

Advanced Putting Metrics in Golf

Kasra Yousefi and Tim B. Swartz *

Abstract

Using ShotLink data that records information on every stroke taken on the PGA Tour, this paper introduces a new metric to assess putting. The methodology is based on ideas from spatial statistics where a spatial map of each green is constructed. The spatial map provides estimates of the expected number of putts from various green locations. The difficulty of a putt is a function of both its distance to the hole and its direction. A golfer's actual performance can then be assessed against the expected number of putts.

Keywords: Bayesian spatial statistics, Professional Golfers' Association, ShotLink data, sports analytics, subjective priors, truncated Poisson.

*Kasra Yousefi is an MSc candidate and Tim Swartz is Professor in the Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Swartz has been partially supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors thank two anonymous reviewers and the Editor whose comments have helped improve the manuscript.

1 INTRODUCTION

The world of sport is littered with statistics. For example, in baseball alone, statistics are kept on *batting average*, *home run totals*, *runs batted in*, *slugging percentage*, *on-base percentage*, *earned run average*, *innings pitched*, *wins*, and *fielding percentage* to name just a few of the more prominent metrics.

Whereas many of the traditional statistics are intuitive, most provide only a partial snapshot of performance, and some statistics can even be misleading. For example, in cricket, the batting statistic known as *strike rate* fails to account for the importance of *dismissals* (Beaudoin and Swartz 2003). As another example, a lofty *save percentage* in baseball may mask the favourable circumstances in which some pitchers enter a game.

With the proliferation of data and the advent of computers, more complex statistics have been proposed which attempt to address critical aspects of performance. For example, the website www.82games.com provides in-depth analyses and statistics related to the National Basketball Association (NBA). Amongst the advanced statistics are a class of statistics that propose to measure the value or contribution of a player or an event relative to what is expected. We refer to these as *relative-value statistics*. For example, in baseball the *VORP* statistic (value over replacement player) attempts to characterize how much a batter contributes offensively or how much a pitcher contributes defensively to the team in comparison to a replacement-level player (Woolner 2002). The comparison is made in terms of runs, the quantity whose value is well understood. *VORP* has proven to be a useful measure in market evaluation. As another example, Chapter 15 of Oliver (2004) discusses relative-value statistics for a given player in the NBA based on team performance with and without the player in the lineup.

In the game of golf, there is a well-known expression that “you drive for show and you putt for dough”. Although the sentiment may not be entirely true, the importance of putting

should not be understated. In golf, putting may be viewed as a game within a game. Once a golfer reaches the *green* (the short grass where the hole is located), a specialized club known as a *putter* is used to stroke the ball into the hole. Naturally, the fewer strokes taken, the better. We note that greens have varying rounded shapes and sizes, where 5,000 square feet may be considered average. With respect to putting, the traditional performance statistic that is commonly reported is the number of *putts per round*. On the PGA (Professional Golfers' Association) Tour, it is generally felt that more than 30 putts per round is an indication of substandard putting performance. However, the total number of putts per round fails to account for the difficulty of putts. For example, a golfer with a high percentage of *greens in regulation*¹ is likely to face more difficult putts than a golfer with a low percentage of greens in regulation. Moreover, a golfer who "chips-in" from off the green will be credited with zero putts. This provides an illusion of good putting with respect to the total number of putts per round statistic since the golfer did not putt.

The inadequacy of the putts per round statistic has been recognized, and in 2011, the PGA Tour began reporting the *strokes gained-putting* statistic. The strokes gained-putting measure falls into the class of relative-value statistics as it attempts to quantify how many strokes a PGA golfer saves relative to other PGA golfers. The statistic is typically reported in shots per round although it is also reported in total putts over a season. For example, in 2011, Luke Donald was the top putter with 0.844 strokes gained per round. Last and number 118 on the list was J.B. Holmes with -0.096 strokes gained per round (or 0.096 strokes lost per round). The key idea behind the strokes gained-putting statistic is that it considers the distance of each initial putt on a green and the expected number of putts that a typical PGA golfer takes from that distance. The methodology was initially developed by Broadie (2008) and was subsequently developed by Fearing, Acimovic and Graves (2011). The calculation

¹Reaching a par 3/4/5 hole in regulation indicates that a golfer has landed on the green in 1/2/3 shots.

of the strokes gained-putting statistic has been facilitated by ShotLink data which records information on every shot taken on the PGA Tour.

In this paper, we propose an enhanced relative-value statistic for putting on the PGA Tour. In addition to taking the distance of a putt into account, we consider additional properties related to the green. For example, it is well-known that a straight uphill putt from 10 feet is easier than an undulating downhill putt from 10 feet. The statistic which we develop uses concepts from the field of spatial statistics. Spatial maps are constructed which provide estimates of the expected number of putts from various green locations. The particular idiosyncrasies of our application result in a novel spatial model that borrows features from both geostatistical models and lattice models as categorized by Cressie (1993). In addition, our spatial model is Bayesian which requires the specification of prior distributions. Bayesian spatial models are considered by Banerjee, Carlin and Gelfand (2003). Given estimates of the expected number of putts from various green locations, we can assess performance by comparing the actual number of putts with the expected number of putts. Our methodology also relies on the availability of ShotLink data.

There have been several recent attempts at using spatial statistics in sports analytics. For example, Shuckers (2011) considers the development of statistics for assessing goaltending based on shot data in the National Hockey League (NHL). In the NBA, shot data based on missile tracking cameras are being used to create spatial maps of preferred shooting locations for individual players (Wilson 2012). Neither of these approaches use informative priors that take the physical layout of the ice/court into account.

Related to our work is a paper by Jensen, Shirley and Wyner (2009) which develops a Bayesian spatial model for the analysis of fielding in Major League Baseball. Whereas our spatial surface is a green with the number of putts observed (1,2,3) from various locations, they consider a field with binary outcomes corresponding to whether catches were made.

