Modeling Time Series of Counts: an Application to Basketball Analytics

António Bernardo de Andrade Amorim Vieira

Instituto Superior Técnico, Lisboa, Portugal

April 2019

Abstract

Basketball statistics has gained relevance over the past 10 years since it provides essential information for team management and coaching decisions. Moreover, the rising importance of the three pointer in the NBA makes it a topic worthy of study. Generally, teams try to mimic strategies deployed by successful peers in order to achieve good results. In that sense, the Houston Rockets, a very successful team in the period under analysis (2012-2017), three pointers time series are compared with those of the Los Angeles Lakers team, a team that struggled in the same span. Recurring to the theory of time series of counts, this study aims at understanding the evolution of the three point time series of the above-mentioned teams. Before modeling, the data-set was carefully analyzed, selecting the most relevant variables and the order of the model applied. This selection is carried out by taking into account the AIC and BIC. The final models selected fit reasonably well, although, both showed an inability to identify downward spikes on the time series. Furthermore, the predictions performed using the models had good quality and only differ marginally from the original three pointer time series.
Keywords: time series of counts, variable selection, order selection, INGARCH, tscount package

1. Introduction

The National Basketball Association (NBA) is the USA Men's major basketball league. The NBA generates over 7 Billion USD a year [15], highlighting how it is a business and not just entertainment. The value of each team in the NBA depends on its location, the roster and, of course, their success.

A management team should then be focused in creating value for its company hence, looking for the right players to sign and the correct culture to instill. Moreover, the location of each franchise can be beneficial although it should not deter management from looking for the other two factors. Throughout the history of the NBA, the teams that won the championships and were successful manage to anticipate trends or set them. In order to do so, it is imperative to understand the team trajectory statistically. By doing so a team is able to design drafting, signing and managing strategies that can help a team win.

The rawest basketball statistics are counts and time-dependent, so a time series-based approach seems the most appropriate for its analysis. Regarding basketball analytics, studies like the one by Hollinger [1] contributed to raising interest in sports analytics and result forecasting. Furthermore, more advanced metrics where created and discussed; see for example Kubatko et al. [2] and Shea et al. [3] and the references therein. As stressed above, forecasting the evolution of three point shooting has become, nowadays, an issue of major importance. This can be witnessed by studies like the one done by Goldman and Rao [4]. With the evolution in technology and in the methodology for sports analytics, forecasting becomes fundamental for coaching and management. The objective is provide facts to support decisions and strategies.

A suitable approach for time series of counts are generalized linear models (GLM), where the serial dependence is incorporated through a link function. This leads to a flexible class of models where, e.g., covariates are easily incorporated, enabling the models to better explain the dynamics of the available information and to provide trustworthy predictions. The thesis will focus on the application of integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) models to characterize the evolution of basketball statistics. To our knowledge, such models have not been previously applied in the field of sports statistics.

A Generalized Linear Model (GLM)-based model is advantageous when compared to the one based on thinning operators since it can easily incorporate covariate effects and negative correlations. Moreover, recent work done by Liboschik et al.

[5] produced a package in the R [6], the $tscount$ package, that enables the user to module, by using likelihood-based estimation methods, count time series following GLM. This software allows for model fitting and assessment, prediction and intervention analysis. Furthermore, the package also allows the analysis of a more general dependent structure instead of the ones obtained with the standard GLM procedures. Following a GLM framework, the resulting models on the package have multiple parameters between covariates, past observation and latent means. To estimate such parameters, and following the work of Liboschik et al. [5], the Quasi Conditional Maximum Likelihood (QML) estimation is used. The QML method is essentially the same as the Maximum Likelihood (ML)maximum likelihood method commonly used in statistics.

The statistics for a team are defined in table 1. To make the distinction of team statistics and the opposing team statistics, a prefix "O-" will be added to each statistical variable relative to the opposing team.

Table 1: Basketball Statistics Definitions

Statistic	Definition
FG	Number baskets scored by a team
FGA	Number of Baskets attempted by a team
FG%	Ration between made baskets and attempts
3P	Number of FG made beyond the 3P line
3PA	Number of three pointers attempted
3P%	Ratio between 3P made and attempted
FT	Number of fouls shots made
FTA	Number of Foul shots attempted
FT%	Ratio between FT made and attempted
ORB	Num. of retrieved balls after missed FG/FT
TRB	Num. of retrieved balls after missed FG/FT
AST	Num. of passes that lead to a made FG
STL	Number of recovered possessions
BLK	Number of deflected FGA that impede scoring
TOV	Number of lost possessions before any FGA
PF	Number of infractions of game rules

This paper comprises five chapters. The Introduction contains the project objectives and its surroundings, as well as a review of the existing literature on the research area. In Chapter 2, background results connected with the problem at hand are introduced and explained in detail. Also, the models and their specifications are described so that the study of the database might be understood. The third chapter extensively describes the database as well as the variable at study. Chapter 4 is for the application of the methodology and the statistical software to the database. Also, the summary of the results is obtained and discussed. Finally, Chapter 5 is devoted to conclusions and to discuss some clues for future research.

