Verifying probability forecasts of categorical outcomes: example of WDL football results in the UEFA Champions League

Jean-Louis Foulley

Introduction

Forecasts made in a wide range of disciplines (**Kahneman D, 2011-Thinking Fast and Slow, Ch 18**)

- Weather and Climate
- Economic and Financial
- Medicine, Diagnostic tests, Epidemics
- International Relations including Military Operations
- Media and Entertainment Market
- Sporting Events…

Dates back to Finley (1884) on whether or not a tornado (Murphy, 1996)

General framework

- Forecast of football matches outcomes
  - ✓Results : Response in Win Draw Loss (WDL or [1],[X],[2]) categories: categorical data
  - ✓Scorelines : {Y(A), Y(B)} goals in match (A vs B): pairs of integers
- Forecast WDL from WDL or SL data,

  SL from SL, WDL and/or other covariates
- Point (Tipsters) vs Probability (RED) Forecast
- See **Review by Reade, Singleton & Brown (2021)**

Objectives of this study

- **Evaluation of probability forecasts of WDL**

  - ✓ For UEFA Champions League (C1)
  - ✓ For matches played during the last 4 seasons: 2017, 2018, 2019, 2020
  - ✓ With probabilities based each season on data from the 3 previous ones: 2017, 2018 & 2019 for 2020
- **Group stage**
  - ✓ 32 teams in 8 independent groups (A, B,…, H) of 4 teams each
  - ✓ Playing 8x12=96 matches from Sept to December

Model/Poisson Regression

The model chosen is a **Poisson Loglinear model** which can be written for matches $m(i,j)\in\mathcal{M}$ between Home team i and Away team j with score-line $\left\{y^{(t)}_{m(ij),1};y^{(t)}_{m(ij),2}\right\}$ at time (t) as:

$$\boxed{y^{(t)}_{m(ij),k} \mid \lambda^{(t)}_{m(ij),k} \sim \mathcal{P}\left(\lambda^{(t)}_{m(ij),k}\right)} \text{ for } k=1 \text{ (home team)}, 2 \text{ (away team)}$$

$$\boxed{\log\lambda^{(t)}_{m(ij),1}=\eta+h+\beta_1\Delta r^{(t')}_{ij}+\beta_2\overline{r}^{(t')}_{ij}}$$

$$\log\lambda^{(t)}_{m(ij),2}=\eta+\beta_1\Delta r^{(t')}_{ji}+\beta_2\overline{r}^{(t')}_{ij}$$

$$\Delta r_{ij}=r_i-r_j;\ \Delta r_{ji}=r_j-r_i\quad \overline{r}_{ij}=½\left(r_i+r_j\right)$$

* η is an intercept, h is the home effect, $r^{(t')}_i$ the ELO rating of team i at time $t'<t$

* Δr_{ij} is the difference in rates between the attacking team i and defending team j,

* $\overline{r}_{ij}$ represents their mean level

Bayesian inference with independent prior distributions are set up on the parameters $\boldsymbol{\theta}$:

$\eta\sim\mathcal{N}\left(\eta_0,\sigma^2_\eta\right)$ $h\sim\left(h_0,\sigma^2_h\right)$ and $\beta_k\sim_{ind}\mathcal{N}\left(b_k,\sigma^2_{\beta_k}\right)$ for $k=1,2$.

Model/Poisson Regression and Bayesian Forecasting

Knowing the posterior distributions of parameters $\boldsymbol{\theta} = \left(\eta, h, \beta_1, \beta_2\right)'$, we can reconstruct the forecasting probabilities of elementary score lines of the future matches:

$$P^f_{m(i,j)}\left(u,v\right) = \Pr\left(Y^f_{m(i,j),1} = u; Y^f_{m(i,j),2} = v \mid \mathbf{y}\right)$$

as **the mean** of the **posterior distribution** of the probability $\Pr\left(Y^f_{m(i,j),1} = u; Y^f_{m(i,j),2} = v\right)$ taken as a product of the marginal ones due the assumption of conditional independence:

$$\boxed{P^f_{m(i,j)}\left(u,v\right) = \mathrm{E}\left(\lambda_{m(i,j),1}\lambda_{m(i,j),2}\left[-(\lambda_{m(i,j),1} + \lambda_{m(i,j),2})\right] / u!v! \mid \mathbf{y}\right)}$$

y=ex ante scorelines (here those of 3 previous seasons) +ELO ratings prior day of play

Probability Scoring Rules

"Predicting is **easy**. Predicting **accurately** is the **hard** bit" Spiegelhalter & Ng, 2009 (One match to go PL, 2009)

Superiority of probability forecasts over categorical ones even economically (Savage, 1971; Winkler & Murphy, 1979)

PSR quantify the quality of a forecast distribution P of a forthcoming, uncertain, event X given

a) -quoted values p of P (ex ante), and b) realized values x of X (ex post)

via a loss function (i.e. penalty) equal to $\boxed{S(p,x)}$

Choose P so as to minimize $\boxed{E_{X\sim q}\left[S(p,X)\right]=S(p,q)}$ e.g $qS(p,1)+(1-q)S(p,0)$

PSR/Brier's score

Brier's (BRS) (Brier, 1950):

$$\boxed{BRS(\mathbf{p}, j) = \sum_{k=1}^{3} \left(p_k - o_{jk}\right)^2 = \left\|\mathbf{p} - \mathbf{o}_j\right\|^2 = \left(\sum_{k=1}^{3} p_k^2\right) - 2p_j + 1}$$

$\mathbf{p} = \left(p_1, p_2, p_3\right)$ stands for the vector of WDL forecasted probabilities

$\mathbf{o}_j = \left(o_{j1}, o_{j2}, o_{j3}\right)$ with $o_{jk} = I\left[j = k\right]$ Kronecker delta, j observed result

Ex $j = 1 \Rightarrow \mathbf{o}_1 = (1,0,0)$, $j = 2 \Rightarrow \mathbf{o}_2 = (0,1,0)$, $j = 3 \Rightarrow \mathbf{o}_3 = (0,0,1)$

BRS depends on probability forecasts of all categories: **non local**

PSR/Brier's score for binary events

Half Brier Score defined for one event A with probability p

$HBS(p, j) = (p - o_j)^2$ .with $o_j = I[j=1]$

$BRS(\mathbf{p}, j) = (p_1 - o_{j1})^2 + (p_2 - o_{j2})^2 + (p_3 - o_{j3})^2$

$$\boxed{BRS(\mathbf{p}, j) = \sum_{k=1}^{3} HBS(p_k, o_{jk}) = HBS_W + HBS_D + HBS_L}$$

