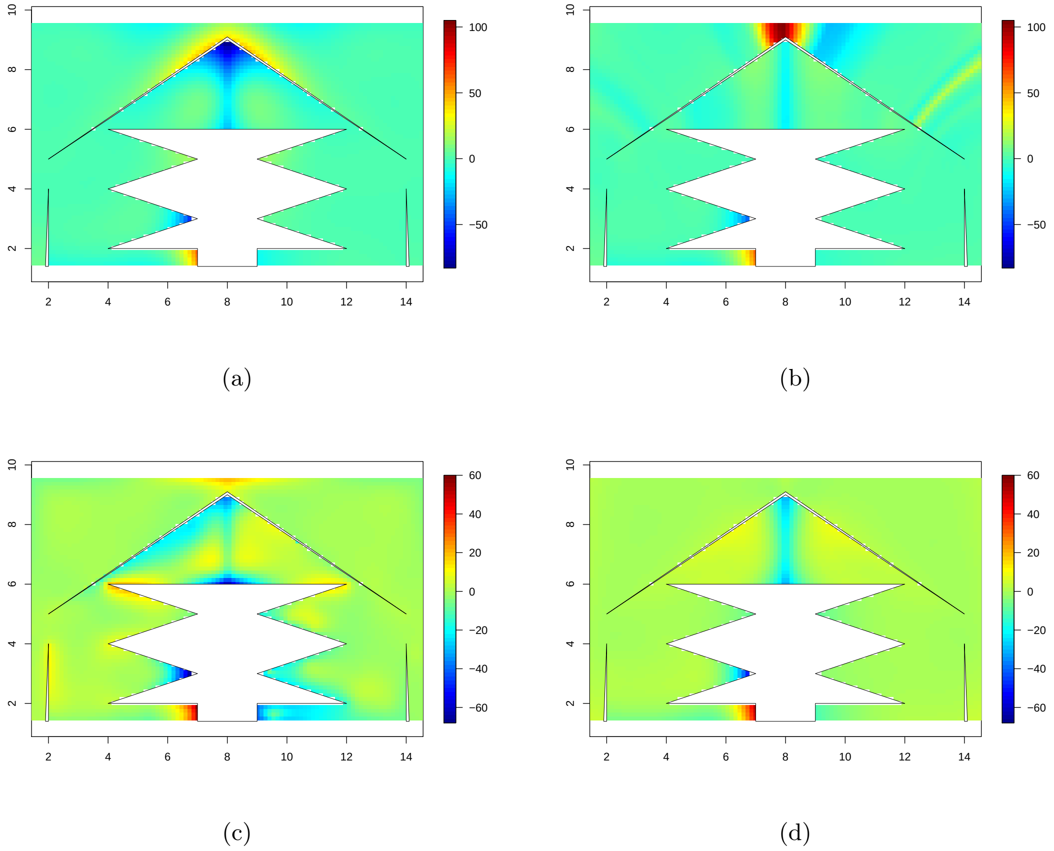
(c)

(d)

(e)

(f)

Figure 4.9: Boxplots of MSE (left) and CV scores (right) for 100 simulations on the data sparse palm function. (a, b) Low noise ($\sigma = 0.5$), (c, d) Medium noise ($\sigma = 9$) and (e, f) High noise ($\sigma = 50$). The plots for low/medium noise have been limited on the y-axis for ease of viewing (MSE, low: GLTPS (1 point at 4000) and SOAP (2 points at 80,000). MSE, medium: GLTPS (1 point at 3000) and SOAP (1 point at 2000). CV, low SOAP (1 point at 15,000) and CReSS (2 points at 7000). CV, medium: SOAP (1 point at 1000)).



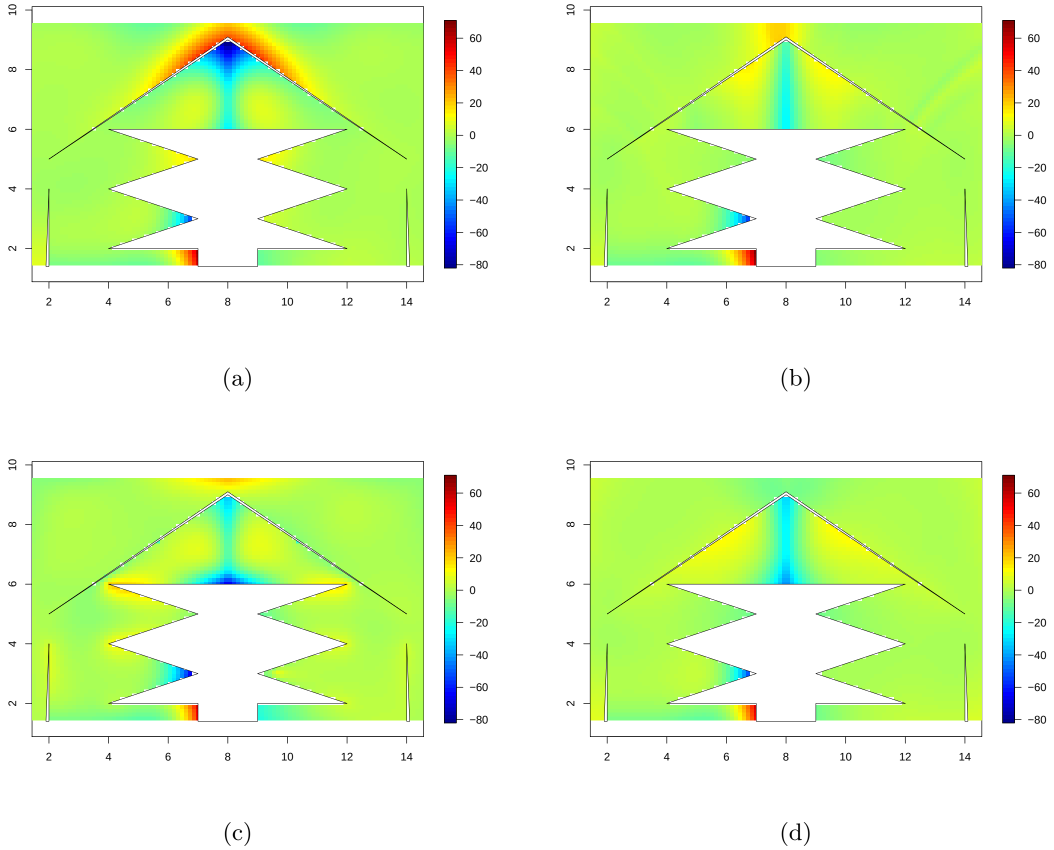
(a)

(b)

(c)

(d)

Figure 4.10: Bias for data sparse low noise a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS



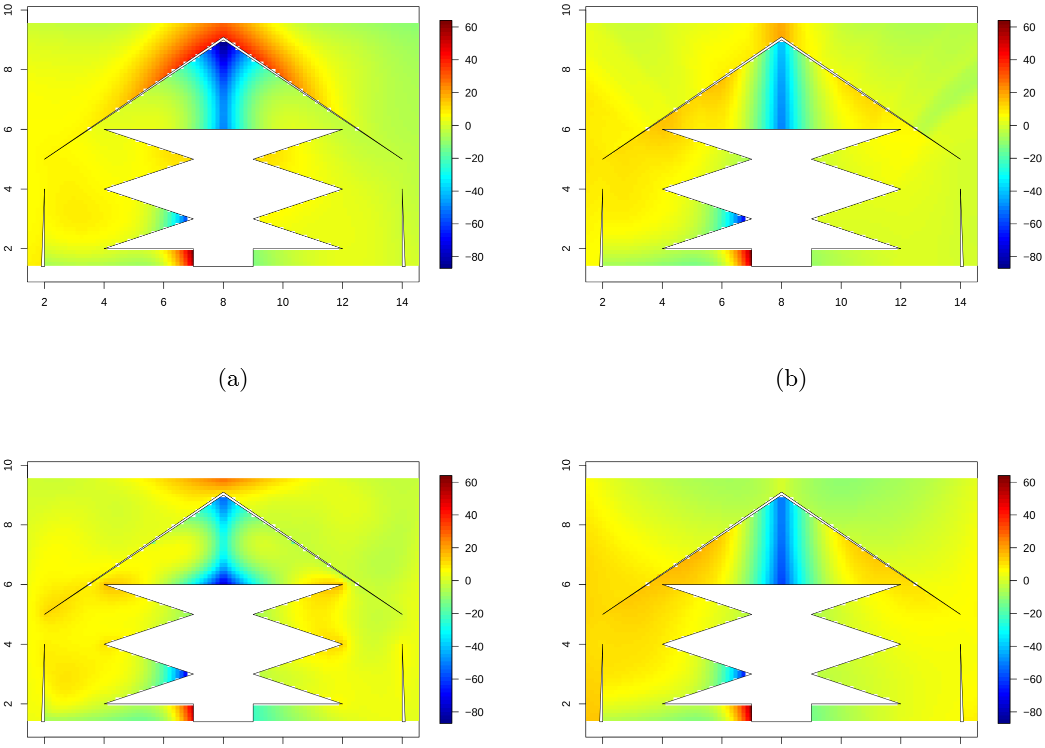
(a)

(b)

(c)

(d)

Figure 4.11: Bias for data sparse medium noise (a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS



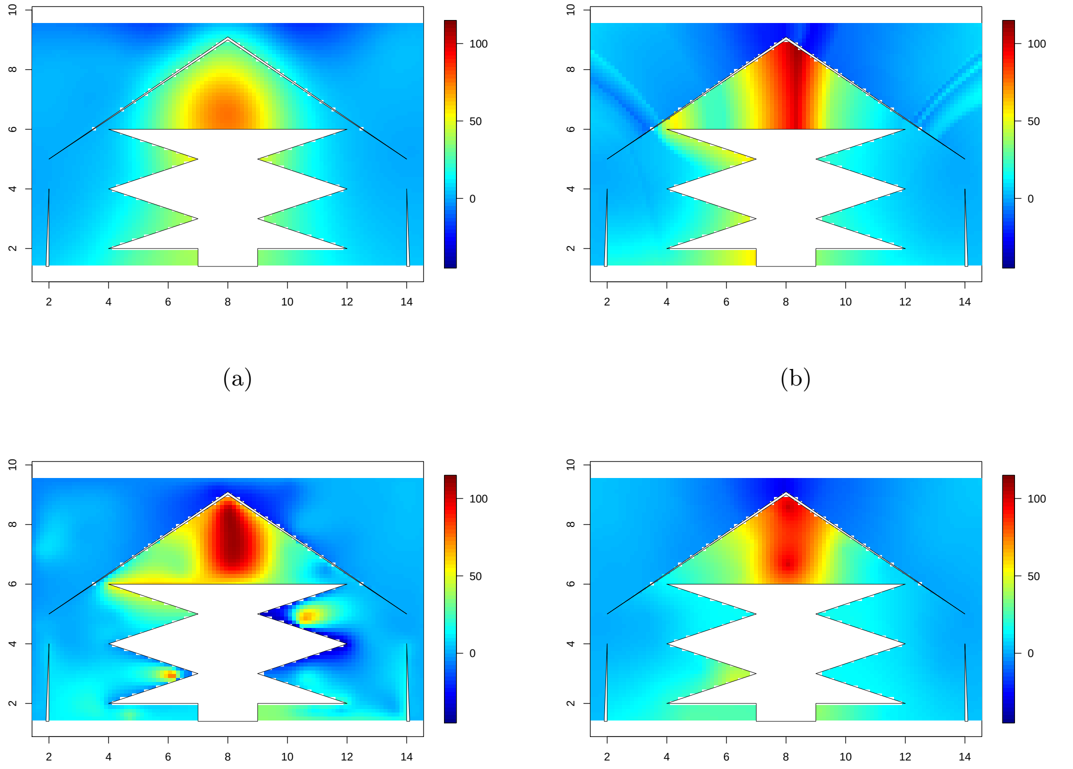
(a)

(b)

(c)

(d)

Figure 4.12: Bias for data sparse high noise, (a) TPS, (b) GLTPS, (c) SOAP and (d) CReSS



(a)

(b)

(c)

(d)

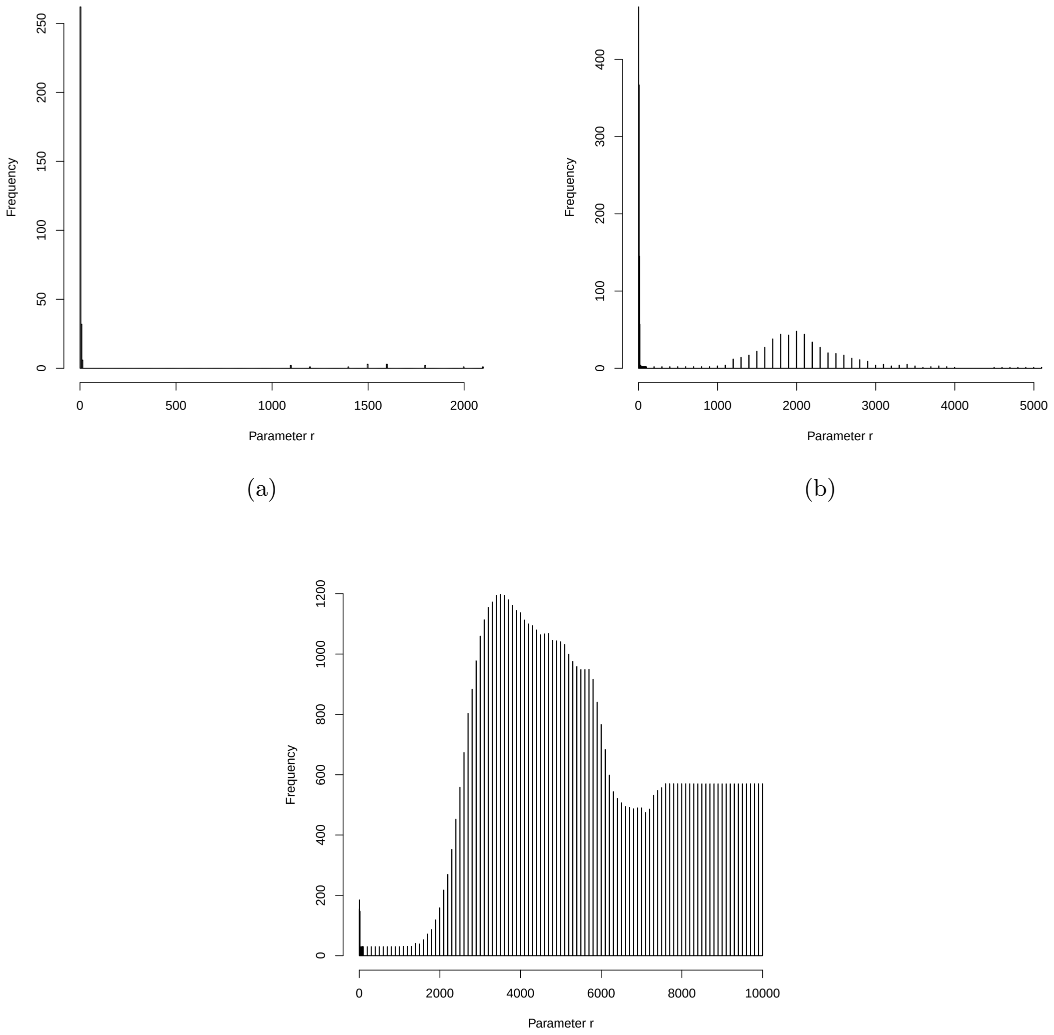
Figure 4.13: Example predictions for low noise, (a) TPS, (b) GLTPS, (c) SOAP, (d) CReSS (iteration 20)

In keeping with the data rich simulation, as noise level increased, the models for averaging became more smooth; the number of models averaged by CReSS and the value of parameter r increased (Table 4.6). However, in contrast to the data rich results there was no clear relationship between the noise level and the number of knots chosen (Tables 4.4 & 4.6). Figure 4.14 shows the distribution of the r values for low, medium and high levels of noise. Many more different values of r were chosen when there was high noise, and there was a definite shift in the distribution towards higher values. However, a number of models were chosen, at high noise, using small values of r.

Similar to the data rich results the MSE score did not necessarily improve because more models were averaged (Figure 4.15). When noise levels were high, between 56 and 1200 models, out of a possible 2010 (15 knot number choices and 134 possible r's), were averaged and Figure 4.15 shows the effect this had on MSE score. There may be some advantage in increasing the number of models averaged to about 600, but thereafter the MSE score increased.

Table 4.6: Parameter choices made by CReSS for each of the three noise levels averaged over 100 realisations of the simulation. The parameters include the mean number of models averaged per realisation, the mean for parameter r and the mean number of knots. Numbers in brackets show the minimum and maximum.

	Low	Medium	High
Mean Number of Models Averaged	3.14	15.96	621.6
	(1, 19)	(1, 83)	(56, 1200)
Mean r	73.15	680.7	5451
	(2, 2100)	(2, 5100)	(2, 10,000)
Mean Number of Knots	39.74	28.85	43.90
	(15, 75)	(10, 80)	(10, 80)



(a)

(b)

(c)

Figure 4.14: Distribution of parameter r chosen using 100 realisations of the simulation for (a) low, (b) medium and (c) high noise levels. The allowed choice of r ranged from 2 to 10,000.

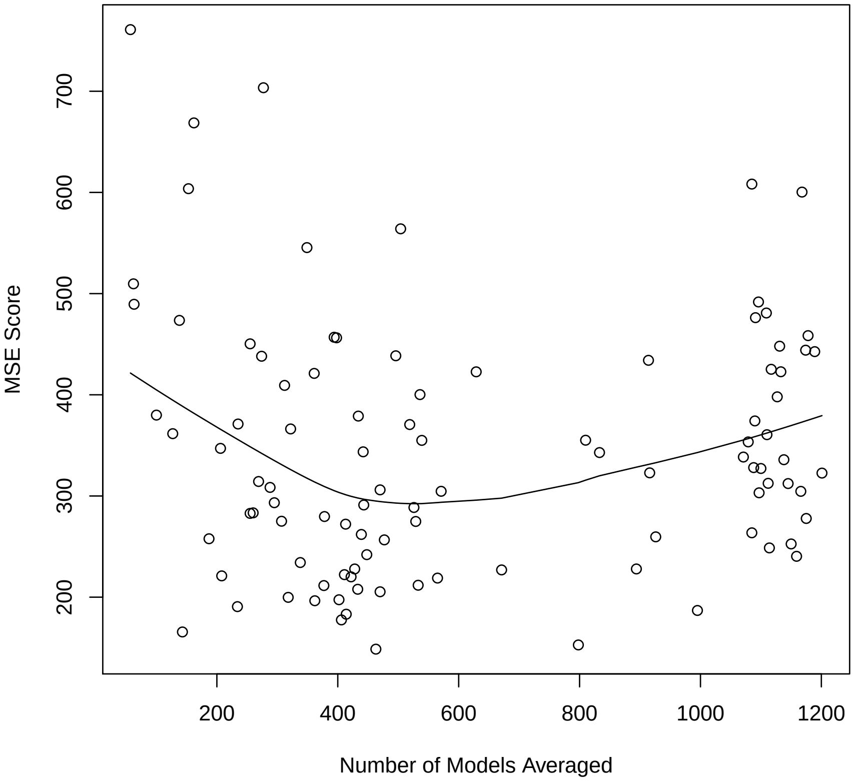


Figure 4.15: Variation in the MSE score with the number of models averaged. The points represent how many models were averaged and the resulting MSE score for each of the 100 simulation realisations. The line represents a locally weighted polynomial regression smooth of the data. The total number of models that could be averaged is 2010.

4.2.4 Discussion

These trials were conducted to determine how the methods perform when there is an island, and how they perform when there is very little data. The performance of TPS, SOAP and CReSS with the data rich complex simulation was similar to their performance with the horseshoe simulation in the Chapter 3. The narrow breakwater highlighted the issue of leakage for the TPS method, and the benefits of SOAP and CReSS in avoiding this are also clearly visible. However, GLTPS suffered from fitting artefacts, seen by visual assessment and caused by reinforcement, rendering the numerical results meaningless. SOAP became quite unstable when there was very little data and the increase in errors around the boundaries highlighted the sensitivity of SOAP parametrisation for boundary loops. CReSS was more stable at all noise levels, in part because it takes advantage of model averaging. In contrast, the high variance in the predictions made by SOAP and GLTPS in the data sparse trials was notable; when they went 'wrong' they went very badly wrong. Therefore, the recommendation is that both the GLTPS and SOAP methods are used with caution on sparse data sets.

The Cross-Validation (CV) method used here (10-fold) was not suitable for distinguishing between methods. It might have been better to use 5-fold CV but the choice is difficult. For 10-fold CV, the model uses lots of data for fitting (90%) and little for predicting (10%) so whilst 10 scores are averaged, the variance of those scores could be high. Five-fold CV averages fewer scores but each score is calculated from 20% of the data making them less variable. If the CV scores were very skewed, the median of the 10 values might be a better choice. AIC$_c$ was used to choose between models during the model averaging process for the CReSS method and picked good models (according to MSE scores) so it must be better at fitting models to out-of-set data than the CV calculated in this analysis.

A closer look at the CReSS results yields some interesting parameter choices. On average, across all noise levels, many more models and much larger values of r were chosen in the data poor trials than in the data rich ones. Given the potential reinforcement problems with global bases it was surprising to find models with such large r selected, but this may be a further indication of the advantages of model averaging. However, even when large values of r predominated, many small values were also chosen, even at high noise, which added some flexibility to the surface. Even greater surface flexibility could be obtained by having a different r for each knot location (rather than for a set of knots), but this might be difficult to implement. Comparing the choice of r results for the data rich simulation with those from the equivalent in the horseshoe simulation (Chapter 3) illustrates how the method adapts to different surfaces. Large, global bases (large r) were chosen for the simple horseshoe and more models were averaged to accommodate this. Small, local bases were chosen for the complex palm simulation, and fewer models were averaged.

CReSS is able to choose the most appropriate size bases for the surface and noise level. Choosing the right value for r seems less important than providing a sufficient range of choices. In the future, the procedure for selecting r (the range and resolution to step over) could be automated, so that the user need not worry about providing values for r. This could be linked with the use of variable r for each knot, mentioned above.

The results from this analysis, together with those from Chapter 3, suggest that CReSS performs best in a wider range of scenarios compared with the other complex methods evaluated here. CReSS can be used successfully on simple or complex regions, when data are rich or sparse, and at low or high levels of noise.

4.3 Case Study: Killer Whale (*Orcinus orca*) Behavioural Study

4.3.1 Introduction

So far this thesis has concentrated on assessing the performance of new statistical methods using simulated data. However, an important assessment of a new method is to examine its performance on real data. In this section CReSS is applied to a killer whale (*Orcinus orca*) data set that was introduced in Chapter 1.

The size of a killer whale population is often estimated using mark recapture analysis of photo-identification data [Ford et al., 2010, Ward et al., 2009]. This can provide accurate estimates of overall population size but a single number cannot reveal how killer whale density varies spatially, or the spatial distribution of different behaviours. However, accurate maps of the distribution of densities and the occurrence of behaviours would make decisions regarding spatial planning better informed.

The data analysed here were collected to aid effective decision making in the conservation of the endangered 'Eastern North Pacific southern resident' killer whale stock, by identifying areas where critical life-history processes such as breeding, weaning or feeding take place. Southern resident killer whales, hereafter referred to as SRKW, consist of three distinct social units (J, K and L) that return each year to feed on salmon returning from the Pacific to spawn. There has been a recent decline of SRKW, which led to the species being listed as endangered under the Canadian Species at Risk Act 2001 [Baird, 2001], and which may have been caused by a decline in prey abundance [Williams et al., 2011] and vessel-based disturbance [Williams et al., 2006, NMFS, 2006]. The main diet of SRKW is the Chinook salmon, *Oncorhynchus tshawytscha* [Ford and Ellis., 2006, Ford et al., 1998], which also happens to be the least common salmonid in the SRKW habitat [Quinn, 2005]. There are a many possible reasons for the decline in chinook salmon stock, for example habitat loss [eg:

Bilby and Mollot., 2008], harvesting [eg: Hoekstra et al., 2007], hydro damming of rivers [eg: Waples et al., 2007] and pollutants [eg: Missildine et al., 2005]. Furthermore, current levels of vessel disturbance (commercial and recreational whale watching), with typically 14-28 whale-watching vessels following a group [Erbe, 2002], have been shown to decrease the time killer whales spend feeding, and to have a lesser effect on other activities such as resting or socialising [Lusseau et al., 2009, Williams et al., 2006]. Lusseau and Higham [2004] make an important observation, that anthropogenic activity does not necessarily affect all behaviours evenly and so information about what is most affected and where certain behaviours take place is key.

One way to mitigate the effects of these anthropogenic activities is to identify areas which are particularly important for specific activities and restrict human activities there. In May 2011 a rule was introduced in inland waters of Washington State to prohibit vessels from approaching within 200 yards of killer whales and from parking in their path; there were also discussions for the establishment of a no-go zone [NMFS, 2011]. More information pertaining to this killer whale stock can be found in a research report by the National Oceanographic and Atmospheric Association [NOAA, 2011]. Since 1982 there has been a Marine Protected Area (MPA) for the Northern RKW population situated in Robson Bight, British Columbia, Canada, which was put in place to protect a rare rubbing behaviour on a particular smooth pebble beach Ford et al. [2000]. This was an obvious area to conserve due to the limitation of available beaches for this behaviour. However, there is no area for SRKWs in which a rare behaviour takes place. Ashe et al. [2010] suggested a candidate MPA site South of San Juan Island to protect feeding areas (Figure 4.16). This site was defined by both local knowledge (interviews with local environmental education coordinators) and spatial assessment of feeding behaviour. There is already an MPA, the Haro Strait exclusion zone [WDFW, 2013], in this area for sea cucumbers and sea urchins but there is no restriction on salmon fishing or other human activities.

A previous spatial assessment by Ashe et al. [2010] of the SRKW data used a simple two-dimensional Thin Plate Spline (TPS) smooth of Latitude and Longitude in a Generalised Additive Model (GAM) framework to predict probability of feeding at particular locations. This section improves on this by accounting for both spatial autocorrelation and the geographic complexity of the study area (multiple islands). The potential problem of leakage seen with extrapolating TPS (Chapters 3 and 4), and the large number of islands present in the region mean that the killer whale dataset benefits from the methods developed in this thesis. Like Ashe et al. [2010] we focus on a simple model that uses a two-dimensional smooth of spatial coordinates to predict the probability of feeding.

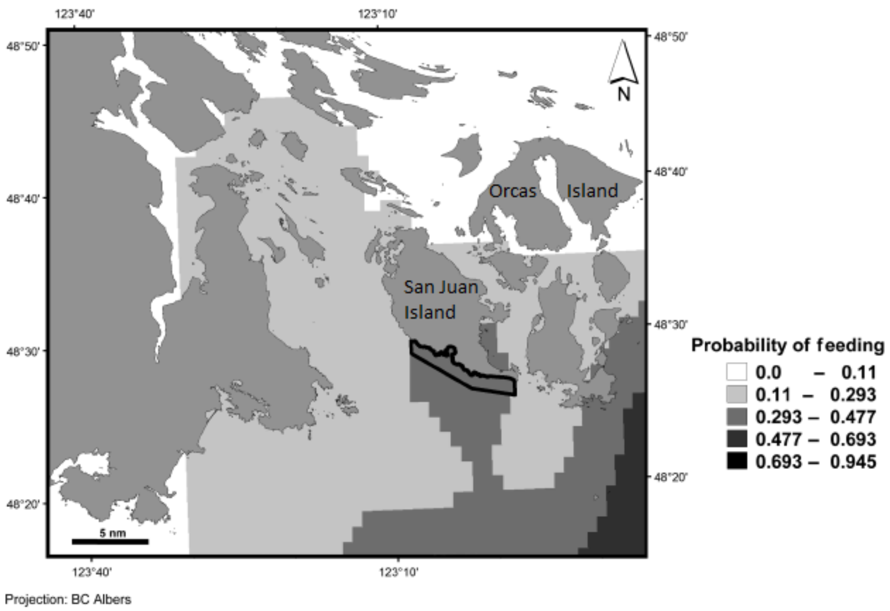


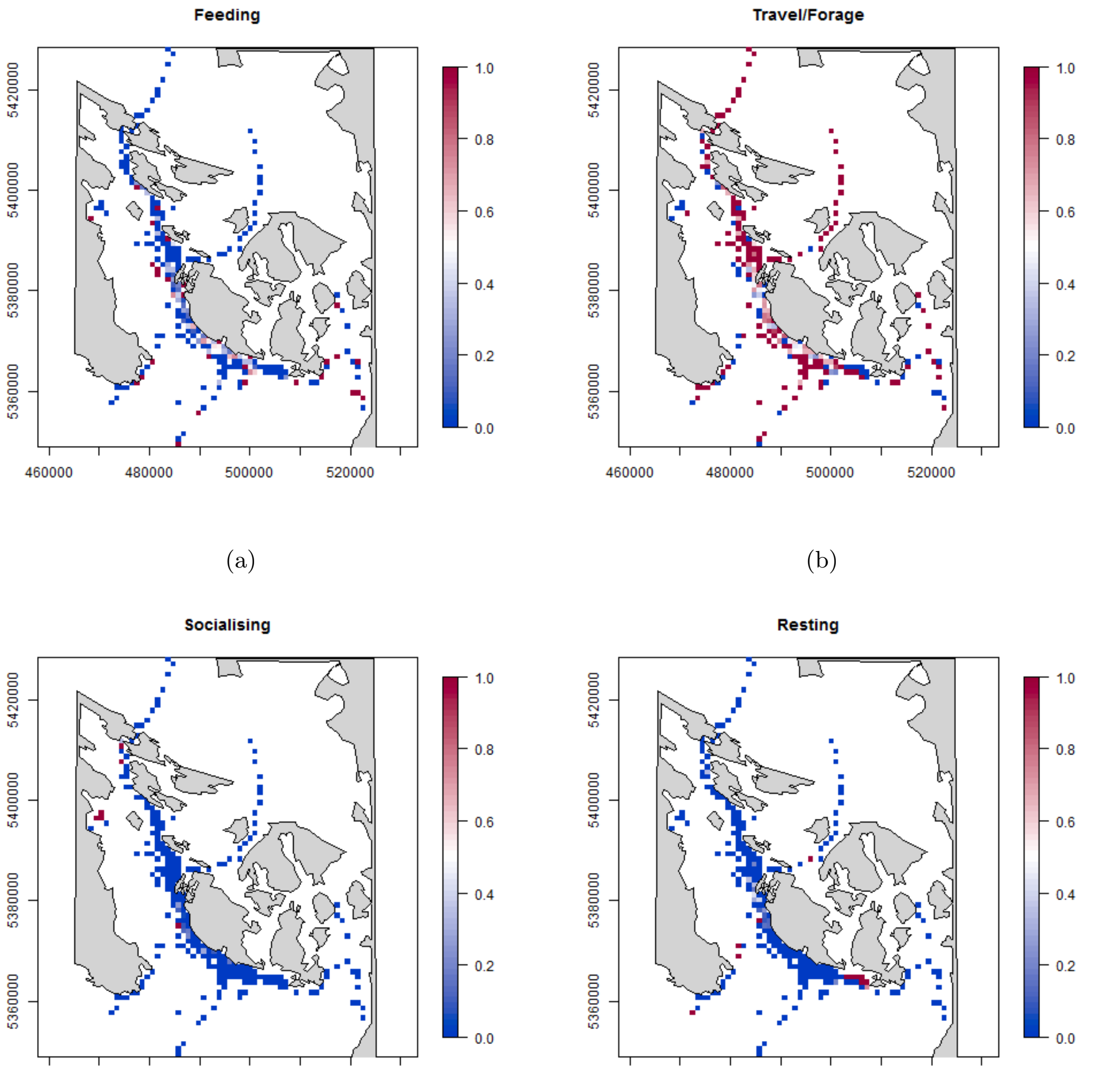
Figure 4.16: Figure 3 from Ashe et al. [2010] showing the predicted probability of feeding by SRKW. The box to the south of San Juan Island has since been proposed as an MPA to protect killer whales feeding.

4.3.2 Methods

The inshore waters around San Juan Island, Washington State (USA) and adjacent Canadian waters (British Columbia) form a complex area of coastline with at least 15 major islands (Figure 4.17). The killer whale data used in this analysis were collected by small boat from May to August 2006 from which five observers searched for killer whales and recorded their location. Once a pod was identified it was followed and the main activity of the pod was recorded every 10 minutes. There were four recorded activity states: travelling/foraging, resting, socialising and feeding. Definitions of each of these states may be found in Ashe et al. [2010].

The data consist of $n = 763$ pod sightings, where pod size ranged from 1 to 50 and all three social groups (J, K and L) were observed during the study. Each observation has an associated binary indicator, for example, $p = 1$ for feeding or $p = 0$ for non-feeding (considered to be travelling/foraging, resting or socialising). Of the 763 data points, travelling and foraging was the most common ($n = 485$) activity and socialising the least common ($n = 28$). Of the remainder, 188 observations were of feeding and 62 of resting. The observed data for each activity state are presented in Figure 4.17 and illustrate proportions rather than the observed binary values; each cell on the plot is approximately 1 km^2 and the colour of each cell represents the mean of the data points recorded within it. Pod size, identification of individuals within each pod and social group (including mixed pods) was also recorded.

This analysis focuses on feeding, given the known effect of anthropogenic disturbance on this activity [Williams et al., 2006, Lusseau et al., 2009]. Therefore the proportion of groups in a feeding state per km^2 was modelled using spatial coordinates as the only covariate. A more descriptive surface might be produced by including other covariates, such as depth or chlorophyll, however the two-dimensional smooth used here was designed to show the potential value of CReSS and to allow a direct comparison with the results of Ashe



(a)

(b)

(c)

(d)

Figure 4.17: Raw proportions for (a) feeding/non-feeding, (b) travel or forage/ not travel or forage, (c) socialising/non-socialising and (d) resting/not resting of killer whales off the West coast of the USA/Canada. The grid cell size is approximately 1 km^2.

et al. [2010]. The coordinates were projected to UTM10U (Universal Transverse Mercator projection), since, at the latitude of the survey area, the scales of latitude and longitude are quite different and this has consequences for the distance metric. The coordinates are subsequently referred to as *Eastings* and *Northings*. Furthermore, due to the repeated measures on killer whale pods and data collection through time, it is likely that there will be correlation within the model residuals. Therefore, Generalised Estimating Equations (GEEs) [Hanley et al., 2003, Liang and Zeger, 1986, Hardin and Hilbe, 2002, Harrison and Hulin, 1989] were used to allow for any autocorrelation in the residuals. This is a common way to deal with autocorrelation, for example Panigada et al. [2008] where GAM based methods were employed to model the mean and with GEEs to generate measures of precision, such as standard errors. The CReSS method is modular and easily implemented in this framework.

4.3.2.1 Generalized Estimating Equations (GEEs)

GEEs were particularly useful in this analysis because the repeated binary measures on the killer whale groups were likely to be spatially and/or temporally auto-correlated. For example, the probability of killer whales feeding at any particular location in space and time are likely to be more similar for points close together in time compared with points distant in time due to environmental/prey conditions. Additional covariate information could be used to explain this but it is often unavailable or unknown. If a pod is feeding at a particular time step, this is likely to increase the likelihood that it will be feeding at the next time step, leading to positive auto-correlation and, if unexplained by the model, sequences of positive or negative residuals, rather than the random scatter assumed under a GLM/GAM. If the assumption of independence of consecutive residuals is violated, because of positive autocorrelation, then this invalidates all model-based estimates of precision (e.g. standard errors). The point estimates from a GEE can be the same as for an equivalent GAM/GLM (depending on the correlation structure chosen for the GEE) but the uncertainty is inflated (for positive autocorrelation) using a GEE.

To check that autocorrelation is present in model residuals, a runs test (e.g. runs.test from the R library lawstat) [Mendenhall, 1982] can be used to test for statistically significant levels of spatio-temporal auto-correlation in model residuals. Generally, data are ordered through time and so the runs test will check for temporal auto-correlation by comparing the number of sequences (runs) of positive or negative residuals with the number that would be expected under the assumption of independence. If positive correlation is present there will be fewer uninterrupted runs (few long strings of positive or negative residuals) than would be expected and results in a negative test statistic and small p-value (p <0.05).

The GEE approach requires the specification of a panel variable. Residuals within panels are correlated but they are assumed to be independent between panels [Hardin and Hilbe, 2002]. GEEs allow the estimation of standard errors to be adjusted for the autocorrelation in the panel residuals. Panels, also known as the blocking structure, can be chosen using information about survey design and/or autocorrelation function (acf) plots [Venables and Ripley, 2002]. The latter illustrate the autocorrelation for a variety of lags between measurements. Data collected at the same point in time are assumed to have identical residuals, correlation =1, and this correlation is then estimated for various time lags between points. In a GEE the nature of the correlation within a panel can either be assumed to follow a particular model chosen by the user (e.g. AR1, Exchangeable) or data based sandwich estimates of variance can be used. Both here, and again in Chapter 7, empirical standard errors were used, so specific details of correlation structures is not included. Hardin and Hilbe [2002] provide a comprehensive review of correlation structures.

An assessment of survey design led to the killer whale social group on any given day being used as the panel variable, since behaviour for the same social group within days, and behaviour for different pods within social groups (on any given day) are likely to be correlated. There were eight social group factor levels (J, K, L and some mixed groups) and 40 survey days. Therefore, *group-day* was used to define the panels within which residuals were permitted to be correlated. There were between 1 and 31 observations on each *group-day* ($j = 1, .., 31$) and 58 *group-days* overall ($i = 1, ..., 58$). The runs test for the killer whale feeding data showed a significant level of positive autocorrelation ($p << 0.01$) and an acf plot, illustrating the mean correlation across panels for each time lag (Figure 4.18(a)), showed the correlation to decay to approximately zero, even within the smallest panels (Figure 4.18(b)). This meant that *group-day* was a suitable panel variable for this data.

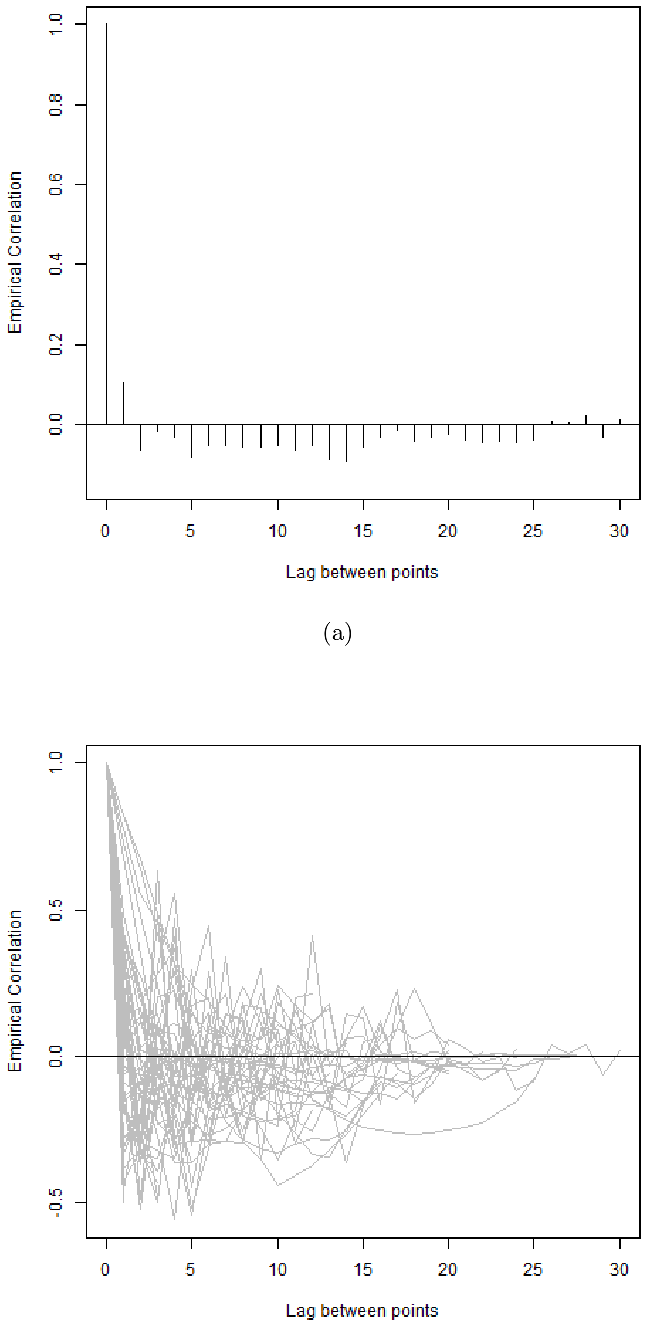
Due to the binary nature of the data, a Binomial based model was used for this analysis with a logit link to ensure predictions lay between 0 and 1. A two dimensional smooth term was used to model the distribution of feeding behaviour inside the GEE framework:

$$\eta_{ij} = \beta_0 + s(\mathbf{X}_{i,j}) \tag{4.1}$$

where η_{ij} represents the additive predictor, s is a smooth function and $\mathbf{X}$ is a matrix ($n \times 2$) of spatial co-ordinates observed for panel i at time j. Many flexible models can result from this specification, and in this analysis a CReSS basis expansion was used for the two dimensional smoother which results in a predictor (with T terms) which is linear in its parameters:

$$s(p_{ij}) = log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \eta_{ij} = \beta_0 + \sum_{t=1}^{T} \delta_t b_{tij} \tag{4.2}$$

where p_{ij} is the probability of a pod feeding for panel i and time j and b_{tij} represents a set of basis functions ($b_{tij}, t = 1, ..., T$) for panel i at time j for a two-dimensional smoother. This equation is similar to the GLM formulation in Equation 2.5, Chapter 2, for a Poisson model with log link function. GEEs can allow for overdispersion by estimating the dispersion parameter (ϕ; Chapter 2), however overdispersion is not possible for binary data [Faraway, 2006] and so $\phi = 1$ for this analysis.



(a)

(b)

Figure 4.18: Auto-correlation plots for the killer whale feeding model residuals. (a) The mean correlation across panels for each time lag. (b) The correlation for each individual panel, where each line is a panel.

4.3.2.2 CReSS

The CReSS method is used inside the GEE framework and so the smooth function in Equation 4.1 and b in Equation 4.2 is the local exponential radial basis function seen in Chapter 3. The GEE enables the estimation of model coefficients and the associated uncertainty (that accounts for the positive correlation in the residuals).

The local basis function requires the input of geodesic distances due to the complex topography in the survey region, in addition to the r parameter which influences the effective range of each radial basis and is permitted to vary across candidate models. For the calculation of geodesic distance, 17 exclusion polygons (defining the coastline) were considered in the analysis and a range of 20 r values ($r_{\text{min}} = 120$, $r_{\text{max}} = 2907$) were considered for each of the candidate models. Model selection criteria were used to discriminate between the different models.

Parameter r dictates the effective range of the radial basis; a large value for r returns a relatively global basis function while a small value for r returns a locally acting basis. Since r is always unknown, multiple values of r were considered and the resulting models were averaged using model weights (Equation 3.2) calculated from the information criterion used to determine model fit. The information criterion used for this analysis is detailed in the next section.

To allow for a range of candidate models with different flexibilities, models with different knot sets consisting of different numbers of knots in each set ($T = 5, 10, ..., 55, 60$) were considered for selection. To maximise spatial coverage for any particular knot number, the knot locations for any given knot set were chosen using a space-filling algorithm [John et al., 1995].

The model for each knot set is re-fitted using different values of r and so the model averaging is calculated over all r and all knot sets. Twelve knot sets were used and 20 r's, thus of 240 models (20×12) were fitted and available for averaging.

4.3.2.3 Prediction and Inference

Model-averaged predictions were obtained by generating predictions onto a grid based on each candidate model and averaging these predictions in line with model weights. There is much debate over the choice of fit statistic to use for GEE models [Pan, 2001a,b]. The fitting procedure is based on a quasi-likelihood, so model selection criteria should also be based on quasi-likelihood, rather than maximum likelihood based scores. Therefore, the weighting procedure was governed using an AIC analogue for GEEs: Quasi-likelihood under the independent model Information Criterion (QICu) [Hardin and Hilbe, 2002, Pan, 2001a] statistic;

$$QICu = -2Q + 2q$$

which for Binomial data has quasi-likelihood, $Q = \sum_{i=1}^{n} \sum_{j=1}^{n_i} y_{ij} \log\left(\frac{p_{ij}}{1-p_{ij}}\right) + \log(1 - p_{ij})$. Here, y is the binary outcome (feeding/not feeding), p are the fitted values evaluated at the quasi-likelihood estimates under the GEE model and q represents the number of estimated coefficients. Based on the QICu scores the associated model weights (w_m) for the m-th model ($m = 1, ..., 240$) were obtained using Equation 3.2. Like AIC, smallest values of QICu are preferred.

Percentile based 95% confidence intervals for each grid cell were also obtained by generating 1000 parametric bootstrap realisations from each GEE based model and averaging these in line with their QICu weights each time. The central 95% of these values across all models were then used to delineate the upper and lower confidence limits for each grid cell.

Diagnostics for binary data are notoriously tricky, however we can assess predictive power of the final model using deviance R^2 and confusion matrices [Pearce and Ferrier, 2000]. The deviance R^2 is calculated as follows:

$$\mathrm{R}^2 = \frac{1 - \exp\left(\frac{D - D_{null}}{n}\right)}{1 - \exp\left(\frac{-D_{null}}{n}\right)}$$

where D is the deviance of the model of interest, D_{null} is the deviance of the null model and n the number of observations. R^2 takes values between zero and one and a value close to one indicates the model fits well to the data.

To construct a confusion matrix, a threshold, p, is chosen to turn the predicted proportions, $\hat{p_{it}}$, into binary feeding (1) or not feeding (0). There are various subjective and objective approaches to determining this threshold, for example, index, data or prediction based methods [see Liu et al., 2005, for a review of methods]. In this analysis the mean of the fitted values was used [Hosmer and Lemeshow, 1989, Cramer, 2003]. This is particularly useful when there is an inequality in the number of zeros and ones in the data. The binary fits were then used to construct a confusion matrix, which specifies the number of ones in the data predicted as ones and the number of zeros predicted as zeros. False positives and false negatives are specified in a similar way (Table 4.7). Each of the cells in the table (a, b, c and d) can be used to calculate three useful indices; sensitivity, specificity and overall prediction success. *Sensitivity* is the proportion feeding correctly predicted as feeding ($a/(a + c)$), *specificity* is the proportion of not-feeding correctly predicted as not-feeding ($d/(b + d)$) and *Overall Prediction Success* (OPS) is the percentage of correctly allocated predictions ($(a+d)/(a+b+d+c)$). OPS can be deceptively high when frequencies of zeros and ones in the data are very different [Pearce and Ferrier, 2000] as we have in this analysis; only a quarter of the data are ones.

4.3.3 Results

The best model (QICu = 795) used 10 knots and $r = 641.8$. However there were 59 models that had a delta QICu<6, above which the weight of the models is effectively zero and they

Table 4.7: The output of a confusion matrix.

		True Values	
		1	0
Predicted	1	True Positive (a)	False Positive (b)
Values	0	False Negative (c)	True Negative (d)

were not part of the set for model averaging. These 59 models ranged from 10 to 45 knots (mean = 17.54) and used values of parameter r from 141.7 to 2907 (mean = 935.3). Figure 4.19 shows the distribution of chosen models with parameter r with each knot number. The weighted average of QICu scores was 796.3.

The deviance R^2 for the averaged models was quite low at 0.141 and the models only explained 9.9% of the deviance. The predicted outcomes were converted into feeding/not feeding (1/0) predictions by assigning each predicted probability a zero or a one using a threshold, above which takes a one and below a zero. The threshold used here was the mean probability of the averaged fits, 0.246. The confusion matrix (Table 4.8) was used to calculate the three indices. The averaged models predicted the probability of feeding to be one, when a one was observed, 66% of the time (sensitivity). Conversely, a zero was predicted 64% of the time a zero was observed (specificity). The overall prediction success was 66%.

Figure 4.20 shows the fitted values and point estimate predictions for the probability of feeding. The fitted values (Figure 4.20(a) and 4.20(b)) correspond reasonably well to the observed values (Figure 4.17). The predicted outcomes were also converted into feeding/not feeding (1/0) predictions using the threshold above. These converted plots are more easily compared with the data and may enable easier delineation of feeding zones. The prediction plots show a high probability of feeding to the far south west and south east of the survey

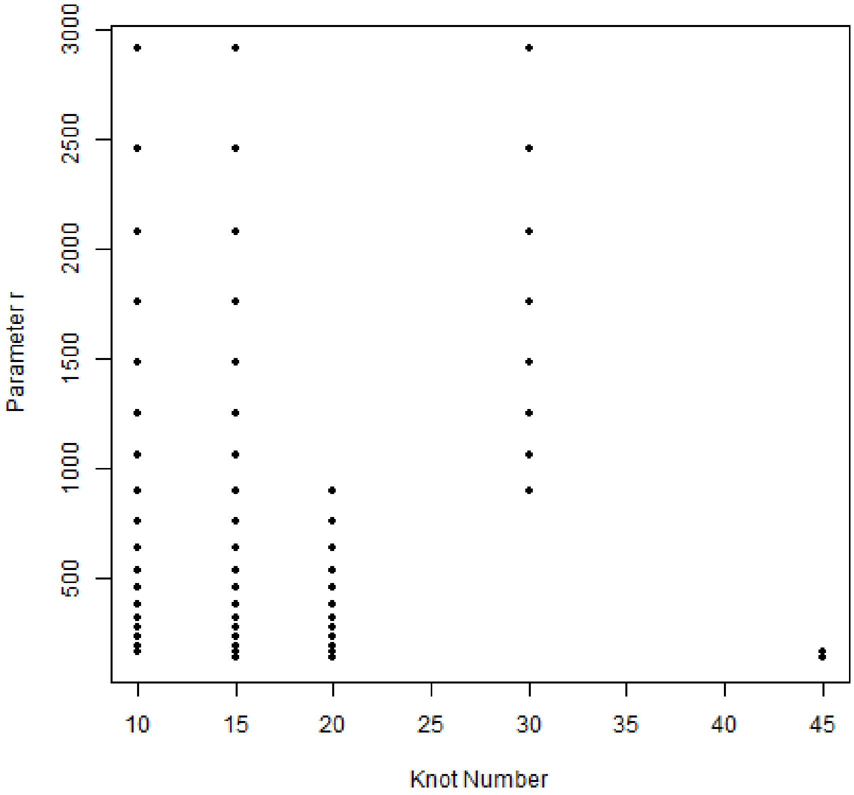
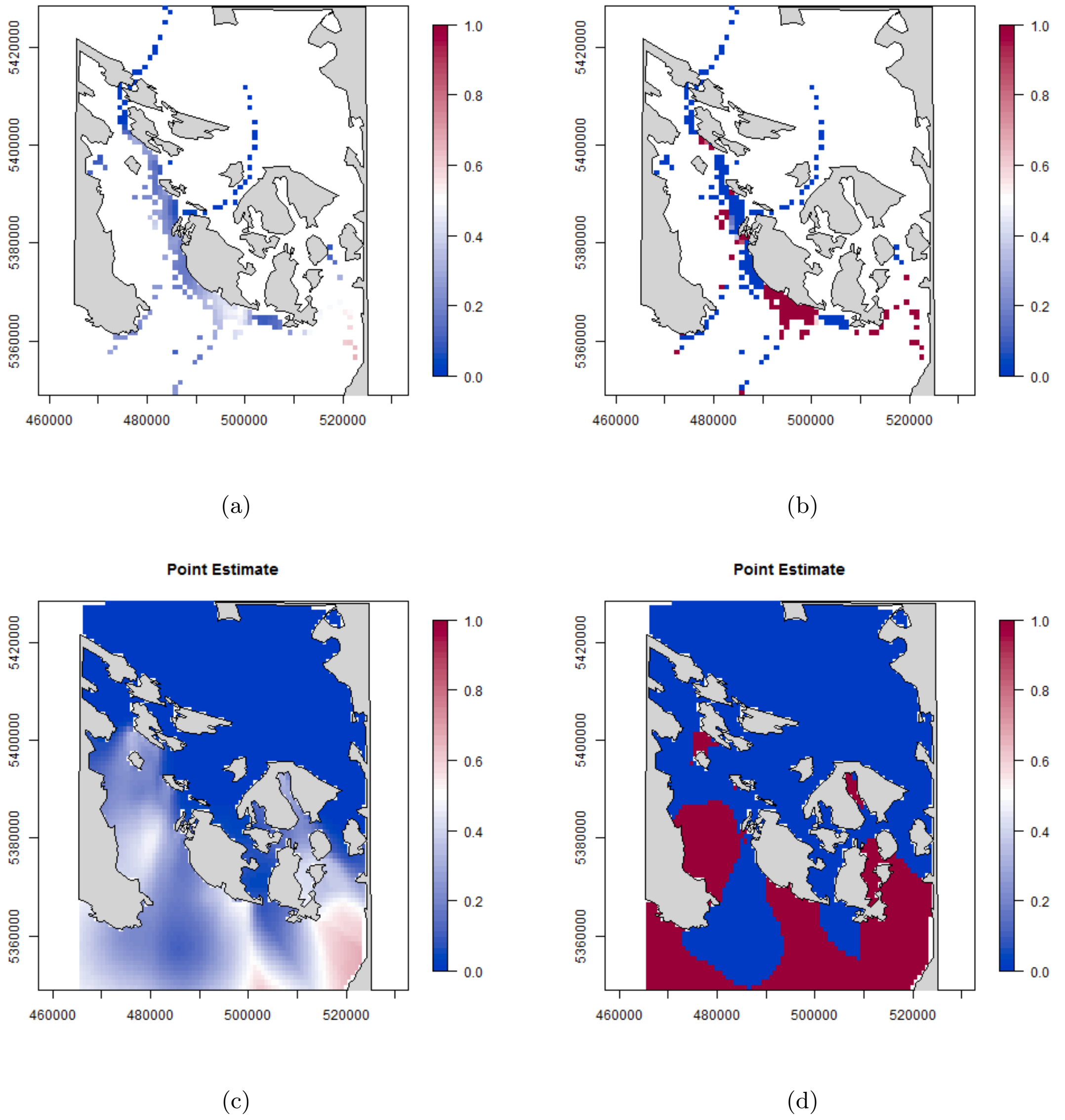


Figure 4.19: Parameter r and knot number of the models in the candidate model set.

region and to the south of San Juan Island (the main island in the centre). Percentile based 95% confidence intervals from the GEEs are shown in Figure 4.21. Much of the pattern seen in the point estimate surface is retained both at the lower and upper confidence limits.

Table 4.8: Confusion matrix for the averaged models.

		True Values	
		1	0
Predicted	1	125	198
Values	0	63	377



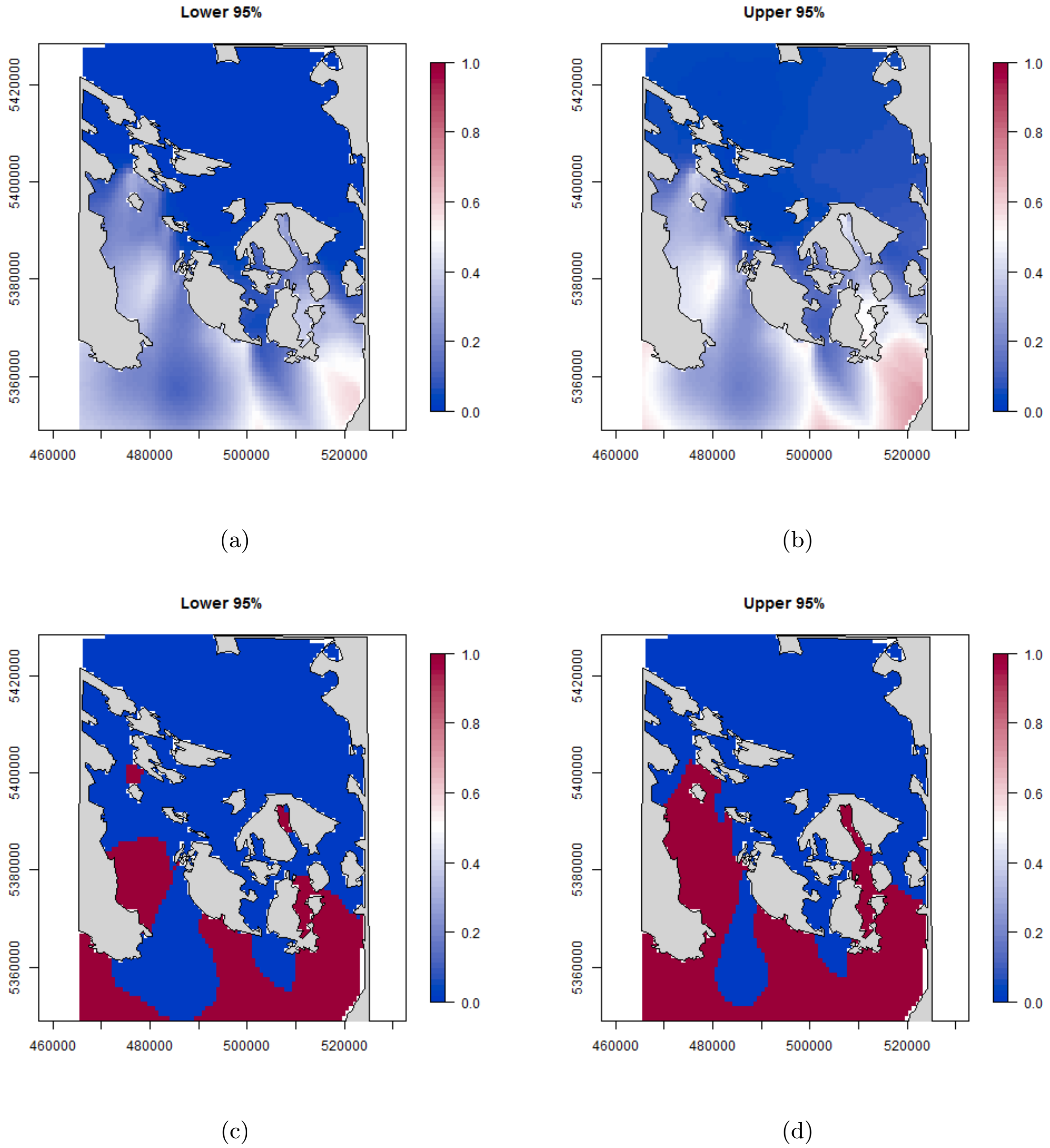
(a)

(b)

(c)

(d)

Figure 4.20: Fitted values (a) and (b) and predictions (c) and (d) for the probability of feeding (1=feeding, 0=not feeding). The probability cut-off for (b) and (d) is 0.246.



(a)

(b)

(c)

(d)

Figure 4.21: Percentile based 95% confidence intervals for predictions for the probability of feeding (1=feeding, 0=not feeding). The probability cut-off for (c) and (d) is 0.246.

Comparison to standard GAM model

Figure 4.22 shows predictions from the CReSS model using the same thresholds as those used in Figure 3 (Figure 4.16) of Ashe et al. [2010]. The surface for CReSS is more flexible than the original but still shows a high probability of feeding to the south of San Juan Island.

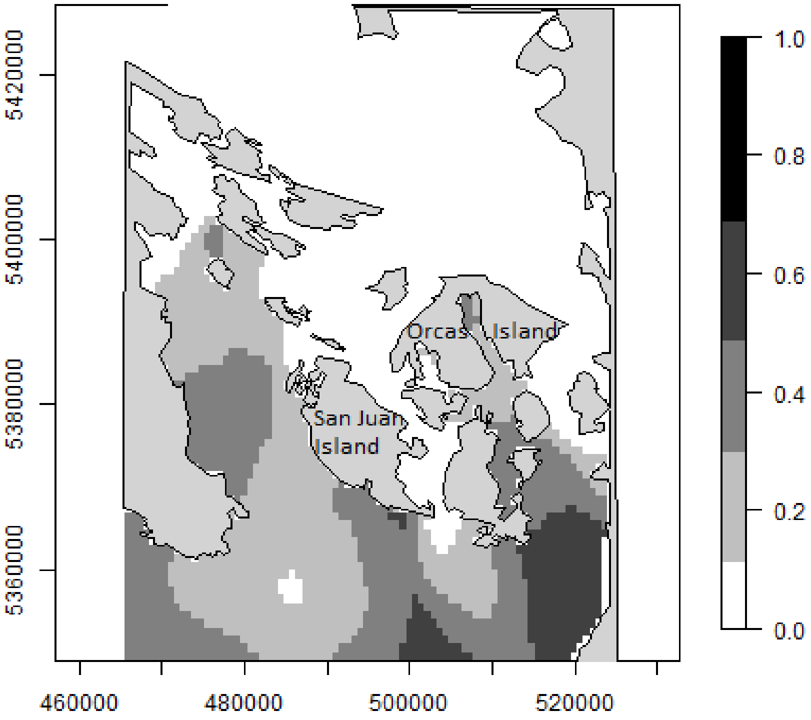


Figure 4.22: CReSS predictions for the probability of feeding (1=feeding, 0=not feeding) using the threshold values in Ashe et al. [2010].

The model presented in Ashe et al. [2010] was re-fitted using gam from the R package mgcv [Wood, 2006] so that uncertainty could be evaluated. This was a binary GAM model containing a two dimensional smooth of space with binomial errors. The AIC score was

807, adjusted $R^2 = 0.122$ and 9.8% of the deviance was explained by the model. Using the confusion matrix (Table 4.9), 59% of the ones were correctly predicted (sensitivity), which was worse than the CReSS model but more of the zeros were correctly predicted (specificity, 71%). The OPS was 68%, but as mentioned above this is likely to be high due to the large number of correctly specified zeros.

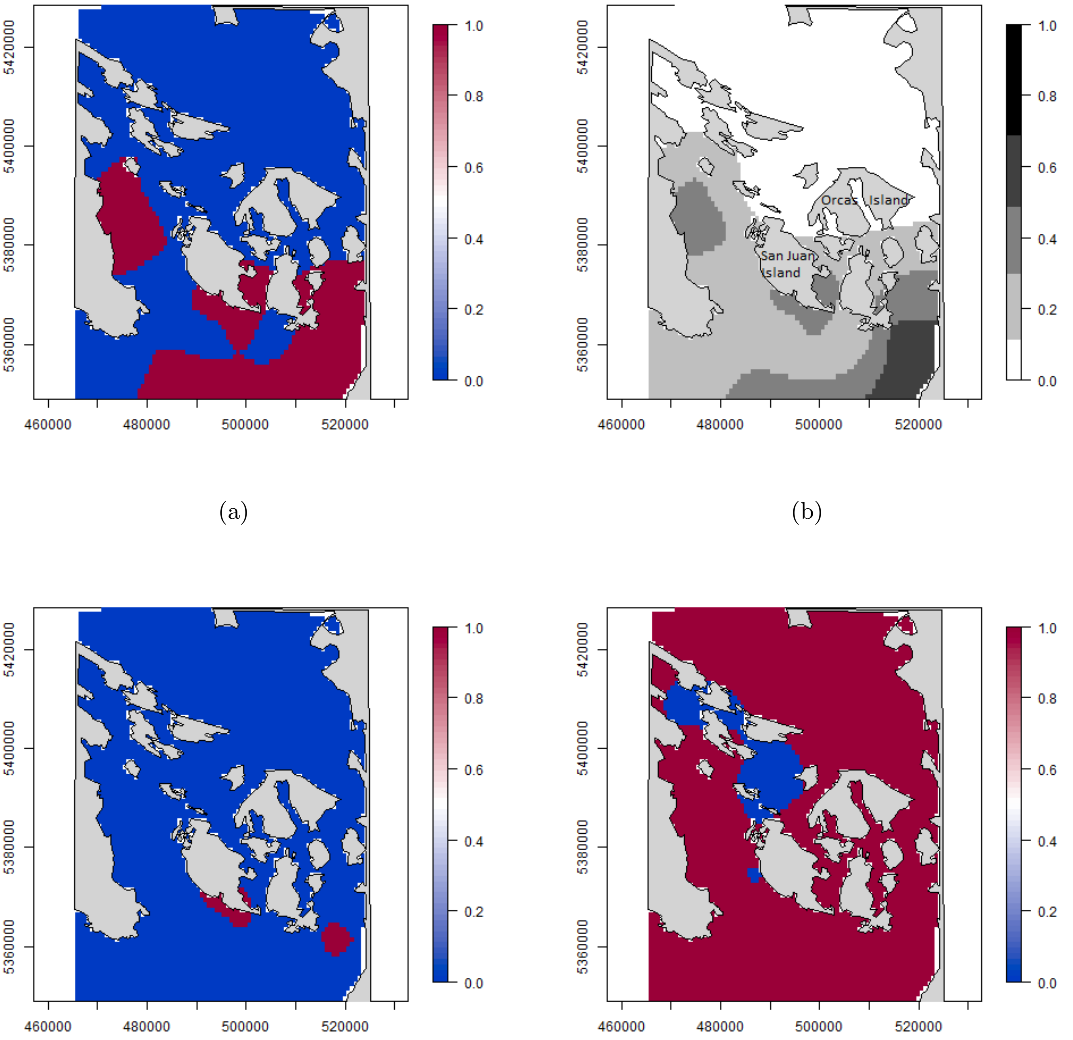
Table 4.9: Confusion matrix for the GAM model.

		True	**Values**
		1	0
Predicted	1	111	164
Values	0	77	411

The predictions here are very similar to those in the original paper, however uncertainty was calculated using 95% confidence intervals (Figure 4.23) and autocorrelation, which can be seen in the residuals, was not accounted for in this analysis. This would lead to wider confidence intervals, so the upper 95% limit could predict feeding for most of the surface. Like the CReSS method, a high probability of feeding is predicted to the south of San Juan Island, however there is also a high probability of feeding to the east of this island that could indicate some 'leakage' of this hotspot. With no data in this area, this cannot be confirmed.

4.3.4 Discussion

These results show that CReSS is a useful and flexible modelling tool for assessing the spatial distribution of behavioural states, such as feeding. Whilst the GAM model had a better specificity and overall prediction success the CReSS model had a better sensitivity (ones predicted as ones). The ones are the observed feeding locations so, in terms of an MPA



(a)

(b)

(c)

(d)

Figure 4.23: Predictions for probability of feeding using the 0/1 threshold (a), the Ashe thresholds (b) and 95% confidence intervals for the probability of feeding (1=feeding, 0=not feeding) (c) and (d).

for feeding behaviour, it is more important to predict these well rather than the non-feeding locations. When there is an inequality in zeros and ones in the data, in favour of zeros, it is easier for a model to predict more of the zeros, particularly for a model, such as GAM, that spreads the modelling effort over the whole survey region. The CReSS model, through model averaging and local bases, is a more focused approach leading to better prediction of ones. The CReSS model also had a better fit to the data using deviance explained and deviance R^2 than the GAM model.

Graphically CReSS produced similar results to those reported in Ashe et al. [2010] and also supported the case for an MPA in the same area. However, CReSS was able to present the uncertainty in the predicted surface and gave a more structured surface than the GAM model. With GAM re-fitted to provide 95% confidence intervals, these were wider than those seen for the CReSS model, and would have been even wider if autocorrelation had been accounted for. A comparison of a GLM with *Eastings* and *Northings* as covariates with an equivalent GEE showed a 27% increase in the standard error of the intercept estimate. Therefore, the upper limit for the GAM predictions could show the probability of feeding to be one across nearly the whole study region. However, even assessing only the lower 95% interval, there is still a high probability of feeding just under San Juan Island for both the GAM and CReSS models. There is some evidence of 'leakage' through the south east of this island from the GAM model, which would be unsurprising given the Euclidean distance metric, but unfortunately there are no data to support this conclusion.

On a statistical note it is interesting that the feeding surface for CReSS is more structured than that for the GAM, which is indicative of the variable smoothness allowed by CReSS across the surface, and is achieved by averaging models with different range parameters. GAMs have only one smoothness parameter that cannot change across the surface, so complex structure in some areas of the data may be smoothed through based on smooth needs in others, creating a simpler probability surface. The range of r parameters selected was surprisingly wide for a surface with so many islands. However, the potential issue of re-enforcement with global bases has not arisen (see Chapter 3 for details on re-enforcement).

A map showing only the probability of feeding is probably not sufficient for identifying a potential MPA. The model use here was based only upon locations were animals were seen and, whilst there may be a high probability of feeding in a particular area, the probability of presence may be very low, leading to an area where killer whales are rarely seen being identified as a hotspot for feeding. Hauser et al. [2007], used kernel density estimates of location data from 1996 to 2001 to identify areas which were used intensively by the three social groups of SRKW. They showed that the area to the south of San Juan Island is commonly used by all three social groups, which spend a disproportionate amount of time there. This suggests that an MPA sited in this area would benefit all three social groups. They also showed that an area to the far west of the study region (the northern parts of the Strait of Juan de Fuca), another area where there was a high probability of feeding seen in the results, was commonly used by the L social group. Hauser et al. [2007] also questioned the behavioural use in these areas and it is hoped that the results here are able to provide a greater insight into this.

The CReSS model suggests a high probability of feeding in an area to the far south east of the study area, however results from [Hauser et al., 2007] imply that this is not commonly visited by SRKW. The raw data indicate that when whales were seen in this area they were either feeding or travelling/foraging (Figure 4.17), which fits with the suggestion that they are infrequently present but travel there to feed. The high feeding probability area to the north, almost entirely surrounded by one island (Orcas Island), does perhaps seem a little unrealistic. There are no data to support the high probability of feeding and no previous studies suggest a high presence of SRKW in this area. This could be a limitation of the model through extrapolation and/or the limited covariates used.

There is a deep channel just to the south and west of San Juan Island with steep sides on the coastal side (see Appendix C), where high densities of killer whales were recorded in both Lucas [2009] and Hauser et al. [2007]. This suggests there may be a relationship between SRKW feeding and water depth. Furthermore, the main prey of SRKW, Chinook salmon, prefer deeper water at night than some other salmonid species [Candy and Quinn, 1999, Walker et al., 2007], so SRKW may prefer such areas for feeding. Candy and Quinn

[1999] also showed that the differences in depth use by Chinook salmon by day and by night were small, which is useful given the sightings for feeding are collected during the day. Thus depth could be one of a number of reasons why the San Juan area (Haro Strait) is a good feeding ground for SRKW. The inclusion of environmental covariates, such as depth, might improve predictions in areas that are shallow such as the area within Orcas Island. However, given this analysis was based only on locations where animals were seen it would be advisable to use the feeding map produced in conjunction with a presence/absence map (that may be based on multiple environmental covariates). If presence within the Orcas Island was low then it is unlikely that an MPA would be sited here based on a high probability of feeding.

