Statistica Sinica Preprint No: SS-2023-0050	
Title	A Continuous-Time Stochastic Process for High-Resolution Network Data in Sports
Manuscript ID	SS-2023-0050
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0050
Complete List of Authors	Nicholas Grieshop, Yong Feng, Guanyu Hu and Michael Schweinberger
Corresponding Authors	Michael Schweinberger
E-mails	mus47@psu.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

A CONTINUOUS-TIME STOCHASTIC PROCESS FOR HIGH-RESOLUTION NETWORK DATA IN SPORTS

Nicholas Grieshop[1], Yong Feng[2], Guanyu Hu[3] and Michael Schweinberger[4]

1 Department of Statistics, University of Missouri-Columbia

2 Department of Economics, University of Missouri-Columbia

3 Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston

4 Department of Statistics, The Pennsylvania State University

Abstract: Technological advances have paved the way for collecting high-resolution network data in basketball, football, and other team-based sports. Such data consist of interactions among players of competing teams indexed by space and time. High-resolution network data are vital to understanding and predicting the performance of teams, because the performance of a team is more than the sum of the strengths of its individual players: Whether a collection of players forms a strong team depends on the strength of the individual players as well as the interactions among the players. We introduce a continuous-time stochastic process as a model of interactions among players of competing teams indexed by space and time, discuss basic properties of the continuous-time stochastic process, and learn

the stochastic process from high-resolution network data by pursuing a Bayesian approach. We present simulation results along with an application to Juventus Turin, Inter Milan, and other football clubs in the premier Italian soccer league.

Key words and phrases: Continuous-time stochastic processes; Relational event data; Soccer games; Spatio-temporal data; Sport analytics.

1. Introduction

Sport analytics has witnessed a surge of interest in the statistics community (see, e.g., Albert et al., 2017), driven by technological advances that have paved the way for collecting high-resolution tracking data in basketball, football, and other team-based sports.

Traditional sport analytics has focused on predicting match outcomes based on summary statistics (Dixon and Coles, 1997; Karlis and Ntzoufras, 2003; Baio and Blangiardo, 2010; Cattelan et al., 2013). In more recent times, the advent of high-resolution tracking data has expanded the role of statistics in sport analytics (Albert et al., 2017) and has enabled granular evaluations of players and teams (Cervone et al., 2014; Franks et al., 2015; Cervone et al., 2016; Wu and Bornn, 2018; Yurko et al., 2019; Hu et al., 2023) along with in-game strategy evaluations (Fernandez and Bornn, 2018; Sandholtz et al., 2020; Nguyen et al., 2023). High-resolution tracking data

fall into two categories: optical ball- and player-tracking data obtained from video footage collected by multiple cameras in sport arenas, and data collected by wearable devices. Some recent papers have used high-resolution tracking data to evaluate the defensive strength of teams (Franks et al., 2015); constructing a dictionary of play types (Miller and Bornn, 2017); assessing the expected value of ball possession in basketball (Cervone et al., 2016; Santos-Fernandez et al., 2022); and constructing deep generative models of spatio-temporal trajectory data (Santos-Fernandez et al., 2022).

As a case in point, we focus on soccer—that is, European football. Soccer is a fast-paced sport that generates high-resolution network data in the form of ball-tracking data indexed by space and time. The statistical analysis of high-resolution network data generated by soccer poses many challenges, including—but not limited to—the following:

1. Scoring a goal in a soccer match is a rare event, and useful predictors are hard to come by: e.g., a soccer team may score 0, 1, or 2 goals during a typical match, and scoring a goal requires a sequence of complex interactions among players of two competing teams.

2. Soccer teams consist of more players and the interactions among the players are more complex than, e.g., in basketball and other team-based sports. The fact that soccer teams are larger than teams in

many other team-based sports implies that the actions of players on the field need to be coordinated. To facilitate coordination, each soccer team adopts a formation, which assigns each player in the team to a specific position (e.g., goalkeeper, striker). Two popular formations of soccer teams, known as 4-4-2 and 3-5-2, are shown in Figure 1 in Supplement A. The chosen formation can affect the defensive and offensive strategies of a soccer team and can hence affect the outcome of a match. In addition, players may have different roles in different formations, and the formations of teams may change during matches.

3. Soccer matches are zero-sum games: One team's gain is another team's loss. For example, if the ball changes hands, one team loses control of the ball while the other team gains control of the ball.

We address the lack of a comprehensive statistical analysis of the network of interactions among soccer players by introducing a continuous-time stochastic process, which helps shed light on

- which player controls the ball and how long, and how ball control depends on the player's attributes (including the player's position in the team's formation and the player's spatial position on the field, provided that the spatial positions of players on the field are known);

- whether a change in ball control is a failure (i.e., the ball is lost to the opposing team) or a success (i.e., the ball remains within the team in control of the ball), and how the probability of a failure or a success depends on attributes of players;
- whether a team on track to winning a match decreases its pace and plays more defensively, while its opponent increases its pace and plays more offensively to change the outcome of the match in its favor;
- unobserved attributes of players that may affect ball control and interactions among players.

1.1 Comparison with non-network models of sport data

In contrast to the literature on basketball and other team-based sports, we do not focus on individual summaries, such as the expected ball possession of individual players (e.g., Cervone et al., 2016; Santos-Fernandez et al., 2022). Instead, we focus on the network of interactions among players, because the performance of a team is more than the sum of the strengths of its players. In other words, a collection of strong players may or may not form a strong soccer team: Whether a collection of players forms a strong soccer team depends on the one hand on the strength of the individual players and on the other hand on how the players interact.

Some recent publications (e.g., Chacoma et al., 2020; Hirotsu et al., 2023; Narizuka et al., 2023) have studied soccer matches by using probabilistic models, but the mentioned publications focus on time-dependent motion processes and ignore the network of interactions among players. By contrast, the proposed stochastic modeling framework focuses on the network of interactions among players and helps incorporate the formations of soccer teams in addition to the spatial distances between players, provided that the spatial positions of players are known.

