GEOFIZIKA VOL. 23 NO. 1 2006

Original scientific paper
UDC 551.501.45

Application of tetrachoric and polychoric correlation coefficients to forecast verification

Josip Juras and Zoran Pasarić

Department of Geophysics, Faculty of Science, University of Zagreb, Zagreb, Croatia

Received 4 October 2005, in final form 4 May 2006

The measure of association in 2×2 ($K \times K$) contingency tables known as tetrachoric (polychoric) correlation coefficient is recalled. These measures rely on two assumptions: 1) there exist continuous latent variables underlying the contingency table and 2) joint distribution of corresponding standard normal deviates is bivariate normal. It is shown that, in practice, the tetrachoric (polychoric) correlation coefficient is an estimate of Pearson correlation coefficient between the latent variables. Consequently, these measures do not depend on bias nor on marginal frequencies of the table, which implies a natural and convenient partition of information (carried by the contingency table), between association, bias and probability of the event and subsequently enables the analysis of how other scores depend on bias and marginal frequencies. Results extended to $K \times K$ tables lead to eventual reduction in dimensionality from K^2 to $2K$. The theoretical findings are illustrated through analysis of real-life, 6×6 contingency tables on verification of quantitative precipitation forecasts.

Keywords: tetrachoric correlation coefficient, contingency table, forecast evaluation

1. Introduction

The history of applying contingency tables to forecast verification, given in detail by Daan (1984) and Murphy (1996), is a rather long one. Nevertheless, during 1990s contingency tables became focal point of several papers (Gandin and Murphy, 1992; Barnston, 1992; Gerrity, 1992; Marzban, 1998). Even though they are most naturally used in verification of forecasts of meteorological events such as severe weather and occurrence of precipitation, in practice contingency tables are even more frequently used to describe the accuracy of forecasts of meteorological elements that are continuous by nature, such as temperature, wind speed, visibility, etc. Moreover, they are also used in forecast verification of meteorological fields (Ward and Folland, 1991; Potts et al., 1996), due to the fact that through contingency tables the pro-

perties of a set of forecasts can be condensed and very clearly displayed, whereas by using the correlation coefficient or the mean square error these properties often remain hidden. On the other hand, although multidimensional means such as scatter diagrams or residual plots can be employed for assessing performance, there are certain situations in which we need single values by which one can continuously monitor development of forecasting methods. As a consequence, a number of scores defined as certain functions of contingency table elements have been introduced into meteorological practice. These scores focus each on certain facet of forecast quality often exhibiting some undesired properties. Here we have: 1) dependence on the probability of the event, which prevents comparison of forecasts between regions with different climate, 2) dependence on bias which may lead to »hedging« in order to improve particular score and 3) dependence on the number of categories, *i.e.* on the dimension of contingency table, which prevents the comparison between tables of different dimensions. In attempt to avoid some of these difficulties as well as to satisfy needs of specific users, over twenty measures for a simple 2×2 contingency table have been proposed. This has resulted in a considerable »confusion« and has certainly obscured the primary goal to continuously monitor the progress in weather forecasting.

In the present paper we recall tetrachoric and polychoric correlation coefficients (TCC and PCC) as measures of association in 2×2 and $K \times K$ contingency tables, respectively. The tetrachoric correlation is proposed by Pearson (1900) as a measure of association between two be-categorical variables. A short history of polychoric correlation is given in Olsson (1979). The essential assumption to be made, the mild one, is that the two variates that are ordered categorical variables (observation and forecast) have come from dichotomizing or polychotomizing underlying continuous variables. These continuous variables are sometimes called latent variables, since they are observed only through category frequencies and not directly. Another assumption, the strong one, requires that latent variables follow, at least approximately, the bivariate normal distribution (BND). Although both coefficients have been widely used in social sciences, it is probably this second assumption that, together with computational difficulties, had prevented wider use of tetrachoric and polychoric coefficients in meteorological practice. However, the BND is not applied to the latent variables directly, but to corresponding standard normal deviates (SND-s) and, as we show, in many cases it can be used to obtain estimate of classical Pearson correlation coefficient between the latent variables. Moreover, it may be argued, as Pearson and Heron (1913) did, that some distribution should be supposed since by the very nature of the problem we do not have any direct information on latent variables. In such a case the BND with its numerous properties is the most natural choice. Gringorten (1971, 1972) first pointed out the possibility of using BND for the formulation of forecasts based on auto regression, which could serve as a reference level in the forecast verification. Juras (1982) used this approach to estimate

probabilities of simultaneous occurrence of weather events at different locations.

