CORRESPONDENCE

Comments on "H. L. Wagner's Unbiased Hit Rate and the Assessment of Categorical Forecasting Accuracy"

IAN T. JOLLIFFE AND DAVID B. STEPHENSON

Exeter Climate Systems, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, United Kingdom

(Manuscript received 15 October 2013, in final form 10 February 2014)

1. Introduction

Many measures of forecast performance for binary deterministic forecasts have been devised and used in atmospheric science. For example, Table 3.3 of Hogan and Mason (2012) has a nonexhaustive list of 18 such measures. Table 3.4 in the same chapter gives properties of these measures, allowing users to make an informed choice of which to use or not to use.

Armistead (2013, hereafter A13) describes a measure, denoted H_u, from behavioral science, which is new to atmospheric science, and advocates its use for deterministic forecasts of multicategory events. Although A13 is to be commended for bringing in ideas from other disciplines, we consider that H_u has several undesirable properties that were not discussed in A13, which potential users should be aware of.

2. Undesirable properties of H_u

A13 concentrates on "binary studies" having only two categories (see Table 1), and so our comments are restricted to this special case. The measure $H_u = a^2/(a+b)(a+c)$, where a, b, and c are as defined in Table 1, was introduced by Wagner (1993) and is called "H. L. Wagner's unbiased hit rate" by A13. It can be interpreted as the product of the hit rate H or probability of detection (POD), $a/(a+c)$; and the frequency of hits (FOH) or success ratio SR, $a/(a+b)$. It can also be written $H_u = H^2/B$, where $B = (a+b)/(a+c)$ is the frequency bias. The measure H_u has several properties that may be undesirable for a user:

- First, H_u is proportional to the square of the "hit rate" H and, so, is often much smaller than H. The hit rate measures how often an occurrence of the event is successfully forecast and SR measures how often a forecast of the event is successful. The measure H_u combines these two measures by taking their product (i.e., $H_u = H \times \text{SR}$). If both H and SR are equal to 0.6, for example, then intuitively the measure of success is 60% whichever way the table is viewed. The product 0.36 seems an unduly pessimistic view of the level of success of the forecasts. The square root of H_u, which is the geometric mean of H and SR, would be preferable.
- Second, H_u can be hedged in some circumstances. In other words, it can be improved by forecasting something other than the forecaster's belief, which is often deemed to be undesirable (Jolliffe 2008). In particular, $c > d$ is a sufficient, though not necessary, condition for H_u to be improved by always forecasting the event to occur. This is similar to the problem identified for Finley's (1884) measure, proportion correct, $(a+d)/n$, which can be improved for his tornado forecasts by never forecasting a tornado.

  Another way to explore whether hedging can improve H_u is to investigate moving a vanishingly small proportion of the "no" row in the table to the "yes" row. In this case a sufficient, but not necessary, condition for improvement is $a/c < 2(a+b)/(c+d)$.
- It is not clear how to interpret H_u as a probability of a compound event. Wagner (1993, p. 16) states that H_u is "an estimate of the joint probability both that a stimulus is correctly identified (given that it is presented), and that a response is correctly used (given that it is used)." In the context of forecasting, this definition of H_u becomes, "an estimate of the joint probability both that an observed event is correctly

Corresponding author address: Prof. Ian Jolliffe, 30 Woodvale Rd., Gurnard, Cowes, Isle of Wight, PO31 8EG, United Kingdom. E-mail: i.t.jolliffe@exeter.ac.uk

DOI: 10.1175/WAF-D-14-00004.1

TABLE 1. Contingency table summarizing the counts recorded for a sample of n deterministic forecasts of a binary event.

Event forecast	Event observed		
	Yes	No	Total
Yes	a (hits)	b (false alarms)	$a + b$
No	c (misses)	d (correct rejections)	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = n$

identified (given that it occurred), and that a forecast of the event is correctly identified, given that the event is forecast.'' However, the joint probability of a hit, given that both of these conditioning outcomes have occurred, can easily be shown to be unity. Alternatively, interpreting the product of the two marginal probabilities as a joint probability would require assuming independence of ''event occurs'' and ''event forecast,'' which would only be true if the forecasts have no skill.

- Third, H_u does not provide a complete description of performance because in general there is more than 1 degree of freedom in the (2 × 2) contingency table. A13 claims that such tables have only 1 degree of freedom. If this were so, there is no need to calculate more than one performance measure, but it is argued by A13 that more than one measure is needed to summarize performance. We agree with this view and hence would prefer to report H and SR separately, rather than the single measure H_u.

  It is appropriate to have 1 degree of freedom if both margins of the table are fixed when collecting data. In the more realistic situation where forecasts are issued and assessed operationally, only the total number of observations n is fixed, so there are 3 degrees of freedom, and three measures are needed to fully describe the table. For example, Stephenson (2000) shows how various measures can be expressed as functions of the three quantities hit rate, false alarm rate $b/(b + d)$, and base rate $(a + c)/n$. In the case of assessing hindcasts it can be argued that the column totals in the (2 × 2) table are fixed, but there are still 2 degrees of freedom.
- The definitions of biased and unbiased are unconventional and somewhat unclear. The conventional definition of unbiased forecasts is that the same number of events is observed as is forecast, so $(a + b) = (a + c)$ and hence the frequency bias $B = 1$. Thus, $H_u = H^2/B$, so H_u is clearly sensitive to B. A13 talks of measures being biased or unbiased, but biased in this context means sensitivity to a form of bias in the forecasts. Wagner (1993) defines measures to be biased if they are sensitive to a different kind of bias, namely response bias, where the latter seems to mean that forecasts are not made equally often for the different forecast categories. Hence, Wagner's bias is a property of performance measures, whereas B is itself a performance measure. Wagner (1993, p. 17) claims that because H_u ''expresses accuracy as proportions of both [forecast] frequency and [observed] frequency it is insensitive to [such] bias.'' It is not clear to us why this is the case, nor why it is important.

For meaningful interpretation and usage of any performance measure, it is important to be aware of any undesirable properties. A13 is to be commended for bringing relevant ideas from behavioral science to the attention of atmospheric scientists, but fails to mention a number of such features, including the small values of H_u compared to what might be expected intuitively of a performance measure, its potential susceptibility to hedging, its lack of an obvious interpretation as the probability of a compound event, its incompleteness as a description of forecast performance, and its sensitivity to frequency bias. However, used carefully and in conjunction with other measures, it may help provide complementary feedback on forecast performance for some users.

REFERENCES

Armistead, T. W., 2013: H. L. Wagner's unbiased hit rate and the assessment of categorical forecasting accuracy. *Wea. Forecasting,* **28,** 802–814, doi:10.1175/WAF-D-12-00047.1.

Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.,* **1,** 85–88.

Hogan, R. J., and I. B. Mason, 2012: Deterministic forecasts of binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley-Blackwell, 31–59.

Jolliffe, I. T., 2008: The impenetrable hedge: A note on propriety, equitability and consistency. *Meteor. Appl.,* **15,** 25–29, doi:10.1002/met.60.

Stephenson, D. B., 2000: Use of the ''odds ratio'' for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232, doi:10.1175/1520-0434(2000)015<0221:UOTORF>2.0.CO;2.

Wagner, H. L., 1993: On measuring performance in categorical judgment studies of nonverbal behaviour. *J. Nonverbal Behav.,* **17,** 3–28, doi:10.1007/BF00987006.