arXiv:2304.01538v1 [stat.AP] 4 Apr 2023

一种用于描述 NBA 篮球得分水平的双重自激泊松模型

Álvaro Briz-Redón[1*]

[1]统计与运筹学系，瓦伦西亚大学，西班牙[*]*alvaro.briz@uv.es*

摘要

本文考虑采用泊松时间序列模型，描述篮球队或球员在比赛层面（赛季内）和分钟层面（比赛内）的投篮命中球数量。为处理数据中存在的时间自相关问题，模型按照 INGARCH(1,1) 设定，引入双重自激结构。为了在比赛内层面估计模型，在贝叶斯框架下采用分而治之程序。利用从2018—2019赛季选取的部分 NBA 球队和球员对该模型进行检验。

关键词 — 贝叶斯推断，NBA 数据，泊松模型，自激模型，体育分析

1 引言

近年来，从方法论和实践角度出发，对职业体育实践所产生的数据集进行统计分析日益受到关注。因此，统计学期刊上关于这一主题的发表数量呈上升趋势（Swartz，2020）。具体而言，篮球数据集受到研究者的广泛关注。相关研究主题包括赛前和比赛中对比赛结果的预测（Shi和Song，2021；Song等人，2020；Tian等人，2020）、对投篮位置空间分布的估计（Yin等人，2020）、投篮选择空间结构的聚类（Hu等人，2021；Yin等人，2022）、对表现变异性的研究（Sandri等人，2020）、对追踪数据的分析（Santos-Fernandez等人，2022），以及网络竞争模型的构建（Horrace等人，2022）。

篮球分析领域另一个备受关注的主题是所谓的“热手”假说。Gilovich等人（1985）的研究使这一假说受到科学界关注。“热手”指球员在一次或多次投篮命中后，成功概率有所提高的现象。研究者对罚球中的这一假说进行了较为深入的研究，而对投篮命中球的研究相对较少。在一篇近期论文中，Lantis和Nesson（2021）发现罚球中存在较小的热手效应，而投篮命中球中则不存在这一效应，尽管现有文献对此问题的研究结果并不一致。

本研究的主要变量是在某一时间划分内球队或球员投篮命中球的数量，这使我们得以分析时间序列数据。由于投篮命中球的数量是整数值变量，泊松模型适合用于分析。具体而言，本文提出一种具有自激结构的泊松模型，用于分析篮球球队或球员得分水平所呈现的时间依赖性。自激模型广泛应用于犯罪学（Mohler等人，2011）和地震分析（Ogata和Katsura，1986）等领域，能够将过程的历史对感兴趣变量的影响纳入考量，即（近期）过去的观测值对其在后续时点的期望值产生的影响。本文考虑一种具有自回归条件异方差结构的泊松模型，该模型通常用于分析经济时间序列（Comte和Lieberman，2000），但近来也开始应用于其他领域，例如犯罪学（Clark和Dixon，2021；Escudero等人，2022）和体育分析（Cerqueti等人，2022）。更确切地说，我们选择双重自激设定，以在两个不同的时间分辨率层级分析数据：博弈层面（一个赛季内）和分钟层级（一场比赛内）。

本文结构如下。第 2 节介绍分析所用的数据集。第 3 节介绍提出的建模方法，以及如何进行模型估计和评估。此外，还展示了该模型结果在时间序列聚类中的辅助用途。随后，第 4 节汇报所考虑数据集的主要结果。最后，第 5 节给出一些总结性评述。

2 数据

本次分析使用了一个包含2018—19赛季NBA所有投篮命中球和投篮未命中球相关信息的数据集。该数据集下载自https://datavizardry.com/2020/01/28/nba-shot-ch

借助 NBA应用程序编程接口，可以构建这类数据集。

（API）。具体来说，对于赛季中的每次投篮命中或未命中，数据集都包含多项变量，包括投篮球员的姓名、该球员所属的球队，以及投篮发生的节次和分钟（节内）。本文的分析仅考虑投篮命中球。此外，每次投篮命中的时间位置都会转换为其发生的比赛分钟。因此，数据集中的每条记录都会被赋予一个介于 1 至 48 之间的整数。加时赛中的投篮命中球已从分析中剔除。

目前，NBA联盟共有30支球队，分为东部和西部两个赛区。赛制包括82场常规赛，以及常规赛结束时各赛区排名前八的球队参加的季后赛。本研究考察了2018-19赛季晋级各自分区半决赛的八支球队：波士顿凯尔特人（BOS）、丹佛掘金（DEN）、金州勇士（GSW）、休斯敦火箭（HOU）、密尔沃基雄鹿（MIL）、费城76人（PHI）、波特兰开拓者（POR）和多伦多猛龙（TOR）。此外，我们还分析了该赛季八名球员的得分水平：布拉德利·比尔（BB）、达米安·利拉德（DL）、多诺万·米切尔（DM）、詹姆斯·哈登（JH）、卡尔-安东尼·唐斯（KAT）、肯巴·沃克（KW）、凯文·杜兰特（KD）和保罗·乔治（PG）。这些球员是2018-19赛季常规赛得分最高的球员，筛选范围为该赛季出场超过75场的NBA球员。下文中，N_g 表示某支球队或球员在该赛季参加的比赛场数。因此，对于所有球队，$N_g = 82$；对于每位球员，$N_g \in \{76, ..., 82\}$。图1展示了多伦多猛龙队和肯巴·沃克的得分模式，具体以每场比赛及赛季分钟内投篮命中数表示。

3 方法论

本研究的方法论目标是定义一种模型，既能捕捉目标变量的时间依赖性（即球队或球员的投篮命中球数），也能捕捉驱动该变量的随机过程可能具有的自激特性。此外，我们希望针对两种不同的时间划分对投篮命中球数进行建模：NBA赛季中的各场比赛，以及NBA比赛中的各分钟。因此，下文将分别讨论数据的赛季内变化和比赛内变化。

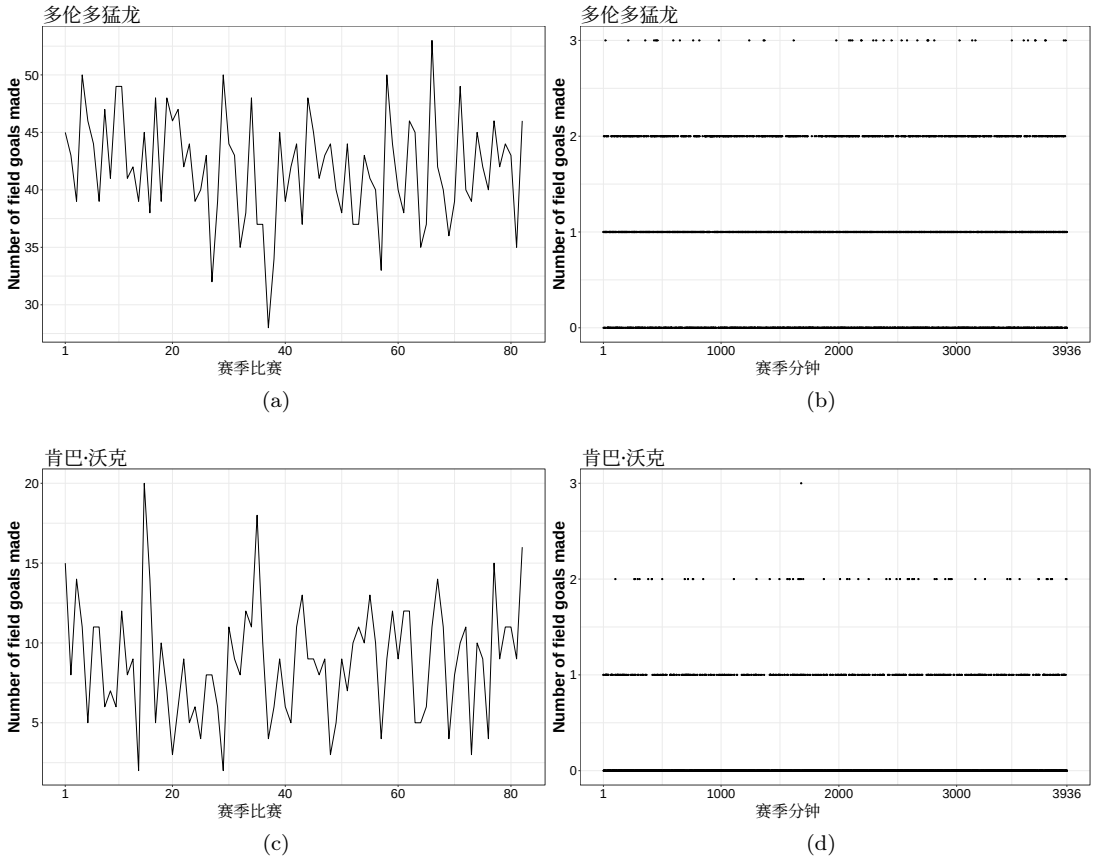


图 1：分析中所考虑的响应变量的图示：每场比赛的投篮命中球数（(a) 和 (c)），以及每分钟的投篮命中球数（(b) 和 (d)）。数据分别对应于多伦多猛龙 2018-19赛季（(a) 和 (b)）以及肯巴·沃克（(c) 和 (d)）

