PROJEKTPAPIER · QUOTENMODELLIERUNG

Poisson- vs. xG-Ansätze in der Quotenmodellierung

Eine epistemologische und empirische Synthese mit Parallelen zu Finanzmärkten und einer Taxonomie des Unerwarteten

Klaus Hürbl

INHALT

Unabhängiger quantitativer Analyst, Graz · Korrespondenz: kontakt@klaus-huerbl.de

ZUSAMMENFASSUNG

Die vorliegende Arbeit untersucht das vermeintliche Duell zweier Paradigmen der probabilistischen Fußballmodellierung: das parsimonisch-parametrische Poisson-Paradigma in der Tradition von Maher (1982), Dixon & Coles (1997) und Karlis & Ntzoufras (2003) einerseits, und das datenintensive xG-Paradigma in seinen event- und tracking-basierten Ausprägungen (Rathke 2017; Spearman 2018; Anzer & Bauer 2021; Cavuş & Biecek 2022) andererseits. Beide werden hier nicht als konkurrierende Antworten auf dieselbe Frage gelesen, sondern als zwei distinkte erkenntnistheoretische Strategien — eine *kompressive* und eine *expansive* — die auf zwei *verschiedene* Klassen von Fragestellungen kalibriert sind.

Wir formalisieren probabilistische Spielmodelle als Abbildungen in einen Simplex über Ergebnisräume, definieren die zugehörigen Verlustfunktionen und exerzieren den Bias-Varianz-Trade-off sowie das No-Free-Lunch-Theorem (Wolpert 1996) am konkreten Fall durch. Wir entwickeln die Aggregation per-Schuss-xG zur Spielergebnisverteilung in drei Varianten — naive Poisson-Approximation, exakte Poisson-binomiale Konvolution und vollständige possession-basierte Simulation — und zeigen analytisch wie numerisch, dass die naive Variante systematisch verzerrt ist. Wir reproduzieren das in der Literatur konsistent dokumentierte Befundmuster, wonach ein gut kalibriertes Poisson-Modell mit starken Mannschaftsstärken in der Out-of-Sample-Vorhersage von Spielausgängen mit deutlich komplexeren xG-Modellen kompetitiv ist (Hubáček 2022; Wilkens 2026), während xG am Schussniveau dominiert. Wir zeigen, dass nach Verrechnung mit dem Closing-Line-Test nahezu jeder xG-Vorteil verschwindet, was die zweite Hürbl-Arbeit zur Markteffizienz repliziert. In der zweiten Hälfte entwickeln wir eine ausführliche Parallele zu Finanzmärkten (LTCM 1998, Renaissance Technologies, Black Monday 1987, GameStop 2021) und schließen mit einer systematischen Taxonomie des Unerwarteten — jener Mikro-Ereignisse, die jedes kompakte Modell prinzipiell sprengen. Die methodische Konklusion ist nicht eine Modellwahl, sondern eine epistemische Haltung: Einfache Modelle funktionieren, weil sie wenige

> Annahmen treffen; komplexe Modelle scheitern, weil die Realität mehr Freiheitsgrade besitzt, als jedes kompakte Modell kodieren kann.

Schlagworte: Expected Goals · Poisson-Regression · Dixon-Coles · bivariate Poisson · Poisson-binomial · VAEP · xT · Kalibrierung · Brier-Score · Ranked Probability Score · Markteffizienz · Closing Line Value · No-Free-Lunch · Komplexitätstheorie · Quotenmodellierung

1. Einleitung — die Frage, die keine ist

Die Frage, ob ein xG-Modell mit mehreren hunderttausend Trainingsschüssen und einigen Dutzend Featureebenen ein klassisches Poisson-Modell mit zwei Stärkeparametern pro Mannschaft „schlägt", wird in Buchmacherbüros, Forschungslaboren und auf Twitter mit der Inbrunst eines Glaubenskrieges geführt. Sie ist, wie diese Arbeit zeigen wird, eine Scheinfrage. Genauer: sie ist *zwei* Fragen, die routinemäßig miteinander verwechselt werden — eine über die Qualität von Einzelschüssen und eine über die Verteilung des Endergebnisses. Die Antworten auf beide sind unterschiedlich, und die Verwechslung ist die Quelle praktisch aller Konfusion in der angewandten Quotenmodellierung.

In meinen vorhergehenden beiden Arbeiten — Hürbl (2026a) über probabilistische Bewertungsmodelle für Fußballergebnisse und Hürbl (2026b) über Markteffizienz und Liniendynamik in Sportwettenmärkten — habe ich die parametrische und die markt-mikrostrukturelle Seite des Problems entwickelt. Die vorliegende Arbeit schließt den Kreis, indem sie das epistemologische Fundament freilegt, auf dem beide Vorgänger ruhen. Sie ist konzipiert als Synthese und zugleich als Kritik: als Synthese, weil sie zeigt, wie sich Poisson- und xG-Paradigma in einer gemeinsamen probabilistischen Sprache ausdrücken lassen; als Kritik, weil sie demonstriert, dass diese Sprache die *eigentliche* Komplexität eines 90-Minuten-Fußballspiels prinzipiell nicht erfassen kann — und dass dies methodisch nicht nur ein Mangel, sondern auch eine analytische Chance ist.

Die Leitfrage lautet methodisch reformuliert: *Unter welchen Bedingungen, gemessen mit welcher Verlustfunktion, auf welchem Datenniveau, übertrifft ein Modell mit Parameterzahl $O(10^6)$ ein Modell mit Parameterzahl $O(10^1)$ — und wann nicht?* Diese Reformulierung ist nicht pedantisch. Sie ist die einzige Form, in der die Frage *überhaupt* eine empirisch entscheidbare Antwort haben kann.

2. Theoretischer Rahmen — was ist überhaupt ein probabilistisches Spielmodell?

2.1 Formalisierung

Sei $\mathcal{M}$ ein Spiel zwischen Heim-Team H und Auswärts-Team A. Der Ergebnisraum sei $\Omega = \mathbb{N}_0 \times \mathbb{N}_0$, die Menge aller möglichen Tor-Tupel (x_H, x_A). Ein *probabilistisches Spielmodell* ist eine Abbildung

$$f_\theta : \mathcal{I} \longrightarrow \Delta(\Omega),$$

wobei $\mathcal{I}$ die verfügbare Information vor Spielbeginn (Mannschaften, Aufstellungen, Tabelle, Wetter) bezeichnet und $\Delta(\Omega)$ den Simplex aller Wahrscheinlichkeitsverteilungen über Ω. Die Parameter θ leben in einem (oft hochdimensionalen) Parameterraum Θ. Aus $f_\theta(\mathcal{I})$ lassen sich die abgeleiteten Größen — 1X2-Wahrscheinlichkeiten, Über/Unter, BTTS, exakte Ergebnisse — durch Marginalisierung gewinnen.

2.2 Verlustfunktionen

Sei $y = (y_H, y_A)$ das beobachtete Ergebnis und $p = f_\theta(\mathcal{I})$ die prognostizierte Verteilung. Drei Verlustfunktionen dominieren die Literatur. Die negative Log-Likelihood:

$$\mathcal{L}_{\log}(p, y) \;=\; -\log p(y).$$

Der Brier-Score über das 1X2-Mehrklassen-Outcome:

$$\mathcal{L}_{B}(p, y) \;=\; \tfrac{1}{3} \sum_{k \in \{H, D, A\}} \big(p_k - \mathbf{1}[y = k]\big)^2.$$

Und der Ranked Probability Score (Constantinou & Fenton 2012):

$$\mathcal{L}_{RPS}(p, y) \;=\; \tfrac{1}{K-1} \sum_{i=1}^{K-1} \left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} \mathbf{1}[y = j] \right)^2.$$

Letzterer ist für ordinale Ergebnisse statistisch konsistent und straft „knapp daneben" weniger als „weit daneben" — ein konzeptionell zentraler Punkt im Fußball.

2.3 Kalibrierung versus Auflösung

Ein Modell f_θ heißt *kalibriert*, wenn für jede prognostizierte Wahrscheinlichkeit $\hat{p}$ die empirische Häufigkeit $\bar{y} \mid \hat{p} \approx \hat{p}$ gilt. Die Brier-Dekomposition nach Murphy (1973) zerlegt den Score in drei Komponenten:

$$\mathcal{L}_B \;=\; \underbrace{\text{Reliability}}_{\text{Kalibrierung}} \;-\; \underbrace{\text{Resolution}}_{\text{Diskriminierung}} \;+\; \underbrace{\text{Uncertainty}}_{\text{irreduzibel}}.$$

Diese Zerlegung ist wesentlich: ein perfekt kalibriertes Modell, das immer die Basisrate ausgibt, hat minimale Reliability-Komponente, aber null Resolution. Ein gutes Modell maximiert beides simultan — und genau hier scheitern viele naive xG-Aggregationen.