Subsequently, we recall a number of nice properties that make the TCC an appealing measure of association. Although these properties do not depend neither on BND nor on latent-variables assumption, they are most easily verified through the BND. However, the latent variables framework implies simple and natural partition of information contained in 2×2 table, between correlation coefficient, bias and one marginal frequency (*e.g.* climatological frequency of the event), with straightforward extension to higher-order tables. Also it suggests that TCC and PCC should be less prone to the three above mentioned deficiencies of scores.

The BND framework may be used to analyze properties of various scores. Since 2×2 table possesses 3 degrees of freedom, one piece of information has to be fixed in order to present results clearly. Here we fix the TCC and analyze four common scores that are used in meteorology, as functions of bias and one marginal frequency. Similar investigations have been done by Barnston (1992) and Potts et al. (1996) using the Monte Carlo method. All analyzed scores are subsequently applied to and compared on 2×2 and 6×6 contingency tables taken from real word.

The conclusions emphasize the need for continuous and detailed monitoring of categorical forecasts, and subsequent evaluation of the scores used.

Table 1. Definition of elements of a contingency table.

		Observation		
		Yes	No	
Forecast	Yes	a	b	P_F
	No	c	d	$1 - P_F$
		P_O	$1 - P_O$	

2. Tetrachoric and polychoric correlation coefficients

2.1. Notation and assumptions: Latent-variables model of forecasting process

In what follows, if not stated otherwise, frequency always stands for the relative frequency. We consider a 2×2 verification problem and the corresponding contingency table of relative frequencies (Table 1). Here a is the relative frequency by which an event is observed and forecasted, $P_O = a + c$ is the frequency by which the event is observed, while $P_F = a + b$ is the frequency by which it is forecasted. P_O and P_F are called the marginal frequencies and a the joint frequency of the contingency table. Obviously, the table is completely

determined by the three frequencies (probabilities), a, P_O, P_F. The bias is denoted by $B = P_F / P_O$.

Having only the contingency table it is implicitly accepted that we are dealing with categorical variables, which in our case are observation (O) and forecast (F). In order to measure a degree of association between them we assume that frequencies of categories are obtained by dichotomizing certain continuous variables, X_O and X_F, underlying the categorical variables O and F. Variables X_O and X_F are called latent variables since, for practical reasons, they are not observed directly but only through their categorical counterparts. The extension of the latent variables framework to $K \times K$ tables for ordered categorical variables is obvious, the only difference being an increased number of thresholds.

One may wonder what are the continuous variables, X_O and X_F, behind a 2×2 contingency table or the corresponding verification problem. On the observation side it might be a particular meteorological element *e.g.* wind speed, precipitation, cloudiness, etc. We say the event is observed if the observed value X_O exceeds some previously fixed threshold value x_O. On the forecast side, the supposed continuous variable might be the judgment probability assessed by the forecaster (subjective forecasting), or it may be the value of the meteorological element itself, if such a value is provided by the forecasting system, *e.g.* numerical model (objective forecasting). The event is forecasted if $X_F > x_F$ for some decision threshold x_F. There is a fundamental difference between observation and decision thresholds in that the former is given by some convention, while the latter is completely at forecaster's disposal.

In order to asses the quality of forecasts it is appealing to estimate the correlation coefficient (r) between X_O and X_F. (For the aspects of forecasting performance that are measured by r – and those that are not measured, *i.e.* ignored – see Murphy (1995)). In order to estimate the correlation coefficient some assumption on latent-variables' distributions must be made, and it is here the BND comes forth. The purpose of the next subsection is to show that very often this is an acceptable assumption.

2.2. Transformation to SND

Any continuous random variable X may be transformed into standard normal variable Z_X by the formula:

$$Z_X = \Phi^{-1}(\Phi_X(X)) \tag{1}$$

Here, the Φ_X is cumulative distribution function (c.d.f.) of X, while Φ is c.d.f. of standard normal distribution. Variable Z_X is called standard normal deviate (SND) corresponding to X.

Obviously, transformation (1) is monotone. Somewhat surprisingly, it is also linear to a high degree, what, of course, depends on the actual distribution of X. So we systematically take variables from gamma and beta families (*e.g.*

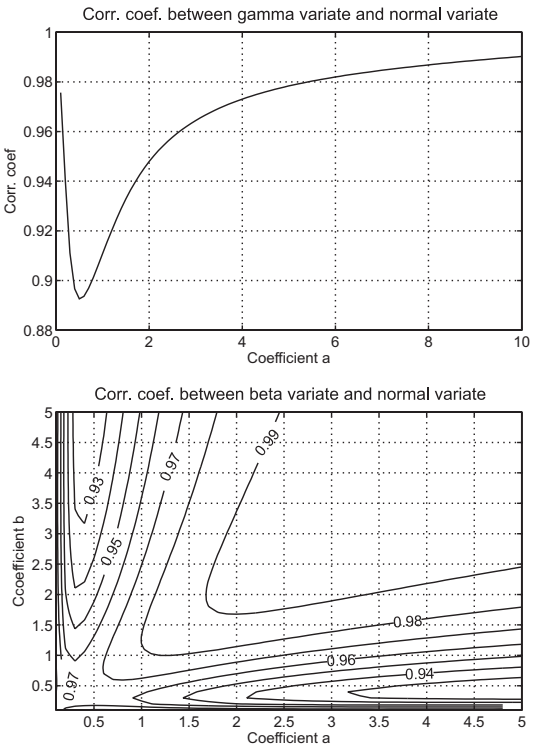

Figure 1. Correlation coefficient between gamma (upper panel) and beta (lower panel) variables and corresponding standard normal deviates as the function of respective parameters.