2. Methodology

2.1. Models for time series of counts

A time series of counts commonly exhibits serial dependence and over-dispersion. As stressed before, the GLM methodology [7] can be applied to cope with these features. Firstly, we denote the count time series by $\{Y_t : t \in \mathbb{N}\}$, by $\{X_t : t \in \mathbb{N}\}$ a time-varying r-dimensional covariate vector, say $X_t = (X_{t,1}, \ldots, X_{t,r})^T$, and $\mathcal{F}_{t-1}$ the filter containing past observations. This r-dimensional vector will incorporate the relevant covariates. Given the randomness of the variables involved, the conditional mean, $E(Y_t|\mathcal{F}_{t-1})$, can be modeled using a stochastic process, with a latent mean process, $\{\lambda_t : t \in \mathbb{N}\}$, such that $E(Y_t|\mathcal{F}_{t-1}) = \lambda_t$. This shapes the model into the following form :

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-j_l}) + \eta^T X_t, \tag{1}$$

where β_k are the autoregressive (AR) parameters, $\eta = (\eta^1, \ldots, \eta^r)^T$ corresponds to the effects of covariates, $g : \mathbb{R}^+ \to \mathbb{R}$ is a link function, $\tilde{g} : \mathbb{N}_0 \to \mathbb{R}$ is a transformation function and $g(\lambda_t)$ is the linear predictor. It is also defined a set Q, as $Q = j_1, \ldots, j_q$, and set a P, as $P = i_1, \ldots, i_p$ for the regression on lagged conditional means $\lambda_{t-j_1}, \ldots, \lambda_{t-j_q}$, where $0 < j_q < \infty$ and $q \in \mathbb{N}_0$, and past observations $Y_{t-i_1}, \ldots, Y_{t-i_p}$, where $0 < i_p < \infty$ and $p \in \mathbb{N}_0$.

The correct choice of link and transformation function will determine the effectiveness of the model fitting. The parameter vector η corresponds to the effects of the selected covariates on the time series. Note that, the set P is created to help the regression over the past observations $Y_{t-i_1}, Y_{t-i_2}, \ldots, Y_{t-i_p}$ and set Q is created to help the regression over past conditional means $\lambda_{t-j_1}, \lambda_{t-j_2}, \ldots, \lambda_{t-j_q}$. The multiplicity of variants of the model, given the choices in link and transformation function, allows for a global application. As a first example, the chosen link function is the identity. The transformation comprises that $g(x) = \tilde{g}(x) = x$. If the effect of covariates is excluded and the assumed distribution is the Poisson the model is defined as :

$$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k} + \sum_{l=1}^{q} \alpha_l \lambda_{t-l}, \tag{2}$$

From (2), it follows that if Y_t, given past, is Poisson-distributed then we obtain the Integer-valued Generalized Autoregressive Conditional Heteroscedasticity (INGARCH) model of order (p,q). These class of models have been studied by Heinen [8], Ferland et al. [9] and Fokianos et al. [10] among others.

Following the model in (2) and by taking the logarithmic function instead of the identity, the model set-up would change considerably. By changing the link function then $g(x) = \log(x)$, $\tilde{g}(x) = \log(x + 1)$ and $\upsilon_t = \log(\lambda_t)$, the resulting model is:

$$\upsilon_t = \beta_0 + \sum_{k=1}^{p} \beta_k \log(Y_{t-k} + 1) + \sum_{l=1}^{q} \alpha_l \upsilon_{t-l}, \tag{3}$$

The case $\tilde{g} = \log(x+1)$ was extensively studied by Fokianos and Tjøstheim [11]. This model properly integrates covariates since INGARCH models (2) require the positivity of the conditional mean process, λ_t thus, neglecting the downward effects of covariates. The log-linear model is an alternative that can capture effects of covariates better than the INGARCH, although, the first one has a multiplicative effect on the response variable and the second an additive effect. The conditional distribution has an impact on the model although not in the same way as the link function. If the assumed distribution is Poisson then $Y_t \mid \mathcal{F}_{t-1} \sim Poisson(\lambda_t)$. An alternative to the Poisson distribution is the negative binomial distribution. Such assumption would allow for the conditional variance to be larger than the mean, which is a common definition for over-dispersion. Since the negative binomial distribution belongs to the class of mixed Poisson distribution, the Poisson is a limiting case of the negative binomial when ϕ tends to infinity. In that case, the conditional variance and the mean will be the same, which, in turn, implies the nonexistence of over-dispersion.

An important feature on the GLM-based models is the integration of covariates. The weight of each covariate has is defined by the regression on past observation and past conditional means. The effect of such covariates can be seen as an internal influence on the data-generating process. An internal covariate effect is felt throughout the process whereas the external is only felt in a certain point. Factoring this in the equation, the model outlook would change. Let $e = (e_1, \ldots, e_r)^T$ be a vector with $e_i = 1$ if the i-th component of the covariate vector has an external effect and $e_i = 0$ otherwise, for $i = 1, \ldots, r$. Now, denoting by $diag(e)$ a diagonal matrix with diagonal elements given by e the model outlook is:

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l \Big(g(\lambda_t - j_l) - \eta^T diag(e) X_{t-j_l} \Big) + \eta^T X_t. \tag{4}$$

2.2. Parameter Estimation

The model parameters coefficients are obtained using the QML estimation. The QML method is more general and more flexible than the conventional ML method since avoids specifying the density function for the conditional mean. The QML estimate is obtained by maximizing a function related to the logarithmic likelihood function instead of maximizing the general log-likelihood function used in the ML method. With respect to the log-linear model 3 , the parameter space defined for it changes slightly in comparison to the parameter space for the INGARCH model since the first one can now integrate negative values. The parameter space is defined as :

$$\Theta = \Big\{ \theta \in \mathbb{R}^{p+q+r+1} : |\beta_1|, ..., |\beta_p|, |\alpha_1|, ..., |\alpha_p|, |\eta_1|, ..., |\eta_r| < 1, \Big| \sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j \Big| < 1 \Big\}. \tag{5}$$