Compared with the continuous-time within-play valuation models of American football in Yurko et al. (2020), the proposed stochastic modeling framework focuses on the pace of soccer matches, who is in control of the ball, whether a change in ball control is a failure or a success, and who secures control of the ball, rather than focusing on action evaluations. As a result, the proposed stochastic modeling framework can provide a more comprehensive understanding of team work in soccer and other team-based sports than the existing literature.

1.2 Comparison with discrete-time models of sport data

State-space models and other discrete-time stochastic processes have been used as predictive models for National Football League (NFL) game

scores and other team-based sports (e.g., Glickman and Stern, 2005; Shaw and Glickman, 2019). By contrast, we focus on continuous-time Markov processes, for at least two reasons.

First, continuous-time Markov processes are natural models of real-world processes where events can occur at any time $t \in [0, +\infty)$, including fast-paced soccer matches.

Second, continuous-time Markov processes can be viewed as discrete-time Markov chains with the time gaps between transitions of the Markov chains filled with Exponential holding times (see, e.g., Chapter 3 of Norris, 1997). In other words, continuous-time Markov processes model when changes take place, and which changes take place. Therefore, continuous-time Markov processes help build richer models than discrete-time Markov processes. For example, in applications to soccer matches, continuous-time Markov processes help shed light on:

- *Clock:* When a change in ball control occurs, and how a change depends on the attributes of the player in control of the ball.

- *Transitions:* Who passes the ball to whom, and how a change in ball control depends on the attributes of the players involved.

1.3 Comparison with relational event models

On mathematical grounds, the closest relatives of the proposed stochastic modeling framework are relational event models (e.g., Butts, 2008; Perry and Wolfe, 2013; Stadtfeld, 2011). Having said that, there are important differences between relational event models and the proposed stochastic modeling framework:

1. *Soccer matches revolve around the ball.* A reasonable stochastic model of soccer matches needs to reflect the fact that soccer matches revolve around the ball: e.g., at any given time t, a single player is in control of the ball and can initiate a relational event (e.g., a pass), and a stochastic model of soccer matches should reflect that. By contrast, relational event models assume that any actor can initiate a relational event at any time: e.g., at any given time t, any employee of a company can send an email to one or more other employees.

2. *Soccer matches are zero-sum games:* One team's gain is another team's loss. As a result, a reasonable stochastic model of soccer matches should distinguish between successful and unsuccessful relational events (e.g., passes), which can affect the outcomes of a match. By contrast, relational event models are not concerned with zero-sum

games and do not distinguish between successful and unsuccessful relational events: e.g., email communications between the employees of a company are not zero-sum games, and the event that employee A sends an email to employee B does not necessarily result in a gain or a loss for employee A or employee B.

3. *The formations of soccer teams and the locations of players on the field can affect the outcome of a match.* By contrast, if an employee of a company considers sending an email, the location of the employee is unimportant: As long as the employee is connected to the World Wide Web, the employee can send an email from any location on planet Earth.

1.4 Structure of paper

We first introduce the data that motivated the proposed stochastic modeling framework (Section 2) and then introduce the stochastic modeling framework (Section 3). A Bayesian approach to learning the stochastic modeling framework from data is described in Section 4, and Bayesian computing is discussed in Section 5. An application to the motivating data is presented in Section 6. Simulation results can be found in Section 7.

2. High-resolution network data

We consider data provided by Hudl & Wyscout (`https://footballdata.wyscout.com/`). The data consists of 380 matches during the 2020/21 season of Serie A, the premier league of the Italian football league system. The data include ball-tracking data, but not player-tracking data. In other words, we know which player is in control of the ball, but we do not know where the players are located on the field.

Figure 2 in Supplement A shows a subset of the data: passes between the players of Juventus Turin (with 4-4-2 formation) and Inter Milan (with 3-5-2 formation). These data are based on the home games of Juventus Turin versus AC Milan and Inter Milan versus AC Milan in 2020/21. The figure reveals that passes depend on the formations of teams. Figure 2(a) in Supplement A shows that the midfield players and defenders of Juventus Turin (with 4-4-2 formation) dominate ball control. By contrast, strikers do not control the ball all too often, but are key to scoring goals and hence winning matches. Figure 2(b) in Supplement A reveals that the midfield players of Inter Milan (with 3-5-2 formation) likewise dominate ball control. In addition, the right wing of Inter Milan plays an important role in Inter Milan's 3-5-2 formation, by passing the ball to the strikers and in so doing helping the team launch counterattacks straight out of the backfield. Other

descriptive summaries, including detailed information on the formations and players of Juventus Turin, Inter Milan, and other soccer clubs in Serie A are presented in Supplement C.

3. Stochastic modeling framework

We introduce a continuous-time stochastic process as a model of soccer matches starting at time $t_0 := 0$ and stopping at time $T \in [90, +\infty)$.

Soccer matches involve two competing teams. Each team consists of 11 players and can substitute up to 5 players during a match, effective 2022/23. Let $\mathcal{T}_{1,t}$ be the set of players of one of the two teams and $\mathcal{T}_{t,2}$ be the set of players of the opposing team at time $t \in [0, T)$. The two sets $\mathcal{T}_{1,t}$ and $\mathcal{T}_{2,t}$ are disjoint, in the sense that $\mathcal{T}_{1,t} \cap \mathcal{T}_{2,t} = \{\}$ for all $t \in [0, T)$. The compositions of the two teams $\mathcal{T}_{1,t}$ and $\mathcal{T}_{2,t}$ can change during a match, because players may be injured; players may be substituted; and the referee may remove players from the field due to violations of rules. We consider changes in the compositions of $\mathcal{T}_{1,t}$ and $\mathcal{T}_{2,t}$ to be exogenous.

3.1 Generic continuous-time stochastic process

We introduce a generic continuous-time stochastic process that captures salient features of soccer matches.

Scoring goals: rare events. We focus on who is control of the ball, whether a change in ball control is a failure or a success, and who secures control of the ball, but we do not model the process of scoring goals. While scoring goals is important for winning matches, the event of scoring a goal is a rare event and useful predictors are hard to come by, because scoring a goal requires a sequence of complex interactions among players of two competing teams. We leave the construction of models for scoring goals to future research and focus here on ball control and interactions among players, which are important for scoring goals and winning matches.

