Leonardo Egidi* and Jonah Gabry

Bayesian hierarchical models for predicting individual performance in soccer

https://doi.org/10.1515/jqas-2017-0066

Abstract: Although there is no consensus on how to measure and quantify individual performance in any sport, there has been less development in this area for soccer than for other major sports. And only once this measurement is defined does modeling for predictive purposes make sense. We use the player ratings provided by a popular Italian fantasy soccer game as proxies for the players' performance; we discuss the merits and flaws of a variety of hierarchical Bayesian models for predicting these ratings, comparing the models on their predictive accuracy on hold-out data. Our central goals are to explore what can be accomplished with a simple freely available dataset comprising only a few variables from the 2015–2016 season in the top Italian league, Serie A, and to focus on a small number of interesting modeling and prediction questions that arise. Among these, we highlight the importance of modeling the missing observations and we propose two models designed for this task. We validate our models through graphical posterior predictive checks and we provide out-of-sample predictions for the second half of the season, using the first half as a training set. We use Stan to sample from the posterior distributions via Markov chain Monte Carlo.

Keywords: graphical posterior predictive checking; hierarchical model; missing observations; players' performance; soccer prediction.

1 Introduction

Compared to the volumes statisticians (professional and amateur) have written about baseball, and to the growing statistical literature on sports like basketball and American football, there has been relatively little published by statisticians about soccer. A few highlights from the limited statistical literature include: Baio and Blangiardo (2010), who use a Bayesian hierarchical model to predict the outcome of individual matches throughout a season in the top Italian league, Serie A; Karlis and Ntzoufras (2000), in which the authors take a frequentist approach to estimating parameters related to the number of goals scored by specific teams; Dixon and Coles (1997), who use a familiar Poisson model for the number of goals between two teams and also consider suitable betting strategies based on their model; and Karlis and Ntzoufras (2009), which is a Bayesian model for the goal differential between two teams using a Skellam (Poisson difference) distribution.

In most of the published statistical research on soccer, including the papers mentioned above, the authors do not focus on modeling the performance of individual players over the course of a season but rather on some aspect of the global result of a match between opposing teams (e.g. goal differential), or on predicting the order of the league table at the end of a season. Relative to sports like baseball (Albert 1992) or American football (Becker and Sun 2016), the performance of individual soccer players is noisy and hard to predict. The dimensions of the pitch combined with the number of players, the difficulty of controlling the ball without the use of hands, and many other factors all contribute to the predictive challenge.

More primitive than the question of how to model player performance is how to *measure* it. Although there is no consensus on how to quantify individual performance in any sport, there has been less development in this area for soccer than for other major sports. And only after measurement is defined does modeling make sense. The oldest procedure for measuring the individual performance in the so called *goal-based* team sports – hockey, soccer and basketball, among others – is the *plus/minus* approach (see Thomas et al. (2013) for some references and recent improvements). A player is rewarded for being in the game when positive events occur for their team and penalized for being in the game when negative events occur. Measuring the individual abilities of players who share the pitch (or ice or court) for much of their time is challenging in any sport, but the rarity of goals in soccer makes the plus/minus system even more problematic.

Although we are interested in modeling the overall performance of individual players, we are not yet convinced that there is an available holistic measure of

***Corresponding author: Leonardo Egidi,** Department of Business, Economics, Mathematics and Statistics 'Bruno de Finetti', University of Trieste, Trieste, TS, Italy, e-mail: leoegidi@hotmail.it, legidi@units.it
Jonah Gabry: Department of Statistics, Columbia University, New York, NY, USA, e-mail: jgabry@gmail.com

individual performance worth modeling. In fact, even as the amount and variety of publicly available soccer data grows – particularly data at the individual player/match level – the interpretability and predictive relevance of that data will remain a question. However, we do suspect that *fantasy* soccer (Lomax 2006; Bonomo, Durán, and Marenco 2014) may be more amenable to modeling given that it has more clearly defined measures of performance. That is, a prediction task for individual fantasy ratings could be well posed and also serve as an example of a possible approach to use in the future when better measures of individual performance in soccer matches become available. The outcome of interest is the fantasy rating of each player in Italy's top league, Serie A, for each match of the 2015–2016 season. We strongly believe that these fantasy ratings may be seen as a proxy for the quality of a player's performance; in fact, they combine a subjective evaluation with an objective factor accounting for specific in-game events. Moreover, given the popularity of such fantasy games, these ratings are themselves an interesting variable to model. In this paper we present and critique several Bayesian hierarchical models (Gelman and Hill 2006; Gelman et al. 2013) designed to predict the results of the Italian fantasy game Fantacalcio. We use RStan (Stan Development Team 2016a), the R (R Core Team 2016) interface to the Stan C++ library (Stan Development Team 2016b), to sample from the posterior distributions via Markov chain Monte Carlo. As far as we can tell from reviewing the literature, there have been no published attempts to use a hierarchical Bayesian framework to address the challenges of modeling this kind of data.

Our central goals are to explore what can be accomplished with a very simple dataset comprising only a few variables (that are freely and easily available), and to focus on a small number of interesting modeling and prediction questions that arise (for instance, those due to the missingness of certain values). For this reason we also gloss over many issues that we believe should be of interest in subsequent research, for instance variable selection, additional temporal correlation structures, and the possibility of constructing more informative prior distributions. Although we restrict our focus to Fantacalcio, the process of developing these models and comparing them on predictive performance does not entirely depend on the idiosyncrasies of this particular fantasy system and is applicable more broadly.

The rest of the paper is structured as follows. We briefly introduce the Italian fantasy soccer game Fantacalcio and we describe our dataset in Section 2. The models we fit to the data are presented in Section 3 with results in Section 4. In Section 5 we carry out a variety of posterior predictive checks as well as out-of-sample prediction tasks. Section 6 concludes. A Supplementary Material file containing data, code and further analysis is provided.

2 Data

In Italy, fantasy soccer was popularized by the brand Fantacalcio edited by Riccardo Albini in the 1990s (`http://www.fantacalcio.it`). At the beginning of the season, Fantacalcio managers are allocated a limited amount of virtual money with which to buy the players that will comprise their roster. After every match in Serie A, the prominent Italian sports periodicals assign each player a rating, a so-called *raw score*, on a scale from one to ten. These are very general and largely subjective performance ratings and there tends not to be much variability in these scores. As a means of systematically including specific in-game events in the ratings, Fantacalcio provides the *point scoring* system. Points are added or deducted from a player's initial raw score for specific positive or negative events during the match.

For player i in match t the total rating y_{it} is

$$y_{it} = \mathsf{R}_{it} + \mathsf{P}_{it}, \tag{1}$$

where R is the raw score and P is the point score. Table 1 lists the game features that contribute to a player's point score P_{it} for a given match. Negative ratings are possible, although not very common. For instance, a goalkeeper with a raw score of three who also allows four goals would have a rating $y_{it} = -1$.

Since it is very rare for a player to participate in all matches, some y_{it} are *missing*, and this may be due to different reasons. First, player i's rating for match t will be missing if the player does not play in the match because of injury, disqualification, coach's decision, or some other reason. In addition, this can occur when a player does not participate in the match for long enough for their impact

Table 1: Bonus/Malus points in Fantacalcio.

Event	Points
Goal	+3
Assist	+1
Penality saved*	+3
Yellow card	−0.5
Red Card	−1
Goal conceded*	−1
Own Goal	−2
Missed penality	−3

The events marked with a * symbol are only applicable to goalkeepers.

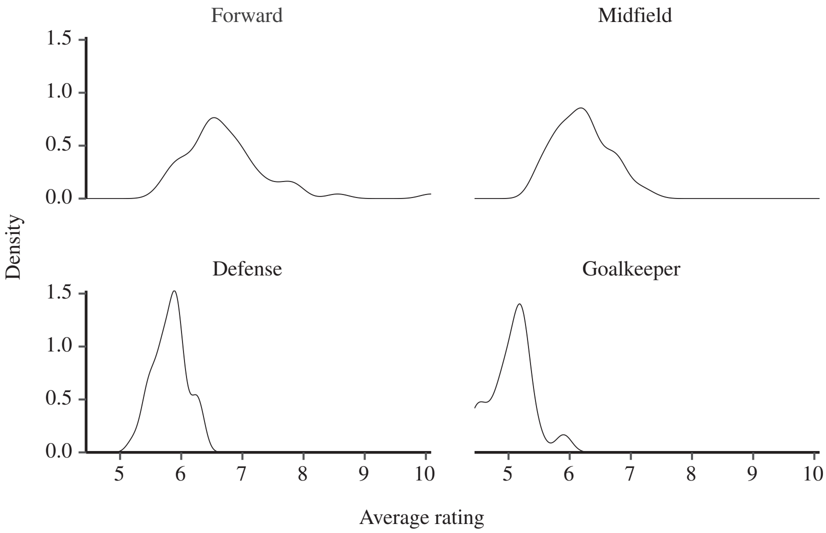

Figure 1: The distributions of average ratings by position.

to be judged by those tasked with assigning the subjective raw score ($\mathsf{R}_{it} = 0$) or for the player to accumulate or lose any objective points ($\mathsf{P}_{it} = 0$).

