**Unveiling True Talent:**

**The Soccer Factor Model for Skill Evaluation**

Paper Track: Soccer

1. **Introduction**

Evaluating a soccer player's performance can be challenging due to the high costs and small margins in recruitment decisions, where team strength can obscure individual skill. **We introduce the Soccer Factor Model** (SFM), which isolates a player’s true skill from team influence. Complementing spatial and temporal analyses ([1], [2]), the SFM focuses on true latent skills.

A key innovation is the introduction of Skill Above Replacement (SAR) and Performance Above Replacement (PAR), adapted from baseball’s WAR metric ( [3], [4]) allowing analysts and managers to make accurate **player comparisons, forecast future performance, and run scenario analyses** that guide roster decisions.

1. **Methods**

The SFM leverages a **Bayesian framework** [5], providing skill estimates along with uncertainties, crucial for decision-making. Though applicable to various sports, our focus is soccer, using **ordered logistic regression to predict goal-scoring potential among strikers**. Player-specific skill and team strength variables are incorporated, and **Hilbert-Space Gaussian Processes** [6] capture nonlinear skill evolution, such as "aging curves".

Our **fully open-source dataset** spans the top four European soccer leagues—Premier League, Bundesliga, Serie A, and La Liga—from 2000/01 to 2023/24, derived from publicly available data scraped from [kicker.de](https://www.kicker.de/). It includes **over 33,000 match observations**, featuring 144 strikers and 13,000+ goals. **All data, code, and results are shared publicly**, enabling transparency and collaboration.

1. **Results**

The SFM provides clear, actionable insights. The new SAR and PAR metrics allow decision-makers to understand **how elite players perform compared to replacement-level players** (RLPs) — players easily sourced from the bench or transfers. Recruiters using the model can either input their own RLPs or apply numeric thresholds to define them.

**SAR strips away team influence to focus purely on a player’s skill, while PAR accounts for the boost a striker receives from their team**. For example, SAR highlights Ronaldo and Messi’s undeniable superiority in raw talent (**Figure 1**), but also highlights uncertainty for others like Haaland, where his true contribution could range from 0.18 to 0.5 goals above replacement per game.

PAR tells a different story. It shows **how team dynamics elevate players** like Ronaldo and Messi, with SAR < PAR indicating that their observed performance is boosted by their strong teams (**Figure 2**). This trend holds for most players displayed here, implying they appear more valuable due to team support. However, Messi, Ronaldo, and Kane maintain almost equal SAR and PAR, showing their brilliance transcends team context.

These metrics complement tools like video analysis and positional data, allowing teams to make more informed decisions that better account for player skill, team dynamics, and game context.

|  |  |
| --- | --- |
| **Figure 1: SAR - All Teams Equal**  A graph with blue and white lines  Description automatically generated | **Figure 2: SAR - PAR**  A graph with blue and white lines  Description automatically generated |

1. **Conclusion**

The Soccer Factor Model enables teams to **evaluate player skill independent of team influence**. The Bayesian SFM provides robust player comparisons, uncertainty quantification, scenario analysis, and future performance forecasts. The model empowers sports professionals to **make informed decisions that reduce player-selection risk**. By fully open-sourcing data and code, we contribute to the analytics community, fostering innovation and broader application in the sports industry.

**References**

|  |  |
| --- | --- |
| [1] | A. Solin and S. Särkkä, "Hilbert Space Methods for Reduced-Rank Gaussian Process Regression," *Stat Comput,* no. 30, pp. 419-446, 2020. |
| [2] | J. Salvatier, T. V. Wiecki and C. Fonnesbeck, "Probabilistic Programming in Python using PyMC3," *Peerj Computer Science,* vol. 2, no. e55, 2016. |
| [3] | W. Spearman, A. Basye, G. Dick, R. Hotovy and P. Pop, "Physics-Based Modeling of Pass Probabilities in Soccer," *Proceedings of the MIT Sloan Sports Analytics Conference,* 2017. |
| [4] | J. Fernandez and L. Bornn, "Wide Open Spaces: A Statistical Technique for Measuring Space Creation in Professional Soccer," *Proceedings of the MIT Sloan Sports Analytics Conference,* 2018. |
| [5] | B. S. Baumer, S. T. Jensen and G. J. Matthews, "openWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball," *Journal of Quantitative Analysis in Sports,* vol. 11, no. 2, pp. 69-84, 2015. |
| [6] | R. Yurko, S. Ventura and M. Horowitz, "nflWAR: A Reproducible Method for Offensive Player Evaluation in Football," *Journal of Quantitative Analysis in Sports,* vol. 15, no. 3, pp. 163-183, 2019. |

is not required, but if you have one please include it here.