Verification of Probabilistic Predictions: A Brief Review

Allan H. Murphy and Edward S. Epstein

The University of Michigan, Ann Arbor
(Revised manuscript received 21 May 1967)

Abstract

The evaluation process is considered in some detail with particular reference to probabilistic predictions. The process consists of several ordered steps at each of which elements (of the process) are identified. Consideration of the purposes leads to the identification of two distinct forms of evaluation: operational evaluation concerned with the value of predictions to the user and empirical evaluation, or verification, concerned with the perfection of predictions, i.e., the association between predictions and observations. Attributes, i.e., desirable properties, of predictions are defined with reference to these purposes, and a number of measures of the attributes for empirical evaluation are considered. An artificial example of comparative verification in which different measures appear to yield contradictory results is used to demonstrate the importance of, and need for, a careful analysis of the evaluation process.

1. Introduction

The evaluation or testing of models is a fundamental part of the (applied) research process (Ackoff, 1962). Ackoff states, in particular, that "in a very fundamental way every model is a predictive instrument" and, as a consequence, "testing the model as a whole consists of testing its ability to predict." Thus, the verification\(^1\) of predictions properly assumes an important role in the field of meteorology. Meteorologists have, of course, been concerned with the practice of forecast verification for many years (Meglis, 1960). Unfortunately, however, considerable controversy surrounds this practice in the meteorological literature (e.g., Brier and Allen, 1951). The controversy is, we believe, the result, in large measure, of the failure on the part of meteorologists to give proper consideration to the nature of the process of evaluation itself.

In this paper we shall consider the evaluation process in some detail, with particular reference to probabilistic predictions. However, this paper does not contain a comprehensive review of either forecast verification in general or the verification of probabilistic predictions. For the former the reader is referred to a paper by Johnson (1957). A review of the latter will be found in a forthcoming report by Murphy (in preparation).

Evaluation process. The evaluation process is, of necessity, an ordered process consisting of several distinct steps. At each step "elements" of the process are identified. First, the purposes for which evaluation is undertaken must be identified. Each purpose will, in turn, define a "form" of evaluation consisting of elements selected at each (subsequent) step. Second, attributes of the predictions must be identified and, subsequently, defined. The attributes represent "desirable" properties of the predictions for a particular purpose. Third, measures must be formulated which ascertain, in a quantitative manner, the extent to which predictions possess particular attributes. Fourth, tests (in a statistical sense) must be developed to permit inferences to be drawn from the results obtained by applying measures to predictions.

We shall be concerned here primarily with the first three steps in the evaluation process. The purposes of evaluation are considered in Section 2. In Section 3 several attributes of probabilistic predictions for a particular purpose are identified and defined. The measures proposed to date for these attributes are described in Section 4. Finally, in Section 5, two hypothetical collections of predictions are compared by means of the measures described in Section 4.

2. Purposes

A number of purposes of evaluation have been identified. Johnson (1957), for example, compiled the following representative list of purposes:

1) to measure the economic value of predictions,
2) to investigate the nature and cause of errors in prediction,
3) to assess the relative value of alternative techniques,
4) to measure the skill of individuals,
5) to perform a quality control service, and
6) to demonstrate the value of national weather services.
Now the distinctions between purposes 3) and 4) and between purposes 1) and 6) are slight, if real. Such distinctions would not, in any case, affect a “framework” for a particular form of evaluation. Thus, we have, in essence, the following four distinct purposes:

1) measurement of the absolute and/or relative “value” of predictions,
2) measurement of the absolute and/or relative “skill” of predictions,
3) performance of a quality control service, and
4) investigation of the nature and cause of errors in predictions.

Of these four purposes only the first has meaning for the decision maker, the user of the predictions. The three remaining purposes have meaning only for the meteorologist (who is concerned with the preparation and presentation of the predictions). This dichotomy of interest or purpose was recognized by Johnson who referred to the economic and scientific purposes for evaluation. We prefer to define and describe this dichotomy in a somewhat different manner.

A decision maker is, in our context, an individual in a meteorological “decision situation.” That is, the decision maker is an individual 1) whose activities or operations are affected to some extent by the “state of the atmosphere,” i.e., by meteorological conditions, and 2) who must select a course of action from a set of available courses of action. The framework for decision making to which we subscribe is a modified version of the personal expected-utility statistical model of the decision situation (Fishburn, 1964). The essential features of this model may be described briefly as follows.