Ultimately, protecting areas that are used as feeding grounds by SRKW may be of little benefit if Chinook salmon stocks continue to decline. Hanson et al. [2010] collected faeces of SRKW, particularly in the San Juan Islands area, and used genetic analysis to identify the specific spawning rivers used by the salmon they had consumed. More recently, Ayres et al. [2012] investigated the effects of vessel disturbance and inadequate prey and suggest that 'identification and recovery of strategic salmon populations are important to effectively promote SRKW recovery'. These studies suggest focusing conservation efforts for the prey species and are a good example of how creation of an MPA could have little effect if other factors are not also taken into consideration.

4.4 Summary

The initial simulations with the palm surface showed that CReSS is effective in areas with islands and in both data rich and sparse areas. Of the other complex methods, GLTPS showed re-inforcement issues (see Section 3.2.2, page 55 for details) leading to questionable surfaces even though numerical results were good. SOAP was difficult to parametrise for this example and gave numerical results that were worse than CReSS, particularly when data were sparse.

The analysis of the feeding distribution of SRKW demonstrated the use of CReSS in practice. It is a flexible modelling tool that can be used for quantitative aspects of spatial conservation planning. It is likely that use of GLTPS or SOAP for this analysis would have led to re-enforcement and parametrisation issues respectively [Scott-Hayward et al., 2013]. The model for this analysis may have been improved by the inclusion of environmental covariates (e.g. depth) and information on the actual distribution of killer whales.

Finally, the knot locations used for both the simulation and case study could be improved. A space-filling design does not easily allow flexibility to vary across the surface. The smoothness of the surface we are trying to approximate may vary, requiring more flexibility, and therefore more knots, in some regions than in others. In the next chapter, we investigate how CReSS can be combined with a spatially adaptive knot placement algorithm to accommodate locally varying complexity more easily.

Chapter 5

Spatially Adaptive Models for Complex Topographies

In certain cases, the space-filling knot placement approach used so far in this thesis does not always achieve good results. Results of previous analyses also suggest that knot placement is of great importance, with knots in different locations giving quite different outcomes. Here we present an extension of a Spatially Adaptive Local Smoothing Algorithm (SALSA; Walker et al., 2010), which is used in combination with the CReSS method to address the knot location side of this model selection issue.

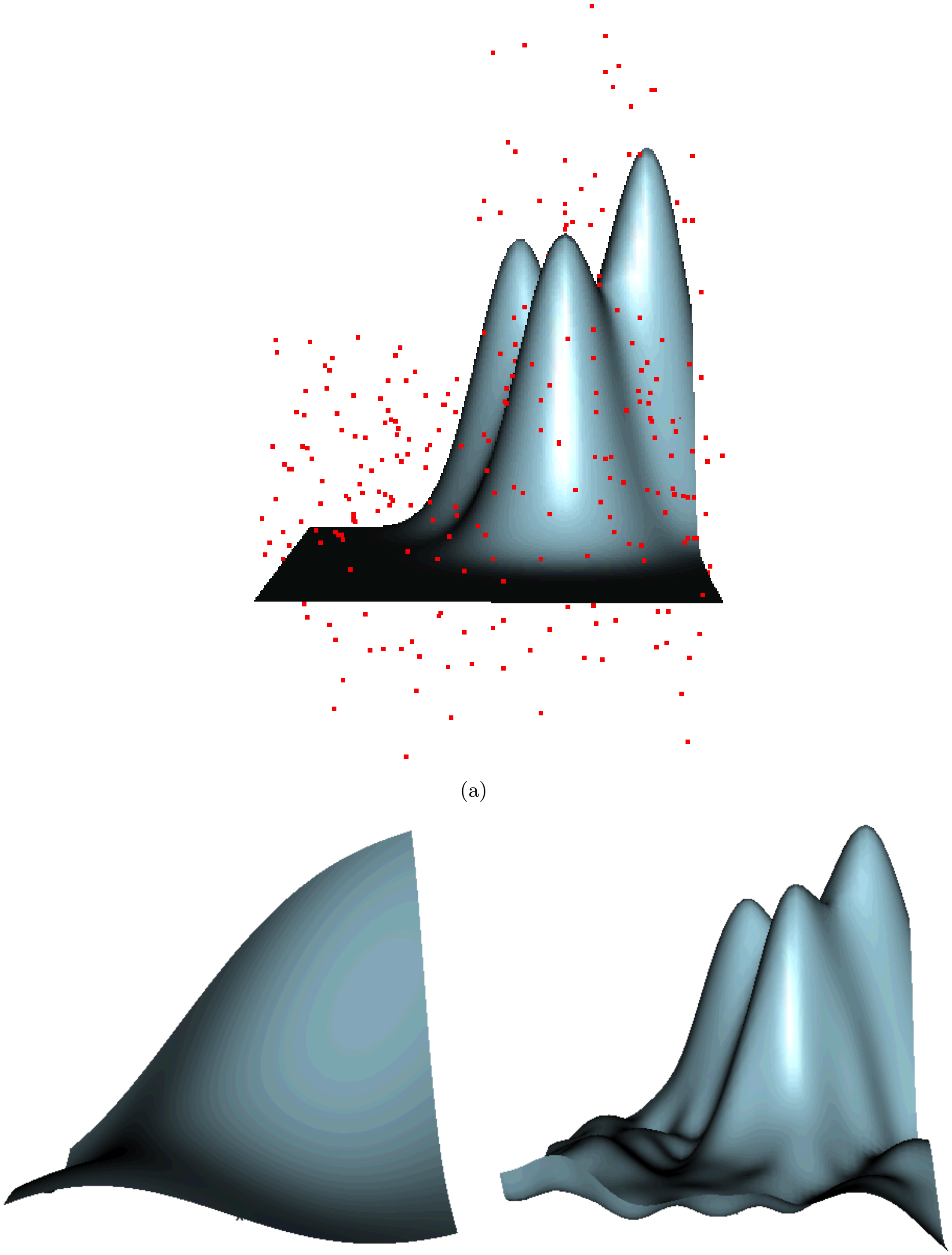
5.1 Introduction

Traditional approaches to smoothing tend to have a single parameter that defines the smoothness across the surface. This means that there is a tendency in some areas of the surface to be over smooth or over wiggly in order to accommodate an average smoothness measure. To illustrate a single smoothing parameter at work, Figure 5.1 shows an example using `gam` from the `mgcv` library [Wood, 2006] in `R`. The underlying function is flat in one part and very wiggly in another. A model of spatial coordinates alone with five degrees of freedom (Figure 5.1(b)) produces a surface that is overly smooth all over, whereas an increase in knots to accommodate the bumpy parts induces bumps in the flat part of the surface (Figure 5.1(c)).

A spatially adaptive approach allows more flexibility in some areas of a surface than others. Varying the smoothness in a one dimensional regression spline is akin to varying the smoothing parameter in a smoothing spline. Pintore et al. [2006] successfully applied spatially adaptive smoothing parameters to traditional one-dimensional test functions, but unfortunately their method was neither general nor well automated. SALSA is the most current method for one-dimensional smoothing and performs as well as, if not better than, competing frequentist methods [Crainiceanu et al., 2007, Ruppert, 2000, Baladandayuthapani and Carroll, 2005, Donoho and Johnstone, 1994, Pintore et al., 2006]. More specifically, the SALSA algorithm uses an adaptive knot-selection approach, with the number and location of the knots being determined in the solution process. Furthermore, it naturally accommodates local changes in smoothness across the covariate range.

Currently there is no version of spatially adaptive knot placement for two-dimensional problems with complex topography. However, there is a method that uses penalised regression splines in a mixed model framework (similar to that of Wang and Ranalli [2007]), where the fixed effects are spatial coordinates, random effects are TPS of spatial coordinates and the coefficients of the random effects are allowed to have spatially variable smoothing parameters [Krivobokova et al., 2008]. The authors also produced an `R` package called `AdaptFit` for fitting these models so here after, this method is referred to as ‘AdaptFit’. There does not seem to be an allowance in this method for topographically complex regions and the use of the TPS basis function suggests that the method is likely to succumb to similar leakage issues to a conventional TPS.

Based on its relative performance with other spatially adaptive one-dimensional methods, SALSA was considered worthy of developing into multiple dimensions for data with complex topography. Further, this model selection routine appeared to fit well with the use of two-dimensional local regression splines in the CReSS method (Chapter 3). SALSA has been adapted, from its one-dimensional form [Walker et al., 2010], in two ways: Firstly, we use a two-dimensional spline basis, and thus knots can move to a set of locations within a two-dimensional coordinate space; secondly, the basis structure used for model fitting is calculated using the Exponential function and geodesic distances (Equation 3.1), as is done in the CReSS method. SALSA is run for a choice of parameter r (Equation 3.1), and the subsequent models are averaged using BIC weights. BIC is used in keeping with authors recommendations from the one dimensional SALSA paper [Walker et al., 2010]. Hereafter, cases where one-dimensional splines are fitted [Walker et al., 2010] are referred to as SALSA1D and the use of two-dimensional splines, SALSA2D.



(a)

(b)

(c)

Figure 5.1: An illustration of fitting smoother-based methods with a single smoothness parameter. (a) the underlying function with noisy data overlaid (red points). (b) a GAM model fitted to the noisy data using five degrees of freedom and (c) a GAM model fitted with 75 degrees of freedom.

5.2 SALSA2D Method

The SALSA2D algorithm uses the same three steps that form the backbone of SALSA1D (Figure 5.2). These steps are *Exchange* (change a knot location), *Improve* (make local knot movements) and *Drop* (remove a knot). However, unlike the one-dimensional version there is no allowance for heteroscedastic errors so there is not a window of half width parameter (refer to Walker et al. [2010] for further details). There are a number of parameters to increase the users ability to control basic characteristics of the final model. These include

`gap` - the minimum allowable geodesic distance between two knots

`maxknots` - maximum allowable number of knots

`minknots` - minimum allowable number of knots

`startknots` - used to calculate the number of start knots

Other inputs required are the data (x_1, x_2, y), a regular grid of knots $(x_{1,t}, x_{2,t})$ and a geodesic distance matrix for distances between data and knots and pairs of knots.

```
SALSA2D:
    Initialise
        Initialise knots $T$ with legal knot locations
    Repeat
        Repeat Exchange step while (fit measure improves)
        Repeat Improvement step while (fit measure improves)
        If ($|T|$ >minknots)
            Perform Drop step
        End If
    While (an improvement in fit measure is made by one of the above steps)
```

Figure 5.2: Pseudocode outlining the structure of SALSA2D (adapted from Figure 1 Walker et al. [2010]), where T is the number of knots used for fitting.

The initialisation process selects legal knot positions at random until the specified number of starting knots is reached or there are no legal positions left. ‘Legal’ knots must be contained within the region of interest, with a minimum gap between knots greater than specified by the `gap` parameter. The initial model is fitted using these knots and a fit criterion (e.g. AIC, BIC, QAIC) is calculated. After initialisation the CReSS-SALSA2D algorithm is made up of four main steps; exchange, improve, drop and model averaging.

- **Exchange** The *Exchange* step allows the solution to move away from a local optimum. Knots are allowed to move to a new position, as near to the maximum residual as possible (but still on the knot grid), or an additional knot is included at this grid position. During this process there must be a legal position (`gap` observed and exclusion zones respected) for a knot to move to or be added. For each move, or addition, the fit statistic is calculated and compared with the step before to see if an improvement can be made. After all possible exchanges have been made, the algorithm moves to the *Improve* step.

- **Improve** The *Improve* step makes local improvements by moving knots around their current position. The algorithm considers relocating each knot, in turn, to each of its eight possible neighbours on the grid, provided the move is legal. For instance, a move to the left may place the knot in an island or within `gap` of another knot, rendering the position illegal. After all knots have been through the *Improve* step, the algorithm moves to the *Drop* step.

- **Drop** The *Drop* step allows for simplification by removing knots as long as the number of knots is greater than `minknots`. Each knot is cycled through in turn to determine if the fit score can be improved by dropping it. As soon as a knot is dropped the algorithm returns to the *Exchange* and *Improve* steps.

Termination of the SALSA2D algorithm occurs if there are no improvements in fit statistic in any of the three steps above. Furthermore, the SALSA2D algorithm terminates if no knots are dropped in the *Drop* step or `minknots` is reached.

- **Model Averaging** SALSA2D arrives at a single model for each choice of parameter r and `gap`. These models are then averaged using fit statistic weights (calculated using Equation 3.2) to obtain a single set of predictions. Where before we used AIC$_c$ to calculate weights, the same equation may be used for BIC or QAIC weights.

5.3 Simulation

The palm simulation (Chapter 4) was used to compare the performance of CReSS-SALSA2D to other methods, in particular, CReSS. Data were randomly chosen from the palm surface, $n = 500$, and a normal errors noise term with standard deviation 0.5 (low), 9 (medium) and 50 (high) were added to the function values. Predictions were obtained on $N = 2518$ points.

For efficiency, possible choices for parameter r were restricted to between 2 and 5, based upon previous simulation results (Section 4.2.2). Furthermore, in line with earlier simulations, the minimum number of knots was set to 10 and the maximum to 100. A grid of 693 possible knot locations was included, containing 543 positions within the region of interest. Other parameters chosen were `startknots` = 24 and `gap` = 0.2, 0.6, 1.14. The grid of knots had a spacing of approximately 0.4 units so a gap of less than 0.4 means a legal knot position can be next to another knot on the grid. Assuming that Euclidean distance equals geodesic distance (meaning there are no exclusion zones nearby), a gap of 0.6 means the eight locations surrounding a knot (top, bottom, left, right and diagonals) are illegal. CReSS-SALSA2D was used to establish knot number and location for each combination of r and `gap`. Thus, with a choice of seven different r and three different gaps, the model averaging step contained 21 models.

5.4 Results

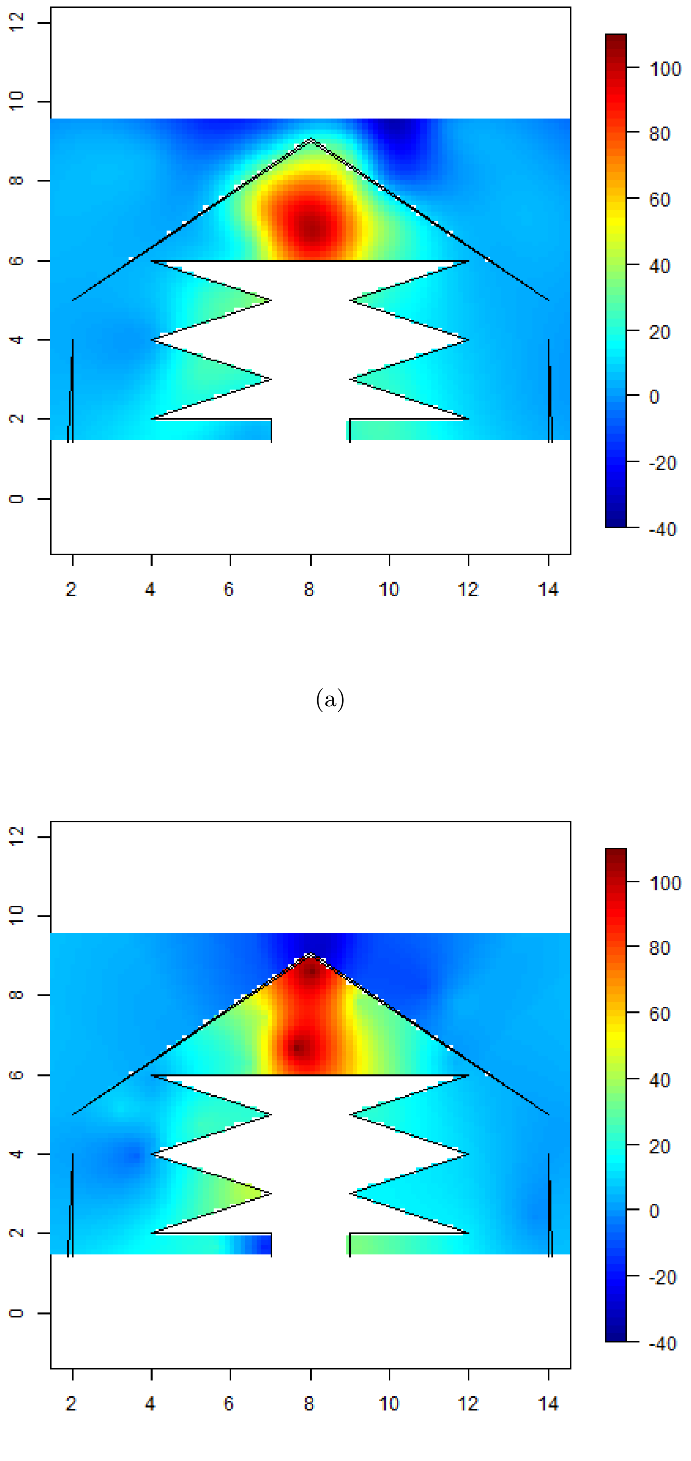
This section mainly shows the results of the simulation study but begins with a quick comparison of the adaptive penalised splines using the `AdaptFit` package with CReSS-SALSA2D.

5.4.1 AdaptFit Comparison

One simulated data set at medium noise was used for this comparison, the results of which made it unnecessary to continue a full set of simulations. To fit the adaptive penalised splines, the maximum number of iterations had to be increased to 100 for the mean function (default=20) and 1,000 for the variance of the random effects (default=50) to achieve convergence. The numbers of knots were left as the default settings (in the absence of any guidance to the contrary), which were 50 knots for the regression function and 12 for the penalty function. The results were very similar to those for the TPS method in Chapter 4 (section 4.2.2) but with a little less leakage across the outer breakwater (Figure 5.3), presumably due to the spatially adaptive nature of the knots. Visually the results for CReSS-SALSA2D are far superior and this was seen numerically too (Figure 5.3 and Table 5.1). The MSE score for the predictions was about seven times that of CReSS-SALSA2D and almost the same as for TPS. The lack of information about boundaries in this method made it very poor on this palm region and so a full simulation was not carried out. The remainder of this section compares CReSS-SALSA2D results with the CReSS simulation from Chapter 4 and discusses results of the two methods in more detail.

Table 5.1: MSE scores for three different methods from one simulation realisation (n=500) at medium noise.

TPS	**AdaptFit**	**CReSS-SALSA2D**
103	101	14



(a)

(b)

Figure 5.3: Predictions for (a) AdaptFit and (b) CReSS-SALSA2D from one simulation realisation (n=500) at medium noise.

5.4.2 Simulation Results

Numerically, the results for low and medium noise CReSS-SALSA2D showed a significantly improved fit ($p < 0.05$; Wilcoxon signed rank test) compared with CReSS (Table 5.2). However, this was only the case for low noise when one simulation realisation, which gave a very bad result, was removed. For this particular realisation the MSE score for the data was 1.10, whereas the MSE score for the predictions was 1324, a thousand times greater. This implies the model fits extremely well to the data but performs poorly when extrapolating (i.e. the model over-fits). For this reason, Table 5.2 shows the results with and without this realisation (represented by Low and Low* respectively) and Figure 5.4a is shown without this result. Figure 5.4 shows boxplots of the differences in MSE scores between the two methods and bias plots for CReSS-SALSA2D. The greatest bias was under the outer breakwater and in the very tips of the fronds on the left hand side. However, the bias in the fronds is much less than that seen for CReSS alone (Figure 5.4 and for example, Figure 4.4).

In order to assess why there was one bad data realisation, the individual result was plotted in Figure 5.5. The prediction problem appeared in the bottom left of the central palm, where the predictions change rapidly from 0 to -910 units. Two knots were placed in this region, but it is an area without any observed data. The rest of the surface shows a very good fit to the underlying function.

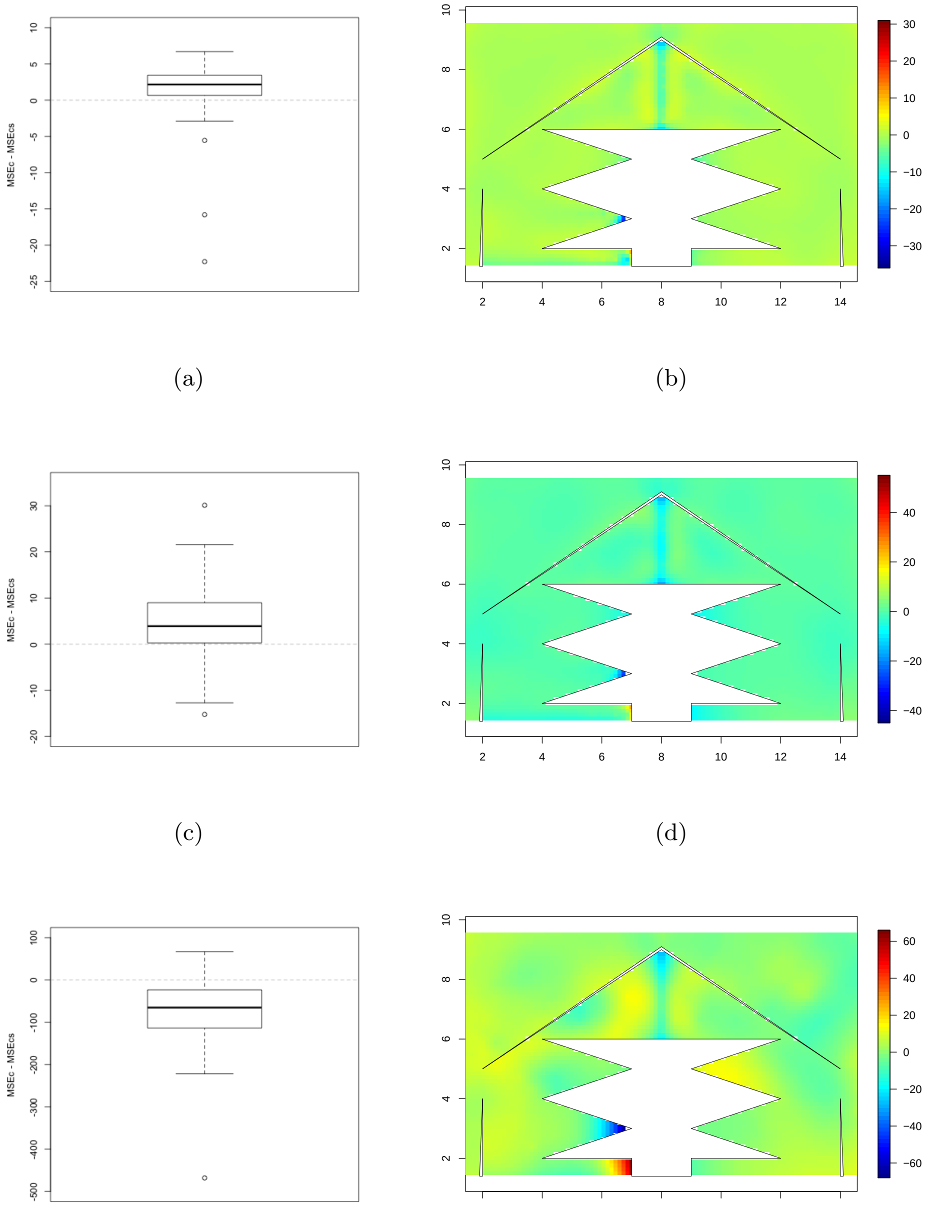
CReSS-SALSA2D does not perform as well as CReSS on data with high noise (Table 5.2 and Figures 5.4(e) and (f)). The mean MSE score was much greater than for CReSS alone. Almost every simulation realisation gave a negative difference in the pairwise comparisons (Figure 5.4(e)), which implies CReSS-SALSA2D usually gives a worse fit. At all noise levels, the variance of the mean MSE scores was larger when SALSA was used.

Figure 5.6 shows where CReSS-SALSA2D moved the knots in comparison with space-filled knots for one single simulation realisation. At low noise, one model was chosen by

Table 5.2: Mean and Median MSE scores and standard deviation for SALSA2D at all noise levels for the palm simulation. Low* represents results with one problematic realisation removed and is therefore for 99 simulations. Results of a wilcoxon signed rank test are represented by * and †, where * indicates CReSS-SALSA2D was significantly better than CReSS and †, CReSS was significantly better.

	CReSS-SALSA2D		CReSS	
Noise Level	μ	σ	μ	σ
Low	19.08	131.84	7.20	2.70
Low*	5.91*	4.33	-	-
Medium	16.79*	7.44	21.97	5.49
High	259.67†	82.92	167.26	38.53

CReSS-SALSA2D and used far fewer knots than space-filled CReSS (27 vs. 90). Furthermore, the knots were mainly located under the hat shaped island and the convex ends of the palm fronds (Figure 5.6(a)). The MSE score improved from 9.14 (CReSS) to 3.25 (CReSS-SALSA2D). At medium noise CReSS model averaged two models of 20 knots (r = 2, 3). CReSS-SALSA2D averaged 5 models with between 12 and 17 knots (r = 2-4). Figure 5.6(b) represents the strongest weighted model of the five (weight = 0.92) with 15 knots. The MSE score improves from 54.61 (CReSS) to 24.51 (CReSS-SALSA2D). There is no plot for high noise since both CReSS and CReSS-SALSA2D averaged more models than at low/medium noise, making graphical representation of knot locations difficult. In general, fewer knots were chosen per model when noise was high, but more models were averaged (Table 5.3).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.4: (a, c, d) Boxplots of differences in MSE score between CReSS and CReSS-SALSA2D. A positive number represents a better model fit for CReSS-SALSA2D. (b, d, f) on the right are bias plots for CReSS-SALSA2D. (a and b) Low noise* ($\sigma = 0.5$), (c, d) Medium noise ($\sigma = 9$) and (e, f) High noise ($\sigma = 50$). * The low noise figure represents only 99 simulations. One extreme case was removed for better comparison.

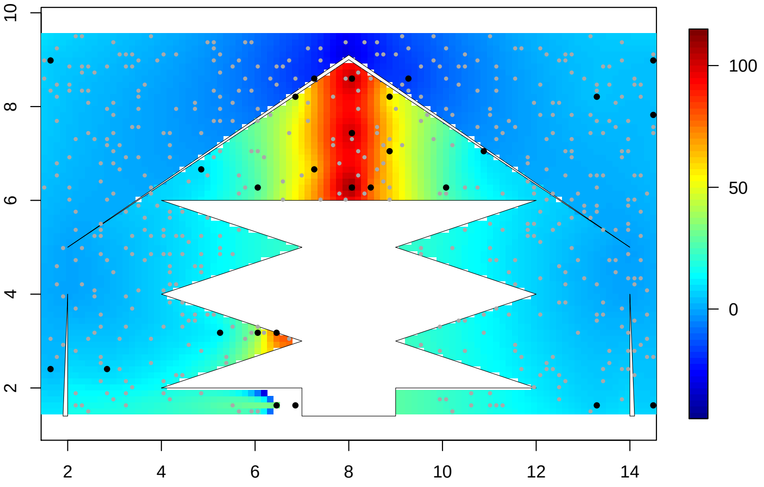
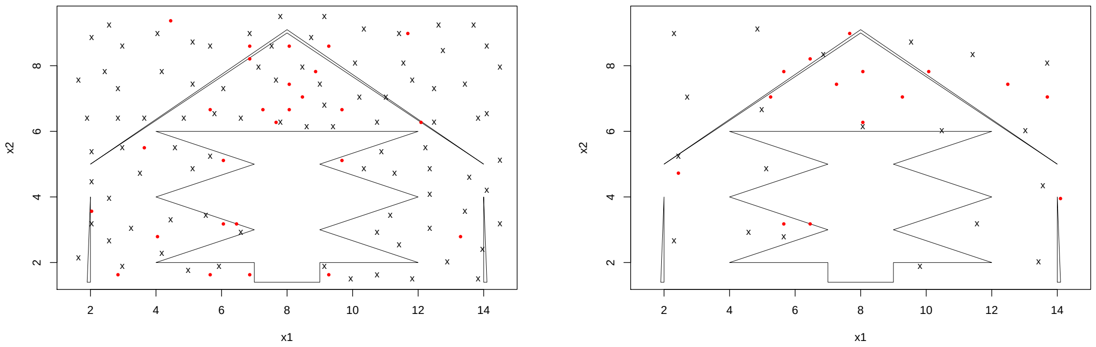


Figure 5.5: The low noise simulation realisation that results in a very bad prediction MSE score. The surface represents the predicted values based upon the model chosen. The grey dots are the data points, the black dots are the knots chosen by CReSS-SALSA2D.



(a)

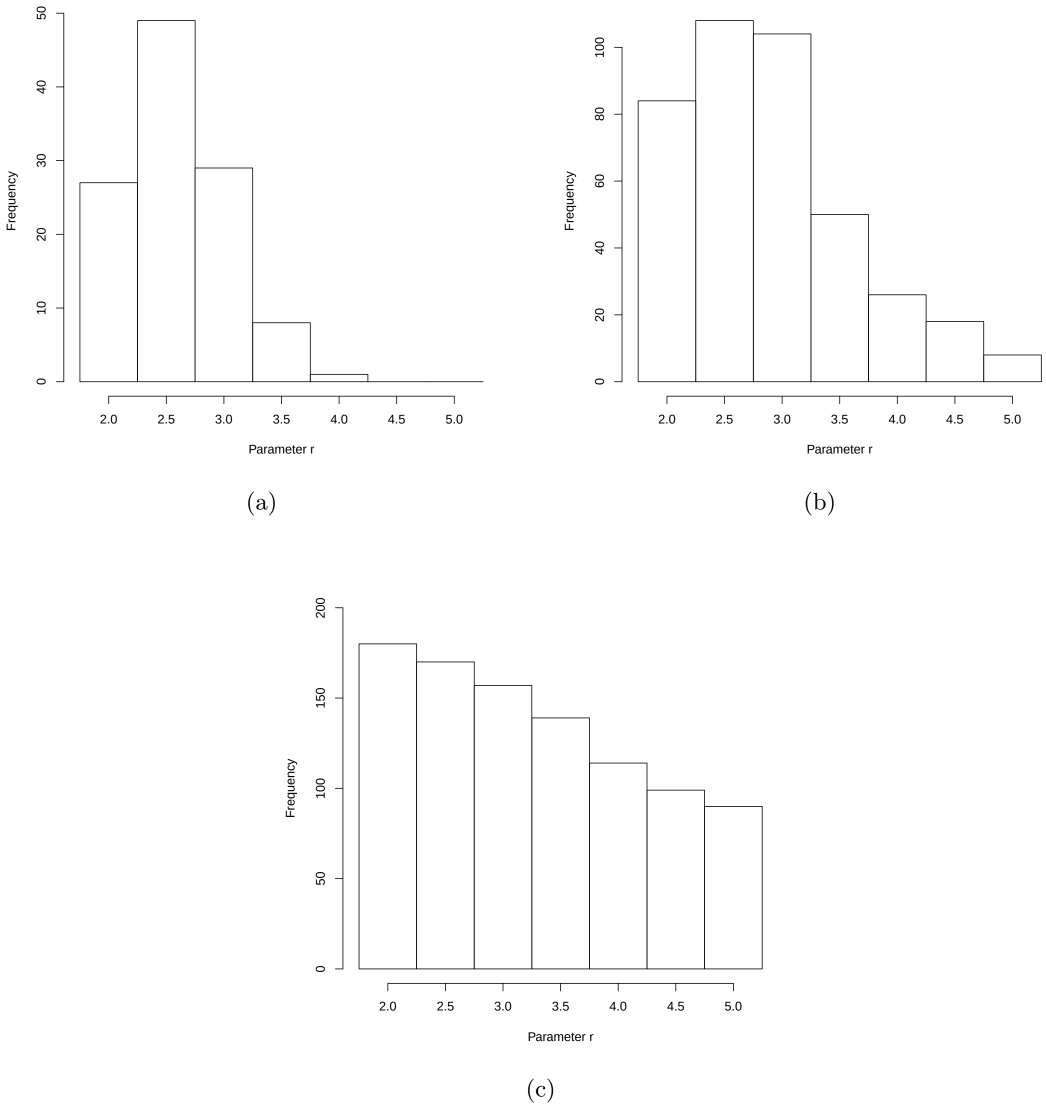
Figure 5.6: A single simulation realisation for low noise (left) and medium noise (right) depicting knot number and location. Black crosses are space-filled knots used in CReSS and red circles are spatially adaptive knot locations chosen using CReSS-SALSA2D.

As noise level increased, the number of models averaged by CReSS-SALSA2D and the value of parameter r increased, while the number of knots decreased (Table 5.3). Figure 5.7 shows the distribution of parameter r for low, medium and high noise. Many more different values of r were chosen for high noise and there was a definite shift in distribution to higher values. However, a number of models using small r were still chosen. There was no striking relationship between the noise level and the number of knots chosen.

We also looked at the behaviour of the MSE score as the number of models averaged increased. Within high noise between 1 and 21 models, of a possible 21 (3 gap choices and 7 possible values for r), were averaged. Figure 5.8 shows the effect this had on MSE score. Generally, as the number of models averaged increased, the MSE score decreased. Therefore, there may be some advantage in increasing the number of models averaged at high noise.

Table 5.3: Parameter choices made by CReSS-SALSA2D for each of the three noise levels averaged over 100 simulation realisations. The parameters include the mean number of models averaged per realisation, the mean for parameter r and the mean number of knots. Numbers in brackets show the minimum and maximum.

	Low	Medium	High
Mean Number of Models Averaged	1.14	3.98	9.49
	(1, 3)	(1, 13)	(1, 21)
Mean r	2.59	2.89	3.26
	(2, 4)	(2, 5)	(2, 5)
Mean Number of Knots	27.18	18.55	11.80
	(22, 38)	(11, 33)	(10, 21)
Mean gap	0.22	0.48	0.66
	(0.2, 0.6)	(0.2, 1.14)	(0.2, 1.14)



(a)

(b)

(c)

Figure 5.7: Distribution of parameter r chosen using 100 simulation realisations for (a) low, (b) medium and (c) high noise levels. The allowed choice of r ranged from 2 to 5. If each r was averaged for all knot numbers in every realisation, the frequency would be 300.

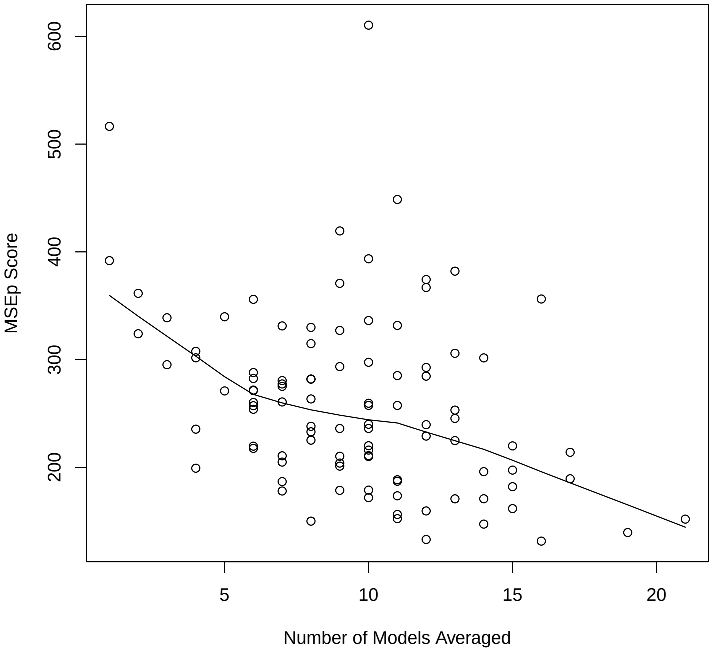


Figure 5.8: Variation in the MSE score with the number of models averaged. The line represents a locally weighted polynomial regression smooth of the data. The total number of models that could be averaged is 21.

5.5 Discussion

Whilst there is an alternative two-dimensional adaptive smoothing method, AdaptFit, there is not one that can deal with topographically complex areas. CReSS-SALSA2D was found to be far superior to AdaptFit on this simulation and, overall, CReSS-SALSA2D showed a marked improvement over CReSS for low and medium noise simulations. Knots were moved to places where we would expect more flexibility to be needed. For example, more knots were placed under the hat-shaped island, where radial bases have to approximate the straight lines of the hotspot, and in the ends of the fronds on the left side where the surface changes rapidly. Fewer knots were placed in the flatter corner areas of the surface. In general CReSS-SALSA2D used fewer knots than CReSS alone.

The reason the problematic low noise realisation was chosen by the algorithm was due to a very good BIC score. The fit to the data was exceptional but the predictions were poor. The BIC score only 'sees' the data and so is based upon the residual sums of squares fit to the data observations. This led to the model over-fitting the data. CReSS-SALSA2D placed two knots in a location outside the range of the data, in two-dimensional geodesic space. Perhaps this was an indication that knots should only be allowed where there is data support. A simple fix would be for the user to check the output and either remove the offending knot(s) and re-fit, or to use a new pseudo-data point to tie down the bases at the edge. The new data point could be a repeat of the last data point, but shifted to the boundary edge. An alternative is to make knot points outside the data range illegal, so that no knot can be placed there, or to have a requirement for data between knots. Further analysis is required to assess why knots are placed in 'bad' positions since it could be an indication that BIC is not the right measure of fit. Since writing this chapter, the SALSA method has been applied to modelling the distribution of Tern species around the UK [Mackenzie and Scott-Hayward, 2012, Mackenzie et al., 2012]. The data collected were in transects (unlike the data cloud presented here) and so the issue of knot placement away from the data became quite apparent. The solution was to make the knot grid fine enough for knots to be represented on all of the transects and then make locations between transects unavailable to the selection process. Additionally, model selection was achieved using five-fold Cross Validation (CV), which is better at choosing models that fit well to data unseen by the model. These changes appeared to reduce the issue of a knot location far away from the data locations falsely raising or lowering the surface in that area.

The results obtained with CReSS-SALSA2D were not good when data with high noise were used. This could be due to over-fitting - the model fits too closely to the noisy data and therefore fails to find a good approximation to the underlying function for the surface. The *Exchange* step in the SALSA algorithm moves a knot to the highest residual; this allows flexibility in areas where the fit is poor but may have a tendency to fit the model too closely to very noisy data points. The raw residuals were used in this simulation, but this is not ideal and the use of standardised residuals would be preferable. These are residuals that have been adjusted for the variance assumed under the model. It is a lengthy process to repeat the simulation using standardised residuals. Running the first 5 simulation realisations for high noise, showed an improvement in a few MSE points over using raw residuals, but not close to the results obtained using CReSS only. This suggests that the use of inappropriate residuals is not the primary cause of the over-fitting at high noise levels. The simulations here used Gaussian errors, and the choice of an appropriate type of residual may be more important when the error distribution is not Gaussian. In the next chapter the use of a quasi-Poisson error distribution is investigated.

There may be other reasons for the problems with CReSS-SALSA2D at high noise. For example, the results in Chapter 4 showed the need for larger r, particularly at high noise, and this could be why CReSS-SALSA2D performed worse at high noise than CReSS. The choice made the simulation more efficient but perhaps too restrictive (i.e. a greater range for r could be allowed). Other factors that may reduce the performance at high noise are that BIC may not be a suitable measure of model fit, and that there may not be enough models to average. We could try another information criterion such as AIC$_c$ or CV, and by averaging over many more models by keeping a track of all the models fitted as part of the knot placement algorithm. This would provide multiple models for each r and gap combination, rather than the 'best' output currently used. As the models are already fitted in the current version of CReSS-SALSA2D, such an approach would require no additional computational effort. Allowing a greater range for r would also increase the number of models available for averaging.

5.5.1 Future of SALSA2D

The SALSA algorithms are a work in progress. To date we have tried to improve fit and speed of the two-dimensional algorithm. The next major development will be to improve the locally adaptive nature of the algorithm. Currently each knot has the same parameter r as every other knot for a single model. A new version of CReSS-SALSA2D could allow r to vary, which would allow each individual knot to act globally (large r) or locally (small r). The plan is to initialise with all knots having r equal to the middle of a predetermined range. Models would then be re-fitted with bigger or smaller values of r for each initialised knot in turn. The model with the lowest BIC score would then be retained. Once each knot has an appropriate r, the *Exchange*, *Improve* and *Drop* steps would be executed. If a knot is moved (*Exchange* step) then r would be re-calculated for the new location. However, since the movements are local in the *Improve* step, the same r may be used for the new knot. If a knot is dropped, r would be re-calculated for all knots to prevent possible gaps in the basis function coverage. At present, the range of parameter r is determined manually, but developments to the CReSS method (Chapter 3) to automate this, could be added to CReSS-SALSA2D.

To speed up the process, SALSA could be initialised at knot locations for a known model fit. For example, the algorithm could be initialised using space-filled knot locations, rather than random ones. This might reduce the number of global moves the algorithm iterates through and thus speed up the process.

A second development would be to combine the SALSA1D and SALSA2D algorithms to allow additional covariates with one- or two-dimensional smooths. This development is at an early stage but it is intended to pave the way for production of an `R` package to run models using SALSA1D and CReSS-SALSA2D and include model selection.

It is also worth noting that there are several other versions of SALSA in development, which include using mixed models, variable radii and extended model averaging.

In order to fully demonstrate their value as a tool for biologists, CReSS-SALSA2D and CReSS need to be applied to real biological datasets. In the next chapter we use both these methods in a more advanced model framework for analysing a large topographically complex data set of cetacean abundance in north-western European waters.

Chapter 6

Case Study: Spatial Analysis of the Joint Cetacean Protocol Data for Harbour Porpoise and Minke Whale in North-western European Waters.

6.1 Introduction

Article II of Council Directive 92/43/EEC on the Conservation of Natural Habitats and of Wild Fauna and Flora (henceforth referred to as the EU Habitats Directive) requires the EU Member states to report on the conservation status of, among others, all cetacean species occurring in their waters every 6 years. This report must contain information on trends in species' range and abundance over the preceding period. The Joint Cetacean

Protocol (JCP) data resource, a collection of survey data from 1969-2010 covering north-western European waters, is one of the databases used by the UK government to provide this information. The data in the JCP have been gathered by various governmental organisations, private sector companies and non-governmental organisations using a variety of surveying techniques. It contains information on sightings of all cetacean species made in this area, but this chapter focuses on the most commonly seen small cetacean, the harbour porpoise (*Phocoena phocoena*) and the most commonly seen baleen whale, the minke whale (*Baelenoptera acutorostrata*). The aim of the analyses reported here is to assess changes in their distribution and abundance in north-western European waters between 1985 and 2010 using spatial density maps.

Harbour porpoise, sometimes known as the common porpoise, are the smallest and most numerous cetacean found in the region of study. Females grow to about 160cm in length and are generally larger than males which grow to about 145cm [Reid et al., 2003]. Typically they occur in small groups of one to three animals and their surfacings are generally inconspicuous. This, combined with their small size, makes detection of this species particularly low in choppy sea states. Palka [1996] found that the detection of harbour porpoise decreased by up to 75% in a Beaufort sea state of 3 compared with Beaufort sea state of 0. They are found mainly in inshore waters [Embling et al., 2010, Marubini et al., 2009] and are reported to have a preference for water depths between 50-100m [MacLeod et al., 2007a, Marubini et al., 2009, Booth, 2010].

According to a report from the Agreement on the Conservation of Small Cetaceans of the Baltic and North Sea [ASCOBANS; Reijnders et al., 2009] entanglement in fishing gear (particularly gillnets) is the greatest threat to harbour porpoises in European waters, and Vinther and Larsen [2004] has suggested that bycatch in Danish waters exceeds a sustainable level. In addition, harbour porpoises appear to be particularly sensitive to acoustic disturbance emanating from shipping noise [Palka and Hammond, 2001], naval exercises [Parsons et al., 2000], marine renewable installations [Teilmann and Cartensen, 2012] and acoustic deterrent devices [Booth, 2010].

In contrast to harbour porpoise, minke whale (family *Balaenopteridae*) grow to a length of 7-8.5m. In the northeast Atlantic, they range from the Barents Sea to Portugal and into the Mediterranean during the summer [Reid et al., 2003]. Their winter range is not known, but is thought to include waters from the southern North Sea to the Straits of Gibraltar [Rice, 1998]. They are often seen singly or in pairs but occasionally, when feeding, they may form groups of 10-15 individuals and associate with other cetaceans such as harbour porpoises [Reid et al., 2003]. The conservation status of the minke whale in the northern hemisphere is listed as of least concern by IUCN. However, like all other cetaceans, it is likely to be affected by chemical pollution and acoustic disturbance [Parsons et al., 1999], not to mention the effects of commercial exploitation in Norwegian and Icelandic waters. It is thought that the main determinant of minke whale distribution is prey distribution [Anderwald et al., 2012, Macleod et al., 2004]. However, depth [Skov et al., 1995, Hooker et al., 1999], sediment type [Naud et al., 2003, Macleod et al., 2004], the location of oceanographic fronts [Kasamatsu et al., 2000, Bjørge, 2001], sea-surface temperature [Anderwald et al., 2012, Kasamatsu et al., 2000, Hamazaki, 2002] and the extent of sea ice [Kasamatsu et al., 2000] have all been shown to be related to distribution. Furthermore, the relationship between distribution and some covariates varies seasonally. Macleod et al. [2004] demonstrated seasonal patterns for prey preference and sediment type and Anderwald et al. [2012] for sea-surface temperature, chlorophyll and prey distribution.

6.1.1 Large scale assessments of cetacean distribution in north-western European Waters

There have been two large scale studies on cetaceans conducted in north-western European waters: the Small Cetacean Abundance in the North Sea (SCANS) surveys I [Hammond et al., 2002] and II [Hammond et al., 2013]. There is also a large database of opportunistic cetacean sightings collected as part of the European Seabirds at Sea project [ESAS

Northridge et al., 1995a, Erratum: Northridge et al., 1995b]. Data from both SCANS-I and ESAS were summarised in the atlas of cetacean distribution in north west European waters compiled by Reid et al. [2003]. SCANS-I and -II and ESAS have all been included as part of the JCP data resource.

The ESAS data were collected mainly from platforms of opportunity (e.g. research vessels, ferries, seismic vessels and oil rig supply vessels) to map the offshore distribution of sea birds in European waters. At the same time, data were collected on cetacean sightings and Northridge et al. [1995a, Erratum: Northridge et al., 1995b] analysed these data to map the distribution and relative abundance of harbour porpoise, minke whale and white-beaked dolphin (*Lagenorhynchus albirostris*). At the time, it was the largest and most comprehensive effort-related sightings database for the north-eastern Atlantic and covered the period from 1979-1990. Sightings were adjusted for detectability, based on seven different sea states, and overlaid on maps of effort data using a 1^o grid. The maps indicate that the main concentrations of porpoise sightings were in the north and central North Sea, west coast of Scotland, southern Irish Sea and south of the coast of Ireland. There was also a seasonal trend, with higher sightings rates in the summer months, peaking in August. The main concentrations of minke whale sightings were off the Hebrides and the north-east coast of England. Sightings rates of minke whale were much lower than for harbour porpoise and peaked in June.

SCANS-I [Hammond et al., 2002] and SCANS-II [Hammond et al., 2013] took place in 1994 and 2005 respectively and were the most comprehensive design-based cetacean surveys to cover north-western European waters. Sightings of all cetacean species, including harbour porpoise and minke whale, from boat and aerial surveys were analysed using Distance sampling for line transect methods [Buckland et al., 2001]. Estimated counts (raw counts inflated for detectability) were analysed using a Generalised Additive Model (GAM) with quasi-Poisson errors, because there was evidence of over-dispersion [Wood, 2006, Hammond et al., 2013].

SCANS-I covered the North Sea, parts of the Baltic Sea, the Channel and the Celtic Sea (Figure 6.1). Hammond et al. [2002] reported that the highest densities of harbour porpoise were found in the central North Sea, whilst the highest densities of minke whale were recorded in the north-western North Sea and in the Celtic Sea (Figure 6.1). SCANS-II extended the SCANS-I survey area to included offshore waters to the west of Scotland and Ireland, the northern Bay of Biscay and the Iberian shelf (http://biology.st-andrews.ac.uk/scans2). Table 6.1 shows the final models for harbour porpoise and minke whale densities from both surveys.

Table 6.1: Models from the SCANS-I (year 1994) and SCANS-II (year 2005) surveys relating harbour porpoise and minke whale density to environmental covariates [Hammond et al., 2013]. Models were fitted using `gam` from the `mgcv` package in `R` [Wood, 2006]. All spatial smooths were restricted to a maximum of 14 degrees of freedom and other covariate smooths to a maximum of 5.

Model	**Term**	**Estimated *df***
Harbour porpoise 2005	s(latitude, longitude)	12.8
	depth	1
	s(distance to coast)	3.3
Minke whale 2005	s(latitude, longitude)	12.9
	s(depth)	4
	s(distance to coast)	3.5
Harbour porpoise 1994	s(latitude, longitude)	12.1
	depth	1
Minke whale 1994	s(latitude, longitude)	12.9
	slope	1

The density surface maps produced for each of the two species and surveys suggest that there had been a marked change in harbour porpoise distribution between 1994 and 2005 from the central North Sea toward the southwest North Sea (Figure 6.1). Over the same period, the main areas of minke whale abundance shifted from the east of Scotland toward the central North Sea, with high densities also being found off the south coast of Ireland

(Figure 6.1).

In 2003, Reid et al. [2003] published an *Atlas of cetacean distribution in north-west European waters* which was based on an analysis of data from ESAS, SCANS-I and the Sea Watch Foundation. The Sea Watch Foundation data spanned the early 1960s to late 1980s and were based on opportunistic sightings collected from land and offshore. Data for 25 species, collected from 1979-1997, were analysed for the atlas to varying degrees of accuracy depending, for the most part, on data sufficiency. The maps for minke whale and harbour porpoise depict the distribution, relative abundance and associated survey effort for each species. However, effort is presented as a background to the maps (similar to the ESAS analysis in 1995), which may allow the casual reader to ascribe undue confidence to areas of apparently high relative density where there is little effort. Thus, as the authors mention, the maps can only provide 'general statements about relative animal densities at a regional level' [Reid et al., 2003]. The areas of highest density for harbour porpoise were to the east of Denmark and in the north-western North Sea, with lower densities off south-west Ireland, south-west Wales and the west coast of Scotland. Harbour porpoise also appeared to show a preference for depths shallower than 100m. The highest relative abundance of minke whales was recorded in the western North Sea, west coast of Scotland and a small area in the central North Sea. There was also an area of high relative abundance off the south coast of Ireland. There was some evidence that minke whales preferred waters less than 200m deep.

This chapter describes a unified analysis of all of the sightings data from harbour porpoises and minke whales contained in the JCP data resource. The analytical tools developed in previous chapters are used to map changes in the spatial distribution of the two species over time, and to identify areas of particularly high and consistent abundance.

Thanks to the Joint Nature Conservation Committee (JNCC) for allowing the use of the JCP data resource in this chapter (T. Dunn *pers. comm.*).

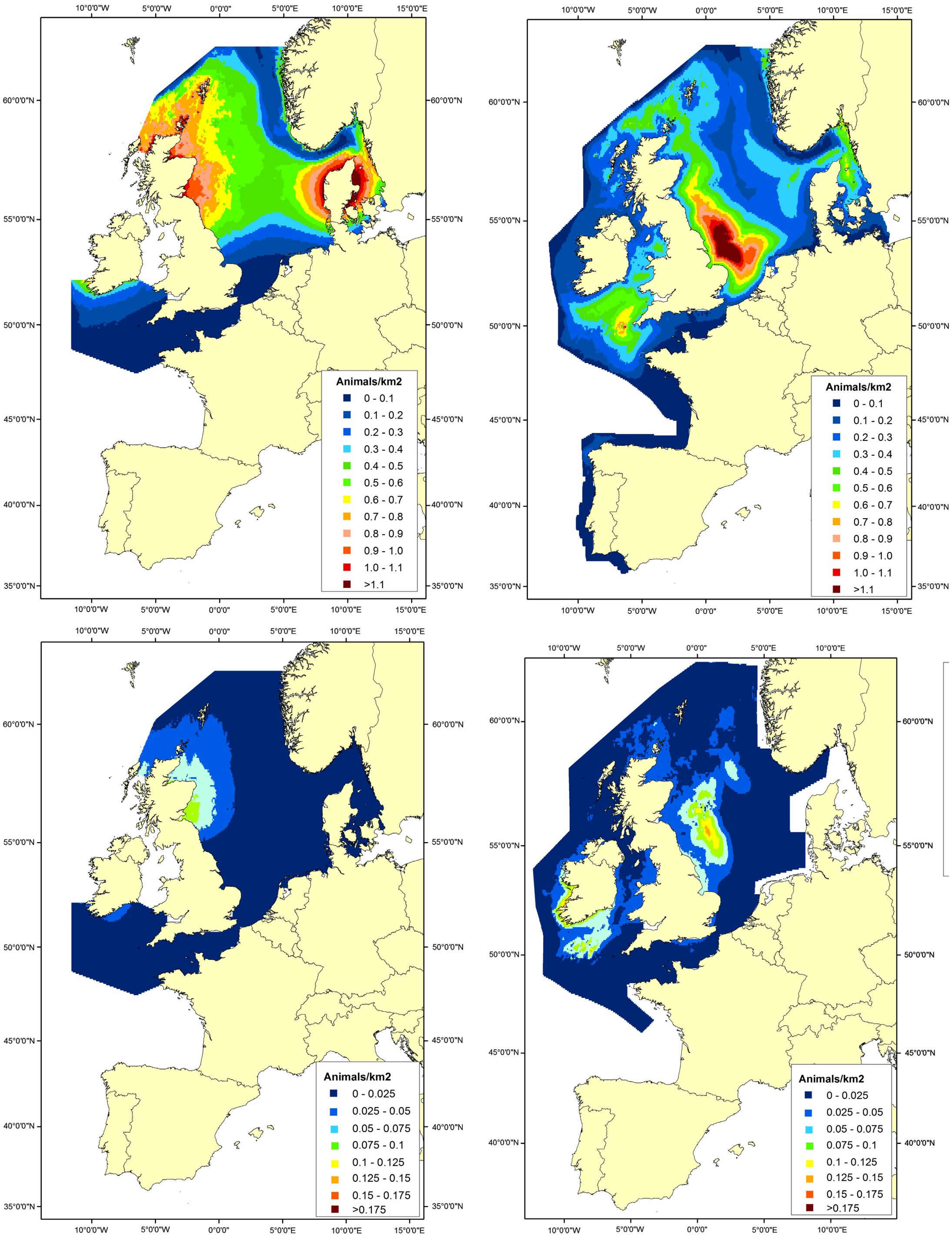


Figure 6.1: SCANS-I (left) and II (right) results for harbour porpoise (top) and minke whale (bottom). Figures taken from Hammond et al. [2013].

6.2 JCP Data Resource

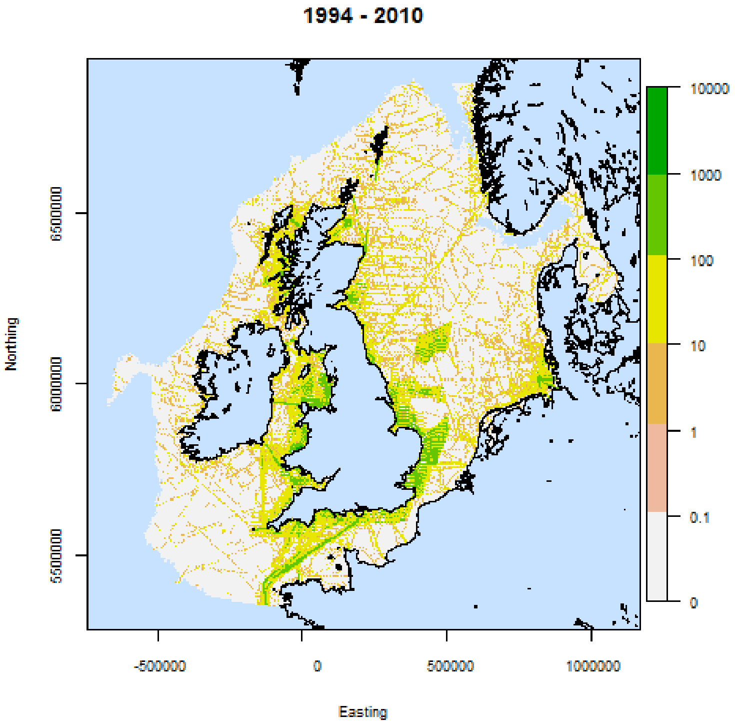
Collation and correction of the data in the JCP data resource involved combining data sets collected from a variety of sources and adjusting the data to account for under-detection (the fact that observers do not see all of the available animals) and availability (the fact that not all animals are at the surface and available for observation). Distance sampling [Buckland et al., 2001] uses the distribution of observed animals to estimate their detectability by creating a detection function, which describes the way in which the probability of detection varies with distance from the trackline for animals at the surface. This function is then used to adjust the raw counts for under-detection. One of the key assumptions of Distance sampling is that all animals on the trackline are available and detected. This assumption is often violated for cetaceans because they may be difficult to detect, even when they are on the trackline, and they may be submerged for long periods. When suitable data (such as observations from two or more independent observers or from telemetry) had been collected, it was possible to estimate the probability of detecting an animal on the trackline. Availability bias was corrected for by using information about diving times. Despite the bias corrections employed, not all biases may have been accounted for and so the estimated abundances from this process are referred to as relative abundances.

Each transect of survey effort was divided into approximately 10 km long segments, and the number of animals detected along this segment, corrected for detectability and availability, was summed to create spatially referenced count data used as input for modelling. The collated data covered the period from 1994 to 2010 and consisted of 88734 segments for harbour porpoise and 131448 segments for minke whale, covering an area of approximately 1.09 million km^2 (Figure 6.2). There are fewer segments for harbour porpoise because data collected when the sea state exceeded Beaufort 2 were not included due to the low detectability of this species in poor sea conditions. Data collected prior to 1994 were not considered for any species due to the small number of sightings and poor spatial coverage.

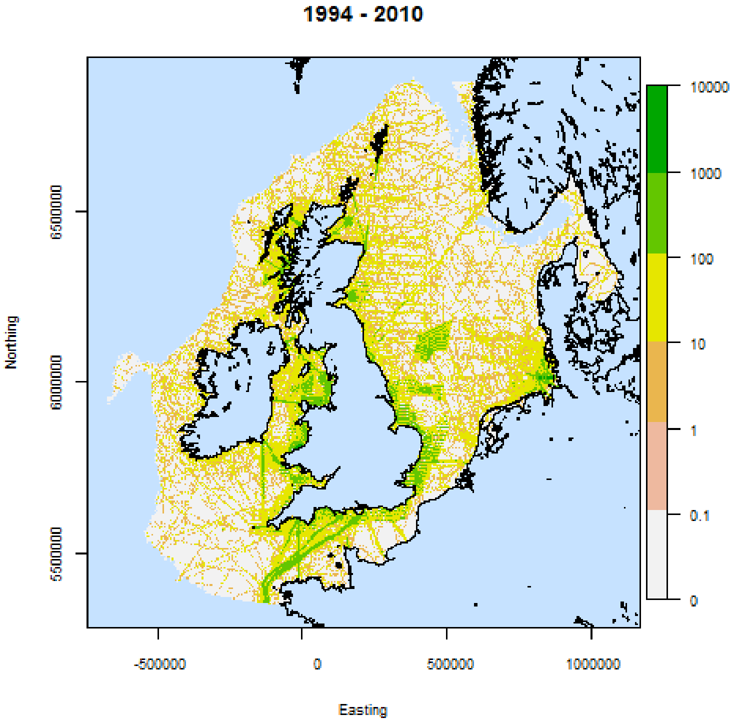
Further details of data collation methods and the corrections employed can be found in Appendix 2 of Paxton et al. [2013].

6.2.1 Explanatory Variables

The environmental covariates used in the analyses were *Depth*, *Slope* and sea surface temperature (SST). *Depth* was either recorded at the time or taken from the ETOPO2: 2 minute resolution relief data available from the National Oceanographic and Atmospheric Administration (NOAA). *Slope* was estimated as a function of the north-south and east-west depth gradients. SST was at 1 degree resolution, weekly averages and was also obtained from NOAA. Each environmental covariate was indexed by geographic location in latitude and longitude. However, for modelling purposes, the coordinates were projected to UTM31U (Universal Transverse Mercator projection) and these are subsequently referred to as *Easting* and *Northing*. Temporal covariates were used to aid identification of any seasonal or long term change; these included day of the year (DoY) and *Year* of survey.



(a)



(b)

Figure 6.2: Effort across all years for harbour porpoise (a) and minke whales (b) available in the JCP resource.

6.3 Methods

6.3.1 Overview

The modelling methods used in this chapter are based upon CReSS (Chapters 3 and 4), for use in topographically complex areas such as this study area (Figure 6.2), and SALSA (Chapter 5) to allow spatially adaptive targeted smoothing. This approach permits smooth functions to be used for each environmental and temporal covariate and the spatial component. Additionally, the spatial component, used to determine similarity between observations, was constructed using at-sea distance (geodesic). In conjunction with these methods a Generalised Estimating Equation (GEE) framework [Hardin and Hilbe, 2002] was implemented to account for any residual autocorrelation (see Chapter 4 for details). Residual autocorrelation was thought likely to be present when considering the survey design, since data from multiple aerial and boat surveys are likely to be spatially and temporally correlated and the model is unlikely to explain this correlation in full. If positive residual autocorrelation is ignored, the uncertainty in the model parameters is underestimated leading to an underestimate of overall model uncertainty. Uncertainty in the entire modelling process was incorporated using a parametric bootstrap technique [Davison and Hinckley, 2007] and GEE based standard errors. This accounted for uncertainty in the detection function modelling and uncertainty in the model parameter estimation, but not model selection uncertainty, since the density surface model was not re-chosen for each bootstrap resample produced as part of the detection function process. Geo-referenced confidence intervals based on the estimated uncertainty were then produced for each density surface to provide a range of plausible surfaces based on the data. An overview of the methods process can be seen in Figure 6.3.

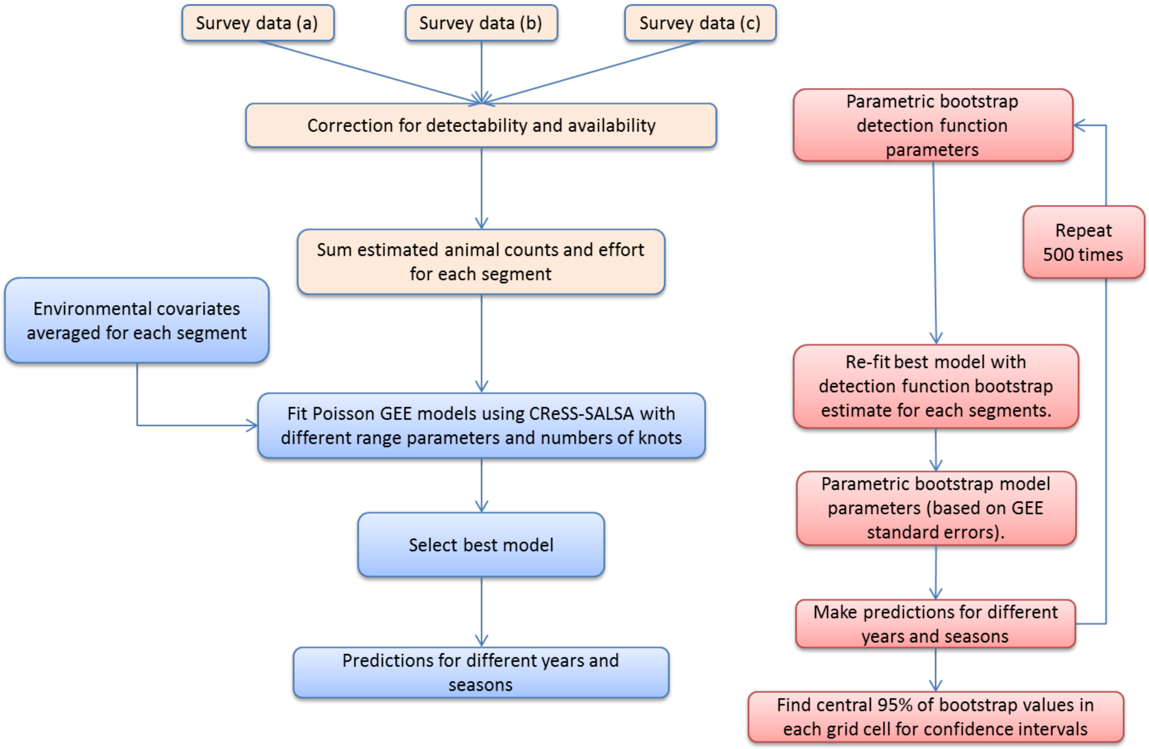

Survey data (a)
Survey data (b)
Survey data (c)
Correction for detectability and availability
Sum estimated animal counts and effort for each segment
Environmental covariates averaged for each segment
Fit Poisson GEE models using CReSS-SALSA with different range parameters and numbers of knots
Select best model
Predictions for different years and seasons
Parametric bootstrap detection function parameters
Repeat 500 times
Re-fit best model with detection function bootstrap estimate for each segments.
Parametric bootstrap model parameters (based on GEE standard errors).
Make predictions for different years and seasons
Find central 95% of bootstrap values in each grid cell for confidence intervals

Figure 6.3: An overview of the data collation (orange), modelling process (blue) and uncertainty estimation (red) for analysis of the JCP data resource.

6.3.2 Modelling Framework: GEEs

Given the methods of data collection, it is likely that the data are correlated in time. Further, if some of the covariates which explain this correlation are missing from the model, then GEEs, described in Chapter 4, section 4.3.2.1, are a suitable framework for modelling the remaining correlation in the residuals. A runs test [Mendenhall, 1982] on residuals from the final harbour porpoise model showed significant levels of positive correlation (H_0: independent residuals, $p << 0.0001$). This justified the consideration of non-independence (based on GEEs), because there are fewer runs of residuals than would be expected (each run is long, resulting in fewer runs) if the residuals were independent. GEEs require a panel variable to be specified, within which the residuals are permitted to be correlated. Conversely, independence is assumed between panels. Based on the survey design and the fact that the data set is a combination of results from multiple surveys, the panel variable (see Chapter 4, section 4.3.2.1 for more on panels) was specified using model residuals belonging to segments from the same day of survey and same observation vessel (survey-day-vessel). These residuals within panels were permitted to be correlated but were deemed independent between survey-day-vessels. This panel variable specification allowed the standard errors to be based on the autocorrelation within each panel. The panel structure led to 4835 panels for harbour porpoise and 6317 panels for minke whale, with the number of points in each panel ranging from 1 to 395.

The data are estimated counts per segment and are non-negative, so a log link function was used. While Poisson errors might be routinely assumed for data of this type (see Chapter 2, section 2.1 for Poisson GLM formulation) the high numbers of zeros in the data means that the expected relationship between the mean and variance for a Poisson model was not likely to hold (i.e. $V(y) >> \mu$). For this reason a dispersion parameter, which forms part of the GEE parameter estimation process, adjusts the variance appropriately.

A varying degree of survey effort contributed to each estimated count and so an offset term was included to model counts per unit effort. The area of each segment (in km^2) was used as an effort term and so the results of any predictions are animals per km^2.

6.3.3 Smoothing Details

The one dimensional covariates were each modelled non-linearly using cubic B-splines [see section 2.2.1 and Faraway, 2006] except for DoY which was modelled using a cyclic cubic regression spline [Wood, 2006]. Cyclic cubic splines have an extra condition that the fitted curve at the boundary knots at either end of the covariate range join smoothly. In the case of DoY, this ensures that day 1 and day 365 do not have a sharp change in relationship. SALSA1D [Walker et al., 2010] was used to choose the number and location of knots for these covariates, but was restricted to an upper bound of $df = 5$ (3 internal knots, 2 boundary knots) to prevent overly complicated models (in much the same way as is done when using the `gam` function in `mgcv`).

The two dimensional spatial smooth was modelled using CReSS, with the flexibility of the surface determined by both the number of knots and the range coefficient (r) for each knot. SALSA2D was used to determine the number and location of these knots. A 60km x 60km grid of points provided good coverage of the survey area from which candidate knots were chosen. SALSA2D was initialised using a space-filling algorithm [Johnson et al., 1990] on the data locations and selected locations were snapped to the candidate knot grid to give starting knot locations. Subsequent knot moves or changes were governed by SALSA2D. However, since the SALSA2D algorithm does not search all possible knot locations, several start points were considered (6, 8, 10 & 12 knots, Table 6.2).

A recent addition to the SALSA2D algorithm described in the discussion of Chapter 5 meant that once the final number and location of knots was selected, an appropriate value for the parameter r could be chosen for each knot in turn. As in knot selection, any changes were governed by a chosen fit criterion. Four candidate values of r were chosen that allowed a variety of local to global smoothing gradients (Table 6.2). The smallest value r_{min} gave basis function values close to zero (very local influence) and r_{max} was chosen to give basis function values close to 1 (global influence). The following formula was used to calculate r_{min} and r_{max}, and thus the range of r:

$$r_{min} = \sqrt{\frac{-g_{min}}{\log(0.05)}}$$

$$r_{max} = \sqrt{\frac{-g_{max}}{\log(0.99)}}$$

where g_{min} is the maximum of either the minimum geodesic distance between pairs of knots or the minimum geodesic distance from a candidate knot to a data point, and g_{max} is the maximum geodesic distance from a knot to a data point. A sequence of values was created using $\log(r_{min})$ and $\log(r_{max})$ and exponentiated to give the sequence of r's used in the analysis (Table 6.2). The log scale enabled the sequence to have more values for locally acting radii than for the globally acting ones. Simulations on the horseshoe and the palm regions using this more general method of calculating r gave similar results to those in Chapters 3 and 4 [Scott-Hayward et al., 2013].

A summary of the modelling parameters used in the analysis for both harbour porpoise and minke whale can be found in Table 6.2.

Table 6.2: Table of model parameters for the JCP analysis for both harbour porpoise and minke whale.

Parameter	**Value**
1D knots	1,2 and 3
2D start points	6, 8, 10 and 12
r	461, 1454, 4585 and 14462
Fit statistic	BIC
Error structure	Poisson

6.3.4 Model Selection Process

Initially, models were fitted to each one-dimensional covariate to establish the strength of any relationship between them and the counts per unit effort on the link scale. These models, together with the uncertainty about the relationships (via percentile based GEE confidence intervals), established an order of 'best' predictors, using an appropriate fit criterion. Any co-linearity between these covariates was also identified at this stage and one or other of the co-linear variables removed. Furthermore, any covariate that exhibited prohibitively large confidence intervals (including infinity) was also eliminated from the next stage. To obtain a single model, the remaining covariates were combined in order of predictive power. Since one of the main aims of the analysis was to assess temporal trends, *Year* was included by default and the remaining covariates were added one by one (conditional on improvement to the model fit) in order of predictive power. In all cases, SALSA1D was used to adjust the knots for each covariate as it was added.

