A BAYESIAN SPATIAL HIERARCHICAL MODEL FOR PUTTING IN GOLF

by

Kasra Yousefi
B.Sc., Simon Fraser University, 2011

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

© Kasra Yousefi 2013
SIMON FRASER UNIVERSITY
Spring 2013

All rights reserved.
However, in accordance with the *Copyright Act of Canada*, this work may be reproduced without authorization under the conditions for "Fair Dealing." Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review and news reporting is likely to be in accordance with the law, particularly if cited appropriately.

APPROVAL

Name: Kasra Yousefi

Degree: Master of Science

Title of project: A BAYESIAN SPATIAL HIERARCHICAL MODEL FOR PUTTING IN GOLF

Examining Committee:

Dr. Carl Schwarz
Professor
Chair

Dr. Tim Swartz
Professor
Senior Supervisor

Dr. David Campbell
Assistant Professor
Committee Member

Dr. Richard Lockhart
Professor
External Examiner

Date Approved: _____

Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the "Institutional Repository" link of the SFU Library website (www.lib.sfu.ca) at http://summit/sfu.ca and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author's written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library
Burnaby, British Columbia, Canada

revised Fall 2011

Abstract

In this project, a novel statistic to evaluate the putting performance of professional golfers is developed. The methodology provided in this paper borrows ideas from Bayesian spatial statistics to determine the expected number of putts at various green locations. After constructing a Bayesian hierarchical model, the necessary derivation and computations of the relevant full conditional distributions are discussed. Data from the 2012 Honda Classic tournament obtained from the ShotLink website is then used to investigate the approach. The Metropolis within Gibbs algorithm is run to generate samples from the posterior distribution with the ultimate goal of determining the posterior mean of expected number of putts at various green locations. A golfer's performance is assessed against the expected number of putts generated from the MCMC run. Finally, the statistic developed in this paper is compared to total putts and the strokes gained-putting statistic. The results indicate that the difficulty of a putt is influenced by both the distance and region from which the shot is taken.

This thesis is dedicated to the loving memory of Leili

"There was the Door to which I found no Key;
There was the Veil through which I might not see:
Some little talk awhile of ME and THEE
There was-and then no more of THEE and ME."

Omar Khayyam

Acknowledgments

As I start writing this section of my thesis, I realize that there are many individuals to whom I owe this achievement. Some of these individuals helped shape my statistical thinking and some others provided me with the mental support I needed to get to where I am today. It would only be fair to thank these people in perhaps the smallest way possible. That is, by mentioning their names in this section of my thesis and letting them know that I deeply appreciate their help and support.

I am forever grateful for having the opportunity to work under the supervision of Dr. Tim Swartz. I will always cherish the memories of working with you and learning from you throughout my time at SFU. Thank you for your kindness and faith in me.

Many thanks to Dr. Richard Lockhart and Dr. David Campbell for agreeing to be in my committee and providing great feedback on this project.

I would like to express my gratitude to the professors at SFU who taught me how to think statistically. I am thankful to Dr. Rachel Altman for constantly challenging me in and outside of her class. Thank you Dr. Joan Hu for your kindness and patience. I enjoyed the lectures of Dr. Richard Lockhart and I am grateful for having the chance of taking a class with him. Great thanks to Dr. Tom Loughin for teaching an amazing course on applied modern statistics. Many thanks to Marie Loughin for improving my statistical writing during my undergraduate degree at SFU and encouraging me to pursue my Master's degree. I would also like to thank Dr. Boxin Tang, Ian Berjovitz and Robin Insley for their support and help.

Many thanks to Saidka for helping me with urgent matters in the shortest time possible. I would also like to thank Kelly and Charlene for their help.

Thanks to my amazing friends in the department of Statistics and Actuarial Science at SFU with whom I shared many laughs and hardships. Fabian, you are an amazing friend and

I cannot thank you enough for always having the time to listen to me in my difficult times. Andrew, I consider you my friend and I will always cherish our interesting chats. I also owe you great thanks for taking the time to help me debug my code for this project at one point. Many thanks to my other fellow classmates for their friendship and support, especially to Jack, Rachel, Megan, Maria, Huijing, Shirin, Dilinuer, Harsha, Zhenhua, Biljana and Nate.

There are three individuals to whom I am indebted deeply. My aunt Floria, thank you for being so supportive of me. My cousin Naz, I will always appreciate your continued kindness and help since my first day in Vancouver. Khosrow, my mentor, thank you for sharing your experiences with me. Thank you for believing in me and listening to me. I am forever indebted to you and I will never forget your support through some of the most difficult times in my life.

Thanks to my parents for giving me their unconditional support and love despite being away from them for so many years. Thank you so much for all you have done for me.

Last but not least I would like to thank my brother, Reza. Reza, your unconditional love and invaluable support has led me to achieve this success. I cannot thank you enough for your encouragement, faith and trust in me. I simply would not have made this far without you. I hope I have made you proud.

Contents

Approval ii

Partial Copyright License iii

Abstract iv

Dedication v

Quotation vi

Acknowledgments vii

Contents ix

List of Tables xi

List of Figures xii

1 Introduction 1

1.1 Putts per Round Statistic 1

1.2 Strokes Gained-Putting Statistic 2

1.3 Objective 3

1.4 Project Outline 3

2 Bayesian Spatial Statistics Models 5

2.1 A Brief Note on Spatial Statistics 5

2.2 Modelling Number of Putts 7

2.3 Enhanced Spatial Putting Statistic based on Bayesian Spatial Models 10

3 Derivations and Computations **12**
3.1 MCMC Methods 13
3.1.1 The Gibbs Sampler 13
3.1.2 The Metropolis-Hastings Algorithm 14
3.2 Derivation of the Full Conditionals 15
3.3 Proposal Density Specifications 16

4 Test Data: The 2012 Honda Classic **20**
4.1 Description of the Data 20
4.2 Detailed Analysis of Hole One - Round Four 21
4.3 Analysis of Round Four 27

5 Discussion **30**

Bibliography **32**

List of Tables

3.1 Putting summaries from the 2012 PGA Tour. 17

4.1 Posterior means and posterior standard deviations for the secondary parameters of interest corresponding to the first hole of the fourth round of the 2012 Honda Classic. 23

4.2 Various putting statistics calculated for the fourth round of the 2012 Honda Classic. 27

List of Figures

2.1 The 8 pie slices with their corresponding slope parameter 9

3.1 The piecewise linear function of distance for putting angles of average difficulty with five knots. The distance r is measured in feet. 18

4.1 The number of putts taken by each of the 76 pro golfers on the first hole of round four of the 2012 Honda Classic. The x and y coordinates are measured in feet. 22

4.2 Posterior draws for σ following iteration 2500. The plot on the top gives the trace plot of σ. The plot on the bottom gives the histogram of σ. 24

4.3 Posterior draws for δ following iteration 2500. The plot on the top gives the trace plot of δ. The plot on the bottom gives the histogram of δ. Overall Metropolis acceptance rate: 32.5%. 24

4.4 Posterior draws for β following iteration 2500. The trace plots are given on the left and the histograms are given on the right hand side. The blue colour indicates a positive β posterior mean and the red colour indicates a negative β posterior mean. 25

4.5 The expected number of putts obtained using the spatial model for a selection of putting locations. 26

4.6 The pair-wise correlation plot for total putts, strokes gained (original) and strokes gained (spatial). 28

Chapter 1

Introduction

All statistics are meaningful in one way or another. When it comes to the world of sport, various statistics are used to evaluate the performance of players. However, some of these statistics are misleading and do not contain all the information necessary to evaluate the performance of players. This project will focus on providing a novel statistic to evaluate the putting performance of golfers as we believe that the current available statistics only provide partial information. To that end, we will first introduce the two commonly used statistics to evaluate the putting performance of golfers and subsequently discuss the shortcomings of these statistics. This motivates us to develop an enhanced statistic that improves upon the commonly used putting statistics.

1.1 Putts per Round Statistic

The putts per round statistic is the traditional statistic used with respect to putting. This metric is calculated by summing the number of putts (number of shots taken by a player from inside the green to get into the hole) for each of the 18 holes in an entire round. If a player chipped in from outside the green, he is credited zero putts for that hole.

This metric is a classic example of a statistic that is misleading as it fails to take the difficulty of putts into consideration. It is obvious that a golfer who hits many greens in regulation will face more difficult putts compared to a golfer who hits only a few greens in regulation and then chips close to the hole. In addition, the golfers who chip in from off the green are credited zero putts. Therefore, evaluation of putting based on total putts does not appear to be a sensible approach.

