Forecast Verification: Part II

Some Commonly Used Verification Metrics

- Non-probabilistic forecasts of discrete predictands: such as the occurrence of heatwaves or extreme precipitation
  - The 2X2 contingency table
  - Metric for 2X2 contingency tables: proportion correct, threat score, hit rate, false alarm rate/ratio
- Now let's look at an example

Wilks 2011

Skill Scores for 2 × 2 Contingency Tables: An Example

Finley chose to evaluate his forecasts using the proportion correct:

PC = (28 + 2680)/2803 = 0.966.

On the basis of this proportion correct, Finley claimed 96.6% accuracy.

Gilbert (1884) pointed out that always forecasting no would produce an even higher proportion correct:

PC = (0 + 2752)/2803 = 0.982,

(a)		Tornados Observed	
		Yes	No
Tornados	Yes	**28**	**72**
Forecast	No	**23**	**2680**
			n = 2803

(b)		Tornados Observed	
		Yes	No
Tornados	Yes	**0**	**0**
Forecast	No	**51**	**2752**
			n = 2803

Finley's tornado forecasts (Finley 1884): tornado forecasts by John Finley, a sergeant in the U.S. Army.

Wilks 2011

Skill Scores for 2 × 2 Contingency Tables: An Example (cont'd)

Employing the threat score gives a more reasonable comparison, because the large number of easy, correct no forecasts are ignored.

Finley's forecasts

(a)		Tornados Observed	
		Yes	No
Tornados	Yes	**28**	**72**
Forecast	No	**23**	**2680**
			n = 2803

TS = 28/(28 + 72 + 23) = 0.228
HSS=0.355

No forecasts

(b)		Tornados Observed	
		Yes	No
Tornados	Yes	**0**	**0**
Forecast	No	**51**	**2752**
			n = 2803

TS = 0/(0 + 0 + 51) = 0
HSS=0

Other metrics for Finley's forecasts:
Bias B = 1.96, approximately twice as many tornados were forecast as actually occurred.
false alarm ratio is FAR = 0.720, 72% of the forecast tornados did not eventually occur.
hit rate H = 0.549, about half of the actual tornados were forecast to occur
false alarm rate F = 0.0262, a very small fraction of the nontornado cases falsely warned of a tornado.

Wilks 2011

Which Metric?

- TS works better than PC for rare events.
- Multiple metrics are available for each type of forecasts, but none of them by itself can provide a complete picture of the forecast verification. It is often necessary to use more than one metric, as there many ways for forecasts to go right or go wrong, and different metrics focus on different aspects.

Continuous Nonprobabilistic Forecasts: MAE and MSE

- Mean Absolute Error (MAE): the arithmetic mean of the absolute values of the differences between the members of each pair.

$$\mathrm{MAE} = \frac{1}{n}\sum_{k=1}^{n}|y_k - o_k|.$$

- Mean Squared Error (MSE): the average squared difference between the forecast and observation pairs.

$$\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2.$$

- *Which metric is more sensitive to large errors?*

Since the MSE is computed by squaring forecast errors, it will be more sensitive to larger errors than will the MAE, and so will also be more sensitive to outliers.

Wilks 2011

Continuous Nonprobabilistic Forecasts (cont'd): correlations

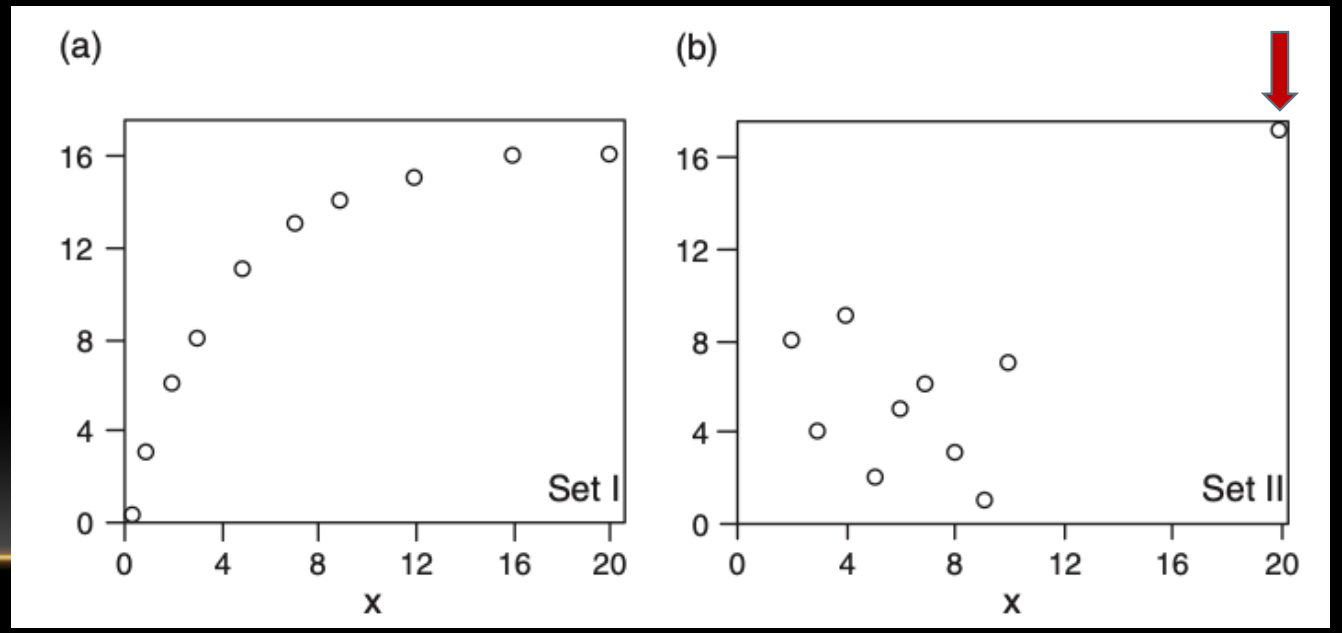
- Pearson correlation: measures the linear relation between the forecast and observation.
- Spearman Rank correlation: measures the degree of similarity between the forecast ranking and the observation ranking.
  - simply the Pearson correlation coefficient computed using the ranks of the data