Wilks, 1995) as X and perform weighted linear regression between X and , the weight being given by the p.d.f. of X. The correlation coefficients obtained are rather high (Figure 1). In case of gamma family the correlation coefficient is always greater than 0.89, and it is even larger in case of beta family.

Gamma family of p.d.f.-s is bounded on the left by zero, the beta family is bounded on two sides, while members of both families vary from symmetric to highly skewed ones. Both distributions are very common in meteorology and, taking into account the above results on linear regression, we may conclude that in practice, transformation (1) would be close to a linear one.

Returning back to the problem of association in contingency table and using the apparent linearity of transformation, we see that correlation coefficients between latent variables X_O and X_F and between corresponding SND-s Z_O and Z_F should be approximately the same. Now in order to estimate the correlation coefficient between the transformed variables it is plausible, if not the only possible – due to complete lack of any additional information, to assume that random vector (Z_O, Z_F) follows the BND (see also Pearson and Heron (1913, p. 177)).

2.3. Tetrachoric correlation

Let $z_O = \Phi^{-1}(P_O)$ and $z_F = \Phi^{-1}(P_F)$ be the standard normal deviates (SND) corresponding to marginal probabilities P_O and P_F, respectively. The tetrachoric correlation coefficient (TCC), introduced by Pearson (1900), is the correlation coefficient r that satisfies

$$a = \int_{z_O}^{\infty}\int_{z_F}^{\infty}\phi(x_1, x_2, r)dx_1dx_2, \tag{2}$$

where $\phi(x_1, x_2, r)$ is the bivariate normal p.d.f.

$$\phi(x_1, x_2, r) = \frac{1}{2\pi\sqrt{1-r^2}}\exp\left[-\frac{1}{2(1-r^2)}(x_1^2 - 2rx_1x_2 + x_2^2)\right]. \tag{3}$$

The lines $x_1 = z_O$ and $x_2 = z_F$ divide this bivariate normal into four quadrants whose probabilities correspond to relative frequencies in the 2×2 table.

Clearly, the SND-s z_O and z_F are uniquely determined by P_O and P_F, respectively. The double integral in (2) can be expressed as (National Bureau of Standards, 1959):

$$a = \frac{1}{2\pi}\int_{\arccos r}^{\pi}\exp\left[-\frac{1}{2}(z_O^2 + z_F^2 - 2z_Oz_F\cos w)\operatorname{cosec}^2 w\right]dw, \tag{4}$$

showing that the joint frequency a is a monotone function of r for any fixed z_O and z_F. It follows that the tetrachoric correlation r is well defined by (2). The same reference contains the tables of joint probabilities of BND. To calculate r from the 2×2 table one may use tetrachoric series expansion of (2) developed by Pearson (1900), the approach that has been adopted by Brown (1977). Programs in FORTRAN and MATLAB that use numerical integration and (4) can be obtained from authors.

Pearson (1900, p. 14) provided a (complicated) expression for large-sample standard error of TCC, that does not use the BND assumption but requires that the entries of the contingency table, that are realizations of certain categorical variables, be mutually independent. Hamdan (1970) showed that the same TCC results from maximum likelihood method and subsequently obtained a simple asymptotic expression for standard error.

In what follows we discuss some properties of the TCC that are valid for any 2×2 table without reference to latent-variables or BND. However, these properties are most easily verified by considering the table has been obtained by dichotomizing some BND. Throughout rest of the paper, the TCC of a 2×2 contingency table, and also polychoric correlation coefficient for $K \times K$ table (next subsection), are denoted by S_r. Since the S_r is a correlation coefficient, it takes values between –1 and 1. Let us discuss some further properties:

1) Random forecast, where joint frequencies are products of corresponding marginal frequencies, is equivalent to $S_r = 0$.

2) For constant forecasts (where $a = b = 0$ or $c = d = 0$) the S_r is undetermined. Indeed, any such table can be approached by a sequence of tables having the same observation threshold and arbitrary but fixed S_r, with forecast threshold tending to plus or minus infinity, as necessary. This reflects the fact that with constant forecast, any association between forecast and observation is absent. Certainly, to produce constant forecasts no skill at all is required.

