DOI 10.1515/jqas-2012-0023 — JQAS 2013; 9(4): 337–345

Garritt L. Page*, Bradley J. Barney and Aaron T. McGuire

Effect of position, usage rate, and per game minutes played on NBA player production curves

Abstract: In this paper, we model a basketball player's on-court production as a function of the percentiles corresponding to the number of games played. A player's production curve is flexibly estimated using Gaussian process regression. The hierarchical structure of the model allows us to borrow strength across players who play the same position and have similar usage rates and play a similar number of minutes per game. From the results of the modeling, we discuss questions regarding the relative deterioration of production for each of the different player positions. Learning how minutes played and usage rate affect a player's career production curve should prove to be useful to NBA decision makers.

Keywords: Bayesian hierarchical model; Career trajectory curves; Gaussian process regression.

***Corresponding author: Garritt L. Page,** Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile, e-mail: page@mat.puc.cl
Bradley J. Barney: Kennesaw State University, Department of Mathematics and Statistics, Kennesaw, GA, USA
Aaron T. McGuire: Capital One, VA, USA

1 Introduction

The National Basketball Association (NBA) is a professional basketball league that is generally considered to employ the most athletic and skilled basketball players in the world. The NBA's global popularity has grown rapidly, and this coupled with the uniqueness of an NBA player's athletic skill has resulted in increasingly large player salaries. Ideally, high producing players would receive higher wages relative to lower producing players. However, being able to determine a player's future production is not easy and there are numerous examples of players whose production and compensation are misaligned. Because uncertainty surrounds future performance, learning how different factors influence a player's future production is of interest. In this paper we focus on three factors: position played, minutes per game, and usage rate. That is, a player's career performance will be modeled as a function of time and will be grouped using the three factors stated. We will refer to these functionals as "production curves."

Two factors that clearly influence game-by-game production are the average number of minutes played and usage rate. The usage rate variable is a modern basketball metric created by John Hollinger that attempts to measure a player's possession use per 40 min (Hollinger 2005; more details are provided in Section 2). In addition to minutes played and usage rate, it seems completely reasonable that career production curves would depend on the position being played as each position in the game of basketball requires a unique skill set (Page, Fellingham, and Reese 2007) and different skills tend to deteriorate with age at different rates.

There has been somewhat of a statistical revolution in today's NBA decision making (Mahoney 2010) and a few studies have focused on career production curves. Ozmen (2012) studied career trends of basketball players in the Turkish Basketball League. In other sports, Berry, Reese, and Larkey (1999) used a Bayesian hierarchical model to compare career production of baseball players, hockey players and golfers regardless of era played. Albert (2009) compared the careers of pitchers in major league baseball. A connected avenue of research is that of investigating how aging impacts performance. For example, Shoals (2010) and Pelton (2010) estimated the age that corresponds to an NBA player's peak performance.

In this paper we employ a hierarchical Bayes model to estimate not only individual player production curves but also the process that generates the individual curves according to position, usage rate, and minutes played per game. The observation and process levels of the hierarchy are modeled using Gaussian process (GP) regressions (Rasmussen and Williams 2006 for GPs and Behseta, Kass, and Wallstrom 2005 for hierarchical GP modeling). Therefore, individual player curves as well as the position by usage rate by minutes per game (hereafter $p\times u\times m$) mean curves are modeled very flexibly. Using a hierarchical model allows us to borrow strength across individuals that belong to the same $p\times u\times m$ category when estimating process level production curves. Unlike other studies about NBA aging, we are not necessarily interested in

estimating a player's "prime" (though this is readily available from this methodology); rather principal interest lies in estimating the entire career production curve.

The remainder of this article is organized as follows. Section 2 contains a description of the data and Section 3 describes the Gaussian process hierarchical regression model employed. Section 4 outlines the results of the study and Section 5 provides a brief discussion.

2 Description of data

Game-by-game summaries (including playoffs) for each retired player drafted in the 1992–1997 NBA drafts and who participated in at least 100 games were collected (which provided a pool of 178 players). The 100 game minimum was instituted to help us enact our binning method which was implemented for computation and interpretation purposes. (This also reduces the noise introduced by the careers of players that contain little information regarding the processes of interest.) The game-by-game summaries collected include typical box-score metrics in addition to a few "modern" basketball metrics. Among these we selected John Hollinger's so called Game Score (Hollinger 2003) as a response variable (more details are provided in the subsequent paragraph). For an aging metric (and as the independent variable) we use percentiles associated with number of games played. Other aging metrics such as player age could have been considered but we feel using games played better represents the stresses that game action applies to the body. Additionally, it allowed us to bin the process in a way that made comparisons and borrowing of strength among and between groups reasonable.