Ball control and interactions among players. We first describe a generic continuous-time stochastic process. We then introduce a specification of the continuous-time stochastic process in Section 3.2 and discuss basic properties of the continuous-time stochastic process in Supplement D.

A generic continuous-time stochastic process of a soccer match starting at time $t_0 \coloneqq 0$ and stopping at time $T \in [90, +\infty)$ takes the following form:

1. At time $t_0 \coloneqq 0$, the referee starts the match. The player who secures control of the ball at time t_0 is chosen at random from the set $\mathcal{T}_{1,t_0} \cup \mathcal{T}_{2,t_0}$ and is denoted by i_1.

2. At time $t_m \coloneqq t_{m-1} + h_m$ $(m = 1, 2, \dots)$, the ball passes from player

$i_m \in \mathcal{T}_{1,t_m} \cup \mathcal{T}_{2,t_m}$ to player $j_m \in \mathcal{T}_{1,t_m} \cup \mathcal{T}_{2,t_m} \setminus \{i_m\}$, where $h_m \sim$ Exponential(λ_{i_m}) and $i_m = j_{m-1}$ $(m = 2, 3, \dots)$. The process of passing the ball from player i_m to player j_m is decomposed as follows:

2.1 The change in ball control is either a failure (indicated by $S_{i_m} = 0$) in that player i_m loses the ball to a player of the opposing team, or is a success (indicated by $S_{i_m} = 1$) in that i_m succeeds in passing the ball to a player of i_m's own team.

2.2 Conditional on $S_{i_m} \in \{0, 1\}$, player i_m cedes control of the ball to player $j_m \in \mathcal{T}_{1,t_m} \cup \mathcal{T}_{2,t_m} \setminus \{i_m\}$, indicated by $i_m \to j_m$.

3. The referee stops the match at time $T \in [90, +\infty)$.

We consider the decision of the referee to stop the match to be exogenous, so that the stopping time $T \in [90, +\infty)$ of the match is non-random. In practice, soccer matches last 90 minutes, but disruptions of matches due to injuries and substitutions of players may result in overtime.

3.2 Specification of continuous-time stochastic process

We introduce a specification of the generic continuous-time stochastic process introduced in Section 3.1, by specifying the distributions of the holding times h_m, the success probabilities $\mathbb{P}(S_{i_m} = s_{i_m})$, and the pass probabilities

$\mathbb{P}(i_m \to j_m \mid S_{i_m} = s_{i_m})$. Basic properties of the resulting continuous-time stochastic process are discussed in Supplement D. Throughout, we denote by $\mathcal{I}_m$ the team of player i_m in control of the ball at time t_m.

Holding time distributions A natural specification of the holding time distributions is

$$h_m \mid \lambda_{i_m} \overset{\text{ind}}{\sim} \text{Exponential}(\lambda_{i_m}).$$

To allow the rate $\lambda_{i_m} \in (0, +\infty)$ of player i_m's holding time h_m to depend on observed attributes of i_m (e.g., the position of i_m in the formation of i_m's team and the location of i_m on the field), we assume that

$$\lambda_{i_m}(\boldsymbol{\omega}) \;\coloneqq\; \exp(\boldsymbol{\omega}^\top \boldsymbol{c}_{i_m}),$$

where $\boldsymbol{\omega} \in \mathbb{R}^p$ is a vector of p parameters and $\boldsymbol{c}_{i_m} \in \mathbb{R}^p$ is a vector of p observed attributes of player i_m.

Success probabilities The probability of a successful pass $\{S_{i_m} = 1\}$ by player i_m can be specified by a logit model:

$$\text{logit}(\mathbb{P}_{\boldsymbol{\alpha},\boldsymbol{\eta}}(S_{i_m} = 1)) \;\coloneqq\; \boldsymbol{\alpha}^\top \boldsymbol{x}_{1,i_m} + \eta_{1,i_m},$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{d_1}$ is a vector of d_1 parameters and $\boldsymbol{x}_{1,i_m} \in \mathbb{R}^{d_1}$ is a vector of d_1 observed attributes of player i_m. The random effect $\eta_{1,i_m} \in \mathbb{R}$ captures the effect of unobserved attributes of player i_m on the success probability.

Pass probabilities The conditional probability of event $\{i_m \to j_m\}$ given $\{S_{i_m} = 0\}$ can be specified by a multinomial logit model:

$$\mathbb{P}_{\boldsymbol{\beta},\boldsymbol{\eta}}(i_m \to j_m \mid S_{i_m} = 0)$$
$$:= \begin{cases} \dfrac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_{2,i_m,j_m} + \eta_{2,j_m})}{\sum_{j \notin \mathcal{I}_m} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_{2,i_m,j} + \eta_{2,j})} & \text{if } j_m \notin \mathcal{I}_m \\ 0 & \text{if } j_m \in \mathcal{I}_m, \end{cases}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{d_2}$ is a vector of d_2 parameters and $\boldsymbol{x}_{2,i_m,j} \in \mathbb{R}^{d_2}$ is a vector of d_2 observed attributes of players i_m and j. The random effect $\eta_{2,j} \in \mathbb{R}$ captures the effect of unobserved attributes of player j on the conditional probability of securing control of the ball.

Along the same lines, the conditional probability of event $\{i_m \to j_m\}$ given $\{S_{i_m} = 1\}$ can be specified by a multinomial logit model:

$$\mathbb{P}_{\boldsymbol{\gamma},\boldsymbol{\eta}}(i_m \to j_m \mid S_{i_m} = 1)$$
$$:= \begin{cases} 0 & \text{if } i_m = j_m \text{ or } j_m \notin \mathcal{I}_m \\ \dfrac{\exp(\boldsymbol{\gamma}^\top \boldsymbol{x}_{3,i_m,j_m} + \eta_{3,j_m})}{\sum_{j \in \mathcal{I}_m \setminus \{i_m\}} \exp(\boldsymbol{\gamma}^\top \boldsymbol{x}_{3,i_m,j} + \eta_{3,j})} & \text{if } i_m \neq j_m \text{ and } j_m \in \mathcal{I}_m, \end{cases}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{d_3}$ is a vector of d_3 parameters and $\boldsymbol{x}_{3,i_m,j} \in \mathbb{R}^{d_3}$ is a vector of d_3 observed attributes of players i_m and j, e.g., whether players i_m and j are friends, whether player i_m passed the ball to player j in the past, whether player i_m received the ball from player j in the past, or the spatial distance between players i_m and j at the time of the pass (provided that the spatial positions of players are known). The random effect $\eta_{3,j} \in \mathbb{R}$ captures the effect of unobserved attributes of player j on the conditional probability of securing control of the ball.

