Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

A SIMULATION MODEL TO ANALYZE THE IMPACT OF DISTANCE AND DIRECTION ON GOLF SCORES

Mark Broadie

Graduate School of Business
Columbia University
New York, NY 10027, USA

Soonmin Ko

Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

ABSTRACT

We develop a simulation model of the game of golf. The model accounts for realistic features of a golf course, including rough, sand, water, and trees, and includes many facets of golfer skill. The model is calibrated to extensive data for amateur and professional golfers. Using the calibrated simulation model we quantify the effect of increased tee shot distance and improved tee shot accuracy on golfer scores. Contrary to previous claims, we find that for long tee shots, directional accuracy has a greater impact on scores than distance.

1 INTRODUCTION

The object of the game of golf is to get the ball in the hole in the fewest number of strokes. Golfers seek to improve by playing, practicing, taking lessons, exercising and buying new equipment, among others. It is of considerable interest to golfers to understand which skill improvements have the biggest effect on scores. In this research, we focus on the improvement, in the form of lower scores, from hitting longer tee shots and from hitting straighter tee shots. For example, if a golfer could hit tee shots with a driver twenty yards further, what would be the change in average score? What would the change in scores be for hitting straighter drives? We develop a simulation model of golf to quantify the effects of changes in golfer skill on score, focusing on tee shots with a driver.

The simulation approach makes it possible to observe the effect of a change in one skill parameter while holding all other factors constant. In the golf context, it is difficult with standard statistical analysis to isolate the effect of a single parameter. For example, a regression of average score on driver distance will not provide a reliable estimate of the impact of changes in driver distance on score because of omitted-variable bias. Golfers who hit the ball further tend to be more skilled at all aspects of the game, so the regression coefficient will overestimate the impact of distance on score, because much of the improvement is due to factors that are correlated with increased distance and not due to distance itself. The inclusion of additional explanatory variables may not resolve the problem because of multicollinearity, i.e., it may be difficult to attribute changes in score to one explanatory variable versus another.

The simulation model we develop includes golf course features and various aspects of the skill of a golfer. The golf course model includes fairways, greens, rough, deep rough, sand, water, trees and out of bounds areas which are found on most golf courses. Fairways are areas of closely mown grass while the grass in the rough is much taller. Shots that start from the fairway are easier to hit and result in a tighter pattern of shot endpoints. It is more difficult to hit the ball from the rough because the grass comes between the club and the ball, and this leads to a more dispersed pattern of shot endpoints. Trees are an important feature of many golf courses because they can cause errant shots to stop abruptly. When trees lies between the ball and the hole, a golfer is often compelled to play a *recovery shot*, i.e., a short shot back to the fairway. Because driver distance and direction affect the likelihood of a shot ending in the rough or more severe hazard, it is important to include these golf course features in order to assess the impact of distance and direction on score. The golfer skill model includes parameters for maximum distance, distance errors and direction errors from various starting conditions (e.g., fairway, rough, sand and recovery positions).

An early golf simulation model was developed in Schied (1999). Scheid's model has four parameters representing aspects of golfer skill, but it does not include rough, sand, trees, or other golf course features. The distance and direction of tee shots

impact the fraction of shots which land in the rough and other more severe hazards, which in turn affects a golfer's score. A more detailed model of the golf course which includes various hazards is necessary to address the questions in this paper.

Golf strokes on a green, an area of finely cut grass which includes the hole, are called putts. Soley (1977) collected and analyzed amateur and professional putting data, mainly in the 1960s and early 1970s. Simulation models of putting, i.e., shots on the green, have been proposed in Gelman and Nolan (2002), Hoadley (1994), and Broadie and Bansal (2008). The putting model we use in this paper has two parameters (representing putting distance and direction errors), is similar to the model of Hoadley (1994), and it provides a good fit to putting data for golfers of all skill levels.

Cochran and Stobbs (1968) were the first to systematically collect and analyze golf shot data. Using data primarily from a professional tournament in 1964, they estimated how the average hole score depended on the distance of the hole, the accuracy of shots from various distances from the fairway and rough, and other similar quantities. They also attempted to answer the question of whether tee shot distance or tee shot directional accuracy has more of an impact on golfer scores. They concluded that "distance from the tee seems to count for more than accuracy" for professional golfers playing a golf course in tournament conditions. However, because they were working with data and not a controlled experiment, they needed to assume that direction errors would not cause a penalty beyond having to play from the rough, i.e., they assumed that increased direction error would not cause additional shots to land in deep rough, behind trees, or out of bounds. But since increased direction error does lead to higher scores because of additional hazards beyond normal rough, it is not clear whether their conclusion is valid.

Perhaps surprisingly, we find that an increase in driver distance, by itself, does not lead to a large reduction in average score. A 20-yard increase in driver distance leads to about a one shot reduction for a golfer with an average score of 100. An improvement in directional accuracy consistent with a 10% increase in fairways hit leads to a 2.6 shot reduction in average score for the same golfer. In contrast to Cochran and Stobbs (1968), we find that for tee shots with a driver, directional accuracy is substantially more important than distance.

The remainder of the paper is organized as follows. In Section 2 we formulate a stochastic dynamic programming model of the game of golf. Details of the model and simulation procedure are given in Section 3. Using extensive data for amateur and professional golfers, we calibrate our model and provide the main numerical results in Section 4.

2 MODEL OVERVIEW

We formulate a model of golf as a stochastic dynamic program. From an initial position (x_0, y_0) on the tee, the golfer attempts to get the ball into the hole (or cup) located at (x_h, y_h) in the fewest number of strokes (or shots). For the first shot, the golfer chooses a target $\mu_1 = (u_1, v_1)$ which represents the intended finishing point of the shot. For short holes, the target may coincide with the hole, but for longer holes the target is chosen within a maximum possible distance that the golfer can hit the ball. Because of physical execution errors and the effects of wind, uneven terrain and other factors, the finishing point of the shot will be random and will not coincide with the target. In general, the starting point of shot k is (x_{k-1}, y_{k-1}) and the random endpoint of shot k is

$$(x_k, y_k) = f(x_{k-1}, y_{k-1}, \mu_k, \varepsilon_k, \theta, \gamma). \tag{1}$$

The randomness of the outcome is contained in ε_k, which may be multidimensional and will depend on the skill parameters of the golfer (θ), the starting point of the shot, the target ($\mu_k = (u_k, v_k)$), and features of the golf hole (γ). The skill parameters of the golfer are denoted $\theta = (\theta_1, \ldots, \theta_s)$, which represent maximum shot distance, distance and direction errors from the fairway, rough, etc. The golf hole, denoted γ, contains the locations of the green, trees, rough, sand, water, and other features and may also include additional details, e.g., the height of trees, the difficulty of the rough, the firmness and speed of the green, etc. For notational simplicity, we typically suppress γ from the parameter list in f. Note that play on the hole is finished once the ball is in the cup, so $(x_h, y_h) = f(x_h, y_h, \mu, \varepsilon, \theta)$, i.e., once the ball is in the hole, it stays in the hole.

