

NUTS AND BOLTS: SPATIAL STATISTICAL MODEL DEVELOPMENT FOR
FISHERIES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Elizabeth Paige Duskey

August 2020

© Elizabeth Paige Duskey

ALL RIGHTS RESERVED

NUTS AND BOLTS: SPATIAL STATISTICAL MODEL DEVELOPMENT FOR FISHERIES

Elizabeth Paige Duskey, Ph.D.

Cornell University 2020

Concerning fisheries, the question of *where* is as critical as the question of *how many*. This dissertation develops, tests, and applies two spatial models to data from one highly mobile, and one sedentary species. It (1) develops a Bayesian hidden Markov model (HMM) to estimate movement and habitat use of diadromous fish species, (2) applies this model to acoustic tracking data from the endangered New York Bight distinct population segment (DPS) of the Atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus*), (3) develops a Bayesian extension of a regression kriging (RK) model in which a penalized spline model estimates large-scale spatial trend and a variogram estimates fine-scale spatial autocorrelation, and (4) applies a modified version of this model to continuous belt transect survey data from Atlantic sea scallops (*Placopecten magellanicus*). The theme of this dissertation is thus the development and thorough testing of novel Bayesian spatial model structures through simulation, followed by application. While the process is itself consistent, the methods and biology and ecology of the species thus analyzed are not. The HMM with nested detection and habitat use models performed very well on simulated data, even with approximately 70% of the data removed. Results of the application to Atlantic sturgeon suggest that while males stage downriver from spawning grounds, females migrate to and from spawning grounds with little evidence of staging. The RK model, while stable, very often gave biased estimates of the variogram parameters. The balance between the penalized spline and the variogram was tenuous at best. That said, it appears to be capable of identifying whether or not fine-scale autocorrelation did indeed exist separate from that of large-scale spatial trend. Results from application of the extended Bayesian model to

Atlantic sea scallops compared to a frequentist and thus sequential application of the penalized spline and variogram models suggest that fine-scale trend may be of a much greater magnitude than originally thought. The results are themselves significant to the species in question. Furthermore, the process of testing and developing a model using biologically and ecologically informed simulations is highly effective, and constrains the infinite realm of possibilities to a manageable set of choices based in reality.

BIOGRAPHICAL SKETCH

Elizabeth Duskey comes from the small town of Camden, OH. She began her academic career as an engineering student at Purdue University in 2008 before transferring to Miami University of Oxford in 2009. There, she earned bachelor's degrees in both mathematics and zoology, studying nonlinear dynamical systems and trophic interactions in freshwater systems. She moved to central NY in 2014 to pursue her Ph.D. at Cornell University, where she ultimately worked under Dr. Patrick J. Sullivan in the Department of Natural Resources, studying and developing spatial statistical methods for fisheries in concert with the New York Department of Environmental Conservation and the National Marine Fisheries Service. She moves to Sweden in the fall of 2020 to study quantitative theoretical ecology under the auspices of a joint venture between the State University of New York College of Environmental Science and Forestry and the Swedish University of Agricultural Sciences.

For my family, who misses me when I'm gone, but who still encourages me to go.

ACKNOWLEDGEMENTS

"...mais *apud me omnia fiunt Mathematicè in Natura*...[...but in my opinion, all things in nature occur mathematically...]" – René Descartes (1640)

First and foremost, I would like to express deepest gratitude to Patrick Sullivan, who took me under his advisement in the winter of 2016, and who encouraged me to leap from the shores of determinism and into the ocean of stochasticity. I would not have nearly the breadth of knowledge or acquaintance had he not so graciously allowed me to stow away on nearly all of his academic trips during my tenure at Cornell. I would also like to thank the other members of my committee. Suresh Sethi, for allowing me to take quantitative lead on a delightful team project alongside my fellow graduate students, and generally encouraging and expanding both my quantitative and my ecological skills. Andy Royle, for being an excellent neighbor, and an excellent sounding board as I pieced my models together. And Dvora Hart, who was a gracious host on multiple occasions, and so very generous with her time and her experience as I navigated the intersection of obscure theorems of probability with the reality of federal fisheries management.

I would also like to thank the New York State Department of Environmental Conservation, Region 3 of New Paltz, NY for allowing me access to their hard won data, and to accompany their employees on several of their field excursions. I would especially like to thank Amanda Higgs, who indulged me with long chats on multiple occasions, and whose intimate, field-based knowledge of the Hudson River and its ecological realities tempered my tendency to hover in the ivory tower.

The Northeast Fisheries Science Center of the National Marine Fisheries Service played host to me several times throughout my time at Cornell University. Their

generosity, and eagerness to share their passion for their work instilled in me a deep and enduring affection for the waters of the Atlantic Ocean. I would like to thank Jui-Han Chang, who always made herself available to answer my questions, to provide data and statistical advice in the wee weekend hours, and to chat about our mutual enjoyment of the macabre.

I would also like to thank the staff, both current and past, of the Department of Natural Resources at Cornell University. Shannon Hovencamp, Brian Hutchinson, Christie Sayre, Sarah Gould, and Melinda von Gordon, who were at times friends, and at other times guides and mentors. Both thanks for their generosity and half-hearted apologies for my endless chatter are due. Half-hearted apologies are also due to Patrick Sullivan and Suresh Sethi, who bore the brunt of my wont to wander the building shoeless and chatter about anything and everything. I would also like to thank Evan Cooch, who both fostered and encouraged my enjoyment of teaching.

Thanks are also due to my fellow graduate students and post-doctoral fellows, especially Kimberly Fitzpatrick, Elise Edwards, Verena Jauß, Taylor Brown, and Ben Marcy-Quay. Your advice, your sympathetic ears, your humor, your willingness to get ice cream at any given time, and your memes were indispensable to my experience in higher education.

Finally, I would like to thank my family, who are the heart and soul of everything I do. My mother, who dared to accompany a fellow southern Midwesterner on an epic journey to the Northeast. My sister, Chandra, for always cheering me onwards, and making me believe I could overcome any obstacle. My brother, Nick, for always making me feel at home, and for reminding me to have fun from time to time. Their spouses, Corey and Kate, and of course their children, Luke, Vince, Fiona, Caiden, Susie, and

Amelia, who all remind me what it truly means to wonder. My grandmother, Doris, with whom I have spent some of the best days of my life. My late grandfather, Ralph, who taught me to love and respect all wild things. The rest of my extended family, grandparents, aunts, uncles, cousins, from Massachusetts to California. And my very best friends, Mattea and Jenn. To all my family, including my friends: I don't ever wonder about the meaning of life; it's you.

INTRODUCTION

Spatial methods arise in fisheries when we consider that population parameters are not uniform throughout a given species' range. The simplest spatial structure divides a species' range into two or more sub-areas, *within* which we assume parameters are uniform. We can estimate these parameters by employing a simple stratification scheme (Cochran, 1977). We can elaborate upon this method by including dynamic and biased movement of fishes between adjacent sub-areas (Beverton and Holt, 2012). We can add additional layers of complexity to this, or any other spatial method, by considering movement or mortality as a function of one or more environmental variables, which are themselves distributed in space. This process can be complicated by the structure of the data itself, especially if it is fishery dependent, and therefore often biased towards regions with higher fish densities. Failure to incorporate the spatial structure of density or abundance indices, and of fishing mortality, can lead to poor management advice (Booth, 2000), and thus to improper management.

Spatial structure can arise directly as a result of movement, or indirectly as a result of spatially dependent mortality and survival. These two sources of spatial variability are not mutually exclusive, and can interact to produce a complex mosaic of fish density or abundance (Planque et al., 2011), length (Petitgas et al., 2003), and sex ratio (Britton and Moser, 1982). Describing these phenomena is a matter of choosing one or more proper methods of analysis. The quantitative methods at our disposal will differ depending on the mobility of the species, and the spatial and temporal scale under consideration. Highly mobile species require a method that explicitly models their location in both space and time, whether discrete or continuous. The spatial distribution of sedentary species is largely unaffected by fine-scale movements, and therefore we must choose from a set of methods that describe aggregation or repulsion

as an explicit or implicit function of environmental or biological variables. In this dissertation, we build and discuss models of both highly mobile and sedentary species. We discuss models that reflect our collective knowledge of the species' biology and ecology.

In Chapter 1, we built a Bayesian hidden Markov model (HMM) to describe movement and habitat use of diadromous fish species. We consider the effects of multiple spatial scales. Diadromous fish are capable of traveling large distances over relatively short periods of time. Behaviors such as feeding and spawning occur on a much finer scale, and are often dependent upon the surrounding habitats. We construct the model as four nested processes: movement, habitat selection, detection, and sampling. This nested structure begs for a Bayesian approach, in which the parameters of a given process may serve as hyperparameters for another. We represent movement among discrete areas of a hypothetical river in two different ways: with a hidden Markov process, and a standard Markov process. The hidden process allows us to provide separate structures for upriver and downriver movement. The standard Markov process produces trajectories with roughly equivalent upriver and downriver movement behavior. We use these processes to generate trajectories of several hypothetical fish, and use the sampling process to mimic the detection capabilities of a static acoustic array. This removes about 70% of the raw simulated data. Even so, we find that the model was capable of teasing apart each parameter used to simulate the data. While it faltered in representing rare occupation of highly unfavorable habitats, it performed very well overall. This novel spatial method for diadromous fishes is an amalgam of several, relatively simple model structures. A Bayesian application allows us to find optimal parameter estimates for all pieces simultaneously. The purpose of this study is not to prove a given model structure is capable of identifying parameters from data generated

from a similar structure. Rather, it is to show that, given the model is an appropriate hypothesis for the structure of movement and habitat of diadromous fishes, it is capable of identifying parameters despite a large percentage of missing data. Given the model's performance, we test it on real diadromous fish data.

In Chapter 2, we apply the model structure developed in Chapter 1 to data on migration of Atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus*) in the Hudson River. We apply it to males and females separately, hypothesizing that, while males linger and stage downriver of historical spawning grounds, females arrive later, and migrate quickly both upriver and downriver. The model serves as evidence in support of both of these hypotheses. It identifies at least one staging areas for males, in addition to a potential spawning area near Troy Dam, which represents the upper limit of Atlantic sturgeon distribution within the Hudson River. We include several habitat variables in the habitat selection portion of the model, including water depth, sediment type, sediment environment, and sediment morphology. We structure the model so that movement occurs at the scale of discrete 15 river kilometer pieces of the river. Habitat selection occurs at the scale of just 1 river kilometer. At this scale, there appears to be no preference for any given sediment type, environment, or morphology. However, both male and female sturgeon appear to prefer relatively shallow depths. Relative, that is, to the maximum available within a given river kilometer. This may reflect an actual preference for shallow waters, or it may reflect the difficulty of detecting adult Atlantic sturgeon in exceptionally deep waters. Curiously, the inclusion of a detection probability model generated extreme autocorrelation of subsequent draws in the Markov chain Monte Carlo (MCMC) process in **JAGS**. This indicates that detection is correlated with another piece of the model, most likely the habitat selection piece. Therefore, we calculate detection probability outside of the model, and find that we

are missing about 73% of all fishes' daily trajectories, a number that is very similar to that used in the detection portion of the simulations of Chapter 1. The model's capability of identifying regions of attraction of the river, indexed by date of arrival, could serve as a powerful tool for identifying habitats and spatial reaches that are in need of protection during the Atlantic sturgeon's historical spawning migration.

In Chapter 3, we shift our attention to a sedentary species. We develop a Bayesian interpretation of a regression kriging model, in which we use a penalized spline to describe trend, and a variogram to describe spatial autocorrelation. This is not a novel model structure, but rather a detailed construction of a method that is relatively well-known. We investigate the situation in which the trend model, that is the penalized spline, describes a large-scale spatial process and the variogram describes a fine-scale spatial process. To do so, we generate a mean from a non-zero Gaussian process, and add autocorrelated residuals that operated at a range that is just a fraction of the range of the data set at large. Generally, in regression kriging, one applies a regression model over one or more auxiliary variables, and subsequently kriges the residuals. However, we take a Bayesian approach, which allows us to estimate all parameters simultaneously via joint probability. This, too, is not entirely without precedent, though exceedingly rare (see Yang and Ng (2019)). Given that we are both regressing and kriging over space, choices regarding the penalized spline have the potential to bias the kriging results. When applying a penalized spline, or the closely related generalized additive model (GAM), one often chooses the number of knots. More knots may potentially allow the penalized spline more freedom to bend to fit the data. Naturally, then, an exceedingly "bendy" penalized spline may interfere with the range and magnitude of the variogram, thus providing a distorted picture of large-scale trend versus fine-scale autocorrelation. We test the effect of the choice of the number of knots on the

interaction of the penalized spline with the range and magnitude of the variogram. We find that knot choice has a significant effect on the parameters of the variogram. Stochasticity in the data seems to have an effect on the estimates of the parameters, as does the dimensionality of the data. While a fickle model choice in some respects, it does appear to be capable of identifying when fine-scale autocorrelation is indeed present. As above, the choice of this model structure itself contains both hypotheses and strong prior beliefs about the structure of the data. It should only be applied when there is reliable evidence that tangible biological or ecological processes are generating separate large-scale and fine-scale spatial trends or autocorrelation. This leads to the following chapter.

In Chapter 4, we develop a Bayesian update of a model first built by Chang et al. (2017). They used a generalized additive model to describe trends in density of Atlantic sea scallops (*Placopecten magellanicus*) with both water depth, and with two-dimensional space. Trends over depth can be explained in part by the interaction of temperatures for physiological optima with predator density. Trends over two-dimensional space can be described in part by oceanic transport of larvae and sediment type. At finer scales, the presence of patchy habitat – such as dead scallop and clam shells, sediment types such as gravel and mud, absence of predators, and currents – improves settlement and survival of larvae, and therefore individuals tend to aggregate. A variogram can describe the spatial range and magnitude of this process. Thus, Chang et al. (2017) kriged the residuals from the GAM. They termed this model structure GAM+OK, and applied to each of several stratified areas within Georges Bank and the Mid-Atlantic Bight on the Atlantic coast of the United States. We update this structure by applying a Bayesian model to estimate both the depth and spatial trends and the variogram of the residuals simultaneously, and compared them to the results

of the GAM+OK method. The range of the variogram, and thus of the fine-scale aggregation, is often similar or significantly smaller as predicted by the Bayesian model than as predicted by the GAM+OK method. The sill of the variogram, and thus the magnitude of the effect of aggregation, is always at least one order of magnitude larger as predicted Bayesian model than as predict by the GAM+OK method. This suggests that, when allowed to interact directly, the model believes fine-scale autocorrelation plays a much more significant role than when acting separately from the GAM. That said, the abundance estimates of both models are similar, with just one gross exception. These differences may be explained in part by the scale of the data. The data source, a belt transect benthic survey with continuous photographic counts and measurements, produced many thousands of data points. These were aggregated to form more manageable data sets, and aggregated further still for the sake of Bayesian estimation via MCMC, which can have punishingly slow run-times. Ultimately, the output from this model is meant to serve as a template for future projection. We intend to generate estimates, add additional weight via growth and additional individuals via stock-recruitment, and use the results as informative priors for subsequent years.

All four of these chapters can be divided into pairs, the first two and the last two, the first of each a story of careful model development and testing, and the second of each a story of model application, with the expected *un*-expected caveats of complex spatial and temporal ecological data. Spatial model development is a matter of piecing together mathematical and statistical structures in a coherent framework, and recognizing that each choice exerts a prior belief, or represents an ecological hypothesis. We have detailed the construction of each model with the express intent that they be immediately reproducible. We have provided examples of each model as tested on simulated data because this allows us to understand the behavior of the models in

a controlled environment, which thus prepares us to understand their behavior when applied to real data sets. Both methods described and tested above add additional power and sophistication to the set of spatial methods available to fisheries scientists.

List of Tables

1.1	Model run details	28
2.1	Biological information of all fish in study	63
2.2	GLM of presence as a function of habitat type	72
2.3	Model run details	74
2.4	Estimated HMM probability transition matrices	76
2.5	Habitat model results for males	80
2.6	Habitat model results for females	81
3.1	Variogram parameters as a function of peakedness	124
4.1	Model run details	161
4.2	All model results	163

List of Figures

1.1	Simulated fish trajectories	11
1.2	Simulation model chart	20
1.3	Posterior densities of transition probabilities of Standard Markov model applied to D-SM	30
1.4	Posterior densities of transition probabilities of Standard Markov model applied to D-HMM	31
1.5	Actual and simulated trajectories for Standard Markov model applied to D-SM	33
1.6	Actual and simulated trajectories for HMM model applied to D-SM . .	33
1.7	RMSE of both Standard Markov and HMM model applied to D-SM . .	34
1.8	Posterior densities of transition probabilities of Standard Markov model applied to D-HMM	35
1.9	Actual and simulated trajectories for Standard Markov model applied to D-HMM	36
1.10	Posterior densities of transition probabilities of HMM model applied to D-HMM	38
1.11	Posterior density of hidden transition matrix of HMM model applied to D-HMM	39
1.12	Actual and simulated trajectories for HMM model applied to D-HMM .	40
1.13	RMSE of both Standard Markov and HMM model applied to D-HMM .	41

1.14	Posterior density of habitat coefficients	42
1.15	Actual and estimated occupancy probability	43
1.16	Posterior density of detection probability	45
1.17	Median estimated trajectory	48
2.1	Hudson River map	57
2.2	Estimation model chart	66
2.3	Posterior density of transition matrices for males	77
2.4	Posterior density of transition matrices for females	79
2.5	Posterior density of habitat coefficients	82
2.6	Occupancy probability as a function of maximum depth	83
2.7	Median path simulated from estimated transition matrices	84
2.8	Posterior density of all trajectories	85
2.9	Example trajectories	85
2.10	Posterior density of detection probability	86
2.11	Sediment type as a function of depth	93
3.1	Two-dimensional mean generating processes	116
3.2	Three-dimensional mean generating processes	116
3.3	Median posterior of all two-dimensional model fits	120
3.4	Posterior density of two-dimensional sill	121
3.5	Posterior density of two-dimensional range	122
3.6	Two-dimensional variogram parameters as a function of peakedness	123
3.7	Two-dimensional variogram parameters and peakedness as a function of the number of knots	125
3.8	Normalized DIC as a function of the number of knots	126
3.9	Median posterior of all three-dimensional model fits	128

3.10	Posterior density of three-dimensional sill	129
3.11	Posterior density of three-dimensional range	130
3.12	Three-dimensional variogram parameters as a function of peakedness .	131
3.13	Three-dimensional variogram parameters and peakedness as a function of the number of knots	132
3.14	Model output sans autocorrelation	134
3.15	Median scallop count with depth and knots	135
3.16	Median scallop count with space and knots	136
3.17	Normalized median estimates of variogram parameters and peakedness	137
3.18	Spline vs. kriging variance	140
4.1	Posterior of probability of presence	164
4.2	Depth in the MAB and GB	166
4.3	Mean density as a function of depth	168
4.4	Mean density as a function of space	170
4.5	Mean density as a combined function of depth and space	171
4.6	Mean kriged density	173
4.7	Total estimated density	174
4.8	Estimated abundance in the MAB and GB	175

Contents

Abstract	i
Biographical Sketch	iii
Dedication	iv
Acknowledgements	v
Introduction	viii
List of Tables	xv
List of Figures	xvi
1 A simulation study of diadromous fish migration and habitat use	1
1.1 Introduction	2
1.2 Methods	6
1.2.1 Data Generation	7
1.2.2 Estimation Model	18
1.2.3 Implementation	25
1.3 Results	28
1.3.1 Movement	29
1.3.2 Habitat Preference	40
1.3.3 Detection and Observation	44
1.4 Discussion	44
1.5 Supplementary Materials	49

2	Migration and habitat use of Hudson River Atlantic sturgeon (<i>Acipenser oxyrinchus oxyrinchus</i>)	53
2.1	Background	54
2.2	Methods	61
2.2.1	Acoustic Tagging	61
2.2.2	Environmental Data	62
2.2.3	The Model	65
2.3	Results	75
2.3.1	Movement	75
2.3.2	Habitat Preference	78
2.3.3	Detection	84
2.4	Discussion	86
2.4.1	Scale	86
2.4.2	Movement	88
2.4.3	Habitat Selection	90
2.4.4	Detection	94
2.4.5	Conclusions and Summary	96
3	Spatial tug of war: kriging spline model residuals	97
3.1	Introduction	98
3.2	Problem Definition	102
3.2.1	Kriging	102
3.2.2	Penalized splines	105
3.3	Simulation Study	108
3.3.1	Estimation Model	108
3.3.2	Data Generation	112
3.3.3	Application	115

3.3.4	Simulation Study Results	119
3.3.5	Three-dimensional	126
3.3.6	No autocorrelation	132
3.3.7	Application	133
3.3.8	Fixed and Random Effects	138
3.4	Discussion	140
4	A Bayesian update of a method to estimate Atlantic sea scallop (<i>Placopecten magellanicus magellanicus</i>) abundance from belt transect surveys	147
4.1	Background	148
4.2	Methods	151
4.2.1	Data Collection	151
4.2.2	Estimation Model	152
4.3	Results	162
4.3.1	Probability of presence	163
4.3.2	Mean Abundance	165
4.3.3	Kriging	169
4.4	Discussion	172

Chapter 1

A simulation study of diadromous fish migration and habitat use

Diadromous fishes are characterized by their often large-scale migrations to and from marine waters. These migratory routes to and from estuaries represent a source of feeding and spawning habitat separate from that used during the majority of the fishes' adult lives. Particularly for spawning fishes, the quality of these habitats may have a profound impact on the overall health and fitness of a given population. For purposes of conservation and recovery, or to maintain a profitable and sustainable fishery, we may wish to predict where the fish are at a given time, and which habitat types are most critical. We can use this information to protect spawning habitats and feeding habitats, to redirect shipping activity, or any number of things conducive to a healthy population. Here, we present a modeling framework to describe the migration and habitat use of diadromous fishes, with which we can estimate habitats and regions within a given river in need of protection. We simulated the migratory routes of several hypothetical fishes using four nested processes: (1) a Markov process for movement (2) a categorical process for habitat preference (3) a binomial process for individual detection, and (4) a basic search process. We then attempted to estimate the parameters of these processes with a hierarchical Bayesian model structure, assuming that data arose from a linear

acoustic array. This framework serves both as a hypothesis for the processes that drive diadromous movement, and as a method by which to predict movement and habitat use for real data sets. We find that the simulation is a good approximation of diadromous behavior, and that the Bayesian model is able to identify and estimate all of the parameters with minimal error. This same structure may be useful for data on individual fish locations throughout a migration.

1.1 Introduction

Many diadromous fishes have seen marked declines in abundance since the late 18th and early 19th century. These fishes are particularly susceptible to declines for several reasons. They often have several small populations specific to a given freshwater system. While the metapopulation may be large, each distinct population is more susceptible to decline and thus extinction, than numbers alone would suggest. They also experience threats in both marine and freshwater systems. Their migrations make them relatively easy to fish, given they pass through narrow bodies of water within easy reach of human exploitation (Limburg and Waldman, 2009).

Aside from direct exploitation, habitat degradation or alteration is often cited as one of the most, if not the most, important source of population decline for diadromous fishes (Jonsson et al., 1999). This includes blockage of natural passages by dams, as well as changes in water quality, bottom substrate, flow, and temperature. These fishes may travel through sections of habitat which serve as little more than a passageway. Even so, habitat degradation and disturbance in these areas can contribute to declines in abundance and fitness (McDowall, 1992).

It can be challenging to manage the effects of habitat degradation and alteration on diadromous species, as their use of habitat changes rapidly in both space and time. As a part of an effective management program, it would be useful to be able to predict, with proper characterization of uncertainty, the location and habitat preference of fishes at a given time during their migration. To make these predictions, we must choose an appropriate modeling framework. Statistical methods in movement ecology have made large strides in recent years, due in part to the availability of large data sets of exceptionally high spatial resolution. In general, these methods are built from some variation on a Markov process, and can take several forms. The random walks, whether simple, biased, or correlated (Codling et al., 2008), introduce relatively simple dependence structure into the estimation process. State-space models (SSM; Jonsen et al. (2003)) allow for estimation of observation error. Hidden Markov models (HMM) allow for estimation of un-observable, or hidden, states (Baum and Petrie, 1966), for example, behavioral modes such as foraging or searching.

In choosing a model structure for diadromous fish migration, we must contend with the often tortuous geometry of the river itself. Most fishes cannot skip across dry land, and are therefore restricted by the boundaries of the river banks. On a broad scale, then, we can think of the migratory routes as linear. McDowall (1993) noted that, in modeling diadromous fish distribution and abundance, we must account for both broad and fine scale processes. On the broadest of scales, these fishes are migrating over very large distances. On a finer scale, particularly during staging and spawning, the fishes are occupying certain habitats, presumably of optimal suitability. The overall pattern of movement is highly directional and non-uniform. Iteroparous diadromous species, for example, reverse direction after spawning, and return to the waters in which they overwinter. Some species temporarily stage prior to spawning (e.g. High et al. (2006);

Godinho et al. (2007); Balazik and Musick (2015)), and may make more than one spawning run. So, in addition to multiple scales, we must also allow for some degree of non-stationary behavior.

Linear, multi-scale, and non-stationary; these are the criteria by which we searched for an appropriate model structure. We ultimately chose two different structures: (1) a standard Markov process and (2) a HMM. In the context of animal movement, the HMM structure is typically used to summarize data as a series of step lengths and turning angles (Michelot et al., 2016), with hidden states comprising of different sets of distributions producing these steps and turning angles. However, this not only requires error relative to step length to be small, it also requires at least two-dimensional positioning. While daily step lengths for diadromous fish are sometimes large, staging and spawning behaviors produce persistence that is only distinguishable at the scale of several hundred meters at best. Therefore, we described each hidden state with a separate probability transition matrix. This allows for a great deal of flexibility in choosing the size and number of spatial regions.

Upon choosing an area of the river according to one of two movement structures described above, each fish chooses a sub-area based on local habitat characteristics. This is a naturally hierarchical process, and thus one that lends itself well to Bayesian analysis. The Bayesian framework allows us to separate the broad scales of movement from the finer scales of habitat use, and then to estimate the parameters of both simultaneously. It also allows us to include some structure to describe the efficiency of our observation processes, given that the probability of detecting a fish by any method is typically less than 1.

We assert that this structure can absorb behaviors modeled explicitly by multi-state Markov processes. These behaviors might include foraging, traveling, or searching (McClintock et al., 2012). Diadromous fishes - and, in fact, other migratory fishes - often exhibit a high degree of fidelity to spawning and staging sites (e.g. Grabowski and Isely (2006); Klimley and Holloway (1999); Adams et al. (2011)). Assuming fidelity, behaviors that produce certain movement patterns will take place in the same regions of the river for most, if not all, individuals. For example, individuals in staging regions will show a relatively high probability of remaining in that spatial state. In contrast, individuals in migratory regions will show a relatively low probability of remaining in that spatial state, with directional pull either upriver or downriver. Therefore, we can infer staging or spawning regions from a high degree of persistence, and migratory routes from very low persistence. This would of course require additional observations in the field for confirmation.

We have built two simulations of diadromous movement and habitat use that attempt to capture the hierarchy of scale described above. Both simulations possess a different degree of structure to help capture non-stationarity of movement. We then estimated the parameters of the data generating processes using Markov chain Monte Carlo (MCMC) in **JAGS** (Plummer, 2017). These simulations are meant to serve both as a hypothesis for the processes driving movement and habitat use, as well as a method of estimation and prediction for use with data on diadromous fish locations. The ultimate goal is to build a general structure by which we can predict the trajectories of diadromous fishes, and quantify their preference for a given set of habitat characteristics on an appropriate scale. By identifying preferred habitats, in both space and time, we hope to contribute to the effective management and recovery of these fishes.

In this paper, we describe each process mathematically, for the sake of clarity and replicability. We demonstrate the processes of simulation, as well as the methods of estimation on the simulated data. We presented a detailed application of these methods to a data set on locations of Atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus*) in the Hudson River in Duskey et al. (2020).

1.2 Methods

We simulate daily movement and occupancy of 200 fishes over 100 days with four nested processes. We summarize the simulation process here, and then describe each step in detail in what follows. Without loss of generality, we discuss the model in reference to anadromous fish. For catadromous fish, simply reverse the direction of migration.

For movement, we use two separate variations on a discrete-time, discrete-space Markov process. All individual fish begin in the same spatial state. First, each fish chooses a subsequent location in the river based only on the previous location on the previous day. For the first 100 fish, we model both upriver and downriver movement with one process, thus assuming that there is no significant staging behavior prior to spawning. For the second 100 fish, we model upriver movement separately from downriver movement, thus including an explicit representation of staging behavior prior to spawning. We assume that all individuals' locations are known on their first day of migration, given that rivers are roughly linear systems and fishes cannot avoid traveling through the mouth. Second, after a fish has chosen from a finite set of discrete spatial states, it chooses a location on a finer spatial scale *within* that state based on the immediate surrounding habitat characteristics. The final two processes are both observational. A fish may be available for detection but not detected because of user

error or equipment failure. A fish may also not be available for detection because it is not in the area sampled in a given time unit. We simulated these data in R (R Core Team, 2018).

The hierarchical structure of the processes described above naturally led to a Bayesian analysis for the purpose of estimation. We use the simulations to explore the hypothesis that these processes are a viable basis for the movement and habitat use of diadromous fishes. We write the model with the intent of determining whether the parameters we used to simulate the data are estimable. If the simulation produces data that walk and talk like diadromous fishes, but the parameters are not or are poorly estimable after applying a somewhat realistic observation process, then we believe the simulation is of little use. We use **JAGS** (Plummer, 2017) to write the model, and the R package *jagsui* (Kellner, 2017) to run the model.

1.2.1 Data Generation

Movement

We assume fish movement is controlled by one of two Markov processes: (1) standard and (2) HMM. A standard Markov process consists of a probability transition matrix – hereafter referred to as a transition matrix – that guides the transitions from one state to another, and an initial distribution. Here, states are discrete spatial reaches in the hypothetical river. The model assumes all fish begin in state 1, and progress from there. State 1 marks the spatial location in which all fish begin their migration. For anadromous fish, this is the mouth of their natal rivers, through which they must pass to reach their spawning grounds. For catadromous fish, this is their overwintering habitat, from which they must travel downriver to reach their natal marine waters. If

a right stochastic transition matrix with N spatial states is given by

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1N} \\ \vdots & \ddots & \vdots \\ p_{N1} & \cdots & p_{NN} \end{bmatrix}$$

where each p_{ij} gives the probability of transitioning from state i to state j , and the initial distribution is given by

$$\alpha = \begin{bmatrix} p_1 & p_2 & \cdots & p_N \end{bmatrix}$$

where p_i gives the probability of beginning the migration in each of N states, then we can simulate the positions for individual $k \in \{1, \dots, K\}$, at discrete time units $t \in \{1, \dots, T\}$ with a multinomial distribution as follows

$$\begin{aligned} s_{k,1} &\sim \text{Multinomial}(\alpha) \\ s_{k,2} &\sim \text{Multinomial}(P[s_{k,1},]) \\ &\vdots \\ s_{k,T} &\sim \text{Multinomial}(P[s_{k,t-1},]) \end{aligned}$$

where s_{kt} is the spatial state occupied by individual k at time t . So, fish decide where to go during time step $t + 1$ based only on their location during time step t . The memorylessness of the Markov process certainly ignores some of the biological reality of complex migratory behavior. However, higher order Markov processes can introduce staggering complexity into the estimation process, and we therefore ignore these in favor of a first order process. No matter the order, the discrete nature of this formulation avoids problems associated with the boundaries between aquatic and

terrestrial habitat. That is, it prevents our hypothetical fish from walking on land, given that we have defined our states as linear sections of the river. Markov processes also have the fortunate capability of modeling non-stationary behavior between separate states. That is, within a given state, the probability distribution governing the direction and magnitude of movement can be radically different than that of movement within other states. The key is to choose a scale at which movement *within* each of these states is, or is approximately, stationary. Of course, in this study, the scale is arbitrary. We revisit the discussion of scale in relation to an actual data set in Duskey et al. (2020).

We have built two variations on the movement process. All other parts of the simulation are identical throughout both variations. Both variations also assume $N = 10$ states, although we invite the reader to change the number of states in the supplemental code if they so desire. The first variation of the model uses a first order Markov process. The second uses an HMM structure with two matrices, one for upriver movement and the other for downriver movement. At each time step, a static 2×2 hidden state transition matrix determines which matrix is guiding the movement of each individual. We describe both variations, and how we built them, below.

Standard Markov: This variation is just as described above, wherein the vector α and the matrix P determine the location of each of $K = 100$ individuals at time step 1, and time steps 2 through $T = 100$, respectively. We choose values for each element of the transition matrix (see Figure 1.3) to simulate bi-directional diadromous movement, with a roughly equal spread of weight around the main diagonal. We show a summary of these 100 simulated fishes' trajectories by spatial state in Figure 1.1, left. Note that, even with the Standard Markov structure, most fish travel to, and remain within, the upper central portion of the river. Note also that, despite this general tendency, some

simulated trajectories do exhibit a typical diadromous pattern of in- and out-migration (see Figure 1.5, left).

HMM: This variation on the Markov process is motivated by the hypothesis that upriver movement is not a mirror of downriver movement. Transition matrices can produce movement in both directions equally with a roughly equal spread of probabilistic weight about the main diagonal. However, it may be that different movement processes guide upriver and downriver migration. This is, of course, moot for semelparous species. For those that exhibit iteroparity, and thus undertake both an in- and out-migration, may follow different patterns during each leg of their journey. For example, adult Atlantic sturgeons (*Acipenser oxyrinchus oxyrinchus*) are iteroparous anadromous fish. They travel upriver to spawn. Prior to spawning, they often stage in portions of the river separate from their spawning grounds (Dovel and Berggren, 1983), and exhibit a certain amount of persistence in portions of the river downriver from the spawning grounds. After spawning, they typically exit the river swiftly, without lingering to feed. Such behavior may necessitate separate models for upriver movement and downriver movement. Any behavior that causes the fish to move in different patterns before and after spawning may lead to the same observed phenomena. Strong and swift upriver and downriver movement may also necessitate two separate matrices. A matrix that guides strongly directional movement naturally will not guide movement in the opposite direction. Our hypothetical fish may find themselves stuck at one end of our river. Thus, we use separate matrices to simulate upriver versus downriver movement, and employ a third matrix to guide the transition between upriver and downriver movement. This necessitates the use of an HMM.

An HMM allows us to simulate hidden states, that is states that are un-observable. When we observe a diadromous fish during its migration, we cannot say definitively

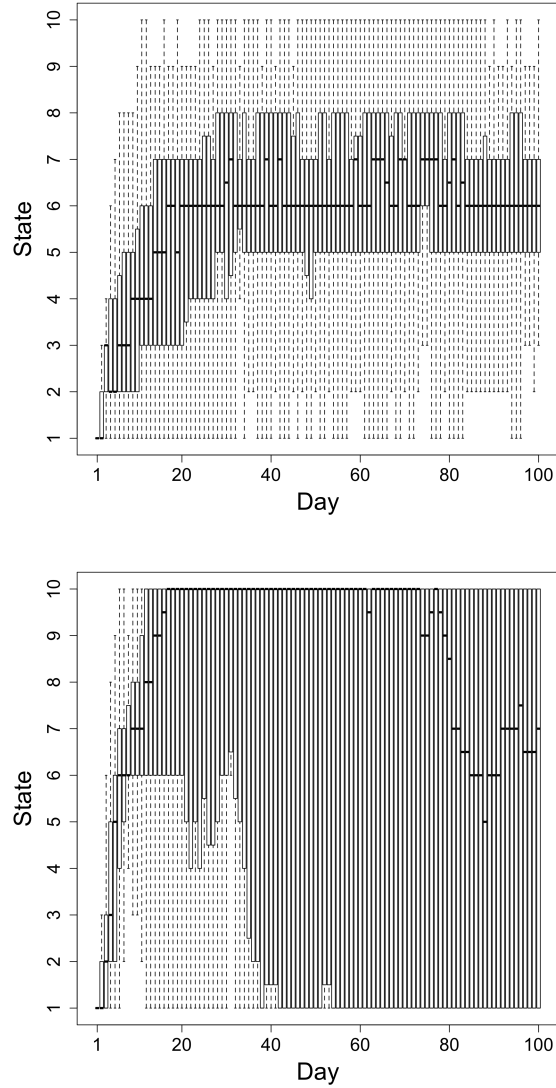


Figure 1.1: Box and whisker plots summarizing the simulated trajectories of all 100 fish which move according to a standard Markov process (above) and all 100 fish which move according to a standard Markov process (below).

whether it is in an upriver or downriver pattern of movement. However, for the purposes of simulation, we can build two separate transition matrices, one where all probabilistic weight is concentrated at or *above* the main diagonal, and another where all probabilistic weight is concentrated at or *below* the main diagonal. To mimic staging, we place relatively high probabilistic weight on transitioning to, and remaining in, intermediate states that lie below the terminal state in the upriver transition matrix (see Figure 1.10). To guide the transition between upriver and downriver movement behavior, we employ a third 2×2 matrix. This is the hidden Markov transition matrix.

Thus, we simulated location on $T = 100$ days for a total of $K = 100$ individual fish using two matrices, each with $N = 10$ states. For each day, we use a 2×2 hidden Markov transition matrix to determine whether each individual is guided by the upriver or downriver matrix. This matrix is given by

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$

where, for example, h_{12} is the probability of transitioning from upriver movement behavior (transition matrix 1) to downriver movement behavior (transition matrix 2). We chose values close to 1 for the main diagonal of H to mimic persistence in a given behavioral movement pattern. We show a summary of these 100 simulated fishes' trajectories by spatial state in Figure 1.1, right. Note that, by day 80, most fish have transitioned to a downriver movement state, and the median position for all 100 simulated trajectories begins to shift downriver from the hypothetical spawning grounds in state 10. This is in sharp contrast to the simulated trajectories given in Figure 1.1, right.

To summarize, we have 100 fish whose movement is guided by a standard Markov process, and thus by a single transition matrix P . We have another 100 fish whose movement is guided by an HMM. In each step, the hidden transition matrix H determines whether each individual is moving upriver or downriver, and thus whether movement is guided by matrix P_1 or P_2 . This data generating process produces $2K \times T$ matrices, with $K = 100$ and $T = 100$. We invite the reader to change the number of fish or the number of days in the supplementary code S.1.

Habitat Preference

In a given time step t , an individual fish k moves according to one of the variations on Markov processes given above. Once it has moved, it then decides where in the spatial state it would like to reside until the next time step. For this, we simulate both a habitat characteristic, and *preference* for that habitat characteristic.

We sub-divide each state into three sub-states. The choice of three is entirely arbitrary, and meant only to demonstrate that the movement and habitat selection processes may easily be given at different scales, as long as habitat selection occurs at a spatial scale equal to or finer than the scale of movement. We simulate a variable habitat characteristic in each of the sub-states. Without loss of generality, we call this characteristic "sand", and give each sub-state a random proportion of sand substrate. We mean for this to demonstrate the capacity of the simulation to account for substrate type or morphology. Minor modifications would allow for the inclusion of additional habitat variables, such as presence or absence of a given food source, water temperature or salinity, hydrology, and so forth. We choose just one arbitrary habitat characteristic to keep the simulation relatively simple and clear. We also assume that the preference for proportion sand substrate remains constant throughout the river,

though we acknowledge that habitat preference may be dynamic throughout an individual's migration. One could easily account for dynamic preference by estimating coefficients of preference separately in different portions of the river. We must caution that additional complexity can add considerable time to the estimation procedure, but this is a common problem in Bayesian analysis and by no means unique to this model structure.

We simulate habitat preference as an exponential function of the proportion sand, as follows

$$w_{n,l} = e^{a_0 + a_1 \text{ sand}_{n,l}}$$

where $w_{n,l}$ is a positive weight associated with occupying sub-state $l \in \{1, \dots, L_n\}$ of state $n \in 1, \dots, N$. The parameters a_0 and a_1 are the intercept and slope, respectively, of the exponential relationship between the weights and the proportion sand in sub-state l of state n . We choose $a_0 = -2$ and $a_1 = 2$, which reflects an affinity for an increase in the proportion sand in a given sub-state (see Figure 1.14). We use these weights to simulate the probability of occupying a given sub-state of a given state as follows

$$q_{n,1:L_n} \sim \text{Dirichlet}(w_{n,1:L_n})$$

where the vector $q_{n,1:L_n}$ gives the probability of occupying any of the L_n sub-states of state n . We use the Dirichlet distribution because, with an input of a vector of ordered positive real numbers of length L_n , its output is a vector of ordered real numbers of length L_n in the interval $(0, 1)$. Within a vector of weights $w_{n,1:L_n}$, those that are greater tend to produce higher numbers within $(0, 1)$. Ultimately, then, those sub-states with a greater proportion sand will produce greater relative weights, and

thus greater relative occupancy probabilities. In our case, we have set $L_n = 3$ for all $n \in 1, \dots, N$, $N = 10$. Therefore, each vector $q_{n,1:3}$ gives the probability of occupying each of the 3 sub-states of state n . Given that we write the spatial state of individual k on day t as $s_{k,t}$, then then we can write the sub-state as

$$\text{sub}_{k,t} \sim \text{Multinomial}(q_{s_{k,t},1:L_n})$$

The value of L_n does not have to be equal for all states, allowing the reader flexibility in assigning discrete sub-states.

The simulation procedure given above allows us to resolve location and habitat preference on a scale at least as fine as that of the movement process and, as we have noted, no larger. The division between movement and habitat selection is deliberate. We hypothesize that an individual fish's trajectory is not distributed uniformly throughout the portions of the river through which it travels, but that it is guided by the habitat available in the immediate vicinity. The extent of this nebulous "vicinity" will vary based on a given species' ability to perceive habitat characteristics over a given range. For fish that can move many river kilometers in a single day, this selection process necessarily takes place on a scale much finer than that of daily movement.

In summary, then, at each time step, each individual fish moves to, or remains in, a given spatial state. Subsequently, it chooses a sub-state based on the habitat characteristics within each of the sub-states. We assign the probability of occupying any sub-state in a given state based on the proportion "sand" within each sub-state. These remain constant throughout the simulation process. The resulting data are vectors of 100 spatial state values and 100 sub-state values for all 200 fish, for a total of 40,000 data points.

Detection Probability

This process, and the following process, are both observational, and are meant to simulate conditions found in the field. Here, we understand detection probability as an interaction between some characteristic of individual fish, and the technology we use to attempt to detect that individual. For example, if we use gill nets of a certain mesh size, fish above a certain size threshold are more likely to entangle in the netting. If we use acoustic tags, fish of a certain size or sex may occupy areas of a river that reduce the strength and reach of the signal or detection range of the equipment. Other mechanisms may interfere with our ability to detect a fish, given that it is in the location in which we have searched and therefore available for detection.

Given that an individual is indeed available for detection, we assume that the probability of detecting any fish on any day is a constant. Mechanisms more complicated than this may play a significant role in detecting an individual. Again, we choose a very simple mechanism for clarity of the simulation and estimation procedure. Note that, if detection probability is close to 1, it may not be necessary to include this process. However, most sampling methods are imperfect, and one may wish to know the extent to which non-detection may affect the data and the conclusions. For our simulations, we choose a detection probability of $p_d = 0.3$. That is, on any given day, we have a 30% chance of detecting each individual, no matter their location. We have found that this is a reasonable value for the detection of riverine fishes with an acoustic array (Duskey et al. 2020). We assume the location of each individual on the first day of their migration is always observed. Or, rather, even if it is un-observed, all individuals must begin at the mouth of the river. We apply the detection probability by simply drawing from a binomial distribution with $p = 0.3$ for each of the remaining 99 days for each of the 200 simulated trajectories. If the result is equal to 0, we remove the

observation, otherwise it remains. On average, then, we remove approximately 70% of the observations.

Search

The search process refers to the areas of the river that we search on any given day t . For our simulations, we assume a search process that covers the entire extent river every day. That is, at no point is there ever an individual that is not *available* for detection. Coupled with the detection probability process above, this mimics an acoustic array with 100%, non-overlapping coverage of the river, each receiver having a constant detection probability of 0.3. This is certainly an over-simplification. Typically, coverage is not 100%, though it may be quite high. If, for example, spatial state 2 contains no receiver, or is not within range of a receiver in an adjacent state, then we would remove all observations in state 2. We invite the reader to choose their own spatial sampling scheme. For example, if we use a mobile tracking unit by which we search only a portion of the river on a given day, the fish occupying areas outside of the portions we search are unavailable for detection. Note, then, that we have two sources of non-detection: that due to the detection probability of an individual, and that due to our sampling scheme. Our final data set consists of a set of locations for a given set of individuals, with missing values. If we do not include an observation process, we have no way of differentiating between the two sources of non-detection.

An additional advantage of the observation process is that we can explore how different sampling schemes might affect our power of estimation. If we have some prior information on individual movement, or even just some prior conjecture, we can construct a hypothesis of movement and habitat preference using the processes described above. Then we can use this observation process to explore what sort of sampling

scheme might best help us to answer whatever question we have in mind. For example, we may be interested in a certain portion of a given population's migratory path. We may find that a systematic scheme that searches the same small range of the river on a repeated schedule is too diffuse to capture some minute behaviors during this portion of migration. We might then choose a scheme that concentrates sampling in a given area at a given time. It is wise to be cautious, though, when translating this simulated work to an actual sampling procedure. It can be just as valuable to know where the fish are not, as to know where they are. We note the potential usefulness of the simulation procedure we have developed to investigate various sampling methods, but a thorough investigation is beyond the scope of this paper. Therefore, we chose a simplified scheme, in which all non-detection is attributed to a failure on the part of the hypothetical receivers, and not on a *lack* of receivers. We built in the capacity for including search schemes both for the sake of the reader, and for the sake of a future study.

1.2.2 Estimation Model

We treat the data generated above as though it were data collected in the field. We use it as input to the estimation procedure detailed in the sections below. The purpose of the procedure is to determine whether the parameters used to generate the data are estimable. Given that the model below is very similar in structure to the data generating process, we are assuming implicitly that this process is known. This is deliberate. We are investigating whether the parameters are estimable, *given* the simulation is an appropriate representation of diadromous movement and habitat use and, most critically, given a relatively poor detection process. If we can show that, even under these circumstances, the model performs well, then we can treat it as a hypothesis for real data sets.

We use Bayesian estimation in order to capture the hierarchical nature of the data generating process. We write the models below in **JAGS** (Plummer, 2017). Note that there are two separate models, one with movement given as a Standard Markov, and one with movement given as an HMM. We present the two estimation models in their entirety in Figure 2.2, including the joint posterior, which Bayes' Theorem states is proportional to the product of the likelihoods and the priors. With the exception of movement, both models are identical. We detail each piece separately, as above.