It is fairly well known that a gold standard "production" metric in basketball does not exist (García and Caro 2011). All metrics – from win shares (Calculating Win Shares 2013) to value over replacement player (VORP) – provide an incomplete picture of a player's worth. With that in mind, we somewhat arbitrarily chose to use John Hollinger's Game Score to represent player game production. Game Score is the following linear combination of common box-score summaries:

$$\begin{aligned}\mathrm{GmSc}=&\mathrm{PTS}+0.4\mathrm{FG}-0.7\mathrm{FGA}-0.4\mathrm{FTM}+0.7\mathrm{ORB}+0.3\mathrm{DRB}\\&+\mathrm{STL}+0.7\mathrm{AST}+0.7\mathrm{BLK}-0.4\mathrm{PF}-\mathrm{TOV},\end{aligned}\tag{1}$$

where PTS denotes total number of points scored in a game, FG denotes the number of field goals made, FGA the number of field goals attempted, FTM the number of free throws made, ORB the number of offensive rebounds, DRB the number of defensive rebounds, STL the number of steals, AST the number of assists, BLK the number of blocks, PF the number of personal fouls, and TOV the number of turnovers. This statistic is a well-known "add the good, subtract the bad" game-by-game production metric that provides a fairly decent evaluation of a player's production during any given game. That said, it has been criticized for overvaluing points scored by poor shooters (Berri 2012), for undervaluing defensive contributions, and for ignoring quality of opponent.

As mentioned in the introduction, for player-specific covariates we used position, usage rate, and average minutes per game. These three variables seem to partition players reasonably well with regards to Game Score making it sensible to borrow strength among the subgroups. Usage rate is defined as follows:

$$\begin{aligned}&\text{Usage Rate}\\&=\frac{[\,\mathrm{FGA}+(\,\mathrm{FTA}\times 0.44)+(\,\mathrm{AST}\times 0.33)+\mathrm{TOV}\,]\times 40\times\text{League Pace}}{(\,\text{Minutes}\times\text{Team Pace}\,)},\end{aligned}\tag{2}$$

(Hollinger 2005: p. 5), with Team Pace defined as

$$\text{Team Pace}=\frac{[\,\mathrm{FGA}-\mathrm{ORB}+\mathrm{TOV}+(\,FTA\times 0.44)]\times 48}{(\,\text{Team Minutes}\times 2\,)}\tag{3}$$

and League Pace defined as the average Team Pace for all NBA teams (p. 1). (In defining Team Pace, all quantities inside the brackets represent combined totals – for the team and its opponents – in games played by the team that season.) We categorized each player's usage rate as being high usage or low usage by comparing the player's average usage rate (across all games) to the corresponding median value among all players considered. We likewise categorized each player's minutes played per game as being either high or low via a comparison to the median value. Thus, a player was categorized as high usage high minute, high usage low minute, low usage high minute, or low usage low minute. Although a player's usage rate and minutes played per game might have changed substantially throughout his career, our method of categorization did not permit the player's group membership to change.

With regards to position, we used a trichotomous categorization. Although there are five designated positions in a basketball game – two guards (a point guard and a shooting guard), two forwards (a small forward and a power forward), and a center, there is some overlap between the five positions. Many players are, for instance,

a guard-forward or a forward-center. Ambiguity in using five positions to classify players arises because many shooting guards and small forwards are indistinguishable depending on their team's lineup decisions. Likewise, centers may play the power forward position depending on an opponent's skill set. Therefore, as is typically currently done, we categorize players as point guards, "wings," and "bigs. " In basketball terms, a "wing" is essentially a guard-forward – "wing" can be used to describe large shooting guards and small forwards. By characterizing centers and power forwards together as "bigs," we can lessen the common problem of classifying players of extremely ambiguous position. The number of players in each of the 12 groups is provided in Table 1.

We borrow strength between players by considering 12 different groups, corresponding to the 12 possible $p\times u\times m$ combinations. Being able to borrow strength between players (and ultimately make comparisons) requires in some sense a matching of career trajectories as players have played a varying numbers of games and minutes. This is of course no easy task. The ideal situation would be for the explanatory variable for each player to more or less represent the same time frame. We believe that one way this can be accomplished is to match the percentiles associated with the number of games played for each player. This naturally leads to considering 100 bins. We use z_{it} to denote the games comprising the tth percentile among all games played by player i. We will use y_{it} to denote the mean Game Score for player i computed across all games in z_{it}. In what follows we will refer to z_{it} as "game percentile." We employ this binning process with the understanding that some players have longer careers than others, so the last percentile for one player might be during his second season while for another player it might be during his fifteenth season. However, we do feel that using usage rate and minutes played to group players does a good job of mitigating this possibility and thus grouped player career trajectories match fairly well (making borrowing of strength sensible).

Table 1 Number of players in each of the j=1, …, 12 $p\times u\times m$ groups.

Usage/Minutes	Bigs	Wings	Point guards	Total
U+/M+	13	17	6	36
U+/M-	13	20	6	39
U-/M+	4	8	3	15
U-/M-	36	26	16	78
Total	31	71	56	178

3 Model

It is reasonable to believe that a player's career Game Score curve as a function of game percentile is not linear. Some player career production curves display an initial increasing trend until a plateau where the player is most productive, eventually followed by a decreasing trend towards the end of the career. This indeed seems to be the case for Glenn Robinson (See Figure 1). However, Shaquille O'Neal displays a brief dip early in his peak period and a relatively long period of decline, while Brevin Knight oscillates between increasing and decreasing production. Thus, it appears as if production curves are highly variable from one player to the next. We anticipate there might be dramatic differences between players based on position, minutes played, and usage rate, and we therefore allow for 12 different group-mean curves, one for each $p\times u\times m$ group. But we also recognize that even players in the same group might exhibit vastly different production curves. In light of this, production curves should be modeled flexibly along with the curves associated with the 12 $p\times u\times m$