Another problem with this metric is its failure to properly distinguish between professional golfers. The PGA (Professional Golfers' Association) reports that in 2012 season the best putter was Jonas Blixt with an average of 27.89 putts per round and Boo Weekley was in the last place (191st) with an average of 30.5 putts per round. As can be seen, there is little to distinguish such professional golfers from one another using this metric.

1.2 Strokes Gained-Putting Statistic

The shortcomings with the putts per round statistic were recognized in a paper by Broadie (2008); as a result he developed a new putting statistic which was later extended by Fearing, Acimovic and Graves (2011). This statistic, which is often referred to as *strokes gained-putting* statistic, became popular and the PGA started reporting it in 2011. Thus far, it has been well received by players as it takes the difficulty of a putt into account. The strokes gained-putting statistic is a relative-value statistic as it quantifies how many strokes a golfer gains relative to other PGA golfers. Strokes gained-putting is often reported as the number of putts gained/lost per round. A positive strokes gained-putting value would indicate above average putting performance and a negative strokes gained-putting value would indicate below average putting performance. For instance, in 2012 season, Brandt Snedeker was the top putter with 0.860 strokes gained per round, and Kyle Thompsons was in the last place (191st) with 1.177 strokes lost per round (or -1.177 strokes gained per round).

According to the PGA, the strokes gained value is determined by subtracting the actual number of putts taken by a player from the average number of putts by a PGA tour player to hole out from the initial location that the putt was taken inside the green. The distances to the holes are obtained using the ShotLink data and the averages are calculated from all putts from the previous season. For instance, the ShotLink data in 2010 shows that the average number of putts it took professional golfers to hole out from seven feet was 1.5 putts. Thus, if a pro golfer took one putt to hole out from 7 feet in 2011 season, his strokes gained value would be 0.5. Further, a golfer's performance is evaluated against the field for an entire round. That is, if the player over the course of the round gained two putts and the average strokes gained for that round was one putt, his strokes gained-putting against the field for that round would be one.

The strokes gained-putting contains more information compared to the traditional statistic used. It allows us to determine the effectiveness of a pro golfer against the field. In addition, it provides information on how easy or difficult a course is compared to other courses. However, the main shortcoming of this approach is that it only considers the distance from the hole to determine the difficulty of a putt and ignores other potential contributing factors.

1.3 Objective

The main goal of this project is to construct an enhanced relative-value statistic that improves upon the strokes gained-putting statistic by taking factors other than distance into account to determine the expected number of putts from various locations on a green. This new metric will feature a spatial aspect that considers the region of the green from which the shot is taken. The greens in various courses often have uneven shapes and it would be sensible to investigate whether taking a shot from certain regions of the green would be more or less difficult than other regions of the green.

The methodology developed in this project borrows ideas from Bayesian spatial statistics. Spatial maps of each green are constructed which provide the expected number of putts from various green locations. The Bayesian part of our model uses a hierarchical structure in which prior distributions are required. The ideas found in Cressie (1993) for geostatistical models, and Banerjee, Carlin, and Gelfand (2004) for Bayesian spatial models are relevant to the model developed in this project. The difficulty of a putt in our model will be a function of both its direction and distance to the hole. Using such a model, the performance of golfers is assessed by comparing the observed number of putts with the expected number of putts. Again, ShotLink data is used to determine the initial location of putts on greens.

1.4 Project Outline

In Chapter 2, a Bayesian spatial model for putting is developed. As this model is non-trivial, the computations utilize Markov chain Monte Carlo (MCMC) methodology for simulation from the posterior distribution. Then, in Chapter 3, the necessary derivation and computations of the relevant full conditional distributions are discussed. It is then seen that the Metropolis within Gibbs algorithm needs to be used as the full conditional distributions are not tractable. In Chapter 4, 2012 Honda Classic data obtained from the ShotLink website is

used as test data for this project. The model developed in this project is then applied to the ShotLink data and the values of our statistic are obtained. We then compare our statistic to the total putts per round statistic and the strokes gained-putting statistic. Finally, Chapter 5 will discuss the implications of the model developed in this paper and will provide a list of potential avenues for future investigation. A more concise companion paper of this project is given by Yousefi and Swartz (2013).

Chapter 2

Bayesian Spatial Statistics Models

This chapter considers the modelling approach required to construct spatial maps that provide estimates of the expected number of putts by a pro golfer from each of the green location realizations. To construct such maps, Bayesian hierarchical modelling is utilized. Such modelling allows us to capture various features of the data and parameters conditionally, one layer of the hierarchy at a time. In this chapter, we will first give a brief discussion of spatial statistics and then we will model the number of putts for our application using spatial statistics.

2.1 A Brief Note on Spatial Statistics

The first law of geography by Tobler (1970) states that "Everything is related to everything else, but near things are more related than distant things". Statistically speaking, this law suggests the existence of positive spatial correlation and weakens the independence assumption of observations (Waller, 2005). This law underlies the importance of considering spatial features of data in various fields. In short, spatial statistics consists of statistical methodologies that consider location and distance when providing inference. The field of spatial statistics has grown in popularity in recent years mainly due to the increase in computing power and the availability of spatial data. Further, as Waller (2005) notes, there has been a great appeal to use Bayesian methods to analyse spatial data primarily through utilizing hierarchical structures. It should be noted that as in other statistical fields, different data structures exist in spatial statistics. Each data structure can help answer a specific inferential question and the analyses will differ based on the data structure available.

Cressie (1993) categorizes spatial data into three types based on the study design and the inferential questions of interest. These three categories are spatial point process, geostatistical, and lattice data.

Spatial point process data consist of observations in a specified study region. In this data structure, locations are realizations of a random process for which we seek inference. With such data, one can answer whether observations are similar at all locations and provide point estimates for the defined study region.

Geostatistical data consist of measurements taken at a fixed number of sites. Note that the sites are chosen by the study design and are not random. Geostatistical data can help researchers predict the outcome at sites where no measurement is taken. For instance, suppose that we take ozone level measurements in Burnaby and then would like to use the data collected in Burnaby to predict the ozone level in Coquitlam where no measurement is taken. The modelling approach required to predict the ozone level measurement in Coquitlam is often referred to as geostatistical modelling. Diggle and Ribeiro Jr. (2007), in their book *Model-based Geostatistics*, give an extensive look regarding such models.

Finally, lattice (also known as regional) data consist of summary measurements for a specified region such as number of deaths due to lung cancer in British Columbia. Note that lattices can have regular or irregular shapes. The analyses relating to such data provide us with accurate estimates of regions with small sample sizes. The idea here is to use empirical Bayesian models (naive and simplistic) or even fully Bayesian models to give such accurate estimates.

It is important to determine the data structure category for the PGA tour putting data. The PGA tour putting data consist of observations in a defined study region (a particular green in a particular round). In addition, the locations are realizations of a random process for which we seek inference. Thus, our data falls into the spatial point process data category. As a result, using the appropriate modelling approach, we will be able to investigate putting proficiency at various green locations by providing the expected number of putts for the set of observations inside a specific green in a specific round. However, one key difference of our model that distinguishes it from other spatial problems is that we only require point estimates (expected number of putts) at the realized locations where putts have taken place.

2.2 Modelling Number of Putts

As we are interested in the expected number of putts, the very first step is to assign a distribution to them. Now, consider a specific green in a specific round of a tournament. Let Z_i denote the number of putts it takes the ith pro golfer to hole out from his initial putting green location where $i = 1, \ldots, n$ and n is the number of players who competed during that round of the tournament. With detailed ShotLink data, we are able to extract the initial putting coordinates (x_i, y_i) on a green for the ith golfer where $i = 1, \ldots, n$. Note that the hole is defined as the origin (0,0) and (x_i, y_i) are the Cartesian coordinates measured in feet relative to the initial putting location for the ith golfer. In addition, let (r_i, θ_i) denote the polar representation of (x_i, y_i). In other words, $r_i^2 = x_i^2 + y_i^2$ and $y_i = r_i sin(\theta_i)$. Suppose further that the data $Z_1, \ldots, Z_n$ are independent with the following distribution:

$$Z_i - 1 \sim \text{Poisson}(exp\{\lambda_i(r_i, \theta_i)\}) \tag{2.1}$$

for $i = 1, \ldots, n$. It can then easily be seen from (2.1) that $E(Z_i) = 1 + e^{\lambda_i}$ for the ith golfer which reasonably depends on the initial putting location (r_i, θ_i).