3) Equality to zero of some off-diagonal element ($b = 0$ or $c = 0$) is equivalent to $S_r = 1$, since in order to have one quadrant empty, the isolines of BND surface (ellipses), must degenerate to the regression line. This stresses that S_r is a measure of association only. Perfect forecast ($b = 0$ and $c = 0$) is equivalent to $S_r = 1$ and bias $B = 1$. Although forecasts without false alarms ($b = 0$) or without missed events ($c = 0$) certainly possess some skill, extreme caution is necessary. First, having a category with zero frequency means either the total number of observation-forecast pairs is not large enough or the assumption on continuous latent variables is not the appropriate one. However, under this assumption zero value of b or c may be replaced with any value lower than $0.5/N$, and corresponding S_r may be calculated, leading to the analysis on how it depends on chosen b or c values. Note also that standard errors are not available in this case (see Pearson and Heron (1913)).

4) The complement symmetry is valid, that is the S_r does not depend on what is chosen to be event vs. non-event. Indeed, this is equivalent to multiplying SND-s z_O and z_F with –1, without changing the underlying BND.

5) The transpose symmetry is valid, that is S_r does not change if forecast and observation are interchanged. This property follows from analogous property of the ordinary correlation coefficient.

In contrast to properties 1–5, next property relies on the latent-variables model of forecasting process.

6) The S_r depends neither on the climatological probability (P_O), nor on the bias (B), since the S_r is correlation coefficient between latent variables that depends on the variables themselves, and not on any thresholds. Thus it may be expected that S_r would facilitate comparison of forecasts over different climatological regions, and discourage the »hedging« that may happen by overforecasting or underforecasting a particular category.

2.4. Polychoric correlation

The situation with $K > 2$ forecast/observation categories results in $K \times K$ contingency tables. The table element, that is the relative frequency of forecasting i-th category while the j-th one is observed, is denoted by P_{ij}, $i,j = 1,..., K$. Marginal frequencies, that are the frequencies of forecast (observation) categories, are denoted by $P_{i.}$ ($P_{.i}$). Then the $K - 1$ probabilities of exceeding

respective observation (forecast) thresholds are denoted by $P_{Oi} = \Sigma_{j>i} P_{\cdot j}$ ($P_{Fi} = \Sigma_{j>i} P_{j\cdot}$), $i = 1,..., K-1$. Analogously to the 2×2 case, biases are defined by $B_i = P_{Fi} / P_{Oi}$, $i = 1,..., K-1$.

The transformation to SND-s is applied exactly as before, resulting in $2(K–1)$ transformed threshold values z_{Oi} and z_{Fi} dividing the categories. Polychoric correlation coefficient (PCC) is correlation coefficient between the SND-s assuming their joint p.d.f. being the bivariate normal.

There are several possibilities to estimate the PCC. Weighted mean of tetrachoric coefficients of two-class tables obtained by partitioning the original $K \times K$ table may be used. Maximum likelihood (ML) method can be applied in two variants. The conditional ML can be used to estimate the TCC for the thresholds z_{Oi} and z_{Fi} (Martinson and Hamdan, 1971), or joint ML can be used to simultaneously estimate the TCC and the thresholds (Olsson, 1979). The former method, implemented in IMSL (1975), is computationally much simpler, while the practical differences between the two methods are small (Olsson, 1979). Both methods produce asymptotic standard errors of respective estimates. Finally, since the data are already categorized, the minimum Chi-Square (MCS) method may be convenient. It has the same asymptotic properties as ML method (Kendall and Stuart, 1973).

3. Examination of some common scores

3.1. Definition of scores and motivation

In this section we examine some common scores for 2×2 verification tables, as functions of the triplet (S_r, B, P_O). Incidentally, from the preceding section it follows that knowledge of (S_r, B, P_O) is sufficient to reconstruct the contingency table uniquely. On the other hand, by using the triplet, information contained in the table is conveniently divided into three independent parts: the tetrachoric correlation coefficient (S_r) measures the association in the table, bias (B) measures the discrepancy between frequencies of forecasts and observations, while the frequency of the event (P_O) measures rareness of the event (rare events usually are more difficult to forecast).

One may ask what is gained by such kind of analysis? We are examining the behavior of certain scores within the simplest and the most well-known theoretical model, namely dichotomous BND. In such a model we know what is the exact amount of association. It is given by the ordinary correlation coefficient (S_r). Besides, it is convenient to express the results in terms of P_O and B, since both quantities have obvious interpretation in the forecasting context.

Among a large number of different measures used in forecast verification, we only consider four of them here. First three are the well known Peirce measure (s_P), Heidke measure (s_H), and square-root of Doolittle measure (s_D).

