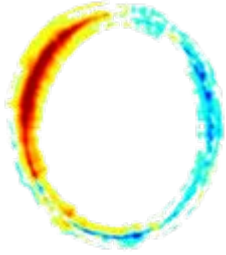
Do High Skill Scores Mean Good Forecasts?

Simon J. Mason

simon@iri.columbia.edu



International Research Institute for Climate and Society
The Earth Institute of Columbia University

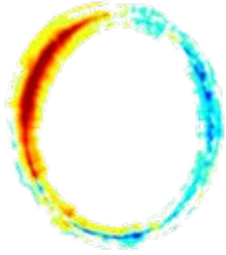
Third International Verification Methods Workshop

Reading, England, 29 January – 2 February, 2007

Do Bad Skill Scores Mean Bad Forecasts?

Simon J. Mason

simon@iri.columbia.edu



International Research Institute for Climate and Society
The Earth Institute of Columbia University

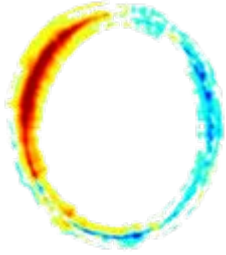
Third International Verification Methods Workshop

Reading, England, 29 January – 2 February, 2007

A Practitioner's Guide to Over-Selling Your Forecasts (and Under-Selling Your Competitor's)

Simon J. Mason

simon@iri.columbia.edu



International Research Institute for Climate and Society
The Earth Institute of Columbia University

Third International Verification Methods Workshop

Reading, England, 29 January – 2 February, 2007

“Tricks of the trade” L. Wilson (2004)

- “How can I get a better (higher) number?”
  - Remove the bias before calculating scores (works really well for quadratic scoring rules) and don’t tell anyone.
  - Claim that the model predicts grid box averages, even if it doesn’t . Make the boxes as large as possible.
  - Never use observation data. It only contains a lot of “noise”. As an alternative,:
    - Verify against an analysis that uses the model being verified as a trial field. Works best in data-sparse areas
    - Use a shorter range forecast from the model being verified and call it observation data.
  - Design a new or modified score. Don’t be bothered by restrictions such as strictly properness. Then the values can be as high as desired.
  - Stratify the verification data using posteriori rules. One can always get rid of pathological cases that bring down the average.
  - When comparing the performance of your model against others, make sure it is your analysis that is used as the verifying one.
  - Always insist on doing the verification of your own products….

Introduction

"I don't know many of the answers, but I do not most of the questions."

W. Macmillan, Oxford University

A few of the questions … and even fewer of the answers.

Introduction

It is not possible to summarise adequately the quality of a set of forecasts in a single number (Murphy 1991).

Skill score problems

Finley's tornado forecasts have received bad press:

OBS.	FORECASTS		
	Tornado	No tornado	*Total*
Tornado	28	23	51
No tornado	72	2680	2752
Total	100	2703	2803

$$\text{Hit score} = \frac{28 + 2680}{2803} = 96.6\%$$

OBS.	FORECASTS		
	Tornado	No tornado	*Total*
Tornado	0	51	51
No tornado	0	2752	2752
Total	0	2803	2803

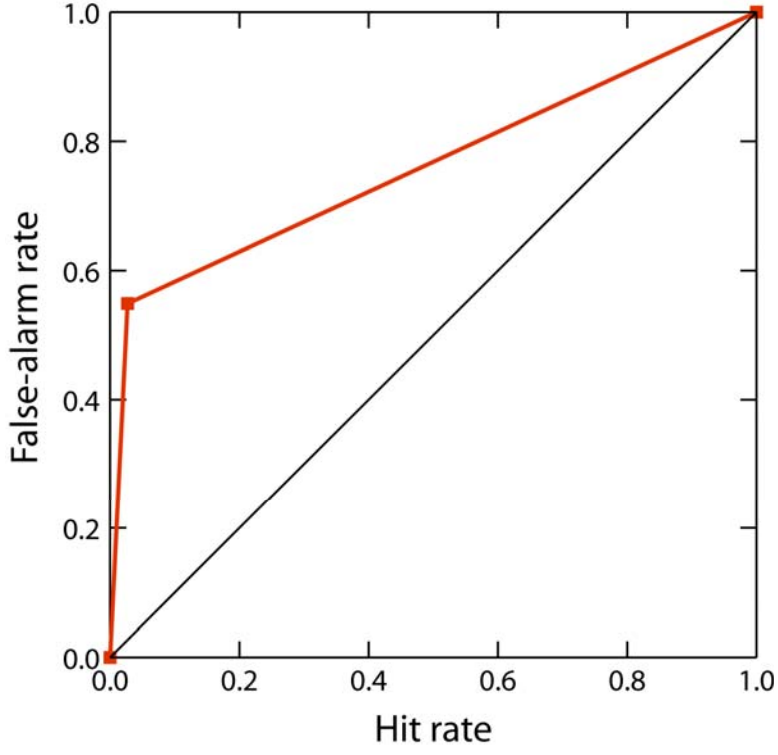
$$\text{Hit score} = \frac{0 + 2752}{2803} = 98.2\%$$