Random effects Let $\boldsymbol{\eta}_i \coloneqq (\eta_{1,i},\, \eta_{2,i},\, \eta_{3,i}) \in \mathbb{R}^3$ and assume that

$$\boldsymbol{\eta}_i \mid \boldsymbol{\Sigma} \overset{\text{iid}}{\sim} \text{MVN}_3(\mathbf{0}_3,\, \boldsymbol{\Sigma}),$$

where $\mathbf{0}_3 \in \mathbb{R}^3$ is the three-dimensional null vector and $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times 3}$ is a positive-definite variance-covariance matrix.

Alternative models It is worth noting that there are other possible approaches to constructing stochastic models of soccer matches. For example, each the two following approaches to constructing models can help shed light on salient aspects of soccer matches:

(a) Assuming player i_m is in control of the ball at time t_m, first deter-

mine whether i_m succeeds in passing the ball to a teamplayer. Then determine which teamplayer j_m receives the ball provided that the pass is a success, otherwise determine which player j_m of the opposing team secures control of the ball provided that the pass is a failure (the approach pursued here).

(b) Assuming player i_m is in control of the ball at time t_m, suppose that i_m first selects a teamplayer k_m and intends to pass the ball to k_m. Then determine whether the intended pass $i_m \to k_m$ succeeds. If the intended pass $i_m \to k_m$ succeeds, set $k_m = j_m$, otherwise select the player j_m who secures control of the ball from the opposing team (an approach suggested by an anonymous referee).

While both approaches can be useful, there are two good reasons for choosing approach (a), that is, the approach pursued here.

First, soccer matches revolve around the ball, so soccer teams wish to retain control of the ball. Thus, the player in control of the ball is first and foremost responsible for passing the ball to a teamplayer—unless the player has the rare opportunity to score a goal. By construction, approach (a) respects the importance of retaining control of the ball.

Second, approach (a) has one advantage over approach (b): If a pass is a failure, we do not observe the intended receiver k_m. Worse, even when

a pass is a success, we may not observe the intended receiver k_m: e.g., i_m may intend to pass the ball to teamplayer k_m, but the ball ends up in possession of some other teamplayer $j_m \neq k_m$ by accident. In fact, instead of observing the intended receiver k_m, we observe the actual receiver j_m, who may or may not be identical to the intended receiver k_m. In other words, the data fall short, in that we do not observe the *intended passes* $i_m \to k_m$, but we observe the *actual passes* $i_m \to j_m$, regardless of whether the passes are failures or successes. As a result, approach (b) would require augmenting the observed passes $i_m \to j_m$ by the unobserved, intended passes $i_m \to k_m$. While it is possible to augment the observed passes $i_m \to j_m$ by the unobserved, intended passes $i_m \to k_m$ using data-augmentation methods, such methods come at additional computational costs compared with approach (a). In addition, there may be statistical costs: It is not clear how much information the data contain about the unobserved, intended passes $i_m \to k_m$.

4. Bayesian learning

We pursue a Bayesian approach to learning the stochastic modeling framework introduced in Section 3 from high-resolution network data.

A Bayesian approach is well-suited to online learning, that is, updat-

ing the knowledge about the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\omega}$, $\boldsymbol{\Sigma}$ and the random effects $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots$ as soon as additional data points roll in. To demonstrate, consider two teams and let $\boldsymbol{x}_1 \coloneqq (h_{1,m}, i_{1,m}, j_{1,m})_{m=1}^{M_1}$ be the outcome of the first match of the two teams (with $M_1 \geq 1$ passes) and $\boldsymbol{x}_2 \coloneqq (h_{2,m}, i_{2,m}, j_{2,m})_{m=1}^{M_2}$ be the outcome of the second match of the two teams (with $M_2 \geq 1$ passes). To ease the presentation, assume that the compositions of the two teams do not change during the first and second match, the 22 players of the two teams are labeled $1, \ldots, 22$, and the random effects are denoted by $\boldsymbol{\eta} \coloneqq (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{22})$. In addition, assume that the outcomes of the first and second match $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfy

$$\begin{aligned} \pi(\boldsymbol{x}_1, \boldsymbol{x}_2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}) &= \pi(\boldsymbol{x}_1 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}) \\ &\quad \times \pi(\boldsymbol{x}_2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}, \boldsymbol{x}_1), \end{aligned}$$

where π denotes a generic probability density function. The conditional

probability density function $\pi(\boldsymbol{x}_1 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta})$ is of the form

$$\begin{aligned}\pi(\boldsymbol{x}_1 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}) \;=\; \prod_{m=1}^{M_1} \Big[& \lambda_{i_{1,m}}(\boldsymbol{\omega}) \, \exp(-\lambda_{i_{1,m}}(\boldsymbol{\omega}) \, h_{1,m}) \\ & \times \; \mathbb{P}_{\boldsymbol{\alpha},\boldsymbol{\eta}}(S_{i_{1,m}} = s_{i_{1,m}}) \\ & \times \; \mathbb{P}_{\boldsymbol{\beta},\boldsymbol{\eta}}(i_{1,m} \to j_{1,m} \mid S_{i_{1,m}} = 0)^{\mathbb{1}(S_{i_{1,m}} = 0)} \\ & \times \; \mathbb{P}_{\boldsymbol{\gamma},\boldsymbol{\eta}}(i_{1,m} \to j_{1,m} \mid S_{i_{1,m}} = 1)^{\mathbb{1}(S_{i_{1,m}} = 1)} \Big] \\ & \times \; \exp\left(-\lambda_{i_{1,M_1+1}}(\boldsymbol{\omega}) \, \left(T_1 - \sum_{k=1}^{M_1} h_{1,k} \right) \right),\end{aligned}$$

assuming that the start time $t_0 \coloneqq 0$ and the stopping time $T_1 \in [90, +\infty)$ of the match are determined by the referee and are both non-random. The function $\mathbb{1}(.)$ is an indicator function, which is 1 if its argument is true and is 0 otherwise. The conditional probability density function $\pi(\boldsymbol{x}_2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}, \boldsymbol{x}_1)$ is of the same form as $\pi(\boldsymbol{x}_1 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta})$, but is based on M_2 passes rather than M_1 passes and can depend on the outcome of the first match $\boldsymbol{x}_1$.