Fourth is the Yule's odds ratio skill score (S_Y), introduced into meteorological practice recently by Stephenson (2000). In our notation these four scores are defined by:

$$s_P = \frac{a - P_O P_F}{P_O(1 - P_O)}, \tag{5}$$

$$s_H = \frac{2(a - P_O P_F)}{P_O + P_F - 2P_O P_F}, \tag{6}$$

$$s_D = \frac{a - P_O P_F}{\sqrt{P_O(1 - P_O) P_F(1 - P_F)}} \tag{7}$$

$$S_Y = \frac{a - P_O P_F}{a[1 - 2(P_O + P_F) + 2a] + P_O P_F} \tag{8}$$

Apparently, all four measures contain the expression $a - P_O P_F$ in the numerator. Pearson (1900) calls it the transfer per unit of the total frequency, since its magnitude measures the divergence of the actual table from the table corresponding to random forecasts. This indicates that all the four scores are measures of association and Pearson has shown that S_r is a function of transfer, too. He proposed that any reasonable measure of association should vanish with transfer, and also it would be beneficial if the measure agrees with correlation coefficient for median divisions ($P_O = P_F = 0.5$), in which case we have (Sheppard, 1898):

$$S_r = \sin\left(\frac{\pi}{2}(4a - 1)\right). \tag{9}$$

Thus, a number of scores that contain sine function have been proposed for evaluation of association in contingency tables. They may be interpreted as attempts to estimate the TCC, (see Johnson and Kotz (1972, p. 119)). In meteorological literature such examples are somewhat rare (see Brooks and Carruthers (1953, p. 238) and Panofsky and Brier (1958, p. 103)).

In what follows, the letter s without indices denotes any of the first three measures s_P, s_H and s_D. In the special case when $P_O = P_F = 0.5$ they reduce to the form $s = 4a–1$, a property which is shared by a large number of scores used in meteorology (Eq. 24 Woodcock, 1976). Thus it is tempting to apply the transform analogous to (9):

$$S = \sin\left(\frac{\pi}{2}s\right) \tag{10}$$

to the first three scores, s_P, s_H, and s_D. The modified measures thus obtained are denoted by S_P, S_H, and S_D, respectively. This transformation brings the

numerical values of original measures closer to the respective values of the tetrachoric correlation coefficient S_r, making the comparison more straightforward, while the essential properties remain unchanged. A short calculation shows that the S_Y score for $P_0 = P_F = 0.5$ is already a sine-like function of the joint frequency (a), and it equals S_r for $S_r = \pm 1$ and 0.

3.2. Comparison of scores

We restrict the analysis to two typical situations. First we fix the amount of association to a reasonably high value ($S_r = 0.85$) that may be expected in modern forecasting systems. Then we fix probability of the event to a relatively low value ($P_O = 0.1$) having in mind rare events. Results for other values are similar.