The decision maker’s objective is to select the “best” course of action. His preferences among the consequences, which result when he selects a course of action and a state of the atmosphere obtains, are expressed in terms of the “operational value” or “utility” of the consequences to him. The decision maker is, of course, uncertain prior to the selection of an action, which state will obtain. We assume that he accepts as his personal probabilities [in the sense of Savage (1954)] the meteorologist’s probabilistic prediction defined on the states. Then the decision maker selects that action for which the expected-utility is a maximum. Thus, utility [in the decision-theoretic sense of von Neumann and Morgenstern (1953)] is the attribute which is indicative of the “value” of predictions to the decision maker. This form of evaluation, i.e., the process of assessing the utility of predictions, is referred to, in this paper, as operational evaluation.

Now, in this context, operational evaluation appears to be relatively straightforward, since only a single attribute, utility, is of concern. However, we must recognize that different decision makers will (even in similar situations) possess different sets of utilities. Utility has meaning only with reference to an individual with an objective in a decision situation (Fishburn, 1964). Thus, a universal measure for operational evaluation does not exist. However, the situation is not impossible since methods can be devised which, at least in part, circumvent this difficulty (Murphy, 1966). Meteorologists should recognize this situation for what it is and realize that they must pay more attention to the nature and properties of decision makers’ utility functions.

Now, consider the point of view of the meteorologist whose immediate objective is, simply, to formulate a prediction system which prepares predictions possessing, in some sense, greater “perfection.” The meteorologist’s ultimate objective is, of course, to formulate a prediction system which prepares perfect predictions. Thus, perfection is the general attribute of the predictions indicative of the “value” of the parent prediction system to the meteorologist. We refer to this form of evaluation, the process of assessing the absolute and/or relative perfection of predictions, as empirical evaluation or verification.

In this paper we shall be concerned with the identification and definition of attributes and with the description of measures for these attributes for empirical evaluation only. The reader interested in operational evaluation is referred to papers by Bagrov (1966), Borgman (1960), Epstein (1962), Glahn (1964), Gringorten (1959), Milly and Bryan (1965), Monin (1962), Murphy (1966), Nelson and Winter (1964), Ogawara (1955), Suzuki (1960), Thompson (1952) and Thompson and Brier (1955) among others.4

3. Attributes

From the point of view of empirical evaluation, what specific properties of probabilistic predictions are “desirable”? The general attribute of concern is, as indicated in Section 2, their perfection. Now a prediction is perfect, in our context, if the prediction assigns probability one to the state of the atmosphere which subsequently obtains. Clearly, in order to identify and define attributes, the elements of the “framework” for empirical evaluation, namely the states of the atmosphere, the predictions and the observations, must be precisely defined.

Suppose the range of the predictand, the individual or composite meteorological element to be predicted, is divided into $G$ mutually exclusive and exhaustive groups, the states of the atmosphere. Then a prediction $p$ is a set of probabilities $(p_1, \ldots, p_G)$, where $p_g$ is the (meteorologist’s) probability that state $g$ obtains and, of course, $p_g \geq 0$ and $\sum p_g = 1$ ($g = 1, \ldots, G$). An observation, on the other hand, is a set $\delta = (\delta_1, \ldots, \delta_G)$, where $\delta_g$ equals one if state $g$ obtains and zero otherwise ($g = 1, \ldots, G$).

4 A comprehensive bibliography on all aspects of probabilistic prediction in meteorology is available on request from the authors of this paper.
Now, note that the sets \( p \) and \( \delta \) are equivalent for a perfect prediction. Thus, in effect, the meteorologist would like the sets \( p \) and \( \delta \) to be as similar, i.e., as closely associated, as possible. Now the notion of “association” appears to possess two particularly relevant definitions which lead to the identification of two distinct attributes—an attribute based on the notion of the association between predictions and observations on an individual basis, and an attribute based on the notion of the association between predictions and observations on a collective basis.