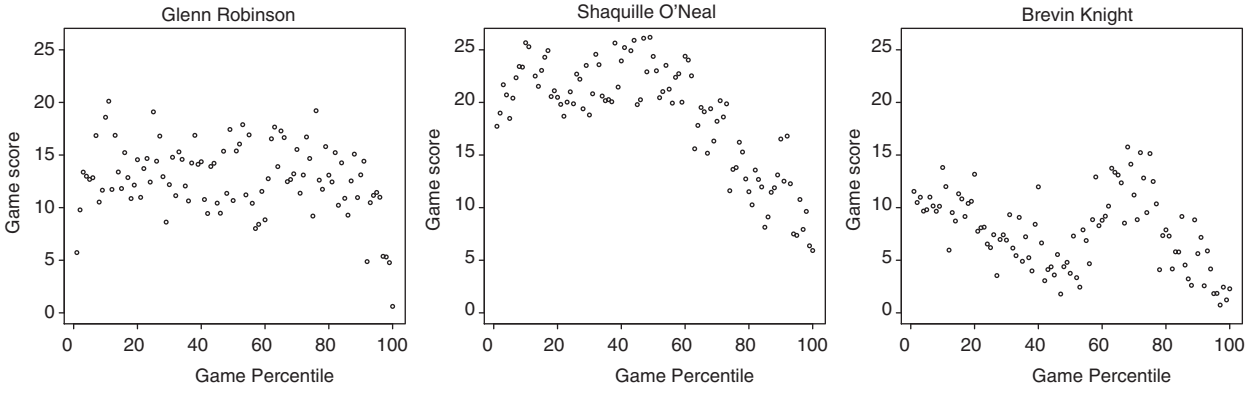

Figure 1 Plots of Hollinger's game score statistic versus game percentile for Glenn Robinson, Shaquille O'Neal, and Brevin Knight.

groups. This can be handled via a hierarchical model where GP regression is employed both at the observation level and process level of the hierarchy. As GPs are central to our modeling approach, they are briefly introduced in the next subsection.

3.1 Gaussian process

We very briefly introduce Gaussian processes (GP) here. For a more general and thorough introduction see Rasmussen and Williams (2006). (In what follows all letters in *bold* font will denote vectors.) Generally speaking, a GP can be considered as a distribution over functions $f(z):\Re\to\Re$ where any finite number of realizations $(f(z_1), f(z_2), \ldots, f(z_T))$ follow a multivariate Gaussian distribution that is completely characterized by a mean function $m(\boldsymbol{z}):\Re^T\to\Re^T$ and positive definite covariance function $K(\boldsymbol{z};\boldsymbol{\theta}):\Re^T\to\Re^{T\times T}$. Here, $\boldsymbol{\theta}$ is a vector of parameters and $\boldsymbol{z}=(z_1, \ldots, z_T)$. Thus, if we let f denote some unknown function with independent variables $z_1, \ldots, z_T$, then f follows a GP (commonly denoted as $f\sim\mathcal{GP}(m(\boldsymbol{z}),K(\boldsymbol{z};\boldsymbol{\theta}))$) if for any $T<\infty$ the realizations of f are distributed as

$$[f(z_1), f(z_2), \ldots, f(z_T)]'\sim N_T(m(\boldsymbol{z}), K(\boldsymbol{z};\boldsymbol{\theta}))$$

What is of principal interest in the current modeling is the estimation of mean function $m(\boldsymbol{z})$ while f is secondary. We make these ideas more concrete relative to the NBA production curve context in the next section.

Selecting an appropriate covariance function can be a difficult model selection exercise when using GPs. For computational convenience and sake of simplicity, we employ the commonly used Gaussian covariance function. The covariance matrix produced by this function has (i, ℓ) element given by the following form

$$\phi\exp\{-\rho(z_i-z_\ell)^2\}, \tag{4}$$

where ρ is a correlation scale parameter that controls how much the functional curve is allowed to vary across time (as $\rho\uparrow$ the curve becomes more linear), and ϕ is a variance parameter that controls the smoothness of the curve (as $\phi\uparrow$ the curve becomes more up and down). Thus for this covariance function $\boldsymbol{\theta}=(\rho, \phi)$.