1,1
2, 2
3, 3
4, 4
……

10, 10

TABLE 3.4 Artificial paired data sets for correlation examples.

Set I		Set II	
x	y	x	y
0	0	2	8
1	3	3	4
2	6	4	9
3	8	5	2
5	11	6	5
7	13	7	6
9	14	8	3
12	15	9	1
16	16	10	7
20	16	20	17



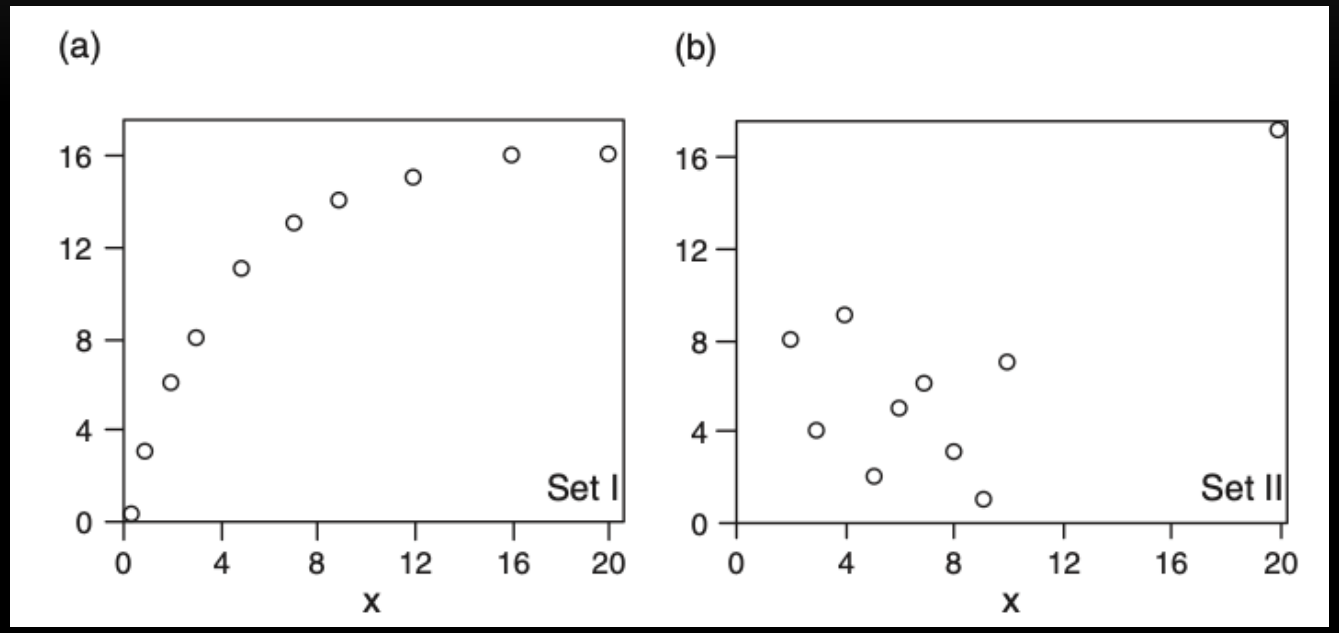
What Pearson and Rank correlations do we have for each dataset?

Wilks 2011

Continuous Nonprobabilistic Forecasts: correlations (cont'd)

TABLE 3.4 Artificial paired data sets for correlation examples.

Set I		Set II	
x	y	x	y
0	0	2	8
1	3	3	4
2	6	4	9
3	8	5	2
5	11	6	5
7	13	7	6
9	14	8	3
12	15	9	1
16	16	10	7
20	16	20	17



Pearson correlation for the data in panel (a) (Set I in Table 3.4) of only 0.88 underrepresents the strength of the relationship, illustrating that this measure of correlation is not robust. The Pearson correlation for the data in panel (b) (Set II) is 0.61, reflecting the overwhelming influence of the single outlying point, and illustrating lack of resistance.

Rank correlation is 1.0 for Set I and 0.018 for Set II

Wilks 2011

References

- Wilks, 2011, “Statistical Methods in the Atmospheric Sciences”, Chapter 7
- Wilks, 2011, “Statistical Methods in the Atmospheric Sciences”, Section 3.5
- Whitaker, J. S., Wei, X., & Vitart, F. (2006). Improving Week-2 Forecasts with Multimodel Reforecast Ensembles, *Monthly Weather Review*, *134*(8), 2279-2284.