Once an appropriate model containing one-dimensional terms was determined, a two-dimensional CReSS smooth of spatial coordinates, with knots selected by SALSA2D, was included in the model, conditional on an improvement in the fit statistic. If the data allowed, and there was an improvement in fit statistic, then an interaction term between the spatial surface and *Year* was also added. This allowed the distribution of animals to change substantially from year to year, rather than the whole surface moving up/down for each year. However, the spatial distribution of effort was extremely patchy, and some areas were only surveyed once or twice making it more difficult to fit an interaction (year-space) for one of the two species analysed.

Model selection was governed by the BIC statistic [Schwarz, 1978, see Chapter 2 for details], rather than based on p-values, because the latter relies on accurate estimates of standard errors. In this case, GEE models were fitted with a working independence correlation structure and empirical standard errors were used for model inference [Hardin and Hilbe, 2002]. This approach to the analysis does not assume that residuals within panels are independent, but uses the observed residual correlation within panels to return standard errors rather than rely on a specified model for the correlation structure. It also returns model coefficients which are identical to those obtained under independence and this equivalence means each model could be fitted as a GAM with Poisson errors, for the purposes of selection, and the final model re-fit as a GEE to estimate uncertainty. The BIC was used to govern both selection of knots and selection of covariates. This guarded against fitting models for the underlying model process that were overly complicated, which could occur using AIC, for example. However, the penalty added per parameter was based on the apparent sample size and not the effective sample size, which is likely to be smaller when autocorrelation is present. If the effective sample size was used then the penalty would probably be smaller and this would likely lead to more complex models being selected.

Five-fold cross-validation was also considered as a model selection alternative but in practice returned overly simplistic models that failed to identify cetacean concentrations that persisted over time. Model selection for GEEs is a current research area and while basic analogues to AIC are available (e.g. QIC$_u$ and QIC$_r$) the use of cross-validation to perform model selection for correlated data is not well studied [Pan, 2001b, Koper and Manseau, 2012].

6.3.5 Model Predictions and Inference

Model predictions were generated for each species on a 5 x 5 km resolution grid covering the survey area outlined in Figure 6.2 for four days (days 45, 136, 227 and 315, representing the four seasons) each year (1994 - 2010), using the best model chosen by BIC.

Percentile based 95% confidence intervals for each grid cell were generated using a two stage parametric bootstrap approach to include uncertainty from the detection function and the spatial modelling processes (see Figure 6.3 for an overview). Empirical (data-based) standard errors were used to represent parameter uncertainty at the modelling stage because these are robust to mis-specification of the correlation structure [Hanley et al., 2003]. Specifically 500 sets of estimated counts ($\hat{N}$) for each segment were generated based on 500 parametric bootstrap realisations of the detection function parameters. The best model, fitted using the real estimated counts and chosen using BIC, was re-fitted 500 times, each time with one of the sets of bootstrapped $\hat{N}$'s using a GEE fitting framework. From each model the variance-covariance matrix of the regression coefficients was then used to generate one parametric realisation of the model parameters. These parameters were then used to predict densities for the grid described above, resulting in 500 sets of predictions - each one relating to a different set of bootstrapped detection function input data ($\hat{N}$s). Confidence intervals were created by finding the lower 2.5th quantile and the upper 97.5th quantile for each of the 500 values in a grid cell.

This work forms part of a larger project [Paxton et al., 2013] for which I was the principle analyst for the spatial modelling and prediction aspects. I did not conduct the analysis for combining multiple sources of data to form the JCP data Resource (the source of the spatial modelling data), nor did I conduct the inference analysis (owing to it requiring SAS software). The points of discussion, in this chapter, both statistical and biological are my own.

6.4 Results

Table 6.3 shows the models selected for each species. Both models contain the same one dimensional covariates (*Year*, *Depth*, *DoY*, *Slope* and *SST*). All the covariates, except *Year* for minke whale, use three knots (which was the maximum permitted) and suggests that more knots might have been allowable. The two-dimensional smooth of *Easting* and *Northing* was low dimensional and chose 12 knots for harbour porpoise and just 11 knots for minke whale. The model for harbour porpoise also contains an interaction term between *Year* and geographic space (s(*Easting*, *Northing*):s(*Year*)). The covariates in the table are presented in the order in which they appeared in the model, and thus their order of importance based on the BIC scores of the individual models.

Results for both species are presented for the years 1994, 2005 and 2010. Plots showing seasonality can be found in Appendix D and the remainder of the time series for each species can be found in Appendix E. Maps corresponding to each of the EU Habitats Directive reporting periods (1994 - 2000, 2001 - 2006, 2007 - 2010), are in Appendix F. Appendices E and F are on the accompanying CD.

6.4.1 Harbour Porpoise

Many (n= 10093) of the 88734 harbour porpoise segments contained non-zero estimated counts, resulting in a high mean sightings rate, relative to other cetacean species, of 0.87 porpoises per km^2. All of the available covariates were successfully fitted with one-dimensional

Table 6.3: Table showing the best selected model, based on BIC, for each species. bs = B-spline, cc = cyclic cubic and CR = CReSS basis. The numbers in brackets are the number of knots selected (for *df* one additional *df* for each boundary knot). X and Y represent the covariates *Easting* and *Northing* and the order of covariates is the order in which they entered the model.

Species	**Model**
Harbour porpoise	bs(Year, 3) + bs(Depth, 3) + cc(DoY, 3) + bs(Slope, 3) + bs(SST, 3) + CR(X, Y, 12) + s(Year, X, Y, 36)
Minke whale	bs(Year, 1) + cc(DoY, 3) + bs(SST, 3) + bs(Slope, 3) + bs(Depth, 3) + CR(X, Y, 11)

smooth functions and selected the maximum flexibility available (3 knots) for the smooth term (Figure 6.4). Co-linearity was not evident at this stage and so the order of best predictors, based on the BIC, can be seen in Table 6.4. *Depth* was the single best covariate for predicting harbour porpoise counts per km^2. Figure 6.4 shows the relationship of each covariate with animal counts. Harbour porpoise show a preference for shallow water < 50m deep, a *slope* of more than 0.5 and a *SST* between 5^{o}C and 15^{o}C. The temporal covariates indicate a sharp decline in density over the prediction region in the late 1990s followed by an increase until about 2008, with some evidence of a decline thereafter, although the confidence intervals become wider after 2007. The relationship with day of the year indicates the lowest numbers of animals were counted in the summer months. However, these are plots of models fitted to a single covariate, and so there is no account of the effect of other covariates on the response. Furthermore, the covariates and models selected were chosen based on their predictive ability, and not a priori based on biological interpretation. Therefore, biological inference from these plots should be measured.

The addition of each covariate to the model gave an improvement in BIC score. The final model contained a two dimensional smooth of *Easting* and *Northing* and an interaction between this two dimensional smooth and *Year*. SALSA2D chose 12 knots, the maximum permitted, for this spatial term and their locations can be seen in Figure 6.5. These knots

Table 6.4: Order of the best single covariate predictors for the harbour porpoise data. Order is calculated based on BIC scores for each of the models where each model contains a single covariate.

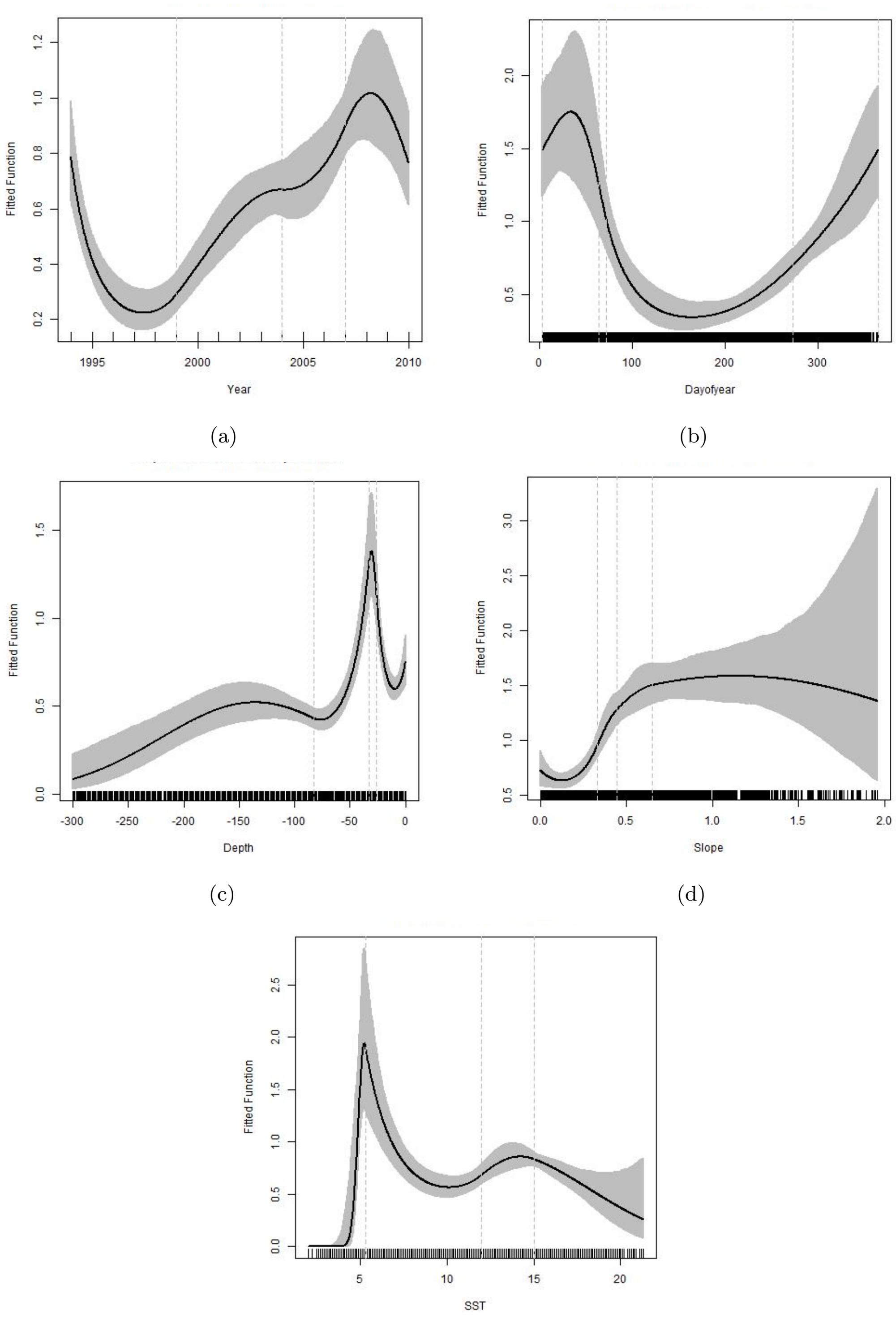
Covariate	**Δ BIC**
Depth	0.00
Day of year	14100
Year	15600
Slope	20500
Sea surface temperature	26400

were fixed across all years.

Predicted density plots for harbour porpoise in the years 1994, 2005 and 2010 are shown in Figures 6.6, 6.7 and 6.8. The raw count data from time periods around the year of prediction are shown alongside the point estimates, together with the GEE-based lower and upper 95% intervals. The colour scale used is taken from Hammond et al. [2013] with some additions at the upper limit to allow for the uncertainty plots. More effort was focused on coastal UK waters toward the end of the study period, which makes the uncertainty in some areas of the study region, for example, the Kattegat/Skagerrak (Figure 6.8(d)), quite high in 2010. If there was no interaction term, then the latter years could borrow strength from good coverage in the early years. However, any temporal shifts in distribution would not be identified. The two areas with the greatest uncertainty in 2010 are the west coast of Ireland and the Kattegat/Skagerrak, where there are no data for the year of prediction or the two preceding years. There seems to be a shift in high density areas from the central North Sea in 1994 to the coast off East Anglia in 2005 and 2010. This shift is also evident in the upper confidence interval surface. The model fits the estimated raw counts well at the beginning and end of the temporal range (Figures 6.6 and 6.8) but not in the mid range (Figure 6.7). In 2005 there are very few non-zero predictions in the North Sea, which appears to be an area of high density in the estimated raw counts. This is likely due to restricted spatial coverage (mostly coastal) in the years pre and post 2005 and the very high estimated observed densities off the west coast of Scotland down through the Irish Sea (Figure 6.9).

Figures 6.10 and 6.11 show harbour porpoise density predictions for winter, spring, summer and autumn 2010. There is a striking lack of coverage of effort in winter and autumn compared with spring and summer. Even in spring and summer the coverage of effort is poor with most effort occurring in UK coastal waters. Figure D.1 in Appendix D shows the effort for each season across all years. It is clear that most of the data (41%) were collected in the summer months. GEE based 95% confidence intervals for these 2010 season density plots are also presented in Appendix D.

Figure 6.12 shows the change in relative abundance for the whole study region during the study period. The total abundance in each year was calculated by summing over each prediction grid cell, which differs to Figure 6.4.1, where one estimate (a mean) was predicted for each year (and year was the only covariate). Here, the confidence intervals are quite wide, making inference difficult, but there is some indication of an increase in numbers from about the year 2000.



(a)

(b)

(c)

(d)

(e)

Figure 6.4: The relationship between each covariate and harbour porpoise counts per km^2 (density on the y-axis); (a) Year, (b) Day of the Year, (c) Depth, (d) Slope and (e) Sea surface temperature. The plots show a cubic B-spline (or cyclic cubic in the case of DoY) with GEE-based 95% confidence intervals (grey shading) for each covariate. The tick marks at the bottom show the distribution of the data and the grey dashed lines show the location of the knots. Three internal knots were selected by SALSA1D for each covariate.

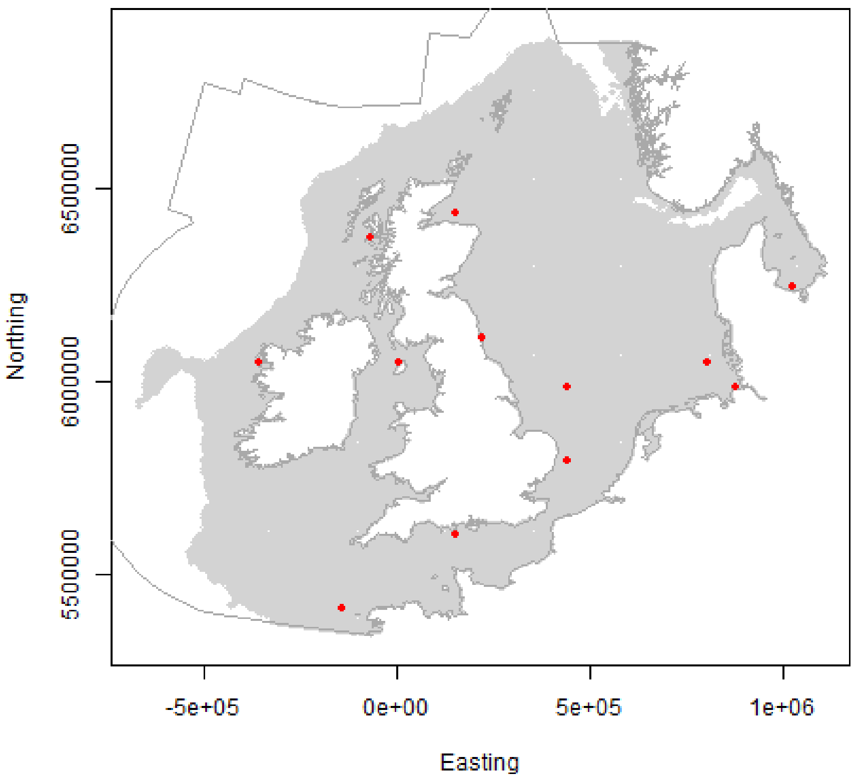
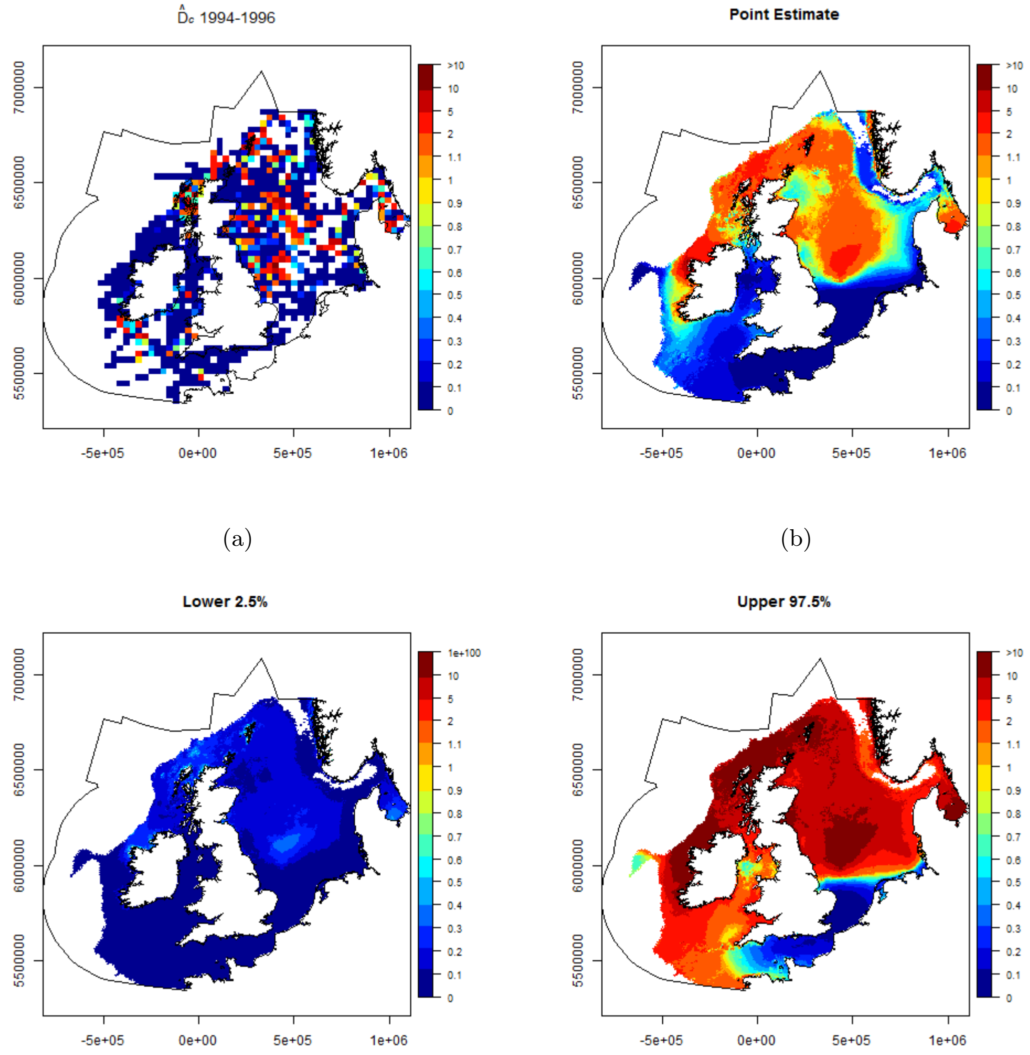


Figure 6.5: A map of the study area showing the prediction grid (grey points) and the knot locations chosen by SALSA2D for the two dimensional smooth of *Easting* and *Northing* in the harbour porpoise model.



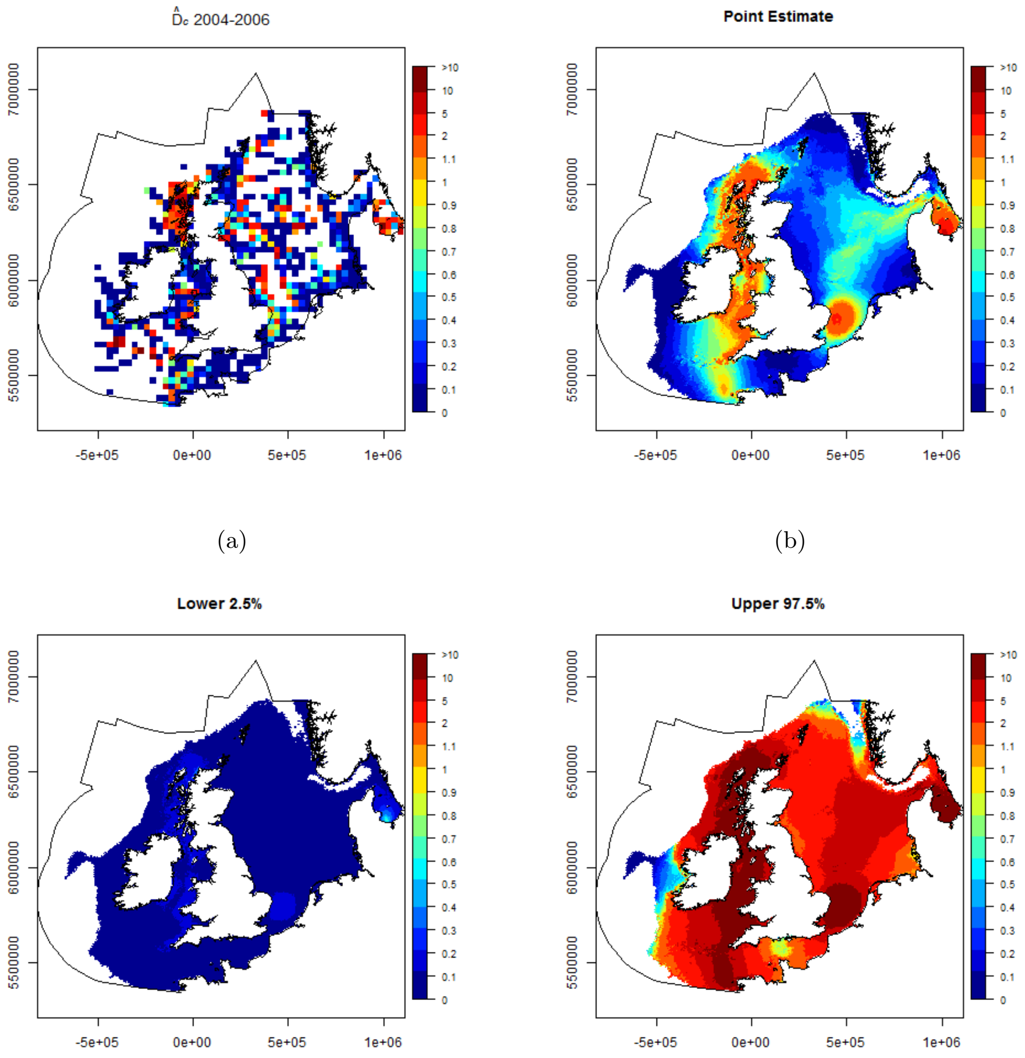
(a)

(b)

(c)

(d)

Figure 6.6: Predicted harbour porpoise densities for summer (day 227) in 1994. (a) The estimated raw densities for summers in 1994 - 1996 that are drawn upon to make predictions for 1994. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



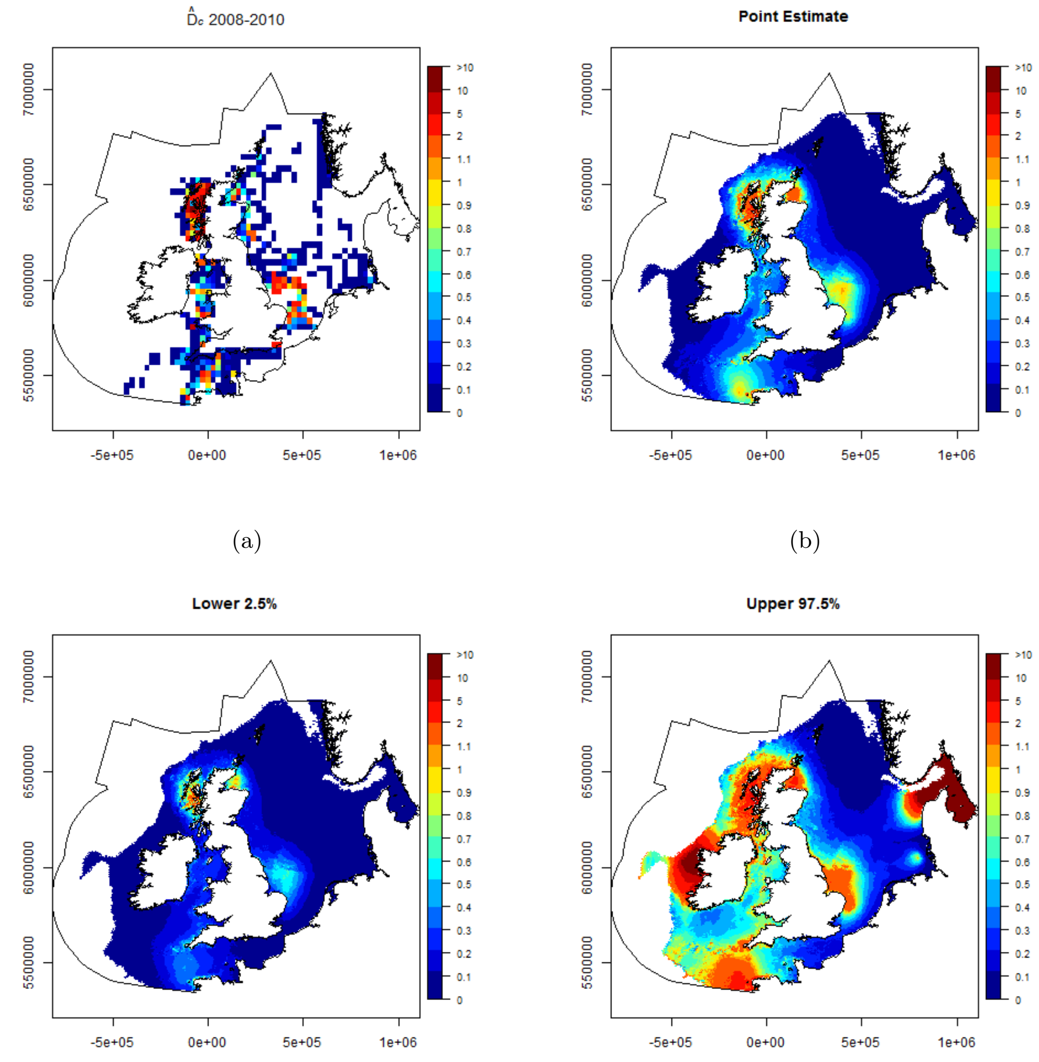
(a)

(b)

(c)

(d)

Figure 6.7: Predicted harbour porpoise densities for summer (day 227) in 2005. (a) The estimated raw densities for summers in 2004 - 2006 that are drawn upon to make predictions for 2005. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



$\hat{D}_c$ 2008-2010

Point Estimate

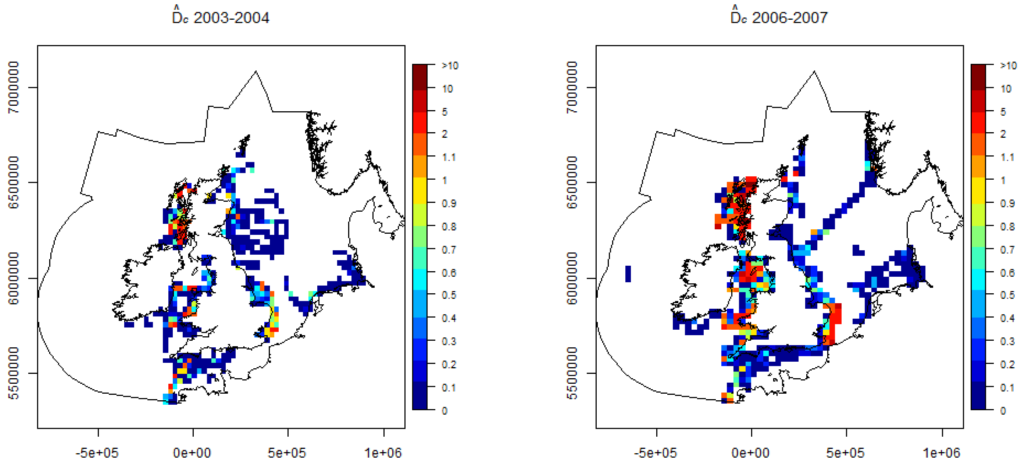
Lower 2.5%

Upper 97.5%

(a) (b)

(c) (d)

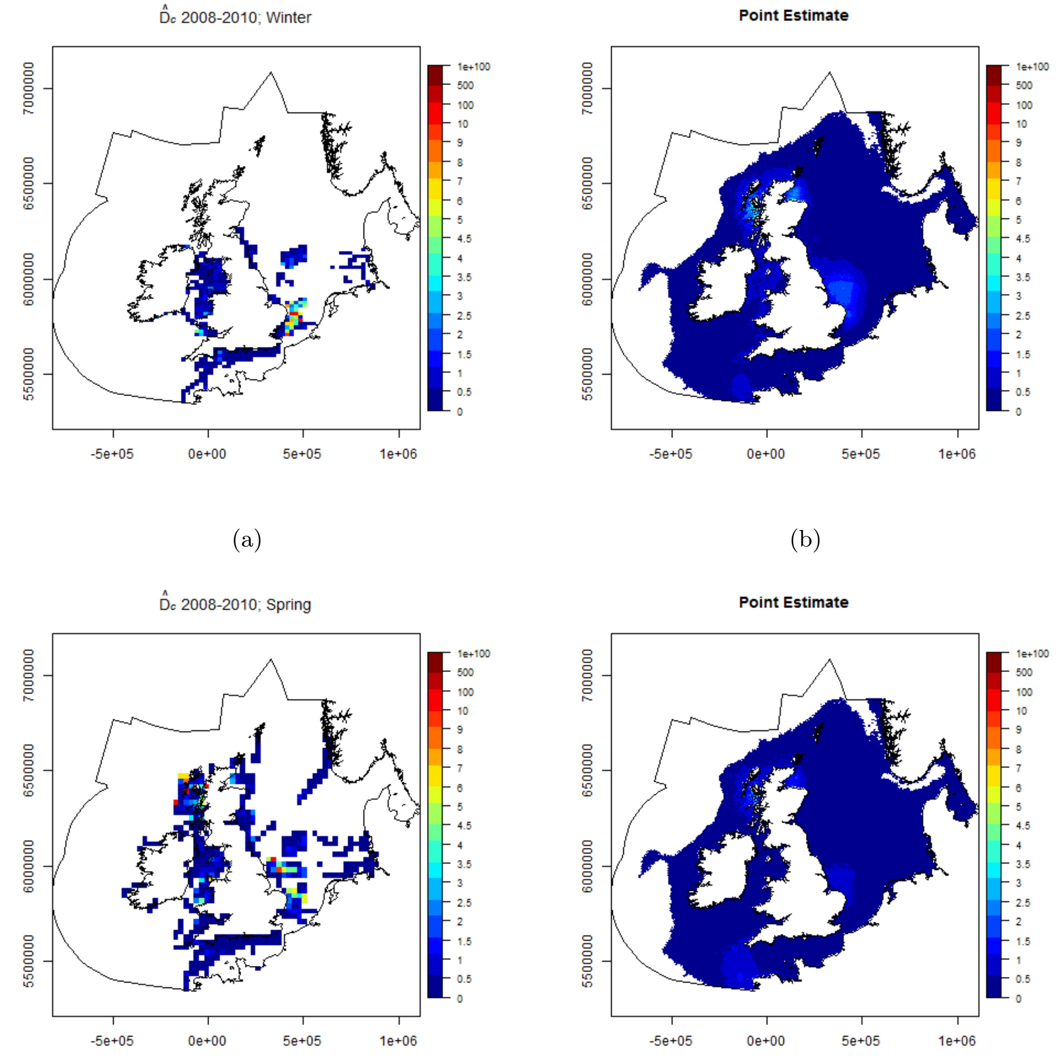
Figure 6.8: Predicted harbour porpoise densities for summer (day 227) in 2010. (a) The estimated raw densities for summers in 2008 - 2010 that are drawn upon to make predictions for 2010. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



(a)

(b)

Figure 6.9: (a) The estimated raw densities for summers in 2003 - 2004 and (b) the estimated raw densities for summers in 2006 - 2007.



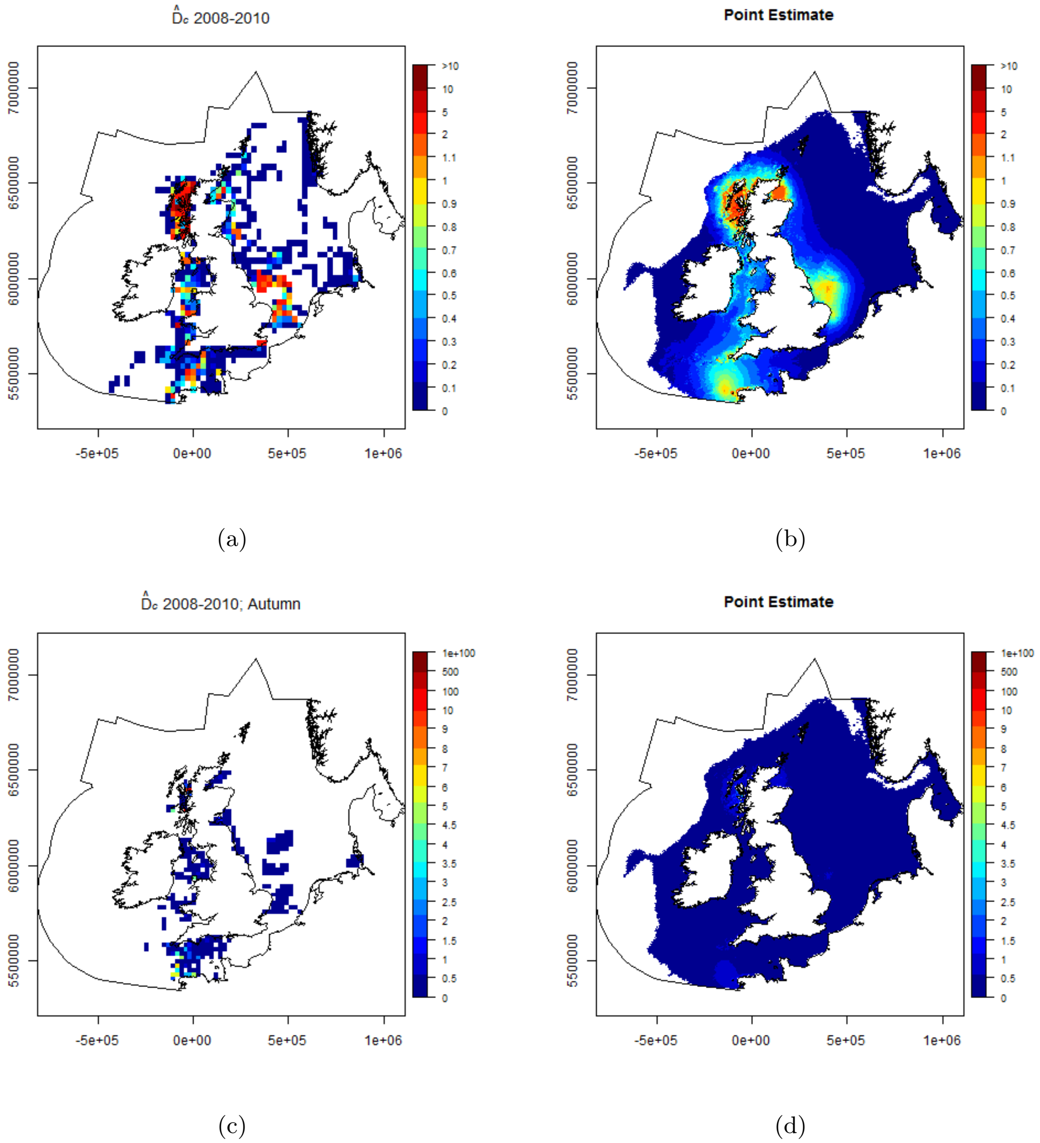
(a)

(b)

(c)

(d)

Figure 6.10: Harbour porpoise densities and predicted densities for 2010 in winter (top) and spring (bottom). The plots on the left are the estimated raw densities for winter or spring in 2008 - 2010 that are drawn upon to make predictions for 2010. The plots on the right are predictions for 2010 in winter and spring. Plots of confidence intervals for these estimates are in Appendix D.



(a)

(b)

(c)

(d)

Figure 6.11: Harbour porpoise densities and predicted densities for 2010 in summer (top) and autumn (bottom). The plots on the left are the estimated raw densities for summer or autumn in 2008 - 2010 that are drawn upon to make predictions for 2010. The plots on the right are predictions for 2010 in summer and autumn. Plots of confidence intervals for these estimates are in Appendix D.

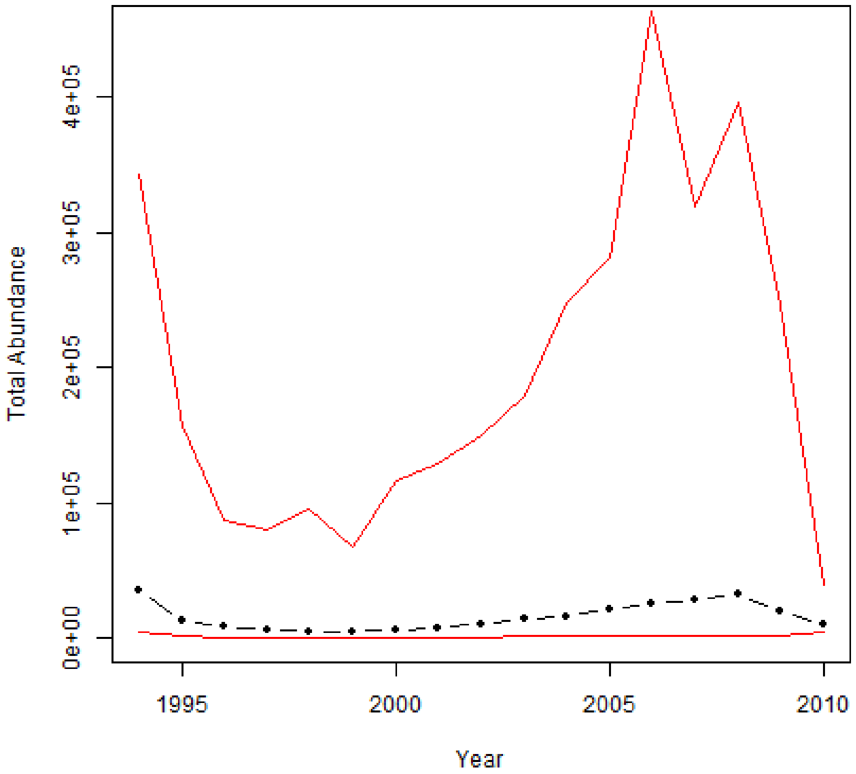


Figure 6.12: Harbour porpoise predictions of relative abundance summed over the whole survey area for each year in the summer (day 227). This excludes the area of the Kattegat and Skagarrak to the east of 8.2 x 10^5 Easting due to very high uncertainty in this region for 2010. Red lines are 95% GEE based percentile intervals from the parametric bootstrap process (Figure 6.3).

6.4.2 Minke Whales

Counts of minke whales were estimated in 1152 out of the 131448 segments, resulting in a relatively low mean sightings rate of 0.022 whales per km^2. All of the available covariates were successfully fitted with one-dimensional smooth functions and all, except *Year*, with the maximum available flexibility (3 knots) for the smooth term (Figure 6.13). Co-linearity was not evident at this stage and the order of best predictors, based on the BIC, can be seen in Table 6.5. DoY was the single best covariate for predicting minke whale counts per km^2. Figure 6.13 shows the relationship of each covariate with animal counts. Minke whales show a preference for seabed *Depth* of 100-200m, *slope* between one and 1.5, and SST of around 14 ^{o}C. The temporal covariates suggested there was a decline in numbers over the survey period and that most animals were seen in the late summer. As discussed in the harbour porpoise results section, it would be unwise to use these figures for biological inference.

Table 6.5: Order of the best single covariate predictors for the minke whale data. Order is calculated based on BIC scores for each of the models.

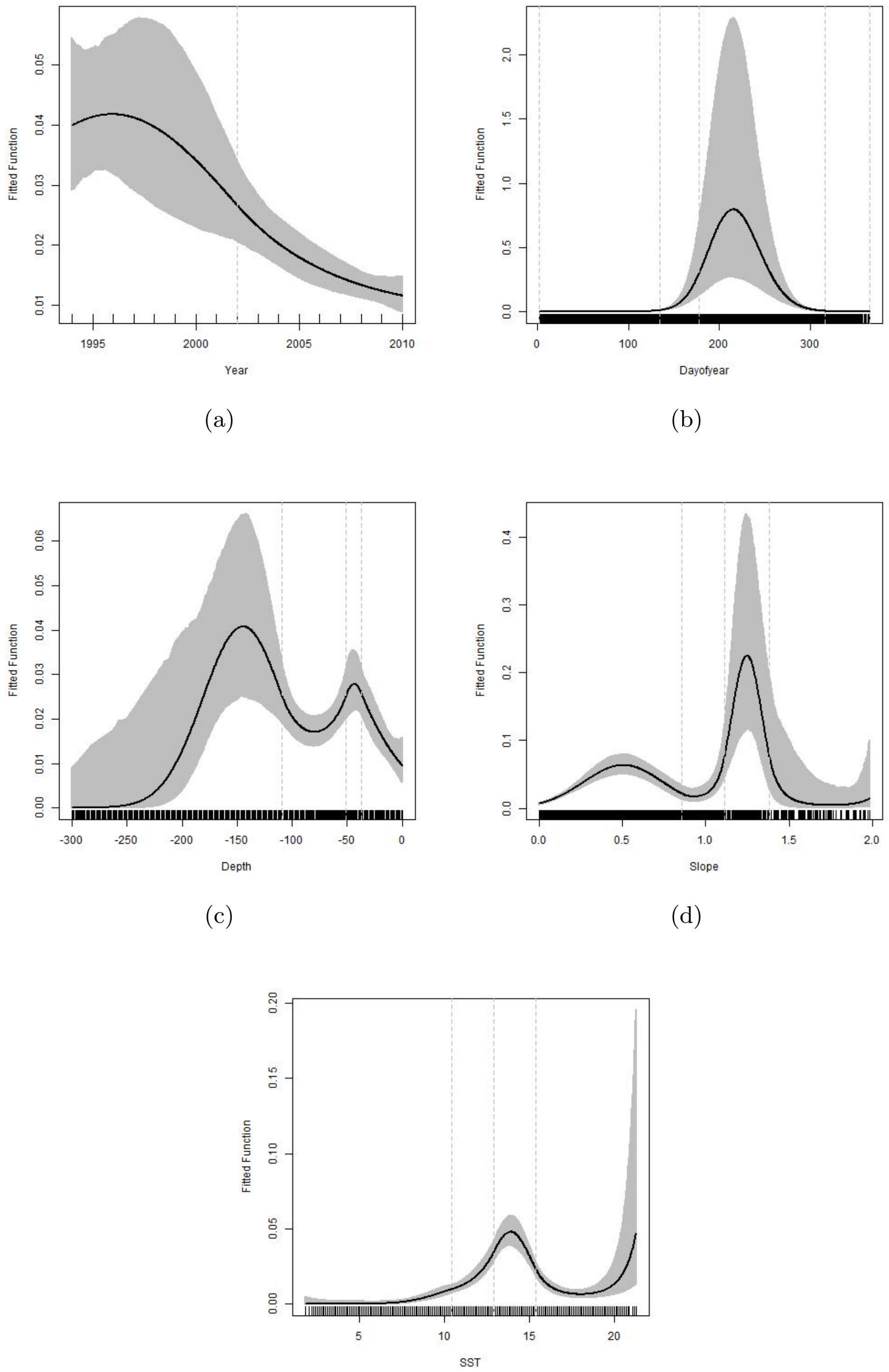
Covariate	**Δ BIC**
Day of year	0.00
Sea surface temperature	434
Slope	631
Year	9650
Depth	12900

The addition of each covariate to the model gave an improvement in BIC score and the final model also contained a two dimensional smooth of *Easting* and *Northing.* Based on BIC scores, a two-dimensional smooth with 11 knots, whose locations were chosen by SALSA2D, was the best model. The SALSA2D algorithm was initialised with 12 knots but removed one knot to improve model fit. The locations of the chosen knots were fixed across years and they are shown in Figure 6.14.

Figures 6.15, 6.16 and 6.17 show predicted plots of the density of minke whale in 1994, 2005 and 2010. The raw count data from the whole study period, together with GEE-based lower and upper 95% percentiles are shown alongside the point estimates. The colour scale used is taken from Hammond et al. [2013] with some additions at the upper limit to allow for the uncertainty plots. The highest densities of minke whale were recorded in the west and north North Sea, off the west coast of Scotland, off the south coast of Ireland and around the Isle of Man. No interaction term was used in the model because this resulted in high uncertainty (unrealistic values) in the extra parameters estimated. The distribution patterns identified therefore persisted throughout the study period, although there was an apparent general decrease in density through time.

The highest densities were recorded off the west coast of Scotland in the summer (Figures 6.18 & 6.19). The mean sightings rate in the whole study area was 0.030 whales per km^2. In the autumn and winter, the sightings rate dropped to 0.0036 whales per km^2 and this is reflected in the low densities shown in Figures 6.18 & 6.19. The best and most even coverage of effort was in spring (29% of total effort) and summer (43% of total effort). However, the coverage is better in winter and autumn than for harbour porpoise because data collected in sea states greater than Beaufort two were included.

Figure 6.20 shows the change in relative abundance for the whole study area over the study period. There is some suggestion that minke whale numbers were at a maximum around the year 2000, but the confidence intervals associated with this time period are wide making inference difficult.



(a)

(b)

(c)

(d)

(e)

Figure 6.13: The relationship between each covariate and minke whale counts per km^2 (density on the y-axis); (a) Year, (b) Day of the Year, (c) Depth, (d) Slope and (e) Sea surface temperature. The plots show a cubic B-spline (or cyclic cubic in the case of DoY) with GEE-based 95% confidence intervals (grey shading) for each covariate. The tick marks at the bottom show the distribution of the data and the grey dashed lines show the location of the knots. Three internal knots were selected by SALSA for each covariate except *Year* which had only one.

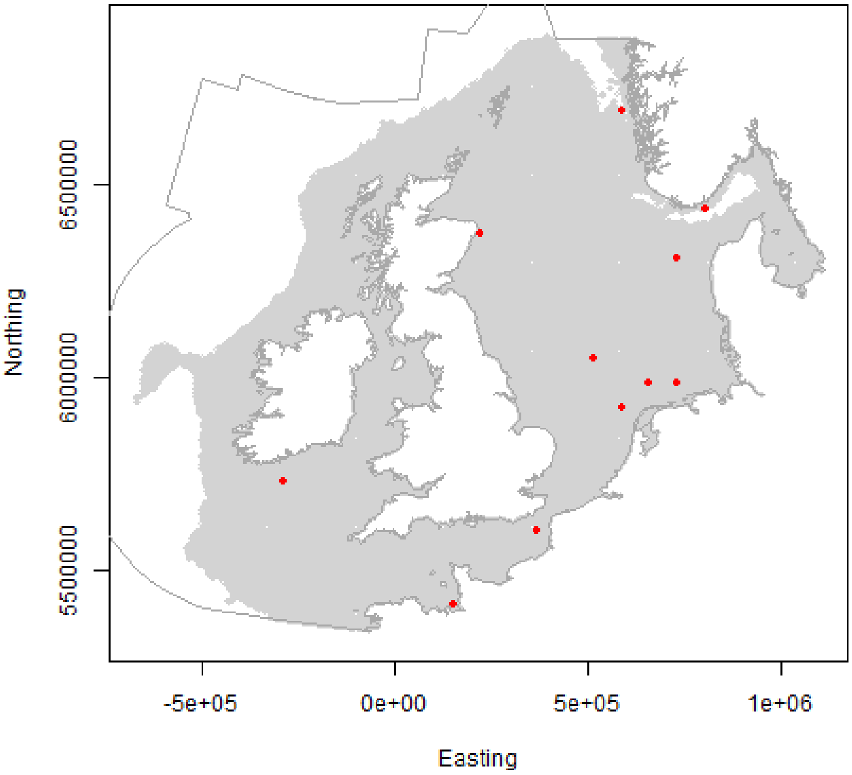
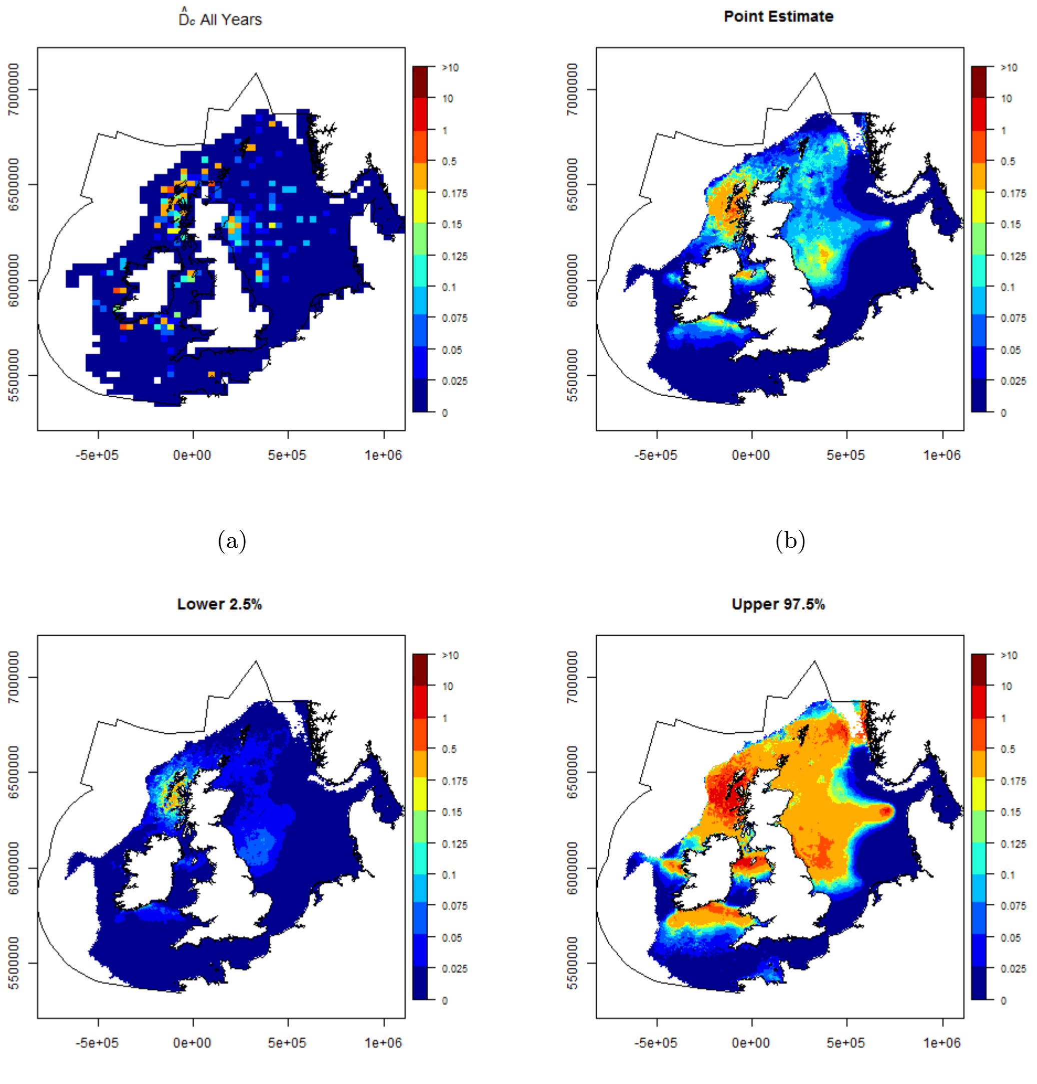


Figure 6.14: A map of the study area showing the prediction grid (grey points) and the knot locations chosen by SALSA2D for the two dimensional smooth of *Easting* and *Northing* in the minke whale model.



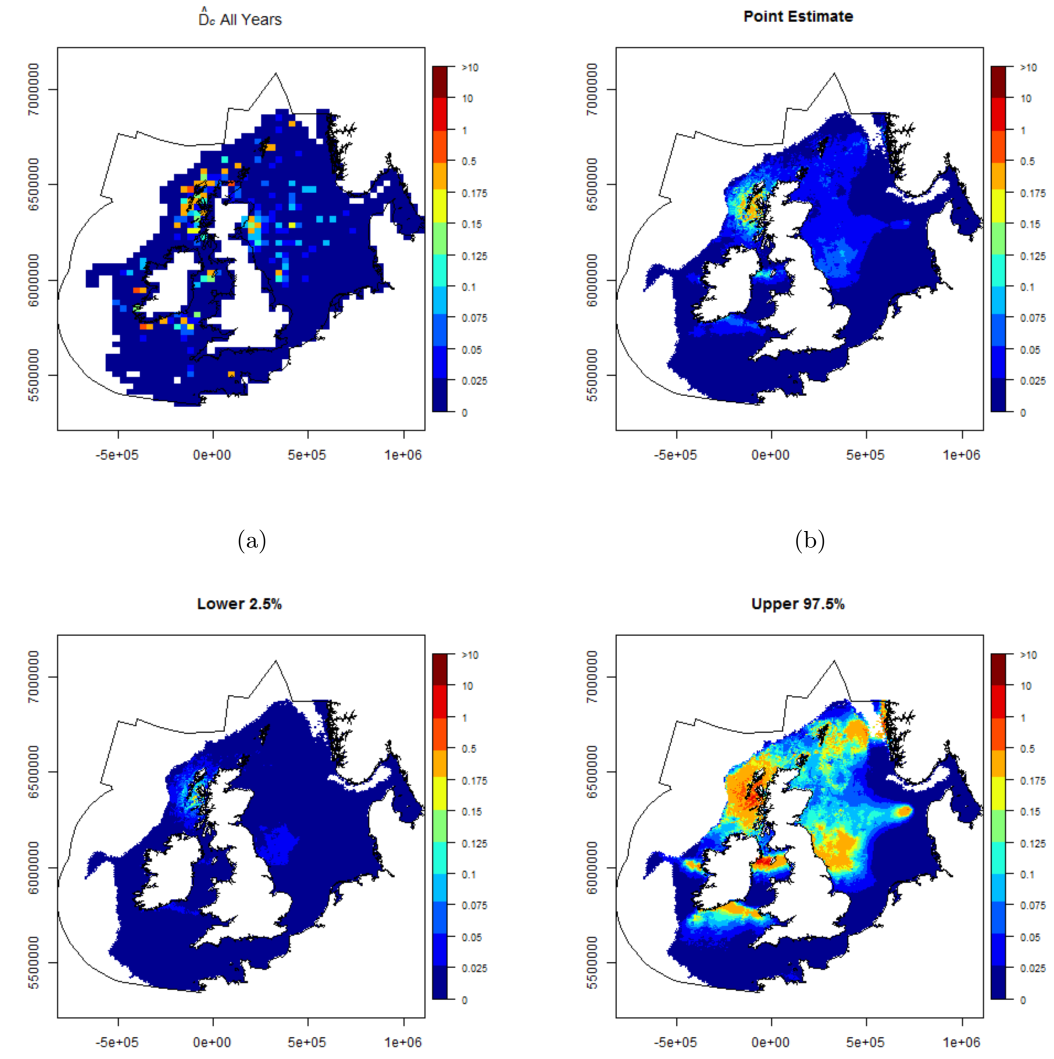
(a)

(b)

(c)

(d)

Figure 6.15: Predicted minke whale densities for summer (day 227) in 1994. (a) The estimated raw densities for all years. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



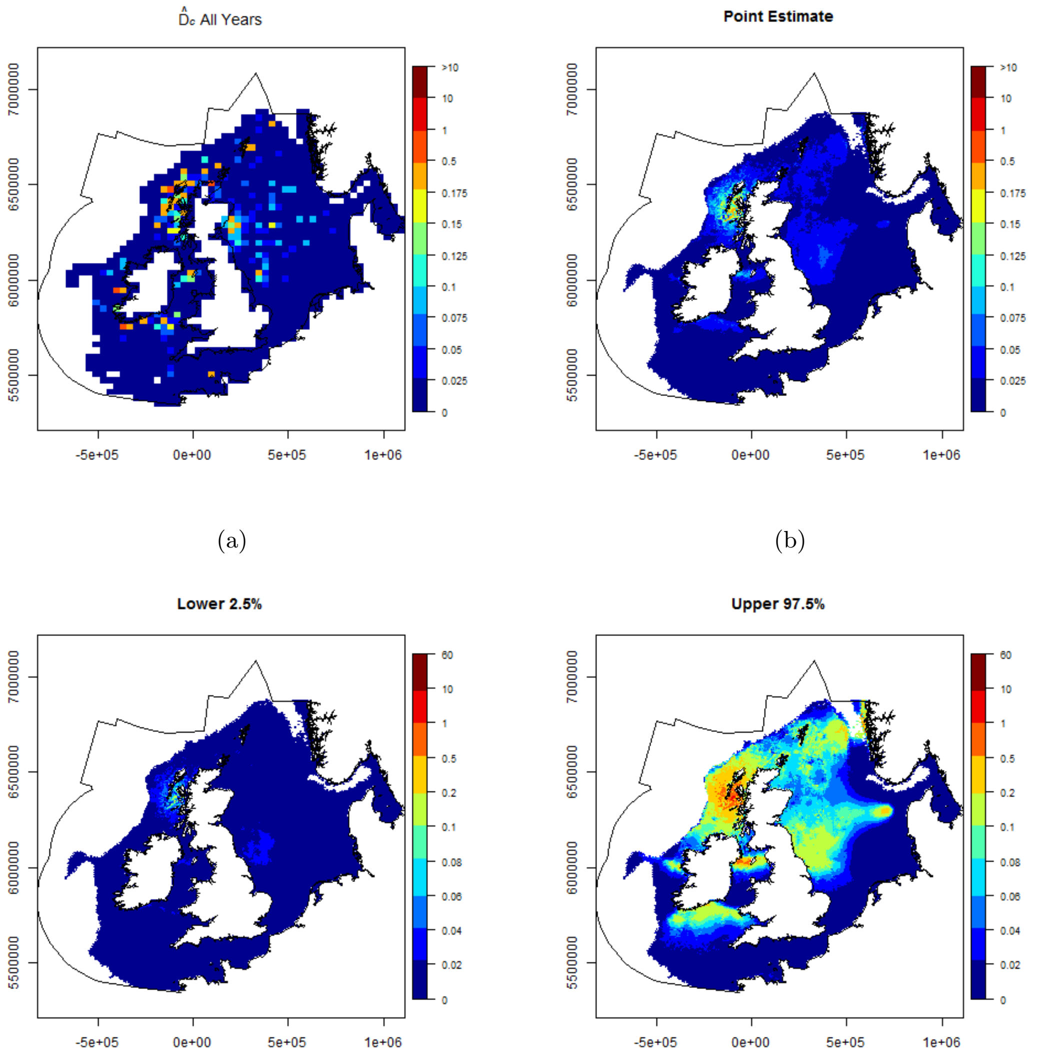
(a)

(b)

(c)

(d)

Figure 6.16: Predicted minke whale densities for summer (day 227) in 2005. (a) The estimated raw densities for all years. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



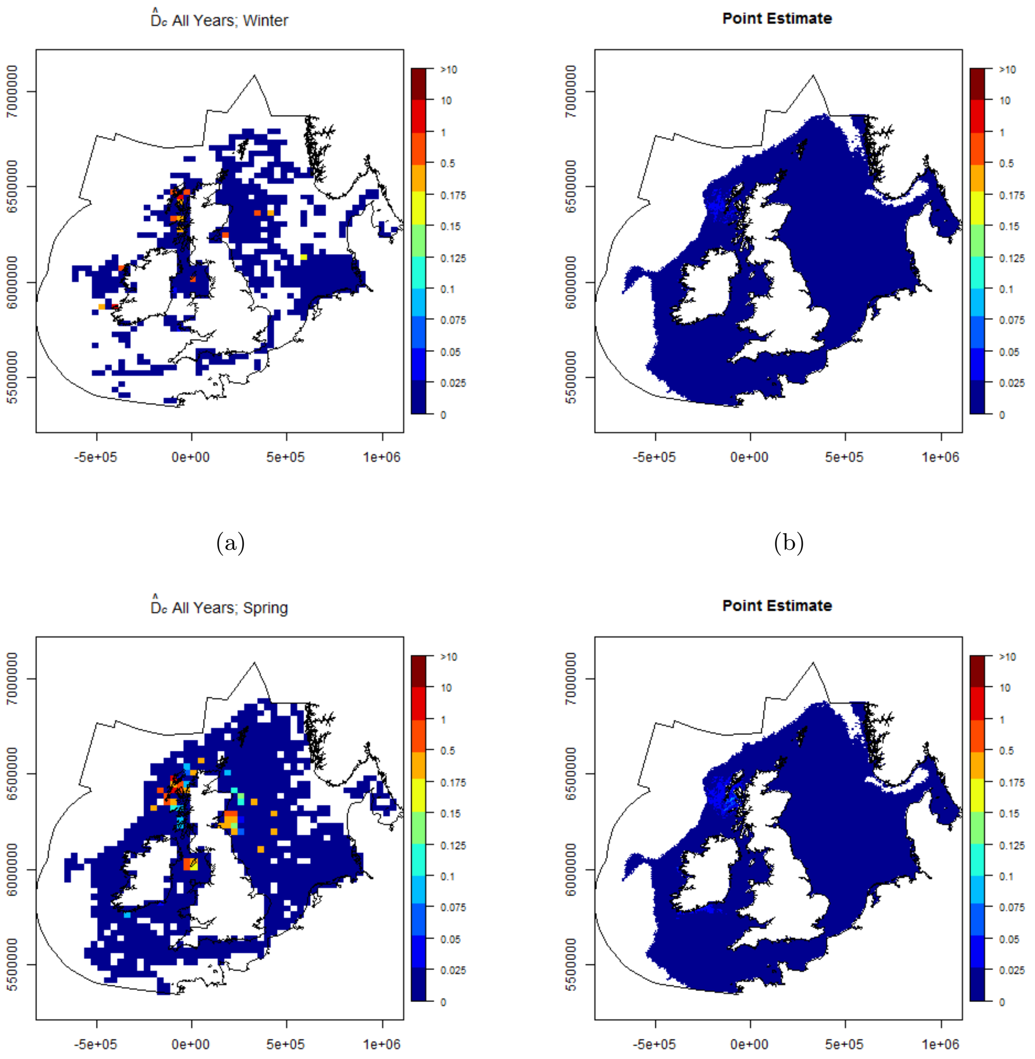
(a)

(b)

(c)

(d)

Figure 6.17: Predicted minke whale densities for summer (day 227) in 2010. (a) The estimated raw densities for all years. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



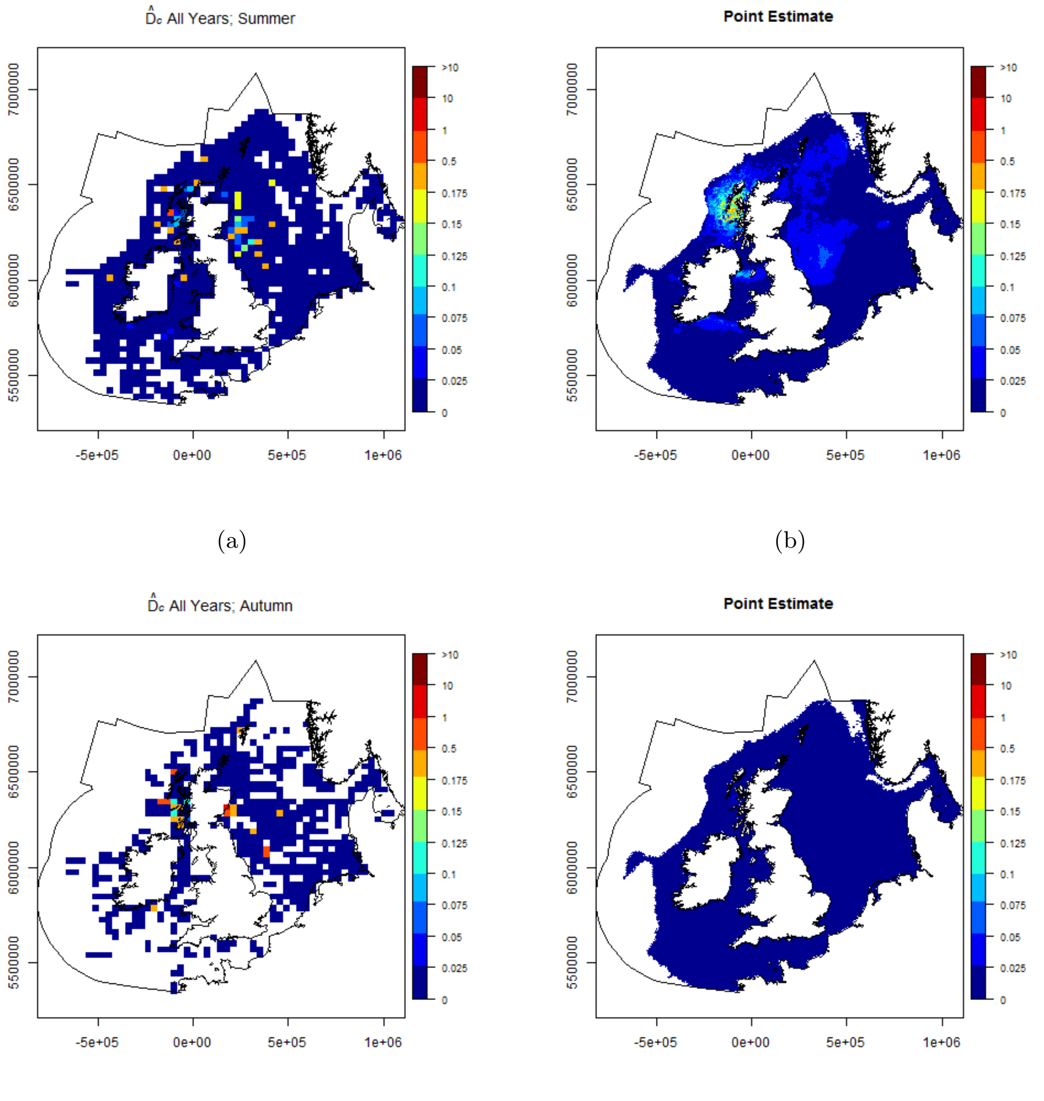
(a)

(b)

(c)

(d)

Figure 6.18: Minke whale densities for 2010 in winter (top) and spring (bottom). The plots on the left are the estimated raw densities for winter or spring in all years. The plots on the right are predictions for 2010 in winter and spring. Plots of confidence intervals for these estimates are in Appendix D.



(a)

(b)

(c)

(d)

Figure 6.19: Minke whale densities for 2010 in summer (top) and autumn (bottom). The plots on the left are the estimated raw densities for summer or autumn in all years. The plots on the right are predictions for 2010 in summer and autumn. Plots of confidence intervals for these estimates are in Appendix D.

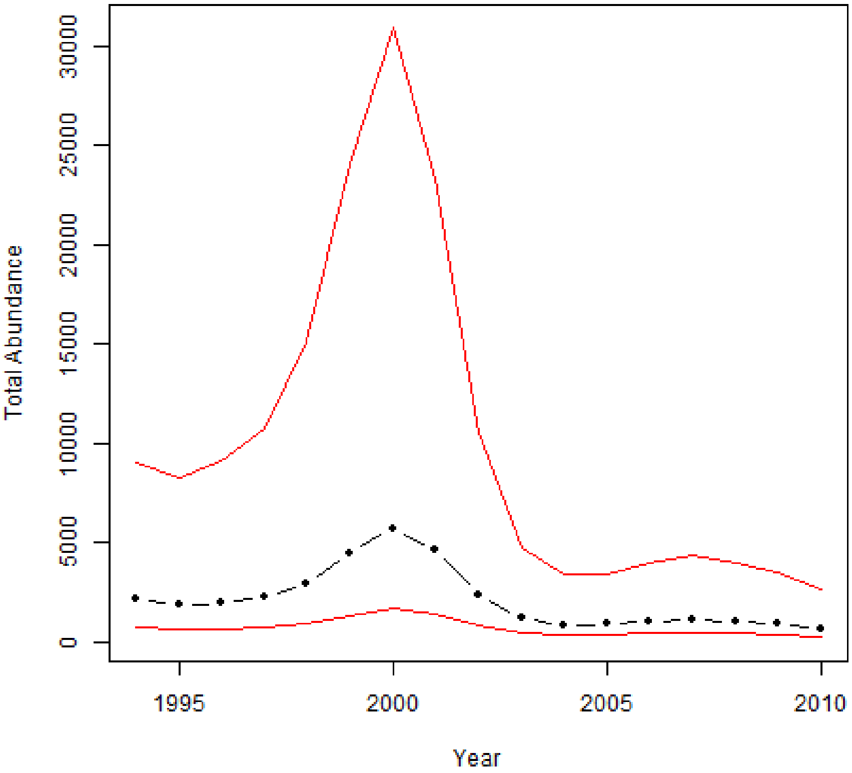


Figure 6.20: Minke whale predictions of relative abundance summed over the whole survey area for each year in the summer (day 227). Red lines are 95% GEE based percentile intervals from the parametric bootstrap process (Figure 6.3).

6.5 Discussion

The aim of this chapter was to assess changes in the range and relative abundance of harbour porpoise and minke whale over the period 1994 - 2010 using methods developed in this thesis (CReSS and SALSA2D). The analysis shows these methods are capable of analysing large quantities of data with multiple covariates in a topographically complex region.

However, modelling these data was a challenging exercise; survey effort was often poor and only a limited number of covariates were available for modelling. For instance, survey effort was limited in some years and there was uneven survey coverage in others. This uneven coverage made it difficult to provide wide-ranging long term estimates with any reliability, because the uncertainty associated with density estimates in poorly sampled areas tends to be much greater than for well sampled areas. The highest levels of sampling effort were exerted in 2010, but even these surveys had poor spatial coverage (relative to the area under study) and the sampling was almost entirely confined to coastal areas. The limited number of covariates available for modelling is likely to have restricted the predictive power of the models. However, additional covariates can only be included if they have spatial and temporal coverage suitable for an analysis of this scale. In particular, the inclusion of more biologically meaningful covariates, such as the presence and direction of tidal currents and prey distribution, might have improved the models ability to explain patterns in the distribution of animals. Embling et al. [2010] showed that tidal currents were a strong predictor of harbour porpoise density off the west coast of Scotland and several authors have suggested that prey distribution is key to explaining the distribution for minke whales [Macleod et al., 2004, Anderwald et al., 2012]. Furthermore, the covariates used in this study, as in many others, probably only served as proxies for the ‘real’ relationship a species has to its environment. Prey availability seems a likely candidate to drive the observed distribution of a species but information on prey distribution is rarely available. Environmental covariates such as depth and sea-surface temperature are likely to affect primary productivity and therefore influence the distribution of prey and their predators.

