MET3220C
Computational Statistics

Forecast Verification

(Chapter 7 of Wilk's book)

Key Points:

1) Background on terminology
2) Contingency tables
3) Quantitative measures of skill

Joint Distributions of Forecasts and Observations

- Consider a forecast y, with I possible values or bins of values.
  - Values or bins are $y_1, y_2, y_3, \ldots, y_I$
- Consider the corresponding observation o, with J possible values or bins of values.
  - Values or bins are $o_1, o_2, o_3, \ldots, o_J$
- The joint distribution of forecasts and observations is written as
  - $p(y_i,o_j) = \Pr\{y_i,o_j\} = \Pr\{y_i \cap o_j\}$; for $i = 1$ to I; $j = 1$ to J
  - This function associates a probability with each of the possible $I \times J$ possible combinations of forecast and observation.
- From the perspective of someone attempting to calibrate (or tune) a forecast, the above equation can be refined as $p(y_i,o_j) = p(o_j|y_i)\, p(y_i)$.
- That is the likelihood of a certain observed value, given a particular forecast. This is called the calibration-refinement (Murphy and Winkler 1987).

From A Seemingly Backward Perspective

- Recall that the joint distribution of forecasts and observations
  - $p(y_i,o_j) = \Pr\{y_i,o_j\}=\Pr\{y_i \cap o_j\}$; for $i = 1$ to I; $j = 1$ to J
- From the perspective of someone applying a forecast, the above equation can be refined as $p(y_i,o_j) = p(y_i|o_j)\, p(o_i)$.
- That is the likelihood of a certain forecast value, given a particular observation. This is called the likelihood-base rate factorization (Murphy and Winkler 1987)

Forecast Verification: How Many Data Do We Need?

- We need enough data to make an adequate estimation of the probability of each of the $I \times J$ possible outcomes.
- Example:
  - Assume that we are rounding temperature forecasts to the nearest degree, and that there are 50 possible outcome for observations, and 60 possible outcomes for the forecast.
  - We then need a minimum of 50×60 times the number of observations needed to make an adequate estimation of a single probability. Why is this a minimum?
  - Recall that the number needed to determine a probability to within a specific measure of error is dependent on the probability.
    - We need to assume the worst in this calculation!
    - Say 1000 points.
  - Then we need 50×60 times 1000 observations = 3,000,000 observations.
    - Collected in a time and space where the statistics do not change!

Forecast Skill

- There are many measures of forecast skill.
  - Different measures emphasize different strengths and weaknesses.
- One of the most common measures is a ‘skill score.’

$$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\%$$

- Where
  - A is a measure of accuracy,
  - A_{perf} is a the measure of accuracy for a perfect forecast, and
  - A_{ref} is reference level to which the accuracy is compared.
- If $A < A_{ref}$, then the skill score is negative,
  - If $A = A_{ref}$, then the skill score is zero,
  - If $A_{perf} > A > A_{ref}$, then the skill score is positive and less than 1,
  - If $A = A_{perf}$, then the skill score is one.
- A_{ref} is often set to the accuracy of a forecast based solely on climatology.

Contingency Tables

- Many measures of skill (or lack of skill) can be described in terms of the odds of outcomes based on a 2x2 contingency table.

Observed Condition

Forecast Condition		Yes	No	
	Yes	Hit (a)	False Alarm (b)	a + b
	No	Miss (c)	Correct Negative (d)	c + d
		a + c	b + d	a + b + c + d = n

- The numbers outside the 2x2 box are the marginal totals.
  - Bottom totals are the marginal totals of the observations.
  - Left totals are the marginal totals of the forecasts
- The sum of the marginal total for observations MUST equal the sum of the marginal totals for the forecasts.
- Both sums are equal to *n*.

Contingency Tables for Probabilities

- A similar 2x2 table can be constructed for probabilities of the joint distribution.

Observed Condition

		Yes	No	
Forecast Condition	Yes	Hit (a/n)	False Alarm (b/n)	(a + b)/n = $p(o_1)$
	No	Miss (c/n)	Correct Negative (d/n)	(c + d)/n = $p(o_2)$
		(a + c)/n = $p(y_1)$	(b + d)/n = $p(y_2)$	n/n=1

- The numbers outside the 2x2 box are the marginal distributions.
  - Bottom totals are the marginal distributions of the observations.
  - Left totals are the marginal distributions of the forecasts
- The sum of the marginal distributions for observations MUST equal the sum of the marginal distributions for the forecasts. MUST equal 1.

Scalar Measure of Skill
Based on The Contingency Table

- Many of the following measures of skill were first developed in medical fields. Applied medicine journals often provide more useful information on the strengths and weaknesses of these techniques than can be found in meteorology discussions.
- Accuracy
  - Accuracy (PC; percentage correct) indicates the fraction of correct forecasts.
  - Correct forecasts are hits and correct negatives
  - These are converted to a fraction by totaling them, and dividing by the sample size.
  - $PC = (a + d) / n$
- What does the value 1 – PC indicate?
- The fraction of wrong forecasts (false alarms and misses).
- PC values near 1 are good, and values near zero are poor. However, if the events being forecast are rare, then PC does not give a good indication of skill in achieving hits. Why?
- Because all the weight is on correct negatives!

Threat Score or Critical Success Index

- The critical success index (CSI) is more appropriate to use that the PC for situations when the event being forecast is rare.
  - Originally proposed by Gilbert (1884). Called ratio of verification.
- The CSI does not consider the number of correct negatives. It compares the number of hits to the number of non-’correct negatives.’
  - CSI = a / (a + b + c)
- A value of zero is appallingly bad.
- A value of one is perfect.

Odds Ratio

- The odds ratio (θ) examines a ratio of odds. In this context, odds are defined as a probability to the compliment of that probability, $p / (1 - p)$.
  - See Stephenson (2000) for the first meteorological application.
- In this case, consider the odds of a hit given that the event does occur, divided by the odds of a false alarm given that the event does not occur.
  - That is, a ratio of the
    - odds of a hit given that the event does occur: $[a / (a + c)] / [1 - a / (a + c)] = a / c$ to
    - odds of a false alarm given that the event does not occur: $[b / (b + d)] / [1 - b / (b + d)] = b / d$
    - $\theta = (a / c) / (b / d) = a d / (b c)$
- This ratio is the product of correct forecasts to the product of incorrect forecasts.
- A value of one is consistent with forecasts that are independent of observations. Larger is better.

False Alarm Rate (FAR)

- The false alarm rate is the chance of false alarm given that the forecast event did not occur.
  - $FAR = b / (a + b)$
  - Smaller values are better.
- FAR is very important when the consequences of a false alarm are serious.
  - Example close schools and offices on the basis of a hurricane forecast.

Others

- There are many other skill scores.
- Wilks' text lists some.
- On behalf of the the WWRP/WGNE Joint Working Group on Verification, the Australian Bureau of Meteorology web site lists many of those used in meteorology, and has good explanations.
  - http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html
  - This web page is a good resource!