Incorporating measurement error in skill assessment

William M. Briggs & Matt Pocernich & David Ruppert

Motivation

- Most verification methods either explicitly or implicitly assume that an observation is accurate. The skill score offered in this presentation incorporates measurement error.
- This skill score incorporates a loss function - the cost of an incorrect forecast. A loss function for a given forecast can be different for different users.

Notation

Observation

$$\begin{aligned} Y &= 1 \ (event) \\ &= 0 \ (no - event) \end{aligned}$$

Forecast

$$\widetilde{X} \in [0, 1]$$

Decision

$$\begin{aligned} \widetilde{X} \rightarrow X &= 1 \ (forecast \ event) \\ &= 0 \ (forecast \ no - event) \end{aligned}$$

Loss

Loss (k)

	$Y = 1$	$Y = 0$
$X = 1$	0	θ
$X = 0$	$1 - \theta$	0

where $\theta \in [0, 1]$ is the cost of forecasting that an event will occur when it doesn't.

- The cost of perfect forecasts need not be 0.
- Relative costs can be established on a monetary basis and then scaled.
- Different users may have different cost values.

Forecasts

Forecast (Decision) – for a calibrated model.

$$X = I(\widetilde{X} > \theta)$$

Climate (base rate)

$$p = P(Y = 1)$$

Optimal Naive Forecast (ONF)

$$ONF = I(p > \theta)$$

Skill Definitions

Skill – in terms of expected loss (risk)

$$E(k^X) < E(k^{ONF})$$

$$Accuracy(X) > Accuracy(ONF)$$

Data

	$Y = 1$	$Y = 0$
$X = 1$	n_{11}	n_{01}
$X = 0$	n_{10}	n_{00}

Skill score with a loss value $\theta = 0.5$

$$K_{\theta=1/2} = \frac{n_{11} - n_{01}}{n_{11} + n_{10}} > 0$$

Quantifying Measurement Error

To quantify measurement error, a “Gold Standard” is needed to verify the accuracy of the observation Y. This gold standard is denoted as W.

$$P(W = 1|Y = 1) = t$$

with perfect observations t = 1.

$$P(W = 1|Y = 0) = u$$

with perfect observations u = 0.

Restrictions

$$t > u$$

$$u < P(X = 1) = \frac{n_{11}}{n_{11} + n_{01}}$$

Skill Score with Measurement Error

Skill score

$$K_{\theta,t,u} = \frac{n_{11}(1-u-\theta(t-u)) - n_{01}(u+\theta(t-u))}{(n_{11}+n_{10})(1-\theta) - n_{++}u(1-\theta)}$$