\[ a. \text{ Validity.} \quad \text{The association between an individual prediction } p \text{ and the relevant observation } \delta \text{ is complete if and only if} \]
\[ p^o = \delta^o \text{ for all } g, \quad (g = 1, \ldots, G), \]
\[ \text{or} \]
\[ p^o - \delta^o = 0 \text{ for all } g, \quad (g = 1, \ldots, G), \]
\[ \text{i.e., if and only if the values assumed by the respective elements of the sets } p \text{ and } \delta \text{ are identical. We will refer to such a prediction as a completely valid prediction. Conversely, a prediction for which} \]
\[ p^o - \delta^o = \pm 1 \text{ for any } g, \quad (g = 1, \ldots, G), \]
\[ \text{will be referred to as a completely invalid prediction. In general, a prediction is neither completely valid nor completely invalid, i.e., in general} \]
\[ 0 < |p^o - \delta^o| < 1 \text{ for all } g, \quad (g = 1, \ldots, G). \]

A prediction for which (1) holds will be referred to as a partially valid prediction. Thus, a “desirable” property of probabilistic predictions (for empirical evaluation) is validity, an attribute based on the notion of the association between predictions and observations on an individual basis.

\[ b. \text{ Bias.} \quad \text{Now, consider the association between predictions and observations on a collective basis. As previously indicated, the association between the predictions in a collection and the relevant observations, on an individual basis, is, in general, incomplete. However, the definition of the term “association” for a collection of predictions leads to the identification of another “desirable” property of predictions, an attribute which indicates that the association between predictions and observations, while incomplete on an individual basis, may be complete on a collective basis. Several “properties” and a number of “statistics” for each property may, of course, be defined for a collection, or a sub-collection of a collection, of predictions. However, only the mean, a familiar member of the class of statistics concerned with “location,” is considered in this paper. The selection of the mean seems particularly appropriate in view of its simplicity and the nature of the elements (of the framework) of concern. This leads to the identification of an attribute which appears to possess a “reasonable” interpretation. Thus, the association between predictions and observations, on a collective basis, is defined in terms of the association of mean values for the predictions in a collection (or a subcollection) and the relevant observations. Consideration of some of the details of the association between predictions and observations on a collective basis has led us to define two specific attributes which are, in reality, simply two aspects of the familiar concept of bias.} \]

Consider a collection of \( N \) predictions \( p_n \), where \( p_n = (p_1, \ldots, p_{\gamma_n}) \) and \( p_n^o \) is the probability that state \( g \) obtains on occasion \( n \), and the \( N \) relevant observations \( \delta_n \), where \( \delta_n = (\delta_1, \ldots, \delta_{\gamma_n}) \) and \( \delta_{n\gamma} \) equals one if state \( g \) obtains on occasion \( n \) and zero otherwise \((g = 1, \ldots, G; n = 1, \ldots, N)\). Although the individual predictions in the collection are, in general, partially valid, the association between predictions and observations may be considered to be complete, on a collective basis, for state \( g (g = 1, \ldots, G) \), if

\[ p^o_n = \delta^o_n, \]
\[ \text{or} \]
\[ p^o_n - \delta^o_n = 0, \quad \text{(2)} \]
\[ \text{where} \]
\[ p^o_n = \frac{1}{N} \sum_n p_{n\gamma}, \quad (n = 1, \ldots, N), \]
\[ \text{and} \]
\[ \delta^o_n = \frac{1}{N} \sum_n \delta_{n\gamma}, \quad (n = 1, \ldots, N), \]
\[ \text{i.e., if the average predicted probability and the observed relative frequency, or sample-of-N climatology, are equal for that state. A collection of predictions for which (2) holds will be referred to, in this paper, as an unbiased collection of predictions in-the-large for state } g (g = 1, \ldots, G). \text{ A collection of predictions for which (2) does not hold will be referred to as a biased collection in-the-large for state } g (g = 1, \ldots, G). \text{ A collection of predictions for which (2) holds for all } g (g = 1, \ldots, G) \text{ will be referred to as a completely unbiased collection in-the-large.} \]

The phrase “in-the-large” appears in the previous paragraph to distinguish that aspect of bias from the aspect to be identified in this paragraph. Suppose that the probabilities which constitute the predictions can assume only \( F \) distinct values. Then, the number of distinct predictions \( R \) is

\[ R = \sum_{f} \binom{F+G-4}{f-1} (F-f+1), \quad (f = 1, \ldots, F). \]

