Assessing competitive balance in the English premier league using a stochastic block model

Nial Friel
@nialfriel
nial.friel@ucd.ie
University College Dublin

Co-authors:
Francesca Basini (Warwick), Vasiliki Tsouli, Ioannis Ntzoufras (Athens).

Outline

← Thread

Bayern & Germany
@iMiaSanMia

The discussion about introducing a play-off system in the Bundesliga in order to bring back excitement to the title race has picked up speed again [Kicker]

20:51 · 09 Feb 22 · Twitter for Android

46 Retweets **204** Quote Tweets **1,231** Likes

Bayern & Germany @iMiaSan... · 1d
Replying to @iMiaSanMia
New DFL CEO Donata Hopfen told Bild: "Of course, the league would be more attractive if it had more competition at

Tweet your reply

Thread

Bayern & Germany
@iMiaSanMia

The discussion about introducing a play-off system in the Bundesliga in order to bring back excitement to the title race has picked up speed again [Kicker]

20:51 · 09 Feb 22 · Twitter for Android

46 Retweets **204** Quote Tweets **1,231** Likes

Bayern & Germany @iMiaSan... · 1d
Replying to @iMiaSanMia
New DFL CEO Donata Hopfen told Bild: "Of course, the league would be more attractive if it had more competition at

Tweet your reply

404 Aa

How can we make Europe's big leagues more competitive?

Michael Cox
Wed, 26 Jan

There are still four months remaining in most European domestic leagues, but the big titles are largely already decided.

It's a situation we've become accustomed to: the rich clubs wrapping up the league by the turn of the year, allowing

European Super League offends principles of competition - Boris Johnson

20 April

GETTY IMAGES

Fans protested against the proposed league outside Chelsea's Stamford Bridge stadium on Tuesday

Plans for a European Super League offend "the basic principles of competition", Boris Johnson has said.

English First division/Premier League – a brief history

- The English football league is one of the oldest football leagues in the world dating back to 1888.
- It grew rapidly with the introduction of a second division in 1892 and today consists of 4 divisions.
- There has been many changes, but perhaps the most significant has been the introduction of the English Premier League in 1992/93.
- This heralded massive increases in revenue. The first TV deal between the Premier League and the television companies generated revenue of around 40 million pounds. This has increased dramatically to 5.14 billion between 2016 and 2019.

An introduction

- Competitive balance is a desirable feature in any professional sports league and encapsulates the notion that there is unpredictability in the outcome of games.
- We develop a stochastic block model framework to facilitate the probabilistic clustering so that teams within a blocks are balanced.
- A key question is assessing the uncertainty in the number of blocks and estimation of the partition or allocation of teams to blocks.
- We apply our model to each season in the English premier league from 1978/79 to 2020/21.
- A key finding of this analysis is evidence which suggests a structural change from a reasonably balanced league to a two-tier league which occurred around the early 2000's.

Herfindahl–Hirschman index of competitive balance

- HHICB is based on assessing a measure of the spread of points share in a given season.
- Suppose that team i scored s_i points over the course of a season in a league involving n teams.
- Define $p_i := s_i / \sum_1^n s_i$ to be the proportion of points achieved by team i.

$$\text{HHICB} = n \sum_{i=1}^{n} p_i^2.$$

- It is therefore simply a measure of the spread of $(p_1, \ldots, p_n)$.
- When each team has an identical proportion of points so that $p_i = 1/n$, then HICB=1.

Relative entropy an index of competitive balance

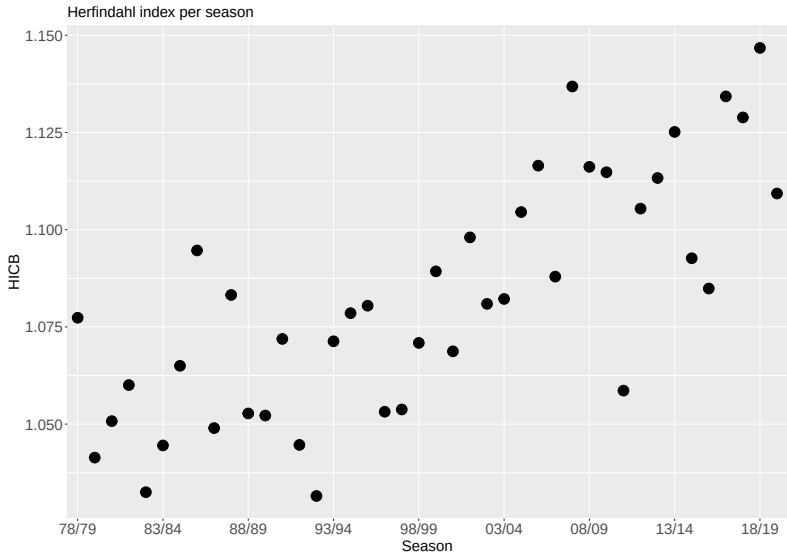
- A natural approach to summarise the proportion of points share among all n teams in a league, $(p_1, \ldots, p_n)$, is to use the concept of entropy, eg, relative entropy:

$$\frac{\sum_{i=1}^{n} p_i \log(p_i)}{\log(1/n)}.$$

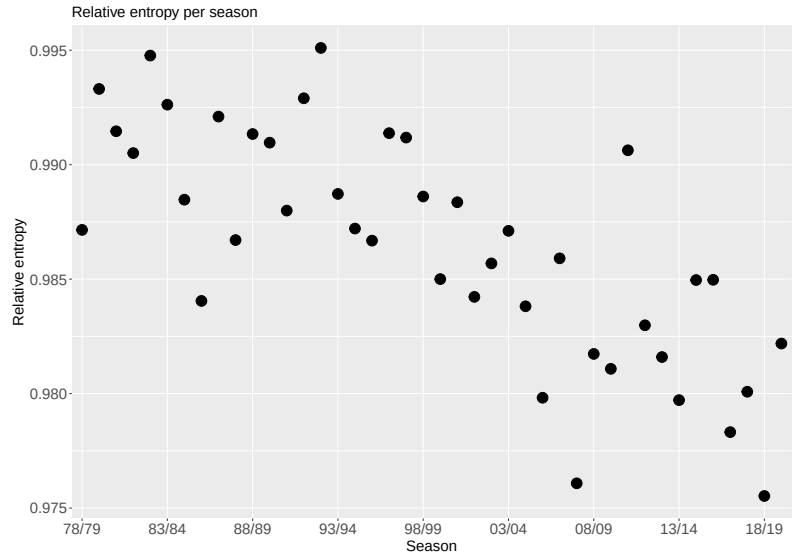
- This statistic takes a maximum value of 1 in the case where p_i, the proportion of points share for team i is $1/n$.

- Lower values of relative entropy suggest a more imbalanced league.

HHICB and Relative entropy: Season 1978/79 to 2019/20.



(a)



(b)

(a) HHICB: increasing over time and so consistent with the hypothesis that the premier league has become more imbalanced over time.

(b) Relative entropy: High values of relative entropy correspond to more balanced league. Again, there is evidence that the premier league has become more imbalanced over time.

Drawbacks of HHCBI, Relative entropy and other approaches

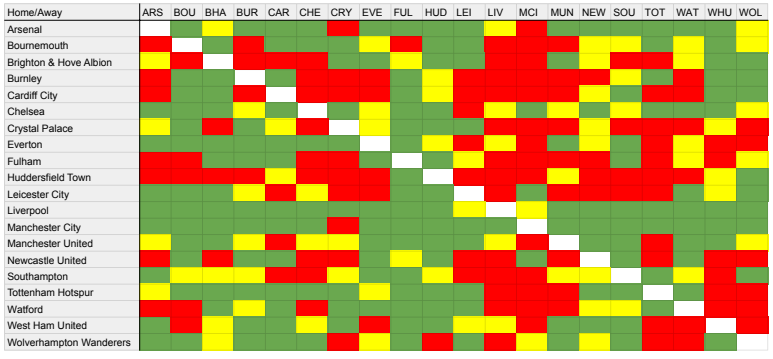
- Limitiations of the previous approaches: they are simply univariate statistics which are not amenable to qualitative conclusions about whether a season is balanced.
- Moreover, if there is evidence to suggest that a season is imbalanced, it would be useful to give an indication of the nature of this imbalance, eg, which teams, if any, are stronger than the rest.
- In short, these approaches do not explicitly model the relational nature of data arising from a league.
- The network model which we propose yields information about the quality of the leagues and the relative competitiveness.
- Generally, the method we propose is more computationally demanding but richer in terms of results and inference we obtain from the final output it offers.