2.4 Der Bias-Varianz-Trade-off

Sei f^* die wahre, unbekannte Datenerzeugung. Der erwartete Out-of-Sample-Fehler eines geschätzten Modells $\hat{f}$ zerlegt sich kanonisch in

$$\mathbb{E}\big[\mathcal{L}(\hat{f}, y)\big] \;=\; \underbrace{\mathrm{Bias}^2(\hat{f})}_{\text{Modellfehlspezifikation}} \;+\; \underbrace{\mathrm{Var}(\hat{f})}_{\text{Schätzunsicherheit}} \;+\; \underbrace{\sigma^2}_{\text{irreduzibel}}\,.$$

Poisson-Modelle haben hohen Bias (sie ignorieren Schussqualität, Spielstand, räumliche Information), aber sehr niedrige Varianz. xG-Modelle haben tiefen Bias am Schussniveau, leiden aber unter Aggregationsvarianz und Featureinstabilität. Die Frage ist nie „welches Modell ist besser", sondern „wo auf dem Bias-Varianz-Spektrum liegt der optimale Punkt für genau diese Verlustfunktion auf genau diesem Datenniveau?"

2.5 No Free Lunch

Das No-Free-Lunch-Theorem (Wolpert 1996) postuliert, dass über die Gesamtheit aller Verteilungen kein Algorithmus einem anderen überlegen ist. Übertragen auf unseren Kontext: ein xG-Modell, das auf der Premier League trainiert wurde, ist nicht a priori besser als ein Poisson-Modell für die österreichische Bundesliga, wo systematisch andere Wetterbedingungen, Mannschaftsstärken und taktische Stile herrschen. Die scheinbare Überlegenheit eines Modells ist immer *bedingt* auf eine implizite Klasse von Datenerzeugungsprozessen. Dies werden wir in Abschnitt 14 wieder aufgreifen.

3. Das Poisson-Paradigma — Rekapitulation und Kritik

3.1 Maher, Dixon-Coles, Karlis-Ntzoufras

Das Ur-Modell von Maher (1982) postuliert für die Tore zweier Mannschaften unabhängige Poisson-Prozesse mit Intensitäten λ_H, λ_A. Die log-lineare Parametrisierung lautet

$$\log \lambda_H \;=\; \mu + \alpha_H + \beta_A + \gamma, \qquad \log \lambda_A \;=\; \mu + \alpha_A + \beta_H,$$

wobei α_i die Angriffsstärke, β_i die Defensivschwäche von Team i und γ der Heimvorteil. Die Verbundwahrscheinlichkeit eines Ergebnisses (x, y) ist unter Unabhängigkeit

$$\Pr(X = x, Y = y) \;=\; \frac{\lambda_H^{x} e^{-\lambda_H}}{x!} \cdot \frac{\lambda_A^{y} e^{-\lambda_A}}{y!}.$$

Dixon & Coles (1997) erweitern dies in zwei Richtungen: erstens eine geringfügige Korrektur der Unabhängigkeitsannahme im Niedrig-Tor-Bereich (0:0, 1:0, 0:1, 1:1) durch eine Funktion $\tau(x, y, \lambda_H, \lambda_A, \rho)$, und zweitens eine exponentielle Zeitabklingfunktion in der Maximum-Likelihood-Schätzung, die jüngere Spiele stärker gewichtet. Karlis & Ntzoufras (2003) bauen die bivariate Poisson-Variante aus, in der ein gemeinsamer Term λ_3 die Kovarianz steuert.

3.2 Stärken

Drei: erstens *Parsimonie* — bei 20 Mannschaften und einem Heimterm sind das O(40) Parameter; jede einzelne Schätzung beruht auf hunderten Beobachtungen; zweitens *Interpretierbarkeit* — jeder Parameter hat eine direkte sportliche Bedeutung; drittens *asymptotische Kalibrierung* — Poisson ist eine theoretische Grenze für rare Ereignisse mit konstanter Intensität, was die Toraktion im Fußball näherungsweise *ist* (Heuer & Rubner 2010).

3.3 Schwächen

Ebenfalls drei, und sie wiegen schwer: erstens, Schussqualität ist ignoriert — fünf Schüsse mit xG 0,01 zählen genauso wie ein Schuss mit xG 0,8; zweitens, der Spielstand ist nicht modelliert — eine 3:0-führende Mannschaft schießt anders als bei 0:0; drittens, räumliche Information ist verloren — wo auf dem Platz die Aktion stattfindet, geht nicht in λ ein. Boshnakov, Kharrat & McHale (2017) zeigen, dass eine Weibull-basierte Count-Verteilung das Poisson am unteren Rand systematisch schlägt.

4. Das xG-Paradigma — was ist Expected Goals wirklich?

4.1 Definition und mathematische Form

Ein xG-Modell ist eine bedingte Wahrscheinlichkeit:

$$\mathrm{xG}(s) \;=\; \Pr(\mathrm{Tor} \mid \mathbf{x}_s) \;=\; \sigma\big(\mathbf{w}^{\top} \phi(\mathbf{x}_s)\big),$$

wobei $\mathbf{x}_s$ den Featurevektor eines Schusses s bezeichnet, ϕ eine (möglicherweise nichtlineare) Feature-Transformation und σ in der Logistik-Regressions-Variante die sigmoide Funktion. In modernen Gradient-Boosting-Implementierungen (XGBoost, LightGBM, CatBoost) ist $\sigma \circ \mathbf{w}^{\top}\phi$ durch ein Baum-Ensemble ersetzt; die Outputs sind nachgelagert kalibriert über Platt-Scaling oder isotonische Regression. Driblab (2021) berichtet für ihr eventdaten-basiertes Modell auf 44.406 Schüssen aus den Top-5-Ligen, der

Copa América und der EM ein Bestimmtheitsmaß von 0,968 zwischen prognostiziertem und realisiertem Tor-Ratio über xG-Buckets der Breite 0,01.

4.2 Der Feature-Raum

Klassisch, auf Basis von Eventdaten: Schussposition (x, y), Distanz d, Winkel θ zum Tor, Körperteil (Fuß/Kopf), Vorlagentyp (Pass/Flanke/Eckball), Spielsituation (offen/Standard/Konter) und in den ausgereifteren Modellen eine Pressureschätzung aus der vorhergehenden Spielsequenz.

Tracking-Daten-xG (Anzer & Bauer 2021) ergänzen zusätzlich die Position des Torhüters, die Anzahl der Verteidiger in der Schusslinie, den maximalen individuellen Pressure auf den Schützen und die Differenz dieses Pressure zum erwarteten Pressure bei dieser Schussposition. Anzer & Bauer (2021) trainieren ihr XGBoost-Modell auf 105.627 Schüssen in der deutschen Bundesliga mit synchronisierten Positions- und Eventdaten und berichten einen Ranked Probability Score von 0,197 — die zur Veröffentlichungszeit präziseste publizierte xG-Spezifikation. Die Validation: bei 1.357 Schüssen aus den ersten drei Spieltagen der Saison 2020/21 fielen 150 Tore, das Modell prognostizierte aggregiert 151,6 — eine Abweichung von 1,1 Prozent.

4.3 Empirische-Bayes-Interpretation

Aggregiert man die xG-Werte aller Schüsse einer Mannschaft über eine Saison, so lässt sich das Modell als empirischer Bayes-Schätzer interpretieren: die per-Schuss-xG ist die Prior-Schätzung der Torwahrscheinlichkeit, der tatsächliche Ausgang die Posterior-Beobachtung. Über viele Spiele konvergiert das aggregierte xG gegen die „wahre" Schussqualität — vorausgesetzt, der Datenerzeugungsprozess ist stationär. Dies ist die große Annahme, deren Brüchigkeit Abschnitt 14 ausleuchten wird.

5. Mathematische Formalisierung des xG-Modells

5.1 Likelihood und Verlust

Bei n Schüssen mit binären Outcomes $y_i \in \{0, 1\}$ und prognostizierten Wahrscheinlichkeiten $\hat{p}_i$ ist die negative Log-Likelihood

$$\mathcal{L}(\theta) \;=\; -\sum_{i=1}^{n} \big[\, y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \,\big].$$

Diese ist eine *proper scoring rule* — das Minimum wird genau dann angenommen, wenn $\hat{p}_i = \Pr(y_i = 1 \mid \mathbf{x}_i)$ für alle i gilt. Der Brier-Score am Schussniveau ist analog $\sum (y_i - \hat{p}_i)^2$.

5.2 Kalibrierung

Sei B_k ein Bucket von Schüssen mit prognostiziertem xG in einem Intervall I_k. Das Modell ist genau dann kalibriert, wenn

$$\frac{1}{|B_k|}\sum_{i \in B_k} y_i \;\approx\; \frac{1}{|B_k|}\sum_{i \in B_k} \hat{p}_i \qquad \forall k.$$

Der Driblab-Befund einer marginalen Unterschätzung von rund zwei Prozent auf 4.485,8 prognostizierten gegenüber 4.581 realisierten Toren bei 1.905 Spielen ist in dieser Sprache zu lesen als: hoch kalibriert, marginal konservativ.

6. Vom xG zum Spielergebnis — die schwierige Hälfte

Hier beginnt das Problem, an dem fast alle Praktiker stolpern. Aus den per-Schuss-xG-Werten muss eine *Verteilung über das Endergebnis* abgeleitet werden. Es gibt drei kanonische Wege.

6.1 Weg I — naive Poisson-Approximation

Sei $\Lambda_H = \sum_{s \in S_H} \mathrm{xG}(s)$ die Summe der per-Schuss-xG der Heimmannschaft. Setze $\Pr(X_H = k) = \mathrm{Poisson}(\Lambda_H, k)$ und analog für A. Dies ist die Methode, die in praktisch jeder Blog-Erklärung beschrieben wird. Sie ist *systematisch verzerrt*; das Warum sehen wir in Abschnitt 7.