6.5.1 Methodological comparisons

The 1995 ESAS study [Northridge et al., 1995a] and the cetacean atlas [Reid et al., 2003] plotted sightings rates with effort overlaid, rather than employing any spatial modelling techniques. For instance, Northridge et al. [1995a] pooled data over 12 years and adjusted sightings rates (animals per km of trackline) for Beaufort sea state and month using a generalised linear model. Similarly, the atlas [Reid et al., 2003] reports sightings rates (individuals sighted per unit time) adjusted for sea-state corrected effort based on approximately 20 years of data and comprised of three data sets. For this reason, the ESAS and the cetacean atlas pool data over time, effectively giving mean sighting rates over the duration of each study period.

The SCANS surveys provide snapshots of cetacean distribution in the north-western European waters in the years in which surveys were conducted. Here the JCP analysis used *Year* as a covariate, which allowed temporal trends to be identified and a better understanding of the animals use of space through time. The full set of temporal distribution maps (1994-2010) was not presented here but can be found in Appendix E, along with mean estimates for each reporting period.

The SCANS surveys were well-designed line transect surveys and so sightings were corrected for detectability using Distance sampling [Buckland et al., 2001]. Spatial models using a GAM were then fitted to the corrected animal counts to produce distribution maps over the survey area. Similar to this JCP analysis, animal counts were modelled using a quasi-Poisson distribution however, only a few covariates were available for model selection, and correlation in the residuals was not considered as part of the modelling. Furthermore, a single smoothing parameter was used across the entire study region, restricting the surfaces to be uniformly flexible across the study area.

This was not the first analysis conducted on the JCP data resource. Two previous preliminary analyses (Phase I and Phase II) were conducted over differing temporal and spatial scales [Paxton and Thomas, 2010, Paxton et al., 2011]. Broadly speaking, each phase consisted of two parts; data collation and spatial modelling. Table 6.6 shows the main differences in the spatial distribution modelling sections of the two phases.

Phase I was analysed by Paxton and Thomas [2010] and was conducted as a test to see if the JCP data resource could be useful. They used a subset of sightings data from 1980-2009 in the Irish Sea. A two stage modelling process was used: the presence versus absence or each species was modelled using a logit-link based Generalised Additive Model (GAM), while the non-zero segments in the data were modelled separately using a GAM with Gamma errors.

Phase II [Paxton et al., 2011] was an extension of Phase I and encompassed a wider geographical area that included the Celtic Sea, the continental shelf to the west of Britain and extended to the longitude of the Hebrides. In this phase, the spatial coordinates were modelled using a CReSS approach underpinned with geodesic distances. This additional model complexity, compared with the Phase I approach, was deemed necessary to avoid unrealistic leakage in hotspots across land forms (and peninsulas) and to permit some areas of the surface to be more flexible than others (based on a choice of locally to globally acting bases for fixed knot locations). This latter issue was thought to be crucial since assuming uniform flexibility across the extended survey area seemed even more unrealistic.

This analysis, Phase III, extended the spatial range even further and covered north-western European waters to the 300m depth contour. In contrast to earlier analyses however, the temporal range was necessarily reduced due to poor effort prior to 1994 in newly considered areas. The modelling approach was improved by using an automated model selection process (SALSA) inside CReSS to include spatially adaptive knot placement. Additionally, an updated version of CReSS allowed automated selection of the ‘range’ parameter and the inclusion of SALSA permitted spatially adaptive knot placement for both one- and two-dimensional covariates. Furthermore, improvements to the SALSA algorithm allowed the 'range' parameter to vary across the surface within a given model. These improvements focused the modelling effort into areas with greatest need and were implemented to prevent over smoothing in highly structured areas of the surface and under smoothing in the flatter areas. In practice, the modelling approach used here reduced the chance of underestimating densities in hotspot areas and overestimating densities in areas where animals were rarely seen (e.g. off-shore areas for a primarily coastal species). Furthermore, the occurrence of 'leakage' was extremely small (limited by the accuracy of the boundary specification) further reducing the chance of high density predictions near coastal regions biasing/being biased by low density areas previously deemed to be nearby and vice versa. If a traditional method, such as an ordinary GAM, was used for modelling these highly uneven surfaces it is easy to see how a hotspot in the data could be smoothed out and be overlooked as a high use area in the fitted surface. This could result in popular areas being excluded from consideration as areas of special conservation interest. While the fitted surfaces for each species are of primary interest, it is also important to give perspective to these predictions by considering the uncertainty in these predictions and the plausible range of geo-referenced values for these underlying surfaces. The treatment of uncertainty in model predictions was improved over the JCP analysis phases. For instance in Phase I, spatio-temporal residual autocorrelation was not modelled explicitly but was included by using a computationally intensive non-parametric bootstrap procedure. This was carried out alongside the modelling process in Phases II and III using GEEs [GEEs; Hanley et al., 2003, see Chapter 4, section 4.3.2.1 for details] which were used to account for correlation in the model residuals, within panels.

Table 6.6: Table of main differences in the spatial density modelling between JCP phases I and II [Paxton and Thomas, 2010, Paxton et al., 2011]. If unspecified, the covariates were fitted using cubic B-splines.

<table>
<tr><th></th><th><b>JCP Phase 1</b></th><th><b>JCP Phase 2</b></th></tr>
<tr><td>Time span</td><td>1980-2009</td><td>1985-2010</td></tr>
<tr><td>Region covered</td><td>Irish Sea</td><td>Celtic Sea and continental shelf to west of Britain and the longitude of the Hebrides</td></tr>
<tr><td rowspan="2">Species Modelled</td><td colspan="2">Harbour porpoise
Minke whale
Bottlenosed dolphin
Common dolphin
Rissos dolphin</td></tr>
<tr><td>-</td><td>White beaked dolphin</td></tr>
<tr><td>Modelling Framework</td><td>Logistic GAM for presence/absence
GAM for non zero segments (Gamma errors)</td><td>GEE with Poisson errors</td></tr>
<tr><td>Covariates</td><td>Spatial coordinates (thin plate spline)
Year
Day of year (cyclic cubic spline)
Survey Mode
Depth
-
-</td><td>Spatial coordinates (CReSS)
Year
Month (discrete)
Availability
Depth
Slope
Sea surface temperature (SST)</td></tr>
<tr><td>Modelling Details</td><td>two stage model
unmodelled spatial correlation dealt with in bootstrap
GCV to choose smoothness</td><td>20 - 100 2D knots were space-filled,
50 range parameters,
Geodesic distances,
QIC_u to choose between models with different knots/range parameters.</td></tr>
<tr><td>Harbour porpoise model</td><td>0/1: s(lon, lat) + s(Year) + s (DoY) + s(Depth) + s(Survey)
Non zero: s(lon, lat) + s(Year) + s(DoY) + s(Survey)</td><td>s(Easting, Northing) + s(Year) + s(Month) + s(Depth) + s(Avail)</td></tr>
<tr><td>Minke whale model</td><td>0/1: s(lon, lat) + s(Year) + s (DoY) + s(Depth) + s(Survey)
Non zero: s(lon) + s(Depth)</td><td>s(Easting, Northing) + s(Year) + s(Avail) + s(SST)</td></tr>
</table>

6.5.2 Technical aspects

This section discusses some of the technical aspects of the modelling process and any improvements that might be made. The analysis presented here shows the use of CReSS-SALSA to be a suitable tool in mapping the distribution of cetacean species using large datasets in topographically complex areas.

Knot selection, as described earlier in this thesis, can have a large effect on the results of spatial modelling. Therefore, SALSA1D and SALSA2D were employed to allow spatially adaptive one- and two-dimensional splines. Knots provide an opportunity to push a surface up or pull it down, and so some knots will be in high density areas and some in low density areas. Unfortunately, one of the limitations of SALSA2D is that the locations of the knots cannot change year to year so these points of flexibility are fixed across time. For models with no interaction term, like that for minke whales, the location of the knots does not vary through time and the knot coefficients simply adjusts the whole surface up or down in a given year. This means that the location of knots is supported by all the data across the whole temporal range, and the limitation of SALSA2D is not an issue. However, for the harbour porpoise model, which contains an interaction term, there are coefficients for every knot-year combination. This allows some parts of the density surface to be pulled up whilst others are pulled down in a given year.

The limitation of SALSA2D is particularly noticeable in the harbour porpoise results where the survey effort is very patchy in some years. The two areas of greatest uncertainty in 2010 (west coast of Ireland and the Skagerrak/Kattegat) coincide with the location of two knots, but there was no survey effort in these regions in any season after 2008. As a result, the model struggles to support a convincing relationship between harbour porpoise density and the covariates in these areas and the uncertainty in the coefficients greatly increases. It is quite likely that the best knot locations are in different places in different years, particularly if there is an interaction effect. Therefore, more work on SALSA2D is required to allow knots to be placed only in areas with good coverage across time, or to allow their locations to be changed from year to year. However, the latter improvement is likely to be computationally expensive, and in some cases the gain in model fit may be relatively small.

The method of selecting the order in which the covariates enter the model could be improved. *Year* was constrained to enter the model first, whether it was a significant covariate or not, because the main focus of the analysis was to detect tends over time. It may have been better to add *Year* as the last covariate to see if there was any unexplained variability left in the model that could be explained by temporal trends. Alternatively, it may have been better, once all one dimensional covariates were added, to go back to the first covariate (in this case *Year*) and see if should still be included and whether the number of knots decreased given all the other covariates in the model. This could be repeated up to the last covariate in the model in a form of forwards and backwards covariate selection procedure.

A slight adaptation to this selection procedure could be to add the covariates in a specific order that relates to previous knowledge about relationships of each species with specific covariates. Covariates known to affect, or be a useful proxy for, distribution could enter the model first, followed by other covariates. For example, previous studies have shown that harbour porpoise prefer shallow water, so *Depth* could be specified to enter the model first and then the next covariate of interest. This would add more biological relevance to the selection of covariates rather than the model fit based procedure used here.

Model selection uncertainty was not included in this analysis. During the knot selection procedures many models are fitted, but only the best, based on fit statistic, is used subsequently. All of the other fitted models are discarded, regardless of how similar they were to the best model. Furthermore, covariates were kept in the model if the fit statistic improved, regardless of how little it may have improved. One way of dealing with model selection uncertainty could be to use model averaging [see Chapter 3 and Burnham and

Anderson, 2010]. This weights all of the models based upon a fit criterion and uses these to calculate weighted predictions for a candidate set of models. The candidate set could include models with or without certain covariates or models with varying knot numbers and range parameters.

One key aspect when interpreting the results of these analyses is at what spatial scale the time-averaged estimates may be reliably interpreted over. The adjusted counts (inputs to the modelling stage) were modelled with spatial smooths and so it is expected that at small spatial scales local fluctuations in the density will be smoothed over. Therefore, there will be systematic over- and under-prediction and non-zero averaged residuals on very small spatial scales. For this reason Paxton et al. [2013] undertook a preliminary analysis to assess at what spatial scale, the averaged residuals approximate zero (little systematic over- or under-prediction; Appendix 5 of Paxton et al. [2013]). They analysed the residuals for a common species, harbour porpoise, and a rare species, Rissos dolphin (*Grampus griseus*), and the results indicate that, predictions in the order of 500-1000km^2 are reliable but at smaller scales, estimates can be biased and absolute residuals relatively large (unreliable inference). The residual analyses can only take place where there is data available so a further warning should be made that in areas of little or no data, inference might be unreliable irrespective of the size of the search area. The areas of high density for harbour porpoise and minke whale and each of the protected areas discussed in the following sections are all greater than 1000km^2. However, many Scottish candidate MPAs are smaller than 500km^2 and thus, using these results at that scale is unwise.

6.5.3 Harbour porpoise distribution

The harbour porpoise results for 1994 show a preference for the central and north-western North Sea, western Scottish and Irish waters and waters to the east of Denmark. In 2005 there was a shift towards the southern North Sea and a greater concentration in the Irish and Celtic Seas down to the coast of France, confirming the findings of the original SCANS

I and II analyses [Hammond et al., 2013]. There appears to have been a general decline in relative abundance by 2010, but the confidence intervals are wide, making inference difficult. The preliminary JCP analyses both found an increasing trend in overall abundance in their respective study regions, with a peak in 2005 and then a decline. This does not match with the overall abundance results seen here but those analyses were for a much smaller study region (Irish and Celtic Seas). Investigation into this smaller region suggests a similar pattern was found in this analysis.

The spatio-temporal interaction in the harbour porpoise analysis makes it difficult to compare the maps presented here with the ones presented in [Reid et al., 2003], but does highlight the issue of pooling data over a number of years. The cetacean atlas map shows a maximum sighting rate over 20 years and therefore indicates that harbour porpoise have a widespread distribution, but with no indication of when that sightings rate was achieved. The JCP resource data pooled over 17 years would give a mean surface that was neither historic nor current distribution, but something in between. The usefulness of this kind of surface, particularly in marine conservation planning is limited. For instance, a recent change in distribution, such as seen here for harbour porpoise, could be masked by historic patterns and lead to the wrong areas being considered important conservation areas.

The three plots corresponding to the three EU Habitats Directive reporting periods (section F.1, Appendix F) indicate three main areas of high density for harbour porpoise that change in importance over time: the west coast of Scotland, the Welsh coast and an area off the coast of East Anglia. Further, these hotspots in the point estimates are also evident in the lower confidence limits for the surface for two of the three areas (west coast of Scotland and the coast of Wales). During reporting period two, the area of high density extends from the west coast of Scotland down through the Irish sea to Lands End, and there is a second localised high density area off the coast of East Anglia. This second hotspot persists in reporting period three but stretches further north into the Thames and Humber shipping forecast areas (see Appendix G for a key to forecast areas). Other hotspots are seen off the west coast of Scotland and west Wales (similar to reporting period one). The area off the coast of Wales was not included in SCANS-I and was identified as a medium density area in SCANS-II. However, both the cetacean atlas [Reid et al., 2003] and the ESAS analysis [Northridge et al., 1995a] show sightings in this area. Furthermore, Baines and Evans [2009] analysed aerial, vessel and vantage point data collected from the Irish Sea and Welsh coast between 1990-2000, and showed high sightings rates of harbour porpoise in this area.

Northridge et al. [1995a] describe a seasonal distributional difference in harbour porpoise sightings rates that suggests high rates occur in the North Sea in winter and spring and a shift to coastal regions of Scotland in summer/autumn. The seasonal pattern for harbour porpoise seen in the analysis reported here is not consistent with this finding and suggests that more harbour porpoises are seen in north-western European waters in the winter/spring than in the summer. The highest densities were recorded in winter, off the west coast of Scotland, in the Moray Firth and off East Anglia, however these estimates are based on small amounts of survey effort. A further reason for this inconsistency might lie in the differing coverage of water depths surveyed across seasons. For example, the mean depth of the data segments in the winter surveys was 28m, which is close to the preferred water depth found in this analysis (Figure 6.4(c)), whereas the mean water depth for the summer segments was 49m. This could have artificially inflated the density of harbour porpoises seen in winter because most of the areas surveyed were in habitat preferred by this species. Generally, the highest densities were found off the west coast of Scotland in all seasons.

Seabed depth has been shown to be important in explaining harbour porpoise distribution in several studies [Booth, 2010, Embling et al., 2010, Embling, 2007, Marubini et al., 2009, Hammond et al., 2013]. However, harbour porpoises have been found frequently in both shallow water [30-60m; Shucksmith et al., 2009, Todd et al., 2009] and deeper shelf waters [> 100m Raum-Suryan and Harvey, 1998, Booth, 2010]. This study suggested the strongest preference for shallow waters with another small peak around 150m and although consistent with what was previously known for this species, the covariate relationships seen in this study should be used with caution. Depth alone is unlikely to drive distribution and was primarily used here as a proxy for un-measured/un-measureable environmental covariates such as prey distribution. It is likely that the distribution of prey or some lower trophic level is determined by depth and this manifests itself in the distribution of porpoises. Recently, Jansen et al. [2012] analysed strandings along the Dutch coast and suggest that porpoises feed offshore on pelagic, schooling species (e.g., poor cod, mackerel, greater sandeel, and sprat) and closer to shore on more benthic and demersal species (e.g., gobies, whiting, herring, and cod). Harbour porpoise are one of the smallest marine mammal predators, with limited energy storage capacity, and it is therefore assumed that they must feed frequently. Therefore, their distribution is likely to be closely linked to that of their prey [Fontaine et al., 2007, Read, 1999].

6.5.4 Minke whale distribution

The maps for minke whale provided by this analysis are similar to the maps of relative abundance in Reid et al. [2003] and Northridge et al. [1995a]. Minke whales are most abundant in the western part of the North Sea (Tyne and Dogger shipping forecast areas; see Appendix G for a key to forecast areas) and west coast of Scotland, with the highest densities in the middle of the North Sea. There is also an area of high density off the south coast of Ireland in both these analyses. SCANS-I and -II [Hammond et al., 2002, 2013] also recorded higher densities in western areas of the North Sea, and SCANS-II recorded high densities off the south coast of Ireland.

The lack of a significant interaction term in the model between space and year indicated that there was no shift in minke whale distribution over time, although there is a suggestion in the raw data (estimated counts), of a shift in distribution southward (Figures in section F.2 Appendix F). This led to low predicted numbers in the Firth of Forth, which was the area of highest density in SCANS-I, and a large number of sightings in the JCP data resource. A model with an interaction term predicted high densities in this area, but the confidence intervals were high over the whole surface making interpretation impossible. Further efforts in this area might result in a lower dimensional interaction term and may produce a fitted surface which more closely resembles the movement of this species over time.

The highest JCP-based density estimates for minke whales were in the summer months, particularly around the west coast of Scotland. Although the data used by Northridge et al. [1995a] included rather few sightings of minke whales from the west coast of Scotland, their maps also suggest that high densities occur there in the spring and summer months.

In terms of total relative abundance over the whole survey area, there is some evidence of a slight increase in minke whale numbers from 2005 to 2008, followed by a decline. However, the confidence intervals on total relative abundance are imprecise so as a result we cannot conclude that there was a significant change in abundance over the 17 year study period.

Depth was not a significant predictor of minke whale density in the SCANS I or JCP Phase II analyses. However, Reid et al. [2003] suggests that minke whale are found predominantly in waters of 200m or less. This conclusion is supported by the analysis in this chapter, although, as with harbour porpoise, the covariate relationships seen in this study should be used with caution. Anderwald et al. [2012] showed there was seasonality associated with depth preference, at least in waters off the west coast of Scotland. This may explain why it is not always selected as a covariate. Furthermore, Anderwald et al. [2012] show that in June minke whales occur in areas where there is a high probability of sandeel occurrence, but do not do so later in the season. Sandeels prefer depths of 30-70m [Wright et al., 2000] and this may account for the peak in minke whale density at this depth range found in this analysis. Later in the season, minke whales on the west coast of Scotland feed on sprat [Anderwald et al., 2012] and pre-spawning herring [Macleod et al., 2004] at deeper depths. Minke whales are a highly mobile species with a highly variable diet. Their preferred prey differs across the north east Atlantic [Haug et al., 1995] and they readily switch diets when the availability of any preferred prey species is low [Macleod et al., 2004, Anderwald et al., 2012]. An interaction term between season/year and space/depth would account for such a temporally and spatially varying distribution and should be included in future analyses of these data.

6.5.5 Identification of Candidate SACs

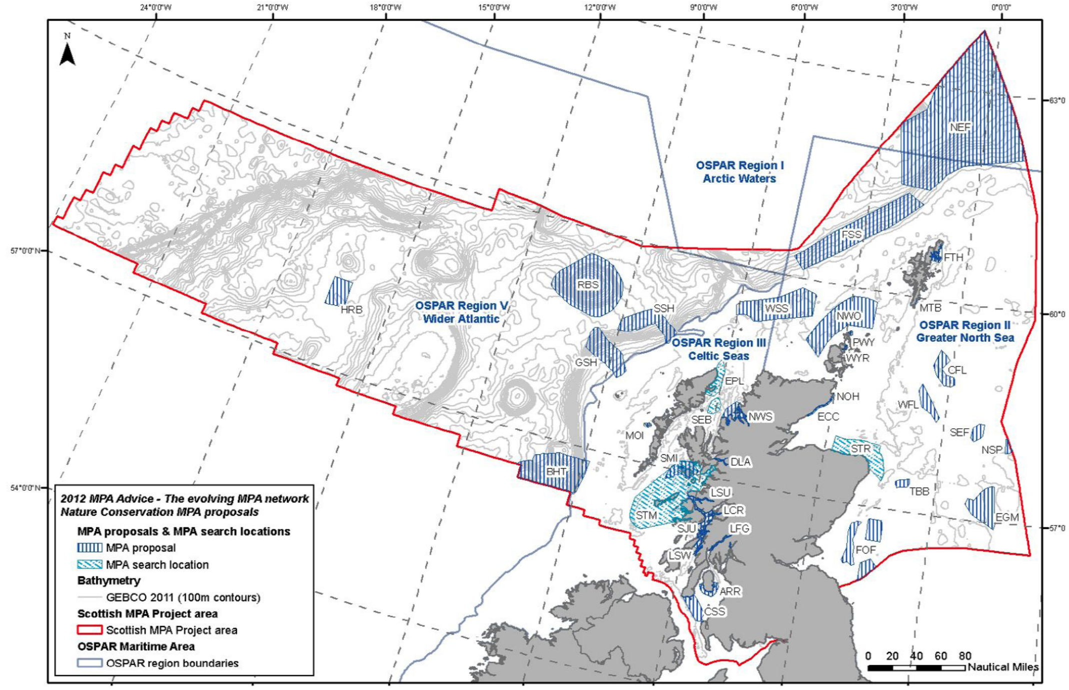
Member states are required to designate Special Areas of Conservation (SAC) for all species that are listed on Annex II of the EU Habitats Directive, which includes harbour porpoise. The guidelines for designation of an SAC state that it must be an area of persistent presence or a high population density (relative to other local areas) or where there is a high ratio of young to adults at certain times of year [Pinn, 2009]. Minke Whales and harbour porpoise are both recognised as a priority species within the UK Biodiversity Action Plan (UK BAP; as of July 2012, now the UK Post-2010 Biodiversity Framework) and are on Annex IV of the Habitats Directive for species in need of strict protection. In this section I describe how the results of this analysis can be used to identify candidate SACs for harbour porpoise, and whilst not a requirement for minke whale I asses if any current or potential SACs would indirectly benefit this species.

At the time of writing, the only marine protected areas for whales dolphins and porpoise around the UK are the Moray Firth SAC, Cardigan Bay SAC and Lleyn Peninsula and Sarnau SAC [Hoyt, 2012]. The Cardigan Bay SAC (off the coast of Wales) (www.cardiganbaysac.org) was designated primarily to protect bottlenose dolphins (*Tursiops truncates*), Atlantic grey seal (*Halichoerus grypus*), river and sea lamphrey (*Petromyzontidae spp.*), reefs, sandbanks and sea caves. The Lleyn Peninsula and Sarnau SAC was designated to protect a variety of habitat types, not the species living within them. However, marine species that indirectly benefit from this area are bottlenose dolphin, otter (*Lutra lutra*) and grey seal. The Moray Firth SAC was designated to protect bottlenose dolphin.

There are currently no sites in UK waters designated primarily to conserve harbour porpoise or minke whale (http://jncc.defra.gov.uk/ProtectedSites/SACselection). The UK government has been told by the European Commission that it must propose some harbour porpoise SACs or face a fine for failing to do so.

Based on the results of this analysis, none of the existing SACs for other marine species or habitats would adequately conserve harbour porpoise or minke whale. The main areas where both lower and upper confidence intervals show high densities of harbour porpoise in summer 2010 are the west coast of Scotland, the Moray Firth and north of the Norfolk coast. However, a species must show persistence in the area considered for SAC status. For harbour porpoise, the west coast of Scotland, the west coast of Wales and the coast of Norfolk, extending northward (section F.1, Appendix F) showed consistently high densities over the period 2007-2010. The Scottish Government is currently considering designating a number of locations as MPAs (Figure 6.21) for species including minke whale but excluding species on Annex II of the Habitat's Directive, which includes harbour porpoise. The most suitable of these MPAs, that might indirectly conserve harbour porpoise, would seem to be the whole of the Skye to Mull (STM) search location and the channel to the north of Skye, which is not a proposed conservation area. The Moray Firth SAC does not extend far enough eastward to capture most of the area used by harbour porpoise in 2010. The highest densities of harbour porpoise off the coast of Wales occurred further south than the boundaries of the two Welsh SACs (section F.1, Appendix F), so the species could be protected either by an extension of the Cardigan Bay SAC or the creation of a new marine conservation zone (MPA equivalent for Wales).

The areas with the highest lower and upper confidence limits for minke whale density in summer 2010 are the west coast of Scotland, central western North Sea (off the coast of Yorkshire) and a small area to the west of the Isle of Man. Because there is no spatial interaction with year in the model, the same areas are identified in all years (section F.2, Appendix F). None of the proposed marine protected areas off the west coast of Scotland



Map projected in Europe Albers Equal Area Conic (Modified Standard Parallels - Standard Parallel 1 = 50.2; Standard Parallel 2 = 58.5). The exact limits of the UK Continental Shelf are set out in orders made under section 1(7) of the Continental Shelf Act 1964 (©Crown Copyright). Coastline ©Crown copyright and database right [2012]. All rights reserved. Ordnance Survey Licence number 100017908. Bathymetry ©GEBCO. NOT TO BE USED FOR NAVIGATION. MPA network ©SNH, JNCC and Marine Scotland. 07.12.12. All rights reserved.

Figure 6.21: Proposed MPA sites and search locations for Scottish territorial waters as of 2012. Image taken from a Marine Scotland report [Marine-Scotland, 2011].

are adequate for minke whale despite this species being included in the Scottish government search criteria. Minke whale are included in the STM and Southern TRench (STR) search locations but again neither are adequate based on this analysis. However, an area similar to that described for harbour porpoise, extending north-eastward from the Skye to Mull search area would encompass the highest density area for minke whale. At the time of writing, there are no proposed protected sites for cetaceans off the Yorkshire coast or around the Isle of Man.

All of the proposed sites for SACs in Scottish and Welsh waters are quite small, given the transient nature of cetaceans and their wide geographic range. Perhaps we should be considering sites that are much larger and appropriate on a global scale. One large conservation area in each place seems more realistic for large transient species that do not seem to have a particular affinity for any one location in these areas.

6.5.6 Future analyses of the JCP data resource

The JCP data resource is an excellent collaborative database and potentially a useful tool for assessing temporal and spatial distributions of cetaceans. However, the quality of any analysis lies in the data which underpins it and, for this analysis, the lack of effort in certain locations and at certain times of year is a major issue. It would be of great benefit if data from other countries could be used to eliminate the coverage gaps in this dataset. For example, surveys are known to have taken place around Germany and Denmark in recent years, and the results of these could improve the models in these areas. Furthermore, there are several types of data that were not included, such as tagging, acoustic and vantage point (static points such as cliff tops or offshore platforms) data. However, inclusion of these is difficult, because of the way each type of data is collected. For vantage point data it is hard to disentangle the effect of declining detectability of animals away from the vantage point and the true distribution of animals in space. Traditional distance sampling methods cannot be used here. However if something is known about habitat preference (for example, from an independent survey) then a new `R` package `nupoint` [Cox et al., 2013] could be used to infer something about the true distribution of animals and allow the inclusion of this type of data into the JCP resource.

This analysis established trends on a large scale. A next step is to identify smaller areas for analysis, for example Scottish waters or specific developer areas. As mentioned earlier, this allows a greater variety of covariates for modelling, which might describe the data better and reduce uncertainty. It might also improve the identification of hotspots. The data could also be re-analysed by dividing the whole study region into many slightly overlapping areas, each of which could be modelled separately. This would allow more or alternative covariates to be used. For example, different environmental covariates could be used to describe animal distributions in off-shore and coastal areas. The results of these separate analyses could then be combined smoothly to create a comprehensive atlas. The following chapter (Chapter 7) describes a method that could be used to achieve this aim.

Data from a number of other cetacean species were analysed as part of the JCP project. The results of these analyses could be compared in order to identify areas where the distributions of different species overlap. For example, data from harbour porpoise and bottlenose dolphins, which are known to attack harbour porpoise [MacLeod et al., 2007b, Ross and Wilson, 1996], could be analysed together to determine the probability of interactions between them. Similarly, data of the distribution of fishing boat track data could be overlaid with the different species maps to identify areas of high interaction between the fishing industry and cetaceans. Since 2005, all vessels in excess of 15m are legally required (under EU legislation 404/2011) to automatically transmit vessel identification, date, time, position, course and speed either hourly or 2 hourly as part of the Vessel Monitoring Scheme (VMS). In the UK, this is coordinated by the Marine Management Organisation (England and Wales), Marine Scotland and the Department of Agriculture and Rural Development (Northern Ireland). Unfortunately, these data are considered personal data and falls under the Data Protection Act, which makes access difficult. Herr et al. [2009] were able to use VMS data from the German Bight region of the North Sea to investigate association and overlap between fisheries and harbour porpoise. They concluded that especially in the summer there was evidence of resource competition with the sandeel fishery and some evidence in spring of overlap with plaice and sole fisheries. The JCP analysis provides results over a much larger area including UK waters, and with access granted to use depersonalised UK VMS data would result in an interesting study of the interactions of UK fishing vessels and cetaceans.

Chapter 7

Combining Density Surfaces

7.1 Introduction

Cetacean species distribution maps are used in various applications such as the identification of proposed Marine Protected Areas (MPAs), environmental impact assessment, monitoring, and the planning of military activities, seismic surveys and offshore renewable energy developments. All of these applications require maps that are comprehensive, up-to-date and include realistic estimates of uncertainty. Often, individual results from existing surveys do not cover the region of interest or are too small to be useful. However, there are many independent surveys of the marine environment taking place all over the world, many of which are documented in the OBIS-SEAMAP database [Read et al., 2011], but each of which has been analysed separately resulting in multiple maps. A number of these maps when looked at together could have enough information to be useful for the applications above, avoiding the need for new, dedicated surveys, which are both expensive and time consuming to implement. Even if a new survey is conducted for a specific area, its utility may be improved if supplemented with other existing mappings of the region.

Even considering overlapping survey data alone, just combining the raw survey data is insufficient, due to different collection techniques and the need to consider uncertainty if inference is to be drawn. Therefore a method is needed that combines the density and precision estimates from multiple analyses to give a single density surface with an aggregate measure of precision. In general, seeking to combine a variety of overlapping density information (estimates and uncertainties) for the purposes of prediction and inference, requires a rigorous statistical approach. The need for such a method is commonplace and some specific examples follow.

The analysis of data from the Joint Cetacean Protocol data resource, described in the previous chapter, showed how difficult it can be to model raw survey data from a variety of sources in a single analysis. If an approach were available, each data set could be modelled separately and the resulting surfaces could be combined into one unifying map to create a species distribution atlas.

For military planning a useful starting point is the database described by Harris [2013], in which survey data are used in combination with habitat suitability to predict the density of marine mammals on a global scale. That database is based on one by Kaschner et al. [2006] who employed a fundamental ecological niche model (described in the General Introduction) based on environmental covariates, such as sea bed depth, sea surface temperature and ice edge association, to obtain global Relative Environmental Suitability (RES) indices. This kind of data is particularly useful for areas/species where there is limited, or no systematic data.

For other applications, such as identifying hotspots or areas for future research, such an index may be suitable by itself. Often though, absolute numbers of animals, rather than relative indices, are required to determine population level consequences of, for example, anthropogenic activities. Thus, Harris [2013] combined the RES index from Kaschner et al. [2006] with information from observed densities from dedicated surveys to produce global marine mammal density estimates. If such databases are to be useful in risk assessment analysis of military exercises, for example, then they must not become fixed at the time point they were created, but must be updated when new information becomes available.

New information could come from recent surveys or the output of a dynamic predictive model [for example Read et al., 2009, Barlow et al., 2009]) using current environmental conditions as inputs. New information may cover too small a geographic area to be useful and may be subject to sampling variability specific to the temporal period of collection. In these cases, combining new densities with the database creates a current database which takes account of the long-term average distribution of animals.

Combined maps can also be used in spatial conservation planning decision support software, such as Marxan [Ball et al., 2009, Ardron et al., 2010], which requires good data coverage. This software cannot distinguish between zeros (locations where no animals were observed) and areas where there was no survey effort; it also shows bias toward data rich areas (Lieberknecht cited in Ardron et al., 2010). If data from multiple surveys are available from regions where reliable distribution maps are required, the most effective way to provide this information is to combine all available survey data for each species of interest and fit a model to 'fill in the gaps' where there is no survey effort. Packages such as Marxan could then be used to generate, evaluate and compare different MPA options using these composite maps.

Williams et al. [2011] recently compiled information on cetacean distribution in the north-eastern Pacific Ocean as part of the process of developing proposals for cetacean-oriented MPAs. The authors highlighted problems in applying Marxan and the need for good data coverage. Their suggested solutions included conducting new, wide ranging surveys, use of RES data only [RES; Kaschner et al., 2006], or extrapolation of densities from RES to un-surveyed areas using relationships between RES and survey data. New surveys are expensive and time consuming, and any kind of extrapolation is a potentially dangerous process, as seen in the killer whale case study in Chapter 4.

It could be argued that only the most recent information available should be considered rather than investing in long-term historic data such as the database by Harris [2013]. There will be situations, of course, where combining density surfaces is deemed not appropriate, for example where there is a large discrepancy between the existing data and the new survey data, and the time lapse between the two is extensive. It could also be argued that the historic data has little relevance to current-day predictions and ought to be abandoned in favour of a very recent survey noting, however, that there is a loss of information about temporal variability, which might be important for some applications.

However, each survey is a snapshot in time and is susceptible to being an unusual estimate due to sampling variability. Long-term historic averages might be more robust to sampling variability but require updating with new information to be considered current.

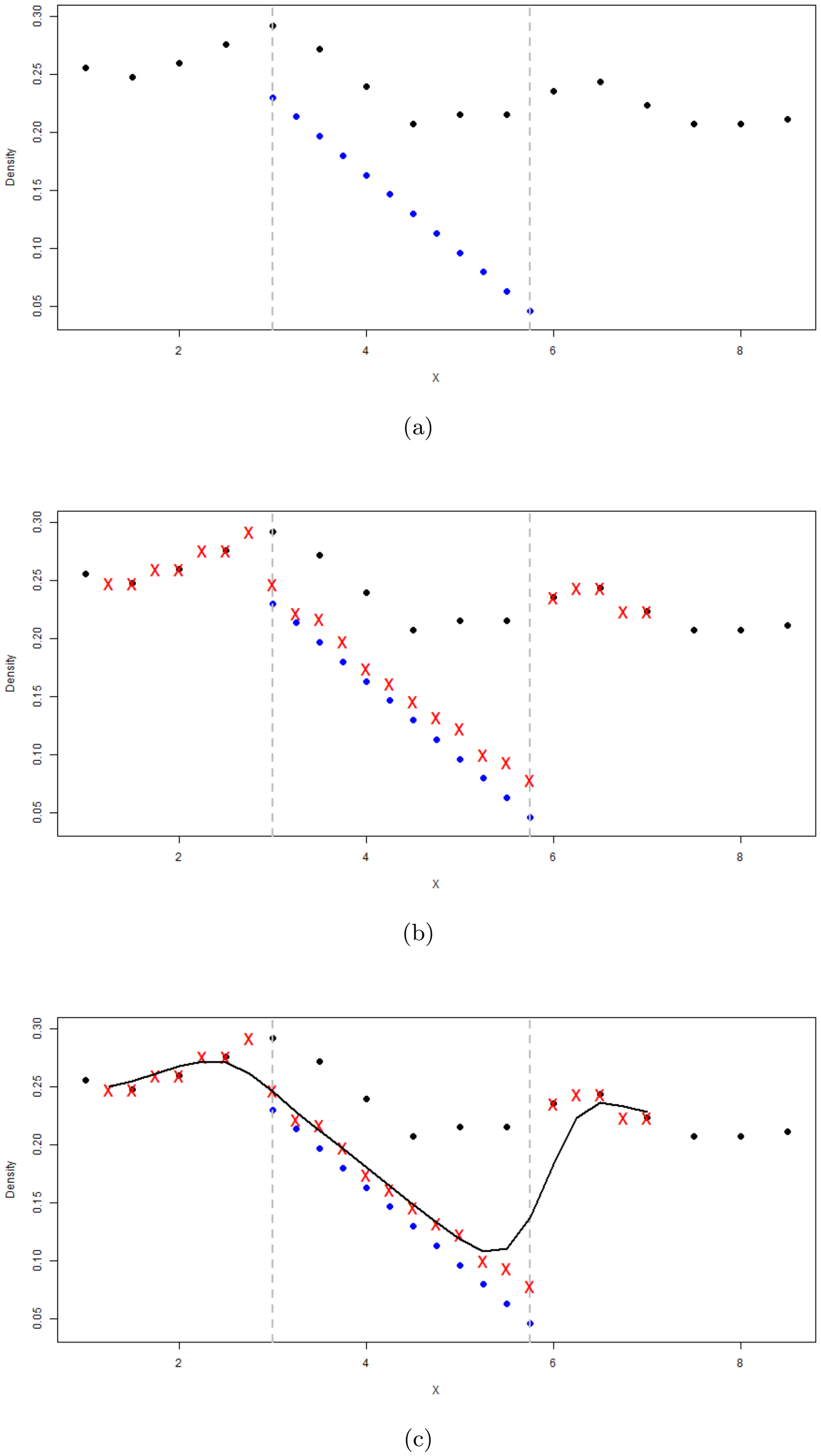
Whilst the case for temporally combining survey data with historical data may be unclear at times, there is most definitely a good spatial argument. For example, if you have survey A and survey B and you need to consider an area covered partly by A alone, B alone and A & B combined. In this case, one or other survey cannot be used alone so it must be decided how best to merge the individual surveys in a statistically robust manner.

Reflecting the issues described above, this chapter describes a method for combining existing density surfaces with new data and provides an example of its use as part of an algorithm called the Dynamic Cetacean Abundance Predictor (DCAP). Section 7.2 describes the methodology for combining competing density estimates (Section 7.2.1) and a smoothing method to smooth the potential transition from a combined surface to the non-overlapping regions (Section 7.2.2). These methods were originally developed for a military planning application, and it is this use that is described here (Section 7.3). The DCAP implementation is described (Section 7.3.1) and the test scenarios used to validate it are presented (Section 7.3.2). The results of the DCAP algorithm are presented and discussed (Section 7.3.3), followed by a discussion of the methods and results (Section 7.4). The work in this chapter formed part of the Marine Mammal Alert, Awareness and Response System (MMAARS) project, conducted by BAE Systems for the US Navy, which was awarded a silver medal in the 2009 BAE Chairman's Awards. I was the principle developer behind the statistical methodologies of the DCAP algorithm part of the MMAARS project.

7.2 General Methods

This section describes a new method for combining two density surfaces both of which have an associated measure of uncertainty. There are two broad problems to solve. The first is to combine the estimates from the two data sources and the second is to smooth the junctions where the two surfaces meet. The two data sets could be the results from two independent surveys or a database, such as the one based on RES values, and a survey. Figure 7.1(a) shows an example involving a coarse resolution density map based on historical data and a fine resolution density map based on survey data that overlap. Although the process described later is for a two-dimensional density surface, for clarity this figure shows the problem in a single dimension. Furthermore, for ease of explanation, the values contained in one density surface are referred to as 'existing data' and the other as 'new data' (representing new data to be added to the current situation, which may be model outputs from survey data or a dynamic predictive model).

Section 7.2.1 describes the process of combining the two data sources (Figure 7.1(b)) and Section 7.2.2 describes the method for smoothing the edges where the two data sources meet. This additional smoothing step is required to prevent un-naturally sharp changes in density between the combined surface and those parts of the existing surface that have not been changed (Figure 7.1(c)).



(a)

(b)

(c)

Figure 7.1: A one dimensional example of two overlapping density surfaces for updating; (a) the existing surface (black) and the new survey results (blue). (b) The mean (weighted by uncertainty) of the two sets of data points (red crosses), shown at the resolution of the new survey data and extended by a buffer into the area covered by existing data. (c) a kernel smooth (black line) with varying bandwidth of the red cross data. The y-axis is density and the x-axis is location.

7.2.1 Combining Competing Estimates at a Point

The aim of this step is to combine two density surfaces; one deemed existing data and the other new data. More specifically, there are two expected values (two estimates of density) to be combined, each with an associated degree of precision. A Bayesian approach is used to combine the estimates considering that there is some existing hypothesis, which is updated through the acquisition of additional information. Here it is assumed that the existing view is our prior (hypothesis) and the new data provides some evidence to modify this, which results in a posterior density (combined existing view and new density data). The Coefficient of Variation (CoV; note this is different to the commonly used CV due to its use in this thesis for Cross Validation) for the prior distribution and the new indicates how informative each source of information is, and thus how much weight each should be given when being combined.

This approach resembles the Best Combined Spatial Predictor (BCSP), which is regularly used in geostatistical analyses Hengl [2009]. For BCSP the two data sets are assumed to be independent and normally distributed. However, in the example presented here the data are assumed to be lognormal and the measure of precision is the CoV, but standard deviation or variance could also be used.

Let y be the new data estimate of density with coefficient of variation CoV_y, and assume that y is lognormally distributed:

$$\log(y) \sim N(\theta; \sigma^2_{\log(y)})$$

Where,

$$\sigma^2_{\log(y)} = \log(\mathrm{CoV}^2_y + 1)$$

The lognormal distribution is commonly used to describe data, such as those obtained from surveys, with a low mean, high variance and values that cannot be negative [Limpert et al., 2001]. Assume a prior distribution for θ, the mean log density, such that,

$$\theta \sim N(\mu_\theta; \sigma_\theta^2)$$

To use this prior we need to specify μ_θ and σ_θ^2 using the information we have in our existing information source: i.e the current estimate of density (μ_0) and its coefficient of variation (CoV_0). Since θ is normal and e^θ is lognormal we can use these estimates to parameterise the distribution of e^θ:

$$e^\theta \sim \log N(\mu_0; (\mathrm{CoV}_0 \times \mu_0)^2) \tag{7.1}$$

Where,

$$\sigma_\theta^2 = \log(\mathrm{CoV}_0^2 + 1)$$

and the mean,

$$\mu_\theta = \log(\mu_0) - \frac{1}{2}\log(\mathrm{CoV}_0^2 + 1)$$

Parameters μ_0 and CoV_0^2 are the densities and squared CoVs in the existing density surface. On the log scale the combination of $\log(y)$ and μ_θ (the new survey density and the existing density on the log scale) equates to a weighted average, where the weightings are given by the level of uncertainty.

The posterior distribution for θ is:

$$\theta | y \sim N(\mu_1; \sigma_1^2)$$

where the posterior mean (μ_1) and variance (σ_1^2) for a Normal distribution with known variance [Murphy, 2007] are defined as

$$\mu_1 = \frac{\sigma_{\log(y)}^{-2} \times \log(y) + \sigma_{\theta}^{-2} \times \mu_\theta}{\sigma_{\log(y)}^{-2} + \sigma_{\theta}^{-2}}$$

$$\sigma_1^2 = \frac{1}{\sigma_{\log(y)}^{-2} + \sigma_{\theta}^{-2}}$$

These are re-expressible as the mean and CoV of θ (new mean and the CoV of the lognormal distribution in Equation 7.1) and thus the density and variance for the combined data:

$$\text{combined density} = e^{(\mu_1 + \frac{\sigma_1^2}{2})}$$

$$\text{combined variance} = (e^{(\mu_1 + \frac{\sigma_1^2}{2})})e^{2\mu_1 + \sigma_1^2}$$

To get back to CoV, take the square root of the combined variance divided by the combined density. These combined estimates become the prior for e^θ (Equation 7.1) for any further modifications from the addition of new information.

Notably, after the densities have been combined there may be an unnaturally sharp transition between the combined density surface and the existing density surface (Figure 7.1(b)). The next section describes a method to alleviate this effect.

7.2.2 Smoothing the Transitions between Datasets

The one-dimensional example in Figure 7.1(b) shows that once the overlapping densities have been combined there may be an unnaturally sharp transition from the new, combined density surface to those parts of the existing surface that have not been updated. This section describes a method to smooth this transition, the results of which can be seen in Figure 7.1(c).

The densities in the area to be smoothed are referred to as y, where some of the y are products of multiple data sources and some may not be. There will always be a potential for a discontinuity when moving from a surface generated by combined data to one based on data from only one source. The aim is to reduce these discontinuities by some sort of local averaging. The data are two dimensional, and denoted by $\{x_{1i}, x_{2i}, y_i; i = 1, ..., n\}$ to which kernel regression smoothing (Section 2.3.4) was applied to smooth the y's. This provides a smoothing parameter, which can be adjusted to decline to zero at some distance from the edge of the combined density region (i.e. to reduce the smooth to an interpolator).

Using the two dimensional kernel smoothing method in Section 2.3.4, the local estimator is $(\mathbf{X^T W X})^{-\mathbf{1}}\mathbf{X^T W y}$, where $\mathbf{X}$ is an $n \times 2$ matrix of spatial coordinates. Matrix $\mathbf{W}$ $(n \times n)$ is the product of two Gaussian kernels, one in each x dimension, of the form

$$w(x_i - x_1; h) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - x_1)^2}{2h^2}} \tag{7.2}$$

where h denotes the smoothing parameter, controlling the width of the kernel function and for simplicity will be considered separately for each x dimension. When h represents the standard deviation of normal densities, then observations within $3h$ of the evaluation point in the covariate axis will contribute to the estimate, and coordinates out with this range have effectively no contribution. Every coordinate at which there is a $\hat{y}$ to be estimated requires an input of an h; in two dimensions this requires specification of an h_1 and h_2 corresponding to x_1 and x_2. We used an Auto-Correlation Function (acf) to find h_1 and h_2 for each coordinate at which there was a $\hat{y}$ to be estimated. The principle is that if the surface is highly variable, the correlation between points is small and the smoothing parameters err on the side of interpolation. Conversely, if the combined densities are at a consistent level, correlation is high and the smoothing will result in a smooth surface. The acf function was calculated as follows:

$$AutoCorr = \min\left(\frac{\sum_{i=1}^{N-k}(y_i - \bar{y})(y_{i+k} - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}\right)$$

where y_i is the density at point i, N is the number of points in the row/column (N.B. data are on a regular grid), $\bar{y}$ is the mean density along that row/column and k is the lag. A range of lags were used where, for example, a lag of 4 calculates the correlation between points 4 steps apart. Since the grid is regular, the lag is proportional to distance so the value of interest k, which minimises the *acf* (i.e. the lag which is required for data points to be uncorrelated), is proportional to the distance at which data points are uncorrelated. The area of combined densities and the single data source are said to be independent at the point where the data points are uncorrelated, and no further smoothing is required. To find the smoothing parameter, h, the number of points that minimises the *acf*, k_{min}, is converted to distance on the covariate axis using the resolution of the data, g_r, and then divided by 3 so that h equates to the standard deviation of the normal density:

$$h = \frac{k_{min} g_r}{3} \tag{7.3}$$

Points $3h$ apart are, therefore, deemed to be independent because it is at this range that points on a normal distribution are considered to be. The *acf* can be used to determine the size of the area outside the boundary of the new survey which will be smoothed to alleviate unnaturally sharp changes in density. This region will be referred to here as the 'buffer zone' and the values of the lag (distance) at the edges of the survey region are used to determine its size. The smoothing parameter in this area decreases from the value of h at the edge of the survey area to effectively zero at the edge of the buffer zone, where the smoothed combined density surface joins the existing density surface. Returning to the one-dimensional example from earlier, Figure 7.1(c) shows the combined densities in the region of the new data and a smooth of these data into the existing data (to the edge of the buffer) to smooth the change in densities between the combined data and existing data. The smoothed estimates, within the buffer/survey region, form the aggregate estimate, which may be subject to modification in the same way with the addition of new information.

This section has described two generic methods for combining and smoothing two sources of spatial density information, one considered to be a map of existing density estimates and associated precision and one considered to be new data. The next section describes a specific implementation of these methods designed to aid in the planning of naval exercises, where potential impacts on local fauna are of interest.

7.3 Example

The example described here formed part of Phase 1B of the Marine Mammal Alert Awareness and Response System (MMAARS) project developed by BAE Systems for the US Navy to help in the mitigation of the potential risk to marine mammals from military sonar [Donovan et al., 2010]. MMAARS uses a risk assessment algorithm, developed previously in conjunction with BAE Systems, which requires up-to-date density surfaces for a large number of cetacean species. These surfaces are derived from a global database of density estimates for 115 marine mammal species developed by Kaschner et al. [2006]. Donovan et al. [2011] extended the work carried out by Kaschner et al. [2006] by establishing a relationship between Kaschner's RES index and observed densities from dedicated surveys. The structure of the data follows specifications defined by the United Kingdom Hydrographic Offices (UKHO) Integrated Water Column product, which has a spatial resolution of 0.5^o grid cells, with Latitude and Longitude fields representing the centre of each grid cell. Each cell contains a density estimate and an associated uncertainty value. Most of the species have just one density map applicable for the entire year, but quarterly estimates are available for 46 species. These 0.5^o density surfaces constitute the existing data information source of the method in section 7.2.1. In the example described here, these are to be combined with simulated data representing the output of an analysis of new survey data.

7.3.1 Implementation: DCAP

DCAP (v0.1.0) is a set of functions written in R [R Development Core Team, 2009], that implement the methods for combining and smoothing density data described in sections 7.2.1 and 7.2.2. This section provides details of their specific implementation in DCAP.

Other than the existing density surface, two types of input to the DCAP software are possible. The user may choose to use the output from a predictive model that provides improved spatio-temporal resolution over specific geographical regions and time periods. The alternative is to use the results from recent surveys of a particular geographic area of interest.

There are three steps to the DCAP algorithm: “initialise” checks the data inputs; “combining” merges the existing density surface with the new survey results; and “smoothing” alleviates any discontinuities between the new and existing density surfaces. An overview of the DCAP algorithm is given in Figure 7.2 and the R code to implement the main features can be found in Appendix H.

```
DCAP:
    Initialise
        Data and constraints check
        Transfer of data to a grid
        Output Figures showing inputs
    Combining Estimates
        Combine existing density surface and new survey data by weighted average
        using coefficients of variation
        Output combined densities and associated coefficients of variation
            Figures showing combined density surface
    Smoothing Edges
        Perform kernel smooth of combined density region and buffer zone
        Output Smoothed densities
            Figures showing buffer zone and combined density surface smoothed
            into the existing density surface
```

Figure 7.2: Overview of the structure of DCAP.

Initialise Step

For the combining and smoothing steps to function correctly, the data must be entered in the correct style. The initialise step checks that the data are described correctly, identifies the type of data present and transfers all existing and new density data to a common, integer-based grid system. This transfer is described in Appendix I. The new survey data may be stratified (single density for a geographic region(s)) or in the form of a density surface, which may be any shape and may include islands. Each grid point must have only one density and one uncertainty value.

Combining Step

This step uses the method for combining density estimates from Section 7.2.1. In the DCAP implementation of the methods the existing data are based upon the RES database. The new data to be combined are either from the results of survey analysis or a dynamic predictive model, the key elements being density estimates and an associated level of precision. The outputs could be used directly, within the region of the merge, or smoothed into the existing data to form an updated database, described next.

Smoothing Step

This step uses the smoothing method described in Section 7.2.2. The time taken to complete this step is important if DCAP is to be used for real time planning of a navy exercise, so some changes to the overall approach should be made to improve efficiency. Firstly, the number of points used to estimate density can be reduced. This is possible since much of matrix $\mathbf{W}$ is sparse, due to the relatively small size of h. We can therefore restrict the number of points used to estimate the density, $\hat{y}_i$, at coordinate $\mathbf{x} = (x_{1,1}, x_{2,1})$. Figure 7.3 shows the weights, $\mathbf{W}$, of one coordinate. The box depicts an area $3h$ x $3h$ within which points have influence on the estimate of $\hat{y}_i$ (the density at the centre of the grid cell, which is the quantity we are interested in).

Another improvement in computational efficiency can be made in the multiple calculation of the *acf*. Rather than computing this for all rows and columns in the data, which

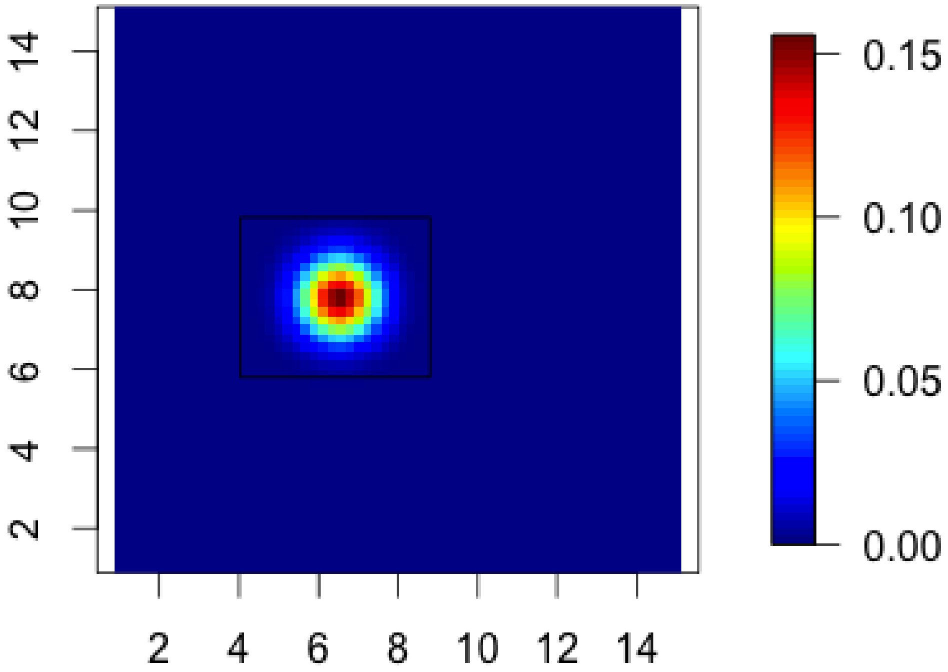


Figure 7.3: Visual depiction of the weights of one coordinate. The box depicts an area $3h$ x $3h$ within which points have influence on the estimate of $\hat{y}_i$ (the coordinate at the centre of the box)

increase in length as the area/resolution of the survey increases, the calculation can be restricted to a subset of the cells and the rest extrapolated. Firstly, the combined data area is made rectangular by including data points from the existing density surface, if needed. Then, with a subset of 10 rows and 10 columns of data (Figure 7.4), the *acf* is used to return an autocorrelation value for each row/column. The *acf* for the first and last row/column, which form the edges of the survey area, is used to calculate the size of the buffer zone. In Figure 7.4 the boundary of the buffer zone (shown in red) is parametrised using r, l, b and t for the right, left, bottom and top dimensions respectively. Parameter r is calculated from the tenth column ($h1_{10}$), l from the first column ($h1_1$), b from the first row ($h2_1$) and t from the tenth row ($h2_{10}$).

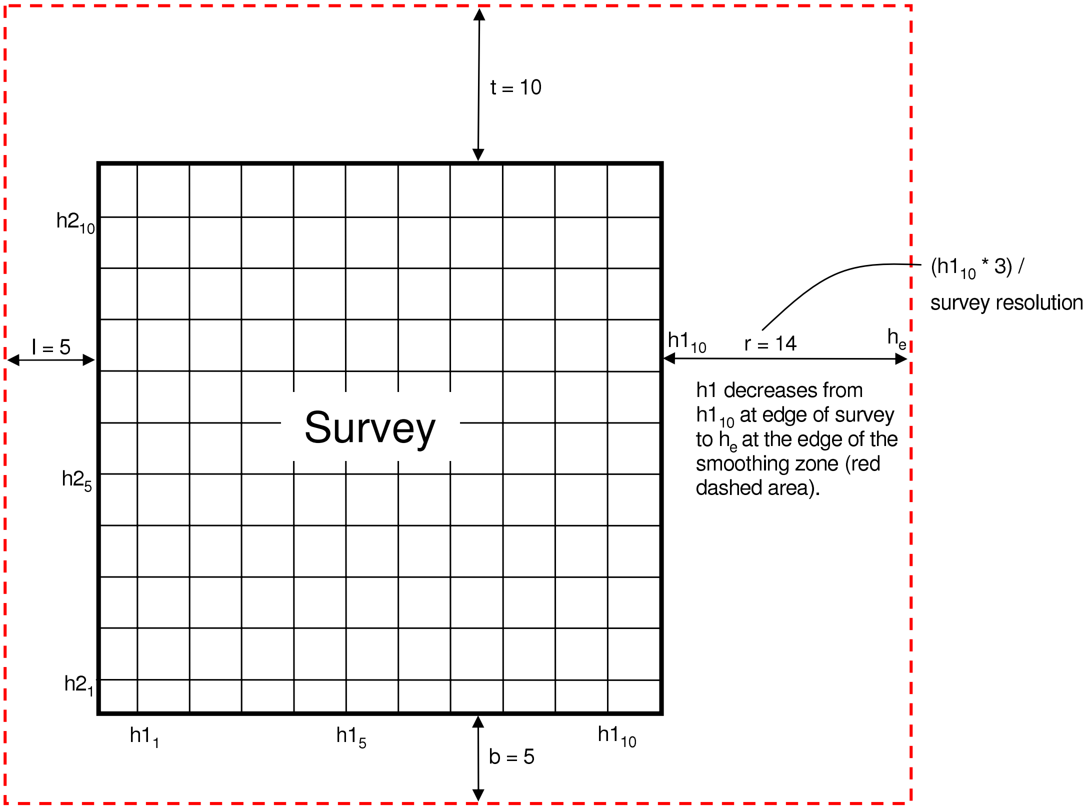
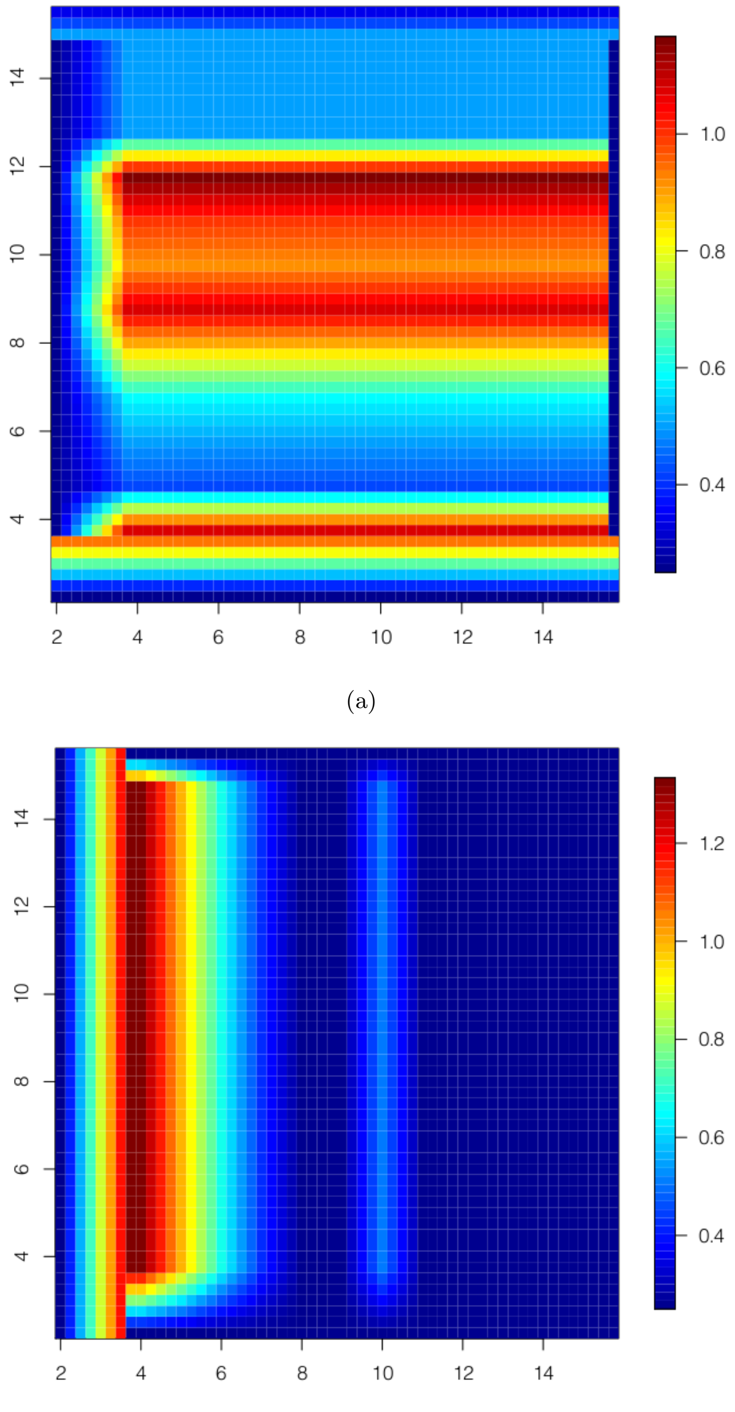


Figure 7.4: An example of the grid used to find the smoothing parameters h_1 and h_2, and the size of the buffer zone. Parameters t, b, l and r are top, bottom, left and right respectively and represent the number of grid points the buffer zone extends to on a particular edge of the squared off survey area.

One value of h is found for each of the 10 rows (h_1) and columns (h_2) and this is linearly extrapolated across the combined density surface (Fig 7.5). Within the buffer zone, the smoothing parameter decreases from the value of h at the edge closest to the new survey area to effectively zero at the edge furthest from it (Figure 7.5), where the smooth density surface (combined and buffer regions) joins the existing density surface. Ideally the corners of the buffer zone should radiate from the survey area, but the current implementation saves computation time by being very simplistic and makes little difference to the final smoothed outcome.

Visual outputs showing the combined density surface, the area used for smoothing and the final smoothed surface are provided. An example output is shown in Figure 7.6 where the existing density surface is shown as half degree grid cells and, the new survey covered an area around the west coast of Ireland and Scotland.



(a)

(b)

Figure 7.5: An example of smoothing parameters, h_1 (a) and h_2 (b) using test case 2e (see section 7.3.2). The buffer zone has dimensions t=3, r=1, b=6 and l=7.

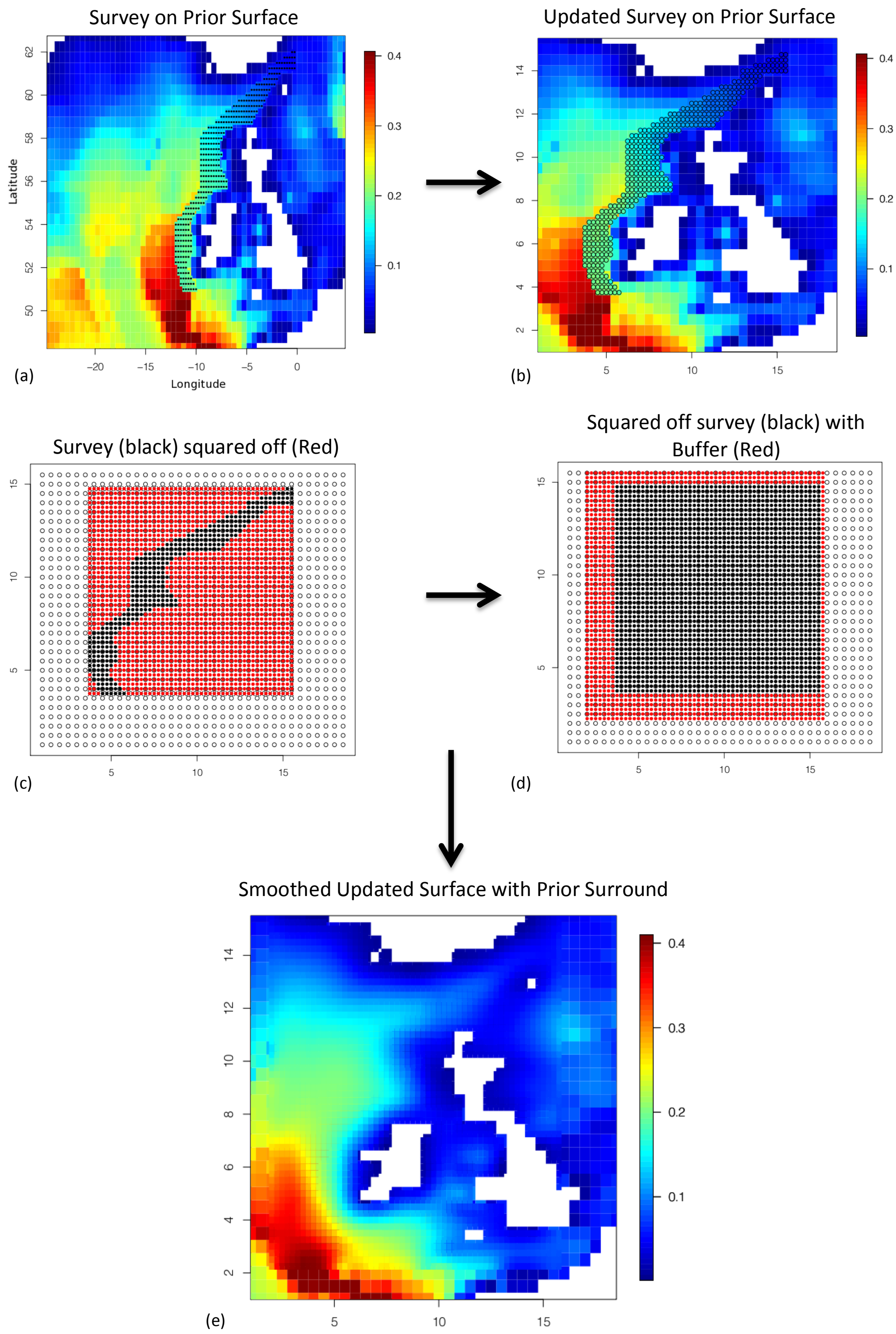


Figure 7.6: An example of the results of applying the combining and smoothing procedures of DCAP to the results of a survey off the west coast of Ireland and Scotland. (a) The existing and new survey densities shown on a Latitude and Longitude grid, (b) the combined densities overlaid on the existing data grid, (c) the area of survey (black) squared off (red) at the resolution of the survey, (d) squared survey area with buffer zone and (e) the combined density surface smoothed into a buffer and embedded in the existing un-updated density surface.

7.3.2 Test Cases

A number of test data sets were generated to assess the performance of the DCAP algorithm (Table 7.1). For most of the tests, the existing density surface was based on data for common dolphins from the UKHO database (*Delphinus* spp.), as this is an abundant species with a global distribution. Results from a number of simulated surveys were generated using Wisp (v1.2.6-1; Zucchini et al., 2007) and then the predicted densities were rescaled to resemble those in the existing density surface. All the simulated survey results had a grid resolution of 0.25^o except when the effects of a change of resolution was being tested. As listed in Table 7.1 surveys 1a-1n confirm that the DCAP application works in simple and complex areas within the Latitude/Longitude coordinate system and with respect to geographical features. Surveys 2a to 2f tested the effect of geometrically complex survey areas, such as complicated coastlines. The shape of these complex areas was based on two real survey areas taken from the Small Cetacean Abundance Survey in the North Sea - phase II (SCANS-II) project [Hammond et al., 2013]. Surveys 4a & 4b tested how the algorithm copes when one of the input datasets (existing data or new data) has an expectation of zero and the other has a positive density. Survey 5 assessed the algorithms performance when there is variable smoothness in two-dimensions, for example, a species whose distribution is strongly influenced by bathymetric feature, such as the shelf edge. In this case, densities will vary smoothly in one dimension but discontinuously in the other. Surveys 6a-6c were included to test whether there are any conflicts when updating a region with data from multiple surveys that are partially overlapping and have different resolutions. Surveys 7a-7c test the algorithm's performance with surveys from the same location but different spatial resolutions, to assess the relationship between computational time and the size of the survey dataset. Two examples, of differing resolutions were used. These were based upon humpback whale (*Megaptera novaeangelis*) sightings around the islands of Hawaii. The density surfaces for these data were also generated using Wisp (v1.2.6-1; Zucchini et al.,

2007) and based on estimated animal density from 4 years of surveys [Mobley, 2008]. The existing density surface was derived from the same database used in the common dolphin examples.

Further description of the test data can be found in a technical report by Burt [2010] and the software testing procedure in a report by Scott-Hayward et al. [2010].

Table 7.1: Survey test data used to assess the performance of the DCAP algorithm. * The northern and southern longitude limit is 60^{o}. S and D indicate whether the survey input is a stratified surface or a density surface. $^{+}$ SCANS-II is the Small Cetacean Abundance survey in the North Sea - phase II.

ID	**Objective**	**Density(D) or Stratified (S)**
	Open Ocean	
1a		D
1b		S
1c		S (2 strata)
	Land	
1d	top	D
1e	bottom	D
1f	right	D
1g	right	S
1h	left	D
1i	islands	D
	Lat/Lon	

– continued from previous page

ID	Objective	Density(D) or Stratified (S)
1j	Northern limit*	D
1k	Southern limit*	D
1l	Crossing the equator	D
1m	Crossing 0^o longitude	D
1n	Crossing -180/180 longitude	D
	Survey Regions - simple geometric shapes	
2a	Diamond	D
2b		S (2 strata)
2c	L-shpae	D
2d	M shape	D
	Survey regions - complicated shapes	
2e	SCANS-II B^+	D
2f	SCANS-II Q^+	D
	Non-overlapping data	D
4a	Existing density surface 0/survey density positive	D
4b	Existing density surface positive/ survey density 0	D
	Variable smoothness in 2D	
5	Step change in survey density	D
	Multiple Surveys	
6a	Several overlapping surveys	D

– continued from previous page

ID	Objective	Density(D) or Stratified (S)
6b		D
6c		D
	Survey grid size (degrees)	
7a	0.25	D
7b	0.1	D
7c	0.05	D
	Humpback data around Hawaii	
3a	0.25^o	D
3b	0.0167^o	D

7.3.3 Results

Simple visual inspection of the output from the DCAP algorithm can provide a good indication that there is nothing overtly wrong with this implementation and the results are consistent with the theory presented. For the combining step, we expect the estimates of density and uncertainty to fall within the original ones; there should be no clearly erroneous values (e.g. infinity) from either the combining or smoothing steps; the smoothed data must return to the values of the existing data at the edge of the smoothed region. Further, the smooth surface would not be expected to remove patterns apparent in the two information sources, unless there was a marked mismatch in precisions. Conversely, under-smoothing would not be expected such that the step between the two surfaces is pronounced.