Notice also that : $HBS(p, j) = j.HBS(p,1) + (1-j)HBS(p,0)$

sum of HBS's for Win, Draw and Loss separately

PSR/Ranked Probability Score (RPS)

$\boxed{RPS(\mathbf{p}, j) = (½)\sum_{k=1}^{2}\left(p_k^* - o_{jk}^*\right)^2}$ Epstein (1969); Constantinou & Fenton (2012)

$\mathbf{p}^* = \left(p_1, p_1 + p_2, \underbrace{p_1 + p_2 + p_3}_{1}\right)$ stands for the **cumulative forecasted probabilities**

Ex: If $j=1 \Rightarrow \mathbf{o}_1 = (1,0,0)$, if $j=2 \Rightarrow \mathbf{o}_2 = (0,1,0)$, if $j=3 \Rightarrow \mathbf{o}_3 = (0,0,1)$

Then $\mathbf{o}_j^* = \left(o_{j1}^*, o_{j2}^*, o_{j3}^*\right)$, $o_{jk}^* = I[j \le k]$ $\mathbf{o}_1^* = (1,1,1)$, $\mathbf{o}_2^* = (0,1,1)$ $\mathbf{o}_3^* = (0,0,1)$,

$$\boxed{RPS(\mathbf{p}, j) = (½)\left[\underbrace{HBS(p_1, o_{j1})}_{HBS(WIN)} + \underbrace{HBS(p_3, o_{j3})}_{HBS(LOSS)}\right]}$$

$o_{j1} = I[j=1]$ and $o_{j3} = I[j=3]$

PSR/ Negative Log Score (NLS)

Negative Logarithm score (NLS) or Ignorance score:

$\boxed{NLS(\mathbf{p}, j) = -\log p_j}$,

-penalizes the observed event j by minus its log probability p_j

-positive value, negatively oriented (the smaller, the better)

-equal to **half the deviance** (D=–2Loglikelihood)

-depends only of probability in category observed j: **Local Score**

-**Strictly proper**

PSR/Example

PSG vs MNU R16 March 6, 2019, Scoreline 1-3

	1	**X**	**2**
RED	0.450	0.270	0.280
JLF	0.551	0.233	0.216
ODP	0.649	0.207	0.144
OBS	**0**	**0**	**1**

Ex: RED

HBS(W)=0.450^2=0.2025 HBS(D)=0.270^2=0.0790 HBS(L)=(1-0.280)^2=0.5184

BRS=0.2025+0.0790+0.5184=0.7999

RPS=0.5(0.2025+0.51840)=0.3604 RPS=0.5(0.7999-0.0790)=0.3604

NLS=-log(280)=1.273

PSR/ Properness

If observed result j is sampled, then $S(\mathbf{p},j)$ **has a distribution** with **expectation**

$\boxed{S(\mathbf{p},\mathbf{q}) = E_{\mathbf{X}\sim\mathbf{q}}\left[S(\mathbf{p},X=j)\right] = \sum_{j=1}^{3} q_j S(\mathbf{p},j)}$ called "Score Function"

where $\mathbf{q} = (q_j)_{1\le j\le 3}$ represents the true distribution of the outcomes

$$\boxed{D(\mathbf{p},\mathbf{q}) = S(\mathbf{p},\mathbf{q}) - S(\mathbf{q},\mathbf{q}) = E_X\left[S(\mathbf{p},X=j) - S(\mathbf{q},X=j)\right]}$$

Proper if $D(\mathbf{p},\mathbf{q}) \ge 0$ "divergence": Negatively Oriented Score (The smaller is better)

Strictly proper if $\boxed{D(\mathbf{p},\mathbf{q}) = S(\mathbf{p},\mathbf{q}) - S(\mathbf{q},\mathbf{q}) \text{ being } 0 \text{ iff } \mathbf{p}=\mathbf{q}}$ (BRS, RPS, NLS)

Proper SR: $S(\mathbf{p},\mathbf{q})$ minimized when forecast distribution $\mathbf{p}$ is the true distribution of the outcome $\mathbf{q}$

PSR/ Properness

Example: Half Brier Score

$$HBS\left(p,q\right)=HBS\left(p,j=1\right)q+BS\left(p,j=0\right)(1-q)$$

$$HBS\left(p,q\right)=(p-1)^2q+p^2\left(1-q\right)=p^2-2pq-q$$

$\boxed{D_{HBS}=(p-q)^2}$ "Epistemic Loss" (All models are wrong, but some are useful, G Box)

$\boxed{HBS(q,q)=q(1-q)}$ "Uncertainty" or "Irreducible Loss": Kull & Flach, 2015

PSR/ Properness of BRS & NLS

BRS: $\boxed{S(\mathbf{q},\mathbf{q}) = 1 - \sum_{k=1}^{3} q_k^2}$ $\boxed{D(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{3} (p_k - q_k)^2 \geq 0}$ OK to be proper

NLS:

$$\boxed{S(\mathbf{q},\mathbf{q}) = \underbrace{-\sum_{k=1}^{3} q_k \ln(q_k)}_{Entropy} = H(\mathbf{q})}$$

$$\boxed{D(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{3} q_k \log \frac{q_k}{p_k} = \underbrace{KL(\mathbf{q}\|\mathbf{p})}_{KullBack-Leibler} \geq 0}$$

NLS, the only local PSR which is strictly proper ZEO proper but not strictly

Two Interpretations of properness

1) Encourage honesty in reporting their probability forecasts (p) (not cheating about p when prior belief is q)
2) Facing a penalty when saying p if the true is q ; with zero penalty when p=q

PSR/Some Improper SR

Linear Score $\boxed{LIN(\mathbf{p}, j) = p_j = \sum_{k=1}^{K} \delta_{jk} p_k}$:

Reward forecast with probability p_j of the observed event j. Makes good sense

but actually not due to overstating probabilities to 0 or 1 if you know the rules.