Representing the outcome of a season as a results matrix

Season 2018/19

Home\Away	ARS	BOU	BHA	BUR	CAR	CHE	CRY	EVE	FUL	HUD	LEI	LIV	MCI	MUN	NEW	SOU	TOT	WAT	WHU	WOL
Arsenal	~	5~1	1~1	3~1	2~1	2~0	2~3	2~0	4~1	1~0	3~1	1~1	0~2	2~0	2~0	2~0	4~2	2~0	3~1	1~1
Bournemouth	1~2	~	2~0	1~3	2~0	4~0	2~1	2~2	0~1	2~1	4~2	0~4	0~1	1~2	2~2	0~0	1~0	3~3	2~0	1~1
Brighton & Hove Albion	1~1	0~5	~	1~3	0~2	1~2	3~1	1~0	2~2	1~0	1~1	0~1	1~4	3~2	1~1	0~1	1~2	0~0	1~0	1~0
Burnley	1~3	4~0	1~0	~	2~0	0~4	1~3	1~5	2~1	1~1	1~2	1~3	0~1	0~2	1~2	1~1	2~1	1~3	2~0	2~0
Cardiff City	2~3	2~0	2~1	1~2	~	1~2	2~3	0~3	4~2	0~0	0~1	0~2	0~5	1~5	0~0	1~0	0~3	1~5	2~0	2~1
Chelsea	3~2	2~0	3~0	2~2	4~1	~	3~1	0~0	2~0	5~0	0~1	1~1	2~0	2~2	2~1	0~0	2~0	3~0	2~0	1~1
Crystal Palace	2~2	5~3	1~2	2~0	0~0	0~1	~	0~0	2~0	2~0	1~0	0~2	1~3	1~3	0~0	0~2	0~1	1~2	1~1	0~1
Everton	1~0	2~0	3~1	2~0	1~0	2~0	2~0	~	3~0	1~1	0~1	0~0	0~2	4~0	1~1	2~1	2~6	2~2	1~3	1~3
Fulham	1~5	0~3	4~2	4~2	1~0	1~2	0~2	2~0	~	1~0	1~1	1~2	0~2	0~3	0~4	3~2	1~2	1~1	0~2	1~1
Huddersfield Town	1~2	0~2	1~2	1~2	0~0	0~3	0~1	0~1	1~0	~	1~4	0~1	0~3	1~1	0~1	1~3	0~2	1~2	1~1	1~0
Leicester City	3~0	2~0	2~1	0~0	0~1	0~0	1~4	1~2	3~1	3~1	~	1~2	2~1	0~1	0~1	1~2	0~2	2~0	1~1	2~0
Liverpool	5~1	3~0	1~0	4~2	4~1	2~0	4~3	1~0	2~0	5~0	1~1	~	0~0	3~1	4~0	3~0	2~1	5~0	4~0	2~0
Manchester City	3~1	3~1	2~0	5~0	2~0	6~0	2~3	3~1	3~0	6~1	1~0	2~1	~	3~1	2~1	6~1	1~0	3~1	1~0	3~0
Manchester United	2~2	4~1	2~1	2~2	0~2	1~1	0~0	2~1	4~1	3~1	2~1	0~0	0~2	~	3~2	3~2	0~3	2~1	2~1	1~1
Newcastle United	1~2	2~1	0~1	2~0	3~0	1~2	0~1	3~2	0~0	2~0	0~2	2~3	2~1	0~2	~	3~1	1~2	1~0	0~3	1~2
Southampton	3~2	3~3	2~2	0~0	1~2	0~3	1~1	2~1	2~0	1~1	1~2	1~3	1~3	2~2	0~0	~	2~1	1~1	1~2	3~1
Tottenham Hotspur	1~1	5~0	1~0	1~0	1~0	3~1	2~0	2~2	3~1	4~0	3~1	1~2	0~1	0~1	1~0	3~1	~	2~1	0~1	1~3
Watford	0~1	0~4	2~0	0~0	3~2	1~2	2~1	1~0	4~1	3~0	2~1	0~3	1~2	1~2	1~1	1~1	2~1	~	1~4	1~2
West Ham United	1~0	1~2	2~2	4~2	3~1	0~0	3~2	0~2	3~1	4~3	2~2	1~1	0~4	3~1	2~0	3~0	0~1	0~2	~	0~1
Wolverhampton Wanderers	3~1	2~0	0~0	1~0	2~0	2~1	0~2	2~2	1~0	0~2	4~3	0~2	1~1	2~1	1~1	2~0	2~3	0~2	3~0	~

(a)



(b)

(a) Cell entries correspond to the result when a home team (row) plays an away team (column).

(b) The results summarised in a results matrix categorising each result as a win, draw or loss.

Representing the outcome of a season as a results matrix

We denote the three categorical variables, "*win*", "*draw*" or "*loss*", by $1, 2$ or 3, respectively.

This leads us to summarise the outcome from a season with a matrix $\mathbf{y}$ which we term a results matrix:

$$\mathbf{y} = \begin{pmatrix} - & y_{12} & \dots & y_{1j} & \dots & y_{1N} \\ y_{21} & - & \dots & y_{2j} & \dots & y_{2N} \\ \dots & \dots & - & \dots & \dots & \\ y_{i1} & y_{i1} & \dots & - & \dots & y_{iN} \\ \dots & \dots & \dots & \dots & - & \dots \\ y_{N1} & y_{N2} & \dots & y_{Nj} & \dots & - \end{pmatrix},$$

where $y_{ij} \in \{1, 2, 3\}, \quad \text{for } i, j = 1, \dots, N; i \neq j.$

Results matrix as an adjacency matrix

- The results matrix $\mathbf{y}$ can be considered as an adjacency matrix of a directed network.
- Each node represents a team and an edge from node i to node j represents the result of the match when team i plays at home against team j, where the edge takes a value in the set $\{1, 2, 3\}$.
- Here the network is dense and complete. This contrasts with the usual sparsity observed in social networks, eg.

This observation is useful as it allows to appeal to statistical models in social network analysis to find structure in the results matrix $\mathbf{y}$.

Outline

Background

A stochastic block model to assess competitive balance

Analysis of over 40 seasons of the English premier league

Erdös-Renyi random graph model

- A simple model for binary random graphs which assumes that $Y_{ij} \sim Bern(p)$. In other words, all edges are assumed to be independent Bernoulli random variables.
- The number of observed edges, E, is a sufficient statistic.
- The likelihood is expressed as

$$f(y|p) = \prod_{i<j} p^{y_{ij}}(1-p)^{1-y_{ij}} = p^{E}(1-p)^{\frac{N(N-1)}{2}-E}$$

Stochastic block models

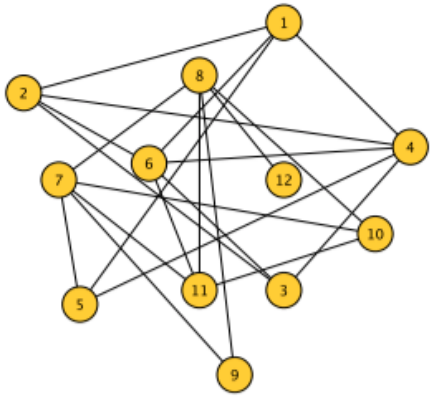
- The Stochastic Block Model (SBM) proposed by (Nowicki and Snijeders, 2001) is a finite mixture model for network data.
- An SBM stochastically paritions the nodes into k blocks. Nodes within each block are modelled as a Erdös-Renyi random graph.
- The nodes are characterised by a latent cluster membership variable.
- The probability of observing a certain edge is determined only by the clustering variables of the two nodes.