Figure 7.6 shows the results from one complete run using survey 2f, which involves a complex survey area. The surface is first combined in the region of the survey area (Figure 7.6b), the output of this process may then be used as an input to another process, or smoothed (Figure 7.6e). The survey area is not rectangular, and so the region must be squared off for smoothing to occur (Figure 7.6c) and before the size of the buffer zone can be estimated (Figure 7.6d). This increases the number of data points for which smoothing must take place and thus increases computational time. Finally, the smoothed surface is embedded in the existing density surface (Figure 7.6e). These results are considered acceptable using the visual inspection specifications noted above.

The original RES based data represents long-term historical averages of distribution both in terms of the environmental covariates feeding into the RES model and the additional density data which spanned a 25 year period. Thus the updated densities achieved using the methods here are useful for keeping such a far-reaching database current in areas that are surveyed further, whilst also maintaining historical information in areas not yet (re-)visited. This means that the database presents information in a predictable form (estimates and precision) for inference, regardless of whether the information is from a single source or from a patchwork of sources. The resolution in the database may also become variable, with the addition of new, finer scale information, rather than the original 0.5^o. Furthermore, even if the initial interest was only in the survey region, insight can be gained into the surrounding area by embedding this into the far-reaching database. Changes in density can be seen in the maps presented here and a similar map showing uncertainty could be produced to show how this varies from the combined region to the existing.

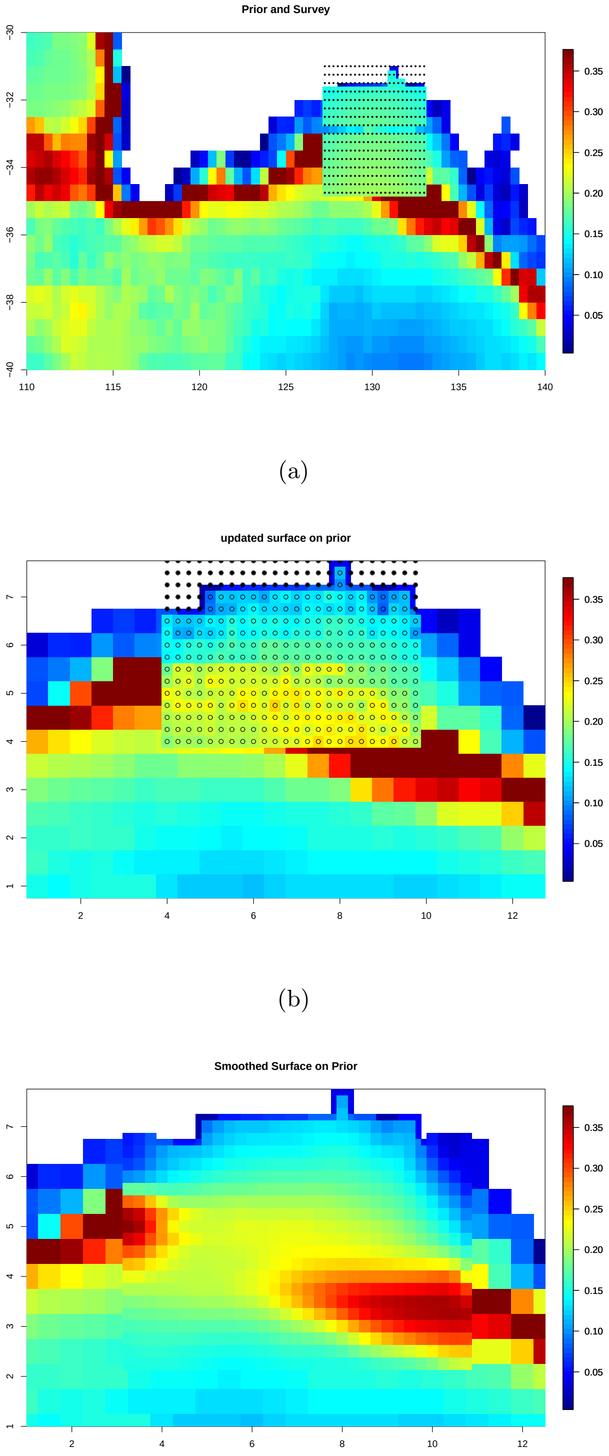
Harris [2013] suggested that the existing database be used with caution in coastal areas and that perhaps data collected on a local scale (fine scale in the region of interest) should be used in conjunction. The methods here provide an excellent way of combining this information to increase the resolution and usefulness around the coast. Figure 7.7 shows the results from a simulated survey in a coastal area (test case 1d). The DCAP algorithm was able to combine the density estimates and smooth the new surface even in the presence of a complex coastline.

The methods for combining estimates and smoothing edges worked as expected but there were a number of technical issues. Each of the test cases was designed to generate types of data input that might cause problems for the DCAP algorithm. A full log of the test results, containing timings, pass/fail and comments for each simulated survey, can be found in Appendix J. The algorithm was considered to have passed a test when all calculations were completed successfully, and the combined and smoothed surfaces were as expected given the theory and data inputs. For both test case 7c and 3b, the high resolution data set was too large for the computer being used and an error message was returned (the computer specifications are given in Appendix J).

The existing distribution could be rapidly combined with stratified survey results since there is no change in resolution to the existing data. The DCAP algorithm took longer to combine the densities if the existing data had "holes" because of the presence of land. Table 7.2 shows the results of increasing the spatial resolution of the test surveys. Unsurprisingly, the more data points (i.e. finer the grid), the longer the algorithm took to complete the combining and smooth procedures.

Table 7.2: Timings for the DCAP algorithm for three data sets of increasing spatial resolution.

ID	**Grid Size (degrees)**	**Number of data points**	**Time (sec.)**
7a	0.25	240	12
7b	0.1	412	1600
7c*	0.05	5096	6400



(a)

(b)

(c)

Figure 7.7: An example of the combining and smoothing procedure of DCAP where the new data come from a survey close to the coast (ID: 1d). (a) The existing density surface, (b) the surface produced by combining the new survey data and the existing surface and (c) the smoothed, combined density surface. Note: the apparent gap in the survey data in (a) is a plotting artifact due to the presence of land.

7.4 Discussion

This chapter has described a process for merging two partially/fully overlapping density surfaces to create a consistent composite map that gives both combined estimates and precisions. The process consists of two stages: first combining the density estimates, whilst accounting for precision, and secondly smoothing the joins between the two combined surfaces. It was illustrated using software developed for the US Navy to allow a global database of cetacean abundance to be combined with new survey data.

One of the most obvious applications of the DCAP algorithm is in assessing the potential impacts of acoustic disturbance associated with military exercises or seismic surveys. The sound produced by these devices may travel hundreds of kilometres from the source [e.g. Nieukirk et al., 2004, Jasny, 2005], potentially causing disturbance over areas that are too large to be covered by individual surveys. Results from multiple surveys, stored in large databases (such as the marine mammal surfaces held in the UKHOs Integrated Water Column product referred to above) are therefore required to assist in risk mitigation. However, it is important that the density surfaces generated using information from these databases are updated when new density information becomes available so that the long-term average is maintained to the present. Furthermore, the precision of the density surface is very important so that this may propagate through to the estimate of precision for any identified risk.

The MMAARS project [Donovan et al., 2010] requires a large scale database of animal densities with associated uncertainty for parts of the risk assessment software to work. Assuming, therefore, that keeping this database up-to-date is also a requirement, one could consider several ways for this to be achieved. Simple means of the existing and new data, or cutting holes and inserting new information are both viable solutions. However, taking means does not consider the uncertainty associated with each density and making holes ignores the historic data, creates edge issues (though one could use the smooth step to alleviate this) and assumes the survey is not an unusual year. The DCAP algorithm is a solution based on statistical theory for combining estimates in light of their precision and the combining of their precisions also. The smoothing step proposed here is a pragmatic use of kernel smoothing with bandwidth varying in line with correlation in the surface to achieve the desired effect. The solution presented is effective as it retains historical information, includes composite estimates of uncertainty and smooths the edges between the existing and new data.

Given the specific application of DCAP shown here, this algorithm has the ability to improve the existing database in a number of ways:

- The historical long-term average densities are kept current through use of recent survey information, whilst maintaining historical data elsewhere to ensure broad coverage. This is a necessity where the mitigation considerations are over very broad areas, outstripping localized survey and modelling information sources.

- Improvement of resolution, assuming the survey resolution is finer than that of the existing data. This is particularly useful in areas around coastlines which are known to be problematic in the existing data [Harris, 2013].

- Our best guesses of existing distribution, in the absence of any previous survey data, can be improved as new information is provided.

- There may be an improvement in precision, particularly if the new survey is in an area that the existing database values are entirely the result of model extrapolations, similar to the previous point. Harris [2013] used survey data to convert relative density into absolute density and if no survey information was available, the absolute densities were predicted. Therefore, the associated uncertainty for these data points was larger in order to reflect this estimation.

7.4.1 Limitations

Computational time may become an issue in real-time planning (e.g. military exercises) when a dynamic predictive model, rather than a survey, is used to update the existing data to predict the present/future distribution of species. Real-time planning is possible if data on current environmental conditions are used as the input to the predictive model. Depending on the size of the update, current density estimates could be produced, visualised and used as input for risk management software very rapidly. However, the test case results for DCAP showed a severe computation problem with large data sets. A waiting time of nearly two hours may not be feasible if decisions on activities are required. Larger datasets (that cover a wider area or are at a finer resolution) are quite likely and would increase the waiting time further. However, the time issue might be improved with the availability of better computing power or further development of the code to run some of the computations in parallel, which often improves computation time. Furthermore, there are several `R` packages for dealing with large datasets, which could be incorporated.

The methodologies described in this chapter cannot distinguish between poor and good quality survey estimates; precise inputs do not necessarily imply good quality data. Therefore, external vetting of these data and the way in which they were analysed is required to ensure that poor inputs have limited or no influence at a given spatial point (N.B. the variance of estimates dictates the influence they have). For example, as described in previous chapters, ignoring positive correlation could lead to levels of uncertainty that are too small. If the input data were the results of such an analysis then those density estimates might get too high a weight when combined with the existing data. Therefore, knowledge of how the input data were analysed is a useful way to assess the quality of the data.

There are some considerations if the database being updated has a temporal aspect. For example, there are 46 species in the database used for the DCAP example for which there are 4 seasonal density estimates. If there are multiple temporal surfaces, such as seasonal density estimates, updating a single temporal surface may lead to discontinuities in time much like we saw with discontinuities in space. A future addition to this methodology would allow smoothing across time as well as space to reduce temporal discontinuities in the database. However it must be noted that if the results of the new survey contain seasonal estimates and each season is updated with the new information, in the same geographic area, this should not be an issue.

No formal tests of the method were conducted. Instead, testing here has been with regards to pragmatic implementation and inspection of some outputs in light of theoretical construction. A simulation process could be used to determine how well the smoothing method compares with some other approach. However this was outside the scope of the project and something to be considered for the future. For example, the current implementation of the smoothing step does not take account of the distribution of land, which is simply modelled as a hole in the data. As was seen in chapters 3 and 4, this could lead to the leakage of predictions around land resulting in biases in the estimates of density in coastal areas. This could have important consequences for the identification of candidate MPAs. It is not a problem if the value $3h$ is smaller than the size of the land hole, because there will be no influence of points across land. However, h is calculated using an automated procedure and we cannot guarantee that this will be the case. The CReSS method described in Chapter 3 has been shown to be an effective way of dealing with the issue of leakage, and the next stage in the development of the methods in this chapter would be to integrate this into the smoothing step.

7.4.2 Potential Applications

So far discussion has centred on using the methods presented here to keep a database up-to-date by utilising new survey information. Another potential use of the algorithm is in a process referred to here as 'quilting'. I define the term quilting as a process used to generate a comprehensive map for a region that is covered by multiple overlapping surveys, none of which covers the entire region. This is not to be confused with the technique of tessellation, which is the tiling of a plane with no overlaps and no gaps. The current version of DCAP can only be used to combine an existing density surface with one additional overlapping survey, however the general methodology presented permits multiple updates. For example, one could combine the first density surface with the density surface from another survey, and then this combined surface could be merged with the results of a second survey, and so on.

There is a general trend [Kaschner et al., 2012] toward using existing data for a purpose for which they were not necessarily intended; to draw inference about large-scale cetacean distribution or trends over time. The key to this is to develop post-hoc methodologies, such as the solution developed in this chapter, to maximise the use of available data. This allows large-scale distributional trends to be identified or conservation issues to be addressed. The 'quilting' process described above is an appropriate method for combining density surfaces from surveys that have been carried out by independent organisations or in situations where it is sensible to model the results from each survey separately. For example, multiple surveys of turtles off the east coast of North America have been carried out by separate groups, and an overall distribution is required to aid management decisions (Borchers *pers. comm.*). Similarly, Williams et al. [2011] identified for the need for comprehensive spatial coverage of cetaceans density estimates in the north-eastern Pacific to identify candidate MPAs. The area covered by most individual surveys in this particular region is small relative to the size of the region itself.

Sometimes it may be appropriate to analyse a large dataset in small sections and combine the results afterward. For example, different covariates, such as tidal state or bottom type, may be good predictors of distribution in inshore areas, but may not be available on a larger geographic area, due to poor coverage, and cannot be included in the model. Thus the DCAP algorithm provides another way of analysing the JCP data resource described in Chapter 6. In that Chapter, data from all available surveys were combined using distance sampling methods and modelled in a single analysis. As an alternative, each survey could be analysed separately both to enable the inclusion of more appropriate covariates and to allow species to have slightly differing relationships with certain covariates depending on geographic location. The results of each survey could then be combined using the methods described in this chapter.

Chapter 8

Conclusions

The main aim of this thesis was to develop several methods that would allow the production of more accurate maps of marine species in areas of complex topography. Such maps are a key component of Species Distribution Modelling (SDM), which is concerned with understanding the factors that determine the distribution of species, and predicting how their distribution may change as a result of natural and anthropogenic factors.

This thesis presents the development and use of two novel methods for spatial smoothing in SDM: the Complex Region Spatial Smoother [CReSS; Chapter 3 and Scott-Hayward et al., 2013] and the bivariate Spatially Adaptive Local Smoothing Algorithm (SALSA2D; Chapter 4). In Chapter 1 I described three different broad applications for maps produced using SDM: estimating temporal trends in distribution, Environmental Impact Assessment (EIA), and spatial conservation planning. Chapters 4 and 6 used CReSS and CReSS-SALSA2D, respectively, to analyse data for two of these applications: the maps produced in Chapter 6 were used to assess changes in the spatial distribution of cetacean species from 1994 to 2010, and the maps of feeding probability for Southern Resident Killer Whales (Chapter 4) were used to identify candidate Marine Protected Areas (MPAs) that would protect important feeding grounds. Furthermore, both CReSS and SALSA2D are now being considered by Marine Scotland as part of a series of workshops on EIAs (the third broad application of SDMs) that will have a target audience of both industry and academia. The aim is to allow practitioners to conduct better assessments and managers to make better-informed decisions.

The thesis dealt with two issues commonly encountered when applying SDM to the distribution of marine organisms: (1) leakage across exclusion zones, defined as a geographic area which an animal may not cross (e.g. land) and (2) adaptive smoothing across a spatial surface. Leakage may occur when SDM is used to model data on a species abundance in areas of complex coastline and is caused when data on one side of an exclusion zone could lead a model to predict that nearby [as the crow flies] areas on the other side are similar in value. Failing to address this can produce misleading results with often severe consequences, such as prioritizing the wrong habitat for protection or constraining human activities in places that are actually not important to wildlife. There are recently developed statistical methods that deal with leakage but they are not suitable for all situations and can be difficult to employ, as explored through the extensive simulations presented in Chapters 3 and 4. Adaptive smoothing allows the smoothness of a density map to vary across the surface, unlike traditional penalised smoothing methods, which only allow one smoothing parameter. If the surface to be modelled is very smooth in one area and very wiggly in another, then a single smoothing parameter will tend to over smooth the wiggles and under smooth the flatter areas. The only existing method for spatially adaptive smoothing [AdaptFit; Krivobokova et al., 2008] does not deal with leakage across exclusion zones, which is an important issue in producing accurate species distribution maps. The solutions provided in this thesis, which are applicable to a variety of ecological problems, are simpler and more appropriate than existing methods.

An additional issue, once density maps have been created, arises from the disconnect between the vast quantity of small-scale, independent, and often overlapping, survey analyses and the requirement for large-scale maps for use in applications such as risk assessment. For example, in regions such as US and European coastal waters, where the number of independently conducted and analysed surveys has increased rapidly in recent years, there is a need for combining maps to allow large-scale management decisions. There are two issues here (1) how to combine competing estimates from differing density surfaces, given their precision, and (2) what to do at the join of the two surfaces to reduce the presence of un-naturally sharp transitions. Currently, there is no published method for dealing with these problems. This thesis contains solutions, based on statistical theory, for both combining of overlapping maps, to provide increased geographic coverage, and smoothing, to avoid any discontinuities in density where different maps join.

The remainder of this chapter summarises the statistical developments in this thesis (Section 8.1), and then explores some potential avenues for further statistical research (Section 8.2). Lastly, the two case studies presented in this thesis are discussed (Section 8.3).

8.1 Statistical Developments

Spatial models should be adjusted to account for the fact that animals have to swim around an island, because better descriptions/predictions of habitat use will lead to better area-based management tools (e.g. critical habitat designation, marine protected areas and risk assessments). Chapter 3 introduced a new method for smoothing in areas with complex topography CReSS [Scott-Hayward et al., 2013] - that addresses some limitations to current methods of smoothing and improves the accuracy of density maps. CReSS deals with issues such as the use of biologically meaningless distances in complex regions, global versus local smoothing and model selection/averaging. Firstly, it introduces a biologically realistic measure of inter-point similarity based on the geodesic distance between points, which reflects the distance an animal must travel between the points. Failing to address this can lead to very biased predictions when the ‘biological’ and Euclidean distances are markedly different. Secondly, CReSS employs a locally varying basis function to accommodate local smoothing requirements and alleviate problems with reinforcement around exclusion zones.

These two modifications are made prior to, or at, the basis function construction stage of the fitting process and therefore allows the basis to be used in a wide variety of statistical models, including maximum likelihood and quasi-likelihood fitting engines. Thirdly, a model-averaging framework is used to reduce sensitivity to small changes in model parameters, such as the basis range parameter or the number of knots.

After a thorough comparison under a variety of simulation settings CReSS was shown to perform as well, or better than other complex region methods (Geodesic Low Rank Thin Plate Splines [GLTPS Wang and Ranalli, 2007] and SOAP film smoothers [Wood et al., 2008]) and much better than Thin Plate Splines [TPS Harder and Desmarais, 1972]. Both GLTPS and SOAP have been previously compared with TPS but no direct comparison between the two has been published until now [Scott-Hayward et al., 2013].

The first simulation in Chapter 3 used a simple, horseshoe-shaped surface, which has already appeared in the literature [Ramsay, 2002]. For this (unrealistically) simplistic trial surface there was little practical difference in the fits between the complex region modelling methods (all were better than TPS) and there was no compelling evidence for any complex region method.

All the methods were then compared using a more topographically complex region that included an island, and with limited and noisy data (Chapter 4). CReSS had the lowest mean squared error at all noise levels and the lowest mean squared error variance across all trials in a sparse data simulation. I concluded that GLTPS should not be used in areas with islands due to its global basis function, and that choosing the dimension of the internal and boundary smooths in SOAP is problematic, particularly with increasing numbers of boundary loops, such as those caused by islands. Using traditional TPS-based techniques that ignore the presence of land, led to biases in predictions, under-estimation of density in hotspots, and overestimation in areas of low density.

Model selection is an important aspect of regression based spatial smoothing when structural components (or generally model complexity) are data-driven, rather than set *a priori.* For example, the location and number of knots or basis functions. CReSS, as it was initially implemented, was based on a space filled design, evenly distributing the knots across the surface [Scott-Hayward et al., 2013]. This is clearly sub-optimal for highly heterogeneous surfaces, where a spatially adaptive approach may be more appropriate. Penalised smoothing methods traditionally may have only one smoothing parameter, for example `gam`, from the `mgcv` library in `R` [Wood, 2006], which makes heterogeneously smooth surfaces overly smooth in wiggly areas and overly wiggly in smooth areas (i.e. locally biased through systematic under-/over-smoothing). Several adaptive approaches exist in one dimension, the most current being the Spatially Adaptive Local Smoothing Algorithm [SALSA; Walker et al., 2010], which uses knot number and location to vary the flexibility of the fitted surface. However, I have found only one approach for two dimensional smoothing, AdaptFit [Krivobokova et al., 2008], and none that deal with complex topography. The development in this thesis furthers the use of SALSA for bivariate smoothing, referred to here as SALSA2D to distinguish it from the original, and may be used in combination with CReSS (CReSS-SALSA2D) to allow for complex topography.

CReSS-SALSA2D performed better than CReSS at both low and medium noise levels. However, CReSS-SALSA2D performed badly at high noise, most likely due to over-fitting or an error in the specification of the range of r, which determines the local nature of the exponential basis. SALSA2D is currently the only option in topographically complex regions and it is being further refined for modelling surfaces with underlying heterogeneous smoothness. For example, I am currently researching ways to refine the selection of r in high noise situations and to ‘factor-interactions’ in the placement of knot locations, an issue identified in Chapter 6.

Current major issues in conservation, such as determining wide range effects of anthropogenic sound on cetaceans [e.g. Knoll et al., 2011], show there is a mismatch between the historical reasons for conducting many surveys and the use for which these surveys are now being considered. For such far-reaching issues there is a real need for large-scale maps to aid in the assessment of risk. There is no clear published method for combining information from multiple sources into a single cohesive information source. Chapter 7 developed methods for combining competing density estimates at a point, given their respective precision, and a smoothing process to alleviate sharp changes between the competing surfaces. Both methods address these problems with statistical tools: the combining method makes use of a Bayesian approach to give a merged density estimate weighted by the precision of the original estimates; the smoothing is evaluated by kernels, where the smoothing parameter is determined by the correlation between data points.

A specific application of these methods, the Dynamic Cetacean Abundance Predictor (DCAP), was developed to form part of software for environmental risk assessment by the US Navy. It is an algorithm created to take a global database of cetacean densities and associated precision, based on a combination of Relative Environmental Suitability (RES) indices and observed densities [Harris, 2013], and uses results from new surveys to keep the database current, specifically, making use of the combining and smoothing steps.

There are a number of current applications where the large-scale composite maps produced by the methods (combining and smoothing) are necessary, as required by software packages used to identify candidate MPAs, such as Marxan [Ball et al., 2009, Ardron et al., 2010]. Other applications require this combination of multiple surveys in order to provide a better understanding of a species distribution over a large geographic range. To emphasise the relevance of the methods in the latter applications, consider two large-scale projects, the Joint Cetacean Protocol (JCP; Chapter 6) and the Protection of Marine Mammals (PoMM).

The JCP project is funded by the UK government and its main aim is to assess changes in the distribution and abundance of seven cetacean species in north-western European waters using the JCP data resource Paxton et al. [2013]. This resource is a collection of survey data from 1969 to 2010 gathered by various governmental organisations, private sector companies and non-governmental organisations using a variety of survey techniques. Similarly, PoMM is a project funded by the European Defence Agency which is being conducted by the Ministries of Defence for Germany, Italy, the Netherlands, Norway, Sweden and the United Kingdom [Knoll et al., 2011]. The aim is to protect marine mammals against the impact of active sonar. One of the main objectives of the project is to create a comprehensive marine mammal database consisting of an encyclopaedia, observations, and maps of species distribution and noise sources in areas of operational interest to European navies.

Both these applications involve combining multiple surveys that were designed differently and can have markedly different outputs, in line with their varied remits. The JCP analysis showed what a complex task it is to do this [Paxton et al., 2013], and in the case of PoMM, it is not clear how the distribution maps will be created. In both cases, DCAP could be used to combine multiple surveys whilst maintaining estimates of the uncertainty associated with the density estimates. The methods for combining density surfaces would be a good solution for PoMM and something to be considered for a future analysis of JCP.

8.2 Future Statistical Developments

Whilst I have been able to demonstrate that the methods developed in this thesis perform well, there is still much that can be achieved through further improvements. The following sections detail some avenues of research for each of the three methods developed in this thesis.

8.2.1 CReSS

It is anticipated that CReSS will be added to the software package `DISTANCE` [Thomas et al., 2010], which is used to analyse distance sampling data [Buckland et al., 2001]. `DISTANCE` was used to generate the data inputs in Chapter 6 and the creation of density surfaces was done separately. However, within `DISTANCE` there is a density surface modelling engine which allows a variety of smoothers. The inclusion of CReSS as an option will improve the accuracy of the species density maps in topographically complex study regions as was seen through extensive simulations in Chapters 3 and 4. The developers of `DISTANCE` estimate there are over 1000 regular users, with 5000-10000 casual users, in over 100 countries, to analyse survey data from many taxa, including birds, mammals, reptiles, amphibians, plants and even litter. The density maps produced from DISTANCE analyses are often used to aid the conservation of species, so their accuracy is important. As a starting point to this process, both CReSS and CReSS-SALSA2D are being turned into an `R` package to make them user-friendly and more widely available. This is currently underway as part of a Marine Scotland contract, which will also include some basic `DISTANCE` analysis.

Another future use of CReSS is in the analyses of data, where the direction of potential movement between points is important. The geodesic distance used by CReSS assumes that the distance from one point to another is the same in both directions. However, a situation could arise where an animal could pass easily from one point to another, but returning could be more difficult. For example, in streams and rivers fish may easily move downstream, but rarely move upstream. This means that upstream data points should influence downstream points but not vice versa. It is not possible to add this information to the current distance matrix, as it contains only one distance for each pair of points. One solution would be to run two models with different distance matrices for each direction, and then use model averaging to produce the final spatially referenced estimates. This could be particularly useful for modelling the distribution of environmental contaminants, for example the leaching of heavy metals into soil, or of marine organisms in areas with strong currents.

8.2.2 SALSA2D

Both CReSS and CReSS-SALSA2D benefit from allowing the range parameter of the CReSS basis (r) to be varied for each knot location; a development which was presented in Chapter 6. Further simulation work needs to be carried out to investigate the robustness of both this approach and SALSA2D, under a range of conditions and using a variety of model selection criteria (e.g. AIC, BIC, CV). Chapters 3 and 4 showed the advantage of model averaging when knots were space-filled, particularly if data are sparse. The capability of SALSA2D to target smoothness to areas where it is particularly required may limit the effectiveness of model averaging, but this is purely speculative and some investigation is required to see if it provides the same advantage that is observed when knots are fixed.

Benchmark functions were used to assess the performance of CReSS and SALSA2D under a variety of scenarios. They allow assessment of model fit to the underlying function, rather than relying on cross-validation techniques where the truth is unknown. The 'palm' simulation, introduced here, provides a more challenging benchmark data set for assessing the performance of two-dimensional spatial modelling techniques than the relatively simple 'horseshoe' [Ramsay, 2002] that is conventionally used for this purpose. More benchmark data sets, particularly ones constructed from biological data, are needed to assess the practical consequences of challenges that often arise in ecological data, such as over-dispersion and autocorrelation.

8.2.3 Combining Density Surfaces

The kernel method for smoothing the junction between two density surfaces is limited because it does not respect biological/geodesic distances. This would be solved through using elements of CReSS. However, of the two processes for combining density surfaces the smoothing step is more computationally expensive and with the inclusion of CReSS is likely to become more so. Nevertheless, the problem is amenable to parallelisation due to several coding loops where the results are independent of one another, for example the *acf* calculations for each grid row/column do not depend on the results of any other row/column and could be computed concurrently. Accessible software tools for parallel computing are now readily available (e.g. `parallel`, `snow` and `multicore` packages in `R`; R Development Core Team, 2009) and multi-core computers (e.g. off-the-shelf 6 core hyperthreaded = 12 effective cores) are commonplace.

8.3 Case Studies

A number of specific case-studies are considered where the applicability of the methods developed here can be viewed in relation to real conservation/biological problems.

In Chapter 3 CReSS was employed to model data on the behaviour of Southern Resident Killer Whales (SRKW) off the North American west coast. These data were collected to produce maps of the distribution of specific behavioural activities, such as feeding or resting. The maps in this thesis represent an improvement on those produced by Ashe et al. [2010] using the same data, because they account for geodesic distances, include an assessment of uncertainty, and allow for correlation in the model residuals. They also illustrate that CReSS is a flexible modelling tool that is transferable to situations where there are error distributions other than Gausian (e.g. quasi-Poisson) or correlated errors and can be used for quantitative aspects of spatial conservation planning. These maps of the probability that whales would be observed feeding, in combination with density maps from Hauser et al. [2007], were used to identify a candidate MPA for SRKW to the south of San Juan Island. However, as mentioned in the General Introduction (Chapter 1), there is more to identifying candidate MPAs than distribution mapping alone. Other considerations are the conservation of prey species, local laws and policies, and human interactions.

In Chapter 6, the techniques developed in Chapters 3 and 5 (CReSS and SALSA2D) were used to model the distribution of harbour porpoise and minke whale in north-western European waters using data from the JCP resource. No off-the-shelf methods could deal with such a large, correlated, topographically complex and heterogeneously smooth data set. CReSS, SALSA2D and SALSA1D [SALSA for univariate smooths; Walker et al., 2010] were used to model animal densities as a function of a variety of environmental and temporal covariates. SALSA1D was used to adjust the flexibility of one-dimensional smooths, which were fitted using cubic B-splines, and CReSS combined with SALSA2D was used to add a two-dimensional smooth of space to the model. Based on the simulation results presented in Chapters 3 - 5, CReSS was used to reduce the chance of underestimating hotspot areas and overestimating density in areas where animals were rarely seen, whilst SALSA reduces the risk of oversmoothing highly structured areas and undersmoothing areas where density was less variable. An update to CReSS, automated selection of r, was also added in this chapter, and simulation results, updated to represent this change, were presented in Scott-Hayward et al. [2013].

The maps produced for harbour porpoise and minke whale were similar to those produced by previous studies [Paxton and Thomas, 2010, Paxton et al., 2011, Hammond et al., 2002, 2013] that analysed subsets of the data. High densities of harbour porpoise were observed off the west coast of Scotland and off the coast of East Anglia in 2010. The area off Scotland was a persistent hotspot for the whole study period (1994-2010), whereas a shift in distribution was identified in the North Sea from central areas to the south west during this time. For minke whale, the main high density areas were on the west coast of Scotland and the western North Sea, but temporal shifts were not investigated due to limitations of the data. The west coast of Scotland appears to be an important area for both species and should perhaps be considered a focal point for further studies of this data resource. Although legislation in Scotland requires the designation of MPAs for harbour porpoise (Marine Scotland Act, 2010), one sited in this region would protect both species. As Ferrier et al. [2002] have noted 'assessments at global or continental scales can help focus attention on broad regions of particular conservation concern, [but] a more detailed assessment is usually required to guide decisions on the actual location of conservation areas'. This view is reflected in the next phase of the JCP project, which aims to identify candidate MPAs for harbour porpoise and bottlenose dolphins.

If I were to re-analyse the JCP data resource to identify candidate MPAs I would consider the 'quilting' method described in the discussion of Chapter 7. The region of interest can be divided into overlapping tiles enabling the effect of environmental covariates, that are not available or appropriate at a larger scale, to be investigated. For example, analyses of tiles near to the coast would use tidal state as a covariate, whereas this might not be appropriate for tiles in open water. Tiles from sub-regions could be assessed individually to identify candidate MPAs and satisfy the suggestion of Ferrier et al. [2002] for a detailed (fine-scale) assessment, or all the tiles could be combined in one unifying map, using the methods from Chapter 7. A single unified map could be used on a small scale for spatial conservation planning and on a larger scale in risk assessment for military or construction activities.

In both the case studies, the issue of autocorrelation was addressed using Generalised Estimating Equations [GEEs; Hardin and Hilbe, 2002]. Autocorrelation is often overlooked or ignored in SDM [Araújo and Guisan, 2006, Hawkins, 2012] and it cannot be addressed with standard implementations of GAMs. Therefore, the confidence intervals associated with the predictions from any models will be too narrow if positive autocorrelation is present and ignored. Combining CReSS or CReSS-SALSA2D with GEEs deals with this issue.

At the start of this thesis I highlighted the need for accurate maps and the need to update existing maps when new information becomes available. Recent legislation (Marine Scotland Act, 2010 and Marine and Coastal Access Act, 2009) requires the creation of a network of MPAs in UK seas to protect biodiversity and geodiversity. The UK and national governments need tools that can be used to identify candidate MPAs. This thesis has shown how biases in the predictions of animal density that result from ignoring geodesic distance or spatial heterogeneity can be reduced, and how data from multiple overlapping surveys can be combined to create unified maps covering large geographic areas.

8.4 Current studies using CReSS and SALSA2D

CReSS-SALSA2D has been used to model the distribution of tern species (*Sternidae spp.*) around the UK [Mackenzie and Scott-Hayward, 2012, Mackenzie et al., 2012] and it is being used to investigate home range distribution of stoats (*Mustela erminea*) and leopards (*Panthera pardus*; Mackenzie, Borchers and Walker *pers. comm.*). The local, radial nature of CReSS and the spatially adaptive nature of SALSA make them particularly suitable for assessing the potential impact of wind farm construction, and there are plans to use them to analyse the potential impacts of Danish wind farms at Nysted, [Petersen et al., 2011] and Röesand in Denmark.

These methods, along with GAMs [Wood, 2006] and GAMMs [Mixed Models; Brown and Prescott, 1999], are currently being assessed by Marine Scotland for use in determining the environmental impact of wind and wave turbine developments on seabird and cetacean species. CReSS and SALSA2D will form part of the software to be taught at workshops on EIAs with a target audience of both industry and academia. The aim is to allow practitioners to conduct better analyses and managers to make better-informed decisions. Furthermore, a suite of benchmark data is being created based on an off-shore line-transect analysis and a vantage point analysis (data is collected from an observer on a cliff-top). The data is both over dispersed and correlated, and describes a variety of impact scenarios (no effect, overall decrease and redistribution of animals).

8.5 Final Remarks

Some of the work presented in this thesis has advanced methods for SDM to address certain key statistical issues. The methods improved the accuracy of maps to designate a candidate MPA for SRKW and establish long-term distributional trends for harbour porpoise and minke whale. However, the methods developed are notably not limited to the applications presented in the case studies; they may be applied to other geographic regions or species (including terrestrial) and, outside the field of biology may be implemented for many spatial regression problems, for example demographic studies [Ramsay, 2002, Marra et al., 2011].

There is a good case for the CReSS and SALSA methods developed here to become standard tools for analysing ecological data and not just in topographically complex regions. CReSS may include geodesic distance, but need not if Euclidean distance is appropriate; may be spatially adaptive, with the inclusion of SALSA-1D and -2D; may use a GEE framework and thus a variety of error distributions (e.g. Binomial, Poisson, quasi-Poisson) and may employ model averaging to improve the robustness of results.

However, uptake of these approaches will require careful dissemination so that they may be understood and applied by scientists not familiar with the details. Development of an `R` package is underway, to ease use, and a workshop is set for late 2013 to teach the methods, initially to a limited number of EIA practitioners but it is hoped to a wider audience in 2014.

Appendix A

Index of Acronymns and Notation

Table A.1: Index of Acronyms

Acronym	Description
ACF	AutoCorrelation Function
AIC	Akaike's Information Criterion
AIC_c	corrected Akaike's Information Criterion
ASCOBANS	Agreement on the Conservation of Small Cetaceans of the Baltic and North Sea
BACI	Before-After-Control-Impact
BAG	Before-After Gradient
BIC	Bayesian Information Criterion
CV	Cross Validation
CoV	Coefficient of Variation
CReSS	Complex Region Spatial Smoother
DCAP	Dynamic Cetacean Abundance Predictor

– continued from previous page

Acronym	**Description**
EIA	Environmental Impact Assessment
ERMC	Environmental Risk Management Capability
ESAS	European Seabirds At Sea
FELS	Finite Element L-Spline
GAM	Generalised Additive Model
GCV	Generalised Cross Validation
GEE	Generalised Estimating Equation
GLM	Generalised Linear Model
GLTPS	Geodesic Low-rank Thin Plate Spline
JCP	Joint Cetacean Protocol
MPA	Marine Protected Area
MSE	Mean Squared Error
NOAA	National Oceanic Atmospheric Administration
OPS	Overall Prediction Score
P-IRLS	Penalised Iteratively Re-weighted Least Squares
PRS	Penalise Regression Spline
QAIC	Quasi likelihood based AIC
QAIC_c	corrected Quasi likelihood based AIC
RES	Relative Environment Suitability
RSS	Residual Sums of Squares

– continued from previous page

Acronym	Description
SAC	Special Area of Conservation
SALSA	Spatially Adaptive Local Smoothing Algorithm
SCANS-II	Small Cetacean Abundance in the North Sea, phase II
SDM	Species Distribution Modelling
SRKW	Southern Resident Killer Whale
SOAP	SOAP film smoothing method
SST	Sea Surface Temperature
TPS	Thin Plate Spline
UK	United Kingdom
UKHO	United Kingdom Hydrographic Office
USA	United States of America

Table A.2: Index of Notation

Parameter	Description	
A	region or domain	
$\hat{\mathbf{b}}$	vector of estimation bias	
$b()$	basis function	
D	deviance	
$d_{i,t}$	Euclidean distance	$(i = 1, \cdots, n)$, $(t = 1, \cdots, T)$

– continued from previous page

Parameter	Description	
$\mathbf{e}_i$	vector of errors	$(i = 1, \cdots, n)$
$g()$	link function	
$g_{i,t}$	Geodesic distance	$(i = 1, \cdots, n)$, $(t = 1, \cdots, T)$
g_r	grid resolution	
$\mathbf{G}$	Geodesic distance matrix	
h	bandwidth/smoothing parameter	
$\mathbf{H}$	Hat matrix	
k	autocorrelation lag	
K	number of estimable parameters $(P + 2)$	
$K()$	kernel smooth	
M	polynomial degree or candidate model set	
N	prediction grid size	
n	number of data points	
P	number of covariates	
R	number candidate models	
r	determines radius of exponential function	
$s()$	smooth function	
$\mathbf{S}$	penalty matrix	
T	number of knots	
w	number of nearest neighbours	

– continued from previous page

Parameter	Description	
$w()$	weights function	
$\mathbf{W}$	weights matrix	$(n \text{ x } n)$
$\mathbf{x}_i$	vector of covariate values	$(i = 1, \cdots, n)$
$\mathbf{X}$	covariate matrix	$(n \text{ x } (P{+}1))$
y_i	response value at data point i	
y_i^*	true function value at data point i	
$\bar{y}$	mean of y	
y_n	true function with noise added	

β	regression model coefficient	
ϵ	error values from a specified distribution	
$\eta()$	linear predictor	
θ	general model parameters	
$\boldsymbol{\kappa}$	knot vector $(\kappa_{1,t}, \kappa_{2,t})$	$(t = 1, \cdots, T)$
λ	Poisson parameter or smoothing parameter	
μ	mean	
$\tilde{\mu}$	median	
τ	knot sets	
σ	standard deviation	
ϕ	dispersion parameter	

Appendix B

Description of Floyd's Algorithm

I use Floyds Algorithm [Floyd, 1962] for the calculation of the shortest distance between data points. Generally speaking, the points are vertices on a graph which may be connected, and have an associated weight, or unconnected. In the case of shortest distance calculation for this thesis, the starting matrix is populated using the Euclidean distance between points. Any distances where the Euclidean connection between the two is invalid, by crossing land for example, are represented by infinity in the matrix.

Imagine four data points with distances between each given in the matrix below. The distance between point one and point four is infinite suggesting that the Euclidean distance between these points is invalid.

$$G(0) = \begin{pmatrix} 0 & 4 & 2 & \infty \\ 4 & 0 & 7 & 1 \\ 2 & 7 & 0 & 4 \\ \infty & 1 & 4 & 0 \end{pmatrix}$$

The general principle is to look at one entry and see if the sum of two others is smaller or larger. If it is smaller then the entry is replaced, otherwise it is left as is.

Round 1:

In the first round, we use the distances from column one and row one (shaded above), to see if the other distances can be made smaller.

First we look at $G(0)_{2,2}$ and compare that distance entry with the sum of the two shaded numbers in the same row/column ($G(0)_{2,1}$ and $G(0)_{1,2}$).

$$G(0)_{2,2} = 0$$

$$G(0)_{2,1} + G(0)_{1,2} = 4 + 4 = 8$$

the zero remains.

then,

$$G(0)_{2,3} = 7$$

$$G(0)_{2,1} + G(0)_{1,3} = 4 + 2 = 6$$

thus, the $G(0)_{2,3}$ entry is replaced with 6.

then,

$$G(0)_{2,4} = 1$$

$$G(0)_{2,1} + G(0)_{1,4} = 4 + \infty = \infty$$

the 1 remains.

then,

$$G(0)_{2,4} = 4$$

$$G(0)_{3,1} + G(0)_{1,4} = 4 + \infty = \infty$$

the 4 remains.

That is the upper half complete (we need not check the zeros) and since the original matrix was symmetric, we can fill in the bottom half the same. So, after round one the updated distance matrix is:

$$G(1) = \begin{pmatrix} 0 & 4 & 2 & \infty \\ 4 & 0 & 6 & 1 \\ 2 & 6 & 0 & 4 \\ \infty & 1 & 4 & 0 \end{pmatrix}$$

Round 2:

Again, we use the shaded row/column to evaluate the shortest distance between two points (an unshaded entry)

First we look at $G(1)_{1,1}$ and compare that distance entry with the sum of the two shaded numbers in the same row/column ($G(1)_{2,1}$ and $G(1)_{1,2}$).

$$G(1)_{2,2} = 0$$

$$G(1)_{2,1} + G(1)_{1,2} = 4 + 4 = 8$$

the zero remains.

then,

$$G(1)_{1,3} = 2$$

$$G(1)_{2,3} + G(1)_{1,2} = 4 + 6 = 8$$

the 2 remains.

then,

$$G(1)_{1,4} = \infty$$

$$G(1)_{2,4} + G(1)_{1,2} = 4 + 1 = 5$$

thus, the $G(1)_{1,4}$ entry is replaced with 5. Data points one and four are not directly connected (represented by ∞), but if you travel through point 2, then you can get there in five units.

then,

$$G(1)_{3,4} = 4$$

$$G(1)_{2,4} + G(1)_{3,2} = 6 + 1 = 7$$

the 4 remains.

That is the upper half complete and since the original matrix was symmetric, we can fill in the bottom half the same. So, after round two the updated distance matrix is:

$$G(2) = \begin{pmatrix} 0 & 4 & 2 & 5 \\ 4 & 0 & 6 & 1 \\ 2 & 6 & 0 & 4 \\ 5 & 1 & 4 & 0 \end{pmatrix}$$

Nothing changes for round 3, so the matrix at the end looks the same:

$$G(3) = \begin{pmatrix} 0 & 4 & 2 & 5 \\ 4 & 0 & 6 & 1 \\ 2 & 6 & 0 & 4 \\ 5 & 1 & 4 & 0 \end{pmatrix}$$

For round four, the sixes are updated to fives.

$$G(4) = \begin{pmatrix} 0 & 4 & 2 & 5 \\ 4 & 0 & 5 & 1 \\ 2 & 5 & 0 & 4 \\ 5 & 1 & 4 & 0 \end{pmatrix}$$

This is one loop through the matrix. The above ‘four round’ process is repeated until a complete set of rounds makes no further changes. In this example, data point one and four were not originally connected. After this loop through the matrix, the distance between points one and four is five units. In this way, from an starting matrix describing which points are connected, we can use Floyds Algorithm to calculate the shortest distance between all data points.

Appendix C

Bathymetry map for the San Juan Islands area.

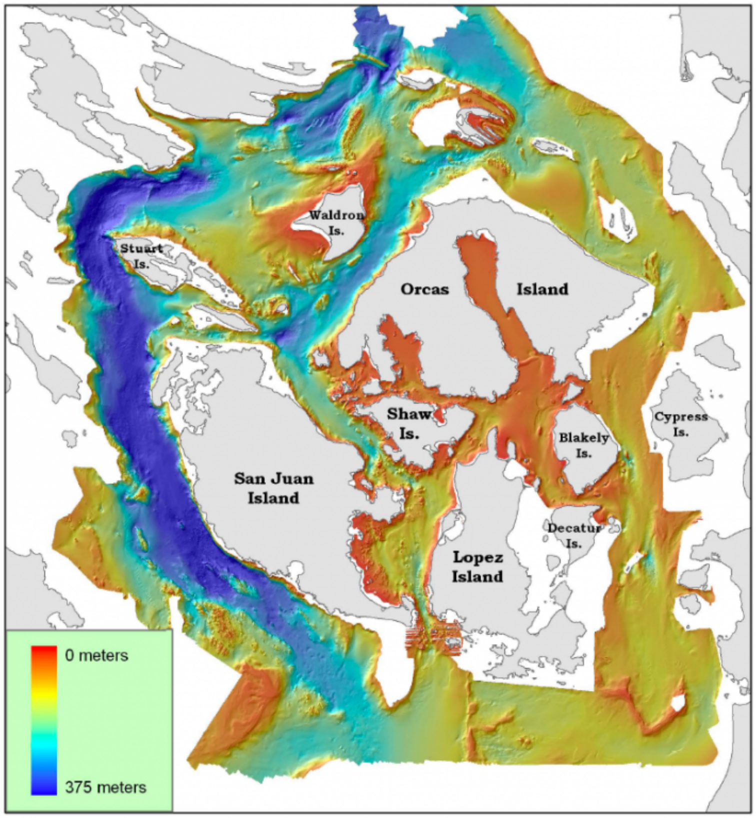
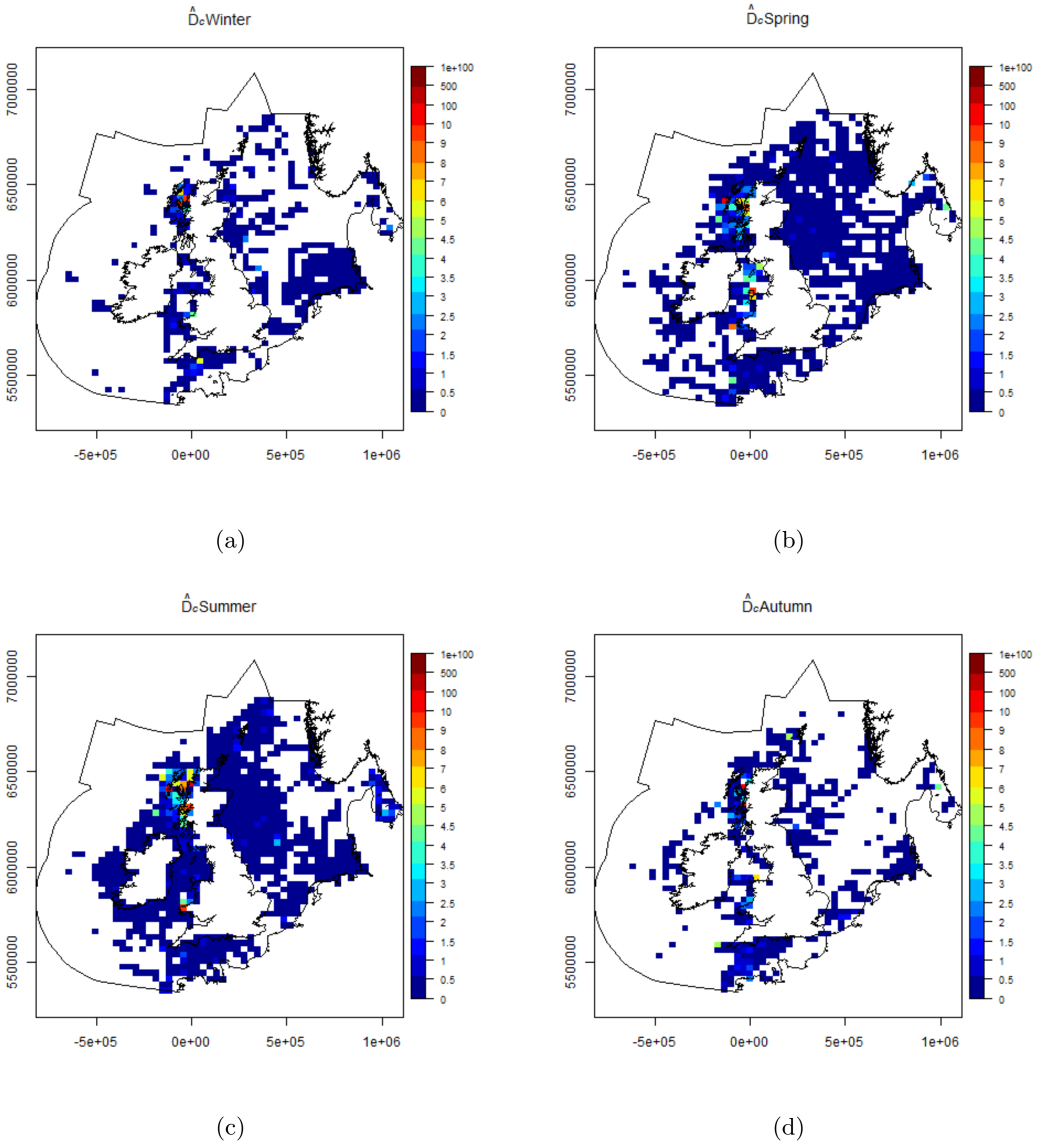


Figure C.1: Seafloor baythmetry map for the San Juan Islands area taken from Greene et al. [2007]

Appendix D

Extra plots of the Joint Cetacean Protocol Analysis - Seasonality

D.1 Harbour Porpoise Seasonal Plots



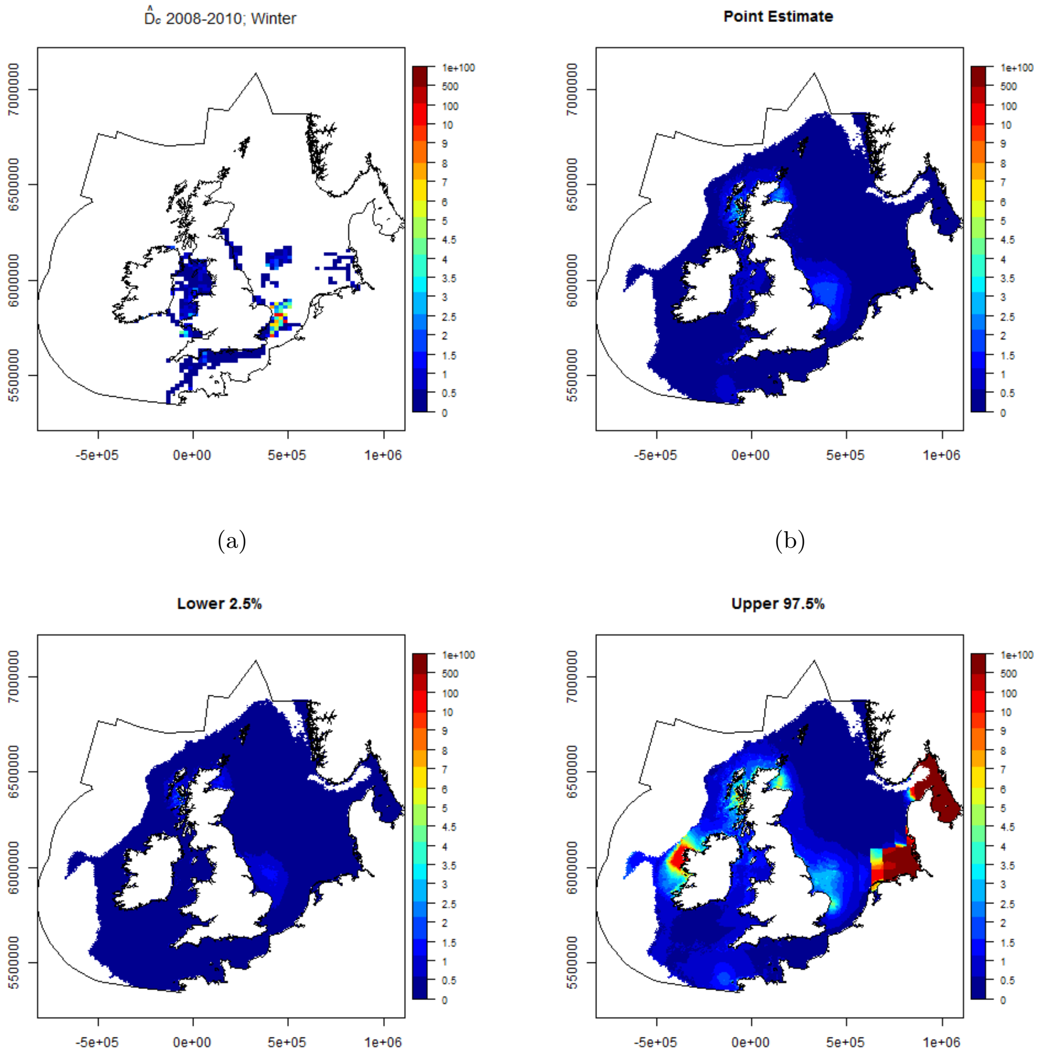
(a)

(b)

(c)

(d)

Figure D.1: Harbour porpoise density data for all years in (a) winter, (b) spring, (c) summer and (d) autumn. 16% of data was collected in winter, 29% in spring, 41% in summer and 14% in autumn.



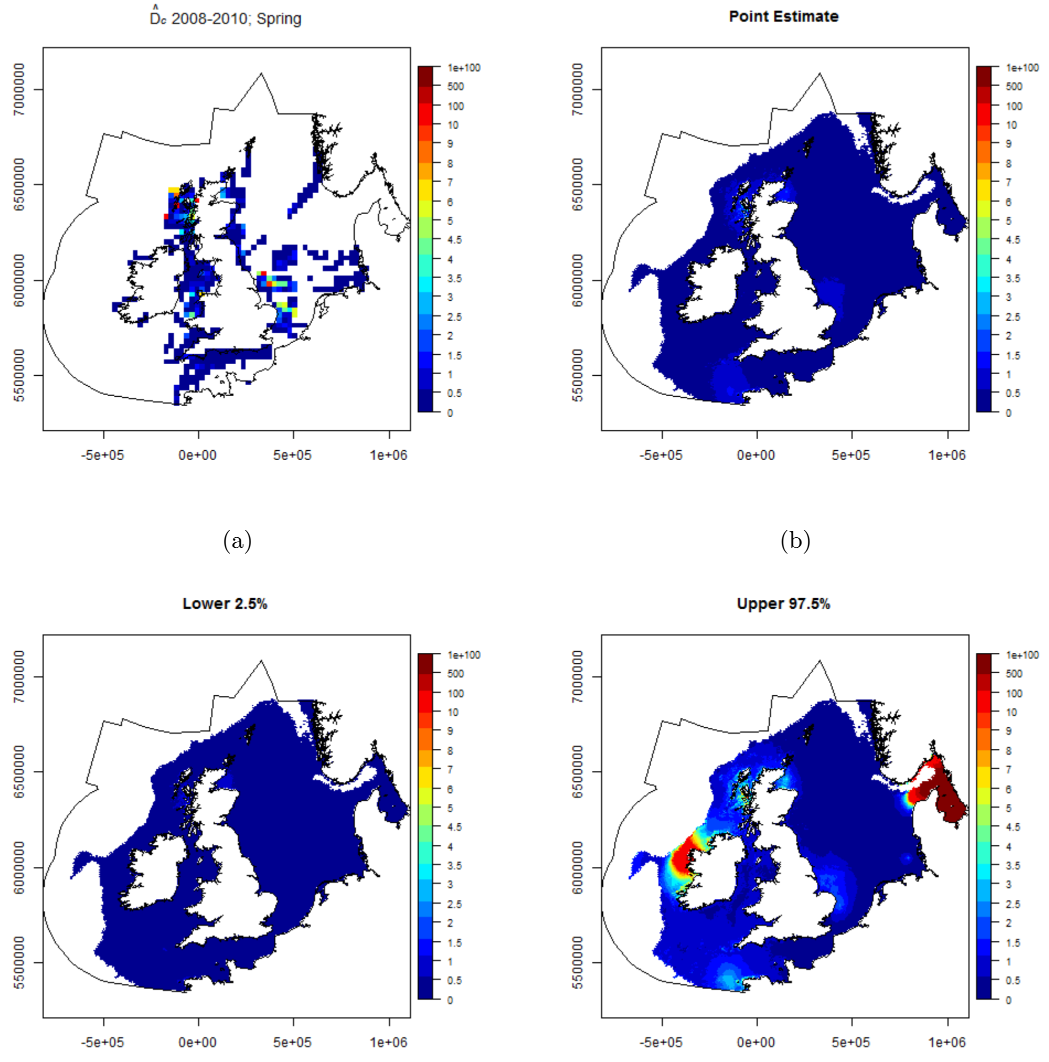
(a)

(b)

(c)

(d)

Figure D.2: Harbour porpoise densities for 2010 in winter. (a) The raw densities for winter in 2008 - 2010 that are drawn upon to make predictions for 2010, (b) point estimates of harbour porpoise density for winter 2010, (c) and (d) are the lower and upper 95% GEE based percentile intervals.



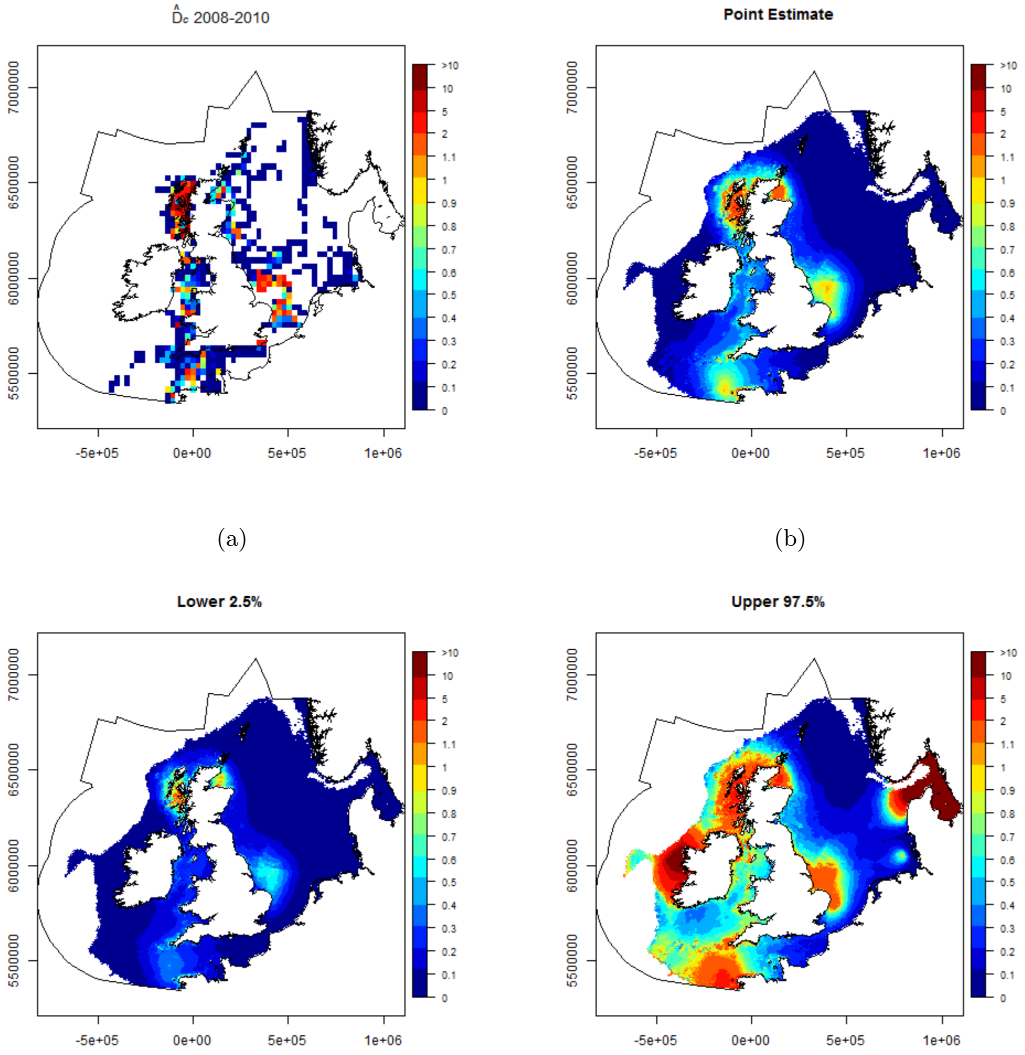
(a)

(b)

(c)

(d)

Figure D.3: Harbour porpoise densities for 2010 in spring. (a) The raw densities for spring in 2008 - 2010 that are drawn upon to make predictions for 2010, (b) point estimates of harbour porpoise density for spring 2010, (c) and (d) are the lower and upper 95% GEE based percentile intervals.



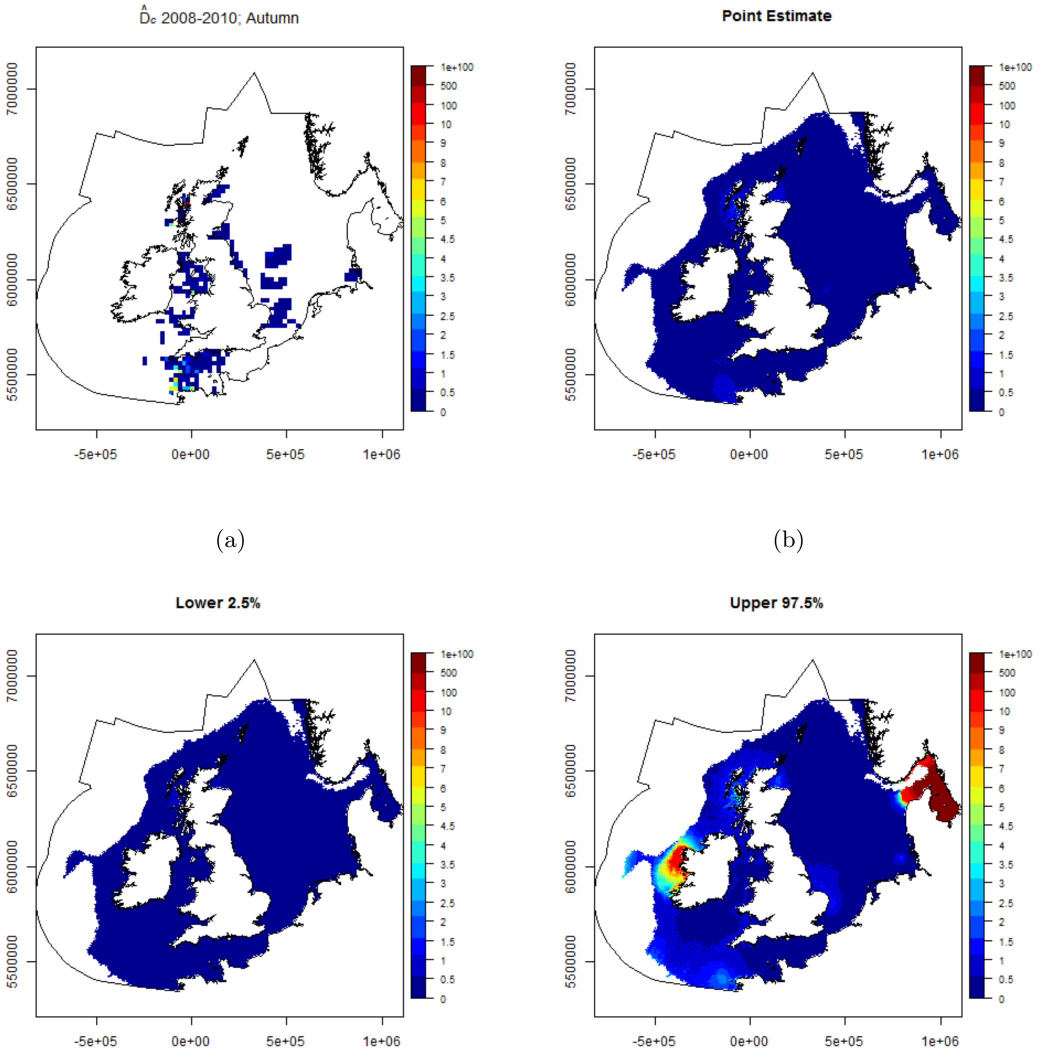
(a)

(b)

(c)

(d)

Figure D.4: Harbour porpoise densities for 2010 in summer. (a) The raw densities for summer in 2008 - 2010 that are drawn upon to make predictions for 2010, (b) point estimates of harbour porpoise density for summer 2010, (c) and (d) are the lower and upper 95% GEE based percentile intervals.



(a)

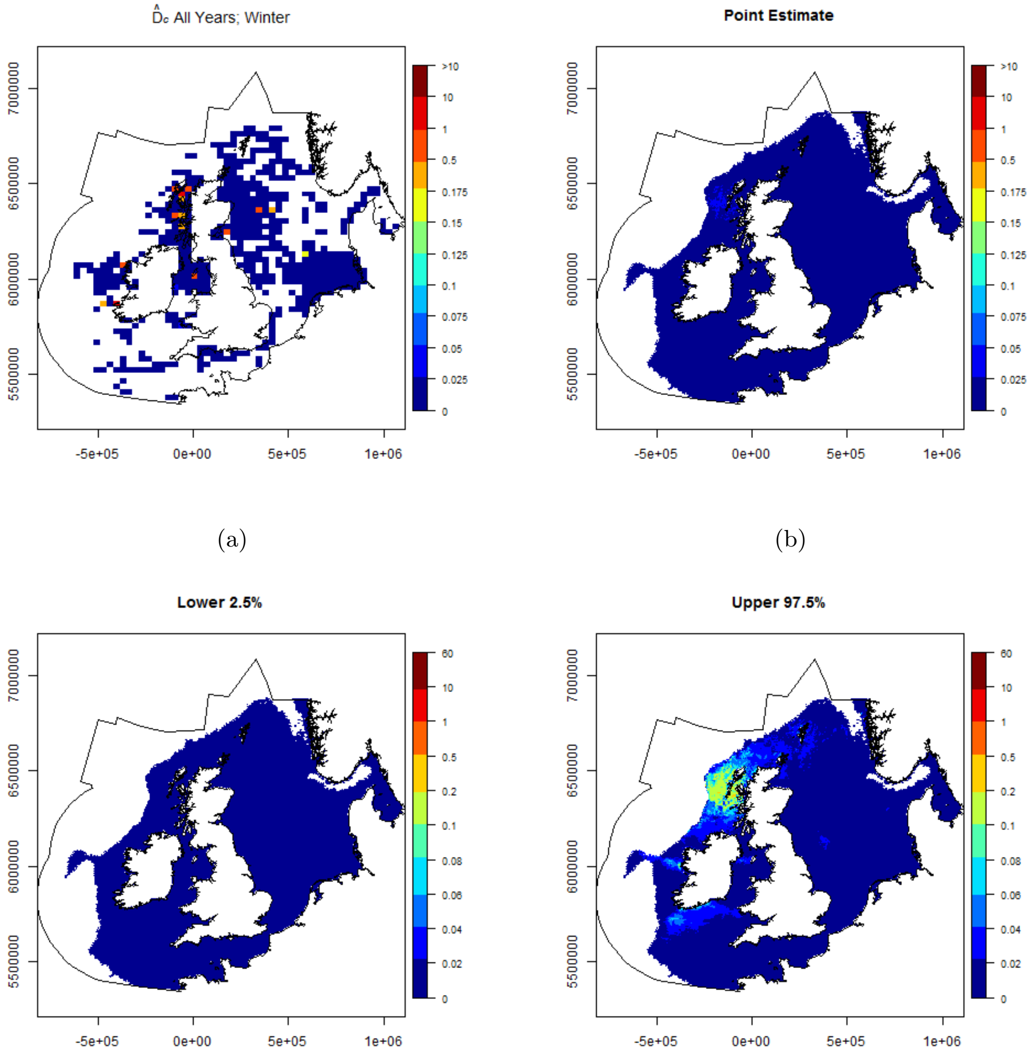
(b)

(c)

(d)

Figure D.5: Harbour porpoise densities for 2010 in autumn. (a) The raw densities for autumn in 2008 - 2010 that are drawn upon to make predictions for 2010, (b) point estimates of harbour porpoise density for autumn 2010, (c) and (d) are the lower and upper 95% GEE based percentile intervals.