$$\text{Hit skill score} = \frac{28 + 2680 - 2752}{2803 - 2752} = -86.3\%$$

Skill score problems

But Finley's forecasts are not so bad; consider the ROC:

OBSERVATIONS	FORECASTS		
	Tornado	No tornado	*Total*
Tornado	28	23	51
No tornado	72	2680	2752
Total	100	2703	2803



$$\text{hit rate} = \frac{28}{51} = 0.549\%$$

$$\text{false-alarm rate} = \frac{72}{2680} = 0.027\%$$

ROC area = 0.761

(trapezium rule)

ROC for deterministic forecasts

Year	Forecast
1984/85	661
1985/86	658
1986/87	573
1987/88	512
1988/89	707
1989/90	692
1990/91	621
1991/92	532
1992/93	584
1993/94	547
1994/95	496
1995/96	713
1996/97	623
1997/98	386
1998/99	728
1999/00	712
2000/01	682
2001/02	671
2002/03	571
2003/04	597

More generally, the comparison of ROC areas for deterministic and probabilistic forecasts often is performed unfairly.

Consider the following retroactive forecasts of DJF seasonal rainfall totals for Lusaka.

In which years would we expect “wet” conditions (wettest 25%, i.e. > 700 mm) to occur?

ROC for deterministic forecasts

Year	Forecast	Probability
1984/85	661	0%
1985/86	658	0%
1986/87	573	0%
1987/88	512	0%
1988/89	707	100%
1989/90	692	0%
1990/91	621	0%
1991/92	532	0%
1992/93	584	0%
1993/94	547	0%
1994/95	496	0%
1995/96	713	100%
1996/97	623	0%
1997/98	386	0%
1998/99	728	100%
1999/00	712	100%
2000/01	682	0%
2001/02	671	0%
2002/03	571	0%
2003/04	597	0%

A common, but highly unfair, strategy is to convert the deterministic forecasts to probabilistic forecasts with probabilities of 0% (if the forecast is for less than 700 mm), or 100% (if the forecast is for more than 100%).

But surely we would be more confident about the season being “wet” given a forecast of 682 mm than one of 386 mm).

ROC for deterministic forecasts

Forecast	Year	Probability
728	1998/99	95%
713	1995/96	90%
712	1999/00	86%
707	1988/89	81%
692	1989/90	76%
682	2000/01	71%
671	2001/02	67%
661	1984/85	62%
658	1985/86	57%
623	1996/97	52%
621	1990/91	48%
597	2003/04	43%
584	1992/93	38%
573	1986/87	33%
571	2002/03	29%
547	1993/94	24%
532	1991/92	19%
512	1987/88	14%
496	1994/95	10%
386	1997/98	5%

A fairer strategy (and one more consistent with the original formulation of the ROC) would be to list the years in order of decreasing forecast rainfall, and to assign a probability of

