Unveiling True Talent: The Soccer Factor Model for Skill Evaluation

Alexandre Andorra
Miami Marlins, PyMC Labs

Maximilian Göbel
Bocconi University

First Draft: April 16, 2024
This Draft: September 30, 2024

Acknowledgments: For helpful comments we thank, without implicating, Chris Fonnesbeck, Ravi Ramineni, Osvaldo Martin, Luciano Paz, Aaron MacNeil, Patrick Ward, Paul Sabin and Luke Bornn.

Key words: soccer analytics, Bayesian statistics, factor models, asset pricing

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 (Spearman et al., 2017; Fernandez and Bornn, 2018), 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 (Baumer et al., 2015; Yurko et al., 2019), allowing analysts and managers to make accurate **player comparisons, forecast future performance, and run scenario analyses** that guide roster decisions.

Methods

The SFM leverages a **Bayesian framework** (Salvatier et al., 2016), 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** (Solin and Särkkä, 2020) 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. 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.

Results

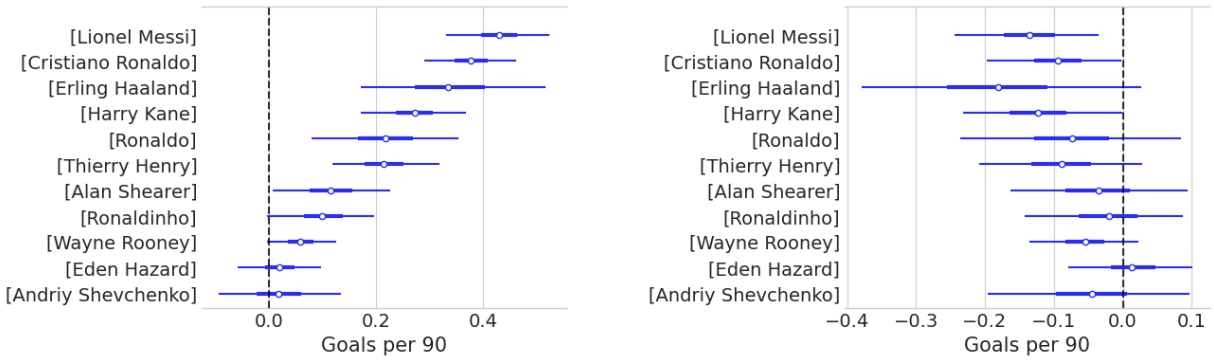
The SFM provides clear, actionable insights. The new SAR and PAR metrics allow decision-makers to understand **how elite players perform compared to replacement-level players** (RLPs) — players easily sourced from the bench or transfers. Recruiters using the model can either input their own RLPs or apply 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 (Fig. 1a), 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 (Fig. 1b). 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: Skill Above Replacement



(a) SAR: All Teams Equal

(b) SAR - PAR

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

Baumer, B. S., Jensen, S. T., and Matthews, G. J. (2015). openWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball. *Journal of Quantitative Analysis in Sports*, 11(2):69–84.

Fernandez, J. and Bornn, L. (2018). Wide Open Spaces: A Statistical Technique for Measuring Space Creation in Professional Soccer. *Proceedings of the MIT Sloan Sports Analytics Conference*.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic Programming in Python using PyMC3. *PeerJ Computer Science*, 2(e55).

Solin, A. and Särkkä, S. (2020). Hilbert Space Methods for Reduced-Rank Gaussian Process Regression. *Stat Comput*, (30):419–446.

Spearman, W., Basye, A., Dick, G., Hotovy, R., and Pop, P. (2017). Physics-Based Modeling of Pass Probabilities in Soccer. *Proceedings of the MIT Sloan Sports Analytics Conference*.

Yurko, R., Ventura, S., and Horowitz, M. (2019). nflWAR: A Reproducible Method for Offensive Player Evaluation in Football. *Journal of Quantitative Analysis in Sports*, 15(3):163–183.