D.2 Minke Whale Seasonal Plots



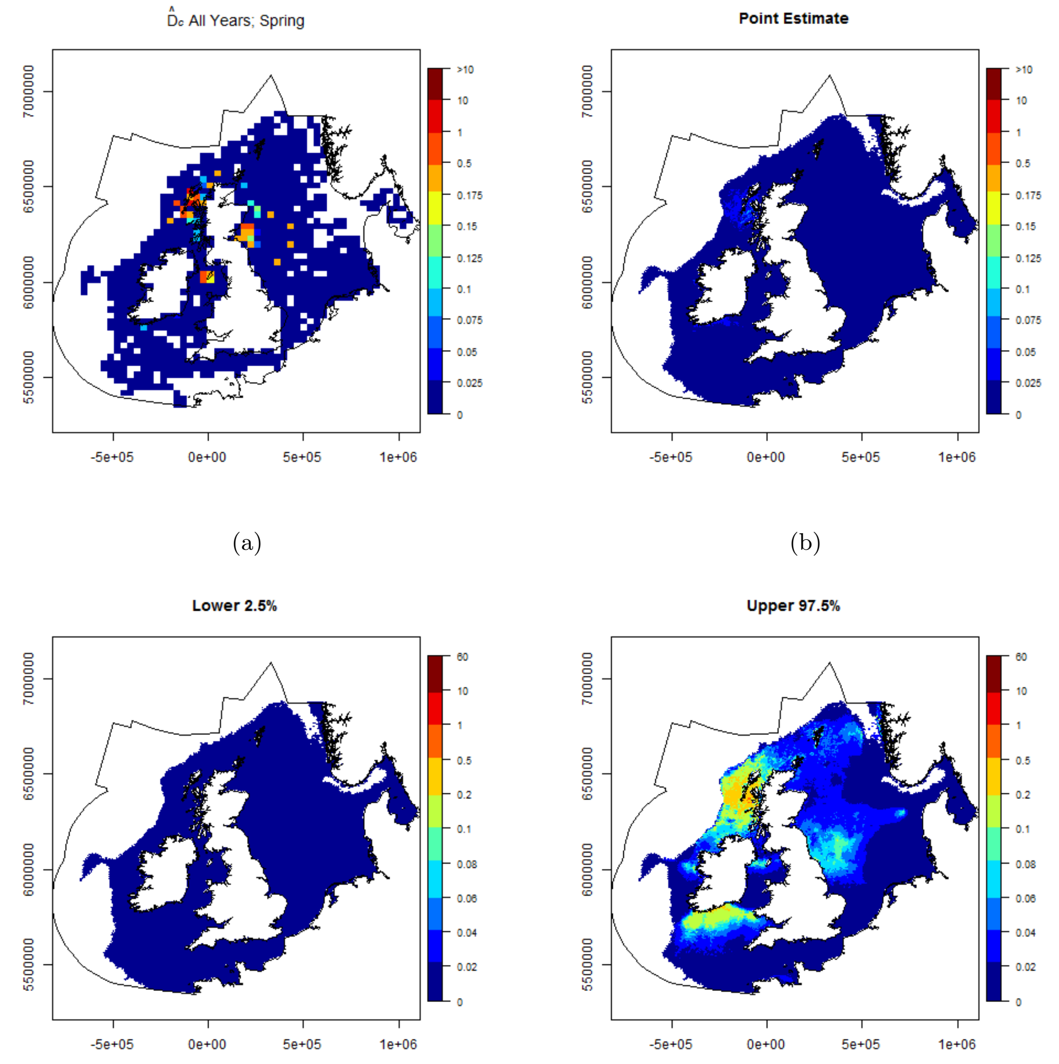
(a)

(b)

(c)

(d)

Figure D.6: Minke whale densities for 2010 in winter. (a) The raw densities for winter in 2008 - 2010 that are drawn upon to make predictions for 2010, (b) point estimates of minke whale density for winter 2010, (c) and (d) are the lower and upper 95% GEE based percentile intervals.



$\hat{D}_c$ All Years; Spring

Point Estimate

Lower 2.5%

Upper 97.5%

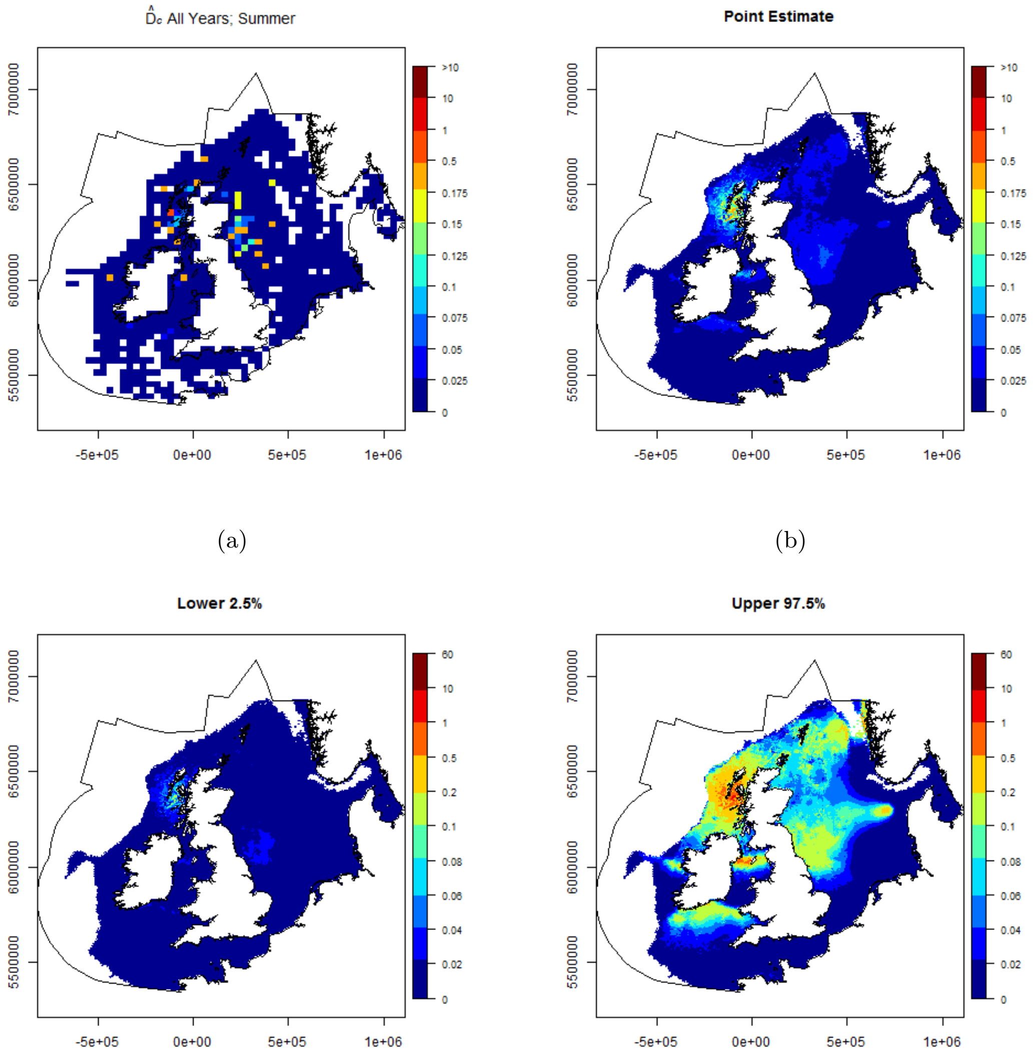
(a)

(b)

(c)

(d)

Figure D.7: Minke whale densities for 2010 in spring. (a) The raw densities for spring in 2008 - 2010 that are drawn upon to make predictions for 2010, (b) point estimates of minke whale density for spring 2010, (c) and (d) are the lower and upper 95% GEE based percentile intervals.



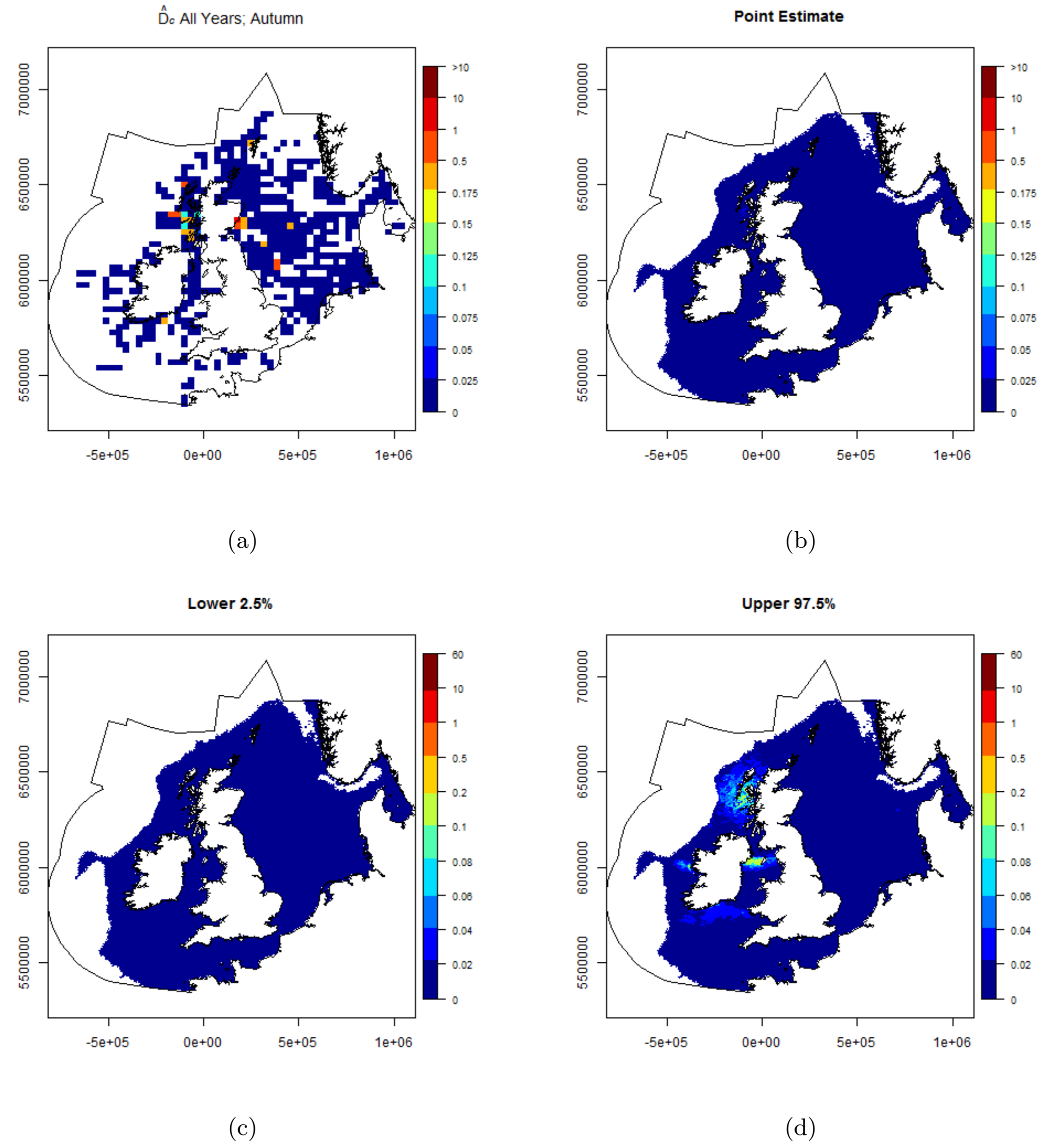
(a)

(b)

(c)

(d)

Figure D.8: Minke whale densities for 2010 in summer. (a) The raw densities for summer in 2008 - 2010 that are drawn upon to make predictions for 2010, (b) point estimates of minke whale density for summer 2010, (c) and (d) are the lower and upper 95% GEE based percentile intervals.



(a)

(b)

(c)

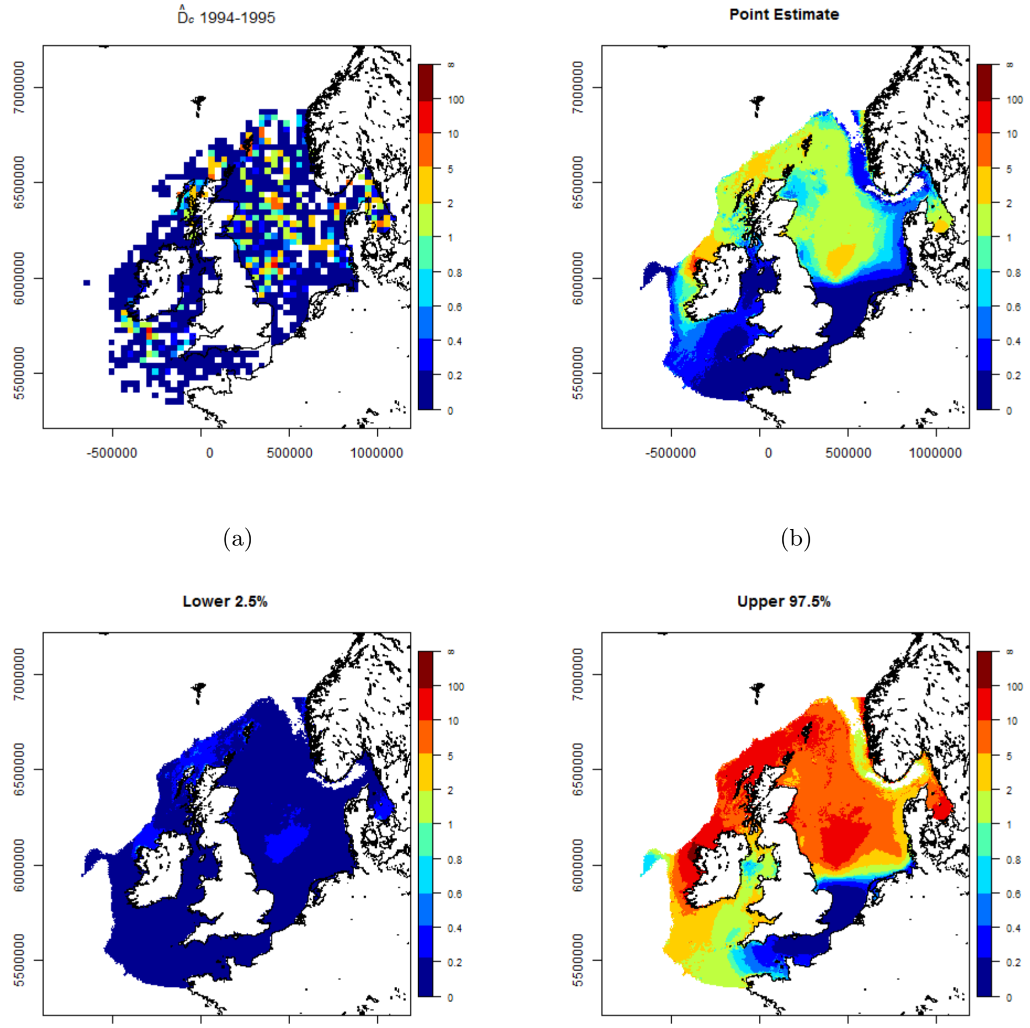
(d)

Figure D.9: Minke whale densities for 2010 in autumn. (a) The raw densities for autumn in 2008 - 2010 that are drawn upon to make predictions for 2010, (b) point estimates of minke whale density for autumn 2010, (c) and (d) are the lower and upper 95% GEE based percentile intervals.

Appendix E

Extra plots of the Joint Cetacean Protocol Analysis - Full time series

E.1 Harbour Porpoise



(a)

(b)

(c)

(d)

Figure E.1: Predicted harbour porpoise densities for summer (day 227) in 1994. (a) The raw densities for summers in 1994 - 1995 that are drawn upon to make predictions for 1994. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.

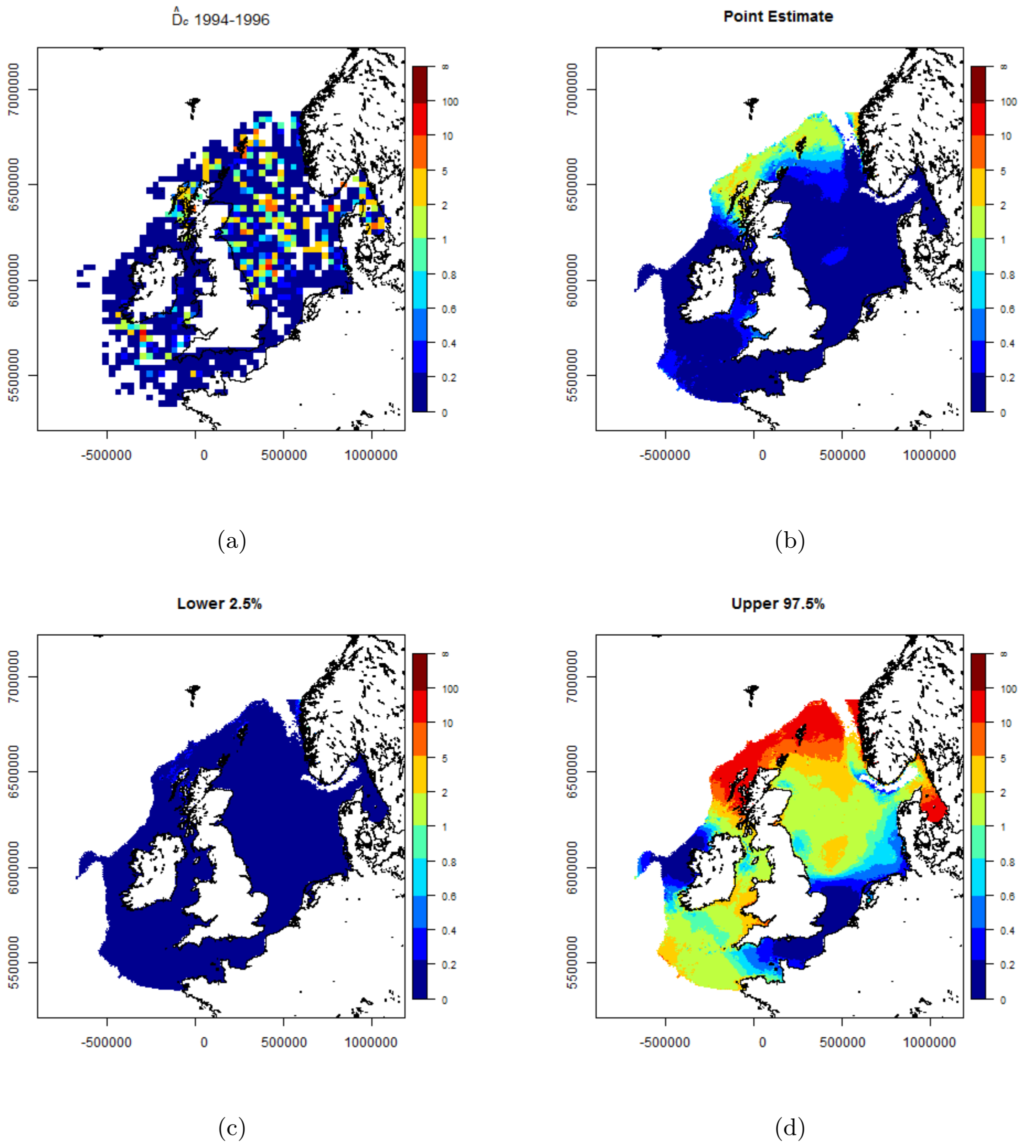
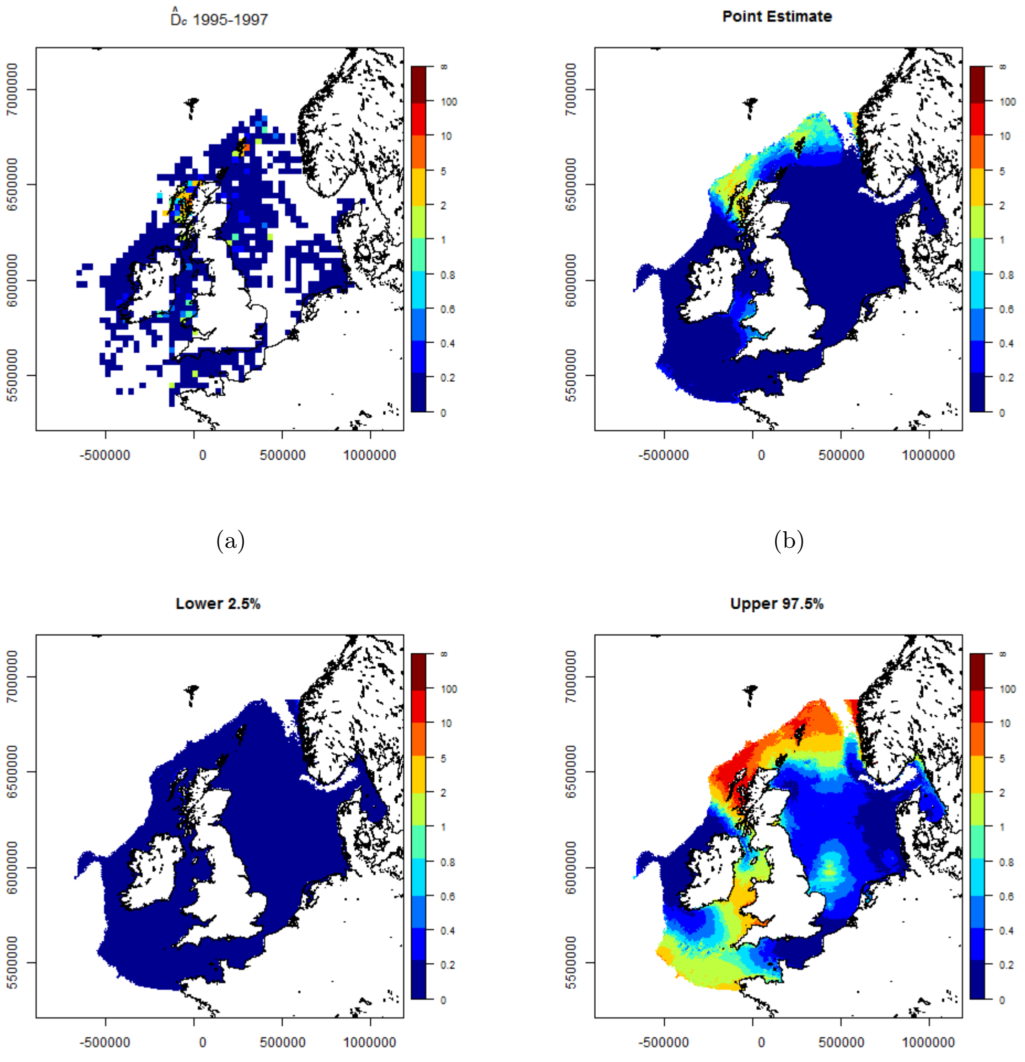


Figure E.2: Predicted harbour porpoise densities for summer (day 227) in 1995. (a) The raw densities for summers in 1994 - 1996 that are drawn upon to make predictions for 1995. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



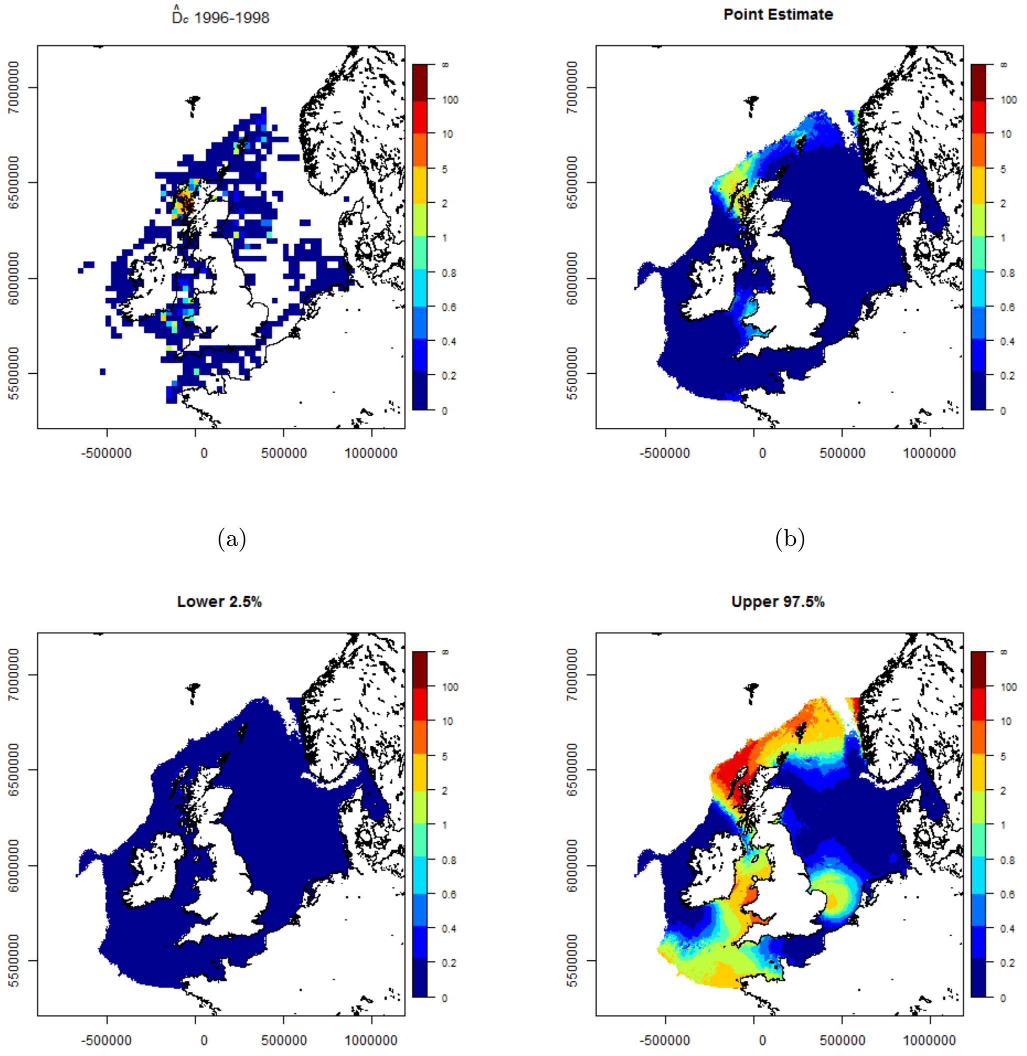
(a)

(b)

(c)

(d)

Figure E.3: Predicted harbour porpoise densities for summer (day 227) in 1996. (a) The raw densities for summers in 1995 - 1997 that are drawn upon to make predictions for 1996. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 96% GEE based percentile intervals.



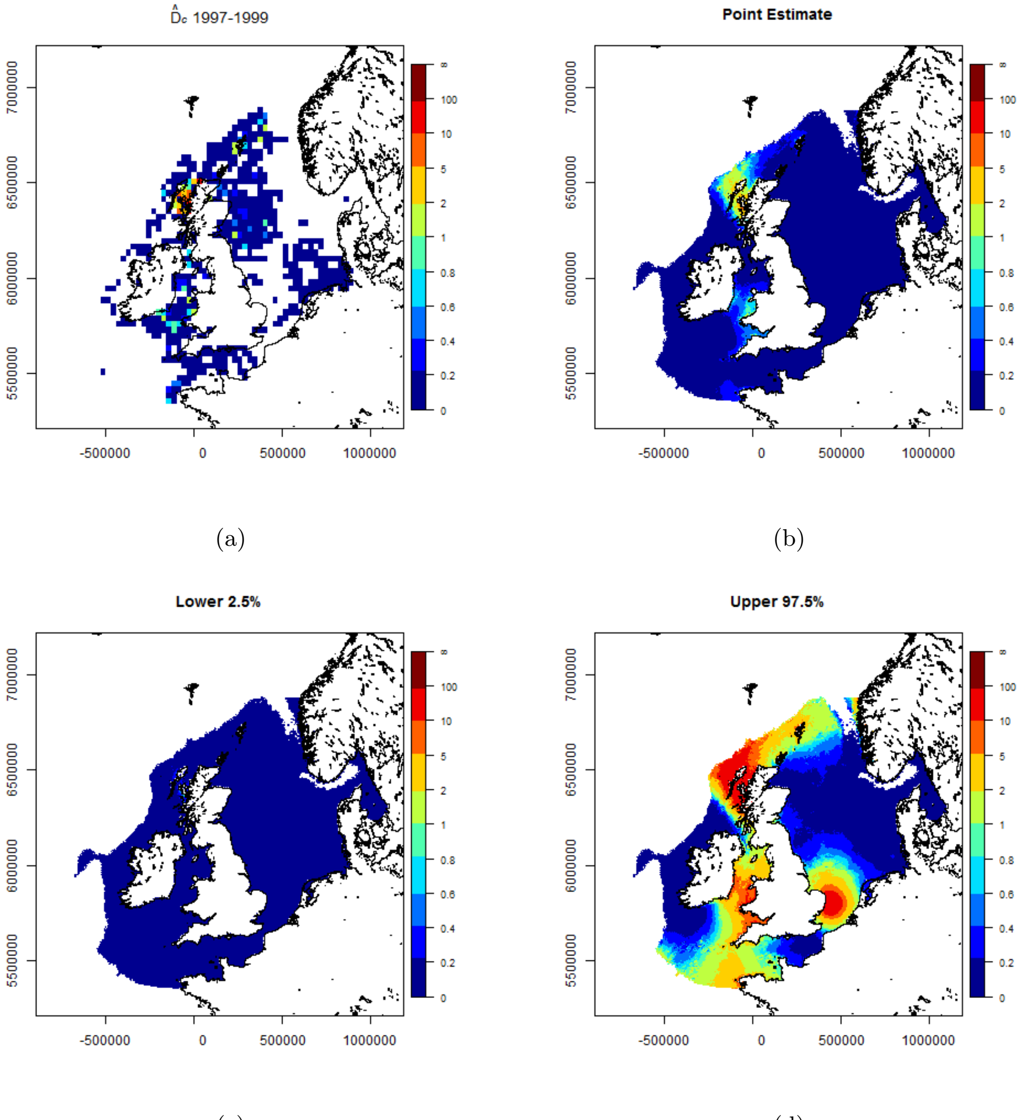
(a)

(b)

(c)

(d)

Figure E.4: Predicted harbour porpoise densities for summer (day 227) in 1997. (a) The raw densities for summers in 1996 - 1998 that are drawn upon to make predictions for 1997. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 97% GEE based percentile intervals.



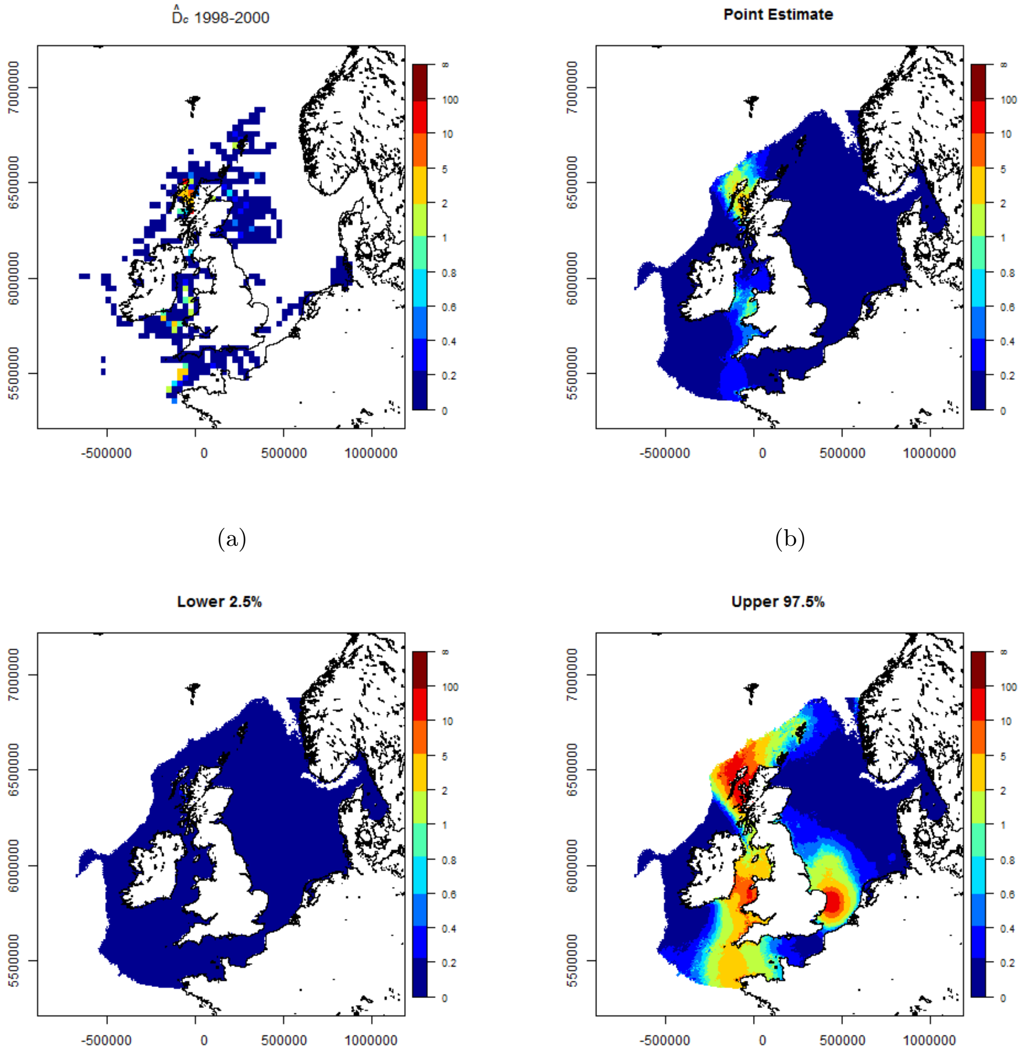
(a)

(b)

(c)

(d)

Figure E.5: Predicted harbour porpoise densities for summer (day 227) in 1998. (a) The raw densities for summers in 1997 - 1999 that are drawn upon to make predictions for 1998. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 98% GEE based percentile intervals.



$\hat{D}_c$ 1998-2000

Point Estimate

Lower 2.5%

Upper 97.5%

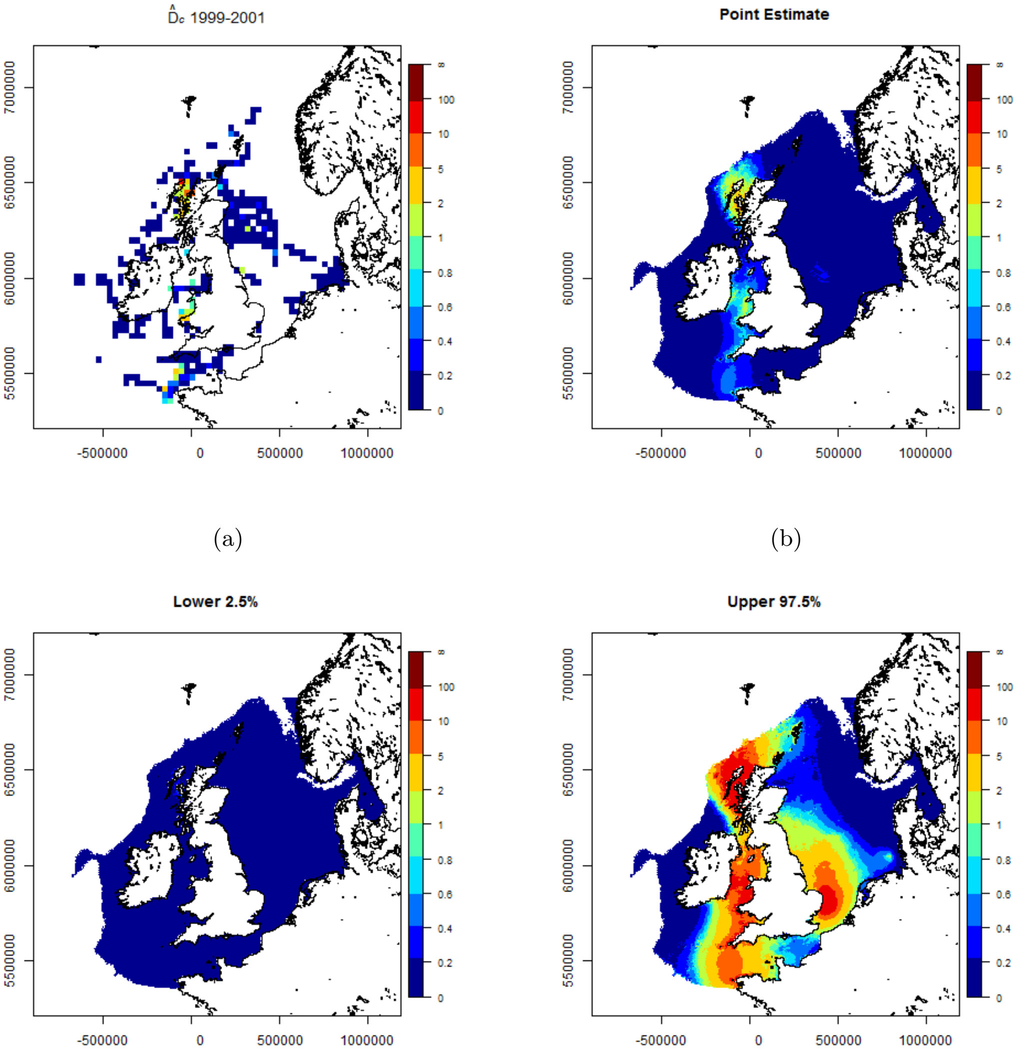
(a)

(b)

(c)

(d)

Figure E.6: Predicted harbour porpoise densities for summer (day 227) in 1999. (a) The raw densities for summers in 1998 - 2000 that are drawn upon to make predictions for 1999. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 99% GEE based percentile intervals.



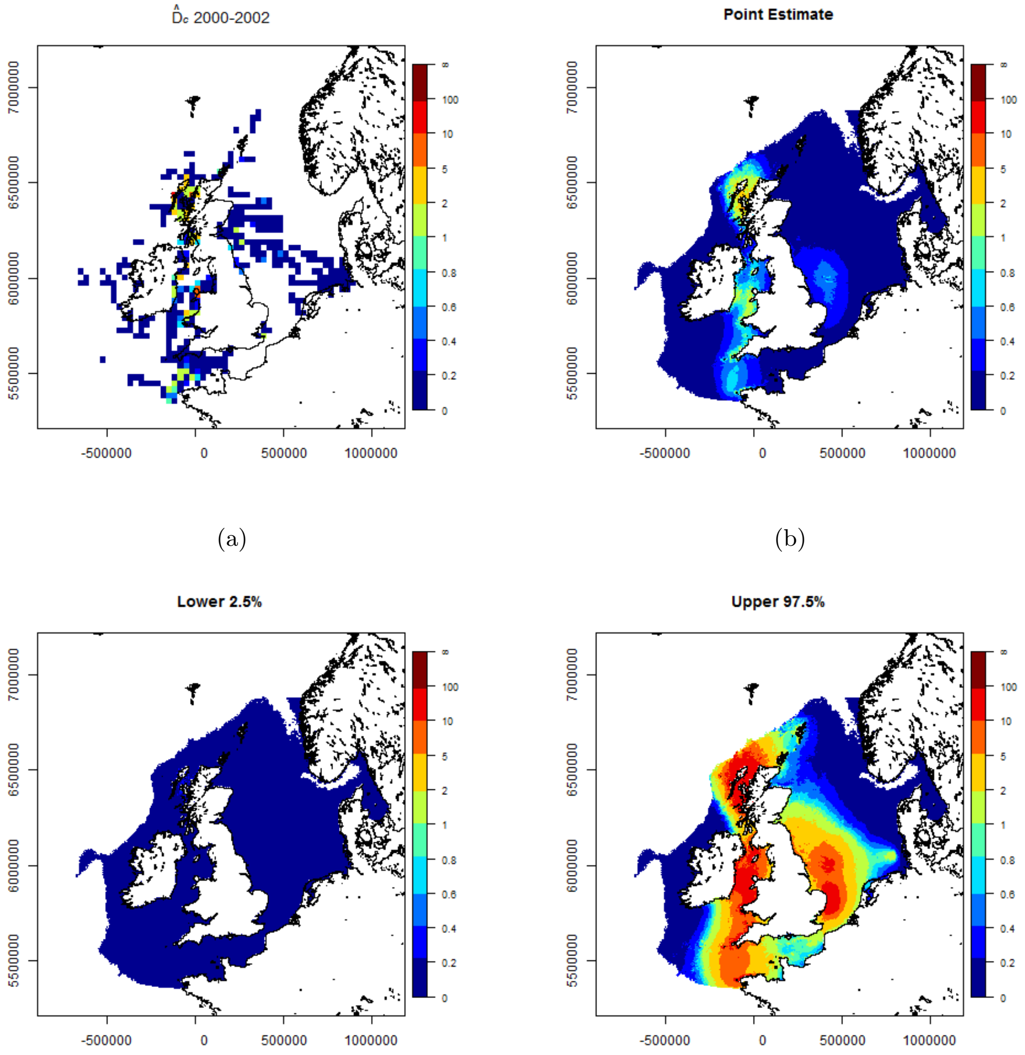
(a)

(b)

(c)

(d)

Figure E.7: Predicted harbour porpoise densities for summer (day 227) in 2000. (a) The raw densities for summers in 1999 - 2001 that are drawn upon to make predictions for 2000. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



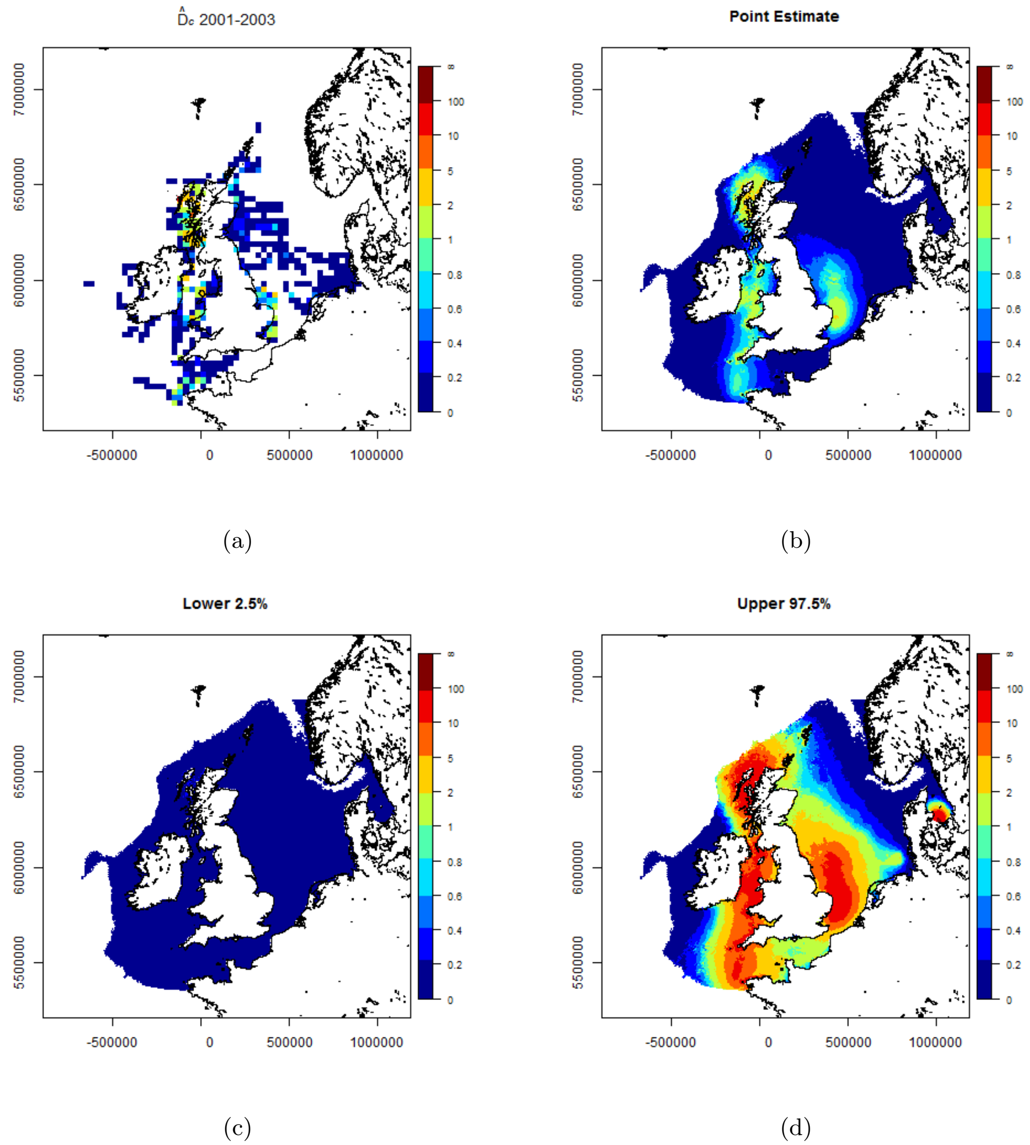
(a)

(b)

(c)

(d)

Figure E.8: Predicted harbour porpoise densities for summer (day 227) in 2001. (a) The raw densities for summers in 2000 - 2002 that are drawn upon to make predictions for 2001. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



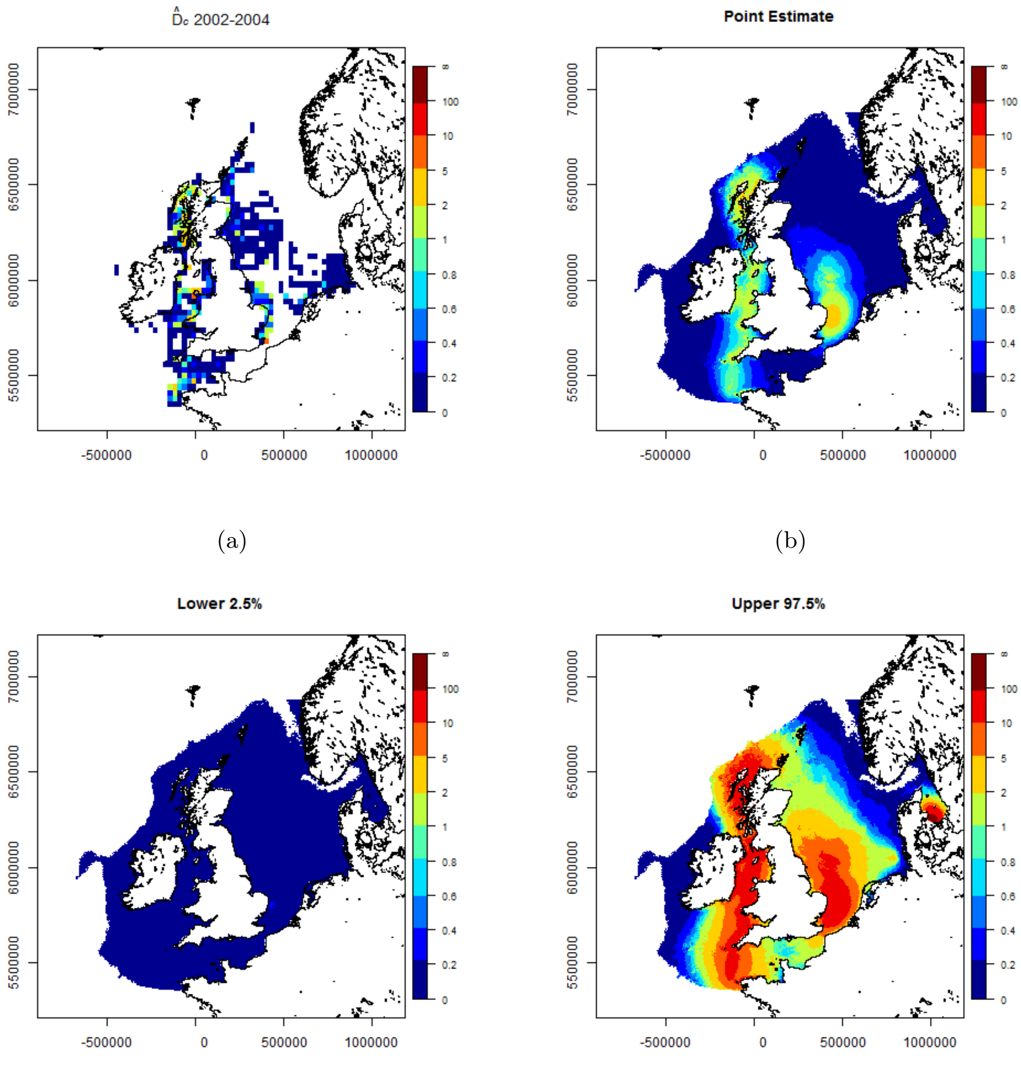
(a)

(b)

(c)

(d)

Figure E.9: Predicted harbour porpoise densities for summer (day 227) in 2002. (a) The raw densities for summers in 2001 - 2003 that are drawn upon to make predictions for 2002. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



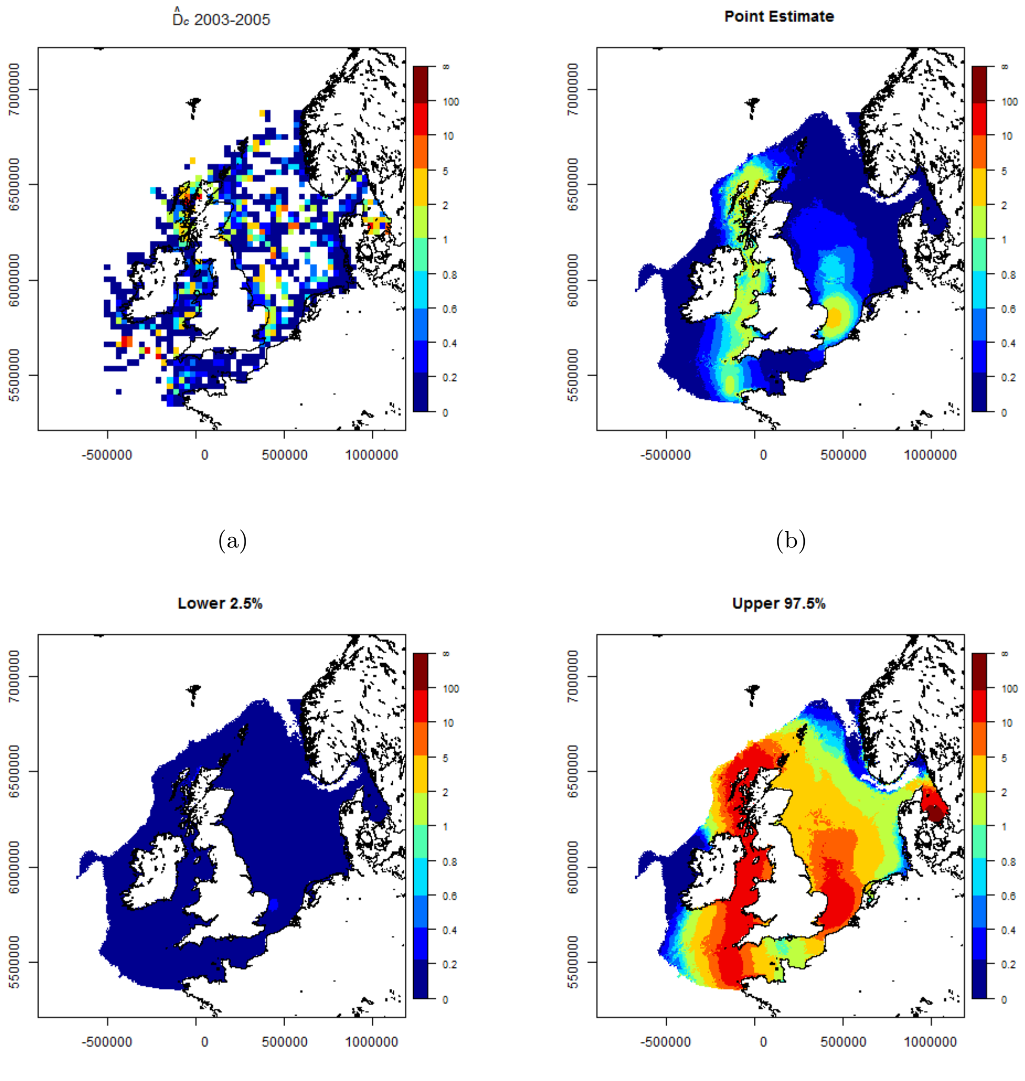
(a)

(b)

(c)

(d)

Figure E.10: Predicted harbour porpoise densities for summer (day 227) in 2003. (a) The raw densities for summers in 2002 - 2004 that are drawn upon to make predictions for 2003. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



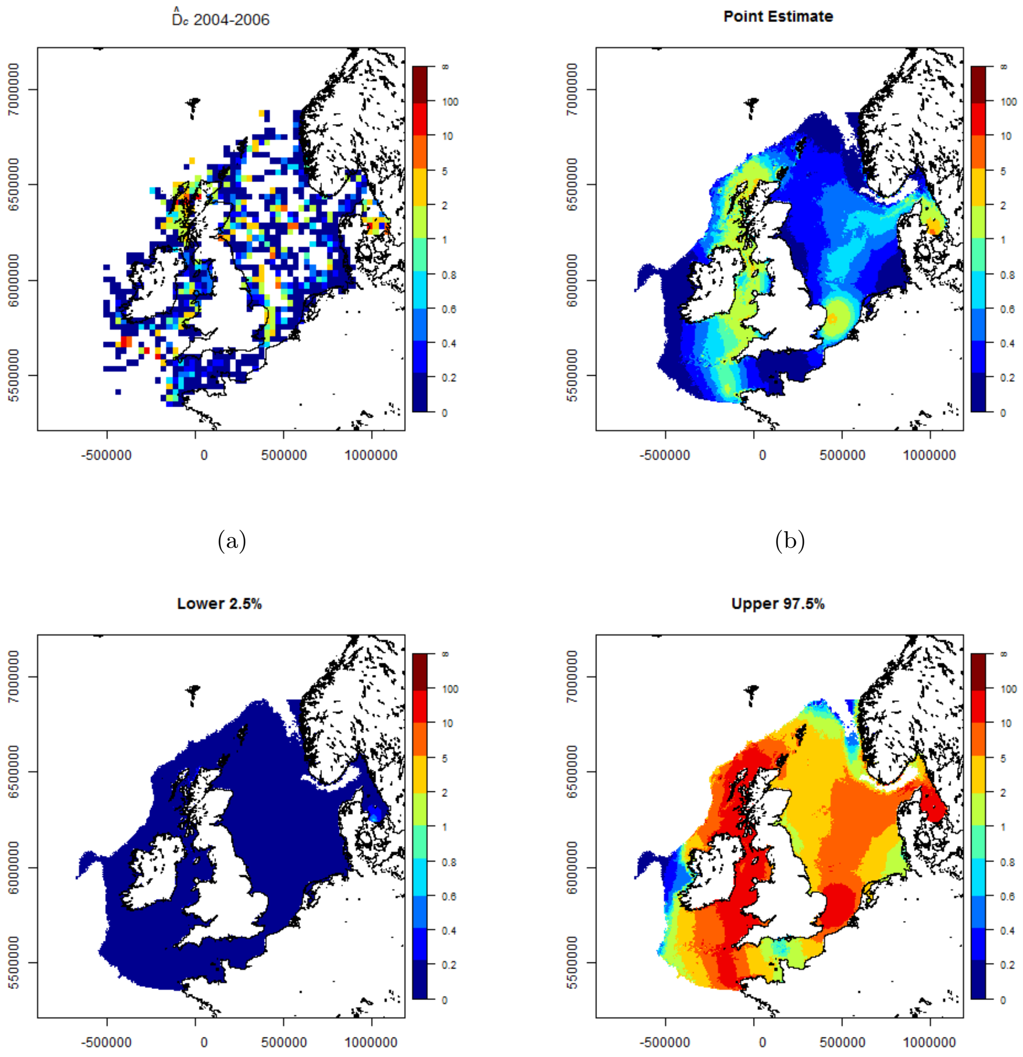
(a)

(b)

(c)

(d)

Figure E.11: Predicted harbour porpoise densities for summer (day 227) in 2004. (a) The raw densities for summers in 2003 - 2005 that are drawn upon to make predictions for 2004. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



$\hat{D}_c$ 2004-2006

Point Estimate

Lower 2.5%

Upper 97.5%

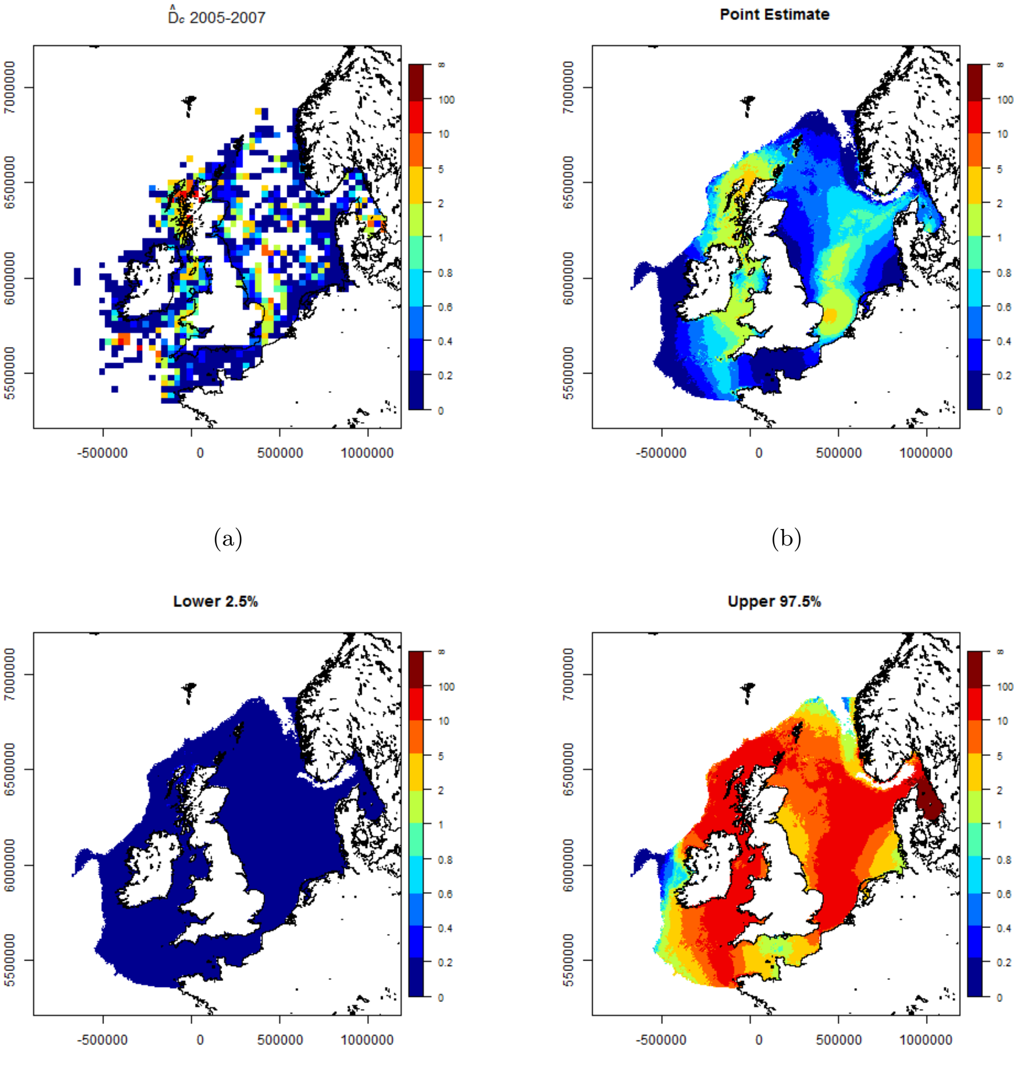
(a)

(b)

(c)

(d)

Figure E.12: Predicted harbour porpoise densities for summer (day 227) in 2005. (a) The raw densities for summers in 2004 - 2006 that are drawn upon to make predictions for 2005. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



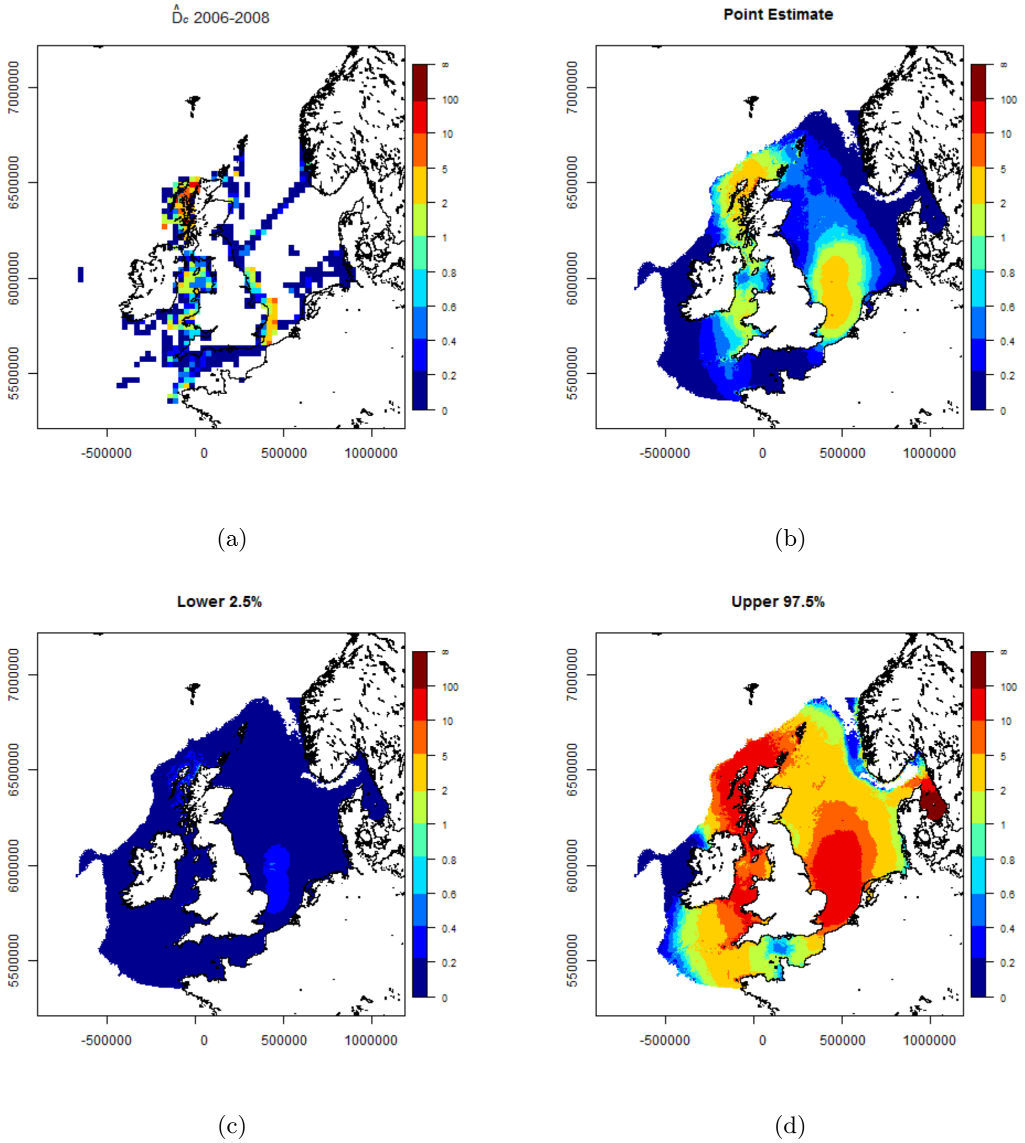
(a)

(b)

(c)

(d)

Figure E.13: Predicted harbour porpoise densities for summer (day 227) in 2006. (a) The raw densities for summers in 2005 - 2007 that are drawn upon to make predictions for 2006. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



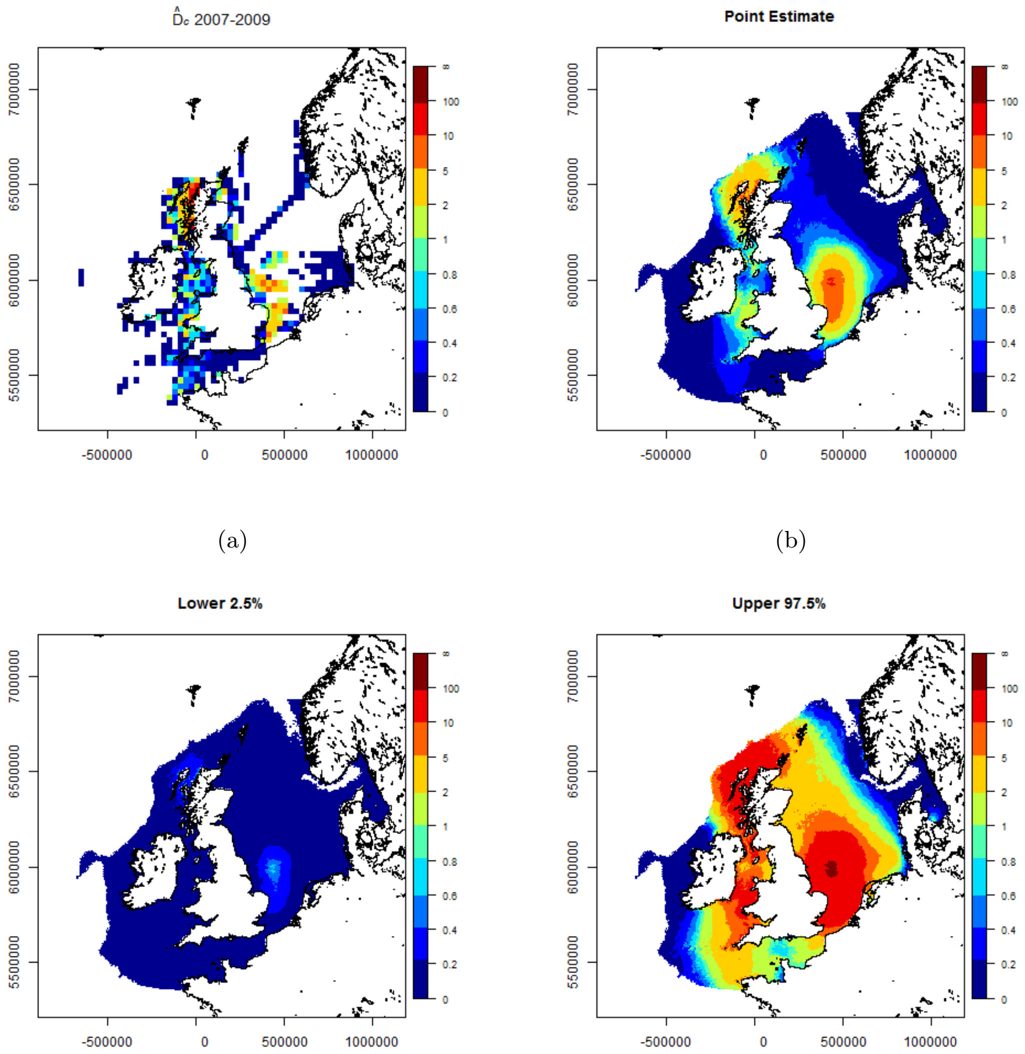
(a)

(b)

(c)

(d)

Figure E.14: Predicted harbour porpoise densities for summer (day 227) in 2007. (a) The raw densities for summers in 2006 - 2008 that are drawn upon to make predictions for 2007. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



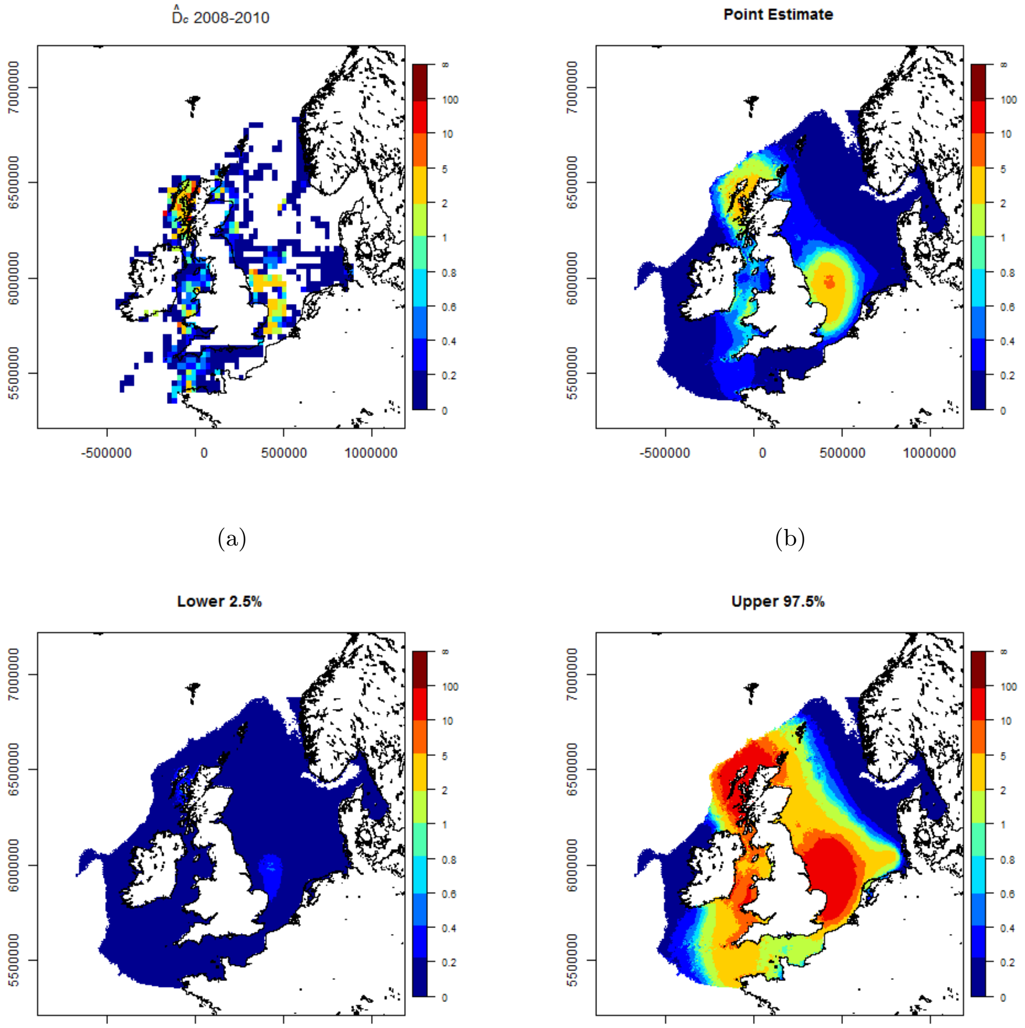
(a)

(b)

(c)

(d)

Figure E.15: Predicted harbour porpoise densities for summer (day 227) in 2008. (a) The raw densities for summers in 2007 - 2009 that are drawn upon to make predictions for 2008. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



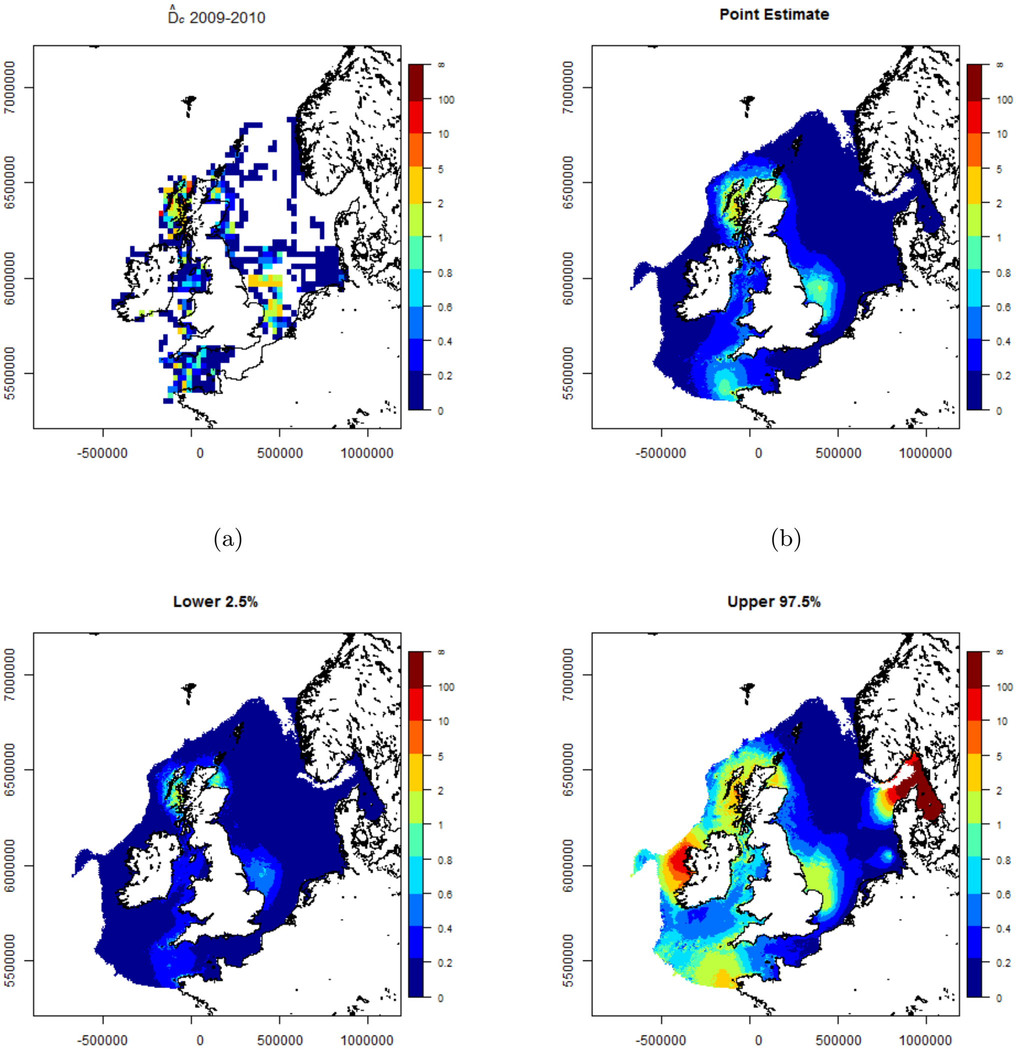
(a)

(b)

(c)

(d)

Figure E.16: Predicted harbour porpoise densities for summer (day 227) in 2009. (a) The raw densities for summers in 2008 - 2010 that are drawn upon to make predictions for 2009. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



(a)

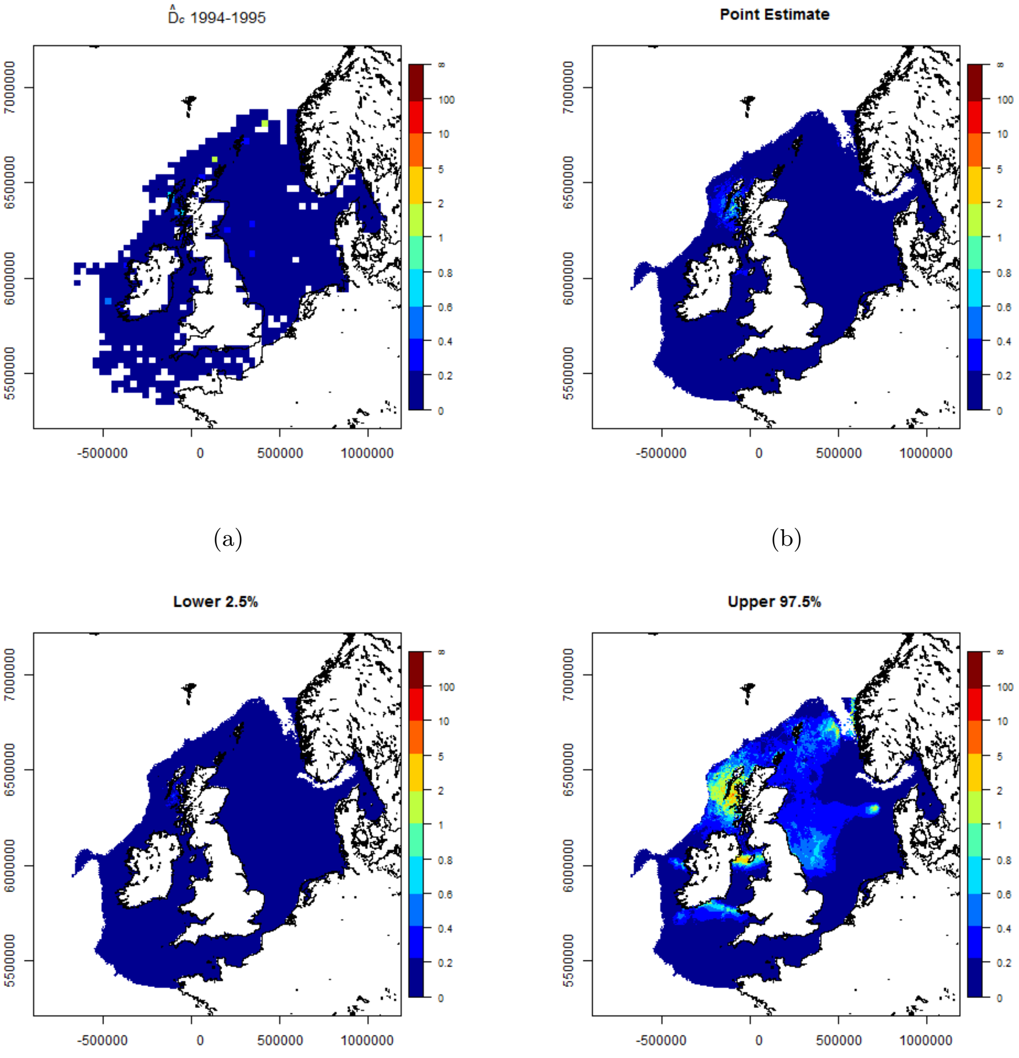
(b)

(c)

(d)

Figure E.17: Predicted harbour porpoise densities for summer (day 227) in 2010. (a) The raw densities for summers in 2009 - 2010 that are drawn upon to make predictions for 2010. (b) Point estimates of harbour porpoise density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.