We assume that random shot errors are independent across shots, i.e., we assume that ε_k and ε_j are independent for $k \neq j$. Similarly, we assume that the golfer selects the target μ_k only depending on the initial position (x_{k-1}, y_{k-1}) of shot k and not on the outcome of previous shots. Hence, we define a policy (or strategy) μ where each target is given by $\mu_k = \mu(x_{k-1}, y_{k-1})$.

Since the object of the game of golf is to holeout in the fewest number of strokes, the "cost" g of any shot which does not start in the hole is one:

$$g(x_k, y_k) = \begin{cases} 1 & \text{if } (x_k, y_k) \neq (x_h, y_h) \quad \text{(i.e., the starting position of shot } k+1 \text{ is not the hole)} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The total number of shots taken on the hole is $\sum_{k=0}^{\infty} g(x_k, y_k)$, which depends on the policy μ through equation (1). A reasonable objective for a golfer is to minimize the expected number of shots to complete a hole. In some circumstances, e.g., with one or two holes left in a tournament, a golfer may want to choose a strategy which maximizes the probability of a low score at the expense of a higher average score. But over the course of one or more 18-hole rounds, golfers typically attempt to choose targets for each shot in order to minimize their average score. So the objective we take to minimize is:

$$J_\mu(x_0, y_0) = E\left[\sum_{k=0}^{\infty} g(x_k, y_k)\right]. \tag{3}$$

We suppress the dependence of J on θ for simplicity. The problem is to find a feasible strategy μ which minimizes the expected score on a hole: $J^*(x_0, y_0) = \min_\mu J_\mu(x_0, y_0)$. A feasible strategy means that the target μ_k for each shot k is reachable by the golfer with a given skill level. For example, if the golfer's maximum tee shot distance is 220 yards, then (roughly speaking) the target cannot be more than this distance from the starting position.

The goal in this paper is not to determine whether golfers play exactly optimal strategies. Since we are interested in determining how scores vary with golfer skill parameters, our first goal is to find a strategy so that the results generated by the model match the data. We solve for a greedy strategy, denoted μ^g, in which golfers ignore the dynamic nature of the solution. In this greedy strategy golfers choose the target for a shot taking into account the possible outcomes of the shot, but they ignore the possible outcomes, targets and course hazards for subsequent shots. Using individual shot data, we validate our model on two levels. First, we check that the model's expected score, $J_{\mu^g}(x, y)$, starting from position (x, y) and following strategy μ^g, matches the average score in the data from the same position. Second, we check that the model's distribution of shot endpoints (x_k, y_k) given a starting point (x_{k-1}, y_{k-1}), target μ_k^g, and skill parameters θ, matches the distribution of shot endpoints in the data. Details of the calibration results are given in Section 4.

3 MODEL DETAILS

In this section, we provide some details of the model and solution methods. We start by describing the model we use for the golf course and the approach used to solve for the golfer's greedy strategy. Then we describe the model for the distribution of shot endpoints which depends on the skill of the golfer and the initial position of the shot, i.e., whether the shot starts on a tee, in the fairway, rough, or sand, etc. The shot endpoint also depends on whether obstacles, such as trees, interfere with the trajectory of the shot. Finally, the endpoint of the shot depends on the amount the ball rolls, which depends on the condition of the course and whether the ball lands on the green, fairway, rough, sand or water (balls usually don't roll very far in water, but they have been known to skip across water on occasion).

3.1 Golf Course Model

The golf course is modeled as eighteen separate holes, with each hole contained in a rectangular subset of $\mathbb{R}^2$. Each hole may contain one or more of these objects: tee area, green, fairway, sand, water, rough, deep rough, out of bounds or trees. With the exception of trees and rough, each object is represented as a polygon, an ordered list of a finite number of points: $((s_1, t_1), \ldots, (s_m, t_m))$. Rough is the remaining area which is not covered by any polygonal object. Except for trees, all objects are non-overlapping. Trees are represented as cylinders with a center, radius and height. Trees can overlap any object, e.g., a point (s, t) can lie in the rough and within a tree. Objects may have additional descriptive parameters. For example, the rough has a parameter which indicates the height of grass in the rough (which determines the amount a ball will roll in the rough), and the green may have parameters to indicate its firmness and speed.

A golf hole map (denoted γ) is a collection of these polygonal objects, trees, and two additional points. The tee, or starting point of the hole, is (x_0, y_0) and the hole (or cup) is the point (x_h, y_h). (Note that "hole" has two meanings: it can refer to the entire playing area or it can refer to the 4.25-inch diameter cup where the ball eventually finishes. The meaning should be clear from the context.) Hole maps were created using Google Earth together with a custom hole editor program written in the Java programming language. A sample representation of a hole is given in Figure 1. This allows the creation of very accurate representations of actual golf courses.

3.2 Greedy Strategy

If the ball is in the hole, the hole is finished and there is no strategy to decide. If the ball is on the green, then the putting strategy from Broadie and Bansal (2008) is used. If the current position (x_{k-1}, y_{k-1}) is not on the green, the Bellman equation

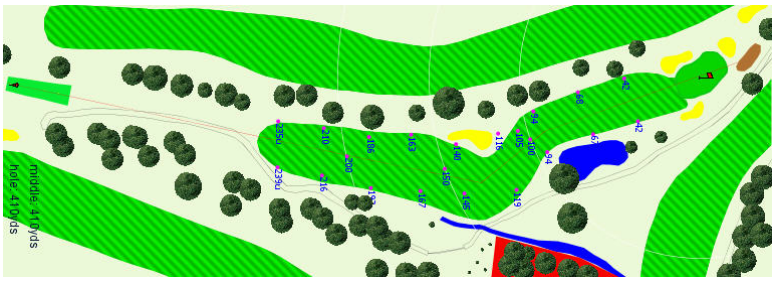


Figure 1: Hole map representation of a golf hole. Fairways are displayed in light and dark green hatching, trees are the circular objects, water is displayed in blue, sand in yellow, deep rough in brown, and out of bounds in red. The tee is on the left and the hole is near the upper right corner.