This parameter space also guarantees that the solution of the process is ergodic and stationary. Christou and Fokianos [12] pointed out that the estimation using the negative binomial assumption of the parameters described above does not depend on the dispersion parameter, ϕ, of the distribution. This enables the use of the QML method when the Poisson is assumed. The dispersion parameter is then estimated separately, for the negative binomial assumption, which is different from a GLM estimation. The best estimators are obtained when there is minimal loss of information. The log-likelihood function, score vector and information matrix are derived conditionally on presample values and on the conditional mean process λ_t at moment zero. The likelihood function and the quasi-maximization of this function would determine the strength of the model. Given a vector of observations $y = (y_1, \ldots, y_n)$, the conditional quasi log-likelihood function, up to a constant, is given by :

$$l(\theta) := \sum_{t=1}^{n} log p_t(y_t; \theta) \simeq \sum_{t=1}^{n} \big(y_t ln(\lambda_t(\theta)) - \lambda_t(\theta) \big), \tag{6}$$

where $p_t(y_t; \theta) := P(Y_t = y | \mathcal{F}_{t-1})$ is the probability density function of a Poisson distribution. The reasoning behind (6) is the use of the log-likelihood function of a misspecified model. The function cannot be oversimplified when maximized in order to maintain the consistency and asymptotically normality of the estimate. This estimate, if the function is designed in order to minimize the loss of information relative to the actual likelihood, only differs slightly to the ML estimate. In that sense, the QML estimator $\tilde{\theta}$ of θ will be the argument that maximizes the log-likelihood function and solves the optimization problem. Therefore,

$$\hat{\theta} := \hat{\theta_n} = arg \max_{\theta \in \Theta} l(\theta). \tag{7}$$

Inspired by the Pearson's $\mathcal{X}^2$-statistic, and, given that $\hat{\lambda}_t$ is the mean of the fitted values resulting from the solution described previously, the disper-

sion parameter ϕ for the negative binomial distribution is estimated by solving the equation:

$$\sum_{t=1}^{n} \frac{(Y_t - \hat{\lambda}_t)^2}{\hat{\lambda}_t + \frac{\lambda_t^2}{\hat{\phi}}} = n - (p + q + r + 1). \quad (8)$$

The variance parameter, σ^2, is estimated as $\hat{\sigma}^2 = \frac{1}{\hat{\phi}}$. The studies by Douc et al. [13] and Sim [14] dissect the use of this estimator for the log-linear case 3. For processes that involve covariates the ideas is:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_{p+q+r+1}\big(0, G_n^{-1}(\hat{\theta}_n, \hat{\sigma}^2) \; G_n^{-1}(\hat{\theta}_n, 0) G_n^{-1}(\hat{\theta}_n, \hat{\sigma}^2)\big), \quad (9)$$

as θ_0, for any n value, represents the true values for the parameters and $\hat{\sigma}^2$ is a consistent estimator for σ^2. This only applies under a couple of assumptions. For deterministic covariates, the assumptions are that the $||X_t|| < c$, where $|| \bullet ||$ denotes the Euclidean norm, the covariate process is bounded and the $lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} X_t X_t^T = A$, where c is a constant and A is a non-singular matrix. For stochastic covariates the assumptions are that $E(X_t)$ and $E(X_t X_t^T)$ exists and $E(X_t X_t^T)$ is a non-singular matrix. All the assumptions for the different types of covariates imply that the information on each covariate grows linearly with the sample size, and, also, the covariates are not linearly dependent.

3. Exploratory Analysis

3.1. Relationship of variables

Identifying the most important variables and understanding their relationships improves the quality of the study.

Table 2: Correlation Table for the Lakers

	FG	FGA	3PA	TRB	ORB	AST	BLK	TOV	OORB	OTOV
FG	1	0.386	0.108	−0.057	−0.015	0.621	0.089	−0.082	0.061	0.171
FGA	0.386	1	0.216	0.308	0.493	0.127	0.027	−0.308	−0.061	0.249
3PA	0.108	0.216	1	0.056	0.044	0.221	0.069	−0.0003	0.030	0.070
TRB	−0.057	0.308	0.056	1	0.579	−0.054	0.081	0.085	−0.145	0.150
ORB	0.015	0.493	0.044	0.579	1	0.175	0.001	0.011	0.114	0.106
AST	0.621	0.127	0.221	0.054	0.175	1	0.186	0.021	0.117	0.083
BLK	0.089	0.027	0.069	0.081	0.001	0.186	1	0.014	0.364	0.011
TOV	0.082	0.308	0.0003	0.085	0.011	0.021	0.014	1	0.067	0.239
OORB	0.061	0.061	0.030	0.145	0.114	0.117	0.364	0.067	1	0.089
OTOV	0.171	0.249	0.070	0.150	0.106	0.083	0.011	0.239	0.089	1

Using the Pearson correlation, for the Lakers (see table 2), a couple of points need to be made. Regarding the FG variable, the AST have the highest correlation with it, at 0.621. It is a testament to the importance of ball movement and passing. There is also a positive correlation with the FGA, of about 0.386, since the increase in tries implies an increase in makes. In turn, the FGA is correlated with the amount of TRB (0.308) and ORB (0.493). The positive correlation is understandable since an offensive rebound represents an extra possessions. The $OTOV$ is also positively correlated with the attempts (0.249). The rebounding variables, TRB and ORB, are heavily correlated with each other as the offensive rebounds are a part of the total rebounds. The blocks, BLK, are positively correlated with the opposition offensive rebounds, at about 0.364. The turnovers are negatively correlated with the number of field goal attempts (-0.308) which makes sense since it leads to a loss of possession.

