# RECREATING THE 2021/22 PREMIER LEAGUE SEASON WITH POISSON PROCESSES


Josiah Wu

Mathematics Student at Imperial College London
Y1 Summer Research Project


## Introduction

- The English Premier League is a well-known professional football league in the UK.
- A Premier League season consists of 20 teams, each playing 38 matches against each other in a round-robin format.
- This project aims to generate simulations to reproduce the 2021/22 Premier League season, due to my dissatisfaction with the disastrous performance of my favourite team (Manchester United) in that season.

## What is a Poisson Process?

The simulations in this project involve modelling with a Stochastic Process called **one-dimensional Poisson Process**. A one-dimensional Poisson Process is a **Counting Process** $\{N(t), t \geq 0\}$, a collection of random variables in which for all $s, t \geq 0$ [1]:

- $N(t) \geq 0$
- $N(t)$ is an integer.
- $s \leq t \implies N(s) \leq N(t)$

A one-dimensional Poisson Process is therefore a Counting Process $\{N(t), t \geq 0\}$ with arbitrary intensity $\lambda(t) > 0$ such that [2]:

- $N(0) = 0$
- For any $m \in \mathbb{N},\ t_0 < t_1 < \cdots < t_m \implies N(t_1) - N(t_0),\ N(t_2) - N(t_1),\ \cdots,\ N(t_m) - N(t_{m-1})$ are independent random variables.
- The number of events occurring within any interval $A = (a, b]$ is randomly distributed by $\mathrm{Poi}(\Lambda)$, where $\Lambda = \int_a^b \lambda(t)dt$.
- If $\lambda(t) = \lambda_0 \ \ \forall t \geq 0$, i.e. the Poisson Process is **homogeneous**, then the time difference between any two consecutive events is randomly distributed by $\mathrm{Exp}(\lambda_0)$.

Note that if $\lambda(t)$ is not constant over time then we say the Poisson Process is **inhomogeneous**.

## Project Outline

As there is no open dataset available on the internet, I web-scraped information for all the goals scored in the 2021/22 Premier League season from the Premier League official website [3], using a Python package named `Selenium`. (If you're interested in how any of this works, please visit the code on my GitHub page. [a])

In this project, I used the data I web-scraped to find the best-fit intensity $\lambda(t)$ for each team. This then can be used to simulate goal-scoring through an inhomogeneous Poisson Process. Finally, one can find the final result of a simulated season by comparing the number of goals scored in matches.

## References


[1] Ross, Sheldon. 1995. Chapter 2: The Poisson Process. Stochastic Process, 2nd Edition. Wiley, pp 59-60.

[2] Tijms, Henk C. 2003. Chapter 1: The Poisson Process and Related Processes. A First Course in Stochastic Models. John Wiley and Sons, pp 2,3,22.

[3] Results. Premier League Official Website. https://www.premierleague.com/results?co=1&se=418&cl=-1

[4] Daley, Daryl J. and Vere-Jones, David. 2003. Chapter 7: Conditional Intensities and Likelihoods. An introduction to the theory of point processes, Volume I, Elementary theory and methods, 2nd Edition. New York, Springer, pp 213.

[5] Stats (v3.6.2) - optim: General-purpose Optimization. RDocumenation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/optim

[6] Chen, Yuanda. 2016. Thinning Algorithm for Simulating Poisson Processes. https://www.math.fsu.edu/~ychen/research/Thinning%20algorithm.pdf

[a] GitHub Repository - https://github.com/jwu29/21-22PLSimulation


## Can goal-scoring be modelled by a homogeneous Poisson Process?

After having collected the data, I suspected that the intensity $\lambda(t)$ is constant. To see if this is true, I first recorded the number of goals scored per match by each team in the 2021/22 season. Then I applied the **Chi-Squared Goodness of Fit Test**, a hypothesis test with the following hypotheses:

- $H_0$: Goals scored per match is randomly distributed by $\mathrm{Poi}(\lambda_0)$ for some $\lambda_0 \in \mathbb{R}$
- $H_1$: Goals scored per match isn't randomly distributed by $\mathrm{Poi}(\lambda_0)$ for all $\lambda_0 \in \mathbb{R}$

If the intensity is indeed constant, then $H_0$ should be concluded for the test. However, looking at the p-values for all 20 teams suggests that there isn't sufficient evidence to conclude $H_0$ for most teams.

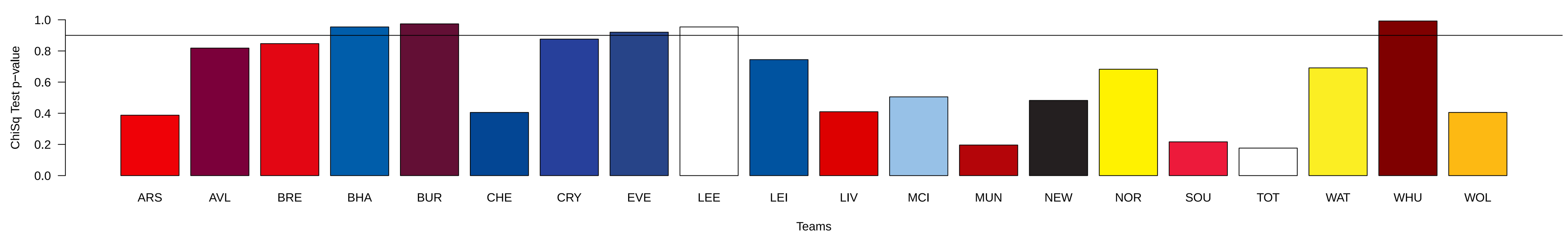


Figure 2: Chi-Squared Test p-values for all teams; the horizontal line represents $p = 0.90$. Note that $H_0$ is accepted when $p \geq 0.90$, and rejected otherwise.