Modeling the missingness is one of focuses of this paper. We return to this issue later in Sections 3.2 (mixture models) and 3.3 (missing data models) when we confront the challenge it poses for our modeling and prediction tasks and consider methods for modeling the missingness that naturally arises in our dataset.

All data for this paper are from the 2015–2016 season of the Italian Serie A and were collected from the Italian publication La Gazzetta dello Sport (`http://www.gazzetta.it`). We use all of the ratings for every player satisfying the following two criteria:

- The player participated in at least a third of matches during the *andata* (the first half of the season). This amounts to dropping players who played in fewer than seven matches in the first half.
- The player participated in the *final* match of the *andata*.

The latter criterion is a simple constraint for considering only those players regularly enrolled in the squad list for the last game of the *andata* – our training set, as explained in Section 5 – which thus belong to the Italian Serie A in the second half of the season with high probability. Professional European soccer leagues allow for a player to be transferred to another team (not necessarily in the same league) at approximately the midpoint of the season, but only a few players in our dataset ended up changing team so we simply set each player's team id in the dataset to their team at the beginning of the season. This avoid the complication of accounting for transfers in the modeling stage, although it would be interesting to investigate this in future work. Our final dataset contains ratings for 237 players (18 goalkeepers, 90 defenders, 78 midfielders, and 51 forwards). Figure 1 displays the distributions of average ratings by position, while Figure 2 shows the bivariate relationship between average rating and the initial standardized price for each player. Although the full season comprises 38 matches for each team, rarely does a player participate in all matches. For the 237 players in our data that meet the two criteria above, the mean number of matches played is 27.5 with a standard deviation of about 7, and 75% of these players missed at least five matches.

Notation for observed data

There are $N = 237$ players and $T = 38$ matches in the dataset. When fitting our models we use only the $T_1 = 19$ matches from the first half of the 2015–2016 Serie A season. The remaining $T_2 = 19$ matches are used later for predictive checks. For match $t \in \{1, \ldots, T\}$, let y_{ijkt} denote the value of the total rating for player $i \in \{1, \ldots, N\}$, with position (role on the team) $j \in \{1, \ldots, J\}$, on a team in team-cluster $k \in \{1, \ldots, K\}$. To ease the notational burden, throughout the rest of the paper the subscripts j and k will often be implicit and we will use y_{it} in place of y_{ijkt}.

The players are grouped into $J = 4$ positions (forward, midfielder, defender, goalkeeper) and $K = 5$ team clusters. The five clusters (Table 2) were determined using the official Serie A rankings at the midpoint of the season. The purpose of the team clustering is both to use a grouping structure that has some practical meaning in

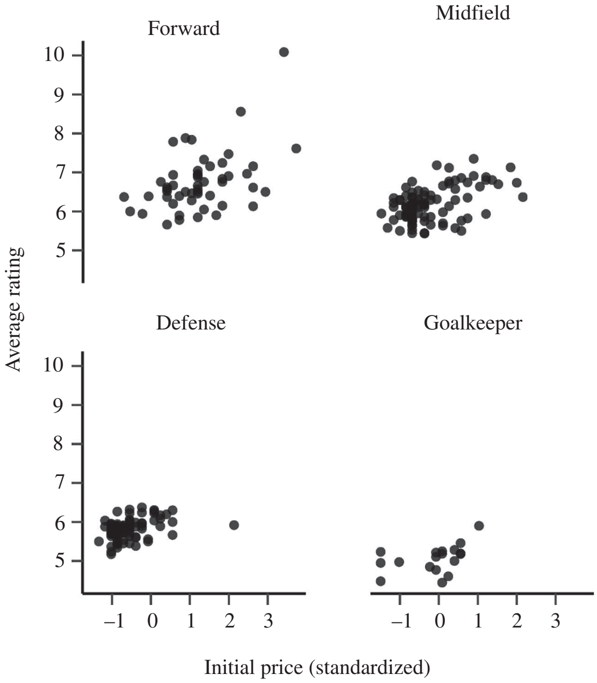

Figure 2: The distributions of average ratings versus initial standardized price.

Table 2: The $K = 5$ team clusters, from weakest to strongest.

Cluster	Teams
1	Palermo, Frosinone, Carpi, Verona
2	Genoa, Sampdoria, Empoli, Udinese
3	Bologna, Chievo, Atalanta, Torino
4	Milan, Fiorentina, Lazio, Sassuolo
5	Juventus, Roma, Inter, Napoli

Group 5 is headlined by Juventus, the top performing team in Serie A for the past several seasons.

this context and also to reduce the computational burden somewhat by including cluster-specific parameters rather than team-specific parameters. We experimented with team-specific parameters but found that it results in models that are slower to fit but that yield similar inferences.

There are only two other variables in our limited dataset. We let $h_{it} = 1$ if player i's team plays match t at its home stadium and $h_{it} = 0$ if the match is played at the opponent's stadium. And we use q_i to denote the initial standardized price for player i. These values are assigned by experts and journalists at the beginning of the season based on their personal judgement and then updated throughout the season to reflect each player's performance (http://www.gazzetta.it/calcio/fantanews/statistiche/serie-a-2015-16/).

3 Models

Notation for model parameters

The notation we use for model parameters is similar to the convention adopted by Gelman and Hill (2006) for multilevel models. According to this, the index variables $j[i]$, $k[i]$ code group membership. For instance, if $j[1] = 4$, then the first unit in the data ($i = 1$) belongs to position group 4. If $k[1] = 3$, then the first unit belongs to team-cluster 3.

We use α_i for individual random effects corresponding to each player $i = 1, \ldots, N$. The parameters γ_k and $\beta_{k,\,t}$ represent, respectively, the team-cluster effect and the team-cluster effect of the team opposing in match t, with $k = 1, \ldots, K$. As already mentioned, in our simplified framework we set the number of team-clusters $K = 5$. We denote by ρ_j the position-specific parameters, with $j = 1, \ldots, J$ and $J = 4$. The standardized prices are multiplied by a slope δ_j, which is allowed to vary across the J positions. Because we are interested in detecting trends in player ratings, we also incorporate the average rating up to the game $t - 1$, $\bar{y}_{i,t-1}$, which is multiplied by a factor $\lambda_{j[i]}$ estimated from the data. In addition, the effect of home and away matches is accounted for in the θ parameter. For the mixture model in Section 3.2, the same average rating $\bar{y}_{i,t-1}$ is also multiplied by a coefficient $\zeta_{j[i]}$ in order to model the probability of participating in the match t. We anticipate that in their posteriors λ and ζ (here denoted as vectors) will be meaningfully different from zero. Since we work in a Bayesian framework, all parameters will be assigned prior distributions, which in turn may depend on hyperparameters that are either fixed or themselves estimated from the data.

3.1 Hierarchical autoregressive model (HAr)

As above, let y_{it} (with indices j and k implied) denote the total rating (1) for player i in match t. For our first model, we code all the missing ratings y as zeros. This makes sense if we are (and, in part, we are) interested in the annual *cumulative* rating of a given player, or of a given subset of players (this is investigated graphically later in Section 5). Or, for instance, experts and scouts may be interested in estimating the number of goals that will be scored by Roma's forwards. Since the number of goals heavily depends on the number of games played, it makes sense to assign a value of zero for any missed matches (unobserved player ratings) as they should not contribute to the total number of goals scored. Later, in Section 3.3, we will take a different approach in which missing values are actually treated as unobserved and we specify a full joint probability model for both the observed and unobserved ratings.

We begin with a standard hierarchical autoregressive model

$$y_{it} \sim \mathsf{Normal}\left(\eta_{it},\ \sigma_y\right), \tag{2}$$

where η_{it} is the linear predictor

$$\eta_{it} = \alpha_0 + \alpha_i + \beta_{k[i],t} + \gamma_{k[i]} + \rho_{j[i]} + \delta_{j[i]} q_i + \lambda_{j[i]} \bar{y}_{i,t-1} + \theta h_{it}, \tag{3}$$

α_0 is the intercept, and σ_y is the standard deviation of the error in predicting the outcome. The term autoregressive is used here for indicating the inclusion of the average rating up to the game $t-1$ in the model. As we are fitting our models using Stan (Stan Development Team 2016b), we follow its convention of parameterizing normal distributions in terms of standard deviation rather than the precision or variance.