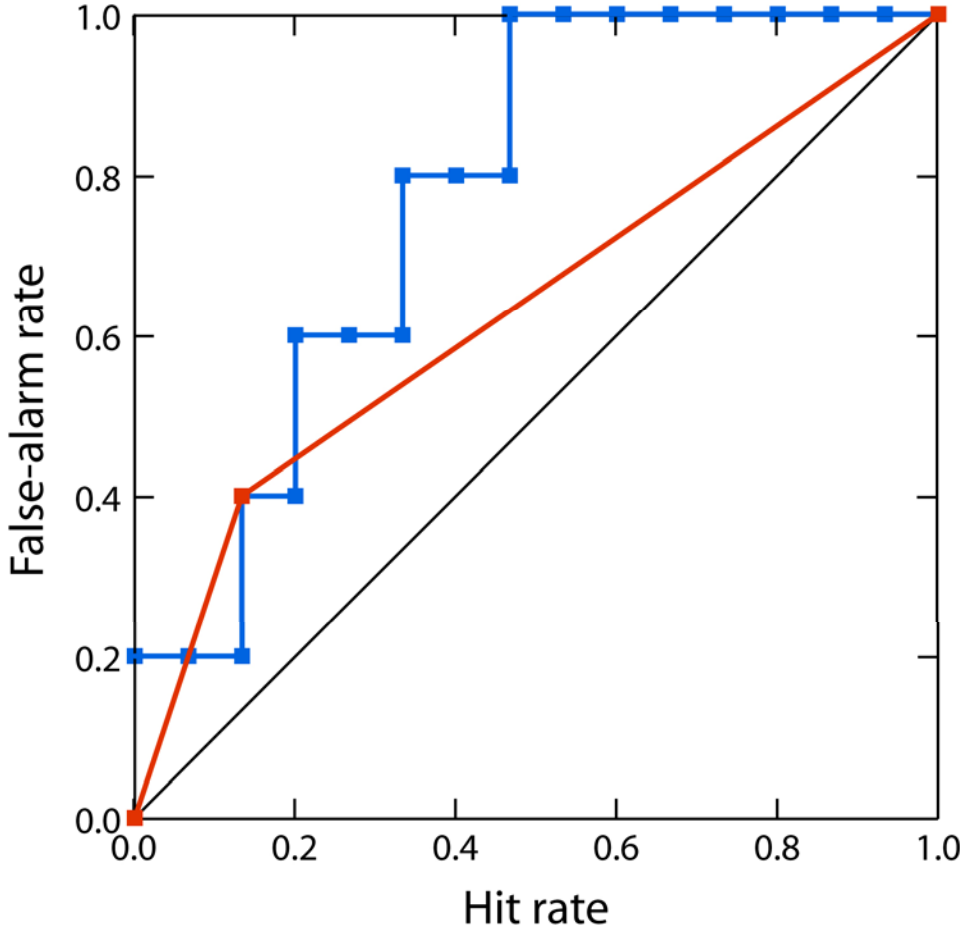
$$(n - r + 1) / (n + 1),$$

where r is the rank of the forecast.

(The actual probability is irrelevant since ROC is insensitive to monotonic transformations of the probabilities.)

ROC for deterministic forecasts

Comparing the ROCs for these two interpretations of deterministic forecasts:



