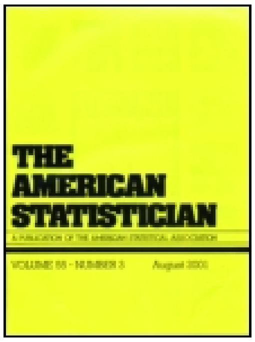


THE AMERICAN STATISTICIAN

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

A Spatial Analysis of Basketball Shot Chart Data

Brian J Reich, James S Hodges, Bradley P Carlin & Adam M Reich

To cite this article: Brian J Reich, James S Hodges, Bradley P Carlin & Adam M Reich (2006) A Spatial Analysis of Basketball Shot Chart Data, The American Statistician, 60:1, 3-12, DOI: 10.1198/000313006X90305

To link to this article: https://doi.org/10.1198/000313006X90305

Published online: 01 Jan 2012.

General

A Spatial Analysis of Basketball Shot Chart Data

Brian J. Reich, James S. Hodges, Bradley P. Carlin, and Adam M. Reich

Basketball coaches at all levels use shot charts to study shot locations and outcomes for their own teams as well as upcoming opponents. Shot charts are simple plots of the location and result of each shot taken during a game. Although shot chart data are rapidly increasing in richness and availability, most coaches still use them purely as descriptive summaries. However, a team's ability to defend a certain player could potentially be improved by using shot data to make inferences about the player's tendencies and abilities. This article develops hierarchical spatial models for shot-chart data, which allow for spatially varying effects of covariates. Our spatial models permit differential smoothing of the fitted surface in two spatial directions, which naturally correspond to polar coordinates: distance to the basket and angle from the line connecting the two baskets. We illustrate our approach using the 2003–2004 shot chart data for Minnesota Timberwolves guard Sam Cassell.

KEY WORDS: Bayesian; Conditionally autoregressive prior; Sports.

1. INTRODUCTION

Statistics and sports have always enjoyed a close connection, as fans use statistics to compare and rank teams and players. More recently, coaches have begun to use statistics for inference as well as description; for example, to decide which pinch hitter to use in a particular baseball situation. The ways in which coaches use information may, however, be hampered by the nature of the data or their preconceived ideas about it.

In basketball, a common and time-honored statistical summary is the *shot chart*. For example, Figure 1 plots the shot chart from the Minnesota Timberwolves versus Milwaukee Bucks game on October 23, 2003. This is a spatial representation of every shot attempted by one player—in this case, Sam Cassell of the Timberwolves. Traditionally, an "O" is coded at the location of every successful shot while an " X" is coded at the location of every unsuccessful shot. Such charts are used in coaching teams as early as middle school, and are becoming even more common now that palm-sized computers can record the location and result of each shot along with covariates, such as game time, players currently on the floor, and so on.

Although shot chart data are rapidly increasing in richness and availability, most coaches still use them purely as descriptive summaries that show the parts of the floor where players prefer to shoot, and how many of these shots actually go in the basket. Such data could be used to make inferences about a player's shooting tendencies and abilities, and thus potentially improve a team's ability to defend opposing players. However, accurate inferences require a statistical model that incorporates important covariates (both time-varying and not) and acknowledges the obvious spatial association in the data. This article's purpose is to analyze shot chart data using hierarchical spatial models and Markov chain Monte Carlo (MCMC) methods, and to reflect on their value to a coach.

Our approach could be illustrated with data from any player with enough shot attempts, but our own spatial location led us to select shot data from the 2003–2004 season of Minnesota Timberwolves guard Sam Cassell. Cassell is a veteran point guard who has played for six NBA teams; he spent the 2003–2004 and 2004–2005 seasons with the Timberwolves and was traded to the Los Angeles Clippers in the summer of 2005. His arrival in Minnesota in 2003, along with guard Latrell Sprewell, coincided with an improvement in the team's fortunes. Along with superstar forward Kevin Garnett, around whom the team had long been built, the Wolves had an excellent regular season and advanced to the NBA Western Conference finals, where they lost in six games to the Los Angeles Lakers. Cassell's season enhanced his reputation as a clutch player who could take over a game in the fourth quarter. For example, one sportswriter wrote, "Cassell is like the planet Venus: the later in the evening it gets, the more

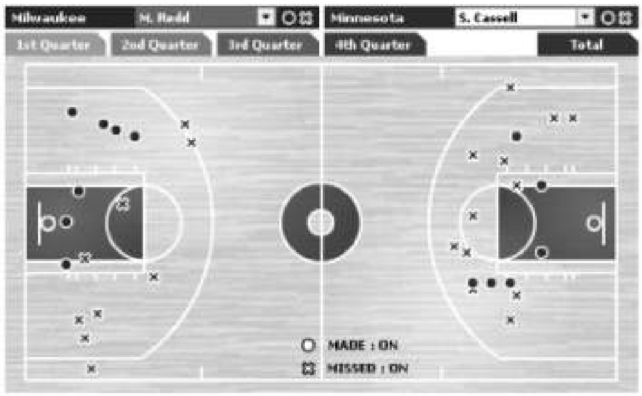


Figure 1. Sam Cassell's shot chart from the October 23, 2003, game against the Milwaukee Bucks (downloaded from espn.com).

Brian J. Reich is Research Associate, Department of Statistics, North Carolina State University, Campus Box 8203, Raleigh, NC 27695 (E-mail: reich@stat.ncsu.edu). James S. Hodges is Senior Research Associate, Bradley P. Carlin is Professor, and Adam M. Reich is a student, Division of Biostatistics, School of Public Health, University of Minnesota, Mayo Mail Code 303, Minneapolis, MN 55455–0392 (E-mail: bjreich@ncsu.edu).

© 2006 American Statistical Association DOI: 10.1198/000313006X90305

visible he becomes to the naked eye," (Powers 2004) while another opined, "The reason for Cassell being better late in games? A simple combination of having the ball on every possession, and then being absolutely, positively, completely unconcerned with whether he falls flat on his face" (Hamilton 2004).

Our data, downloaded from espn.com, are 90% complete: of the 1,270 shots Cassell took during the 2003–2004 season, we have full information for $N = 1{,}139$. Six games' data were unavailable, and a few shots were recorded in game summaries but not on the shot chart. For each shot, we have game clock time elapsed since Cassell's last shot (excluding time on the bench), its location in polar coordinates, that is, feet from the basket and angle from the center line, its result (make or miss), and 10 binary covariates (Table 1). Some of these covariates are dichotomized continuous covariates, for example, BEHIND is derived from the game's current score. Binary covariates improve computational efficiency immensely at a relatively low cost in lost information; we comment further below.

Table 1 anticipates that the absence of two players, Garnett and Sprewell, may affect Cassell's shooting. Garnett, Sprewell, and Cassell scored 64% of Minnesota's points in 2003–2004, a higher percentage than any other trio in the league. No other Timberwolves player scored more than 7% of the team's points, so no other player's presence is likely to affect Cassell's shooting.