Movement

Standard Markov: We use the Dirichlet distribution as a prior for each row of the transition matrix. Given that each row of this matrix is a vector of probabilities that sum to 1, the Dirichlet is a natural choice. The gamma distribution is the conjugate prior for the weights of a Dirichlet distribution. Thus, we estimate all N^2 elements with the following priors

$$w_{i,j} \sim \text{Gamma}(1, 1)$$

$$p_{i,1:N} \sim \text{Dirichlet}(w_{i,1:N} + 0.1)$$

where i gives the state from which an individual is transitioning and j gives the state to which an individual is transitioning. Note that we add an offset of 0.1 to the Dirichlet distribution. Weights near zero can cause the calculation of the density of the Dirichlet to generate non-finite numbers. A small offset typically prevents long model runs from ending in painful error messages. We find that any number below 0.1 may still occasionally produce non-finite numbers. Significantly larger numbers may artificially decrease the variance of the estimates. The vectors $p_{i,1:N}$ stack to form the transition matrix P .

$$Pr(\text{sub}, s, D | \theta) \propto Pr(\text{sub} | s, \theta) Pr(s | \theta) Pr(D | \theta) P(\theta)$$

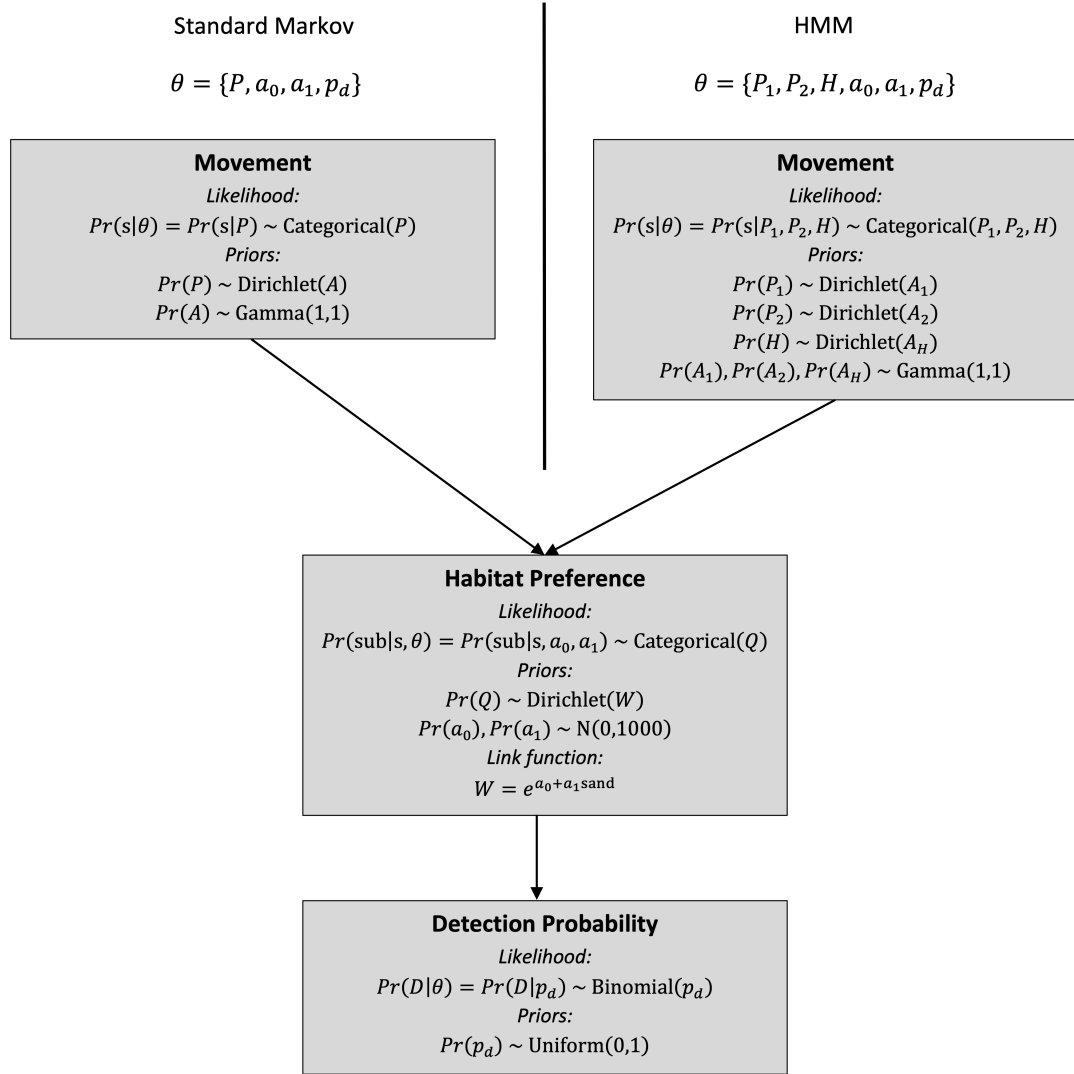


Figure 1.2: A flow chart describing the structure of the estimation model. The full joint posterior is given at the head, in which sub are the $K \times T$ matrices of sub-state locations, s are the $K \times T$ matrices of spatial state locations, D are the $K \times T$ matrices indicating whether an individual k is detected on day t . The set of parameters θ depends on the model. We define the model hyperparameters within the text; we use simple capital letters here for the sake of brevity and cleanliness.

We use the following likelihoods to train the transition matrix

$$s_{k,t} \sim \text{Cat}(P[s_{k,t-1},])$$

where $s_{k,t}$ is given as above for each individual k and for all time steps $t \neq 0$. We use Cat – the categorical distribution – in **JAGS** in place of the multinomial distribution, as it is generally more efficient.

HMM: As in the Standard Markov variation on the Markov process, we use the Dirichlet distribution with gamma priors placed upon the weights for the estimation procedure, the difference being that we now have three matrices to estimate. The priors for the elements of the upriver and downriver transition matrices are given by

$$w_{h,i,j} \sim \text{Gamma}(1, 1)$$

$$p_{h,i,1:N} \sim \text{Dirichlet}(w_{h,i,1:N} + 0.1)$$

where the index $h \in \{1, 2\}$ gives the hidden state and i and j still give states from and to which an individual is transitioning, respectively. Similarly, the priors for the elements of the 2×2 hidden transition matrix are given by

$$w_{i,j} \sim \text{Gamma}(1, 1)$$

$$h_{i,1:2} \sim \text{Dirichlet}(w_{i,1:2} + 0.1)$$

Then we can write the likelihood

$$s_{k,t} \sim \text{Cat}(P_{r_{k,t}}[s_{k,t-1},])$$

where $r_{k,t} \in 1, 2$ is the hidden state of individual k at time t , and is given by

$$r_{k,t} \sim \text{Cat}(H[r_{k,t-1},])$$

The matrices P_1 and P_2 are the transition matrices. We assume arbitrarily that all fish begin in movement pattern 1, and thus their movement is guided by transition matrix P_1 . The number of hidden states is also somewhat arbitrary, but is driven by a specific hypothesis of diadromous movement. In this case, we believe movement is driven by two separate processes that we call upriver and downriver movement. We could easily estimate a third movement process, and attempt to separate the probabilistic structure driving spawning behavior. However, for relative simplicity and clarity, we restrict ourselves to two. We provide the model with no additional information regarding the structure of the matrices. We simply tell our model that we believe there are two separate matrices for movement, and ask it to calculate both.

Habitat Preference

We assume all habitat variables – i.e. proportion "sand" – within each of the sub-states are observed without error. We use the exponential habitat preference function above to generate weights within the model using these data, as follows

$$\ln(w_{n,l}) = a_0 + a_1 \text{ sand}_{n,l}$$

We use the following priors for the slope and intercept

$$a_0 \sim N(0, 1000)$$

$$a_1 \sim N(0, 1000)$$

The set of all draws from a_0 and a_1 serve as hyperparameters for the Dirichlet priors on the probabilities of occupying the sub-states, as follows

$$q_{n,1:L_n} \sim \text{Dirichlet}(w_{n,1:L_n} + 0.1)$$

Recall that the vectors $q_{n,1:L_n}$ give the probability of occupying each of the L_n sub-states in state n . Recall also that, in our case, the number of sub-states is $L_n = 3$ for all spatial states $n \in 1, \dots, N$, $N = 10$. We invite the reader to change the number of sub-states as they desire in the supplementary code S.1. The likelihood is given by

$$\text{sub}_{k,t} \sim \text{Cat}(q_{s_{k,t},1:L_n})$$

where, again, $\text{sub}_{k,t}$ is the sub-state location of fish k on day t . Recall that $s_{k,t}$ is the state location of fish k on day t . Given that the model is nested, the indexing can be complicated. However, we note that sub-state occupancy is simply a matter of dividing up a probability of 1 amongst all sub-states. This structure serves as a hypothesis, first that there is a spatial preference among diadromous fishes as they migrate upriver or downriver, and that habitat type is the mechanism driving this preference. That is, fish choose a location based on the habitat it offers. We assume implicitly that all 200 hypothetical individuals in our study exhibit the same preference, regardless of time or location. In reality, the choice may differ among species, among ages, and among sex. The habitat preference piece of the model can easily be modified to estimate separate parameters indexed by these or any other auxiliary variables.

Detection Probability

We can generate a $K \times T$ matrix of 1's and 0's that indicate whether individual $k \in 1, \dots, K$, $K = 100$ is detected on day $t \in 1, \dots, T$, $T = 100$. Call this matrix

$$D = \begin{bmatrix} d_{11} & \dots & d_{1T} \\ \vdots & \ddots & \vdots \\ d_{K1} & \dots & d_{KT} \end{bmatrix}$$

where $d_{kt} = 1$ if we detected individual k on day t and where $d_{kt} = 0$ if we did not.

To estimate detection probability within the model, we use the following prior

$$p_d \sim \text{Uniform}(0, 1)$$

and the following likelihood

$$d_{k,t} \sim \text{Binomial}(1, p \text{ search}_{k,t})$$

where

$$\text{search}_{i,j} = \begin{cases} 0, & s_{k,t} \notin \text{range}_t \\ 1, & s_{k,t} \in \text{range}_t \end{cases}$$

where, again, $s_{k,t}$ is the spatial state occupied by individual k on day t . We use $\text{search}_{k,t}$ to denote whether fish k is or is not available for detection on day t , and assign a value of 1 or 0, respectively. This is dependent upon the area of the river that we search on day t , which we call range_t . Note the dependence of $\text{search}_{k,t}$ on $s_{k,t}$, which we also estimate. Since the estimation of location contains error, in addition to the error inherent in estimation, p_d is doubly burdened by error. However, if we do not include whether or not a fish is available for detection on a given day, we necessarily bias

detection low, and may give an inaccurate summary of the efficiency of our sampling procedure.

Search

The search process is simply the method by which we choose range_t , and thus determines whether or not a fish is available for detection on each day t . The structure of the search process is entirely a matter of choice, and thus serves only as data input.

1.2.3 Implementation

As noted above, we simulate two separate data sets, one with movement arising from a Standard Markov process, and one with movement arising from an HMM. We use a total of $K = 100$ individuals on $T = 100$ days for both variations on the Markov process, for a total of 200 individuals. We set separate random seeds for both data sets for replicability. Simulation in R is pseudo-random, and setting the seed determines the starting point in the random tables. Using our chosen seed will replicate our data exactly. We note here that we tried 30 separate random seeds, all with similar results. Around half of those seeds resulted in model runs with a high degree of autocorrelation in the MCMC sampling procedure. Standard methods of reducing autocorrelation in MCMC - including longer burn-in, higher thinning rate, and longer chains - are typically sufficient for convergence, if computationally time-consuming.

Recall that we simulated two separate data sets, one with movement arising from a standard Markov process, and one with movement arising from an HMM. The Standard Markov estimation model is certainly more parsimonious than the HMM estimation model, having just N^2 movement parameters. The HMM model, by contrast, has $2 \cdot N^2 + 2 \cdot 2$ parameters, accounting for the two movement transition matrices, as well

as the 2×2 hidden transition matrix guiding the choice between upriver and downriver movement. While the data generation process for the HMM variation on the Markov process is more complicated, we want to test whether a standard Markov process is just as capable of describing movement data generated by an HMM. Therefore, we run both variations of the model on both data sets, for a total of four model runs. We compare the Deviance Information Criteria (DIC) of the Standard Markov model output with that of the HMM model output for both data sets. We also compare the ability of each model to estimate the known spatial states of each individual by calculating the root mean square error (RMSE) of each individual's estimated location on each day. This is given as

$$\text{RMSE}_{s_{k,t}} = \sqrt{\sum_{i=1}^M \frac{(\hat{s}_{k,t,i} - s_{k,t})^2}{M}}$$

where $\hat{s}_{k,t,i}$ is the estimated location of individual k on day t on the i^{th} MCMC sample. We obtain a total RMSE for each individual as follows

$$\text{RMSE}_{s_k} = \sum_{i=1}^T \text{RMSE}_{s_{k,i}}$$

That is, we sum over all $T = 100$ days to obtain a statistical summary of the estimates for each individual's entire trajectory. We hypothesize that the output of the HMM variation on the estimation model applied to the data generated by the HMM variation on the movement process will have a lower DIC and a lower RMSE_{s_k} for all k individuals compared to the Standard Markov estimation model output. Similarly, we hypothesize that the output of the Standard Markov variation on the estimation model applied to the data generated by the standard Markov variation on the movement process will have a lower DIC and a lower RMSE_{s_k} for all k individuals compared to the HMM

estimation model output. This, in a mathematical display of like dissolves like.

Recall that the habitat and detection portions are identical throughout both variations on the model; only the movement portion differs. We present the model run details in Table 4.1. All parameters in all four model runs converge, according to the Gelman-Rubin diagnostic - i.e. $\hat{R} < 1.1$ (Gelman and Rubin, 1992). We choose to track the estimated locations of each fish on each day for both model runs, that is, the values for $s_{k,t}$, which can add considerable time and computational burden to the estimation procedure. However, it can be difficult to replicate the posterior distribution of the location of each fish on each day by repeated sampling. We find that using the posterior of the estimated transition matrices to estimate the posterior of each fish's trajectory overestimated error, often by orders of magnitude. Therefore, we chose to track the values of $s_{k,t}$ within the MCMC procedure. These values typically take many tens of thousands more iterations to converge than any of the other parameters. Visual inspection of the chains of non-converged values of $s_{k,t}$ reveal that this arises from brief forays into states far from the median of the posterior, where the median is always at either extreme of the river. This is especially prevalent when using the model with the HMM variation on the movement process, where very infrequent switches to a separate behavioral state, upriver or downriver, result in just a few iterations straying far from the median. Given that the Gelman-Rubin diagnostic is essentially a measure of the potential information gain of additional samples, we assert that it may not be an appropriate descriptor of the convergence of all the values of $s_{k,t}$. Even so, we chose to run the model until all estimates of $s_{k,t}$ converged. Readers with larger data sets may face a much greater computational burden than we face with our simplified data sets. While the consequences of *not* running the model until all $s_{k,t}$ converge may be elevated uncertainty - as brief but large deviations from the median of the posterior

Table 1.1: Run details for all models and all variations. Variation refers the variation on the movement process, either Standard Markov or HMM. Data refers to the data set, defined in the text. Time is in minutes. We ran all chains in parallel on separate cores.

Variation	Data	Chains	Iterations	Burn-in	Time	DIC
Standard Markov	D-SM	3	15,000	5,000	22.1	25117.0
	D-HMM	3	80,000	30,000	174.2	25082.4
HMM	D-SM	3	90,000	30,000	116.2	21029.6
	D-HMM	3	90,000	30,000	190.5	20366.0

gain more importance with less samples - this may be acceptable in the face of the additional hours or days necessary to achieve convergence.

1.3 Results

For clarity and brevity, we note here that we refer to the data simulated from the Standard Markov variation of the movement model as D-SM, and that we refer to the data simulated from the HMM variation of the movement model as D-HMM. The values we use to simulate the data are all given above, and in the figures cited below. We first present the results of the movement portion of the models. We discuss both models applied to the data set D-SM, and then discuss both models applied to data set D-HMM. Then, given that the habitat preference piece and the detection probability piece are identical throughout both models, we discuss the results of these pieces from all four models together.

1.3.1 Movement

D-SM

The Standard Markov movement model performs remarkably well when applied to data set D-SM. Despite the high percentage of missing values - about 70% in data set D-SM given a p_d of 0.3 - the model is able to capture the entire structure of the probability transition matrix (Figure 1.3). All but one of the non-zero entries in the actual probability transition matrix are contained in the 95% highest posterior density intervals of the estimated probability transition matrix. There are several zeroes in the off diagonals of the actual transition matrix. The median estimates for all of these values are less than 0.01, the majority of them one or two orders of magnitude lower.

The HMM model applied to data set D-SM is able to determine that only one movement process is at play. In essence, it switches off the downriver estimation procedure by setting the upriver to downriver transition probability - the element h_{12} in the matrix H given above - to a number close to zero. In effect, then, no fish's movement is ever guided by the downriver transition matrix P_2 . The posterior distributions of each element of P_2 matches the prior, indicating no connection to the data, while those of the upriver probability transition matrix P_1 are nearly identical to those estimated by the Standard Markov model applied to data set D-SM (Figure 1.4).

We rank the estimated trajectories for all $K = 100$ hypothetical fish by $RMSE_{s_k}$, and present the actual and estimated trajectories with the lowest, middle, and highest $RMSE_{s_k}$ for both the Standard Markov model applied to data set D-SM (Figure 1.5) and the HMM model applied to data set D-SM (Figure 1.6). It is clear that frequent observations distributed homogeneously through time produce the best and most reliable estimates of the individual trajectories. Thus, while the characterization

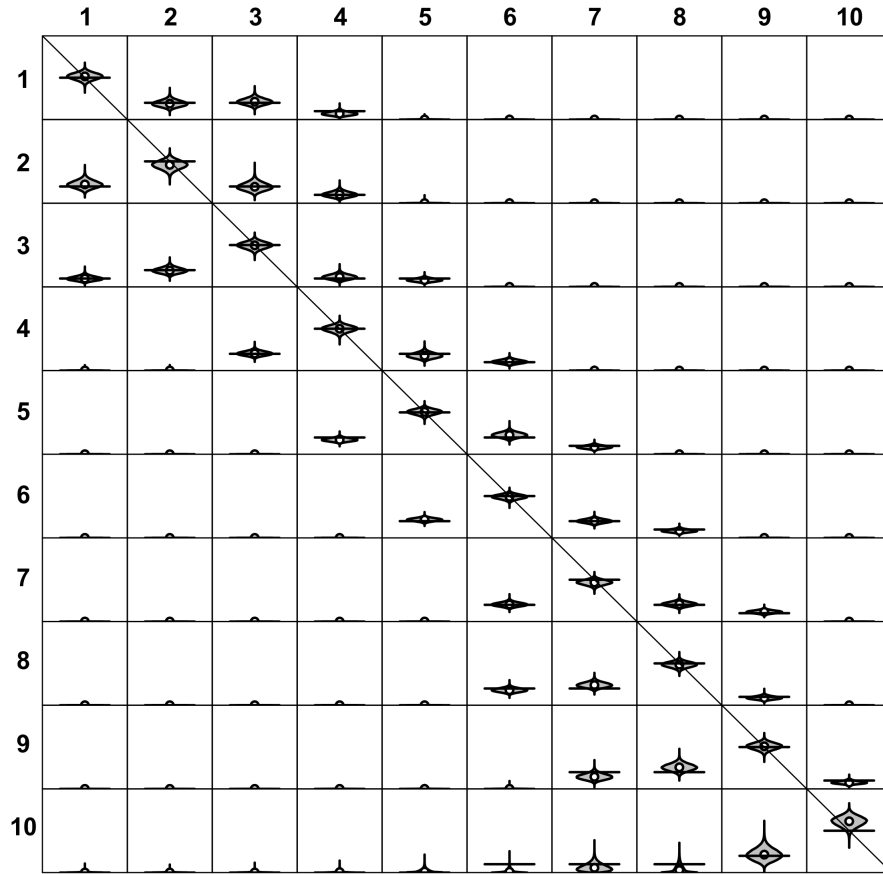


Figure 1.3: Violin plots of the posterior densities of transition probabilities from model with Standard Markov movement applied to data set D-SM. The y-axis of all 100 individual plots ranges from 0 to 1. The labels to the left of the plot give the state from which an individual is transitioning, and the labels above the plot give the state to which an individual is transitioning. Open circles give the median, black horizontal bars give the actual value used to simulate the data. The solid diagonal line is drawn along the main diagonal of the transition matrix. Values above represent upriver movement, and values below represent downriver movement.

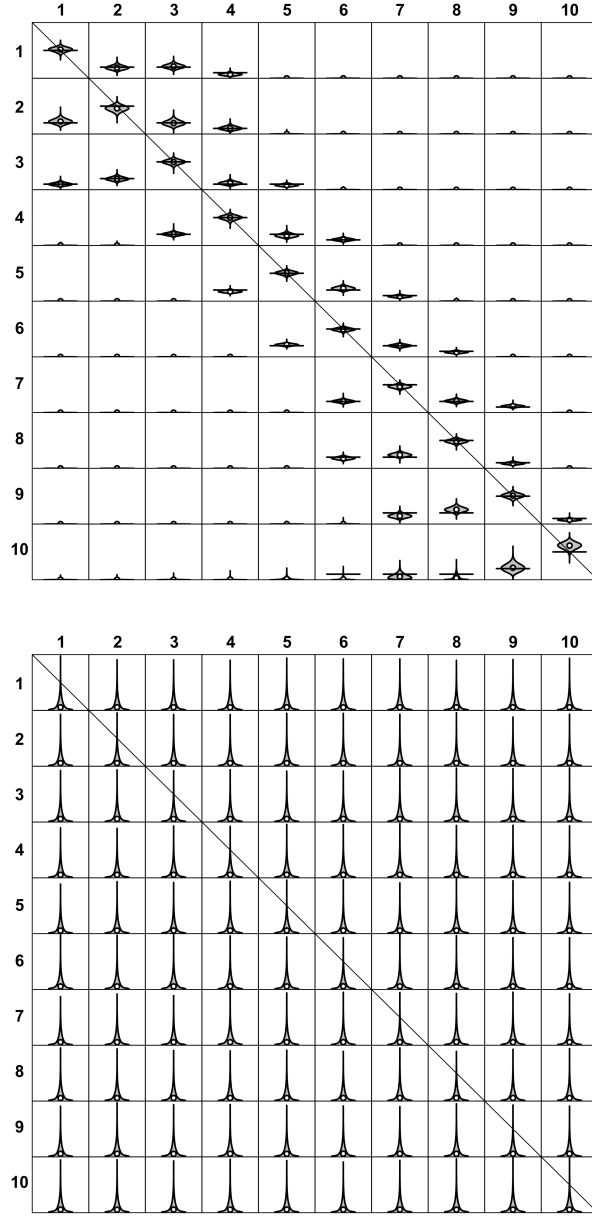


Figure 1.4: Violin plots of the posterior densities of transition probabilities from the model with HMM movement applied to data set D-SM. Top: upriver transition matrix. Bottom: downriver transition matrix. The y-axis of all 100 individual plots ranges from 0 to 1. The vertical labels to the left of the plots give the state from which an individual is transitioning, and the horizontal labels above the plots give the state to which an individual is transitioning. Open circles give the median, black horizontal bars give the actual value used to simulate the data. The violins themselves are rotated kernel density estimates. The solid diagonal line is drawn along the main diagonal of the transition matrix. Values above represent upriver movement, and values below represent downriver movement.

of the process that governs movement may be excellent, individual trajectories may be difficult to estimate if there are long stretches of time - in this case, on the order of about 20 days or more - with no observations. Counter to our hypothesis, the $RMSE_{s_k}$ values of the estimates from the standard Markov model applied to data set D-SM are not consistently lower than those from the HMM model applied to data set D-SM. Rather, they are nearly equivalent, with the $RMSE_{s_k}$ of the estimates from the standard Markov model applied to data set D-SM slightly lower than those of the HMM model applied to data set D-SM for 57 individuals, and the reverse case for the other 43 (Figure 1.7). The DIC of the Standard Markov variation of the model applied to data set D-SM is greater than that of the HMM variation of the model applied to data set D-HMM. In this case, DIC is clearly not a reliable measure of model parsimony. Given the HMM model rendered 100 parameters obsolete, we would be wise to remove them, despite the DIC. While neither $RMSE_{s_k}$ nor DIC are definitively in favor of the Standard Markov model over the HMM model, it clearly has a built-in mechanism to determine whether the additional computational burdens induced by the HMM model are necessary.

D-HMM

The standard Markov movement model applied to data set D-HMM produces a curious hybrid of the actual upriver and downriver probability transition matrices (Figure 1.8). Estimated transition probabilities from lower spatial states – that is, downriver states 1, 2, 3, etc. – are generally biased towards the actual downriver values, while those from higher spatial states, or upriver states, are generally biased towards the actual upriver values. The tension between the actual upriver and downriver values is unsurprising given the Standard Markov model operates under the assumption that there is only one process guiding movement. As above, we use the full posterior distri-

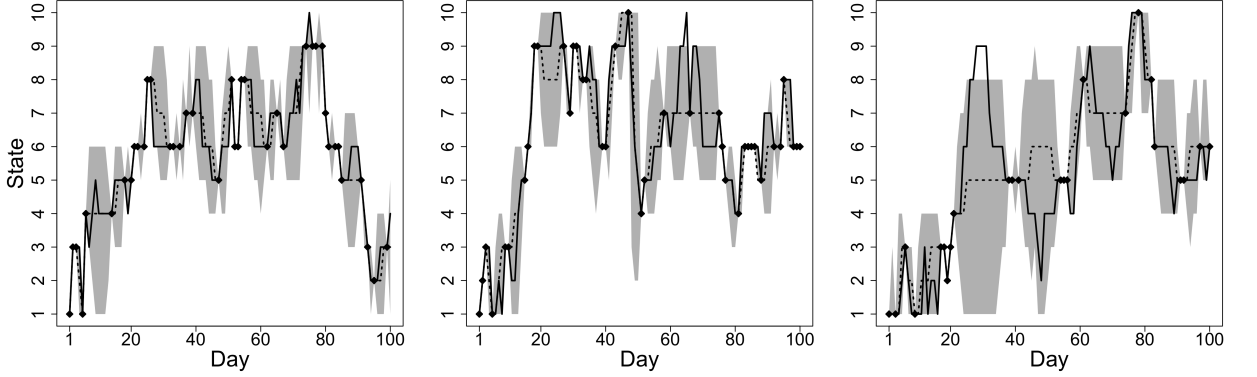


Figure 1.5: Actual and estimated trajectories for data set D-SM with the standard Markov variation of the movement model. From left to right: trajectories with the lowest, middle (i.e. 50th), and highest $RMSE_{s_k}$. Solid black lines give the actual trajectory. Dotted lines give the mean estimated trajectory. Shaded regions are 95% highest posterior density intervals. Diamonds show locations of observations i.e. values that remained after we implemented the detection probability and search processes.

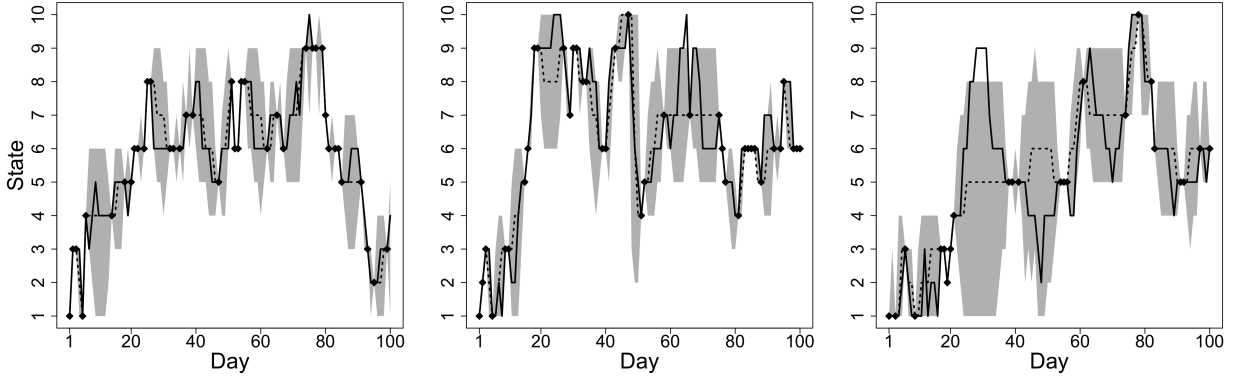


Figure 1.6: Actual and estimated trajectories for data set D-SM with the HMM variation of the movement model. From left to right: trajectories with the lowest, middle (i.e. 50th), and highest $RMSE_{s_k}$. Solid black lines give the actual trajectory. Dotted lines give the mean estimated trajectory. Shaded regions are 95% highest posterior density intervals. Diamonds show locations of observations i.e. values that remained after we implemented the detection probability and search processes.

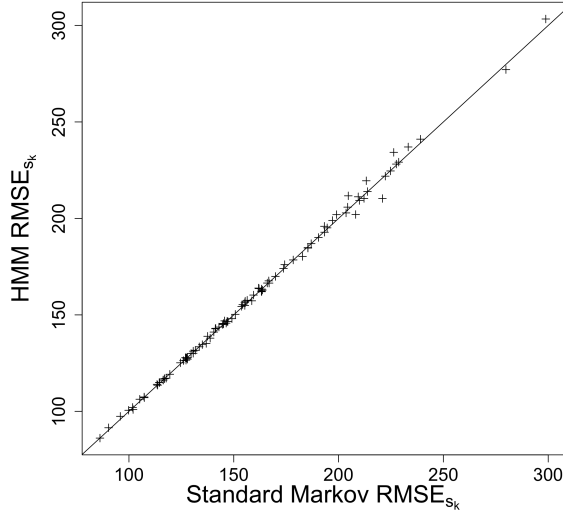


Figure 1.7: The RMSE_{s_k} of the Standard Markov model versus that of the HMM model applied to data set D-SM. The diagonal line is a 1 : 1 line.

bution of this transition matrix to build the posterior of the 100 simulated trajectories in data set D-HMM. However, given the sheer size of the array produced by the model, we were unable to save the model files in their entirety on a standard laptop computer. We were forced to calculate the mean, median, and 95% highest posterior density intervals of the individual trajectory estimates. We use the following equation, which we call root square error, or RSE, simply for differentiating from RMSE.

$$\text{RSE}_{s_k} = \sqrt{\sum_{i=1}^T \frac{(\hat{s}_{k,i} - s_{k,t})^2}{T}}$$

where here $\hat{s}_{k,i}$ is the median estimate of the state location for individual k on day i . We rank based on the RSE_{s_k} (Figure 1.9).

We make similar inferences from these results as from those given in Figures 1.5 and 1.6. Namely that more, and more homogeneous, observations give better and less uncertain estimates of the trajectories. We do note that the consequences of miss-

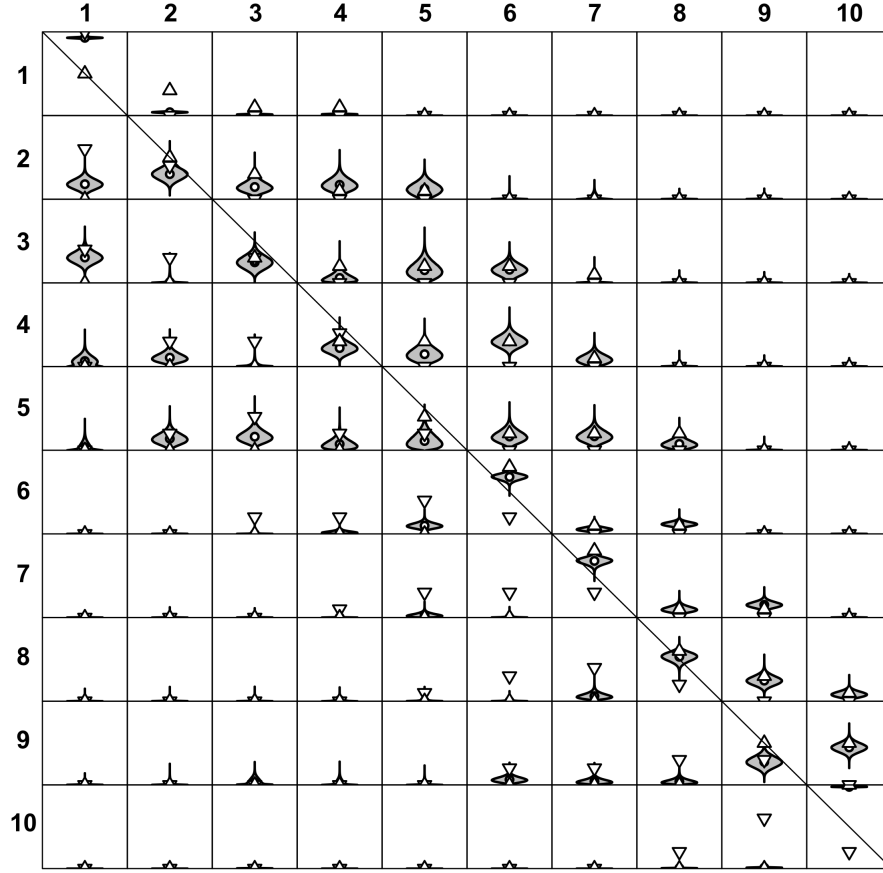


Figure 1.8: Violin plots of the posterior densities of transition probabilities from the model with Standard Markov movement applied to data set D-HMM. The y-axis of all 100 individual plots ranges from 0 to 1. The vertical labels to the left of the plot give the state from which an individual is transitioning, and the horizontal labels above the plot give the state to which an individual is transitioning. Open circles give the median, open triangles give actual upriver probability transition matrix values, and open inverted triangles give actual downriver probability transition matrix values. The violins themselves are rotated kernel density estimates. The solid diagonal line is drawn along the main diagonal of the transition matrix. Values above represent upriver movement, and values below represent downriver movement.

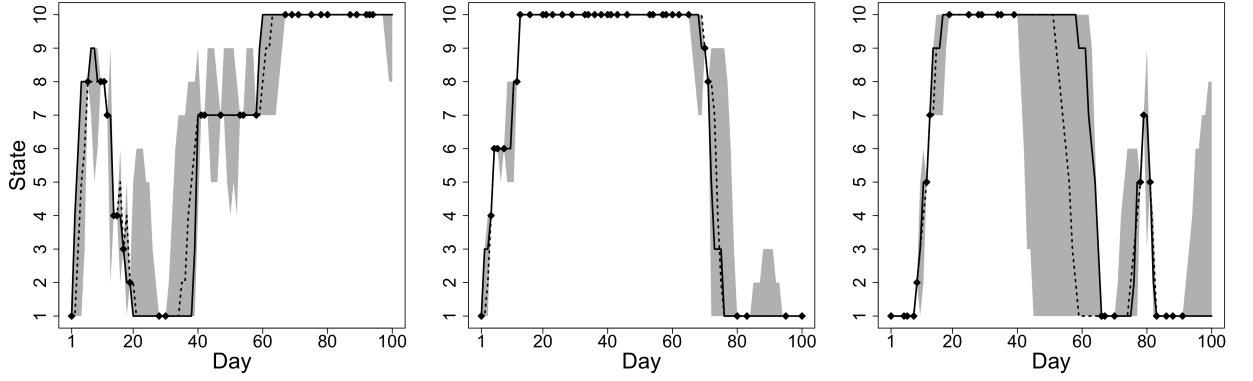


Figure 1.9: Actual and estimated trajectories for data set D-HMM with the Standard Markov variation of the movement model. From left to right: trajectories with the lowest, middle (i.e. 50th), and highest RSE_{s_k} . Solid black lines give the actual trajectory. Dotted lines give the mean estimated trajectory. Shaded regions are 95% credible intervals. Diamonds show locations of observations i.e. values that remained after we implemented the detection probability and search processes.

ing observations for these estimates are more severe. By this we mean that just a few consecutive days of non-observation produces extremely wide 95% highest posterior density intervals. This is clearly a result of the relatively high uncertainty in the estimates of the transition probabilities shown in Figure 1.8.

The HMM model performs well when applied to data set D-HMM. It is able to distinguish between upriver and downriver movement, and to capture the structure of both the upriver and downriver transition matrices (Figure 1.10). The model is also able to capture the structure of the hidden transition matrix (Figure 1.11). However, some biases are clearly present. For data set D-HMM, the model is generally able to estimate the upriver transition probabilities with greater precision and accuracy than those of the downriver transition matrix. After attempting many simulations - upwards of 200 throughout the development and testing of this model - we find that this is generally a function of the efficiency of the detection and observation processes. About 70% of the individual daily locations are unobserved in data set D-HMM. Higher detection

probabilities produce estimates of both transition matrices with lower uncertainty. We also find that this is a function of the structure of the hidden transition matrix. We choose very high diagonal values for this matrix - 0.98. That is, individuals in a given behavioral state tend to remain in that state. Given both this persistence, and that we started all 100 individuals in upriver movement, each individual generally spends more time moving upriver than downriver. After trying several random seeds, we find that, on average, all 100 fish over 100 days spend roughly 60% of their time moving upriver. In data set D-HMM, the fish spend about 56% of their time in an upriver behavioral state. Longer observation periods produce more even results, with about 50% spent in both upriver and downriver movement behavior. For real world applications, this is simply a caution against truncating the observation program before the fish have completed their out- or in-river migration, depending on the nature of their diadromy.

As above, we use the transition matrices to build the posterior distribution of the estimated trajectories, and rank based on RSE_{s_k} (Figure 1.12). The mean estimated trajectories are fairly reasonable. However, during periods of non-observation, the uncertainty quickly grows to encompass most, if not all, of the river. This uncertainty is a function of the relatively broad 95% highest posterior density intervals for the transition probabilities. It is also a function of the hierarchical nature of the process. There is error in both the determination of the behavioral state, and of movement. This propagation makes frequent observations that are homogeneous in time even more imperative to the estimation procedure. Note that, when we apply the Standard Markov model to data set D-HMM, there is somewhat greater uncertainty in the estimation of daily trajectories, as summarized by RSE_{s_k} (Figure 1.13). There is a cluster of points below the 1 : 1 line in Figure 1.13, indicating that, in most cases, RSE_{s_k} is lower for the HMM model applied to data set D-HMM than for the Standard Markov model

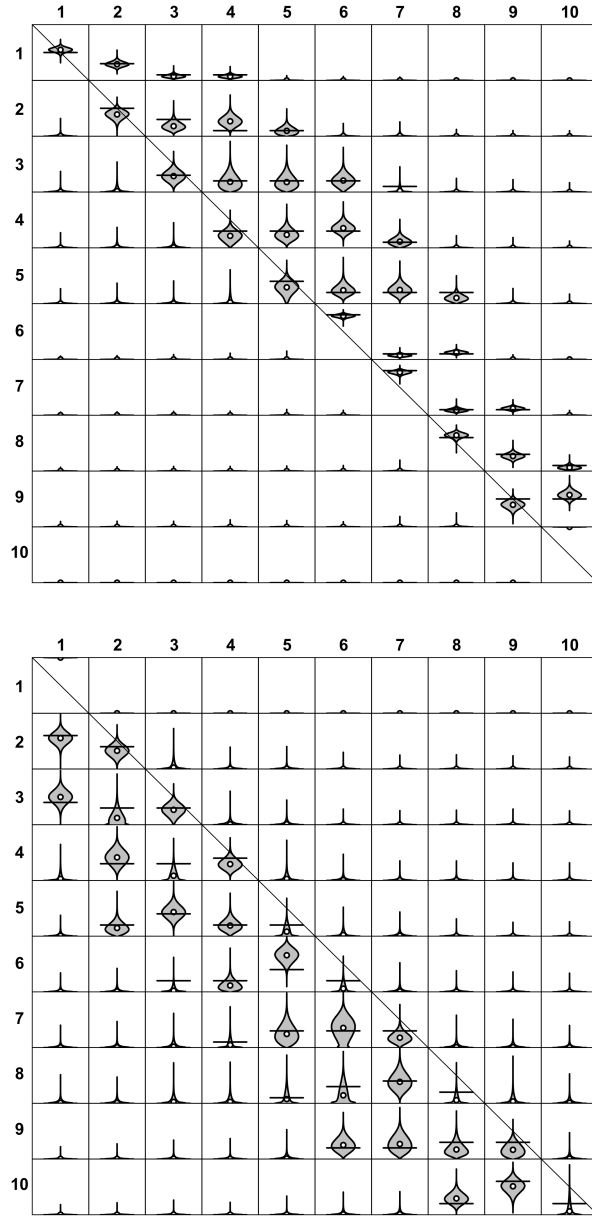


Figure 1.10: Violin plots of the posterior densities of transition probabilities from the model with HMM movement applied to data set D-HMM. Top: upriver transition matrix. Bottom: downriver transition matrix. The y-axis of all 200 individual plots ranges from 0 to 1. The labels to the left of the plots give the state from which an individual is transitioning, and the labels above the plots give the state to which an individual is transitioning. Open circles give the median, black horizontal bars give the actual value used to simulate the data. The solid diagonal line is drawn along the main diagonal of the transition matrix. Values above represent upriver movement, and values below represent downriver movement.

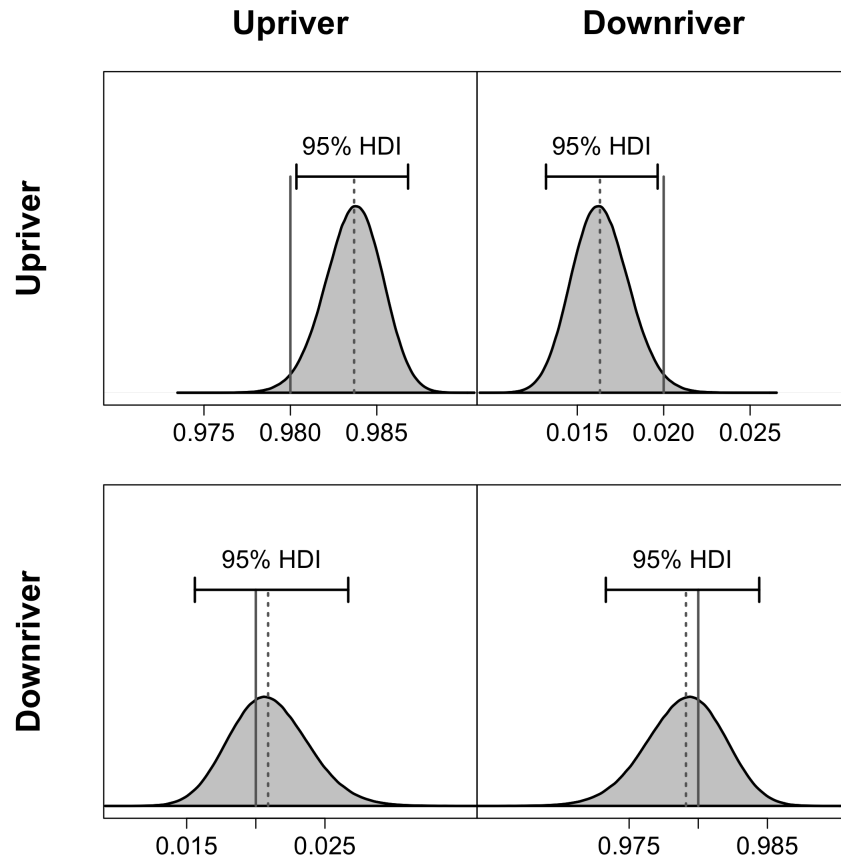


Figure 1.11: Posterior distributions of the elements of the hidden transition matrix. Floating horizontal brackets give the 95% highest posterior density interval. Vertical gray bars give the actual values from which we simulated data set D-HMM. The labels along the left of the plots give the state from which an individual is transitioning at any given time step. The labels along the top give the state to which an individual is transitioning. Note that we started all individuals in the upriver movement state. Values above represent upriver movement, and values below represent downriver movement.

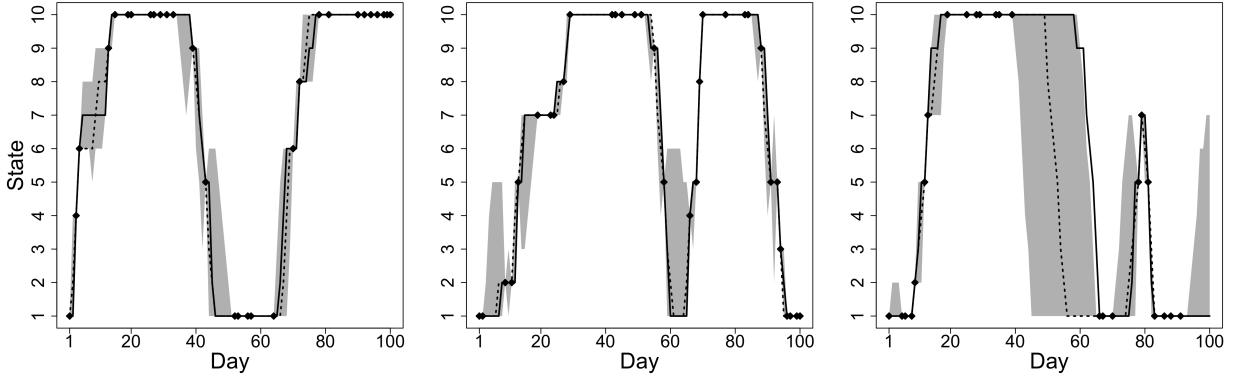


Figure 1.12: Actual and estimated trajectories for data set D-HMM with the HMM variation of the movement model. From left to right: trajectories with the lowest, middle (i.e. 50th), and highest $RMSE_{s_k}$. Solid black lines give the actual trajectory. Dotted lines give the mean estimated trajectory. Shaded regions are 95% credible intervals. Diamonds show locations of observations i.e. values that remained after we implemented the detection probability and search processes.

applied to data set D-HMM.

1.3.2 Habitat Preference

Recall that habitat preference is simulated as a simple logistic function of the proportion of so-called "sand" substrate in a given sub-state. We denote a_0 as the intercept of this function, and a_1 as the slope. We have two estimates of these values for both data sets D-SM and D-HMM, one from each variation of the movement model, and thus four in total (Figure 1.14). While the model for habitat preference is identical in both variations, the posterior densities of the coefficients are not quite identical through all four applications. All four model runs estimate that, roughly, $a_0 \cong -a_1$. Median estimates are similar throughout, with $a_0 \cong -a_1 \cong -3.5$. To compare, we use both the actual slope and intercept values of $a_0 = -a_1 = -2$ and the approximate estimated slope and intercept to generate 10,000 samples over 11 values ranging from 0 to 1 of proportion sand substrate from the Dirichlet distribution, and summarize the results in Figure 1.15. Note that, below values at or very close to a proportion sand

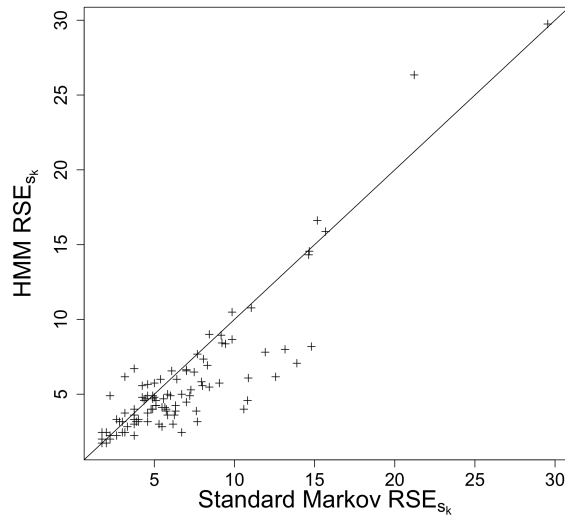


Figure 1.13: The RSE_{s_k} of the Standard Markov model versus that of the HMM model applied to data set D-HMM. The diagonal line is a 1 : 1 line.

equal to 1, the resulting median occupancy probabilities are similar amongst both the actual and approximate estimated coefficients. At lower values of proportion sand, the model underestimates the median to a lesser degree as the proportion sand falls to 0, and underestimates uncertainty to a greater degree. The effect, then, may be to predict a higher or lower degree of fidelity to certain habitat types than the fish in fact exhibit. In this case, the model believes that, as proportion sand in a given sub-state approaches 0, there is a near-zero probability that the fish will occupy that sub-state. In reality, while low, the probability may be appreciable enough that, on rare occasions, the fish may in fact occupy that sub-state despite its unfavorable habitat. From this perspective, it is not surprising that the model underestimates uncertainty in unfavorable sub-states, as with imperfect detection (recall $p_d = 0.3$), we reduce the probability that we will observe rare occurrences which is, by definition, small to begin with.