Original skill score

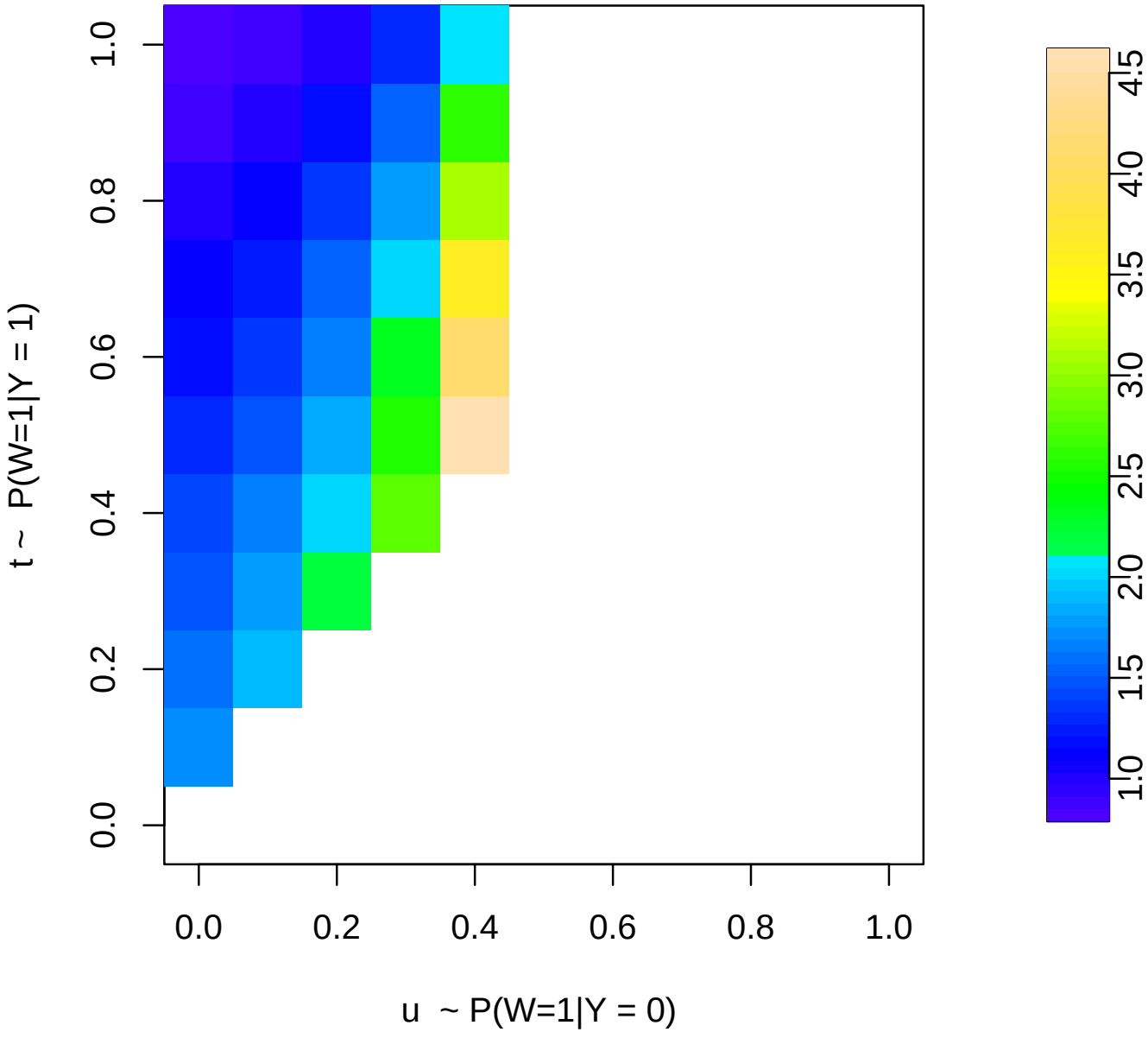
$$K_{\theta=1/2,t=1,u=0} = \frac{n_{11}-n_{01}}{n_{11}+n_{10}}$$

Likelihood Ratio Statistic with asymptotic distribution $\widetilde{\chi}^2$ distribution with 1 df.