Alternative: Spherical node $SN(\mathbf{p}, j) = p_j / \|\mathbf{p}\|_2$

Power Loss $\boxed{PLS(\mathbf{p}, j) = \sum_{k=1}^{K} |p_k - \delta_{ik}|^r}$ $\delta_{ik} = I(j = k);\ r > 0$

Improper for $r \neq 2$, especially for $r = 1$ (absolute loss)

$$ALS(\mathbf{q},\mathbf{q}) = 2\sum_{k=1}^{K} q_k (1 - q_k) \quad D_{ALS}(\mathbf{p},\mathbf{q}) = 2\sum_{k=1}^{K} q_k (q_k - p_k)$$

PSR/Properties

- Orientation:
  - ✓ Penalty: Negatively (BRS, RPS, NLS)
  - ✓ Reward: Positively (LIN, ZEO)
- Locality
  - ✓ Local depends only of what is observed (NLS, ZEO) consistent with likelihood principle (only observed values are relevant in inference)
- Sensitive to Distance
  - ✓ favors adjacent categories eg RPS (Constantinou & Fenton, 2012: argument for soccer PSR)
- Properness
  - ✓ Invariant property by affine transformation
  - ✓ Improper (LIN, ALS)
  - ✓ Proper (BRS, RPS, NLS, ZEO)
  - ✓ Strictly proper (BRS, RPS, NLS)

PSR/ Score function

If observed result j resorts from sampling, then $S(\mathbf{p}, j)$ **has a distribution** with **expectation**

$$\boxed{S(\mathbf{p},\mathbf{q}) = E_X\left[S(\mathbf{p}, X = j)\right] = \sum_{j=1}^{3} q_j S(\mathbf{p}, j)}$$

where $\mathbf{q} = \left(q_j\right)_{1\leq j\leq 3}$ and $q_j = \Pr\left(X = j\right)$ represent the true distribution of the outcome

$S(\mathbf{p},\mathbf{q})$ **estimated by the empirical score** on a sample of matches $m = 1\ldots, M$ with

probability forecasts $\mathbf{p}_m = \left(p_{m,1}, p_{m,2}, p_{m,3}\right)$ and observed result x_m

$$\boxed{\overline{S} = M^{-1}\left[\sum_{m=1}^{M} S(\mathbf{p}_m, x_m)\right]}.$$

PSR: Overall results for C1 Group stage

Probability scores and their skill forms pertaining to POR probability forecasts of C1 match outcomes (4 group-stage seasons, 2017 to 2020): option Bayes plug-in

Seasons	Focus on	BRS	RPS	NLS	ZEO
17-20	POR	0.5397	0.1774	0.9148	0.5781
	BOD	0.5120	0.1650	0.8722	0.6042
	HOM	0.6540	0.2261	1.0792	0.4352
	POR vs *HOM*	0.1748	0.2153	0.1644	0.1429
	BOD vs *HOM*	0.2171	0.2702	0.2070	0.1690
	POR vs BOD	*-0.0541*	*-0.0752*	*-0.0426*	*-0.0261*

POR: POisson Regression
BOD: Betting Odds as Three Way Odds implied Probabilities : mean of 10 to 12 bookmakers odds
HOM: Home Effect Implied Probability (constant over matches) eg (0.48, 0.20, 0.32) for WDL respectively

BRS: Brier score; RPS: Ranked Probability score. NLS: Negative Log score; ZEO: Zero-One score
Skill forms: BRS*=1-BRS(F)/BRS(Ref); RPS*=1-RPS(F)/RPS(Ref)
NLS*=NLS(Ref)-NLS(F) according to Tippett et al (2017); ZEO*=ZEO(F)-ZEO(Ref)

PSR: Results for Poisson vs Davidson (WDL)

Table : Probability scores and their skill forms pertaining to probability forecasts of C match outcomes via Poisson regression (POR) and Davidson's (DAV) models

Season	Focus on	BRS	RPS	NLS	ZEO
POISSON	MOD	**0.5406**	**0.1779**	**0.9164**	**0.5836**
	HOM	0.6550	0.2336	1.0803	0.4349
	Skill	0.1747	0.2384	0.1639	0.1487
DAVIDSON	MOD	**0.5404**	**0.1776**	**0.9164**	**0.5830**
	HOM	0.6564	0.2312	1.0803	0.4323
	Skill	0.1767	0.2318	0.1639	0.1507

Same legend as in table 1

Uncertainty of PSR/ Expected vs Plug-in PSR

How to cope with uncertainty in estimating forecasting probabilities in the value of PSR?

The **average of the posterior discrepancy distribution** is **a better summary** than the **discrepancy of the point estimate (plug-in)**: Gelman et al (2004), Plummer (2008)

For **NLS**, $S(\mathbf{p}_\theta, j) = -\log p_j$, the expected S_{EX} is $\boxed{E_{post}\left[S(\mathbf{p}_\theta, j)\right] = -E_{post}\left[\log p_{\theta,j}\right]}$

vs the plug-in version S_{PI} $\boxed{S(\overline{\mathbf{p}}_\theta, j) = -\log E_{post}\left(p_{\theta,j}\right)}$.

As the log is a **concave function**, Jensen's inequality implies $\boxed{S_{EX} \geq S_{PI}}$

Uncertainty in forecasting probabilities **penalizes their measure of efficiency**.

A-Overall

Seasons	Focus on	BRS	RPS	NLS
Plug-In	POR	0.5397	0.1774	0.9148
	HOM	0.6540	0.2261	1.0792
Expected	POR	*0.5406*	*0.1779*	*0.9164*
	HOM	*0.6550*	*0.2336*	*1.0803*

Bias : $\overline{B}_{EX\ vs\ PI} = M^{-1}\sum_{m=1}^{M} B\left(\mathbf{p}_{\theta,m}, x_m\right) = \boxed{\mathrm{O}(1/\mathrm{n}_1)}$ $\quad n_1$ = size of training sample

Distribution oriented (DO) verification

Distribution-Oriented (DO) Approach: Murphy & Winkler (1987), Murphy (1997)

Based on joint distribution $\left[P, X\right]$ of Forecast P and Outcome X factored in 2 ways

1) $\boxed{\left[P, X\right] = \underbrace{\left[P\right]}_{REF} \underbrace{\left[X \mid P\right]}_{CAL}}$: Calibration Refinement (CR)

2) $\boxed{\left[P, X\right] = \underbrace{\left[X\right]}_{UNC} \underbrace{\left[P \mid X\right]}_{LIK}}$: Likelihood Base rate (LB)