is:

$$J^*(x_{k-1},y_{k-1}) = \min_{\mu_k} E\left[1+J^*(f(x_{k-1},y_{k-1},\mu_k,\varepsilon_k,\theta))\right] \tag{4}$$

The optimal target for shot k is the solution of:

$$\min_{\mu_k} E\left[J^*(f(x_{k-1},y_{k-1},\mu_k,\varepsilon_k,\theta))\right] = \min_{\mu_k} E\left[J^*(x_k,y_k) \mid x_{k-1},y_{k-1},\mu_k,\theta\right]$$

Suppose there is an *average score* function $H(x,y)$ which satisfies $J^*(x,y) \approx aH(x,y)+b$ for some constants a and b with $a>0$. Then

$$\min_{\mu_k} E\left[J^*(x_k,y_k) \mid x_{k-1},y_{k-1},\mu_k,\theta\right] \approx \min_{\mu_k} E\left[aH(x_k,y_k)+b \mid x_{k-1},y_{k-1},\mu_k,\theta\right] \tag{5}$$

$$= a\min_{\mu_k} E\left[H(f(x_{k-1},y_{k-1},\mu_k,\varepsilon_k,\theta))\right]+b \tag{6}$$

$$\approx a\left(\min_{\mu_k} \frac{1}{N}\sum_{i=1}^{N} H(f(x_{k-1},y_{k-1},\mu_k,\varepsilon_k^{(i)},\theta))\right)+b \tag{7}$$

The approximation in (7) comes from replacing the random variable ε_k by an equiprobable N-point approximation $\varepsilon_k^{(i)}$, $i=1,\ldots,N$. Equation (7) says that the golfer chooses a target $\mu_k=(u_k,v_k)$ for shot k to minimize the average remaining score over the N equiprobable shot endpoints. Denote the solution to (7) for shot k by μ_k^g.

In most situations we found that it was adequate to take $N=1$ and to solve the simpler problem

$$\min_{\mu_k} H(\mu_k). \tag{8}$$

where $\mu_k = f(x_{k-1},y_{k-1},\mu_k,\varepsilon_k^{(1)},\theta)$. Here the golfer chooses the target $\mu_k=(u_k,v_k)$ assuming that the endpoint (x_k,y_k) of shot k will be the target while ignoring the randomness in the outcomes. The interpretation is that golfer chooses the target as if the shot will be executed perfectly. In this approach, if the hole is within the golfer's range, the golfer will aim at the hole (i.e., choose the target $\mu_k=(x_h,y_h)$). For longer shots on straight holes, it means the golfer will choose the target which is straight down the fairway and as close to the hole as possible. This approach ignores potential asymmetries in the placement of hazards (e.g., when there is rough to the right of the green but water to the left of the green) or asymmetries in the shot distribution (e.g., when poor shots are more likely to finish short of the target).

The focus of this paper is to determine the effect of changes in golfer skill on average scores and is not to determine whether golfers are playing optimal strategies. The latter problem is more difficult because the strategy used by golfers is unobservable. In our case, we assume golfers follow a simple and reasonable strategy, and then we directly check whether the simulation model output matches the data, in effect, verifying that the data is generated *as if* golfers are following the greedy strategy.

3.3 Solving for the Greedy Target

The greedy target problem (8) is to find the feasible target which minimizes the average score function $H(\mu_k)$. Recall that a feasible target is simply a location in the plane $\mathbb{R}^2$ that is within a specified distance from the starting position (given by one of the golfer skill parameters $\theta_i = D$). In our model, the maximum distance D depends on the location of the initial position, e.g., whether the initial position is on the tee, in the fairway, rough, sand, deep rough, or in a recovery position (i.e., a direct path to the hole is blocked by trees). The greedy target problem (8) can be written in more detail as:

$$\begin{array}{ll} \underset{(u_k,v_k)}{\text{minimize}} & H(u_k,v_k) \\ \text{subject to:} & ||(x_{k-1},y_{k-1})-(u_k,v_k)|| \le D \end{array} \tag{9}$$

where initial position of the shot is (x_{k-1},y_{k-1}) (assumed not to be on the green), the target is $\mu_k = (u_k,v_k)$ and $||(x_{k-1},y_{k-1})-(u_k,v_k)|| = \sqrt{(x_{k-1}-u_k)^2+(y_{k-1}-v_k)^2}$ is the standard Euclidean distance norm. This is a straightforward optimization problem that can be solved with a brute force search procedure over a grid of candidate targets. The number of points in the grid is specified in the simulation procedure. Candidate targets are first checked for feasibility. The objective is evaluated at all of the remaining feasible candidate targets and the best target is recorded.

The success of this greedy procedure relies on finding a suitable function $H(x,y)$. The form of the average score function H that we use is:

$$H(x,y) = \begin{cases} h_0(d_h(x,y)) & \text{if } (x,y) \text{ is in the green,} \\ h_1(d_h(x,y)) & \text{if } (x,y) \text{ is in tee or fairway,} \\ h_2(d_h(x,y)) & \text{if } (x,y) \text{ is in rough,} \\ h_3(d_h(x,y)) & \text{if } (x,y) \text{ is in sand,} \\ h_4(d_h(x,y)) & \text{if } (x,y) \text{ is in deep rough,} \\ h_5(d_h(x,y)) & \text{if } (x,y) \text{ is in recovery,} \end{cases} \tag{10}$$

where $d_h(x,y) = ||(x,y)-(x_h,y_h)||$ is the distance from the initial position to the hole. Note that special situations, such as the initial position lying in water or out of bounds, are handled separately. The functions $h_i(d)$, $i = 0,\ldots,5$, are continuous piecewise linear functions of d, typically with three pieces (more pieces are used in h_0). Evaluating the function H requires answering yes or no to the questions "Does (x,y) lie in a given area (e.g., rough, sand, etc.)?" Since objects are represented as polygons these questions can be answered quickly using standard algorithms (see, e.g., chapter 21 in Press et al. (2007)). The coefficients of the piecewise linear functions are estimated using data and standard statistical methods. We choose H to represent the average score function of a highly skilled golfer with a zero handicap (called a scratch golfer).