The individual-level, position-level, and team-cluster-level parameters are given hierarchical normal priors,

$$\begin{aligned}
\alpha_i &\sim \mathsf{Normal}\left(0,\ \sigma_\alpha\right), \quad i = 1, \dots, N \\
\gamma_k &\sim \mathsf{Normal}\left(0,\ \sigma_\gamma\right), \quad k = 1, \dots, K \\
\beta_k &\sim \mathsf{Normal}\left(0,\ \sigma_\beta\right), \quad k = 1, \dots, K \\
\rho_j &\sim \mathsf{Normal}\left(0,\ \sigma_\rho\right), \quad j = 1, \dots, J
\end{aligned} \tag{4}$$

with weakly informative prior distributions for the remaining parameters and hyperparameters,

$$\begin{aligned}
\alpha_0 &\sim \mathsf{Normal}\,(0, 5) \\
\theta &\sim \mathsf{Normal}\,(0, 5) \\
\delta_j &\overset{iid}{\sim} \mathsf{Normal}\,(0, 5), \quad j = 1, \dots, J \\
\lambda_j &\overset{iid}{\sim} \mathsf{Normal}\,(0, 1), \quad j = 1, \dots, J \\
(\sigma_\theta, \sigma_\alpha, \sigma_\gamma, \sigma_\beta, \sigma_\rho) &\overset{iid}{\sim} \mathsf{Normal}^+(0, 2.5), \\
\sigma_y &\sim \mathsf{Cauchy}^+(0, 5),
\end{aligned}$$

where Normal^+ and Cauchy^+ denote the half-Normal and half-Cauchy distributions. See Gelman et al. (2006) and Gelman (2016) for a discussion related to the choice of these priors for the scale parameters. Note that centering the individual-level, the team-cluster-level, and the position-level parameters in (4) at μ_α, μ_γ, μ_β, and μ_ρ would make the model nonidentifiable, because a constant could be added to each of these hyperparameters without changing the predictions of the model. This is the motivation for centering these prior distributions at zero and including the global intercept.

A previous sensitivity analysis with different input values for these hyperparameters suggested no variation in the posterior estimates. However, other researchers and/or sports experts may have a particular instinct about these priors, and be more subjective in their elicitation. A deep illustration on subjective priors in sports analytics is provided by Silva and Swartz (2016), where the prior elicitation is driven by the instincts of the experts and of the gambling websites in a logistic regression framework.

3.2 Mixture model (MIX)

Even if we found that some players have a tendency to be ejected from matches due to red cards, for instance, or tend to suffer injuries at a high rate, it would still be very challenging to arrive at sufficiently informative probability distributions for these events. Even with detailed player histories over many seasons, it would be hard to predict the number of missing matches in the current season. Nevertheless, we can try to incorporate the *missingness* behavior intrinsic to the game into our models. Assuming that it is very rare for a player to play in every match during a season, we can try to model the overall propensity for missingness. A general way of doing this entails introducing a latent variable, which we denote V_{it} and define as

$$V_{it} = \begin{cases} 1, & \text{if player } i \text{ participates in match } t, \\ 0, & \text{otherwise.} \end{cases}$$

If for each player i we let $\pi_{it} = Pr\left(V_{it} = 1\right)$, then we can specify a mixture of a Gaussian distribution and a point mass at 0 (Gottardo and Raftery 2008)

$$p\left(y_{it} \mid \eta_{it}, \sigma_y\right) = \pi_{it}\, \mathsf{Normal}\left(y_{it} \mid \eta_{it}, \sigma_y\right) + \left(1 - \pi_{it}\right) \delta_0, \tag{5}$$

where δ_0 is the Dirac mass at zero and η_{it} is the same linear predictor as before. The probability π_{it} is modeled using a logit regression,

$$\pi_{it} = \mathrm{logit}^{-1}\left(p_0 + \zeta_{j[i]} \bar{y}_{i,t-1}\right), \tag{6}$$

which takes into account predictors that are likely to correlate with player participation. The variable $\bar{y}_{i,\,t-1}$ is the average rating for player i up to match $t-1$, and p_0 is an intercept for the logit model. How to model π_{it} could be the subject of entire papers, but better models would likely require variables beyond what we have in our dataset (e.g. injury histories). Our simplistic model will suffice for our purposes of exploring what we can do with only this dataset. For the new parameters introduced in (6) we use the weakly informative priors

$$\begin{aligned}
p_0 &\sim \mathsf{Normal}\,(0, 2.5), \\
\zeta_j &\overset{iid}{\sim} \mathsf{Normal}\,(0, 1), \quad j = 1, \dots, J.
\end{aligned}$$

The models for the group-level parameters and the hyperpriors are the same as in 3.1.

3.3 Refitting the HAr model accounting for missing data

As we have already mentioned, it is difficult to deal with the issue of missing data in such a way as to yield a reasonable estimate of the cumulative ratings over a season. The MIX model may be seen as a natural attempt at modeling the missingness, while, to ease the problem, in the initial HAr model missing values were treated as zeros and not modeled. We have already speculated about the legitimacy of this approach, but we are only partially interested in the cumulative rating over the entire season and are also interested in assessing the predictive accuracy of our models game by game. That is, we also want to answer the question: how will a player perform if they play in the match? One way to do this is by treating each missing player rating as an unknown parameter rather than somewhat arbitrarily fixing it at zero. As broadly outlined in Gelman et al. (2013), Bayesian inference draws no distinction between missing data and parameters, so the target distribution is the joint posterior distribution of the missing data and other model parameters conditional on the observed data.

Let y represent the complete data we could have observed in the absence of missing values; we split our data matrix into two subsets, $y = (y^{obs}, y^{mis})$, where y^{obs} denotes the observed values and y^{mis} denotes the missing values. We also define I to be the inclusion matrix such that $I_{it} = 1$ if y_{it} is observed and $I_{it} = 0$ if y_{it} is missing. In this setup, y^{obs} are data and y^{mis} are parameters. For convenience, we specify our new augmented model as

$$y_{it} = \begin{cases} y_{it}^{obs}, & \text{if } I_{it} = 1 \\ \xi_{it}, & \text{if } I_{it} = 0, \end{cases} \qquad i = 1, \ldots, N, \ t = 1, \ldots, T \tag{7}$$

where y_{it}^{obs} is an observed rating for player i in match t and each ξ_{it} is a parameter. The same idea is then incorporated into the HAr model from 3.1. We refer to this modified model as the HAr-mis model.

4 Results

4.1 Estimates

We fit the models via Markov chain Monte Carlo using RStan Stan Development Team (2016a), the R interface to the Stan C++ library Stan Development Team (2016b), and monitored convergence as recommended in Stan Development Team (2016c). Figure 3 shows the parameter estimates.

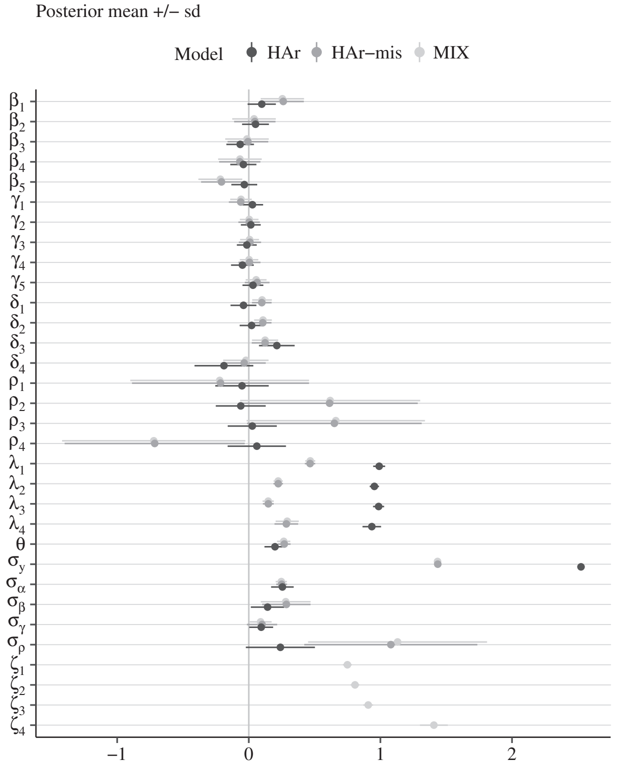

Figure 3: Posterior means ± standard deviations for the model parameters common to the HAr, HAr-mis, and MIX models. $\beta_{k,t}$ and γ_k are the parameters for the opposing team-cluster in match t and the player's team-cluster ($k = 1$ the weakest, $k = 5$ the strongest). The parameters δ_j (coefficients on initial price), λ_j (coefficients of the lagged average rating) and ρ_j all vary by position (1 = Forward, 2 = Midfield, 3 = Defender, 4 = Goalkeeper). θ is the coefficient for the home/away predictor. σ_y is the individual-level standard deviation and the other σ's are the hierarchical standard deviation parameters. For the MIX model, the ζ's are the coefficients on the lagged average rating from (6).