3.2 Hierarchical specification of the model

As stated earlier, we use z_{it} to denote the games in the ith player's tth career length percentile, t=1,…, 100. Since the entries of $\boldsymbol{z}_i=(z_{i1}, \ldots, z_{i100})$ are equal for each player, for notational convenience in what follows we use $\boldsymbol{z}$ to denote each player's game percentile vector. Also, we use y_{it} to denote the ith player's average production during games in z_{it}. We use x_i to denote a player-specific variable indicating to which of the j=1, …, 12 $p\times u\times m$ groups player i belongs. Each player's Game Score is then modeled as a function of game percentile as follows

$$y_{it}=\beta_{0i}+f_i(z_t)+\varepsilon_{it} \text{ with } \varepsilon_{it}\overset{ind}{\sim}N(0, \sigma_i^2). \tag{5}$$

Recognizing that inevitably some players are more productive than others, we include a player specific "intercept" β_{0i} to account for some of the systematic variability in average player performance over time. To equation (5) we add the following hierarchical structure

$$f_i(\boldsymbol{z})\,|\,x_i=j\sim\mathcal{GP}(m_j(\boldsymbol{z}), K_j(\boldsymbol{z};\phi_j,\rho_j)), \tag{6}$$

$$m_j(\boldsymbol{z})\sim\mathcal{GP}(\boldsymbol{0}, K(\boldsymbol{z};s,r)), \tag{7}$$

where $m_j(\boldsymbol{z})$ and $K(\boldsymbol{z};\phi_j,\rho_j)$ denote the jth group's mean and covariance functions with s and r being user supplied. We are now modeling each player's production curve and the mean curve for each of the j=1, …, 12 $p\times u\times m$ groups using GP regressions. This allows us to estimate group curves while borrowing strength across players that we expect to have somewhat similar production curves. To make the MCMC algorithm employed (details of which are in the next section) numerically stable, we standardized $\boldsymbol{z}$ to the unit interval by dividing each entry of $\boldsymbol{z}$ by 100. We note that this does not affect the shape of the mean curve, so we use the original units in $\boldsymbol{z}$ when interpreting the results. To complete the Bayesian model we assigned fairly diffuse ($Ga(1, 1)$) independent priors for $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_{12})$ and $\boldsymbol{\rho}=(\rho_1,\ldots,\rho_{12})$, a diffuse Gaussian prior ($N(0, 100)$) for β_{0i} and set s=100 and r=1 which provides prior covariance matrices with variances equal to 100.

A few comments of the features that accompany this model are warranted. The methodology is extremely flexible in estimating production curves. No rigid polynomial forms have been assumed. The uncertainty of production curve form has been incorporated in all parameter estimates. Also, a natural mechanism of Bayesian hierarchical modeling is that the estimates for individuals are shrunk to the overall mean.

3.3 Computation

The joint posterior distributions arrived at through GP regression are intractable analytically. Because exact

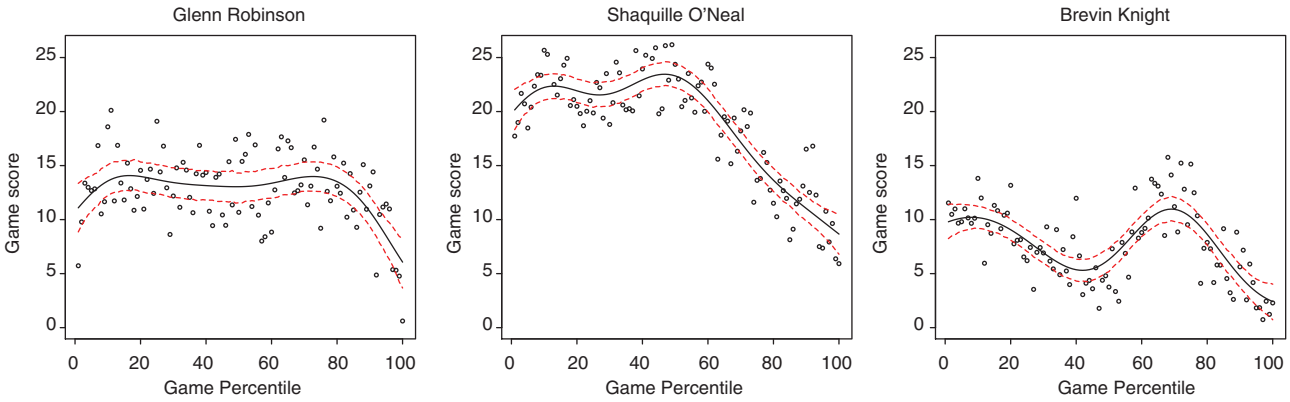

Figure 2 Individual fitted curves and 95% mean curve credible intervals for Game Score corresponding to Glenn Robinson, Shaquille O'Neal, and Brevin Knight.

sampling is not possible, we resort to MCMC methods to sample from the joint posterior distribution. Specifically, we use a hybrid Gibbs sampler (Geman and Geman 1984) and Metropolis-Hastings (M-H) algorithm (Metropolis et al. 1953 and Hastings 1970). Gibbs steps can be used to update the $f(\cdot)_i$'s, the $m(\cdot)_j$'s, and the β_{0i}'s. To update