In the next subsections, we provide details on the simulation of shot endpoints in equation (1). The shot endpoint distribution depends on the level of golfer skill and on the condition of the starting point, i.e., whether the shot starts on the tee, fairway, rough, sand, etc.

3.4 Shot Simulation

The core of our golf simulation is the routine for generating the random result a single golf shot given a starting position, target, golfer skill parameters, and course parameters. To start, we focus on shots that start off of the green (putts are treated separately) and we ignore the possibility of the ball hitting a tree. A well-struck golf shot will travel through the air and then bounce and roll on the ground before coming to rest. The amount of bounce and roll depends on the ball's landing velocity and angle, the firmness of the landing area, the height of the grass, and other factors. We will return to the modeling of bounce and roll later.

It is convenient to simulate shot distances and directions which are then transformed to shot endpoints. Given an initial ball position (x_{k-1},y_{k-1}) and target $\mu_k = (u_k,v_k)$, the shot endpoint (x_k,y_k) is simulated as follows. Let d_μ represent the distance to the target: $d_\mu = ||(x_{k-1},y_{k-1})-(u_k,v_k)||$. Define α_μ to be the angle between the vector from the initial position to the target and horizontal, i.e., $\cos(\alpha_\mu) = (u_k - x_{k-1})/d_\mu$ and $\sin(\alpha_\mu) = (v_k - y_{k-1})/d_\mu$. Suppose a random angle α and

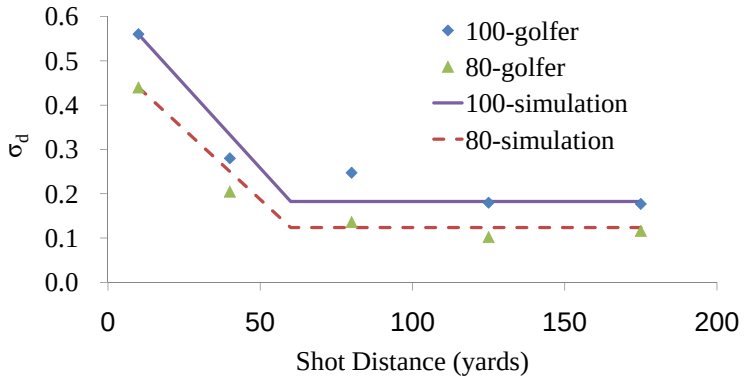


Figure 2: Relative distance errors, σ_d, as a function of initial distance to the hole for the 80-golfer and 100-golfer groups from the fairway.

random shot distance l are generated according to specified distributions. The shot endpoint is (x_k, y_k) with

$$x_k = x_{k-1} + l\cos(\alpha_\mu - \alpha) \tag{11}$$
$$y_k = y_{k-1} + l\sin(\alpha_\mu - \alpha) \tag{12}$$

The convention is that $\alpha > 0$ corresponds to shots that finish to the right of the target.

We generate the random angle according to $\alpha \sim N(0, \sigma_\alpha^2)$. The interpretation of a zero mean angle is that golfers do not systematically miss to the left or right of the target. The parameter σ_α represents the directional skill of the golfer, with larger values generating "wilder" shots. We find that the normality and zero mean assumptions are generally consistent with the data.

We generate the random shot distance by $l = d_\mu(1 + \sigma_d X)$ where X is a random variable with standard deviation one and satisfying $X \geq -1$. The parameter σ_d represents the *relative* distance error of the golfer. We take α and X to be independent, which is consistent with the data. For putts we take X to be a standard normal random variable. For shots other than putts, X often has a negative mean (i.e., shots on average end up short of the target), negative skewness and excess kurtosis. The main reason is that the maximum distance a shot can travel with a given club is limited by the physical skill of the golfer, but poor shots may travel very short distances (shots colorfully referred to as "fat," "chili-dipped" or "chunked") leading to negative means and negatively skewed distance distributions. Another reason is that many golf holes have fairways leading up to the green and rough or worse trouble beyond the green, so golfers often prefer mishit shots to finish short of the target rather than beyond the target. When the ball starts close to the hole, e.g., within sixty yards, the mishits of amateur golfers can also travel relatively far beyond the target (shots referred to as "thin" or "skulled"). For these reasons, we typically do not take X to be normally distributed for shots other than putts.

Our model for distance error, X, is taken to be a mixture of two distributions, where the first distribution represents a "good" shot and the second a "bad" shot:

$$X = \begin{cases} X_1 & \text{with probability } p & \text{"good shot"} \\ X_2 & \text{with probability } 1-p & \text{"bad shot"} \end{cases} \tag{13}$$

The mixture distribution approach allows use to incorporate skewness and excess kurtosis in the simulation model. Details are omitted for space reasons.

Of primary importance are the following qualitative features of the shot distance model. First, distance errors, e.g., as summarized by σ_d, are larger for less-skilled golfers. Second, distance errors are larger from the rough than the fairway or tee, and even larger from the sand. Finally, relative distance errors are nearly constant for distances greater than sixty yards, but increase as the distance to the hole decreases. Figure 2 illustrates the dependence of σ_d on the distance to the hole in the data and shows the piecewise linear approximation used in the simulation.

3.5 Recovery Shot and Tree Hit

As direction error increases shots are more likely to miss the fairway and land in the rough, hit a tree, or go out of bounds. A ball which hits a tree does not travel as far and leaves the golfer with a longer shot to the hole. Even if the ball does not hit a tree, the path to the hole on the next shot may be blocked by trees, a situation referred to as a recovery shot. In this case, the golfer is often forced to play a short shot out of the trees toward the fairway, rather than playing directly to the hole. So a wild tee shot can hinder the golfer's next shot in a number of ways. In order to assess the effect of direction error, it is important to model the likelihood of a tree being hit on a shot and to determine whether a ball is in a recovery position.

Until now, we've described the shot endpoint as a function of the distance and direction of the shot. In our simulation, the final endpoint of the shot may be considerably shorter than the shot distance random variable because of the presence of trees. Using the position and radius of trees in our golf course model, we incorporate tree hits and recovery shots in the simulation. The exact algorithms are omitted for space reasons. However, since our database contains information on tree hits and recovery shots, the simulation results can be calibrated to the data.

3.6 Carry and Roll

In addition to trees affecting the final endpoint of a shot, the endpoint may differ because of varying amount of roll. For example, a tee shot which lands on the fairway may travel much further than a similar shot which lands in the rough, because the higher grass in the rough reduces the amount that the ball will roll. A similar shot which lands in water will typically not roll at all.