Stochastic block models: Model specification

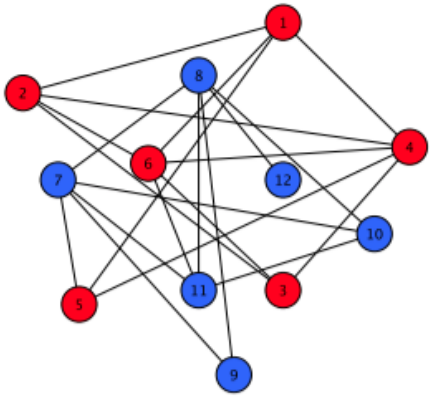
- The SBM assumes that the nodes are partitioned in k clusters.
- The cluster membership of each node i is denoted by $z_i \in \{1, \ldots, K\}$ where $P(z_i = g) = \alpha_g, \;\; g = 1, \ldots, k$.
- Given $z_i = g$ and $z_j = h$, the probability of observing an edge y_{ij} is given by $p_{gh} \in [0, 1]$.
- The $k \times k$ matrix of block connection probabilities is denoted by $P = \{p_{gh}\}$.
- Given $z_i = g$ and $z_j = h$, an edge is then drawn from a Bernoulli distribution with probabililty p_{gh}:

$$y_{ij} | (z_i = g, z_j = h) \sim Bern(p_{gh}).$$

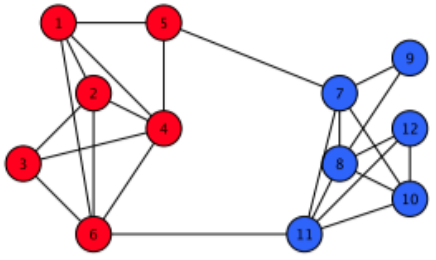
An example of an SBM



(a) Observation X^n



(b) Unobserved communities



(c) Community detection

One of the thing that you don't know... is the number of things that you don't know!

From a statistical perspective a source of uncertainty is the number of blocks/clusters, k!

- Statisticians are used to dealing with this issue! (Although it is non-trivial in the case of SBMs).
- In principle, one could apply a raft of well-studied approaches:
  - Expectation-Maximisation
  - Reversible-jump MCMC, alá Richardson and Green (1998).
  - ...

Objective of our SBM

- The aim of our stochastic block model is to partition the N teams in a league, into K blocks in such a way that the probability of a win, draw or loss for the home team is, broadly speaking, similar when any two teams in the same block play against one and other.

- But equally, that the probability outcome when any team from one block plays at home against another team from a different block also has a similar probability of a win, draw or a loss.

- Crucially, the probability of a given outcome depends on the blocks to which each team are assigned. One of the key objective is to infer the most likely value of K. In particular, if we deem that $K = 1$ has most support, then we have some evidence that the league is balanced.

Stochastic block model: two main assumptions

- For a K block (or cluster) model, each node (or team), i for $i = 1, \ldots, N$, belongs to one of the blocks with membership or allocation label, $z_i \in \{1, \ldots, K\}$.

- The distribution of $\mathbf{y} = (\mathbf{y_{ij}})_{1 \leq i \neq j \leq N}$ is assumed to be conditionally independent given the latent variable of cluster memberships, $\mathbf{z} := (z_1, \ldots, z_N)$.

Multinomial likelihood model

- Suppose $z_i = k$ and $z_j = l$.
- Let's suppose that the probabililty of a win,draw or a loss when a team in block k plays at home against a team in block l is

$$\underline{p}^{kl} = (p_1^{kl}, p_2^{kl}, p_3^{kl}).$$

- We then model the outcome y_{ij} as a multinomial random variable, $Multi(y_{ij}; 1, \underline{p}^{kl})$,

$$P(y_{ij} = \omega | z_i = k, z_j = l, \underline{p}^{kl}, K) = p_\omega^{kl}, \quad \omega = 1, 2, 3.$$

Multinomial likelihood model

- Suppose $z_i = k$ and $z_j = l$.
- Let's suppose that the probabililty of a win,draw or a loss when a team in block k plays at home against a team in block l is

$$\underline{p}^{kl} = (p_1^{kl}, p_2^{kl}, p_3^{kl}).$$

- We then model the outcome y_{ij} as a multinomial random variable, $Multi(y_{ij}; 1, \underline{p}^{kl})$,

$$P(y_{ij} = \omega | z_i = k, z_j = l, \underline{p}^{kl}, K) = p_\omega^{kl}, \quad \omega = 1, 2, 3.$$

- Eg, team i, Spurs are assigned to block 1 and play at home against team j, Crystal Palace who are assigned to block 3.
- Suppose that $\underline{p}^{13} = (0.6, 0.3, 0.1)$.
- Data: **Spurs** $2 - 0$ **Crystal Palace**.

$$f(y_{ij} = 1 | z_i = 1, z_j = 3, \underline{p}^{13}, K) = 0.6.$$

Distribution for the allocation vector $\mathbf{z}$

We assume that the entries of $\mathbf{z}$ are independent and identically distributed following a multinomial distribution:

$$z_i|\theta, K \overset{iid}{\sim} Multi(1, \theta = (\theta_1, \theta_2, \ldots, \theta_K)), \text{ for } i = 1, \ldots, N,$$

where $P(z_i = k|\theta, K) = \theta_k$ is the probability that node i belongs to cluster k, $\theta_k > 0$, $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \theta_k = 1$.

Thus, the distribution of the partition of the N nodes into K clusters conditional on $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$ is:

$$\pi(\mathbf{z}|\theta, K) = \prod_{i=1}^{N} Multi(z_i; 1, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{I(z_i=k)}.$$

Prior for θ:

We assume a vague conjugate prior for the vector θ following a Dirichlet distribution of dimension K with vector of concentration parameters γ:

$$\theta|K \sim Dir(\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k, \ldots, \gamma_K)).$$

We set all concentration parameters equal to $\gamma_0 = 1$ yielding a uniform prior.

Blocks interaction probabilities:

$$\mathbf{p} = \begin{pmatrix} \underline{p}^{11} & \underline{p}^{12} & \dots & \underline{p}^{1l} & \dots & \underline{p}^{1K} \\ \underline{p}^{21} & \underline{p}^{22} & \dots & \underline{p}^{2l} & \dots & \underline{p}^{2K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \underline{p}^{k1} & \underline{p}^{k2} & \dots & \underline{p}^{kl} & \dots & \underline{p}^{kK} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \underline{p}^{K1} & \underline{p}^{K2} & \dots & \underline{p}^{Kl} & \dots & \underline{p}^{KK} \end{pmatrix}$$

where,

$$\underline{p}^{kl} = (p_1^{kl}, p_2^{kl}, p_3^{kl}) \text{ and } \sum_{\omega=1}^{3} p_\omega^{kl} = 1,$$

for all $k = 1, \dots, K$ and $l = 1, \dots, K$.

As before, p_1^{kl}, p_2^{kl} or p_3^{kl} is the probability that a team allocated to block k playing at home against a team allocated to block l, wins, draws or loses, respectively.

Distribution of the relational pattern of $\mathbf{y}$:

We model the observation y_{ij} conditional on the latent allocations z_i, z_j as a multinomial distribution,

$$\begin{aligned} f(y_{ij}|z_i, z_j, \mathbf{p}, K) &= \prod_{k=1}^{K}\prod_{l=1}^{K} Multi(y_{ij}; 1, \underline{p}^{kl})^{I(z_i=k)I(z_j=l)} \\ &= \prod_{k=1}^{K}\prod_{l=1}^{K}\left\{\prod_{\omega=1}^{3}\left(p_{\omega}^{kl}\right)^{I(y_{ij}=\omega)}\right\}^{I(z_i=k)I(z_j=l)} \end{aligned}$$

for $i, j = 1, \ldots, N,\ i \neq j$.

Thus,

$$\begin{aligned} f(\mathbf{y}|\mathbf{z}, \mathbf{p}, K) &= \prod_{i=1}^{N-1}\prod_{\substack{j=1 \\ j\neq i}}^{N} f(y_{ij}|z_i, z_j, \mathbf{p}, K) \\ &= \prod_{i=1}^{N-1}\prod_{\substack{j=1 \\ j\neq i}}^{N}\prod_{k=1}^{K}\prod_{l=1}^{K}\left\{\prod_{\omega=1}^{3}\left(p_{\omega}^{kl}\right)^{I(y_{ij}=\omega)}\right\}^{I(z_i=k)I(z_j=l)} . \quad (1) \end{aligned}$$

Prior for the block interaction probabilities:

We assume that the entries of $\mathbf{p}$ are mutually independent and that each $\underline{p}^{kl}$ follows a conjugate prior from a 3-dimensional Dirichlet distribution:

$$\underline{p}^{kl} \sim Dir(\beta = (\beta_1, \beta_2, \beta_3)), \quad \text{for } k = 1, \ldots, K, \quad \text{and } l = 1, \ldots, K.$$

We set all the hyperparameters $\beta_1, \beta_2, \beta_3$ to 1 leading to a uniform distribution.