At first glance, it is seen that the Poisson distribution is a good choice as it is tractable and its support matches our application. However, one major problem with using the Poisson distribution is that λ_i belongs to $\mathbb{R}$ which is not sensible for our application. For instance, in the case that $\lambda_i = 0$, $Pr(Z_i = 1) = Pr(Z_i = 2) = .37$ indicating that two-putts are as common as one-putts from a putting location where the expected number of putts is two. All golfers would agree that for such a putt, the probability of two-putting should exceed the probability of one-putting. To address this issue, we consider a variation of the model in (2.1) where $Z_1, \ldots, Z_n$ are assumed independent with

$$Z_i - 1 \sim \text{truncated-Poisson}(exp\{\lambda_i(r_i, \theta_i)\}) \tag{2.2}$$

For our application, we impose the truncation such that $Pr(Z_i \geq 4) = 0$ for $i = 1, \ldots, n$. The truncated-Poisson seems to be sensible for the PGA tour putting application. It can be seen that $lim_{\lambda \to \infty} Pr(Z = 3|\lambda) = 1$ which indicates that there are locations on the green where the probability of three-putting approaches one. Further, $lim_{\lambda \to -\infty} Pr(Z = 1|\lambda) = 1$ which indicates that there are green locations (possibly very close to the hole) where the probability of one-putting approaches one.

The next layer of the hierarchical modelling requires specifying the distribution of $\lambda = (\lambda_1, \ldots, \lambda_n)'$. We borrow ideas from Diggle et al. (1998), and Besag et al. (1991) to propose

the following distribution for λ:

$$\lambda \sim MVN(\mu, \sigma^2 V) \tag{2.3}$$

where $\mu = (\mu_1, \ldots, \mu_n)'$. Further, let $V = (v_{ij})$ be the Gaussian covariance function such that

$$v_{ij} = exp\{-\delta^2 ||(x_i, y_i) - (x_j, y_j)||^2\} \tag{2.4}$$

where $||.||$ denotes the Euclidean distance and $\delta > 0$. The specification of the variance-covariance matrix in (2.4) is a common choice (Bannerjee et al., 2004) as it guarantees positive-definiteness. The diagonal elements of V equal one, and for distant spatial locations, $v_{ij} \to 0$. In addition, the Gaussian covariance function defined in (2.4) assigns greater correlation to parameters λ_i and λ_j that are spatially close while it assigns smaller correlation to parameters λ_i and λ_j that are spatially distant.

The next step in our modelling approach is to specify a prior distribution for the vector μ in (2.3). The main idea here is that a putt on a given line to the pin should have a smaller apriori probability of being made compared to a shorter putt along the same line. As discussed, λ_i relates to the probability of a putt being made from the initial putting location (x_i, y_i) such that smaller λ_i values indicate less difficult putts. Further, μ_i represents the prior mean of λ_i. In our model, we divide each green into 8 "pie slices" emanating from the hole as illustrated in Figure 2.1. The polar covariate θ_i represents the region in which the putt of the ith golfer resides. Such an approach allows for putts to share common features within the same region. Further, the relative difficulty of putts within the same region is influenced by the distance r_i of the putt to the hole. Thus, we define μ_i as follows:

$$\mu_i = g(r_i + \beta^{(\theta_i)} r_i) \tag{2.5}$$

where g is an increasing piecewise linear function and $\beta^{(\theta_i)}$ is mapped to one of $\beta_1, \ldots, \beta_8$ with respect to the slice corresponding to θ_i. Note that $\beta_1, \ldots, \beta_8$ represent different slopes and account for varying difficulty between the 8 putting angles. After some experimentation with respect to the number of observations available for a particular green, it was found that 8 slices is small enough to provide stable parameter estimation and is large enough to yield realistic varying difficulty of putting angles. The specification of μ_i in (2.5) indicates that shorter putting distances r_i lead to smaller values of μ_i and longer putting distances r_i lead to larger values of μ_i. Further, (2.5) provides a nice interpretation for β. A slice

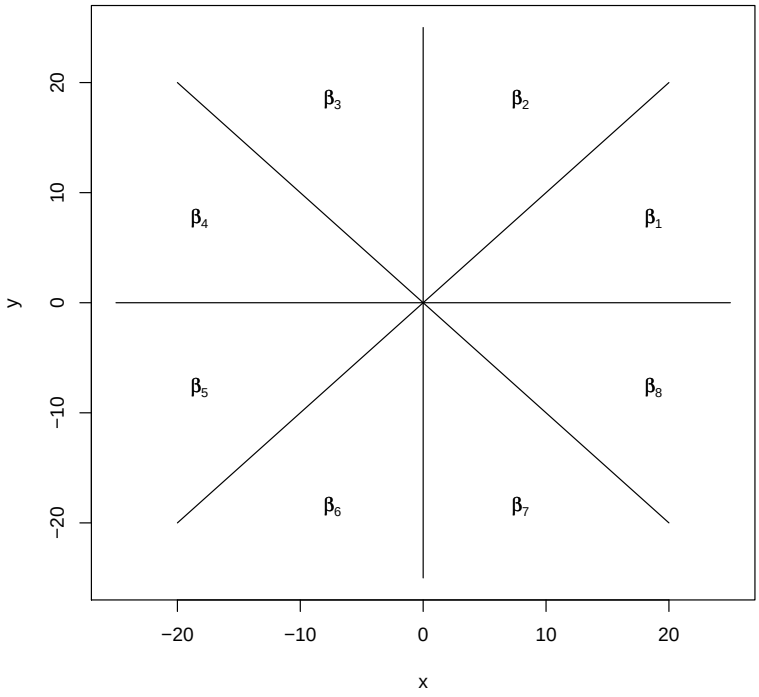

Figure 2.1: The 8 pie slices with their corresponding slope parameter

of normal putting difficulty has $\beta = 0$ while $\beta > 0$ and $\beta < 0$ denote more difficult and less difficult putting regions, respectively. For instance, $\beta = -0.15$ represents a putt that is equivalent in difficulty to a normal putt shortened in length by 15%. The specification of knots for the piecewise linear function g will be given in detail in Chapter 3.

We now assign a distribution to $\beta_1, \ldots, \beta_8$ to complete the modelling requirements. Assume $\beta_1, \ldots, \beta_8$ are independent with

$$\beta_j \sim \text{Normal}(0, \sigma_\beta^2) \tag{2.6}$$

where $j = 1, \ldots, 8$. The hyperparameter σ_β is fixed at 0.2 to cover a wide range of possible β_j values. Note that the β_js are centred at 0 (the average value), with differences accounting

for more and less difficult putting regions. Further, we set:

$$\begin{aligned} [\sigma] &\propto 1/\sigma \\ [\delta] &\propto 1 \end{aligned} \tag{2.7}$$

where $[\cdot]$ is standard notation for the probability density function. The distributions in (2.7) are reference priors. Both of the unknown parameters are constrained to the positive real line.

2.3 Enhanced Spatial Putting Statistic based on Bayesian Spatial Models

As discussed before, our ultimate goal is to determine the expected number of putts from the green location realizations. Here, the metric to evaluate the putting performance of a golfer based on the Bayesian spatial model developed in this chapter is given.

We define the performance measure of the ith golfer on a particular hole for a particular round of a golf tournament as

$$\hat{E}(Z_i|\lambda_i) \approx E(Z_i|\hat{\lambda}_i) - Z_i \tag{2.8}$$

such that $E(Z_i|\hat{\lambda}_i)$ is obtained under (2.2) as follows where $\hat{\tau}_i = exp\{\hat{\lambda}_i\}$:

$$\begin{aligned} E(Z_i - 1|\hat{\lambda}_i) &= \frac{\sum_{z_i=1}^{3}(z_i-1)\hat{\tau}_i^{z_i-1}e^{-\hat{\tau}_i}/(z_i-1)!}{\sum_{z_i=1}^{3}\hat{\tau}_i^{z_i-1}e^{-\hat{\tau}_i}/(z_i-1)!} \\ &= \frac{0\times\hat{\tau}_i^0e^{-\hat{\tau}_i}/0! + 1\times\hat{\tau}_ie^{-\hat{\tau}_i}/1! + 2\times\hat{\tau}_i^2e^{-\hat{\tau}_i}/2!}{\hat{\tau}_i^0e^{-\hat{\tau}_i}/0! + \hat{\tau}_ie^{-\hat{\tau}_i}/1! + \hat{\tau}_i^2e^{-\hat{\tau}_i}/2!} \\ &= \frac{e^{-\hat{\tau}_i}(\hat{\tau}_i + \hat{\tau}_i^2)}{e^{-\hat{\tau}_i}(1 + \hat{\tau}_i + \hat{\tau}_i^2/2)} \\ &= \frac{\hat{\tau}_i(1+\hat{\tau}_i)}{1 + \hat{\tau}_i + \hat{\tau}_i^2/2} \end{aligned}$$

Thus, we have that $E(Z_i|\hat{\lambda}_i)$ is:

$$E(Z_i|\hat{\lambda}_i) = 1 + \frac{\hat{\tau}_i(1+\hat{\tau}_i)}{1 + \hat{\tau}_i + \hat{\tau}_i^2/2} \tag{2.9}$$

It can immediately be seen that the metric in (2.8) gives relative strokes gained on the hole under our model. Further, a positive value of this metric is interpreted as above average putting performance while a negative value indicates below average putting performance on the hole. Note that a player who chipped-in will not have a relative strokes gained value for that hole. This is a sensible approach as the player who chips-in does not make a putting attempt.