For all models, the β, γ and δ vectors are almost all shrunk towards their grand mean 0, with little variability. For the position-specific vector ρ, the HAr-mis and MIX models estimate slightly positive values (approximately 0.5) for midfielders (ρ_2) and defenders (ρ_3), while for the HAr model these parameters are shrunk close to zero. The goalkeeper effect (ρ_4) is slightly positive for the HAr model but clearly negative for the HAr-mis and MIX models. For all models these position-level parameters have larger posterior uncertainties than the other parameters. All three models recognize a slight advantage due to playing at home ($\theta > 0$). Also in Figure 3 we see that for the λ's, the coefficients on the lagged average ratings, the estimates obtained from the HAr model are much larger than those obtained under the HAr-mis and MIX models, which again give nearly identical estimates. Since for every match day t these coefficients are multiplied by the lagged average rating $\bar{y}_{i,t-1}$, we believe that the larger λ estimates from the HAr model are the result of coding the missing values as zeros.

For the MIX model only, there are also additional parameters $\zeta_1, \ldots, \zeta_4$ that scale the lagged average rating in the logit model (6). These parameters are all positive – which corresponds to the intuition that higher ratings are associated with higher probabilities of participating in the next match – and they also exhibit non-negligible variation across positions (for goalkeepers, ζ_4, the estimated association is strongest).

4.2 Inference through hypothetical data

In this section we give an example of a more interesting comparison focusing on simulating hypothetical players rather than comparing parameter estimates.

Comparing parameter estimates across models is standard practice, but we are more interested in the implications of the parameters for the outcome variable rather than the parameters themselves. For our purposes it should be more informative to simulate outcomes under each of the models for players differing only in their position. We can then directly compare the variability in the ratings for these hypothetical players. Note that comparing predictions rather than parameter estimates would be even more essential if we were fitting logistic regression models (or other GLMs) rather than Gaussian linear models.

We predict ratings for several players at different positions on the field and with the average position price in virtual money, all on the same cluster team, all playing against the same cluster team, and all playing at their home stadium. Figure 4 shows the predicted ratings from each of the models for 19 new matches and $N = 237$ (the size of our dataset) hypothetical players. For the HAr model, the variability within positions appears to be large when compared with the same variability in the predictions from the the HAr-mis model and, a bit less, the MIX model. Moreover, the different positions appear quite distinct according to the HAr-mis model, less under the MIX model and quite overlapped under the HAr model. The predicted values for the HAr model are shrunk together and turn out to be much too low for each position. This failure of the HAr model can be explained by the fact that it treats missed matches as zeros and, then, it will tend to favor players with fewer zeros.

Conversely, the simulations from the HAr-mis model are more clearly separated into strata corresponding to the different positions and the hierarchy of positions is correct: forwards tend to register the highest simulated ratings, then midfielders, defenders, and goalkeepers. The MIX model is less able to clearly separate the positions in the predictions but it does get the correct ordering on average. As expected, it also predicts a non-negligible number of zeros (missing values).

Here we only show the comparison made by varying a player's position, but analogous visualizations can be made to explore the effect of changing other variables.

5 Posterior predictive checks and predictions

Now that we have estimated all of the models, we turn our attention to evaluating the fit of the models to the observed data as well as the predictive performance of the models on hold-out data. We use the 19 match days comprising the first half of the Serie A season – the *andata* – as training data, and for every player in the dataset we make in-sample predictions for those 19 matches as well as out-of-sample predictions for the remaining 19 matches – the *ritorno*. As usual in a Bayesian framework, the prediction for a new dataset may be directly performed via the posterior predictive distribution for our unknown set of observable values. Following the notation of Gelman et al. (2013), we denote by $\tilde{y}$ a generic unknown observable. Its distribution conditional on the observed y is

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}, \theta|y)\, \mathrm{d}\theta = \int_{\Theta} p(\tilde{y}|\theta)\, p(\theta|y)\, \mathrm{d}\theta, \qquad (8)$$

where the independence of y and $\tilde{y}$ conditional on θ is assumed. We are also implicitly conditioning on the observed predictors. Sampling from this posterior predictive distribution will allow us to both assess the fit of the model to observed data and also make out-of-sample predictions that average over the posterior.

5.1 In-sample posterior predictive checks

To assess how well the models fit the training data, for each draw of the parameters from the posterior distribution we draw a dataset from the posterior predictive

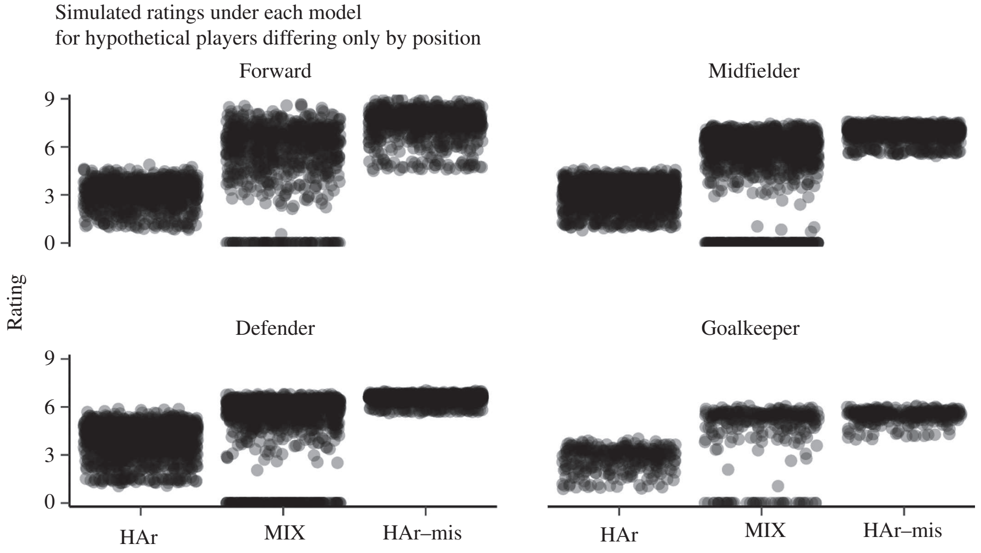

Figure 4: Predicted ratings of hypothetical players differing only in their position. Predictions from each of the three models are shown for 19 matches for each of 237 players (the size of our dataset), all playing at home ($h_{it} = 1$), all playing on a team in cluster $k = 3$ against an opponent in cluster $k = 3$, with standardized average position price $\bar{q}_{j[i]}$, $j = 1, \ldots, J$, $i = 1, \ldots, N$.

distribution of the outcome under each of the models. We should expect the in-sample predictive performance to be better than performance on out-of-sample prediction tasks (Gelman, Hwang, and Vehtari 2014; Vehtari, Gelman, and Gabry 2017). Figure 5 shows an example of a graphical posterior predictive check focusing on the *cumulative* ratings for each player over the matches in the training data. For illustration purposes, here we only show the results for one team, Napoli, but equivalent plots could be made analogously for all the other teams. The dashed black lines represent the observed values, while the red, green, blue lines represent predictions from the HAr, HAr-mis and MIX models, respectively.

For many of the players all of the models make reasonable predictions. However, for players with many missed matches the HAr and MIX models outperform the HAr-mis model (see the plot for El Kaddouri, for instance). The HAr-mis model will perform well on many of the predictive tasks, but it is not designed to predict in-sample cumulative ratings. The cumulutative rating is very sensitive to the number of missing values, but for each missing value the HAr-mis will predict a plausible rating for if the player had played instead of a zero.

Figure 6 provides a different graphical check of the model fitting. Each row of plots shows the distribution of a test statistic $T(y^{rep})$ computed over the replicated datasets y^{rep} generated from the posterior predictive distribution under each of the models. The vertical black lines indicate the value of $T(y)$, the statistic computed from the observed data. If we consider the distributions of these statistics – mean, median, minimum, maximum, and standard deviation – we immediately notice that the three models differ in their ability to replicate many of these features of the data. According to the mean, the median and standard deviation, the MIX model seems to be best at capturing these aspects of the training data.

In the fourth row we can see that the HAr model severely underestimates the minimum rating in the data, the HAr-mis model predicts a reasonable distribution of the minimum, and for the MIX model the distribution for the minimum is highly concentrated around 0, which is due to the nature of the model.

On the other hand, it is the MIX and the HAr-mis models that substantially underestimate the maximum rating, while the HAr model is able to predict plausible maximums when compared to the observed value. However, Figure 7 reveals that although the HAr-mis model fails to predict the overall maximum, it does predict reasonable maximum values for defenders and goalkeepers. Its failure to reproduce the maximums for the forwards and midfielders is explained by the rarity of their maximums (17 and 14, respectively) in the training data. Only one rating as high as 17 was observed in the first half of the season and there were only three ratings of at least 17 observed over the full season (about 1 in every 2000 observed ratings). To allow the HAr-mis model to predict such extreme values it may be possible to use a t-distribution instead of a Gaussian model, but for our purposes in this paper the ability of a model to replicate these very rare ratings is not so essential.