"Likelihood" that a forecast would have been issued from a given outcome, reversed logic as compared to 1)

Decomposition of BS/ Calibration-Refinement factorisation

Let **X** be the binary outcome of the event H, D or A with probability **q** ;

P the random variable probabilistic forecast of X taking values **p.**

Taking the **Half-Brier Score** defined as a loss function as $S(P,X)=(P-X)^2$,

$$\mathbb{E}[S(P,X)] = \underbrace{\mathrm{Var}(X)}_{UNC} - \underbrace{\mathbb{E}_P\left\{\left[\mathbb{E}_X(X \mid P) - \mathbb{E}_X(X)\right]^2\right\}}_{RES} + \underbrace{\mathbb{E}_P\left\{\left[\mathbb{E}_X(X \mid P) - P\right]^2\right\}}_{REL}$$

1) **Uncertainty (**UNC) equal to the variance of the outcome that is out of control of the forecaster,
2) **Resolution** (RES) referring to the variability between the conditional expectations of the observed outcomes given their forecasts $\mathbb{E}_X(X \mid P)$,
3) **Reliability** (REL) or Calibration (CAL) measuring how close the outcomes for a given forecast are from their forecasts

The Murphy estimation of UNC, RES, REL

If the **forecasts take a few K distinct values** $\{p_k, k=1,\dots,K\}$ with n_k occurrences of binary outcomes X, then one can just use sample means

$$\hat{\mathbb{E}}(X \mid P = p_k) = \bar{X}_k = X_{k+}/n_k;\ X_{k+} = \sum_{i=1}^{N} I(p_i = p_k) X_i;\ \bar{X} = \left(\sum_{i=1}^{N} X_i\right)/N.$$

The Murphy (1973) decomposition is fully applicable without restrictions:

$$\boxed{REL = N^{-1}\sum_{k=1}^{K} n_k\left(\bar{x}_k - p_k\right)^2},\ \boxed{RES = N^{-1}\sum_{k=1}^{K} n_k\left(\bar{x}_k - \bar{x}\right)^2},\ \boxed{UNC = \bar{x}\left(1-\bar{x}\right)}.$$

Reliability: Binning & Counting

In fact, facing too **many** distinct forecast values.

The **forecasts** have to be distributed into **intervals named bins** $B_1,..B_d,..,B_D$

Forecasts and outcomes are **averaged within bins**

$\boxed{\text{INT}}$: Intervals; $\boxed{\text{QUA}}$: Quantiles: $\boxed{\text{ISO}}$-Regression

Choice of D, n_D : LOO (Broecker, 2012) Type 1& 2 E (Gweon et al. 2019)

Arbitrariness in defining intervals and quantiles

Reliability/ Binning/IsoRegression

Bins automatically determined by the pool-adjacent-violators (**PAV**) algorithm applied to

Nonparametric isotonic regression for estimating

the conditional $q_p = \Pr\left(X = 1 | P = p\right)$ probabilities

by minimizing the regression MSE with respect to D:

$$\boxed{\mathrm{MSE}_{ISO} = \sum_{d=1}^{D} \sum_{i=1}^{N} I\left(p_i \in \left[b_d, b_{d+1}\right]\right)\left(q_d - p_i\right)^2}$$

under the constraints of **isotonicity** (q_d estimation is a non-decreasing function of the original p_i's).

see Dimitriadis, Gneiting & Jordan (2021

CR decomposition/current expression

Alternative decomposition to avoid inconsistencies in the Murphy decomposition, use of **3 score functions** pertaining to **3 types of forecast**

1) $\boxed{\mathbb{E}[S(P,X)]}$ 2) $\boxed{\mathbb{E}[S(Q_P,X)]}$ 3) $\boxed{\mathbb{E}[S(\pi,X)]}$

1) $\underbrace{P}_{\text{Issued}}$ 2) $\underbrace{Q_P = E(X \mid P = p)}_{\text{"Calibrated"}}$ 3) $\underbrace{\pi = E(X)}_{\text{"Climatological"}}$

$$\boxed{\mathbb{E}[S(P,X)]} = \underbrace{\mathbb{E}[S(P,X) - S(Q_P,X)]}_{REL} - \underbrace{\mathbb{E}[S(Q_P,X) - S(\pi,X)]}_{RES} + \underbrace{\mathbb{E}[S(\pi,X)]}_{UNC}$$

CR factorization/expression via divergence functions

Ensures Right HS=Left HS

Ensures **non negativity** of REL, RES for Proper SR

$\boxed{\mathbb{E}_{X|P=p}[S(p,X)-S(q_P,X)]=D(p,q_P)\geq 0}$; $\boxed{\mathbb{E}_{X|P}[S(\pi,X)-S(q_P,X)]=D(\pi,q_P)\geq 0}$

Applicable to any proper scoring rule (Dawid, 186; Broecker, 2012)

$2N\left[\bar{L}(\mathbf{p})-\bar{L}(\hat{q}_p)\right]$ is the loglikelihood ratio statistic contrasting

i) The original forecast procedure (ex ante)

ii) The (re)calibration procedure or model (ex post)

In addition, it has an asymptotic Chi-square distribution with #DF equal to the # of parameters specifying this model.

URR decompostion for Hwin, Draw, Awin

Calibration-Refinement Factorization of Brier's score pertaining to HomeWin (a), Draw (b) under two forecasting procedures: Poisson regression model (POI) and Odds (ODD)

a-HWIN	BRS	SKI(%)	B-TEST	MET	REL	RES
POI	0.1849	24.8	1.035 [0.309]	INT	0.0035 (1.4)	0.0644 (26.2)
				QUA	0.0030 (1.2)	0.0639 (26.0)
				ISO	**0.0116 (4.7)**	**0.0725 (29.5)**
ODD	0.1732	29.5	2.099 [0.147]	INT	0.0041 (1.7)	0.0766 (31.2)
				QUA	0.0048 (2.0)	0.0774 (31.5)
				ISO	**0.0122 (4.9)**	**0.0847 (34.5)**

UNC=0.2458

b-DRAW	BRS	SKI(%)	B-TEST	MET	REL	RES
POI	0.1849	1.4	3.995 [0.045]	INT	0.0010 (0.5)	0.0036 (1.9)
				QUA	0.0031 (1.7)	0.0058 (3.1)
				ISO	**0.0099 (5.3)**	**0.0125 (6.7)**
ODD	0.1820	**3.0**	2.102 [0.147]	INT	0.0011 (0.6)	0.0066 (3.5)
				QUA	0.0005 (0.3)	0.0060 (3.2)
				ISO	**0.0048 (2.6)**	**0.0092 (4.9)**

UNC=0.1875

Skill (SKI) defined as SKI=(BRSref -BRS)/ BRSref where BRSref=UNC so that SKI=(RES-REL)/UNC

B-TEST: Brier-Score Test for departure of its expectation from that induced by the null hypothesis of perfect forecast calibration expressed with its corresponding statistic and P-value within brackets

URR decomposition via Log Loss

For the logloss score $L(P,X) = -[X\log(P) + (1-X)\log(1-P)]$.