The proposed enhanced putting statistic for an entire round of golf for the ith pro golfer would involve a summation of (2.8) over all 18 holes (if he did not chip-in) in that particular round. The enhanced putting statistic for a given tournament would involve summation of 72 terms (18 holes $\times$ 4 rounds) assuming there were no chip-ins. Similarly, season averages of pro golfers can be calculated. Finally, the enhanced putting statistic is adjusted against the field as is done for strokes gained-putting statistic which was discussed in section 1.2. Such an adjustment will allow us to compare strokes gained-putting and our spatial putting statistic properly.

Finally, it should be noted that we only require the expected number of putts at a green location realization unlike other spatial problems (Cressie, 1993) as discussed in section 2.1. This leads to a simplification of our inferential problem. In the next chapter, we will fit the models specified here in an attempt to determine the posterior density of the expected number of putts.

Chapter 3

Derivations and Computations

In this chapter, we derive the posterior density of the expected number of putts using the specified distributions in Chapter 2. The notation $[B \mid A]$ denotes the density (or probability mass function) of B given A. Accordingly, we have:

$$\begin{aligned}[\lambda, \beta, \sigma, \delta \mid Z] &= [Z \mid \lambda, \beta, \sigma, \delta] \cdot [\lambda, \beta, \sigma, \delta] \\ &= [Z \mid \lambda] \cdot [\lambda \mid \beta, \sigma, \delta] \cdot [\beta, \sigma, \delta].\end{aligned}$$

In the above expression $[Z \mid \lambda, \beta, \sigma, \delta]$ reduces to $[Z \mid \lambda]$ as Z follows the truncated-Poisson distribution (2.2) and only depends on λ. Further, we make the reasonable assumption that β, σ and δ are independent from one another. Thus, we have:

$$[\lambda, \beta, \sigma, \delta \mid Z] = [Z \mid \lambda] \cdot [\lambda \mid \beta, \sigma, \delta] \cdot [\beta] \cdot [\sigma] \cdot [\delta] \tag{3.1}$$

Note that the above posterior density has $n + 10$ dimensions as λ has n dimensions, β has 8 dimensions, σ has one dimension and δ has one dimension. In the next step, we replace the densities in (3.1) with the parametric distributions specified in Chapter 2. Therefore, (3.1) becomes:

$$\begin{aligned}[\lambda, \beta, \sigma, \delta \mid Z] \quad \propto \quad & \prod_{i=1}^{n} \frac{e^{\lambda_i (Z_i - 1)} e^{-e^{-\lambda_i}}}{e^{-e^{-\lambda_i}}(1 + e^{\lambda_i} + e^{2\lambda_i}/2)} \cdot \frac{e^{-\frac{1}{\sigma^2}(\lambda-\mu)' V^{-1} (\lambda-\mu)}}{\sigma^n |V|^{1/2}} \\ & \cdot \prod_{j=1}^{8} e^{-\frac{1}{2\sigma_\beta^2}(\beta_j - \beta_0)^2} \cdot \frac{1}{\sigma}\end{aligned}$$

$$= \prod_{i=1}^{n} \frac{e^{\lambda_i(Z_i-1)}e^{-e^{-\lambda_i}}}{e^{-e^{-\lambda_i}}(1+e^{\lambda_i}+e^{2\lambda_i}/2)} \cdot \frac{e^{-\frac{1}{\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)}}{\sigma^{n+1}|V|^{1/2}} \cdot \prod_{j=1}^{8} e^{-\frac{1}{2\sigma_\beta^2}(\beta_j-\beta_0)^2} \tag{3.2}$$

where V and μ_i are given in (2.4) and (2.5), respectively.

It can be seen from (3.2) that the complexity and dimensionality of the posterior density prevents straightforward interpretation. Hence, posterior summaries such as posterior means or posterior standard deviations are required for interpretation. However, these quantities are not tractable and cannot be derived analytically in our model. Further, obtaining analytic approximations to the posterior density does not appear to be feasible. As a result, we turn to Monte Carlo integration techniques that are easy to program and can often handle high-dimensional models with ease. We will use Markov chain Monte Carlo (MCMC) methods to approximate the posterior summaries in our model. To that end, let us give a brief introduction to MCMC methods and then discuss our computational approach.

3.1 MCMC Methods

It is quite common in Bayesian statistics to come across parameters of interest that take the form of intractable integrals. Further, in the case of high dimensional models, it is standard practice to utilize MCMC methods. In short, MCMC methods allow us to simulate draws that are approximately from the desired posterior density. It is then easy to approximate the posterior summaries using these draws. However, the main issue that requires our attention is to ensure that convergence to the desired posterior density is achieved. In practice, this is accomplished by judging the sampled output through use of plots and numerical summaries (Carlin and Louis, 2009). Here, we will discuss only the two most common MCMC algorithms. These algorithms are the Gibbs sampler and Metropolis-Hastings.

3.1.1 The Gibbs Sampler

In a Gibbs sampling algorithm, we need the full conditional distributions of the model parameters. Full (or complete) conditional distributions are the distributions of the parameters of interest conditioned on all the other parameters and the data Z. For our model, they are

derived in section 3.2. As Carlin and Louis (2009) note, it has been shown that under mild conditions, the knowledge of full conditional distributions allow us to determine the joint density uniquely. Suppose now that we have a model with l parameters, $a = a_1, \ldots, a_l$. Let $[a_i \mid \cdot\,]$ for $i = 1, \ldots, l$ denote the full conditional distribution for the ith parameter. The Gibbs sampling algorithm can be implemented if the full conditional distributions are known densities (e.g. Normal) or it is easy to generate samples from them through methodologies such as rejection sampling. The description of the Gibbs sampling algorithm is as follows:

The Gibbs Sampling Algorithm:

1. Set starting values $a_2^{(0)}, \ldots, a_l^{(0)}$
2. For $s = 1, \ldots, S$, repeat:
   - Generate $a_1^{(s)}$ from $[a_1 \mid a_2^{(s-1)}, a_3^{(s-1)}, \ldots, a_l^{(s-1)}, z]$
   - Generate $a_2^{(s)}$ from $[a_2 \mid a_1^{(s)}, a_3^{(s-1)}, \ldots, a_l^{(s-1)}, z]$
   - $\vdots$
   - Generate $a_l^{(s)}$ from $[a_l \mid a_1^{(s)}, a_2^{(s)}, \ldots, a_{l-1}^{(s)}, z]$

We can then use the generated draws to estimate posterior quantities of interest.

3.1.2 The Metropolis-Hastings Algorithm

Suppose that some or all of the full conditional distributions are unfamiliar densities. Then, the Gibbs sampling algorithm cannot be used and we consequently turn to Metropolis-Hastings which is applicable when some of the underlying densities of the full conditional distributions are unknown. One major part of Metropolis-Hastings involves specifying a candidate (proposal) distribution. Evans and Swartz (2000) state that a good candidate distribution would wander around the support of the target density unpredictably and completely. Let $q(x \mid y)$ denote the proposal density which specifies the probability density of going to x given that the chain is currently at y. For a one-dimensional parameter a whose full conditional distribution is an unknown density, the implementation of the Metropolis-Hastings algorithm is as follows:

The Metropolis-Hastings Algorithm:

1. Set an initial value $a^{(0)}$
2. For $s = 1, \ldots, S$, repeat:
   - Generate $a^\star$ from $q(\cdot \mid a^{(s-1)})$
   - Compute the acceptance ratio $AR = \frac{[a^\star|z]q(a^{(s-1)}|a^\star)}{[a^{(s-1)}|z]q(a^\star|a^{(s-1)})}$
   - Accept $a^\star$ as $a^{(s)}$ with probability $\min(AR, 1)$. Otherwise, set $a^{(s)} = a^{(s-1)}$ if $a^\star$ is not accepted.