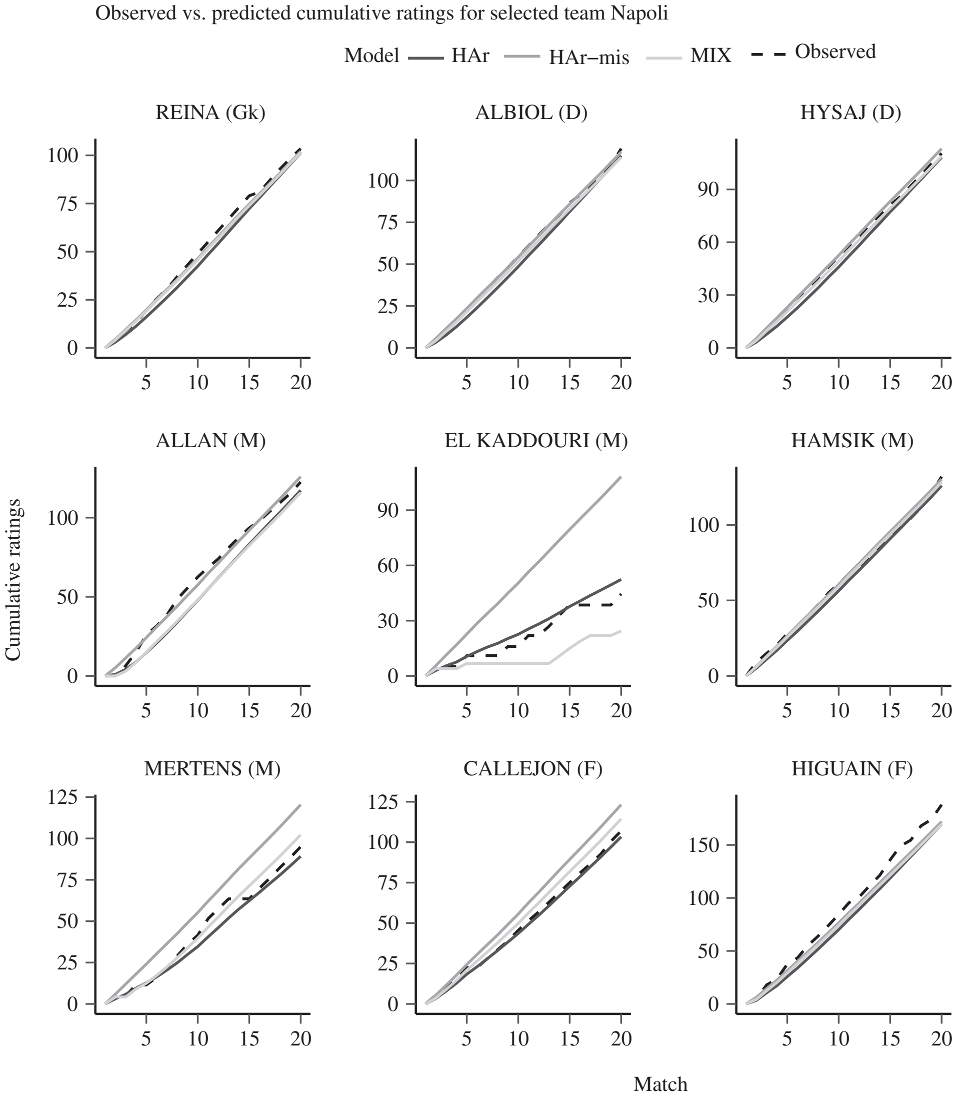

Figure 5: Observed vs. median predicted cumulative ratings for selected team Napoli during the first half of the 2015–2016 Serie A season.

5.2 In-sample and out-of-sample calibration

We are also interested in the calibration of the models on both the training and hold-out data. In Figures 8, 9, and 10 we display the median predictions and 50% posterior predictive intervals under the HAr, MIX and HAr-mis models for our selected team Napoli, overlaying the observed data points. In a broader analysis we could plot and analyze these graphs for each team in Serie A under each of the models.

In a well-calibrated model we expect half of the observed values to lie outside the corresponding 50% intervals. By this measure we can see in the plots that the HAr-mis and MIX model have decent but not excellent calibration, since for many of the players – particularly the goalkeeper and defenders – the 50% intervals cover more than 50% of the observed (blue) points. Conversely, for the volatile superstar Higuaín (an outlier even among forwards) only a many fewer points fall inside the intervals. Although the HAr model seems to be generally better calibrated, its main flaw consists in overestimating the defenders (and some other players) in the second part of the season, as already alluded to Section 4.2. Furthermore, the HAr model appears to identify an increasing trend in the ratings that is not actually supported by the data. As will be clear in Section 5.3, the out-of-sample predictions from the HAr model will in fact tend to be unreliable, while the MIX and the HAr-mis models tend to both better detect the best players *on average*.

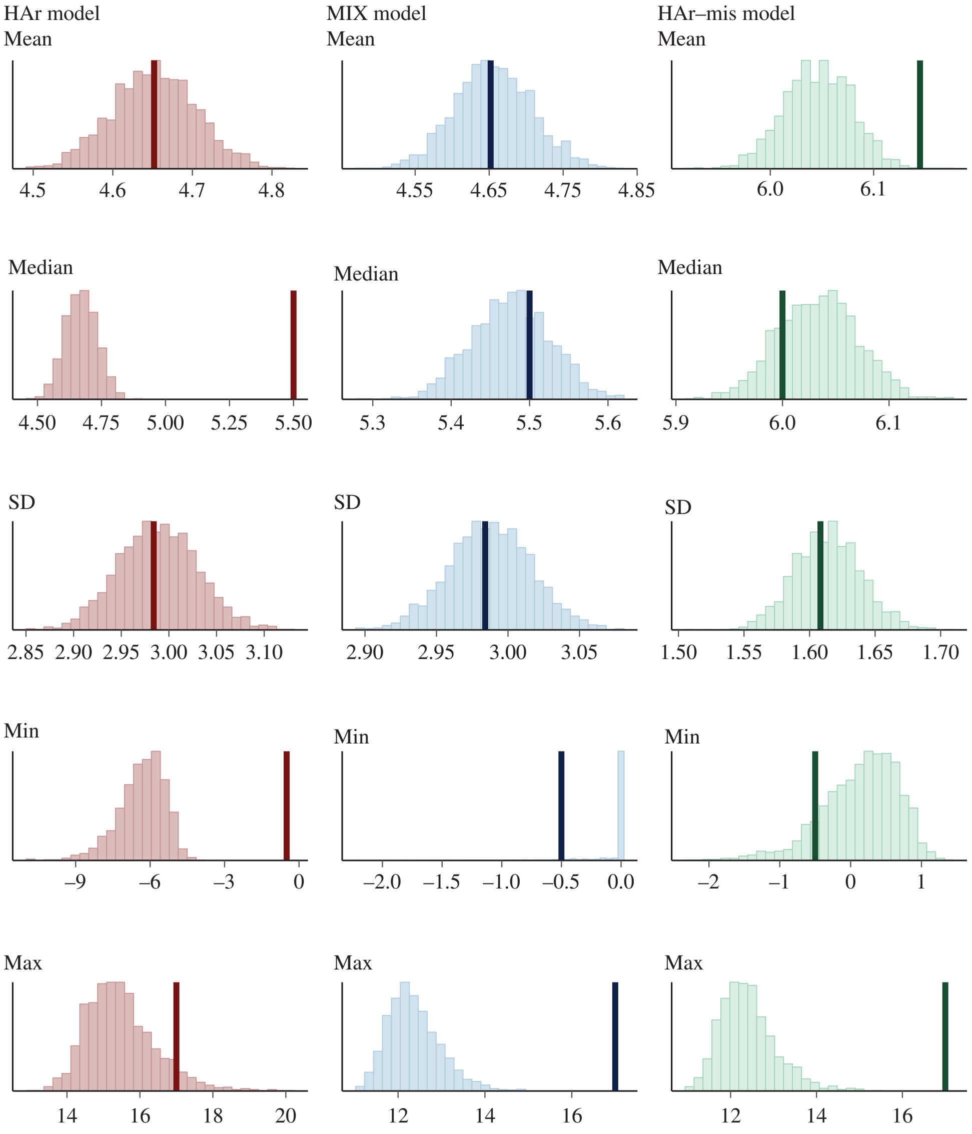

Figure 6: In-sample posterior predictive checks of test statistics for the HAr, MIX and HAr-mis models. For a particular test statistic T the plots show $T(y^{rep})$ (histogram) and $T(y)$ (thick vertical line). Each column corresponds to one of the three models, and each row to a different statistic T (mean, median, sd, minimum, maximum). We can see that the HAr model predicts much lower minimum values than the observed minimum. On the other hand, under the MIX model the distribution for the minimum is highly concentrated around zero.