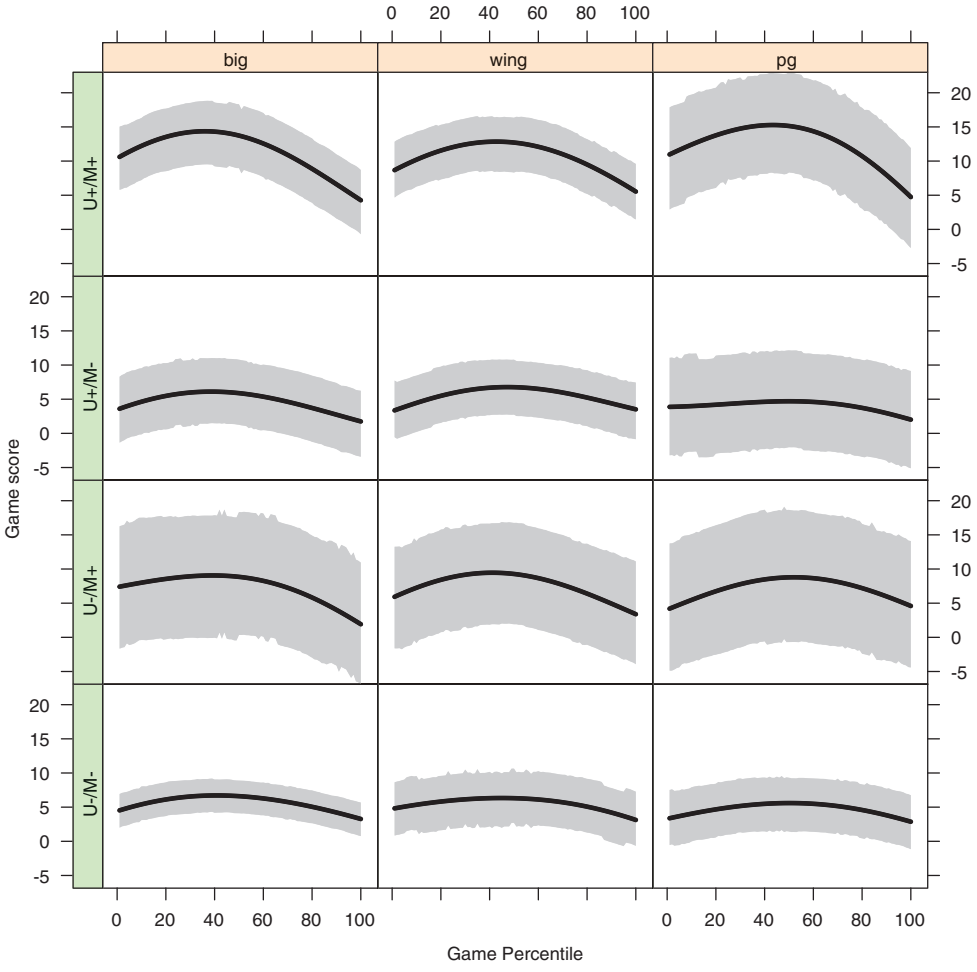

Figure 3 Process level mean production curves with 95% posterior credible bands corresponding to Game Score for each of the 12 $p \times u \times m$ categories. In the row labels, the first symbol depicts usage rate (+for high, –for low), and the second depicts minutes per game.

ϕ and ρ a random walk M-H step with Gaussian proposal distributions was used. Convergence was assessed using history plots of 3000 MCMC draws. These were collected by running three independent chains until 1000 MCMC iterates were obtained after a burn-in of 10,000 and a thinning of 50. (Figures 6 and 7 in the Appendix provide trace plots for a few randomly selected parameters.)

4 Results

Results are presented graphically. We briefly highlight fitted curves for individual players and variability among the β_{0i}'s, but focus the majority of our attention on the estimates of the 12 $m_j(\cdot)$'s. As a means of shedding some light on the variability of the β_{0i}'s we calculated the standard deviation of the 178 β_{0i} posterior means. This standard deviation is only 1.57. Thus the β_{0i}'s do not capture much of the player-to-player variability.

4.1 Estimated curves for individual players

As a means to show individual estimated curves are reasonable, Figure 2 provides fitted curves with 95% credible bands corresponding to the three players highlighted in Figure 1 (note that z_t were transformed to their original units). The fits are all quite sensible.

Table 2 Estimated mean (sd) game percentile where maximum Game Score is achieved for j=1,…, 12 $p\times u\times m$ groups.

Usage/Minutes	Bigs	Wings	Point Guards
U+/M+	36.0 (3.5)	42.7 (3.9)	43.2 (4.7)
U+/M-	38.6 (5.1)	47.4 (4.5)	42.2 (19.7)
U-/M+	36.6 (12.4)	41.1 (5.4)	52.2 (10.8)
U-/M-	41.1 (4.2)	44.7 (7.1)	50.0 (7.5)

4.2 Average group production curves and comparisons

Figure 3 displays the mean curves and error bands for each of the 12 $p\times u\times m$ categories for Game Score. Notice that the uncertainty associated with production curves is not the same across groups; the uncertainty is especially large for point guards and for players with high minutes but low usage rates. This is to be expected as the number of players in each group varies.