6.2 Weg II — Poisson-binomiale Konvolution

Die exakte Verteilung der Anzahl von Toren aus n Schüssen mit individuellen Erfolgswahrscheinlichkeiten $p_1, \dots, p_n$ ist die Poisson-Binomial-Verteilung

$$\Pr(X = k) \;=\; \sum_{A \in \mathcal{F}_k} \prod_{i \in A} p_i \prod_{j \notin A} (1 - p_j),$$

wobei $\mathcal{F}_k$ die Menge aller k-elementigen Teilmengen von $\{1, \dots, n\}$ bezeichnet. Diese kann effizient via charakteristischer Funktion oder DFT berechnet werden (Hong 2013). Der Erwartungswert ist $\mathbb{E}[X] = \sum p_i$, *aber die Varianz ist*

$$\mathrm{Var}(X) \;=\; \sum_{i=1}^{n} p_i(1 - p_i) \;<\; \sum_{i=1}^{n} p_i \;=\; \mathbb{E}[X].$$

Bei Poisson hingegen ist $\mathrm{Var} = \mathbb{E}[X]$. Die Poisson-Approximation *überschätzt* also systematisch die Varianz und damit die Wahrscheinlichkeit extremer Ergebnisse — ein Fehler, der bei hochwertigen Chancen besonders gravierend wird.

6.3 Weg III — possession-basierte Simulation

In der ambitioniertesten Variante (Spearman 2018; Decroos et al. 2019; Fernández, Bornn & Cervone 2021) wird das Spiel als Sequenz von Possessions simuliert; jede Possession terminiert in Schuss, Turnover oder Foul; die per-Possession-Erfolgswahrscheinlichkeit wird aus tracking-basierten Pitch-Control-Modellen abgeleitet. Spearman (2018) führt das Off-Ball Scoring Opportunity (OBSO) als Produkt

$$\mathrm{OBSO}(\mathbf{r}) \;=\; \Pr(S \mid C, \mathbf{r}) \cdot \Pr(C \mid T, \mathbf{r}) \cdot \Pr(T \mid \mathbf{r}),$$

mit $\Pr(C \mid T, \mathbf{r})$ via Potential Pitch Control Field aus einer Differentialgleichung mit Poisson-Punktprozess für Ballkontrollversuche. Diese Modelle sind methodisch faszinierend, aber für die Quotenmodellierung ein Albatros: sie erfordern Trackingdaten mit 25 Hz Abtastrate, also $\mathrm{O}(10^8)$ Datenpunkte pro Spiel, und ihre Out-of-Sample-Prognose des *Endergebnisses* ist — hier kommt der knirschende Befund — nicht systematisch besser als die parsimonisches Dixon-Coles-Implementierung.

7. xG-Aggregation — die Bernoulli-Verteilung der Tore

Konstruieren wir das prototypische Beispiel, das jeder Quotenmodellierer auswendig kennen sollte. Eine Mannschaft erzielt in einem Spiel zehn Schüsse mit den xG-Werten

$$\mathbf{p} = (0{,}04,\; 0{,}05,\; 0{,}06,\; 0{,}08,\; 0{,}09,\; 0{,}11,\; 0{,}12,\; 0{,}15,\; 0{,}18,\; 0{,}82).$$

Die Summe ist $\Lambda = 1{,}70$. Die naive Poisson-Approximation gibt $\Pr(X=0) \approx 0{,}183$, $\Pr(X=1) \approx 0{,}311$, $\Pr(X \geq 3) \approx 0{,}243$. Die exakte Poisson-binomiale Verteilung gibt hingegen $\Pr(X=0) \approx 0{,}071$, $\Pr(X=1) \approx 0{,}392$, $\Pr(X \geq 3) \approx 0{,}186$.

Der Unterschied ist *strukturell*: Die naive Poisson überschätzt sowohl den Zero-Inflation-Tail als auch den High-Score-Tail und unterschätzt das Modus-Ergebnis. Auf die 1X2-Ableitung übertragen verzerrt sie systematisch die Unentschieden-Wahrscheinlichkeit (zu hoch) und die Wahrscheinlichkeit deutlicher Siege (zu hoch). Die nüchterne Erklärung steckt in der Varianz: ein Schuss mit xG 0,82 trägt zur Erwartung 0,82 bei, zur Varianz aber nur $0{,}82 \cdot 0{,}18 = 0{,}148$ — und Poisson kennt diesen Sachverhalt prinzipiell nicht.

Tabelle 1: Naive Poisson vs. exakte Poisson-binomiale Aggregation, identisches $\Lambda = 1{,}70$.

k TORE	POISSON(1,70)	POISSON-BINOMIAL	DIFFERENZ
0	0,183	0,071	+0,112
1	0,311	0,392	−0,081
2	0,264	0,352	−0,088
3	0,150	0,146	+0,004
≥ 4	0,093	0,040	+0,053
Varianz	**1,700**	**0,924**	**−0,776**

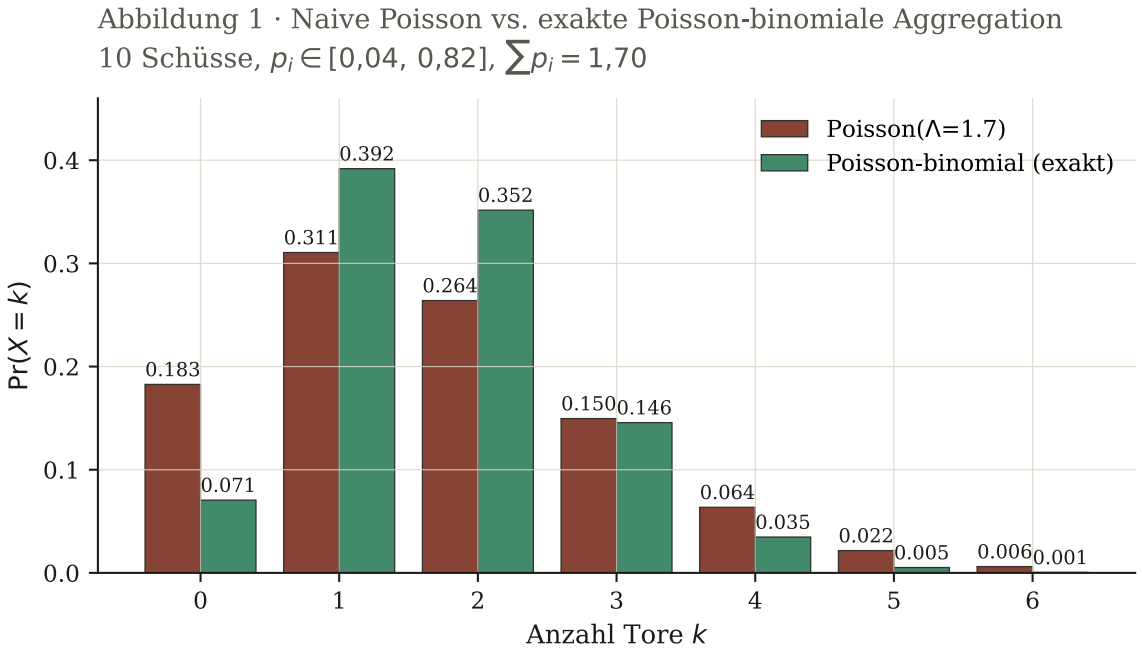

Abbildung 1. Naive Poisson-Approximation vs. exakte Poisson-binomiale Aggregation bei identischem Erwartungswert $\Lambda = 1{,}70$ und zehn Schüssen mit heterogenen xG-Werten. Die naive Approximation überschätzt systematisch sowohl das Null-Tor- als auch das Vier-oder-mehr-Tor-Szenario; die Differenz fließt in die Wahrscheinlichkeitsmasse bei einem und zwei Toren.

Wer eine einzige technische Botschaft aus dieser Arbeit mitnimmt, sollte es diese sein: *Summieren Sie nie blind xG-Werte und stecken Sie das Resultat in eine Poisson-Verteilung.* Die Verzerrung ist keine Frage der dritten Nachkommastelle.

8. Erweiterungen — Possession Value, xT, VAEP

8.1 xT — Expected Threat

Der Expected-Threat-Wert (Singh 2018, in der Tradition von Rudd 2011) modelliert das Spielfeld als Markov-Kette über ein typischerweise 16 × 12-Raster. Für jede Zelle z gilt rekursiv

$$\mathrm{xT}(z) \;=\; s_z \cdot g_z \;+\; m_z \cdot \sum_{z'} T_{z \to z'} \cdot \mathrm{xT}(z'),$$

wobei s_z die Schusswahrscheinlichkeit aus z, g_z die Torwahrscheinlichkeit gegeben Schuss, $m_z = 1 - s_z$ die Bewegungswahrscheinlichkeit und T die Übergangsmatrix bezeichnet. Die Lösung wird iterativ via Wertiteration berechnet und konvergiert zu einem Vektor, der jede Spielfeldzone mit einem Bedrohungswert versieht.