3.1 INGARCH 模型

INGARCH模型（Ferland等人，2006）是一种时间序列模型，对应于整数值广义自回归条件异方差（GARCH）过程；该过程此前由Bollerslev（1986）提出。如果 $\{Y_t\}_{t\in\mathcal{T}}$ 表示按时间划分 $\mathcal{T}$ 索引的计数时间序列，我们可以假设 Y_t 服从速率为 λ_t 的泊松分布，并对该速率进行建模。在INGARCH(p,q)模型下， λ_t 的结构为

$$
\begin{aligned}
Y_t &\sim Po(\lambda_t) \\
\lambda_t &= d + \sum_{i=1}^{p} \kappa_i \lambda_{t-i} + \sum_{j=1}^{q} \eta_j Y_{t-j}.
\end{aligned}
$$

这类模型将过去观测值的影响与时间序列期望值的影响结合起来，而后者可能取决于外生协变量信息。 η_i表示随机过程中的重复现象，而 κ_j则控制在没有重复现象时 Y_t 的衰减。出于实际应用方面的考虑，该模型有一个广泛使用的特例，即 INGARCH(1,1) 设定，其表达式为

$$Y_t \sim Po(\lambda_t)$$

$$\lambda_t = d + \kappa\lambda_{t-1} + \eta Y_{t-1}.$$

Heinen（2003）已确立该模型的主要性质。在$\kappa + \eta < 1$的条件下，INGARCH(1,1)模型是平稳的，其无条件均值为

$$E(Y_t) = \mu = \frac{d}{1-(\kappa+\eta)},$$

而 Y_t 的无条件方差为

$$Var(Y_t) = \frac{\mu(1-(\kappa+\eta)^2+\eta^2)}{1-(\kappa+\eta)^2}.$$

因此，有 $Var(Y_t) \geq E(Y_t)$，这表明该模型能够捕捉观测值的过度离散（只要 $\eta \neq 0$），尽管其（条件）定义为泊松模型。具体而言，模型的过度离散程度是 η 的递增函数，在较小程度上也是 κ的递增函数。最后，无条件自相关函数为

$$Corr(Y_t, Y_{t-r}) = \frac{\eta(1-\kappa(\kappa+\eta))(\kappa+\eta)^{r-1}}{1-(\kappa+\eta)^2+\eta^2}. \tag{1}$$

3.2 双重自激泊松模型

3.2.1 比赛层级模型

本文提出一种双重自激泊松模型，用于同时描述NBA赛季中一支篮球队/球员在各场比赛中的投篮命中球数，以及该赛季每场比赛各分钟内的投篮命中球数。设 $\{Y_g\}_{g\in\{1,\ldots,N_g\}}$ 为时间序列，表示一支篮球队/球员在NBA赛季中参加的 N_g 场比赛里投篮命中球的数量；设$\{Y_{gm}\}_{g\in\{1,\ldots,N_g\},\ m\in\{1,\ldots,48\}}$ 为时间序列，表示一支

篮球队/球员在相应比赛的（$N_g \cdot 48$）分钟内投篮命中球的数量，这些比赛共计 N_g 场，由该球队或球员参加（对于球队，由于 $N_g = 82$，时间序列的长度为 $82 \cdot 48 = 3936$）。

如果我们假设 $Y_g \sim Po(\lambda_g)$，其中 λ_g 表示第 g 场比赛的得分水平，那么 INGARCH模型假设，篮球球队/球员在一场比赛中的投篮命中球期望数量取决于先前的期望值（ λ_g）和先前的观测值（ Y_g）。这使得模型能够考虑所分析数据的自激特性。此外，INGARCH模型的结构允许纳入固定效应和随机效应。本文采用以下 INGARCH(1,1) 结构对 Y_g进行建模：

$$
\begin{aligned}
&Y_g \sim Po(\lambda_g) \\
&\lambda_g = \exp(\alpha_S + \alpha_{\text{Home}} I_{Home_g=1}) + \kappa_S \lambda_{g-1} I_{g>1} + \eta_S Y_{g-1} I_{g>1},
\end{aligned}
\tag{2}
$$

其中， α_S 是赛季级截距参数， α_{Home} 表示主场作战对响应变量的影响， κ_S 表示先前期望值 λ_{g-1} 对 λ_g 的影响，而 η_S 表示先前观测值 Y_{g-1} 对 λ_g 的影响。项 $I_{g>1}$ 表示一个指示函数，当 $g > 1$时取值为 1。对于 $g = 1$ ，我们无法在 λ_1 的设定中纳入自激参数（因为对于 $g = 1$没有观测到该过程的历史数据），因此在估计中仅包含 α_S 和 α_{Home} 参数。

3.2.2 分钟级模型

第二步，考虑采用类似的 INGARCH 结构，用于描述一支球队/球员在整个 NBA赛季中每分钟投中的投篮命中球的期望数量。此外，将根据（2）对 λ_g 进行建模的项纳入 λ_{gm} 的设定中，从而使我们能够考虑球队/球员在比赛 g 中的（自激发）得分水平。具体而言，采用以下表达式对 λ_{gm} 进行建模

$$
\begin{aligned}
&Y_{gm} \sim Po(\lambda_{gm}) \\
&\lambda_{gm} = \exp\left(\alpha_G + \sum_{Q \in \{1,2,3,4\}} \sum_{H \in H_Q} \alpha_{QH} I_{QH_{gm}=1} + \frac{\lambda_g}{48}\right) + \kappa_G \lambda_{gm-1} I_{m>1} + \eta_G Y_{gm-1} I_{m>1}.
\end{aligned}
\tag{3}
$$

参数 α_G、 κ_G 和 η_G 在比赛层面表示式 (3) 中与 α_S、 κ_S 和 η_S 在式 (2) 中相同的内容。值得注意的是，比赛 g、 λ_g 的建模期望值除以 48，以纳入 λ_{gm} 的设定。这一修正是必要的，因为 λ_g 表示的是每场比赛的预期得分水平

而非每分钟的预期得分水平。术语 $I_{m>1}$ 表示一个指示函数，当 $m > 1$ 时取值为 1。与比赛层级模型的情况类似，我们无法将自激参数纳入 λ_{g1} 的设定，因此只有 α_G 和 α_{QH} 参与其估计。

因此，这种建模方法呈现出一种层级结构：比赛 g 中第 m 分钟的预期得分水平取决于比赛 g的（总体）预期得分水平。由于 λ_g 和 λ_{gm} 均采用自激设定，因此本文其余部分将完整模型称为双重自激泊松模型。严格来说，具有双重自激结构的是分钟级模型，而比赛层级模型更适合描述为自激（INGARCH(1,1)）模型。尽管如此，本文仍统一使用“双重自激泊松模型”这一术语，因为比赛层级模型的输出也参与了分钟级建模过程的估计。

3.3 模型估计

比赛层级模型和分钟级模型均在贝叶斯框架下拟合。因此，截距参数（α_S和 α_G）及固定效应参数（α_{Home}和 α_{QH}）均采用无信息先验，先验分布为 $N(0, 1000)$。此外，自激效应对应的参数（η_S、κ_S、 η_G和 κ_G）采用均匀先验，先验分布为 $U(0, 1)$。模型借助用于贝叶斯推断的 NIMBLE系统（de Valpine 等，2017）进行拟合；该系统基于马尔可夫链蒙特卡罗（MCMC）程序。