Table 1. Description of the Binary Covariates Measured at the Time of Each Shot Attempt

Name	*The covariate equals one when:*	*Frequency*
NOKG	Kevin Garnett is not in the game	0.13
NOLS	Latrell Sprewell is not in the game	0.21
HOME	The game is played in Minnesota	0.55
NOREST	The Timberwolves had less than 2 days off since their last game	0.72
2HALF/OT	Second half or overtime	0.48
BEHIND	The Timberwolves are losing	0.37
BLOCK	The opponent averages more than 4.8 blocks per game	0.48
FGPALL	The opponent allowed a field goal percentage under 44%	0.50
MISSLAST	Cassell missed his previous shot	0.47
TEAMFGA	The Timberwolves took more that 80 shots in the current game	0.51

Our statistical models also consider the NOKG × NOLS interaction term because Cassell's performance may only be affected by both players' absence. Team success is known to be affected by HOME and NOREST, so these covariates are included as predictors of Cassell's shooting. The current game situation, represented by 2HALF/OT, BEHIND, and 2HALF/OT × BEHIND, could influence offensive strategy and thus Cassell's shot selection. MISSLAST is used to test conjectures relating to the so-called "hot hand" theory that past results affect current shot selection and success (Larkey, Smith, and Kadane 1989; Gilovich, Valone, and Tversky 1985; Tversky and Gilovich 1989; Albright 1993). Other predictors of Cassell's shooting are BLOCK and FGPALL, which represent the opponent's defensive prowess, and TEAMFGA which measures the game's tempo.

The rest of the article is as follows. Section 2 considers relatively simple models for the time between Cassell's shots. Although time between shots is not recorded in a shot chart, which is our primary focus, these data help paint a complete picture of Cassell's shooting tendencies. These models do not have spatial or temporal random effects, but offer a quick, simple look at an aspect of the data that does not seem to justify more intensive analysis. Section 3 turns to analyzing factors affecting Cassell's choice of locations for taking shots. Here spatial correlation is important; we use conditionally autoregressive (CAR) models that allow for differential smoothing in the two relevant spatial coordinates, distance from the basket and angle relative to the center line. Section 4 considers shot success, again accounting for covariates and the similarity of outcomes from shots taken at proximate locations. The last two sections use a recently developed model evaluation tool, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, and van der Linde 2002) to choose among competing models. Finally, Section 5 summarizes the findings and suggests avenues for future research in this area.

2. SHOT FREQUENCY

Cassell attempted an average of 15.0 shots per game and the median game time between his shots, excluding game time when he was not in the game, was 1.6 minutes. To assess factors associated with his shot frequency, we used a multiple linear regression with the log of the time between shot attempts as the response, main effects for each covariate in Table 1, interaction effects for $\mathrm{NOKG} \times \mathrm{NOLS}$ and $\mathrm{BEHIND} \times \mathrm{2HALF/OT}$, and homogeneous error variance. We chose diffuse normal priors with mean 0, variance 1,000 for the mean parameters and an $\mathrm{IG}(0.01, 0.01)$ prior for the error variance, where $\mathrm{IG}(\alpha_1, \alpha_2)$ denotes the inverse gamma distribution with density proportional to $(\sigma_e^2)^{-(\alpha_1+1)} \exp\left(-\alpha_2/\sigma_e^2\right)$. Although Daniels and Kass (1999) showed that $\mathrm{IG}(\epsilon, \epsilon)$ priors may not be truly "noninformative" in the sense that changing ϵ may lead to substantive changes in the posterior for σ_e^2, such changes typically do not impact the mean parameters' posteriors which are our primary focus.

The covariates NOKG, NOLS, and BEHIND can change between shot attempts; because their values are recorded only at the time of each shot, we do not have the exact times when these covariates changed values. One possible remedy is an error-in-covariates model, where the known binary (0/1) covariate X_{ki}

Table 2. Posterior Summaries of the Natural Exponentials (e^{b_k}) of the Coefficients from the Regression on Log Time Between Shots

	Posterior median	*Posterior 95% CI*
INT	1.28	(1.07, 1.52)
2HALF/OT	1.22	(1.05, 1.40)
BEHIND	1.13	(0.96, 1.32)
BEHIND × 2HALF/OT	0.88	(0.69, 1.10)
NOREST	1.13	(0.99, 1.28)
NOKG	0.90	(0.71, 1.11)
NOLS	0.90	(0.76, 1.06)
NOKG × NOLS	1.10	(0.77, 1.54)
HOME	1.09	(0.97, 1.22)
TEAMFGA	0.92	(0.81, 1.03)
BLOCK	1.04	(0.93, 1.16)
FGPALL	0.99	(0.87, 1.11)
MISSLAST	1.01	(0.91, 1.13)
NOKG + NOLS + NOKG × NOLS	0.88	(0.69, 1.10)

is replaced by a distribution on the interval (0,1) that depends on the observed value of X_k at the times of the $(i-1)$th (previous) and ith (current) shot. Specifically, the value of the kth covariate for the ith shot is

$$\begin{cases} P_{ki} * X_{ki} + (1 - P_{ki}) * U_{ki} & \text{if} \quad X_{ki} = X_{ki-1} \\ U_{ki} & \text{if} \quad X_{ki} \neq X_{ki-1} \end{cases}, \quad (1)$$

where the X_{ki} is the value of the kth binary covariate at the time of the ith shot, $P_{ki} \sim$ Bern(0.95), and $U_{ki} \sim$ Unif(0,1), independent across k and i. Under this model, X_k is replaced by a random draw representing the fraction of time between shots that the covariate was 1. We assume there is a 95% chance that the covariate was constant between the two shots if a covariate is the same at the times of consecutive shots. If the covariate switched between shots, we let the proportion of the time between shots when the covariate was equal to one follow a Uniform(0,1) distribution. However, this error-in-covariates model gave results similar to a simpler model using each covariate's value at the time of the shot, so we present only the latter.

Table 2 shows the posterior medians and 95% confidence intervals of the exponentials of the regression parameters (e^{b_k}), that is, the fitted multiples of the median time between shots relative to the zero condition. Only the interval for 2HALF/OT does not cover 1.0; the posterior median time between shots is 22% *longer* in the second half and overtime compared to the first half. Of the 1,139 shots in this dataset, 551 (48%) were taken after halftime and only 256 (22%) were taken in the fourth quarter, fewer than any other quarter—surprising considering Cassell's reputation as a confident performer who asserts himself in the fourth quarter, as extolled, for example, by Powers (2004). A possible explanation for this unexpected result is that, as one of the league's best teams in 2003–2004, the Wolves were often holding the ball to protect a lead late in the game. However, BEHIND and BEHIND × 2HALF/OT were not significantly associated with Cassell's shot frequency (Table 2).

Other predictors of time between shots showed mild trends. The posterior median of time between shots was 13% longer when playing with less than two days rest between games, 9% longer when playing at home, and 8% shorter in games that the Timberwolves attempted more than 80 shots. Cassell became more aggressive when either Garnett or Sprewell was on the bench because he was forced to shoulder more of the scoring; the posterior median of time between shots was 10% shorter when either Garnett or Sprewell was on the bench, and 12% shorter when both Sprewell and Garnett were on the bench.