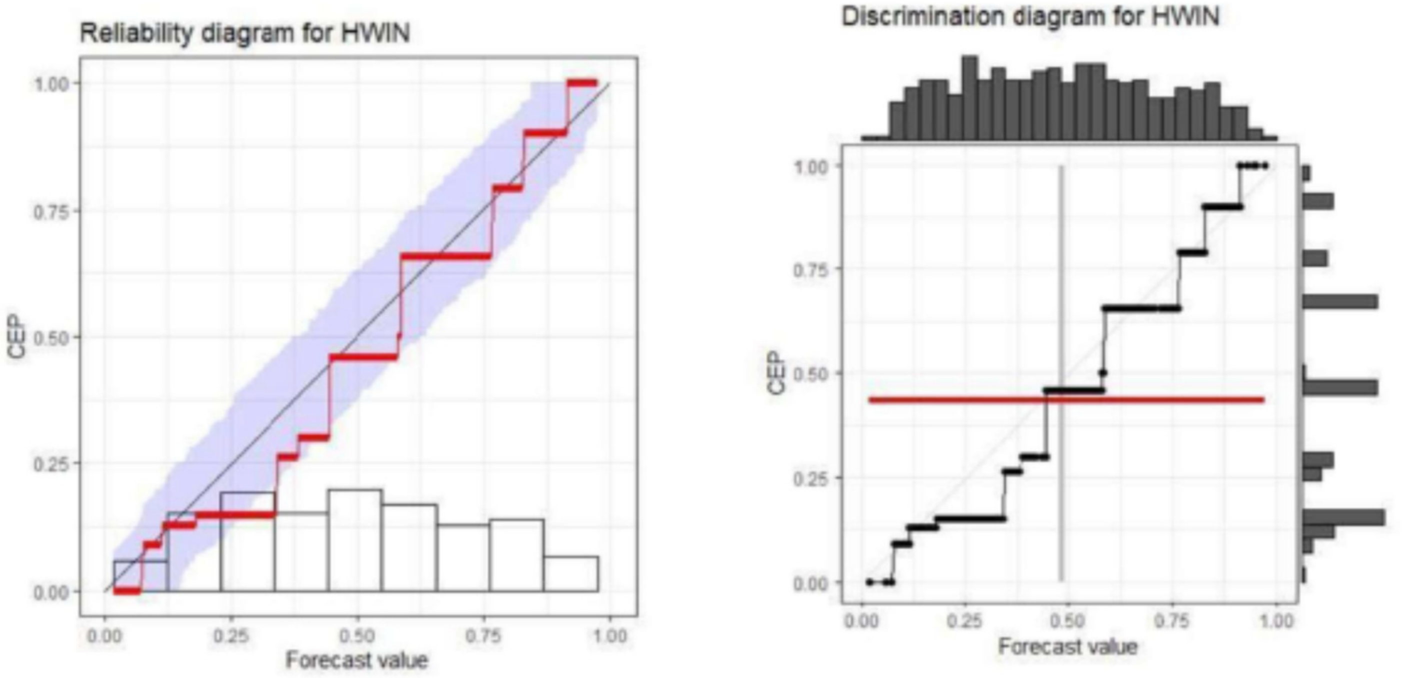
$REL = \bar{L}(\mathbf{p}) - \bar{L}(\hat{q}_p)$ $RES = \bar{L}(\bar{\pi}) - \bar{L}(\hat{q}_p)$ and the statistic $2N \text{ x } REL \rightarrow \chi^2_D$

Table: Calibration-Refinement Factorization of **Log Loss score** (LLS) pertaining to HomeWin, Draw and AwayWin under a Poisson regression model (POI).

POI	LLS	SKI(%)	TEST-REL	REL	RES	UNC
HWIN	0.5509	19.5	10.53 [0.23]	0.0137 (2.0)	0.1475 (21.5)	0.6846 (100)
DRAW	0.5560	**1.1**	6.72 [0.24]	0.0090 (1.6)	0.0151 (2.7)	0.5623 (100)
AWIN	0.5071	18.6	6.77 [0.56]	0.0088 (1.4)	0.1248 (20.0)	0.6231 (100)

TEST-REL(LL) =2NxREL has a asymptotic Chi-square distribution with DF= No bins (here 8,5,8) and its corresponding P-value ; SKI=1-(BRS/UNC)