In our application, the relevant spatial features are the distance and the direction to the hole whereas Jensen, Shirley and Wyner (2009) include covariates such as the velocity of the batted ball, the distance travelled by the ball and the direction of travel by the fielder. One of the major differences between the two analyses is that we model a typical PGA Tour golfer and then compare differences between expected results for the typical player and observed results for a given player. On the other hand, Jensen, Shirley and Wyner (2009) model each fielder individually with a fielder effect.

We also remark on the paper by Reich, Hodges, Carlin and Reich (2006) which is related to our work. Here, the authors consider various Bayesian spatial analyses of basketball shot data corresponding to Sam Cassell during the 2003-2004 NBA season. As in our work, they take distance and shooting angle (from the basket) into account. In one of their models, outcomes (i.e. misses and makes) corresponding to shooting attempts are modeled using Bayesian logistic regression where the court is divided into 122 regions. Reich et al. (2006) spatially smooth regression parameters using CAR (conditionally autoregressive) prior distributions. Interesting covariates are considered such as the presence/absence of specific players on the court.

In section 2, we develop a novel Bayesian spatial statistics model where we take various features of putting into account. In particular, we propose a prior distribution which implies that a putt on a given line to the hole should have a greater probability of being made than a longer putt along the same line. As our model is non-trivial, the computations rely on Markov chain Monte Carlo (MCMC) methodology for simulation from the posterior distribution. An overview of the computations is provided in section 3 where Metropolis within Gibbs steps are utilized. We then create a spatial map for the number of expected putts taken from each putting location with respect to sample data from the 2012 Honda Classic. Comparisons are made with the intermediate calculations used in the strokes gained-putting statistic. As

expected, we observe that factors other than distance play a role in the difficulty of putts. We conclude with a discussion of future research directions in section 4.

2 SPATIAL STATISTICS MODELS

In a Bayesian hierarchical setting, modeling is often facilitated by thinking about data and parameters conditionally, one level of the hierarchy at a time. In our application, we wish to create a spatial map which provides estimates of the expected number of putts by PGA Tour golfers from each of the realized putting locations.

Prior to describing the modeling details, our modeling framework begins with the specification of the distribution of the number of putts (the data) taken from various initial locations on the green. The distribution of the number of putts corresponding to the i th putting location is characterized by a parameter λ_i . The parameter λ_i is an unknown and is a quantity that we wish to estimate. There are aspects of the λ_i for which we have intuition, and we assign prior distributions to the λ_i based on our physical understanding of putting. For example, we would like putts i and j to have similar difficulty if they are spatially near one another. With λ_i and λ_j strongly correlated a priori, “learning” about parameter λ_i borrows from the learning of λ_j , and vice-versa. We would also like to assign a prior distribution whereby λ_i and λ_j are “close” if the corresponding putts are of similar length but perhaps from different directions. We are essentially smoothing the parameters spatially. In assigning prior distributions to the λ_i that take into account our prior knowledge, there remain unknowns in these distributions which are similarly characterized by secondary parameters which are sometimes referred to as hyperparameters. These hyperparameters themselves have distributions and we assign prior distributions to these quantities. The specification of distributions on the various layers of parameters where we make use of our underlying

knowledge is referred to as *hierarchical modeling*.

Using ShotLink data, consider a particular green in a particular round of a tournament. The data consist of Z_1, \dots, Z_n where Z_i is the number of putts that it takes from the i th putting location, $i = 1, \dots, n$. The i th putting location corresponds to where the i th golfer's shot landed on the green. However, we emphasize that the subscript i used in modeling refers to the location and the characteristics of the location, and not the characteristics of the i th golfer. For the purposes of modeling, all PGA golfers are assumed to have the same putting ability. We are also able to access the ShotLink data to extract useful covariates (x_i, y_i) , $i = 1, \dots, n$. Here, (x_i, y_i) are the Cartesian coordinates measured in feet of the i th putting location where the pin (i.e. hole) is defined as the origin. For convenience, we can also transform (x_i, y_i) to its polar representation (r_i, θ_i) . In our initial step of modeling, we first assume Z_1, \dots, Z_n independent with

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

for $i = 1, \dots, n$. In (1), the expected number of putts $E(Z_i) = 1 + e^{\lambda_i}$ sensibly depends on the putting location given by (r_i, θ_i) . Therefore, we interpret λ_i as the difficulty of the putting location (r_i, θ_i) for PGA golfers where larger values of λ_i correspond to more difficult putting locations. Note that the Poisson is a tractable distribution and its support is correct for our application. Also, the Poisson distribution is defined for any $\lambda_i \in \mathfrak{R}$. However, it is clear that the Poisson distribution is inappropriate for modeling in this application. For example, it is difficult to imagine how the conditions of a Poisson process are even approximately satisfied. With respect to our application, the Poisson distribution assigns too much probability to $\text{Prob}(Z_i = 1)$ when λ_i is negative. For example, when $\lambda_i = 0$, we have $\text{Prob}(Z_i = 1) = \text{Prob}(Z_i = 2) = 0.37$ which implies that one-putts are as common as two-putts from a putting location where the mean is two putts. In spatial statistics, the

Poisson distribution is often used for lattice problems where the number of counts refer to some region i rather than a particular location (Cressie 1993).

Although we initially considered a model based on (1), we prefer a variation where Z_1, \dots, Z_n are assumed independent,

$$Z_i - 1 \sim \text{truncated-Poisson}(\exp\{\lambda_i(r_i, \theta_i)\}) \quad (2)$$

and the truncation is imposed such that $\text{Prob}(Z_i \geq 4) = 0$ for $i = 1, \dots, n$. The truncation has the benefit of defining a structure where variance is less than the mean. In the PGA Tour putting application, the truncated Poisson appears to be realistic. For example, $\lim_{\lambda \rightarrow -\infty} \text{Prob}(Z = 1 \mid \lambda) = 1$; i.e. there are locations on the green (e.g. extremely close to the pin) where the probability of sinking the putt approaches 1.0. Also, $\lim_{\lambda \rightarrow \infty} \text{Prob}(Z = 3 \mid \lambda) = 1$; i.e. there may be locations on the green (e.g. very far from the pin with a tricky slope) where the probability of three-putting approaches 1.0.