3. SHOT LOCATION

3.1 Model

We now turn to the modeling the association between the covariates in Table 1 and the floor locations from which Sam Cassell took shots. Because the game is focused on a fixed point (the basket), it is natural to refer to locations on the court using not ordinary rectangular (x, y) coordinates, but *polar* coordinates, that is, the distance from the basket in feet, and the angle from the line connecting the two baskets. As shown in Figure 2, we divided the court into an 11×11 grid based on distance and angle. This grid resolution was chosen because the dataset contained 11 distinct angle values and categorizing distance into 11 categories seemed to capture the complexity of the shot chart data while making our analysis computationally feasible. The entire semicircle within two feet of the basket is defined as Region 1. This region is unlike the others in that many of the shots taken from this region are in transition or immediately following a rebound. Therefore, we model Region 1 as disconnected from the rest of the grid, that is, as having no neighbors.

A statistical analysis using spatial methods offers several advantages compared to inspecting plots of raw percentages, such as Figure 3(a)'s plot of Cassell's 2003–2004 shot locations. Figure 3(a) has several cells with no shot attempts, mostly in the corners which most offensive players try to avoid. Spatial smoothing avoids shot probabilities of zero, which are clearly incorrect. The data's spareness is even more troublesome when estimating the effect of covariates on shot location. Our models for shot location include as many as four covariates; given that 48 of the 122 cells have four or fewer shots attempts, borrowing strength from neighboring cells is very important for obtaining stable estimates of covariate effects. Finally, statistical modeling provides a means for measuring uncertainty and testing the significance of each covariate's association with shot location.

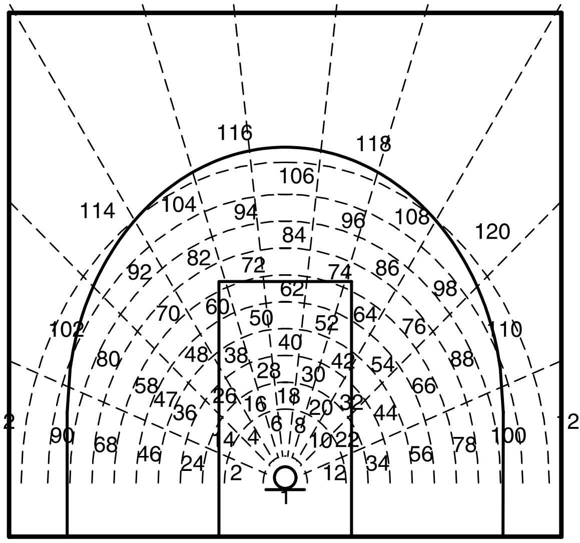


Figure 2. The spatial grid.

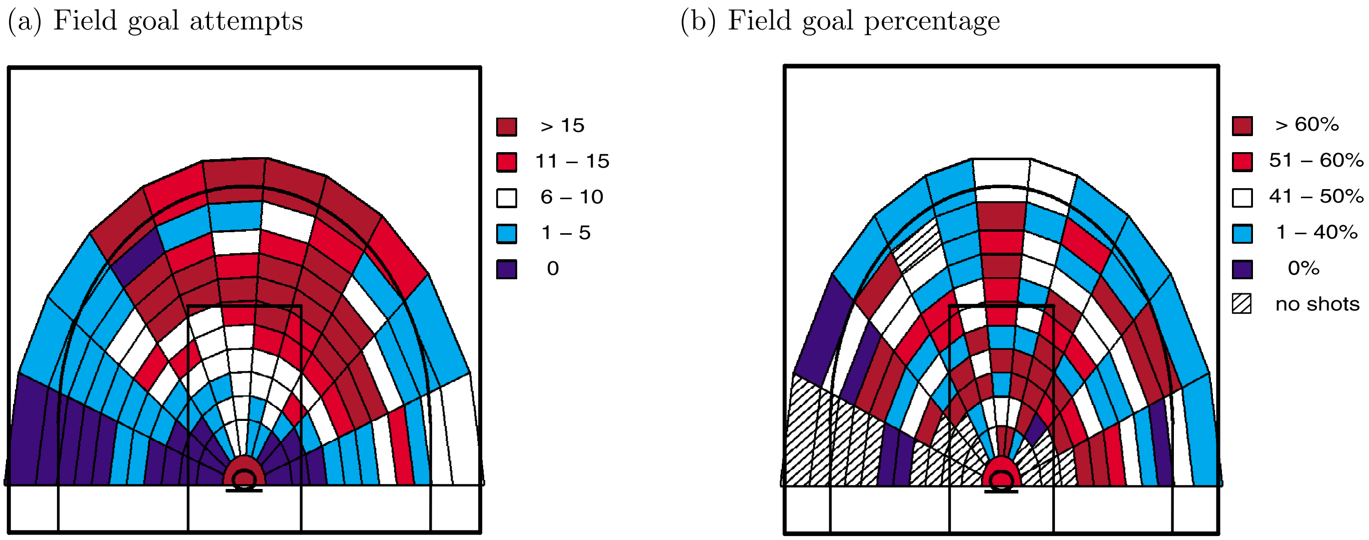


Figure 3. Observed field goal attempts and percentage by region.

Let $y_i \in \{1, \ldots, p\}$ be the region of the ith shot, and let $y_i|\boldsymbol{\theta}(\boldsymbol{\eta}_i)$ follow a multinomial$(\theta_1(\boldsymbol{\eta}_i), \ldots, \theta_p(\boldsymbol{\eta}_i))$ distribution where $\boldsymbol{\theta}(\boldsymbol{\eta}_i) = (\theta_1(\boldsymbol{\eta}_i), \ldots, \boldsymbol{\theta}_p(\boldsymbol{\eta}_i))$ and $\theta_j(\boldsymbol{\eta}_i)$ is the probability that the ith shot was taken from the jth region. The cell probabilities depend on the ith shot's p-vector of linear predictors $\boldsymbol{\eta}_i = \log A + \mathbf{x}_i\mathbf{b}$. Log$A$ is the p-vector of offset terms with $\log A_j$ equal to the log of the area of the jth region in x-y coordinates (Figure 2); these are necessary because the regions in Figure 2 have different areas. Associated with the ith shot is a q-vector of covariates, $\mathbf{x}_i$, and each of the p regions has its own set of q regression coefficients, $\mathbf{b}_{\circ j} = (b_{1j}, \ldots, b_{qj})'$ for Region j, which are related to the $\theta_j(\boldsymbol{\eta}_i)$ through the formula

$$\theta_j(\boldsymbol{\eta}_i) = \frac{\exp(\eta_{ij})}{\sum_{l=1}^{p} \exp(\eta_{il})} = \frac{\exp(\log \mathrm{A}_j + \mathbf{x}_i'\mathbf{b}_{\circ j})}{\sum_{l=1}^{p} \exp(\log \mathrm{A}_l + \mathbf{x}_i'\mathbf{b}_{\circ l})}. \tag{2}$$

This multinomial logit model (McFadden 1974; Cooper and Nakanishi 1988) is a common model for multinomial data and reduces to ordinary logistic regression if there are just two regions. Let $\mathbf{b}_{k\circ} = (b_{k1}, \ldots b_{kp})$ be the p-vector of regression parameters associated with the kth covariate, $k = 1, \ldots, q$. Since adding a constant to $\mathbf{b}_{k\circ}$ does not change (2), the constraint $\mathbf{b}_{\circ 1} \equiv \mathbf{0}$ is needed to ensure identification. Under this parameterization, b_{kj} is positive if the presence of the kth covariate increases the probability of a shot coming from region j relative to Region 1 (directly under the basket).