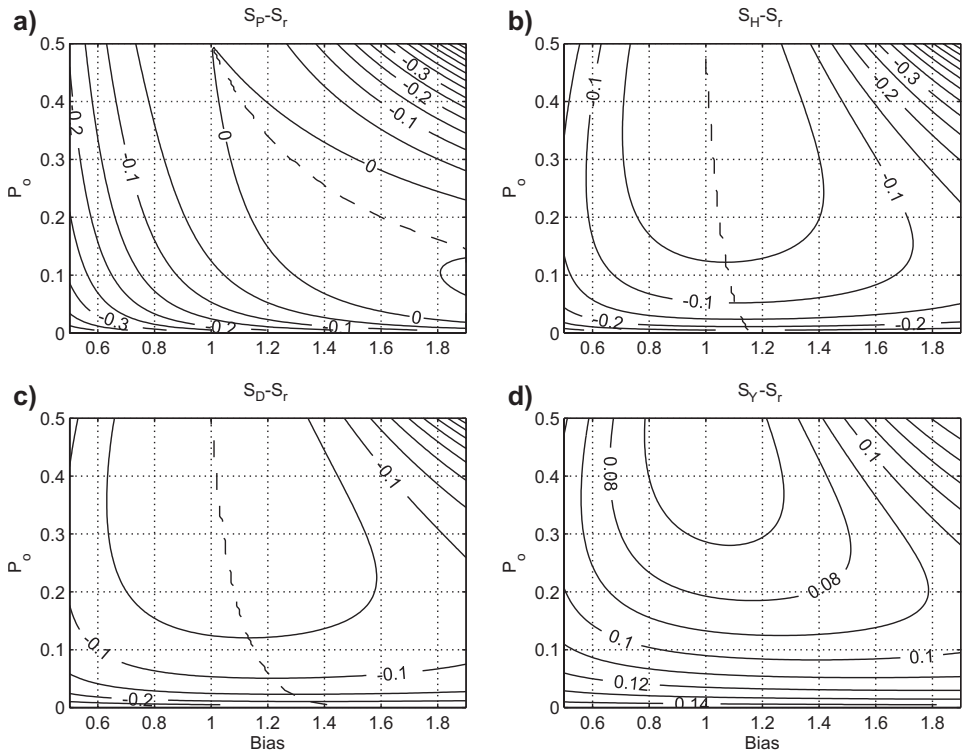

Figure 2. Modified Pierce (a), Heidke (b) and Doolittle (c) score, and the Yule's (d) score as a function of bias (B) and probability of event (P_O), for a fixed value of the TCC, $S_r = 0.85$; plotted are the differences between respective scores and the S_r. Contour interval equals 0.05 for graphs (a) – (c) and 0.01 for graph (d). The dashed curves on plots (a) – (c) denote the loci of maxima of respective fields with respect to B for a fixed P_O.

Fig. 2 shows how the four considered scores depend on P_O and B when the TCC is fixed to $S_r = 0.85$. Actually, the differences between the scores and $S_r = 0.85$ are plotted in order to make the figure more easy to read. As it was already noted, the first three measures are equal to S_r if marginal probabilities are equal to one half, that is for $P_0 = 0.5$, $B = 1$. The difference $S_Y - S_r$ at the same point, although does not equal zero, exhibits a minimum. Apparently, all the four scores are close to S_r for moderate biases ($0.8 < B < 1.4$) and for the events that are not too rare ($P_O > 0.15$).

One can notice that scores S_P, S_H, S_D decrease along the line $B = 1$ as the probability of the event, P_O, decreases. This decrease, that is very similar for all three measures, makes them unsuccessful when they are applied to rare events. On the contrary, the score S_Y increases along the line $B = 1$, as P_O decreases.

Fig. 2 clearly shows that for values of P_O close to 0.5 (upper edge of the graph) the first three scores decrease as the bias diverges from B = 1. In other words, for large P_O all three measures penalize the existence of bias in forecasts in almost the same way, which may be considered reasonable. Unfortunately, the measures do not keep this property at small values of P_O (rare events). In this case, the Peirce score favors overforecasting ($B > 1$) as has already been reported by Stansky et al. (1989), Doswell et al. (1990) and Marzban (1998), while the Heidke and Doolittle scores become insensitive to bias. The weakly expressed maxima of S_H and S_D for small P_0 are found not at $B = 1$ but at somewhat larger values of bias (dashed lines on graphs (b) and (c)). Apparently, these two scores become practically equal for biases between 0.8 and 1.2. The isolines of S_Y look quite similar to those of S_H and S_D. However, the S_Y score increases in all directions starting from $B = 1$, $P_0 = 0.5$. For large P_0 it favors biases that are less than one, and even more the biases greater than one (overforecasting). For small P_0 it becomes more or less insensitive to bias.

Fig. 3 shows differences between individual scores and the S_r as a function of bias and S_r itself, the frequency of event, P_O, being fixed to 0.1. For all scores the differences are zero when S_r equals zero and the same is valid for the S_Y score when S_r equals one. The differences are biggest for medium values of TCC, $0.3 < S_r < 0.7$. Dashed lines on graphs (a) – (c) denote the loci of maxima of respective differences as a function of bias for every (constant) value of TCC. It is seen again that in the case of relatively low frequency of the event, Peirce score favors overforecasting. Almost horizontal isolines suggests that in the same situation the other three scores show little sensitivity to the bias. As before, Heidke and Doolittle scores slightly favor overforecasting, while the Yule's score slightly favors underforecasting (maximal values are reached at $B = 0$).

4. An example of the 2 × 2 contingency table

Table 2 contains two contingency matrices for categorical forecasts of fog with a lead time of 3 hours (Kruizinga et al., 1989). They describe the overall