To get a sense of the adequacy of the truncated-Poisson distribution for the given application, we considered 2012 data as provided at www.pgatour.com. In Figure 1, we provide a barplot of the observed percentages of the number of putts taken from 5 to 10 feet. This is compared with the percentages arising from the fitted truncated-Poisson(0.61) distribution. We remark that the fit is not as good at larger distances, where alternative distributions may be considered. We comment on this further in point 5 of the Discussion.

Having specified the probability distribution of the data in (2), we note that the difficulty of the n putting locations is characterized by the unknown parameter vector $\lambda = (\lambda_1, \dots, \lambda_n)'$. Our interest is therefore focused on λ and a hierarchical Bayesian formulation involves assigning prior distributions to these unknown parameters. Ideally, subjective priors are assigned which take into account our physical understanding of the parameters. In the

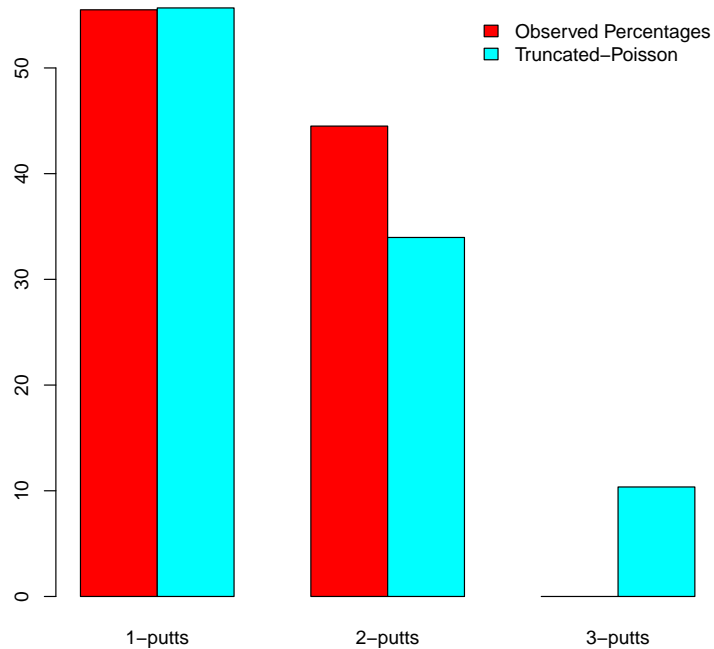


Figure 1: The observed percentages of the number of putts taken from 5-10 feet in 2012 compared with percentages given by a fitted truncated-Poisson.

spirit of Besag et al. (1991) and Diggle et al. (1998), we propose

$$\lambda \sim \text{Normal}_n(\mu, \sigma^2 V) \tag{3}$$

where $\mu = (\mu_1, \dots, \mu_n)'$. A motivation is to smooth the vector λ whereby $|\lambda_i - \lambda_j|$ is “small” with high probability when putting locations i and j are spatially “close”. Our specification of the variance-covariance matrix in (3) is a standard choice (Bannerjee et al. (2004)) which

assures positive-definiteness. We let $V = (v_{ij})$ be the Gaussian covariance function where

$$v_{ij} = \exp\{ -\delta^2 \| (x_i, y_i) - (x_j, y_j) \|^2 \} \quad (4)$$

and $\| \cdot \|$ denotes Euclidean distance. In (4), we require $\delta > 0$. We have used the covariates (x_i, y_i) corresponding to the i th initial putting location to assign a greater correlation to parameters λ_i and λ_j whose putting locations are spatially close. With putting locations i and j that are spatially close, λ_i and λ_j will tend to have similar values. In the case of spatially distant locations, $v_{ij} \rightarrow 0$. And for diagonal entries, $v_{ii} = 1$. The matrix V is known to provide a smooth surface where there are strong spatial correlations within a small range of distances.

A secondary motivation in the hierarchical model concerns putting locations i and j that are roughly the same distance from the pin but are not spatially close (and consequently weakly dependent ipriori). We want to impose a structure where the mean values μ_i and μ_j are similar. A novel aspect of the spatial modeling exercise involves the prior specification of the vector μ in (3). Stated somewhat differently, the idea which we wish to implement is that a shorter putt on a given line to the hole should have a greater probability apriori of being made than a longer putt along the same line. Recall that λ_i relates to the probability of a putt being made from location (x_i, y_i) where larger values of λ_i characterize more difficult putting locations. Specifically,

$$\text{Prob}(Z_i = 1) = \frac{1}{1 + e^{\lambda_i} + e^{2\lambda_i}/2} .$$

And also recall that μ_i is the prior mean of λ_i . Our approach divides the spatial map into 8 “pie slices” emanating from the pin (see Figure 2). The rationale is that there are typically undulations in putting greens and that by dividing a green into slices, putts within the same

slice will be impacted by the terrain in a similar manner. Although the choice of 8 slices is somewhat arbitrary, we neither want too few slices (resulting in within-slice heterogeneity) nor too many slices (resulting in few observations per slice.) The polar covariate θ_i defines the slice in which the i th putting location resides. Again, the idea is that putts within the same slice share common features with respect to the putting terrain, and that the relative difficulty is only affected by the length r_i of the putt. Accordingly, we set

$$\mu_i = g(r_i + \beta^{(\theta_i)} r_i) \tag{5}$$

where $\beta^{(\theta_i)} \in \mathfrak{R}$ is mapped to one of β_1, \dots, β_8 according to the slice corresponding to θ_i and g is increasing piecewise linear. The knots, slopes and ordinates for the piecewise linear function g are described in the appendix and are based on historical data. From (5), we see that the mean difficulty μ_i of the i th putting location is affected by both the length of the putt r_i and the slice in which it resides. Specifically, $\mu_i \in \mathfrak{R}$ where longer putting distances r_i yield larger values of μ_i . The different slopes β_1, \dots, β_8 accommodate varying difficulty amongst the 8 putting angles. We have experimented with the numbers of slices. We have found that 8 slices is sufficiently small to yield stable parameter estimation, and yet it is sufficiently large to provide realism in the varying difficulty of putting angles. The parametrization (5) provides an appealing interpretation for β where $\beta = 0$ denotes a slice of typical difficulty. Slices with $\beta > 0$ and $\beta < 0$ represent more difficult and less difficult putting angles respectively. For example, $\beta = 0.1$ represents a putt that is equivalent in difficulty to a typical putt extended by 10% in length.

