Simulating Tracking Data to Advance Sports Analytics Research

Demonstration Track

David Radke*
Chicago Blackhawks
Chicago, USA
dradke@blackhawks.com

Kyle Tilbury*
University of Waterloo
Waterloo, Canada
ktilbury@uwaterloo.ca

ABSTRACT

Advanced analytics have transformed how sports teams operate, particularly in episodic sports like baseball. Their impact on continuous invasion sports, such as soccer and ice hockey, has been limited due to increased game complexity and restricted access to high-resolution game tracking data. In this demo, we present a method to collect and utilize simulated soccer tracking data from the Google Research Football environment to support the development of models designed for continuous tracking data. The data is stored in a schema that is representative of real tracking data and we provide processes that extract high-level features and events. We include examples of established tracking data models to showcase the efficacy of the simulated data. We address the scarcity of publicly available tracking data, providing support for research at the intersection of artificial intelligence and sports analytics.

KEYWORDS

Agent-based Simulation, Sports Analytics

ACM Reference Format:

David Radke* and Kyle Tilbury*. 2025. Simulating Tracking Data to Advance Sports Analytics Research: Demonstration Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Using advanced analytics to inform decision making in sports has been an increasing trend over recent decades. This has been particularly prominent in discrete episodic sports like baseball and has revolutionized how front offices operate [2, 7]. However, the adoption of analytics in invasion style sports has been relatively limited due to more complex multiagent problems [12] and limited public access to data [9].

Most publicly available datasets in invasion sports record *event* data, or play-by-play data, indicating when an on-ball/puck event occurs, players involved, and location; however, event data fails to accurately capture the state of the game during the event (i.e., location of all other players). At the highest levels, professional ice hockey, basketball, and association football (soccer) leagues collect *tracking* data throughout games, providing the location and velocity of players and the ball/puck multiple times per-second. Joined with event data, tracking data provides more insight into the state of a game when players choose to take actions (or not to take actions) and has supported research at the intersection of multiagent systems and invasion games, such as measuring player coordination [10, 13], improving or evaluating joint actions [15, 16], and evaluating individual player actions [3].

Despite the crucial potential that tracking data has to support research at understanding inter-player behavior, public access to existing tracking data remains limited. Tracking datasets are typically protected by bargaining agreements, only giving access to a select few entities involved such as teams, broadcasters, or gambling companies. As a result, most research at the intersection of artificial intelligence (AI) and sports utilizes event data, limiting the complexity and usability of model architectures.

In this demo, we present a system to collect simulated ball and player tracking data from the Google Research Football (GRF) reinforcement learning (RL) environment [5] in a way that emulates the schema of real-world ball and player tracking data. We provide access to a pre-collected dataset that consists of 3,000 simulated soccer games [6] and provide a process to extract event data and different segments of continuous gameplay. Additionally, our demonstration includes two examples of popular models for player and team evaluation across different invasion sports: expected goals (xG) and pitch control. We show that the simulated dataset not only is able to support models typically built using real tracking data, but also offers an accessible solution for advancing research in sports analytics and multiagent systems. We include our code and pre-recorded dataset (`repository`) and our demo video (`video`).

2 TRACKING DATA AND SCHEMA

Most current tracking systems deployed across professional and amateur invasion sports record *center of mass* locations for each player using computer vision or hardware-based systems. This represents each player and the ball/puck as a three-dimensional coordinate on the playing surface that is referenced using unique player and team identification keys. While GRF has the ability to represent the game state as a rendered image, the headless state space of the environment represents players and the ball using similar center of mass coordinates. We utilize the headless representation of the state to record GRF data that imitates the schema of real world invasion game tracking datasets.