Now, consider the sub-collection, from the collection of \( N \) predictions, of the \( N_f \) predictions for which

\[ p_n = p^o, \]
\[ \text{or} \]
\[ (p_{1n}, \ldots, p_{\gamma_n}) = (p^o_1, \ldots, p^o_{\gamma_n}), \]
where \( p_{g} \) is the (constant) probability assigned to state \( g \) when the prediction is \( p' \) (\( g = 1, \ldots, G; n = 1, \ldots, N; r = 1, \ldots, K \)). The association between the predictions in such a sub-collection and the relevant observations may be considered to be complete if for that sub-collection and state \( g \) (\( g = 1, \ldots, G \)),

\[
p_{g} = \delta_{g},
\]
or

\[
p_{g} - \delta_{g} = 0,
\]

where

\[
\delta_{g} = \frac{1}{N} \sum_{n} \delta_{g n}, \quad (n \in N),
\]

i.e., if the (constant) probability \( p_{g} \) and the observed relative frequency \( \delta_{g} \) are equal. A sub-collection of predictions for which (3) holds will be referred to as an unbiased sub-collection in-the-small for state \( g \) (\( g = 1, \ldots, G \)). A sub-collection of predictions for which (3) does not hold will be referred to as a biased sub-collection in-the-small for state \( g \) (\( g = 1, \ldots, G \)). A sub-collection of predictions for which (3) holds for all \( g \) (\( g = 1, \ldots, G \)) will be referred to as a completely unbiased sub-collection in-the-small.

c. 'Validity' and 'sharpness'. Bross (1953) identified two "standards" (attributes) for probabilistic predictions, namely 'validity' and 'sharpness.' The attribute 'validity' as defined by Bross is, in essence, equivalent to the attribute bias-in-the-small as defined in this paper.

A prediction is 'sharp' according to Bross if, on the other hand, the prediction 'discriminates' among the relevant states. The 'sharpness' of a prediction is related to the "information" (Shannon and Weaver, 1949) contained in the probabilities which constitute the prediction. A prediction is considered to be completely 'sharp' if the probability \( p_{g} \) (\( g = 1, \ldots, G \)) equals zero or one. Now, an attribute based on the notion of information (alone) is concerned exclusively with the behavior of the prediction, i.e., the observation is not taken into consideration. Thus, 'sharpness' is not an attribute of proper concern for empirical evaluation, the purpose of which is to assess the absolute and/or relative association between predictions and observations. However, a measure of the attribute 'sharpness' will be described in the next section.

d. 'Reliability' and 'resolution'. Sanders (1958, 1963) partitioned the probability score (Brier, 1950) into two complementary scores, each of which was considered to be a measure of a different 'aspect' (attribute) of the predictions. Sanders referred to these attributes initially (1958) as 'reliability' and 'resolution,' but subsequently (1963) he adopted Bross's terms 'validity' and 'sharpness,' respectively. The attribute 'reliability' as defined by Sanders is, in essence, equivalent to the attribute 'validity' as defined by Bross and the attribute bias-in-the-small as defined in this paper. However, the attribute 'resolution' as defined by Sanders is not equivalent to the attribute 'sharpness' as defined by Fross, for the measure of 'resolution' proposed by Sanders is concerned with the behavior, for a collection of predictions, of sub-collections of observations, the definition of which depends upon the predictions (refer to Section 3), while, as noted previously, the attribute 'sharpness' is concerned exclusively with the behavior of predictions. Although the attribute 'resolution' is concerned with both predictions and observations, the former enter in an indirect manner. Thus, the attribute 'resolution' is not considered an attribute of proper concern for empirical evaluation.

4. Measures

The following measures are described, in brief, in this section:

1) probability score (Brier, 1950),
2) 'reliability' and 'resolution' scores (Sanders, 1958),
3) information quantity (Shannon and Weaver, 1949; Bross, 1953),
4) information ratio (Holloway and Woodbury, 1955),
5) 'validity' measure (Miller, 1962),
6) 'skill' score (Gringorten, 1965),
7) distance measures (Epstein and Murphy, 1965).

Some of the measures have been formulated with respect to specific attributes, while other measures have been formulated without particular reference to any attribute. We will consider, in particular, the relevance of these measures for empirical evaluation.