The posterior of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta}$ based on the outcome of the first

match $\boldsymbol{x}_1$ is proportional to

$$\begin{aligned}\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta} \mid \boldsymbol{x}_1) \;&\propto\; \pi(\boldsymbol{x}_1 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}) \\ &\times\; \pi(\boldsymbol{\eta} \mid \boldsymbol{\Sigma})\, \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}),\end{aligned}$$

where $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma})$ is the prior of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}$. The prior of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}$ is described in Section 5.

As soon as the outcome of the second match $\boldsymbol{x}_2$ is observed, the knowledge about $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta}$ in light of $\boldsymbol{x}_2$ can be updated as follows:

$$\begin{aligned}&\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta} \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \\ \propto\;& \pi(\boldsymbol{x}_1, \boldsymbol{x}_2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta})\, \pi(\boldsymbol{\eta} \mid \boldsymbol{\Sigma})\, \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) \\ \propto\;& \pi(\boldsymbol{x}_2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}, \boldsymbol{x}_1)\, \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta} \mid \boldsymbol{x}_1).\end{aligned}$$

In other words, as soon as the outcome of the second match $\boldsymbol{x}_2$ is observed, we can update the knowledge about $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta}$ in light of $\boldsymbol{x}_2$ via $\pi(\boldsymbol{x}_2 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}, \boldsymbol{x}_1)$, with the knowledge about $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta}$ prior to the second match $\boldsymbol{x}_2$ being quantified by $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta} \mid \boldsymbol{x}_1)$, the posterior based on the outcome of the first match $\boldsymbol{x}_1$. As a result, a Bayesian approach is a natural approach to updating knowledge about the stochastic modeling framework as additional data points roll in. More than two teams with can be handled, and multiple matches in parallel.

5. Bayesian computing

While a Bayesian approach to learning the stochastic modeling framework introduced in Section 3 from high-resolution network data is natural, the posterior $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \boldsymbol{\eta} \mid \boldsymbol{x})$ of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\omega}$, $\boldsymbol{\Sigma}$ and the random effects $\boldsymbol{\eta}$ based on the outcome of a match $\boldsymbol{x}$ is not available in closed form. We approximate the posterior by using Markov chain Monte Carlo methods, by sampling from the full conditional distributions of the parameters and the random effects:

$$
\begin{aligned}
\pi(\boldsymbol{\alpha} \mid \boldsymbol{x}) &\propto \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\eta};\, \boldsymbol{x})\, \pi(\boldsymbol{\alpha}) \\
\pi(\boldsymbol{\beta} \mid \boldsymbol{x}) &\propto \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\eta};\, \boldsymbol{x})\, \pi(\boldsymbol{\beta}) \\
\pi(\boldsymbol{\gamma} \mid \boldsymbol{x}) &\propto \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\eta};\, \boldsymbol{x})\, \pi(\boldsymbol{\gamma}) \\
\pi(\boldsymbol{\omega} \mid \boldsymbol{x}) &\propto \mathcal{L}(\boldsymbol{\omega};\, \boldsymbol{x})\, \pi(\boldsymbol{\omega}) \\
\pi(\boldsymbol{\eta} \mid \boldsymbol{x}) &\propto \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\eta};\, \boldsymbol{x})\, \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\eta};\, \boldsymbol{x})\, \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\eta};\, \boldsymbol{x})\, \mathcal{L}(\boldsymbol{\Sigma};\, \boldsymbol{\eta}) \\
\pi(\boldsymbol{\Sigma} \mid \boldsymbol{\eta}) &\propto \mathcal{L}(\boldsymbol{\Sigma};\, \boldsymbol{\eta})\, \pi(\boldsymbol{\Sigma}),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\eta};\, \boldsymbol{x}) &\propto \prod_{m=1}^{M} \mathbb{P}_{\boldsymbol{\alpha},\boldsymbol{\eta}}(S_{i_m} = s_{i_m}) \\
\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\eta};\, \boldsymbol{x}) &\propto \prod_{m=1}^{M} \mathbb{P}_{\boldsymbol{\beta},\boldsymbol{\eta}}(i_m \to j_m \mid S_{i_m} = 0)^{\mathbb{1}(S_{i_m}=0)} \\
\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\eta};\, \boldsymbol{x}) &\propto \prod_{m=1}^{M} \mathbb{P}_{\boldsymbol{\gamma},\boldsymbol{\eta}}(i_m \to j_m \mid S_{i_m} = 1)^{\mathbb{1}(S_{i_m}=1)} \\
\mathcal{L}(\boldsymbol{\omega};\, \boldsymbol{x}) &\propto \prod_{m=1}^{M} [\lambda_{i_m}(\boldsymbol{\omega})\, \exp(-\lambda_{i_m}(\boldsymbol{\omega})\, h_m)] \exp\left(-\lambda_{i_{M+1}}(\boldsymbol{\omega})\, \left(T - \sum_{k=1}^{M} h_k\right)\right) \\
\mathcal{L}(\boldsymbol{\Sigma};\, \boldsymbol{\eta}) &\propto \prod_{i=1}^{n} \det(\boldsymbol{\Sigma}^{-1})^{1/2} \exp\left(-\frac{1}{2}\, \boldsymbol{\eta}_i^{\top}\, \boldsymbol{\Sigma}^{-1}\, \boldsymbol{\eta}_i\right),
\end{aligned}
$$

assuming that $\boldsymbol{x}$ is the outcome of a single soccer match with $M \geq 1$ passes starting at time $t_0 = 0$ and stopping at time $T \in [90, +\infty)$; note that both the start time t_0 and the stopping time T are non-random. We assume that the prior factorizes according to