Prior for K:

We treat the number of blocks or clusters as a random variable and choose a probability mass function for K which is distributed as a zero-truncated Poisson random variable with $\lambda = 1$ restricted to $1 \leq k \leq K_{max}$, where K_{max} is an user specified upper limit on the plausible number of blocks.

$$\pi(K|K > 0) = \frac{Poi(1)}{1 - Poi(K = 0)} = \frac{1}{K!(e-1)}. \tag{2}$$

Therefore this prior probability mass function is proportional to $\frac{1}{K!}$.

Bayesian model

We can write the joint posterior distribution as

$$\pi(\mathbf{z}, \mathbf{p}, \theta, K|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{p}, \mathbf{z}, K)\pi(\mathbf{p}|K)\pi(\mathbf{z}|\theta, K)\pi(\theta|K)\pi(K).$$

We are mainly interested in the marginal distribution of the latent allocation vector and number of blocks:

$$\begin{aligned}
\pi(\mathbf{z}, K|\mathbf{y}) &= \int_{\Theta}\int_{\mathbf{P}} f(\mathbf{y}|\mathbf{p}, \mathbf{z}, K)\pi(\mathbf{p}|K)\pi(\mathbf{z}|\theta, K)\pi(\theta|K)\pi(K)d\theta d\mathbf{p} \\
&= \int_{\mathbf{P}} f(\mathbf{y}|\mathbf{p}, \mathbf{z}, K)\pi(\mathbf{p}|K)d\mathbf{p} \times \int_{\Theta} \pi(\mathbf{z}|\theta, K)\pi(\theta|K)d\theta \times \pi(K) \\
&= f(\mathbf{y}|\mathbf{z}, K)\pi(\mathbf{z}|K)\pi(K).
\end{aligned}$$

We can analytically integrate both expressions above because of conjugacy.

Bayesian model

$$\pi(\mathbf{z}, K|\mathbf{y}) \propto \prod_{k=1}^{K}\prod_{l=1}^{K}\Gamma(3)\frac{\prod_{\omega=1}^{3}\Gamma(N_{kl}^{\omega}+1)}{\Gamma(\sum_{\omega=1}^{3}(N_{kl}^{\omega}+1))}\cdot\prod_{k=1}^{K}\Gamma(n_k+1)\frac{\Gamma(K)}{\Gamma(N+K)}\times\frac{1}{K!},$$

where we define

$$N_{kl}^{\omega} = \sum_{i=1}^{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} I(y_{ij}=\omega)I\left(z_i=k\right)I\left(z_j=l\right),$$

for $\omega = 1, 2, 3$ and for $k, = 1, \ldots, K$ and $l, = 1, \ldots, K$. Also,

$$n_k = \sum_{i=1}^{N} I\left(z_i = k\right),\ \ k = 1, \ldots, K.$$

- N_{kl}^{ω} counts the number of times that the outcome ω was observed for all games involving a team allocated to block k playing at home against a team allocated to block l.
- While n_k accounts for the number of teams allocated to block k.

MCMC scheme

The algorithm is based on three move types:

MK: Metropolis move to insert or remove an empty cluster. This move changes the current state of K but not the allocation vector $\mathbf{z}$.

M-GS: Metropolis-within-Gibbs move that updates all components of the allocation vector $\mathbf{z}$ but does not change the number of clusters.

AE: Metropolis-Hastings move to absorb or eject a cluster. This move affects both $\mathbf{z}$ and K.

Label switching

- The allocation vector $\mathbf{z}$ is not identifiable by the model. This is because the likelihood is invariant to permutations of the labels of $\mathbf{z}$.
- We use the relabelling algorithm of Carpeneto and Toth (1980).

Outline

Analysis of over 40 seasons of the English premier league

- Here we analyse each season in turn from 1978/79 to 2020/21.
- Our interest is to explore if the league has become more or less competitive over this time.
- We are also interested to explore the question of whether there is evidence of a *big-six* group of teams.
- To begin we explore in detail the 2018/19 season as this allows us to present some of the salient features of the model.

Analysis of the 2018/19 season

The output of the MCMC algorithm yielded the following estimates of the posterior probability for different values of K.

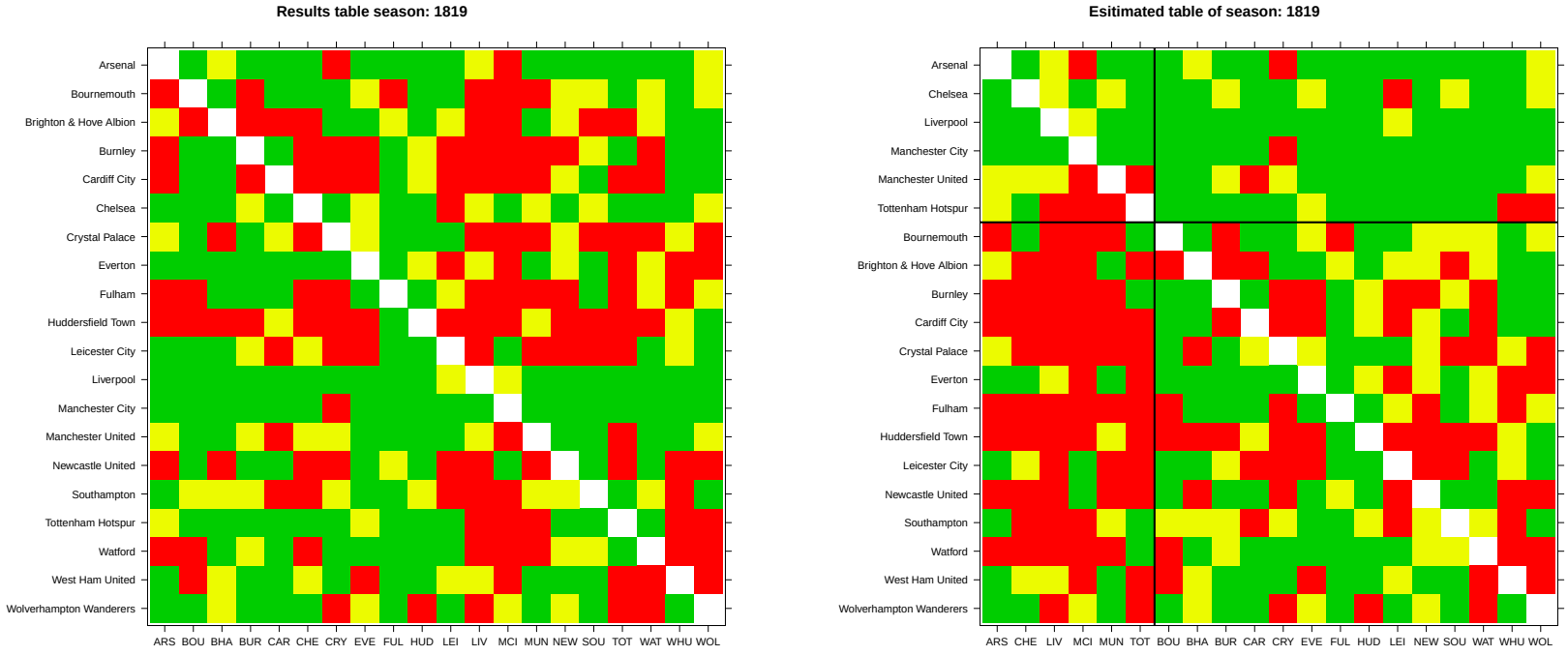
K	1	2	3	4
$P(K\|\text{data})$	0.0	0.97	0.02	0.01

Strong evidence that a 2 block model has most posterior support.