Each measure is defined for a collection of \( N \) predictions.

a. Probability score. The probability score \( PS \) is defined as

\[
PS = \frac{1}{N} \sum_{n} \sum_{g} (p_{g n} - \delta_{g n})^{2}, \quad (g = 1, \ldots, G; n = 1, \ldots, N).
\]

Note that

\[
0 \leq PS \leq 2.
\]

\( PS \) equals zero if each prediction is categorical and correct, i.e., if the probability assigned to the state that obtains on each occasion is one, while \( PS \) equals two if each prediction is categorical and incorrect, i.e., if the probability assigned to a state that does not obtain on each occasion is one. Thus, the probability score is a measure of the validity of predictions.\(^6\) The smaller \( PS \) the greater the validity of predictions.

---

\(^6\) Epstein and Murphy (1965) have shown that the probability score is proportional to the mean square distance, in a geometric framework, between points representing individual predictions and observations (refer to the description of the distance measures in this section).
b. 'Reliability' and 'resolution' scores. A modified version of the probability score has been partitioned into two separate scores, namely a 'reliability' score and a 'resolution' score, by Sanders (1958). Sanders considers each probability a distinct prediction. Thus, for a collection of \( M (= N G) \) predictions, Sanders' probability score \( PS \) is defined as

\[
PS = \frac{1}{M} \sum_{m=1}^{M} (p_m - \delta_m)^2, \quad (m=1, \cdots, M),
\]

where \( p_m \) is the probability assigned to the "predictand" on "occasion" \( m \), and \( \delta_m \) is one if the predictand obtains on occasion \( m \) and zero otherwise. Note that

\[ 0 \leq PS \leq 1. \]

Sanders then considers that the probability \( p_m \) \((m=1, \cdots, M)\) assumes only \( F \) distinct values (refer to Section 3). Thus, the collection of \( M \) predictions is divided into \( F \) sub-collections, where \( M_f \) is the number of predictions in the collection for which \( p_m = p_f \) (a constant) \((f=1, \cdots, F; m=1, \cdots, M)\). Then, Sanders shows that \( PS \) can be expressed as

\[
PS = \frac{1}{M} \sum_{f=1}^{F} M_f (p_f - \bar{\delta}_f)^2 + \frac{1}{M} \sum_{f=1}^{F} M_f \bar{\delta}_f (1 - \bar{\delta}_f),
\]

where

\[
\bar{\delta}_f = \frac{1}{M_f} \sum_{m=1}^{M_f} \delta_m, \quad (m \in M_f).
\]

Sanders (1958) referred to the first and second terms on the right-hand side of (4) as measures of the 'reliability' and 'resolution,' respectively, of the predictions, i.e.,

\[
PS_{rel} = \frac{1}{M} \sum_{f=1}^{F} M_f (p_f - \bar{\delta}_f)^2, \quad (f=1, \cdots, F),
\]

and

\[
PS_{res} = \frac{1}{M} \sum_{f=1}^{F} M_f \bar{\delta}_f (1 - \bar{\delta}_f), \quad (f=1, \cdots, F).
\]

Note that

\[ 0 \leq PS_{rel} \leq 1, \]

and

\[ 0 \leq PS_{res} \leq \frac{1}{M}. \]

\( PS_{rel} \) equals zero if and only if \( p_f = \bar{\delta}_f \) for all \( f \), while \( PS_{res} \) equals zero if \( \bar{\delta}_f \) equals zero or one for all \( f (f=1, \cdots, F) \).

\( PS_{rel} \) is a measure of the attribute bias in-the-small, as is evident from a comparison of (3) and (5). As indicated previously, \( PS_{res} \) cannot be considered to measure directly a "desirable" property of predictions for empirical evaluation. However, since Sanders has shown that \( PS_{res} \) represents the difference between a measure of validity \( (PS_v) \) and a measure of bias in-the-small \( [PS_v(RET)] \), \( PS_{res} \) may be considered a "measure" of the extent to which a collection of predictions can be unbiased in-the-small and yet be only partially valid.

c. Information quantity. The information quantity \( I \), as a verification measure, is defined as

\[
I = -\frac{1}{N} \sum_{n=1}^{N} \sum_{g=1}^{G} p_{gn} \ln p_{gn}, \quad (g=1, \cdots, G; n=1, \cdots, N).
\]

Note that \[ 0 \leq I \leq \ln G. \]

\( I \) equals zero if each prediction is categorical and \( I \) equals \( \ln G \) if \( p_{gn} = 1/G \) for all \( g \) and \( n \) \((g=1, \cdots, G; n=1, \cdots, N)\). The information quantity is a measure of the 'sharpness' of predictions. The smaller \( I \) the 'sharper' the predictions.