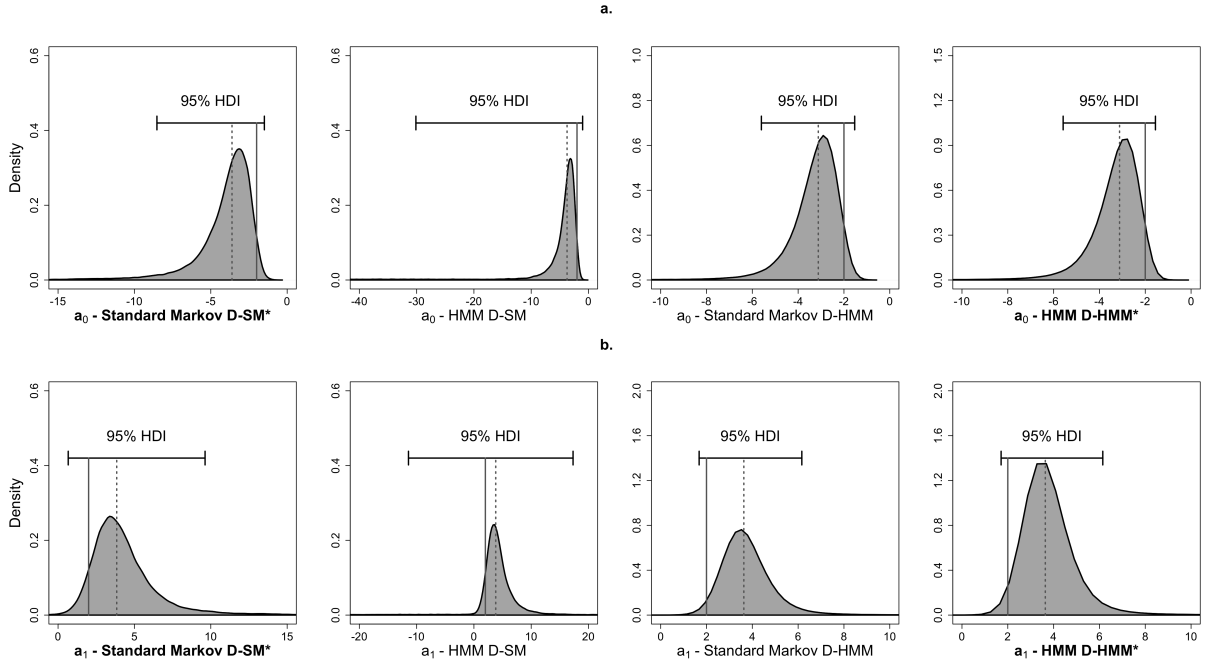


Figure 1.14: Posterior density of: **a.** the intercept (a_0) and **b.** the slope (a_1) of the habitat preference model. Floating horizontal brackets give the 95% highest posterior density intervals. Solid vertical lines give the actual values from which we simulated each respective data set. Broken vertical lines give the median of the posterior density. Abscissa labels give the model applied (either Standard Markov or HMM) and the data set to which it is applied (D-SM or D-HMM). Bold and starred abscissa labels indicate that the model applied is equivalent to the simulation process used to generate the data set.

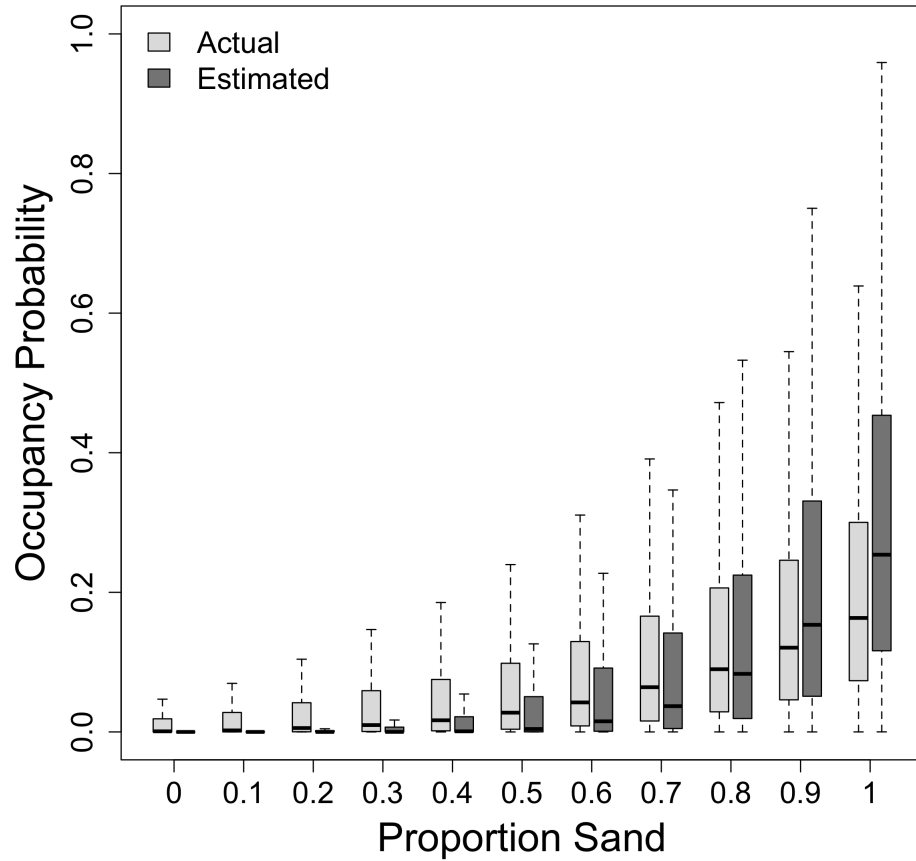


Figure 1.15: Box and whisker plots summarizing 10,000 samples from the Dirichlet distribution over 11 values of proportion sand substrate using weights generated by the equation $\ln(w) = a_0 + a_1 \text{ sand}$. The actual values are $a_0 = -a_1 = -2$. The estimated values are approximate at $a_0 = -a_1 = -3.5$. The plots all exclude the outliers.

Relative to those values we use to simulate the data, the performance of the model is somewhat weak. Actual values are consistently over- or under-estimated, and highest posterior density intervals are broad. However, empirically, the actual and median estimated occupancy probabilities are nearly identical (not shown), suggesting that the model is performing well empirically, but failing to properly identify the underlying process. Improving detection probability would certainly alleviate some of this error. However, in cases where this is not practical, one might at least attempt to achieve a high contrast in available habitats covered by the sampling scheme.

1.3.3 Detection and Observation

The detection and observation processes used to simulate the data used in this paper are very simple, and so estimates of detection probability are straightforward. The minor discrepancies observed in the median estimates from the actual value are due to the randomness introduced during the simulation procedure, and not to any systematic issues with the model. This is why *both* models applied to data set D-SM, and both models applied to data set D-HMM, under- and over-estimate the detection probability, respectively (Figure 1.16).

1.4 Discussion

Diadromous behavior can be difficult to describe, mathematically. Staging and spawning may produce a high degree of non-stationarity in fishes' day to day movements that complicate both summary and prediction. However, for purposes of conservation and restoration, it is often critical that we have at least a broad understanding of both timing, and how diadromous fishes use different parts of the river as they migrate up- or down-river to spawn. Traditional models of habitat use are inadequate, as the fishes

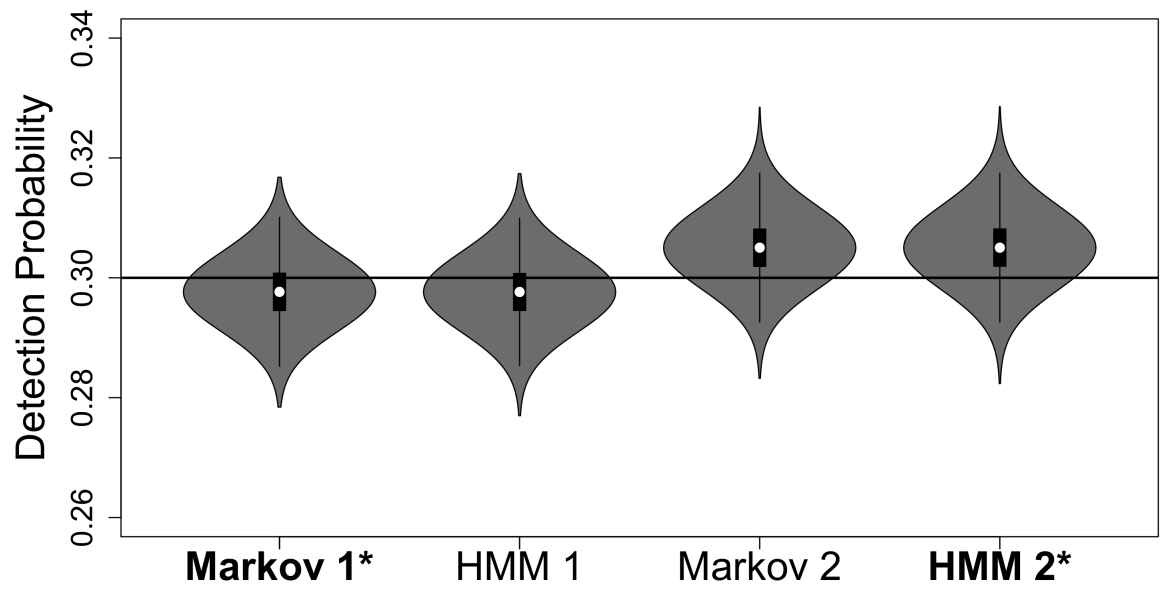


Figure 1.16: Posterior densities of detection probability for all four model runs. Bold and starred abscissa labels indicate that the model applied is equivalent to the simulation process used to generate the data set. The solid horizontal line is the actual detection probability used to simulate all data sets.

are, by definition, on the move, and both their surroundings and their needs change throughout their spawning run. One must explicitly model both the movement and the habitat selection process on scales appropriate to the species in question. The novel model structure presented here allows for as much complexity as is necessary, or as is allowed by the resolution of the data. After examining the output, we recommend applying the Hidden Markov model, or HMM, to data on iteroparous diadromous fish movement and occupancy. Though it may introduce more complexity than is perhaps necessary, the model is clearly capable of distinguishing between populations that do and do not exhibit evidence of extensive staging, highly directional movement, or other behaviors that might produce different patterns during upriver and downriver migration. If the results indicate that the second transition matrix has been switched off, one may then run the simpler Standard Markov model, which takes considerably less time (see Table 4.1). Note that Table 4.1 shows that the DIC of the HMM model is lower in both cases when applied to either data set D-SM or D-HMM. However, given the additional time and computational power required to estimate a second transition matrix which is effectively switched off during the estimation procedure, Standard Markov is clearly more appropriate for data set D-SM, as expected. Thus, in this case, DIC may not be the most appropriate measure by which to select a model. Rather, we suggest allowing the HMM model to perform selection on itself. If it believes two separate classes of movement are unnecessary, it will behave as though it were only asked to calculate one.

The primary limitation of the model structures described above is that they are an empirical description of the movement process, and give no mechanistic explanation for the drivers of movement within the river, or persistence within a given state. Whereas we describe movement and habitat selection as separate, nested processes, the impetus

for occupying a given state may in fact be identical to the impetus for moving to that state to begin with. This limits the ability of the model to predict the trajectory of a given fish without several observations along the way. However, given the fidelity of many diadromous fishes to their natal staging and spawning grounds, good estimates of an empirical description of movement may be adequate to summarize overall movement throughout the river. From the estimated transition matrices, we can give estimates of the median path. That is, we can simulate several - in this case, 10,000 - trajectories, and calculate the median position on each day, along with the 2.5% and 97.5% quantiles. If we think of data sets D-SM and D-HMM as distinct populations, then Figure 1.17 (left) suggests that we focus management efforts mid-river for D-SM and upriver for D-HMM. For D-SM, we might also note that elevated human activity at either extreme of the river would have minimal effect on the population once migration is well underway, provided the activity does not have one or more indirect effects mid-river. For D-HMM, the spread of fish across the river at any given time is slightly more broad, and may in fact be present in any part of the river at any given time. Even so, elevated activity near the mouth of the river may have a minimal effect on the population as a whole, if most are concentrated upriver. Note, however, that if we think of these 10,000 trajectories as predictive medians and intervals for one individual beginning in state 1, we only have a vague inclination as to where they might be at a later date. Whether or not this is adequate depends on our objectives.

There are clear biases in the estimation of occupancy as driven by habitat characteristics. We have hypothesized that this is due in part to the lack of observation of rare occurrences in non-sandy habitats. This may also be due to the structure of the equations themselves. The slope and intercept of the occupancy function are correlated with one another by nature, and thus multiple values may produce the same

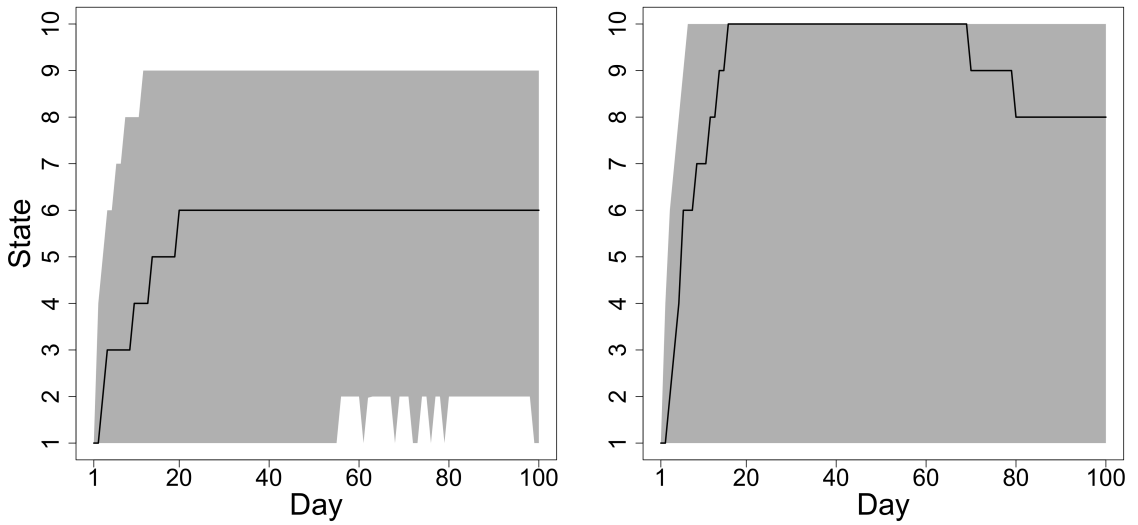


Figure 1.17: Median trajectories calculated from 10,000 trajectories simulated from: (left) the estimated median transition matrix of the Variation 1 model applied to data set D1 and (right) the estimated median transition matrices of the Variation 2 model applied to data set D2. The solid line is the median trajectory. The shaded regions represent the 2.5% and 97.5% quantiles of the simulated trajectories.

general outcome. We find that, while estimated values are offset from actual values (see Figure 1.14), the model is able to estimate the actual occupancy probabilities very well (not shown). They are typically accurate to within a few thousandths of the probability used to simulate the data. However, despite the accuracy, there is some degree of imprecision evident in the posterior distributions. This may, again, represent the model's strength as a descriptor of *average* behavior, and not of any individual's behavior. While averaging complex behavior may mask systematic variability among individuals, we believe that, in the case of animals with strong site fidelity, averaging can be interpreted as a strength. This, of course, is dependent upon our objectives. We wish to characterize general habitat use in order to guide use and conservation of an entire river system, and therefore find this structure to appropriate.

The conclusions drawn above demonstrate the utility of the model, given of course that our assumptions about diadromous fishes are correct. However, the model structure does not require very limiting assumptions: spatial and temporal dependence in movement, the existence of habitat preference, and site fidelity, which presents statistically as a common distribution for movement. If, for a given diadromous species, there is good evidence supporting these assumptions – and we find to be typically true – then this model structure may be very useful. In addition to its ability to summarize multiple classes of movement as well as the median path, Figures 1.5 and 1.12 also demonstrate that estimated transition matrices may allow one to predict the location of an observed individual over relatively short time intervals. The model also allows for a separation of spatial scale largely unprecedented in the literature. This consideration of scale is similar to the separation of temporal scale given by Leos-Barajas et al. (2017), which allows for the model to reflect the coarseness of the data source, without data removal. Our structure is similar in that it allows discrete models to be scaled as finely as the data source, while accounting for large-scale movement, all without untenable complexity and model run-times. We have built our model with diadromous fish in mind, but assert that it is appropriate for any group of organisms exhibiting both site fidelity and spatial boundedness during their migration. We present an application of this model to Atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus*) in Part II of this paper, and demonstrate the ability of the model to isolate staging behavior, as well as to identify potential new spawning grounds.

1.5 Supplementary Materials

S1. Basic JAGS model code for Variation 2 model, described above, written in R and assigned to file "simhmm.jags". Remove the hidden transition matrix code, and the

first of the 3 dimensions of the transition matrices, to create the Variation 1 model.

```
cat("model{

# Priors for upriver and downriver transition matrices with nstates states
for(i in 1:nstates) {
  for(j in 1:nstates) {
    aup[i,j] ~ dgamma(1,1)
    adn[i,j] ~ dgamma(1,1)
  }
  trn[1,i,1:nstates] ~ ddirich(aup[i,] + 0.1)
  trn[2,i,1:nstates] ~ ddirich(adn[i,] + 0.1)
}

# Priors for hidden transition matrix with 2 states
for(i in 1:2) {
  for(j in 1:2) {
    ahmm[i,j] ~ dgamma(1,1)
  }
  hmm[i,1:2] ~ ddirich(ahmm[i,1:2] + 0.1)
}

# Estimate states of nfish fish over ndays days from transition matrix
for(i in 1:nfish) {

# All fish begin in hidden state 1 i.e. upriver migration
m[i,1] <- 1

# Hidden state m and observable state s of fish i on day k
for(k in 2:ndays) {
  m[i,k] ~ dcat(hmm[m[i,k-1],])
  s[i,k] ~ dcat(trn[m[i,k],s[i,k-1],])
}
```

```

    }
  }

# Estimate detection probability
for(i in 1:nfish) {
  for(j in 1:ndays) {

    # Check that fish 's location was in any of the search areas
    for(k in 1:length(rg[,j])) {
      srb[i,j,k] <- ifelse(s[i,j] == rg[k,j], 1, 0)
    }

    # This will be >0 if fish was searchable
    srbsum[i,j] <- sum(srb[i,j,])

    # Check that fish was searchable
    search[i,j] <- ifelse(srbsum[i,j] > 0, 1, 0)

    # Binomial likelihood
    det[i,j] ~ dbin(p*search[i,j], 1)
  }
}

# Detection prior
p ~ dunif(0,1)

# Estimate occupancy at the substate scale
for(i in 1:nstates) {

  # Log link for occupancy probability of pcs substates as a function of sand
  for(j in 1:pcs) {

```



```

      log(w[i,j]) <- (a0 + a1*sand[i,j])
    }

    # Estimates of occupancy probability
    pocc[i,1:pcs] ~ ddirich(w[i,1:pcs] + 0.1)
  }

  # Grid scale occupancy of each fish on each day drawn from categorical
  for(i in 1:nfish) {
    for(j in 1:ndays) {
      grid[i,j] ~ dcat(pocc[s[i,j],])
    }
  }

  # Logistic regression priors – occupancy
  a0 ~ dnorm(0, 0.001)
  a1 ~ dnorm(0, 0.001)

}, file = "simhmm.jags")

```

Chapter 2

Migration and habitat use of Hudson River Atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus*)

One of the difficulties of modeling habitat use behaviors of diadromous fish is that they are, by definition, on the move. Ideally, any model structure we choose to describe diadromous migration will include an explicit representation of movement, in addition to a representation of habitat preference. In this paper, we present an application of a hierarchical Bayesian model that accounts for diadromous movement with a hidden Markov process, habitat preference with a logistic model of occupancy probability, as well as imperfect observation with a basic post-hoc summary of detection probability. We apply the model to acoustic detections of Hudson River spawning Atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus*), with separate applications for males and females. We find that males exhibit evidence of staging behavior in areas downriver of historical spawning grounds, whereas females arrive later, and tend to move quickly upriver and downriver with no evidence of staging. We also find that, on average, both males and females exhibit a preference for areas of the river with a maximum depth of $< 10\text{m}$. However, on rarer occurrences, both males and females exhibit a preference for areas with a maximum depth of $> 50\text{m}$. Detection probability of both males and females is roughly 0.27. This model represents a general structure by which we can summarize

acoustic detections of migratory fishes, and a tool that we can use to address and generate hypotheses concerning the migration of the Atlantic sturgeon of the Hudson River.

2.1 Background

The Atlantic sturgeon (*Acipenser oxyrinchus oxyrinchus*) are an ancient species of fish distributed in coastal regions of the Atlantic ocean. They were once an abundant and valuable fishery, yielding both flesh for meat – nicknamed Albany beef (Stegemann, 1994) - and eggs for caviar. The Atlantic sturgeon are distributed along the east coast of Canada and the United States, from Maine to Florida. They are categorized into distinct population segments (DPS) that exhibit evidence of a meta-population structure, migrating to separate natal spawning grounds and migrating back to the continental shelf of the Atlantic. North to south, the distinct population segments are: Gulf of Maine, New York Bight, Chesapeake Bay, Carolina, and South Atlantic. Nation wide, the fishery landings peaked in the United States in 1890, but had collapsed by 1905, owing to enormous catches when fish were at high density as they passed in and out of rivers for their spawning runs (Smith, 1985). After additional fishing pressure throughout the 1900's resulted in precipitous declines in population size, Atlantic sturgeon in the United States were placed on a forty year fishing moratorium beginning in 1998, both within state waters and in the United States' exclusive economic zone (EEZ; ASMFC (1998)). In this paper, we are concerned with the Atlantic sturgeon of the Hudson River. The population of Atlantic sturgeon in the Hudson River belong to the New York Bight distinct population segment, which was listed as endangered in 2012 (National Oceanic and Atmospheric Administration, 2012) after several years of low numbers and only marginal growth. Our objective is to identify an optimal model

structure to identify what the Endangered Species Act (ESA) refers to as critical habitat – which it defines as specific areas with biological or physical features essential to the conservation of the species contained within the geographical area it occupies at the time of listing – for the Atlantic sturgeon of the Hudson River.

Attempting to identify critical habitat for an anadromous species such as the Atlantic sturgeon of the Hudson River generates several unique challenges. Diadromy of any kind exposes fish to a wide number of sources of mortality. When traveling upriver or downriver en masse, fish are forced through relatively narrow channels, and are therefore easy to capture (Limburg and Waldman, 2009). While capture is currently limited to bycatch, fishes may also come into close contact with shipping activity, and be vulnerable to ship strikes and other indirect disturbances. In fact, ship strikes are one of the greatest sources of anthropogenic mortality for Atlantic sturgeon (ASMFC, 2017). Identifying and summarizing the migratory route of these fish in the Hudson River may ultimately help us to reduce these sources of mortality. Atlantic sturgeon undertake large scale migrations, but they have also been observed to have a fairly high degree of persistence, or residency, in areas of the Hudson River over shorter time scales (Dovel and Berggren, 1983). This indicates the need for a model structure that operates at both fine and broad scales. Atlantic sturgeon in both the Hudson River and other systems have shown remarkable year-to-year consistency in their spatial usage of their natal rivers (Dovel and Berggren (1983); Collins et al. (2000), Amanda Higgs, pers. comm.). Therefore, data limited to just one or a few years may be adequate to describe Atlantic sturgeon's staging and spawning grounds. Males mature and begin migrating upriver to spawn at around 11-20 years of age, while females mature at 20-32 years (Taub, 1990; ASMFC, 2017). In the Hudson River, spawning grounds typically lie between river kilometers 56 and 132 (Dovel and Berggren, 1983), although

there is evidence for spawning grounds up to river kilometer 182 (Van Eenennaam et al. (1996); see Figure 2.1). These same authors stated that, while females exhibit deliberate movement in and out of the river, males tend to linger. This is consistent with our own observations, and suggests the need for either a separate model for the sexes, or random effects that account for sex. For movement, then, we require a structure that operates at multiple scales, on the separate sexes, and that may operate on just one or a few years of data.

As the sturgeon recover, it has been noted that the juvenile age classes contribute more to population biomass than all adult classes (Gross et al., 2002). The life of a juvenile begins with a successful spawning run. The eggs of adult sturgeon are adhesive, and adhere to hard substrates. The success of a spawning run thus depends on the ability of adults to migrate to natal regions of the river with appropriate habitat and substrate type, to spawn undisturbed, and to return to the sea, where they reside until the next spawning period. It is clear, then, that we must also identify and manage particular habitats and habitat variables that drive occupancy in a given area at an appropriate scale. Breece et al. (2018) find that temperature, depth, and spatial location all have an effect on Atlantic sturgeon movement and residency in Delaware Bay. Thus, there is a precedent for identifiable effects of habitat type on occupancy and persistence of fishes within a given region of a river system. In addition to a model structure that can describe movement and occupancy at multiple spatial scales, we also wish to choose a structure that can relate occupancy to habitat type. Assuming discrete information about a subset of the adults' locations in both space and time, we have a broad choice of tools with which to summarize and predict movement and habitat use according to the requirements described above. These range from simply day to day summaries of mean location, to sophisticated models describing fish movement and

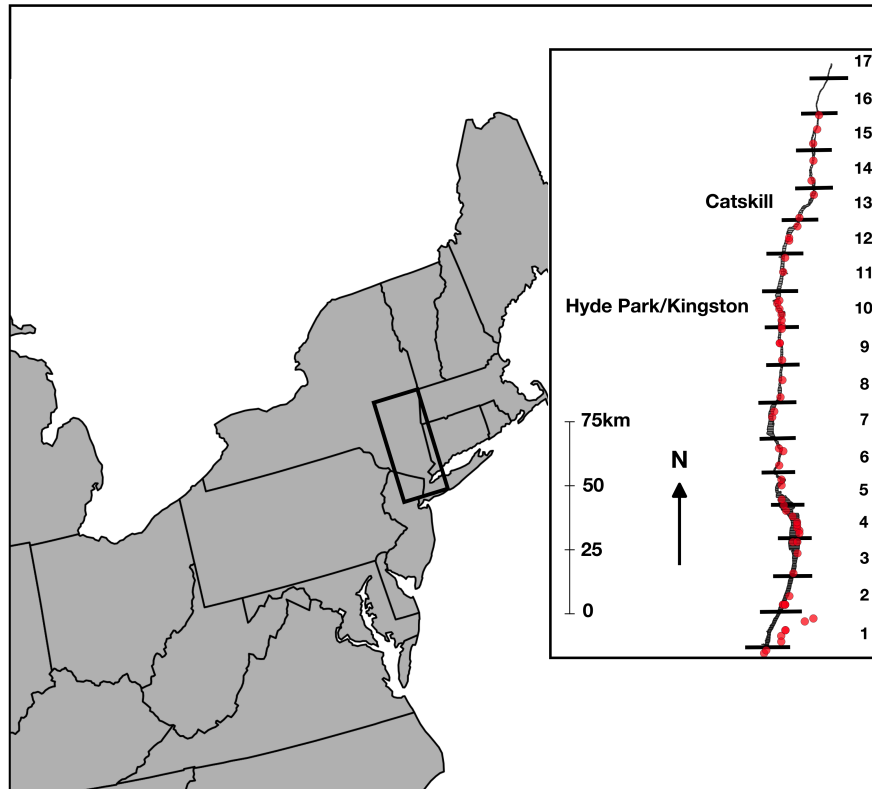


Figure 2.1: A map of the northeastern United States, with the Hudson River as an inset. Red circles show the locations of the acoustic receivers. Each area of the river encased above and below by a solid horizontal bar are treated as separate spatial "states", marked by number. Each of these states is further sub-divided into 15 sub-states, which are marked by river kilometers. The mouth of the river, not pictured, lies at the southern end, and is called state 1. It is given as a state in the model, but is not sub-divided, and is not considered as a part of the habitat selection model. The remaining 16 states progress sequentially upriver, terminating near the Troy Dam, which lies at the upper extreme of this figure.

occupancy.

When choosing a structure to model animal occupancy and habitat use, one must take into account their behavior and their environment, as well as the nature of the data one has collected. In our case, we have acoustic detections of a highly mobile animal along a one-dimensional path. That is, receivers cannot pinpoint individuals within a given river kilometer. We are also constrained by the boundaries of the river, and so must either choose a continuous structure that respects the shoreline as a spatial limit, or a discrete structure that ignores geographical details. Our objective is to not only choose a structure optimal for our data on Atlantic sturgeon in the Hudson River, but also to choose a structure easily generalizable to other diadromous species. We wish to describe not only movement, but also habitat use, detection, and observation. We wish for this model to be only as complex as the data allow, and to give good, interpretable summaries of occupancy and habitat use throughout the observation period. Movement over relatively long time scales may be difficult to model, as the behavioral drivers of movement may change, and therefore some method of describing behavioral switching may be necessary. Migrating iteroparous and anadromous species of fish are a quintessential example of behavioral switching. They travel long distances over a short period of time, and then encamp upon reaching their natal spawning grounds. After spawning, they then travel in the opposite direction. There may be other complex behaviors – including staging, multiple incursions into the river, and movement within the river – that require similarly sophisticated models.

Multiple researchers have used mixtures of multiple random walks and hidden Markov models (HMM) to describe separate behavioral classes of movement (Morales et al. (2004), Roberts et al. (2004), Franke et al. (2006)). These examples classify behaviors as a composite of step length and turning angle. However, data collected via a

one-dimensional array provide information that is by definition linear, and therefore a concept of turning angle is moot. Koster et al. (2018) describe a generalized linear mixed effect model (GLMM) that accounts for variables such as flow regime as drivers of directional (upstream or downstream) movement in a riverine system. We are interested in the Atlantic sturgeons' usage of space, and only in the direction of movement insofar as it affects the amount of time fish spend in a given area of the river. Schönfisch and Kinder (2002) use the concept of cellular automata to divide a river into a series of uninterrupted flows divided by permeable dams. They assign each dam a permeability, or probability of passage, and fish migrate upriver according to these values of permeability, which they infer from counts of Atlantic salmon (*Salmo salar*) in each area. They propose that downriver migration might be modeled by adding a second graph of the river, with a different set of permeabilities, connected perhaps at the estuary. In our case, we might attach a second graph at the Troy Dam, an impermeable barrier at river kilometer 246 of the Hudson River. There are no other barriers below the Troy Dam, and so any other permeabilities would represent the draw of a given region, rather than the ability of the fish to pass from one area to the next. This general structure is more reflective of our needs, but still lacks the ability to identify separate behaviors, beyond a direct comparison of upriver and downriver permeabilities.

While spatially explicit individual or agent based models are powerful tools that have been used to reproduce complex behaviors of migrating riverine fishes (Jager and DeAngelis (2018); Snyder et al. (2019)), we ultimately chose to assume that each individual's movement arises from a common probability distribution. This choice is motivated in part by our desire to interpolate missing observations in our data, and also to give a straightforward probabilistic uncertainty to these interpolations. It is also motivated by the site fidelity exhibited by many sub-populations of Atlantic sturgeon.

Melnychuk et al. (2017) use a spatially explicit capture recapture (SECR) model to summarize transition probabilities among areas within the New York Bight. We chose not to estimate transition probabilities as a function of environmental covariates, the rationale of which is explained in the methods below. That said, the motivation is the same, to provide a statistical summary of migration patterns in an effort to determine how they might overlap with unnatural sources of mortality.

When modeling animal migration, we must choose between four general structures: discrete or continuous time, and discrete or continuous space. We chose discrete time and discrete space. Continuous time models typically use a geometric or exponential distribution to model waiting times between events. In this case, the time an animal spends in a given state before transitioning to the next. Given a low detection probability, this would more accurately describe the waiting time between detections, and it may be difficult to non-observation to non-detection or to a lack of availability for detection. While the methods behind first order HMM are memoryless, we assert that there is indeed a certain kind of memory is assuming that all fish move according to the same statistical distribution. In the case of anadromous fish, we might consider this assumption akin to a shared memory of their natal river waters. Models of habitat use often compare available habitat to habitat used, and consider any disproportion between choice and availability to be evidence of positive or negative selection (Neu et al., 1974). This general technique has, in fact, been used to quantify habitat use of Atlantic sturgeon (Dunton et al., 2010). We chose, instead, to incorporate habitat variables directly into probability of occupancy with a log-link function. This simple addition allows for a straightforward probabilistic interpretation of habitat preference.

Our broader objectives were to develop a single model that could describe all stages of diadromous migration. In this paper, our objectives with respect to the Atlantic

sturgeon of the Hudson River were to interpolate the missing data points, as well as to determine whether males and females exhibit different movement behaviors. We apply the model described in Duskey et al. (2020) and interpret the output with respect to known and unknown attributes of Atlantic sturgeon habitat preference and migration. This will ultimately help us to identify critical habitat of the Atlantic sturgeon within the Hudson River.

2.2 Methods

2.2.1 Acoustic Tagging

We collected adult Atlantic sturgeon in the ocean off the coast of Delaware and implanted acoustic tags (VEMCO V-16, 6-H, 69.0 kHz; $\approx$ 6.5 year estimated battery life) in those fish suitable for the procedure (i.e. $> 1.5\text{m FL}$). We deployed an array of acoustic receivers (VEMCO Ltd. VR2W) within the Hudson River (Figure 2.1). We deployed most of these in April and May, and retrieved them in late October. Exact methods are described in Kazyak et al. (2020). We receive data as a tag number associated with a given receiver at a given time. We associate each receiver's detection range with just one river kilometer, given that adverse weather, dense boat traffic, and geometric tortuosity of the river can all severely limit detection range. We limit data for analysis to that which includes at least three observations at a receiver no more than twenty minutes apart. Of the nearly five million individual observations, this quality assurance process removes just 1%. For those individuals observed in multiple river kilometers on a single day, we record the very last observation. We translate these data into a detection history for each individual i on day j as s_{ij} in a matrix. The entries s_{ij} are listed as "states", which are defined visually in Figure 2.1. Each state comprises 15 river kilometers. This choice is somewhat arbitrary, and represents a balance between

the resolution needed to represent movement at a fine enough scale for management, and the resolution needed to estimate parameters at a large enough scale to reduce the number of parameters in the model. It also represents the minimal scale at which we can reliably place fish given a detection at a given receiver, which improves the reliability of occupancy as a function of habitat variables. Given that the fish were tagged in marine waters, we are unconcerned with the phenomenon of fallback (Frank et al., 2009) due to stress of the tagging procedure affecting our results. We consider only those fish that migrated up the Hudson River in the year 2015 in this study. Of these fish, 33 are male, and 12 are female, with mean total (fork) length of 1,914mm (1,691mm) and 2,231mm (1,982mm) (Table 2.1). The distribution of total length of both males and females adhere very well to assumptions of normality. Biases in sex ratio of tagged individuals may result from differences in migration timing between the sexes, and not from actual differences in the population (Dovel and Berggren, 1983). Only 6 of the fish considered in this study were tagged in 2015, all of which are male. None of these appear to exhibit any differences in migration timing or pattern compared to the other 27 males.

2.2.2 Environmental Data

We supplement our model with three environmental data sources: Hudson River bottom substrate type, sediment environment, bottom morphology, and depth. Substrate type reflects grain size, while sediment environment reflects the behavior of the bottom currents. All of these data were collected by, or in response to, the Hudson River Estuary Program or the Hudson River Estuary Program Action Plan. The Hudson River bottom substrate data were based on grain size analysis of sediment cores and sediment grabs, and supplemented by sidescan sonar (Bell et al., 2006). Analysis reveals

Table 2.1: Biological information of all 45 fish considered in this study. Mean total (fork) length for males (mm) = 1914 (1691). Mean total (fork) length for females (mm) = 2231 (1982).

Capture Date	Total Length (mm)	Fork Length (mm)	Sex
4/5/2009	1790.00	1560.00	male
4/10/2009	2030.00	1860.00	male
4/13/2009	1770.00	1600.00	male
4/26/2009	2100.00	1880.00	male
4/2/2010	1860.00	1680.00	female
4/15/2010	1620.00	1380.00	male
4/15/2010	1910.00	1650.00	male
4/15/2010	1900.00	1720.00	male
4/16/2010	2070.00	1850.00	male
4/20/2010	2550.00	2300.00	female
4/21/2010	1800.00	1610.00	male
4/3/2011	1920.00	1700.00	male
4/3/2011	1900.00	1670.00	male
4/18/2011	2300.00	2080.00	female
4/19/2011	2370.00	2110.00	female
4/19/2011	1750.00	1590.00	male
4/20/2011	1980.00	1770.00	male
4/21/2011	1950.00	1740.00	female
4/26/2011	2130.00	1860.00	female
4/27/2011	2040.00	1790.00	female
5/5/2011	2040.00	1780.00	male
3/29/2012	1700.00	1520.00	male
4/3/2012	1750.00	1520.00	male
4/3/2012	1830.00	1650.00	male
4/4/2012	2040.00	1760.00	male
4/4/2012	1930.00	1700.00	male
4/4/2012	1890.00	1680.00	male
4/5/2012	1800.00	1640.00	male
4/14/2012	2250.00	2020.00	female
4/12/2012	2110.00	1900.00	male
4/18/2012	2010.00	1710.00	male
4/20/2012	2030.00	1740.00	female
4/20/2012	2610.00	2360.00	female
4/10/2013	1990.00	1740.00	male
4/26/2013	2160.00	1890.00	female
4/28/2014	2520.00	2210.00	female
5/2/2014	2160.00	1900.00	male
5/5/2014	2060.00	1800.00	male
5/5/2014	1900.00	1680.00	male
4/14/2015	1980.00	1720.00	male
4/18/2015	1870.00	1610.00	male
4/21/2015	1820.00	1620.00	male
4/23/2015	1920.00	1630.00	male
4/26/2015	1900.00	1720.00	male
4/26/2015	1930.00	1680.00	male

there are nine sediment types: all possible two-way combinations of mud, sand, and gravel. Details of collection and type definition are given in Bell (2008b). We simplify these data by pooling all sediment types dominated by mud, sand, or gravel, resulting in just three sediment types. While this masks complexity that may play a role in habitat selection, the parsimony ultimately improves the performance of the model. Sediment environment is described either as erosional, depositional, or dynamic, based on acoustic and bathymetric data of the bottom substrate (Bell et al., 2006). The defining attributes of each of these environment types is described in Bell (2008a). Morphology data are interpreted from the bathymetry of the Hudson River with a horizontal accuracy of $\pm 1\text{m}$ (Bell et al., 2006). The depth and slope of the Hudson River's bottom substrate are placed into one of the following categories: channel floor, channel wall, channel terrace, pinnacle, marginal flat, marginal bank, and bar. The defining attributes of each of these morphology types is described in Hudson River National Estuarine Research Reserve (2008). Depth data are collected as 1m bathymetric contours with a Simrad EM 3000 system referenced to NAVD 1988 (Marine Sciences Research Center, State University of New York at Stony Brook, 2007). Details of collection and processing are described in Flood (2007). In addition to bathymetry and attributes of the sediment, we include the area in square meters of each subsection of the river as a covariate, standardized by the largest area observed to reduce the magnitude of the estimated coefficients. We recognize that a detection in a given area is not necessarily an indication that a fish is making significant use of the surrounding habitat. To attempt to circumvent this, we also include a variable we call "index", which gives the sequential river kilometer within each state, ordered from downriver to upriver. Given that the system is linear, fish must pass through all areas below any given river kilometer to reach it. For this reason, we record each fish's *last* observed location on a given day, and include "index" in the habitat selection process. While

there is still room for spurious correlation, we believe we have minimized it by taking the actions described above. We summarize all environmental data at the scale of one river kilometer. Substrate type, sediment environment, and bottom morphology are all given as proportions of a given river kilometer. Depth is given both as mean depth, and as maximum depth. We calculate maximum depth as the mean of the maximum values in each possible composite habitat type. There are thus a total of 17 environmental covariates that we consider in this model. We process all spatial data in QGIS (QGIS Development Team, 2009).

2.2.3 The Model

We use three nested model structures to describe (1) movement, (2) occupancy, and (3) detection of Atlantic sturgeon in the Hudson River, which we summarize in Figure 2.2. We describe each of these three pieces below, and include the structure of both their priors and likelihood.

Movement

We use a Hidden Markov model, hereafter referred to as HMM, to summarize movement. A hidden Markov process with M hidden states consists of an initial distribution $m = [q_1, \dots, q_M]$ and an $M \times M$ probability transition matrix that we call H . The initial distribution m gives the probability of being in any of the M states at time step $t = 1$. The probability transition matrix H gives the probability h_{ij} of transitioning from state i to state j . Each of the M states are described by an observable stochastic process. In our case, the observable outcomes are the locations of each fish at a given time $t \in 1, \dots, T$. We have chosen to model movement as discrete in both space and time, and thus each of M hidden states are themselves Markov processes with $N \times N$ probability transition matrices that we call P_m . We define the N states as separate

$$Pr(\text{sub}, s, D | \theta) \propto Pr(\text{sub} | s, \theta) Pr(s | \theta) Pr(D | \theta) P(\theta)$$

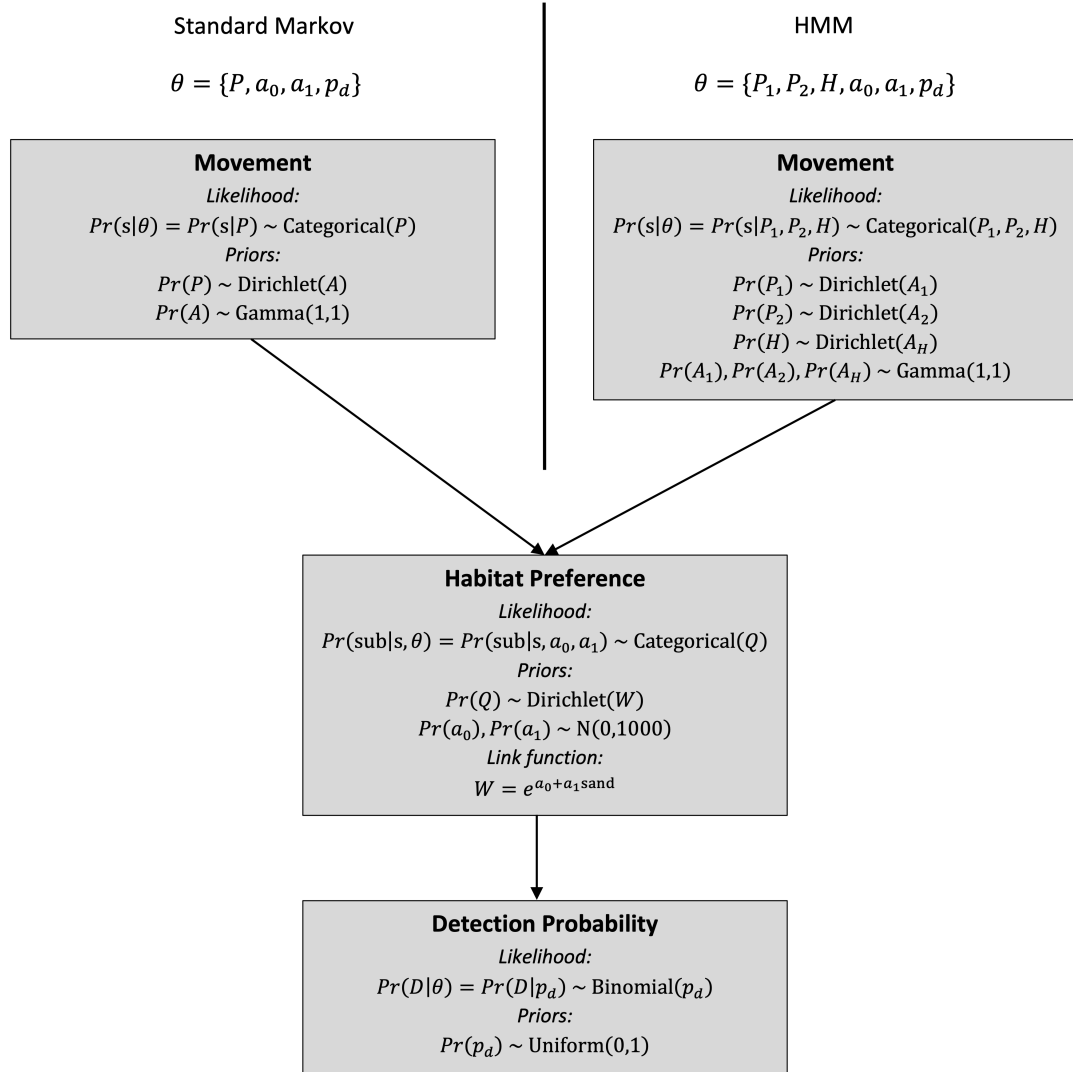


Figure 2.2: A flow chart describing the structure of the estimation model. The full joint posterior is given at the head, in which sub are the $K \times T$ matrices of sub-state locations, s are the $K \times T$ matrices of spatial state locations, D are the $K \times T$ matrices indicating whether an individual k is detected on day t , and where θ is the set of all model parameters. We define the model hyperparameters within the text; we use simple capital letters here for the sake of brevity and cleanliness.

spatial regions in the river (Figure 2.1). Movement occurs at the scale of 15 river km discrete spatial sections, of which there are 17. The matrix P_m gives the probability $p_{m,ij}$ of moving from spatial region i to spatial region j . Therefore, hidden states are separate classes of movement behavior, requiring matrices $[P_1, \dots, P_M]$ to describe M different patterns of movement. Hypothetically, we could estimate as many hidden states, and thus patterns of movement, as we like. The number that we choose reflects our hypothesis about the structure of migration within the river. We have chosen to estimate just two hidden states, and hypothesize that these behaviors will generally describe (1) upriver migration and staging and (2) downriver migration and exit to marine waters. Thus, the hidden matrix is of the form

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$

where, without loss of generality, we call hidden state 1 upriver movement and hidden state 2 downriver movement. Then, for example, element h_{12} gives the probability of transitioning from upriver to downriver movement behavior. We use a Dirichlet prior for both rows of the hidden transition matrix, as the input is an ordered vector of, in this case, 2 positive real weights, and the output is then an ordered vector of 2 real numbers in the interval $(0, 1)$ that sum to 1. We write this as follows

$$w_{i,j} \sim \text{Gamma}(1, 1)$$

$$p_{i,1:2} \sim \text{Dirichlet}(w_{i,1:2} + 0.1)$$

where $i, j \in 1, 2$ give the hidden state i from which an animal is transitioning and the hidden state j to which an animal is transitioning in a given time step. We use gamma priors for the weight hyperparameters. We hypothesize that the probabilistic

weight will be concentrated at the main diagonal, that is fish will tend to persist in a given hidden state or movement pattern. The observable states are the 17 15 river km sections. The probability transition matrices describing movement are then

$$P_1 = \begin{bmatrix} p_{1,1,1} & \cdots & p_{1,1,17} \\ \vdots & \ddots & \vdots \\ p_{1,17,1} & \cdots & p_{1,17,17} \end{bmatrix}$$

$$P_2 = \begin{bmatrix} p_{2,1,1} & \cdots & p_{2,1,17} \\ \vdots & \ddots & \vdots \\ p_{2,17,1} & \cdots & p_{2,17,17} \end{bmatrix}$$

where element $p_{h,i,j}$ gives the probability of transitioning to state j from state i given one of two hidden states h . Therefore, during the estimation procedure, we assume each fish's positional history arises from common probability and hidden transition matrices. We use Dirichlet priors for these matrices as well, which we write as

$$w_{h,i,j} \sim \text{Gamma}(1, 1)$$

$$p_{h,i,1:17} \sim \text{Dirichlet}(w_{h,i,1:17} + 0.1)$$

where each element $p_{h,i,j}$ is as defined above.

We assume that each fish begins in hidden state $m = 1$, reflecting our belief that each fish must first travel upriver before it can travel downriver. We only estimate location for each individual on days falling *after* the first day of observation. Then we can write the likelihoods as

$$s_{k,t} \sim \text{Categorical}(P_{r_{k,t}}[s_{k,t-1},])$$

where $s_{k,t}$ is the spatial state location of individual k on day $t \in \xi_k, \dots, T$. We call the first day of observation for individual k ξ_k . The variable $r_{k,t}$ is the hidden state i.e. movement behavior of individual, given by

$$r_{k,t} \sim \text{Categorical}(H[r_{k,t-1},])$$

Recall that $r_{k,\xi_k} = 1$ for all individuals.

In summary, the model first determines whether fish are in hidden state 1 or 2 – upriver or downriver movement behavior – and then moves the fish within the river according to probability transition matrix P_1 or P_2 , respectively. We restrict the maximum daily movement of Atlantic sturgeon by forcing the priors for the probability transition matrices to concentrate within 6 states (90 river kilometers) of any given state. We do not force probabilities outside of this range to be exactly zero, allowing for rare long-range movements. We find in Duskey et al. (2020) that this model structure is capable of determining whether more than one class of movement behavior, or hidden state, is necessary. The hidden transition matrix can effectively switch off the second, or downriver, matrix by setting the probability h_{12} of transitioning from upriver to downriver to a number very close to zero. We find that this occurs when movement data are generated by a single transition matrix that produces more or less symmetric movement in and out of the river. That is, when there is no noticeable staging behavior. When transition probabilities are concentrated roughly equally around the main diagonal, this can produce movement upriver and downriver in the short term. In the long term, a single matrix will tend to produce a stable distribution of fish throughout the river. That is, it is more so an empirical description of movement than a comprehensive summary.