Table 3: Correlation Matrix for the Rockets

	FG	FGA	3PA	TRB	ORB	AST	BLK	TOV	OORB	OTOV
FG	1	0.471	0.139	0.076	0.031	0.714	0.058	0.182	0.028	0.090
FGA	0.471	1	0.449	0.288	0.453	0.299	0.0003	0.328	0.056	0.200
3PA	0.139	0.449	1	0.015	0.090	0.348	0.078	0.136	0.004	0.154
TRB	0.076	0.288	0.015	1	0.575	0.020	0.146	0.144	0.035	0.165
ORB	0.031	0.453	0.090	0.575	1	0.086	0.036	0.006	0.010	0.039
AST	0.714	0.299	0.348	0.020	0.086	1	0.052	0.091	0.031	0.037
BLK	0.058	0.0003	0.078	0.146	0.036	0.052	1	0.025	0.263	0.017
TOV	0.182	0.328	0.136	0.144	0.006	0.091	0.025	1	0.073	0.226
OORB	0.028	0.056	0.004	0.035	0.010	0.031	0.263	0.073	1	0.029
OTOV	0.090	0.200	0.154	0.165	0.039	0.037	0.017	0.226	0.029	1

Regarding the Rockets' correlation matrix (see table3), the FG variable is highly correlated with AST, at 0.714, and also with the number of FGA in the game. This resembles the Lakers case although, the Rockets FG presents a higher correlation for both variables. One explanation would be the the tendency for ball movement and passing. The FGA is positively correlated with the amount of FG (0.471), $3PA$ (0.449), ORB (0.463) and AST (0.299). The first two, since they are a part of the calculation of total attempts, are trivially correlated. Similar to the Lakers, the ORB are correlated with the FGA as they represent extra possessions. This statistic is also negatively correlated with turnovers (-0.328) since losing possession of the ball prevents attempts at the basket. The ORB are also correlated to the FGA (0.453) and the total number of rebounds (0.575). Lastly, the own team's turnovers are negatively correlated with the team's field goals attempts (-0.328) and positively correlated with the opposition number of turnovers.

3.2. Time series analysis

The three point statistic, $3P$, is the target time series under analysis. The summary statistic for $3P$ is:

Table 4: Summary Statistics

Variable	Lakers	Rockets
n	410	410
mean	8.27	11.31
variance	10.50	15.13
median	8	11
trimmed	8.10	11.11
mad	2.97	4.45
min	1	3
max	19	24
skew	0.53	0.48
kurt	0.04	-0.17
se	0.16	0.19

The table above (4) shows that the Rockets present a significant higher average of $3P$ scored and standard deviation, suggesting the values are more dispersed compared to the Lakers. The next step is to analyze the time series plots. The frequency for the $3P$ time series was 82 observations,

equaling the number of games per team per season. An NBA season starts in October and ends near June of the following year. There is a short break during the season between the 40th and 50th game, due to the All-Star Game, which will be relevant when analyzing the time series. For the Lakers:

Figure 1: Lakers time series of counts

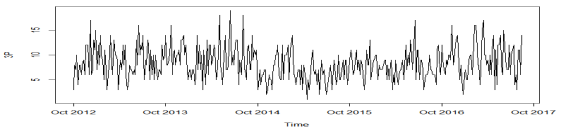
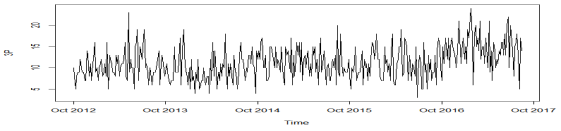


Figure 1 shows no consistent trend from season to season nevertheless, within each year, there are some in-season trends. In the first season, the Lakers changed coaches twice before hiring Mike D'Antoni for the rest of the season which, explains the low values in the first games of the first season and, its subsequent rise since the hiring, given the coach is known for valuing the three point shot.At the start of the 3^{rd} seasonthe coach was fired and replaced and the Lakers experienced the lowest values of $3P$, specially before February 2015. Afterward, the team recovered steadily although in 4^{th} season the $3P$ stagnated and the trend ended. This disparity in values from season to season can also be seen in the passage from the 2^{nd} to the 3^{rd} season. This value disrupton might be due to: roster changes, long term-injuries and, coaching changes.

It is worth noting that there are several spikes in the last thirds of all seasons. These spikes coincide with the All-Star Break that provides the players almost a two week rest period. During it the teams perfect their strategy and rest their players which, in turn, improves the team's performance and thus their statistics totals in the short-term. Notwithstanding, the team performance is unsustainable since fatigue eventually kicks back in and the performance recedes.

Figure 2: Rockets time series of counts



Regarding the Rockets (Figure 2), there are noticeable differences in comparison to the Lakers. The scales and the several spikes along the plot suggests that the data in the Rockets case is more dispersed, which is verified by the standard deviation. The first season marks the arrival of the star player James Harden to the Rockets. The team went through an adaptation period that is noticeable as there is a slight trend upwards in the first part of the first season. In the second season, the same happened with Dwight Howard. The difference was the consistency of the values around the mean since, in the latter part of the second season the values are not as dispersed. Moreover, the values were consistent from the 2^{nd} season to the 3^{rd} season. The 3^{rd} season was the most consistent given the little deviation from the mean and, the nonexistence of high-valued spikes. This coincides with the highest winning season for the Rockets (56 wins) and the lowest roster turnover. In the fourth season, the coach was fired and replaced and although the season started with an apparent consistency in the values, there was a significant downward trend post All-star. Afterward, the Rockets hired coach Mike D'Antoni which prompted a clear upward trend from the start of the fifth season onward. The mean for this season is noticeably higher than the other seasons and the maximum value of $3P$ scored is attained in this stretch in time.

3.3. Model Order Selection

The model order is obtained by fitting a simpler model with no external covariates effect, as in model (2) and (3). Recurring to iterations and using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as a selection criteria several possible models are evaluated. The lower those criterion are the better the model will be.