In the case that the proposal distribution is symmetric, the acceptance ratio reduces to $[a^\star \mid z]/[a^{(s-1)} \mid z]$. Further, it is quite common in practice to choose a proposal density such that $q(a^\star \mid a^{(s-1)}) = q(a^\star)$. These proposal densities are known as independent proposals. It is also instructive to assess the acceptance rate (the fraction of proposal draws that are accepted) of the Metropolis-Hastings algorithm. An acceptance rate between 0.25 and 0.5 for a random walk is often desirable (Robert and Casella, 2009).

With these MCMC methods at our disposal, we now turn to the joint posterior density of our model given in (3.2). The full conditional distributions of our parameters are derived in the next section.

3.2 Derivation of the Full Conditionals

We now apply simple algebra to obtain the full conditional distributions of our model from the joint posterior density given in (3.2) as follows:

$$\begin{array}{lcl} [\lambda_i \mid \cdot\,] & \propto & \frac{e^{\lambda_i(Z_i-1)}}{(1+e^{\lambda_i}+e^{2\lambda_i}/2)} \cdot \exp\{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)\} \\ [\beta_j \mid \cdot\,] & \propto & \prod_{j=1}^{8} \exp\{-\frac{1}{2\sigma_\beta^2}(\beta_j-\beta_0)^2\} \cdot \exp\{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)\} \\ \left[\sigma^2 \mid \cdot\,\right] & \sim & \text{Inverse-Gamma}\left(\frac{n-1}{2}, \frac{2}{(\lambda-\mu)'V^{-1}(\lambda-\mu)}\right) \\ [\delta \mid \cdot\,] & \propto & |V|^{-1/2} \exp\{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)\} \end{array} \tag{3.3}$$

Note that σ^2 follows the Inverse-Gamma since:

$$
\begin{aligned}
[\sigma^2 \mid \cdot] &\propto \frac{\exp\{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)\}}{\sigma^{n+1}} \\
&\propto \left(\frac{(\lambda-\mu)'V^{-1}(\lambda-\mu)}{2}\right)^{\frac{n-1}{2}} \cdot \frac{1}{\Gamma(\frac{n-1}{2})} \cdot \frac{\exp\{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)\}}{(\sigma^2)^{\frac{n-1}{2}+1}} \\
&\sim \text{Inverse-Gamma}\left(\frac{n-1}{2}, \frac{2}{(\lambda-\mu)'V^{-1}(\lambda-\mu)}\right)
\end{aligned}
$$

We observe that drawing samples from σ is straightforward. It is simply done by generating random variates v from the corresponding Gamma distribution and then setting $\sigma = 1/\sqrt{v}$. The other distributions in (3.3) are unknown. Hence, we use the so-called *Metropolis within Gibbs algorithm* (Gilks, Richardson and Spiegelhalter, 1996) as our implementation involves a Gibbs step (generating samples from σ) and Metropolis-Hastings steps for generating samples from the posterior densities of the other parameters. As discussed in section 3.1.2, the Metropolis-Hastings steps require the specification of proposal densities. Thus, we shift our focus to introduce reasonable proposal densities for the unknown parameters.

3.3 Proposal Density Specifications

To simulate λ_i draws, the 2012 PGA tour putting data up to and including the Ryder Cup on September 30, 2012 is considered. The data are summarized in Table 3.1 which were obtained from the website www.pgatour.com. Note that the table gives the putting performance by pro golfers at a distance of r feet from the hole and the resultant expected number of putts. Recall from section 2.3 that the expected number of putts is given by $E(Z|\lambda) = 1 + \tau(1+\tau)/(1+\tau+\tau^2/2)$ where $\tau = e^{\lambda}$. Further, λ depends on the distance to the hole r and the directional angle θ. To determine plausible λ values, we solve the above expression for τ. This yields:

$$
\tau = \frac{2 - E(Z \mid \lambda) \pm \sqrt{(E(Z \mid \lambda) - 2)^2 - 2\,(E(Z \mid \lambda) - 3)(E(Z \mid \lambda) - 1)}}{E(Z \mid \lambda) - 3} \tag{3.4}
$$

Putting Distance r (in feet)	Proportion of One-Putts	Two-Putts	Three-Putts	E(Z)	λ
07.5	0.554	0.441	0.005	1.45	-0.722
12.5	0.298	0.694	0.008	1.71	-0.165
17.5	0.180	0.804	0.016	1.84	0.071
22.5	0.114	0.861	0.025	1.91	0.192

Table 3.1: Putting summaries from the 2012 PGA Tour.

We can use (3.4) and the E(Z) values in the fifth column of Table 3.1 to obtain values of λ. Note that E(Z) in Table 3.1 represents average putting conditions. The computed λ values are given in the sixth column of Table 3.1. Now, recall that $\mu_i = g(r_i + \beta^{(\theta_i)} r_i)$ is the mean of λ_i as defined in (2.5). For instance, when $r = 7.5$, we have $-0.722 = g(7.5)$. We now use the values given in Table 3.1 to obtain the knots for the piecewise linear function g for putting angles of average difficulty. We have:

$$\mu_i = g(r_i + \beta^{(\theta_i)} r_i) = \begin{cases} -4.600 + 0.705(r_i - 2.0 + \beta^{(\theta_i)} r_i) & 2.0 \le r_i < 7.5 \\ -0.722 + 0.111(r_i - 7.5 + \beta^{(\theta_i)} r_i) & 7.5 \le r_i < 12.5 \\ -0.165 + 0.047(r_i - 12.5 + \beta^{(\theta_i)} r_i) & 12.5 \le r_i < 17.5 \\ 0.071 + 0.024(r_i - 17.5 + \beta^{(\theta_i)} r_i) & 17.5 \le r_i < 22.5 \\ 0.192 + 0.019(r_i - 22.5 + \beta^{(\theta_i)} r_i) & 22.5 \le r_i \le 40.0 \end{cases}$$

where observations $r_i < 2$ are set to $r_i = 2$ and observations $r_i > 40$ are set to $r_i = 40$. The piecewise linear function g gives a subjective estimate of the mean for a putt with an average difficulty ($\beta^{(\theta_i)} = 0$). Further, the slopes are obtained by using the endpoints. For instance, the slope for the first knot is obtained from $(-0.722 + 4.600)/(7.5 - 2) = 0.705$. The plot of the piecewise linear function is provided in Figure 3.1 where $\beta^{(\theta_i)} = 0$.

It should be noted that other than the piecewise linear function g, various functional forms were considered. However, it was found that the function g gives a preferable fit given the knots. Further, distance and angle of the putt are incorporated nicely into the function. Recall from section 2.2 that $\beta^{(\theta_i)} r_i$ has an appealing interpretation. For instance, $\beta = -0.15$ represents a putt that is equivalent in difficulty to a standard putt shortened in length by 15%.

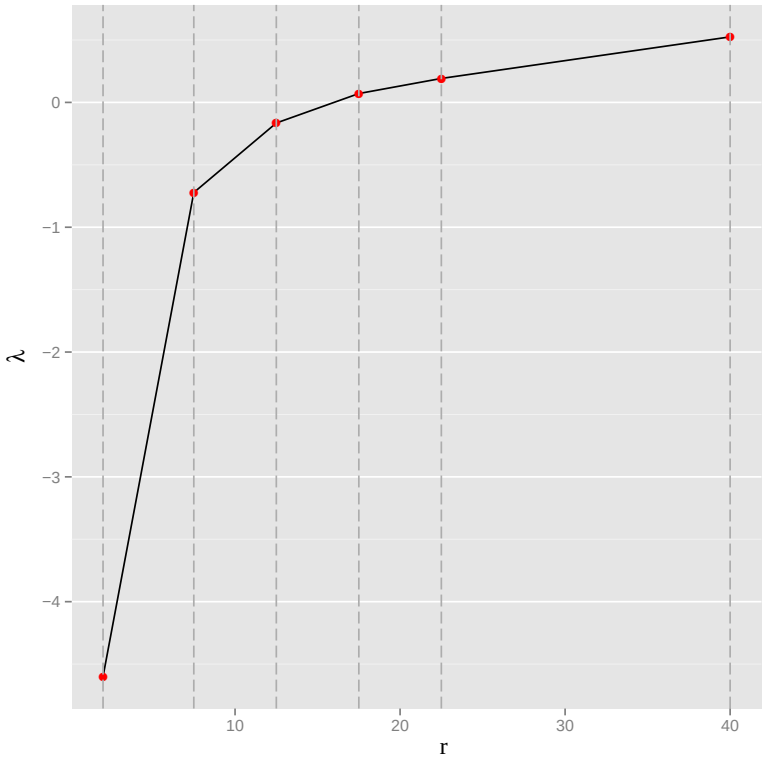

Figure 3.1: The piecewise linear function of distance for putting angles of average difficulty with five knots. The distance r is measured in feet.