Habitat Preference

Occupancy, or habitat selection, occurs in this model at the scale of 1 river km. After a fish moves to one of 17 observable spatial states according to the behavior dictated by its hidden state, it must then choose amongst 15 sub-states, which are defined as all river kilometers within a given observable state, ordered from downriver to upriver. This choice is guided by a simple logistic function that relates the 17 environmental covariates described above to the probability of occupying any one of the 15 sub-states. If we organize the covariate data in a three-dimensional array X with elements x_{ijk} giving the value of covariate i for sub-state j of state k , then we can calculate weights for occupying sub-state j of state k as

$$\ln(w_{jk}) = I(\boldsymbol{\alpha} X_{1:18,j,k})^T$$

$$\boldsymbol{\alpha} = [a_0, a_1, \dots, a_{17}]$$

$$I = [1, i_1, \dots, i_{17}]$$

where I is a vector of binomial variables that serve as numerical on/off switches for the influence of each covariate on the weights w_{jk} . The mean of the posterior distribution of these binomial variables can be treated as an inclusion probability, that is the probability that each covariate lends non-zero influence to the weights. We consider here a posterior inclusion probability ≥ 0.75 to be sufficient evidence that a given covariate is a significant predictor of the weights, and therefore of occupancy. The vectors $w_{1:15,k}$ each contain 15 positive real numbers. We then use these vectors as parameters in the Dirichlet distribution to produce vectors with 15 elements in the interval $(0, 1)$, as follows

$$q_{1:15,k} \sim \text{Dirichlet}(w_{1:15,k})$$

Then, we can write the likelihood

$$\text{sub}_{k,t} \sim \text{Categorical}(q_{1:15,k})$$

where $\text{sub}_{k,t}$ is the sub-state location of individual k on day $t \in \xi_k, \dots, T$. The elements within each of these vectors sum to 1, thereby producing probabilities of occupying a given sub-state within a given observable state.

Since acoustic receiver coverage within the Hudson River is both non-uniform and incomplete (see Figure 2.1), there is a concerning possibility that habitat selection within a given state will reflect accidental bias in receiver placement with respect to habitat type, rather than any actual preference on the part of the Atlantic sturgeon. We run a generalized linear model on receiver presence or absence as a response to the habitat variables in the river kilometer in which the receivers appear. We find that only the pinnacle habitat type – isolated peaks rising at least 4 meters from the surrounding channel floor – is predictive of receiver presence (Table 2.2). However, this is not systematic bias in receiver placement, but rather is a reflection of the extreme rarity of the pinnacle bottom morphology type. According to the Hudson River morphology metadata (Hudson River National Estuarine Research Reserve, 2008) the pinnacle morphology represents isolated peaks elevated 4 meters above the surrounding substrate in just two regions: Tappan Zee (river kilometer 44) and southern Haverstraw Bay (river kilometer 57). Given that acoustic receivers happen to appear in these regions, receiver presence then appears to be highly correlated with the pinnacle morphology type. Though we still include this variable in the model, we interpret the results concerning pinnacles with care.

Table 2.2: Generalized linear model (GLM) with all habitat types as the predictors, and presence or absence of an acoustic receiver as the response. Values for substrate type (mud, sand, and gravel) and values for bottom morphology (channel floor, channel wall, channel terrace, pinnacle, marginal flat, marginal bank, and bar) each sum to 1 and are therefore collinear. Thus, coefficients were only calculated for all but one of each. Boldface and star indicates a significant p-value with $\alpha = 0.05$.

	Coefficient	σ^2	P-value
Intercept	-1.93	2.65	0.47
Mean Depth	0.09	0.13	0.47
Max Depth	0.04	0.06	0.49
Index	0.01	0.03	0.66
Mud	-1.13	1.92	0.56
Sand	0.22	1.66	0.89
Gravel	—	—	—
Erosion	-0.64	1.13	0.57
Deposition	-1.29	0.97	0.19
Channel Floor	-1.01	2.33	0.66
Channel Wall	0.64	2.32	0.78
Channel Terrace	2.18	2.54	0.39
Pinnacle	100.46	47.72	0.04*
Marginal Flat	2.50	3.28	0.45
Marginal Bank	-2.22	5.07	0.66
Bar	—	—	—
Standardized Area	0.30	0.67	0.66

In summary, after a fish moves on a 15 river km scale, it then chooses amongst each of the 15 river km based on the available habitat, which we define in relation to depth and substrate. The associated parameters represent *relative* preference. That is, a negative coefficient indicates that fish prefer a lower proportion of sediment type, or shallower depth, relative to that which is available. A positive coefficient indicates just the opposite.

Detection Probability

The model generates posterior distributions of the estimated location of each fish on each day. We assume observed locations are recorded without error. Given the spatial scale of the model, we believe this assumption is warranted. Un-observed locations are interpolated by the movement and occupancy processes described above. We call fish k 's location on day $t \in \xi_k, \dots, T$ $s_{k,t}$. If, during a given iteration of the model, $s_{k,t}$ falls in one of the river kilometers in which there is an acoustic receiver, we say that the fish is available for detection and we set a variable $\text{search}_{k,t} = 1$. If we observe fish k on day t , then we also set a variable $d_{k,t} = 1$. For iteration n of the model, then, we have defined

$$p_n = \frac{\sum_{i=1}^S \sum_{j=1}^T \text{detect}_{ij}}{\sum_{i=1}^S \sum_{j=1}^T \text{search}_{ij}}$$

which gives the probability of detecting any of K fish on any of T days. During model development, we attempted to include this procedure within the model itself, but it generated major autocorrelation in the MCMC process. When we compared output from the model that included detection probability, and the model that did not, it became clear that the former was attempting to adjust the estimates of s_{ij} to estimates of the detection probability. That is, a certain number of the values of s_{ij}

Table 2.3: Run details for model runs on both male and female data. Iterations includes burn-in. Time is in minutes. We ran all chains in parallel on separate cores.

Model	Chains	Iterations	Burn-in	Thinning	Time
Males	3	1,000,000	100,000	100	1089.2
Females	3	1,000	4,000	1	9.7

were forced into areas containing an acoustic receiver. This had a profound impact on the estimates of both habitat selection and movement, one too pathological to solve with additional iterations, and increased thinning. Therefore, we calculate all values of p_n post-hoc. We believe that the effects of including detection probability on estimated spatial states is worth extensive study, but it is outside of the scope of this paper to consider here.

Implementation

Given that the hidden Markov process, the observable Markov processes, and the habitat selection process are all nested, and therefore hierarchical, we use a Bayesian hierarchical model to estimate the parameters. We have written the model in **JAGS** (Plummer, 2017) and use the function "autojags" in the package *jagsUI* (Kellner, 2017) to run the model from *R* (R Core Team, 2019a). The "autojags" function pauses the MCMC periodically to check for convergence ($\hat{R} < 1.1$) amongst all parameters. When it achieves convergence, the function halts sampling, and summarizes the output. We run the model separately on male and female Atlantic sturgeon, implicitly hypothesizing that migratory behavior for males and females is fundamentally different (Table 4.1).

2.3 Results

2.3.1 Movement

For male Atlantic sturgeon, the probabilistic weight of the hidden transition matrix is concentrated in the main diagonal, with median values near 0.95 and credible intervals above 0.9 (Table 2.4). This indicates that there are two classes of movement behavior, and that there is a high degree of persistence in both. That is, fish moving upriver tend to continue upriver for several days before switching to a downriver movement behavior. For the upriver transition matrix, most of the probabilistic weight is concentrated at or above the main diagonal, indicating that this matrix does indeed guide upriver movement (Figure 2.3, upper). This matrix also guides high persistence within a few states in particular: 7 (river kilometers 76–90), 11 (river kilometers 136–150), and 13 (river kilometers 166–180). The latter of the two represent known spawning regions in Newburgh, Hyde Park, and Catskill. The former, state 7, may represent an additional spawning region, or a staging area for males. Though not as pronounced, there is additional persistence in state 16 (river kilometers 211–225), which may represent yet another spawning region. Overall, the upriver probability transition matrix has very little weight concentrated below the diagonal, and therefore reflects highly directional movement which, given that the model calculated a second transition matrix with the majority of the probabilistic weight concentrated below the main diagonal (Figure 2.3, lower), could not be adequately described by a single matrix. The downriver matrix of for males is somewhat different in that there is some weight above the main diagonal, reflecting a more leisurely downriver migration. There are no states with very high and dense posteriors along the main diagonal, and therefore none with very high persistence. According to these estimates, then, a male Atlantic sturgeon generally moves into the river with very little fallback, remaining in one or more of three basins of attraction

Table 2.4: Hidden Markov probability transition matrices for both males and females, shown as follows: Median (95% HDI).

	Upriver	Downriver
	<i>Males</i>	
Upriver	0.95 (0.92,0.97)	0.05 (0.03,0.08)
Downriver	0.03 (0.00,0.08)	0.97 (0.92,1.00)
	<i>Females</i>	
Upriver	1.00 (0.98,1.00)	0.00 (0.00,0.02)
Downriver	0.63 (0.10,1.00)	0.37 (0.00,0.90)

for several days. Once transitioning into a downriver behavioral mode, the male may swim back and forth in the river, gradually moving downriver before finally exiting.

For female Atlantic sturgeon, the median and 95% credible intervals for the hidden Markov probability transition matrix values indicate that there is only one class of movement behavior (Table 2.4). Persistence in the upriver movement state is very high, with a median value very near 1. Transition probabilities within the downriver class of movement oscillate between 0 and 1, reflecting their prior distributions, given they have little to no relation to the data. It is clear from the downriver matrix that there is no coherent secondary class of movement (Figure 2.4). The posterior densities are all reflective of their priors, the concentration of weight within 6 states of a given state reflecting the daily movement restriction. The terms upriver and downriver thus have little meaning when applied to these two probability transition matrices. All aspects of movement are governed by the upriver transition matrix (Figure 2.4). The high degree of persistence in state 7 (river kilometers 76 – 90) is suggestive of spawning, as females do not typically stage (Dovel and Berggren, 1983). Persistence in the known spawning regions in state 11 (river kilometers 136 – 150) is only moderately high, perhaps reflecting rapid spawning and departure. Persistence in state 13 (river

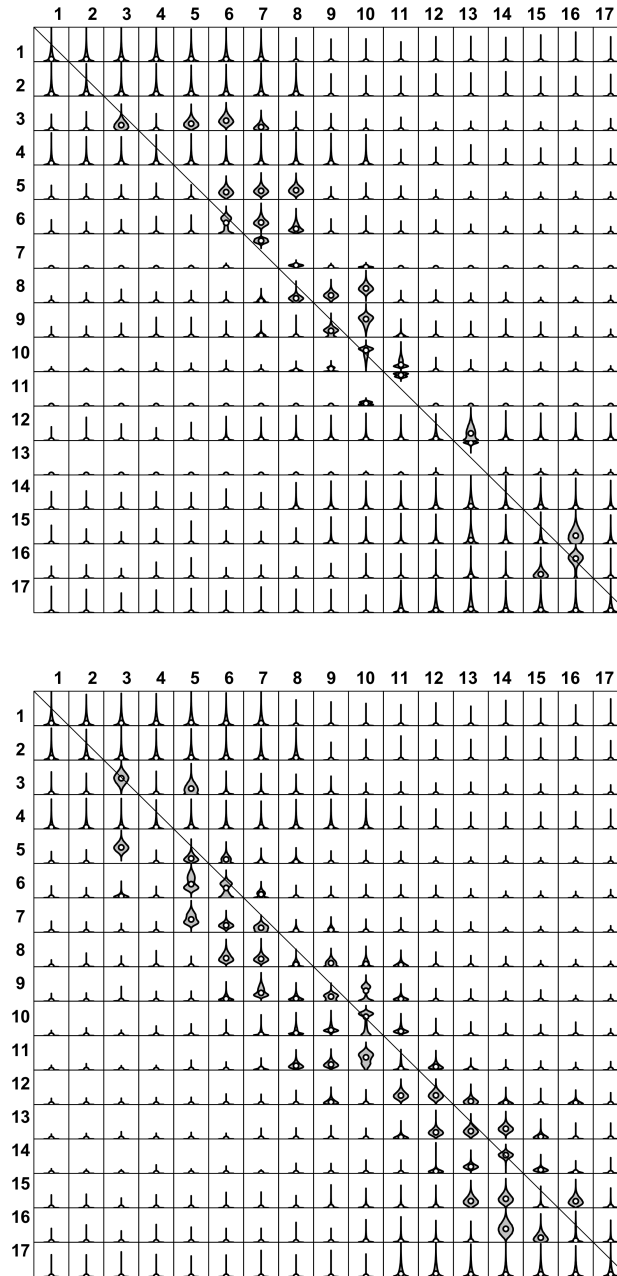


Figure 2.3: Violin plots of the posterior densities of transition probabilities from the model applied to male Atlantic sturgeon. Upper: upriver transition matrix. Lower: downriver transition matrix. The y-axis of all 578 individual plots ranges from 0 to 1. The labels to the left of the plots give the state from which an individual is transitioning, and the labels above the plots give the state to which an individual is transitioning. Open circles give the median. The solid diagonal line is drawn along the main diagonal of the transition matrix.

kilometers 166 – 180) is very low, with most weight concentrated in movement to the state below. This could also reflect rapid spawning and departure. The movement of an average female Atlantic sturgeon is more difficult to describe than a male's. One can draw a number of paths with such a symmetric matrix with just one major basin of attraction.

2.3.2 Habitat Preference

Habitat preference model results for both sexes are remarkably similar. At the river kilometer scale, only mean and maximum depth appear to play any appreciable role in selection (Tables 2.5 and 2.6). We have chosen a minimum inclusion probability of 0.75 as evidence that a given covariate is a significant predictor of occupancy. In this case, only maximum depth remains. Recall that we calculate maximum depth as an average of all maximum depths recorded for all composite habitat types within a given river kilometer. Contrary to our expectations, both male and female sturgeon tend, on average, to choose river kilometers with lower relative maximum depths. The posterior density of the maximum depth coefficient does overlap zero, but most weight is concentrated near -0.5 , indicating that the probability of occupying a given river kilometer decreases as the average maximum depth increases (Figure 2.5). The posterior density of the coefficient of mean depth (not shown) is also skewed slightly negative for both males and females (see Tables 2.5 and 2.6)), indicating a slight average preference for shallow mean depths. However, when we consider the relationship of occupancy probability in a given sub-state with maximum depth over the entire 95% high posterior density interval, we find that, on the upper extreme, some fish may in fact prefer waters with a maximum depth exceeding 40m (Figure 2.6), assuming habitat of all depths within the range of 1 – 50m is available. The maximum depth in the Hudson River below the Troy Dam is approximately 50m. Thus, while males and females appear to

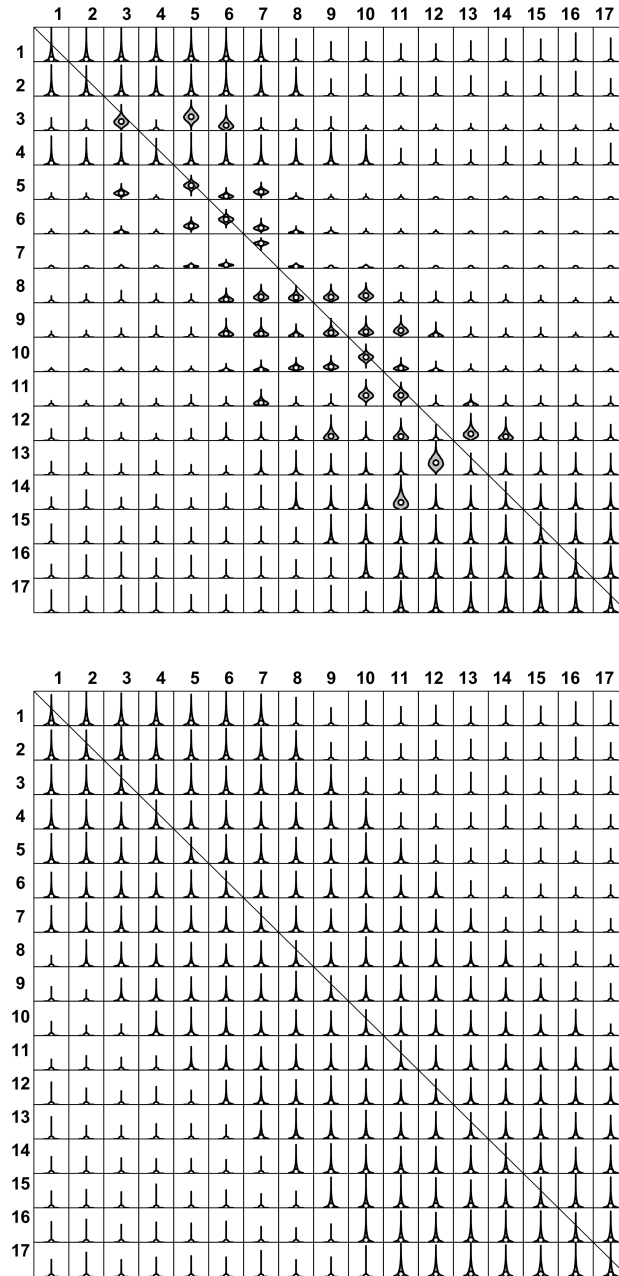


Figure 2.4: Violin plots of the posterior densities of transition probabilities from the model applied to female Atlantic sturgeon. Upper: upriver transition matrix. Lower: downriver transition matrix. The y-axis of all 578 individual plots ranges from 0 to 1. The labels to the left of the plots give the state from which an individual is transitioning, and the labels above the plots give the state to which an individual is transitioning. Open circles give the median. The solid diagonal line is drawn along the main diagonal of the transition matrix.

Table 2.5: Habitat sub-model results for male Atlantic sturgeon. The header "Inclusion" is inclusion probability.

	Inclusion	Median	σ^2	95% HDI
Intercept	—	-0.05	0.31	(-0.66, 0.57)
Mean Depth	0.65	-0.25	0.35	(-0.87, 0.45)
Max Depth	0.88*	-0.45	0.28	(-0.96, 0.16)
Index	0.51	-0.12	0.31	(-0.72, 0.50)
Mud	0.50	-0.01	0.32	(-0.63, 0.60)
Sand	0.50	-0.01	0.32	(-0.63, 0.60)
Gravel	0.50	0.00	0.32	(-0.63, 0.63)
Erosion	0.51	-0.00	0.32	(-0.62, 0.62)
Deposition	0.50	-0.01	0.32	(-0.62, 0.62)
Dynamic	0.50	-0.01	0.31	(-0.63, 0.59)
Channel Floor	0.51	-0.01	0.32	(-0.62, 0.62)
Channel Wall	0.50	-0.01	0.32	(-0.62, 0.61)
Channel Terrace	0.50	-0.00	0.32	(-0.60, 0.64)
Pinnacle	0.50	0.00	0.32	(-0.62, 0.62)
Marginal Flat	0.50	-0.01	0.32	(-0.63, 0.61)
Marginal Bank	0.50	-0.00	0.32	(-0.61, 0.63)
Bar	0.50	-0.00	0.32	(-0.61, 0.63)
Standardized Area	0.50	-0.02	0.31	(-0.65, 0.59)

prefer shallow waters on average, the model supports the hypothesis that, on occasion, fish may prefer very deep waters, or at times, exhibit no preference for depth at all.

We use the median estimates of the hidden transition matrices, the probability transition matrices, and occupancy probability for each river km to simulate 1,000 trajectories for both males and females. We summarize all 1,000 trajectories for both males and females in Figure 2.7. After a brief period of swift upriver migration, males pause for about six days near river km 80 (state 7). Then, males continue to travel upriver and concentrate near river km 120 (state 9). Once males reach this region of the river,

Table 2.6: Habitat sub-model results for female Atlantic sturgeon. Inclusion is inclusion probability.

	Inclusion	Median	σ^2	95% HDI
Intercept	—	-0.04	0.32	(-0.69, 0.56)
Mean Depth	0.63	-0.21	0.34	(-0.83, 0.47)
Max Depth	0.85*	-0.40	0.28	(-0.90, 0.24)
Index	0.49	-0.09	0.31	(-0.65, 0.56)
Mud	0.50	-0.01	0.32	(-0.62, 0.63)
Sand	0.50	-0.01	0.32	(-0.61, 0.64)
Gravel	0.50	0.00	0.32	(-0.61, 0.64)
Erosion	0.50	-0.00	0.32	(-0.64, 0.62)
Deposition	0.50	-0.00	0.31	(-0.62, 0.59)
Dynamic	0.50	-0.01	0.32	(-0.62, 0.63)
Channel Floor	0.50	-0.01	0.31	(-0.63, 0.59)
Channel Wall	0.50	0.00	0.32	(-0.64, 0.60)
Channel Terrace	0.49	-0.00	0.33	(-0.66, 0.63)
Pinnacle	0.49	-0.00	0.31	(-0.62, 0.63)
Marginal Flat	0.50	0.01	0.32	(-0.63, 0.61)
Marginal Bank	0.50	-0.00	0.32	(-0.59, 0.64)
Bar	0.50	0.00	0.32	(-0.60, 0.65)
Standardized Area	0.50	-0.04	0.31	(-0.65, 0.56)

there is a relatively equal spread of individuals 40 river km's above and below river km 120, perhaps representing those individuals still staging, and those individuals traveling further upriver to spawn elsewhere. Females migrate upriver just as swiftly, but with no noticeable staging behavior (Figure 2.7, right). Then, they remain largely in river km 80 (state 7), with a highly uneven spread of individuals up to about river km 160 (state 12). This may represent a tendency to linger in staging areas, with brief sojourns upriver to spawn. In general, estimated parameters suggest males use a larger portion of the river, while females concentrate more intensely downriver.

We also track the estimated daily location of each fish in the model output. This can place a substantial computational burden on a model with several fish, and so one may wish to track just a few representative individuals. We combine the posterior

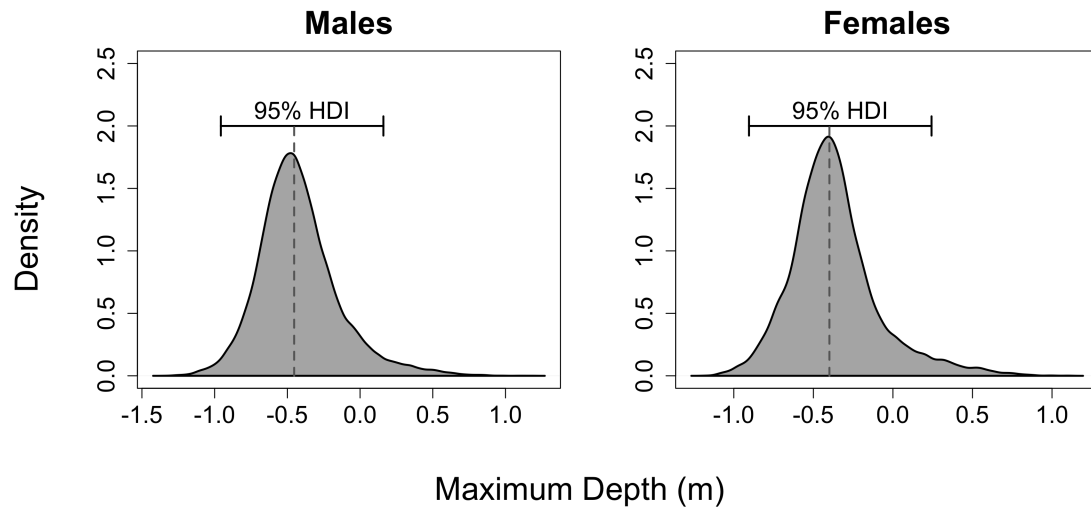


Figure 2.5: Posterior density of maximum depth coefficient for males (left) and females (right). Floating horizontal bars mark the 95% high posterior density intervals (HDI. Broken vertical gray lines are median estimates.)

densities of the estimated daily locations of all males and all females (Figure 2.8). Most males appear between Julian days 125 and 135 (May 5, 2015 to May 15, 2015) while most females appear between Julian days 140 and 150 (May 20, 2015 to May 30, 2015). Both males and females spread rapidly throughout the river once they enter. However, most males congregate in known spawning regions between Julian days 140 and 160 (May 20, 2015 to June 9, 2015). This coincides with the influx of females. Several males linger on into the fall, while most females leave in July. All males and females we consider in this study were caught and captured between late March and early June, and may be unrepresentative of the entire spawning population. Given the sex ratio of male:female Atlantic sturgeon is 2.75, it is especially likely that female migration and spawning behavior may be representative of early spawners, rather than females as a whole. We also provide trajectories of three individual males and three individual females (Figure 2.9). The tendency of males to linger, and the contrasting tendency of females to travel swiftly upriver and downriver are both clear. A subset of

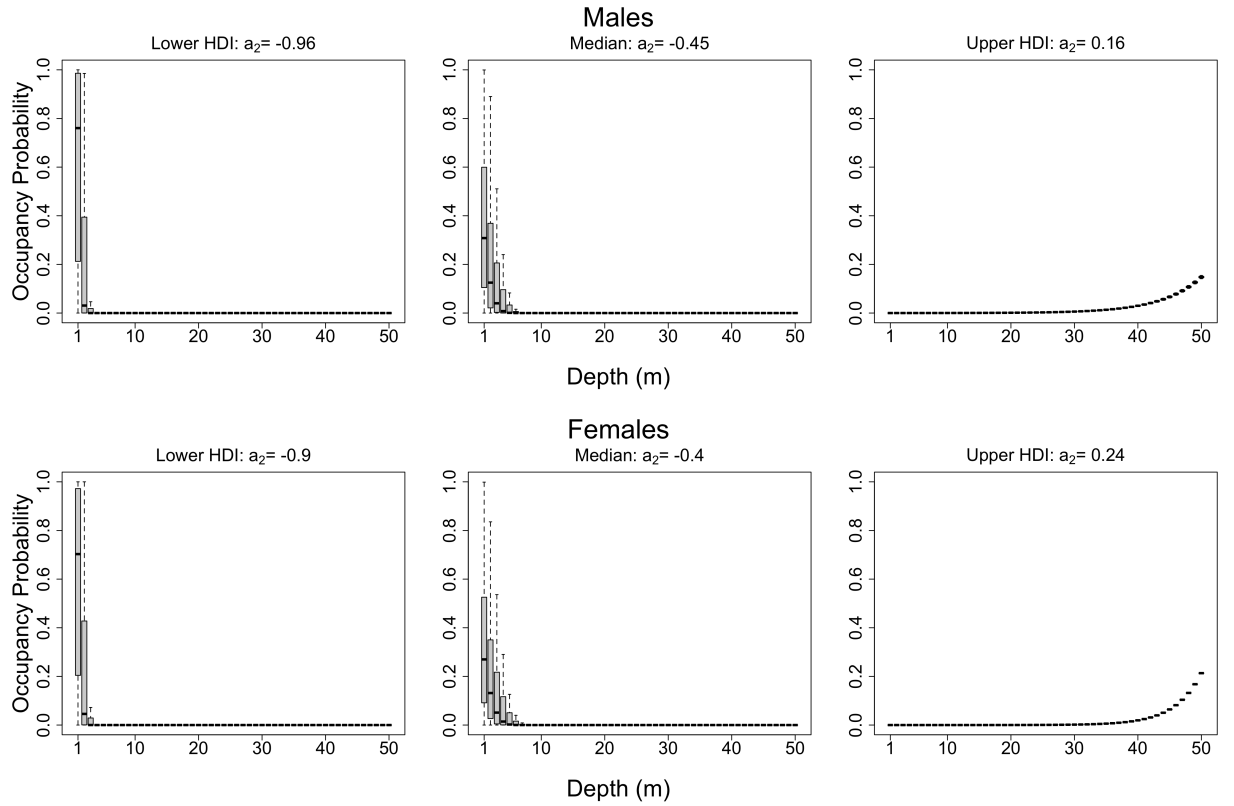


Figure 2.6: Box and whisker plots summarizing 10,000 samples of predicted occupancy probability for males (upper) and females (lower) for sub-states with maximum depths ranging from 1 to 50 meters in increments of 1m. Labels at the top of each figure give the maximum depth coefficient used to generate the samples: from left to right, the lower extreme, the median, and the upper extreme of the highest posterior density interval.

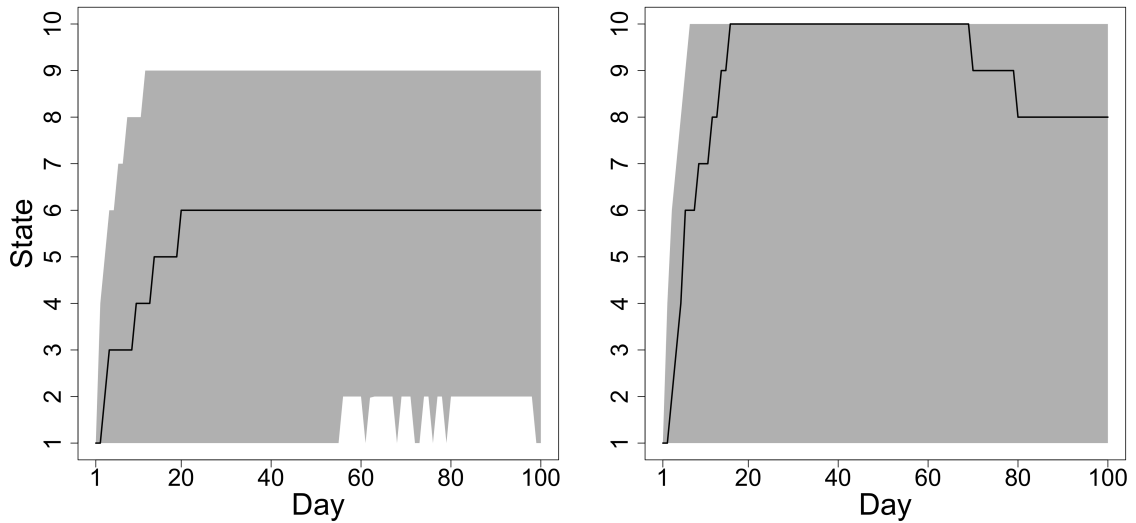


Figure 2.7: Median path of 1,000 simulated 50-day trajectories for males (left) and females (right). All individuals began in river kilometer 1 and in upriver migration. The solid black lines are the daily median for all fish. Shaded regions are 2.5% (lower) and 97.5% (upper) quantiles.

individuals of both sexes tend to linger in state 7 (river km's 76 – 90) at the tail end of their migration. Whether this is indicative of spawning, or feeding post-spawning, is unknown.

2.3.3 Detection

We calculate detection probability from each iteration of both models as described above. This produces a post-hoc posterior density of detection probability for both males and females (Figure 2.10). Median detection probability of males (0.271) is slightly less than the median detection probability of females (0.275), but a simple t-test suggests that there is no significant difference between these two distributions. Recall that these values are calculated based on the assumption that a given Atlantic sturgeon is only available for detection if it observed, or estimated, to be in a river km in which there is at least one acoustic receiver. If the receiver's detection range is greater

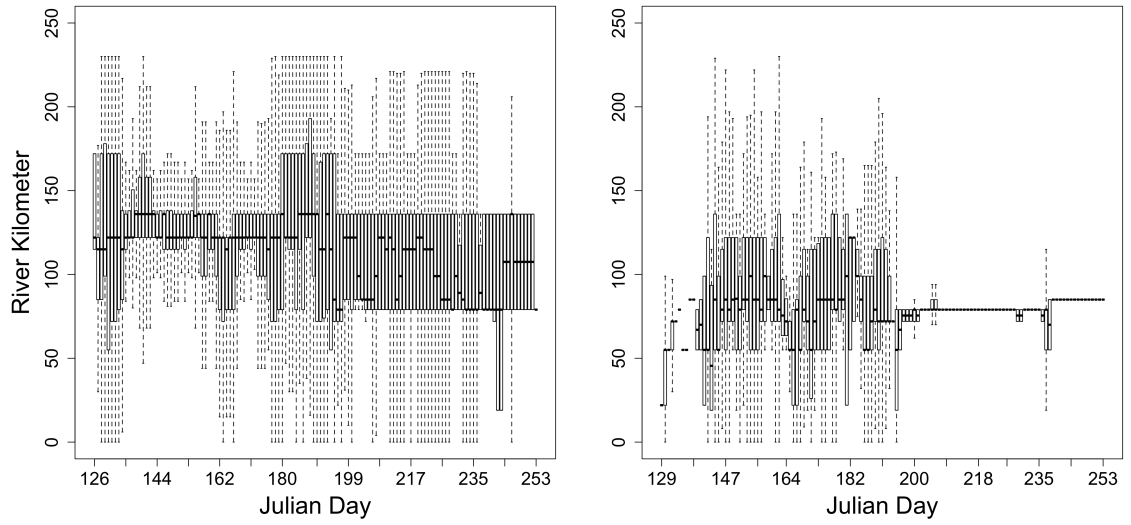


Figure 2.8: Combined posterior density of the estimated location of all males (left) and all females (right).

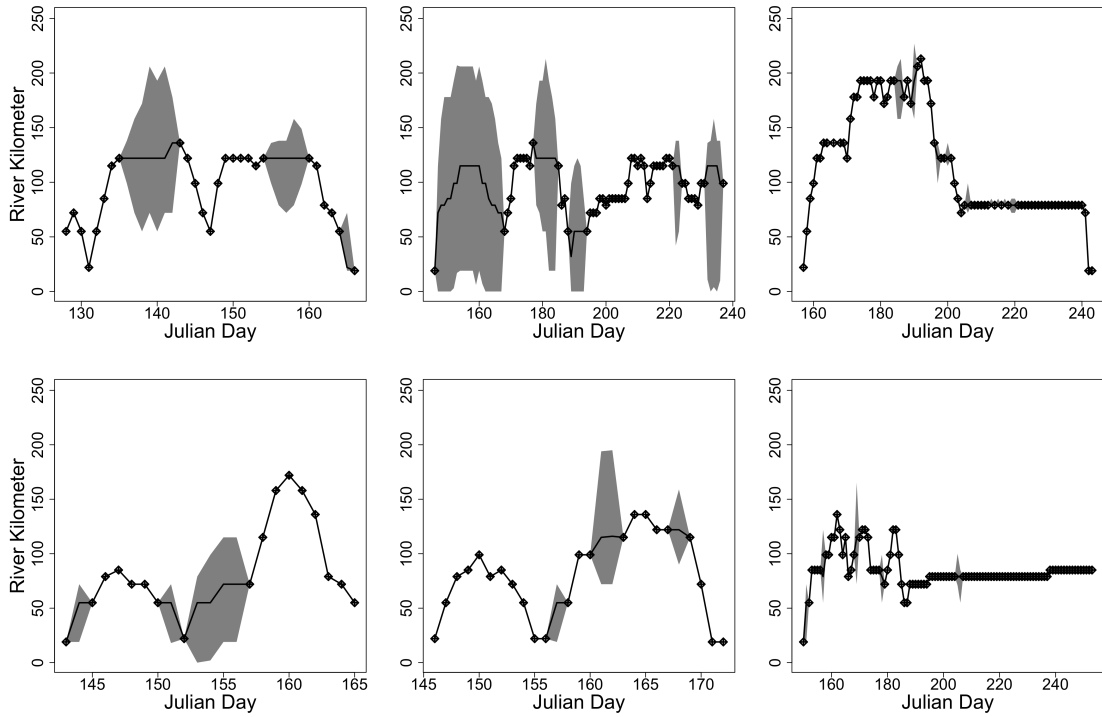


Figure 2.9: Estimated trajectories of three male (upper) and female (lower) Atlantic sturgeon. Solid black lines are median estimates. Shaded regions are 95% high posterior density intervals (HDI). Cross-hatched diamond symbols are observations.

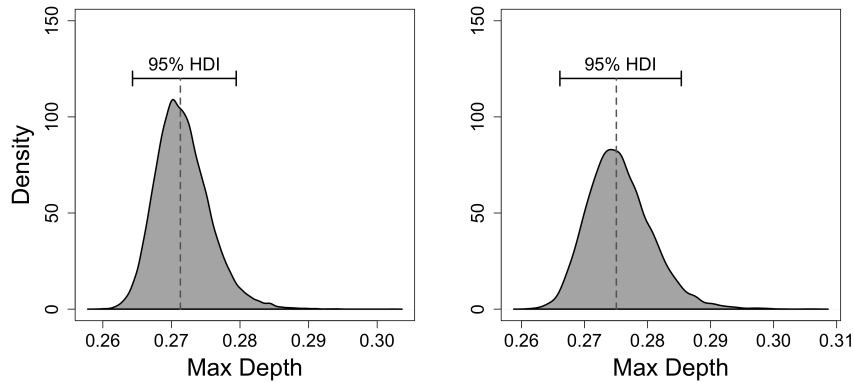


Figure 2.10: Post-hoc posterior density of males (left) and females (right). Floating horizontal bars mark the 95% high posterior density intervals (HDI). Broken vertical gray lines are median estimates.

than one river km, these values may be overestimates. If, by contrast, the receiver's detection range is less than one km, which is more likely, then these values may be underestimates. We note here that, if one does not first take into account whether a given fish is available for detection based on the range of the acoustic receivers, then these probabilities are estimated to be closer to 0.1.

2.4 Discussion

2.4.1 Scale

The model structure presented here is both a novel approach to summarizing complex diadromous behavior in fishes, as well as a study of the particular habits of the Atlantic sturgeon of the Hudson River. We assert that one of the model's most useful features is the separation of movement and habitat selection at scales appropriate to each process. Diadromous fish migrate long distances, often several hundred kilometers, but they still use and occupy habitats at a fine scale as they stage, and choose areas in which to spawn and possibly feed. Biologically speaking, during migration, an animal's

perceptive range is greatly exceeded by the distance they travel in a given time step. If we choose a model that is discrete in both space and time, then we may attempt to account for this in one of three ways: (1) decreasing the size of the spatial states, (2) decreasing the length of the time step, or (2) modeling movement and habitat selection as separate processes. Decreasing the size of the spatial states is by far the most punishing in terms of the increase in computational burden to the model. The number of parameters increases as a square of the number of spatial states. Decreasing the length of the time step may necessitate a decrease in the number of spatial states as step size decreases. It may also not be warranted by the structure of the data. Modeling movement and habitat preference as nested processes allows us to consider larger time steps, and therefore larger spatial expanses. Increasing the number of spatial sub-states in the occupancy sub-model results in a linear increase in computational time. There is, however, a limit to the fineness of scale at which one can model occupancy. This is imposed both by the methods of data collection, as well as by the scale at which animals move in a given time step. Our data arise from detection at acoustic receivers with a broad enough detection range that occupancy at scales smaller than a few hundred meters is ill advised. While we do observe individuals at multiple adjacent receivers within one day, we very rarely observe individuals at more than one state in a given day. In 2015, there were on average about three acoustic receivers in each spatial state that are spread fairly evenly within the states (see Figure 2.1). Therefore, we feel that our choices of scale for both movement and habitat selection are appropriate for Atlantic sturgeon in the Hudson River.

We want to emphasize here that choice of scale is crucial both to an accurate reflection of the data source, and as a reflection of the objectives of the study. With respect to the Atlantic sturgeon, we wish to interpolate missing data points, identify

regions and habitats with high concentrations of fish in both time and space, and to describe differences in migratory behavior between males and females. These objectives all operate at fairly broad spatial scales.

2.4.2 Movement

The median estimated path of Atlantic sturgeon in the Hudson River quickly converges in the central portion of the river, with some differentiation between males and females (Figures 2.7 and 2.8). However, it is clear from examination of the trajectories that each individual may follow a unique path through the river (Figure 2.9). This variability is captured by the variance in the probabilistic weight assigned to the elements of the transition matrices (Figures 2.3 and 2.4). Males show a tendency to linger briefly in staging areas prior to traveling upriver to known spawning regions. On average, males appear to linger for roughly ten days. Coincidentally, females tend to arrive approximately ten days after males (Figure 2.8). Thus, males' tendency to linger downriver may reflect staging, acclimating to freshwater, the wait for an appreciable number of females, or all of the above. Previous evidence has suggested that males arrive earlier, and stay longer in their natal rivers than do females (Dovel and Berggren, 1983). This certainly appears to be the case with these fish.

The hidden Markov portion of the movement sub-model identifies two separate classes of movement behavior for males, and just one for females (see Table 2.4). This provides quantitative evidence for our hypothesis that, while males stage downriver of spawning regions, females typically travel upriver, spawn quickly, and travel downriver in a symmetric pattern, with little to no staging behavior. We can see evidence for this behavior in the few male and female trajectories we have shown in Figure 2.9. Whereas males linger in spawning regions for 10 to 20 days, female trajectories peak briefly in

spawning regions before traveling downriver. We can also glean from these trajectories the effect of non-observation on estimated trajectories. As in Duskey et al. (2020), as the number of days between observations increases, uncertainty also quickly increases. When a trajectory is unknown, both male and female fish are generally quickly drawn to central areas in the river. Thus, this model structure is not necessarily ideal for predicting future trajectories. Instead, it excels at two things: first, at interpolating an individual's location between observations; and second, at summarizing basins of attraction in the river in the probability transition matrices. The three regions of the Hudson River showing greatest persistence for males are located at known staging and spawning regions. Transition probabilities for females are more diffuse, with just one area of attraction in state 7 (river kilometers 76 – 90). However, we believe the perceived persistence in state 7 is driven by just one individual female, whose trajectory is shown at the far right in Figure 2.9 (bottom). Given our results arise from just twelve females, it is unknown whether this behavior is common. The most common behavior can best be described as two incursions into the river, typically terminating at separate parts of the river. Whether this represents two spawning events, or the arrival of additional adults and a tendency to group, is unknown.

Atlantic sturgeon in the Hudson River exhibit a high degree of site fidelity from year to year. Some fish show such remarkable similarity in their year to year migratory paths, that investigators in the field can identify some individuals based just on their location on a given Julian day (Amanda Higgs pers. comm.). Site fidelity is not necessarily universal to Atlantic sturgeon. In the Altamaha River system in Georgia, Ingram and Peterson (2016) find that downriver migration is typically swift and direct. They find no evidence of site fidelity, however, whereas we find that fish for which we have multiple consecutive years of data tend to travel to the same regions year after

year, with remarkable fidelity (data not published). These regions are: near Hyde Park and Kingston, from river kilometers 136 to 148; and at or above Catskill, around river kilometer 182. Whether the difference in site fidelity between the New York Bight distinct population segment and the South Atlantic distinct population segment is due to differences between the rivers, or between the populations, is unknown. Ingram and Peterson (2016) also found no evidence for multiple spawning runs in the South Atlantic distinct population segment, whereas the New York Bight distinct population segment show evidence of runs in both the spring and the fall. This is likely due to differences in temperature and flow between the Hudson and Altamaha Rivers.

In Duskey et al. (2020), we state that estimated positions, as well as median estimated paths, can guide recreational and industrial usage of the river. Given the spread of adults of both sexes throughout the Hudson River, there are scarcely areas that can support heavy boat traffic without chance of striking adult Atlantic sturgeon. While males concentrate in the central portion of the river, thus largely fleeing from river kilometers 0 – 80, this coincides with the first influx of females. While, generally speaking, river kilometers 0 – 50 house far fewer spawning adult Atlantic sturgeon than the central river km's, it is of course a region through which they pass if they are to reach their natal spawning grounds. In that sense, any portion of a river below the spawning areas might also be classified as critical habitat.

2.4.3 Habitat Selection

For most, the inclusion probability is nearly exactly 0.5, with a median coefficient estimate near zero. That is, they are all switched on as often as they are switched off, and serve as little more than noise. However, at the one river km scale, it appears that depth is the only variable that has an appreciable effect on habitat selection, with

an inclusion probability of 0.88 and 0.85 for males and females, respectively (Tables 2.5 and 2.5). This is not unexpected. Given that observations of Atlantic sturgeon occupying deep channel habitat in the field, we suspected they might show a preference for river km's with deeper habitats. However, most of the posterior density falls below zero for both males and females (Figure 2.5), which indicates that, on average, Atlantic sturgeon gravitate towards river kilometers with lower *relative* average maximum depth. The distinction between relative and absolute is critical. When we simulate from the model, it suggests that Atlantic sturgeon always choose the shallowest waters available (Figure 2.6) As we have noted, a small portion of the highest posterior density interval of the maximum depth habitat preference coefficient falls above zero. At times, then, both males and females may gravitate the areas with deep channels.

That said, a preference for shallow waters is not at all unprecedented. Dunton et al. (2015) find that Atlantic sturgeon in coastal waters prefer shallow depths ($< 20\text{m}$). Stein et al. (2004) find that coastal bycatch of Atlantic sturgeon is concentrated in areas with depths of less than 60m, as well as in areas with sand and gravel substrate, which they suggest is indicative of a concentration of preferred food sources. Within the Hudson River, sand and gravel sediment types are concentrated at depths of roughly 20m or less (Figure 2.11). Sediment environment and bottom morphology have a less clear relationship with depth, though it is apparent that depositional environments do not appear until maximum depths exceed 10m and increase thereafter, and that channel walls begin to dominate as depth increases. Thus, a preference for shallower maximum depths within the Hudson River is perhaps not contradictory of previous evidence, and may be a composite of preferences for habitat types highly correlated with depth. Behavior and physiological needs are, of course, different during spawning migrations than during marine migration, but the extent of this difference is generally

unknown. Peak physiological performance of adult spawning Atlantic sturgeon occurs at water temperatures of about 16-21 degrees C (ASMFC, 2012). They have shown preference for temperatures within this range (Dunton et al., 2010). In 2015, many of the observed migrations take place between Julian days 120 through 160, during which time mean surface temperatures typically fall in or below this range (unpublished data). Thus, drivers of occupancy at depth may not be physiological, but rather trophic. Preferred food sources – for example, oligochaetes and gammarids – have been shown to aggregate in submerged aquatic vegetation (Strayer (2007)), which peaks at depths of 10 – 20m (unpublished data). It has been said that Atlantic sturgeon do not feed during spawning migrations, but that they do feed post-spawning (Smith, 1985), which may drive the apparent average draw to shallow waters.

The apparent draw to shallower depths may also result from using just one set of coefficients throughout the entire Hudson River. As needs change throughout the migration, those habitat covariates driving selection may also change. For example, male Atlantic sturgeon may seek out shallower habitats flush with preferred prey items in staging portions of the river. When they move upriver to spawn, they may instead follow the females to sandy or gravely substrate, to which eggs will adhere. As they travel back downriver, they may again seek prey items in shallow waters, or may show little to no preference at all as they migrate indiscriminately towards marine waters. Additional coefficients will add additional time to a model already burdened by a high number of parameters. One can associate different sets of coefficients with different regions of the river, which will serve as a hypothesis that the processes driving selection differ spatially. We have yet to run this version of the model ourselves. We simply note here that average maximum depth and mean depth are both correlated with most other habitat covariates. Thus, a preference for shallower depths may be reflective of

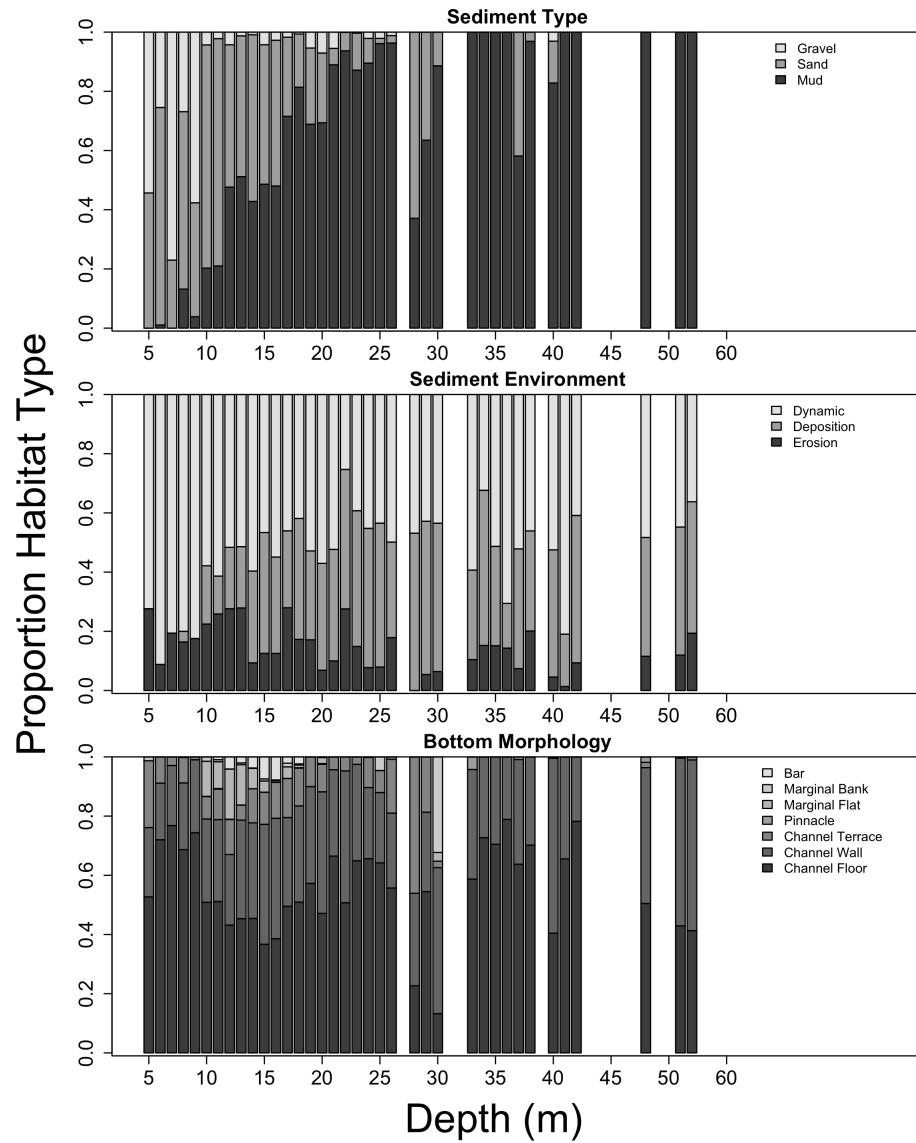


Figure 2.11: Stacked bar plots of proportion habitat type as a function of depth (m): sediment type (upper), sediment environment (middle), and bottom morphology (lower). Each bar is a summary over a 1m interval. Missing bars indicate we collected no data in the associated depth range.

multiple drivers of habitat selection, differing in both space and time.