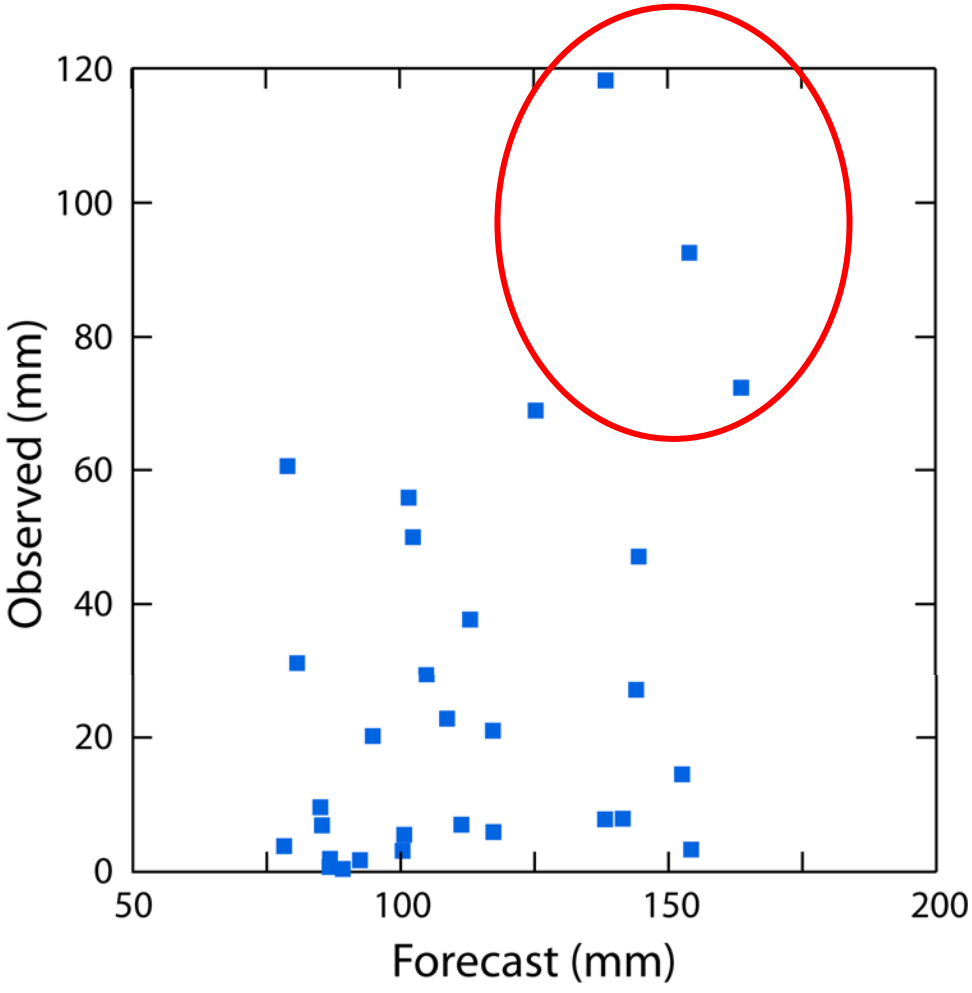
ROC area = 0.773 (continuous)

ROC area = 0.633 (simple)

Tricks of the trade

1. Choose a skill score and / or reference strategy to give you the best possible value and your opponent the worst possible value.

Distributional assumptions



The correlation between ensemble mean JFM seasonal rainfall forecasts for Kalbarri and observed values is 0.391 (30 years).

If the three wettest years (1971, 1975, 2000) are omitted, the correlation drops to 0.079.

Is all the skill contributed by only 10% of the cases? Note that we do not know these years a priori.

Tricks of the trade

1. Choice of skill score and / or reference strategy.
2. Use well-known scores when the distributional assumptions of such scores are violated.

Signal and noise

The correlation between the ensemble mean JFM seasonal rainfall forecasts for Kalbarri and the observed values is greater than the skill of predicting an additional ensemble member; i.e., the skill exceeds the potential predictability!

But is the signal-to-noise ratio the correct way to estimate potential predictability for precipitation?

Signal and Noise

The signal typically is measured by decomposing the ensemble variance into signal and noise terms:

inter-ensemble variance, $\text{signal} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{x}_{i\bullet} - \bar{x}\right)^2$

intra-ensemble variance, $\text{noise} = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2$

But if the variance of the ensemble increases with the ensemble mean then the noise term is over-estimated.

Signal and Noise

To get the noise term the ensemble mean needs to be removed by division not subtraction, and the signal needs to be rescaled to get the correct ratio.

$$\text{noise}' = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\left(x_{ij}-\bar{x}_{i\bullet}\right)}{\bar{x}_{i\bullet}}\right)^2 \quad ?$$

$$\text{signal}' = \frac{n^2}{n-1} \quad ?$$

Are we (slightly?) underestimating the potential predictability of precipitation?

Tricks of the trade

1. Choice of skill score and / or reference strategy.
2. Violate distributional assumptions.
3. Underestimate the potential predictability to demonstrate that you are closer to the maximum possible skill than you really are.

P-values

But given a verification score how do we know whether the score's value is good, especially if the score is some abstract number like the Gerrity score that has no obvious interpretation?

A commonly used method to assess whether a verification score's value is "good" is to calculate the probability that a value at least as good as that observed could have been achieved given completely useless forecasts.