The multinomial logit model with flat prior for $\mathbf{b}$ can be motivated by an assumed underlying heterogeneous Poisson point process. Conditioning on the total number of shot attempts yields a multinomial model and induces a negative posterior correlation between the shot frequencies of each pair of cells. By contrast, a spatially referenced prior encourages a positive association between neighboring cells' shot intensities. Besag (1974) introduced the conditionally autoregressive (CAR) distribution, which is often used as a prior in spatial modeling of areal data. The CAR distribution can be defined by its conditional distributions: if $\mathbf{b}_{k\circ}$, the spatially varying coefficient of covariate k, is given a CAR(τ_k) prior, $b_{kj}|b_{kl,l\neq j}$ is normal with mean $\bar{b}_{kj}$ and precision (inverse variance) $\tau_k m_j$, where $\bar{b}_{kj}$ is the mean of $\mathbf{b}_{k\circ}$ at the m_j neighbors of region j, and $\tau_k > 0$ controls the degree of smoothing of each b_{kj} towards its neighbors.

Figure 2 shows two distinct types of neighbors: *angle neighbors*, adjacent regions that are the same distance from the basket (e.g., Region 47 is an angle neighbor of Regions 46 and 48), and *distance neighbors*, adjacent regions that are the same angle from the middle of the court (e.g., Region 47 is a distance neighbor of Regions 36 and 58). We expect $\mathbf{b}_{kj}$ to be similar to the $\mathbf{b}_{k\circ}$ at regions neighboring region j, but the amount of smoothing of the two types of neighbors may be different if, say, the covariate only influences the distance of a shot, but not the angle. This *two neighbor relation CAR* (2NRCAR) distribution was mentioned by Besag and Higdon (1999) and developed by Reich, Hodges, and Carlin (2004). It extends the CAR distribution to allow a different amount of smoothing of distance and angle neighbors by adding a parameter $\beta_k \in (0, 1)$ which controls the relative influence of the two neighbor types. Under a 2NRCAR(τ_k, β_k) prior, the distribution of $b_{kj}|b_{kl,l\neq j}$ is normal with mean

$$E(b_{kj}|b_{kl,l\neq j}) = \bar{b}_{akj}\frac{m_{aj}\beta_k}{m_{aj}\beta_k + m_{dj}(1-\beta_k)} + \bar{b}_{dkj}\frac{m_{dj}(1-\beta_k)}{m_{aj}\beta_k + m_{dj}(1-\beta_k)}$$

and precision $\tau_k\,(m_{aj}\beta_k + m_{dj}(1-\beta_k))$, where $\bar{b}_{akj}$ and $\bar{b}_{dkj}$ are the mean of $\mathbf{b}_{k\circ}$ at region j's m_{aj} angle neighbors and m_{dj} distance neighbors, respectively. The conditional prior mean of $b_{kj}|b_{kl,l\neq j}$ is a weighted average of $\bar{b}_{akj}$ and $\bar{b}_{dkj}$, with β_k determining the weight of each neighbor type and τ_k controlling the overall smoothness of $\mathbf{b}_{k\circ}$. This prior simplifies to the usual CAR prior if $\beta_k = 0.5$; $\beta_k > 0.5$ indicates stronger smoothing of angle neighbors, while $\beta_k < 0.5$ indicates stronger smoothing of distance neighbors.

Each β_k is given an independent Uniform(0,1) prior and the τ_k are given independent Gamma(0.01, 0.01) priors, where Gamma(α_1, α_2) denotes the gamma distribution with density proportional to $\tau_k^{\alpha_1 - 1}\exp(-\alpha_2\tau_k)$. For each fit in this article,

Table 3. Summary of Covariate Selection for the Shot Location Model. Each covariate listed in the left column corresponds to a vector of p = 122 parameters, Region 1 under the basket and 121 other regions with a 2NRCAR prior.

Covariates	*Dimension of* **b**	*DIC*	p_D
INT	122	9773.4	74.6
INT, 2HALF/OT	244	9762.0	91.2
INT, HOME	244	9766.8	86.8
INT, NOREST	244	9768.3	85.6
INT, BLOCK	244	9772.2	83.0
INT, TEAMFGA	244	9774.1	87.0
INT, FGPALL	244	9774.3	92.4
INT, NOLS	244	9777.2	83.2
INT, MISSLAST	244	9777.3	80.0
INT, BEHIND	244	9777.8	85.1
INT, BEHIND × 2HALF/OT	244	9778.0	80.3
INT, NOKG	244	9778.1	80.4
INT, NOKG × NOLS	244	9778.9	80.0
INT, HOME, 2HALF/OT, NOREST, BLK	610	9744.8	124.4

30,000 MCMC samples from the full posterior were drawn with the first 5,000 discarded as burn-in. Metropolis-Hastings updates were used for the b_{kj} and the β_k with normal and beta candidate distributions, respectively. The candidate distributions were tuned during the burn-in period to give Metropolis-Hastings acceptance rates between 0.35 and 0.50. The τ_k follow gamma full conditional distributions and were updated using Gibbs sampling. Convergence was monitored by comparing draws of the τ_k, β_k, and several b_{kj} from three parallel chains. All computing was done in R (Ihaka and Gentleman 1996).

To compare the models according to their fit to the data, we use the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002). Deviance is defined as $D(\mathbf{b}) = -2\log(f(y|\mathbf{b})) = -2\sum_{i=1}^{N}\sum_{j=1}^{p} I(y_i = j)\log(\theta_j(\boldsymbol{\eta}_i)) = -2\sum_{i=1}^{N}\log(\theta_{y_i}(\boldsymbol{\eta}_i))$, where $I(y_i = j) = 1$ if $y_i = j$, and 0 otherwise. The DIC statistic is DIC $= \overline{D} + p_D$, where $\overline{D} = E(D(\mathbf{b})|y)$ and $p_D = \overline{D} - D(E(\mathbf{b}|y))$, the expectation being taken with respect to the joint posterior. Spiegelhalter et al. (2002) showed that $\overline{D}$ measures the model's fit to the data, while p_D measures its complexity (its "effective number of parameters"). DIC is thus a penalized likelihood criterion, a hierarchical-model version of the Akaike information criterion (AIC). Models with smaller DIC are preferred.

As a final note, the use of dichotomized covariates (Table 1) is computationally important. For example, the use of BEHIND allows the likelihood to be written as the product of just two multinomial densities: one for shots taken when in the lead, and one for shots taken when behind. Using the actual difference would lead to a likelihood that is the product of 1139 multinomial densities, one for each shot, which is currently not practical with MCMC methods.