For the Lakers and after several iterations it could be seen that the models with a logarithmic link function (3) have always a better score in both criteria. The reduction on the AR and moving average (MA) parameters was more denounced in the BIC since it penalizes over-parameterized models. Moreover, the negative binomial assumption for the distribution yields a better score in the AIC in every iteration. On the other hand, the Poisson yields better results in the BIC in the majority of iterations although the difference in the remaining scores is minimal. The AIC is prone to favor bigger, over-parameterized models whereas the BIC is prone to favor overly simple models. Nonetheless, given the data-set size ($n \geq 400$) the BIC is considered reliable for decision criteria.

Depending on link function chosen, the best model for the Lakers can be represented in the following two forms. For a logarithmic link function, let $\upsilon_t = log(\lambda_t)$ then the model will have the form:

$$\upsilon_t = \beta_0 + \beta_1 \log(Y_{\text{t-1}} + 1) + \alpha_l \upsilon_{t-1}, \qquad (10)$$

whereas for a identity link function the model would

look like:

$$\lambda_t = \beta_0 + \beta_1 Y_{\text{t-1}} + \alpha_l \lambda_{t-1}, \qquad (11)$$

where $Y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t)$ or $Y_t|\mathcal{F}_{t-1} \sim NegBin(\lambda_t, \phi)$ depending on the assumption for the conditional distribution is used.

Similarly to the Lakers case, the addition of parameters to the initial model resulted in an increase in the values of the criterion. The simpler the iterations the better the scores for the BIC and AIC were.

Let $\upsilon_t = log(\lambda_t)$ then the best base model for the rockets will have the form:

$$\upsilon_t = \beta_0 + \sum_{k=1}^{2} \beta_k \log(Y_{\text{t-k}} + 1) + \alpha_1 \upsilon_{t-1}, \qquad (12)$$

whereas for the log-linear model variation. For the identity link function the model would be:

$$\lambda_t = \beta_0 + \sum_{k=1}^{2} \beta_k Y_{\text{t-k}} + \alpha_1 \lambda_{t-1}, \qquad (13)$$

where $Y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t)$ or $Y_t|\mathcal{F}_{t-1} \sim NegBin(\lambda_t, \phi)$ depending on the conditional distribution considered.

3.4. Variable selection

The criteria for variable selection will be similar to the model order selection. The iteration process for variable selection will resemble a forward induction. Firstly, the $3P$ correlation with the relevant statistics will be checked.

Table 5: Correlation with 3-pointers for the Lakers

	Lakers	Rockets
FG	0.333	0.409
FGA	0.015	0.218
3PA	0.693	0.705
3P%	0.772	0.688
TRB	0.096	0.062
ORB	0.125	0.083
AST	0.451	0.569
BLK	0.134	0.014
TOV	0.073	0.102

By table (5) it is important to point out the high correlation of $3PA$ with $3P$, at about 0.693 for the Lakers and 0.705 for the Rockets. The conversion rate, $3P\%$, is calculated by $\frac{3P}{3PA}$, has a high correlation (0.772 for the Lakers and 0.688 for the Rockets) with the $3P$ variable and is a great indicator efficiency. However, due to multicollinearity problems both $3PA$ and $3P\%$ will be excluded from any regression. The FG and $3P$ have also a high correlation (0.333 for the Lakers and 0.409 for the Rockets). The AST variable has a strong connection to FG following the correlation tables (2,3). Naturally then, the AST have also strong correlation with $3P$ at about 0.451 for the Lakers and 0.569 for the Rockets. It is also worth mentioning the small positive correlation between $3P$ and BLK for the Lakers (0.134) which suggests that a good defense can translate into a good offense. Finally, there is a small negative correlation between the TOV and $3P$ as the loss of possessions decreases the possibilities of attempting and scoring a three pointer. For model fitting, the variables to be included and tested in the base model will be the ones with higher correlation, namely AST, FG, BLK and ORB for the Lakers and FG, FGA, AST, TOV for the Rockets.

Following the correlation analysis and using an iteration process the variables are introduced in their respective base models individually at first. For the Lakers case, the starting point is a log-linear (1,1) or a INGARCH (1,1). The best scoring iteration would determine the first variable chosen. And then the process would repeat until all combinations are evaluated. For the Lakers, the model will include the AST and FG variables.

In the Rockets case, the starting point was a log-linear (2,1) or an INGARCH (2,1). After all the iterations were evaluated it was noted that the more complex models had worse results than the simpler models. As such, the AST was the only covariate integrated.

3.5. Link function and conditional distribution

The conditional distribution for the response variable can be a Poisson or a Negative Binomial distribution. Using the PIT histogram the choice is performed. For the Lakers, the Poisson and the Negative Binomial assumption presented similar scores. The only difference came through the choice in the link function. The reason is the fact that the dispersion parameter of the negative binomial distribution could not be estimated. In this case, $Y_t|\mathcal{F}_{t-1} \sim NegBin(\lambda_t, 0)$ since $\phi = 0$. Therefore, in the Lakers case the assumed distribution will be the Poisson, $Y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t)$. The PIT histograms for both link functions show no signs of over-dispersion nor under-dispersion as they are all close to uniformity.