Analysis of the 2018/19 season

	Points	P(top block)
Manchester City	98	0.98
Liverpool	97	0.98
Chelsea	72	0.92
Tottenham Hotspur	71	0.81
Arsenal	70	0.90
Manchester United	66	0.85
Wolverhampton Wanderers	57	0.23
Everton	54	0.07
Leicester City	52	0.01
West Ham United	52	0.02
Watford	50	0.00
Crystal Palace	49	0.00
Bournemouth	45	0.00
Newcastle United	45	0.00
Burnley	40	0.00
Southampton	39	0.00
Brighton & Hove Albion	36	0.00
Cardiff City	34	0.00
Fulham	26	0.00
Huddersfield Town	16	0.00

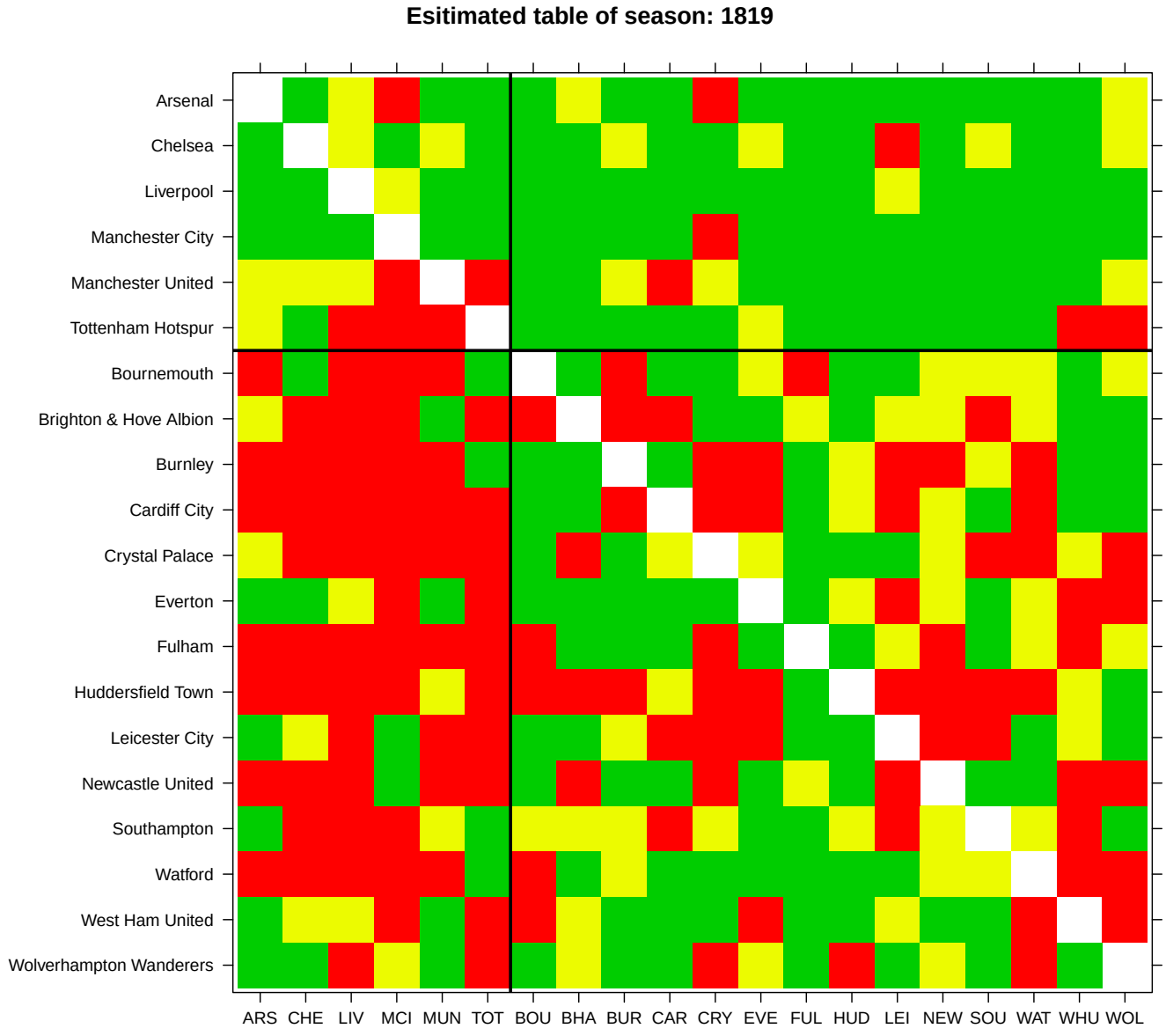
Analysis of the 2018/19 season



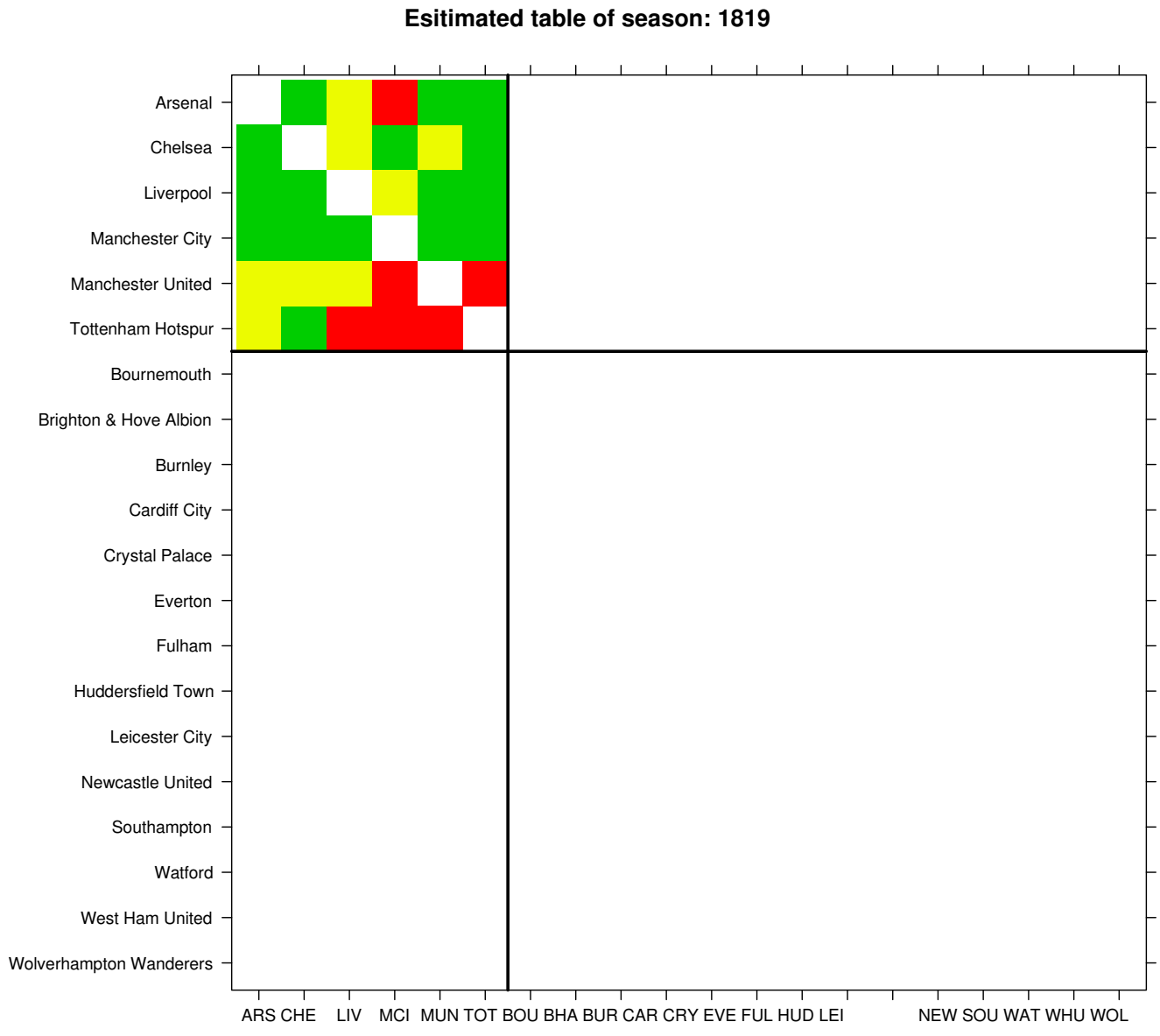
(a) (b)

Figure: (a) Teams listed in alphabetical order. (b) Teams listed by most likely block membership, a posteriori. The solid horizontal and vertical lines (which also coincides with final league position) separates each team in their most likely block.