此时需要指出，拟合所提出的模型在计算上要求很高，因此我们采取了一些策略进行模型估计。首先，分别估计比赛层级模型和分钟级模型。我们先估计比赛层级模型，以便计算 $g = 1$、.....、 N_g 对应的 λ_g 的后验均值。随后，将 λ_g 的后验均值（记作 λ_g）作为分钟级模型中的偏置项：

$$
\begin{aligned}
&Y_{gm} \sim Po(\lambda_{gm}) \\
&\lambda_{gm} = \exp\left(\alpha_G + \sum_{Q\in\{1,2,3,4\}} \sum_{H\in H_Q} \alpha_{QH} I_{QH_{gm}=1} + \frac{\hat{\lambda}_g}{48}\right) + \kappa_G \lambda_{gm-1} I_{m>1} + \eta_G Y_{gm-1} I_{m>1},
\end{aligned}
\tag{4}
$$

随后对其进行估计。为了拟合这一模型（它在计算层面上构成了主要挑战），我们采用分而治之策略。我们考虑以下

观测值划分:

$$\{Y_{gm}\}_{g\in G} = \bigcup_{k=1}^{K} \{Y_{gm}\}_{g\in G_k},$$

其中， G 表示 NBA 球队/球员参加的所有比赛的集合， $\{G_k\}_{k=1}^{K}$ 表示 G的不重叠划分。在所有情况下，均有 $m \in \{1, ..., 48\}$，因此，来自某场具体比赛的数据不会在不同划分之间共享。为合并与各子集 G_k对应的感兴趣参数的后验分布，我们计算由 MCMC 模拟得到的所有后验分布所构成集合的 Wasserstein 重心（Cuturi and Doucet，2014）。Ou等人（2021）证明了该方法具有良好的理论性质。因此，若 p（$\theta|\{Y_{gm}\}_{g\in G_k}$）表示基于子集 G_k得到的模型参数（θ）的后验分布，则 θ 的合并赛季层级后验分布计算如下

$$\bar{p}(\theta|\{Y_{gm}\}_{g\in G}) = WB(p(\theta|\{Y_{gm}\}_{g\in G_1}), ..., p(\theta|\{Y_{gm}\}_{g\in G_K})),$$

其中， WB 表示 Wasserstein 重心。在本文余下部分，我们将$\bar{p}$（$\theta|\{Y_{gm}\}_{g\in G}$）称为相应参数的重心后验分布。

3.4 模型比较

模型比较采用了 Watanabe 和 Opper（2010）提出的 Watanabe–Akaike信息准则（WAIC）。WAIC 是一种衡量贝叶斯模型拟合优度的指标，同时会根据模型所涉及的有效参数数量考虑其复杂度。对于一组基于相同数据拟合的模型，WAIC 可用于直接进行模型比较。通常，WAIC 值最小的模型表现最佳，也就是说，它在拟合度和复杂度之间实现了最佳平衡。针对比赛层级和分钟层级，分别将每个所提出的模型与不包含自激效应的基线模型进行比较。结果部分概述了各情况下所采用的基线模型结构。

3.5 得分水平的聚类

作为一项次要分析，我们建议使用双重自激泊松模型的输出，对所分析的球队或球员进行层次聚类。最近，Cerqueti等人（2022年）提出了一种基于拟合INGARCH(1,1)模型的时间序列新模糊聚类方法。在我们的案例中，我们建议使用计算重心后验分布时所采用的Wasserstein距离，根据某些模型

参数来衡量球队或球员之间的相似性。具体而言，我们尤其关注模型自激分量中涉及的参数。这样便可以计算球队或球员集合的不相似度矩阵，并采用该领域常用的方法对这些元素进行层次聚类。具体来说，我们采用Sokal和Michener（1958年）提出的非加权组平均法（UPGMA），该方法已在R函数hclust中实现。

3.6 软件

本研究使用 R 编程语言（R 核心团队，2021年）进行分析。具体而言，使用了 R软件包 ggplot2（Wickham，2016年）、nimble（de Valpine 等，2017）、transport（Schuhmacher 等，2022年）和 T4transport（You，2022年）。

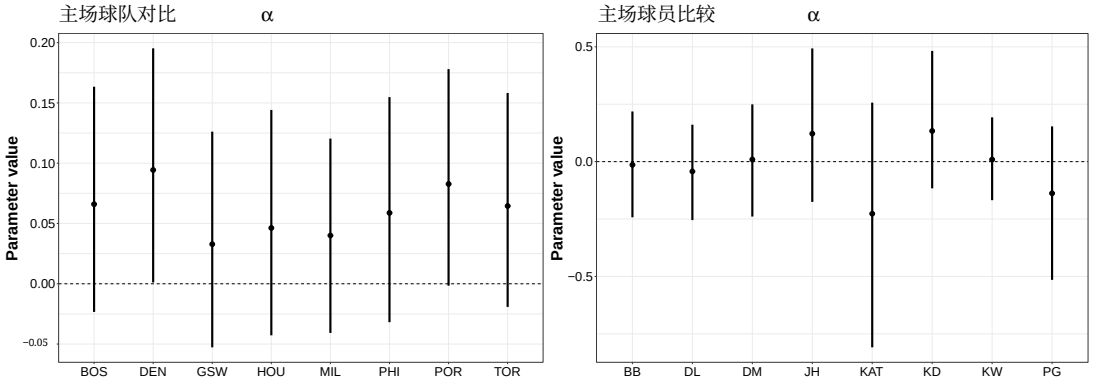
4 结果

本节讨论基于所提出的建模框架，对选定的球队和球员进行分析所得的主要结果。首先，我们分析两个模型中固定效应的估计值，即比赛层级模型中的主场比赛效应，以及分钟级模型中由半节划分所体现的时间效应。其次，我们分析模型中纳入的自激效应，包括比赛层级和分钟层级的自激效应，这正是本研究的主要动机。

4.1 主场效应

主队是否在比赛获胜概率方面具有优势，是一个受到广泛研究的问题。总体而言，既有研究表明这种效应确实存在，不过其大小取决于所分析的球队（Pollard 和 Gómez，2007）以及职业联赛（国家）（Gómez 和 Pollard，2011）。本文分析球队的得分水平是否取决于其是否在主场比赛。对于纳入分析的球员，也研究了这一点，因为表示比赛是在主场还是客场进行的变量同样被纳入球员模型。因此，图2汇总了主场比赛对每场比赛投篮命中球数的影响估计值（通过参数 α_{Home}衡量），包括球队层面（图2a）和球员层面（图2b）。球队层面的结果表明，主场比赛可能会增加每场比赛的投篮命中球数，因为所有情况下 α_{Home} 的后验均值都大于0。不过，结合相应的95%可信区间来看，只有 DEN 和 POR 显示出明显影响，

如图2a所示。对于 DEN，该区间不包含0；而对于 POR，该区间的下限略低于0。就所分析的球员而言，没有观察到任何人具有主场优势，如图2b所示。在所研究的球员中，只有 JH 和 KD 的后验均值大于0，而 KAT 和控球后卫（PG）则呈现相反趋势，尽管这些结果不能视为具有统计学意义。最后，其余四名受分析球员的后验分布都集中在0附近，这表明主场比赛在得分水平方面既没有优势，也没有劣势。



(a) (b)

图2： α_{Home}的估计值（根据参数的后验分布均值计算），以及所研究球队（a）和球员（b）对应的95%可信区间

4.2 比赛内时间效应

比赛内时间效应由所采用的分钟级建模方法中涉及的 α_{QH} 参数给出。对于该模型，我们采用了前文提到的分而治之估计策略。具体而言，对于每支待分析的球队和每位球员，我们针对整个 NBA 赛季中的以下四个比赛子集分别拟合了四个分钟级模型：$G_1 = \{1, ..., 21\}$、 $G_2 = \{22, ..., 42\}$、 $G_3 = \{43, ..., 62\}$和 $G_4 = \{63, ..., N_g\}$。图3和图4分别展示了各子集中各球队和球员的 α_{QH}估计值，据此可以分析得分水平是否存在比赛内时间变异。值得再次指出的是，这些估计值应与第一节前半段（参考水平）进行比较；该时间区间包含比赛开始后的前六分钟。 α_{QH} 参数的大多数估计值为负，这表明比赛进行到第6分钟之后的得分水平低于此前。所研究的不同球队总体上都呈现这一趋势，尽管图3所示的大多数95%可信区间都包含0。