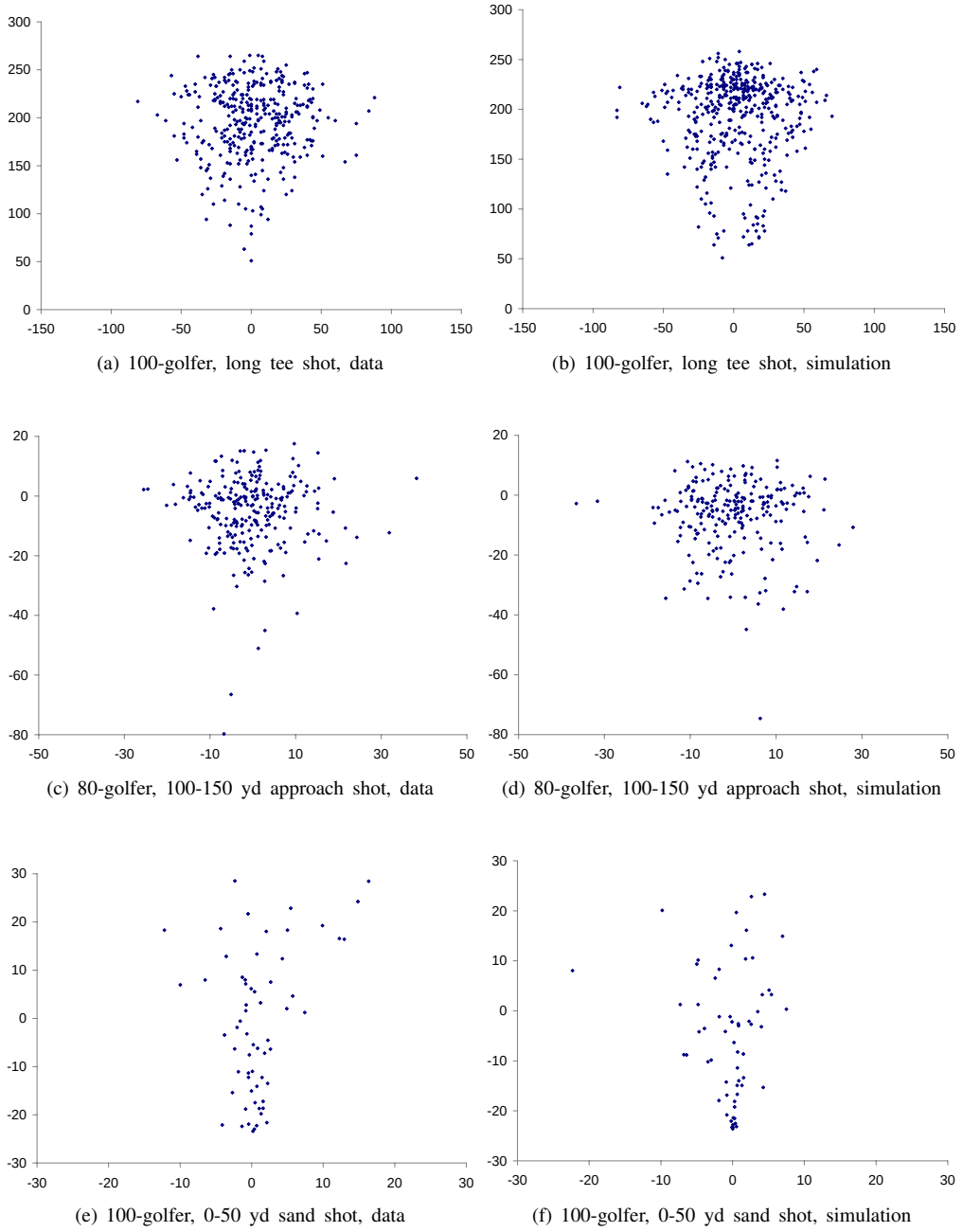
Each shot trajectory is divided into two parts, the trajectory in the air called *carry*, and the trajectory on the ground called *roll*. The amount of roll depends on the ground conditions (e.g., green, fairway, rough or sand) and on the ball's velocity and angle when it hits the ground. These impact conditions depend on the ball's initial velocity, angle, and spin rate which are determined by the club used and the speed of the golfer's swing. Although the physics are complicated, the main effects of a shot landing on the fairway are these. The carry fraction, i.e., carry length divided by the total shot length, increases with the total shot length for shots hit from a tee with a driver. This happens because long shots land with a greater impact angle and roll less. For shots hit from the fairway, which land on the fairway, the carry fraction decreases with the total shot length. Shorter shots are hit with clubs with higher loft, leading to larger initial angles, large impact angles and less roll. For shots hit from the rough, which land on the fairway, the carry fraction is less than for similar shots from the fairway. Shots hit from the rough have less spin and roll more after they land. The carry fractions used in the simulation are based on results from Werner and Greig (2000).

4 RESULTS

The data used in the development of the simulation model comes from two sources, Golfmetrics and ShotLink^{TM}. Golfmetrics is a custom software application that was created to capture and store golfer shot data. The database currently has over 55,000 shots from over 160 (primarily amateur) golfers during 2005-2009. See Broadie (2008) for more details on the Golfmetrics program and data. The PGA TOUR provided shot data for all shots at all PGA tournaments between 2004 and 2008. Their data was collected using the ShotLink^{TM} system. In these results, we focused on two groups of amateur golfer: the 80-golfer group, representing rounds with an 18-hole total score in the range 78-82 and the 100-golfer group, representing rounds with an 18-hole total score in the range 96-107.

First we illustrate the empirical distributions of shot patterns for the two golfer categories for homogenous subgroups of shots. We compare the shot patterns from the data with similar shot patterns generated from the simulation model. These graphical results are useful to get a flavor for the distributions that are simulated. Then in order to validate the simulation model, we provide more quantitative measures of the fit of the simulation model to the data.

For a given golfer, or golfer skill level, shot distributions depend on the initial position of the ball (e.g., fairway, rough or sand), so the data are separated into relatively homogeneous subgroups for presentation and analysis. For example, the initial position might be restricted to lie in the fairway between d_1 and d_2 yards from the hole. Figure 3 shows data and simulated patterns for representative subgroups. In order to present a number of different shots on a single graph, the long tee shots are shifted and rotated so that start position is (0,0) and the hole is in the direction of the vertical axis. Approach and sand shots are shifted, rotated and also scaled so the plotted distance from the start position to the hole is the same. The general similarity of the data and simulated shots is evident in Figure 3. However, in order to verify that the simulation accurately represents the golfer, we present additional quantitative results next.