## Modelling the Intensity Function

From the last section we can conclude that the intensity function cannot be constant. However, I made an educated guess that every team's intensity function is periodic every match, and is a multiple of the PDF for the Beta distribution. The resulting intensity function is

$$\lambda(t; \alpha, \beta, \theta) = \theta\Big[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(t \bmod 1)^{\alpha-1}(1-(t \bmod 1))^{\beta-1}\Big]$$

It can be shown that for any intensity $\lambda(t) \in \mathbb{R}$, the log-likelihood of intensity [4] is given by

$$log\big[L(x_1, \cdots, x_n)\big] = \sum_{i=1}^{n} log\ \lambda(x_i) - \int_A \lambda(s)ds$$

By using the built-in `optim` function in R, it's possible to find optimal values of $\alpha, \beta, \theta$ which maximizes the log-likelihood [5].

## How to simulate one-dimensional Poisson Processes

Simulating a one-dimensional homogeneous Poisson Process (1D-HPP) within the interval $(0, t]$ requires drawing samples $(x_1, \cdots, x_k)$ from $X \sim \mathrm{Exp}(\lambda_0)$, where $k = sup\{\sum_{n=1}^{q} x_n \leq t : q \in \mathbb{Z}^+\}$. Then a realization of 1D-HPP is denoted by $\vec{v} = (y_1, y_2, \cdots, y_k) = (x_1, x_1 + x_2, \cdots, \sum_{n=1}^{k} x_n)$ [6].

However, simulating a 1D inhomogeneous Poisson Process (1D-IHPP) is not as straight forward. It involves an algorithm called the **thinning algorithm** [6], which works if $\lambda(t)$ is bounded above:

1. Let $M = sup_{t\geq 0}\ \lambda(t)$. Simulate a 1D-HPP in $(0, t]$ with intensity $M$, yielding $\vec{v} = (y_1, \cdots, y_k)$.
2. Draw sample $(u_1, \cdots, u_k)$ from $\mathrm{Unif}(0, 1)$; for $1 \leq i \leq k$, if $u_i > \lambda(y_i)/M$, delete $y_i$ from $\vec{v}$; else, continue.
3. Return $\vec{v}$.

## Putting it all together...

We can now use the optimal parameters to simulate results of 38 matches for all 20 teams in the 2021/22 season. Figure 3 depicts the result of one particular simulation.

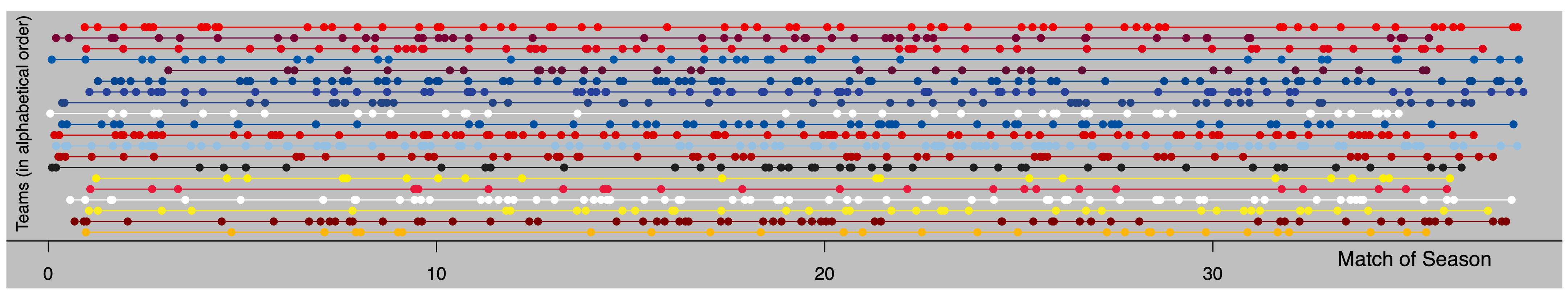


Figure 3: A graph showing all goals scored in a simulated 21/22 Premier League season, where the dots represent the goals.

I then used the results to generate the final league table for the 2021/22 season 100 times. It turns out the model is accurate in predicting the top four teams in the league, but not for other positions. However, I believe the model could be improved by making the multiplier $\theta$ in the intensity function more dependent on the opponent - for example, if a team faces a difficult opponent then the multiplier $\theta$ should be smaller.

| | Champions | 2nd Place | 3rd Place | 4th Place | 5th Place | 6th Place | 7th Place | 8th Place | 9th Place | 10th Place |
|---|---|---|---|---|---|---|---|---|---|---|
| Projected Standings | Man City (51%) | Liverpool (38%) | Chelsea (33%) | Tottenham (25%) | Arsenal (23%) | Leicester (14%) | Man United (11%) | Man United (16%) | Cry. Palace (15%) | Cry. Palace (12%) |
| | Liverpool (42%) | Man City (34%) | Tottenham (12%) | Chelsea (19%) | Tottenham (17%) | Tottenham (13%) | Arsenal (11%) | Arsenal (11%) | West Ham (11%) | Everton (11%) |
| | Chelsea (2%) | Leicester (5%) | Leicester (11%) | West Ham (13%) | Leicester (15%) | West Ham (12%) | Leicester (11%) | Aston Villa (11%) | United (10%) | Arsenal (11%) |
| Actual Standings | Man City | Liverpool | Chelsea | Tottenham | Arsenal | Man United | West Ham | Leicester | Brighton | Wolves |

Figure 4: Table of Projected Standings. See [a] for complete table.