To complete the model specification, we assign β_1, \dots, β_8 independent $\text{Normal}(0, \sigma_\beta^2)$. The β_j are centered about the average value, with differences accounting for more difficult and less difficult putting angles. The hyperparameter $\sigma_\beta = 0.2$ is specified to cover plausible

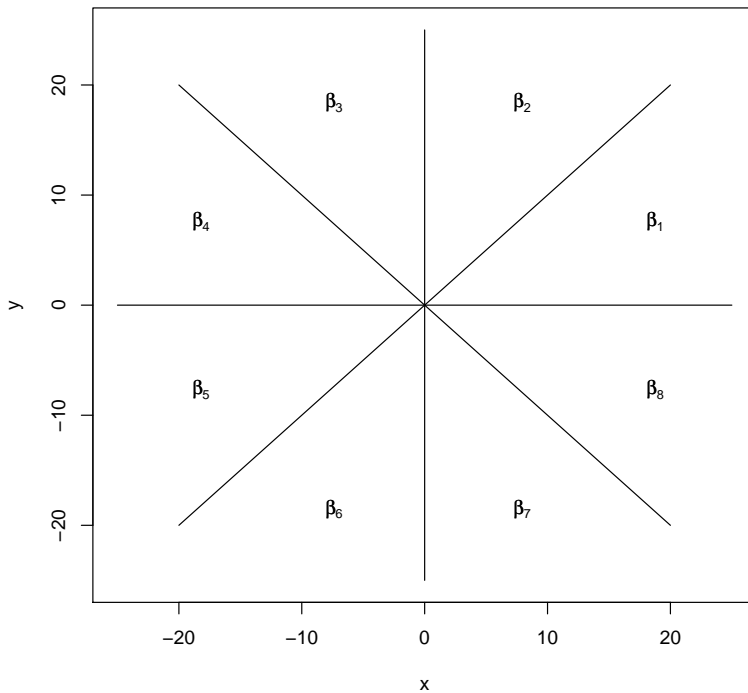


Figure 2: The 8 pie slices with their corresponding parameter. Within a slice, a putt has decreasing probability of success as the distance to the pin r_i increases.

values of the β_j . We also set

$$\begin{aligned} [\sigma] &\propto 1/\sigma \\ [\delta] &\propto 1 \end{aligned} \tag{6}$$

where $[\cdot]$ is generic notation for the probability density function. The distributions in (6) are standard reference priors where both are constrained to the positive real line.

2.1 Advanced Putting Statistics based on Spatial Models

Having specified the spatial models above, it is necessary to fit the models (section 3) whereby parameter estimates are obtained. The estimates which are the most important to us are the expected number of putts from the realized green locations.

Accordingly, under (1), we calculate $E(Z_i | \lambda_i = \hat{\lambda}_i) = 1 + \hat{\tau}_i$ where $\hat{\tau}_i = \exp\{\hat{\lambda}_i\}$. Under (2), we instead calculate $E(Z_i | \lambda_i = \hat{\lambda}_i) = 1 + \hat{\tau}_i(1 + \hat{\tau}_i)/(1 + \hat{\tau}_i + \hat{\tau}_i^2/2)$ where $\hat{\tau}_i = \exp\{\hat{\lambda}_i\}$. For the i th golfer on the given hole for the given round of tournament golf, his performance measure is therefore given by

$$E(Z_i | \lambda_i = \hat{\lambda}_i) - Z_i \tag{7}$$

which represents relative strokes gained on the hole. Recall that $E(Z_i | \lambda_i)$ represents the average number of strokes for PGA golfers from the i th putting location with difficulty characterized by λ_i and that Z_i is the actual number of strokes taken from the i th putting location by the i th golfer. The statistic (7) relates actual performance to expected performance. A positive value of (7) indicates above average performance on the hole whereas a negative value indicates below average performance. Our proposed advanced putting statistic for a round of golf for the i th golfer would therefore involve a summation of (7) over all 18 holes. For a tournament statistic, the summation would involve 72 terms corresponding to four rounds of 18 holes of golf. In a tournament, 72 spatial maps would need to be created. (Note that new hole locations are used for each round in a tournament.) Season averages might similarly be calculated. As is done for the strokes gained-putting statistic, we adjust the spatial statistic against the field (<http://wrongfairway.com/tag/strokes-gained-putting>).

We emphasize that our inferential problem only requires the expectation of Z at the realized locations where putts have taken place. This simplifies inference and also distin-

guishes our application from geostatistical problems (Cressie 1993) where spatial estimates are required at locations other than those corresponding to sample data.

3 COMPUTATIONS AND ANALYSIS

After specifying the model components, the standard first exercise in a Bayesian application is an attempt to express the posterior distribution. We use $[A | B]$ to generically denote the density of A given B . Following the distributional assumptions given in section 2, the posterior density takes the form

$$\begin{aligned}
[\lambda, \sigma, \beta, \delta | Z] &= [Z | \lambda, \sigma, \beta, \delta] \cdot [\lambda, \sigma, \beta, \delta] \\
&= [Z | \lambda] \cdot [\lambda | \sigma, \beta, \delta] \cdot [\sigma, \beta, \delta] \\
&= [Z | \lambda] \cdot [\lambda | \sigma, \beta, \delta] \cdot [\sigma] \cdot [\beta] \cdot [\delta]
\end{aligned} \tag{8}$$

which has dimension $n + 1 + 8 + 1 = n + 10$ and where the main parameters of interest $\lambda_1, \dots, \lambda_n$ characterize the difficulty of the n putting locations. Substituting the parametric distributions, (8) reduces to