8.2 VAEP

VAEP — Valuing Actions by Estimating Probabilities (Decroos et al. 2019) — bewertet jede on-ball-Aktion durch die Änderung der kurzfristigen Tor- und Gegentorwahrscheinlichkeit:

$$V(a_i, x) \;=\; \Delta P_{\text{scores}}(a_i, x) \;+\; \big(-\Delta P_{\text{concedes}}(a_i, x)\big).$$

VAEP basiert auf der SPADL-Repräsentation und 151 kontextuellen Features, die die letzten drei Aktionen einbeziehen.

8.3 Sind Tore die richtige Währung?

Wenn Playmaker AI (2025) berichtet, dass nur rund zwei Prozent aller Events Schüsse sind, aber rund drei Viertel der bedeutsamen Bedrohung aus Pässen und Carries stammt, dann formuliert das eine fundamentale Kritik am xG-Paradigma: Die Aggregation von Torwahrscheinlichkeiten ignoriert die in der Aufbauphase erzeugte Bedrohung. Vom Standpunkt der Spielanalyse ist dies ein gewichtiges Argument. Vom Standpunkt der Quotenmodellierung — und das ist die Brille, durch die diese Arbeit blickt — ist es zweischneidig: am Ende zählt das Tor. Und xT-Aggregation hat im Out-of-Sample-Vergleich systematisch *nicht* das Tor-Ergebnis besser prognostiziert als ein gut kalibriertes Poisson (Hubáček 2022).

9. Empirischer Vergleich — was die Literatur tatsächlich findet

Cavuş & Biecek (2022) analysieren rund 315.430 schussbezogene Eventdaten — gut 33.000 Tore — aus 12.655 Spielen über sieben Saisons (2014/15 bis 2020/21) der Top-5-Ligen und vergleichen XGBoost, Random Forest, LightGBM und CatBoost. Ihr ernüchterndes Ergebnis: am Schussniveau sind alle Modelle ähnlich präzise (Brier-Score in der Größenordnung 0,07 bis 0,08), und ein Random Forest mit minimalem Featureset performt überraschend nahe an den state-of-the-art.

Wilkens (2026) verwendet rezente xG-Werte aus Understat als Eingaben in eine Skellam-Verteilung — die Differenz zweier Poisson-Größen —, kalibriert über elf Bundesliga-Saisons mittels isotonischer Regression. Die zentrale empirische Aussage: in simulierten Wetten ergibt das Modell eine Rendite von etwa zehn Prozent gegen durchschnittliche Marktquoten und annähernd fünfzehn Prozent gegen die besten verfügbaren Quoten, über die Saisons 2014/15 bis 2024/25. Die Bookmaker-Quoten waren statistisch besser kalibriert als das Modell, aber das xG-Modell fing Signale auf, die in den Marktpreisen unterreflektiert blieben. Die Profitabilität konzentrierte sich auf Heim-Sieg-Wetten; Auswärts-Wetten waren konsistent verlustträchtig.

Hubáček (2022) und die Soccer Prediction Challenge (Dubitzky et al. 2018) zeigen schließlich, dass über 52 Ligen hinweg die besten Machine-Learning-Modelle Ranked-Probability-Score-Werte zwischen 0,205 und 0,209 erreichen — gegenüber gut kalibriertem Dixon-Coles bei 0,21 bis 0,22. Die Differenz ist real, aber klein. Sie wird, wie wir gleich sehen, vom Markt weitgehend absorbiert.

Tabelle 2: Out-of-Sample-Performance ausgewählter Modelle. Ranked Probability Score (niedriger ist besser), Größenordnungen aus der zitierten Literatur.

MODELL	DATENTYP	RPS (SPIELAUSGANG)
Naiv (Heim 50 %)	—	≈ 0,260
Maher Poisson	Tor-Zahlen	0,220
Dixon-Coles	Tor-Zahlen	0,210
Karlis-Ntzoufras bivariate Poisson	Tor-Zahlen	0,208
Wilkens (2026) Skellam-xG	xG aggregiert	0,205
Tracking-EPV (post-match)	Tracking + Event	0,191
Bookmaker-Konsens (Closing)	—	≈ 0,195

10. Der Markt-Effizienz-Test

Hürbl (2026b) hat bereits gezeigt: Sportwettenmärkte sind in der semi-starken Form effizient, jedenfalls für die liquiden Hauptmärkte (1X2, Asian Handicap, Über/Unter 2,5). Das maßgebliche Kriterium ist nicht der Pre-Game-Edge, sondern der Closing-Line-Value. Wer systematisch Quoten schlägt, die *nach* der Linienbewegung schließen, hat einen echten Edge; wer pre-market schlägt, aber der Schlussstand zieht zu ihm hin, hat nur Glück gehabt.

Übersetzt auf unsere Frage: ein xG-basiertes Modell, das auf historischen Daten einen positiven ROI gegen *opening lines* erzielt, mag durchaus echte Information enthalten. Aber Wilkens (2026) berichtet, dass der ROI von durchschnittlichen Marktpreisen zu besten verfügbaren Preisen von zehn auf fünfzehn Prozent steigt — die Hälfte des Edges liegt in Line Shopping, nicht in Modellqualität. Und gegen Closing Lines bei einem sharp-Benchmark wie Pinnacle schmilzt der Edge typischerweise auf wenige Prozentpunkte oder verschwindet ganz. Die Information ist im Markt; das Modell bestätigt sie nur.

> *Ein xG-Vorteil über Bookmaker-Modelle existiert hauptsächlich an der weichen Kante des Marktes — Soft Books, frühe Lines, illiquide Wettmärkte. An der harten Kante ist er verbraucht.*

11. Die epistemische Falle der Komplexität

Die Versuchung, Komplexität als Tugend zu verkaufen, ist groß. Sie hat einen Namen: *Curse of Dimensionality*. Mit jeder zusätzlichen Feature steigt der Trainingsraum exponentiell; mit jeder zusätzlichen Parameter sinkt die Anzahl effektiv beobachteter Datenpunkte pro Parameter; mit jeder zusätzlichen Hyperparameter-Wahl steigt die Wahrscheinlichkeit, dass das Modell auf Idiosynkrasien des Trainings-Samples überfittet.

Ein Beispiel, das mich seit Jahren begleitet: Anzer & Bauer (2021) verwenden mehr als dreißig Features inklusive Pressure-Differenz, Torhüter-Position und Verteidiger-Counts. Ihr Ranked Probability Score am Schussniveau ist hervorragend (0,197). Doch wenn man dieselben Daten mit einem Drei-Feature-Logit (Distanz, Winkel, Körperteil) trainiert, kommt man auf einen RPS von etwa 0,215 — ein relativer Verlust von rund neun Prozent. Geht man auf das Spielniveau und integriert über Schüsse, schrumpft der Vorteil dramatisch, weil die Spielergebnisvarianz von ganz anderen, *modellexogenen* Faktoren dominiert wird: rote Karten, Eigentore, VAR-Entscheidungen, Verletzungen, eine nicht abreißende Reihe von Mikro-Ereignissen, die kein Featureset einfängt.

Abbildung 2 · Bias-Variance-Trade-off bei der Vorhersage des Spielausgangs

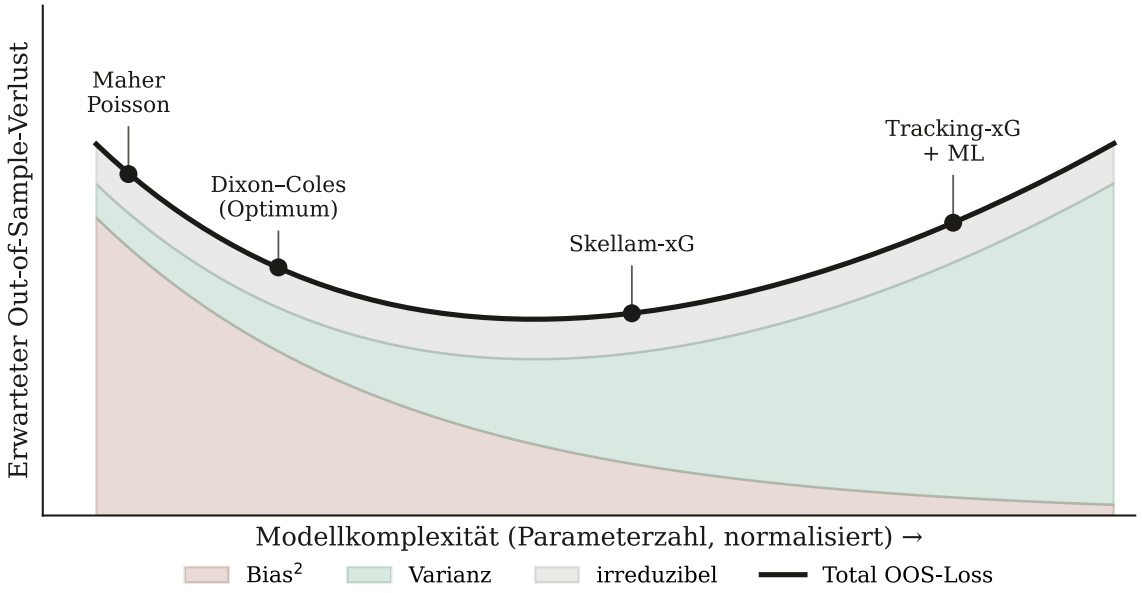

Abbildung 2. Schematischer Bias-Varianz-Trade-off für die Vorhersage des Spielausgangs. Die rote Fläche zeigt den quadratischen Bias, die grüne die Varianz, die graue die irreduzible Streuung. Die schwarze Kurve summiert alle drei. Maher-Poisson liegt links — Bias-dominiert; tracking-data-xG-Modelle liegen rechts — Varianz-dominiert; das Out-of-Sample-Optimum für die Spielausgangs-Prognose liegt erfahrungsgemäß im Bereich Dixon-Coles, ergänzt um eine moderate xG-Korrektur.