Now, we use the information given in Table 3.1 to give a proposal density for λ_i. For instance, when $r = 7.5$, we have $\tau = 0.486$, and as a result $\lambda = -0.722$. Thus, in the case that $r_i = 7.5$ corresponding to the ith pro golfer, we consider $\lambda_i \sim \text{Normal}(-0.722, 0.4)$ as the proposal density where the variance is relatively large to cover plausible λ_i values. Note that Normal distribution is symmetric and this choice of proposal density leads to a simplification of the acceptance ratio in the Metropolis-Hastings step.

Next, we consider $\text{Normal}(0, \sigma_\beta^2)$ as the proposal distribution for β_j. The $\text{Normal}(0, \sigma_\beta^2)$ is also the prior distribution for β_j, $j = 1, \ldots, 8$ as defined in (2.6).

Finally, for generation of random variates from δ, we take the constraint that $\delta > 0$ into account. Therefore, we use a proposal density of $\text{Gamma}(0.25, 1)$. This attempts to cover aposteriori plausible values of δ near 0.25. The variance is large enough to cover a wide range of δ values.

Having specified the model, the algorithm, and the proposal densities, we move on to fit this model to the test data obtained from the ShotLink website. In the next chapter, we give a description of the data and the steps taken to transform it into the required format. We fit our model and obtain the putting statistic arising from our model. The spatial putting statistic is then compared to the commonly used putting statistics discussed in Chapter 1.

Chapter 4

Test Data: The 2012 Honda Classic

In this chapter, we first give a description of the data. Next, we discuss the early difficulties with the data and the steps taken to transform it into the required format. We then fit the models discussed in Chapter 3 and assess the convergence of the parameters. Finally, we compute the putting statistic arising from our model and compare it to total putts per round and the strokes gained-putting statistic.

The test data in our analysis were obtained from the ShotLink website and consist of the results from the 2012 Honda Classic held at the PGA National Championship course. This tournament took place on March 1-4, 2012 in Palm Beach Gardens, Florida.

4.1 Description of the Data

The data consist of all the shots taken by players who competed in the tournament and give the x and y coordinates of the shots along with the location from which the shot was taken (e.g. fairway or green). Further, the distance to the hole is provided. However, the number of putts is not provided in the data. As a result, an R program was written to extract the total number of putts taken and the starting location (x, y) of the first putt on the green for each hole and each golfer.

One other missing piece in the data are the coordinates of the holes in the tournament. It was found that the hole coordinates were not recorded. We used the Euclidean distance to determine the coordinates of the hole. This is possible since we have the x and y coordinates of at least two separate shots and their distance from the hole. Newton's Method for solving a pair of equations was used to find the coordinates of the hole. To illustrate this better, suppose the coordinates of the hole are (x_h, y_h). Further, consider the coordinates of two

other putts as (x_1, y_1) and (x_2, y_2) that are distances d_1 and d_2 from the hole, respectively. Note that (x_1, y_1), (x_2, y_2), d_1 and d_2 are all known from the data. Consequently, we have the following equations:

$$f_1(x_h, y_h) = (x_1 - x_h)^2 + (y_1 - y_h)^2 - d_1^2 = 0$$
$$f_2(x_h, y_h) = (x_2 - x_h)^2 + (y_2 - y_h)^2 - d_2^2 = 0$$

We then arrange them into the following form:

$$F(\upsilon) = \begin{pmatrix} f_1(x_h, y_h) \\ f_2(x_h, y_h) \end{pmatrix} = \begin{pmatrix} (x_1 - x_h)^2 + (y_1 - y_h)^2 - d_1^2 \\ (x_2 - x_h)^2 + (y_2 - y_h)^2 - d_2^2 \end{pmatrix} = 0$$

where $\upsilon = (x_h, y_h)'$

Newton's Method can then be easily applied to the above equation. After applying Newton's Method, the hole coordinates were found within six digits of accuracy. Once the hole coordinates were found, we transformed the coordinates of all the shots for a particular hole so that they would be centred around zero.

Having transformed the data into the required format, we fit our model. For illustration purposes, we only analyze the first hole of the fourth round of the tournament in detail and then discuss the overall results of our model for the entire fourth round.

4.2 Detailed Analysis of Hole One - Round Four

Figure 4.1 gives the number of putts taken by each of the 76 pro golfers on the first hole of round four from the 2012 Honda Classic. It is observed that the probability of a one-putt increases as the length of the putt from the hole decreases. We further observe that putts in the first quadrant seem to be more difficult compared to other quadrants as three-putts are more common in the first region compared to other regions.

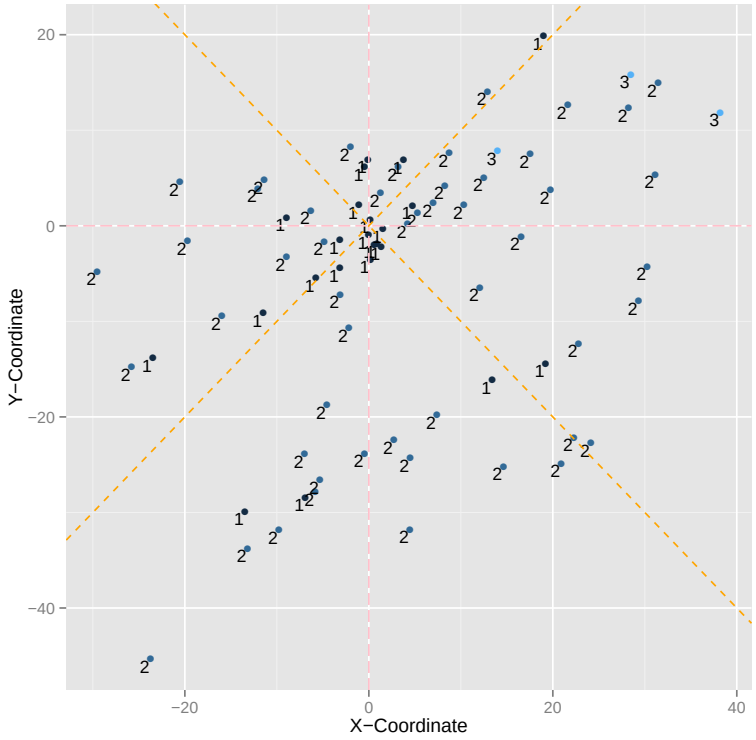

Figure 4.1: The number of putts taken by each of the 76 pro golfers on the first hole of round four of the 2012 Honda Classic. The x and y coordinates are measured in feet.

Next, we use the data from the first hole of the fourth round to fit our spatial model. Table 4.1 gives the Markov chain estimates for δ, σ and β. The Metropolis within Gibbs implementation was run for 25000 iterations which took approximately 10 hours of computation on a Mac Pro workstation. Further, we used the first 2500 iterations as burn-in. We observe from Table 4.1 that the largest β value is β_1 which corresponds to the first region. This is in agreement with our original belief that the first quadrant is more difficult than other quadrants. It is also observed that the posterior standard deviations for all secondary parameters are relatively small compared to their posterior means. This is desirable since it conveys that there is information in the data regarding these parameters.

Parameter	Post Mean	Post Std Dev
δ	0.082	0.021
σ	2.384	0.856
β_1	0.151	0.076
β_2	-0.084	0.043
β_3	-0.063	0.039
β_4	-0.078	0.049
β_5	-0.081	0.043
β_6	0.079	0.061
β_7	0.088	0.057
β_8	0.087	0.048

Table 4.1: Posterior means and posterior standard deviations for the secondary parameters of interest corresponding to the first hole of the fourth round of the 2012 Honda Classic.