Figure 4 depicts the same process level production curves as Figure 3, but including the 12 curves in the same plot aids in comparison across groups. It appears that players with high minutes per game tend to experience more dramatic changes over the course of their careers than players with low minutes. This is particularly apparent for players who, in addition to high minutes, have high usage rates. There is also some indication that players with high minutes and high usage rates tend to

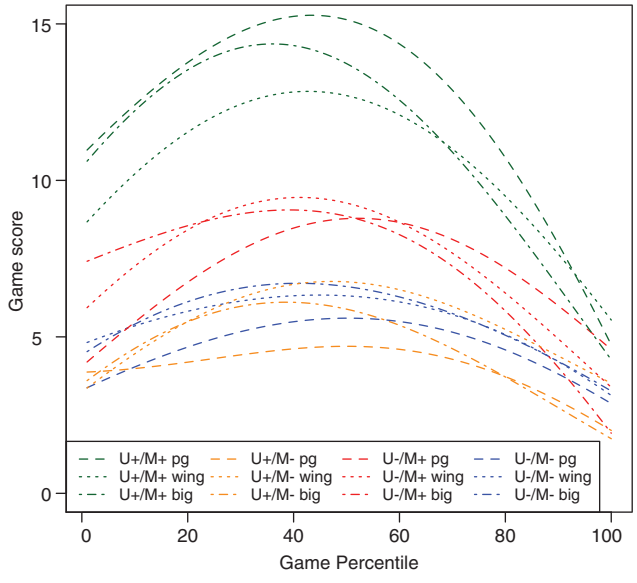

Figure 4 Process level mean production curves corresponding to game score for each of the 12 $p\times u\times m$ categories. In the legend, the first symbol depicts usage rate (+for high, –for low), and the second depicts minutes per game.

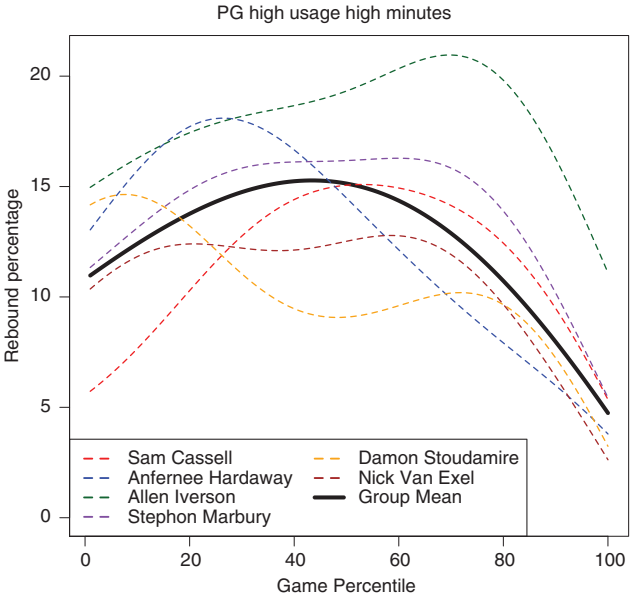

Figure 5 Individual mean player production curves for the six high usage, high minutes point guards. The solid line depicts the group average.

reach their peak earlier. One possible explanation is that they more rapidly accumulate wear and tear on their bodies with each game because of the time and energy exerted. Though not as pronounced, there also appears to be a tendency for bigs to peak slightly earlier than wings and point guards, and for point guards to peak slightly later than the others.

Table 2 provides an estimated mean game percentile for which the maximum game score is achieved. As indicated by Figure (4) the point guard position generally tends to attain peak performance around the 50th percentile, while for wings this occurs around the 45th and for bigs it is around the 40th. That said, the variability is quite large in a few categories and therefore conclusions must be drawn cautiously.

Figure 5 displays the individual production curves for the six high usage, high minutes point guards included in our analyses. It can be seen that there is some variation in the overall curve heights, yet there is a tendency for these players to have a gradual increase followed by a rapid decline. This overall tendency supports our choice of categories because, imperfect as the choice may be, the individual curves have fairly similar heights and shapes.

5 Conclusions

We have posed a hierarchical GP regression model that flexibly estimates career production curves for 12 $p \times u \times m$ categories. The general finding from this is that aging trends do impact natural talent evaluation of NBA players. Point guards appear to exhibit a different aging pattern than wings or bigs when it comes to Hollinger's Game Score, with a slightly later peak and a much steeper decline once that peak is met. The fact that the average point guard spends more of his career improving his game score is not lost on talent evaluators in the NBA, and as such, point guards are given more time on average to perform the functions of a rotation player in the NBA. Also, regardless of position, players with high minutes tend to have greater fluctuation in Game Score