$$G_{\theta,t,u} = G_{\theta,t,u}(n_{11}, n_{01}, t, u)$$

Relation between u and t

Example Forecast = c(1000,100,100,1000), theta = 0.5



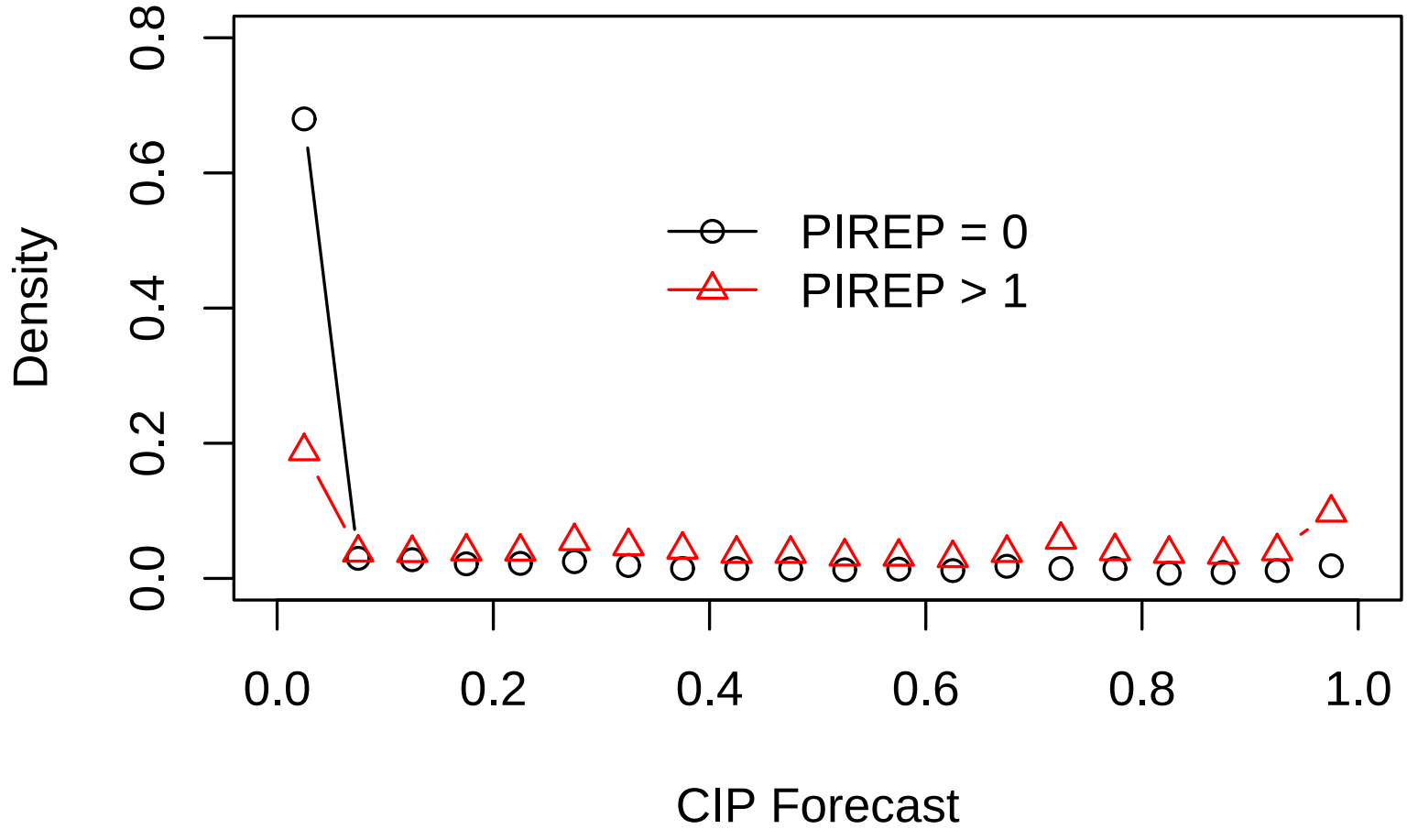
PIREP/ Current Icing Potential (CIP) Example

Forecast Current Icing Potential (CIP)

Observation PIREPs are recorded as a categorical description of icing conditions. When conditions were greater than “trace”, it is assumed that icing is present.

Gold Standard Observation Matching data from specially equipped research, a gold standard for PIREPs was estimated. After trying a range of optimal matching methods, the most agreement was used to calculate values for u and t.

PIREPs from Winter 2003



Note: Icing forecast is not calibrated.

Gold Standard from Research Aircraft Information.

Pirep Gold standard data

	$Y = 1$	$Y = 0$
$W = 1$	43	17
$W = 0$	10	4

Pirep forecast data - Converting forecast to a yes forecast if

$$\widetilde{X} > 0.5, X = 1$$

	$Y = 1$	$Y = 0$
$X = 1$	4028	798
$X = 0$	5161	5267

Summary of CIP skill scores

Pirep data

$$\widehat{P}(W = 1|Y = 1) = t = 0.8113$$

$$\widehat{P}(W = 1|Y = 0) = u = 0.8095$$

$$K_{1/2,u=0} = 0.342, \quad G = 420, \quad \text{p-value} = 0$$

$$K_{1/2,u=0.1} = 0.227, \quad G = 102, \quad \text{p-value} = 0.$$

Conclusion

- Since the sampling is based on the location of the observations, not the location of the forecasts, we do not get a representative sample of of the forecast necessary for the conditional probabilities.
- Likewise, tornados are certainly not all observed (at least at the time of Finely) leaving the total number of tornados (n_{+*}), undercounted.
- This is a measurement error that is beyond this measure of misclassification.
- Possibly a tool to address what is a well known and discussed problem with PIREPs.

Reference

Assessing the skill of yes/no forecasts. by W.M. Briggs and D. Ruppert (2004a) *Reubmitted to Biometrics*.

Relation between K_θ and forecast to binary conversion threshold.