We further assess the convergence of the secondary parameters by use of plots. Figure 4.2 gives the trace plot and histogram of σ draws from the Inverse-Gamma. We see that it is moving around quite well and there is no indication of lack of convergence..

Figure 4.3 gives the trace plot and histogram of δ draws. The acceptance ratio was found to be 32.5% which is reasonable. We see that it covers a plausible range and we see no indication as to lack of convergence.

We finally assess the convergence of β draws. Figure 4.4 summarizes the trace plots and histograms for $\beta_1, \ldots, \beta_8$. The acceptance ratio was found to be 25.5 % which is reasonable. We further see that there is no indication of lack of convergence. The histograms again suggest that the first quadrant is a more difficult putting direction.

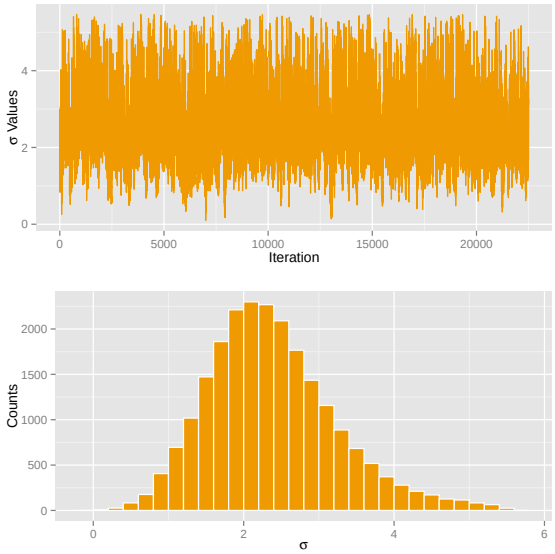

Figure 4.2: Posterior draws for σ following iteration 2500. The plot on the top gives the trace plot of σ. The plot on the bottom gives the histogram of σ.

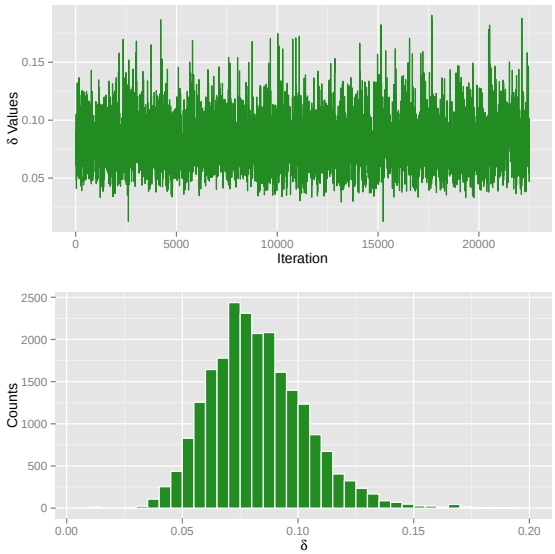

Figure 4.3: Posterior draws for δ following iteration 2500. The plot on the top gives the trace plot of δ. The plot on the bottom gives the histogram of δ. Overall Metropolis acceptance rate: 32.5%.

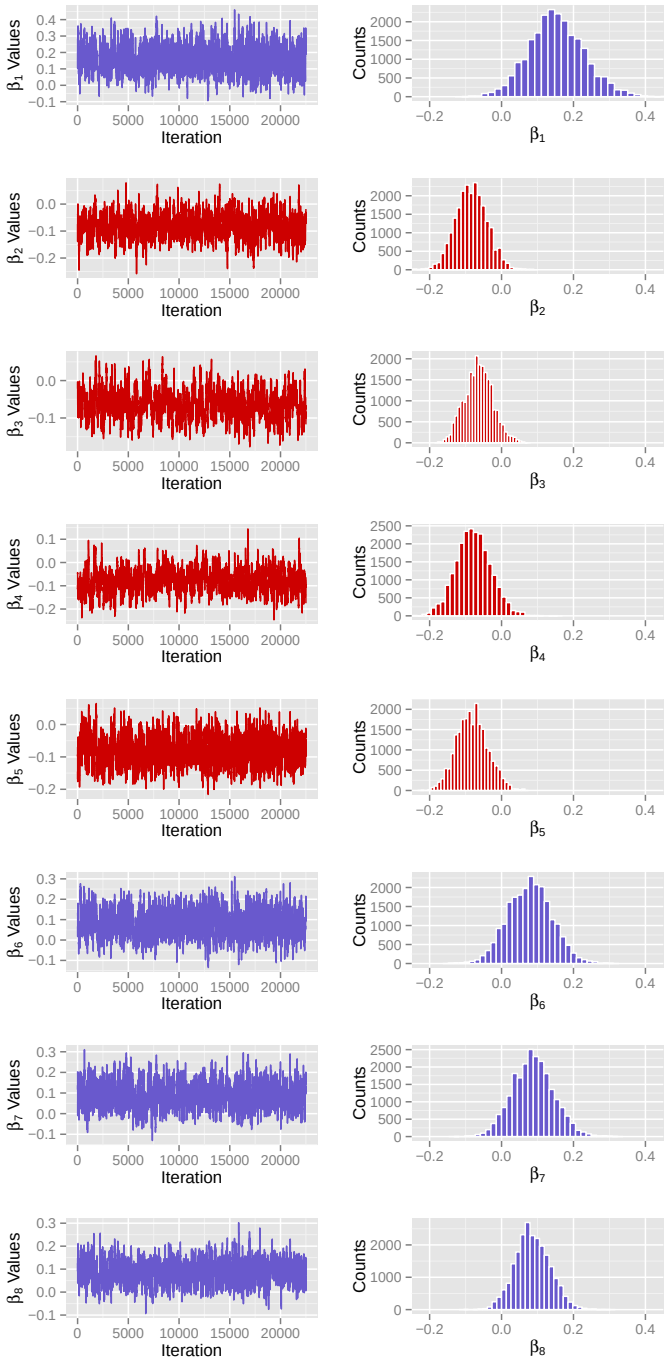

Figure 4.4: Posterior draws for β following iteration 2500. The trace plots are given on the left and the histograms are given on the right hand side. The blue colour indicates a positive β posterior mean and the red colour indicates a negative β posterior mean.

We now provide the fitted spatial map that gives the expected number of putts $E(Z_i \mid \hat{\lambda}_i)$ where $\hat{\lambda}_1, \ldots, \hat{\lambda}_{76}$ are the estimated posterior means of the primary parameters $\lambda_1, \ldots, \lambda_{76}$. Figure 4.5 gives the fitted spatial map:

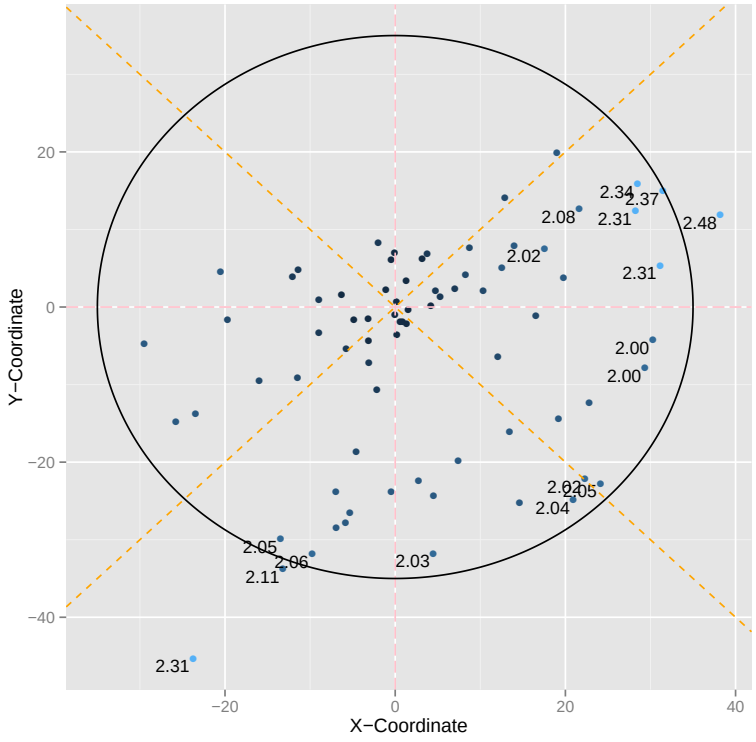

Figure 4.5: The expected number of putts obtained using the spatial model for a selection of putting locations.