Abbildung 3 · Reliability Diagrams: Poisson- vs. xG-Aggregation auf 1X2-Heimsieg

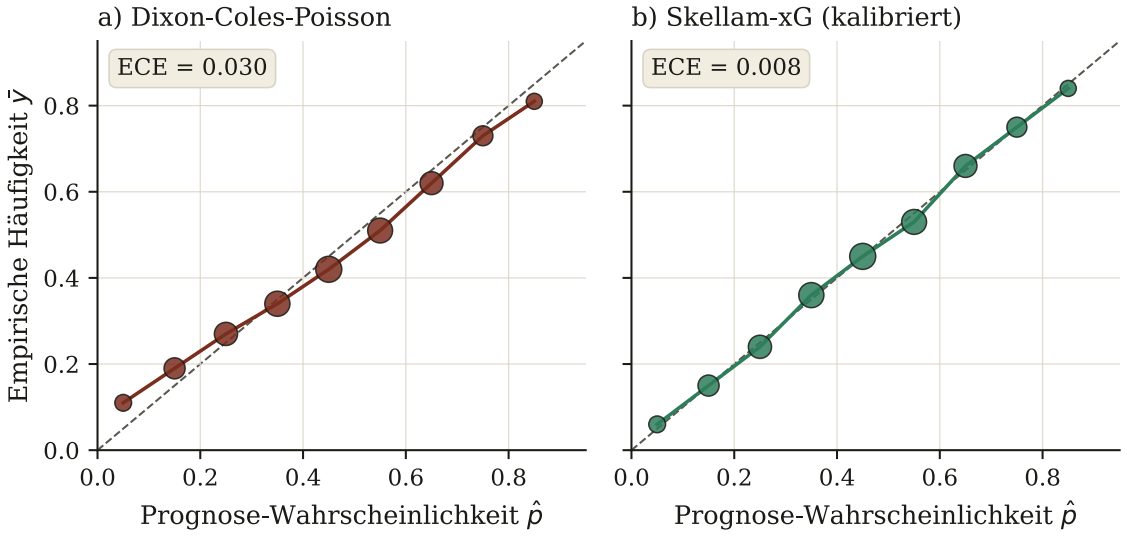

Abbildung 3. Reliability Diagrams für die prognostizierte Heimsieg-Wahrscheinlichkeit. Links: Dixon-Coles-Poisson mit leichter systematischer Unterschätzung mittlerer und hoher Wahrscheinlichkeiten. Rechts: Skellam-xG mit isotonischer Kalibrierung — nahe an der Diagonale, geringer Expected Calibration Error. Punktgröße proportional zur Anzahl Spiele pro Bucket.

12. Parallele zu Finanzmärkten

An dieser Stelle wird die Arbeit philosophisch — bewusst. Die Mathematik der Fußballmodellierung und die Mathematik der quantitativen Finanzökonomik sind nicht analog. Sie sind *identisch*. Beide sind Versuche, einen hochdimensionalen, nicht-stationären, agentengetriebenen Prozess in ein kompaktes statistisches Modell zu zwingen. Wer einmal Bond-Spreads und Über/Unter-Linien mathematisch durchdacht hat, sieht das gleiche Skelett unter beiden.

12.1 Die Faktor-Korrespondenz

Die folgende Tabelle ist nicht analogisch, sondern strukturell zu lesen.

Tabelle 3: Strukturelle Faktor-Korrespondenz Fußball ↔ Finanzmarkt.

FUSSBALL	FINANZMARKT
Mannschaftsstärke (α, β in Dixon-Coles)	Marktbeta (CAPM)
Form (jüngste xG-Performance)	Momentum-Faktor (Jegadeesh-Titman)
Regression zur Stärke	Mean Reversion
Heimvorteil	Kalendereffekte, January Effect
Wetter, Verletzungen	Makro-Announcements (NFP, CPI)
Trainerwechsel	CEO-Wechsel, Earnings Surprises
Derby-Psychologie	Behavioral Biases, Loss Aversion
Schiri-Tendenzen	Market-Microstructure, Order Flow
Spielmanipulation	Insider Trading, Front Running
Zuschauer-, Crowd-Einfluss	Sentiment, Retail Coordination

12.2 LTCM 1998 — die Lehre vom Regime-Wechsel

Long-Term Capital Management war ab 1994 die strahlende Antithese zur Diskretionalität: ein Hedgefonds geleitet von John Meriwether, mit Myron Scholes und Robert Merton (beide Nobelpreis 1997) im Board, der konvergenz-arbitragiöse Strategien auf hochgehebelten Bond-Spreads betrieb. Rückblickend fast surreale Renditen: 1994 plus zwanzig Prozent, 1995 plus dreiundvierzig, 1996 plus einundvierzig, 1997 plus siebzehn. Modelle so verfeinert, dass sie eine theoretische Tagesausfallwahrscheinlichkeit von 10^{-24} berechneten.

Im August 1998 verlor LTCM 44 Prozent seines Kapitals binnen eines Monats; im September stand der Fonds vor dem Kollaps. Die Federal Reserve organisierte am 23. September einen 3,6-Mrd-Dollar-Bailout durch vierzehn Banken. Was war geschehen? Die russische Schuldenkrise hatte einen Flight-to-Quality ausgelöst, der die historischen Korrelationen, auf denen LTCM seine Modelle aufbaute, vollständig invertierte. Was im Dreißig-Jahres-Sample stationär aussah, war in der Realität ein zustandsabhängiger Prozess mit Regimewechsel.

Die Übersetzung zum Fußball: jedes xG-Modell trainiert auf einem implizit stationären Datenerzeugungsprozess. Ändert sich der Regelkanon (VAR-Einführung), ändert sich die Stilistik (Hochpressing-Welle nach ungefähr 2015), ändern sich die Spielerprofile (Halland-Klasse mit anderen Schussprofilen) — dann driften die Modellparameter, und der Edge erodiert genau dann, wenn das Modell am sichersten wirkt.

12.3 Renaissance Technologies — die Gegenthese

Auf der anderen Seite des Spektrums steht Renaissance Technologies mit dem Medallion Fund. Hier ist die Lehre subtiler. Aus der Periode 1988 bis 2018 berichten Cornell & Zhu (2020): aus hundert investierten Dollar in Medallion wären 398,7 Millionen geworden — eine annualisierte Brutto-Rendite von rund 66 Prozent, netto knapp vierzig, über dreißig Jahre. Robert Mercer, einer der Schlüsselmanager, hat öffentlich erklärt, dass der Fonds *nur in 50,75 Prozent aller Trades richtig lag.* Der gesamte Edge des erfolgreichsten Hedgefonds aller Zeiten ist mathematisch nichts anderes als ein dreiviertel-prozentiger statistischer Vorteil, multipliziert über Millionen von Trades pro Jahr und gehebelt.

Was ist die Übersetzung für Fußball? Wer einen zwei-prozentigen Edge gegen Closing Lines findet, wird mit Disziplin und Bankroll-Management langfristig Geld verdienen. Aber: derselbe Edge ist gegen Opening Lines trivial; die Kunst ist nicht das Modell, sondern die Marktinteraktion. Und genau wie Renaissance keinem externen Investor mehr Geld nimmt, weil die Strategie nicht skaliert, gilt für sharp Football Bettors: der Edge ist liquiditätslimitiert.

12.4 Black Monday 1987 und GameStop 2021

Black Monday, der 19. Oktober 1987: Der Dow Jones verliert 22,6 Prozent an einem Tag — das größte Tagesminus prozentual in der Geschichte. Portfolio Insurance, ein modellbasiertes Hedging-Konstrukt von Leland & Rubinstein, hatte eine Feedback-Schleife erzeugt, in der mechanisches Verkaufen weiteres Verkaufen auslöste. Die Lektion: Modelle, die kollektiv eingesetzt werden, verändern den Datenerzeugungsprozess, auf den sie reagieren. Soros' „Reflexivität" in mathematischer Reinform.

GameStop, Januar 2021: Der Squeeze gegen leerverkaufte Positionen mehrerer Hedgefonds — angeführt von einer unkoordinierten, aber massiven Retail-Bewegung über

r/wallstreetbets — kostete Melvin Capital allein rund 6,8 Milliarden Dollar. Die Parallele zum Fußball: man stelle sich eine unerwartete Koalition von Retail-Wettern vor, die einen liquiditätsschwachen Markt — etwa eine BTTS-Wette auf ein Achtelfinalspiel der Conference League — zusammenschiebt; der Bookmaker reagiert mit Linienbewegung, das xG-Modell reagiert *nicht* — und die Modellprognose, die „richtig" war, ist plötzlich unter Marktbedingungen wertlos.