$$\begin{aligned}
[\lambda, \sigma, \beta, \delta | Z] &\propto \prod_{i=1}^n \frac{e^{\lambda_i(Z_i-1)} e^{-e^{-\lambda_i}}}{e^{-e^{-\lambda_i}} (1 + e^{\lambda_i} + e^{2\lambda_i}/2)} \cdot \frac{e^{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)}}{\sigma^n |V|^{1/2}} \cdot \prod_{j=1}^8 e^{-\frac{1}{2\sigma_j^2} \beta_j^2} \cdot \frac{1}{\sigma} \\
&= \prod_{i=1}^n \frac{e^{\lambda_i(Z_i-1)}}{(1 + e^{\lambda_i} + e^{2\lambda_i}/2)} \cdot \frac{e^{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)}}{\sigma^{n+1} |V|^{1/2}} \cdot \prod_{j=1}^8 e^{-\frac{1}{2\sigma_j^2} \beta_j^2}
\end{aligned} \tag{9}$$

where $\mu_i = g(r_i + \beta^{(\theta_i)} r_i)$ and V is given in (4).

Whereas the posterior density (9) provides the full description of parameter uncertainty given observed data, the complexity of (9) is such that posterior summaries are needed for

interpretation. Posterior summaries are typically simple quantities such as posterior means and posterior standard deviations. In this application, these quantities take the form of intractable integrals.

Based on the complexity of the posterior density in (9), it seems that a sampling based methodology is the only feasible way to approximate posterior summaries. Details associated with a Markov chain implementation are provided in the appendix.

3.1 A Test Case: The 2012 Honda Classic

The data used in our analysis were taken from the 2012 Honda Classic, held March 1-4, 2012 at the PGA National Champion course in Palm Beach Gardens, Florida. The Champion course at PGA National is known for its overall difficulty and for its undulating greens. ShotLink data were extracted for the final (fourth) round of the tournament. For each hole and for each golfer, the data consist of the starting location (x, y) of the first putt on the green and the total number of putts taken.

For illustration purposes, we examine the first hole of the final round in detail. Figure 3 provides the number of putts taken by each of the 76 golfers on the first hole of the final round. As expected, we observe that the probability of sinking a putt (i.e. a one-putt) decreases as the length of the putt from the hole increases. We also observe that putts in the first quadrant are more difficult. For example, when compared to the other quadrants, three-putts are more common in the first quadrant.

We next fit the spatial model using the data taken from the first hole of the final round. The Metropolis within Gibbs algorithm was run for 25000 iterations where the first 2500 iterations were used as burn-in. This required approximately 10 hours of computation on a Mac Pro workstation. Convergence was assessed using standard diagnostic tests (e.g. trace plots, use of multiple chains, etc.) In Table 1, we provide the Markov chain estimates for the

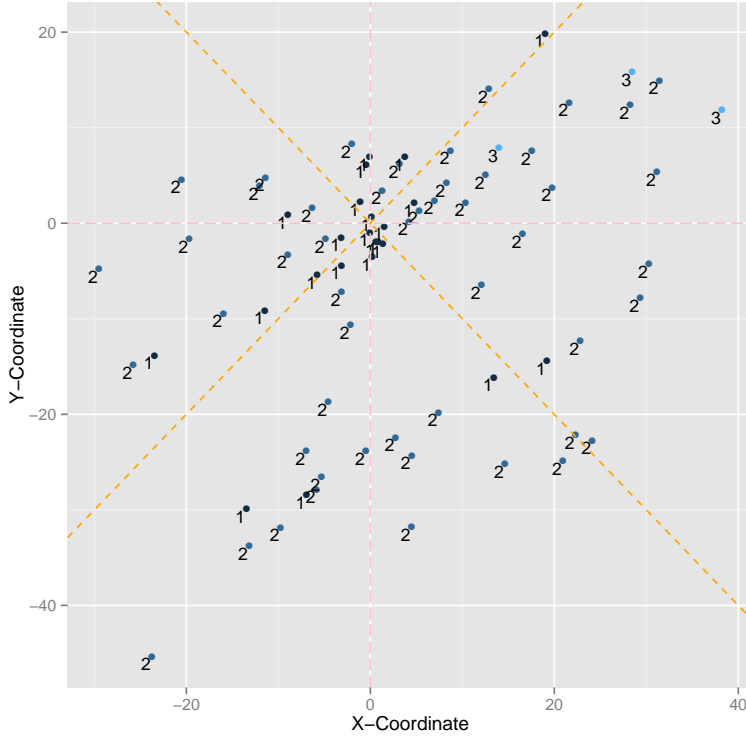


Figure 3: The number of putts taken by each of the 76 golfers on the first hole of the final round of the 2012 Honda Classic. The x and y coordinates are measured in feet.

secondary parameters of interest. They appear to agree with our intuition. In particular, the largest β value is β_1 , and this corresponds to the first quadrant where the putting terrain is believed to be more difficult. We also note that the posterior standard deviations are not large when compared to the posterior means. This suggests that there is substantial information in the data concerning the secondary parameters. Note that we investigated the pairwise correlation between the β 's and found no significant correlations.

In Figure 4, we focus on the primary parameters $\lambda_1, \dots, \lambda_n$ by plotting the expected number of putts $E(Z_i | \lambda_i = \hat{\lambda}_i)$ for a selection of putting locations where $\hat{\lambda}_1, \dots, \hat{\lambda}_n$ are the corresponding estimated posterior means. The expected value estimates are believed to be

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

accurate to within one digit in the last decimal place. We observe several appealing features:

- within a quadrant, the expected number of putts increases as the distance from the hole increases
- the expected number of putts is similar when the putting locations are spatially close
- the expected number of putts is greater in the first quadrant than in other quadrants when comparing putts with the same radii