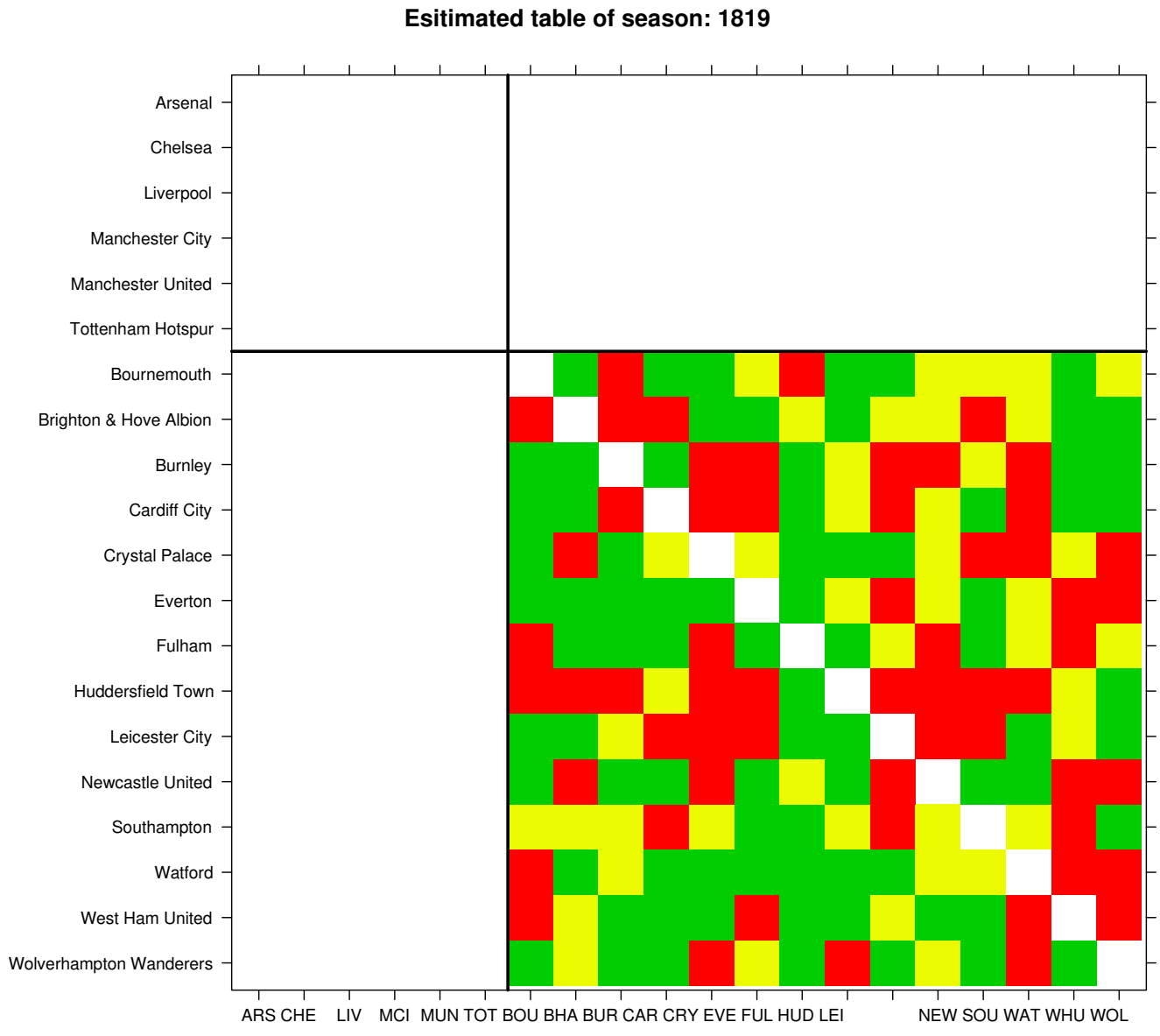
Analysis of the 2018/19 season



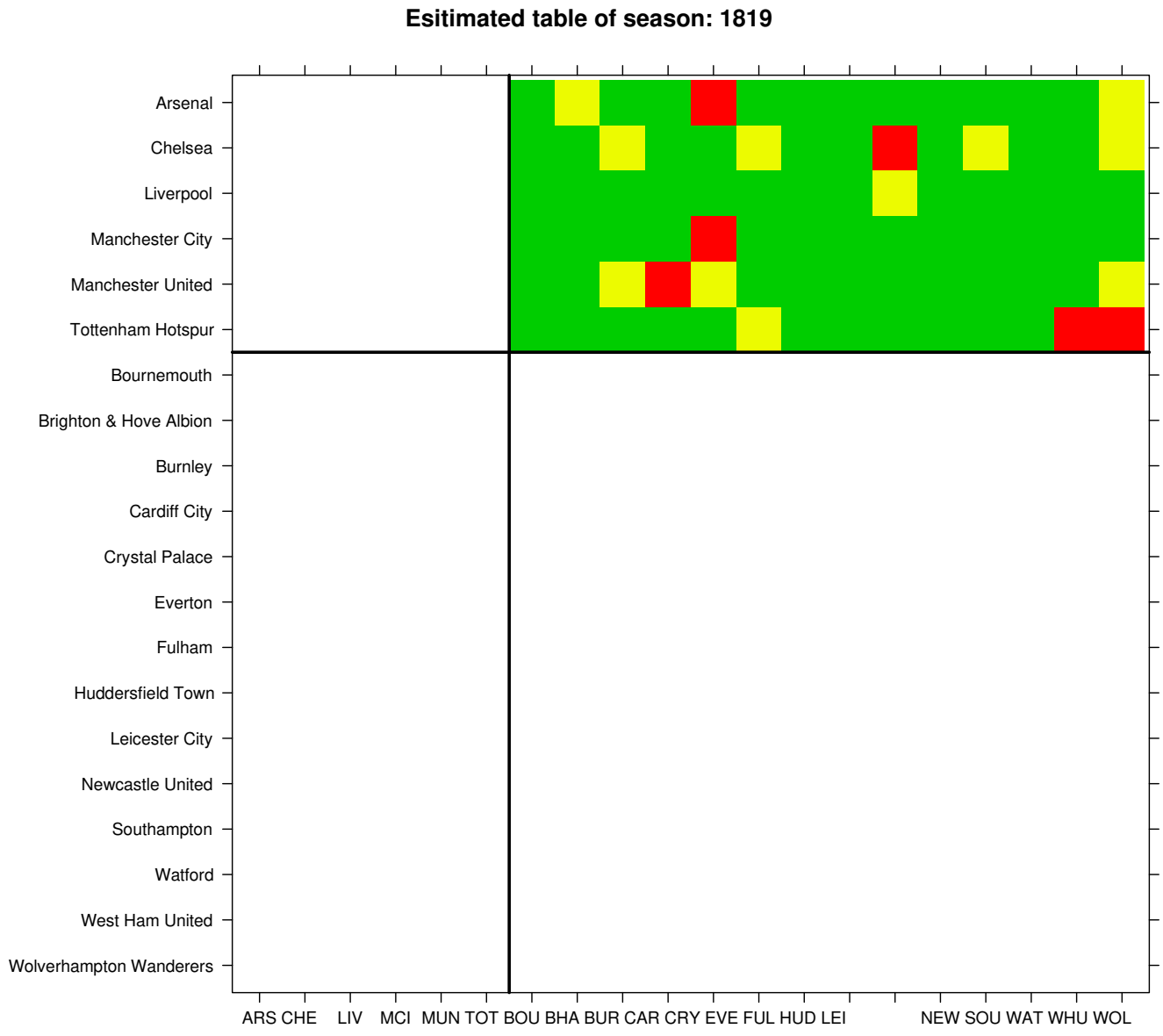
Analysis of the 2018/19 season



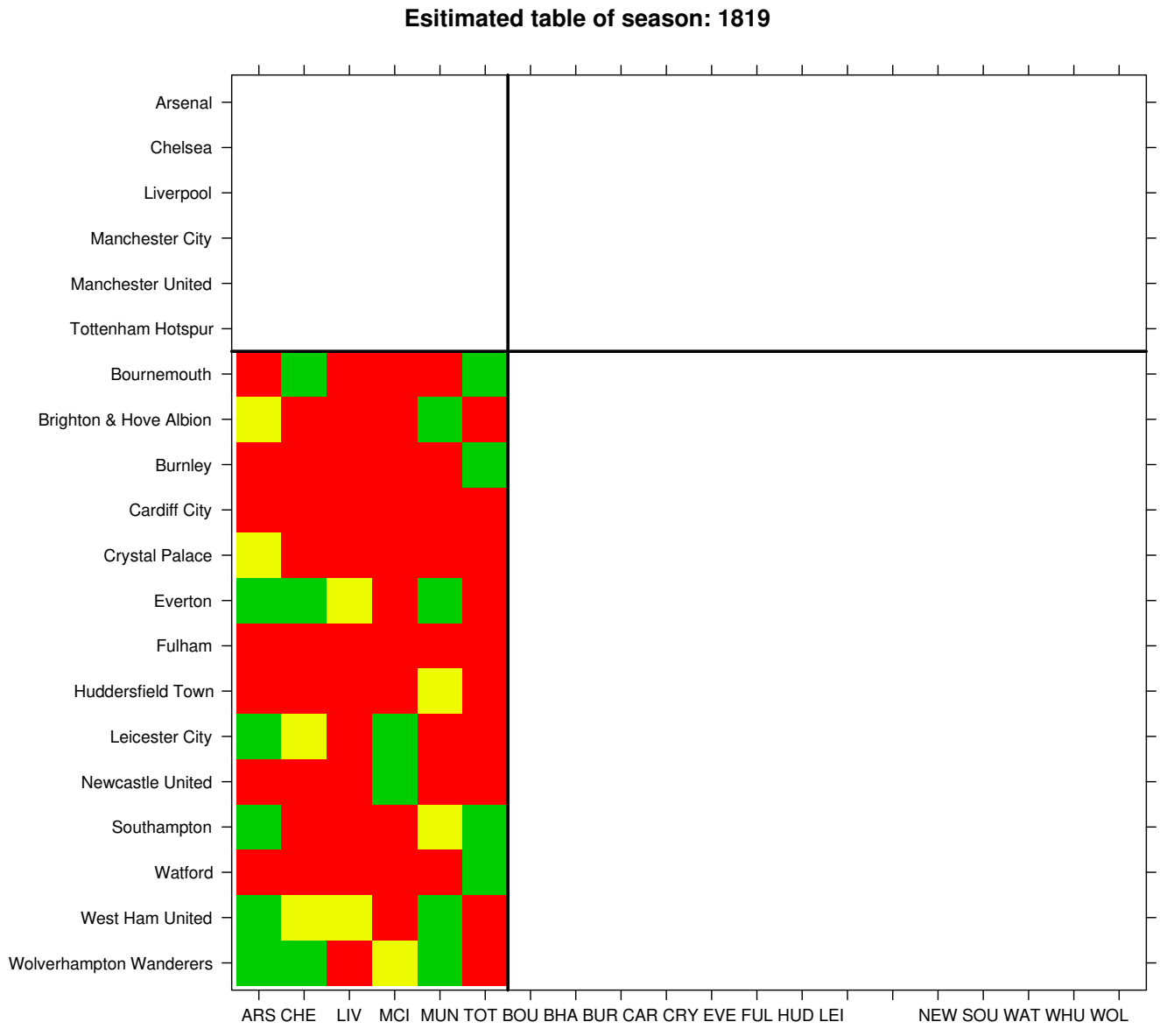
Analysis of the 2018/19 season



Analysis of the 2018/19 season



Analysis of the 2018/19 season



Did Tottenham hotspur overachieve in 2018/19?

	Tottenham	Arsenal	Manchester United
Tottenham	–	1 – 1	0 – 1
Arsenal	4 – 2	–	2 – 0
Manchester United	0 – 3	2 – 2	–

- This indicates that Tottenham's record against both teams (1 Win, 1 Draw, 2 Losses) was identical to Manchester United's but much worse that Arsenal's record against both teams (2 Wins, 2 Draws).

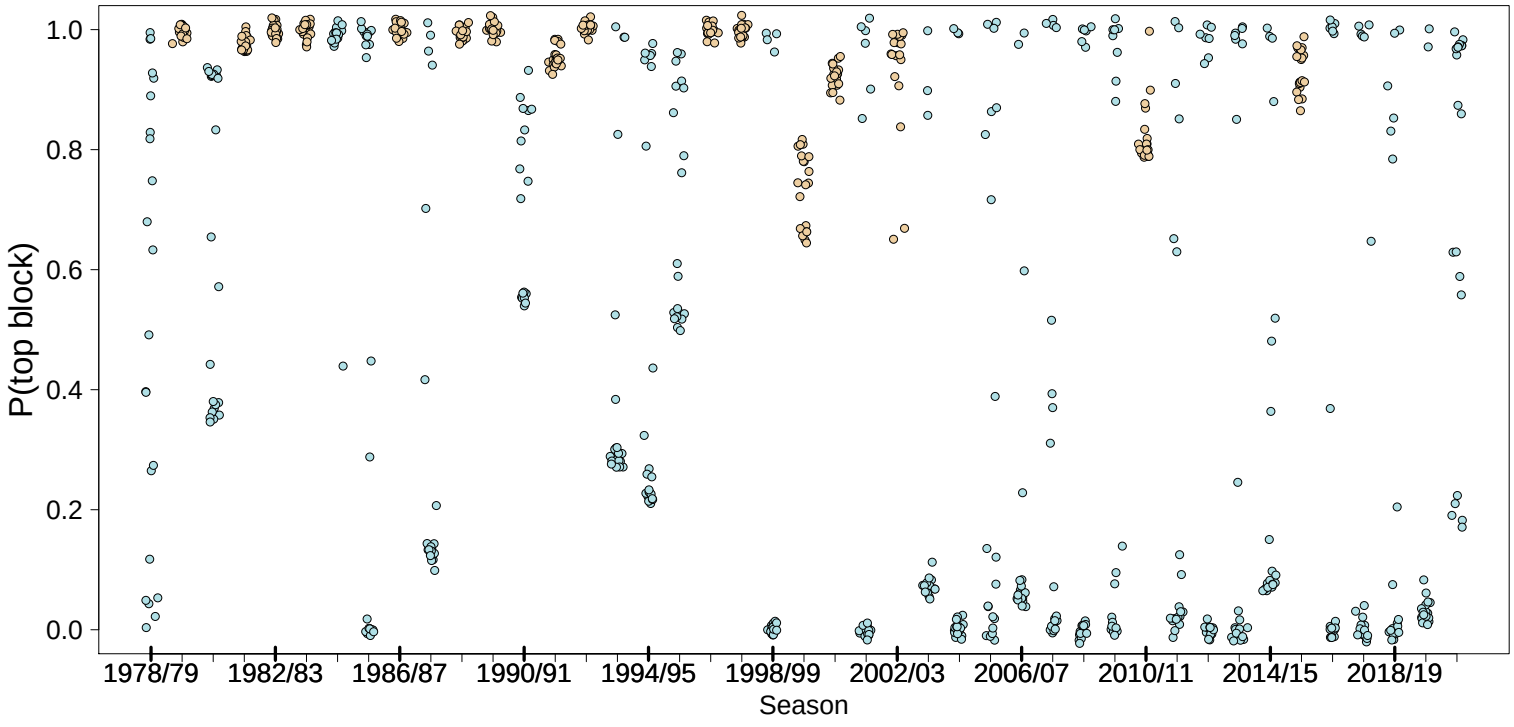