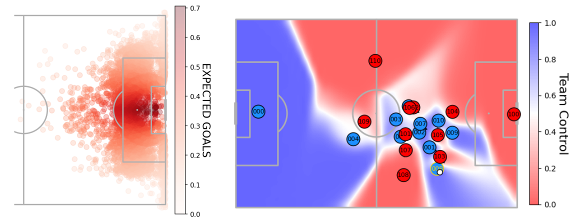
We store the recorded tracking data in a schema analogous to real invasion game tracking data [8]. Each game in our configuration lasts for 3,000 timesteps. We record center of mass data for all 22 agents on the field and the ball at each timestep, referred to as *entities*. Thus, each timestep contains 23 consecutive rows of data logging information on each entity for that moment in time,

*These authors contributed equally to this work.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

such as identification keys, coordinate locations of the entity in the environment, team affiliation, role on the team, and their current velocity. Agents are defined with one of eight roles (i.e., positions) within their team, representing different agent *types* that make up a larger team structure. Agent types and team structures in sporting domains represent similar areas of research across the broader field of multiagent systems [1, 11]. We extract a boolean variable from the environment to indicate if an agent is in possession of the ball.

We include a pre-collected dataset of tracking data from 3,000 simulated GRF games [6]. Outlined below, our demo provides the ability to collect more tracking data, extract high level features from the raw tracking data, and example models initially published with real football tracking data.



(a) Expected goals. **(b) Pitch control.**

Figure 1: Examples of expected goals (xG) and pitch control models with our simulated dataset.

3 COLLECT DATA AND EXTRACT EVENTS

Collecting Simulated Data. Our demo first involves a process to collect simulated data from the GRF environment [5]. Raw tracking data shows the positions and velocities of all players and the ball at each timestep of the game.

Extracting Events and Stints. While raw tracking provides a rich view of the state space, various models for invasion game sports involve determining key moments when players make decisions. We provide a system that extracts higher level *event* information from the raw tracking data such as passes, receptions, shots, turnovers, and interceptions. Passes and receptions indicate instances where a team transfers possession of the ball from one agent (i.e., player) to a teammate. Transitions that are not to a teammate are labeled as turnovers and interceptions. Shots represent events when an agent moves the ball in an attacking direction and either scores a goal, puts the ball out of play beyond the attacking goal line, or transfers possession to the opposing goaltender. Furthermore, many invasion game models utilize continuous segments of gameplay known as *stints*. We include a process to automatically identify stints from the raw tracking data and assign them unique identifiers. Extracting these actionable events and features provides the groundwork for more advanced analyses.

4 BUILDING MODELS WITH THE DATA

Event Data: Expected Goals (xG). Expected goals (xG) models aims to estimate the probability that a shot results in a goal. xG models are common across sports with goaltenders as a way to assign value to both shooters and goaltenders by comparing the predicted probability of a goal with the shot outcome. These models typically rely on information encoded in event data to model the environment of a shot, such as angle and distance of the shot location. While more advanced xG models can be developed with tracking data, we include features from our event extraction process to show the feasibility of that process.

We build an xG model using shot angle and distance from the goal. We align the shot direction so that all shots are modeled as shooting towards the *right* goal (i.e., highest x-coordinate direction) and measure shot location as coordinates of where the shot is taken from. We train a logistic regression model from the set of goals and non-goals using polar coordinates from the center of the opposing goal. Figure 1a shows the results of our logistic regression model on shots extracted from the pre-recorded recorded dataset. Each dot represents a shot location from either team and darker red represents higher probability of the shot being a goal. Consistent with work utilizing real soccer data, we observe a monotonic decline in scoring probability as the shot location moves further from the goal at increased angles [13].

Tracking Data: Pitch Control. A common advanced model in invasion sports analytics using tracking data is called *Pitch Control* [14]. Given a state of the game represented by tracking data, pitch control aims to quantify the probability that a player would possess the ball if the ball were moved to any location of the pitch. The union of teammates represents the probability that either team would gain possession of the ball. The concept of pitch control builds upon traditional Voronoi diagrams [4] to include information on player velocity, ball travel speed, and player control time in measuring player and team spatial control over the pitch. Pitch control models have been used to understand player passing abilities and decisions, including how players take actions to maximize pass completion while also maximizing the pitch *value* with on- and off-ball movements [3, 13, 14].