This probability is called a *p*-value.

P-values

Calculating p-values: all methods involve defining a distribution of scores under the null hypothesis of no skill. There are a number of ways of obtaining this distribution:

1. *Exact theoretical distribution*: e.g., binomial for hit rates, *U* for ROC area
2. *Approximate theoretical distribution*: e.g., Student's *t* for correlation, gaussian for ROC area.
3. *Empirical distribution*: using permutation methods.
4. *Empirical distribution*: using artificial series.

P-values

Sample

No.	Obs 1	Obs 2	No.	For 1	For 2
1	-0.08	-0.23	1	0.28	0.16
2	1.56	1.59	2	0.77	0.87
3	0.58	0.41	3	0.44	0.34
4	0.90	0.92	4	0.59	0.71
5	-0.21	-0.55	5	0.37	0.19

Permutation 1

No.	Obs 1	Obs 2	No.	For 1	For 2
4	0.90	0.92	1	0.28	0.16
1	-0.08	-0.23	2	0.77	0.87
3	0.58	0.41	3	0.44	0.34
5	-0.21	-0.55	4	0.59	0.71
2	1.56	1.59	5	0.37	0.19

Permutation 2

No.	Obs 1	Obs 2	No.	For 1	For 2
5	-0.21	-0.55	1	0.28	0.16
4	0.90	0.92	2	0.77	0.87
1	-0.08	-0.23	3	0.44	0.34
3	0.58	0.41	4	0.59	0.71
2	1.56	1.59	5	0.37	0.19

All four methods assume that all the forecast-observation pairs are independent of other forecast-observation pairs. If this assumption is invalid, the permutation procedure may be modifiable to account for the dependence by block sampling.

P-values

Sample

No.	Obs 1	Obs 2	No.	For 1	For 2
1	-0.69	-0.70	1	-0.28	-0.30
2	-0.08	-0.23	2	0.28	0.16
3	1.56	1.59	3	0.77	0.87
4	0.58	0.41	4	0.44	0.34
5	0.90	0.92	5	0.59	0.71
6	-0.21	-0.55	6	0.37	0.19

Permutation 1

No.	Obs 1	Obs 2	No.	For 1	For 2
3	1.56	1.59	1	-0.28	-0.30
4	0.58	0.41	2	0.28	0.16
1	-0.69	-0.70	3	0.77	0.87
2	-0.08	-0.23	4	0.44	0.34
5	0.90	0.92	5	0.59	0.71
6	-0.21	-0.55	6	0.37	0.19

Permutation 2

No.	Obs 1	Obs 2	No.	For 1	For 2
3	1.56	1.59	1	-0.28	-0.30
4	0.58	0.41	2	0.28	0.16
5	0.90	0.92	3	0.77	0.87
6	-0.21	-0.55	4	0.44	0.34
1	-0.69	-0.70	5	0.59	0.71
2	-0.21	-0.23	6	0.37	0.19

If there is temporal dependence then temporal block sampling will need to be applied as well.

P-values

To obtain *p*-values for probabilistic verification scores where the probabilities are discrete, a permutation procedure may not be valid.