3.2 Results

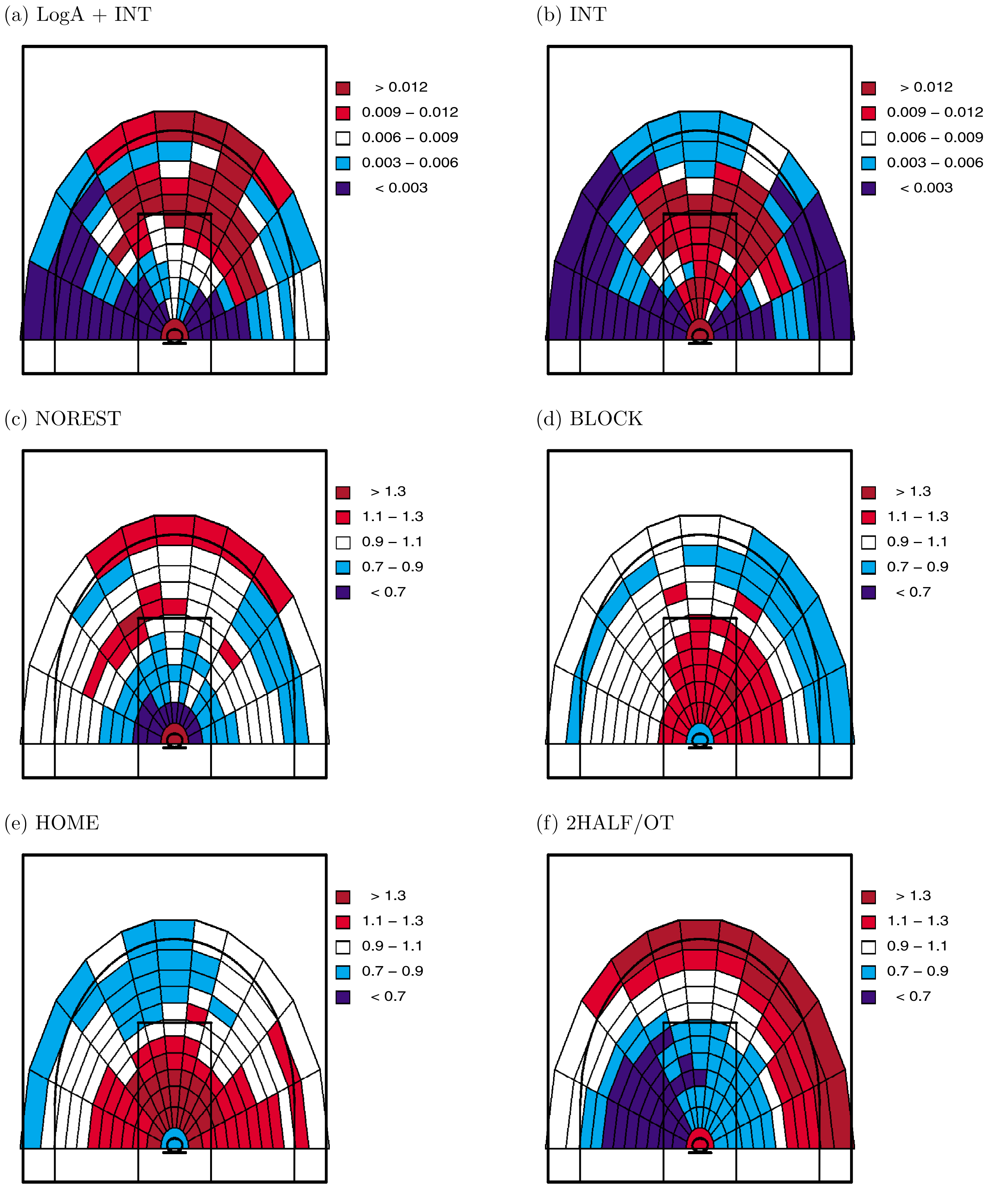
Table 3 contains DIC and p_D for each model listed. Comparing row 1, for the intercept-only model, to rows 2–13, for the models with the intercept and one covariate, shows that four covariates (2HALF/OT, HOME, NOREST, and BLOCK) improved DIC compared to the intercept-only model. The full model—with a p-vector of parameters associated with the intercept and each of the 12 covariates in Table 3 (1586 b_{kj}, 13 τ_k, and 13 β_k)—is very large, so only the four covariates that improved DIC compared to the intercept-only model were included in the "final model" (Table 3's bottom row), which has the smallest DIC in Table 3. The final model still has 620 parameters (610 b_{kj}, 5 τ_k, and 5 β_k) but because $\mathbf{b}$ is smoothed by the 2NRCAR prior, the effective model size is $p_D = 124.4$.

The covariates FGPALL, NOKG, NOLS, MISSLAST, and BEHIND were not included in the final analysis because their inclusion did not improve the DIC compared to the intercept-only model. As with shot frequency (Section 2), removing Garnett or Sprewell from the lineup does not significantly influence Cassell's shot location. It is somewhat surprising that his shot selection is not affected by changes in the coach's play-calling when Garnett or Sprewell is on the bench. It is also surprising that neither the duration between shots (Section 2) nor shot location is associated with the score. Conventional wisdom holds that when a team is behind, its players often try to shoot more quickly and settle for longer shots, although this tactic may be less relevant in the NBA because the 24-second shot clock prohibits long possessions at any time during the game.

Many authors have investigated the so-called "hot hand" theory that the probability of making a shot is related to the result of recent shot attempts. Though some disagree on this point, most studies have found no relationship. However, enough athletes believe in the hot hand that it could influence their shot selection. For example, a player who has made a few consecutive shots may believe he is "hot" and thus be more likely to attempt shots he would normally avoid. However, our analysis found little relationship between MISSLAST and the location of Cassell's current shot. Also, Section 2 shows that MISSLAST was not a significant predictor of time between shots, and Section 4 shows that MISSLAST was not a significant predictor of shooting percentage. Of course, this result applies only to Sam Cassell, a confident NBA veteran, and may not hold more generally.

Table 4. Summary of the Final Shot Location Model

	Posterior median and 95% CI of τ_k	*Posterior median and 95% CI of* β_k	*Number of the covariate's 121 unconstrained* b_{kj} *parameters for which the 95% CI does not cover 0*
INT	1.88 (1.17, 3.26)	0.19 (0.06, 0.45)	121
BLOCK	33.69 (8.98, 139.73)	0.61 (0.14, 0.93)	26
2HALF/OT	17.80 (6.35, 88.45)	0.61 (0.15, 0.96)	5
HOME	17.22 (6.05, 67.11)	0.48 (0.11, 0.90)	1
NOREST	23.18 (5.79, 180.27)	0.63 (0.04, 0.97)	0



(a) LogA + INT

(b) INT

(c) NOREST

(d) BLOCK

(e) HOME

(f) 2HALF/OT

Figure 4. Summary of the final shot location model. Panels (a) and (b) plot estimated baseline shot intensity (a) without adjusting for cell area, that is, $\theta\,(logA + b_{INT\cdot})$'s posterior mean, and (b) adjusting for cell area, that is, $\theta\,(b_{INT\cdot})$'s posterior mean. Panels (c)–(f) show the estimated relative risks of shot attempts due to each covariate, that is, $[\theta\,(logA + b_{INT\cdot} + b_{k\cdot})\,/\,\theta\,(logA + b_{INT\cdot})]$'s posterior mean, $k \in \{NOREST, BLOCK, HOME, 2HALF/OT\}$.