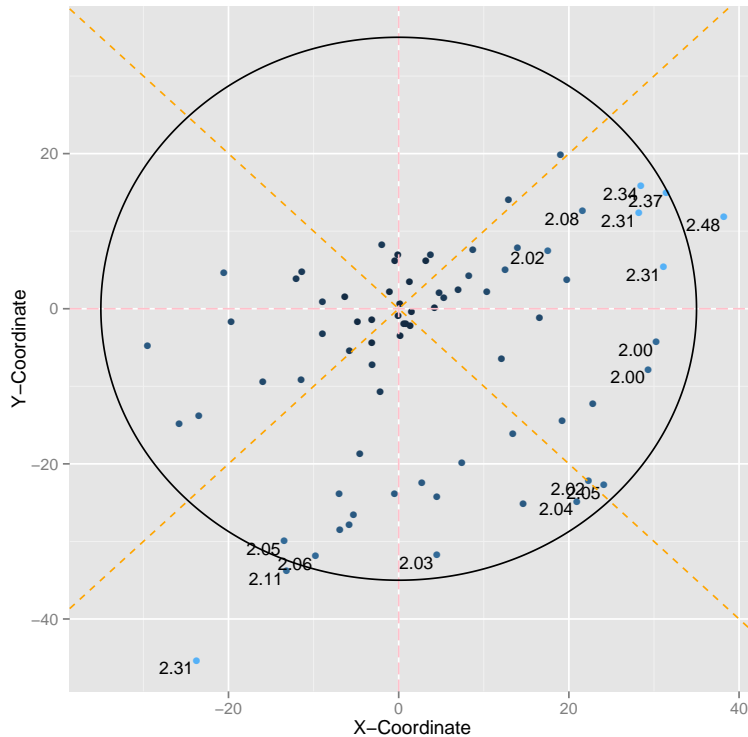


Figure 4: For a selection of putting locations i , the expected number of putts $E(Z_i | \lambda_i = \hat{\lambda}_i)$ obtained using the spatial model.

We now turn to the fitting of the spatial model for all 18 holes in the fourth round of the 2012 Honda Classic. Recall that model (2) is based on a truncated-Poisson distribution where four-putts are viewed as an impossibility. Consequently, data for which the observed number of putts $Z_i \geq 4$ are converted to $Z_i = 3$. Fortunately, four-putts are very rare, and in this dataset, involving $76 \cdot 18 = 1368$ putting opportunities, there was only one observed four-putt.

To get a sense of the utility of the spatial approach, we calculate various statistics for the fourth round. These statistics are recorded in Table 2 for the top 11 finishers in the tournament. We first observe that the total number of putts in the round varies from 25 to

32. As discussed previously, this statistic is not a good measure of putting proficiency as it does not account for the initial location of the ball on the green. The strokes gained-putting statistic has been adjusted for the field where the fourth column of Table 2 was obtained from http://media1.pgatourhq.com/reports/R20120101_LeadersStatisticalSummary.pdf. With the strokes gained-putting statistic, we observe that all of the 11 golfers putted above average (i.e. positive values). This is not surprising as these are the top 11 finishers and it is well-known that putting is a key component of success. According to the strokes gained-putting statistic, we observe that Tiger Woods had the best round of putting amongst the golfers in Table 2 where he was more than three strokes better than average. When we compare the strokes gained-putting statistic to the enhanced spatial statistic developed in this paper, we observe general agreement. Using the spatial model, Tiger Woods also had the best round of putting amongst the golfers in Table 2, but we note that the spatial model suggests that he was nearly four strokes better than average. The differences between the original strokes gained-putting statistic and the spatial strokes gained-putting statistic indicate that factors other than distance (e.g. undulation of the greens) also affect the difficulty of putts.

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

We would like to investigate further the differentiation between strokes gained (spatial) and strokes gained (original). From the last two columns of Table 2, it seems that Tom Gillis may have putted from relatively difficult directions. The entries suggest that he is an additional 2.4 strokes better than the field when the spatial aspect is considered. Conversely, it seems that Kevin Stadler may have putted from relatively easy directions. For each of these golfers, we have 18 values which are the posterior means of the β 's corresponding to the slices of their initial green locations. These values are used to produce the boxplot in Figure 5. Indeed, we observe that Stadler's β values tend to be smaller indicating that he generally putted from easier directions.

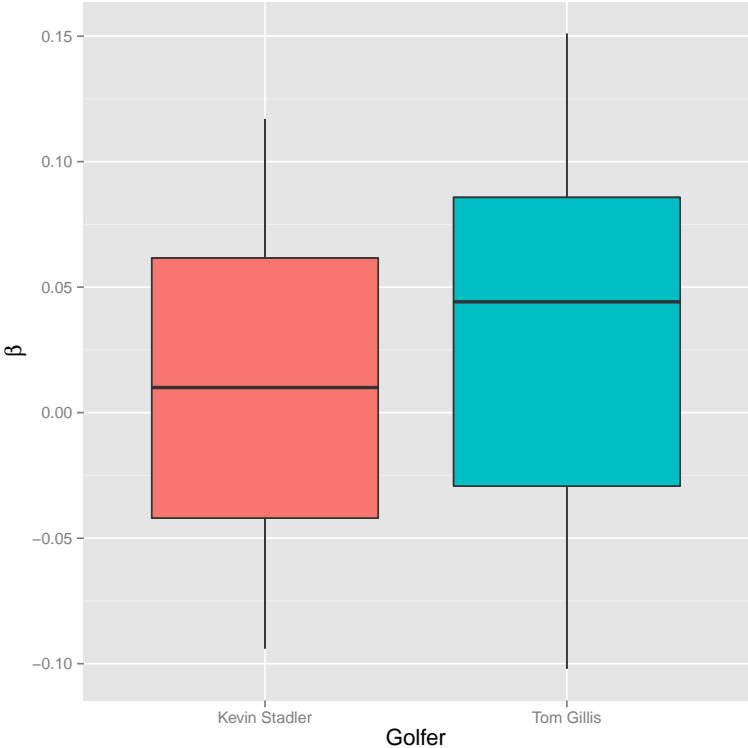


Figure 5: Boxplot of the posterior means of the β 's corresponding to the slices of the initial green locations for Kevin Stadler and Tom Gillis.