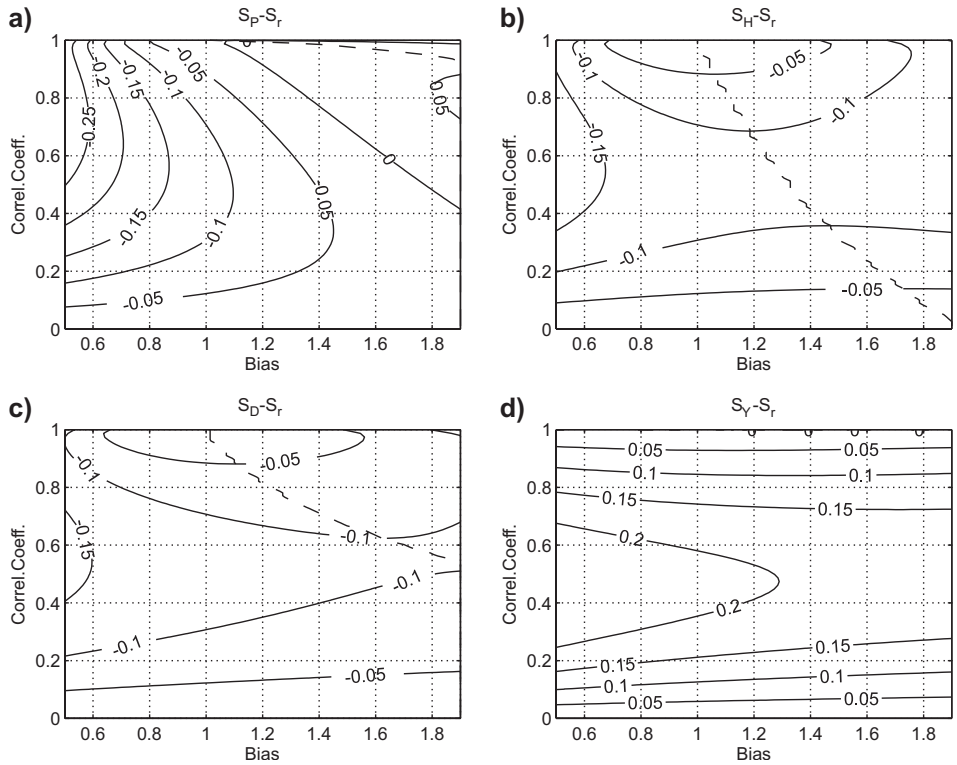

Figure 3. Modified Pierce (a), Heidke (b) and Doolittle (c) score, and the Yule's (d) score as a function of bias (B) and tetrachoric correlation coefficient (S_r), for a fixed probability of event, P_O = 0.1; plotted are the differences between respective scores and S_r. Contour intervals are 0.05, while the meaning of the dashed curves is analogous to that in Fig. 2.

quality of statistical and persistence forecasts, where the latter ones serve as control forecasts.

Let us use the contingency table of statistical method to give an explicit calculation of the TCC. The threshold between observation categories is defined by the frequency of the event P_O = 0.061, and the corresponding SND value is z_O = –1.546. Similarly for the forecast categories we have P_F = 0.141 and z_F = –1.075. The joint probability that the fog was forecasted and observed is a = 0.048. By solving equation (2) for r we find the TCC of statistical method, S_r = 0.81.

If one adopts s_P as the measure of skill, then s_P = 0.69 for the statistical method and s_P = 0.54 for the persistence model suggests the former method is having higher skill. Ranking on S_P scores with values 0.89 and 0.75 leads to the same conclusion. The S_r value for the statistical method is 0.81, and for the persistent method it is 0.90, indicating that the association between

Table 2. Contingency tables for very short-range forecasts of fog based on a statistical model and persistence, after Kruizinga et al. (1989, Table 6).

Statistical				Persistence			
Fcst.\Obs.	Yes	No	P_F	Fcst.\Obs.	Yes	No	P_F
Yes	0.048	0.093	0.141	Yes	0.033	0.013	0.046
No	0.013	0.846	0.859	No	0.027	0.927	0.954
P_O	0.061	0.939		P_O	0.061	0.939	
	s_P = 0.69	S_P = 0.89			s_P = 0.54	S_P = 0.75	
	s_D = 0.48	S_D = 0.68			s_D = 0.61	S_D = 0.82	
		S_Y = 0.94				S_Y = 0.97	
		S_r = 0.81				S_r = 0.90	
	Bias = 2.33				Bias = 0.76		

forecasts and observations in the latter method is better than in the former. It is not rare that various measures give various rankings for a specific set of forecasts, since each measure gauges its own facet of forecast quality. In this example the conflict between S_r and S_P is due to the fact that the S_P (and s_P) measure favors overforecasting of rare events, and its relatively large value is due to a large bias ($B > 2$). The S_r is a measure of association only and it does not depend on bias.

One can find that statistical forecasts are still better for those users who are susceptible to undetected events. However, the persistent forecasts, which are based only on the most recent observation of visibility, could be easily adjusted to the needs of these users. Namely, the fog should be forecasted not only in the case when it is already present, but more often, when visibility is less than, say, 3 kilometers. For such a modified autoregressive forecasting system, S_r would remain the same but s_P score would probably increase to a value even greater than the value for statistical forecasts. This example illustrates that S_r, like the ordinary correlation coefficient, indicates a »potential value« of forecasts as it was noted by Murphy and Epstein (1989) and Murphy (1995) for anomaly and ordinary correlation coefficients.