Lastly, acoustic receiver detection may be limited by depth. That is, Atlantic sturgeon residing in deep waters may go undetected. We calculate mean maximum depth of estimated location for un-detected fish available for detection i.e. within range of any receiver. We compare these values to mean maximum depth of observed location for detected fish. We find that the mean maximum depth of un-detected males is slightly lower than that of detected fish by about 2m (19m vs. 21m), and that of un-detected females is lower by about 4m (23m vs. 27m. This suggests that we are more likely to detect individuals in slightly deeper waters. This is likely just a reflection of the fact that boat traffic, and therefore interference with acoustic receivers, is most dense near the mouth of the river, where waters are shallower than waters further upriver.

It is clear, then, that the habitat sub-model is only capable of providing as much information as it is given. We expected perhaps a moderate preference for certain substrate types, but these are typically only expressed at the terminus of a spawning migration. We thus recommend estimating multiple coefficients of habitat preference. For fish with some measure of site fidelity, it is best to index these by area. For fish without evidence of site fidelity, it may be best to index these by time, or perhaps by hidden state, or movement behavior.

2.4.4 Detection

The estimated detection probability of the acoustic array for both males and females is about 0.27. This is very similar to the detection probability of 0.3 that we used to simulate data in Duskey et al. (2020). We are surprised by the severity of the autocorrelation in the MCMC induced by the inclusion of detection probability within

the model. The model seems to be incapable of fitting the hidden and probability transition matrices to the paths of the sturgeon whilst simultaneously fitting even a simple model of detection probability to the data. When we include detection probability, the estimates of the movement shift, often dramatically, which suggests that the model is attempting to force the unobserved locations to fit a given detection probability. We find it somewhat disconcerting that a simple hierarchical calculation of detection probability operates with such a heavy hand. This may be unique to our model structure, or it may be that it is not uncommon for estimated movement and locations to be so affected by detection. Given that joint posterior densities allow for nested model estimates to have profound effects upon one another, we suspect that it is, in fact, the latter. If one has a method by which to confirm detection probability in the field, then detection probability may be of great help to estimate unknown locations. For example, if one uses a device with a known time-indexed detection probability of 0.5 in a given area, then one would expect to detect about 50% of the animals that pass through that area in a given time step. This serves as a useful constraint. An appropriate model structure should not, then, estimate that an animal appears in this area 100 times, when it is only observed 10 times. However, if detection probability is unknown, it may be unwise to include it as a parameter *within* the model itself, but rather to calculate outside of the model structure, if possible. We find no issue with detection probability in Duskey et al. (2020). However, in this paper, we also assume all areas within the hypothetical river are covered by some sort of array, and remove observations evenly throughout. For more complicated situations, where only a portion of a given habitat is covered, the effects of including detection probability may be severe, as in our model. We assert that this is a problem worthy of additional study, but do not address it further here.

2.4.5 Conclusions and Summary

The novel model structure presented above is a useful method by which to build a statistical summary of movement and habitat selection. It is capable of identifying and describing different classes of movement behavior. Its method of separation of scale may both reduce computational burden, and provide a more coherent description of the actual process of movement and occupancy. Detection probability may be a problematic player within the model structure, but may provide additional useful constraints if known. Estimating detection probability while taking both estimated position and acoustic receiver coverage allowed for what is most likely a more accurate summary of the receivers' actual ability to detect a fish, given it is in range. We have built this model specifically for diadromous fish moving in a largely linear system. However, it could easily be adapted for non-linear systems, and would be particularly useful for identifying evidence of separate classes of discrete movement behavior. We apply the model to acoustic receiver data from 45 tagged Atlantic sturgeon, and find evidence for two separate classes of movement behavior in males, summarized broadly as upriver with pronounced staging and downriver. We find no such evidence for females, indicating a lack of staging. We also identify regions in the Hudson River most flush with Atlantic sturgeons on a day-to-day basis. We find that average maximum depth is an important predictor of habitat selection. We will carry out additional model runs to determine if habitat preference varies by spatial region.

Chapter 3

Spatial tug of war: kriging spline model residuals

The smoothing spline class of models, including penalized splines such as the generalized additive model (GAM), allow us as ecological modelers to describe non-linear relationships between response and predictor variables of interest. A problem arises when the response variable is spatially autocorrelated. When we apply a spline class model to our data, we assume implicitly that the residuals are independent. However, our data often violate this assumption. A potential solution to this problem is to apply ordinary kriging to the model residuals in order to account for spatial dependence. This is a fairly common method used throughout the natural sciences. It is most often applied step-wise, in which the spline model is applied first, and ordinary kriging is applied to the spline model results. We have developed an implementation of this model in a Bayesian framework in **JAGS**, in which the spline model and the variogram are fit simultaneously via Markov Chain Monte Carlo (MCMC). In essence then, the spline bends to minimize residuals while the variogram attempts to explain whatever remains. We have often found that, upon model convergence, the behavior of both the spline and residuals is highly dependent upon the choice of the number of knots included in the spline. Generally, as knots increase, the importance of the variogram diminishes. This relationship is complicated by the structure of the data. We describe

these model characteristics, and present an application of the Bayesian model to a series of simulated data sets.

3.1 Introduction

Kriging, developed in part by D.G. Krige and G. Matheron in the mid-20th century (Cressie, 1990), and penalized splines, developed in part by Kimeldorf and Wahba (1970) and given an historical context by Wright (1977), are two common methods of spatial interpolation. A general overview of spatial interpolation methods can be found in Myers (1994). These methods were developed in the context of mining and engineering, respectively, but we use and analyze them here to describe the spatial distribution and abundance of more or less sedentary organisms. In the context of ecological applications, penalized splines may describe a nonlinear trend in abundance with space, while kriging may describe aggregative or territorial behavior that presents as a certain degree of autocorrelation. While both were developed as methods of interpolation, kriging and penalized splines are fundamentally different. Penalized splines seek to approximate the general shape of a process, while kriging attempts to fit the data with as much accuracy as possible, serving as an interpolator for areas in which a given investigator has no data. That is, analysts choose penalized splines to describe a nonlinear trend, and kriging to describe stochastic variability. It is easy to see that penalized splines and kriging are similar. In fact, we can render the predictive output of kriging and penalized splines equivalent by sensible choice of the smoothing parameter, described below by the Greek letter π (Matheron et al., 1981; Dubrule, 1983). Kriging interpolates via weighted sum of the observations, and these weights are determined by the calculated distance between an unobserved location and all observed locations. Splines interpolate by fitting to observed data points, and the influence of each data

point on the fitted line or surface diminishes as one travels further from its location. Cressie and Wikle (2015) noted that the penalized spline is essentially kriging with a fixed covariance function. While it is clearly possible to force these models into equivalency, the philosophy underlying each method is different, particularly if we investigate the structure of the variance. Penalized splines operate by minimizing the squared deviation from the estimated mean, which is itself a fixed non-parametric function of the data. Kriging operates by minimizing the variance of the estimator, which is a unique function of the data at every point. Thus penalized splines operate on the assumption of fixed effects, while kriging operates on the assumption of random effects. The practical outcome of this distinction is that kriging is both precise and accurate at a local scale, whereas penalized splines provide a good description of trend underlying apparent noise in the data.

There is a great deal of literature in which investigators compare the fit of penalized splines and kriging in a variety of contexts (Dubrule, 1984; Voltz and Webster, 1990; Laslett, 1994; Altman, 2000; Da Costa et al., 2018). Unsurprisingly, the outcome of kriging is generally as good as or better than penalized splines. As competitors, kriging has a clear advantage. Which of either is employed should depend upon the nature of the motivating question. Spatial trend is driven by one or more broad-scale processes, and choosing a penalized spline as an estimator reflects a desire to empirically characterize how abundance or density respond to these processes. Spatial autocorrelation is driven by whether the presence of one or more individual exerts an attractive or repellent force on other individuals, and typically operates on a relatively fine-scale. The choice of a kriging model, then, reflects a desire to characterize this covariance amongst individuals in space. We, however, are interested in those cases in which penalized splines and kriging are treated as partners. At the end of the 20th

century, there was a paradigm shift in statistical thought, in which it was considered that some model residuals should no longer be classified as unstructured nuisance, but rather as a set of autocorrelated variables arising from a mechanism separate from that producing the mean (Opsomer et al., 2001). This invited an updated model structure, in which separate models were applied to describe the mean, and the autocorrelation observed in the residuals. This is meant to be applied in cases in which data are nonstationary, a situation noted by Cressie (1986) as the rule rather than the exception. Simple kriging assumes the mean is constant, while ordinary kriging estimates a local mean for each observation.

Thus arose universal and regression kriging. Universal kriging, or kriging with a trend, assumes that the mean is given by a deterministic function of space, often a polynomial. Regression kriging, alternatively termed kriging with external drift, assumes the mean is given by a stochastic function of one or more auxiliary variables. These auxiliary variables are also typically spatial in nature, such as depth or elevation, temperature, or habitat type. One then simultaneously krige the residuals that result in an iterative reweighting process. This is a fairly common procedure (Alsamamra et al., 2009; Wu and Li, 2013; Ndongo et al., 2016). Given the discussion above, we are interested in the case in which both penalized spline regression and kriging are applied subsequently to the data as a function of the spatial coordinates. Regression kriging or kriging with external drift is preferable when the residuals are independent of the auxiliary variables (Hengl et al., 2007). At first glance, regressing upon the spatial coordinates, and kriging upon the residuals, seems to violate this principle. However, we are considering situations in which the trend guiding abundance throughout the region is driven by a separate process than that which guides local spatial autocorrelation. That is, the coordinates in the regression serve as a proxy for broad spatial drivers, while kriging

accounts for aggregation on a finer scale, driven by a different process. In this sense, the local residual is, in fact, independent of broad scale trend, and depends only upon local abundance. Caution is required, however, as is a good understanding of the mechanistic drivers of the process. Arguably, the choice of this model is stronger prior information than any individual choice of distribution we might make for the parameters. If the residuals are dependent upon the auxiliary information, regression kriging amounts to an empirical attempt at a better model fit to the data, absent any sound statistical basis.

This procedure is not entirely unprecedented. Kammann and Wand (2003) merged additive models and kriging to produce estimates of birth weight in Cape Cod, Massachusetts. These aptly named “geoadditive” models appear to be an effective method by which to account for additional, non-spatially coded information. Chang et al. (2017) applied generalized additive models to both spatial coordinates and water depth, and then kriged the residuals to account for relatively fine-scale aggregation. Given that penalized splines can produce equivalent estimates, what, then, is the use of applying both model structures? Note that, if we apply simple or ordinary kriging, we must assume the mean is zero or a non-zero constant, respectively. If we apply universal kriging, we can use a deterministic function to factor out the mean process. This could include an interpolating spline. If we choose kriging, the motivation for first applying a spline model must be that we suspect a general trend over the entire space, with error. That is, the trend operates on a broader scale. A similar logic applies in the reverse direction. If we apply a spline model, the motivation for subsequently applying a kriging model must be that we suspect spatial autocorrelation on a finer scale. How much finer is not a matter of objectivity. It is a matter of the processes that drive the observations.

The purpose of the paper is two-fold. First, to present the penalized spline-kriging model structure and adapt it for a Bayesian framework. Fitting both models simultaneously with a joint probability distribution allows the penalized spline and the variogram to interact directly, which then allows us to identify any potential problems, both philosophical and practical. Second, we undertake a simulation study to demonstrate the behavior of penalized splines and variograms when acting in concert to fit to observations. Ultimately, we aim to develop a guiding set of principles as to when this model structure is and is not appropriate, and the strength of the conclusions that can be drawn from the results.

3.2 Problem Definition

3.2.1 Kriging

We begin with a familiar problem. We would like to estimate the value of a stochastic process Z at the point ϵ_0 as a function of the observations $\epsilon^T = (\epsilon_1, \dots, \epsilon_N)$. If we restrict ourselves to linear predictors, then we have a function of the form

$$\hat{\epsilon}_0 = Z(\epsilon_0) = \sum_{i=1}^N w_i \epsilon_i = \mathbf{w}^T \boldsymbol{\epsilon}$$

where

$$\sum_i^N w_i = \mathbf{w}^T \mathbf{1} = 1$$

Requiring that the weights $\mathbf{w}$ sum to 1 guarantees that the estimator $Z(\epsilon_0)$ is unbiased, assuming the mean is constant. We want to choose $\mathbf{w}$ so as to minimize the estimation variance. If we write the covariance of any of the two observations as $\text{Cov}(\epsilon_i, \epsilon_j) = C_{ij}$,

then we can write the variance-covariance matrix of the observations ϵ as

$$C = \begin{bmatrix} C_{11} & \dots & C_{1N} \\ \vdots & \ddots & \vdots \\ C_{N1} & \dots & C_{NN} \end{bmatrix}$$

and write the vector of covariances of the observations ϵ with ϵ_0 as $D^T = (\text{Cov}(\epsilon_1, \epsilon_0), \dots, \text{Cov}(\epsilon_N, \epsilon_0))$,

then we can write the estimation variance equation as

$$\begin{aligned} \text{Var}(\epsilon_0) &= E[(\epsilon_0 - \hat{\epsilon}_0)^2] \\ &= \text{Var}(\epsilon_0) + \text{Var}(\hat{\epsilon}_0) - 2\text{Cov}(\hat{\epsilon}_0, \epsilon_0) \\ &= \sigma_{\epsilon_0}^2 + \text{Var}(\mathbf{w}^T \epsilon) - 2\text{Cov}(\mathbf{w}^T \epsilon, \epsilon_0) \\ &= \sigma_{\epsilon_0}^2 + \sum_{i=1}^N \sum_{j=1}^N w_i w_j C_{ij} - 2 \sum_{i=1}^N w_i D_i \\ &= \sigma_{\epsilon_0}^2 + \mathbf{w}^T C \mathbf{w} - 2\mathbf{w}^T D \end{aligned}$$

with the restriction that $\mathbf{w}^T \mathbf{1} = 1$. We can use the method of Lagrange multipliers to introduce this restriction into the estimation variance equation as follows

$$\sigma_{\epsilon_0}^2 + \mathbf{w}^T C \mathbf{w} - 2\mathbf{w}^T D + 2\lambda(\mathbf{w}^T \mathbf{1} - 1)$$

Note that the additional term will be zero when $\mathbf{w}^T \mathbf{1} = 1$. When we take partial derivatives with respect to $\mathbf{w}$ and λ , and set to zero, we find that

$$\mathbf{w} = C^{-1}(D - \lambda \mathbf{1})$$

and that

$$\lambda = \frac{\mathbf{1}^T C^{-1} D - 1}{\mathbf{1}^T C^{-1} \mathbf{1}}$$

Thus, the weights w are determined entirely by the structure of the variances and covariances. In order for these equations to be solvable, the variance-covariance matrix must be positive definite. The familiar equations relating the estimation variance to distance between points, or variograms, are designed with this restriction in mind.

Now, we are interested in the more specific problem in which the observations ϵ are the residuals of a penalized spline fit to observations Y . If we fit a model to Y with the design matrix B , given generally by

$$Y = B\hat{\beta} + \epsilon$$

then we can kriging the residuals ϵ by fitting a variogram to the variance-covariance structure of the residuals. Call B_i the i th row of the design matrix B . Then, the covariance of any two residuals is given by

$$\begin{aligned} \text{Cov}(\epsilon_i, \epsilon_j) &= \text{Cov}(y_i - B_i\hat{\beta}, y_j - B_j\hat{\beta}) \\ &= E[(y_i - B_i\hat{\beta}) - E[y_i - B_i\hat{\beta}]][(y_j - B_j\hat{\beta}) - E[y_j - B_j\hat{\beta}]] \\ &= E[(y_i - B_i\hat{\beta}) - (B_i\hat{\beta} - B_i\hat{\beta})][(y_j - B_j\hat{\beta}) - (B_j\hat{\beta} - B_j\hat{\beta})] \\ &= E[(y_i - B_i\hat{\beta})((y_j - B_j\hat{\beta}))] \\ &= \text{Cov}(y_i, y_j) \end{aligned}$$

Thus, assuming the model is unbiased at all observations, the covariance structure of the observations is equivalent to the covariance structure of the residuals. This is precisely where our familiar problem diverges. If the model is unbiased at all observations, kriging of the residuals is clearly not necessary, and serves only to improve the fit of the model at the expense of any sound statistical reasoning. Already, our first principle arises:

Principle 1: Apply kriging to penalized spline model residuals when you expect or observe local bias in the residuals of the model results.

The term "local" is, of course, subject to each reader's judgment, but is generally meant to describe biases that occur over just a fraction of the entire range of the process. The same guidelines for kriging also apply. That is, we appeal to the variogram when we wish to estimate the value of a stochastic process at points over which there are no observations. In our case, the process is fine-scale autocorrelation remaining in the residuals of a penalized spline model fit.

3.2.2 Penalized splines

Informally, splines are piecewise polynomial functions stitched together at locations called knots, the number and placement of which we may choose. We use splines to interpolate unknown values within a given range of observations. Penalized splines apply a penalty to the sum of squared errors proportional to some measure of curvature of the polynomial functions to prevent overfitting to the observations. The penalty is often given as the integral of the square of the second derivative scaled by a constant π , as follows

$$\sum_{i=1}^N (y_i - \hat{f}(x_i))^2 + \pi \int_{\min(\mathbf{x})}^{\max(\mathbf{x})} \hat{f}''(x) dx$$

The cubic spline is the minimizer of this entire system, which uses basis functions f of the form $z_k = ||x - k||^3$, with k indexing each of the knots, typically placed at data quantiles, or spaced evenly throughout the range of the data. The thin plate spline, which is the default in the package **mgcv** (Wood, 2003), is the multi-dimensional generalization of the cubic spline, and uses radial basis functions of the form $z_k = r^2 \log(r)$ where $r = ||x - k||$. We use only cubic splines and thin plate splines throughout this paper, given that they are in such common use. We can write the basis functions,

whether they be cubic or thin plate, in matrix form

$$B = \begin{bmatrix} X & Z_K \end{bmatrix}$$

where the matrix X contains the parametric variables, including the intercept, and the matrix Z_K contains the non-parametric basis functions. We can write the vector of coefficients, both parametric and non-parametric, as $\beta^T = (\beta_0, \dots, \beta_{K+m})$, where m is the number of parametric variables and K is the number of knots. If we call the residuals $\epsilon^T = (\epsilon_1, \dots, \epsilon_N)$, then we can write the system of equations as

$$Y = B\beta + \epsilon$$

and the penalized error function as

$$(Y - B\beta)^T(Y - B\beta) + \pi\beta^T\Omega\beta$$

The elements of Ω are often given by $\Omega_{ij} = \int_{\min(\mathbf{x})}^{\max(\mathbf{x})} f_i''(x)f_j''(x)dx$ where the functions f_i and f_j are taken from the matrix B . The second derivative is a natural expression of the curvature inherent in each function f . If we take the partial derivative with respect to β and set to zero, we find that

$$\hat{\beta} = (B^T B + \pi\Omega)^{-1} B^T Y$$

a variation of the familiar regression formula. As $\pi \rightarrow 0$, the penalized spline becomes an interpolator, and as $\pi \rightarrow \infty$, it becomes a simple parametric regression as the basis coefficients of the non-parametric function approach zero. Note that

$$\epsilon = Y - B\hat{\beta}$$

and the direct connection between the curvature of the spline, guided by the parameter π , and the character of the residuals becomes clear. Note also that the structure of the design matrix B is determined by the choice of the number of knots. It is clear from these formulas that, if we were to manually decrease π and to increase the number of knots, local bias would decrease, and therefore the importance of the variogram would diminish, reflected in the estimates of the sill and the range. Thus, we arrive at our second principle:

Principle 2: There is a direct trade-off between the curvature of the penalized spline as a function of the number of knots, and the structure of the variance-covariance matrix as specified by the variogram.

For those applying the models sequentially, and using generalized cross-validation to select the smoothing parameter π (Wood, 2011), the penalized spline takes precedence. The fitted variogram may then only be informed by what remains, and the local bias therein. Any fitting scheme without proper re-weighting of the coefficients of the penalized spline by the covariances provided by the variogram allows for no communication between the two model structures, and is therefore generally inadvisable. We believe that, in most cases, it is best to fit both models simultaneously via methods of joint probability estimation, and present a model by which to do so below. Here, the promised tug-of-war arises as the penalized spline bends to explain the mean while the variogram sweeps away the residuals. While it is possible to estimate the optimal number and placement of knots (Lindstrom, 1999, 2002), we choose to manipulate the number of knots manually. In the simulation study that follows, we investigate the effects of knot choice on the tug between the penalized spline and the variogram. We hypothesize that, as the number of knots increases, the degree of autocorrelation in the residuals decreases, and therefore the importance of the variogram decreases,

reflected as a decrease in the estimates of both the sill and the range. Should this prove plausible, another principle will arise.

3.3 Simulation Study

We have built a Bayesian model by which to apply joint probability estimation to the parameters of the penalized spline and the variogram simultaneously. We first describe the structure of the model, followed by the process of data generation. We describe all functions in detail in order to improve the reproducibility of the models, in concert or individually.

3.3.1 Estimation Model

We apply the methods of Crainiceanu et al. (2005) to produce estimates of a low rank spline using the package **jagsUI** (Kellner, 2019). The motivation of low rank spline is to remove parametric variables to reduce autocorrelation between the coefficients of the basis functions. In practice, autocorrelation in the model structure induces autocorrelation in successive draws of Markov Chain Monte Carlo (MCMC) chains. For the cubic spline, this means that the basis functions

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=4}^{K+3} \beta_k ||x - \kappa_k||_+^3$$

become

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=2}^{K+1} \beta_k ||x - \kappa_k||_+^3$$

where K is the number of knots, κ_k is the k^{th} knot, and the double vertical bars represent the Euclidean 2-norm, or distance between the observations x and the knots. The subscript $+$ is a function that sets all non-positive values to zero. This means that

each smooth function operates only at or above the linear coordinate of its associated knot. Similarly, for thin plate splines, the basis functions

$$f(x) = \beta_0 + \beta_1 x + \beta_2 y + \sum_{k=3}^{K+2} \beta_k r^2 \ln(r)$$

become

$$f(x) = \beta_0 + \sum_{k=1}^K \beta_k r^2 \ln(r)$$

where

$$r = ||x - \kappa_k||$$

where the double vertical bars are, again, the Euclidean 2-norm or distance between the observations x and the knots. Using the clever reparameterization described in Crainiceanu et al. (2005), we can set the coefficients of the basis function $b_K = \Omega^{1/2} \beta_K$, where the subscript K is meant to reference all coefficients of the basis functions. Say $\mathbf{b}^T = (\beta_{\text{parametric}}, b_K)$, where the subscript parametric to only those coefficients of the parametric variables. Then we can rewrite the estimation equation as

$$Y = D\mathbf{b} + \epsilon$$

where

$$D = \begin{bmatrix} X & Z \end{bmatrix}$$

where

$$Z = Z_K \Omega^{-1/2}$$

We can then use the following priors for these coefficients

$$\beta_{\text{parametric}} \sim N(0, 10^4)$$

$$b_K \sim N(0, \sigma^2)$$

$$\sigma^2 \sim T(10, 1)[0, c]$$

where the distribution given by $T(10, 1)$ is the non-central t-distribution with non-centrality parameter 10 and 1 degree of freedom. Gelman et al. (2006) recommends non-central t priors for variance parameters, and we have found that they improve mixing and convergence of the MCMC process. In practice, we have also found that mixing and convergence improves drastically by setting the variance of the parametric coefficients equal to σ^2 . This may result in scaling issues, as σ^2 contains information about both the variance of the coefficients and the penalty π . Readers may wish to try both formulations, and evaluate both fit and convergence. The constant c represents an upper limit for the variance. One can easily leave the variance unbounded. However, this can produce severe autocorrelation both amongst the penalized spline coefficients, and between these coefficients and the parameters of the variogram. We have found that imposing a limit on the variance reduces the curvature of the penalized spline as the coefficients are forced closer to zero, essentially giving the variogram a head start. We assert that the major reduction in autocorrelation outweighs the slight reduction in curvature of the penalized spline, although we caution that this may induce the appearance of mechanistic autocorrelation where there is none.

With a careful choice of constant c , the likelihood, with observations Y , is then

$$\mu = e^{Db+\text{offset}}$$

$$Y \sim \text{Poisson}(\mu)$$

where the offset is the area over which the observations Y were collected. The Poisson distribution is commonly used for organism abundance data, but we can easily modify this model to accommodate over- or under-dispersion by using the negative binomial distribution.

We calculate the residuals from the estimates μ as follows

$$\epsilon = Y - \mu$$

Within the model, it is only necessary to estimate the parameters of the variogram. The weights for each new estimation point can be calculated outside of the model. We place the following priors on the parameters of the variogram

$$\text{sill} \sim T(20, 1)[0,]$$

$$\text{range} \sim T(0, 1)[0,]$$

Through trial and error, we find that these priors produce similar estimates to those produced by the more common so-called uninformative normal priors with zero mean and large variance, but with superior mixing and convergence. The inclusion of the term $[0,]$ indicates that we truncate the distribution at zero. We calculate the covariance

matrix from these parameters as follows

$$C_{ij} = \text{sill } e^{\frac{-||x_i - x_j||}{\text{range}}}$$

This is the exponential variogram, which we use throughout. This could easily be modified to include other variogram equations. We have found that including a nugget induces autocorrelation into the estimation of the variogram parameters, and so we omit it in both data generation, and in estimation, in order to be able to run many models in a relatively short period of time. The likelihood is then

$$\epsilon \sim MVN(\nu, C)$$

where MVN refers to the multivariate normal distribution. We set the prior for the mean vector ν as

$$\nu \sim N(\beta_\nu, \sigma_\nu^2)$$

$$\beta_\nu \sim N(0, 10^4)$$

$$\sigma_\nu^2 \sim T(10, 1)$$

With common mean β_ν , this is defined as ordinary kriging. If we set $\beta_\nu = 0$, then this is simple kriging. Estimating β_ν can be difficult in MCMC, as it also induces autocorrelation. We again chose to set $\beta_\nu = 0$ for the sake of running several models in a relatively short period of time.

3.3.2 Data Generation

We generate a total of twenty data sets for this study. Ten are two-dimensional, with each data point described by an x-coordinate and an observed count. The other ten are

three-dimensional, with each data point described by an x-coordinate, a y-coordinate, and an observed count. In general, for both the two- and three-dimensional data sets, we use a non-negative function to generate a global mean, add fine-scale autocorrelated errors, and then exponentiate the sum to generate a set of positive mean values for a Poisson distribution. For the two-dimensional data sets, we generate the mean with the following function

$$f(\mathbf{x}) = m \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x_i - \mu_{\mathbf{x}})^2}{2\sigma^2}}$$

This is the familiar Gaussian function, with the multiplicative constant m meant to increase or decrease the magnitude of the values generated by f applied to each x-coordinate $x_i \in \mathbf{x}$. Inherent in this formulation as written above is that the *RHS* is applied independently to each of the, say, n $x_i \in \mathbf{x}$. We always set the value μ_x equal to the geometric center of the data, indicating that a peak of mean count occurs in the central region. We then generate autocorrelated residuals for each of the n mean data points as follows

$$g(\mathbf{x}) = \Gamma \epsilon$$

where the $n \times 1$ vector ϵ is a set of *IID* random Gaussian variables with mean zero and user chosen variance σ . The $n \times n$ lower triangular matrix Γ is given by

$$\Gamma = \text{chol}(C)$$

where the function $\text{chol}(\cdot)$ refers to the Cholesky decomposition and where

$$C = \text{nugget} + \text{sill} e^{\frac{-D}{\text{range}}}$$

The elements d_{ij} of the $n \times n$ matrix D are given by $\|x_i - x_j\|$, the Euclidean 2-norm, or distance. We generate the observations Y as follows

$$Y \sim \text{Poisson}(e^{f(\mathbf{x})+g(\mathbf{x})})$$

Note that exponentiation ensures that any negative numbers generated by adding $g(\mathbf{x})$ to $f(\mathbf{x})$ will be placed on a non-negative scale. We use the same parameters for the variogram for all two-dimensional data sets: a nugget of zero, a sill of 100, and a range set to 10% of the maximum range of the data. We set $n = 50$ data points.

The generation of the three-dimensional data sets is nearly identical. The vector $\mathbf{x}$ now contains elements of the form (x_i, y_i) . We use a multivariate Gaussian function to generate the mean process as follows

$$f(\mathbf{x}) = m \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{x}})\Sigma^{-1}(\mathbf{x}-\mu_{\mathbf{x}})^T}$$

The multiplicative constant m has the same meaning as above. We treat each mean value as independent by setting $\Sigma = \sigma^2 I$. The value $\mu_{\mathbf{x}}$ is again placed at the geometric center of the range of the data. The autocorrelated residuals are generated just as above, producing observations Y at each of the ordered pairs in $\mathbf{x}$. For the three-dimensional case, the variogram parameters are as follows: a nugget of zero, a sill of 100, and a range set to 15% percent of the maximum range of the data. We set $n = 49$ data points on a 7×7 grid.

We generate ten data sets for both the two- and three- to explore the effect of the magnitude and range of peakedness or maximum abundance on the balance between the curvature of the penalized spline, and the range and sill of the variogram. We

accomplish this by varying the value of the variance σ^2 of the mean generating functions. Decreasing the variance produces a higher peak over a narrower range, which we refer to as peakedness. Given that both the two- and three-dimensional data sets have a mean generating process at or near zero at the upper and lower spatial extremes, peakedness can also be interpreted by the maximum count (Figures 3.1 and 3.2). We also generate two additional data sets: a two-dimensional data set and a three-dimensional data set, both with $\sigma^2 = 0.5$, and with no autocorrelated residuals. In order to explore the effect of knot choice on the outcome of the balance between the curvature of the penalized spline and the parameters of the variogram, we ran nine models on each of the 20 data sets with autocorrelated residuals, with the number of knots range from 2 to 10. In order to determine the nature of the balance between the penalized spline and the variogram in the absence of autocorrelated residuals, we also ran these nine models on each of the two- and three-dimensional data sets with no autocorrelated residuals. For all but one data set, we set the constant c , the upper limit of the estimated variance of the penalized spline coefficients, to 1. For the three-dimensional data set with a mean variance of 0.05 (Figure 3.2 center), we set $c = 10$. Thus, we have a total of 198 model runs, summarized and discussed below.

3.3.3 Application

In addition to the simulated data, we also apply a modified model structure to belt transect survey data on density of Atlantic sea scallops (*Placopecten magellanicus*) in Georges Bank off the coast of Massachusetts. The data arise from a series of continuous images of the sea floor. Data collection methods are described in Chang et al. (2017). The data are given as counts of individuals in a subset of the images, aggregated over 5km. Scallop density tends to peak at intermediate depths as a result of both water temperature and predation (Hart and Chute, 2004; Hart, 2006; Shank

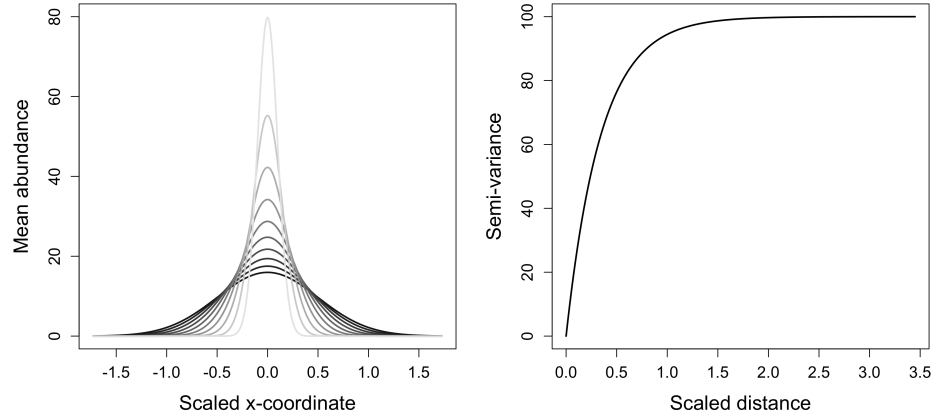


Figure 3.1: Mean generating process (left) and variogram used to generate autocorrelated residuals (right) for two-dimensional simulations. The variance σ^2 of the mean generating process decreases evenly from a value of 0.5 to 0.1 as the colors range from dark to light. We use the same variogram for all two-dimensional simulations.

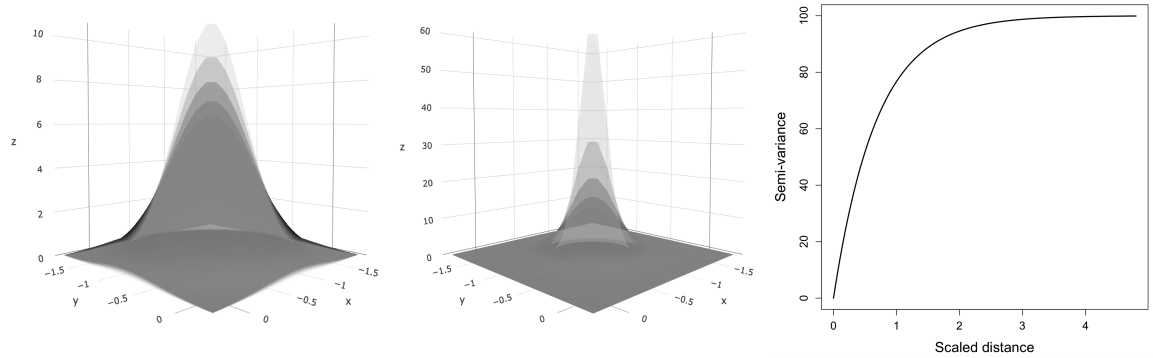


Figure 3.2: Mean generating process (left and center) and variogram used to generate autocorrelated residuals (right) for two-dimensional simulations. The first five and last five data sets are separated for visibility. Note the change in scale. The diagonals of the variance $\Sigma = \sigma^2 I$ of the mean generating process decreases evenly from a value of 0.5 to 0.3 (left) and 0.25 to 0.05 (center) as the colors range from dark to light. We use the same variogram for all three-simulations.

et al., 2012). At broad spatial scales, density may be driven by sediment type or current transport (Thouzeau et al., 1991). At fine scales, adults tend to aggregate, which may increase fertilization success (Stokesbury and Himmelman, 1993). Therefore, separate processes drive broad- and fine-scale spatial patterns, making the description of sea scallop abundance an ideal candidate for this model structure.

We use an additive model structure to combine a cubic spline regression over the relationship of abundance with depth, which is highly correlated with latitude, and a thin plate spline regression over the relationship of abundance with spatial coordinates. Thus, we use the methods above to produce two design matrices D_d and D_c , the first for depth and the second for spatial coordinates. We combine to form the design matrix

$$D = \begin{bmatrix} D_d & D_c \end{bmatrix}$$

Note that this results in a matrix that includes columns two intercepts. Without loss of generality, we remove one of these columns to model a global intercept.

Data for all regions are zero- inflated. Thus, in addition to including both a cubic and a thin plate spline, we also include a zero-inflation, or hurdle, model, in which we treat probability of presence as a second order polynomial function of latitude, as follows.

$$\pi_i = \alpha_0 + \alpha_1 y_i + \alpha_2 y_i^2$$

where π_i is the probability of a non-zero observation at the i^{th} sampling location, and y_i is the latitude of the i^{th} sampling location. We then set the following priors for the

coefficients of this function

$$\alpha_0 \sim T(1)$$

$$\alpha_1 \sim T(1)$$

$$\alpha_2 \sim T(1)$$

where $T(1)$ is the t-distribution with one degree of freedom. The peak at zero in this distribution helps to prevent debilitating autocorrelation with the other parameters of the model.

Lastly, rather than a Poisson, we use a negative binomial distribution to account for under- or over-dispersion in abundance. This results in a likelihood of the form

$$Y \sim NB(p, r \pi)$$

where r is a measure of under- or over-dispersion and where

$$p = \frac{r}{r + \mu}$$

Recall that we calculate μ as the matrix product of the design matrix D and the coefficients $\mathbf{b}$, each of which has a normal prior with mean zero, and variance σ^2 , which is itself a hyperparameter with a t-distribution with one degree of freedom and non-centrality parameter 10, truncated at zero and some upper limit c , which we set equal to one.

We vary the knots of both the cubic and the thin plate spline from two to ten, including all combinations, for a total of 81 model runs. We expect the sill, the range, and the maximum estimated count to fall as the number of knots of both penalized

splines increases.

3.3.4 Simulation Study Results

Two-dimensional

It is clear that the median curvature of the penalized spline changes as the number of knots increases (Figure 3.3). This is especially apparent in for the case in which the variance of the mean generating process is set to $\sigma^2 = 0.5$ (Figure 3.3, top left). As the number of knots increases, the curvature identifies two areas of attraction. For the remainder of the data sets, the model identifies just one central area of attraction. However, as the number of knots increases to 10, the estimate of mean maximum count also increases, up to nearly twice that estimated by a model with just 2 knots. For example, in the simulated data set with variance set to $\sigma^2 = 0.01$ (Figure 3.3, bottom left), the estimated maximum rises from about 50 to nearly 100. Simultaneously, the median estimate of the sill of the variogram decreases from 700 to less than 400 (Figure 3.4). Similarly, the range decreases from a median estimate of about 0.11 to 0.05 (Figure 3.5). This suggests a close correspondence between an increase in the curvature, expressed as the maximum of the central peak, and a decrease in both the range and sill of the variogram, as expected.

We can see this phenomenon more clearly if we examine peakedness as a function of normalized mean maximum count estimates relative to normalized sill and range (Figure 3.6). When we run simple linear models on the normalized median estimates of the sill and the range as a function of the normalized mean maximum count, we find that, on average, a 100% increase in the mean maximum count results in a 63% decrease in the sill and an 85% decrease in the range (Table 3.1). In general, then, the curvature of the penalized spline increases more quickly with an increase in the

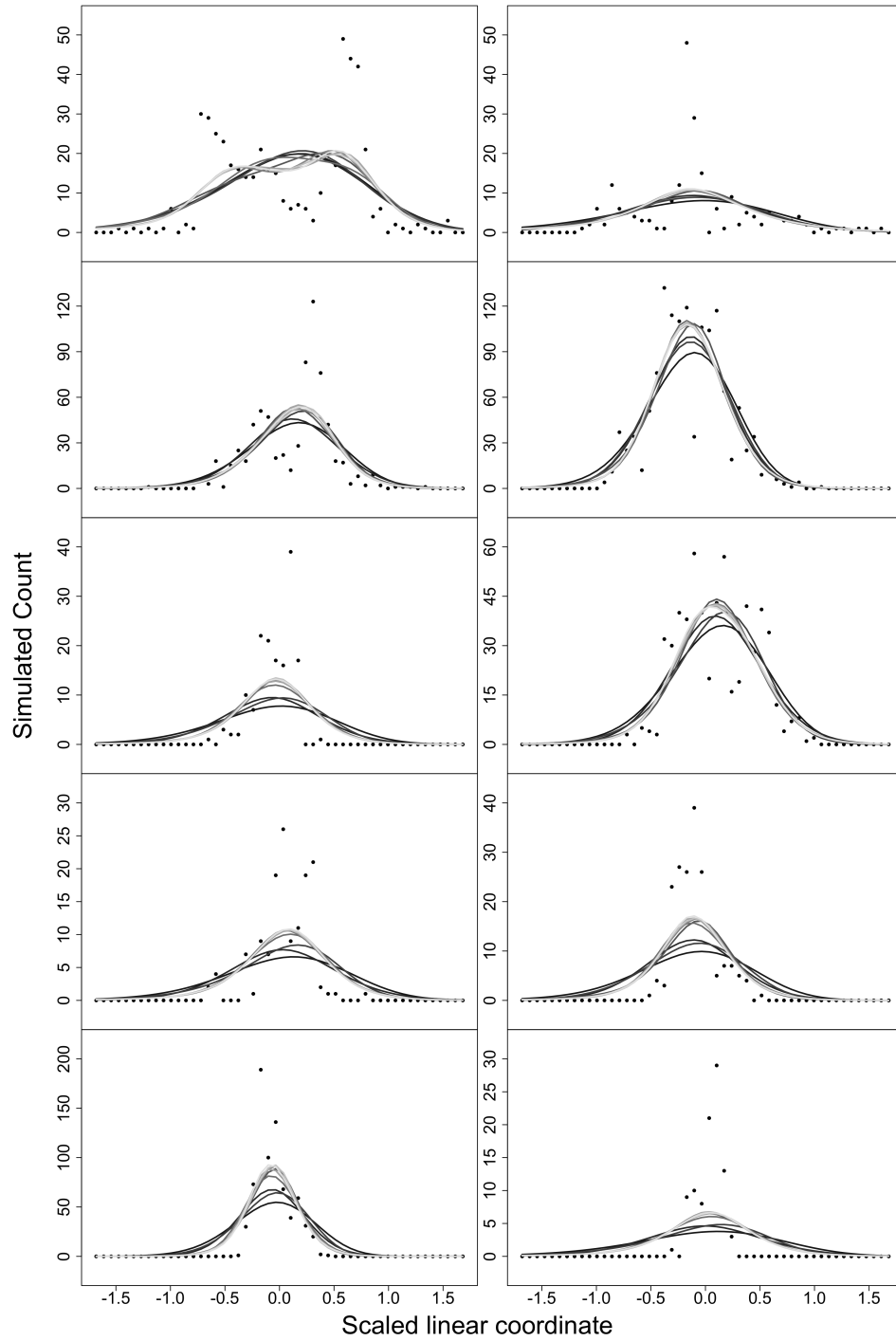


Figure 3.3: The median of all nine penalized spline model fits to all ten two-dimensional data sets. From left to right, and top to bottom, the variance σ^2 of the mean generating process decreases from 0.5 to 0.1. Observations are shown as points. The median of the penalized spline fits are solid gray lines, with the number of knots ranging from 2 to 10 as colors range from dark to light.

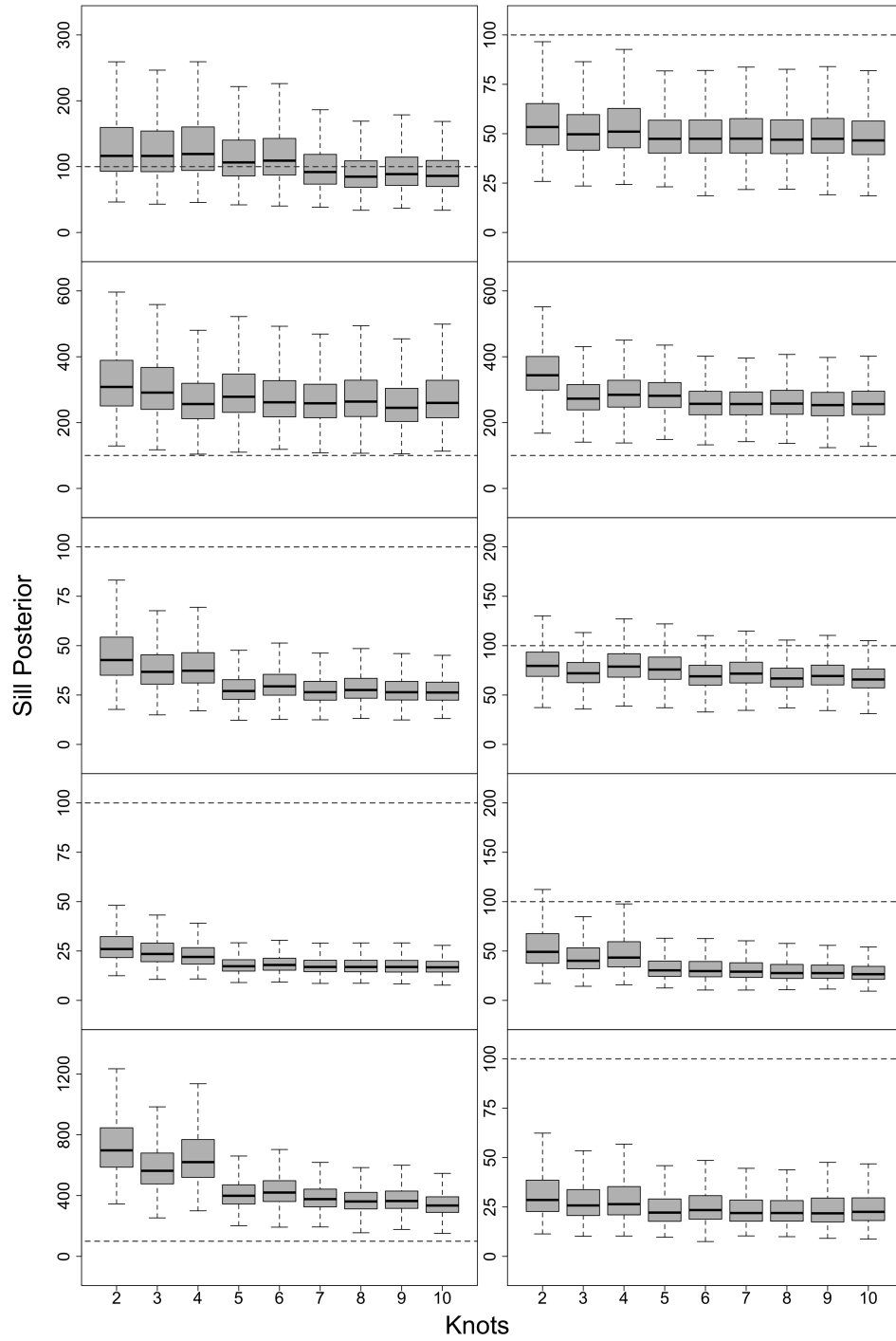


Figure 3.4: The posterior distribution of the sill of the variogram for all nine models applied to all ten data sets. From left to right, and top to bottom, the variance σ^2 of the mean generating process decreases evenly from 0.5 to 0.1. Broken horizontal lines give the actual sill value of 100 used to generate all ten data sets. It is only shown in plots in which the scale of the posterior of the sill is at or near 100.