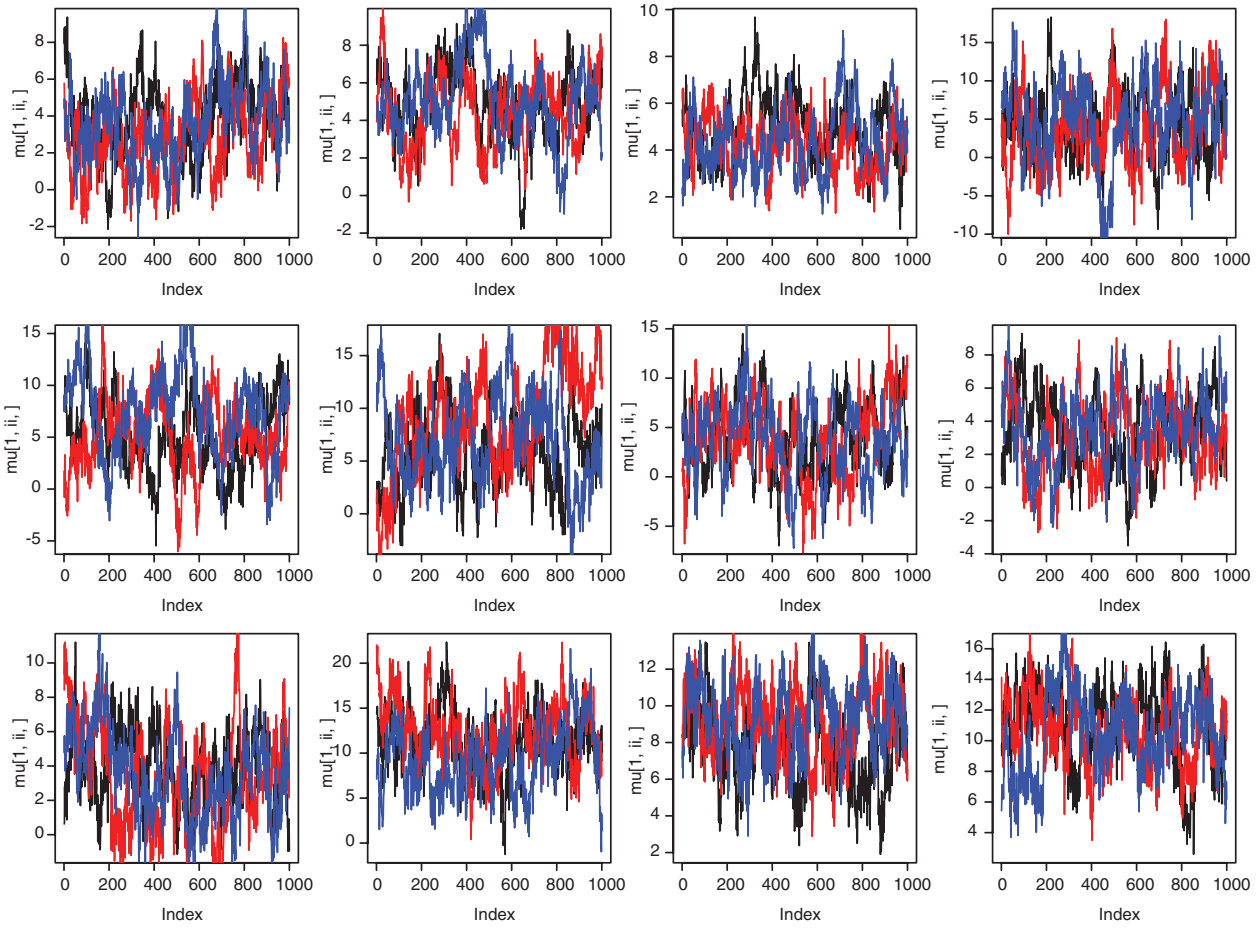

Figure 6 History plots of MCMC iterates for the first entry of $\mu(\cdot)_j$, j=1, …, 12.

production over time. While these are the general results of interest, the difference between individual curves is also of note.

The information gained from this modeling exercise is a nice starting point in the study of NBA player production curves. Obviously production curves are functions of much more than the three covariates that we considered. Factors such as quality of teammates, coaching philosophy and injury surely influence production (Fearnhead and Taylor 2011). Including these variables as they become available could potentially improve inference. Additionally, future research could focus on estimating career timepoints that correspond to slope inflections of a production curve in addition to incorporating a "team role" variable. For NBA decision-makers mulling over new contracts, research of this nature could be very valuable in determining contracts on-the-margins, and is without question a worthy subject of future study

Acknowledgments: The authors would like to thank the editor, associate editor, and anonymous referees for comments that improved the manuscript. The first author's work was partially funded by grant FONDECYT 11121131.

Appendix A

MCMC iterate history plots

We provide a few arbitrarily selected history plots of MCMC iterates (see Figures 6 and 7). The first set corresponds to the first entry of $\mu(\cdot)_j$ for the twelve groups, and the second set corresponds to ρ_j for the twelve groups. The general indication these plots provide is that the MC chains have reached their equilibrium states.