不过，所考察赛季的不同子集之间也存在明显差异。例如，DEN仅在赛季的某些阶段表现出较低的平均得分水平。另一方面，所分析的球员在 Q_{21} 和 Q_{41} 阶段的得分水平较低，而且在大多数情况下，这一结果具有统计学意义。这可能尤其是因为对 Q_{21}而言，这一比赛阶段通常是这些球员的休息时间。通过对数据集分段来估计模型，除了降低计算复杂度之外，还能观察到估计值在整个赛季中的变化。为了获得 α_{QH}的赛季层面估计值，我们可以根据每个参数及每支球队/每位球员对应的四个后验分布，求得重心后验分布。下一节将以模型的自激效应为例说明这一过程。

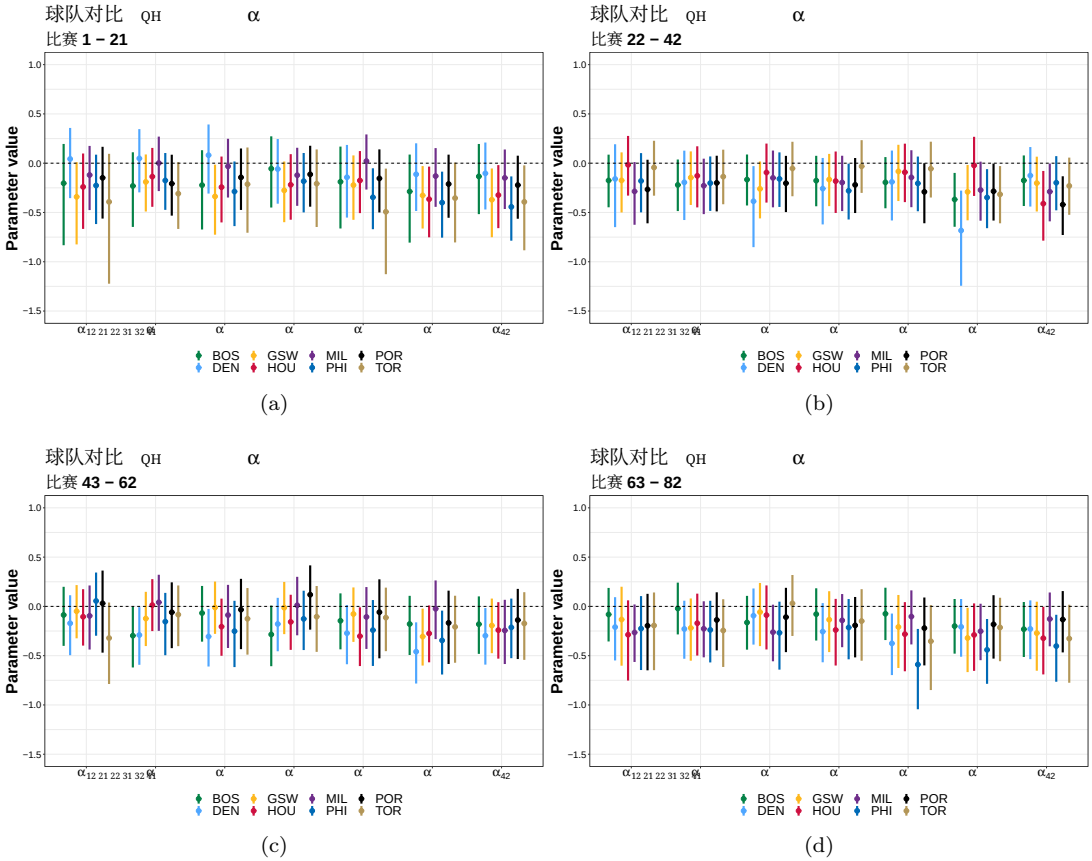


图3： α_{QH}的估计值，计算为参数后验分布的均值，以及研究中各球队的相关95%可信区间。针对模型推断所考虑的赛季各阶段，分别给出了相应的估计值

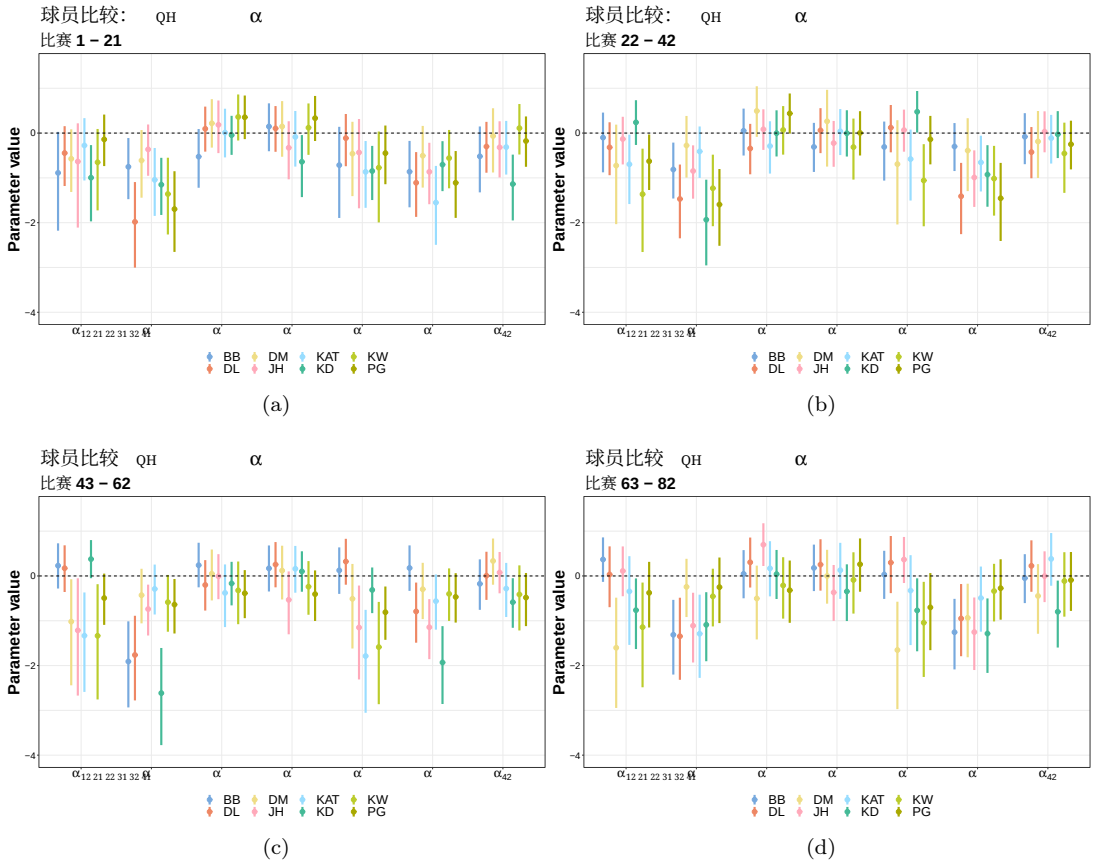


图4： α_{QH}的估计值，计算为参数后验分布的均值，以及所研究球员的相关95%可信区间。模型推断所考虑的赛季各阶段均提供了估计值

4.3 自激效应

本节给出了模型在赛季内和比赛内两个层面上自激效应的结果。作为总结，表1展示了两支球队和球员的 η_S（η_G）以及 κ_S（κ_G）的后验均值，以及相应的 $\eta_S+\kappa_S$（$\eta_G+\kappa_G$）估计值。所有情况下，这些数值均小于1，而根据 MCMC 抽样值推导出的 P（$\eta_S+\kappa_S<1$）（P（$\eta_G+\kappa_G<1$））估计值均等于1，这证实了模型的平稳性，而模型的平稳性由条件 $\eta_S+\kappa_S<1$（$\eta_G+\kappa_G<1$）保证。