(a) 100-golfer, long tee shot, data

(b) 100-golfer, long tee shot, simulation

(c) 80-golfer, 100-150 yd approach shot, data

(d) 80-golfer, 100-150 yd approach shot, simulation

(e) 100-golfer, 0-50 yd sand shot, data

(f) 100-golfer, 0-50 yd sand shot, simulation

Figure 3: Comparison of shot patterns in the data and simulation. Golfer data on the left; simulated results on the right. Panels (a) and (b): Tee shots start at (0,0). Panels (c) and (d): Approach shots start at (0,-125) with the hole at (0,0). Panels (e) and (f): Sand shots start at (0,-25) with the hole at (0,0).

Table 1 presents summary statistics for 18-hole rounds for actual data and simulated results. Most importantly, the simulation matches (within statistical error) the average 18-hole score for a golfer group. However, in order to have confidence in the validity of the simulation results when the parameters change, it is important that more detailed statistics also match. The green in regulation (GIR) statistic represents the fraction of par-i holes in which the golfer's ball is on the green in $i-2$ shots or less, for $i = 3, 4$, or 5. The par-3, par-4, and par-5 statistics represent the average score on par-3, par-4, and par-5 holes, respectively. Table 1 shows that the simulation results match the data for both the 100-golfer and 80-golfer groups

for these statistics. In addition, we check whether the average number of shots in an 18-hole round in various categories (e.g., putts, long shots, short shots, sand shots, and recovery shots) match.

In order to check whether the quality of putts, long shots, short shots, etc., match between the data and simulation, we define the *shot value* measure. In theory, the quality of a single shot k is $J^*(x_{k-1},y_{k-1}) - J^*(x_k,y_k) - 1$, which measures whether the shot decreased the golfer's average score by more or less than one, which is the cost of a single shot. Because the function J^* is not known, we measure the quality of a shot using the known function $H(x,y)$ and define shot value as:

$$SV(x_{k-1},y_{k-1},x_k,y_k) = H(x_{k-1},y_{k-1}) - H(x_k,y_k) - 1. \tag{14}$$

Because H represents the average score function for a scratch golfer (a highly-skilled amateur), the average shot value for the 80-golfer group will be negative for most or all groups of shots. The average shot value for the 100-golfer group will be tend to be more negative, while PGA TOUR golfers will have positive average shot values. Since shot value measures the quality of a shot relative to a benchmark scratch golfer, we check whether the average shot values in the data match average shot values from the simulation. As shown in Table 1, in almost all cases the simulation results match the data, within statistical error.

Table 1: Round statistics comparing data and simulation results for the 100-golfer and 80-golfer groups. GIR stands for greens in regulation. The labels # long, # short, # sand and # recov refer to the average number of long shots, short shots, sand shots and recovery shots, respectively, in an 18-hole round. The average shot value for shots in a subgroup is denoted SV. The numbers in the parentheses are the standard errors of the data. The standard errors of the simulation results are much smaller and are omitted.

	100-golfer			80-golfer		
	Data	(Std err)	Sim	Data	(Std err)	Sim
Score	100.0	(0.7)	100.3	80.0	(0.1)	80.2
GIR (%)	12.8	(0.9)	11.5	45.9	(0.9)	42.5
Par-3	4.33	(0.06)	4.32	3.47	(0.03)	3.41
Par-4	5.74	(0.05)	5.83	4.65	(0.02)	4.69
Par-5	6.85	(0.11)	6.80	5.36	(0.04)	5.41
# Putts	34.7	(0.5)	34.9	32.0	(0.3)	31.5
Putt SV	-0.12	(0.01)	-0.11	-0.05	(0.01)	-0.05
# Long	34.1	(0.6)	35.1	30.2	(0.2)	30.9
Long SV	-0.41	(0.02)	-0.39	-0.08	(0.01)	-0.07
# Short	16.7	(0.5)	19.5	10.8	(0.2)	12.6
Short SV	-0.30	(0.02)	-0.27	-0.09	(0.01)	-0.07
# Sand	2.2	(0.3)	2.1	1.7	(0.1)	1.5
Sand SV	-0.43	(0.07)	-0.40	-0.13	(0.04)	-0.04
# Recov	4.4	(0.4)	4.3	1.8	(0.1)	1.2
Recov SV	-0.21	(0.04)	-0.23	-0.03	(0.03)	0.04

Additional shot-level simulation validation results are given in Table 2. As before, shots are grouped by distance to the hole and initial position of the ball. "Approach 100-150" refers to shots that start between 100 and 150 yards from the hole and "Sand Shot" refers to shots that start from the sand (between 0 and 50 yards from the hole). "Long tee shot" refers to tee shots on par-4 and par-5 holes, which golfers typically use a driver to obtain maximum shot distance. Table 2 shows that the simulation results match the data, within statistical error, in most cases.

These results show that the simulation model output is consistent with data for both golfer groups. Now we can vary simulation parameters and be reasonably confident that the resulting simulated scores are indicative of scores that would be obtained by real golfers with those skills or equipment. In particular, we vary the golfer's driver distance, as measured by the 75th percentile of long tee shot distance (or driver distance). We also vary the golfer's long tee shot directional accuracy (or driver directional accuracy), as measured by σ_α, the standard deviation of directional error. When we vary these parameters, it is important to note that they only apply to long tee shots with a driver. That is, the distances of other long shots (e.g., when the golfer is on the fairway or rough and is far from the hole) is not changed. Similarly, the directional accuracy of shots other than tee shots with the driver is not changed.

Table 2: Shot statistics for data and simulation results for the 100-golfer and 80-golfer groups. SV stands for shot value. % green, % sand and % frwy stand for the percentage of shots landing on green, sand, and fairway, respectively. % recov is the percentage shots which lead to recovery situations on the next shot. 75% dist is 75th percentile of shot distance. σ_α is the standard deviation of direction error. The numbers in the parentheses are the standard errors of the data. The standard errors of the simulation results are much smaller and are omitted.