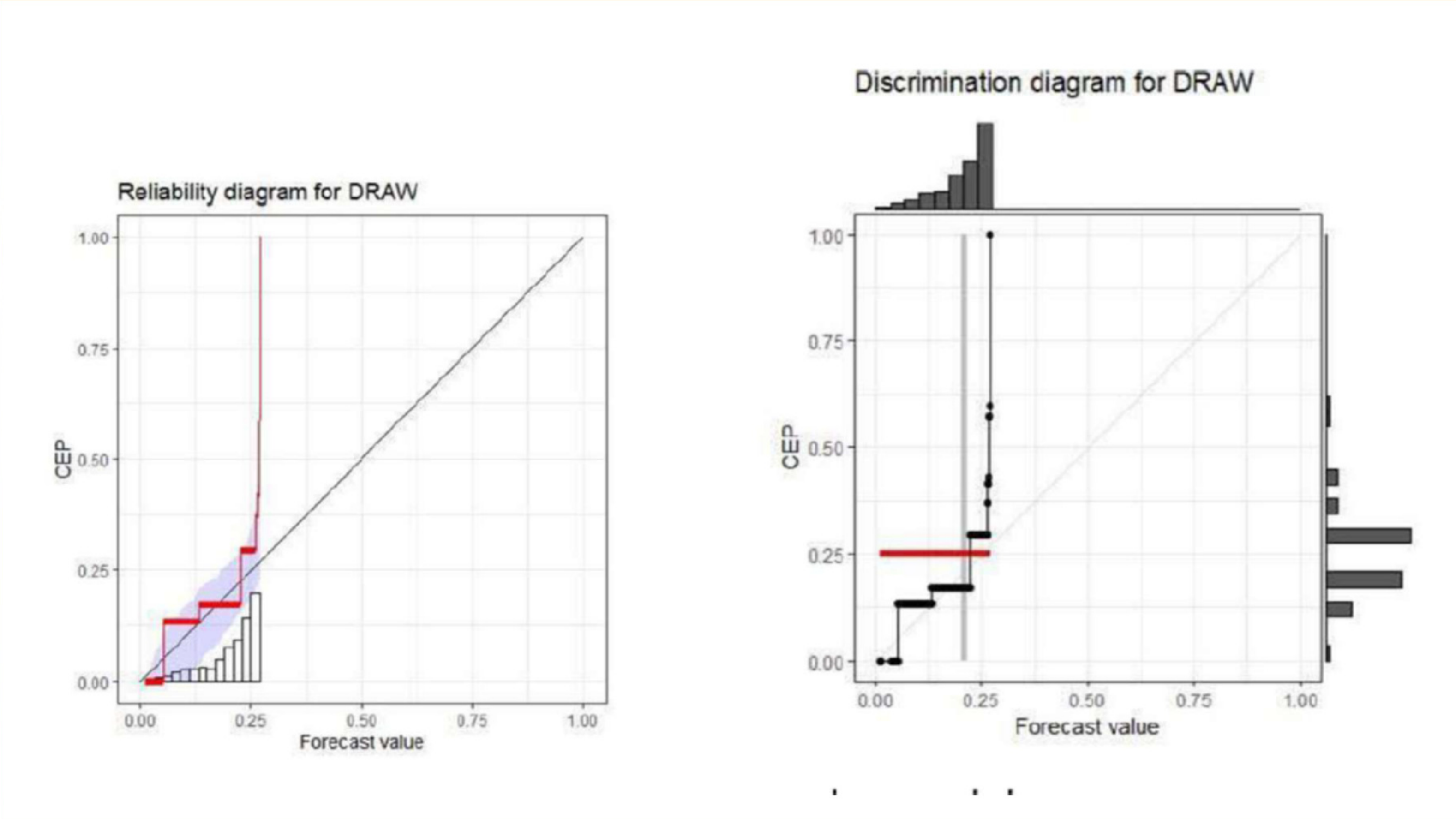
Reliability diagrams for Hwin via Iso-Regression



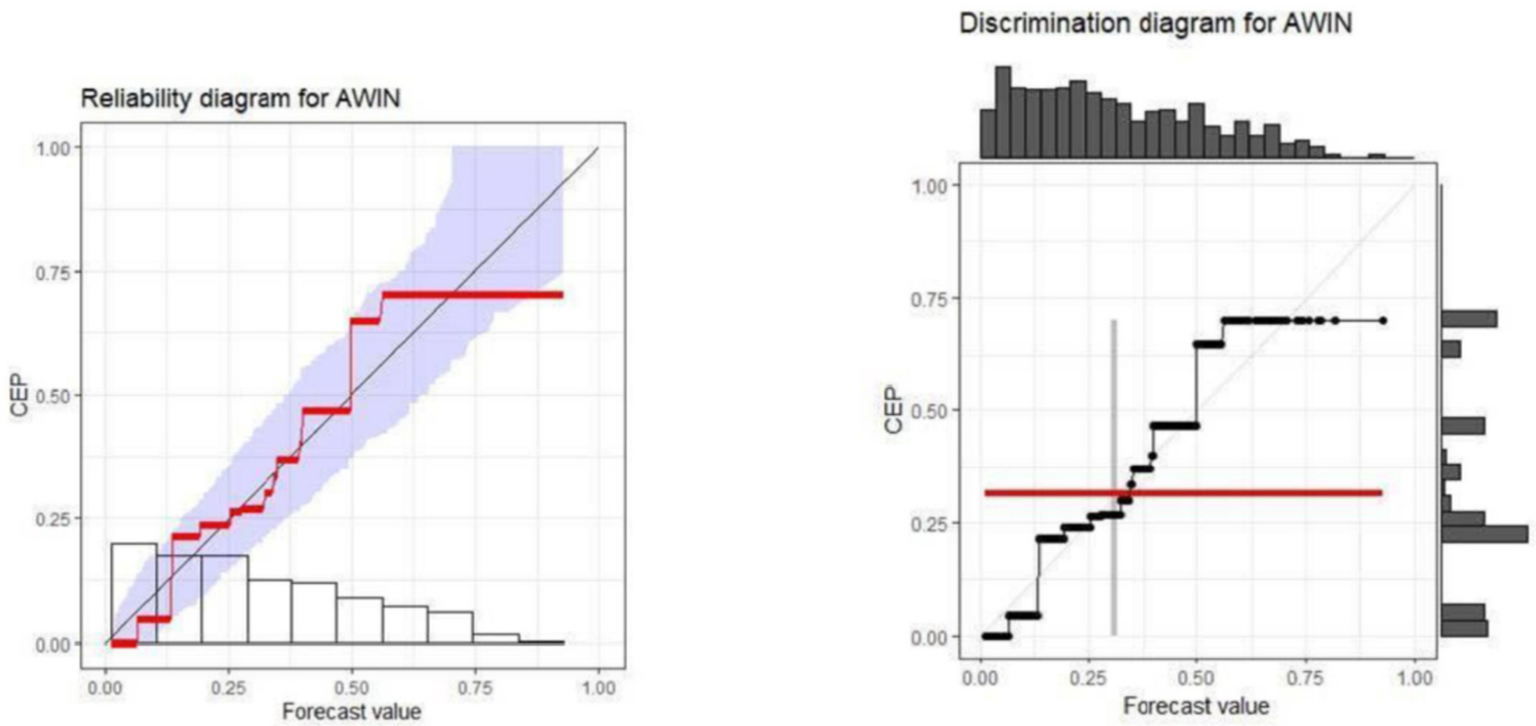
1-HomeWin Poisson Iso-Reg

3-HomeWin Poisson Iso-Reg

Reliability diagrams for Draw via Iso-Regression



Reliability diagrams for Awin via Iso-Regression



Figure/ Reliability Diagrams for AwayWin Probability Forecasts with plots of the Conditional Probability Events (CEP) against the Forecast Probability Values

