

Statistical Power -
A Neglected Topic in
Forecast Verification

Ian Jolliffe

University of Aberdeen

Jacqueline Potts

Biomathematics and Statistics Scotland

Outline of Talk

1. Statistical Significance and Power for Verification Measures
2. Some Standard Verification Measures
3. A Case Study - Mean Sea Level Pressure on a Spatial Grid
 - Real Data
 - Simulations
 - Other Verification Measures
4. Concluding Remarks

Three possible steps in verification

Suppose we have a set of forecasts and a corresponding set of verifying observations. We may follow the three steps:

1. Calculate the value of one or more measures of how good the forecasts are
2. Assess the statistical significance of the calculated value(s)
3. Investigate the statistical power of the measure(s)

1. is commonplace, and is sometimes followed by 2., but 3. is rarely done.

Statistical Significance and Power

- Statistical significance - given a value of a verification measure how likely is it that this value could have arisen by chance if the forecasts have no skill? To calculate this we need to know the the probability distribution of the measure under the *null hypothesis* of no skill.
- Power - if the forecasts have genuine skill how likely is it that our verification measure will be able to detect it? This requires knowledge of the probability distribution of the measure under an *alternative* hypothesis. Whereas the null hypothesis is simple, often the alternative is not.

Some Standard Verification Measures

Suppose our forecasts are made at a grid of n points, that the forecasts are $\hat{x}_i, i = 1, 2, \dots, n$ and the corresponding verifying observations are $x_i, i = 1, 2, \dots, n$. Let $\bar{\hat{x}}$ and \bar{x} be the averages over the grid of the forecasts and verifying observations respectively. Also let c_i represent 'climatology' at the i th gridpoint, and \bar{c} the average of the c_i over the grid. The following are commonly used verification measures:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{x}_i - x_i)^2$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(\hat{x}_i - \bar{\hat{x}})}{(\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (\hat{x}_i - \bar{\hat{x}})^2)^{\frac{1}{2}}}$$

$$ACC = \frac{\sum_{i=1}^n ((x_i - c_i) - (\bar{x} - \bar{c}))((\hat{x}_i - c_i) - (\bar{\hat{x}} - \bar{c}))}{(\sum_{i=1}^n ((x_i - c_i) - (\bar{x} - \bar{c}))^2 \sum_{i=1}^n ((\hat{x}_i - c_i) - (\bar{\hat{x}} - \bar{c}))^2)^{\frac{1}{2}}}$$

$$ACC_u = \frac{\sum_{i=1}^n (x_i - c_i)(\hat{x}_i - c_i)}{(\sum_{i=1}^n (x_i - c_i)^2 \sum_{i=1}^n (\hat{x}_i - c_i)^2)^{\frac{1}{2}}}$$

Case study - Mean Sea Level Pressure(MSLP)

We consider two data sets, both giving MSLP for the first 4 half-months of the year for the 30 years 1951-1980. 'Climatology' is the average pressure field for these 30 years averaged over all 4 half-months.

- North Atlantic Region - 99 gridpoints on the $5^\circ \times 10^\circ$ grid covering $30^\circ N - 70^\circ N$ and $60^\circ W - 40^\circ E$
- U.K. region - 30 gridpoints on the same grid covering $45^\circ N - 65^\circ N$ and $30^\circ W - 20^\circ E$

Null distributions

It is possible to estimate null distributions of our verification measures (or others) in two ways.

- use asymptotic expressions for means and variances of measures together with a Gaussian assumption
- calculate values of the measures from two sets which are (almost) unrelated such as data from the same half-month in different years

The spatial correlation in the data poses a major problem in evaluating null distributions. Comparison of these two approaches suggests that there are only 6 or 7 independent pieces of information in the 99 North Atlantic grid-points, and only 3 or 4 in the UK region.

Because of these small effective sample sizes a Gaussian assumption for the null distribution is highly dubious for many measures.

For some measures (for example correlations) known non-Gaussian distributions are more appropriate but the problem of spatial dependence is still present.

Estimating power

The empirical null distributions of the verification measures can be used to determine thresholds for the measures beyond which we reject the null hypothesis of no skill.