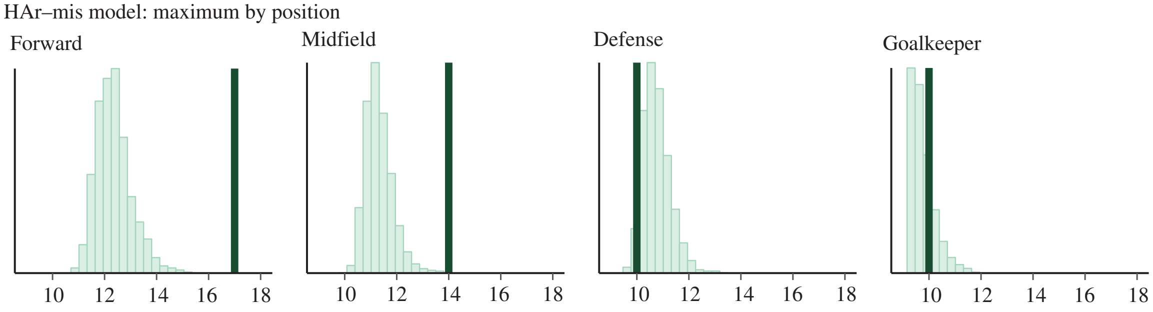

Figure 7: Posterior predictive check for $T(y) = \max$ over different positions for the HAr-mis model. The thick vertical line is the observed value.

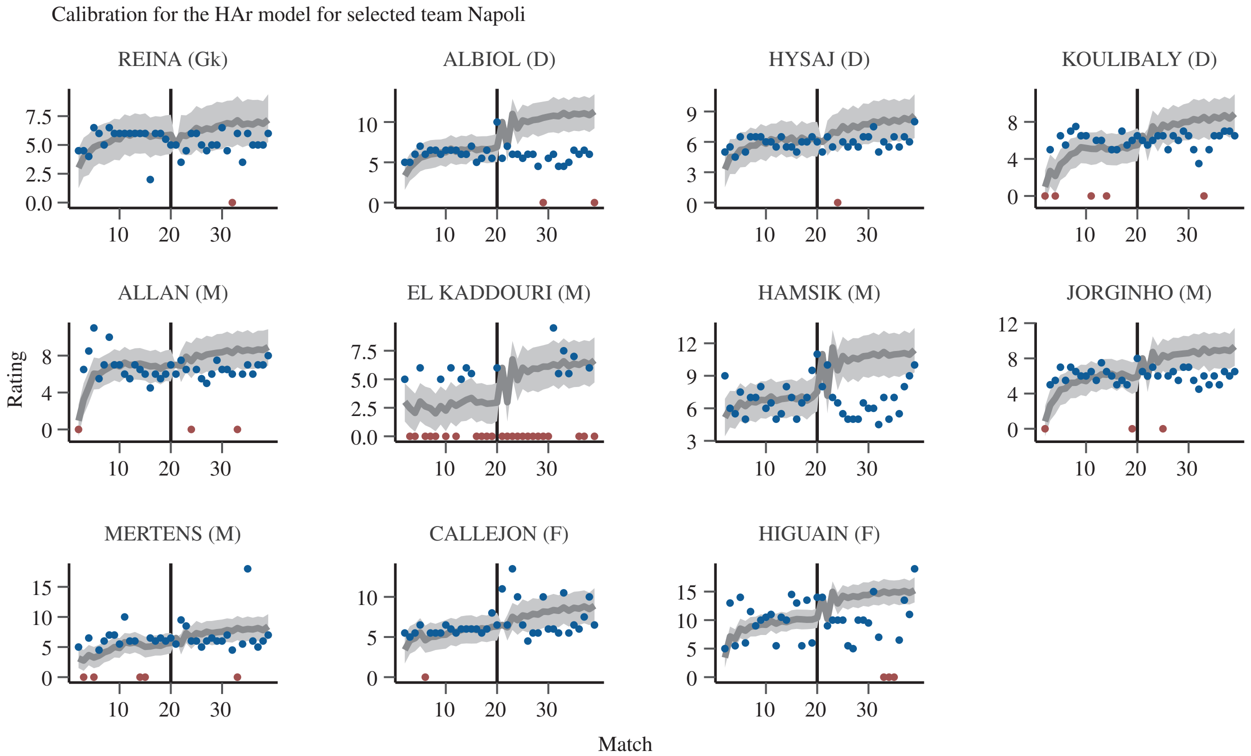

Figure 8: Calibration check for the HAr model for selected team Napoli. Blue points are observed values y^{obs}, red points are the zeros. The light gray ribbons represent 50% posterior predictive intervals and the overlaid dark gray lines are the median predictions. The vertical black lines separate the in-sample predictions from the out-of sample predictions.

5.3 Out-of-sample predictive checks

RMSE on hold-out data

For out-of-sample prediction we fit the models over the $T = 19$ matches in the first half of the season and then generate predictions for the $T^{\star} = 19$ matches in the second half of the season. For each player $i = 1, \ldots, N$ and for each posterior predictive simulation $s = 1, \ldots, S$ we compute the root mean square error (RMSE) over the matches $T + 1, \ldots, T + T^{\star}$ in the held out data (corresponding to matches 20 through 38 of the season),

$$\text{RMSE}_i^{(s)} = \sqrt{\frac{\sum_{t=T+1}^{T+T^{\star}} \left(\tilde{y}_{it}^{(s)} - y_{it}\right)^2}{T^{\star}}}. \qquad (9)$$

In the above equation $\tilde{y}_{it}^{(s)}$ is the sth simulation from the posterior predictive distribution of the predicted rating for player i at match t, and y_{it} is the corresponding observation. From this we obtain an RMSE *distribution* for each player.

Averaging over the simulations for each player and then averaging over players within positions we compute

$$\overline{\text{RMSE}}_j = \frac{\sum_{i=1}^{\#(i \in j)} S^{-1} \sum_{s=1}^{S} \text{RMSE}_i^{(s)}}{\#(i \in j)}, \quad j = 1, \ldots, J,$$

where $\#(i \in j)$ is the number of observations of position group j. Figure 11 shows these position-average RMSE values under each of the three models. The trend is similar across all models and suggests that our predictive ability is worst for forwards. Comparing across models, the missing data models (MIX and HAr-mis) do better than the HAr model, with the HAr-mis performing best. The (good) performance of this model is due to the fact that it does not predict future missing values. That is, the RMSEs computed above are then averaged over the missing values in the second part of the season. This plot is further confirmation that modeling the missing values is important for predictive accuracy on hold-out data.

It is worth noting that in a dynamic framework, where the models could be updated between matches, the RMSEs would almost certainly be much lower than the RMSEs computed for the second half of the season in one batch. For instance, fitting our models at time t and projecting for

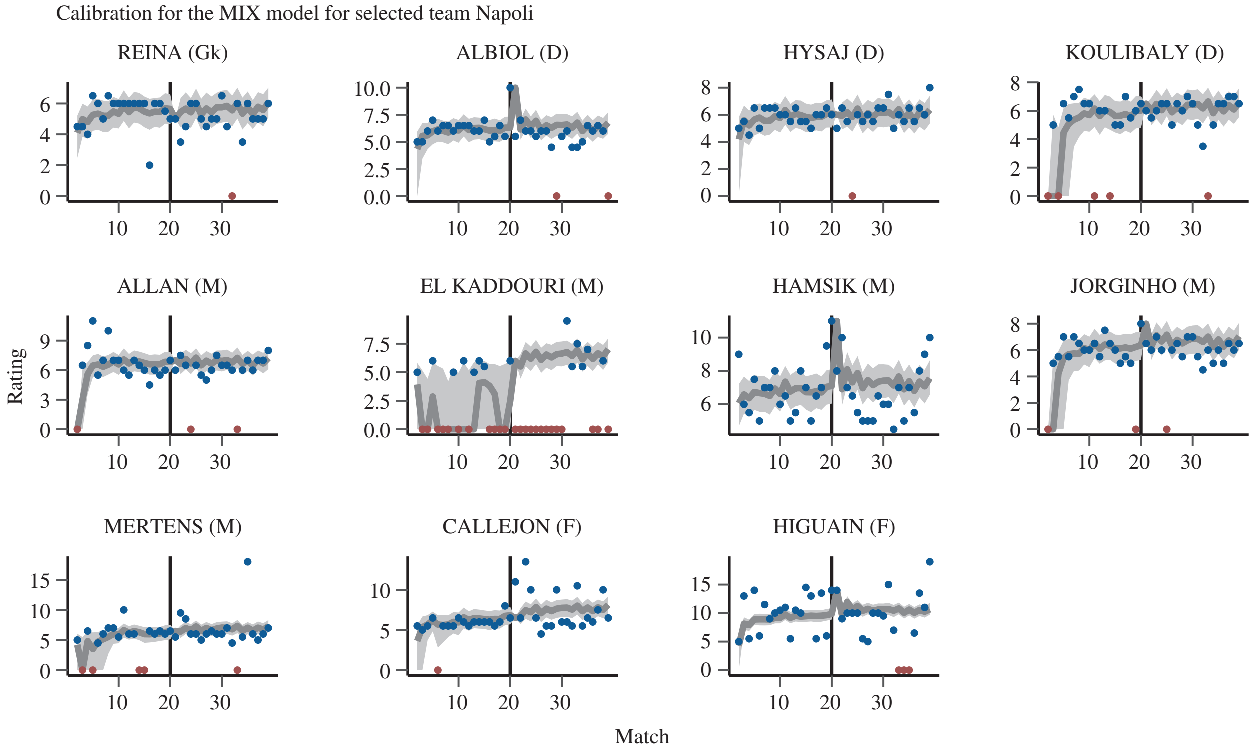

Figure 9: Calibration check for the MIX model for selected team Napoli. Blue points are observed values y^{obs}, red points are the missing values. The light gray ribbons represent 50% posterior predictive intervals and the overlaid dark gray lines are the median predictions. The vertical black lines separate the in-sample predictions from the out-of sample predictions.