E.2 Minke Whale



(a)

(b)

(c)

(d)

Figure E.18: Predicted minke whale densities for summer (day 227) in 1994. (a) The raw densities for summers in 1994 - 1995 that are drawn upon to make predictions for 1994. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.

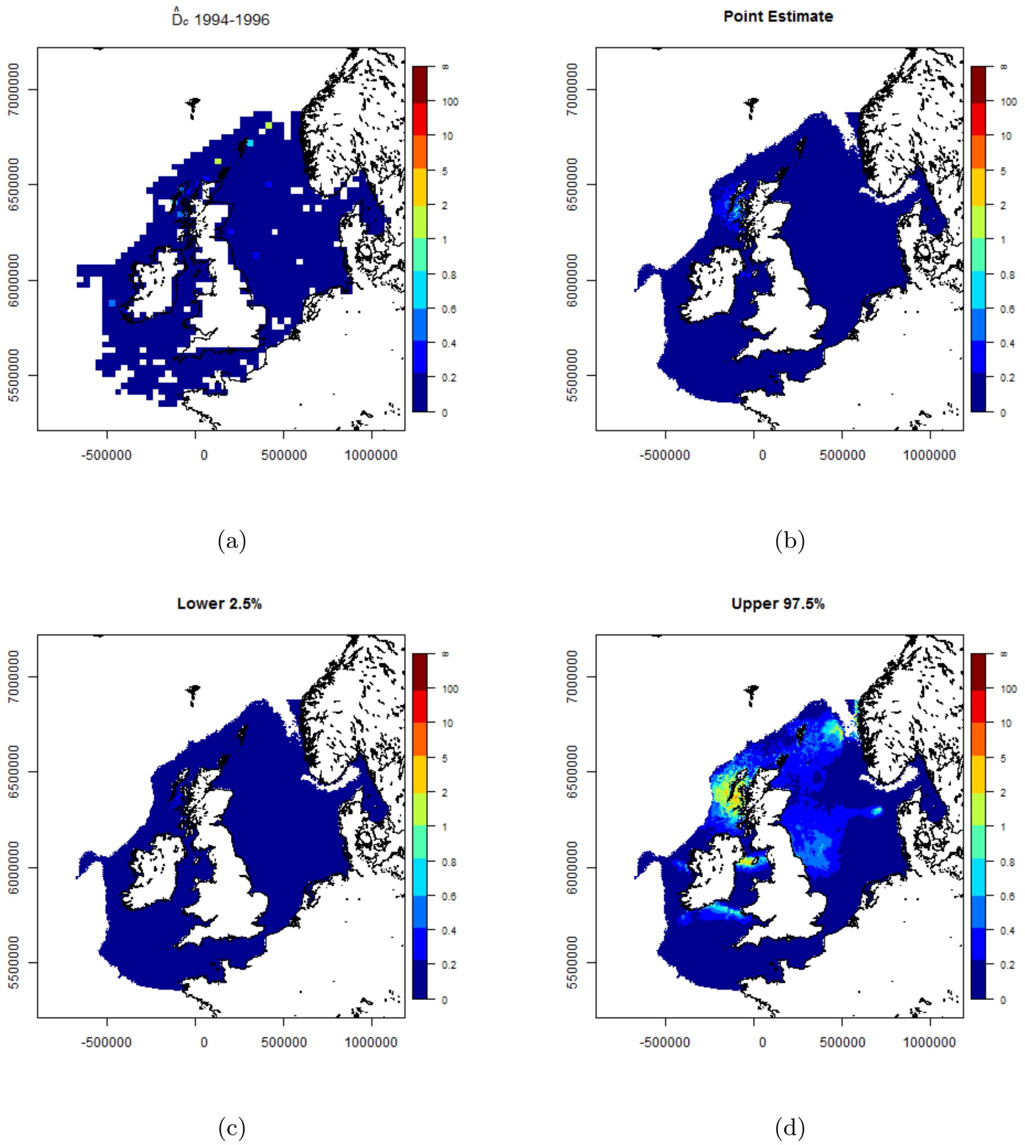
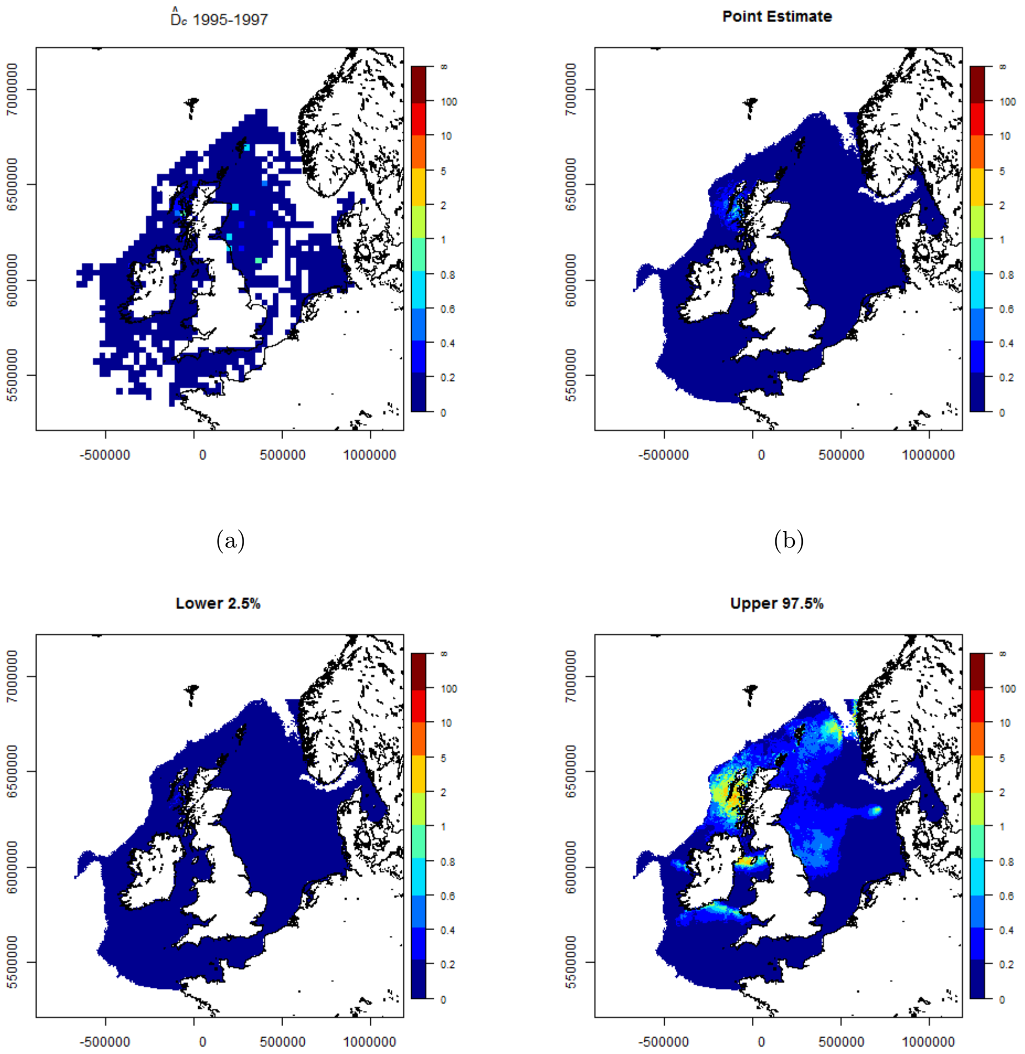


Figure E.19: Predicted minke whale densities for summer (day 227) in 1995. (a) The raw densities for summers in 1994 - 1996 that are drawn upon to make predictions for 1995. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



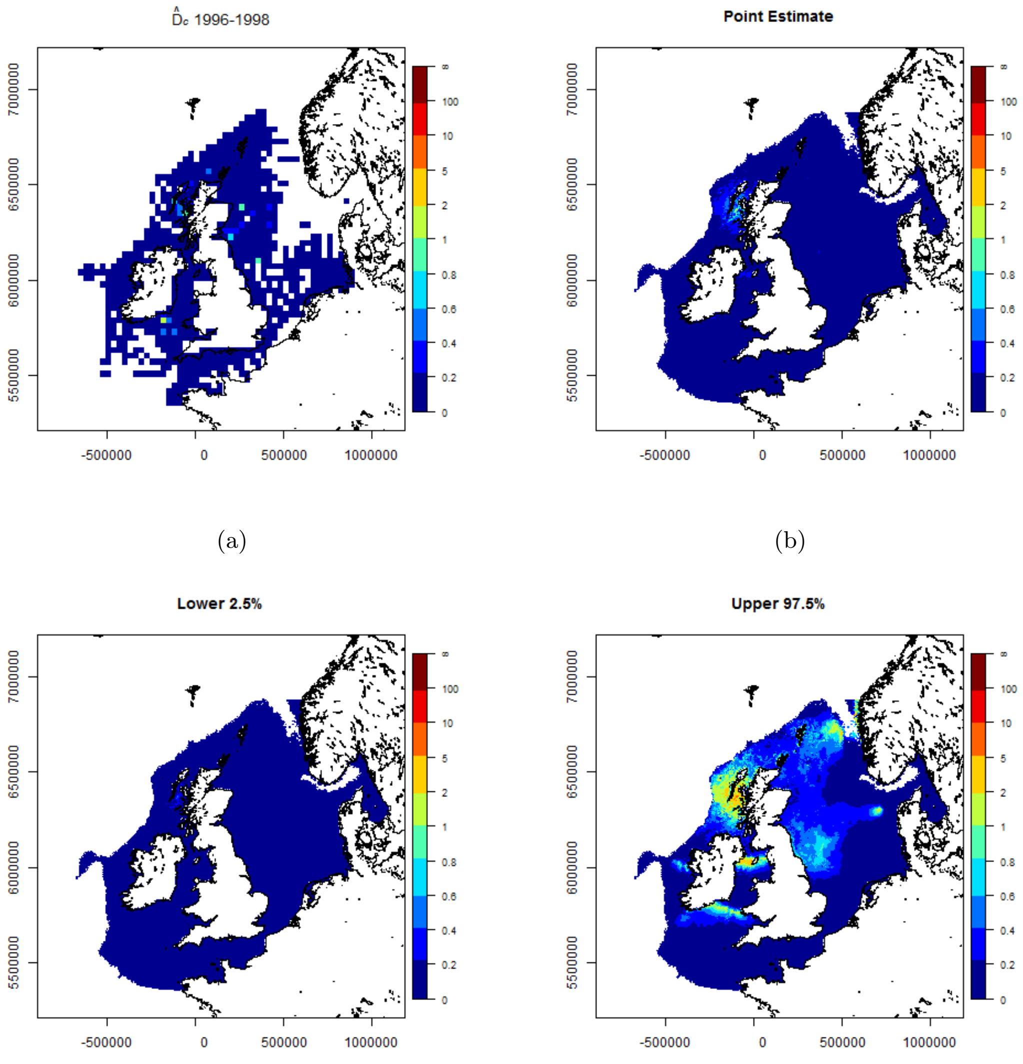
(a)

(b)

(c)

(d)

Figure E.20: Predicted minke whale densities for summer (day 227) in 1996. (a) The raw densities for summers in 1995 - 1997 that are drawn upon to make predictions for 1996. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 96% GEE based percentile intervals.



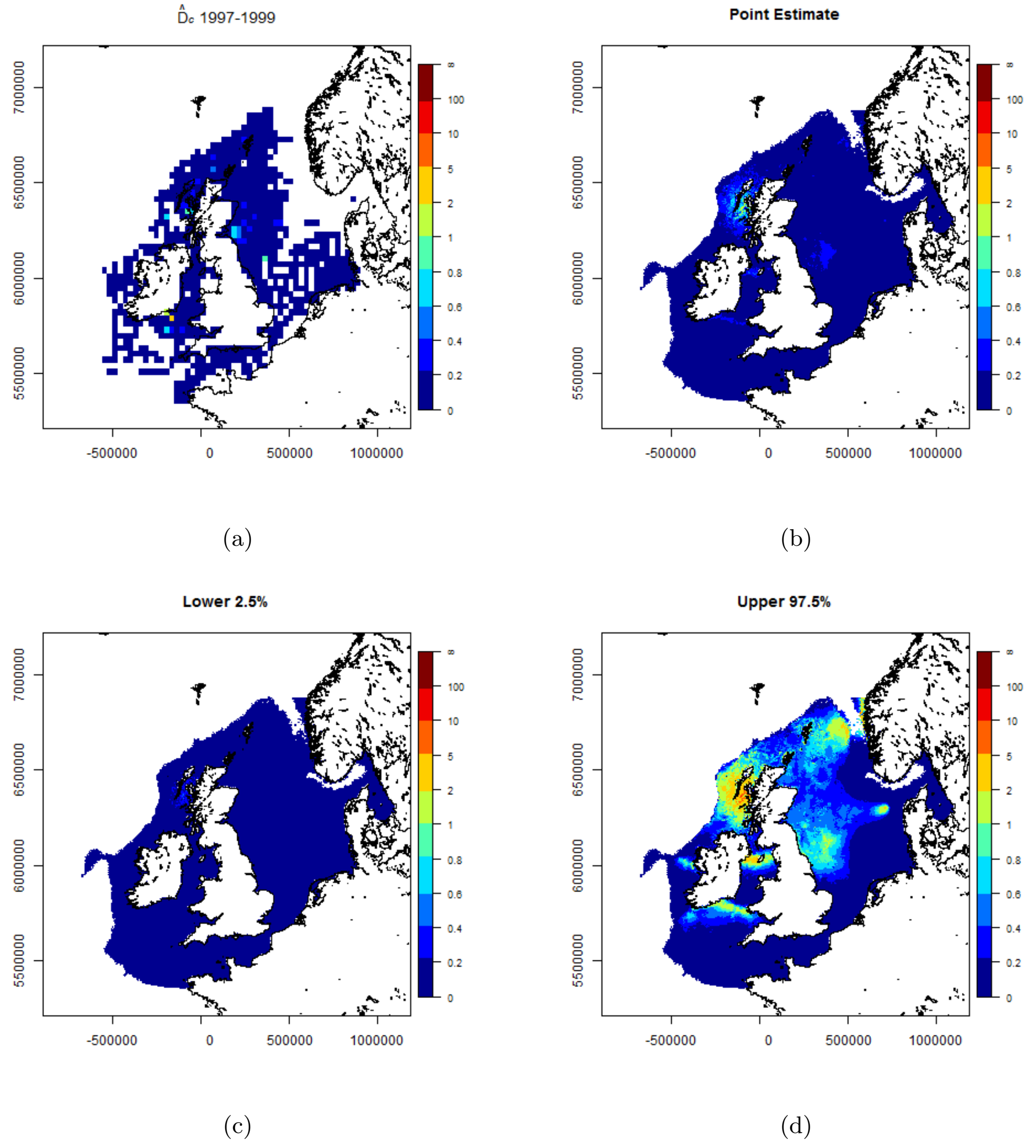
(a)

(b)

(c)

(d)

Figure E.21: Predicted minke whale densities for summer (day 227) in 1997. (a) The raw densities for summers in 1996 - 1998 that are drawn upon to make predictions for 1997. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 97% GEE based percentile intervals.



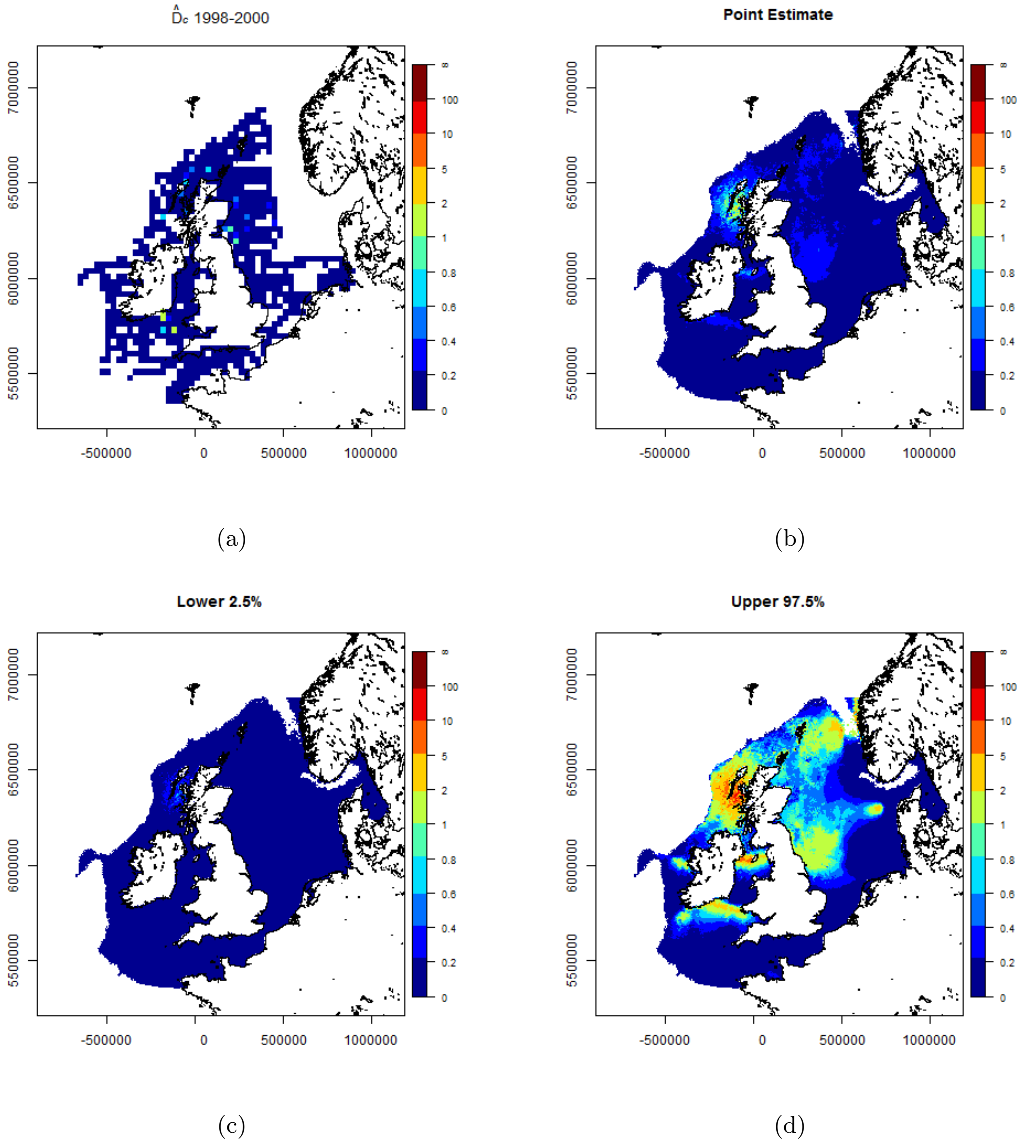
(a)

(b)

(c)

(d)

Figure E.22: Predicted minke whale densities for summer (day 227) in 1998. (a) The raw densities for summers in 1997 - 1999 that are drawn upon to make predictions for 1998. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 98% GEE based percentile intervals.



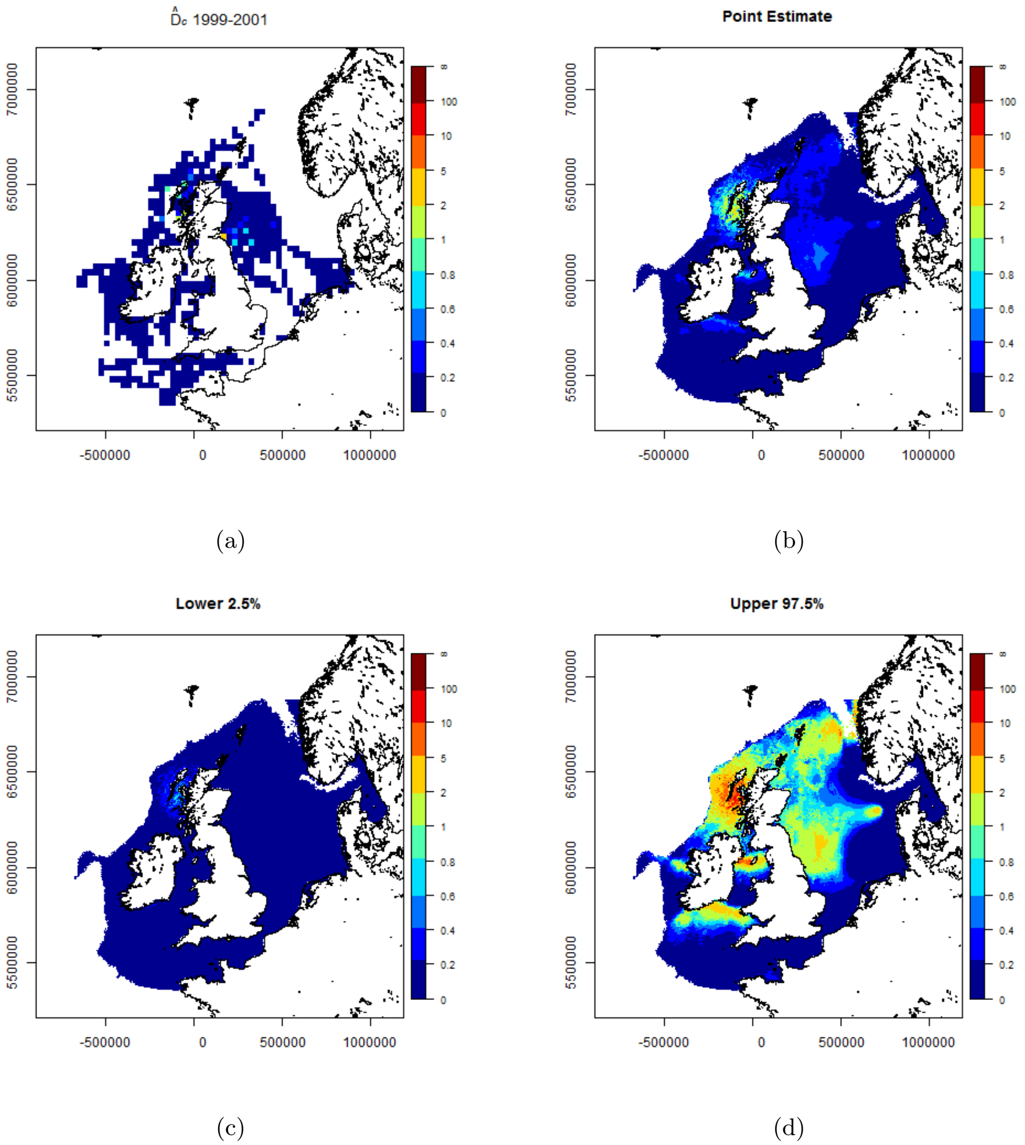
(a)

(b)

(c)

(d)

Figure E.23: Predicted minke whale densities for summer (day 227) in 1999. (a) The raw densities for summers in 1998 - 2000 that are drawn upon to make predictions for 1999. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 99% GEE based percentile intervals.



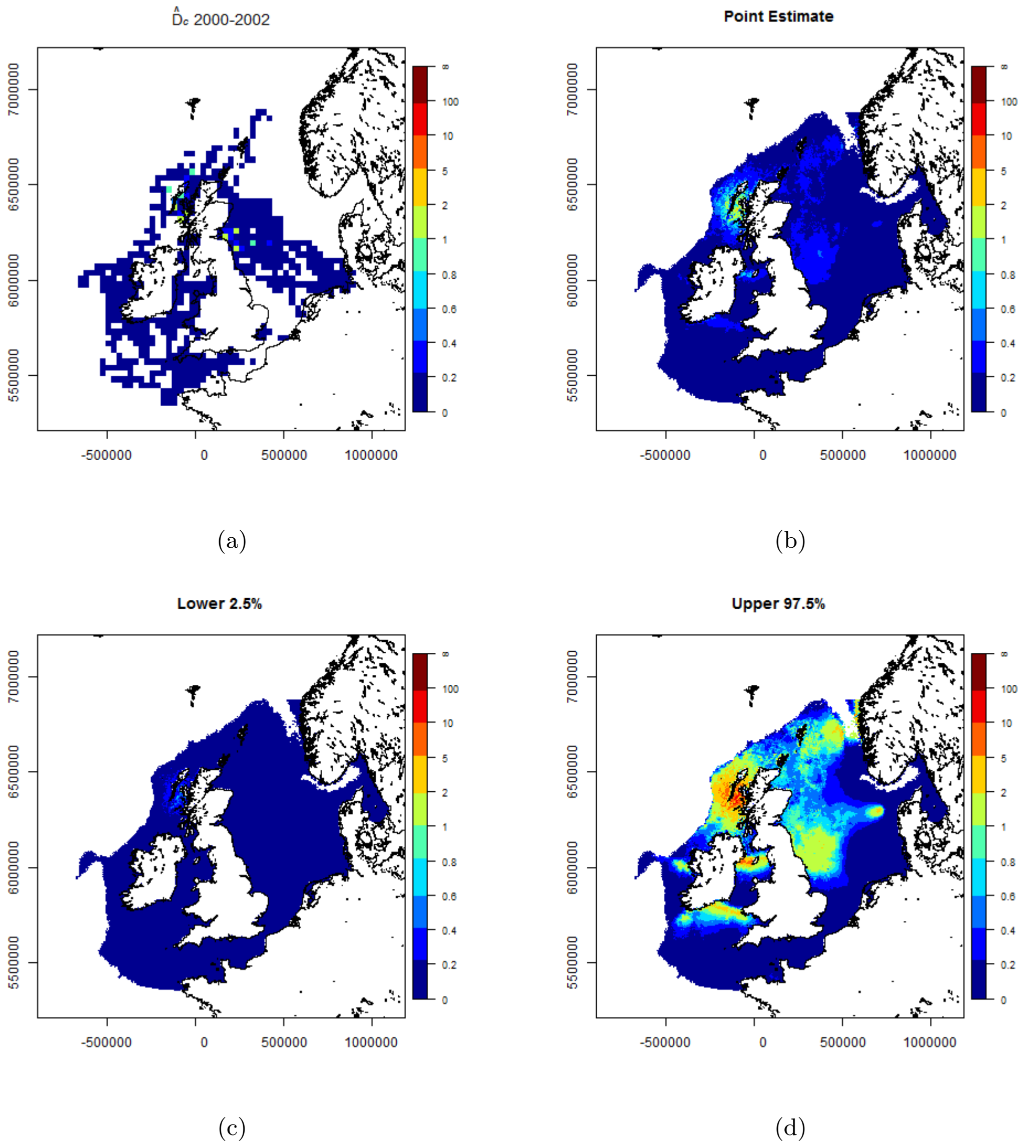
(a)

(b)

(c)

(d)

Figure E.24: Predicted minke whale densities for summer (day 227) in 2000. (a) The raw densities for summers in 1999 - 2001 that are drawn upon to make predictions for 2000. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



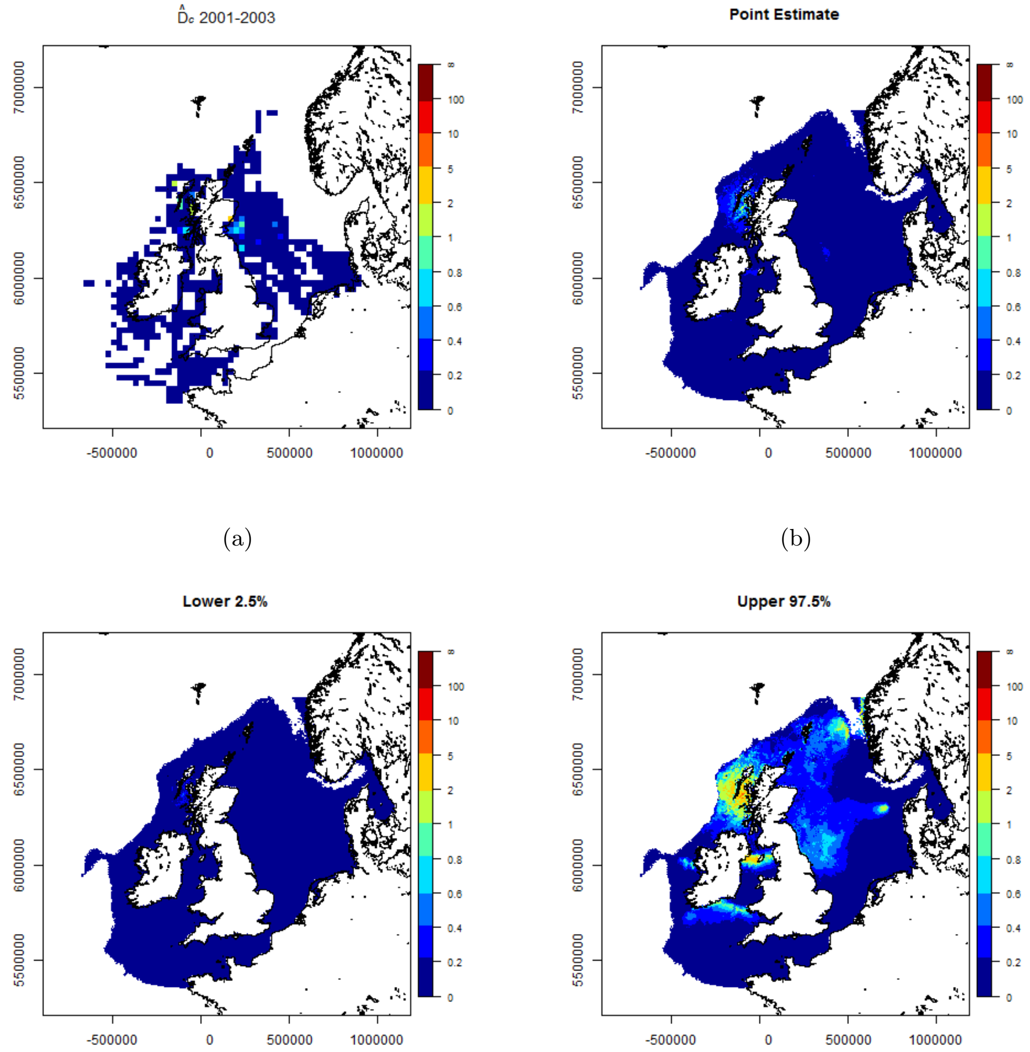
(a)

(b)

(c)

(d)

Figure E.25: Predicted minke whale densities for summer (day 227) in 2001. (a) The raw densities for summers in 2000 - 2002 that are drawn upon to make predictions for 2001. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



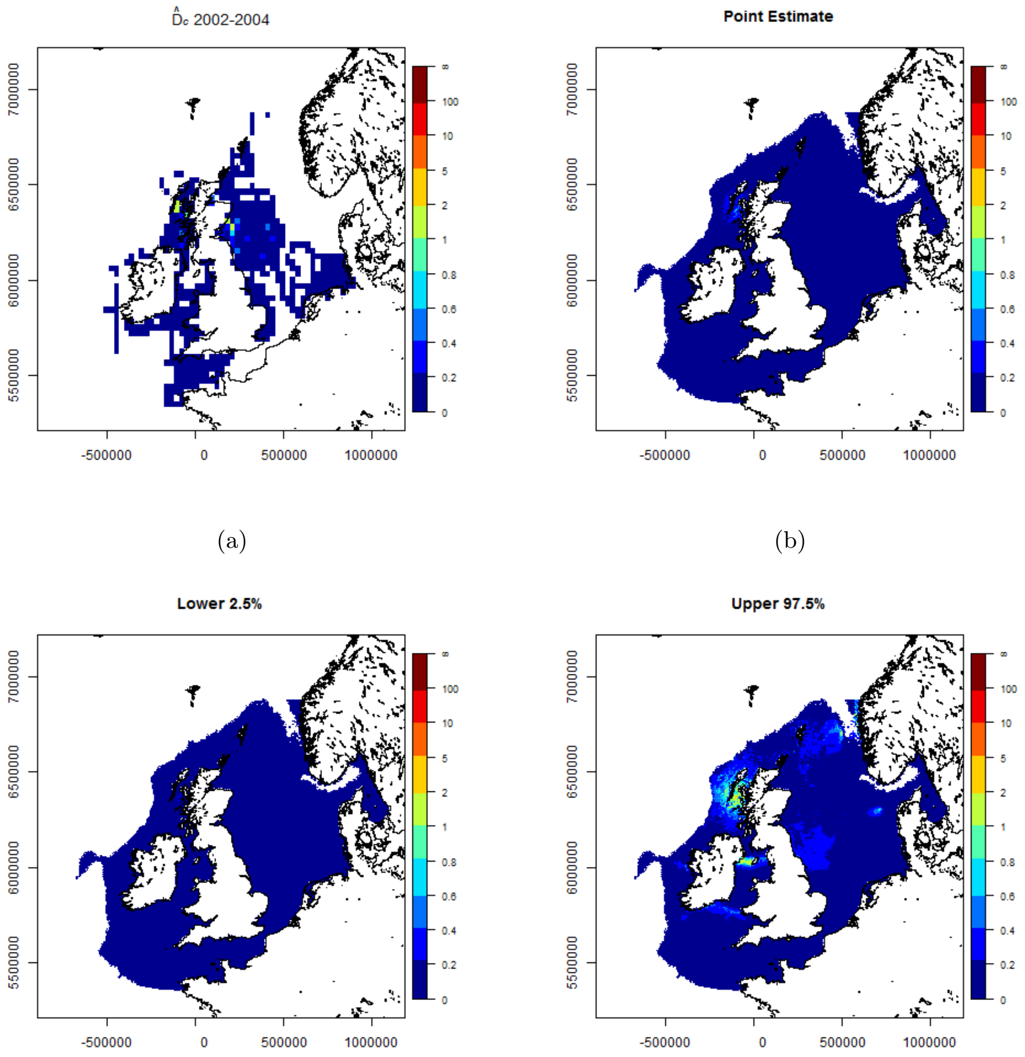
(a)

(b)

(c)

(d)

Figure E.26: Predicted minke whale densities for summer (day 227) in 2002. (a) The raw densities for summers in 2001 - 2003 that are drawn upon to make predictions for 2002. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



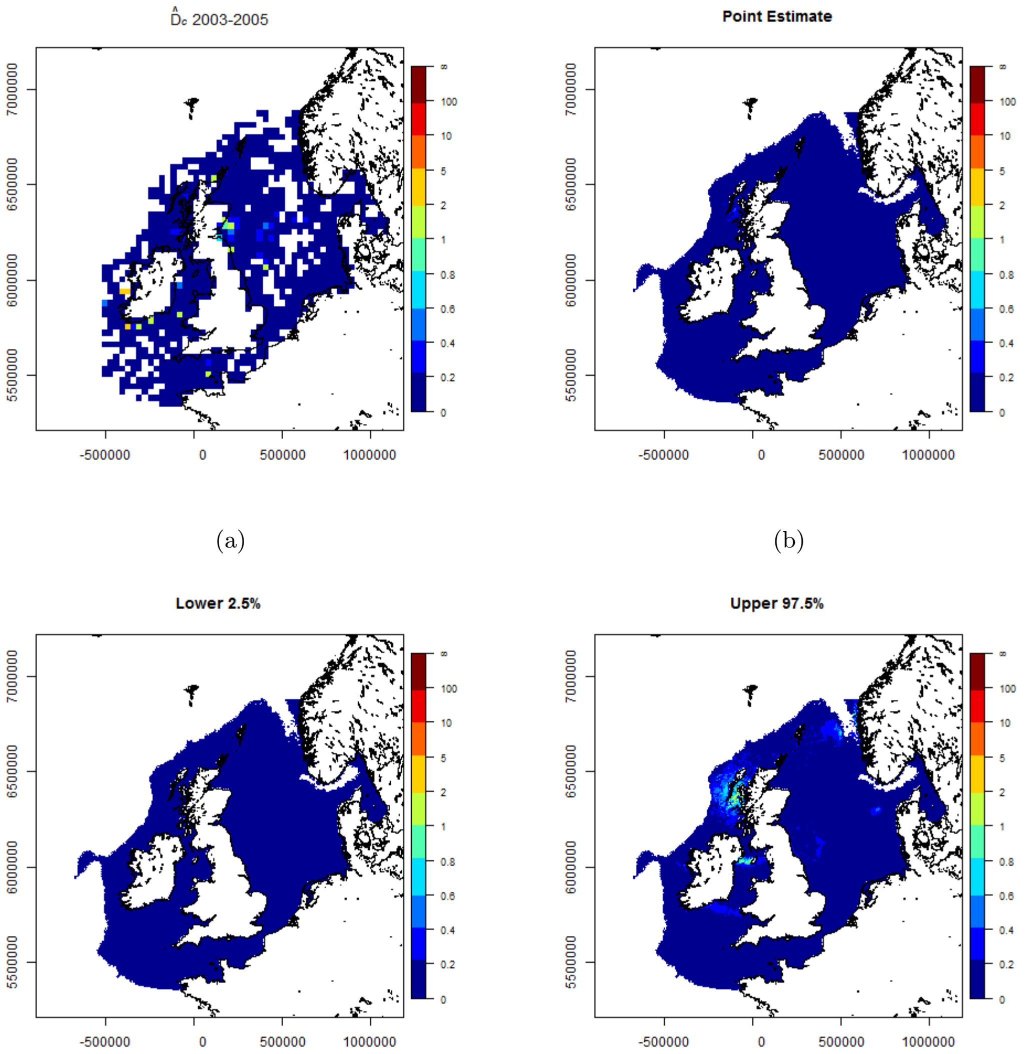
(a)

(b)

(c)

(d)

Figure E.27: Predicted minke whale densities for summer (day 227) in 2003. (a) The raw densities for summers in 2002 - 2004 that are drawn upon to make predictions for 2003. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



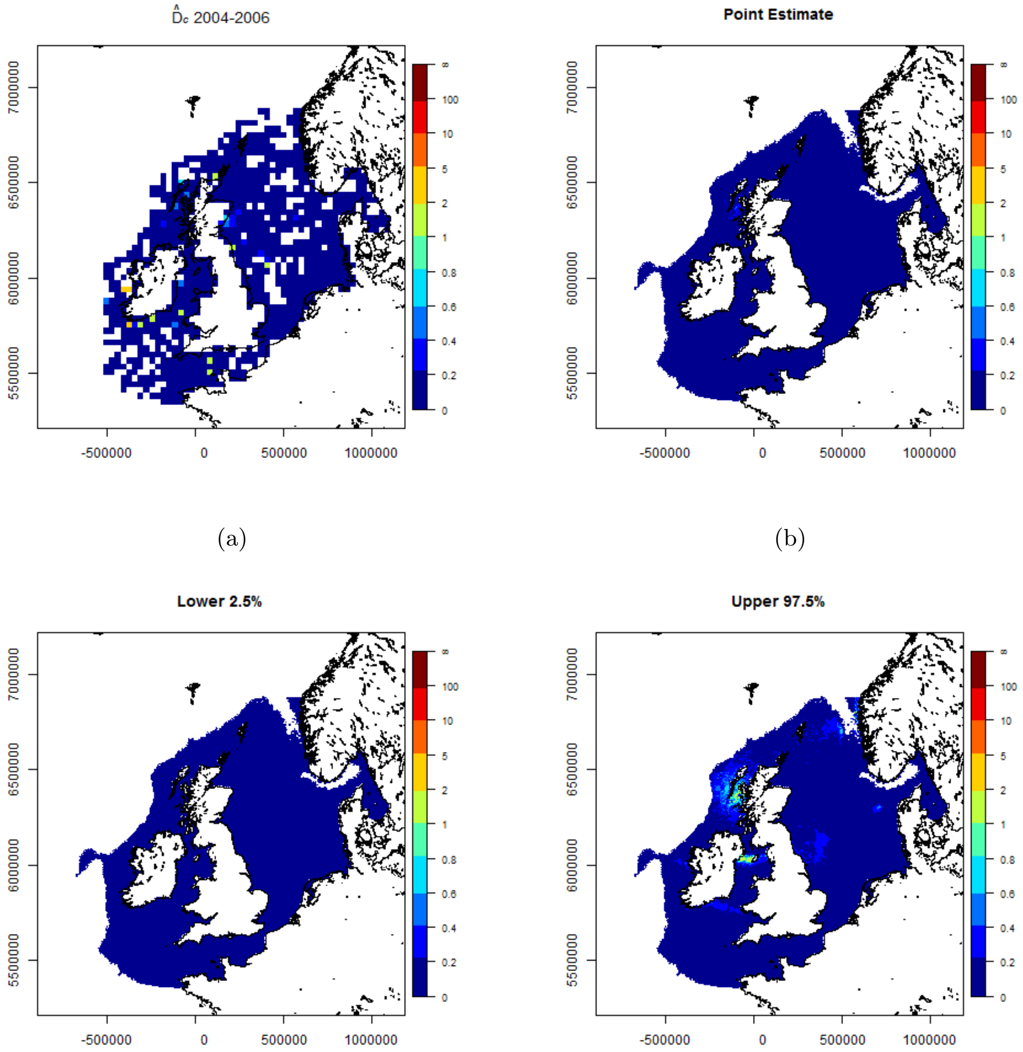
(a)

(b)

(c)

(d)

Figure E.28: Predicted minke whale densities for summer (day 227) in 2004. (a) The raw densities for summers in 2003 - 2005 that are drawn upon to make predictions for 2004. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



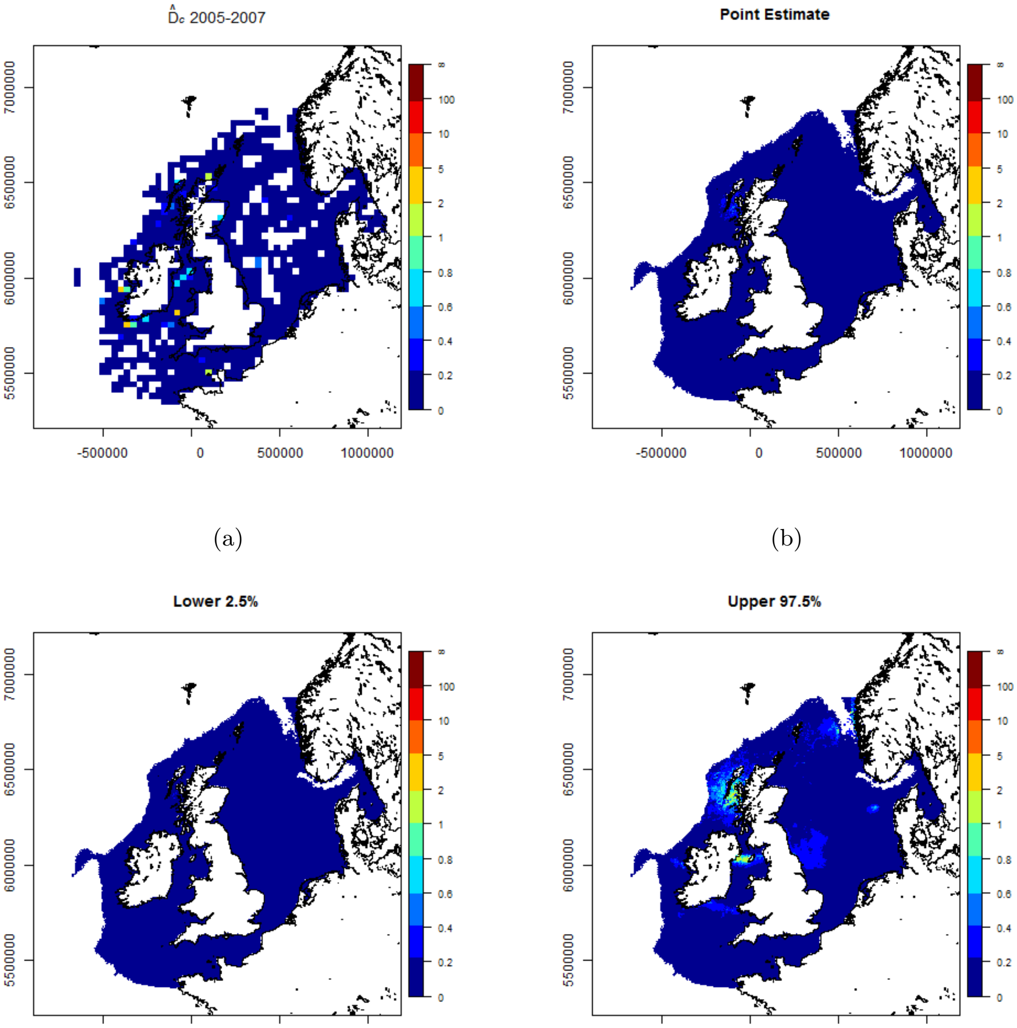
(a)

(b)

(c)

(d)

Figure E.29: Predicted minke whale densities for summer (day 227) in 2005. (a) The raw densities for summers in 2004 - 2006 that are drawn upon to make predictions for 2005. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



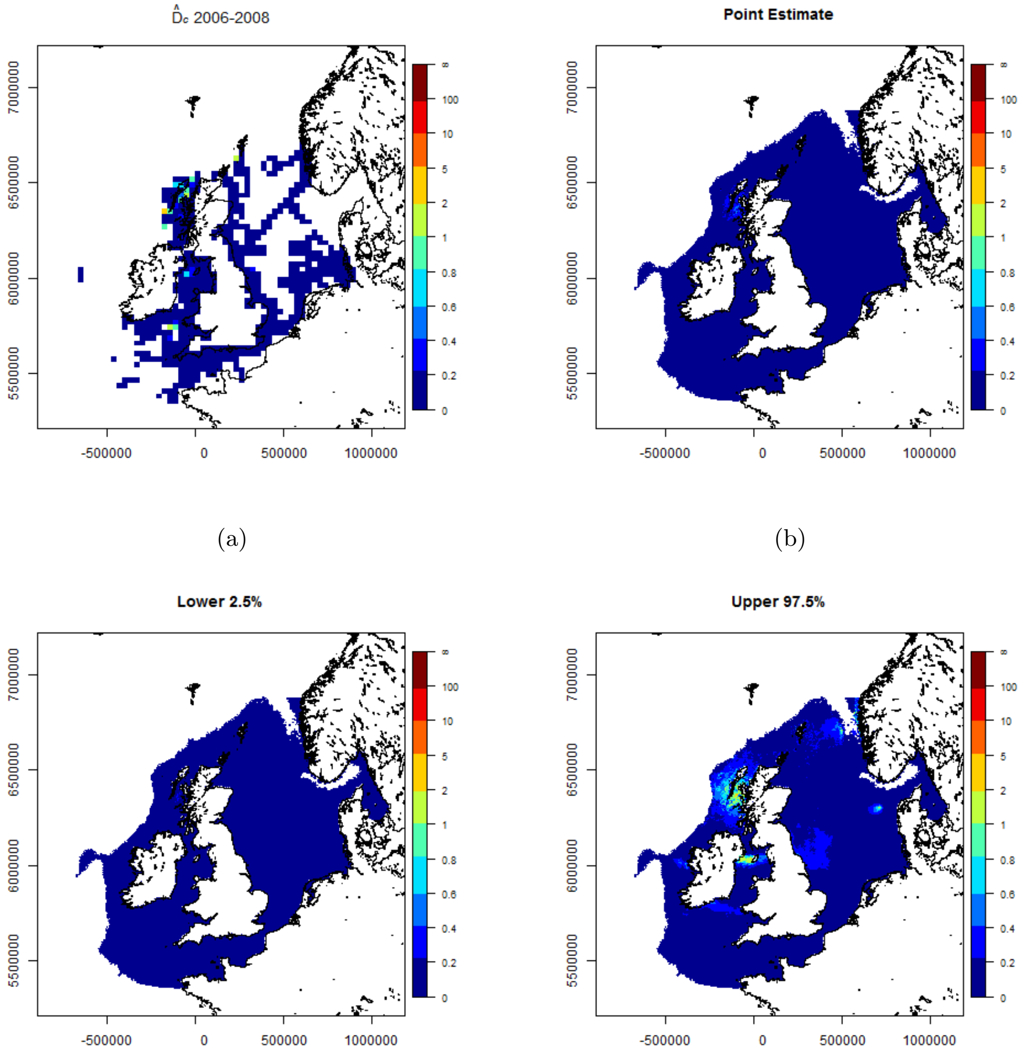
(a)

(b)

(c)

(d)

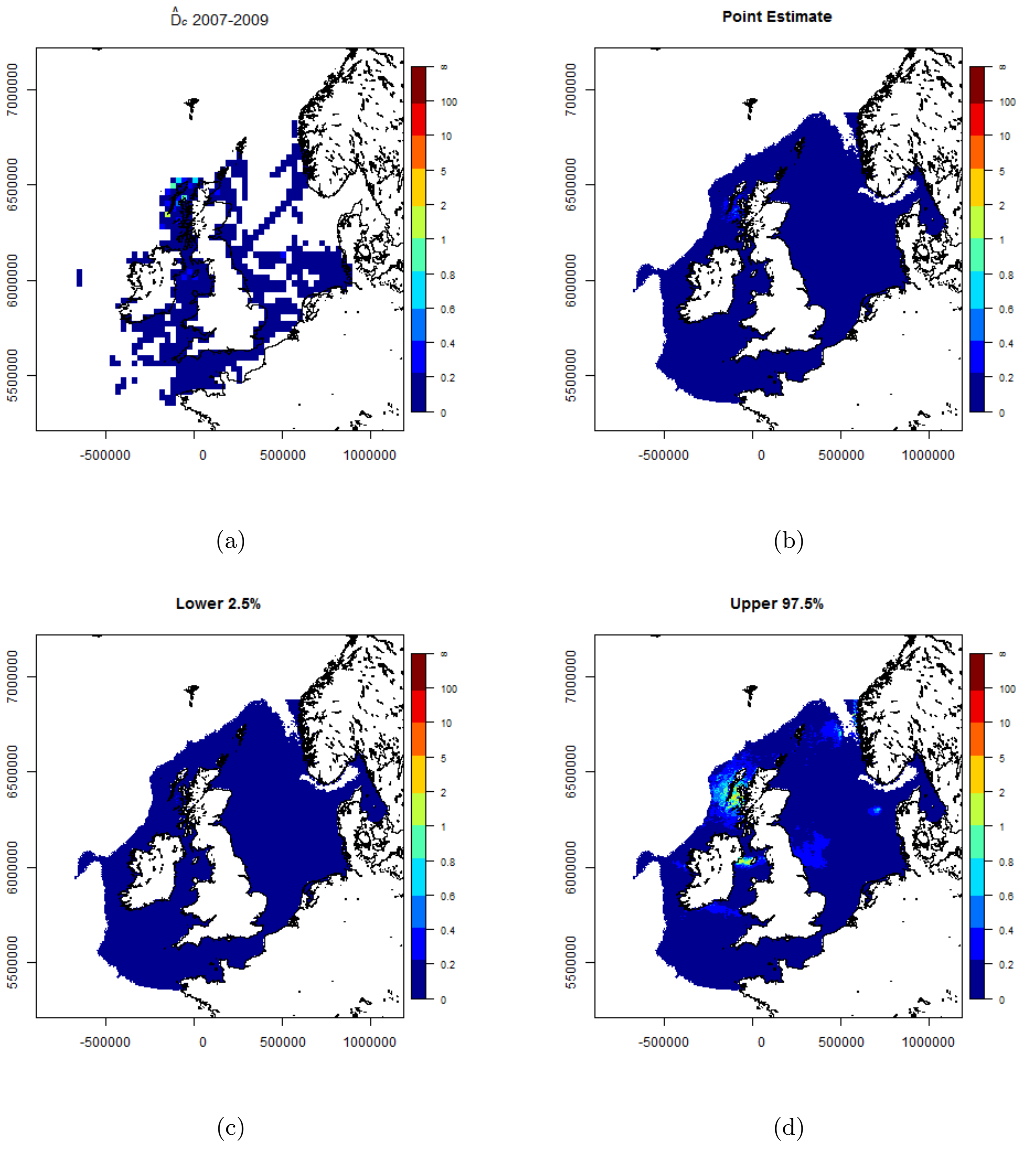
Figure E.30: Predicted minke whale densities for summer (day 227) in 2006. (a) The raw densities for summers in 2005 - 2007 that are drawn upon to make predictions for 2006. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



(a) (b)

(c) (d)

Figure E.31: Predicted minke whale densities for summer (day 227) in 2007. (a) The raw densities for summers in 2006 - 2008 that are drawn upon to make predictions for 2007. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



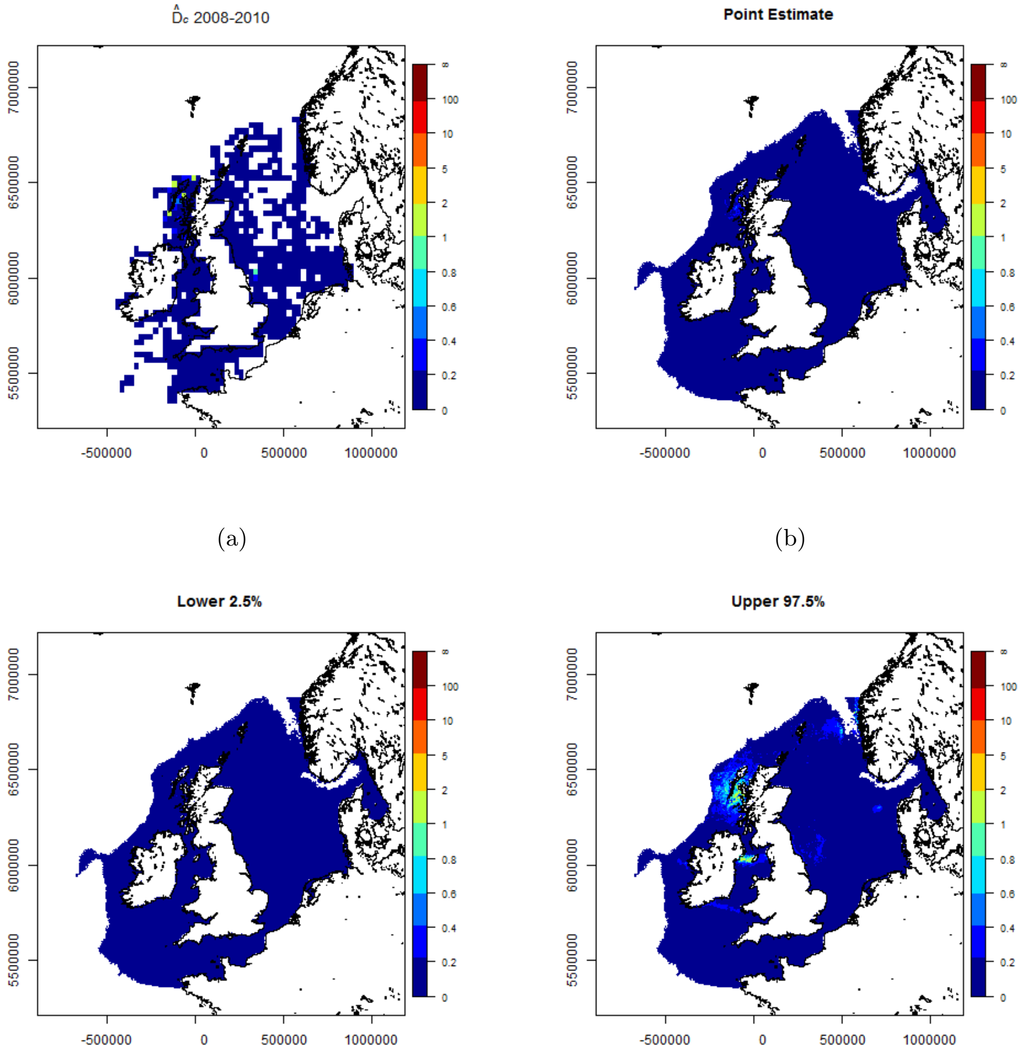
(a)

(b)

(c)

(d)

Figure E.32: Predicted minke whale densities for summer (day 227) in 2008. (a) The raw densities for summers in 2007 - 2009 that are drawn upon to make predictions for 2008. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



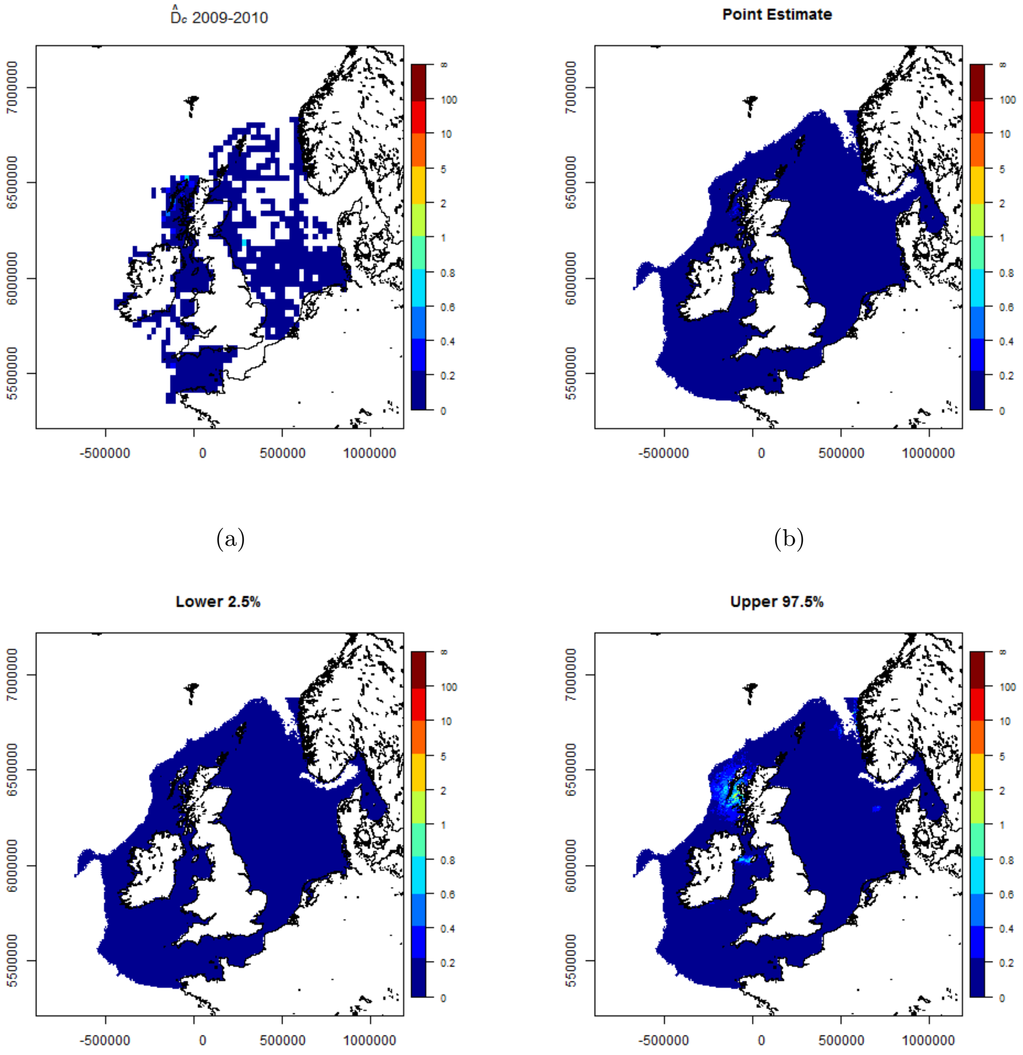
(a)

(b)

(c)

(d)

Figure E.33: Predicted minke whale densities for summer (day 227) in 2009. (a) The raw densities for summers in 2008 - 2010 that are drawn upon to make predictions for 2009. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.



(a)

(b)

(c)

(d)

Figure E.34: Predicted minke whale densities for summer (day 227) in 2010. (a) The raw densities for summers in 2009 - 2010 that are drawn upon to make predictions for 2010. (b) Point estimates of minke whale density. (c) and (d) are the lower and upper 95% GEE based percentile intervals.

Appendix F

Extra plots of the Joint Cetacean Protocol Analysis - Reporting periods

F.1 Harbour Porpoise

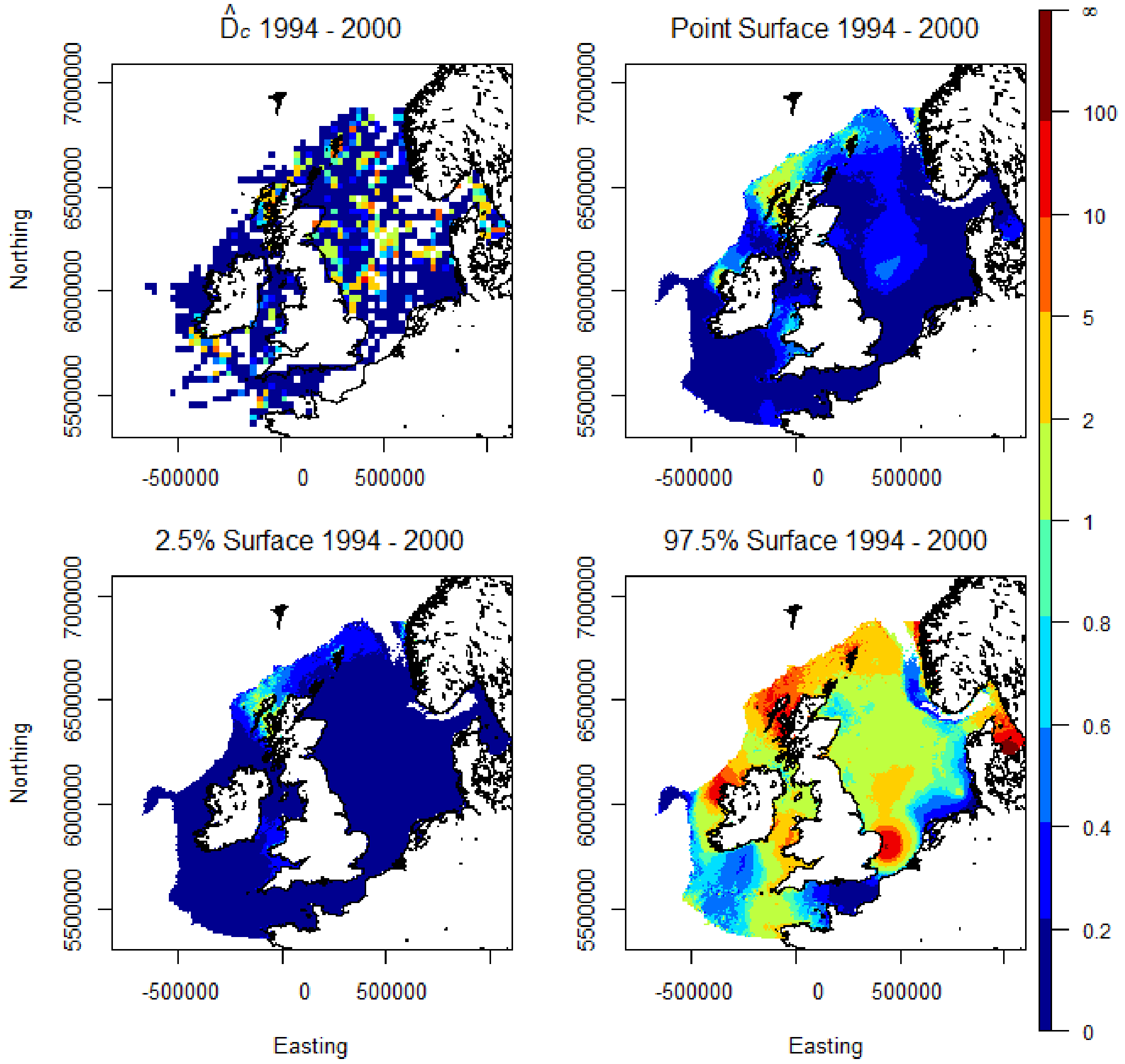


Figure F.1: Predicted harbour porpoise densities for summer (day 227) in reporting period one (1994-2000). (top left) The raw densities for summers in 1994 - 2000 that are drawn upon to make predictions. (top right) Point estimates of harbour porpoise density. (bottom left) and (bottom right) are the lower and upper 95% GEE based percentile intervals.