The final model has slightly smaller DIC (9744.8) than the model with the same covariates but with $\beta_k = 0.5, k = 1, \ldots, q$ (DIC = 9749.1). Table 4 summarizes the posteriors of the smoothing parameters $\{\tau_k, \beta_k\}$ in the final model. The posterior median of the smoothing precision τ_k is smallest for the intercept, indicating more spatial variation in the intercept than in any of the coefficients. For the intercept, β_k has posterior median 0.19 and its 95% confidence interval does not cover 0.5, that is, distance neighbors appear to be significantly more similar than angle neighbors. For BLOCK, 2HALF/OT, and NOREST, the posterior median of β_k is greater than 0.5 (stronger smoothing of angle neighbors), suggesting these covariates have more influence on the distance of the shot than the angle. However, the data provide little information about β_k, whose posterior confidence intervals cover 0.50 and indeed most of the unit interval.

Figure 4 illustrates the effect of each covariate's effect on shot location. As the intercept's small τ_k indicated, there is little smoothing of the intercept parameters. The plot of $\boldsymbol{\theta}(\log A + b_{\mathrm{INT}\cdot})$'s posterior mean (Figure 4(a)), that is, the estimated baseline shot intensity, resembles the plot of the observed number of shot attempts (Figure 3(a)). Figure 4(b) plots the estimated baseline shot intensities adjusted for the different area of the grid squares; predictably, this shifts mass toward the basket where the grid squares are smaller. For each region other than Region 1, the posterior 95% confidence interval of the $b_{\mathrm{INT}j}$ does not cover zero and the posterior mean is negative. That is, Cassell's favorite region is right at the basket. His next favorite type of shot is the mid-range shot from 12–20 feet. From all distances, he prefers shots from the center of the court and has a preference for shooting from left of the center line. Cassell's preference for shooting from the left side of the court does not necessarily indicate he prefers dribbling with his left hand; investigating dribbling tendencies requires data that are not included in shot charts, for example, the number of dribbles taken with each hand.

Although NOREST was included in the final model, the 95% confidence intervals for its b_{jk} cover zero at all 122 regions (Table 4). When playing on less than two days rest, Cassell takes more lay-ups, mid-range shots from the right, and three-point shots from the top of the key (Figure 4(c)). Also, he takes fewer shots in the lane from 2–10 feet.

Table 5. Summary of Covariate Selection for the Shot Success Model. Each covariate listed in the left column corresponds to a vector of p = 122 parameters, Region 1 under the basket and 121 other regions with a 2NRCAR prior.

Covariates	*Dimension of* **b**	*DIC*	p_D
INT	122	1574.3	9.8
INT, NOKG	244	1565.7	36.2
INT, NOKG × NOLS	244	1572.0	24.8
INT, TEAMFGA	244	1573.3	18.1
INT, BEHIND	244	1574.9	19.6
INT, HOME	244	1575.0	21.4
INT, NOREST	244	1576.9	15.9
INT, FGPALL	244	1577.1	15.1
INT, NOLS	244	1578.0	16.1
INT, MISSLAST	244	1578.1	15.1
INT, BEHIND × 2HALF/OT	244	1579.5	16.9
INT, BLOCK	244	1581.5	19.6
INT, 2HALF/OT	244	1581.8	20.2
INT, NOKG, NOKG × NOLS, TEAMFGA	488	1564.4	50.7

For BLOCK, 2HALF/OT, and HOME, some regions had 95% confidence intervals that did not cover zero (Table 4). The presence of a good shot blocker affects the way Cassell finishes his drives to the basket; against opponents that block a lot of shots, Cassell takes fewer lay-ups and is forced to take more shots from 2–15 feet (Figure 4(d)). Thus, teams with a quality shot-blocker might do well to defend Cassell slightly differently—say, by retreating only to the 3–15 foot area (instead of all the way to the basket) when they are beaten on a drive.

It is well known that teams tend to win more often at home than on the road. This was true for every team in the NBA in 2003–2004. For Cassell, the home-court advantage does not seem to influence the game time between his shots (Section 2) or his field goal percentage given the shot's location (as will be shown in Section 4), but it *does* appear to affect his shot location: Figure 4(e) shows an increase in frequency of short shots from the lane at home. The plot of observed field goal percentage by region in Figure 3(b) shows that he is generally successful from this area, so this tendency is probably good for the Timberwolves.

In the second half, Cassell's affinity for long shots becomes stronger (Figure 4(f)), especially from the left side of the court. His percentage of shots from more than 20 feet increases from 16% (94/588) in the first half to 26% (144/557) in the second half. Although second-half strategy is affected by the score, the percentage of Cassell's second-half shots taken from greater than 20 feet was similar whether the Wolves were ahead or behind (98/370 = 26% when ahead, 46/181 = 25% when behind). Combined with Section 2's finding that the time between Cassell's shot attempts increases in the second half, this suggests that in the second half Cassell may have deferred to the league's most valuable player in 2003–2004, Kevin Garnett, creating fewer shots himself and taking more jump shots when the defense collapsed on Garnett.

4. SHOT SUCCESS

4.1 Model

In this section, we assume shot location is fixed and analyze the probability that Cassell makes a shot from each spot on the floor. To correctly model the covariates' effect on field goal percentage, their effect should be allowed to vary spatially. For example, when the Wolves are trailing late in the game, a defense may emphasize preventing three-point shots, and thus Cassell's field goal percentage may decrease on the perimeter and increase near the basket.

Accordingly, we model shooting percentage with separate logistic regressions at each of the $p = 122$ regions in Figure 2 and spatially smooth the regression parameters for each covariate. Let $z_i = 1$ if Cassell made the ith shot and 0 otherwise, and let z_i have a Bernoulli(π_i) distribution where $\mathrm{logit}(\pi_i) = \log(\pi_i/(1-\pi_i)) = \mathbf{x}_i \mathbf{b}_{\circ y_i}$. Here, y_i is the region of the ith shot, and $\mathbf{b}_{\circ y_i}$ is the q-vector of regression parameters associated with region y_i. In this model, the constraint $\mathbf{b}_{\circ 1} = \mathbf{0}$ is no longer necessary. As in Section 3, each of the $\mathbf{b}_{k\circ}$, $k = 1, \ldots, q$, has a 2NRCAR(τ_k, β_k) prior which differentiates between smoothing of angle and distance neighbors. It