- This is also consistent with Arsenal being estimated a higher posterior membership to the top block than Tottenham despite having a lower overall league position.

Analysis of over 40 seasons of the Premier League

Season	Number of clusters 1	2	3	4
78/79	1.84	96.23	1.93	0.00
79/80	97.78	2.19	0.03	0.00
80/81	33.20	66.36	0.44	0.00
81/82	97.94	2.04	0.01	0.00
82/83	99.83	0.17	0.00	0.00
83/84	99.24	0.76	0.00	0.00
84/85	44.02	55.56	0.42	0.00
85/86	0.00	99.88	0.12	0.00
86/87	99.62	0.38	0.00	0.00
87/88	14.69	85.02	0.29	0.00
88/89	99.31	0.68	0.00	0.00
89/90	98.65	1.32	0.03	0.00
90/91	48.50	50.85	0.65	0.00
91/92	95.42	4.57	0.01	0.00
92/93	98.97	1.03	0.00	0.00
93/94	29.71	69.55	0.73	0.01
94/95	22.67	73.90	3.42	0.02
95/96	56.10	43.75	0.15	0.00
96/97	99.72	0.28	0.00	0.00
97/98	98.34	1.66	0.00	0.00
98/99	0.41	99.43	0.15	0.00
99/00	62.52	37.13	0.36	0.00