$$
\pi(\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \boldsymbol{\gamma},\, \boldsymbol{\omega},\, \boldsymbol{\Sigma}) \;=\; \pi(\boldsymbol{\alpha})\, \pi(\boldsymbol{\beta})\, \pi(\boldsymbol{\gamma})\, \pi(\boldsymbol{\omega})\, \pi(\boldsymbol{\Sigma}),
$$

with marginal priors of the form

$$
\begin{aligned}
\alpha_k &\overset{\text{iid}}{\sim} N(0, 10^2), \quad k = 1, \ldots, d_1, \qquad \beta_k \overset{\text{iid}}{\sim} N(0, 10^2), \quad k = 1, \ldots, d_2 \\
\gamma_k &\overset{\text{iid}}{\sim} N(0, 10^2), \quad k = 1, \ldots, d_3, \qquad \omega_k \overset{\text{iid}}{\sim} N(0, 10^2), \quad k = 1, \ldots, p,
\end{aligned}
$$

where $N(0, 10^2)$ is a Gaussian with mean 0 and variance $10^2 = 100$. To specify the prior of the variance-covariance matrix $\mathbf{\Sigma}$ of the random effects, we decompose $\mathbf{\Sigma}$ according to

$$\mathbf{\Sigma} \coloneqq \begin{pmatrix} \sigma_{\eta_1} & 0 & 0 \\ 0 & \sigma_{\eta_2} & 0 \\ 0 & 0 & \sigma_{\eta_3} \end{pmatrix} \mathbf{\Lambda} \begin{pmatrix} \sigma_{\eta_1} & 0 & 0 \\ 0 & \sigma_{\eta_2} & 0 \\ 0 & 0 & \sigma_{\eta_3} \end{pmatrix},$$

where $\mathbf{\Lambda} \in [-1, +1]^{3\times 3}$ is a correlation matrix. We then assume that $\mathbf{\Lambda} \sim \text{LKJcorr}(2)$ has a Lewandowski-Kurowicka-Joe (LKJ) distribution with parameter 2 and $\sigma_{\eta_k} \overset{\text{iid}}{\sim} \text{Exponential}(1)$ $(k = 1, 2, 3)$.

To sample from the full conditionals, we use Markov chain Monte Carlo methods implemented in `R` package `rstan` (Stan Development Team, 2023). Since the stochastic modeling framework leverages exponential-family distributions as building blocks (e.g., Bernoulli, Multinomial, Exponential, and multivariate Gaussians), we do not have more numerical issues than other exponential-family models, such as generalized linear models, Gaussian and non-Gaussian graphical models (Efron, 2022).

6. Application

We use the stochastic modeling framework introduced in Section 3 to analyze the data described in Section 2. We focus on the matches of four soccer teams during the 2020/21 season of Serie A, the premier league of the Italian football league system:

- Juventus Turin (Juventus F.C.; 15,832 observations);
- Inter Milan (Internazionale Milano; 13,564 observations);
- Crotone (Crotone S.r.l.; 8,125 observations);
- Fiorentina (ACF Fiorentina; 8,107 observations).

Juventus Turin and Inter Milan belong to the most storied Italian soccer clubs, while Crotone and Fiorentina were mid- and low-level teams during the 2020/21 season, respectively. The numbers of observations mentioned above refer to the total numbers of passes during the 2020/21 season, aggregated over all matches played by the selected teams with the dominant formation. The selected teams have in common that all of them were proficient users of the 4-4-2 formation (Juventus Turin) or the 3-5-2 formation (Inter Milan, Crotone, Fiorentina).

We use the following specification of the stochastic modeling framework:

- **Module 1 (M1):** The Exponential model of the holding times h_m uses the following covariates: position-specific indicators of who is in control of the ball and indicators of whether the player's team is on track to winning or losing the match (i.e., the player's team has scored at least one more goal or one less goal than its opponent, respectively).

- **Module 2 (M2):** The logit model of the probability of a successful pass $\{S_{i_m} = 1\}$ uses the following covariates, in addition to an intercept: the length of the pass in terms of two-dimensional Euclidean distance; an indicator of whether player i_m initiates the pass in the opposing team's half of the field; an indicator of whether the ball ends up in the opposing team's third of the field; an indicator of whether the pass is a forward pass; an indicator of whether the pass is an air pass; indicators of whether the player's team is on track to winning or losing the match (i.e., whether the player's team has scored at least one more goal or one less goal than its opponent, respectively); and a position-specific random effect.

- **Module 3 (M3):** The multinomial logit model of the conditional probability of event $\{i_m \to j_m\}$ given $\{S_{i_m} = 1\}$ uses the following predictors: the graph distance between players i_m and j_m—defined

as the length of the shortest path between i_m and j_m—based on the nearest-neighbor graph in Figure 3 in Supplement A; the number of times j_m received the ball prior to the m-th pass; and a position-specific random effect.

It would be interesting to include more features into the multinomial logit model of the conditional probability of event $\{i_m \to j_m\}$ given $\{S_{i_m} = 1\}$, e.g., the spatial positions of players and additional network features. That said, we do not have data on the spatial distances between players and additional network features. Note that these limitations are limitations of the data, not the model: The model can incorporate spatial distances between players as well as additional network features. In addition, note that we focus here on all matches involving the four mentioned teams with the dominant formation, but we do not use the data of the opposing teams. As a consequence, we do not specify the conditional probabilities of events $\{i_m \to j_m\}$ given $\{S_{i_m} = 0\}$. Last, but not least, note that we use position-specific rather than player-specific random effects, because the data do not include complete information about which position is filled by which player.