4.3.1 赛季内自激效应

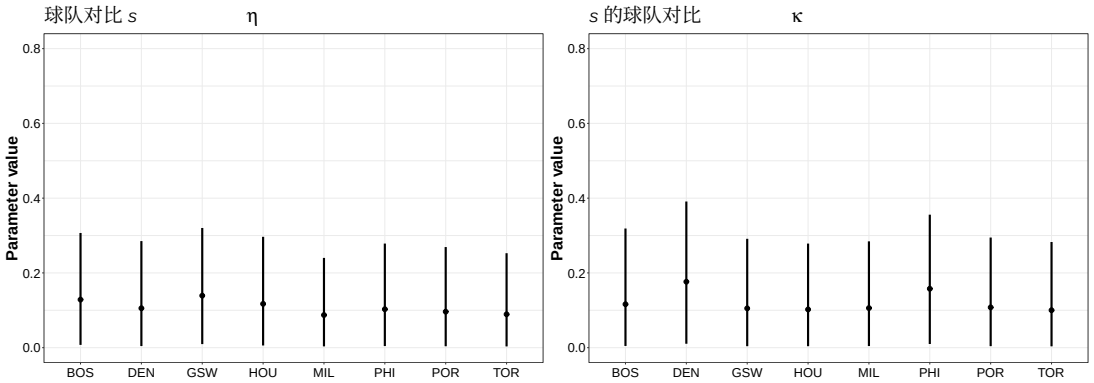
关于博弈层面的自激效应研究，我们必须重点关注模型参数 η_S 和 κ_S 的估计。图5和图6分别展示了所分析球队（图5）和球员（图6）对应的后验均值及其95%

	比赛级模型			分钟级模型		
球队/球员名称	$\hat{\eta}_S$	$\hat{\kappa}_S$	$\hat{\eta}_S+\hat{\kappa}_S$	$\hat{\eta}_G$	$\hat{\kappa}_G$	$\hat{\eta}_G+\hat{\kappa}_G$
波士顿凯尔特人	0.13	0.12	0.24	0.03	0.16	0.20
丹佛掘金	0.11	0.18	0.28	0.03	0.30	0.33
金州勇士	0.14	0.11	0.24	0.04	0.23	0.27
休斯敦火箭	0.12	0.10	0.22	0.03	0.25	0.29
密尔沃基雄鹿	0.09	0.11	0.19	0.04	0.22	0.26
费城76人	0.10	0.16	0.26	0.04	0.26	0.30
波特兰开拓者	0.10	0.11	0.20	0.03	0.27	0.30
多伦多猛龙	0.09	0.10	0.19	0.03	0.29	0.33
布拉德利·比尔	0.14	0.24	0.38	0.05	0.15	0.21
达米安·利拉德	0.05	0.24	0.29	0.06	0.14	0.20
多诺万·米切尔	0.10	0.27	0.37	0.06	0.51	0.57
詹姆斯·哈登	0.26	0.27	0.54	0.05	0.17	0.21
卡尔-安东尼·唐斯	0.16	0.58	0.74	0.09	0.32	0.41
肯巴·沃克	0.08	0.11	0.19	0.09	0.48	0.57
凯文·杜兰特	0.20	0.19	0.39	0.04	0.14	0.18
保罗·乔治	0.28	0.20	0.48	0.06	0.18	0.24

表 1：所分析的各球队和球员的 η_S、 κ_S 和 η_G 及 κ_G 的后验均值。条件 $\hat{\eta}_S+\hat{\kappa}_S<1$ ($\hat{\eta}_G+\hat{\kappa}_G<1$) 表示比赛级（分钟级）模型的平稳性

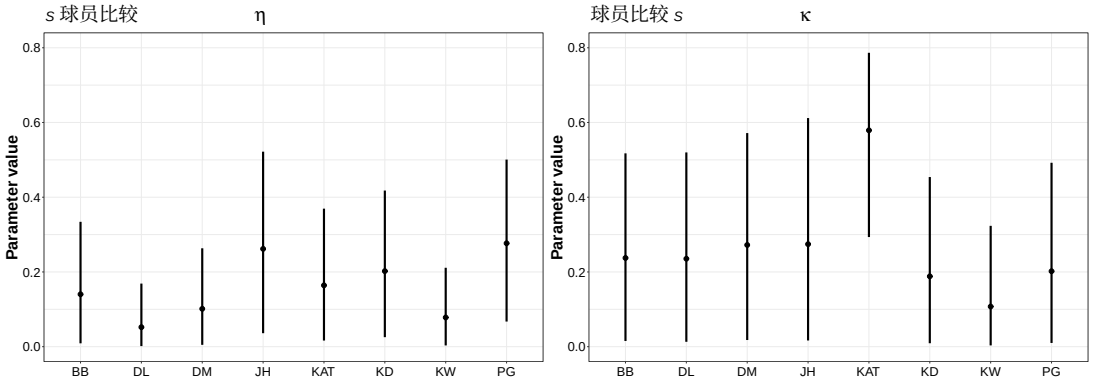
可信区间。这两个参数在各球队中的估计值（ η_S 和 κ_S ）几乎没有差异，不过 DEN 和 PHI 的 κ_S 后验均值较高，值得一提。球员之间的情况则大不相同，他们在这两个参数上的估计值差异显著。具体而言，JH 和控球后卫（PG）的 η_S 高于其他球员，而 KAT 的 κ_S则尤为突出。同时，DL 和 KW 等球员的 η_S估计值较小，KW 的 κ_S 估计值也较小。为了直观展示 η_S 和 κ_S 的估计值如何反映这些球员不同的得分倾向，图7根据 η_S 和 κ_S的估计值，展示了拟合模型得出的自相关函数估计（依据公式（1））。该自相关函数针对滞后值 $r=1$，.....，10 进行估计，表示 Y_g 与 Y_{g-r}之间的相关系数。具体而言，图7显示，JH 和控球后卫（PG）在 $r=1$处的自相关函数估计值接近0.3，是所分析球员中最高的，这与这两名球员较大的 η_S 估计值相对应。我们还可以看到，KAT 的比赛层面得分水平的时间依赖性持续时间更长；该球员的 κ_S估计值最高。无论如何，我们应注意到，这些估计值对应的可信区间彼此重叠，因此对球员间差异的判断应谨慎。尽管如此，博弈层面的模型分析使我们得出结论：尽管比赛层面的得分动态并不能很好地刻画所研究的球队，但有些

联盟中最优秀的得分手在这方面表现出明显差异。



(a) (b)

图5： η_S和 κ_S的估计值，计算方式为参数后验分布的均值，并给出了所研究各球队对应的95%可信区间。



(a) (b)

图6：研究球员的 η_S和 κ_S的估计值（通过相应参数的后验分布均值计算）及其95%可信区间

4.3.2 比赛内自激效应

比赛内自激效应的研究基于通过分钟级模型估计的 η_G 和 κ_G 参数。我们根据前文提到的NBA赛季四个子集，计算这些参数的重心后验分布，从而得到这些参数唯一的赛季层面估计值。例如，图8展示了波士顿凯尔特人每个赛季子集对应的η_G 后验分布，以及由此得到的重心后验分布。