The estimates of the expected values are believed to be accurate to within one digit in the last decimal place. We also monitored and assessed the convergence of $\lambda_1, \ldots, \lambda_{76}$, but we are not providing the trace plots and histograms of the λ parameter due to its high dimensionality.

From Figure 4.5, we observe a number of appealing features. First, within a quadrant, the expected number of putts decreases as the distance from the hole decreases. Second, the expected number of putts for spatially close putts is similar. Finally, when comparing putts with the same radii, we observe that the expected number of putts in the first quadrant is greater compared to other quadrants.

4.3 Analysis of Round Four

To give a ranking of the players in the fourth round based on the spatial model developed in this paper, we fitted our model to each of the 18 holes in the fourth round of the tournament. The first step involved converting four-putts to three-putts due to the definition of our model in (2.2) which is based on a truncated-Poisson distribution where four-putts are not permitted. This modelling approach was used as four-putts are very rare. Further, in our data set, we observed only one four-putt out of $76 \times 18 = 1368$ putting opportunities in the fourth round.

Table 4.2 provides the total putts per round statistic and the strokes gained-putting statistic (original) compared to our spatial strokes gained-putting statistic in the fourth round for the top 11 finishers in the tournament.

Golfer	Finishing Position	Total Putts	Strokes Gained (Original)	Strokes Gained (Spatial)
Rory McIlroy	1	28	3.0	3.0
Tiger Woods	2	26	3.2	3.9
Tom Gillis	2	30	0.6	3.0
Lee Westwood	4	28	1.3	1.8
Charl Schwartzel	5	32	1.1	2.1
Justin Rose	5	30	1.1	1.5
Rickie Fowler	7	26	2.0	3.0
Dicky Pride	7	26	2.7	2.8
Graeme McDowell	9	29	0.8	1.2
Kevin Stadler	9	25	3.1	2.4
Chris Stroud	9	25	2.6	1.9

Table 4.2: Various putting statistics calculated for the fourth round of the 2012 Honda Classic.

Recall from section 1.1 that the total putts per round statistic is not a good measure to evaluate putting performance of professional golfers as it does not take the initial location of the ball on the green into consideration. On the other hand, the strokes gained-putting statistic is a much better measure to evaluate putting proficiency of pro golfers. The stokes gained-putting statistic in the fourth column of the above table was obtained from http://media1.pgatourhq.com/reports/R20120101_LeadersStatisticalSummary.pdf.

The strokes gained-putting statistic has been adjusted for the field. The strokes gained-putting statistic values show that the top 11 finishers putted above average since the values

are all positive. With the strokes gained-putting statistic, we observe that Tiger Woods had the best putting performance amongst the top 11 finishers as he was three strokes better than average.

With our spatial putting statistic, we observe that Tiger Woods also had the best putting performance. However, our spatial putting statistic suggests that Tiger Woods was four strokes better than average. The one stroke difference between our statistic and the original strokes gained-putting statistic maybe due to Tiger Woods taking shots in more difficult regions.

We can also calculate the correlation between the total putts, strokes gained (original) and strokes gained (spatial) from Table 4.2. Figure 4.6 summarizes our findings:

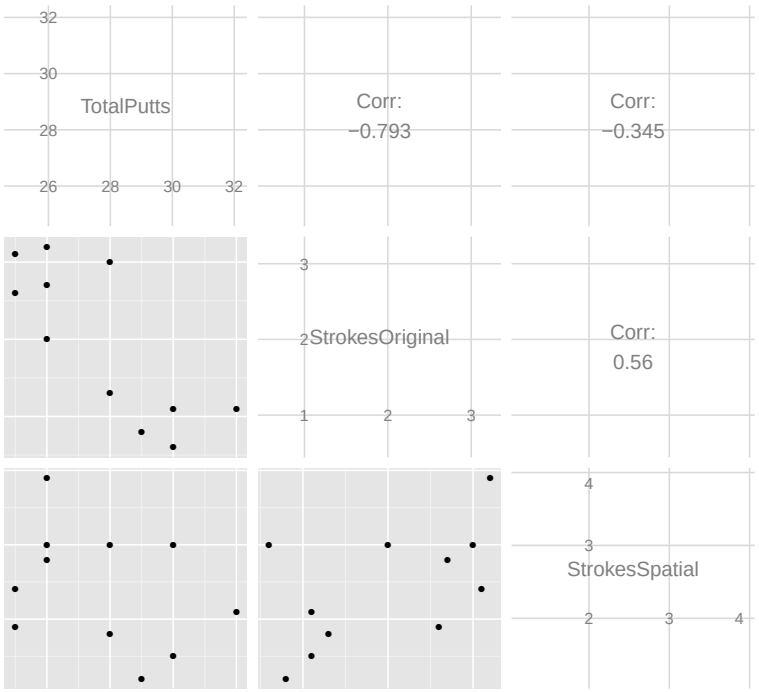

Figure 4.6: The pair-wise correlation plot for total putts, strokes gained (original) and strokes gained (spatial).

We see from Figure 4.6 that there is a positive correlation (0.56) between strokes gained (original) and our strokes gained (spatial). Further, we can test the null hypothesis that correlation coefficient between strokes gained (original) and strokes gained (spatial) is zero. This yields a p-value of 0.3 meaning that we fail to reject the null hypothesis.

Overall, we observe general agreement when comparing the strokes gained-putting statistic and our spatial strokes-gained putting statistic. However, we also observe some key differences. For instance, Tom Gillis has the worst putting performance with respect to the strokes gained-putting statistic amongst the top 11 finishers while he is tied for second best putting performance with respect to our spatial strokes gained-putting statistic. These key differences suggest that potential factors other than distance such as direction are influencing the difficulty of putts.

Chapter 5

Discussion

In this project, we introduced a new metric to evaluate the putting performance of professional golfers. The statistic developed in this paper has a spatial aspect that assesses the difficulty of putts by considering both distance and direction from the hole. The computation of this statistic is facilitated by ShotLink data that records the location on the green for all putts.

The comparison of the spatial statistic developed in this project and the strokes gained-putting statistic indicates that difficulty of putts may be influenced by factors other than distance. This result opens up a new area of research for future investigations as we have only touched the surface here.

First, it is possible to obtain the estimates of expected number of putts for a varying numbers of slices and observe whether the results differ significantly or not. It may also be preferable to determine the slices (Figure 2.1) on a hole by hole basis.

Further, it is quite possible not to adapt a splitting procedure and rather treat the angle difficulty $\beta^{(\theta_i)}$ as a continuous parameter.

Second, the spatial Bayesian model developed in this paper is such that it fits a model to each hole in the fourth round separately. It may be desirable and possible to develop a more complex model that is strengthened by borrowing information from all four rounds of a tournament. Such a modelling approach may lead to more stable and improved parameter estimates.

Finally, in this project, we estimated the expected number of putts from the observed putting green locations. However, it may be possible to draw inference for putting difficulty at other locations. Such inference may help pro golfers to strategize the location of their approach shots to the green.

Bibliography

[1] Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC: Boca Raton, Florida.

[2] Besag, J., York, J. and Mollie, A. (1991). "Bayesian image restoration, with two applications in spatial statistics (with discussion)", *Annals of the Institute of Statistical Mathematics*, 43(1): 1-59.

[3] Broadie, M. (2008). "Assessing golfer performance using golfmetrics", In *Science and Golf V: Proceedings on the 2008 World Scientific Congress of Golf*, D. Crews and R. Lutz (editors), Energy in Motion Inc, Mesa, Arizona, 253-262.

[4] Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis, Third Edition*, Chapman and Hall/CRC: Boca Raton, Florida.

[5] Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition*, John Wiley and Sons: New York.

[6] Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*, Springer: New York.

[7] Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). "Model-based geostatistics (with discussion)", *Journal of the Royal Statistical Society, Series C*, 47(3): 299-350.

[8] Evans, M. and Swartz, T.B. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford University Press: New York.

[9] Fearing, D., Acimovic, J. and Graves, S.C. (2011). "How to catch a Tiger: Understanding putting performance on the PGA Tour", *Journal of Quantitative Analysis in Sports*: 7(1), Article 5.

[10] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (editors) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London.

[11] Tobler, W. (1970). "A computer movie simulating urban growth in the Detroit region", *Economic Geography*: 46, 234-240.

[12] Waller, L. A. (2004). "Bayesian thinking in spatial statistics", Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA.

[13] Yousefi, K. and Swartz, T.B. (2013). "Advanced Putting Metrics in Golf", *submitted for publication.*