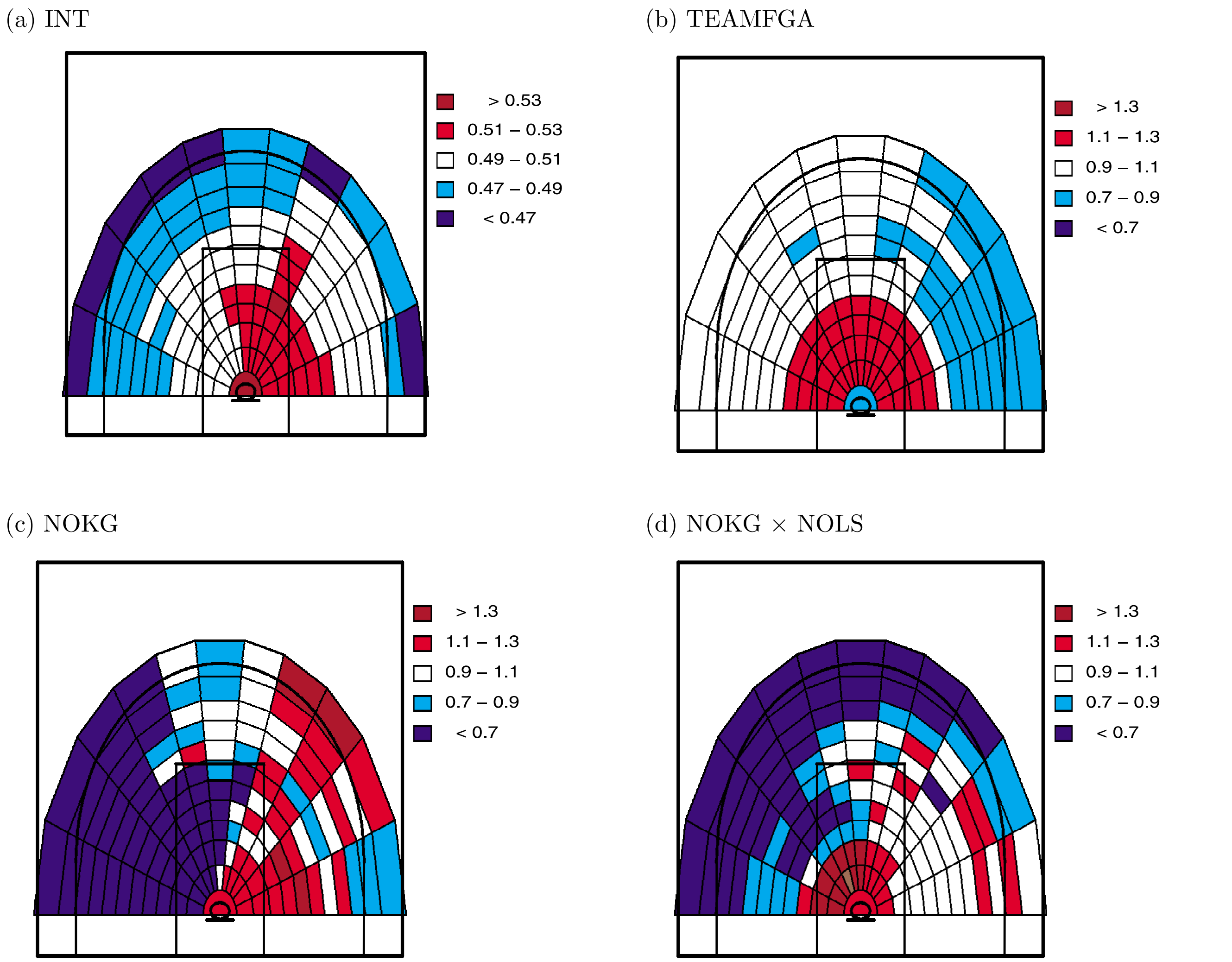


Figure 5. Summary of the final shot percentage model. Panel (a) plots the posterior mean baseline shooting percentage, that is, logit$^{-1}(b_{INT\,j})$. Panels (b) and (c) plot the posterior mean relative risk of making a shot due to TEAMFGA and NOKG, respectively, that is, logit$^{-1}(b_{INT\,j} + b_{kj})$ /logit$^{-1}(b_{INT\,j})$, $k \in \{TEAMFGA; NOKG\}$. Panel (d) plots the posterior mean relative risk of making a shot when both Garnett and Sprewell are on the bench compared to when only Garnett is on the bench, that is, logit$^{-1}(b_{INT\,j} + b_{NOKGj} + b_{NOKG\times NOLSj})$/logit$^{-1}(b_{INT\,j} + b_{NOKGj})$.

seems reasonable that shot success will be affected more by distance than angle: changing the angle of a shot has little effect on shooting technique, while longer shots are generally more difficult than short shots. The 2NRCAR prior allows for this fit by letting the intercept's β_k be greater than 0.50.

4.2 Results

Table 5 shows that the additions of NOKG, NOKG×NOLS, and TEAMFGA improved DIC compared to the intercept-only model. The final model with these three covariates has the smallest DIC in Table 5. The DIC of the final 2NRCAR model, 1564.4, is similar to or smaller than other, simpler models with the same three covariates, including the standard CAR model with $\beta_k = 0.5$ for $k \in \{\mathrm{INT}, \mathrm{NOKG}, \mathrm{NOKG} \times \mathrm{NOLS}, \mathrm{TEAMFGA}\}$ (DIC = 1564.2), the model ignoring angle with $b_{k\cdot}$ constant within rows of Figure 2, $k \in \{\mathrm{INT}, \mathrm{NOKG}, \mathrm{NOKG} \times \mathrm{NOLS}, \mathrm{TEAMFGA}\}$ (DIC = 1571.1), and the model ignoring shot location altogether with $b_{k1} = \cdots = b_{kp}$, $k \in \{\mathrm{INT}, \mathrm{NOKG}, \mathrm{NOKG} \times \mathrm{NOLS}, \mathrm{TEAMFGA}\}$ (DIC = 1579.7).

Figure 5(a) plots each region's posterior mean baseline shot-success percentage, that is, logit$^{-1}(b_{\mathrm{INT}j})$. In contrast with the shot location model of Section 3, most of the region-to-region variation in the observed field goal percentage (Figure 3(b)) is smoothed away (Figure 5(a)). The posterior of $\mathbf{b}$ is quite smooth; the final model's p_D is 50.7, far less than the dimension of $\mathbf{b}$ (488) and the p_D of any of the shot location models of Section 3. That is, there is less spatial variation in shot success than in shot intensity.

Figure 5(a) shows that Cassell's estimated baseline shot success percentage varies with both distance *and* angle. Cassell's observed field goal percentage decreases with distance (65/109 = 59% for lay-ups and dunks, 159/305 = 52% from 0–14 feet, 233/490 = 48% from 15–20 feet, and 99/237 = 42% from over 20 feet). Figure 5(a) also shows that his shooting percentage is generally higher on the left side of the court, especially from mid- and short-range. Since β_{INT}'s posterior 95% interval, (0.08,

Table 6. Summary of the Final Shot Success Model

	Posterior median and 95% CI of τ_k	Posterior median and 95% CI of β_k	Number of the covariate's 122 b_{kj} parameters for which the 95% CI does not cover 0
INT	51.0 (10.9, 272.9)	0.51 (0.08, 0.93)	1
NOKG × NOLS	0.9 (0.1, 30.0)	0.62 (0.05, 0.98)	8
NOKG	0.7 (0.1, 58.9)	0.42 (0.01, 0.91)	3
TEAMFGA	15.7 (4.1, 80.4)	0.61 (0.05, 0.98)	2

0.93), covers 0.50 (Table 6), we cannot conclude that Cassell's shooting percentage changes more with distance than with angle.