因此，图9和图10展示了所分析球队（图9）和球员（图10）两个参数的重心后验均值及其相应的95%可信区间

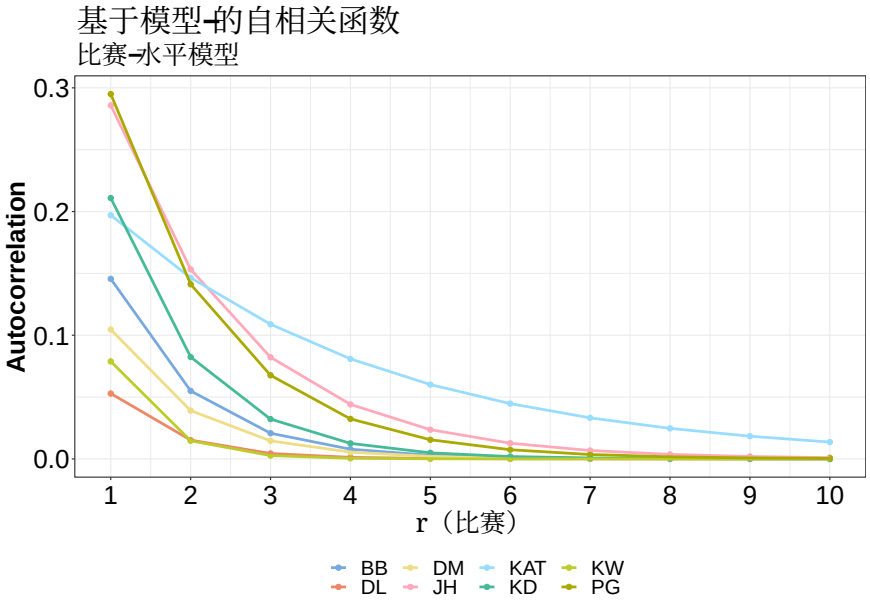


图7：基于为每位受分析球员拟合的比赛级自激模型对自相关函数进行估计。

。与比赛层级模型的情况一样， η_G和 κ_G在球队之间的变异小于在球员之间的变异。根据从重心后验分布中获得的赛季层面估计值，我们可以指出，KAT和KW的 η_G值最大，而DM、KAT和KW的 κ_G 估计值最大。这一结果表明，就重复性而言，这三名球员呈现出更为独特的得分模式。图11展示了分钟级模型（遵循（1））估计的自相关函数，其中考虑了滞后值 $r = 1$、.....、10（以分钟计），表示 Y_{gm} 与 Y_{gm-r}之间的相关系数。显然，分钟级模型中的自相关函数值低于比赛层级模型中的相应值。不过，它有助于理解每分钟得分水平之间时间依赖性的大小，以及球员之间现有的差异。

4.4 所提出的模型与基线模型的比较

表 2 展示了本文提出的双重自激设定下拟合的各比赛级和分钟级模型的 WAIC 值，以及用于比较的基线模型的 WAIC 值。这些基线模型排除了所建议模型中自激分量涉及的所有参数。具体而言，比赛级基线模型仅考虑主场效应，而分钟级基线模型仅纳入

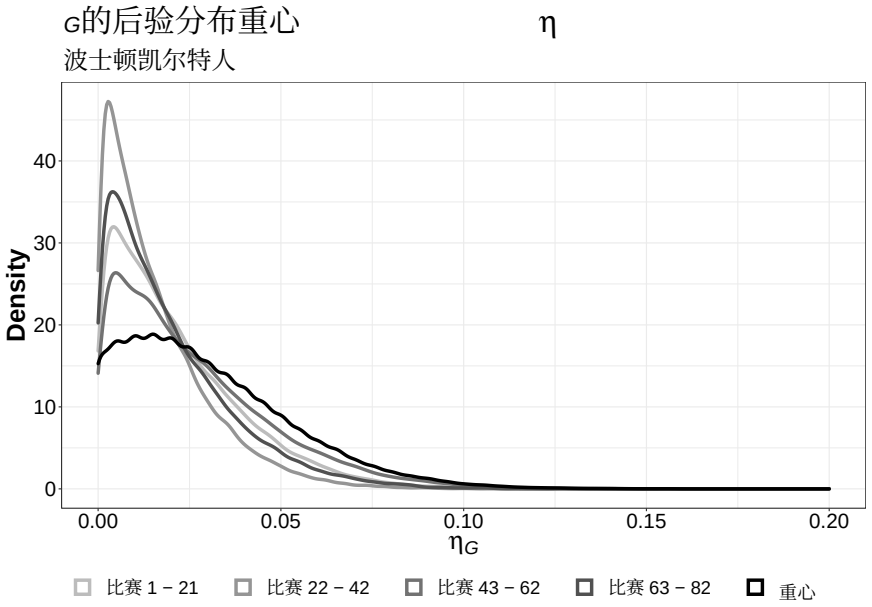


图8：与波士顿凯尔特人及分析所考虑的赛季各阶段相对应的 η_G 后验分布，以及基于Wasserstein距离的这些后验分布的重心

由 α_{QH} 给出的比赛内时间效应。此外，比赛层级基线模型提供的 λ_g 后验均值被纳入分钟层级基线模型，作为偏置项。首先，比赛级模型的结果表明，在大多数情况下，采用自激结构对比赛层面的得分水平进行建模并无益处。具体而言，除 JH、KAT 和控球后卫（PG）这三名球员外，WAIC 均未得到改善。这一结果与这些球员在 η_S 和 κ_S 参数上具有较高估计值的情况一致。其次，分钟级模型所得的 WAIC 值表明，总体而言，这种建模方法在球队和球员层面都能带来更好的结果。事实上，对于赛季中的许多球员／球队和阶段，双重自激模型都改善了 WAIC。尤其是其中一些对象，在赛季至少两个阶段都有所改善。对于 KAT，除最后一个阶段外，所有阶段都观察到了改善。这再次表明，该球员的得分水平具有更强的自激程度。

4.5 得分水平聚类

最后，我们根据模型中自激分量相关参数的估计值，对球员进行了层次聚类。图 12 展示了该组球员对应的树状图，聚类依据为后验分布的重心估计

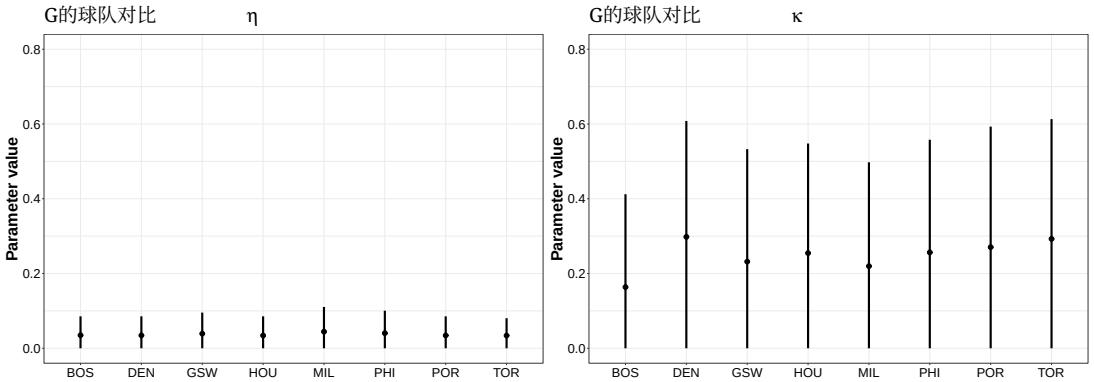


图9： η_G的估计值（a）和 κ_G的估计值（b），通过相应参数的重心后验分布均值计算；以及研究中各球队对应的95%可信区间

，其中包括 η_S 和 κ_S。此外，还列出了通过计算每对球员后验分布之间的 Wasserstein 距离得到的差异矩阵。出于示例说明的目的，我们选择了 η_S 和 κ_S ，因为它们是球员之间变异性最大的自激参数，而非分钟级模型对应的参数。基于同样的原因，本分析比较的是球员而非球队，因为在 η_S 和 κ_S 方面，各球队之间的差异很小。对于 η_S，在该参数上取值最高的球员 JH 和控球后卫，也是在相应后验分布方面最相似的球员。不过，参数 η_S的球员间差异并不大。相比之下，对于 κ_S，KAT 的后验分布与其他球员相距甚远，这一点也反映在树状图（图 12c）以及与不相似度矩阵对应的热力图（图 12d）中。

5 讨论

本文提出了一种双重自激泊松模型，用于分析表示篮球球队/球员投篮命中数的时间序列数据，并考虑两个不同的时间域。从这个意义上说，不能排除可变时间单元问题（Cheng和Adepeju，2014年）的存在，因为选择比赛和分钟作为参照时间单元虽然自然，却带有任意性。选择更大的比赛内时间单元可能会更为合适。