It is interesting to ask how much can be gained through a good round of putting. From Table 2, it appears that an exceptional round of putting can lower a golfer’s score (relative to the field) by as many as three strokes. In the 17 PGA stroke play tournaments of 2013 prior to May 1, the margin of victory ranged from zero strokes (decided in a playoff) to four strokes, with an average margin of victory of 1.8 strokes. Clearly, three strokes saved in a round by putting is a meaningful performance.

4 DISCUSSION

This paper introduces an enhanced metric for the evaluation of putting proficiency on the PGA Tour. The approach is novel in that it assesses the difficulty of putts by considering both the distance and the orientation with respect to the pin. The methodology relies on the development of a spatial statistics model and is facilitated by ShotLink data which records the position on the green for all putts.

Whereas the approach appears promising and provides new insights with respect to putting proficiency, we consider our work to be an initial exploration of spatial dependencies on putting. We have identified at least six avenues for future investigation.

1. It may be preferable to determine the slices (Figure 2) on a hole by hole basis. Since the slices characterize directional difficulty, it would not be ideal if a slice contained a ridge that affected putts in only a portion of the slice. It is possible to both vary the number of slices and rotate the angles to improve uniformity within slices. Related to this, it would be good to have additional information on the shape and the orientation of the greens. The ShotLink data only reveal coordinates for each putting location. Greens are not generally circular, and such information would provide accurate shapes of the spatial maps and allow a golfer to better relate the map to the actual green.

2. The model is currently fit for each of the 18 holes in each round of a tournament. It may be possible to consider more complex models where information concerning a green can be borrowed over the four rounds of a tournament. This may improve parameter estimation.
3. To quantify how much better the “great” putters are than average putters, it would be interesting to calculate our strokes gained-putting spatial statistic for an entire season, and to produce standard errors. This would provide more insight on the value of putting in the overall game of golf.
4. Although we calculate the expected number of putts from the actual putting locations on the greens, it may be possible to infer putting difficulty at other locations. Such historical maps could be useful for PGA Tour professionals who strategize the location of their approach shots to the green.
5. As remarked in section 2, the fit of the truncated-Poisson for intermediate putting distances was less than ideal (e.g. Figure 1). Although the Poisson distribution has been used extensively in spatial statistics, it may be preferable to consider alternative distributions for the number of putts defined on the integers 1, 2, 3. Ideally, such a distribution would be tractable and be characterized by a single parameter which describes the difficulty of a putt. Using historical putting percentages from various distances, we have experimented with the probability mass function $p_i(z_i)$ where $p_i(1) = \lambda_i$, $p_i(2) = 1 - \lambda_i - e^{-41\lambda_i}$ and $p_i(3) = 1 - p_i(1) - p_i(2)$ with appropriate constraints on λ_i . Of course, a different prior specification would be required with alternative distributions.
6. Although our motivation was the introduction of a spatial component to model putting, alternative models could be considered. For example, as suggested by a Reviewer, the

outcomes corresponding to individual putts might be modeled as a function of the location and also the golfer. Golfer effects could then be analyzed instead of modeling all PGA golfers as identically distributed. It may also be possible to consider every putt on a green, instead of only the initial putts. Such an approach would introduce dependencies between successive putts. Consequently, it would seem natural to extend the putting outcomes from Bernoulli data to the resting locations of the intermediate putts.

5 REFERENCES

- Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC: Boca Raton, Florida.
- Beaudoin, D. and Swartz, T.B. (2003). “The best batsmen and bowlers in one-day cricket”, *South African Statistical Journal*, 37(2): 203-222.
- 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.
- 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.
- Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition*, John Wiley and Sons: New York.
- 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.
- 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.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (editors) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London.
- Jensen, S.T., Shirley, K.E. and Wyner, A.J. (2009). “Bayesball: A Bayesian hierarchical model for evaluating fielding in Major League Baseball”, *The Annals of Applied Statistics*, 3(2), 491-520.

- Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis*, Potomac Books: Washington.
- Reich, B.J., Hodges, J.S., Carlin, B.P. and Reich, A.M. (2006). “A spatial analysis of basketball shot chart data”, *The American Statistician*, 60(1), 3-12.
- Shuckers, M.E. (2011). “DIGR: A defense independent rating of NHL goaltenders using spatially smoothed save percentage maps”, *MIT Sloan Sports Analytics Conference*, March 4-5, 2011, Boston, MA.
- Wilson, M. (2012). “Moneyball 2.0: How missile tracking cameras are remaking the NBA”, <http://www.fastcodesign.com/1670059>.
- Woolner, K. (2002). “Understanding and measuring replacement level”, In *Baseball Prospectus 2002*, J. Sheehan (editor), Brassey’s Inc: Dulles, Virginia, 55-66.

6 APPENDIX

We provide the details associated with the Markov chain Monte Carlo implementation briefly discussed in section 3.

In a Markov chain approach, it is typical to first consider the construction of a Gibbs sampling algorithm. In a Gibbs sampling algorithm, we require the full conditional distributions of the model parameters. A little algebra yields the following full conditional densities:

$$\begin{aligned}
[\lambda_i | \cdot] &\propto \frac{e^{\lambda_i(Z_i-1)}}{(1+e^{\lambda_i}+e^{2\lambda_i}/2)} \cdot e^{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)} \\
[\sigma^2 | \cdot] &\propto \text{Inverse-Gamma}\left(\frac{n-1}{2}, \frac{2}{(\lambda-\mu)'V^{-1}(\lambda-\mu)}\right) \\
[\beta_j | \cdot] &\propto \exp\left\{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)\right\} \cdot \prod_{j=1}^8 e^{-\frac{1}{2\sigma_j^2}\beta_j^2} \\
[\delta | \cdot] &\propto |V|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\lambda-\mu)'V^{-1}(\lambda-\mu)\right\}
\end{aligned} \tag{10}$$

Referring to the full conditional distributions in (10), we observe that sampling σ is straightforward. Most statistical software packages facilitate generation of a random variate v from the required Gamma distribution, and we then set $\sigma = 1/\sqrt{v}$.