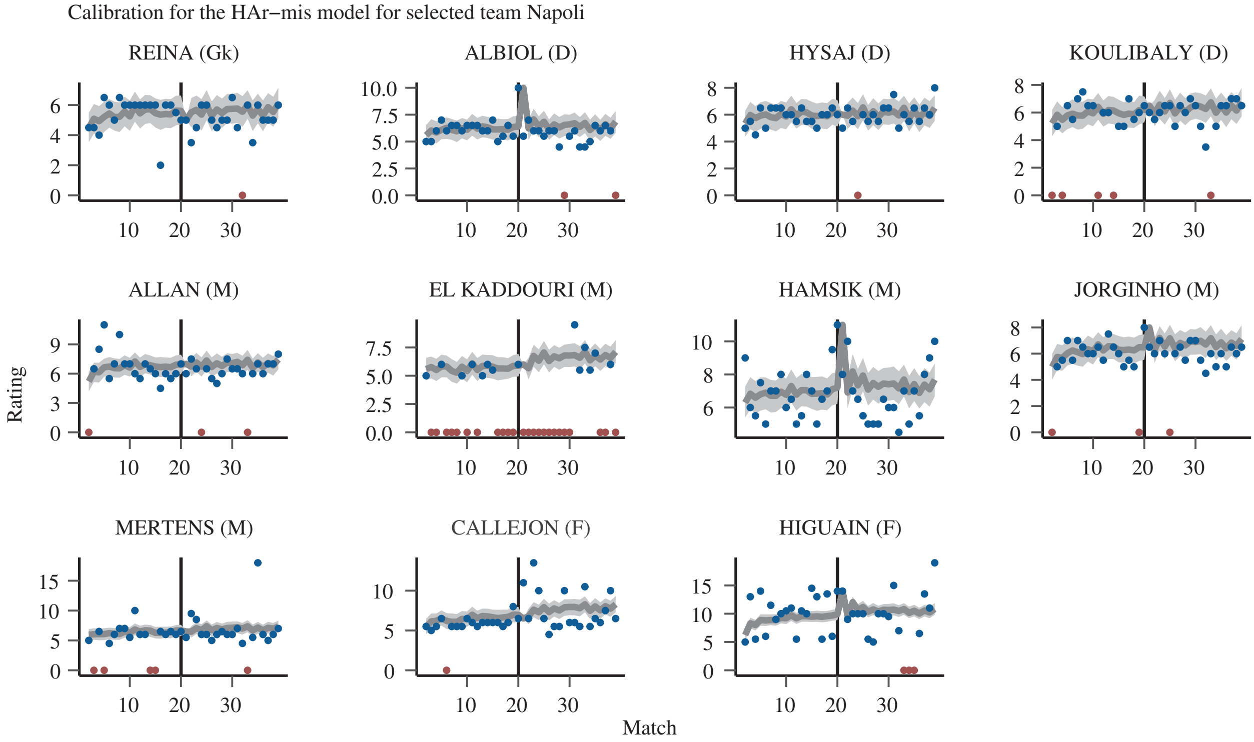

Figure 10: Calibration check for the HAr-mis model for selected team Napoli. Blue points are observed values y^{obs}, red points are the missing values. The light gray ribbons represent 50% posterior predictive intervals and the overlaid dark gray lines are the median predictions. The vertical black lines separate the in-sample predictions from the out-of sample predictions.

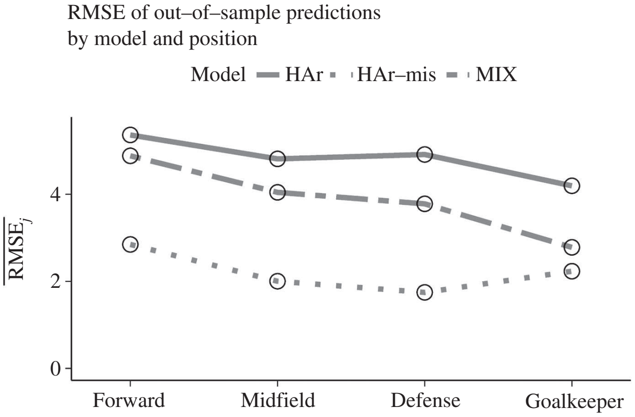

Figure 11: Average RMSE for the different positions for each model. The trend is the same across models: better predictions are obtained for goalkeepers, followed by defenders, midfielders, and finally forwards. The HAr-mis and MIX models register the lowest RMSE.

match $t+1$, we could account for the disqualification of certain players, injuries, etc. If we know in advance that a player is disqualified for the next match we would have $y_{i,t+1} = \tilde{y}_{i,t+1} = 0$, and the corresponding RMSE would be zero.

Roster selection

Based on average predicted ratings for the held-out data from the second half of the 2015–2016 Serie A season, Figure 12 displays the best teams of eleven players that can be assembled from the available players according to each of the models using their posterior medians. Also shown is the best team assembled using the observed ratings from the same set of matches. Here we assume that, in addition to a single goalkeeper, a team is comprised of four defenders, three midfielders, and three forwards. This is a common structure, although certainly many other formations are also used. As is evident at a first glance, the predictions obtained from the HAr model are quite inefficient. As we saw in the calibration plots in Figure 8, the HAr model tends to overestimate the player ratings, and we can see here that the projected ratings for the top players are quite far from their averages computed from the observed ratings in the hold-out data.

The rosters assembled based on the predictions from the HAr-mis and MIX models are identical except for the ordering of the players within the positions. Four of the eleven players (Acerbi, Pogba, Hamsik, Higuaín) from the team based on the actual ratings are included in the HAr-mis and MIX teams and, of the players that don't match, several are close. Dybala, the third best forward according to the models, is also rated highly (fifth best) according to the observed ratings. Rudiger, the second best defender according to both models, also has high observed mean rating (the eighth best among the 90 defenders). And Bonucci, one of the defenders included based on the observed ratings is also ranked highly by the HAr-mis model (ninth best) and MIX model (eleventh best).

Informally, this is further evidence that modeling the missingness allows us to obtain better out-of-sample predictions. Unlike the HAr model, the rosters selected by the MIX and the HAr-mis models appear to be quite competitive, which confirms the better performance we saw earlier in both the RMSE and the calibration comparisons.

6 Discussion

Although we are interested in our predictions for their own sake, our primary goal in this paper has been of an exploratory rather than confirmatory nature. Given the lack of published research on modeling this kind of data within a Bayesian framework, we hope our proposed models and process will be useful to other researchers interested in working on individual-level predictions in the presence of noisy soccer data.

We proposed various hierarchical models for predicting player ratings and fit them according to two different scenarios: in the first scenario the HAr treated the missing values as zeros; in the second scenario the MIX and the extended HAr-mis models allow for modeling the missing values themselves. We think the second framework is appealing in theory and we found in practice that the predictive performance is good both in-sample and out-of-sample. The HAr-mis and the MIX models yield similar posterior estimates, but they differ in their prediction ability, as suggested by the RMSE and calibration plots. The HAr-mis provides a simplicistic estimate conditioned on playing a given game, but it does not model the propensity to miss a game. We would suggest using the MIX model for practical purposes, since it naturally allows for the inclusion of more predictors and covariates associated with the probability of missing a game. Furthermore, there is not an appreciable loss of utility adopting the MIX model for assembling a good roster, which is the main task for each manager.