得分水平是否存在传染性是一个备受关注的问题，通常会从热手假说的角度探讨；该假说关注投篮或罚球的命中率

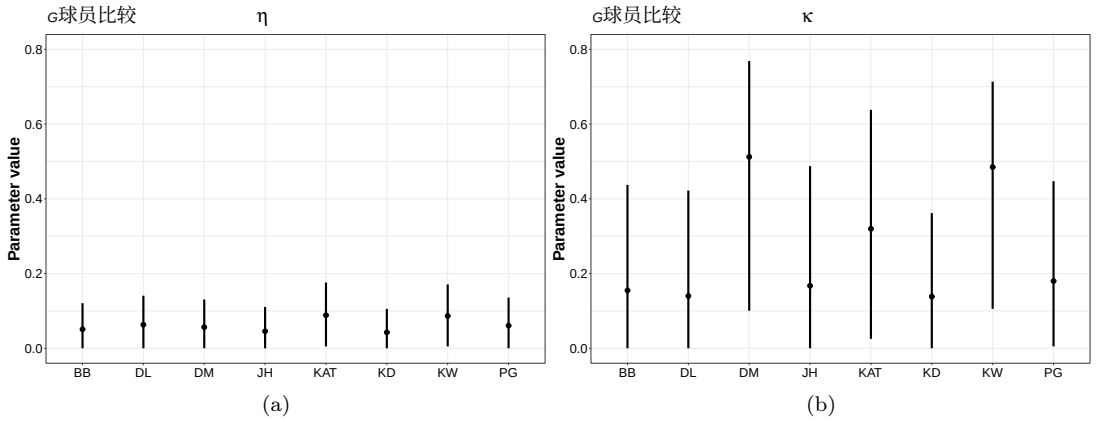


图10： η_G的（a）和 κ_G的估计值，计算方法为相应参数重心后验分布的均值；同时给出研究中球员的95%可信区间

。然而，本文关注的是一般意义上的得分水平，也就是球队或球员的得分效率。研究这一变量的自激特性，是另一种分析方法，也能反映球队或球员在赛季或比赛中可能经历的得分动态。结果并未表明某些球员或球队存在此类自激动态，但该模型对其他球员或球队则适用。具体而言，本文采用的建模方法有助于识别所考察球员之间不同的得分模式。根据关键参数对应后验分布之间的Wasserstein距离对球队或球员进行聚类，也证明是一种成本低廉的选择。

沿此方向开展的后续研究有助于揭示球队内部的得分模式。特别是，对球队所有球员得分模式对应的多元时间序列进行分析，或可揭示部分球员之间是否存在球队内部的传染动态；这一思路可借鉴Horrace等人（2022年）近期研究的“团队化学反应”概念。

最后值得指出的是，所提出的模型目前仅用于解释性分析，旨在描述和理解一些NBA球队及球员的得分模式。也值得进一步探讨将该模型用作比赛中的预测工具，应用场景包括篮球及其他适合采用该模型的体育项目。

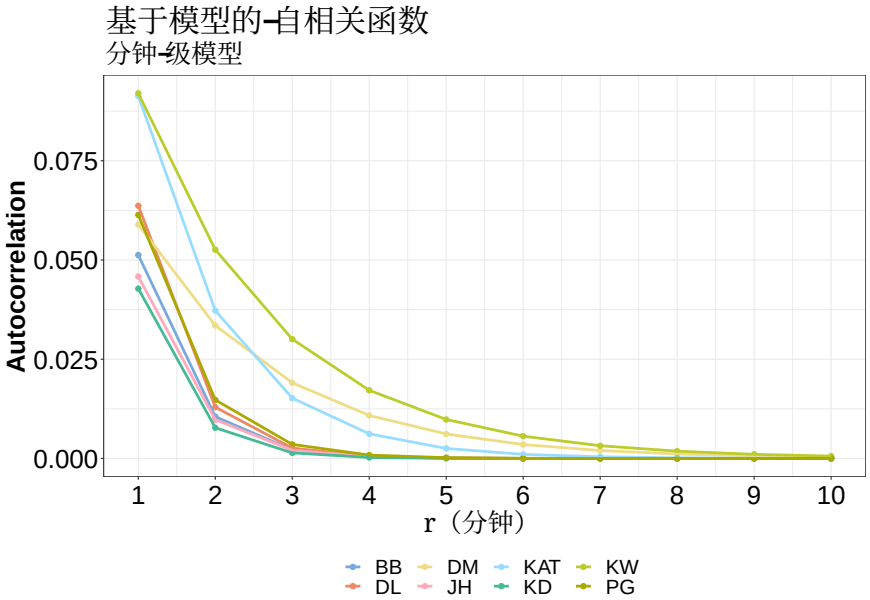


图11：基于为所分析的每位球员拟合的分钟级自激模型估计的自相关函数

	比赛级模型		分钟级模型							
			比赛 1-21		比赛 22-42		比赛 43-62		比赛 63-82	
球队/球员名称	基线	DSE	基线	DSE	基线	DSE	基线	DSE	基线	DSE
波士顿凯尔特人	499.56	504.30	2258.38	2261.91	**2293.57**	**2292.99**	**2197.91**	**2197.75**	2210.12	2210.74
丹佛掘金	502.74	504.78	**2275.48**	**2274.34**	2262.05	2262.48	2213.70	2214.11	2160.58	2161.39
金州勇士	512.05	515.93	2303.43	2304.84	**2305.82**	**2305.29**	2251.63	2252.79	**2201.46**	**2200.68**
休斯敦火箭	497.80	501.79	2227.13	2228.85	2205.88	2206.22	2141.46	2141.75	**2158.17**	**2157.49**
密尔沃基雄鹿	507.45	511.67	2362.24	2364.40	**2275.33**	**2275.03**	**2218.26**	**2217.96**	2205.60	2205.77
费城76人	507.70	509.82	2209.47	2210.44	2304.81	2306.04	2218.63	2218.92	**2160.79**	**2160.62**
波特兰开拓者	498.68	503.48	2260.74	2264.26	**2258.70**	**2258.57**	**2200.71**	**2200.42**	2223.38	2223.71
多伦多猛龙	501.63	507.02	2318.89	2321.09	**2251.91**	**2251.21**	2193.42	2193.75	2183.80	2184.56
布拉德利·比尔	407.88	410.32	937.75	939.77	1106.79	1107.53	1041.53	1041.79	939.01	939.90
达米安·利拉德	395.01	399.41	954.11	955.56	922.74	922.78	933.75	933.88	815.00	815.57
多诺万·米切尔	400.25	404.12	957.35	960.22	**945.64**	**944.44**	**1020.03**	**1019.94**	712.67	713.10
詹姆斯·哈登	**391.25**	**388.63**	**1023.76**	**1023.11**	1173.90	1175.32	1147.19	1148.66	**838.13**	**837.73**
卡尔-安东尼·唐斯	**394.67**	**382.11**	**893.64**	**885.82**	**1019.17**	**1018.14**	**1031.85**	**1031.55**	721.25	721.91
肯巴·沃克	452.34	457.64	1005.01	1008.07	**927.24**	**927.03**	**988.95**	**988.39**	986.18	986.55
凯文·杜兰特	383.84	386.39	1081.21	1082.63	**1005.51**	**1003.00**	939.18	939.73	682.43	684.19
保罗·乔治	**377.04**	**376.63**	922.72	924.32	1044.72	1044.80	**1028.94**	**1026.65**	**729.18**	**728.60**