Figure 3: Lakers PIT histogram with logarithmic link function

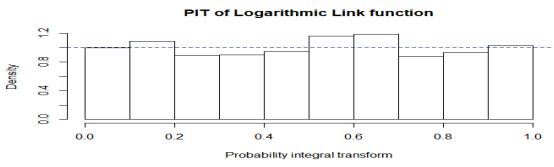
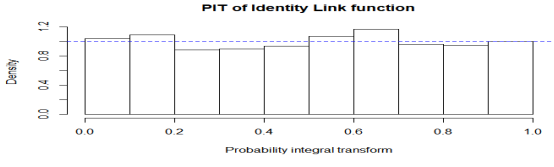


Figure 4: Lakers PIT histogram with identity link function



For the Rockets, the dispersion parameter for the negative binomial distribution was almost surely zero and thus, the assumed conditional distribution is a Poisson. To check the presence of over or under-dispersion the PIT histograms are used.

Figure 5: Rockets PIT histogram with logarithmic link function

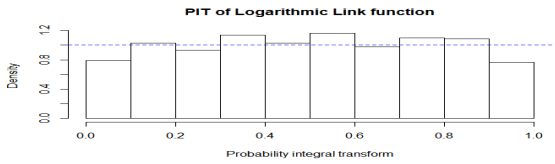
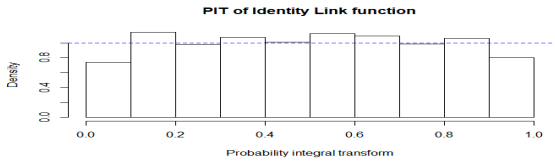


Figure 6: Rockets PIT histogram with identity link function



Both PIT histograms show some signs of over-dispersion since they have a upside down U-shape. The disparity in the scales shows that the logarithmic link function provides a slightly better result although the conditional distribution might not follow a true Poisson distribution. Therefore, the chosen conditional distribution for the response variable is a Poisson, $Y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t)$.

The link function used can be either a logarithmic or an identity function. Using the scoring defined in 3 the choice is made. For the Lakers case:

Table 6: Lakers scoring board

Scores	Log Poisson	Identity Poisson
logarithmic	2.435	2.445
quadratic	0.101	0.100
spherical	0.318	0.316
rankprob	1.576	1.595
dawseb	3.055	3.078
normsq	0.956	0.978
sqerror	7.895	8.097

For the Lakers, the scoring table 6 shows that in every criterion the logarithmic link function has better score, except for the Brier score. Moreover, the logarithmic model has better scores in the AIC and BIC. Therefore, the logarithmic function is the chosen link function. From analyzing the scoring table

Table 7: Rockets scoring board

Scores	Log Poisson	Identity Poisson
logarithmic	2.530	2.531
quadratic	0.093	0.093
spherical	0.305	0.304
rankprob	1.732	1.734
dawseb	3.232	3.236
normsq	0.830	0.831
sqerror	9.623	9.617

(7) we concluded that the logarithmic link function presents the best results among almost all scoring criterion except for the spherical score. This prompts the decision of using the logarithmic function on the Rockets model.

4. Results

4.1. Model Fitting

The objective is to fit a model similar to (4). By using the package $tscount$, the order and the variables selected before the models are fitted. Both models have a logarithmic link function and assume a Poisson conditional distribution for the response values. The modeling will use a test set of 400 of the 410 observations in the dataset for each team. Denote by $\psi_t = log(\lambda_t)$, the model for the Lakers time series of counts would then be :

$$\psi_t = 1.056 + 0.2162\big(log(Y_{t-1}+1)\big) - 0.1351\psi_{t-1} + 0.0283(AST) + 0.0069(FG) \quad (14)$$

The intercept and β_1 have confidence intervals, at a 95%, that do not contain zero which is indicative of their relevance to the model. Nonetheless, the confidence interval for α_1 contains zero which might suggest its irrelevancy. However, as discussed previously decreasing the AR parameter would result in a loss of information. The intercept has a significant impact over the other parameters of the model at about 1.06 and the past observation, β_{t-1}, has a significant positive impact in the λ_t of 0.216. Notwithstanding, the past mean, α_1, has a small negative effect on the outcome. One possible reason for it is that the higher the past values are, the harder it is to maintain the "hot streak", meaning high consecutive values in the number of threes. The AST and FG variables have a low impact on the response variable however, it is worth noting that the the AST confidence interval does not contain zero proving its importance. The variable FG has a low estimate value and its confidence interval contains zero which suggest a lack of relevancy. However, its exclusion wouldn't improve the quality of the model.

Denote by $\psi_t = log(\lambda_t)$, the model would then be :

$$\psi_t = 0.7297 + 0.0641\big(log(Y_{t-1}+1)\big) + 0.1228\big(log(Y_{t-2}+1)\big) + 0.1715\psi_{t-1} + 0.0355(AST) \quad (15)$$

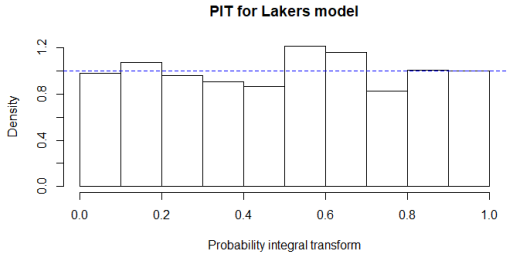
The Rockets final model presented in 15 has a logarithmic link function and the assumed distribution is a Poisson. The intercept in the Rockets case is lower than the Lakers one, at 0.7297, and its confidence interval at 95% does not contain zero cementing its importance. The past observation at time $t-1$, Y_{t-1}, has a low estimate and its confidence interval contains zero which might suggest its lack of importance. This estimate has a significant lower value compared to the Lakers model.