As expected, we found that a player's position is, in most cases, an important factor for predicting the Fantacalcio ratings. However, it is somewhat counterintuitive that the inferences from these models suggest that the quality of a player's team, the opposing team, and the initial fantasy price do not account for much of the

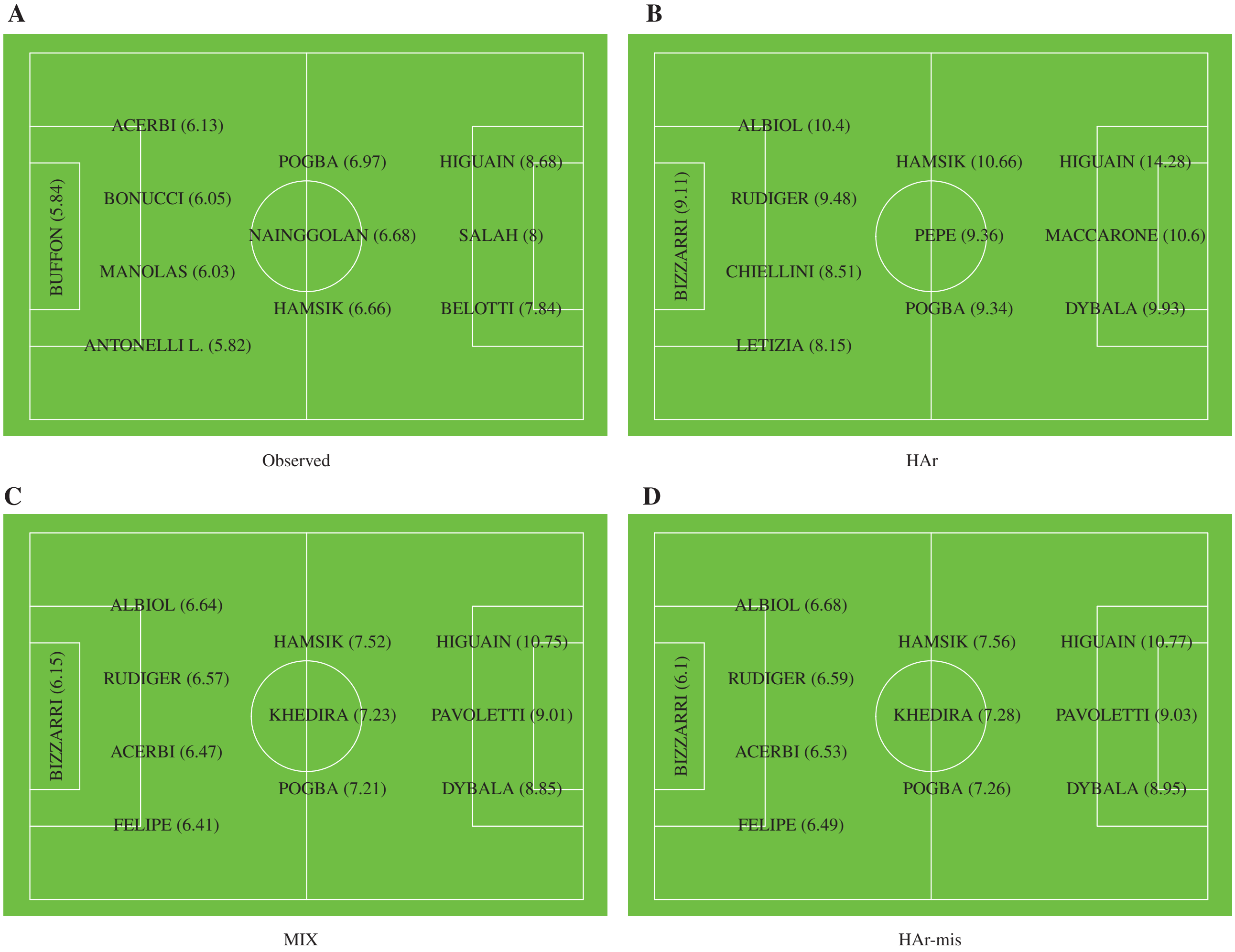

Figure 12: Best teams according to out-of-sample prediction of average player ratings for the HAr, MIX and HAr-mis model (Panels B, C, D) compared to the observed best team (Panel A) for the second part of the season. The averaged ratings are computed for those players who played at least 15 matches in the second half of the season.

variation in the ratings (net of the other variables). It is also notable that the association between the current and lagged performance ratings – expressed by the average lagged rating – is slightly different from zero after accounting for the other inputs into the models. Future research should consider whether other functional forms for describing associations over time are more appropriate, to what extent the inclusion of additional information in the models (e.g. injury data) improves the predictive performance, and if more informative priors can be developed at the position and team levels of the models. As is, the models may be over-shrinking these parameters. Another question to assess in the future is the division into training and testing datasets. In this paper we split the season in half, but these models should also be useful dynamically, using data available through match day t to predict rating for match day $t + 1$.

The recent successes in the soccer analytics industry are due in large part to the increasing number of available metrics for analyzing and describing the game. However, even as the amount and variety of publicly available soccer data grows – particularly data at the individual player/match level – the interpretability and predictive relevance of that data will remain a question. In fact, it is not straightforward to identify whether a player is or is not performing well – or collecting more point scores in the Fantacalcio framework – based on metrics such as the total distance run over the course of a match, the number (or percentage) of passes successfully completed, the total number of shots, the number of shots on target, or the number of “dangerous” attacks. According to our current knowledge, the only attempt at using these and many other metrics for measuring player performance is the OPTA index, which positively weights certain game features (e.g. goals, assists, shots, minutes) and negatively weights others (e.g. missed passes, yellow cards, missed goals, etc.). At least we are not aware of other attempts but we do not have proprietary information about what teams and other companies are doing (see www.optasports.com for further details about the firm and its activity). Despite its appeal, the weighting used for the index appears not to be formulated using statistical methodology and tools like principal component analysis, cluster analysis, or any kind of regression analysis.

Compared to attempts like the OPTA index, our ratings may be crude approximations to player performance since they gloss over many games events. But the formulation of an index based on as many variables as possible for describing the players' performances has not been the aim of this paper. The attractiveness of our approach – not necessarily all of our particular choices in model construction but our approach in general – is that it is based on a coherent statistical framework: we have an outcome variable y (the player rating) that is actually available, probability models relating the outcome to predictors, the ability to add prior information into an analysis in a principled way, and the ability to propagate our uncertainty into the predictions by drawing from the posterior predictive distribution. Our approach is also transparent, fits naturally into powerful statistical frameworks for model criticism (e.g. posterior predictive checking), and can easily be modified by anyone who has different ideas about the form of the relationship between the outcome and predictors.

References

Albert, J. 1992. "A Bayesian Analysis of a Poisson Random Effects Model for Home Run Hitters." *The American Statistician* 46:246–253.

Baio, G. and M. Blangiardo. 2010. "Bayesian Hierarchical Model for the Prediction of Football Results." *Journal of Applied Statistics* 37:253–264.

Becker, A. and X. A. Sun. 2016. "An Analytical Approach for Fantasy Football Draft and Lineup Management." *Journal of Quantitative Analysis in Sports* 12:17–30.

Bonomo, F., G. Durán, and J. Marenco. 2014. "Mathematical Programming as a Tool for Virtual Soccer Coaches: A Case Study of a Fantasy Sport Game." *International Transactions in Operational Research* 21:399–414.

Dixon, M. J. and S. G. Coles. 1997. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46:265–280.

Gelman, A. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian analysis* 1:515–534.

Gelman, A. 2016. "Prior Choice Recommendations Wiki !" URL http://andrewgelman.com/page/2/.

Gelman, A. and J. Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. Chapman & Hall/CRC, London.

Gelman, A., J. Hwang, and A. Vehtari. 2014. "Understanding Predictive Information Criteria for Bayesian Models." *Statistics and Computing* 24:997–1016.

Gottardo, R. and A. E. Raftery. 2008. "Markov Chain Monte Carlo with Mixtures of Mutually Singular Distributions." *Journal of Computational and Graphical Statistics* 17:949–975.

Karlis, D. and I. Ntzoufras. 2000. "On Modelling Soccer Data." *Student* 3:229–245.

Karlis, D. and I. Ntzoufras. 2009. "Bayesian Modelling of Football Outcomes: Using the Skellam's Distribution for the Goal Difference." *IMA Journal of Management Mathematics* 20:133–145.

Lomax, R. G. 2006. "Fantasy Sports: History, Game Types, and Research." Pp. 383–392 in *Handbook of Sports and Media*, editor by A. A. Raney and J. Bryant. Routledge, London, UK.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Silva, R. M. and T. B. Swartz. 2016. "Analysis of Substitution Times in Soccer." *Journal of Quantitative Analysis in Sports* 12:113–122.

Stan Development Team. 2016a. "RStan: The R Interface to Stan, version 2.14.1." URL http://mc-stan.org.

Stan Development Team. 2016b. "The Stan C++ library, version 2.14.0." URL http://mc-stan.org.

Stan Development Team. 2016c. *Stan Modeling Language User's Guide and Reference Manual, Version 2.14.0*. URL http://mc-stan.org/.

Thomas, A., S. L. Ventura, S. T. Jensen, and S. Ma. 2013. "Competing Process Hazard Function Models for Player Ratings in Ice Hockey." *The Annals of Applied Statistics* 7(3):1497–1524.

Vehtari, A., A. Gelman, and J. Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing* 27:1413–1432.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/jqas-2017-0066).