Posterior sensitivity checks and posterior predictive checks can be found in Sections 6.1 and 6.2, respectively: The posterior sensitivity checks suggest that the posterior is not too sensitive to the choice of prior, while the

posterior predictive checks indicate that model-based predictions match the observed data. Tables 9 and 10 in Supplement E present posterior summaries of the model parameters, based on the 2020/21 matches of Fiorentina, Crotone, and Inter Milan (with 3-5-2 formation) and Juventus Turin (with 4-4-2 formation). Among other things, these results suggest that the rate at which players pass the ball is reduced when the team is on track to winning a match, compared to scenarios in which the team is neither on track to winning nor losing a match (holding everything else fixed). By contrast, when on track to losing a match, the rate at which players of Juventus Turin and Inter Milan pass the ball is reduced, while the rate at which players of Fiorentina and Crotone pass the ball is not reduced. There is an additional observation suggesting that the modus operandi of Juventus Turin and Inter Milan is different from the modus operandi of Fiorentina and Crotone: Starting a pass in the opponent's half of the field does not increase the probability of a successful pass among Fiorentina and Crotone players, but it does increase the probability of a successful pass among Juventus Turin and Inter Milan players. Taken together, these results suggest that the modus operandi of Juventus Turin and Inter Milan differs from the modus operandi of Fiorentina and Crotone, warranting more research into how these and other soccer teams operate and how the

modus operandi affects match outcomes. That said, we hasten to point out that we cannot make causal statements about how soccer teams operate. Causal inference for soccer and other team-based sports is a challenging but promising direction for future research, as we discuss in Section 8.5.

Among the position-specific effects, it is worth noting that the length of time the goal keeper controls the ball tends to be lower than the length of time other positions control the ball. This observation makes sense, because the goal keeper has an incentive to remove the ball from the penalty area as soon as possible, so that the opposing team cannot gain control of the ball in the penalty area and score an easy goal.

6.1 Posterior sensitivity checks

To assess the sensitivity of the posterior to the choice of prior, we consider the following three priors:

- Prior 1:

$$\alpha_k \overset{\text{iid}}{\sim} N(0, 5^2), \quad k = 1, \ldots, d_1, \qquad \beta_k \overset{\text{iid}}{\sim} N(0, 5^2), \quad k = 1, \ldots, d_2$$
$$\gamma_k \overset{\text{iid}}{\sim} N(0, 5^2), \quad k = 1, \ldots, d_3, \qquad \omega_k \overset{\text{iid}}{\sim} N(0, 5^2), \quad k = 1, \ldots, p;$$

- Prior 2, used in Section 6:

$$\alpha_k \overset{\text{iid}}{\sim} N(0, 10^2), \quad k = 1, \ldots, d_1, \qquad \beta_k \overset{\text{iid}}{\sim} N(0, 10^2), \quad k = 1, \ldots, d_2$$
$$\gamma_k \overset{\text{iid}}{\sim} N(0, 10^2), \quad k = 1, \ldots, d_3, \qquad \omega_k \overset{\text{iid}}{\sim} N(0, 10^2), \quad k = 1, \ldots, p;$$

- Prior 3:

$$\alpha_k \overset{\text{iid}}{\sim} N(0, 15^2), \quad k = 1, \ldots, d_1, \qquad \beta_k \overset{\text{iid}}{\sim} N(0, 15^2), \quad k = 1, \ldots, d_2$$
$$\gamma_k \overset{\text{iid}}{\sim} N(0, 15^2), \quad k = 1, \ldots, d_3, \qquad \omega_k \overset{\text{iid}}{\sim} N(0, 15^2), \quad k = 1, \ldots, p;$$

where $N(0, 5^2)$, $N(0, 10^2)$, and $N(0, 15^2)$ are Gaussians with mean 0 and variances $5^2 = 25$, $10^2 = 100$, and $15^2 = 225$, respectively. The random effects prior is described in Section 5 and is the same under all three priors. The posteriors under these priors are similar, as can be seen by comparing Tables 9 and 10 with the corresponding tables in Supplement F.

6.2 Posterior predictive checks

Using the posterior draws generated in Section 6, we compare model-based predictions of the waiting times between passes and the proportions of successful passes to the observed waiting times and the observed proportions of successful passes by Inter Milan, Crotone, and Fiorentina during the 2020/21 season. The model-based predictions (i.e., posterior predictions)

are shown in Figure 4 in Supplement G and match the observed data.

7. Simulation results

We simulate data from the stochastic modeling framework specified in Section 6. We choose the data-generating parameters of the model so that the simulated data mimic the Inter Milan data in Section 6. We simulate 100 short soccer seasons, each with 1,000 passes. To estimate the model from the 100 simulated soccer seasons, we leverage the Bayesian approach described in Section 5, using the prior described in Section 5. We present in Figure 5 in Supplement H aggregated simulations results based on all 100 simulated soccer seasons. In addition, we present the data-generating parameters along with posterior summaries of the parameters based on one of the 100 simulated soccer seasons in Table 15 in Supplement H. The figure and table demonstrate that the posterior means of the parameters cluster around the data-generating parameters.

8. Discussion

We view the proposed stochastic modeling framework as a first step to modeling soccer matches and other team-based sports as space- and time-indexed network processes and hope that it will stimulate future research.

To stimulate future research, we conclude with a short discussion of open questions and directions for future research.

8.1 Model specification

The deluge of high-resolution network data generated by soccer and other team-based sports implies that there are many possible features that may be relevant for predicting ball control, goals, and match outcomes. The specific features used in Section 6 make sense as a starting point, but sound model selection procedures and more data are needed to shed light on which features are useful for predicting ball control, goals, and match outcomes.

In addition, the proposed stochastic modeling framework includes player-specific random effects $\boldsymbol{\eta}_i \in \mathbb{R}^3$, which are correlated within players i but are shared across soccer matches. Since the proposed stochastic modeling framework is already fairly complex, we stick to the player-specific random effects. More advanced latent process models—e.g., multilevel models with position- and team-specific random effects and other more complex latent process models—constitute an interesting direction for future research.

8.2 Causal inference

While impressive progress has been made on the foundations of causal inference (e.g., Peters et al., 2017; Imbens and Rubin, 2015; Pearl, 2009), causal inference for soccer and other team-based sports poses challenges (e.g., Hall et al., 2002; Price et al., 2022; Dona and Swartz, 2023). First, conducting experiments in soccer is hard. Thus, causal inference needs to rely on observational rather than experimental data. Second, the outcomes of interest may be player-specific (e.g., scoring goals) or team-specific (e.g., winning matches) or both. Third, the outcomes of players and teams may depend on the outcomes of other players in the same team as well as the opposing team. As a result, there can be interference (Hudgens and Halloran, 2008; Sävje et al., 2021; Li and Wager, 2022), complicating causal inference. Last, but not least, soccer matches are network-, space-, and time-dependent processes, and stochastic processes aspiring to emulate them will have to reflect the complexity of these real-world processes. As a consequence, causal inference for soccer and other team-based sports is a challenging but promising direction for future research.