d. Information ratio. The information ratio \( IR \) is defined as

\[
IR = \left[ \sum_{n=1}^{N} \ln \left( \frac{p_{n}\ast n}{p_{n}\ast n} \right) \right] \left[ \sum_{n=1}^{N} \ln \left( \frac{1}{p_{n}\ast n} \right) \right],
\]

\((n=1, \cdots, N), \)

where \( p_{n}\ast n \) is the probability assigned to the state that obtains on occasion \( n \) and \( p_{n}\ast n \) is the climatological probability of the state that obtains on occasion \( n \). Note that\[ -\infty < IR \leq 1. \]

\( IR \) "equals" minus infinity if \( p_{n}\ast n = 0 \) on any occasion; \( IR \) equals zero if \( p_{n}\ast n = p_{n}\ast n \) on all occasions; and \( IR \) equals one if \( p_{n}\ast n = 1 \) on all occasions. The information ratio can be shown to be a relative measure of "partial" validity, i.e., \( IR \) measures the validity of the probability assigned to the state that obtains relative to the validity of the relevant climatological probability. For a particular set of climatological probabilities, the larger \( IR \) the greater the partial validity of the predictions.

e. 'Validity' measure. The 'validity' measure \( VM \) is defined as

\[
VM = \sum_{g=1}^{G} \sum_{g'=1}^{G-1} \rho_{g\bar{g}'} U_{g\bar{g}'} U_{g'\bar{g}}, \quad (g, g' = 1, \cdots, G-1),
\]

where \( \rho_{g\bar{g}'} \) is the element in row \( g \) and column \( g' \) of the matrix \( \rho^{-1} \), where \[ \theta = (\rho_{g\bar{g}'}) \]

and

\[
\sum_{g=1}^{G} \sum_{g'=1}^{G-1} \rho_{g\bar{g}'} \left\{ \left[ \sum_{n=1}^{N} p_{n}(1-p_{n}) \right] \left[ \sum_{n=1}^{N} p_{n}(1-p_{n}) \right] \right\}^{1}, \quad (g \neq g'),
\]

\[
1, \quad (g = g'),
\]

\((g, g' = 1, \cdots, G-1; n = 1, \cdots, N).\]

\footnote{Under the assumption that 0<\( p_{n}\ast n < 1 \), \( (g^* = 1, \cdots, G) \), \( IR \) does not exceed one.}
Further,

\[ U_{o'} = \frac{U_o - E(U_o)}{[V(U_o)]^{1/2}}, \quad (g = 1, \ldots, G), \]

where

\[ U_o = \sum_{n} \delta_{on}, \quad (n = 1, \ldots, N), \]

\[ E(U_o) = \sum_{n} p_{on}, \quad (n = 1, \ldots, N), \]

\[ V(U_o) = \sum_{n} p_{on}(1 - p_{on}), \quad (n = 1, \ldots, N). \]

\[ U_{o'}, E(U_{o'}) \text{ and } V(U_{o'}) \text{ possess, of course, similar definitions. Note that} \]

\[ VM \geq 0. \]

The ‘validity’ measure is a measure of the attribute bias in-the-large. The smaller \( VM \) the less the bias in-the-large of the predictions.

\( f. \) ‘Skill’ score. The ‘skill’ score \( SS \) is defined as

\[ SS = \frac{1}{NG} \sum_{n} \sum_{g} SS_{gn}, \quad (g = 1, \ldots, G; n = 1, \ldots, N), \]

where

\[ SS_{gn} = \begin{cases} 
1/p_{og} \quad & \text{if } p_{gn} > p_{og} \text{ and } \delta_{gn} = 1 \\
1/(1-p_{og}) \quad & \text{if } p_{gn} < p_{og} \text{ and } \delta_{gn} = 0 \\
0 \quad & \text{otherwise}
\end{cases} \]

Note that\(^8\)

\[ SS \geq 0. \]

The ‘skill’ score is a relative measure of the attribute validity. For a particular set of climatological probabilities, the greater \( SS \) the greater the validity of the predictions relative to the validity of the climatological probabilities. However, note that \( SS_{gn} \), the ‘skill’ score on occasion \( n \), assumes, at most, \( G(G^{-1}G+1)+1 \) different values \( (n=1, \ldots, N) \). Thus, \( SS \) is not continuous (in the mathematical sense) and, as such, is not a particularly appropriate measure for empirical evaluation.