Calibration via a logistic regression

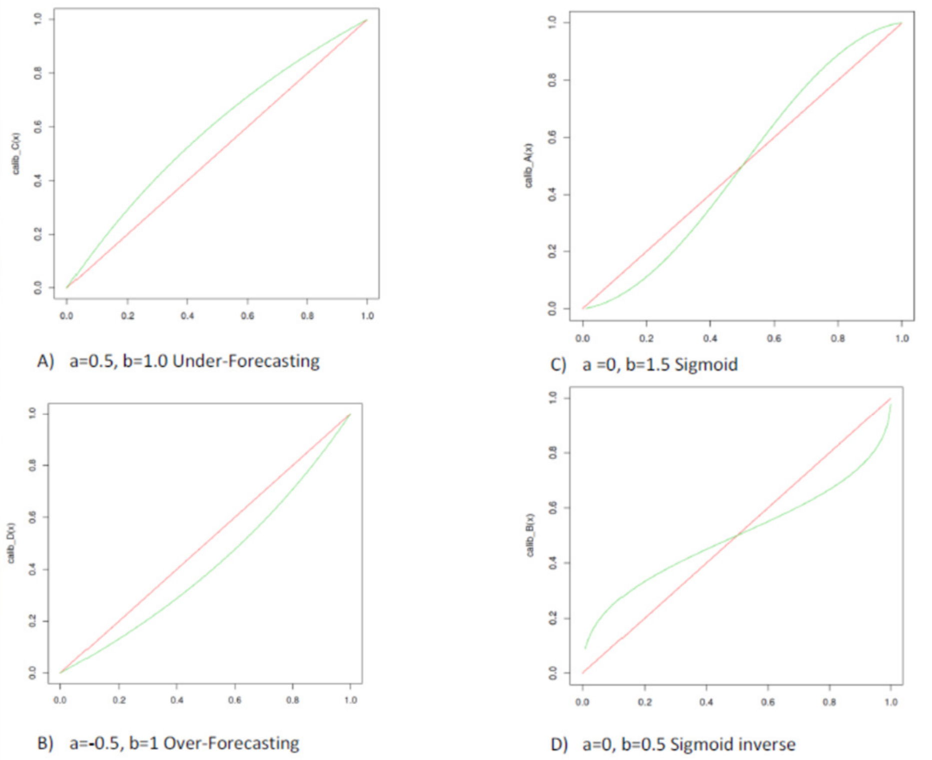


Fig Pattern of calibration as a function of intercept (a) and slope (b) in logistic regression $\boxed{\text{logit}(q_p) = a + b\,\text{logit}(p)}$

Calibration via a logistic regression on logit

Table 2: Calibration analysis via fitting a logistic model of the probability of Homewin, Draw and Awaywin (AWIN) on the logit of its probabilistic forecast under a Poisson regression model (POI)

Category	Criterion	Estimation	SE	T-Statistics	DF	P-value
Homewin	intercept	-0.259	0.119	4.700	1	0.030
	slope	1.113	0.129	0.765	1	0.382
	D0 vs D1	423.085 vs 417.489		5.596	2	0.061
Draw	intercept	0.153	0.466	0.108	1	0.742
	slope	0.932	0.346	0.039	1	0.843
	D0 vs D1	426.981 vs 422.981		4.000	2	0.135
Awaywin	intercept	0.076	0.149	0.261	1	0.610
	slope	1.053	0.134	0.156	1	0.693
	D0 vs D1	389.458 vs 389.176		0.282	2	0.870

Intercept (α) and slope (β) of the logit regression model with their estimation and standard error (SE). Deviance D(k)=-2L(k) where L(k) is the loglikelihood of the null model (0: $\alpha=0$; $\beta=1$) vs the unspecified parameter model (1: $\alpha \neq 0$; $\beta \neq 0$); T-statistics: Wald for intercept=0 and slope=1; Deviance differences ΔD=D0-D1 and their corresponding degrees of freedom (DF) and P-values

Decomposition of BS/Likelihood-base rate factorization

Murphy and Winkler (1987) also gave the dual decomposition of Calibration-Refinement

$$\boxed{\mathbb{E}[S(P,X)] = REF - DIS + CB2}$$

1) Refinement (REF) equal to $\boxed{\mathrm{Var}(P)}$, the variance of probabilistic forecasts also known as Sharpness,

2) **Discrimination** (DIS) equal to $\boxed{\mathrm{Var}_X\left[\mathbb{E}_P(P \mid X)\right]}$ i.e., characterizing the difference between conditional distributions of forecasts given the outcomes X, beneficial

$$\mathrm{Var}_X\left[\mathbb{E}_P(P \mid X)\right] = \boxed{\left[\mathbb{E}_P(P \mid X=1) - \mathbb{E}_P(P \mid X=0)\right]^2} \mathrm{Var}_X(X)$$

3) $CB2 = \mathbb{E}_X\left\{\left[\mathbb{E}_P(P \mid X) - X\right]^2\right\}$ is the dual of reliability labelled as Conditional Bias type 2 by Bradley et al, (2003).

LB factorization/ Distribution of P given X=0 & X=1

Table: Characteristics of conditional distributions of probability forecasts given the outcomes under two Forecasting procedures: Poisson regression (POI) and Odds Probabilities (ODD)