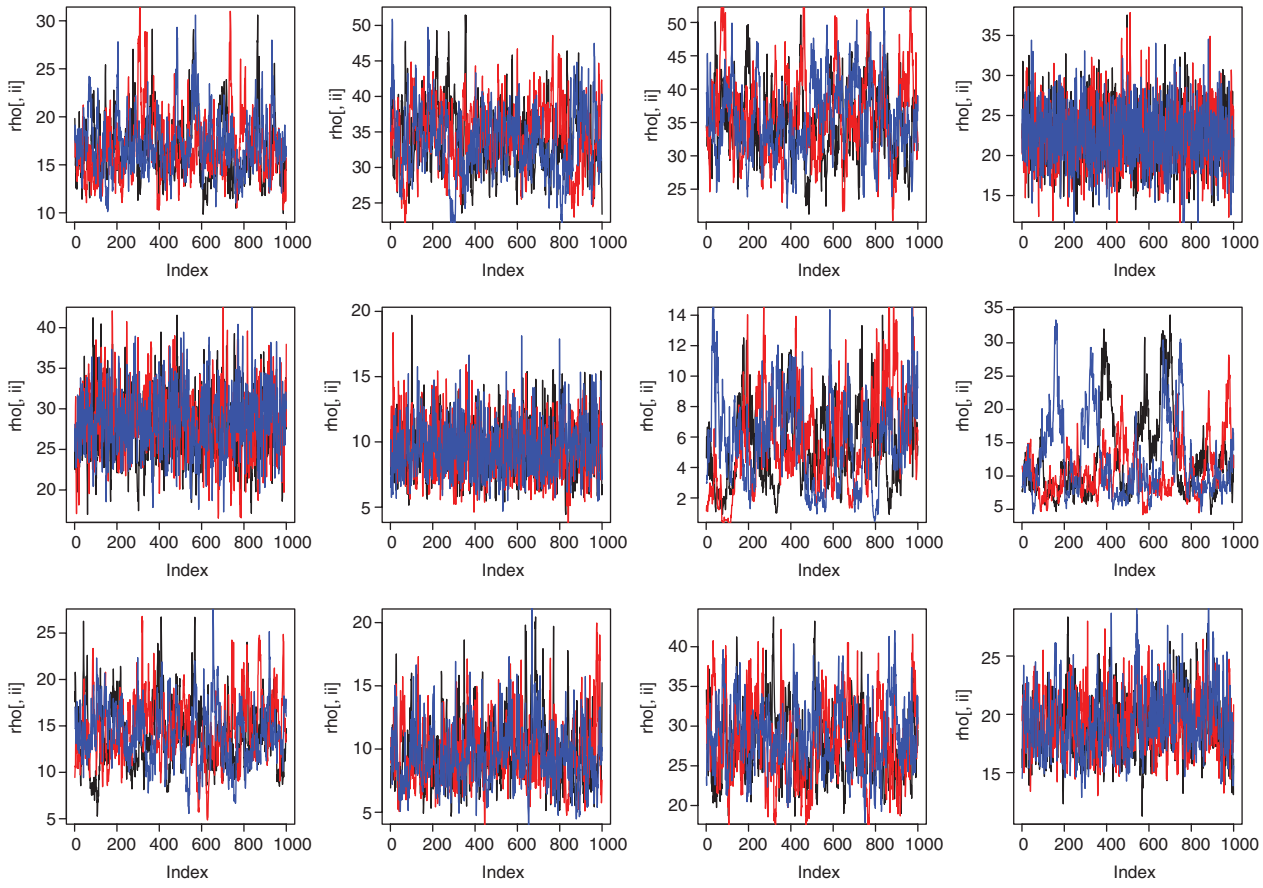

Figure 7 History plots of MCMC iterates for ρ_j, j=1, …, 12.

References

Albert, J. 2009. "Is Roger Clemens' Whip Trajectory Unusual?" *Chance* 22(2): 8–20.

Behseta, S., R. E. Kass, and G. L. Wallstrom. 2005. "Hierarchical models for assessing variability among functions." *Biometrika* 92: 419–434.

Berri, D. 2012. "David Berri educates us on John Hollinger." (URL: http://www.3sob.com/december-2012) david-berri-educates-us-on-john-hollinger/5515.

Berry, S. M., C. S. Reese, and P. D. Larkey. 1999. "Bridging Different Eras in Sports." *Journal of the American Statistical Association* 94: 661–676.

"Calculating Win Shares." 2013. (URL: http://www.basketball-reference.com/about/ws.html).

Fearnhead, P. and B. M. Taylor. 2011. "On Estimating the Ability of NBA Players." *Journal of Quantitative Analysis in Sports* 7(3): article 11.

García, J. A. M. and L. M. Caro. 2011. "A Stakeholder Assessment of Basketball Player Evaluation Metrics." (URL: http://hdl.handle.net/10045/16880).

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741.

Hastings, W. K. 1970. "Monte Carlo Methods using Markov Chains and their Applications." *Biometrika* 57: 97–109.

Hollinger, J. 2003. Pro Basketball Prospectus, Brassey's Inc.

Hollinger, J. 2005. Pro Basketball Forecast, Potomac Books, Inc.

Mahoney, R. 2010. "NBA's Statistical Revolution Bringing Real Change, More Winning." Retrieved January 10, 2013 (URL: http://probasketballtalk.nbcsports.com/2010/03/12/nbas-statistical-revolution-bringing-real-change-more-winning).

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. "Equations of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21: 1087–1091.

Ozmen, M. U. 2012. "Foreign Player Quota, Experience and Efficiency of Basketball Players." *Journal of Quantitative Analysis in Sports* 8(1): article 6.

Page, G. L., G. W. Fellingham, and C. S. Reese. 2007. "Using Box-Scores to Determine a Position's Contribution to Winning Basketball Games." *Journal of Quantitative Analysis in Sports* 3(4): article 1.

Pelton, K. 2010. "Rethinking NBA Aging." Retrieved January 10, 2013 (URL: http://basketballprospectus.com/article.php?articleid=896).

Rasmussen, C. E. and C. K. I. Williams. 2006. Gaussian Processes for Machine Learning, Massachusetts Institute of Technology, MIT Press.

Shoals, B. 2010. "For prime Out Loud!." Retrieved January 10, 2013 (URL: http://www.aolnews.com/2010/01/20/for-prime-out-loud).