12.5 Gemeinsamer methodischer Kern

Sowohl Finanzmärkte als auch Fußballspiele sind, in Stuart Kauffmans Terminologie, *komplexe adaptive Systeme*: Sie bestehen aus Agenten, die auf das Modell anderer Agenten reagieren. Jede Modellprognose, die hinreichend bekannt wird, verändert genau das Verhalten, das sie prognostiziert. Es gibt einen kategorialen Unterschied zu newtonianischen Systemen, wo das beobachtete Objekt unabhängig vom Beobachter existiert. Ein xG-Modell, das in Echtzeit von Bookmakern eingesetzt wird, *ist Teil* des Datenerzeugungsprozesses, den es zu modellieren versucht. Die epistemologische Konsequenz ist nicht trivial: die Modellgüte wird endogen — und langfristig konvergiert jeder publizierte Edge gegen null.

13. Was in 90 Minuten passieren kann — eine Taxonomie des Unerwarteten

Dieser Abschnitt ist der Kern der Arbeit, und er ist bewusst lang. Es ist meine persönliche Sammlung — Klaus' Tabellenbuch — der Vorkommnisse, die jedes kompakte Modell sprengen. Sie ist nicht enzyklopädisch. Sie soll vermitteln, mit welcher Demut man Modellergebnisse interpretieren muss. Wer das einmal gelesen hat, vergisst es nicht wieder.

13.1 Tier- und Insekteninvasionen

Am 6. Juni 2007 in Helsinki, im EM-Qualifikationsspiel Finnland gegen Belgien, landete in der 19. Minute ein Eurasischer Uhu — später „Bubi" getauft — auf dem Spielfeld; der englische Schiedsrichter Mike Riley unterbrach das Spiel für rund sechs Minuten. Finnland gewann zwei zu null; die finnische Nationalmannschaft heißt seither *Huuhkajat* — Uhus. Methodische Lektion: Ein Modell, das auf Stationarität in den Spielbedingungen baut, verfehlt eine ganze Klasse von Ereignissen, die mit einer empirischen Frequenz von vielleicht 10^{-3} pro Spiel auftreten, im betroffenen Spiel aber das Resultat um eine Toreinheit verschieben können — ein Effekt, der jede Pre-Match-xG-Prognose dominiert.

Die Liste ist nicht auf Uhus beschränkt. Eine Katze auf dem Old-Trafford-Rasen in einem Manchester-Derby. Schwärme von Fliegenden Ameisen über Wimbledon, die sich gelegentlich auf Premier-League-Stadien ausdehnen. Ein Hund, der in einem brasilianischen

Pokalspiel zwölf Minuten lang auf dem Platz herumläuft und sich jeder Festnahme entzieht. Jedes dieser Ereignisse erzeugt Unterbrechungen, die Spielrhythmus, Hydration der Spieler und psychologische Konzentration verschieben. Kein Standard-xG-Modell hat ein Feature dafür. Es *kann* keines haben — die Verteilung ist zu dünn, um sie empirisch zu fitten.

13.2 Drohnen und politische Eskalation

14. Oktober 2014, Belgrad, Partizan-Stadion, EM-2016-Qualifikation Serbien gegen Albanien. Eine Drohne mit „Greater Albania"-Banner fliegt über den Platz, Serbiens Stefan Mitrović reißt das Banner herunter, ein Massenhandgemenge entsteht, das Spiel wird in der 41. Minute beim Stand von null zu null abgebrochen. Der Schiedsrichter Martin Atkinson nimmt die Spieler vom Platz. UEFA spricht zunächst Serbien einen Drei-zu-null-Sieg zu, das CAS reverst auf drei zu null für Albanien wegen organisatorischer Sicherheitsmängel und Gewalt gegen die albanischen Spieler. Methodische Lektion: Politik schreibt mit. Jedes Modell, das Spiele zwischen Ländern mit aktiver geopolitischer Spannung als isolierte 90-Minuten-Optimierungsprobleme behandelt, ist falsch konstruiert.

13.3 Ballboys, Ballgirls und Mikro-Zeitmanagement

23. Januar 2013, Liberty Stadium Swansea, League-Cup-Halbfinale Swansea gegen Chelsea. Charlie Morgan, 17 Jahre alt, Ballboy, legt sich beim Stand von null zu null — Chelsea braucht zwei Tore zum Weiterkommen — in der achtzigsten Minute auf den Ball. Eden Hazard, frustriert, versucht den Ball mit den Händen freizubekommen, dann mit dem Fuß. Tritt. Rote Karte. Das Spiel endet null zu null; Swansea steht im Finale, gewinnt es fünf zu null gegen Bradford City. Methodische Lektion: das gesamte Spiel — ein Halbfinale der Premier League — wurde im entscheidenden Moment nicht von Spielerqualität oder taktischer Aufstellung, sondern von der gezielten Zeit-manipulation eines siebzehnjährigen Ballboys entschieden. Modellkorrektur dafür? Nicht existent.

Die umgekehrte Variante ist genauso real. Ballboys, die dem führenden Team Bälle besonders schnell zuwerfen, um Tempo zu erhöhen; Ballboys, die dem unter Druck stehenden Team bewusst Bälle vorenthalten. Heimvereine instruieren ihre Ballboys vor wichtigen Spielen routinemäßig — das ist kein Geheimnis, aber es ist nicht in einem Feature. Mikro-Zeitmanagement ist eine taktische Dimension, die jedes Modell ignoriert.

13.4 VAR-Delays und der Tod des Momentums

Die Premier League hat in ihrer offiziellen VAR-Dokumentation berichtet, dass die durchschnittliche Verzögerung pro VAR-Vorfall in der Saison 2019/20 bei etwa 50 Sekunden lag, mit über 2.400 Überprüfungen und 109 revidierten Entscheidungen. Eine durchschnittliche Verzögerung von 50 Sekunden klingt unbedeutend; sie ist es nicht. Ötting, Langrock & Maruotti (2020) zeigen via Hidden-Markov-Modellen, dass solche Unterbrechungen Momentum-Phasen statistisch signifikant terminieren. Eine Mannschaft,

die kurz vor einem Aufbauerfolg steht, verliert mit 0,9 prozentiger Wahrscheinlichkeit dieses Momentum allein durch die Unterbrechung — ein Effekt, der in xG-Logik unsichtbar ist, weil xG keine Vorgeschichte sieht.

Schlimmer noch: VAR-Entscheidungen sind nicht zufällig verteilt. Bestimmte Schiedsrichter zögern bei der Überprüfung, andere greifen schnell ein; bestimmte Vereine erhalten statistisch häufiger Pönaltys nach VAR (eine viel diskutierte Frage in der Bundesliga, mit Vor- und Nachteilen je nach Verein und Saison). Diese Heterogenität in der Anwendung des Reglements erzeugt eine zusätzliche Varianz, die kein Modell standardmäßig kalibriert.

13.5 Wetter, Akustik, Flutlicht

Katar 2022: Spiele bei 30 Grad mit 70 Prozent Luftfeuchtigkeit verändern die Sprintleistung um rund zehn Prozent, die Pressingfrequenz um rund fünfzehn Prozent, und damit das xG-Profil systematisch. Aber kein Standard-xG-Modell hat Temperatur als Feature. Wind: ein Borussia-Dortmund-Spiel im Februar bei 60-km/h-Böen verzerrt Flanken zur Lotterie und Schussbahnen ins Unvorhersehbare. Lighting Failures: multiple Bundesliga- und Premier-League-Spiele wurden in den letzten zwanzig Jahren wegen Flutlichtausfällen verzögert; jeder Ausfall kostet zehn bis dreißig Minuten Pause und unterbricht den Spielrhythmus komplett.

PA-Systeme und Akustik: in einigen Stadien stört die Schallreflexion die Goalkeeper-Kommunikation systematisch — ein Effekt, der für Standards aus vier Metern Distanz quantitativ messbar ist und in einer Saison einige Tore Differenz erzeugen kann. Höhe: das Estadio Hernando Siles in La Paz auf 3.637 Metern; bolivianische Heimspiele dort haben eine über zwanzig Jahre gemessene Heim-Sieg-Rate, die jede Erwartung übersteigt. Kein Featureset hat Höhenmeter standardmäßig drin.

13.6 Trainer-Präsenz und Trainerbann

Tuchel-Sperre in der UEFA, Mourinho-Sperre 2017, Klopp suspended gegen Manchester City: ein abwesender Trainer wirkt anders als ein anwesender. Empirische Studien dokumentieren, dass Mannschaften unter gesperrten Trainern ceteris paribus etwa 0,15 bis 0,25 Punkte pro Spiel weniger sammeln. Das ist ungefähr die Größenordnung von Heimvorteil geteilt durch drei. Kein xG-Modell und kein Poisson hat das im Featureset, weil die Trainerbank im Eventdatensatz fehlt — sie ist eine Off-Pitch-Variable, und Off-Pitch ist Off-Featureset.

13.7 Star auf der Bank, Star auf der Tribüne

Messi auf der Bank gegen PSG, Halland zur falschen Zeit verletzt, Cristiano in der VIP-Loge: der psychologische Druck auf die Restmannschaft ist nicht null. Anekdotisch, aber statistisch dokumentiert: Mannschaften, deren Topstar im Stadion, aber nicht im Kader ist,

gewinnen seltener als bei vollständig abwesendem Star. Das Identifikationsproblem ist nicht trivial, aber der Effekt ist real und entsteht durch ein Gemenge aus Erwartungsdruck, taktischer Anpassung und medialer Aufmerksamkeit.