	100-golfer						80-golfer					
	Short 0-100			Approach 100-150			Short 0-100			Approach 100-150		
	Data	Std err	Sim	Data	Std err	Sim	Data	Std err	Sim	Data	Std err	Sim
SV	-0.30	(0.02)	-0.27	-0.36	(0.03)	-0.34	-0.09	(0.01)	-0.07	-0.02	(0.02)	-0.03
% Green	71.5	(1.6)	72.9	32.5	(3.8)	25.9	86.4	(0.8)	87.7	63.8	(1.7)	50.6
% Sand	2.6	(0.6)	3.0	11.7	(2.1)	10.9	1.7	(0.3)	1.4	7.4	(1.0)	8.0
% Frwy	6.8	(0.9)	6.6	15.0	(2.4)	16.1	2.7	(0.5)	2.6	7.7	(1.4)	11.6
% Recov	0	(0)	0	0	(0)	0	0	(0)	0	0	(0)	0

	Approach 150-200			Sand Shot			Approach 150-200			Sand Shot		
	Data	Std err	Sim	Data	Std err	Sim	Data	Std err	Sim	Data	Std err	Sim
SV	-0.45	(0.04)	-0.49	-0.43	(0.07)	-0.40	-0.06	(0.01)	-0.03	-0.13	(0.04)	-0.04
% Green	18.0	(2.5)	15.9	50.8	(7.0)	55.0	43.7	(2.1)	39.4	72.6	(2.8)	84.2
% Sand	9.0	(1.8)	7.3	16.9	(4.9)	13.8	11.4	(1.4)	7.8	4.2	(0.8)	3.6
% Frwy	12.6	(3.0)	15.6	3.1	(0)	1.6	12.1	(1.1)	12.2	1.1	(2.3)	0.7
% Recov	1.2	(0)	0.8	0	(0)	0	0	(0)	0	0	(0)	0

Long Tee Shot

	100-golfer			80-golfer		
	Data	Std err	Sim	Data	Std err	Sim
SV	-0.43	(0.03)	-0.37	-0.09	(0.01)	-0.09
75% Dist	225.0	(3.1)	224.1	250.1	(1.8)	250.6
σ_α	7.9	(0.6)	7.9	5.5	(0.2)	5.5
% Frwy	43.9	(2.1)	43.2	53.7	(1.4)	49.8
% Penalty	7.8	(1.6)	6.0	1.8	(0.3)	3.2
% Recov	22.5	(2.4)	19.9	11.6	(0.9)	7.5

The main results are presented in Table 3. If a 100-golfer's driver distance increases by 20 yards, his average score will decrease by about 1.2 shots on a typical course. (Table 2 shows that the 75th-percentile driver distance for the 100-golfer group is 225 yards and 250 yards for the 80-golfer group.) An increase of 80 yards only leads to an average score decrease of about three shots. Perhaps surprisingly, an increase in driver distance, by itself, does not lead to a large reduction in average score, because more drives miss the fairway and end up in recovery situations or out of bounds. The reduction in scores from the increase in length is partially offset by the worse condition of some of the shots. The improvement from a 20-yard increase for the 80-golfer is smaller – only 0.5 shots on average.

If a 100-golfer's driver directional accuracy improves by two degrees, the average score will decrease by about 2.6 shots. For reference, Table 2 shows that the driver directional accuracy for the 100-golfer group is 7.9° and 5.5° for the 80-golfer group. A 2° decrease in driver directional error leads to 10% more fairways being hit, i.e., an increase from about 43% to 53%. More importantly, the fraction of recovery shots decreases by about 5% (from 20% to 15%) and the fraction of tee shots which finish out of bounds decreases from 4.4% to 2%. The reduction of shots which end up "in trouble" has a greater impact on average score. For an 80-golfer, a 20-yard increase in driver distance leads to one-half shot reduction in average score. A one-degree improvement in directional accuracy (which gives an 8% increase in fairways hit) leads to about one shot reduction in average score. Additional simulation results (not presented) show that the results from parameter changes are approximately additive. For a 100-golfer on a typical course, a 20-yard increase in driver distance and a 2-degree improvement in directional accuracy would lead to a 3.8 shot decrease in average score (1.2 shots from extra distance and

Table 3: Changes in average 18-hole scores for 100-golfer and 80-golfer groups. Frwy is the % on fairway for long tee shots. GIR is the % greens in regulation. Recov is the percentage of recovery shots that follow from long tee shots. OB is the percentage of out of bounds shots that follow long tee shots. The *rough only* results refer to a hypothetical golf course with no trees, no OB, no sand and no water. Distance changes refer to changes in the 75th-percentile driver distance. Direction changes refer to changes in σ_α, the standard deviation of driver directional accuracy.