To determine power we see how often the threshold is exceeded in circumstances in which the two data sets being compared are related. This can be done either

- using simulated data with known structure
- empirically using data known to be related

Simulated data

Spatial patterns displayed by MSLP can be modelled rather well by multivariate normal (Gaussian) distributions. Hence we generate pairs of related fields from a MND to investigate power.

Let \mathbf{c} be the vector of climatology for our n gridpoints and \mathbf{x} , \mathbf{y} corresponding vectors for two realisations of the field. The matrix Σ is a diagonal matrix of standard deviations of \mathbf{x} and \mathbf{y} and \mathbf{R} is the corresponding correlation matrix. We can write $\mathbf{R} = \mathbf{L}\mathbf{L}'$ where \mathbf{L} is lower triangular.

Finally suppose that ϵ , ξ are each vectors of n independent $N(0,1)$ random variables and $\eta = \rho\epsilon + \sqrt{(1 - \rho^2)}\xi$, where ρ is a scalar lying between zero and one.

Then generate

$$\mathbf{x} = \mathbf{c} + \Sigma\mathbf{L}\epsilon; \quad \mathbf{y} = \mathbf{c} + \Sigma\mathbf{L}\eta$$

The scalar ρ is the correlation between corresponding elements of \mathbf{x} and \mathbf{y} and as it increases so does the dependence between the fields defined by \mathbf{x} and \mathbf{y} .

Results

Results are based on 10000 simulated pairs of data sets. A 'W' indicates a weighted version of a verification measure, with weights based on the area of a gridbox represented by a grid point - a gridbox at $30^{\circ}N$ has more than twice the area of one at $70^{\circ}N$. The unweighted results are based on standardised data (divided by standard deviation) whereas no such standardisation is done for the weighted data.

A second table presents some empirical results in which the dependent pairs of fields are the fields observed in consecutive half-months. There are 90 such pairs in our data sets, and we present the percentage of rejections out of 90 for two thresholds.

Percentage rejections (out of 10000) for simulated dependent data (5% critical values)

Score	Region	$\rho = 0.2$	$\rho = 0.6$	$\rho = 0.9$
WMSE	NA	10	43	98.0
Wr	NA	11	45	95
WACC	NA	13*	55*	98.3*
$WACC_u$	NA	12	46	95
WMSE	UK	8	27	85
Wr	UK	11	36	86*
WACC	UK	11*	37*	85
$WACC_u$	UK	10	31	77
MSE	NA	11	51	99.7*
ACC_u	NA	13*	56*	98.5
MSE	UK	8	29	91*
ACC_u	UK	10*	34*	84

Percentage rejections (out of 90) for dependent (consecutive half-month) data

Score	Region	5% level	10% level
WMSE	NA	11	21
Wr	NA	11	20
WACC	NA	10	28*
<i>WACC_u</i>	NA	14*	26
WMSE	UK	8	17
Wr	UK	4	15
WACC	UK	9*	20*
<i>WACC_u</i>	UK	9*	20*
MSE	NA	10	20
<i>ACC_u</i>	NA	18*	32*
MSE	UK	9	17
<i>ACC_u</i>	UK	12*	18*

From the tables we see

- There is less power for the UK region than the North Atlantic. Unsurprising, as we'd expect greater power for larger data sets
- There is greater power for WACC than from either MSE or 'ordinary' correlation, except for large values of ρ . Intuitively, given that dependence is introduced via correlation in the simulated data, we might expect MSE to do less well there
- There is little to choose between weighted and unweighted measures

A further empirical comparison was done using as 'dependent' data ensembles of 30-day forecasts made at 6-hour intervals. The three conclusions above hold for these data too.

Other measures

Measures other than those defined above were also investigated.

- S1 - compares E-W and N-S pressure *differences* for the two maps being compared. Power is slightly better than those reported above, perhaps because, unlike the other measures, we are taking spatial position into account
- Measures, such as LEPS, based on comparing probability distributions for predicted and observed MSLP. Power is generally worse than for the other scores
- Measures based on the principal components (EOFs) of MSLP fields. Power is similar to that for other measures

Concluding remarks

- Little has been done to investigate the power of various verification measures
- The results above give some information but they are limited in scope. In particular the simulations examine only one particular type of dependence between fields, and field data are only one of many classes of data that are forecast
- Different measures quantify different aspects of similarity between forecasts and verifying data. Hence different measures are likely to be powerful for different alternative hypotheses