Despite that, its importance is high as excluding the term would result in a loss of information. The past observation at time $t-2$, Y_{t-2}, has a higher estimate compared to the first one, of 0.1228, and its confidence interval does not contain zero. The moving average component of the model, α_{t-1}, has a positive estimate of 0.1715 and a confidence interval that contains zero. Nevertheless, this variable cannot be removed from the model due to the loss of information inherent with it. Finally, the covariate added, AST, has confidence interval that does not contain zero and has a positive impact on λ_t. The AST has the higher estimate, of 0.0355, compared to the Lakers which might be because of the higher correlation between $3P$ and AST in the Rockets case compared to the Lakers case.

4.2. Model Assessment

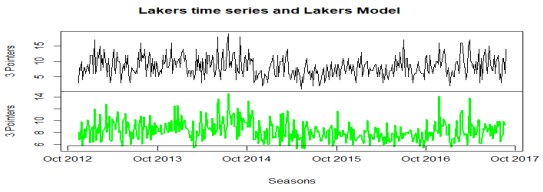
The quality of the model is be accessed using PIT histograms and the plots comparing the real values of the time series and the fitted values.

Figure 7: Lakers PIT histogram for the Lakers model



The PIT histogram (7) for the Lakers model shows that there isn't either overdispersion nor underdispersion. In fact, the PIT histogram is close to uniformity which is a testament to the correct calibration of the model.

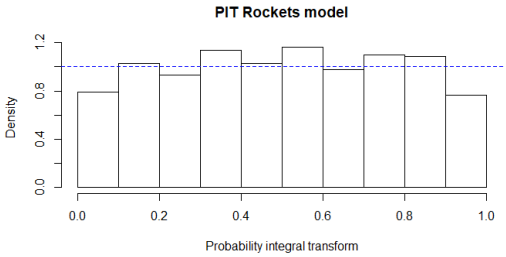
Figure 8: Lakers original time series (black) and the fitted values of the model(green)



The plot above (8) shows that the fitted values, represented in the green line, capture most of the trends that occur either within a season or between them. The model fitted properly showcases the upward spikes in values of threes with the exception of the spikes after the All-Star break during season 4. However, the same cannot be said for the downward spikes in value since, the decrease in value for $3P$ isn't captured by the model across time consistently. This is recognizable in the post All-Star break of the third season, around February 2014, where there is a high dense region of downward spikes in the real-time series and none in the model one. It can be seen that the fitted values remain close to the mean at the end of the third season which, further proves that the model cannot properly capture the downward trends. Moreover, the model fitted values present decreases in values of a far lower magnitude compared to the real ones. There is, nonetheless, a downward trend in the beginning of the second season that the model was able to identify.

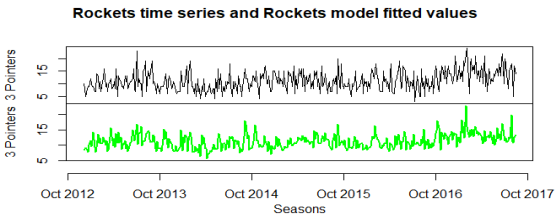
In the case of the Rockets the PIT histogram reveals some problems from the model fitting as seen below:

Figure 9: PIT histogram for the Rockets model



In the figure above (9), the histogram shows an upside-down U-shape. This is a sign of overdispersion of the predictive distribution. Since there is a significant distance to the uniform distribution then it suffices to say that there is an incorrect calibration to the model.

Figure 10: Rockets original time series (black) and the fitted values of the model(green)



The model fitted values showcase the major trends from the original time series without, however, the same sensitivity as one can see in (10). Also, the model fitted values present a lower dispersion compared to the real-time series. A possible explanation is that the weights of the model estimates for the Rockets are not high enough to cause significant changes in values. One example is the third season where, there is a high dispersion in the real values. Despite that, the model fitted values present a low dispersion and the spikes

generated have a relatively small magnitude compared to the real value time series of counts. Moreover, the upward evolution in three pointers from the end of the fourth season to the beginning of the fifth season is more accentuated in the real-time series in comparison to the model fitted values. This might suggest that the model does not react as quickly to variations. Another deficiency is the inability to capture the downward spikes. In the real-time series there are several values that drop well below the mean but the model fitted couldn't capture those low values which is similar to the Lakers case.

4.3. Prediction

In this section, the plots will show the real values of the time series, that where stored in a test set, in a green line and, the predicted values generated by the model, in a red line. The objective is to verify if the predictions are close to reality even though since, the model response variable is λ_t, this prediction are random Poisson values.

Figure 11: Lakers prediction plot with the real values in green and the predicted values in red

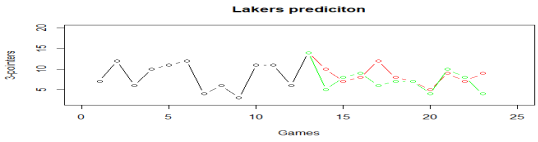
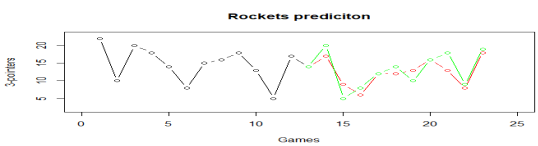


Table 8: Lakers table of predicted values and real values

Prediction	14	10	7	8	12	8	7	5	9	7	9
Real Values	14	5	8	9	6	7	7	4	10	8	4

Given the table (8) and plot (8) above, some of the assertions made previously can be quantitatively explained. In the first step, there is a decrease in both values however, in the real-time series the decrease was of 9 three pointers instead of only 4. The following two predictions have a minimal difference to the real values. However, the fourth predicted value suggests a rise in value to 12 when in reality, it drops to 6. The rest of the prediction, apart from the last value, is almost identical to the real values. Lastly, the last value predict predicts increase of 2 three pointers whereas in reality the value drops by 4 three pointers.