Each covariate in the final model has some b_{kj} whose posterior 95% interval does not cover zero (Table 6). Games in which the Timberwolves attempted more than 80 shots are likely to be fast-paced games with more shot attempts coming off fast breaks. In these games, Cassell's shooting percentage for lay-ups and from the left corner decreased and his shooting percentage from short range increased (Figure 5(b)).

Cassell's field goal percentage generally decreased when Garnett left the game (Figure 5(c)). Combining shots from all locations, Cassell shot 49.7% (492/989) when Garnett was on the floor and 42.1% (64/152) when Garnett was out of the game. However, Cassell's shooting percentage actually increased for some regions on the left side of the court. The cause of this discrepancy may be that Garnett prefers to play on the left side of the court; when he is out of the game, the defense may now focus less attention on that area, allowing Cassell to shoot a higher percentage from this side.

Figure 5(d) plots the relative risk of making a shot when both Garnett and Sprewell are on the bench compared to making a shot when only Garnett is on the bench, that is, the effect of removing Sprewell when Garnett is already out of the game. When Sprewell leaves the game, Cassell's field goal percentage increases near the basket and decreases from the perimeter, especially from the right side of the court. This suggests that both Garnett and Sprewell create easy perimeter shots for Cassell from the right side of the court.

5. CONCLUSIONS

This article developed a new statistical model for analyzing basketball shot charts, accounting for important covariates and spatial correlation in the data and raising shot charts from a purely descriptive tool to a fully inferential one. Our results allow coaches to replace conventional wisdom, anecdotal evidence, and casual sports talk with empirical conclusions. As in many other fields, our hierarchical model provides discipline to counteract the natural human tendency to find patterns in virtually any data, even when patterns are absent. Our models capture the spatial element in shooting, and the 2NRCAR allows differential smoothing for shots taken from similar distances as opposed to similar angles.

For Sam Cassell's 2003–2004 data, we found the 2NRCAR superior to the standard CAR for shot location data, but its complexity was not justified for shot success. We also discovered useful predictors of both location and field goal percentage. The absence of Garnett or Sprewell did not significantly change Cassell's shot selection (frequency or location), but he did generally shoot a lower percentage when they were out of the game, perhaps because more of his shots were contested when the defense did not have to concentrated on Garnett and Sprewell. This supports the conventional wisdom that good players "make each other better." Against good shot-blocking teams, Cassell took fewer lay-ups and more shots from 2–15 feet. Also, in the second half he took fewer shots overall and a higher percentage of jumpshots. However, other covariates (BEHIND, FGPALL, and MISSLAST) were not important. Finally, Cassell shot somewhat better from the left side of the court, and especially so when Garnett was in the game.

Devising a plan for defending Cassell using these results is complicated by the fact that the opposing coaches were already making defensive adjustments during the games from which we have data. However, teams frequently try out new defensive strategies and our analysis provides some ideas that could be tested in a game. For example, our results show that while Cassell's shot frequency and shot success percentage are not affected by the presence of a shot blocker, Cassell takes fewer layups and more shots from 3–15 feet when playing against a good shot-blocking team. It is not obvious all players respond this way to a shot blocker: some players may take fewer shots overall, others may take more three-point shots, and others may take the same shots but make a lower percentage near the basket. This information could help a coach position his interior defenders further away from the basket against Cassell than other perimeter players.

Our analysis is primarily intended as illustrative, and more elaborate (and possibly more useful) analyses are easy to envision. For example, shot frequency, location, and success could be analyzed simultaneously. This could improve the analysis of, say, a post player's shot selection because shots taken immediately after the previous shot are more likely to be short shots resulting from offensive rebounds. Also, adding temporal correlation (in addition to spatial correlation) to our models should be possible using additional game-specific random effects. A reanalysis of the shot frequency data incorporating day-specific random effects, distributed either independently or AR(1), did not improve DIC. Still, accounting for temporal correlation may be important when modeling shot location or success, for example, to capture the effect of nagging minor injuries. Our areal data methods also require aggregation of the shot chart data to the 122 regions in Figure 1; using a geostatistical or point process model on the exact spatial locations may provide better results. In this vein, Hickson and Waller (2003) used a Poisson point process model to look for differences in the spatial distribution of Michael Jordan's made and missed shots during the 2001–2002 season. Finally, a coach facing the Timberwolves would need to know how to defend not only Sam Cassell, but also Garnett and Sprewell *at the same time*. This suggests a *multivariate* version of our analysis, perhaps using a multivariate conditionally autoregressive (MCAR) approach (Gelfand and Vounatsou 2003; Jin, Carlin, and Banerjee in press). Standard (1NR) versions of this model are available in the WinBUGS language (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml), while 2NR versions await development.

[Received February 2005. Revised October 2005.]

REFERENCES

Albright, S.C. (1993), "A Statistical Analysis of Hitting Streaks in Baseball," *Journal of the American Statistical Association*, 88, 1175–1183.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society*, Series B, 36, 192–236.

Besag, J., and Higdon, D. (1999), "Bayesian Analysis of Agricultural Field Experiments" (with discussion), *Journal of the Royal Statistical Society*, Series B, 61, 691–746.

Cooper, L. G., and Nakanishi, M. (1988), *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness*, Boston: Kluwer.

Daniels, M. J., and Kass, R. E. (1999), "Nonconjugate Bayesian Estimation of Covariance Matrices and its use in Hierarchical Models," *Journal of the American Statistical Association*, 94, 1254–1263.

Gelfand, A. E., and Vounatsou, P. (2003), "Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis," *Biostatistics*, 4, 11–25.

Gilovich, T., Vallone, R., and Tversky, A. (1985), "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology*, 17, 295–314.

Hamilton, B. (2004), "Just Call Cassell Mr. Clutch," *Saint Paul Pioneer Press*, May 9, 2004.

Hickson, D. A., and Waller, L. A. (2003), "Spatial Analyses of Basketball Shot Charts: An Application to Michael Jordan's 2001–2002 NBA Season," Technical Report, Department of Biostatistics, Emory University.

Ihaka, R., and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Jin, X., Carlin, B. P., and Banerjee, S. (in press), "Generalized Hierarchical Multivariate CAR Models for Areal Data," *Biometrics*.

Larkey, P. D., Smith, R. A., and Kadane, J. B. (1989), "It's Okay to Believe in the 'Hot Hand,' " *Chance*, 2, 22–30.

McFadden, D. (1974), *Conditional Logit Analysis of Qualitative Choice Behavior. Frontiers of Econometrics*, ed. P. Zarembka, New York: Academic Press.

Powers, T. (2004), "Geezers Gamble Pays Off Big Time," *Saint Paul Pioneer Press*, Feb 4, 2004.

Reich, B. J., Hodges, J. S., and Carlin, B. P. (2004), "Spatial Analysis of Periodontal Data using Conditionally Autoregressive Priors Having Two Types of Neighbor Relations," Research Report 2004–2004, Division of Biostatistics, University of Minnesota.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society*, Series B, 64, 583–639.

Tversky, A., and Gilovich, T. (1989), "The 'Hot Hand': Statistical Reality or Cognitive Illusion?" *Chance*, 2, 31–34.