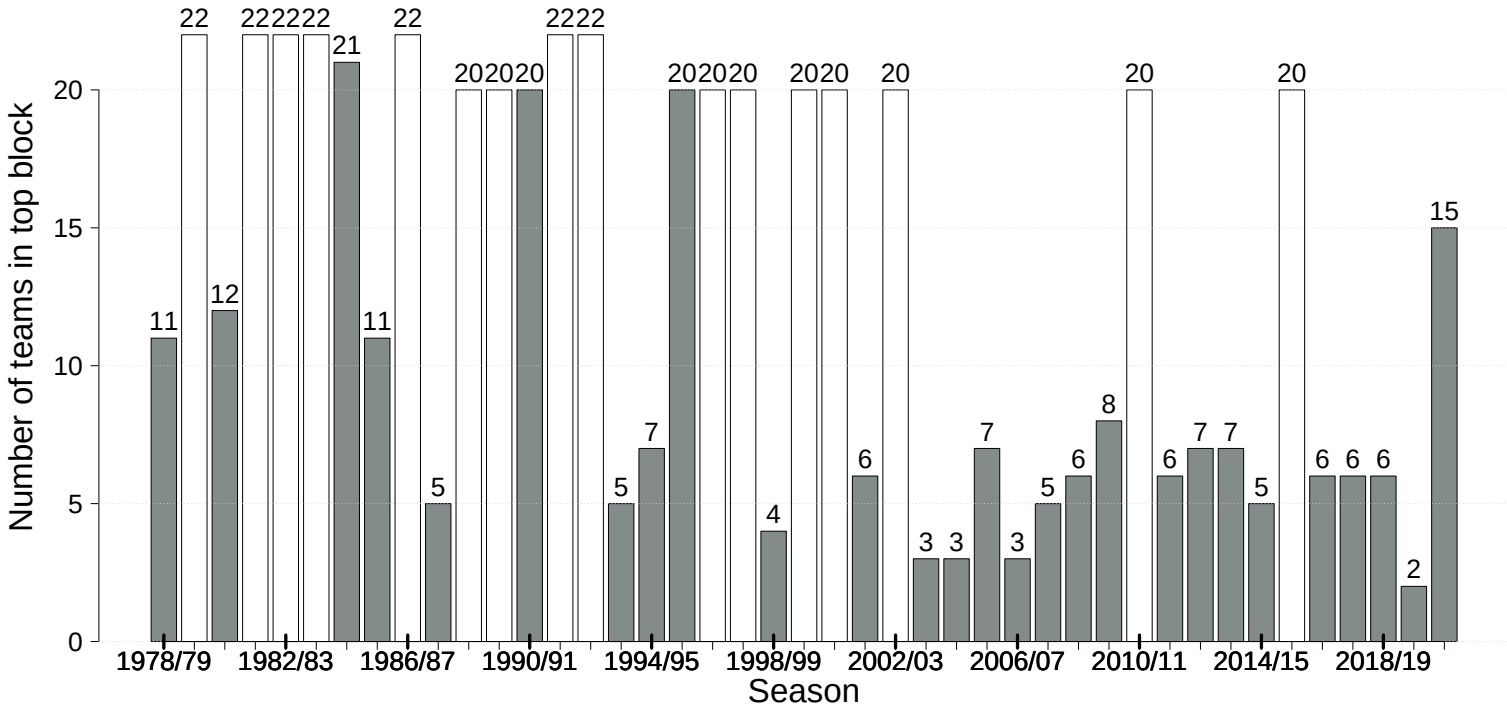
Season	Number of clusters 1	2	3	4
00/01	89.61	9.67	0.68	0.04
01/02	0.27	99.38	0.36	0.00
02/03	58.55	40.85	0.59	0.00
03/04	4.90	90.37	4.68	0.04
04/05	0.00	99.87	0.13	0.00
05/06	0.29	97.22	2.38	0.11
06/07	3.01	94.84	2.13	0.02
07/08	0.00	94.11	5.79	0.09
08/09	0.00	99.30	0.69	0.01
09/10	0.08	94.71	5.14	0.08
10/11	80.94	19.00	0.05	0.00
11/12	1.95	96.30	1.74	0.01
12/13	0.00	99.61	0.39	0.00
13/14	0.00	98.85	1.16	0.00
14/15	9.12	87.32	3.55	0.02
15/16	78.58	21.23	0.18	0.02
16/17	0.00	99.11	0.89	0.00
17/18	0.00	98.07	1.92	0.01
18/19	0.00	98.11	1.87	0.02
19/20	1.47	97.02	1.51	0.00
20/21	15.86	81.85	2.25	0.04

Probability of membership of each team to the top block



- The posterior allocation probability of belonging to the strongest group of teams over the 42 seasons under study.
- For each season the colour indicates whether the league was partitioned into a single cluster (sand colour) or two (light blue). Each point represents the estimated posterior probability for a team.

Posterior estimate of the size of the strongest block



Barplot displaying the posterior estimate of the number of teams allocated to the strongest block each season. Single and two block seasons are coloured white and grey, respectively. This illustrates that the size of the strongest block has generally decreased during the second half of the study period. Last season was an exception though!

Summary of the results

- The results indicate there was most support each season, a posteriori, for either a one block or a two block model, however the number of seasons where a two block model has most support, a posteriori, increases considerably over the past two decades.

- In particular, over the first half of this study period there is no strong support for either model. In fact for some seasons there is broadly equal posterior support for either model.

- Since around 2000 there is typically most support from a two block model, but further that the posterior probability for a two block model is over 0.8 for almost every season since 2003/04, providing strong evidence that the league has become more competitively imbalanced since then.

Evidence for the emergence of a *big-six* groups of teams

MAP allocation of teams to the strongest block ✓or not ✗ per season.

	Arsenal	Chelsea	Liverpool	Man City	Man Utd	Tottenham	Additional teams
20/21	✓	✓	✓	✓	✓	✓	9 further teams
19/20	✗	✗	✓	✓	✗	✗	
18/19	✓	✓	✓	✓	✓	✓	
17/18	✓	✓	✓	✓	✓	✓	
16/17	✓	✗	✓	✓	✓	✓	
15/16							
14/15	✓	✓	✗	✓	✓	✗	
13/14	✓	✓	✓	✓	✓	✓	Everton
12/13	✓	✓	✓	✓	✓	✓	Everton
11/12	✓	✓	✗	✓	✓	✓	Newcastle
10/11							
09/10	✓	✓	✓	✓	✓	✓	Everton, Aston Villa
08/09	✓	✓	✓	✗	✓	✗	Everton, Aston Villa
07/08	✓	✓	✓	✗	✓	✗	Everton
06/07	✓	✓	✗	✗	✓	✗	
05/06	✓	✓	✓	✗	✓	✓	Blackburn, Newcastle
04/05	✓	✓	✗	✗	✓	✗	
03/04	✓	✓	✗	✗	✓	✗	
02/03							
01/02	✓	✓	✓		✓	✗	Leeds, Newcastle

- The composition of the strongest block of teams has been stable since 2009/10 containing the six big teams almost each season.
- Prior to this, Man City were never allocated to the strong block, while Tottenham Hotspur were only allocated to it in 2005/06.
- Each of the other four teams (with the exception of Liverpool for some seasons) have been in the strongest block over this period.

Conclusions

- Our analysis of the English Premier League, our analysis has uncovered evidence that, broadly speaking, the league was quite balanced from around 1980 to 2000.
- However, subsequent to that, there is strong evidence that the league has become more imbalanced since from 2003/04 we see an emergence of league seasons where two blocks are most probable, a posteriori.
- In addition, our analysis suggests the emergence of a so-called *big-six* teams (Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, Tottenham Hospur) since around 2010 as during this time period all six teams, with only a few exceptions have always been present in the strongest block of teams.
- Season 2020/21 appears to be an outliers of sorts resulting in a much larger top block than in recent seasons. It was impacted by covid. Some games were played behind closed doors.

Possible extensions to this framework

- The SBM does not directly model for the number of goals scored by either team.
- In fact, there is a literature which have developed statistical models for football match data beginning with Dixon and Coles (1997), where a Poisson GLM framework is used to model the number of goals scored by either team.
- This has been extended by several authors, including Karlis and Ntzoufras (2003) to the bivariate Poisson setting.
- There are many more extensions of this Poisson GLM framework.

The final whistle!

Assessing competitive balance in the English Premier League for over forty seasons using a stochastic block model
Basini et al.
`https://arxiv.org/abs/2107.08732`

`https://github.com/basins95/Football_SBM`