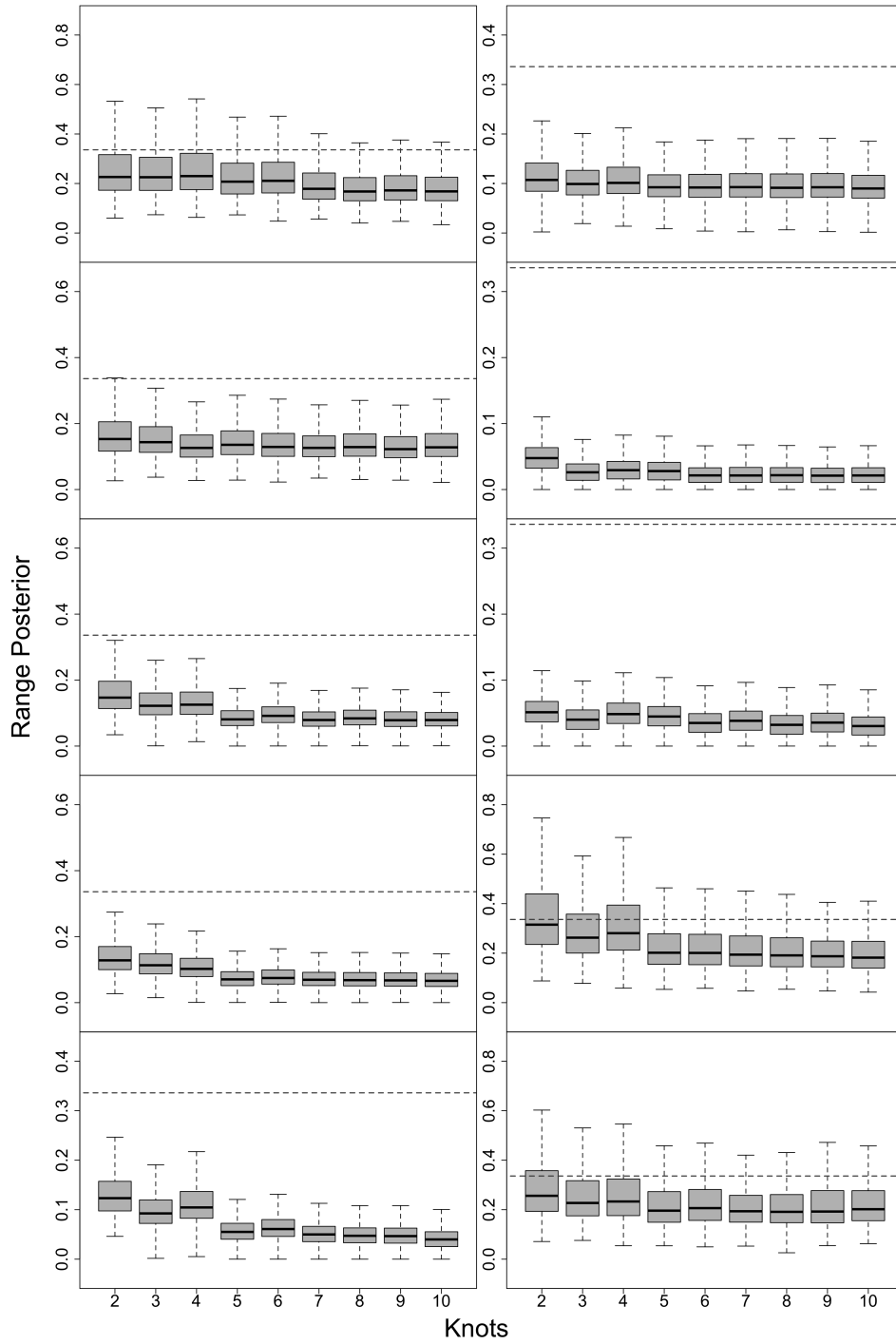


Figure 3.5: The posterior distribution of the range of the variogram for all nine models applied to all ten two-dimensional data sets. From left to right, and top to bottom, the variance σ^2 of the mean generating process decreases evenly from 0.5 to 0.1. Broken horizontal lines give the actual range value of 0.34 used to generate all ten data sets.

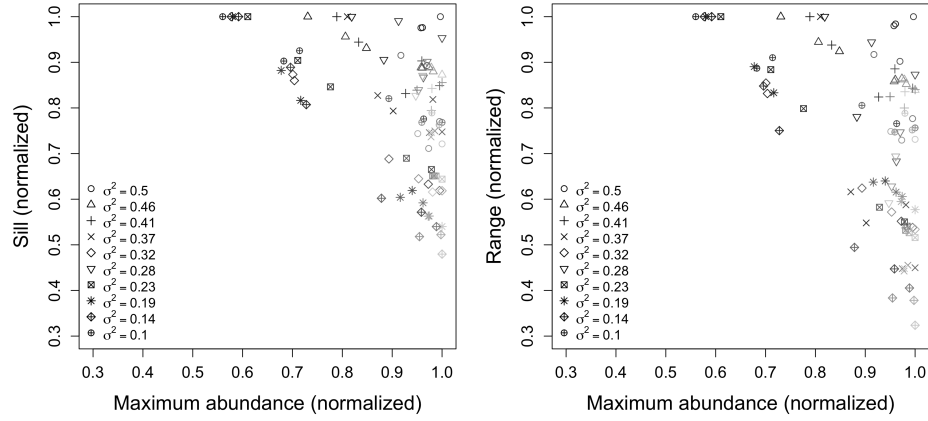


Figure 3.6: Normalized estimates of the sill (left) and range (right) as a function of normalized estimates of maximum count or peakedness for all two-dimensional data sets. The number in the legend indicates the variance of the mean generating process. The colors of the symbols range from dark to light as knots increase, identical to those used in other figures.

number of knots than the sill decreases. In turn, the sill decreases more quickly than the range. This is not entirely consistent across all data sets. For example, when we run simple linear models on the normalized estimated range and sill as a function of the mean maximum count on the results from each data set individually, we find that, while the trends are negative, there is no significant relationship for the data sets generated with variance $\sigma^2 = 0.5$ and $\sigma^2 = 0.28$ (Figure 3.3, top left and center right; Figure 3.6; Table 3.1). However, all relationships between the median estimate of the sill and range, and the mean maximum count, are negative.

In addition to the increase in the curvature, and the decrease in the sill and the range, there is a clear phenomenon in which the sill and the range decrease, and then level out (Figures 3.4 and 3.5). In Figure 3.6, this presents as a clustering of light-colored points in the bottom right. For the curvature of the penalized spline, this is also somewhat clear in Figure 3.3, in which light-colored lines cluster at higher peaks. This is most clear in Figure 3.7. This suggests that, as the number of knots increases,

Two-dimensional				
σ^2	Sill	P	Range	P
0.50	-1.15	0.508	-1.10	0.514
0.46	-0.45	<0.001*	-0.56	<0.001*
0.41	-0.73	0.003*	-0.77	0.002*
0.37	-1.12	0.002*	-2.41	0.002*
0.32	-0.90	<0.001*	-1.10	<0.001*
0.28	-0.58	0.163	-1.43	0.121
0.23	-0.92	<0.001*	-1.26	<0.001*
0.19	-1.06	<0.001*	-0.98	<0.001*
0.14	-1.22	<0.001*	-1.56	<0.001*
0.10	-0.52	<0.001*	-0.53	<0.001*
All	-0.63	<0.001*	-0.85	<0.001*

Three-dimensional				
σ^2	Sill	P	Range	P
0.50	0.02	0.886	0.27	0.429
0.45	0.08	0.268	0.27	<0.001*
0.40	-1.55	<0.001*	-0.25	0.002*
0.35	-1.11	0.005*	-1.79	0.001*
0.30	0.11	0.123	0.51	0.005*
0.25	-0.56	<0.001*	-0.07	0.739
0.20	-1.39	<0.001*	0.35	0.090*
0.15	2.86	<0.001*	3.96	0.017*
0.10	1.33	<0.001*	3.16	0.231
0.05	-1.26	<0.001*	-0.23	<0.001*
All	-0.13	0.297	-0.50	0.023*

Table 3.1: Slope and p-values of the linear relationship between the normalized median estimates of the sill and range as a function of the normalized mean maximum count. Asterisks by p-values indicate significance with $\alpha = 0.05$. The rows marked "All" represent all data combined.

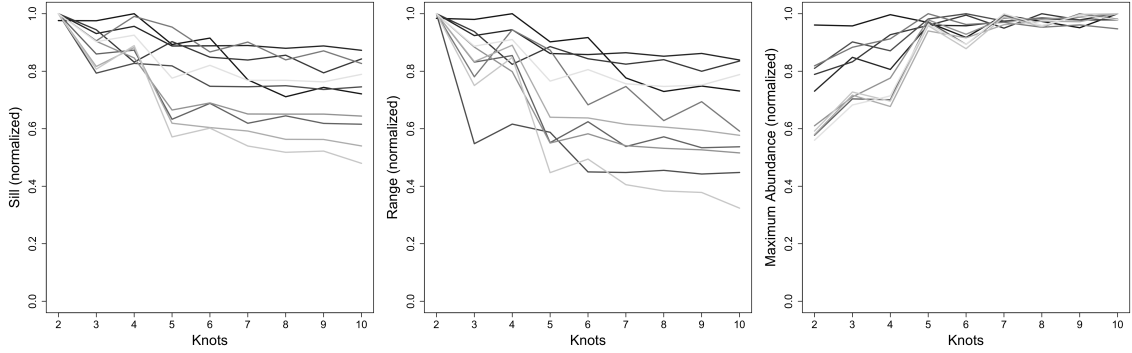


Figure 3.7: Median estimates of the sill (left), the range (center), and the maximum count (right) as a function of the number of knots of the penalized spline applied to the two-dimensional data sets. Colors range from dark to light as the variance of the mean generating process decreases.

the relationship between the curvature of penalized spline and the parameters of the variogram stabilizes. The sill and the range tend to stabilize at lower numbers, relative to the maximum, as the variance of the mean generating process decreases. Similarly, the maximum count tends to stabilize at a higher number, relative to the maximum, as the variance decreases. Note that, in general, the model tends to over- or under-estimate the sill of the variogram, and consistently under-estimates the range.

We calculate the deviance information criterion (DIC) of all two-dimensional model runs, and note the minimal DIC. The DIC, while suffering from both practical and theoretical troubles – including lack of consistency and a reportedly weak theoretical basis – is still in wide use throughout the biological and ecological sciences (Spiegelhalter et al., 2014). We use it here just as others might use it, to choose the optimal number of knots. For the two-dimensional model runs, we find that the DIC tends to fall as the number of knots increases, mirroring the fall in the sill and the range (Figure 3.8, left). While repeated model runs show that the DIC is somewhat inconsistent, it typically does not differ by more than 5 points between each run. Therefore, we consider differences between two model runs of less than 5 to be insignificant, and err

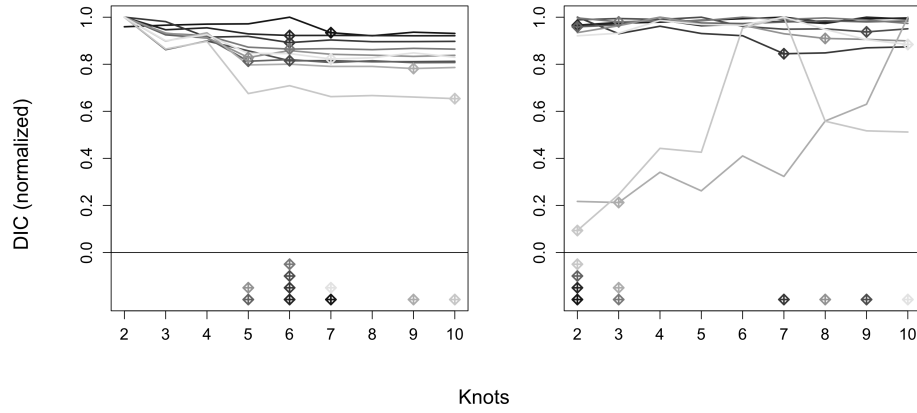


Figure 3.8: Normalized deviance information criterion as a function of the number of knots of the penalized spline applied to the two-dimensional data sets. Colors range from dark to light as the variance of the mean generating process decreases. Symbols on each line mark the model with the minimal DIC. These symbols are repeated below the horizontal line for additional readability. Note that, in order for the DIC of one model to be considered truly less than another, the absolute difference must be at least 5.

on the side of simplicity by choosing the model with the lower number of knots. When we apply this procedure to model runs on all ten two-dimensional data sets, we find the optimal number of knots typically falls between 5 and 7. Only two models have an optimal number of knots greater than 7, both of which arise from a relatively low variance of the mean generating process ($\sigma^2 = 0.19$ and $\sigma^2 = 0.14$, see Figure 3.1).

3.3.5 Three-dimensional

The median curvature of the penalized spline appears to have a more complicated relationship with the number of knots in the three-dimensional case (Figure 3.9). In some cases, there is a clear rise in peakedness or maximum count as the number of knots increases (Figure 3.9, bottom right). In others, median curvature appears to hold steady (Figure 3.9, top left). The relationship between the number of knots and the sill, and the number of knots and the range, is similarly complicated (Figures 3.10

and 3.11). In some cases, the sill does decline, in some it increases, and in others it remains the same. There is no clear pattern as the variance of the mean generating process increases. The estimates of the range are similar, with some peculiar behavior for variance $\sigma^2 = 0.15$ and $\sigma^2 = 0.1$ (Figure 3.11) in which the range increases precipitously. The median estimates of the range rise to many orders of magnitude higher than the entire range covered by the data sets. This would clearly have no biological meaning, and may indicate that the model is attempting to fit outliers.

When we examine the normalized mean maximum count estimates relative to normalized sill and range, we find that, while there is some similarity to the pattern in the two-dimensional case (see Figure 3.7), there are gross exceptions. When we examine Figure 3.12, we see similar behavior to that seen in Figure 3.6, wherein lighter colors cluster at the bottom right, indicating decreases in the sill and range relative to increases in the mean maximum count. However, the results are not as clear. When we run simple linear models on the normalized median estimates of the sill and range as a function of the normalized mean maximum count, we find that, on average, there is a 12% decrease in the sill and a 50% decrease in the range. Only the latter of these is significant (Table 3.1). However, when we run these models on the results from each data set individually, half show significant decreases in the sill, ranging from 56% to 155% relative to a 100% decrease in the mean maximum count. Two show significant increases. Three show significant decreases in the normalized estimate of the range, ranging from 23% to 179%. Four show significant *increases* in the range, ranging from 27% to 396%. There is no apparent pattern in those data sets that show increases versus decreases in the sill or the range. Those sets that show increases in the sill tend also to show decreases in the range, with just one exception (Table 3.1, Three-dimensional, $\sigma^2 = 0.20$).

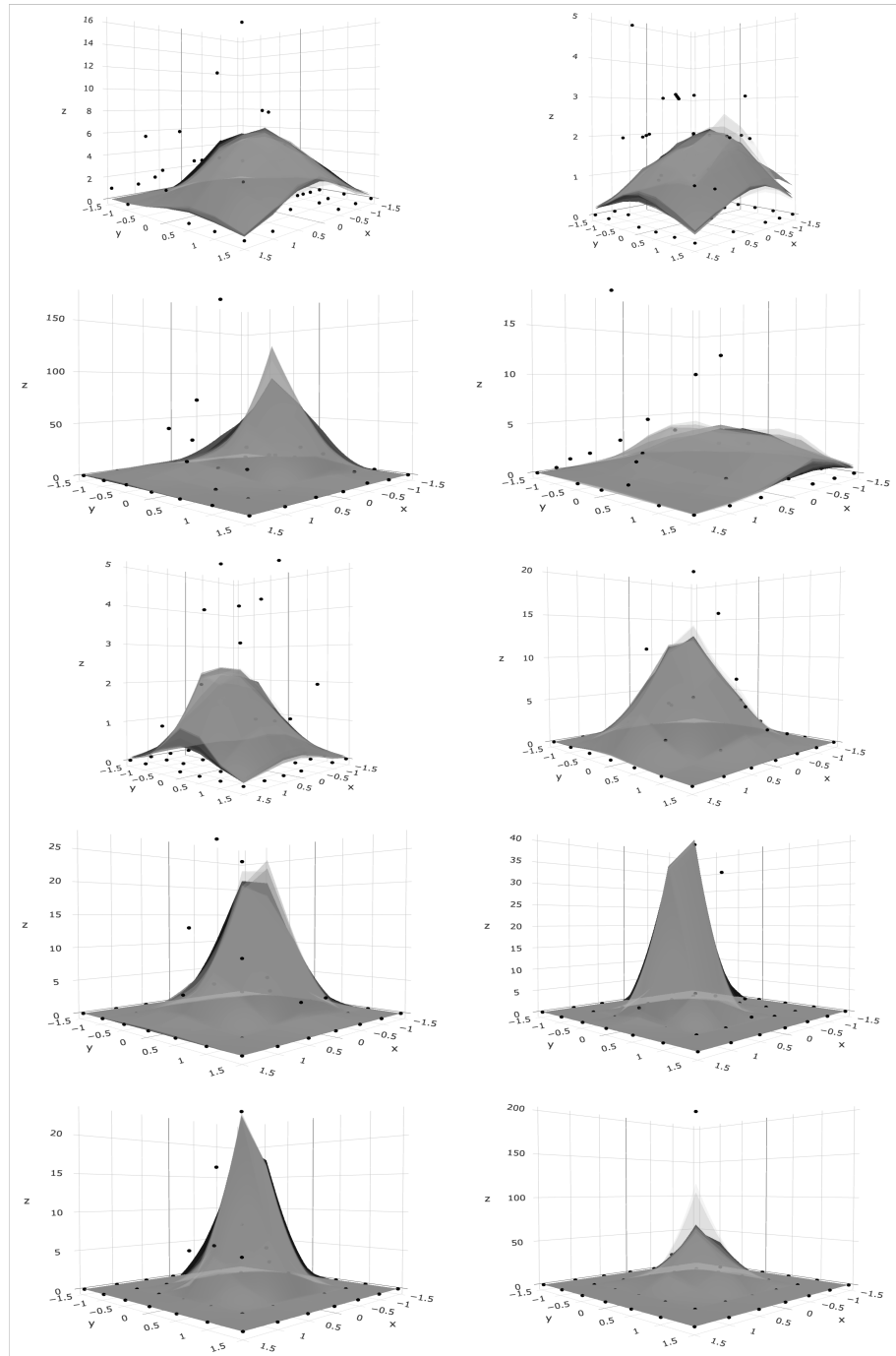


Figure 3.9: The median of all nine penalized spline model fits to all ten three-dimensional data sets. From left to right, and top to bottom, the variance σ^2 of the mean generating process decreases from 0.5 to 0.05. Observations are shown as points. The median of the penalized spline fits are transparent gray surfaces, with the number of knots ranging from 2 to 10 as colors range from dark to light.

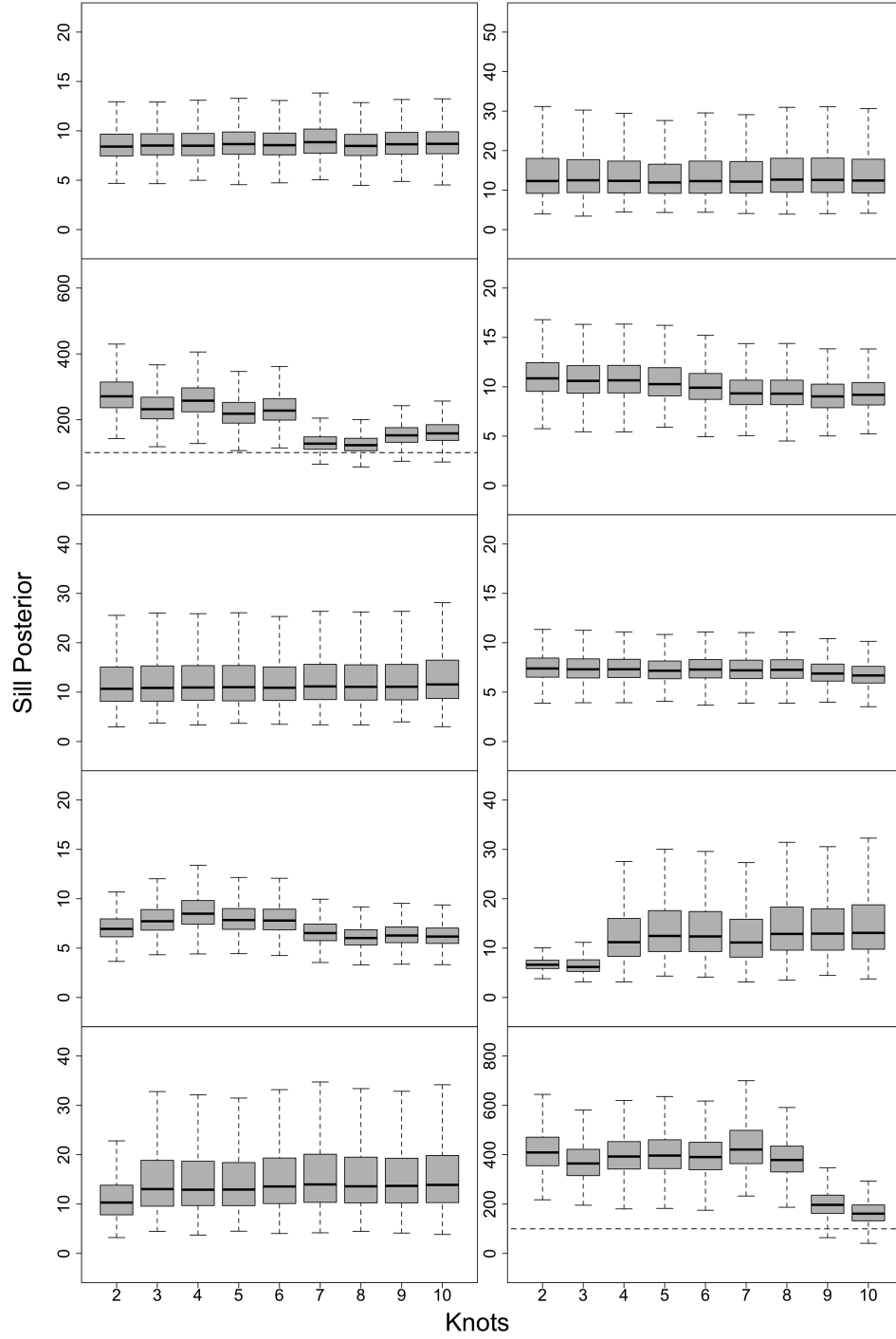


Figure 3.10: The posterior distribution of the sill of the variogram for all nine models applied to all ten three-dimensional data sets. From left to right, and top to bottom, the variance σ^2 of the mean generating process decreases evenly from 0.5 to 0.05. Broken horizontal lines give the actual sill value of 100 used to generate all ten data sets. It is only shown in plots in which the scale of the posterior of the sill is at or near 100.

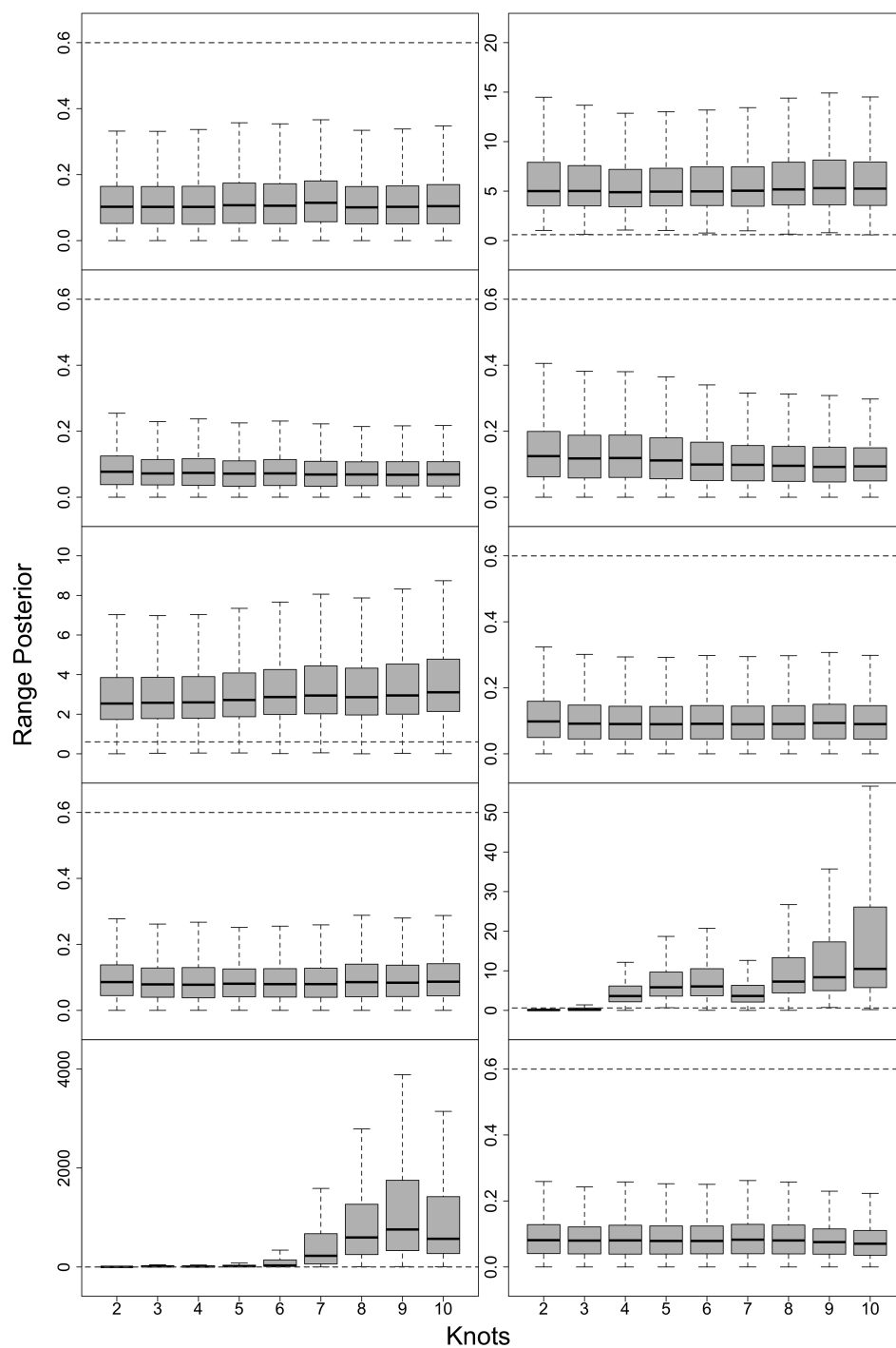


Figure 3.11: The posterior distribution of the range of the variogram for all nine models applied to all ten three-dimensional data sets. From left to right, and top to bottom, the variance σ^2 of the mean generating process decreases evenly from 0.5 to 0.05. Broken horizontal lines give the actual range value of 0.6 used to generate all ten data sets.

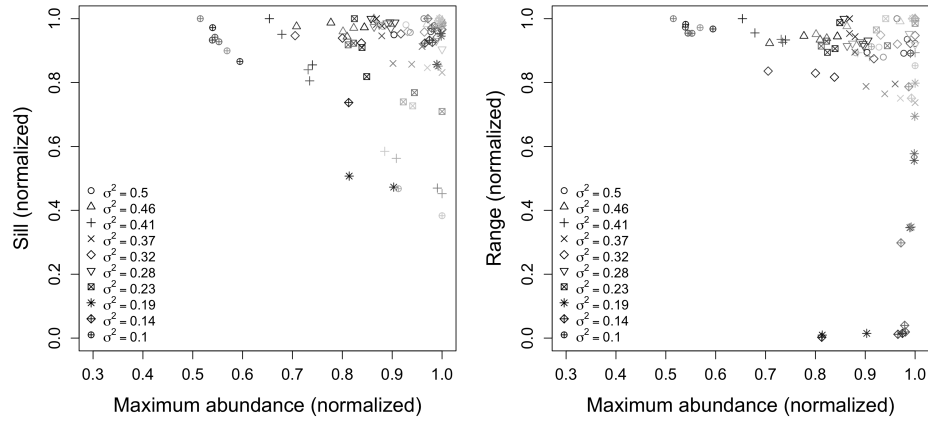


Figure 3.12: Normalized estimates of the sill (left) and range (right) as a function of normalized estimates of maximum count for all three-dimensional data sets. The number in the legend indicates the variance of the mean generating process. The colors of the symbols range from dark to light as knots increase, identical to those used in other figures.

There is moderate evidence that, as the number of knots increases, the relationship between the curvature of the penalized spline and the parameters of the variogram stabilizes (Figure 3.13). This is especially apparent in the median estimates of the sill (Figure 3.13, left). A few data sets show increases in the sill, that then stabilize near their maximum. The phenomenon in which the mean maximum count increases, and then stabilizes, does appear to hold steady for the three-dimensional case, despite the relatively erratic behavior of the variogram. While inconsistent, there certainly *is* a relationship between the curvature of the penalized spline and the parameters of the variogram. In general, the data we simulate in the three-dimensional case are less smooth than the two-dimensional case. This may be due in part to the fact that each point in the three-dimensional space is influenced by neighbors in the both the x- and the y-direction. Whatever the cause, the data itself may be the source of the unpredictable relationship between the spline and the variogram.

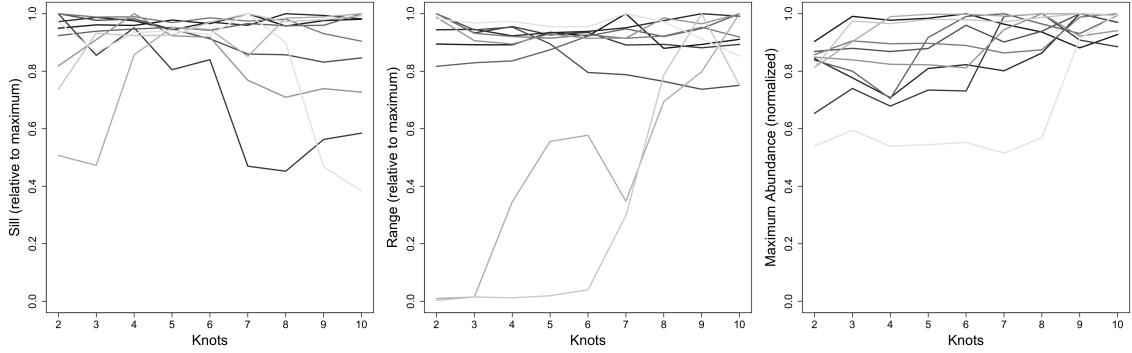


Figure 3.13: Median estimates of the sill (left), the range (center), and the maximum count (right) as a function of the number of knots of the penalized spline applied to the three-dimensional data sets. Colors range from dark to light as the variance of the mean generating process decreases.

We also calculate the deviance information criterion (DIC) of all three-dimensional model runs, and note the minimal DIC. As above, we consider differences between two model runs of less than 5 to be insignificant, and choose the model with the lower number of knots. The optimal number of knots for six of the three-dimensional simulated data sets is just 2 or 3 (Figure 3.8). The other four fall at 7 knots or above, with no consistency related to the variance of the mean generating process.

3.3.6 No autocorrelation

When we apply the model structure to the data sets without autocorrelation, we find that the results for the two-dimensional data set are reminiscent of the results applied to the data sets with autocorrelation, with a decline in the sill and the range, and an increase in the maximum count (Figure 3.14, left, and see Figures 3.3, 3.4, 3.5, and 3.7). However, the decline is precipitous, and the stability of the balance between the penalized spline and the parameters of the variogram is marked. By "stability", we mean that, as the number of knots increases, the parameters of the penalized spline and of the variogram reach a point past which any changes with the addition of more

knots is insignificant. These parameters settle at very low values, suggesting very little autocorrelation at a very narrow range. This is highly suggestive that this model fit is simply an alternative expression of the variance of the solution at each point.

By contrast, the model structure applied to the two-dimensional data set does not decline. Rather, it starts and remains at one level (Figure 3.14, right). The sill is low, but the range is relatively high, covering the range of the entire data set. This is also suggestive that this is simply an expression of the variance. In both cases, the results confirm that, when applied to a data set with no structured autocorrelation, the model will identify little to no autocorrelation.

3.3.7 Application

Visually, it is simplest to examine the change in the median estimated counts with depth and with two-dimensional space, separately. We do this by splitting the design matrix D into its component pieces, D_d and D_c , and multiplying by the associated coefficients to arrive at two respective components of the estimate, $\hat{\mu}_d$ and $\hat{\mu}_c$. The final estimate is then of course, $\hat{\mu} = \hat{\mu}_d \hat{\mu}_c$. In all cases, the median estimate as a function of depth alone falls and then rises (Figure 3.15). This relationship is fairly consistent across all choices for the number of knots of the thin plate spline. While peakedness increases, it does not do so consistently across the number of knots of the cubic spline. The estimated maximum count instead more or less oscillates around a central value.

The median estimate as a function of two-dimensional space only rises in some cases, and falls in others (Figure 3.16). While the relationship of count with depth is relatively similar across all models, the relationship of count with two-dimensional space varies widely across choices of the knots for the number of knots of the cubic spline.

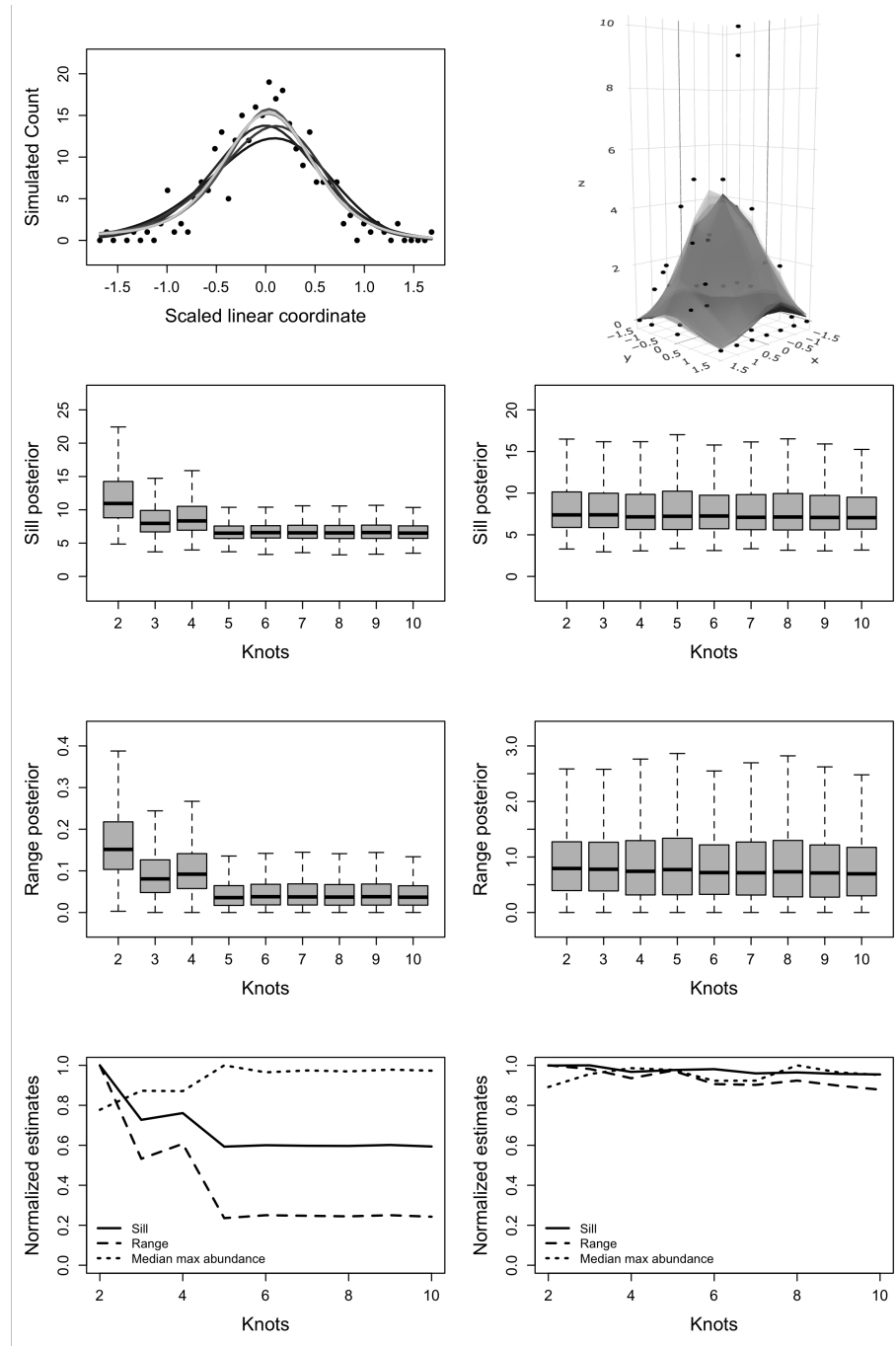


Figure 3.14: All output from the model structure applied to two-dimensional (all left) and three-dimensional (all right) data sets with no spatial autocorrelation.

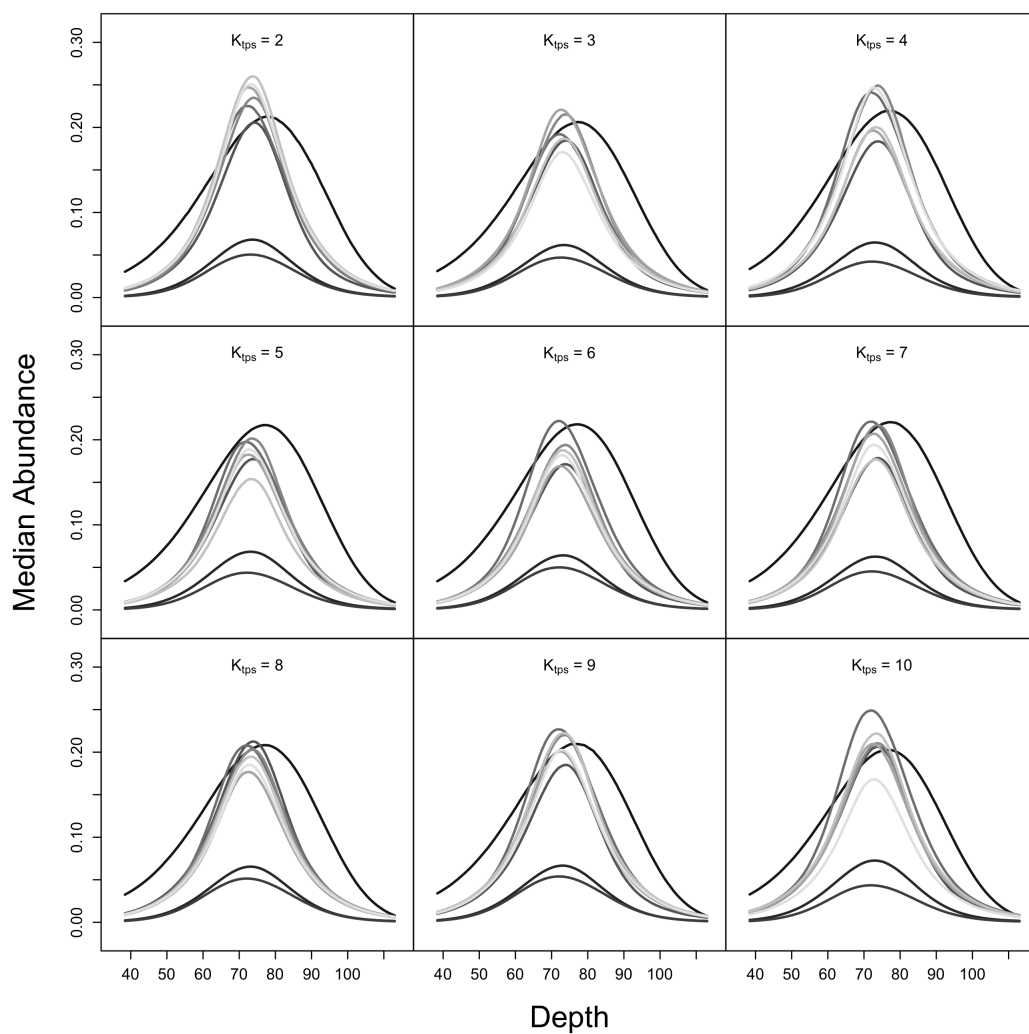


Figure 3.15: Median posterior estimate of scallop count as a function of depth. The value of K_{tps} listed above each of the nine plots gives the number of knots of the thin plate spline. Here, colors range from dark to light within each plot as the number of knots of the cubic spline ranges from two to ten.

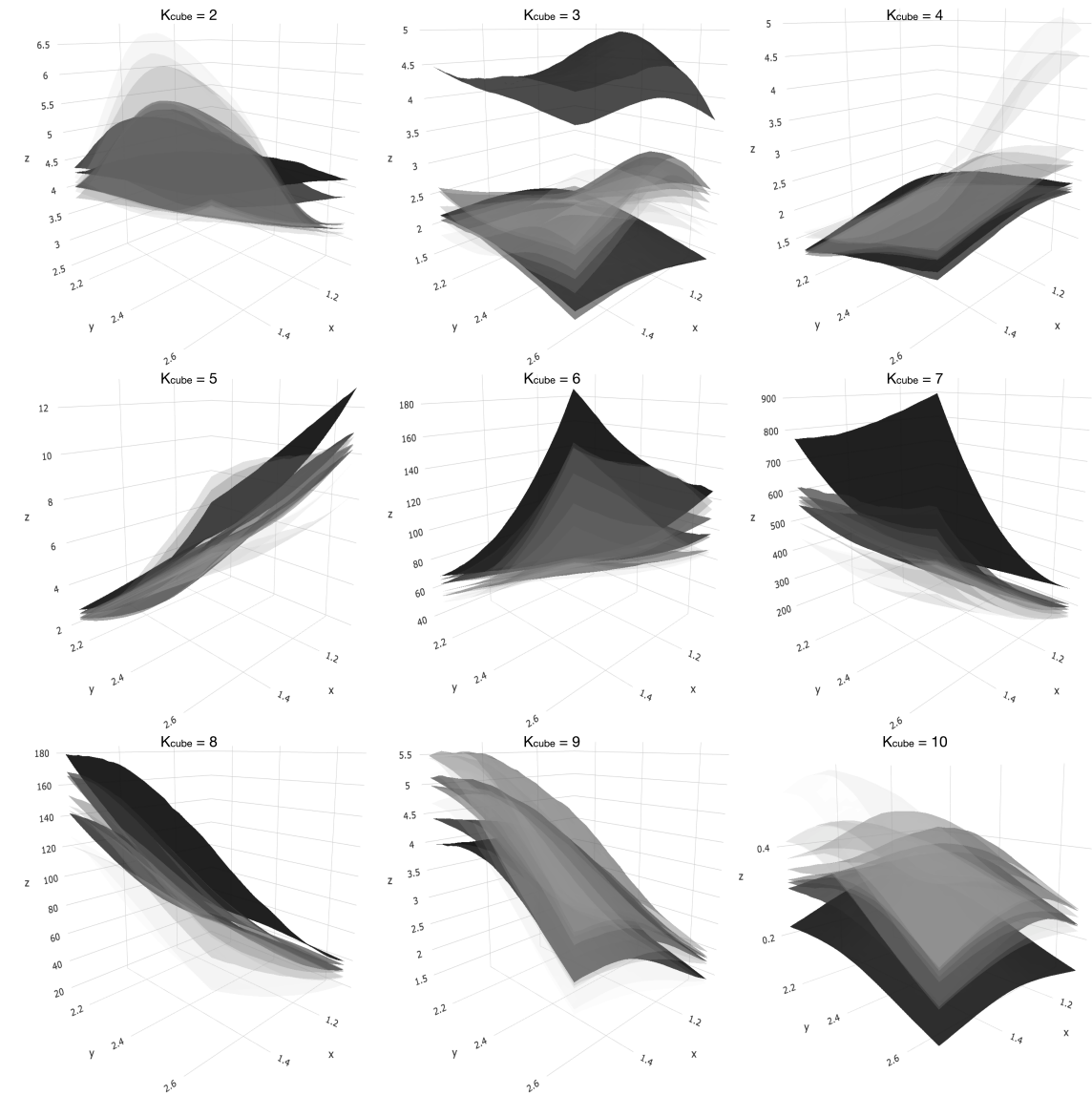


Figure 3.16: Median posterior estimate of scallop count as a function of depth. The value of K_{cube} listed above each of the nine plots gives the number of knots of the cubic spline. Here, colors and opacity range from dark and high to light and low within each plot as the number of knots of the thin plate spline ranges from two to ten.

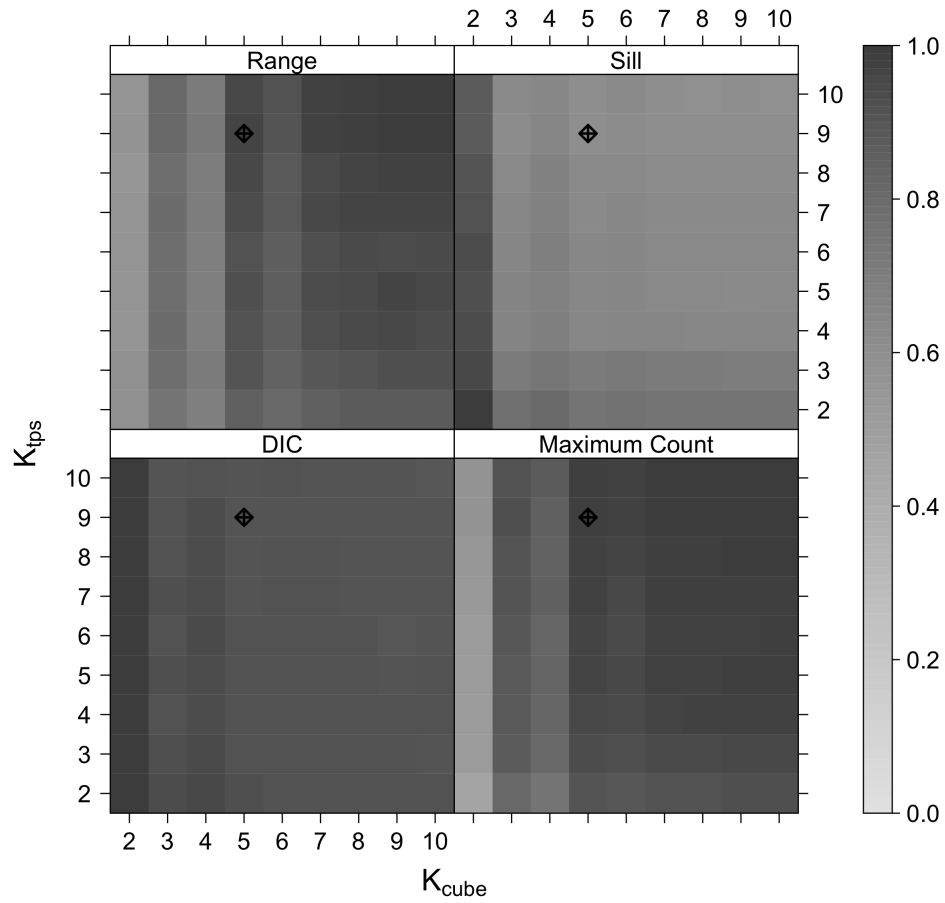


Figure 3.17: Color level plots of the normalized median estimates of sill, range, and maximum count – which we interpret as a measure of peakedness – and of DIC as a function of the number of knots the cubic spline (K_{cube}) and of the thin plate spline (K_{tps}). The diamond symbol with crosshairs marks the optimal model as chosen by the DIC.

The sill of the variogram tends to fall by about half with an increase of knots in both the cubic and the thin place spline (Figure 3.17). This is expected. The range, however, tends to rise, and much more so with the number of knots of the cubic spline, rather than the thin plate spline. That is, the relationship with depth appears to have more power over the range than the relationship with two-dimensional space.

According to the DIC, the optimal model is one with 5 knots for the cubic spline and 9 knots for the thin plate spline (see Figure 3.17). The DIC also appears to be affected by the number of knots of the cubic spline – or the relationship of count with depth – more so than number of knots of the thin plate spline. This is also true of the estimated maximum count, which we treat as a summary of peakedness in the simulation study. We note here that, in the case of the cubic spline, the maximum count does not appear to be quite so correlated with peakedness. However, it does appear to be correlated with peakedness in the case of the thin plate spline, which reflects in the estimated maximum count in relation to the number of knots of the thin plate spline (see Figure 3.17, bottom right).

The model for probability of presence does not vary much with knot choice, with probability increasing from a mean of about 0.4 to 0.8 as latitude increases from one extreme of the region to the other. The dispersion factor, however, does increase with the number of knots of both the cubic and the thin plate spline, from a value of about 0.6 to about 0.8 (not shown). This indicates a lessening degree of aggregation inherent in the distribution of the scallops as the number of knots increases which is, in turn, attributed to other factors, such as variability with depth, two-dimensional space, and fine-scale aggregation.