Acknowledgements

We acknowledge support by the U.S. National Science Foundation in the form of NSF awards SES-224129223058 and DMS-2412923 (GH) and the U.S. Department of Defense award ARO W911NF-21-1-0335 (MS). We are indebted to an anonymous Associate Editor and two anonymous referees for numerous constructive suggestions that greatly improved the manuscript.

References

Albert, J., M. E. Glickman, T. B. Swartz, and R. H. Koning (2017). *Handbook of Statistical Methods and Analyses in Sports*. Boca Raton, FL: Chapman & Hall/CRC.

Baio, G. and M. Blangiardo (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics 37*(2), 253–264.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodolgy 38*, 155–200.

Cattelan, M., C. Varin, and D. Firth (2013). Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(1), 135–150.

Cervone, D., A. D'Amour, L. Bornn, and K. Goldsberry (2014). POINTWISE: predicting points and valuing decisions in real time with NBA optical tracking data. In *Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA*, Volume 28, pp. 1–9.

Cervone, D., A. D'Amour, L. Bornn, and K. Goldsberry (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association (Applications and Case Studies) 111*(514), 585–599.

Chacoma, A., N. Almeira, J. I. Perotti, and O. V. Billoni (2020). Modeling ball possession dynamics in the game of football. *Physical Review E 102*(4), 042120.

Dixon, M. J. and S. G. Coles (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 46*(2), 265–280.

Dona, N. E. and T. Swartz (2023). A causal investigation of pace of play in soccer. *Statistica Applicata – Italian Journal of Applied Statistics 35*, 1–29.

Efron, B. (2022). *Exponential Families in Theory and Practice*. Cambridge, MA: Cambridge University Press.

Fernandez, J. and L. Bornn (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *Sloan Sports Analytics Conference*, Volume 2018.

Franks, A., A. Miller, L. Bornn, and K. Goldsberry (2015). Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics 9*(1), 94–121.

Glickman, M. E. and H. S. Stern (2005). A state-space model for national football league scores. In *Anthology of Statistics in Sports*, pp. 23–33. SIAM.

Hall, S., S. Szymanski, and A. S. Zimbalist (2002). Testing causality between team performance and payroll: The cases of major league baseball and English soccer. *Journal of Sports Economics 3*(2), 149–168.

Hirotsu, N., K. Inoue, K. Yamamoto, and M. Yoshimura (2023). Soccer as a Markov process: modelling and estimation of the zonal variation of team strengths. *IMA Journal of Management Mathematics 34*(2), 257–284.

Hu, G., H.-C. Yang, Y. Xue, and D. K. Dey (2023). Zero-inflated Poisson model with clustered regression coefficients: Application to heterogeneity learning of field goal attempts of professional basketball players. *Canadian Journal of Statistics 51*, 157–172.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*, 832–842.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(3), 381–393.

Li, S. and S. Wager (2022). Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics 50*, 2334 – 2358.

Miller, A. C. and L. Bornn (2017). Possession sketches: Mapping NBA strategies. In *Proceedings of the 2017 MIT Sloan Sports Analytics Conference*, pp. 1–12.

Narizuka, T., K. Takizawa, and Y. Yamazaki (2023). Validation of a motion model for soccer players' sprint by means of tracking data. *Scientific Reports 13*(1), 865.

Nguyen, Q., R. Yurko, and G. J. Matthews (2023). Here comes the strain: Analyzing defensive pass rush in American football with player tracking data. *arXiv preprint arXiv:2305.10262*.

Norris, J. R. (1997). *Markov Chains*. Cambridge: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2 ed.). Cambridge: Cambridge University Press.

Perry, P. O. and P. J. Wolfe (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 75*, 821–849.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge: MIT Press.

Price, K., H. Cai, W. Shen, and G. Hu (2022). How much does home field advantage matter in soccer games? A causal inference approach for English premier league analysis. *arXiv preprint arXiv:2205.07193*.

Sandholtz, N., J. Mortensen, and L. Bornn (2020). Measuring spatial allocative efficiency in basketball. *Journal of Quantitative Analysis in Sports 16*(4), 271–289.

Santos-Fernandez, E., F. Denti, K. Mengersen, and A. Mira (2022). The role of intrinsic dimension in high-resolution player tracking data—insights in basketball. *The Annals of Applied Statistics 16*, 326–348.

Sävje, F., P. Aronow, and M. Hudgens (2021). Average treatment effects in the presence of unknown interference. *The Annals of Statistics 49*, 673–701.

Shaw, L. and M. Glickman (2019). Dynamic analysis of team strategy in professional football. *Barça Sports Analytics Summit 13*.

Stadtfeld, C. (2011). *Events in Social Networks. A Stochastic Actor-oriented Framework for Dynamic Event Processes in Social Networks.* Ph. D. thesis, Karlsruhe Institute of Technology. Download: uvka.ubka.uni-karlsruhe.de/shop/download/1000025407.

Stan Development Team (2023). RStan: the R interface to Stan. R package version 2.21.8.

Wu, S. and L. Bornn (2018). Modeling offensive player movement in professional basketball. *The American Statistician 72*(1), 72–79.

Yurko, R., F. Matano, L. F. Richardson, N. Granered, T. Pospisil, K. Pelechrinis, and S. L. Ventura (2020). Going deep: models for

continuous-time within-play valuation of game outcomes in American football with tracking data. *Journal of Quantitative Analysis in Sports 16*(2), 163–182.

Yurko, R., S. Ventura, and M. Horowitz (2019). nflWAR: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports 15*(3), 163–183.

Department of Statistics, University of Missouri-Columbia

E-mail: grieshopn@missouri.edu

Department of Economics, University of Missouri-Columbia

E-mail: yong.feng@mail.missouri.edu

Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston

E-mail: guanyu.hu@uth.tmc.edu

Department of Statistics, The Pennsylvania State University

E-mail: michael.schweinberger@psu.edu