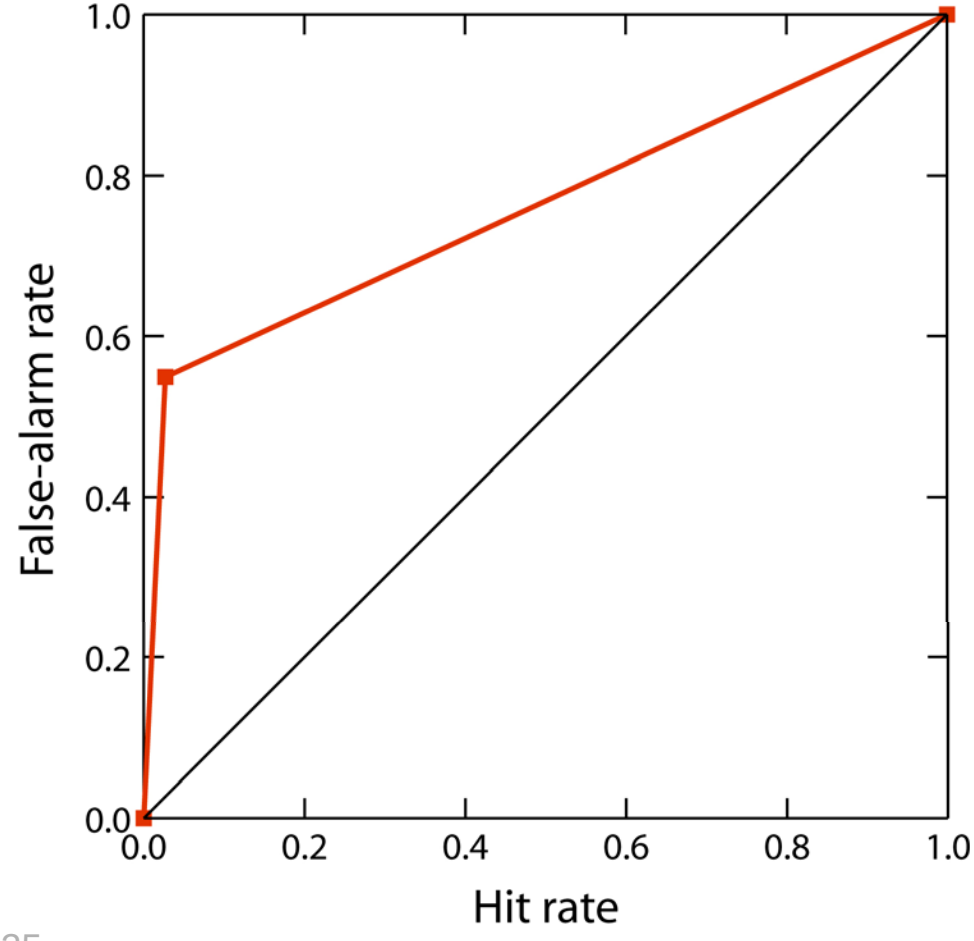
Consider the following forecasts, both of which have an identical and minimum possible *p*-value (the pairing of observations and forecasts is the best possible *given* the forecasts). But forecast set A is clearly preferable to set B, and we can imagine a third set of probabilities that would improve on both.

Should one regenerate probabilities? Should the dependence between the ensemble members be considered to reproduce the sharpness?

Obs	1	1	0	0	0
A	90	80	15	10	5
B	70	60	30	20	10

P-values

Finley's forecasts do generate a good ROC:



ROC area $= 0.761$

(trapezium rule)

p-value $= 0.794 \times 10^{-11}$

P-values

P-values indicate only how confident we can be that our forecasts have *some* skill; the actual amount of skill that we may have could be exceedingly small.

So a small *p*-value allows us only to say:

“I am very confident that I do not have no skill.”

Tricks of the trade

1. Choice of skill score and / or reference strategy.
2. Violate distributional assumptions.
3. Underestimate the potential predictability.
4. If the sample is large, avoid showing the score and show only the *p*-value (or if unavoidable, show an obscure and abstract score),. Forecasts with very marginal skill can be made to look spectacularly good.

P-values

So, if “I am very confident that I do not have no skill”, how much skill do I have?

The sample score provides *one* indication of the skill. But is this value correct?

Confidence intervals

If we had a different set of forecasts the calculated score will vary from the sample score even if the skill of the forecasts is unchanged. The calculated score is therefore only an estimate of the 'real' score. It would be helpful to know how sensitive the score is to the sample; if the score is sensitive the uncertainty in the estimate will be high.

A recommended way of indicating uncertainty is to calculate confidence intervals. (Confidence intervals can also be used as an alternative to *p*-values).

Confidence intervals

There are many ways of calculating confidence intervals. Some of the most commonly used procedures include:

1. *Exact theoretical distribution*: e.g., binomial for hit rates

2. *Approximate theoretical distribution*: e.g., Student's *t* for ROC area.

3. *Empirical distribution*: using bootstrap methods.

As with the *p*-values, all three methods assume that all the forecast-observation pairs are independent of other forecast-observation pairs.