Figure 12: Rockets prediction plot with the real values in green and the predicted one in red



In the Rockets case (12), the first predicted value understates the real increase in value compared to reality. Similarly, the second predicted value produces a downward spike of a lower magnitude compared to the real value. Across the plot, it can be seen that the red dots aren't as dispersed as the green dots. It is worth mentioning that, the fifth and sixth predicted values suggest a consecutive increase in values whereas, in reality, the real-time series increases its value first and decreases after. For the last values predicted, the predicted values are close to reality with exception if the eighth value predicted which, estimated a decrease in the value when in reality there was a slight increase. This can be numerically seen in the table below (9).

Table 9: Rockets Table of predicted values and real values

Prediction	14	17	9	6	12	12	13	16	13	8	18
Real Values	14	20	5	8	12	14	10	16	18	9	19

5. Conclusions

The purpose of this analysis was the study of the evolution basketball statistics and its subsequent prediction. The $3P$ time series of counts were studied and described for both teams. The order of the base models was chosen for both teams using the AIC and BIC. The variables chosen using the same criteria were then introduced in the final models. For the Lakers model, the AST and FG variable was used and for the Rockets model only the AST. Moreover, the choice for the link function and the conditional distribution for the response variable was done based on several assessment tools. Both models used a logarithmic link function and assumed a Poisson distribution for the conditional distribution. Afterward, a model similar to 4 was fitted using an R-package called $tscount$. For the model fitting, it was used a train set containing only 400 of the 410 observation possible for both the time series and the covariates used. Subsequently, the model was assessed and tested in order to check the quality of the model fitted using PIT histograms, Marcal plots and ACF plots. Also the fitted values were directly compared to the real valued time series. Finally, a prediction was made using such model and then the values were interpreted. The prediction were then checked against the test set that was created during the model fitting part.

The models obtained in chapter 4 properly describe the evolution of $3P$ over the 5 seasons. The type of model employed allows for the integration of covariates that greatly improves the assessment tools used. Both models presented of at least one AR parameter in the models, consistent with the theory that one game affects the following one, which leads to a concept of "form". A team that scored an high value of $3P$ will probably score an

high value the next game. Another similarity is the presence of the AST variable in both models, a stat correlated with ball sharing and passing. The variables from the opposing team had practically no correlation with the outcome of the response variable and provided no relevant information to the study. Moreover, defensive statistics didn't improve the models results despite having residual correlation with $3P$. It is worth mentioning that the model of the most successful team during this period, Houston Rockets, had a higher AST estimate. On the other hand, the Lakers model presented one estimate that was negative, the MA parameter. This might reflect the inability to sustain a high level of three pointers over an extended period. Moreover, both models were able to capture and identify the majority of the upward spikes in the values as well as the trends within the season and between them. Lastly, the prediction made using the models estimated proved to be quite effective. Most of the values only had marginal differences compared to the original value of the response variable and, only in a few step-ahead predictions had major errors. Given the results of the study undertaken it can be said that the application of integer time series models to basketball statistics garners some relevant results. The application of said models to more statistics besides $3P$ might give insight on how a team can improve, by changing its game strategy, and help managerial decisions when acquiring new players. In the case of $3P$, if the team is focused in passing the ball it is more probable that they would score a higher number of $3P$. Also the managers are able to see if the teams are able or not of maintaining a high performance.

Despite the relevant results there are still some issues. The models were unable to identify parameters with a significantly negative impact in the response variable, which, consequentially resulted in an inability to identify downward spikes. The model variations was also not as high as the original time series. The data-set displayed some multicolinnearity problems ans well as dependency issues. Although perfect dependence is not attainable in this sport, given the residual dependence of all action in-game, a solution to this problems would be the access to bio-metric data for example.

References

[1] Hollinger, J. (2005). *Pro basketball forecast*. Potomac Books Inc., ISBN: 978-1574889628.

[2] Kubatko, J., Oliver, D., Pelton, K. and Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. Journal of Quantitative Analysis in Sports 3. *Journal of Quantitative Analysis in Sports* **3**.

[3] Shea, S. M. and Baker, C. E. (2013). *Basketball analytics: Objective and efficient strategies for understanding how teams win. Advanced Metrics*. CreateSpace Independent Publishing Platform, ISBN: 978-1492923176.

[4] Goldman, M. and Rao, J. M. (2013). Live by the Three, Die by the Three? The Price of Risk in the NBA. *MIT Analytics Conference*, 1–15.

[5] Liboschik, T., Fokianos, K. and Fried, R. (2017). tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. *Journal of Statistical Software* **82**, 1–51.

[6] R Core Team (2016). *R – A Language and Environment for Statistical Computing.*

[7] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* **135**, 370–384.

[8] Heinen, A. (2003). Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model. *SSRN Electronic Journal*.

[9] Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-Valued GARCH Process. *Journal of Time-Series Analysis* **27**, 923–942.

[10] Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association* **104**, 1430–1439.

[11] Fokianos, K. and Tjøstheim, D. (2011). Log-Linear Poisson autoregression. *Journal of Multivariate Analysis* **102**, 563–578.

[12] Christou, V. and Fokianos, K. (2014). Quasi-Likelihood Inference for Negative Binomial time series models. *Journal of Time Series Analysis* **35**, 55–78.

[13] Douc, R., Doukhan, P. and Moulines, E. (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications* **123**, 2620–2647.

[14] Sim, T. (2016). Maximum Llkelihood estimation in partially observed Markov models with applications to time series of counts. *Statistics Theory [stat.TH]. Télécom ParisTech.*

[15] *Total NBA league revenue* from 2001/02 to 2016/17 (in billion U.S. dollars)*, 2018. Statista.