The remaining distributions in (10) are nonstandard statistical distributions, and we therefore introduce Metropolis steps, sometimes referred to as “Metropolis within Gibbs” steps for variate generation (Gilks, Richardson and Spiegelhalter 1996). A general strategy in Metropolis is to introduce proposal distributions which facilitate variate generation and yield variates that are in the “vicinity” of the full conditional distributions. For the generation of λ_i , we consider putting data obtained from the 2012 PGA Tour up to and including the Ryder Cup on September 30, 2012. The data were obtained from the website www.pgatour.com and are summarized in Table 3 by considering the median putting performance by PGA Tour professionals at a distance of r feet from the pin. From section 2.1 of the paper, we recall that the expected number of putts is given by $E(Z | \lambda) = 1 + \tau(1 + \tau)/(1 + \tau + \tau^2/2)$ where $\tau = \exp\{\lambda\}$. Since $\lambda(r, \theta)$ is a function of the distance to the pin r and the directional angle θ , we equate the values in the fifth column of Table 3 to $E(Z | \lambda)$ from which plausible values of λ can be derived for a specified distance r to the pin. For example, when $r = 17.5$, we obtain $\lambda = 0.071$. These plausible values of λ can then be used in the development of proposal densities. For example, if $r_i = 17.5$ corresponding to the i th golfer, we consider the proposal density $\lambda_i \sim \text{Normal}(0.071, 0.04)$ where the variance is conservatively large relative to plausible values of λ_i .

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

Table 3: Putting summaries from the 2012 PGA Tour and the resultant expected number of putts.

For the generation of β_j , we consider the $\text{Normal}(0, \sigma_\beta^2)$ proposal distribution. Recall that the $\text{Normal}(0, \sigma_\beta^2)$ distribution is also the prior distribution for β_j , $j = 1, \dots, 8$. Matching the

prior with the proposal results in a simplification of the corresponding Metropolis acceptance ratio. Recall from (3) that λ_i has mean μ_i and that $\mu_i = g(r_i + \beta^{(\theta_i)}r_i)$ according to (5). Referring to the case of $r = 17.5$ in Table 1 and the above considerations, this suggests

$$0.071 = g(17.5)$$

corresponding to putting angles of average difficulty. Using similar constraints at other distances r , we obtain knots for the piecewise linear function g . To be precise, we set

$$g(r_i + \beta^{(\theta_i)}r_i) = \begin{cases} -4.600 + 0.705(r_i - 2.0 + \beta^{(\theta_i)}r_i) & 2.0 \leq (1 + \beta^{(\theta_i)})r_i < 7.5 \\ -0.722 + 0.111(r_i - 7.5 + \beta^{(\theta_i)}r_i) & 7.5 \leq (1 + \beta^{(\theta_i)})r_i < 12.5 \\ -0.165 + 0.047(r_i - 12.5 + \beta^{(\theta_i)}r_i) & 12.5 \leq (1 + \beta^{(\theta_i)})r_i < 17.5 \\ 0.071 + 0.024(r_i - 17.5 + \beta^{(\theta_i)}r_i) & 17.5 \leq (1 + \beta^{(\theta_i)})r_i < 22.5 \\ 0.192 + 0.019(r_i - 22.5 + \beta^{(\theta_i)}r_i) & 22.5 \leq (1 + \beta^{(\theta_i)})r_i < 40.5 \end{cases}$$

where values $(1 + \beta^{(\theta_i)})r_i < 2$ are set to $(1 + \beta^{(\theta_i)})r_i = 2$ and values $(1 + \beta^{(\theta_i)})r_i > 40$ are set to $(1 + \beta^{(\theta_i)})r_i = 40$. To get a feeling for the piecewise linear function g , Figure 6 provides a plot of g versus r at the maximum posterior mean $\beta_1 = 0.151$ corresponding to the first quadrant (green line) and at the prior mean $\beta = 0.0$ (red line). We observe that a putt from any given distance is more difficult in the first quadrant than on average.

For the generation of δ , we need to be aware of the constraint $\delta > 0$. We consider the proposal distribution $\text{Gamma}(0.25, 1.0)$. This is based on a subjective estimate $\delta = 0.25$ and a sufficiently large variance to capture the true value of δ .

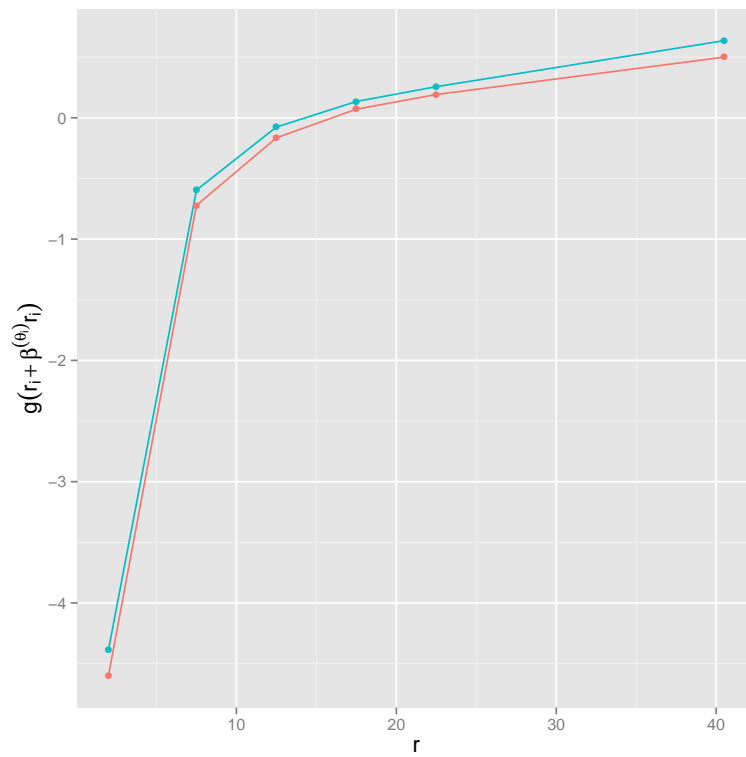


Figure 6: The piecewise linear function g evaluated at the maximum posterior mean $\beta_1 = 0.151$ (green line) and at $\beta = 0$ (red line).