Confidence intervals

Sample

No.	Obs 1	Obs 2	No.	For 1	For 2
1	-0.08	-0.23	1	0.28	0.16
2	1.56	1.59	2	0.77	0.87
3	0.58	0.41	3	0.44	0.34
4	0.90	0.92	4	0.59	0.71
5	-0.21	-0.55	5	0.37	0.19

Bootstrap 1

No.	Obs 1	Obs 2	No.	For 1	For 2
1	-0.08	-0.23	1	0.28	0.16
1	-0.08	-0.23	1	0.28	0.16
3	0.58	0.41	3	0.44	0.34
4	0.90	0.92	4	0.59	0.71
4	0.90	0.92	4	0.59	0.71

Permutation 2

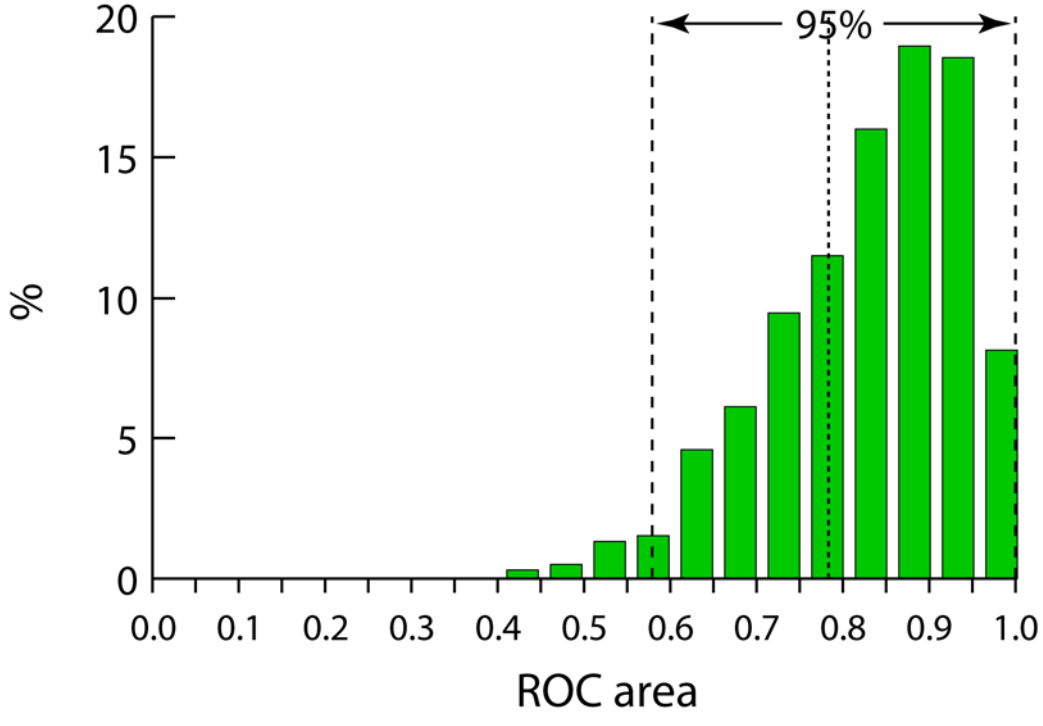
No.	Obs 1	Obs 2	No.	For 1	For 2
2	1.56	1.59	5	0.37	0.19
2	1.56	1.59	5	0.37	0.19
2	1.56	1.59	5	0.37	0.19
4	0.90	0.92	4	0.59	0.71
5	-0.21	-0.55	5	0.37	0.19

A *bootstrap* procedure involves resampling with replacement (compare with the permutation procedure in which the object is to generate useless sets of forecasts).

Block temporal resampling could be applied if there is temporal dependence.

Confidence intervals

The best way to obtain confidence intervals for most verification procedures is by bootstrapping.



Note:

1. the sample score can be biased (some bootstrap procedures adjust for this);
2. the distribution of skill scores generally will be skewed.

Tricks of the trade

1. Choice of skill score and / or reference strategy.
2. Violate distributional assumptions.
3. Underestimate the potential predictability.
4. Show *p*-values if your sample size is large.
5. Show confidence intervals only if they suit your purposes.

Conclusion

“Lies, damn lies, and verification.”