3.3.8 Fixed and Random Effects

As noted above, one major difference between the penalized spline and the variogram is the representation of variance. The variance of the penalized spline is representative of deviation from a fixed mean and varies only with the magnitude of the mean, whereas the variance of the kriged residuals are realizations of a random process occurring at each location. The variance of the mean estimated from the penalized spline is

equivalent to the variance of the kriged residuals at the locations of the observations, which is easily derivable from the fact that the residuals at the observation locations are *known*. Therefore, given that

$$\text{Var}(\hat{\mu} + \hat{\epsilon}) = \text{Var}(\hat{\mu}) + \text{Var}(\hat{\epsilon}) - 2 \text{Cov}(\hat{\mu}, \hat{\epsilon})$$

where here we call the mean estimate $\hat{\mu}$ and the residual estimate $\hat{\epsilon}$, the variance of the predictions at the observation locations is zero. However, when we predict at new locations, the variance is an increasing function of the distance from nearby observations, and the magnitude of the variance is a function of the magnitude of the residuals. What constitutes as "nearby" is dependent upon the structure of the data, and of the estimates of the range and sill of the variogram.

This phenomenon is most easily visible in the two-dimensional results. We use the results from the model with two knots applied to the two-dimensional data set with a mean generating process with a variance of $\sigma^2 = 0.5$ (see Figure 3.1, left) to predict counts at one thousand new locations (Figure 3.18, left). These locations are spaced evenly over the entire spatial range of the observations. It is clear that, while the magnitude of the variance of the predicted mean arising from the spline model and the predicted residuals arising from the kriging model is similar, there is fine-scale variability present in the variance of the predicted residuals that is not present in the variance of the predicted mean (Figure 3.18, center). The variance of the sum of the predicted mean and the predicted residuals rises as the distance from the nearest observation increases (Figure 3.18, right). The variance peaks exactly halfway between two adjacent observations, and the number at which it peaks is clearly a function of the magnitude of the residuals.

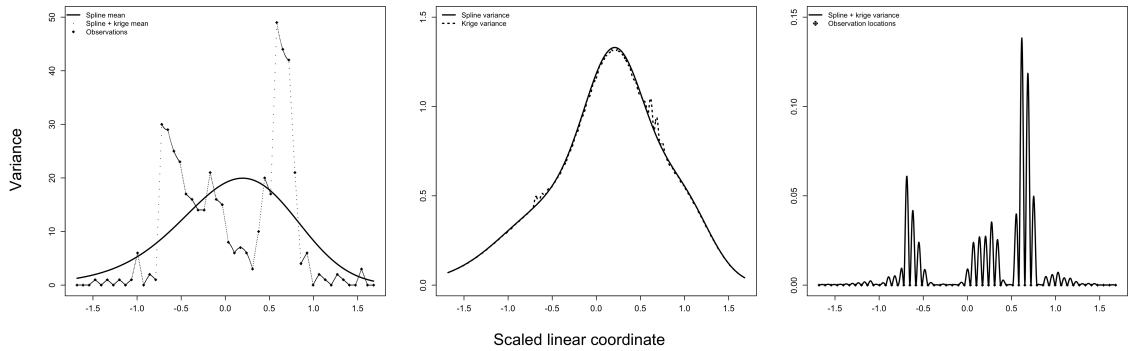


Figure 3.18: Mean predictions (left) and prediction variance of the mean counts arising from the spline model and the residuals arising from the krige model (center) and of the sum of the counts arising from both models (right) at one thousand new locations. Locations are spaced evenly across the entire range of the two-dimensional data sets. Predictions arise from a model with two knots applied to the two dimensional data set with a mean generating process with a variance of $\sigma^2 = 0.5$.

3.4 Discussion

Applying a penalized spline to a spatial process and kriging the residuals is a method by which we can describe both broad and fine-scale trends in organism abundance. It has a recent precedent grounded in several decades of spatial modeling and research. It is clear from the results presented above that the model is oftentimes capable of distinguishing between broad-scale and fine-scale trends. In the two-dimensional case, the model consistently reaches a stable balance between the fit of the penalized spline and the parameters of the variogram, however biased it may be relative to the processes we use to generate the data. This, after a rather precipitous decline in the estimates of the sill and the range. The curvature inherent in the data source, defined numerically as the variance of the mean generating function, and practically as peakedness or maximum count, seems to affect this outcome. As the variance decreases, and thus as the mean maximum count increases, the drop in the sill and the range as the number of knots increases is generally more severe. This suggests that it is difficult for the model to distinguish between the mean generating process and autocorrelation as the

source of high abundance peaks.

The knots, then, can serve as a prior belief concerning the nature of the data generating process. If we believe that spatial patterns of abundance are driven primarily by some broad-scale spatial variable, then we might choose a larger number of knots, say 10 or more. If we believe that peaks in abundance are driven by finer scale, aggregative processes, we can limit the number of knots. It seems clear that, even if we manipulate the number of knots, the model is capable of determining whether or not there is fine-scale spatial autocorrelation in the data, separate from the broad-scale trend. This suggests that we might use this model to test the hypothesis that there is indeed fine-scale aggregation. However, given that the estimates of the sill and the range of the variogram do fall, often by a large percentage, when we increase the number of knots, we must be cautious in our interpretation of the output.

In general, the model tends to over- or under-estimate the actual value of the sill we use to generate the data, and to consistently under-estimate the value of the range. Whether the model under- or over-estimates the sill is inconsistent, and therefore is likely based both in the stochasticity built into the data generating process, as well as in the nature of the contest between the penalized spline and the variogram. That it consistently under-estimates the range suggests the model attributes more fine-scale variability to the mean generating process – and thus to the penalized spline – than we actually attribute during the generation process. In this case, then, the flexibility of the penalized spline artificially diminishes the scale of the fine-scale autocorrelation. If we have prior information about the scale of aggregation, we might combat this by fixing the range, or by assigning the range an informative prior distribution. This distribution may arise from prior studies on the observed range of aggregation in the study area. Given that the range and sill of the variogram fall in concert, this may also

prevent the choice of the number of knots from deflating the magnitude of the spatial autocorrelation.

In the three-dimensional case, there is less consistency, but still some evidence that the model is capable of reaching a stable relationship between the penalized spline and the parameters of the variogram. This relationship is complicated by the fact that the consistent decrease in variance of the mean generating process does not produce a consistent increase the observed maximum count. This is likely due in large part to the addition of a third dimension, which increases the number of quantities having an influence on any given point. This produces some very high peaks in the simulated counts, which are likely the source of the peculiar behavior of the variogram with respect to those data sets constructed with a mean generating process having a very low variance. This may be the case in which the model benefits most from a fixed or informative range. Otherwise, left unfettered, the model may go to literal extremes to fit the observations.

Outside of those cases with absurd ranges an order of magnitude or more greater than the entire range of the data set, the thin plate spline model also tends to consistently under-estimate the range, as well as the sill. Even in the absurd cases, the DIC suggests a model with just 2 or 3 knots is optimal (Figure 3.8, right), where the estimates of the range are less egregious, if not entirely reasonable. It is of note that the pattern of change in the DIC for the three-dimensional models follows the pattern of change in the range much more closely than the pattern of change in the sill (see Figure 3.13, left and center). When we examine the components of the DIC, we find that the increase in the range mirrors an increase in the number of effective parameters, whereas the deviance tends to fall. In this sense, the DIC is fairly robust to absurd model results.

While these results are promising, the trouble with the initial decline in the parameters of the variogram remains. If we use this model to test a hypothesis concerning the range and magnitude of spatial autocorrelation, we may easily underestimate one or both parameters, unless we restrict the number of knots. If we deliberately choose a higher number of knots, we must recognize that this choice is potentially a strong source of prior information. At best, this choice structures a hypothesis, assuming that the curvature follows a given pattern, and seeking the resulting range and magnitude of the autocorrelation. At worst, it is guided by a simple desire to see the mean curve rise in prominence. We can defer to the DIC or some other information criterion in cases where there is only very weak or very little prior information. In cases for which we do have some prior information, we can supplement, or even overrule, the strictly analytical choice. We can also note the change in the parameters of the penalized spline in the variogram, and choose a number of knots past which there is no longer a significant change, "significant" being perhaps more than a 5% change or more. Given the results above, and these insights, we arrive at the following principle.

Principle 3: The outcome of applying a penalized spline and kriging the residuals is highly dependent upon the choice of the number of knots.

This is counter to some conventional wisdom concerning the use of penalized splines, which suggests that the number of knots is relatively inconsequential beyond some minimum, outside of modest increases in error and computation time as the number of knots increases (Ruppert, 2002). This may be appropriate for penalized spline in isolation. We suggest that it may be best to run the model with several different choices for the number of knots, and to investigate the outcome of the variogram fit. This, not to arrive at some optimal fit by trial and error, but to investigate the nature of the competition between the penalized spline and the variogram for one's particular data

set. Given that, even with a low number of knots, the model tends to underestimate the range, perhaps the most biologically interpretable of the variogram parameters, it is clear that the penalized spline and the variogram are, in fact, in competition. Absent strong prior knowledge, it may be difficult, if not impossible, to distinguish between overfitting of the penalized spline, and an actual biological or ecological process. If we choose to increase the number of knots, we are implicitly assuming that the noise we perceive is due to the mean process. Whether intentionally or not, the number of knots becomes a fairly strong piece of prior information for a Bayesian model. In the case of the data set with $\sigma^2 = 0.5$ of the two-dimensional simulations (see Figure 3.3, top left), we shift from a uni-modal to a bi-modal distribution. Without a good ecological understanding of the reason for this distribution, we may mistakenly identify two regions of attraction, which are in fact a result of a strong autocorrelative – or, rather, aggregative - process masking the fact that the strongest region of attraction lies somewhere between these two regions. In general, without a good understanding of the data source, the balance between the mean process and the degree of autocorrelation may not be in any sense mechanistic.

That said, it does appear that the model can provide convincing evidence that there is a separate, fine-scale autocorrelative process to begin with, or lack thereof (see Figure 3.14). The range and magnitude may easily be mis-identified. However, we do not want to undervalue evidence that this separation of scale exists. This may be a powerful tool, if used with the understanding that the exact nature of the processes may not be identifiable with this model alone. This is also demonstrable with the application to Atlantic sea scallop belt transect survey data with a variation of the model described in Chang et al. (2017), and given here in a Bayesian context. The sill behaves as expected, with an increase following an increase in the number of knots (see Figure

3.17, top right). The range, however, rises with the number of knots (see Figure 3.17, top left), from about 5% of the range of the data set to about 10%. Given that the increase follows primarily from an increase in the number of knots in the cubic spline or depth relationship, we might surmise that said relationship might explain some of the variability at very fine scales. In this case, the range of the variogram might increase to account for variability at much broader scales, though still significantly smaller than the range of the entire data set. Given the sheer knowledge base concerning the sea scallop fishery on the United States' Atlantic coast, we can use prior information to guide our choice of model. In this case, the observed scale of aggregation, typical median density updated by well-established growth equations, and so forth. We can use this information in concert with the model recommended by the DIC in cases in which information is paltry, or parameters are less well studied.

One must also take care to note that the prediction variance is driven by the structure of the residuals. The residuals are, of course, determined by the structure of the penalized spline model and how it varies amongst each iteration of the MCMC, and therefore the overall magnitude of the variance of the spline model and the kriging model are very similar (Figure 3.18, center). Therefore, the variance of the sum is driven by fine-scale stochasticity, arising from error in the predictions. This differs a great deal from the variance of either model in isolation, and gives us our final principle.

Principle 4: The variance of the predictions arising from applying a penalized spline and kriging the residuals is the equivalent to the variance of the prediction error.

This is not a new result, but rather a demonstration of a relationship noted in Hengl et al. (2003).

In summary, then, we hypothesized that an increase in the number of knots of the penalized spline would result in a decline in the range and sill of the variogram. We find this to be true, with a few exceptions. We surmised that, despite the conceptual difficulty that this decline presents, the model may still be useful if the relationship between the penalized spline and the variogram stabilizes. It appears to do so, although the three-dimensional case is fraught by more complex data, and thus a more complex outcome. One can combine an analytical and Bayesian approach to model selection by considering both information criteria, as well as prior information about the species in question. Overall, with cautious intent, this model structure, may prove an exceptionally useful addition to the statistical toolbox.

Chapter 4

A Bayesian update of a method to estimate Atlantic sea scallop (*Placopecten magellanicus magellanicus*) abundance from belt transect surveys

Chang et al. (2017) compare both the accuracy and the precision of several modeling methods to estimate abundance from a belt transect survey of Atlantic sea scallops (*Placopecten magellanicus*). They find that describing broad scale trends over water depth and space with a generalized additive model and using ordinary kriging on the residuals to assess spatial autocorrelation – termed GAM+OK – performs best in terms of both accuracy and precision. We build upon these results by adapting the model structure for a Bayesian framework. This allows us to estimate the parameters of the GAM and the variogram simultaneously. We apply the model to the 2015 belt transect survey of Atlantic sea scallops in Georges Bank and the Mid-Atlantic Bight off the eastern coast of the United States. In general, Bayesian model predictions agree with the predictions of the frequentist GAM+OK method, with few exceptions. One notable exception – CA2_{sinGeorgesBank} – *–currently arises from a significant error in the depth maps provided, and awaits correction. However, scale autocorrelation among observations is far higher than that suggested by its frequentist counterpart. The OK method is more accurate, as subsequent application of the GAM and then the OK might favor trends*

4.1 Background

The belt transect surveys employed by the Northeast Fisheries Science Center (NEFSC) collect continuous, photographic data on benthic communities off the Atlantic coast of the United States. These data are, by nature, both non-random and spatially autocorrelated. If we consider, specifically, the Atlantic sea scallops (*Placopecten magellanicus*), then we must also account for their tendency to aggregate, which ultimately results in a high percentage of zeroes in the data, or zero-inflation. Chang et al. (2017) test several modeling methods, which include ordinary kriging (OK), and two variations on regression kriging (RK) with generalized additive models (GAM's). They find that regression kriging with a generalized additive model, termed GAM+OK and applied by applying the GAM and OK subsequently, performs best amongst all model-based approaches, in terms of both accuracy and precision. We expand upon this model here by estimating the parameters of the GAM and OK jointly in a Bayesian framework.

The GAM+OK method is employed by applying a generalized additive model to density as function of depth, and as a function of two-dimensional spatial coordinates. They also include a hurdle model to describe the degree of zero-inflation. They use the residuals of this model to perform ordinary kriging. Conceptually, this model describes a trend with depth, and in space. Atlantic sea scallops tend to occupy intermediate depths. At shallow depths, maximum bottom temperatures may exceed the temperatures at which the scallops can survive (Hart and Chute, 2004). In deeper waters, scallop recruitment is inhibited both by limited food resources and, in some areas, predation by *Astropecten* sea stars (Hart, 2006). In space, broad trends may follow sediment type or tidal and other current transport (Thouzeau et al., 1991). At finer spatial scales, scallops may enhance their own recruitment directly by aggregating

as fecund adults (Stokesbury and Himmelman, 1993). Therefore, each piece of this model is designed to describe a spatial biological processes at different scales. A challenge exists, however, in that the scales of the regression and geostatistical models may overlap. For example, when we consider preference for depth, water depth is itself related to the spatial coordinates.

With the GAM+OK method, the model first finds an optimal fit to depth and to two-dimensional space. As a natural result of fitting methods, the models will attempt to find an unbiased fit of the smoothing functions to the data, thus attributing a majority of the observed spatial distribution of density and abundance to depth or spatial trend, and therefore to effects such as temperature, predator density, and broad-scale processes such as sediment type and transport. A variogram is then fitted to the residuals that remain, describing what relatively fine-scale autocorrelation amongst observations remains after applying the GAM. In a Bayesian framework, we can find an optimal fit to all pieces simultaneously. However, this can produce an infinite estimation space and thus induce collinearity, as noted by Ginsbourger et al. (2009). We can mitigate this by clever choice of prior distributions, and by application of model restrictions that reflect prior knowledge of the system. For example, we can restrict the curvature of the depth portion of the model by limiting the magnitude of the variance of centered depth model coefficients, thereby reflecting a generally smooth relationship of density with depth as a function of temperature and predator density.

In general, besides those advantages common to all Bayesian statistics – including non-asymptotic estimation and probabilistic interpretability – there are two advantages of translating the GAM+OK model to a Bayesian framework specific to Atlantic sea scallops. First, while subsequent application of the GAM and then the OK has proven both accurate and precise (Chang et al., 2017), overfitting and therefore bias in the

estimates can result due to the lack of a reweighting scheme by which an optimal fit to both the GAM and the OK are found simultaneously. Bayesian methods provide a natural way by which to find the joint posterior probability of all parameters simultaneously, provided we specify a formal connection between the fit of the GAM to the mean and the variance-covariance matrix determined by the parameters of the variogram. Second, Bayesian methods provide a sound theoretical framework in which to quantify and incorporate prior knowledge. There are many years of survey data on the Atlantic sea scallop, and a great deal of biological data concerning the animals' physiology and behavior. This includes extensive knowledge of growth and reproductive rates. For purposes of both priors and projection, each years' survey data is highly correlated with the next. Therefore, a Bayesian update of the GAM+OK method is a natural next step in development.

In this paper, we present the Bayesian model, its results, and both its potential faults and the theoretical gains. Ultimately, we would like to use this model to produce informative estimates for subsequent years, augmented by one year of growth and reproduction. Informative priors can be robust to inaccurate information (Depaoli, 2014), avoid problems associated with non-informative priors when data are relatively sparse (Lenk and Orme, 2009), and can easily be adjusted to reflect the degree of confidence in one's past information (Horrocks and Rueffer, 2014). This could serve as a powerful framework for analysis of transect survey information.

4.2 Methods

4.2.1 Data Collection

The Woods Hole Oceanographic Institution (WHOI), the Northeast Fisheries Science Center (NEFSC), and commercial fishermen developed the Habitat Camera Mapping System (HabCam) to survey benthic communities. They collected continuous photographic data in Georges Bank (GB) in 2011, and in both GB and the Mid-Atlantic Bight (MAB) in 2012-2019. Two cameras produce roughly six photographs per second as they are towed along pre-designed transects at approximately 5-7 knots. The system is designed to maintain a height of about 2m off the sea floor. Researchers evaluated a subset of these photographs for Atlantic sea scallop abundance and shell height, producing tens of thousands of observations. We aggregate these data at the scale of 5km to reduce the number of observations and thereby reduce the computational burden on the models. We stratify both GB and the MAB to produce areas with similar depth profiles, for seven areas in GB, five areas in the MAB, and thus twelve areas in total. We scale both the UTM coordinates and the depth of all observations for input into the model as follows.

$$y_i = \frac{y_i - \mu_Y}{\sigma_Y}$$

where $y_i \in Y$, and where μ_Y and σ_Y are the mean and standard deviation, respectively, of the UTM coordinates or of depth. Scaling prevents the exponential link functions from generating very large numbers. We use these scaled coordinates and depths in the following estimation model.

4.2.2 Estimation Model

We built a Bayesian extension of the GAM+OK method in **JAGS** (Plummer, 2017). We manually built the code necessary for both generalized additive models and kriging in R (R Core Team, 2019b). We execute all model runs using the package **jagsUI** (Kellner, 2019).

Generalized additive model

A generalized additive model can be written

$$\mu_{y_i} = \beta_0 + \sum_{j=1}^m f_j(x_{ij})$$

where the functions $f_j(x)$ are unknown smooth functions on a set of predictors x_i . Typically, these smooth functions are penalized splines, a series of polynomial function stitched together at locations called knots, the fit of which is penalized by some measure of the curvature. The penalty is scaled by a parameter λ , which can be either chosen at will or estimated. The default in the package **mgcv** (Wood, 2003) are the thin plate splines, the multi-dimensional generalization of the cubic splines, in which the scaling parameter λ is estimated via generalized cross validation. Informally, the thin plate splines smooth with a radial basis function, and penalize the error function with the scaled sum of the second derivatives. The second derivative is a concise measure of the degree of curvature. Chang et al. (2017) use the default splines for their analysis, and so we use a similar structure here. They use a cubic spline to smooth over depth, and a thin plate spline to smooth over the coordinates.

For our analysis, we use a low rank, truncated basis for the cubic spline, as given in Crainiceanu et al. (2005). This can be written

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=2}^{K_d+1} \beta_k ||x - \kappa_k||_+^3$$

where x is scaled depth, K_d is the number of knots, β_k are the basis function coefficients, and κ_k is the k^{th} knot. The double vertical bars refer to Euclidean distance. The subscript $+$ is a function that sets all negative output to zero. Low rank is achieved by removing the parametric terms x^2 and x^3 . This reduces autocorrelation amongst the coefficients. We space knots evenly throughout the range of the data. Say we have n observations. If we call Ω the $n \times K_d$ matrix of second derivatives, with elements given by $\Omega_{ij} = \int_{\min(x)}^{\max(x)} f_i''(x) f_j''(x) dx$, then we can follow Crainiceanu et al. (2005) and re-parameterize the coefficients as $b = (\beta_0, \beta_1, b_{K_d})$ where $b_{K_d} = \Omega^{1/2} \beta_{K_d}$. The subscript K_d refers to all K_d coefficients of the non-parametric basis functions. If we call all K_d basis functions applied to all n observations Z_{K_d} , then we can write the design matrix

$$B = \begin{bmatrix} \beta_0 & \beta_1 & Z_{K_d} \end{bmatrix}$$

where β_0 is an $n \times 1$ column vector repeating the value β_0 , and where β_1 is an $n \times 1$ column vector repeating the value β_1 . We then set $Z_d = Z_{K_d} \Omega^{-1/2}$. This gives the final design matrix for the depth portion of the model as follows.

$$D_d = \begin{bmatrix} \beta_0 & \beta_1 & Z_d \end{bmatrix}$$

We perform a similar set of steps to parameterize the spatial portion of the generalized additive model. We include a two-dimensional thin plate spline on abundance as a function of spatial coordinates in order to describe broad-scale trends. While this may

seem counter-intuitive, given that the variogram models spatial autocorrelation, Chang et al. (2017) intend for this portion of the model to describe broad spatial trend as a function of environmental variables. In a previous paper, Duskey et al. (2020) argue that a Bayesian model structure that includes simultaneous estimation of a penalized spline, and of a variogram on the residuals, is capable of identifying fine-scale spatial autocorrelation separate from a large-scale, mean generating process. The model may underestimate the sill and the range as the number of knots, and thus as the flexibility of the surface, increases. Thus, these parameters may not have a straightforward and literal interpretation. Rather, they may serve as something of a litmus test as to whether or not there is appreciable fine-scale autocorrelation.

We use the full rank representation of the thin plate spline, which is written

$$f(x) = \beta_0 + \beta_1 a + \beta_2 b + \sum_{k=3}^{K_c+2} \beta_k r^2 \ln(r)$$

where $x = (a, b)$ are scaled UTM coordinates, K_c is the number of knots, β_k is as above, and where

$$r = ||x - \kappa_k||$$

The knots κ_k are now given as ordered pairs. The basis functions belong to the class of radial basis functions, in which the output depends only upon the distance between points. In this case, low rank may be achieved by removing the higher order parametric terms a and b . However, we find that these terms induce little autocorrelation amongst the parameters. When we combine the depth portion of the generalized additive model with the spatial portion, we also remove β_0 . If we call all K_c basis functions applied

to all n observations Z_{K_c} , then we can write the design matrix

$$D_c = \begin{bmatrix} \beta_2 & \beta_3 & Z_{K_c} \end{bmatrix}$$

Without loss of generality, we rename β_1 and β_2 , β_2 and β_3 to prevent confusion when we combine both functions. Then, β_2 and β_3 are $n \times 1$ column vectors repeating the values β_2 and β_3 , respectively. We then reparameterize Z_{K_c} as Z_c just as above, following the methods of Crainiceanu et al. (2005), and obtain $b_{K_c} = (\beta_2, \beta_3, b_{K_c})$. However, in this case, the elements of Ω are given by $\Omega_{ij} = \|\kappa_i - \kappa_j\|$, that is, by the distance between each of K_c knots. We can then write the combined design matrix as

$$\begin{aligned} D &= \begin{bmatrix} \beta_0 & \beta_1 & Z_c & \beta_2 & \beta_3 & Z_d \end{bmatrix} \\ &= \begin{bmatrix} D_d & D_c \end{bmatrix} \end{aligned}$$

This is an $m \times n$ matrix, where $m = 2 + K_d + 2 + K_c$. The coefficients are given by $\mathbf{b} = (b_{K_d}, b_{K_c})$. Then, with abundance observations Y , we can write the model

$$E[Y] = D\mathbf{b}$$

where f is a given link function. We use a log-link function, and therefore

$$E[Y] = e^{D\mathbf{b}}$$

We can generate the matrix D before we run the model. Therefore, all that remains to be estimated within **JAGS** are the coefficients b . We set the following priors for

b_{K_d} and b_{K_c} , and their associated variances

$$b_{K_d} \sim N(0, \sigma_d^2)$$

$$b_{K_c} \sim N(0, \sigma_c^2)$$

$$\sigma_d^2 \sim T(10, 1)[0, 1]$$

$$\sigma_c^2 \sim T(10, 1)[0, 1]$$

where the distribution $T(10, 1)[0, 1]$ is the non-central t-distribution with non-centrality parameter 10, as recommended by Gelman et al. (2006) for hierarchical variance parameters. We constrain the variance to a maximum of 1. This constrains the curvature of the cubic spline and the thin plate spline, producing a broad trend, and preventing severe autocorrelation of the coefficients with the parameters of the variogram. This reflects the strong prior belief that the relationship between mean Atlantic sea scallop abundance and both depth and space is very broad in scale based on previous surveys (Hart (2006), Chang et al. (2017)). It also serves as a method to increase the number of effective samples per actual sample in an MCMC chain, and thereby drastically reduce computation time.

We use a negative binomial likelihood for mean Atlantic sea scallop abundance, as follows.

$$\mu = e^{Db + \text{offset}}$$

$$p = \frac{r}{r + \mu}$$

$$Y \sim NB(p, r\pi)$$

The offset is the area, in square meters, of each observation $y_i \in Y$. The parameter r is commonly interpreted as a measure of under- or over-dispersion. Values of r greater than 1 imply aggregation, while values less than 1 imply regularity. Values at or near 1 suggest that the variance scales with the mean. We gave r the following prior.

$$r \sim \text{Gamma}(1, 1)$$

with shape and rate parameters both set to 1. The parameter π is a measure of the probability of presence. As noted in Chang et al. (2017), Atlantic sea scallop abundance data tends to be zero-inflated. We can model presence, π_i , for the i^{th} observation as a function of space. Atlantic sea scallop presence tends to correlate with depth. Thus, we model presence as a polynomial function of the coordinate direction in which depth experiences the greatest change. We make this decision based on investigation of each of the twelve areas in GB and the MAB. This can be written

$$\pi_i = \alpha_0 + \alpha_1 c_i + \alpha_2 c_i^2$$

where we give the coefficients the following priors

$$\alpha_0 \sim T(0, 1)$$

$$\alpha_1 \sim T(0, 1)$$

$$\alpha_2 \sim T(0, 1)$$

Probability of presence is thus a second order function of either the x- or y-coordinate.

The likelihood is

$$\text{Presence}_i \sim \text{Binomial}(\pi_i)$$

where

$$\text{Presence}_i = \begin{cases} 0 & \text{if } y_i = 0 \\ 1 & \text{if } y_i > 0 \end{cases} \quad (4.1)$$

In summary, we manipulate the design matrix in order to center the coefficients of the generalized additive model. The smoothing penalty and its associated scaling parameter λ is absorbed by the variance. We chose a negative binomial likelihood for abundance to account for under- or over-dispersion, offset by area to account for density. We include probability of presence to model zero-inflation. This produces a mean μ , which we use to calculate the residuals as follows.

$$\epsilon = Y - \mu$$

We input the residuals into the calculation of the variogram.

Kriging

Kriging is a common method of spatial interpolation. Given the residuals ϵ , we wish to predict the residual at a new point, given by ϵ_0 , as a weighted sum of the observed residuals ϵ_i as follows.

$$\hat{\epsilon}_0 = Z(\epsilon_0) = \sum_{i=1}^N w_i \epsilon_i = \mathbf{w}^T \boldsymbol{\epsilon}$$

where

$$\sum_i^N w_i = \mathbf{w}^T \mathbf{1} = 1$$

Forcing the weights w_i to sum to 1 ensures that the estimator $Z(\epsilon_0)$ is unbiased. We can find a precise and accurate fit to the residuals by minimizing the variance

$$\text{Var}(\epsilon_0) = E[(\epsilon_0 - \hat{\epsilon}_0)^2]$$

Ultimately, the weights of the variogram are a function of the variance-covariance matrix of the observed residuals, and of the covariance of the observations ϵ_i with the unobserved quantity ϵ_0 . This can be written

$$\mathbf{w} = C^{-1}(D - \lambda \mathbf{1})$$

where the elements of C are given by $C_{ij} = \text{Cov}(\epsilon_i, \epsilon_j)$, the elements of D are given by $D_i = \text{Cov}(\epsilon_i, \epsilon_0)$, and λ is a Lagrange multiplier, ensuring that the weights $\mathbf{w}$ sum to 1. The structure of C is typically given by one of several variograms, a strictly positive definite function which asserts that data points that are closer together are more closely related, that is, have a greater covariance, or alternatively, a lower semivariance. Chang et al. (2017) use a Matérn variogram. However, Bessel functions are unavailable in **JAGS**, and so we use the closely related exponential variogram, written as

$$\gamma(h) = \text{nugget} + (\text{sill} - \text{nugget})(1 - e^{-\frac{h}{\text{range}}})$$

where $h = \|\cdot\|$, the Euclidean distance between any two points. Then, we can write the covariance between any two residuals, ϵ_i and ϵ_j as

$$\text{Cov}(\epsilon_i, \epsilon_j) = \text{nugget} + \text{sill} e^{-\frac{\|\epsilon_i - \epsilon_j\|}{\text{range}}}$$

Therefore, within **JAGS**, it is only necessary to estimate nugget, sill, and range of the variogram. We choose not to estimate a nugget. This serves to reduce the autocorrelation amongst the parameters of the variogram. Practically, the nugget describes very fine-scale variability, which may be interpreted as process or measurement error. Setting the nugget to zero assumes there is none in the residuals. Effectively, we are assuming that there is no co-variability at scales finer than the minimum distance -

about 5km - between points in our observations. While this is suspect at best, we have accounted for some variability at finer scales in the generalized additive model. Therefore, when used in concert with a GAM, ignoring the nugget effect may be acceptable. Given that inclusion of a nugget also produces autocorrelation in subsequent MCMC draws of the GAM parameters in fact suggests that the nugget and one or more parameters of the GAM are attempting to explain the same phenomenon. Therefore, we have just three parameters to estimate for the variogram, as well as n means for each of the observations, with the following priors.

$$\begin{aligned}\text{sill} &\sim N(0, 1000) \\ \text{range} &\sim T(0, 1)[0,] \\ \sigma_\epsilon^2 &\sim T(10, 1)[0, 1] \\ \mu_{i_\epsilon} &\sim N(0, \sigma_\epsilon^2)\end{aligned}$$

where σ_ϵ^2 is the global variance of the residuals. Then we calculate the variance-covariance matrix C from the exponential co-variogram as above, and arrive at the likelihood

$$Y \sim MVN(\mu_Y, C)$$

where

$$\mu_Y = \boldsymbol{\mu}_\epsilon + \mu$$

where $\boldsymbol{\mu}_\epsilon = \mu_{1_\epsilon}, \dots, \mu_{n_\epsilon}$, where μ is the mean generated by the GAM above, and MVN is the multi-variate normal distribution. We can write the estimated totals

$$\hat{Y} = \hat{\mu}_\epsilon + \hat{\mu}$$

where $\hat{\mu}_\epsilon$ and $\hat{\mu}$ are estimates produced by the output of the **JAGS** model.

Area	Region	Observations	Burn-in	Iterations	Thin	Samples	Time (min)
DMV	MAB	80	10000	25000	5	9000	104
ET	MAB	122	50000	310000	50	15600	3564
HC	MAB	154	5000	35000	5	18000	927
HCsr	MAB	158	2000	12000	2	15000	307
LI	MAB	106	5000	15000	5	6000	185
CA2_N	GB	35	1000	10000	1	27000	9
CA2_S	GB	150	10000	30000	5	12000	806
GSC_E	GB	224	1000	4000	1	9000	261
GSC_S	GB	214	10000	45000	5	21000	1006
GSC_W	GB	198	10000	30000	5	12000	1417
NF	GB	86	5000	30000	5	15000	74
SF	GB	152	50000	55000	5	3000	766

Table 4.1: **JAGS** model run details. Areas are: Delmarva (DMV), Elephant Trunk (ET), Hudson Canyon (HC), Hudson Canyon southern rim (HCsr), Long Island (LI), Closed Area 2 North (CA2_N), Closed Area 2 South (CA2_S), Great South Channel East (GSC_E), Great South Channel South (GSC_S), Great South Channel West (GSC_W), Northern Flank (NF), and Southern Flank (SF). Observations is the number of observations. Burn-in is the number of samples we discard at the beginning of the model run. Iterations is the total number of iterations, including burn-in. Thin is the amount of thinning i.e. a thin of 5 indicates we only kept every 5th sample. Samples is the total number of samples, after burn-in and thinning. We run each model in parallel on three cores of a standard laptop computer. All parameters of all model runs converged, with a Gelman-Rubin diagnostic of $\hat{R} < 1.1$.

Implementation

We run the Bayesian model, which we refer to as Bayes+OK, in **JAGS** (Plummer, 2017) using the package **jagsUI** (Kellner, 2019) in R (R Core Team, 2019b). We use the function *autojags()*, which allows the user to run the model in increments, until the model converges, reducing unnecessary run-time. The details of all model runs are given in Table 4.1. We run all models with 10 knots for the depth smoother and 10 knots for the spatial coordinate smoother.

4.3 Results

Overall, the Bayes+OK model proves stable and efficient, converging in relatively few iterations in each area of both the Mid-Atlantic Bight (MAB) and Georges Bank (GB) (Table 4.1). In most cases, the abundance results agree fairly well with the results of the GAM+OK given by Chang et al. (2017). However, the Bayes+OK model produces particularly high estimates of total abundance in Closed Area 2 South (CA2_S) and Great South Channel South (GSC_S) (Table 4.2). Even where overall abundance results agree, the range and sill parameters of the variogram are quite different. Ranges are large in the Bayes+OK model in just three cases, with the rest typically smaller by at least half. The sill parameters of the variogram are much greater, typically by multiple orders of magnitude. That is, the Bayes+OK model believes that there is a very high degree of spatial autocorrelation at very fine scales. The GAM appears to explain very large-scale variability. Alternatively, the GAM+OK accounts for the same phenomena by attributing a low level of autocorrelation at larger scales. Also of note is that the coefficient of variation (CV) of abundance estimates are typically lower for the Bayesian model, at least by half, and often by an order of magnitude. The gross exception is area CA2_S, with a CV of 45%, five times the CV of the GAM+OK method. Note that Chang et al. (2017) use data aggregated at a scales of 0.75km, whereas we use data aggregated at a scales of 5km. Some differences may be induced by this scale of aggregation.

Below, we present the results of each piece of the model individually, followed by summary totals in each area of each region.

Area	Region	GAM+OK					Bayes+OK			
		Abundance	CV	Range	Sill	Nugget	Abundance	CV	Range	Sill
DMV	MAB	2445	0.12	7318	0.20	0.00	2504	0.05	1976	871
ET	MAB	8577	0.09	4947	5.10	6.01	9857	0.01	1992	4913
HC	MAB	2808	0.15	6718	0.36	0.20	2463	0.09	1728	1770
HCsr	MAB	2767	0.08	12384	0.08	0.05	2031	0.03	1577	443
LI	MAB	1949	0.13	704	0.11	0.00	1830	0.03	3276	235
Total	MAB	18547					18684			
CA2_N	GB	198	0.38	11989	0.11	0.09	136	0.03	4265	331
CA2_S	GB	1043	0.08	42676	0.80	0.35	10757	0.45	366	779
GSC_E	GB	416	2.59	788	14.75	13.56	440	0.29	943	819
GSC_S	GB	16118	0.47	46733	152.09	22.02	23526	0.03	1911	2420
GSC_W	GB	1465	0.28	12684	0.38	0.31	988	0.03	1309	1372
NF	GB	109	0.55	178410	0.12	0.16	107	0.16	2898	116
SF	GB	523	0.18	3037	0.04	0.06	603	0.07	884	22
Total	GB	20684					36556			

Table 4.2: Results of frequentist GAM+OK (left) and Bayesian model (right). Abundance is given in millions. Range is given in meters. Recall that we did not estimate a nugget in the Bayesian model, and so none is shown.

4.3.1 Probability of presence

We begin by presenting the results of the probability of presence model. Recall that we model presence as a second order polynomial function of a cardinal direction. Given that presence tends to correlate with depth, we chose the cardinal direction for each area of each region by determining in which - East-West or North-South - the depth changed most. In most cases, it is clear that there is a strong relationship between cardinal direction and probability of presence (Figure 4.1). There are exceptions. In the Mid-Atlantic Bight, Delmarva (DMV) and Elephant Trunk (ET) both have a consistently high probability of presence throughout their respective areas. In Georges Bank, Closed Area 2 North (CA2_N) has a consistently low probability of presence throughout the area. Otherwise, each area tends to have a varied and complex relationship of presence with location, often represented by a strong preference for a central region, or an extreme. Unsurprisingly, the estimates for probability of presence for Closed Area 2 South (CA2_S) are as uncertain as its estimates of abundance, with very broad 95% high posterior density intervals.

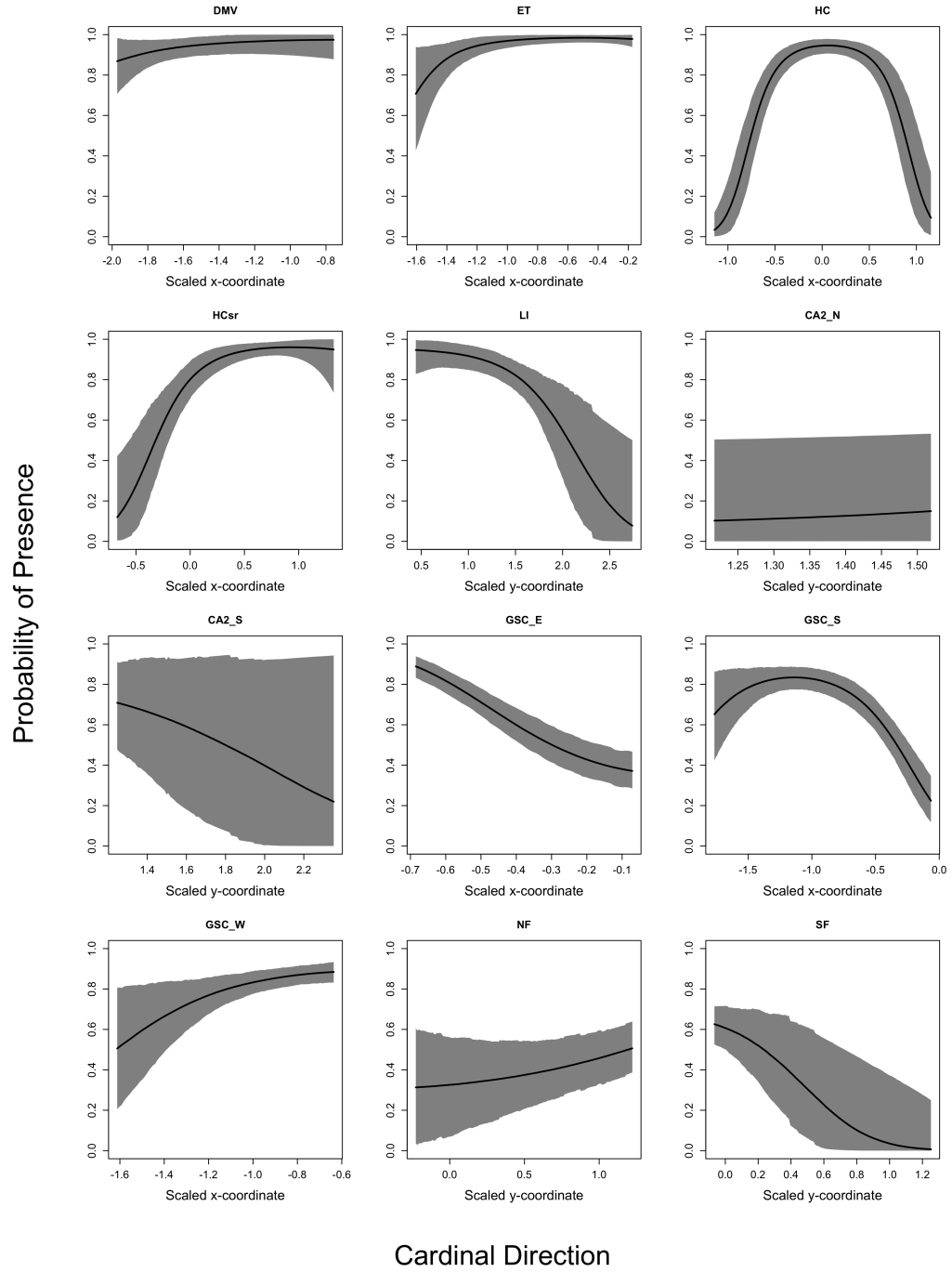


Figure 4.1: Probability of presence as a function of cardinal direction for each area of each region. Abbreviations are given as in Table 4.1, and in the same order as both Table 4.1 and Table 4.2, from left to right and top to bottom. The cardinal direction is listed below the x-axis of each individual plot. Shaded areas are 95% high posterior density intervals. The lower and upper extremes of the x-axis represent the minimum and maximum value of the given cardinal direction within that area.

Our intention was to allow cardinal coordinate to serve primarily as a proxy for depth. Visually, we can associate the results shown in Figure 4.1 with those in Figure 4.2, in which we present the depth. In some areas, as in Hudson Canyon (HC), it is clear that the probability of presence is greatest at intermediate depths of about 65 meters or so. In other areas, the preference appears to be for shallow depths of about 50 meters at the northern extreme of the Southern Flank (SF). In the case of Delmarva (DMV) and Elephant Trunk (ET), there are clear and substantial differences in depth throughout both areas, yet little apparent preference for one depth or another, with only slight evidence of avoidance of the western extreme.

4.3.2 Mean Abundance

We model mean abundance with a generalized additive model (GAM), as a cubic function of depth, and a radial function of two-dimensional spatial coordinate. Note that we obtain abundance estimates by modeling density, and multiplying the results by observed area, in m^2 . We can attempt to understand the contribution of both by splitting the design matrix D into its depth and coordinate components, given above as D_d and D_c , respectively. Then, we can arrive at two components of estimated mean density, given as

$$\hat{y}_d = e^{b_{K_d} D_d}$$

$$\hat{y}_c = e^{b_{K_c} D_c}$$

Then

$$\hat{y} = \hat{y}_d \cdot \hat{y}_c$$

Note the lack of offset, and the use of a lowercase y to signify that we are now considering density. As each estimation area is exactly $750 \times 750 \text{ m}$, all values are

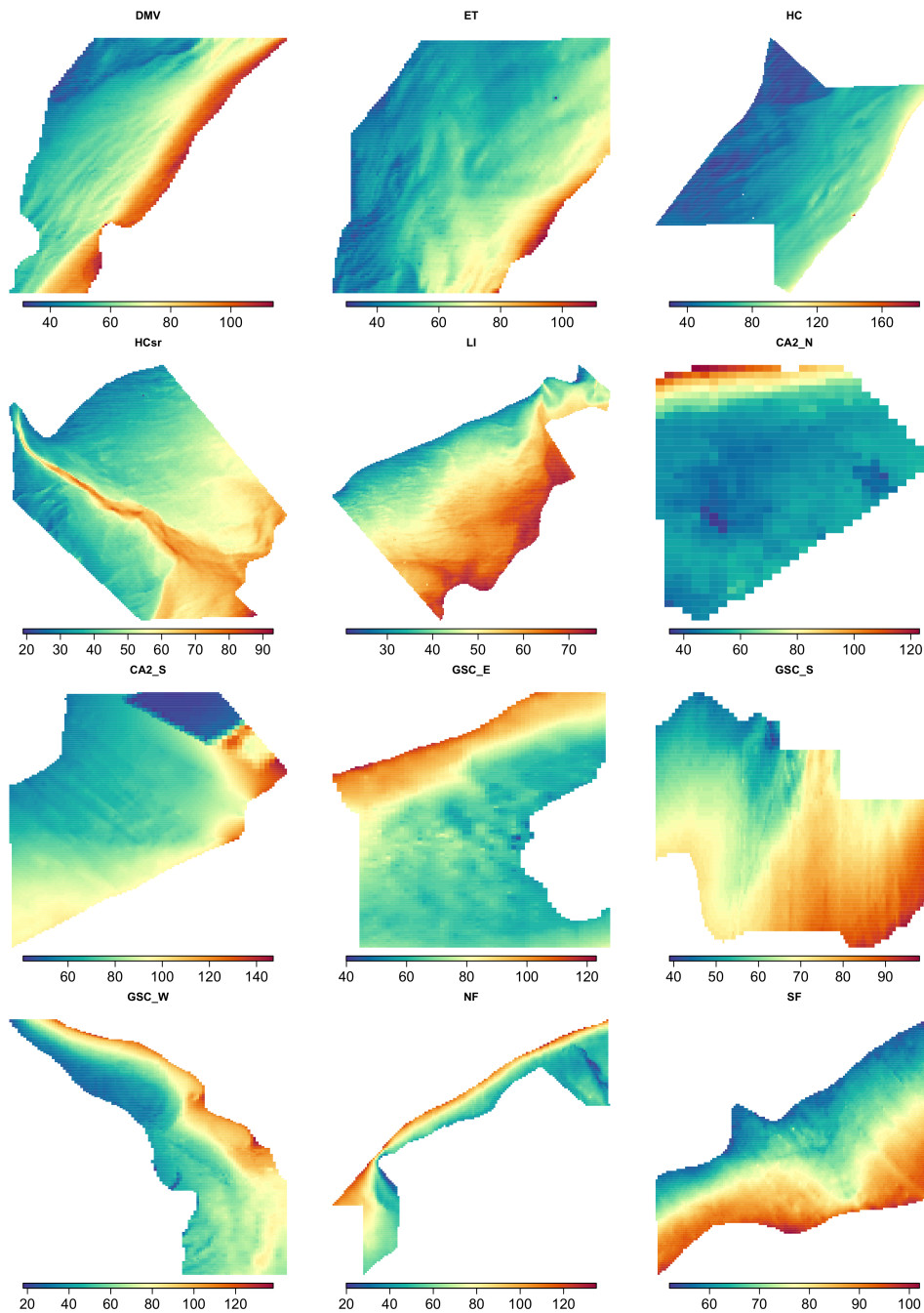


Figure 4.2: Observed depth (m) in each area of each region. Abbreviations are given as in Table 4.1, and in the same order as both Table 4.1 and Table 4.2, from left to right and top to bottom.

proportional to abundance, with identical proportionality factor. We first consider the impact of depth alone on density, absent observed depth in each spatial area of each region (Figure 4.3). In general density peaks at intermediate values, with values approaching zero at the extremes. This is fairly universal throughout each area of each region. The 95% high posterior density intervals are fairly broad, indicating a certain degree of uncertainty in the results. We observe the greatest spread in Closed Area 2 South (CA2_S) of Georges Bank, which is consistent with the high CV (Table 4.2) of the abundance estimate. We can also observe that about half of the predictive depths falls outside of the range of observed depths. It is, at best, inadvisable to predict outside of the range of an additive model. It is especially problematic in this area, as values appear to be increasing significantly just as the observation range terminates at about 97m. That said, both the spatial smoother and kriging do have the potential to mitigate this effect, given that some observations do appear in the general area in which these extreme depths appear. In general, mean density estimates for depth alone are below one, which scales estimates from coordinates alone downward as a function of depth, more so in shallow or deep waters.