Figure 1b shows pitch control for an individual timestep of a simulated game. Players are divided into blue and red teams and each player has a unique identification number. The player in possession of the ball (player 008) is highlighted with a gold circle. Blue regions represent areas where the blue team has a higher probability of recovering the ball if it were moved to those locations and red represents the inverse for the red team. White locations represent areas where possession probability is close to 50% for either team.

5 CONCLUSION

This demonstration presents a data collection and analysis process for simulated player and ball tracking data using the Google Research Football environment. We show processes to extract additional features from the raw tracking data and show how real tracking-based models can be developed with this simulated data. Like real tracking data, several limitations exist in this dataset such as the lack of pose estimation, some omitted ball possessions, and various edge cases in the event extraction that are areas for future work. With the lack of publicly available tracking data for most researchers, this demonstration offers a tangible way to help forward research in artificial intelligence and sports analytics.

REFERENCES

[1] Stefano V Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66–95.

[2] Ramy Elitzur. 2020. Data analytics effects in major league baseball. *Omega* 90 (2020), 102001.

[3] Javier Fernández, Luke Bornn, and Daniel Cervone. 2021. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning* 110, 6 (2021), 1389–1427.

[4] S Kim. 2004. Voronoi analysis of a soccer game. *Nonlinear Analysis: Modelling and Control* 9, 3 (2004), 233–240.

[5] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 4501–4510.

[6] Kyle Tilbury, David Radke. 2024. Simulated Football Dataset of 3,000 Games. https://drive.google.com/drive/folders/1PZ8b-ftnqhIqMV0qnkTB_LWtCnBjhTdD. Accessed 2024-12-17.

[7] Michael Lewis. 2004. *Moneyball: The art of winning an unfair game.* WW Norton & Company.

[8] Shayegan Omidshafiei, Daniel Hennes, Marta Garnelo, Zhe Wang, Adria Recasens, Eugene Tarassov, Yi Yang, Romuald Elie, Jerome T Connor, Paul Muller, et al. 2022. Multiagent off-screen behavior prediction in football. *Scientific reports* 12, 1 (2022), 8638.

[9] Devin Pleuler. 2024. Unexpected Origins and the Fermi Paradox. https://www.centralwinger.com/p/unexpected-origins-and-the-fermi. Accessed: 12-02-2024.

[10] David Radke, Tim Brecht, and Daniel Radke. 2022. Identifying Completed Pass Types and Improving Passing Lane Models. In *Linköping Hockey Analytics Conference*. 71–86.

[11] David Radke, Kate Larson, and Tim Brecht. 2022. Exploring the Benefits of Teams in Multiagent Learning. In *IJCAI*.

[12] David Radke and Alexi Orchard. 2023. Presenting multiagent challenges in team sports analytics. *AAMAS* (2023).

[13] William Spearman. 2018. Beyond expected goals. In *Proceedings of the 12th MIT sloan sports analytics conference*. 1–17.

[14] William Spearman, Austin Basye, Greg Dick, Ryan Hotovy, and Paul Pop. 2017. Physics-based modeling of pass probabilities in soccer. In *Proceeding of the 11th MIT Sloan Sports Analytics Conference*, Vol. 1.

[15] Karl Tuyls, Shayegan Omidshafiei, Paul Muller, Zhe Wang, Jerome Connor, Daniel Hennes, Ian Graham, William Spearman, Tim Waskett, Dafydd Steel, et al. 2021. Game Plan: What AI can do for Football, and What Football can do for AI. *Journal of Artificial Intelligence Research* 71 (2021), 41–88.

[16] Zhe Wang, Petar Veličković, Daniel Hennes, Nenad Tomašev, Laurel Prince, Michael Kaisers, Yoram Bachrach, Romuald Elie, Li Kevin Wenliang, Federico Piccinini, et al. 2024. TacticAI: an AI assistant for football tactics. *Nature communications* 15, 1 (2024), 1906.