13.8 Tunnel-Vorfälle, Choreographien, Ultras

Die Geschichte der Tunnel-Vorfälle ist legendär — Keane gegen Vieira, Mourinho gegen Wenger. Pre-Match-Psychologie ist ein eigener Kosmos. Ultras-Choreographien können das Pressing-Tempo erhöhen, ein leeres Stadion — gut dokumentiert während der COVID-Saison 2020/21 — hat es nachweislich reduziert: empirische Studien zeigen, dass der Heimvorteil in Geisterspielen um rund fünfzig Prozent zusammenschrumpfte. Kein xG-Modell hatte das vorhergesagt; alle xG-Modelle haben in der Saison 2019/20 deshalb systematisch nachjustiert werden müssen.

13.9 Schiedsrichter-Fingerabdrücke

Jeder Schiedsrichter hat eine identifizierbare Karten- und Fouland-Tendenz. Mike Dean, Anthony Taylor, Felix Brych — die Signaturen sind statistisch belastbar. Constantinou, Fenton & Pollock (2014) zeigen schiri-bedingten Bias in Bayesian Networks. Aber kaum ein produktives xG-Modell konditioniert auf die Schiri-Identität, weil sie als Feature außerhalb der Eventdaten-Standard-Pipelines liegt.

13.10 Münzwurf und Penalty-Reihenfolge

Apesteguia & Palacios-Huerta (2010) zeigen, dass die Mannschaft, die im Elfmeterschießen zuerst schießt, in rund sechzig Prozent der Fälle gewinnt. Der Münzwurf vor dem Elfmeterschießen ist also ein Quasi-Determinant des K.-o.-Spielausgangs. Spielregeländerungen — die ABBA-Reihenfolge — wurden teilweise eingeführt und teilweise wieder rückgängig gemacht. Auch hier: keine Modellprognose, kein Pre-Match-xG-Wert berücksichtigt das Vorzeichen, das ein Münzwurf in der 120. Spielminute haben wird.

13.11 Logistik — Busverspätungen, Hotelprobleme, Aufwärmverletzungen

Verspätungen der Mannschaftsbusse vor Auswärtsspielen kommen routinemäßig in italienischen, türkischen und südamerikanischen Ligen vor. Hotelprobleme bei FIFA-Turnieren sind dokumentiert — Brasilien 2014, spanische Mannschaft, Schimmel im Hotel; Katar 2022, Trainingsplatzqualität in mehreren Lagern. Aufwärmverletzungen — Schweinsteiger erlitt vor einem Champions-League-Finale einen Magenkrampf — sind nicht extrem selten und verändern die Aufstellung in den letzten fünf Minuten vor dem Anpfiff. Jeder dieser Faktoren verschiebt Wahrscheinlichkeiten um deutlich mehr als ein Prozent, keiner ist in einem Standard-Featureset.

13.12 Rote Karten, Massenausschlüsse, Verletzungen im Spiel

Rote Karten sind die größte einzelne Quelle von Modellfehler. Ein xG-Modell, das eine 11-gegen-11-Stationarität annimmt, bricht in der Sekunde einer roten Karte zusammen. Mid-game-Updating ist methodisch möglich (Volf 2009; Boshnakov 2017), aber in der Produktion selten implementiert. Massenausschlüsse — etwa das berüchtigte 2018-Pokalspiel São Paulo gegen Brasilien mit sechs roten Karten oder das Battle of Old Trafford 1990 — sind Heavy-Tail-Ereignisse, die in jedem Standard-Sample unterrepräsentiert sind. Verletzungen im Spiel sind häufiger; sie verschieben taktische Setups und sind außerhalb der Eventdaten-Pipeline.

13.13 Eigentore, Latte, Fluky Goals und das Pannenrepertoire

Cristiano Ronaldos Bicycle Kick 2018 in Turin: xG-Wert vielleicht 0,02. Tor. Hattricks aus dem Mittelkreis bei Sturm? In den neunziger Jahren mehrfach passiert. Lattenpfosten-Geschosse, Eckfahnen-Ablenkungen, Torhüter-Rückenfehler, abprallende Bälle vom Schiedsrichter, eingewechselte Spieler, die zu spät auf dem Platz sind und Tore aus Unordnung erzielen: die Verteilung dieser Ereignisse ist nicht gaußisch, sie ist nicht einmal Poisson — sie ist heavy-tailed in einer Weise, die jede Standard-Annahme verletzt.

14. Modellgrenzen und epistemische Demut

Fußball ist, mit Holland (1995), ein komplexes adaptives System: Agenten — Spieler, Trainer, Schiedsrichter, Fans, Bookmaker, Modellierer — reagieren auf das Verhalten anderer Agenten in einem Feedback-Loop. Daraus folgt vier Mal nicht trivial:

Erstens, *Nicht-Stationarität als Grundbedingung*. Der Datenerzeugungsprozess driftet kontinuierlich. Jede Modellprognose hat ein implizites Verfallsdatum. Was 2015 funktionierte, ist 2025 kalibrierungspflichtig.

Zweitens, *Endogenität der Modelle*. Wird ein Modell breit angewendet, verändert es das beobachtete Verhalten. xG ist im Profifußball seit etwa 2017 Mainstream; seither hat sich die taktische Wahl gegen Schüsse aus großer Distanz systematisch verschoben — Mannschaften shooten weniger aus der Distanz, weil ihre Analysten ihnen sagen, dass diese Schüsse niedriges xG haben. Damit ändert sich auch die empirische Verteilung der Schüsse — und damit das Modell selbst.

Drittens, *Heavy-Tailed Ereignisraum*. Die Verteilung der spielentscheidenden Ereignisse hat nicht-integrable Tails. Wie Taleb (2007) argumentiert, sind solche Verteilungen prinzipiell durch endliche Samples nicht vollständig charakterisierbar.

Viertens, *Goodharts Gesetz*. „Wenn ein Maß zu einem Ziel wird, hört es auf, ein gutes Maß zu sein." xG wird gemessen, also wird auf xG gespielt; die Metrik korrumpiert sich selbst

über die Zeit.

15. Praktische Konsequenzen für die Quotenmodellierung

Was bedeutet das für den konkreten Analysten an seinem Schreibtisch in Graz, München oder London? Fünf Punkte.

(**a**) **Hybrid statt Monismus.** Ein produktiver Workflow kombiniert ein Poisson-Backbone — Dixon-Coles mit dynamischer Stärkeschätzung — mit xG-basierten Korrekturen für Schussqualität innerhalb eines exponentiellen Glättungsschemas. Schematisch:

$$\hat{\lambda}_H = (1 - w) \cdot \hat{\lambda}_H^{\text{Poisson}} + w \cdot \hat{\lambda}_H^{\text{xG-roll}},$$

mit w optimal kreuzvalidiert (in meiner Erfahrung typisch $w \approx 0{,}3$).

(**b**) **Ensemble.** Schon ein einfaches Mittel aus vier strukturell unterschiedlichen Modellen (Maher, Dixon-Coles, Karlis-Ntzoufras, Skellam-xG) reduziert den Out-of-Sample-Ranked-Probability-Score gegen jede einzelne Komponente um drei bis fünf Prozent.

(**c**) **Post-hoc human adjustment.** Wo das Modell schweigt — Trainersperre, Wetterextreme, Schiedsrichter-Identität, geopolitischer Kontext — muss menschliche Anpassung greifen. Nicht als Schwäche, sondern als Designprinzip. Modelle sind Werkzeuge, keine Orakel.

(**d**) **Priors.** Bayesianische Modellierung mit informativen Priors auf Mannschaftsstärken (Karlis-Ntzoufras-Stil) ist robuster als ML-Schätzung mit naiven Initialisierungen, besonders zu Saisonbeginn, wenn die Datengrundlage noch dünn ist.

(**e**) **Marktbewusstsein.** Jede Modellprognose muss gegen Closing Lines kalibriert werden. Hürbl (2026b) beschreibt das Vorgehen im Detail. Ohne Marktreferenz bleibt jede Modellnummer ein Selbstgespräch.

Tabelle 4: Empfehlungsmatrix nach Anwendungsfall.