		100-golfer					**80-golfer**				
Typical course		Score	Frwy	GIR	Recov	OB	Score	Frwy	GIR	Recov	OB
Distance	-40	2.6	4.7	-3.9	0.8	-1.1	-	-	-	-	-
	-20	1.2	3.2	-2.0	0.1	-0.4	0.7	6.0	-3.3	-0.7	-0.2
	20	-1.2	-4.0	2.6	0.8	0.5	-0.5	-6.6	2.9	1.7	0.2
	40	-1.8	-9.7	4.6	3.3	0.5	-1.0	-12.9	4.8	3.6	0.6
	60	-2.3	-14.1	6.7	4.2	1.4	-1.6	-20.7	6.8	4.5	1.6
	80	-3.1	-18.7	8.8	6.1	2.0	-2.0	-24.6	8.8	4.1	2.6
Direction	-3	-3.9	17.3	2.9	-7.5	-3.3	-2.9	31.1	7.9	-4.5	-2.2
	-2	-2.6	10.1	1.8	-4.7	-2.4	-2.1	18.0	5.5	-3.0	-1.9
	-1	-1.4	4.5	0.9	-2.0	-1.1	-1.1	7.8	2.6	-1.4	-1.2
	1	0.9	-3.7	-0.3	1.4	1.4	0.9	-6.3	-1.6	1.2	1.4
	2	1.9	-6.9	-0.9	2.7	2.5	1.9	-10.6	-3.7	2.0	2.9
	3	3.0	-9.7	-1.5	3.7	3.9	2.7	-14.1	-4.9	2.4	4.5
Base case		100.3	42.7	11.5	19.9	4.4	80.2	50.1	42.5	7.5	2.2
(Std err)		(0.1)	(0.2)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)
No trees		Score	Frwy	GIR	Recov	OB	Score	Frwy	GIR	Recov	OB
Distance	-40	2.8	4.7	-4.9	0	-2.0	-	-	-	-	-
	-20	1.6	3.2	-2.7	0	-0.9	1.4	3.4	-5.3	0	-0.3
	20	-2.0	-2.8	4.2	0	0.3	-0.9	-5.4	4.0	0	0.1
	40	-3.3	-7.0	7.7	0	0.8	-2.0	-10.4	8.0	0	0.3
	60	-4.7	-10.2	11.8	0	1.5	-3.4	-17.0	12.5	0	1.7
	80	-6.2	-14.8	16.1	0	3.0	-4.2	-19.1	16.4	0	3.3
Direction	-3	-2.5	17.2	2.3	0	-5.0	-1.9	30.9	5.1	0	-3.4
	-2	-1.7	10.4	1.5	0	-3.2	-1.4	17.8	3.3	0	-2.9
	-1	-0.9	4.9	0.6	0	-1.6	-0.7	7.7	1.4	0	-1.7
	1	0.7	-3.3	-0.3	0	1.6	0.8	-6.5	-1.3	0	1.7
	2	1.5	-6.3	-0.6	0	3.3	1.3	-11.0	-2.4	0	3.4
	3	2.4	-8.7	-1.4	0	4.7	2.1	-14.7	-3.5	0	5.3
Base case		97.0	44.1	13.3	0	7.0	78.4	53.7	47.5	0	3.5
(Std err)		(0.1)	(0.2)	(0.1)	(0)	(0.1)	(0.1)	(0.2)	(0.2)	(0)	(0.1)
Rough only		Score	Frwy	GIR	Recov	OB	Score	Frwy	GIR	Recov	OB
Distance	-40	3.6	4.9	-5.8	0	0	-	-	-	-	-
	-20	1.8	3.0	-3.2	0	0	1.0	3.7	-5.3	0	0
	20	-1.7	-2.5	4.2	0	0	-1.3	-4.9	5.3	0	0
	40	-3.4	-6.7	8.6	0	0	-2.6	-9.8	11.1	0	0
	60	-5.1	-10.2	13.8	0	0	-4.2	-16.7	16.1	0	0
	80	-6.8	-14.4	19.9	0	0	-5.2	-19.1	20.4	0	0
Direction	-3	-0.8	18.7	1.8	0	0	-1.0	32.8	3.6	0	0
	-2	-0.5	11.0	1.3	0	0	-0.6	19.7	2.2	0	0
	-1	-0.2	5.0	0.5	0	0	-0.3	8.8	0.9	0	0
	1	0.2	-3.8	-0.4	0	0	0.1	-6.2	-0.6	0	0
	2	0.4	-7.4	-0.6	0	0	0.3	-11.4	-1.3	0	0
	3	0.5	-9.7	-0.9	0	0	0.4	-15.3	-2.0	0	0
Base case		91.2	41.4	14.7	0	0	76.3	51.7	50.0	0	0
(Std err)		(0.1)	(0.2)	(0.1)	(0)	(0)	(0.04)	(0.2)	(0.2)	(0)	(0)

2.6 shots from better accuracy) with a 6% increase in fairways hit from 43% to 49% (a 4% reduction from extra distance and a 10% increase from better accuracy).

These results would differ for courses with wider fairways and fewer trees and other hazards. To isolate these effects and illustrate the range of results that might be expected for other courses, Table 3 presents results for a course with no trees, but still with water, sand, deep rough, and out of bounds hazards. In this case, a 20-yard increase in driver distance reduces a 100-golfer's average score by 2.0 shots (more than the 1.6 shots on a typical course). A 2° improvement in driver accuracy reduces a 100-golfer's average score by 1.7 shots (less than the 2.6 shots on a typical course). When there is less trouble, distance becomes relatively more important than accuracy.

Table 3 also presents results for a more extreme "rough only" case where the course does not have any trees, out of bounds, sand or water. The 80-golfer's average score decreases from 80 on a typical course to 76.3 on the rough-only course. A 20-yard increase in driver distance reduces the 80-golfer's average score by an additional 1.3 shots in the rough-only case, compared to a reduction of 0.5 shots for a typical course. When the only trouble is rough, distance improvements are relatively more important than accuracy improvements.

5 ACKNOWLEDGMENT

Thanks to Lou Lipnickey for programming and development of the Golfmetrics software used in this project. Thanks to the PGA TOUR for providing the data collected with ShotLink^{TM}. This research was supported in part by a grant from the United States Golf Association.

REFERENCES

Broadie, M. 2008. Assessing golfer performance using golfmetrics. In *Science and Golf V: Proceedings of the World Scientific Congress of Golf*, ed. D. Crews and R. Lutz, 253–262. Mesa, Arizona: Energy in Motion, Inc.

Broadie, M., and M. Bansal. 2008. A simulation model to analyze the impact of hole size on putting in golf. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Moench, and O. Rose, 2826–2834: The Society for Computer Simulation.

Cochran, A., and J. Stobbs. 1968. *Search for the perfect swing: The proven scientific approach to fundamentally improving your game*. Chicago, Illinois: Triumph Books.

Gelman, A., and D. Nolan. 2002. A probability model for golf putting. *Teaching Statistics* 24:93–95.

Hoadley, B. 1994. How to improve your putting score without improving. In *Science and Golf II: Proceedings of the World Scientific Congress of Golf*, ed. M. Farrally and A. Cochran, 186–192. London: E & FN Spon.

Press, W., S. Teuklosky, W. Vetterling, and B. Flannery. 2007. *Numerical recipes: The art of scientific computing*. 3rd ed. Cambridge, UK: Cambridge University Press.

Schied, F. 1999. Yardage rating by the curve. In *Science and Golf III: Proceedings of the 1998 World Scientific Congress of Golf*, ed. M. Farrally and A. Cochran, 371–376. UK: Human Kinetics.

Soley, C. 1977. *How well should you putt? a search for a putting standard*. Soley Golf Bureau.

Werner, F., and R. Greig. 2000. *How golf clubs really work and how to optimize their designs*. Origin Inc.

AUTHOR BIOGRAPHIES

MARK BROADIE is the Carson Family Professor of Business at the Graduate School of Business, Columbia University. His research focuses on issues in financial engineering, with a particular focus on the design and analysis of efficient Monte Carlo methods for security pricing and risk management. His email address is <mnb2@columbia.edu> and his webpage is <www.columbia.edu/~mnb2/broadie/>.

SOONMIN KO is a doctoral student in the Industrial Engineering and Operations Research department at Columbia University. His research interests are in the areas of financial engineering and dynamic programming. His email address is <sk2822@columbia.edu>.