\( g. \) Distance measures. The distance measures are a class of measures defined in the context of a geometric framework in which (Euclidean) distances between predictions and observations are natural measures for empirical evaluation.

The distance measure for the attribute validity \( v \) is defined\(^9\) as

\[ v = \frac{1}{N} \sum_{n} v_{n}, \quad (n = 1, \ldots, N), \]

where

\[ v_{n} = \left[ -\frac{G-1}{G} \sum_{g} \sum_{g'} (p_{gn} - \delta_{gn})(p_{g'n} - \delta_{g'n}) \right]^{1/2}, \quad (g, g' = 1, \ldots, G; g \neq g'). \]

Note that

\[ 0 \leq v \leq \left( \frac{G-1}{G} \right)^{1/2}. \]

Thus, \( v \) equals zero if each prediction is completely valid, i.e., categorical and correct, and \( v \) equals \( \left( (G-1)/G \right)^{1/2} \) if each prediction is completely invalid, i.e., categorical and incorrect.

The distance measure for the attribute bias in-the-large \( b_L \) is defined as

\[ b_L = \left[ -\frac{G-1}{G} \sum_{g} \sum_{g'} (\tilde{p}_{og} - \tilde{\delta}_{og})(\tilde{p}_{g'o} - \tilde{\delta}_{g'o}) \right]^{1/2}, \quad (g, g' = 1, \ldots, G; g \neq g'), \]

(refer to Section 3 for the definitions of \( \tilde{p}_{og} \) and \( \tilde{\delta}_{og} \)). Note that

\[ 0 \leq b_L \leq \left( \frac{G-1}{G} \right)^{1/2}. \]

Thus, \( b_L \) equals zero if the collection of predictions is completely unbiased in-the-large.

The distance measure for the attribute bias in-the-small \( b_S \) is defined as

\[ b_S = \frac{1}{N} \sum_{n} \sum_{r} \left[ -\frac{G-1}{G} \sum_{g} \sum_{g'} (p_{gn} - \delta_{gn})(p_{g'n} - \delta_{g'n}) \right]^{1/2}, \quad (g, g' = 1, \ldots, G; g \neq g'; r = 1, \ldots, R) \]

(refer to Section 3 for the definitions of \( N, p_{gn} \text{ and } \delta_{gn} \)). Note that

\[ 0 \leq b_S \leq \left( \frac{G-1}{G} \right)^{1/2}. \]

Thus, \( b_S \) equals zero when the collection of predictions is completely unbiased in-the-small.

The distance measures can, of course, be defined on the unit interval for a particular predictand, i.e., for a particular value of \( G \), simply by dividing the measures, as defined, by \( \left( (G-1)/G \right)^{1/2} \).

5. Verification: an illustrative example

To illustrate the importance of the identification and definition, and subsequent careful consideration, of the elements of the evaluation process, we will consider the following example.

Consider a predictand, the range of which has been divided into an odd number (greater than or equal to
five) of states, and two probability prediction systems, Alpha and Beta, each of which prepares, on a number of occasions, unimodal, symmetric predictions about the "median" state. We will consider two very special collections of predictions. Suppose that each prediction system assigns the same probability to the median state, and that this state is subsequently observed, on each occasion. Further, suppose that, on each occasion, the predictions prepared by system Alpha are at least as "concentrated" as those prepared by system Beta, i.e., suppose that, on each occasion, the sum of the probabilities assigned by system Alpha to the states in any symmetric interval about the median state is equal to or greater than the sum of the probabilities assigned to these states by system Beta.

Now, we ask the question, which prediction system is "better"? To assist us to answer this question we shall use the measures described in Section 4 to compare the two collections of predictions. The results of this "comparative verification" are indicated in Table 1.

At first glance the results of the comparative verification appear confusing. There is an obvious lack of agreement among the measures as to which prediction system is "better." However, if we recall that the measures are concerned with different attributes of the predictions, then, at least in this situation, much of the confusion disappears. The probability score and the distance measure \( v \), both measures of the attribute validity, indicate that Beta is the "better" prediction system. The information ratio, a relative measure of partial validity, indicates that the prediction systems are equally "good." The distance measure \( b_L \), a measure of the attribute bias in-the-large, indicates that prediction system Beta is "better." The information quantity, on the other hand, a measure of the attribute "sharpness" (actually not an attribute of proper concern for empirical evaluation), indicates that prediction system Alpha is "better." However, the results (in this example) are clearly not contradictory, since measures of different attributes should not necessarily be expected to yield the same results.