表 2：本文提出的双重自激模型（DSE）与不包含比赛级和分钟级自激效应的基线模型所得的 WAIC 值。粗体数值对应于双重自激模型优于基线模型的球队/球员。

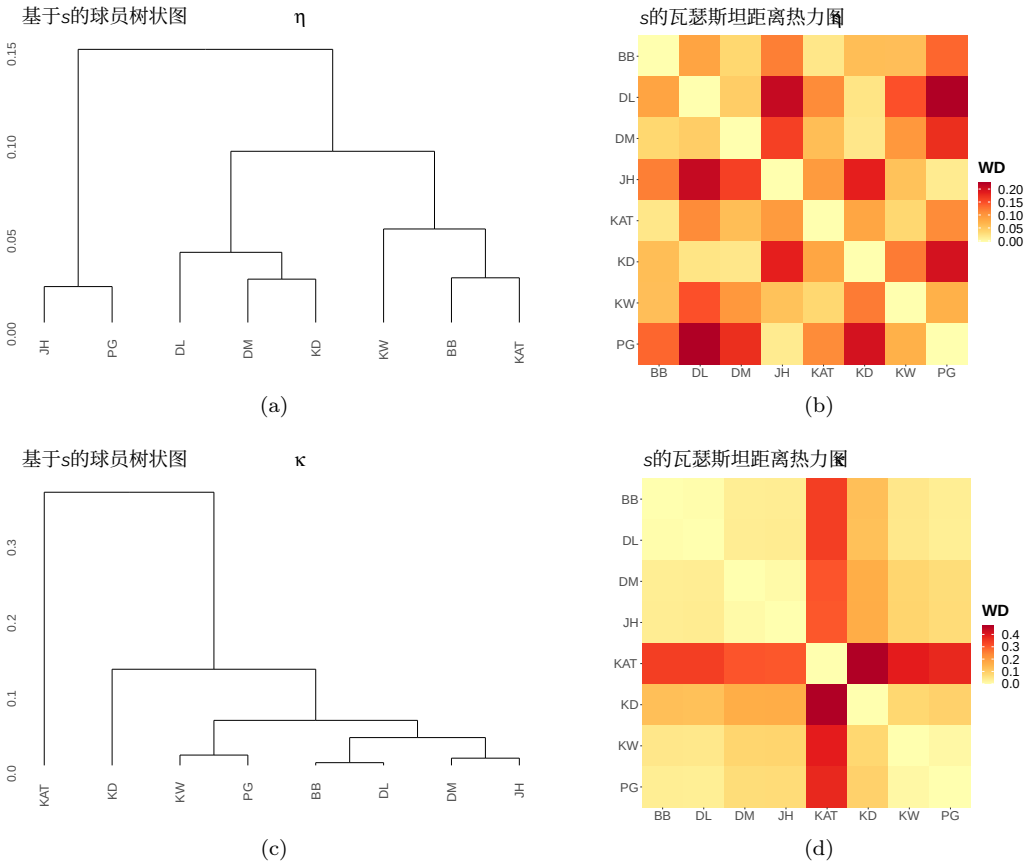


图 12：基于 η_S (a)和 κ_S (c)参数的球员树状图，以及球员之间以 η_S(b)和 κ_S (d)的瓦瑟斯坦距离（WD）表示的差异矩阵（以热力图形式呈现）

数据可用性声明

用于分析的NBA数据集和用于模型拟合的主要R代码将发布在 https://github.com/albrizre/NBA_DSE。

资金支持

本研究未获得公共、商业或非营利部门任何资助机构的专项资助。

利益冲突

作者声明不存在利益冲突。

参考文献

Bollerslev, T. (1986年). 广义自回归条件异方差。《计量经济学杂志》, 31(3):307–327.

Cerqueti, R., D'Urso, P., De Giovanni, L., Mattera, R., and Vitale, V. (2022年). 基于INGARCH的模糊聚类：计数时间序列及其足球应用。《机器学习及其应用》, 第100417页。

Cheng, T. and Adepeju, M. (2014年). 可变时间单元问题（MTUP）及其对时空聚类检测的影响。《*PLOS ONE*》, 9(6):e100465.

Clark, N. J. and Dixon, P. M. (2021年). 一类空间相关自激统计模型。《空间统计学》, 43:100493.

Comte, F. and Lieberman, O. (2000年). 多元GARCH过程中的二阶非因果性。《时间序列分析杂志》, 21(5):535–557.

Cuturi, M. 和 Doucet, A.（2014年）。Wasserstein重心的快速计算。载于《国际机器学习大会》，第685–693页。PMLR。

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., 和 Bodik, R.（2017）。使用模型编程：利用 NIMBLE 为通用模型结构编写统计算法。计算与图形统计学杂志, 26(2)：403–413。

Escudero, I., Angulo, J. M., 和 Mateu, J.（2022年）。一种具有广义自回归条件异方差结构的犯罪计数空间相关模型。*Entropy*，24(7)：892。Ferland, R., Latour, A., 和 Oraichi, D.（2006）。整数值 GARCH 过程。《时间序列分析杂志》，27(6)：923–942。Gilovich, T., Vallone, R., 和 Tversky, A.（1985）。篮球中的热手现象：对随机序列的错误认知。《认知心理学》，17(3)：295–314。Gómez, M. A. 和 Pollard, R.（2011）。欧洲首都球队的篮球主场优势减弱。《欧洲运动科学杂志》，11(2)：143–148。Heinen, A.（2003）。时间序列计数数据建模：自回归条件泊松模型。可在 *SSRN 1117187* 获取。

Horrace, W. C., Jung, H. 和 Sanders, S.（2022年）。NBA中的网络竞争与团队化学反应。《商业与经济统计杂志》，40(1)：35–49。

Hu, G., Yang, H.-C. 和 Xue, Y.（2021年）。职业篮球运动员投篮选择的贝叶斯群组学习。*Stat*，10(1)：e324。

Lantis, R. 和 Nesson, E.（2021年）。热手投篮：对 NBA 投篮和罚球投篮中的“热手现象”的分析。《体育经济学杂志》，22(6)：639–677。

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. 和Tita, G. E.（2011）。犯罪的自激点过程建模。《美国统计学会杂志》，106(493)：100–108。

Ogata, Y. 和 Katsura, K.（1986年）。用于地震数据分析的线性参数化强度点过程模型。《应用概率杂志》，23(A):291–310。Ou, R.、Sen, D. 和 Dunson, D.（2021年）。通过分而治之实现时间序列的可扩展贝叶斯推断。arXiv 预印本 arXiv:2106.11043。Pollard, R. 和 Gómez, M. A.（2007）。根据球队实力分析不同篮球联赛中的主场优势。载于《伊比利亚篮球研究大会》，第4卷，第61–64页。R 核心团队（2021年）。R：一种用于统计计算的语言和环境。

Sandri, M.、Zuccolotto, P.、Manisera, M. 等（2020年）。篮球中投篮表现变异性与队友互动的马尔可夫切换建模。《皇家统计学会会刊：C辑》，69(5):1337–1356。

Santos-Fernandez, E., Denti, F., Mengersen, K., 和 Mira, A.（2022年）。内在维数在高分辨率球员追踪数据中的作用——对篮球的洞见。应用统计年鉴, 16(1):326–348。

Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F., 和 Schmitzer, B.（2022年）。*transport*：最优传输方案和 *Wasserstein*距离的计算。R软件包版本 0.12-4。

Shi, J. 和 Song, K.（2021年）。一种基于离散时间有限状态马尔可夫链的 NBA篮球比赛进行中预测模型。统计通信：模拟与计算, 50(11):3768–3776。

Sokal, R. R. 和Michener, C. D.（1958年）。一种评估系统关系的统计方法。堪萨斯大学科学通报, 38:1409–1438。

Song, K., Zou, Q., and Shi, J.（2020年）。NBA篮球比赛得分和表现统计数据建模。统计通信：模拟与计算，49(10):2604–2616。

Swartz, T. B.（2020年）。我的体育论文应该发表在哪里？美国统计学家，74(2):103– 108。

Tian, X., Gao, Y., and Shi, J.（2020年）。利用逆高斯过程对篮球比赛建模。统计通信：模拟与计算, 第1–11页。

Watanabe, S. and Opper, M.（2010年）。奇异学习理论中贝叶斯交叉验证与广泛适用信息准则的渐近等价性。机器学习研究杂志，11(12)。

Wickham, H.（2016年）。*ggplot2*：优雅的数据分析图形。施普林格纽约出版社。

Yin, F.、Hu, G. 和 Shen, W.（2022年）。运用贝叶斯矩阵聚类方法分析职业篮球投篮出手情况。计算与图形统计学杂志，第 1–23 页。

Yin, F.、Jiao, J.、Hu, G. 和 Yan, J.（2020年）。空间齐次点过程的贝叶斯非参数估计：NBA 投篮位置的空间分析。*arXiv* 预印本 *arXiv:2011.11178*。

You, K.（2022年）。*T4transport*：计算最优传输工具。R软件包版本 0.1.1。