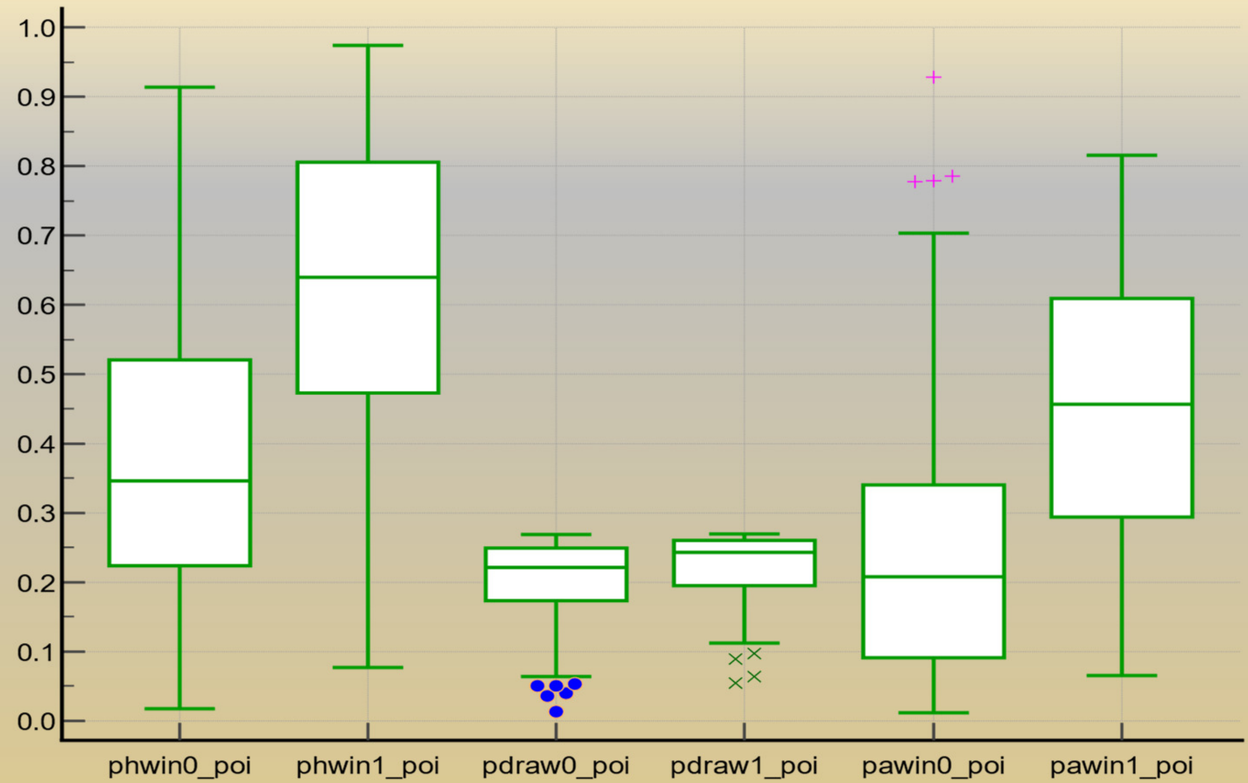
Method		Home Win		Draw		Away Win	
		POI	ODD	POI	ODD	POI	ODD
Sample sizes		217-167		288-96		263-121	
Mean %	X=0	37.88	35.01	20.22	20.97	24.30	23.24
	X=1	63.08	62.73	22.34	23.89	44.84	48.62
	Dif 1-0	**24.20**	**27.71**	**2.02**	**2.93**	**20.54**	**25.38**
Wilcoxon	Z	9.93	10.76	3.59	3.55	9.09	10.13
	P-val	*<0.0001*	*<0.0001*	*0.0002*	*0.0002*	*<0.0001*	*<0.0001*
KS	D	*0.473*	*0.511*	*0.236*	*0.236*	*0.447*	*0.521*
	P-val	*<0.0001*	*<0.0001*	*0.0007*	*0.0007*	*<0.0001*	*<0.0001*
C-statistic	Estimation	**0.795**	**0.820**	**0.622**	**0.624**	**0.789**	**0.820**

Sample sizes of forecasts having X=0 vs X=1 respectively; Z: Normal approximation of the Wilcoxon-test with one sided P value

KS: Kolmogorov-Smirnoff two sample test on Max [F(X=0)-F(X=1)] Empirical Distribution Functions

C-statistic: Harrell's concordance index varying from 0.5 (no discrimination) to 1 (perfect discrimination) equal to AUC (area under the ROC curve

LB factorization/ Distribution of P given X=0 & X=1



Discussion

- Complementary results not shown here on
  - ✓ ROC curves plot of TPR (sensitivity) against FPR (1-specificity) across varying thresholds on forecasts with AUC
- Yates' decomposition alternative to LB
- Application to UEFA, C1
  - ✓ Good results in terms of REL, RES, DIS for Hwin and A win
  - ✓ Lack of RES and DIS for Draw
- Extension of CR decomposition to J multiple category state
- Scoring Rules for parameter inference

  $$\hat{\theta} = ArgMax_{\theta}\bar{S}\left(p_{\theta}\right)$$

  - ✓ Ex: Hyvarinen Score
  - ✓ Minimum Contrast Estimators (Birgé & Massard, 1993)

Probability Scoring Rules/References

- Concepts coming from Meteorological Research
  - ✓ Murphy AH & Winkler RL framework (Winkler, 2006)
  - ✓ Wilks D (2011)
- Broecker J. https://www.reading.ac.uk/maths-and-stats/about/team/jochen-broecker.aspx
- Dawid AP
  - ✓ 1986. Probability forecasting. Encyclopedia Stat Science
  - ✓ 2014. Theory & Application of proper scoring rules
- Gneiting T & Raftery A (2007)
  - ✓ Strictly proper scoring rules, JASA,102, 359

References

- Foulley, J.-L. (2020) “Assessment of probability forecasts of football outcomes: example of the UEFA Champions League”. 40th Virtual International Forecasting Symposium 2020 26-28 oct 2020
- https://www.researchgate.net/publication/344888100_Assessment_of_probability_forecasts_of_football_outcomes_example_of_the_UEFA_Champions_League
- Foulley, J.-L. (2021). “More on verification of probability forecasts for football outcomes: score decompositions, reliability, and discrimination analyses”. arXiv:2106.14345, DOI: 10.13140/RG.2.2.12988.16001
- https://www.researchgate.net/publication/352816631_More_on_verification_of_probability_forecasts_for_football_outcomes_score_decompositions_reliability_and_discrimination_analyses

Acknowledgements

- **Dr James Reade** (Department of Economics, University of Reading, UK) for hosting my forecasts on his website and the challenging confrontation with RED.
- **Dr Gilles Celeux** (Institut de mathématiques d'Orsay, Université Paris-Sud & Inria Orsay, France) for regular discussions on sport outcome events & analysis.
- **Lo Ndeye** (Université Paris-Sud, Orsay, Fr) for early works on modelling forecasts.
- **Timo Dimitriadis** & **Ph Bastien** (Loreal) for guidance in implementing and running the CORP software
- **Dr Maxime Taillardat** (CNRM UMR 3589, Météo-France, Toulouse, France) for his comments and explanations on the scoring rules theory.