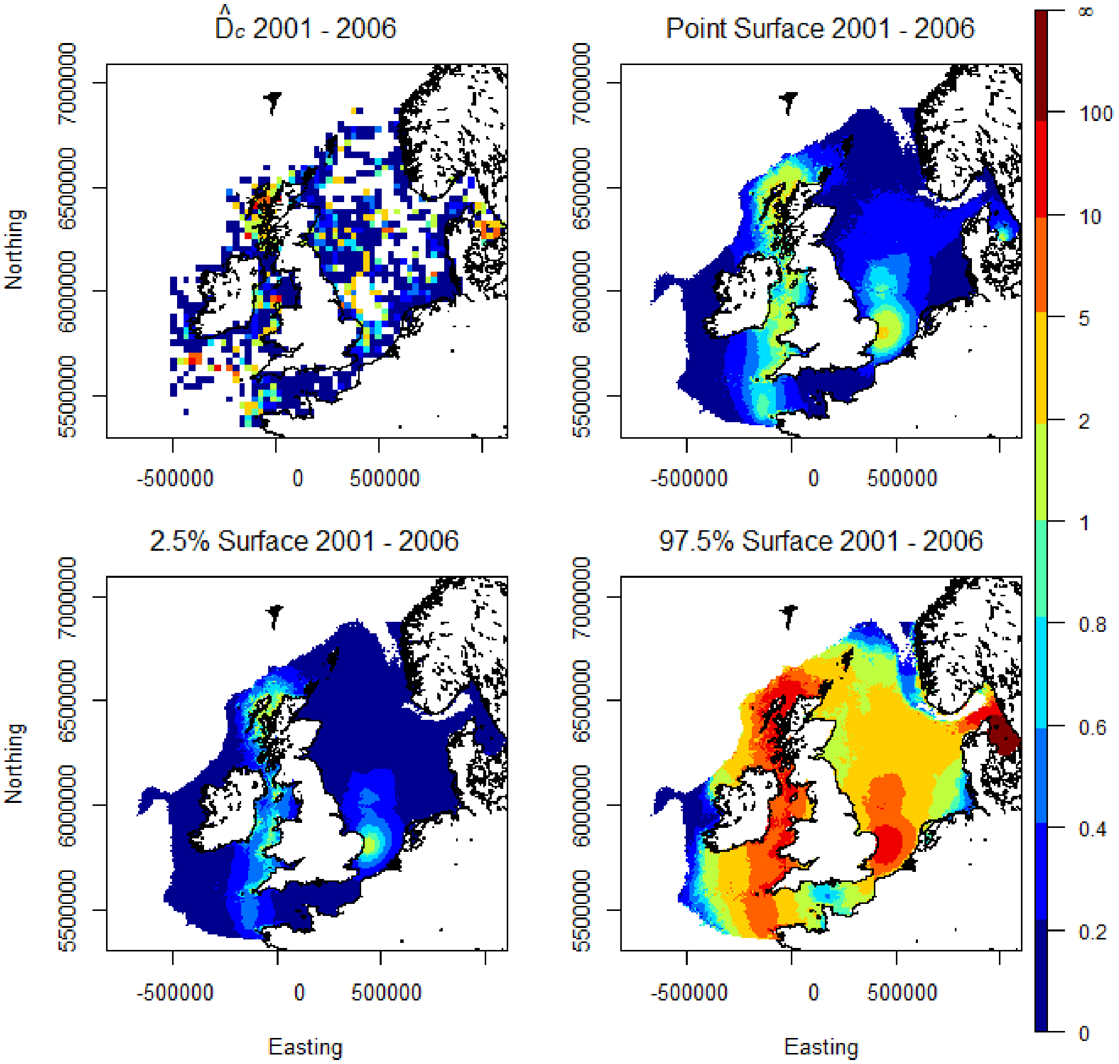


Figure F.2: Predicted harbour porpoise densities for summer (day 227) in reporting period two (2001-2006). (top left) The raw densities for summers in 2001 - 2006 that are drawn upon to make predictions. (top right) Point estimates of harbour porpoise density. (bottom left) and (bottom right) are the lower and upper 95% GEE based percentile intervals.

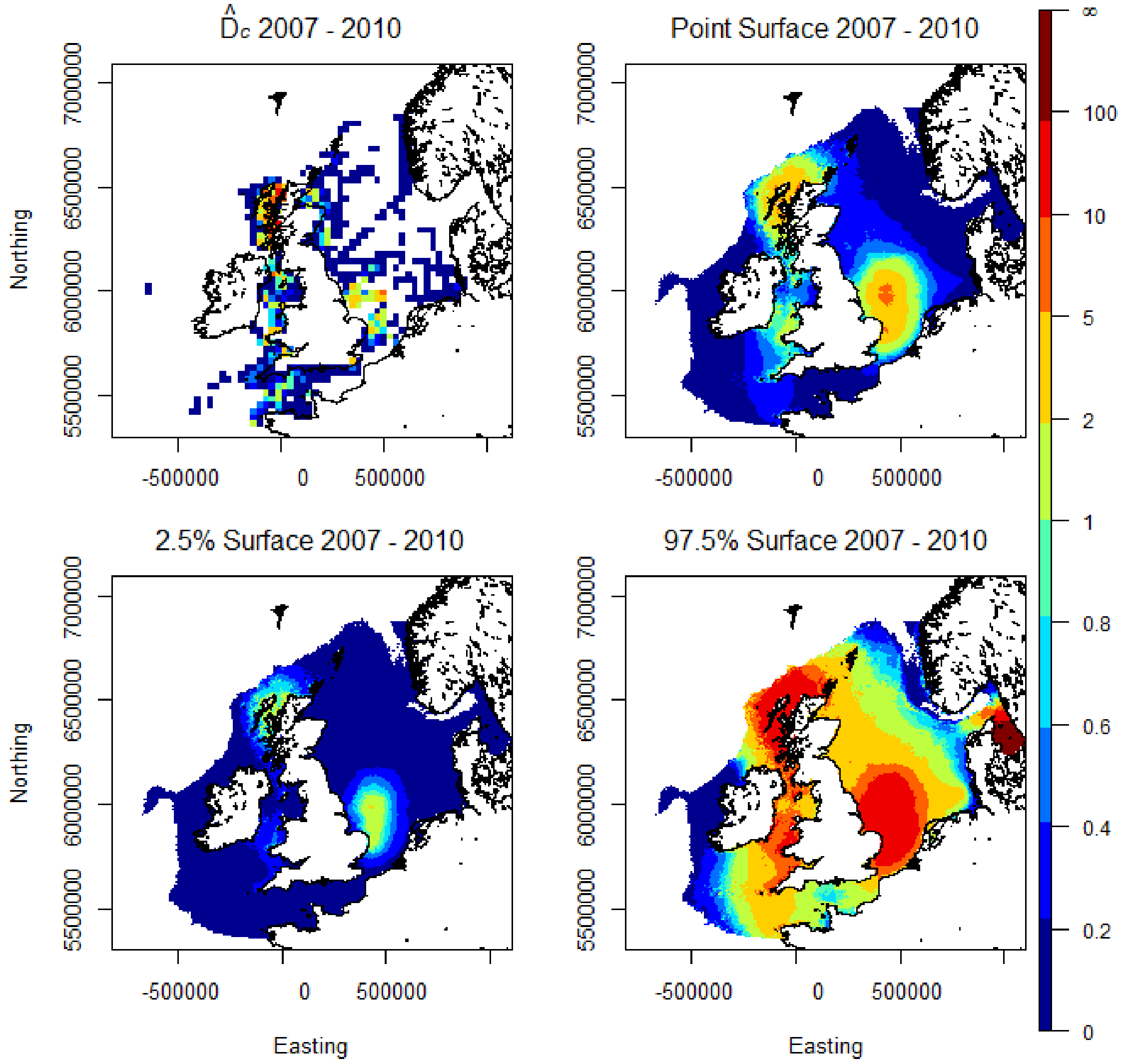


Figure F.3: Predicted harbour porpoise densities for summer (day 227) in reporting period three (2007-2010). (top left) The raw densities for summers in 2007 - 2010 that are drawn upon to make predictions. (top right) Point estimates of harbour porpoise density. (bottom left) and (bottom right) are the lower and upper 95% GEE based percentile intervals.

F.2 Minke Whale

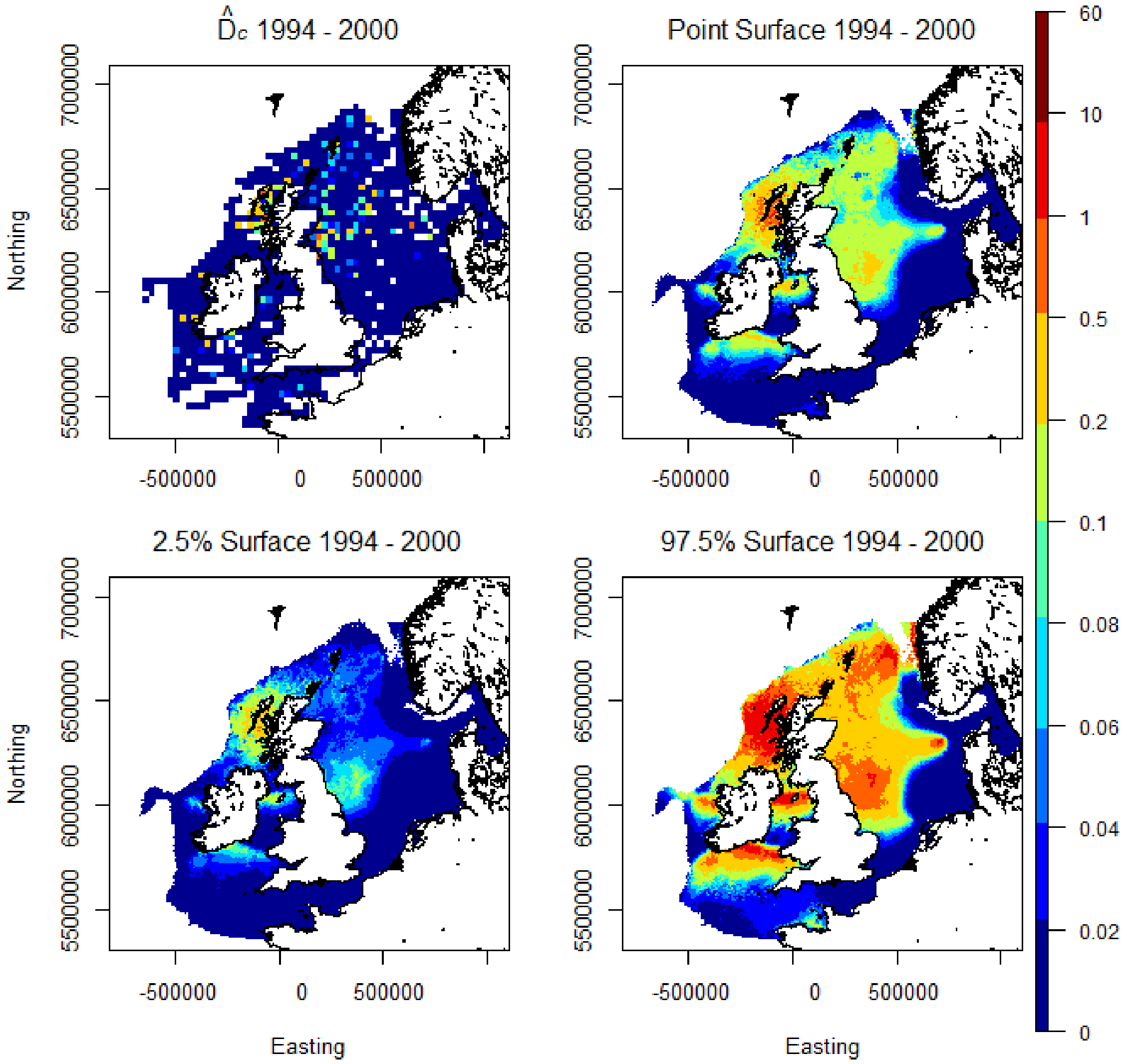


Figure F.4: Predicted minke whale densities for summer (day 227) in reporting period one (1994-2000). (top left) The raw densities for summers in 1994 - 2000 that are drawn upon to make predictions. (top right) Point estimates of minke whale density. (bottom left) and (bottom right) are the lower and upper 95% GEE based percentile intervals.

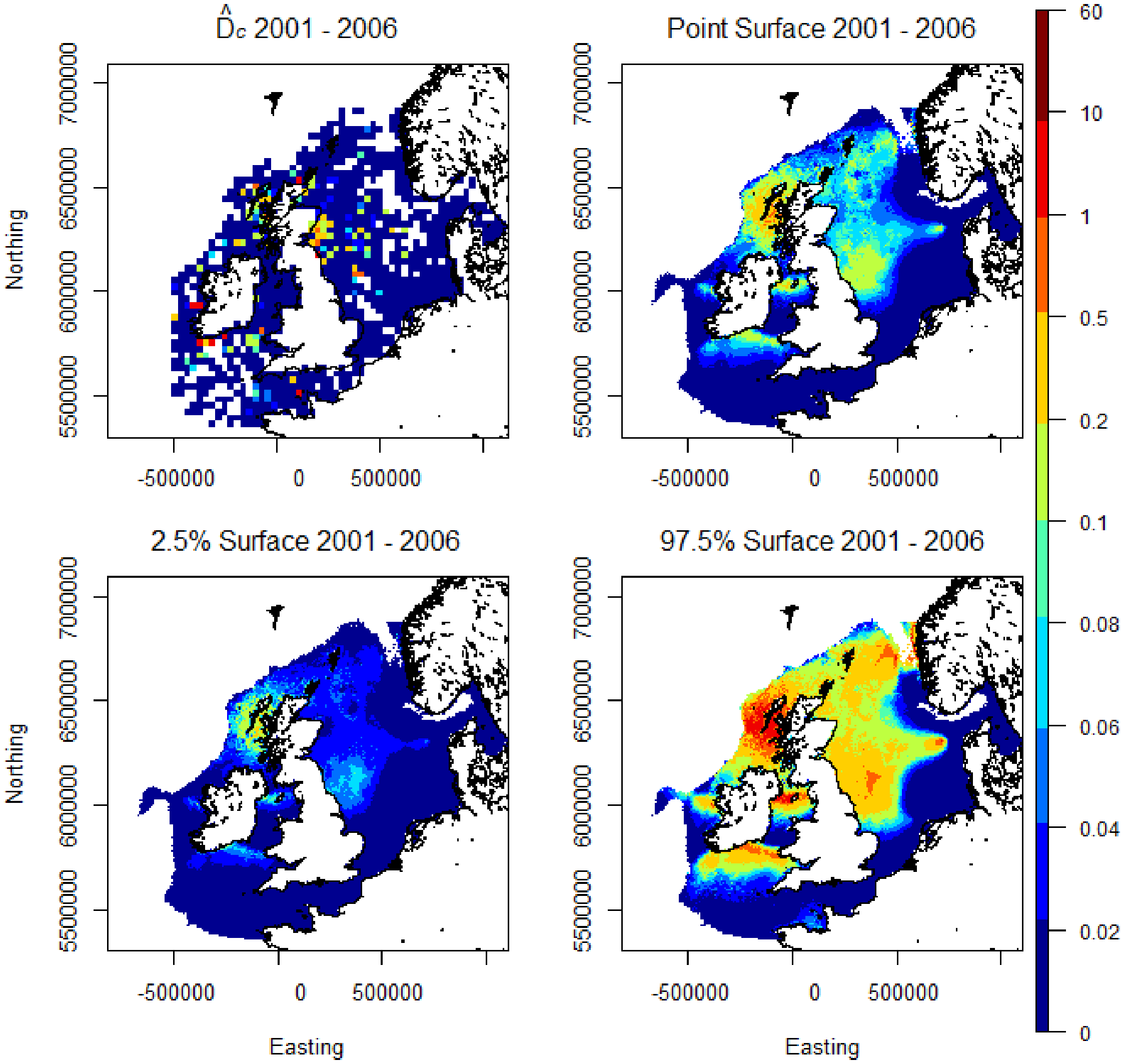


Figure F.5: Predicted minke whale densities for summer (day 227) in reporting period two (2001-2006). (top left) The raw densities for summers in 2001 - 2006 that are drawn upon to make predictions. (top right) Point estimates of minke whale density. (bottom left) and (bottom right) are the lower and upper 95% GEE based percentile intervals.

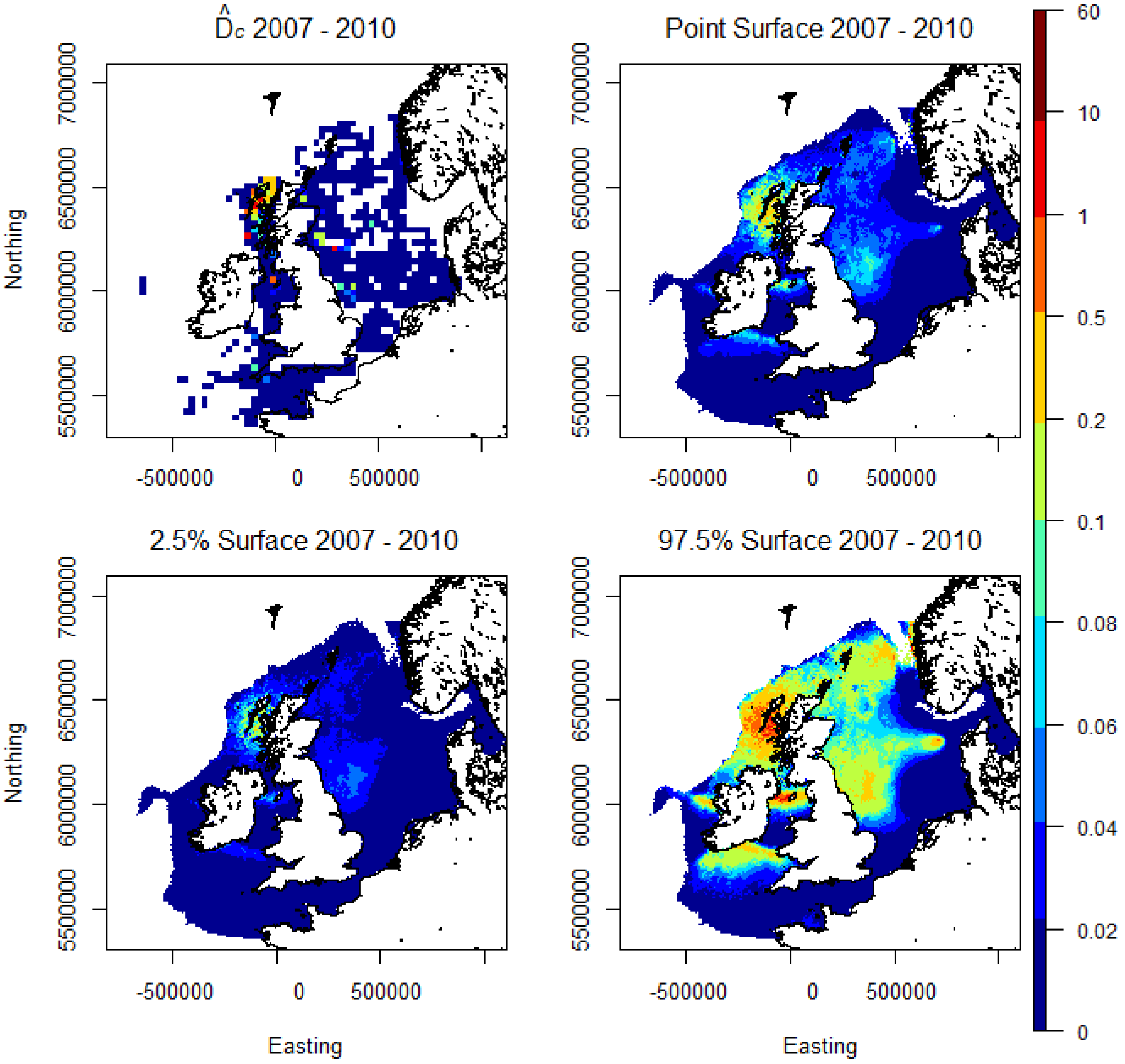


Figure F.6: Predicted minke whale densities for summer (day 227) in reporting period three (2007-2010). (top left) The raw densities for summers in 2007 - 2010 that are drawn upon to make predictions. (top right) Point estimates of minke whale density. (bottom left) and (bottom right) are the lower and upper 95% GEE based percentile intervals.

Appendix G

UK Shipping Forecast Areas

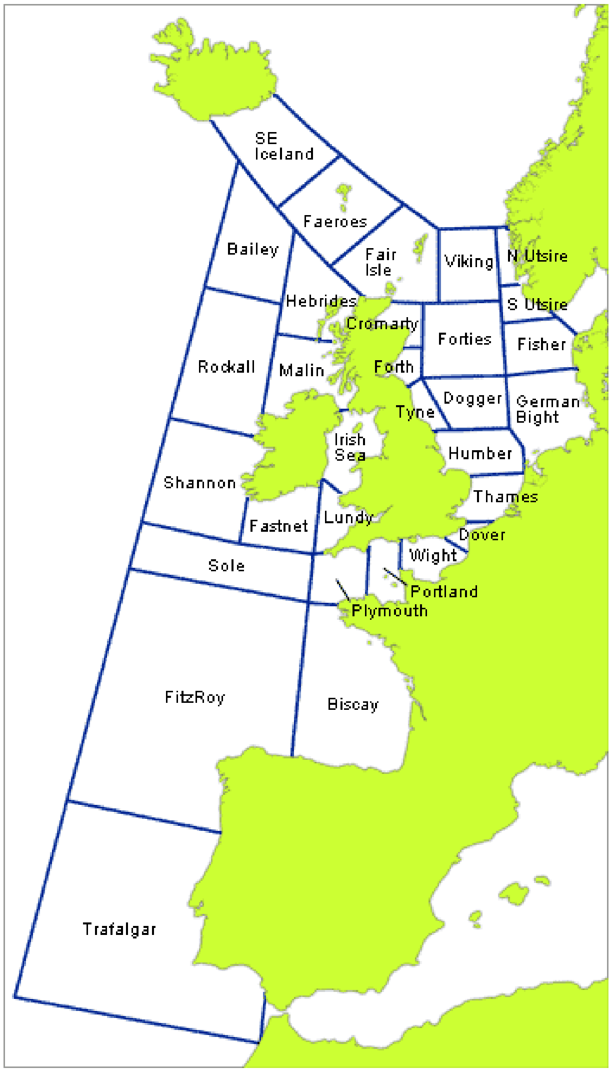


Figure G.1: Map of UK shipping forecast areas taken from www.metoffice.gov.uk

Appendix H

R Code for Combining Density Surfaces Chapter

H.1 Combining Code using the Bayesian Update Procedure

```
# this function evaluates the ‘posterior’; it updates the existing density surface
# with the information associated with the new data
#
# inputs:
#   data_new_y     = new data
#   data_new_cv    = new data, cv values (between 0 and 25)
#   data_prior_y   = prior data (existing data)
#   data_prior_cv  = prior data (existing data), cv values (between 0 and 25)
#
evaluatePosterior<- function(data_new_y, data_new_cv, data_prior_y, data_prior_cv){

  no_el<- length(data_prior_y)
  post_y<- vector(length=no_el)
  post_sigma<- vector(length=no_el)
  post_cv<- vector(length=no_el)
  for(i in 1:no_el){

    y = data_new_y[i]

    if(is.na(y)==T){
      post_y[i]<- NA
      post_sigma[i]<- NA
      post_cv[i]<- NA
    }
    else{
      if(y==0){
        y<- 1e-20
      }
      sigma2logy = log(data_new_cv[i]^2+1)    # variance of log(y) (taken as known)
      # prior mean and cv of lognormal
      # prior mean for theta=mean[log(y)]
      mutheta = log(data_prior_y[i]) - 0.5*log(data_prior_cv[i]^2+1)
      sigma2theta = log(data_prior_cv[i]^2+1)   # prior variance for theta=mean[log(y)]

      # calculate posterior on log scale
      taulogy = 1/sigma2logy                     # precision of logy
      tau0 = 1/sigma2theta                       # precision of mean[log(y)]
      taupost = taulogy + tau0                   # posterior precision of mean[log(y)]
      # posterior mean of mean[log(y)]
      mupost = (taulogy*log(y) + tau0*mutheta)/taupost
      # posterior variance of mean[log(y)]
      sigma2post = 1/taupost
      # convert posterior of  mean[log(y)] to posterior of exp{mean[log(y)]}
      # (i.e. posterior on scale of y)
      estpost = fromNorm2logNorm(list(mun=mupost, sigma2n=sigma2post))

      post_y[i]<- estpost$muln
      post_sigma[i]<- estpost$sigma2ln
      post_cv[i]<- sqrt(estpost$sigma2ln)/estpost$muln   # calculate posterior cv of y
    }
  }
  return(list(post_y=post_y, post_cv=post_cv, post_sigma=post_sigma))
}
```

H.2 Smoothing Code

```
#
# -- 2D Nonparametric Regression (Kernel Smoothing) --
# [A.W. Bowman and A. Azzalini, "Applied Smoothing Techniques for Data Analysis",
#  Oxford, Clarendon Press, 1997, pag. 53]
#

# this function calculates the regression estimate for each point ‘p’ and
# bandwiths ‘h1’ and ‘h2’;
#
#  inputs:
#    p        = point (coordinates) at which to calculate the regression estimate
#    y        = original value of the parameter associated with the point ‘p’
#    loc      = locations (matrix of coordinates)
#    Y        = data to be smoothed
#    h_values = bandwidths
#    ind      = index of points to use to calculate the regression estimate
#
evaluate2DNonParamRegression<- function(p, y, loc, Y, h_values, ind){

  # number of points to use for the calculation
  no_el<- length(ind)

  # define the ‘design matrix’ (or ‘X matrix’)
  x_mx<- matrix(NA, no_el, 3)
  x_mx[,1]<- rep(1, no_el)
  # distances in X direction
  x_mx[,2]<- loc[ind,1] - p[1]
  # distances in Y direction
  x_mx[,3]<- loc[ind,2] - p[2]

  # check for column of zeros in ‘x_mx’ which would lead to singularities;
  # in this case, return the original value
  if(length(which(x_mx[,2]==0))==dim(x_mx)[1] | length(which(x_mx[,3]==0))==dim(x_mx)[1]){
    y
  }
  else{
    # define the ‘weight matrix’ (or ‘W matrix’)
    w_mx<- matrix(0, no_el, no_el)

    # calculate the weights
    w1<- evaluateGaussianKernel(p[1], loc[ind,1], h_values[1])
    w2<- evaluateGaussianKernel(p[2], loc[ind,2], h_values[2])
    diag(w_mx)<- (w1 * w2)

    # make kernel estimate
    solve((t(x_mx) %*% w_mx %*% x_mx)) %*% t(x_mx) %*% w_mx %*% Y[ind]
  }
}
#
#
# this function carries out the kernel smoothing
#
# inputs:
#   gpoints  = grid points (locations)
#   data_y   = data Y to be smoothed
#   data_h   = bandwidths
#   data_id  = IDs associated with the data Y
#   hole_id  = IDs associated with the holes
#
doKernelSmoothing<- function(gpoints, data_y, data_h, data_id, hole_id){

  no_el<- length(data_y)
  no_cols<- length(unique(gpoints[,1]))
  no_rows<- length(unique(gpoints[,2]))

  # define data structure for the smoothed data
  data_sm_y<- vector(length=no_el)

  for(i in 1:no_el){
    if(is.na(data_y[i])){
      data_sm_y[i]<- NA
    }
    else{
      h_values<- data_h[i,]
      # find IDs of points in box around point of interest ‘i’
      roi_id <- evaluateROI_SFD(c(no_cols, no_rows), i, round(max(h_values)*3/res_min))
      if(is.null(hole_id)==F){
        # make sure any holes in the data are given NA’s in corresponding vector of ID’s
        for(j in 1:length(roi_id)){
          for(s in 1:length(hole_id)){
            if(data_id[roi_id[j]]==hole_id[s]){
              roi_id[j]<- NA
              break
            }
          }
        }
      }
      # if there are not enough points to smooth just use the original data
      if(length(na.omit(roi_id))<4){
        data_sm_y[i]<- data_y[i]
      }
      else{
        data_sm_y[i]<- evaluate2DNonParamRegression(gpoints[i,], data_y[i],
                                                    gpoints, data_y,
                                                    h_values, na.omit(roi_id))[1]
      }
    }
  }
  return(data_sm_y)
}
```

Appendix I

DCAP Pre-processing Method

This is a general description of the pre-processing method for DCAP (v0.1.0), Chapter 7.

- Data enters the algorithm as .csv files. For density data there is one prior file and one survey file each containing a grid of regularly space points with an associated density and CoV. For stratified data, there is one prior file containing a regular grid of points. The survey file contains a row for each strata and consists of a strata ID, density and CoV. There is also a boundary file that denotes the polygon for each strata (separated by NA's).
- The first check is to see if there are points that cross the date line from -180^o to 180^o. If so then the data is stitched together at the date line and only increasing values of degrees are allowed.
- The prior surface may be quite large so to aid computation, the relevant section of prior surface is extracted out for working with. For the DCAP algorithm, the section is the size of the survey area plus 3 degrees in all directions.
- Next, the data is transformed to a grid where the bottom left corner is point (1,1), this point also has associated ID=1. The IDs increase by row, for example, ID=2 on a grid of 0.25 resolution would be (1.25, 1).
- Search for holes in the survey data, which indicate land.
- The new data structure is (gridx, gridy, density, CoV, ID)
- Repeat transformation for prior data and insert NA's for land
- Put both data sets at the same resolution so that they overlap exactly. For example, each data point at 0.5 resolution becomes 4 points at 0.25 resolution. The density remains the same as the single point

for the 4 points but the CoVs change. The CoV for each new point is: $\text{CoV}_{new} = \sqrt{(}\text{CoV}_{old}/\text{number}$ of new cells).

- The two data sets are now ready for updating.

Appendix J

DCAP Log

Table J.1: DCAP log. All timings relate to use of a computer with the following specifications: Windows Dual Core, 2.40 GHz CPU, 2.00 GB RAM.

ID	**Objectives**	**INPUT FILES**		**DCAP Version**	**Computation Time**	**Test Result**	**Notes / Comments**
		Prior survey	**Survey data**		**(s)**	**(Pass/Fail)**	
1a	**Open ocean**	Prior1	Surveyds1	0.3.0	12.42	P	
1b			Surveyst1	0.3.0	3.5	P	
1c			Surveyst2	0.3.0	3.5	P	
	Land						
1d	Top	Prior3	Surveyds3	0.3.0	31.83	P	
1e	Bottom	Prior4	Surveyds4	0.3.0	31.69	P	
1f	Right	Prior1	Surveyds1a	0.3.0	17.04	P	
1g			Surveyst4	0.3.0	1.58	P	
1h	Left		Surveyds1b	0.3.0	17	P	
1i	Islands		Surveyds1f	0.3.0	29.87	P	
	Lat/Lon						
1j	Northern limit1	Prior2c	Surveyds2c	0.3.0	22.45	P	
1k	Southern limit1	Prior2d	Surveyds2d	0.3.0	22.48	P	
1l	Crossing the equator	Prior2	Surveyds2	0.3.0	22.45	P	
1m	Crossing 0o longitude	Prior2a	Surveyds2a	0.3.0	23.29	P	
1n	Crossing 180/180o longitude	Prior2b	Surveyds2b	0.3.0	23.75	P	
	Survey regions						
	- simple geometric shapes						
2a	Diamond	Prior1	Surveyds1c	0.3.0	9.1	P	
2b			Surveyst3	0.3.0	1.25	P	
2c	L shape		Surveyds1d	0.3.0	12.69	P	
2d	U shape		Surveyds1e	0.3.0	11.8	P	
	- complicated shapes						
2e	SCANS-II B	Prior1	Surveyds1g	0.3.0	87.98	P	
2f	SCANS-II Q		Surveyds1h	0.3.0	1657.7	P	
	Non-overlapping data						
4a	Prior density 0/survey density positive	Prior2	Surveyds2f	0.3.0	31.63	P	
4b	Prior density positive/survey density 0	Prior1	Surveyds1j	0.3.0	11.61	P	
	Variable smoothness in 2D						
5	Step change in survey density	Prior1	Surveyds1i	0.3.0	14.3	P	
	Multiple updates						
6a	Several overlapping surveys	Prior1	Surveyds1k	0.3.0	13.05	P	
6b			Surveyds1l	0.3.0		F	unable to store/retrieve output
6c			Surveyds1m	0.3.0		F	from previous update
	Survey size						
7a	Increase resolution of ds (0.1)	Prior1	Surveyds1n	0.3.0	411.96	P	
7b	(0.05)		Surveyds1o	0.3.0	5095.86	P	
7c	(0.01)		Surveyds1p	0.3.0		F	Dataset too big (160000 points)
	Demonstration data						
3a	Humpback data around Hawaii	Priordemo	Surveydsdemo1	0.3.0	35.88	P	
3b			Surveydsdemo2	0.3.0		F	Dataset too big (51895 points)

Bibliography

H. Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60:255–265, 1973.

M. C. Andersen, H. Adams, B. Hope, and M. Powell. Risk analysis for invasive species: general framework and research needs. *Risk Analysis*, 24(4):893–900, 2004.

P. Anderwald, P. Evans, R. Dyer, A. Dale, P. Wright, and A. Hoelzel. Spatial scale and environmental deterinants in minke whale habitat use and foraging. *Marine Ecology Progress Series*, 450:259–274, 2012.

H. G. Andrewartha and L. C. Birch. The distribution and abundance of animals. *The distribution and abundance of animals.*, 1954.

M. Araújo and A. Guisan. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33:1677–1688, 2006.

M. B. Araújo, R. G. Pearson, W. Thuiller, and M. Erhard. Validation of species-climate impact models under climate change. *Global Change Biology*, 11(9):1504–1513, 2005.

J. A. Ardron, H. Possingham, and C. Klein, editors. *Marxan good practices handbook.* University of Queensland, St Lucia, Queensland, Australia, and Pacific Marine Analysis and Research Association, Vancouver, British Colombia, Canada, 2 edition, 2010.

E. Ashe, D. Noren, and R. Williams. Animal behaviour and marine protected areas: incorporating behavioural data into the selection of marine protected areas for an endangered killer whale population. *Animal Conservation*, 13(2):196–203, 2010.

K. L. Ayres, R. K. Booth, J. A. Hempelmann, K. L. Koski, C. K. Emmons, R. W. Baird, K. Balcomb-Bartok, M. B. Hanson, M. J. Ford, and S. K. Wasser. Distinguishing the impacts of inadequate prey and vessel traffic on an endangered killer whale (*Orcinus orca*) population. *PLOS One*, 7(6):e36842, 2012.

M. Baines and P. Evans. Atlas of the marine mammals of Wales. CCW monitoring report no. 68. Technical report, 2009.

R. Baird. Status of killer whales, *Orcinus orca*, in Canada. *Canadian Field Naturalist*, 115: 676–701, 2001.

V. Baladandayuthapani and R. Carroll. Spatially adaptive bayesian penalized splines (P-splines). *Journal of Computational and Graphical Statistics*, 14(2):378–394, 2005.

I. Ball, H. Possingham, and M. Watts. Marxan and relatives: Software for spatial conservation prioritisation. In A. Moilanen, K. Wilson, and H. Possingham, editors, *Spatial conservation prioritisation: Quantitative methods and computational tools*, pages 185–195. Oxford University Press, UK, 2009.

J. Barlow, M. Ferguson, E. Becker, J. V. Redfern, K. Forney, I. Vichis, P. Fiedler, T. Gerrodette, and L. Ballance. Predictive modelling of cetacean densities in the eastern Pacific Ocean. Technical report, National Oceanic and Atmospheric Administration, Southwest Fisheries Science Center, 2009.

C. Barton, D. Jackson, P. Bloor, and Z. Crutchfield. Using a before-after-gradient design to determine post-construction effects of an offshore wind farm on birds: preliminary results (poster). In R. May and K. Bevanger, editors, *Conference on wind energy and wildlife impacts*, 2011.

R. E. Bilby and L. A. Mollot. Effects of changing land use patterns on the distribution of coho salmon (*Oncorhynchus kistutch*) in the puget sound region. *Canadian Journal of Fisheries and Aquatic Science*, 65:2138Äê2148, 2008.

A. Bjørge. *How persistent are marine mammal habitats in an ocean of variability?*, pages 63–91. Kluwer Academic/Plenum Publishers, London, 2001.

C. Booth. *Variation in habitat preference and distribution of harbour porpoises west of* S*cotland*. PhD thesis, University of St Andrews, 2010.

A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis; The kernel approach with S-Plus Illustrations*, volume 18 of *Oxford Statistical Science Series*. Oxford University Press, 2003.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

H. Brown and R. Prescott. *Applied Mixed Models in Medicine.* J. Wiley & Sons, West Sussex, England, 1999.

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model selection: An integral part of inference. *Biometrics*, 53:603–618, 1997.

S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. T. Thomas. *Introduction to Distance Sampling: Estimating abundance of biological populations.* Oxford University Press, Oxford, 2001.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A practical information-theoretic approach.* Springer, 2002.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A practical information-theoretic approach.* Springer, 2^{nd} edition, 2010.

M. Burt. Generation of test data sets. Technical report, RUWPA, University of St. Andrews, 2010.

C. Camphuysen, A. Fox, M. Leopold, and I. K. Petersen. Towards standardised seabirds at sea census techniques in connection with environmental impact assessments for offshore wind farms in the uk a comparison of ship and aerial sampling methods for marine birds, and their applicability to offshore wind farm assessments. *Koninklijk Nederlands Instituut voor Onderzoek der Zee Report commissioned by COWRIE*, 2004.

J. Candy and T. Quinn. Behavior of adult chinook salmon (*Oncorhynchus tshawytscha*) in british columbia coastal waters determined from ultrasonic telemetry. *Canadian Journal of Zoology*, 77:1161–1169, 1999.

G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2009.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):pp. 829–836, 1979. ISSN 01621459.

M. J. Cox, D. L. Borchers, and N. Kelly. nupoint: An R package for density estimation from point transects in the presence of non-uniform animal density. *Methods in Ecology and Evolution*, 2013.

C. Crainiceanu, D. Ruppert, R. Carroll, A. Joshi, and B. Goodner. Spatially adaptive bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 16(2):265–288, 2007.

J. Cramer. *Logit models: from economics and other fields.* Cambridge University Press, 2003.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, Volume 31(4):377–403, 1979.

A. Davison and D. Hinckley. *Bootstrap Methods and Their Application.* Cambridge, 2007.

C. de Boor. *A Practical Guide to Splines: Revised Edition.* Springer, 2001.

D. Donoho and J. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

C. Donovan, L. A. S. Scott-Hayward, L. Milazzo, and D. L. Borchers. MMAARS algorithm report. Technical report, University of St. Andrews sub-contract number BIS-09-001-STAU-001 for BAE Systems, Sensor Systems, 2010.

C. Donovan, K. Kaschner, C. Harris, R. Wiff, N. Quick, and J. Harwood. Combining survey data and an index of habitat suitability to estimate marine mammal density at a global scale. *In prep.*, 2011.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):115–121, 1996.

H. Elith, J., C. P. Graham, P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J Hijmans, F. Huettmann, J. R Leathwick, A. Lehmann, J. Li, L. G Lohmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

J. Elith, S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. A statistical explanation of maxent for ecologists. *Diversity and Distributions*, 17(1):43–57, 2011.

J. I. Ellis and D. C. Schneider. Evaluation of a gradient sampling design for environmental impact assessment. *Environmental Monitoring and Assessment*, 48(2):157–172, 1997.

C. Embling. *Predictive models of cetacean distributions off the west coast of* S*cotland,*. PhD thesis, School of Biology, 2007.

C. Embling, P. Gillibrand, J. Gordon, J. Shrimpton, P. Stevick, and P. Hammond. Using habitat models to identify suitable sites for marine protected areas for harbour porpoises (*Phocoena phocoena*). *Biological Conservation*, 143:267–279, 2010.

C. Erbe. Underwater noise of whale-watching boats and potential effects on killer whales (*Orcinus orca*), based on an acoustic impact model. *Marine Mammal Science*, 18:394–418, 2002.

J. Fan. Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, 21:196–216, 1993.

J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* CRC Press, 2006.

S. Ferrier, G. Watson, J. Pearce, and M. Drielsma. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation*, 11(12):2275–2307, 2002.

R. W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5:345, 1962.

M. Fontaine, K. Tolley, U. Siebert, S. Gobert, G. Lepoint, J.-M. Bouquegneau, and K. Das. Long-term feeding ecology and habitat use in harbour porpoises *Phocoena phocoena* from Scandinavian waters inferred from trace elements and stable isotopes. *BMC ecology*, 7 (1):1, 2007.

J. Ford, G. Ellis, and K. Balcomb. *Killer Whales: The natural history and genealogy of Orcinus orca in British Columbia and Washington State.* University of British Columbia press, second edition, 2000.

J. Ford, G. Ellis, P. Olesiuk, and K. Balcomb. Linking killer whale survival and prey abundance: Food limitation in the oceans' apex predator? *Biology Letters*, 6:139–142, 2010.

J. K. B. Ford and G. M. Ellis. Selective foraging by fish eating killer whales, *Orcinus orca* in British Columbia. *Marine Ecology Progress Series*, 316:185–199, 2006.

J. K. B. Ford, G. M. Ellis, L. G. Barrett-Lennard, A. B. Morton, R. S. Palm, and K. Balcomb III. Dietary specialization in two sympatric populations of killer whales (*Orcinus orca*) in coastal British Columbia and adjacent waters. *Canadian Journal of Zoology*, 76: 1456–1471, 1998.

B. Fornberg, T. Driscoll, G. Wright, and R. Charles. Observations on the behavior of radial basis function approximations near boundaries. *Computers & Mathematics with Applications*, 43(3-5):473–490, 2002.

A. Fox, M. Desholm, J. Kahlert, T. K. Christensen, and I. Krag Petersen. Information needs to support environmental impact assessment of the effects of European marine offshore wind farms on birds. *Ibis*, 148(s1):129–144, 2006.

J. Franklin. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, 19(4):474–499, 1995.

J. Franklin and J. A. Miller. *Mapping species distributions: spatial inference and prediction*, volume 338. Cambridge University Press Cambridge, 2009.

J. H. Friedman and B. W. Silverman. Flexible parsimonious smoothing and additive modelling. *Technometrics*, 31:3–21, 1989.

R. Furrer, D. Nychka, and S. Sain. *fields: Tools for spatial data*, 2010. URL `http://CRAN.R-project.org/package=fields`. R package version 6.3.

A. M. Gormley, E. Slooten, S. Dawson, R. J. Barker, W. Rayment, S. du Fresne, and S. Bräger. First evidence that marine protected areas can work for marine mammals. *Journal of Applied Ecology*, 49(2):474–480, 2012.

P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalised Linear Models (A roughness penalty approach).* Chapman and Hall, England, 1994.

R. H. Green. *Sampling design and statistical methods for environmental biologists.* John Wiley & Sons Inc., 1979.

H. Greene, B. Dieter, C. Endris, H. Lopez, L. Murai, and M. Erdey. Geology and seafloor bathymetry of the San Juan Islands. Center for Habitat Studies, Moss Landing Marine Laboratories., 2007.

C. Gu. *Smoothing Spline ANOVA models.* New York: Springer Verlag, 2002.

M. Guillemette and J. K. Larsen. Post development experiments to detect anthropogenic disturbances: the case of sea ducks and wind parks. *Ecological Applications*, 12(3):868–877, 2002.

A. Guisan and W. Thuiller. Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009, 2005.

A. Guisan and N. Zimmerman. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135:147–186, 2000.

A. Guisan, T. Edwards, and T. Hastie. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157:89–100, 2002.

B. S. Halpern, S. Walbridge, K. A. Selkoe, C. V. Kappel, F. Micheli, C. D'Agrosa, J. F. Bruno, K. S. Casey, C. Ebert, H. E. Fox, R. Fujita, D. Heinemann, H. S. Lenihan, E. M. P.

Madin, M. T. Perry, E. R. Selig, M. Spalding, R. Steneck, and R. Watson. A global map of human impact on marine ecosystems. *Science*, 319(5865):pp. 948–952, 2008.

T. Hamazaki. Spatiotemporal prediction models of cetacean habitats in the mid-western North Atlantic ocean (from Cape Hatteras, North Carolina, USA to Nova Scotia, Canada). *Marine Mammal Science*, 18(4):920–939, 2002.

P. S. Hammond, P. Berggren, H. Benke, D. L. Borchers, A. Collet, M. P. Heide-Jorgensen, S. Heimlich, A. R. Hiby, M. F. Leopold, and N. Oien. Abundance of harbour porpoise and other cetaceans in the North Sea and adjacent waters. *The Journal of Applied Ecology*, 39:361–376, 2002.

P. S. Hammond, K. Macleod, P. Berggren, D. L. Borchers, L. Burt, A. Cañadas, G. Desportes, G. P. Donovan, A. Gilles, D. Gillespie, et al. Cetacean abundance and distribution in European Atlantic shelf waters to inform conservation and management. *Biological Conservation*, 164:107–122, 2013.

J. Hanley, A. Negassa, M. Edwardes, and J. Forrester. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Practice of epidemiology*, 157(4): 364–375, 2003.

M. Hanson, R. Baird, J. Ford, J. Hempelmann-Halos, D. V. Doornik, J. Candy, C. Emmons, G. Schorr, B. Gisborne, K. Ayres, S. Wasser, K. Balcomb, K. Balcomb-Bartok, J. Sneva, and M. Ford. Species and stock identification of prey consumed by endangered southern resident killer whales in their summer range. *Endangered Species Research*, 11:69–82, 2010.

R. L. Harder and R. N. Desmarais. Interpolation using surface splines. *Journal of Aircraft*, 9:189–191, 1972.

J. Hardin and J. Hilbe. *Generalized Estimating Equations*. Chapman & Hall/CRC, 2002.

C. Harris. Supporting documentation for predicted density data: Available from SMRU Ltd. Technical report, University of St. Andrews, 2013.

D. Harrison and C. Hulin. Investigations of absenteeism: Using event-history models to study the absence-taking process. *Journal of Applied Psychology*, 74:300–316, 1989.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer Series in Statistics, second edition, 2009.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models.* Chapman & Hall, 1990.

T. Haug, H. Gjøsæter, U. Lindstrøm, and K. T. Nilssen. Diet and food availability for north-east Atlantic minke whales (*Balaenoptera acutorostrata*), during the summer of 1992. *ICES Journal of Marine Science: Journal du Conseil*, 52(1):77–86, 1995.

D. Hauser, M. Logsdon, E. Holmes, G. VanBlaricom, and R. Osborne. Summer distribution patterns of southern resident killer whales *Orcinus orca*: core areas and spatial segregation of social groups. *Marine Ecology Progress Series*, 351:301–310, 2007.

B. Hawkins. Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography*, 39: 1–9, 2012.

N. E. Heckman and J. O. Ramsay. Penalized regression with model-based penalties. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 28:241–258, 2000.

T. Hengl. *A practical guide to geostatistical mapping of environmental variables.* Amsterdam, http://spatial-analyst.net/book, second edition, 2009.

H. Herr, H. O. Fock, and U. Siebert. Spatio-temporal associations between harbour porpoise, *Phocoena phocoena*, and specific fisheries in the German Bight. *Biological Conservation*, 142(12):2962–2972, 2009.

A. H. Hirzel and G. Le Lay. Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, 45(5):1372–1381, 2008.

J. M. Hoekstra, K. K. Bartz, M. H. Ruckelshaus, J. M. Moslemi, and T. K. Harms. Quantitative threat analysis for management of an imperiled species: Chinook salmon (*Oncorhynchus twhawytscha*). *Ecological Applications*, 17(7):2061–2073, 2007.

S. K. Hooker, H. Whitehead, and S. Gowans. Marine protected area design and the spatial and temporal distribution of cetaceans in a submarine canyon. *Conservation Biology*, 13 (3):pp. 592–602, 1999.

D. Hosmer and S. Lemeshow. Logistic regression for matched case-control studies. *Applied logistic regression*, 2:223–259, 1989.

E. Hoyt. *Marine Protected Areas For Whales, Dolphins and Porpoises.* Earthscan, UK, 2005.

E. Hoyt. *Marine Protected Areas for Whales, Dolphins and Porpoises: A world handbook for cetacean habitat conservation and planning.* Routledge, 2012.

C. M. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

O. E. Jansen, G. M. Aarts, K. Das, G. Lepoint, L. Michel, and P. J. Reijnders. Feeding ecology of harbour porpoises: stable isotope analysis of carbon and nitrogen in muscle and bone. *Marine Biology Research*, 8(9):829–841, 2012.

M. Jasny. *Sounding the depths II: The rising toll of sonar, shipping and industrial ocean noise on marine life.* Natural Resources Defense Council, 2005.

P. W. M. John, M. E. Johnson, L. M. Moore, and D. Ylvisaker. Minimax distance designs in two-level factorial experiments. *Journal of Statistical Planning and Inference*, 44:249–263, 1995.

M. Johnson, L. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148, 1990.

F. Kasamatsu, P. Ensor, G. Joyce, and N. Kimura. Distribution of minke whales in the Bellingshausen and Amundsen Seas (60 W-120 W), with special reference to environmental/physiographic variables. *Fisheries Oceanography*, 9(3):214–223, 2000.

K. Kaschner, R. Watson, A. W. Trites, and D. Pauly. Mapping world-wide distributions of marine mammal species using a Relative Environmental Suitability (RES) model. *Marine Ecology Progress Series*, 316:285–310, 2006.

K. Kaschner, N. Quick, R. Jewell, R. Williams, and C. Harris. Global review and gap analysis of cetacean line-transect survey coverage. *PLOS One*, 7(9), September 2012.

M. Kelly and R. K. Meentemeyer. Landscape dynamics of the spread of sudden oak death. *PE & RS- Photogrammetric Engineering & Remote Sensing*, 68(10):1001–1009, 2002.

M. Knoll, R. Dekeling, M. Stifani, P. Kvadsheim, K. Liddell, S.-L. Gunnarsson, T. Johansson, G. Pavan, N. Nordlund, F. Benders, T. Zwan, S. Ludwig, D. Lorenzen, R. Kreimeyer, and I. Nissen. Protection of Marine Mammals (PoMM) European Defence Agency (EDA). *Poster at 4th International Conference on the Effects of Sound in the Ocean on Marine Mammals*, Amsterdam, September 2011.

N. Koper and M. Manseau. A guide to developing resource selection functions from telemetry data using generalized estimating equations and generalized linear mixed models. *Rangifer*, 32(2):195–204, 2012.

T. Krivobokova, C. M. Crainiceanu, and G. Kauermann. Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, 17(1):1–20, 2008.

A. Lehmann, J. Overton, and M. Austin. Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity and Conservation*, 11:2085–2092, 2002.

K.-Y. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi: 10.1093/biomet/73.1.13. URL `http://biomet.oxfordjournals.org/content/73/1/13.abstract`.

E. Limpert, W. Stahel, and M. Abbt. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352, 2001.

C. Liu, P. Berry, T. Dawson, and R. Pearson. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28:385–393, 2005.

M. Louzao, J. Bécares, B. Rodríguez, K. D. Hyrenbach, A. Ruiz, J. Arcos, et al. Combining vessel-based surveys and tracking data to identify key marine areas for seabirds. *Marine Ecology Progress Series*, 391:183–197, 2009.

J. Lucas. Feeding ecology of "southern resident" killer whales (*Orcinus orca*): Benthic habitat and spatial distribution. Master's thesis, The Evergreen State College, 2009.

D. Lusseau and J. Higham. Managing the impacts of dolphin-based tourism through the definition of critical habitats: the case of bottlenose dolphins (*Tursiops* spp.) in Doubtful Sound, New Zealand. *Tourism management*, 25(657-667), 2004.

D. Lusseau, D. Bain, R. Williams, and J. Smith. Vessel traffic distupts the foraging behaviour of southern resident killer whales *Orcinus orca*. *Endangered Species Research*, 6: 211–221, 2009.

M. L. Mackenzie and L. Scott-Hayward. Analysis of Tern data using spatially adaptive smoother-based models for Papa Westray and the Pentland Firth. Commissioned by JNCC. Technical report, University of St. Andrews. Contract Number: C10-0206-0387, March 2012.

M. L. Mackenzie, L. Scott-Hayward, and C. G. Walker. Distribution modelling of tern *sterna* sp. using aerial survey transect data for (i) Greater Thames and (ii) Liverpool Bay areas. Technical report, University of St. Andrews. Contract Number: C10-0206-0387, 2012.

C. MacLeod, C. Weir, C. Pierpoint, and E. Harland. The habitat preferences of marine mammals west of Scotland (UK). *Journal of the Marine Biological Association of the United Kingdom*, 87:157–164, 2007a.

K. Macleod, R. Fairbairns, A. Gill, B. Fairbairns, J. Gordon, C. Blair-Myers, and E. C. M. Parsons. Seasonal distribution of minke whales, *Balaenoptera acutorostrata*, in relation to physiography and prey off the Isle of Mull, Scotland. *Marine Ecology Progress Series*, 277:263–274, 2004.

R. MacLeod, C. MacLeod, J. Learmonth, P. Jepson, R. Reid, R. Deaville, and G. Pierce. Mass-dependent predation risk and lethal dolphin-porpoise interactions. *Proceedings of the Royal Society B: Biological Sciences*, 274(1625):2587–2593, 2007b.

R. Maggini, A. Guisan, and D. Cherix. A stratified approach for modeling the distribution of a threatened ant species in the Swiss National Park. *Biodiversity and Conservation*, 11(12):2117–2141, 2002.

R. Maggini, A. Lehmann, N. E. Zimmermann, and A. Guisan. Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography*, 33(10):1729–1749, 2006.

Mainstream. Neart na Gaoithe Proposed offshore wind farm scoping report - appendices. Technical report, Mainstream Renewable Power, www.nearthnagaoithe.com, 2009.

B. Manly, L. McDonald, D. Thomas, T. McDonald, and W. Erickson. *Resource selection by animals: statistical analysis and design for field studies.* Nordrecht, Netherlands: Kluwer, 2002.

Marine-Scotland. Marine Protected Areas in Scotland's seas. Guidelines on the selection of MPAs and development of the MPA network. 2011.

G. Marra, D. L. Miller, and L. Zanin. Modelling the spatiotemporal distribution of the incidence of resident foreign population. *Statistica Neerlandica*, 66(2):133–160, 2011.

F. Marubini, A. Gimona, P. Evans, P. Wright, and G. Pierce. Habitat preferences and interannual variability in occurrence of the harbour porpoise *Phocoena phocoena* off northwest Scotland. *Marine Ecology-Progress Series*, 381:297–310, 2009.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 2^{nd} edition, 1989.

T. L. McDonald, W. P. Erickson, and L. L. McDonald. Analysis of count data from before-after control-impact studies. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 262–279, 2000.

G. K. Meffe. Conservation science and public policy: only the beginning. *Conservation Biology*, 13(3):463–464, 1999.

W. Mendenhall. *Statistics for Management and Economics*. Duxbury Press, Boston, fourth edition, 1982.

B. R. Missildine, R. J. Peters, G. ChinÄêLeo, and D. Houck. Polychlorinated biphenyl concentrations in adult chinook salmon (*Oncorhynchus tshawytscha*) returning to coastal and puget sound hatcheries of washington state. *Environmental Science and Technology*, 39(18):6944–6951, 2005.

J. Mobley. Aerial surveys for marine mammals performed in support of USWEX exercises 200. Technical report, Environmental Division, US Pacific Fleet, 2008.

M. L. Morrison, W. M. Block, M. D. Strickland, B. A. Collier, and M. J. Peterson. *Wildlife study design*. Springer, 2008.

K. P. Murphy. Conjugate bayesian analysis of the gaussian distribution. Technical report, University of British Colombia, 2007.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 10: 186–190, 1964.

M. Naud, B. Long, J. Brêthes, and R. Sears. Influences of underwater bottom topography and geomorphology on minke whale (*Balaenoptera acutorostrata*) distribution in the Mingan Islands (Canada). *Journal of the Marine Biological Association of the UK*, 83 (04):889–896, 2003.

S. L. Nieukirk, K. M. Stafford, D. K. Mellinger, R. P. Dziak, and C. G. Fox. Low-frequency whale and seismic airgun sounds recorded in the mid-Atlantic Ocean. *The Journal of the Acoustical Society of America*, 115:1832, 2004.

L. Nilsson. Habitat selection, food choice, and feeding habits of diving ducks in coastal waters of south Sweden during the non-breeding season. *Ornis Scandinavica*, pages 55–78, 1972.

H. Nix, J. McMahon, and D. Mackenzie. *The potential for pigeon pea in* A*ustralia.* P*roceeedings of* P*igeon* P*ea.*, chapter Potential areas of production and the future of pigeion pea and other grain legumes in Australia. University of Queensland, Queensland, Australia, 1977.

NMFS. Endangered and threatened wildlife and plants: Endangered status for southern resident killer whales. In *Federal Register*, volume 20 (222), pages 69903–69912. National Marine Fisheries Service, 2006.

NMFS. Protective regulations for killer whales in the northwest region under the endangered species act and marine mammal protection act. In *Federal Register*, volume 76 (72), pages 20870–20890. National Marine Fisheries Service, 2011.

NOAA. Southern resident killer whales research update. Technical report, National Oceanic and Atmospheric Administration, 2011.

S. Northridge, M. Tasker, A. Webb, and J. Williams. Distribution and relative abundance of harbour porpoises (*Phocoena phocoena* L.), white-beaked dolphins (*Lagenorhynchus albirostris* Gray), and minke whales (*Balaenoptera aucutorostra* Lacepede) around the British Isles. *ICES Journal of Marine Science*, 52(1):55–66, 1995a.

S. Northridge, M. Tasker, A. Webb, and J. Williams. ERRATUM: Distribution and relative abundance of harbour porpoises (*Phocoena phocoena* L.), white-beaked dolphins (*Lagenorhynchus albirostris* Gray), and minke whales (*Balaenoptera aucutorostra* Lacepede) around the British Isles. *ICES Journal of Marine Science*, 52(1):55–66, 1995b.

D. Nychka and N. Saltzman. Design of air-quality monitoring networks. In D. Nychka, L. Cox, and W. Piegorsch, editors, *Case Studies in Environmental Statistics, Lecture Notes in Statistics*, pages 51–75. Springer Verlag, New York, 1998.

D. Palka. Effects of Beaufort sea state on the sightability of harbor porpoises in the Gulf of Maine. *Reports of the International Whaling Commission*, 46(575-582), 1996.

D. Palka and P. Hammond. Accounting for responsive movement in line transet estimates of abundance. *Canadian Journal of Fisheries and Aquatic Science*, 58(4):777–787, 2001.

W. Pan. Akaikes's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125, 2001a.

W. Pan. Model selection in estimating equations. *Biometrics*, 57:529–534, 2001b.

S. Panigada, M. Zanardelli, M. MacKenzie, C. Donovan, F. Melin, and P. Hammond. Modelling habitat preferences for fin whales and striped dolphins in the Pelagos Sanctuary (western Mediterranean Sea) with physiographic and remote sensing variables. *Remote Sensing of Environment*, 112(8):3400–3412, 2008.

R. L. Parker and J. A. Rice. Discussion of "some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B.W. Silverman. *Journal of the royal Statistical Society, Series B*, 47:40–42, 1985.

E. Parsons, J. Shrimpton, and P. Evans. Cetacean conservation in northwest Scotland: perceived threats to cetaceans. *European Research on Cetaceans*, 13:128–133, 1999.

E. Parsons, I. Birks, P. Evans, J. Gordon, J. Shrimpton, and S. Pooley. The possible impacts of military activity on cetaceans in West Scotland. *European Research on Cetaceans*, 14: 185–190, 2000.

C. G. M. Paxton and L. T. Thomas. Phase one data analysis of Joint Cetacean Protocol data. Technical report, University of St. Andrews. Contract Number: C09-0207-0216 for the Joint Nature Conservation Committee, 2010.

C. G. M. Paxton, M. Mackenzie, M. Burt, E. Rexstad, and L. T. Thomas. Phase II data analysis of Joint Cetacean Protocol data resource. Technical report, University of St. Andrews: Contract Number: C11-0207-0421 for Joint Nature Conservation Committee, 2011.

C. G. M. Paxton, M. Mackenzie, M. Burt, E. Rexstad, and L. T. Thomas. Revised phase III data analysis of Joint Cetacean Protocol data resource. Technical report, University of St. Andrews: Contract Number: C11-0207-0421 for Joint Nature Conservation Committee, 2013.

J. Pearce and S. Ferrier. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133:225–245, 2000.

R. Pearson, T. Dawson, P. Berry, and P. Harrison. SPECIES: a spatial evaluation of climate impact on the envelope of species. *Ecological Modelling*, 154(3):289–300, 2002.

R. G. Pearson, T. P. Dawson, and C. Liu. Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, 27(3):285–298, 2004.

I. Petersen, M. MacKenzie, E. Rexstad, M. Wisz, and A. Fox. Comparing pre- and post-construction distributions of long-tailed ducks, *Clangula hyemalis*, in and around the nysted offshore wind farm, denmark : a quasi-designed experiment accounting for imperfect detection, local surface features and autocorrelation. Technical Report 2011-1, University of St. Andrews, 2011.

I. K. Petersen, T. Christensen, J. Kahlert, M. Desholm, and A. Fox. Final results of bird studies at the offshore windfarms at Nysted and Horns Rev, Denmark. report to Dong Energy and Vattenfall A/S. Technical report, National Environmental Research Institute, Ronde, Denmark, 2006.

S. J. Phillips, M. Dudík, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83. ACM, 2004.

S. J. Phillips, R. P. Anderson, and R. E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3):231–259, 2006.

E. Pinn. Threshold for designation of special areas of conservation for harbour porpoise and other highly mobile, wide ranging marine species. Technical report, Joint Nature Conservation Committe, 2009.

A. Pintore, P. Speckman, and C. Holmes. Spatially adaptive smoothing splines. *Biometrika*, 93(1):113–125, 2006.

J. Potts and J. Elith. Comparing species abundance models. *Ecological Modelling*, 199: 153–163, 2006.

T. Quinn. *The Behavior and Ecology of Pacific Salmon and Trout.* University of Washington Press, Seattle, WA, USA, 2005.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer, New York, 2^{nd} edition, 2005.

T. O. Ramsay. *A Bivariate Finite Element Smoothing Spline Applied To Image Registration.* PhD thesis, Dept. of Mathematics and Statistics, Queen's University, Canada, 1999.

T. O. Ramsay. Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1):307–319, 2002.

S. L. Rathbun. Spatial modelling in irregularly shaped regions: Kriging estuaries. *Environmetrics*, 9:109–129, 1998.

K. L. Raum-Suryan and J. T. Harvey. Distribution and abundance of and habitat use by harbor porpoise, *Phocoena phocoena*, off the northern San Juan Islands, Washington. *Fishery Bulletin*, 96(4):808–822, 1998.

M. J. Rayner, M. N. Clout, R. K. Stamp, M. J. Imber, D. H. Brunton, and M. E. Hauber. Predictive habitat modelling for the population census of a burrowing seabird: A study of the endangered Cook's petrel. *Biological conservation*, 138(1):235–247, 2007.

A. Read, P. Halpin, B. Best, E. Fujioka, C. Good, L. Hazen, E. LaBrecque, S. Qian, and R. Schick. Predictive spatial analysis of marine mammal habitats. Technical report, Duke University Marine Laboratory, North Carolina, 2009.

A. Read, P. Halpin, L. Crowder, B. Best, and E. Fujioka. OBIS-SEAMAP: Mapping marine mammals, birds and turtles. World Wide Web, 2011. URL `http://seamap.env.duke.edu`.

A. J. Read. Harbour porpoise *Phocena phocena* (linnaeus. 1758). *Handbook of Marine Mammals*, 6:323–355, 1999.

J. Reid, P. Evans, and S. Northridge. Atlas of cetacean distribution in north-west European waters. Technical report, Joint Nature Conservation Committee, Peterborough, 2003.

P. Reijnders, G. Donovan, A. Bjørge, K.-H. Kock, S. Eisfeld, M. Scheidat, and M. L. Tasker. ASCOBANS conservation plan for harbour porpoises (*Phocoena phocoena* L.) in the North Sea. Technical report, 6th Meeting of the Parties to ASCOBANS, 2009.

C. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10:177–183, 1967.

J. E. Reynolds III, H. Marsh, and T. J. Ragen. Marine mammal conservation. *Endangered Species Research*, 7:23–28, 2009.

D. Rice. Marine mammals of the world: systematics and distribution. *Spec Publ 4. Society for Marine Mammalogy*, 1998.

H. Ross and B. Wilson. Violent interactions between bottlenose dolphins and harbour porpoises. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 263 (1368):283–286, 1996.

J. T. Rotenberry, K. L. Preston, and S. T. Knick. GIS-based niche modeling for mapping species' habitat. *Ecology*, 87(6):1458–1464, 2006.

D. Ruppert. Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics*, 42(2):205–223, June 2000.

D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):pp. 735–757, 2002.

D. Ruppert, M. Wand, and R. Carroll. *Semiparametric Regression.* Cambridge University Press, 2003.

R. Rust, D. Simester, R. Brodie, and V. Nilikant. Model selection criteria: An investigation of relative accuracy, posterior probabilities, and combinations of criteria. *Management Science*, 41:322–333, 1995.

P. Sanchez, M. Demestre, L. Recasens, F. Maynou, and P. Martin. Combining GIS and GAMs to identify potential habitats of squid *Loligo vulgaris* in the Northwestern Mediterranean. *Essential Fish Habitat Mapping in the Mediterranean*, pages 91–98, 2008.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

L. Scott-Hayward, L. Milazzo, and C. Donovan. MMAARS project - software testing report. Technical report, CREEM, University of St. Andrews, 2010.

L. Scott-Hayward, M. L. Mackenzie, C. R. Donovan, C. G. Walker, and E. Ashe. Complex Region Spatial Smoother (CReSS). *Journal of Computational and Graphical Statistics*, 2013. doi: 10.1080/10618600.2012.762920.

R. Shucksmith, N. H. Jones, G. W. Stoyle, A. Davies, and E. F. Dicks. Abundance and distribution of the harbour porpoise (*Phocoena phocoena*) on the north coast of Anglesey, Wales, UK. *Journal of the Marine Biological Association of the United Kingdom*, 89(05): 1051–1058, 2009.

B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47: 1–52, 1985.

H. Skov, J. Durinck, F. Danielsen, and D. Bloch. Co-occurrence of cetaceans and seabirds in the northeast Atlantic. *Journal of Biogeography*, pages 71–88, 1995.

E. Smith. BACI *design.* E*ncyclopedia of* E*nvironmetrics.* Wiley: New York, 2002.

C. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 13: 689–705, 1985.

J. Teilmann and J. Cartensen. Negative long term effects on harbour porpoises from a large scale offshore wind farm in the Baltic - evidence of slow recovery. *Environmental Research Letters*, 7(4), 2012.

J. Teixeira and J. Arntzen. Potential impact of climate warming on the distribution of the Golden-striped salamander, *Chioglossa lusitanica*, on the Iberian Peninsula. *Biodiversity and Conservation*, 11:2167–2176, 2002.

C. D. Thomas, A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. Erasmus, M. F. De Siqueira, A. Grainger, L. Hannah, et al. Extinction risk from climate change. *Nature*, 427(6970):145–148, 2004.

L. Thomas, S. Buckland, E. Rexstad, J. L. Laake, S. Strindberg, S. L. Hedley, J. R. Bishop, T. A. Marques, and K. P. Burnham. Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology*, 47:5–14, 2010.

V. L. Todd, W. D. Pearse, N. C. Tregenza, P. A. Lepper, and I. B. Todd. Diel echolocation activity of harbour porpoises (*Phocoena phocoena*) around North Sea offshore gas installations. *ICES Journal of Marine Science: Journal du Conseil*, 66(4):734–745, 2009.

A. Underwood. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. *Journal of experimental marine biology and ecology*, 161(2): 145–178, 1992.

UNEP. Global synthesis: A report from the regional seas conventions and action plans. Marine Biodiversity Assessment & Outlook Series. United Nations Environment Programme: Regional Seas, UNEP Regional Seas Programme, October 2010.

W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, fourth edition, 2002.

M. Vinther and F. Larsen. Updated estimates of harbour porpoise (*Phocoena phocoena*) bycatch in the Danish North Sea bottom-set gillnet fishery. *Journal of Cetacean Research Management*, 6:19–24, 2004.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

C. Walker, M. Mackenzie, C. Donovan, and M. O'Sullivan. SALSA - a Spatially Adaptive Local Smoothing Algorithm. *Journal of Statistical Computation and Simulation*, 81(2): 179–191, 2010.

R. Walker, V. Sviridov, S. Urawa, and T. Azumaya. Spatio-temporal variation in vertical distributions of pacific salmon in the ocean. *North Pacific Anadromous Fish Commission*, 4:193–201, 2007.

H. Wang and M. G. Ranalli. Low-rank smoothing splines on complicated domains. *Biometrics*, 63(1):209–217, 2007.

R. S. Waples, R. W. Zabel, M. D. Scheuerell, and B. L. Sanderson. Evolutionary responses by native species to major anthropogenic changes to their ecosystems: Pacific salmon in the Columbia River hydropower system. *Molecular Ecology*, 17:84–96, 2007.

E. Ward, E. Holmes, and K. Balcomb. Quantifying the effects of prey abundance on killer whale reproduction. *Journal of Applied Ecology*, 46(632 - 640), 2009.

G. S. Watson. Smooth regression analysis. *Sankhya (Indian Journal of Statistics); Series A*, 26:539–572, 1964.

WDFW. Washing department of fishing and wildlife. World Wide Web, 2013.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

R. Williams, D. Lusseau, and P. Hammond. Estimating relative energetic costs of human disturbance to killer whales (*Orcinus orca*). *Biological Conservation*, 133:301–311, 2006.

R. Williams, K. Kaschner, E. Hoyt, R. Reeves, and E. Ashe. Mapping large-scale spatial patterns in cetacean density: Preliminary work to inform systematic conservation planning and MPA network design in the northeastern Pacific. Report for WDCS, the Whale and Dolphin Conservation Society, 2011.

S. N. Wood. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62: 413–428, 2000.

S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, 2006.

S. N. Wood. *soap: Soap film smoothing*, 2010. R package version 0.1-5.

S. N. Wood, M. V. Bravington, and S. L. Hedley. Soap film smoothing. *J. R. Statist. Soc. B*, 70(5), 2008.

P. Wright, H. Jensen, and I. Tuck. The influence of sediment type on the distribution of the lesser sandeel, *Ammodytes marinus*. *Journal of Sea Research*, 44(3):243–256, 2000.

N. E. Zimmermann, T. C. Edwards, C. H. Graham, P. B. Pearman, and J.-C. Svenning. New trends in species distribution modelling. *Ecography*, 33(6):985–989, 2010.

W. Zucchini, D. L. Borchers, M. Erdelmeier, E. Rexstad, and J. Bishop. *WISP v1.2.6-1*. Institut fur Statistik und Okonometrie, Geror-August-Universitat Gottingen, Platz der Gottinger Seiben 5, Gottingen, Germany, 2007.