This example illustrates the importance of the identification, and subsequent consideration, of the elements of the evaluation process. Of course, even the most elaborate framework cannot be expected to eliminate all the "problems" in such a process. In our framework for empirical evaluation, for example, the meteorologist must still select the attributes, and subsequently the measures, which he considers "appropriate."

6. Summary and conclusions

We have considered two purposes for the evaluation of predictions: 1) the assessment of their absolute and/or relative "skill," or "value" to the meteorologist, and 2) the assessment of their absolute and/or relative "utility," or "value" to the decision maker. These purposes, in turn, define two forms of the evaluation process, which we have referred to as empirical evaluation, or verification, and operational evaluation, respectively.

For operational evaluation, considered very briefly here, only a single attribute, or "desirable" property, of the predictions is of concern, their utility. The "value" of a prediction is measured in terms of the utility, to the decision maker, of the consequence that results. However, since the measure of the attribute utility will be different for different decision makers and for different decision situations, a universal measure for operational evaluation does not exist.

For empirical evaluation several attributes, and measures for these attributes, were described. The relevant attributes and measures are indicated in Table 2.

The measures we have described which do not appear to fit into our framework for empirical evaluation need not necessarily be discarded. These measures simply do not appear to be relevant for our attributes. In particular, they (i.e., the information quantity and the "resolution" score) seem to be appropriate for attributes relative to another purpose, a purpose concerned with the internal characteristics of prediction systems which are, in some sense, independent of the association between predictions and observations. However, we have chosen not to investigate this problem further.

The example considered served to illustrate our contention that evaluation becomes most difficult when one tries to make it appear simple. The problem of evaluation, as a whole, will become more tractable and less confusing when we have obtained general agreement on the need for precise definition of the purposes for evaluation and the attributes to be measured for each purpose. As the example indicates, there need be no

---

Table 1. Comparative verification of prediction systems Alpha and Beta.

<table>
<thead>
<tr>
<th>Measure</th>
<th>&quot;Better&quot; prediction system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability score</td>
<td>Beta</td>
</tr>
<tr>
<td>Information quantity</td>
<td>Alpha</td>
</tr>
<tr>
<td>Information ratio</td>
<td>Alpha and Beta equally &quot;good&quot;</td>
</tr>
<tr>
<td>Distance measure ( v )</td>
<td>Beta</td>
</tr>
<tr>
<td>Distance measure ( b_L )</td>
<td>Beta</td>
</tr>
</tbody>
</table>

*Without additional knowledge, e.g., knowledge of the probabilities assigned to each state on each occasion and/or knowledge of the relevant climatological probabilities, the relative values of certain measures described in Section 4 cannot be specified.

Table 2. Attributes and measures for empirical evaluation.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>( PS, (IR,SS), * v )</td>
</tr>
<tr>
<td>Bias in-the-large</td>
<td>( VM, b_L )</td>
</tr>
<tr>
<td>Bias in-the-small</td>
<td>( PS, (REL), b_S )</td>
</tr>
</tbody>
</table>

*The measures IR and SS are both relative measures of the attribute validity. Furthermore, the former is only a partial measure of validity, while the latter is not continuous (in the mathematical sense).
such thing as a “best” collection of predictions (even for a particular purpose), for “best” has meaning only with reference to a specific attribute.

We could, of course, have carried this line of reasoning further. Not only can measures appropriate to different attributes (for a particular purpose) yield “conflicting” results, but different measures of the same attribute may also yield such results. For example, the measures $PS_s$(REL) and $b_s$, both measures of the attribute bias in-the-small, are distinct, and thus could yield different results in a comparative verification. This situation is similar to that in operational evaluation (in which utility is the attribute of concern) where, although the proper measure in a particular situation is unequivocal, no unique universal measure exists. In the case of those attributes appropriate to empirical evaluation, the choice of a measure is not unequivocal, it is always arbitrary. Thus, let us not persist in arguments concerning which measure is “best,” when the important considerations are the attribute it measures and the reason for making the measurement.

REFERENCES