When we consider the impact of two-dimensional spatial coordinates alone on density, we find that most mean density estimates are greater than one, with some exceptions in Hudson Canyon southern rim (HCsr), Closed Area 2 South (CA2_S), Great South Channel East (GSC_E), Great South Channel West (GSC_W), and Southern Flank (SF) (Figure 4.4). This is especially of note in Closed Area 2 South (CA2_S), with a relatively low contribution from the spatial smoother, and an extremely high, if not somewhat pathological, depth smoother (Figure 4.3). On the opposite extreme, the spatial smoother predicts very high densities in the central area of Delmarva (DMV), in the northeast of Closed Area 2 North (CA2_N), and the northwest of Great South

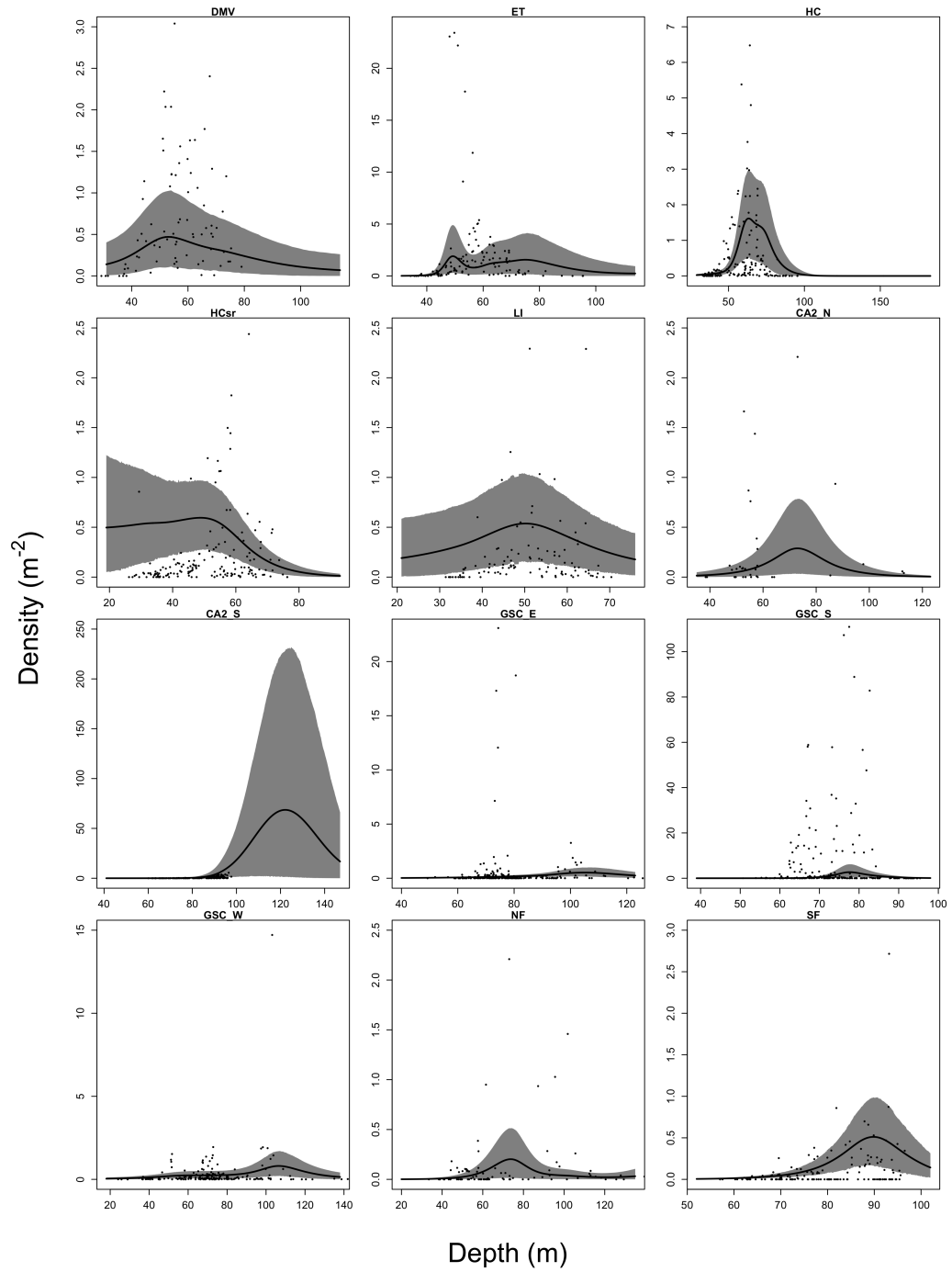


Figure 4.3: The relationship of mean density (m^{-2}) to depth. Abbreviations are given as in Table 4.1, and in the same order as both Table 4.1 and Table 4.2, from left to right and top to bottom. Shaded areas are 95% high posterior density intervals. Points are all density observations within a given area. The lower and upper extremes of the x-axis represent the minimum and maximum value of depth within that area.

Channel South (GSC_S) (Figure 4.4). The cardinal coordinates are, of course, an indirect measure of other variables that vary over space. This could include depth, which was our intention, or other variables, such as presence of competitors or predators.

Combining the probability of presence, the depth smoother, and the spatial smoother, we arrive at mean density throughout each area of each region (Figure 4.5). The combination is easy to visualize. Consider Hudson Canyon (HC). First, note that probability of presence is high in the north-south strip (Figure 4.1). Second, note that density peaks at intermediate depths of about 65m (Figure 4.3) and that these depths are common in the southwest-northeast strip (Figure 4.2). Lastly, note that density peaks near the southwest. All these things combined, we arrive at a fairly narrow peak of density in the south-central portion of Hudson Canyon. Here, all pieces appear to play a roughly equal role. By contrast, in Elephant Trunk (ET), the spatial smoother appears to play a dominant role in the mean (Figure 4.4). Similarly, the depth smoother clearly plays a dominant role in Closed Area 2 South (CA2_S; Figure 4.3). One can follow the trajectory of the results from each piece of the model to arrive at the mean density estimates for any area from either region.

4.3.3 Kriging

The results for the parameters of the variogram are given in Table 4.2. Note that, in most cases, some or all of the survey track is visible within the estimates (Figure 4.6). There may be at least two explanations for this phenomenon. First, the visibility of the tracks is a representation of the fact that the estimated range is small relative to the distance between adjacent tracks, and thus additional information becomes available as we approach the tracks. The information is very strong, given that, with a nugget set to zero, the kriging portion of the model is more or less serving as an exact

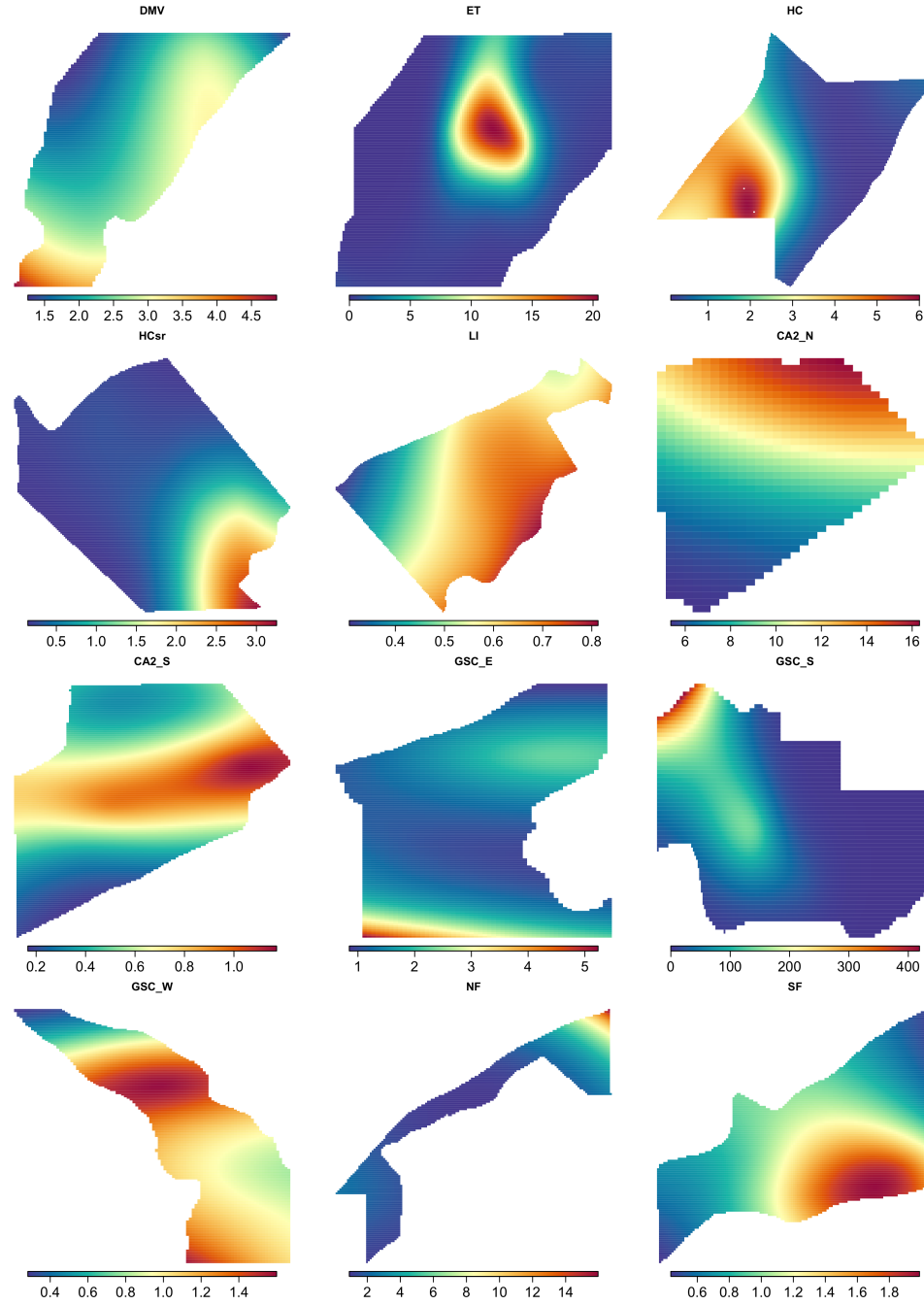


Figure 4.4: The relationship of mean density (m^{-2}) to two-dimensional spatial coordinates. Abbreviations are given as in Table 4.1, and in the same order as both Table 4.1 and Table 4.2, from left to right and top to bottom.

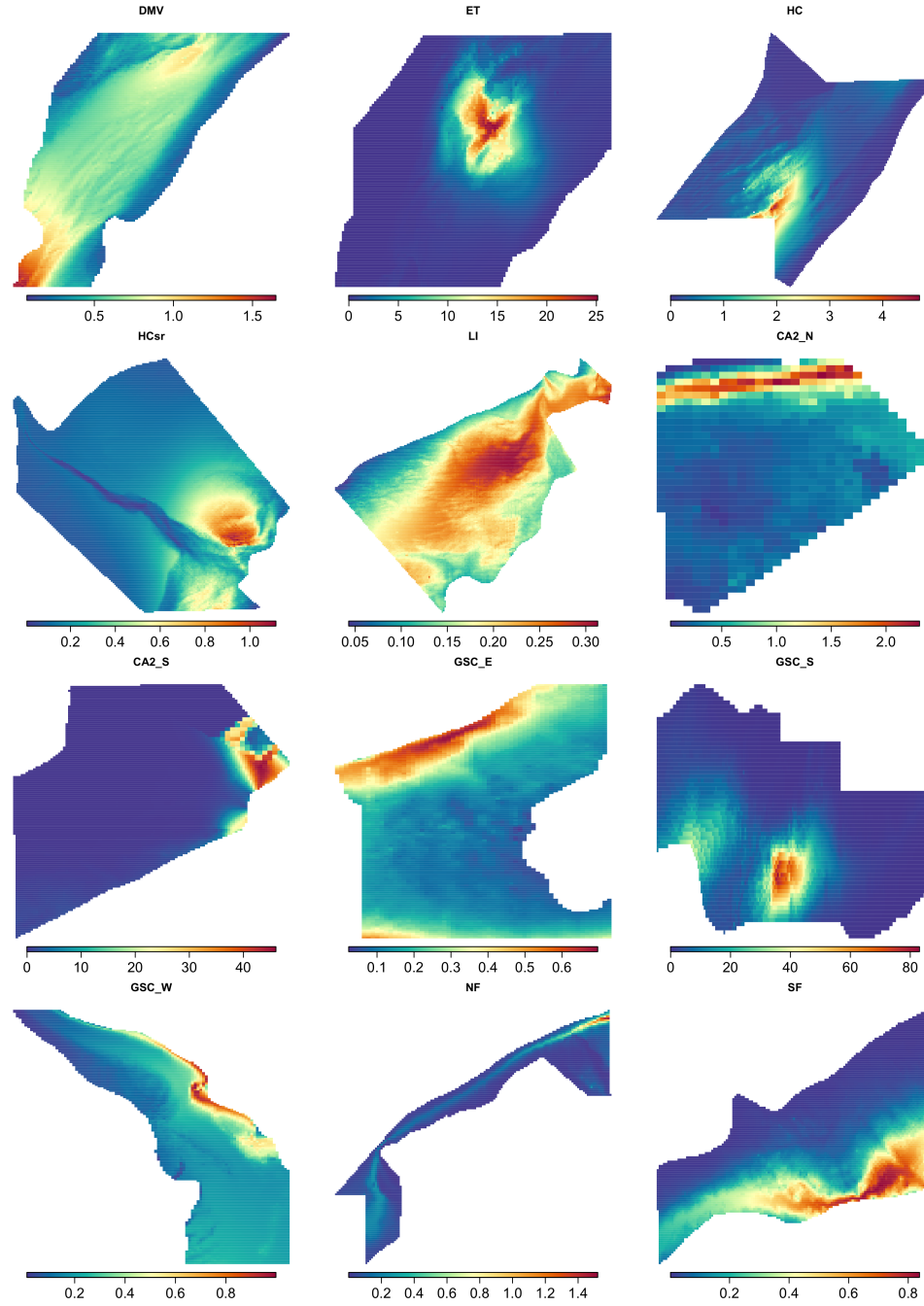


Figure 4.5: Mean density (m^{-2}) as a function of both depth and two-dimensional spatial coordinates. Abbreviations are given as in Table 4.1, and in the same order as both Table 4.1 and Table 4.2, from left to right and top to bottom.

interpolator for the observations, a problem noted in Ginsbourger et al. (2009). While kriging is indeed a method of spatial interpolation, it is designed to estimate spatial *trends*, with a localized interpretation of the variance. Second, the broad scale of the data may mask much of the fine-scale autocorrelation process, as the estimator reverts towards the mean provided by the GAM as we move away from the observations.

The final estimates of the totals are obtained by adding the residuals in Figure 4.6 to the mean values obtained in Figure 4.5. This gives Figure 4.7. In general, the spatial structure of the density is one of hotspots surrounded by very low density, which agrees with our interpretation of spatial Atlantic sea scallop density as zero-inflated and spatially autocorrelated. When we consider the entirety of the Mid-Atlantic Bight or of Georges Bank (Figure 4.8), we can see that the most extreme hotspots of scallop density are in just one or a few areas within each region. In general, these hotspots tend to agree with observations in the HabCam survey data of unusually high density.

4.4 Discussion

In general, the model appears to converge well. There are some performance issues in the output, namely very low CV's, though this is a natural consequence of using a model-based approach, in which we make assumptions concerning the variability. We might also attribute this in part to the variogram, which in most cases takes precedence. While we may cause this by forcing a fairly low variance, and thus a high precision, on the generalized additive model (GAM) coefficients, the lack of any restrictions on the variance produces severe autocorrelation in subsequent draws of the Markov Chain Monte Carlo (MCMC) chains. We are essentially forcing over-smoothing by controlling the variance, or rather the precision, but this prevents pathological interference between the GAM and the variogram. This suggests that both model structures are in fierce

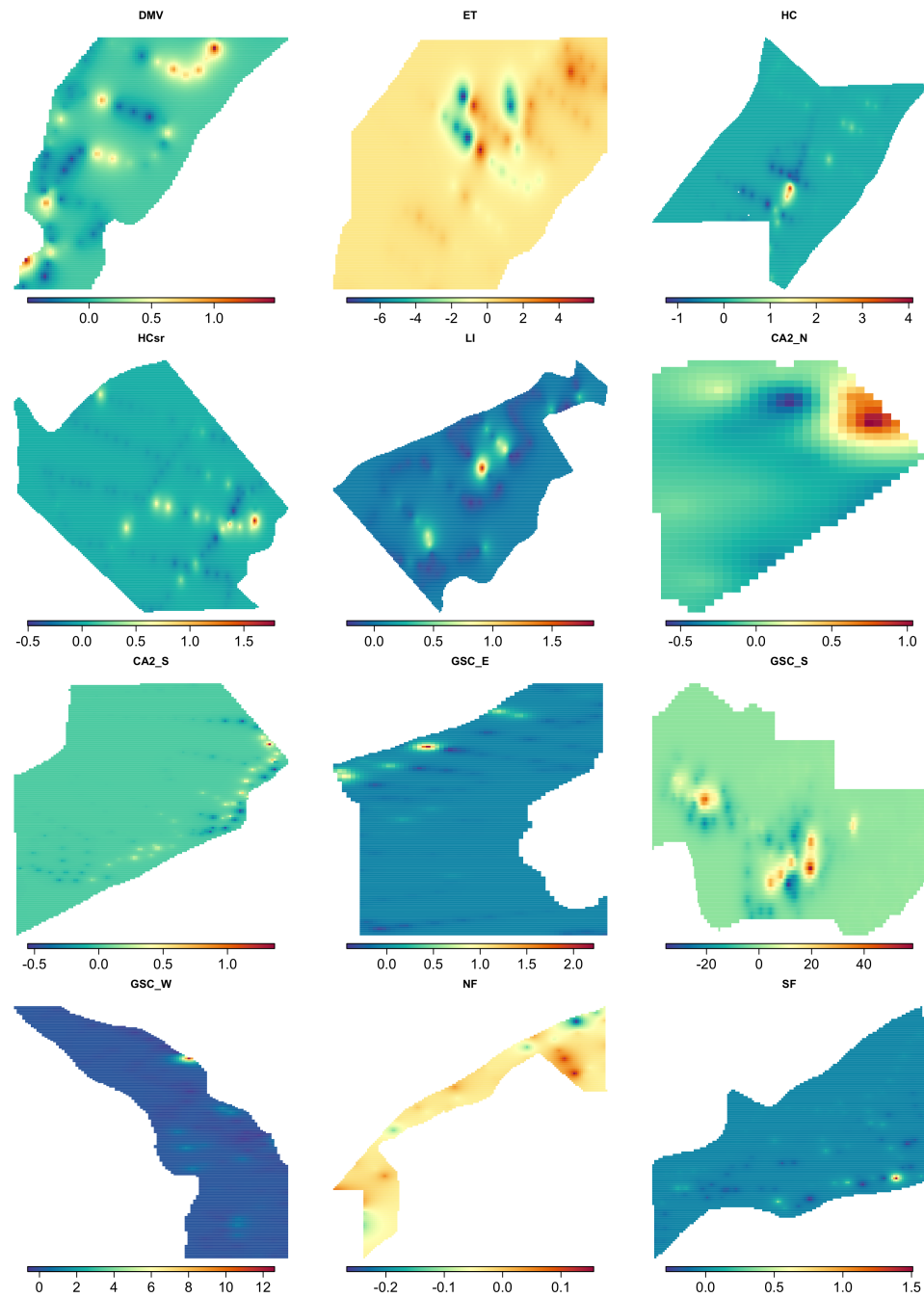


Figure 4.6: Mean kriged density (m^{-2}) residuals. Abbreviations are given as in Table 4.1, and in the same order as both Table 4.1 and Table 4.2, from left to right and top to bottom.

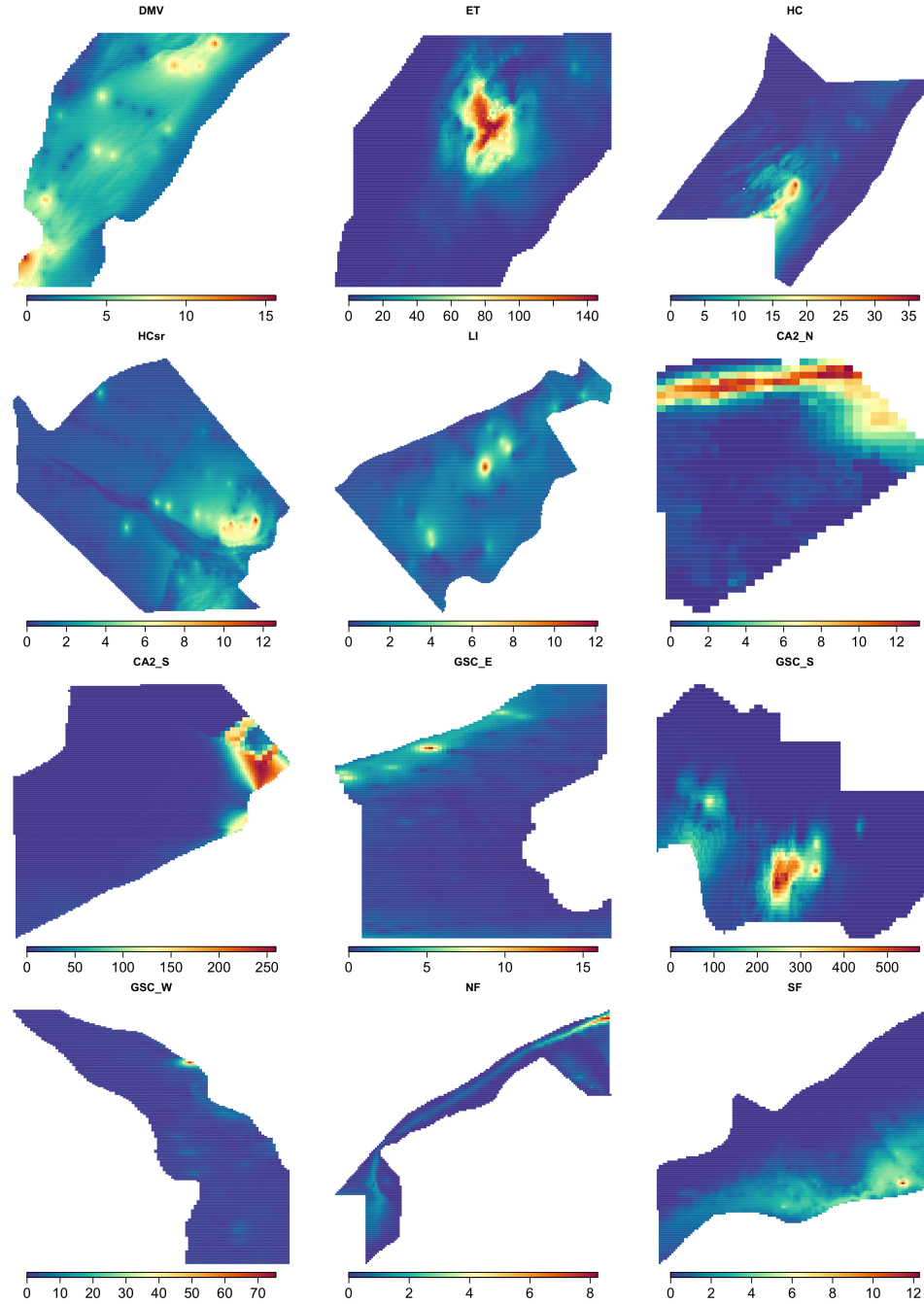


Figure 4.7: Total estimated density (m^{-2}), including both the mean and the kriged residuals. Abbreviations are given as in Table 4.1, and in the same order as both Table 4.1 and Table 4.2, from left to right and top to bottom.

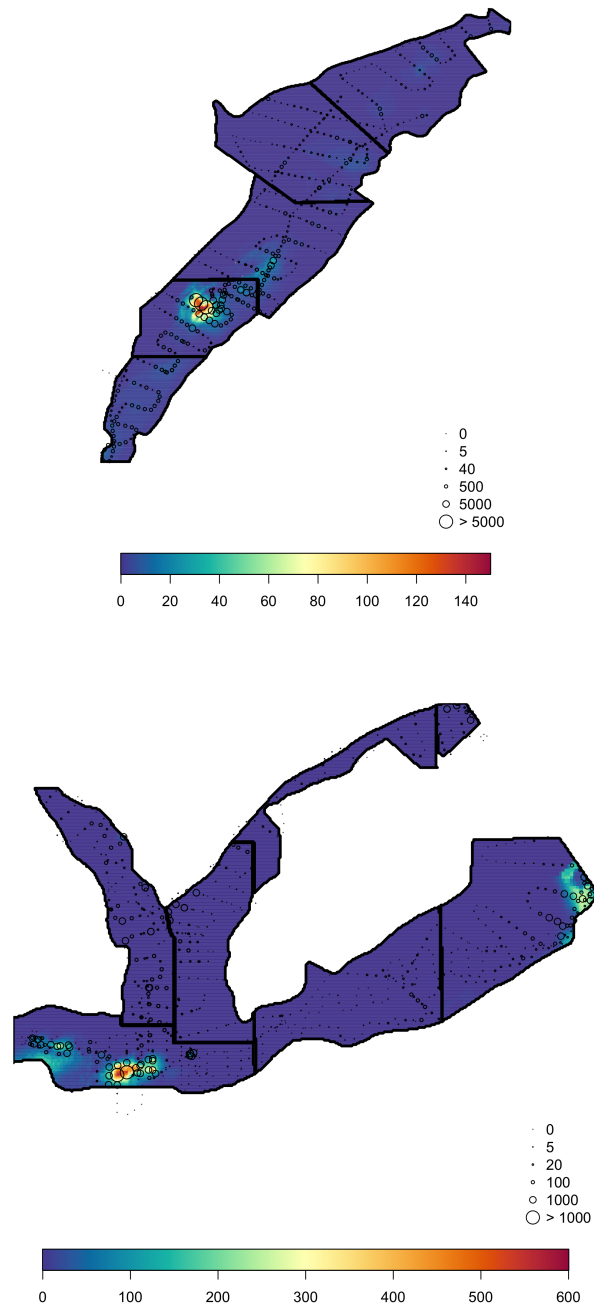


Figure 4.8: Abundance estimates for the Mid-Atlantic Bight (above) and for Georges Bank (below), in hundreds of thousands. Included are all observations in each area of each region. Each pixel is 750×750 meters.

competition and that, without some prior information concerning which dominates fine-scale variability, there is a large solution space with many potential solutions. Fortunately, in our case, we do have some prior information on the mechanisms guiding spatial density.

That said, there is still a delicate balance between choosing model priors that give convergence in a reasonable amount of time and choosing model priors based on known quantities. Our current parameterization may err on the side of sacrificing some realism for swift convergence. During model development, we found that, in some cases, increasing the variance of the GAM model coefficients did increase the range and decrease the sill of the variogram. While they are, as noted, still lower and higher, respectively, than those produced by the GAM+OK method, they are perhaps more reasonable. This was much more common in areas with large, dense aggregations, such as Elephant Trunk (ET; see Figure 4.7), where the range increases from about 2000m to about 8000m as we remove restrictions on the GAM. However, this was not the case in areas such as Delmarva (DMV; see Figure 4.7), with a range roughly equivalent to that produced by the GAM+OK method (Table 4.2). Removing restrictions entirely still results in the same sill and range, suggesting a robust estimate. One can potentially arrive at an infinite number of solutions based on these or other restrictions within the model and therefore, interpretations must be treated with caution, and tempered by expert knowledge. In our case, we recognize that the spatial scales suggested by the Bayesian estimates of the sills and ranges of the variogram are finer than those of the frequentist method because the Bayesian GAM operates at a broader scale, thus leaving the variogram to identify much stronger autocorrelation at a finer scale, and vice versa (Table 4.2). In future runs, we may wish to tweak the variance of the coefficients, as well as inform the priors of the ranges, until they arrive at estimates more consistent

with observations. While GAM coefficients have no *direct* biological interpretation, the aggregate estimates do of course have a clear interpretation, which we can validate with observations (see Figure 4.3). The sill and especially the range of the variogram *do* have a direct biological interpretation. We can use prior information concerning aggregative behavior as a check on kriging model output.

Broad-scale aggregation may also cause some of the discrepancies between the Bayesian and the GAM+OK methods (Table 4.2). However, given that computation time of the Bayesian model increases exponentially with the number of observations, we were forced to aggregate during the process of development. We are searching for a quantitative interpretation of aggregative behavior that we have observed on scales far smaller than 5km. A scale of 0.75km may still mask some of this behavior, but not to the extent of a scale of 5km. Chang et al. (2017) find that the scale of aggregation, alternatively termed segment length, had a significant effect on model output and performance. While we recognize that, in the case of an exponential variogram, the effective range – defined as the distance at which the semi-variance is 95% of the estimated sill – is exactly three times the range, some of the estimated ranges are still, at most, at or just slightly above the scale of the aggregated data, with Closed Area 2 South (CA2_N; Table 4.2) the greatest exception. Note that the effective range of a Matérn variogram is exactly twice the range.

Of more interest, perhaps, is the sill of the variogram which, even in cases where the ranges or effective ranges are on similar or nearly identical scales, are many times larger than the sills predicted by the GAM+OK method. This may be in part a result of the low estimates of the ranges. As the range shrinks, it may explain very strong, very fine-scale autocorrelation. Thus, the Bayesian model interprets the magnitude of spatial autocorrelation is very high. Meanwhile, even when we remove restrictions

on the variance, the depth and spatial smoothers remain very smooth indeed, in the case of Delmarva, for example. This may suggest that, when run simultaneously, the variogram, and thus spatial autocorrelation, plays a much more significant role than that suggested by a structure in which a GAM and OK are performed subsequently. One might also be concerned that the relatively low number of knots of the GAM may give the variogram an advantage. However, Duskey and Sullivan (2020) find that the relationship between a penalized spline and a variogram tends to stabilize between two and ten knots, especially when smoothing over a one-dimensional variable. This is less consistent when smoothing over two-dimensional space, but still apparent in some cases. Thus, ten knots appears to be entirely adequate to allow the GAM to compete effectively with the variogram.

In summary, any restriction serves as a strong prior belief, and so we cannot say that we have confirmed that each of these processes – depth preference, spatial trend, small-scale aggregation – described by the GAM and by the variogram act on the exact scales predicted by either the Bayesian or the GAM+OK method. However, we apply these restrictions because we do have good, prior information about Atlantic sea scallop distribution, abundance, and ecology, based in part on previous surveys, so this is precisely the situation for which informative priors were designed. Generally speaking, the model does behave well when we include these restrictions, which is some evidence that it is an appropriate structure. The fact that it also offers similar estimates with very different output for the parameters suggests that, in actuality, more of the process is attributable to spatial autocorrelation than originally thought. This is the primary advantage of joint probability estimation, arriving at an optimal solution for all pieces of the model simultaneously. The primary disadvantage at present is computation time. That said, now that the model is developed, additional cores may be used to run the

model at smaller aggregative scales. However, aggregation is certainly necessary to some extent with any model, as any structure would struggle with the thousands of data points, let alone a Bayesian model that relies on MCMC. Laplace approximation methods (Lindgren et al., 2015) or alternative repeated sampling methods to MCMC that reduce autocorrelation may reduce computation time (Girolami and Calderhead, 2011). Additional or different model restrictions may also improve estimation.

Overall, then, while care must be taken in aggregation as we predict outside of our observable range, we believe this is a strong and viable model structure, when use of prior information is employed. The Bayes+OK model performs well, typically providing similar estimates to those provided by the GAM+OK model, yet with very different characterization of the fine-scale autocorrelation. In general, the Bayes+OK model describes autocorrelation on a much finer scale than the GAM+OK model, and with a much greater magnitude. Given that the Bayes+OK model includes a reweighting scheme by which both the parameters of the GAM and of the variogram are estimated simultaneously, it is likely a more biologically accurate representation of the processes driving patterns of abundance and density. This model could be a powerful tool, when combined with models of growth and recruitment to project forward. Subsequent Bayesian model runs may become more efficient as we offer up better prior estimates.

Bibliography

- Adams, A. J., J. E. Hill, and C. Samoray (2011). Characteristics of spawning ground fidelity by a diadromous fish: a multi-year perspective. *Environmental Biology of Fishes* 92(3), 403–411.
- Alsamamra, H., J. A. Ruiz-Arias, D. Pozo-Vázquez, and J. Tovar-Pescador (2009). A comparative study of ordinary and residual kriging techniques for mapping global solar radiation over southern Spain. *Agricultural and Forest meteorology* 149(8), 1343–1357.
- Altman, N. (2000). Theory & methods: Kriging, smooth, both or neither? (with discussion). *Australian & New Zealand Journal of Statistics* 42(4), 441–461.
- ASMFC (1998). Amendment 1 to the interstate fishery management plan for Atlantic sturgeon. *Management Report* (31).
- ASMFC (2012). Habitat addendum IV to amendment 1 to the interstate fishery management plan for Atlantic sturgeon. *Management Report*.
- ASMFC (2017). Atlantic sturgeon benchmark stock assessment.
- Balazik, M. T. and J. A. Musick (2015). Dual annual spawning races in Atlantic sturgeon. *PloS one* 10(5), e0128234.
- Baum, L. E. and T. Petrie (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics* 37(6), 1554–1563.

- Bell, R. (2008a). Hudson river estuary sediment environment map (nysdec) metadata. *Lamont-Doherty Earth Observatory of Columbia University*.
- Bell, R. (2008b). Hudson river estuary sediment type map (nysdec) metadata. *Lamont-Doherty Earth Observatory of Columbia University*.
- Bell, R., S. Carbotte, F. Nitsche, W. Ryan, and Lamont-Doherty Earth Observatory of Columbia University (2006). Hudson river estuary sediment environment map (nysdec). *Hudson River Estuary Program*.
- Bell, R., W. Ryan, S. Carbotte, F. Nitsche, C. McHugh, and Lamont-Doherty Earth Observatory of Columbia University (2006). Hudson river estuary sediment cores map (nysdec). *Hudson River Estuary Program*.
- Bell, R., W. Ryan, S. Carbotte, F. Nitsche, C. McHugh, Lamont-Doherty Earth Observatory of Columbia University, and Marine Sciences Research Center, State University of New York at Stony Brook (2006). Hudson river estuary morphology - new york state (nysdec). *Hudson River Estuary Program*.
- Beverton, R. J. and S. J. Holt (2012). *On the dynamics of exploited fish populations*, Volume 11. Springer Science & Business Media.
- Booth, A. J. (2000). Incorporating the spatial component of fisheries data into stock assessment models. *ICES Journal of Marine Science* 57(4), 858–865.
- Breece, M. W., D. A. Fox, and M. J. Oliver (2018). Environmental drivers of adult atlantic sturgeon movement and residency in the delaware bay. *Marine and Coastal Fisheries* 10(2), 269–280.
- Britton, R. H. and M. E. Moser (1982). Size specific predation by herons and its effect

- on the sex-ratio of natural populations of the mosquito fish *Gambusia affinis baird* and *girard*. *Oecologia* 53(2), 146–151.
- Chang, J.-H., B. V. Shank, and D. R. Hart (2017). A comparison of methods to estimate abundance and biomass from belt transect surveys. *Limnology and Oceanography: Methods* 15(5), 480–494.
- Cochran, W. (1977). Sampling techniques (3rd edn.,) wiley. New York.
- Codling, E. A., M. J. Plank, and S. Benhamou (2008). Random walk models in biology. *Journal of the Royal Society Interface* 5(25), 813–834.
- Collins, M. R., T. I. Smith, W. C. Post, and O. Pashuk (2000). Habitat utilization and biological characteristics of adult atlantic sturgeon in two south carolina rivers. *Transactions of the American Fisheries Society* 129(4), 982–988.
- Crainiceanu, C., D. Ruppert, and M. P. Wand (2005). Bayesian analysis for penalized spline regression using winbugs. *Journal of Statistical Software* 14(14), 1–24.
- Cressie, N. (1986). Kriging nonstationary data. *Journal of the American Statistical Association* 81(395), 625–634.
- Cressie, N. (1990). The origins of kriging. *Mathematical geology* 22(3), 239–252.
- Cressie, N. and C. K. Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.
- Da Costa, J., F. Chainet, B. Celse, M. Lacoue-Negre, C. Ruckebusch, and D. Espinat (2018). Comparing kriging, spline, and mlr in product properties modelization: Application to cloud point prediction. *Energy & Fuels* 32(4), 5623–5634.

- Depaoli, S. (2014). The impact of inaccurate “informative” priors for growth parameters in bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal* 21(2), 239–252.
- Dovel, W. and T. Berggren (1983). Atlantic sturgeon of the hudson estuary, new york. *New York Fish and Game Journal* 30(2), 140–172.
- Dubrule, O. (1983). Two methods with different objectives: splines and kriging. *Journal of the International Association for Mathematical Geology* 15(2), 245–257.
- Dubrule, O. (1984). Comparing splines and kriging. *Computers & Geosciences* 10(2-3), 327–338.
- Dunton, K. J., A. Jordaan, D. O. Conover, K. A. McKown, L. A. Bonacci, and M. G. Frisk (2015). Marine distribution and habitat use of atlantic sturgeon in new york lead to fisheries interactions and bycatch. *Marine and Coastal Fisheries* 7(1), 18–32.
- Dunton, K. J., A. Jordaan, K. A. McKown, D. O. Conover, and M. G. Frisk (2010). Abundance and distribution of atlantic sturgeon (*acipenser oxyrinchus*) within the northwest atlantic ocean, determined from five fishery-independent surveys. *Fishery Bulletin* 108(4), 450.
- Flood, R. (2007). Hudson river estuary bathymetric contours metadata. *Marine Sciences Research Center, State University of New York at Stony Brook*.
- Frank, H. J., M. E. Mather, J. M. Smith, R. M. Muth, J. T. Finn, and S. D. McCormick (2009). What is “fallback”? metrics needed to assess telemetry tag effects on anadromous fish behavior. *Hydrobiologia* 635(1), 237–249.
- Franke, A., T. Caelli, G. Kuzyk, and R. J. Hudson (2006). Prediction of wolf (*canis*

- lupus) kill-sites using hidden markov models. *Ecological modelling* 197(1-2), 237–246.
- Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis* 1(3), 515–534.
- Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7(4), 457–472.
- Ginsbourger, D., D. Dupuy, A. Badea, L. Carraro, and O. Roustant (2009). A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments. *Applied Stochastic Models in Business and Industry* 25(2), 115–131.
- Girolami, M. and B. Calderhead (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(2), 123–214.
- Godinho, A. L., B. Kynard, and H. P. Godinho (2007). Migration and spawning of female surubim (*pseudoplatystoma corruscans*, pimelodidae) in the são francisco river, brazil. *Environmental Biology of Fishes* 80(4), 421–433.
- Grabowski, T. B. and J. J. Isely (2006). Seasonal and diel movements and habitat use of robust redhorses in the lower savannah river, georgia and south carolina. *Transactions of the American Fisheries Society* 135(5), 1145–1155.
- Gross, M. R., J. Repka, C. T. Robertson, D. H. Secor, and W. Van Winkle (2002). Sturgeon conservation: insights from elasticity analysis. In *American Fisheries Society Symposium*, Volume 28, pp. 13–30. Citeseer.

- Hart, D. and A. Chute (2004). Essential fish habitat source document: sea scallop. *Placopecten magellanicus, life history and habitat characteristics*.
- Hart, D. R. (2006). Effects of sea stars and crabs on sea scallop *Placopecten magellanicus* recruitment in the mid-atlantic bight (usa). *Marine Ecology Progress Series* 306, 209–221.
- Hengl, T., G. B. Heuvelink, and D. G. Rossiter (2007). About regression-kriging: From equations to case studies. *Computers & geosciences* 33(10), 1301–1315.
- Hengl, T., G. B. Heuvelink, and A. Stein (2003). Comparison of kriging with external drift and regression kriging. ITC Enschede, Netherlands.
- High, B., C. A. Peery, and D. H. Bennett (2006). Temporary staging of columbia river summer steelhead in coolwater areas and its effect on migration rates. *Transactions of the American Fisheries Society* 135(2), 519–528.
- Horrocks, J. and M. Rueffer (2014). A bayesian approach to estimating animal density from binary acoustic transects. *Computational Statistics & Data Analysis* 80, 17–25.
- Hudson River National Estuarine Research Reserve (2008). Hudson river estuary morphology - new york state (nysdec) metadata. *Hudson River Estuary Program*.
- Jager, H. I. and D. L. DeAngelis (2018). The confluences of ideas leading to, and the flow of ideas emerging from, individual-based modeling of riverine fishes. *Ecological Modelling* 384, 341–352.
- Jonsen, I. D., R. A. Myers, and J. M. Flemming (2003). Meta-analysis of animal movement using state-space models. *Ecology* 84(11), 3055–3063.
- Jonsson, B., R. S. Waples, and K. Friedland (1999). Extinction considerations for diadromous fishes. *ICES Journal of Marine Science* 56(4), 405–409.

- Kammann, E. and M. P. Wand (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52(1), 1–18.
- Kellner, K. (2017). *jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses*. R package version 1.4.9.
- Kellner, K. (2019). *jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses*. R package version 1.5.1.
- Kimeldorf, G. S. and G. Wahba (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* 41(2), 495–502.
- Klimley, A. P. and C. Holloway (1999). School fidelity and homing synchronicity of yellowfin tuna, *thunnus albacares*. *Marine Biology* 133(2), 307–317.
- Koster, W., D. Crook, D. Dawson, S. Gaskill, and J. Morrongiello (2018). Predicting the influence of streamflow on migration and spawning of a threatened diadromous fish, the australian grayling *prototroctes maraena*. *Environmental management* 61(3), 443–453.
- Laslett, G. M. (1994). Kriging and splines: an empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association* 89(426), 391–400.
- Lenk, P. and B. Orme (2009). The value of informative priors in bayesian inference with sparse data. *Journal of Marketing Research* 46(6), 832–845.
- Leos-Barajas, V., E. J. Gangloff, T. Adam, R. Langrock, F. M. Van Beest, J. Nabe-Nielsen, and J. M. Morales (2017). Multi-scale modeling of animal movement and

- general behavior data using hidden markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics* 22(3), 232–248.
- Limburg, K. E. and J. R. Waldman (2009). Dramatic declines in north atlantic diadromous fishes. *BioScience* 59(11), 955–965.
- Lindgren, F., H. Rue, et al. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software* 63(19), 1–25.
- Lindstrom, M. J. (1999). Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics* 8(2), 333–352.
- Lindstrom, M. J. (2002). Bayesian estimation of free-knot splines using reversible jumps. *Computational statistics & data analysis* 41(2), 255–269.
- Marine Sciences Research Center, State University of New York at Stony Brook (2007). Hudson river estuary bathymetric contours. *Hudson River Estuary Program*.
- Matheron, G. et al. (1981). Splines and kriging; their formal equivalence.
- McClintock, B. T., R. King, L. Thomas, J. Matthiopoulos, B. J. McConnell, and J. M. Morales (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecological Monographs* 82(3), 335–349.
- McDowall, R. (1992). Particular problems for the conservation of diadromous fish. *Aquatic Conservation: Marine and freshwater ecosystems* 2(4), 351–355.
- McDowall, R. (1993). Implications of diadromy for the structuring and modelling of riverine fish communities in new zealand. *New Zealand journal of marine and freshwater research* 27(4), 453–462.
- Melnychuk, M. C., K. J. Dunton, A. Jordaan, K. A. McKown, and M. G. Frisk (2017). Informing conservation strategies for the endangered atlantic sturgeon using acoustic

- telemetry and multi-state mark–recapture models. *Journal of Applied Ecology* 54(3), 914–925.
- Michelot, T., R. Langrock, and T. A. Patterson (2016). movehmm: an r package for the statistical modelling of animal movement data using hidden markov models. *Methods in Ecology and Evolution* 7(11), 1308–1315.
- Morales, J. M., D. T. Haydon, J. Frair, K. E. Holsinger, and J. M. Fryxell (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology* 85(9), 2436–2445.
- Myers, D. E. (1994). Spatial interpolation: an overview. *Geoderma* 62(1-3), 17–28.
- National Oceanic and Atmospheric Administration (2012). Endangered and threatened wildlife and plants; threatened and endangered status for distinct population segments of atlantic sturgeon in the northeast region. *Fed. Reg.* 77, 5879–5912.
- Ndongo, F. B., A. Fasso, E. Porcu, and J. Mateu (2016). Recent developments of functional kriging.
- Neu, C. W., C. R. Byers, and J. M. Peek (1974). A technique for analysis of utilization-availability data. *The Journal of Wildlife Management*, 541–545.
- Opsomer, J., Y. Wang, and Y. Yang (2001). Nonparametric regression with correlated errors. *Statistical Science*, 134–153.
- Petitgas, P., J. Massé, P. Grellier, and P. Beillois (2003). Variation in the spatial distribution of fish length: a multi-annual geostatistics approach on anchovy in biscay 1983-2002. *ICES CM 2003*, 14.

- Planque, B., C. Loots, P. Petitgas, U. Lindstrøm, and S. Vaz (2011). Understanding what controls the spatial distribution of fish populations using a multi-model approach. *Fisheries Oceanography* 20(1), 1–17.
- Plummer, M. (2017). Jags version 4.3.0 user manual [computer software manual]. Retrieved from sourceforge.net/projects/mcmc-jags/files/Manuals/4.x.
- QGIS Development Team (2009). *QGIS Geographic Information System*. Open Source Geospatial Foundation.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- R Core Team (2019a). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- R Core Team (2019b). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Roberts, S., T. Guilford, I. Rezek, and D. Biro (2004). Positional entropy during pigeon homing i: application of bayesian latent state modelling. *Journal of theoretical biology* 227(1), 39–50.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics* 11(4), 735–757.
- Schönfisch, B. and M. Kinder (2002). A fish migration model. In *International Conference on Cellular Automata*, pp. 210–219. Springer.
- Shank, B. V., D. R. Hart, and K. D. Friedland (2012). Post-settlement predation by sea stars and crabs on the sea scallop in the mid-atlantic bight. *Marine Ecology Progress Series* 468, 161–177.

- Smith, T. I. (1985). The fishery, biology, and management of atlantic sturgeon, *acipenser oxyrhynchus*, in north america. *Environmental Biology of Fishes* 14(1), 61–72.
- Snyder, M. N., N. H. Schumaker, J. L. Ebersole, J. B. Dunham, R. L. Comeleo, M. L. Keefer, P. Leinenbach, A. Brookes, B. Cope, J. Wu, et al. (2019). Individual based modeling of fish migration in a 2-d river system: model description and case study. *Landscape Ecology* 34(4), 737–754.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van der Linde (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 485–493.
- Stegemann, E. C. (1994). Sturgeon-the king of freshwater fishes. *Conservation-ist* 49(1), 18–23.
- Stein, A. B., K. D. Friedland, and M. Sutherland (2004). Atlantic sturgeon marine distribution and habitat use along the northeastern coast of the united states. *Transactions of the American Fisheries Society* 133(3), 527–537.
- Stokesbury, K. D. and J. H. Himmelman (1993). Spatial distribution of the giant scallop *placopecten magellanicus* in unharvested beds in the baie des chaleurs, quebec. *Marine Ecology-Progress Series* 96, 159–159.
- Strayer, D. L. (2007). Submersed vegetation as habitat for invertebrates in the hudson river estuary. *Estuaries and Coasts* 30(2), 253–264.
- Taub, S. H. (1990). *Fishery management plan for Atlantic sturgeon (Acipenser oxyrhynchus oxyrhynchus)*. Number 17. ASMFC.

- Thouzeau, G., G. Robert, and S. J. Smith (1991). Spatial variability in distribution and growth of juvenile and adult sea scallops *placopecten magellanicus* (gmelin) on eastern georges bank (northwest atlantic). *Marine Ecology Progress Series*, 205–218.
- Van Eenennaam, J., S. Doroshov, G. Moberg, J. Watson, D. Moore, and J. Linares (1996). Reproductive conditions of the atlantic sturgeon (*acipenser oxyrinchus*) in the hudson river. *Estuaries* 19(4), 769.
- Voltz, M. and R. Webster (1990). A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *Journal of soil Science* 41(3), 473–490.
- Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* 65(1), 95–114.
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1), 3–36.
- Wright, I. W. (1977). Spline methods in statistics. Technical report, NORTH CAROLINA UNIV AT CHAPEL HILL DEPT OF STATISTICS.
- Wu, T. and Y. Li (2013). Spatial interpolation of temperature in the united states using residual kriging. *Applied Geography* 44, 112–120.
- Yang, P. and T. L. Ng (2019). Fast bayesian regression kriging method for real-time merging of radar, rain gauge, and crowdsourced rainfall data. *Water Resources Research* 55(4), 3194–3214.