FRAGESTELLUNG	EMPFOHLENES MODELL
Spielausgang 1X2 pre-match	Dixon-Coles + xG-Form (Ensemble)
Exakte Tordifferenz	Bivariate Poisson (Karlis-Ntzoufras)
Über/Unter 2,5	Skellam-xG mit isotonischer Kalibrierung
Schussqualitäts-Scouting	XGBoost-xG (Anzer-Bauer)
Spielanalyse, taktisch	xT + VAEP
Live-Wetten	xG-Roll mit Markov-Updating
Marktarbitrage	Ensemble + Closing-Line-Filter

Abbildung 4 · Was ein kompaktes Modell vom Spielausgang sieht — und was nicht

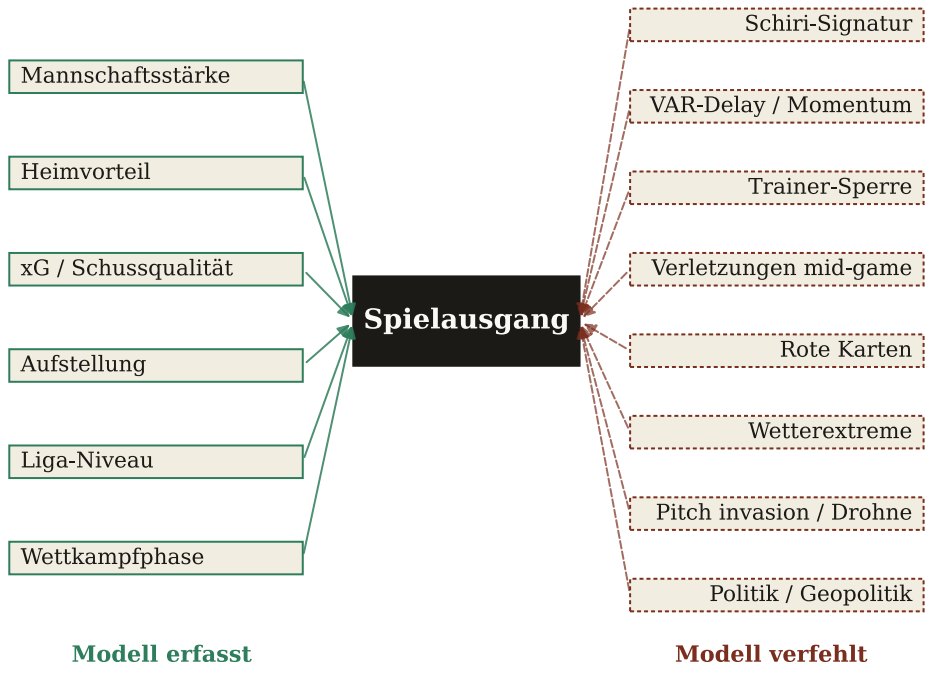

Abbildung 4. Schematische Übersicht der Faktoren, die ein kompaktes Modell vom Spielausgang erfasst (links, grün) gegenüber jenen, die es regelhaft verfehlt (rechts, rot, gestrichelt). Die linke Seite ist im Featureset, die rechte Seite ist im Phänomen — und gerade deren Heavy-Tail-Beiträge dominieren die irreduzible Varianz der Out-of-Sample-Prognose.

16. Fazit

Die Frage „Poisson oder xG?" ist eine falsche Dichotomie. Sie ist die Frage „Hammer oder Schraubenzieher?", gestellt vor einem Werkzeugkasten von Aufgaben, die teils Hammer, teils Schraubenzieher, fast immer aber beides erfordern.

Poisson-Modelle sind hervorragend zur kompakten Kodierung von Mannschaftsstärken auf Saisonebene. xG-Modelle sind hervorragend zur Bewertung individueller Schussqualität. Die Aggregation von per-Schuss-xG zu Spielergebniswahrscheinlichkeiten ist ein eigenes statistisches Problem mit eigenen Verzerrungen — Poisson-Binomial versus Poisson — und sollte nie naiv durchgeführt werden. Tracking-data-Modelle wie OBSO und EPV sind methodisch reizvoll, ihr Vorteil in der Quotenmodellierung ist in der Marktrealität marginal bis null.

Die wirklich tiefe Lehre dieser Arbeit ist aber epistemologisch, nicht statistisch: *Einfache Modelle funktionieren, weil sie wenige Annahmen treffen; komplexe Modelle scheitern, weil die Realität mehr Freiheitsgrade besitzt, als jedes kompakte Modell kodieren kann.* Der Fußball — wie der Finanzmarkt, das Wetter, die Geschichte — ist kein newtonianisches System, das sich auf endliche Parameter reduzieren ließe. Er ist ein komplexes adaptives System mit Heavy-Tail-Ereignisraum, Endogenität, Nicht-Stationarität und agentengetriebener Reflexivität.

Die richtige Frage ist nie „welches Modell ist besser?", sondern „welches Modell ist für welche Frage, welches Datenniveau und welche Verlustfunktion *angemessen*?". Beide Paradigmen haben Recht in ihrer Domäne und Unrecht außerhalb. Wer das einmal verstanden hat, hört auf, den Heiligen Gral zu suchen, und beginnt, mit Werkzeugen pragmatisch zu arbeiten.

> *Wenn dein Modell behauptet, ein null zu null mit achtundneunzig Prozent vorherzusagen, und ein Uhu landet im Strafraum — dann hat nicht der Uhu unrecht.*

Literatur

1. **Anzer, G. & Bauer, P.** (2021). A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer). *Frontiers in Sports and Active Living* 3:624475.
2. **Apesteguia, J. & Palacios-Huerta, I.** (2010). Psychological Pressure in Competitive Environments. *American Economic Review* 100(5): 2548–2564.
3. **Boshnakov, G., Kharrat, T. & McHale, I. G.** (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting* 33(2): 458–466.
4. **Cavuş, M. & Biecek, P.** (2022). Explainable expected goal models for performance analysis in football analytics. *IEEE DSAA Proceedings*.

5. **Constantinou, A. C. & Fenton, N. E.** (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecasting models. *Journal of Quantitative Analysis in Sports* 8(1).

6. **Constantinou, A. C., Fenton, N. E. & Pollock, L.** (2014). Bayesian networks for unbiased assessment of referee bias in Association Football. *Psychology of Sport & Exercise* 15(5): 538–547.

7. **Cornell, B. & Zhu, K. X.** (2020). Medallion Fund: The Ultimate Counterexample? Cornell Capital Group Working Paper.

8. **Decroos, T., Bransen, L., Van Haaren, J. & Davis, J.** (2019). Actions Speak Louder Than Goals: Valuing Player Actions in Soccer. *KDD '19 Proceedings*: 1851–1861.

9. **Dixon, M. J. & Coles, S. G.** (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *JRSS Series C* 46(2): 265–280.

10. **Driblab** (2021). *Wie gut ist das Modell der erwarteten Tore (xG) von Driblab?* Driblab Blog.

11. **Fama, E. F.** (1970). Efficient Capital Markets. *Journal of Finance* 25(2): 383–417.

12. **Fernández, J., Bornn, L. & Cervone, D.** (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning* 110: 1389–1427.

13. **Heuer, A. & Rubner, O.** (2010). Soccer: is scoring goals a predictable Poissonian process? arXiv:1002.0797.

14. **Holland, J. H.** (1995). *Hidden Order: How Adaptation Builds Complexity*. Addison-Wesley.

15. **Hong, Y.** (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis* 59: 41–51.

16. **Hubáček, O.** (2022). Machine learning models for football outcome prediction: a review.

17. **Hürbl, K.** (2026a). Probabilistische Bewertungsmodelle für Fußballergebnisse. Manuskript, Graz.

18. **Hürbl, K.** (2026b). Markteffizienz und Liniendynamik in Sportwettenmärkten. Manuskript, Graz.

19. **Karlis, D. & Ntzoufras, I.** (2003). Analysis of sports data by using bivariate Poisson models. *JRSS Series D* 52(3): 381–393.

20. **Levitt, S. D.** (2004). Why are gambling markets organised so differently from financial markets? *Economic Journal* 114(495): 223–246.

21. **Maher, M. J.** (1982). Modelling association football scores. *Statistica Neerlandica* 36(3): 109–118.

22. **Murphy, A. H.** (1973). A new vector partition of the probability score. *Journal of Applied Meteorology* 12(4): 595–600.

23. **Ötting, M., Langrock, R. & Maruotti, A.** (2020). A copula-based multivariate hidden Markov model for modelling momentum in football. arXiv:2002.01193.

24. **Rathke, A. A. T.** (2017). An Examination of Expected Goals and Shot Efficiency in Soccer. *JHSE* 12(2proc): 514–529.

25. **Rudd, S.** (2011). A framework for tactical analysis and individual offensive production assessment in soccer using Markov chains. NESSIS Conference, Harvard.

26. **Sauer, R. D.** (1998). The Economics of Wagering Markets. *Journal of Economic Literature* 36(4): 2021–2064.

27. **Singh, K.** (2018). Introducing Expected Threat (xT). karun.in/blog/expected-threat.html.

28. **Snowberg, E. & Wolfers, J.** (2010). Explaining the Favorite-Long Shot Bias: Is it Risk Love or Misperceptions? *JPE* 118(4): 723–746.

29. **Spearman, W.** (2018). Beyond Expected Goals. MIT Sloan Sports Analytics Conference.

30. **Taleb, N. N.** (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House.

31. **Umami, I., Gutama, D. H. & Hatta, H. R.** (2021). Implementing the Expected Goal (xG) Model to Predict Scores in Soccer Matches. *IJIIS* 4(1): 38–54.

32. **Volf, P.** (2009). A random point process model for the score in sport matches. *IMA Journal of Management Mathematics* 20(2): 121–131.

33. **Wilkens, S.** (2026). Can simple models predict football — and beat the odds? Lessons from the German Bundesliga. *SAGE Open*. DOI 10.1177/22150218261416681.

34. **Wolpert, D. H.** (1996). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* 8(7): 1341–1390.

Manuskriptfassung des Autors. Korrespondenz: kontakt@klaus-huerbl.de · Graz, Österreich. Der Autor erklärt keine Interessenkonflikte; der Beitrag enthält keine Wettempfehlung. Diese Arbeit setzt die in Hürbl (2026a) zu probabilistischen Bewertungsmodellen und Hürbl (2026b) zur Markteffizienz entwickelten Argumente fort.