5. Multicategorical tables

Recently meteorologists have returned to the problem of finding an optimal scoring matrix for multicategorical contingency tables (Gandin and Murphy, 1992; Gerrity, 1992; Barnston, 1992; Potts et al., 1996). All the quoted papers start from the assumption that each of the unskilled forecast strategies, such as randomly choosing a forecast category or always forecasting the

same category, must receive a zero score. This premise seems plausible, but it also has opponents (Rousseau, 1980). Many of the national verification summaries, especially those for the forecasts with a large lead time, indicate that categories around prevailing conditions are more often used than the categories for rare, extreme events. It seems that the subjective (or official) criteria for evaluation of forecasts differs from the principles of equitability.

Gandin and Murphy (1992) have shown that there is an unlimited number of $K \times K$ scoring matrices that satisfy the principles of equitability. They pointed out that for a 2×2 contingency table and symmetric scoring matrices, the s_P score and its monotonic transformations are the only scores that satisfies these principles (see also (Marzban and Lakshmanan, 1999)). Gerrity (1992) has introduced a subset of Gandin and Murphy scoring matrices that depend on marginal probabilities of observations, only. He also showed that the resulting equitable skill score (ESS) for a $K \times K$ contingency table can be obtained by averaging the s_P scores computed for the K–1 two-class tables generated by partitioning the original contingency table at its K–1 thresholds. It follows that ESS should retain the same (undesired) properties of the s_P score that were identified in Section 3.

The S_P decreases rapidly when the frequencies of observed categories are far from one half (Fig. 2). This is the reason why for the same degree of association among forecasts and observation (as measured by S_r), the ESS decrease as the number of categories increase. This property of ESS, identified already by Barnston (1992), makes it difficult to compare the contingency tables of different dimensions, and even the contingency tables of the same dimensions, if they have different marginal frequencies. This probably motivated Barnston to investigate the connection between ESS and ordinary correlation coefficient for tables with different dimensions (see Fig. 3a–d of his paper). The results of Section 3 (Fig. 3) on S_P score show that the ESS score would also prefer overforecasting of outermost categories that are less frequently observed.

5.1. Fictitious examples

In this subsection some equitable scores are compared to S_r assuming a situation with three equiprobable categories of observation. In such situation any 3×3 matrix of joint probabilities of some standard BND with all the marginal probabilities equal to 1/3 could be interpreted as the performance matrix of some set of forecasts. On the other side any such matrix can be used for construction of an equitable scoring matrix, just by subtracting 1/9 from each element, dividing subsequently the whole matrix by the sum of diagonal elements and finally multiplying by three (since, the equitability here implies sums of rows to be zero, while sum of diagonal elements must be three). An example of such an equiprobable performance matrix that is obtained from BND with $r = 0.71$ is given in Table 3a. The corresponding scoring matrix

Table 3. (a) Joint probabilities of BND for three equiprobable categories with r = 0.71, (b) the equitable scoring matrix obtained from (a) as is explained in the text and (c) the scoring matrix for three equiprobable categories from Gerrity (1992).

(a)			(b)			(c)		
0.219	0.093	0.022	1.278	–0.222	–1.057	1.25	–0.25	–1
0.093	0.147	0.093	–0.222	0.443	–0.222	–0.25	0.5	–0.25
0.022	0.093	0.219	–1.057	–0.222	1.278	–1	–0.25	1.25

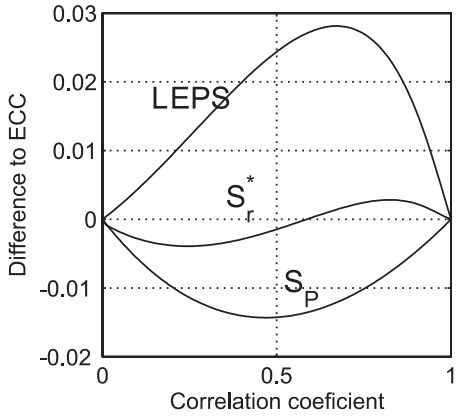

Figure 4. Equitable skill scores calculated for a series of 3×3 contingency tables with equiprobable categories generated from the standard BND with varying correlation coefficients; plotted are the differences between scores and the underlying correlation coefficient. Curves denoted by S_r^*, S_P and LEPS are obtained by applying the scoring matrix (b) and (c) from Table 3 and the scoring matrix of Potts et al. (1996, Table 1), respectively.

(Table 3b) differs only slightly from the scoring matrix obtained by Gerrity's method (Table 3c). We applied both matrices (b) and (c), as well as the linear error in probability space (LEPS) scoring matrix derived by Potts et al. (1996, Table 1), suitably normalized, to the series of equiprobable 3×3 tables generated form BNDs with varying correlation coefficients. The differences between respective scores and the underlying correlation coefficient are shown at Fig. 4. Estimates obtained by matrix (b) differ from the true correlation coefficient by not more than 0.005. Actually the matrix (b), *i.e.* the value $r = 0.71$ behind it, was chosen such that the corresponding score reconstructs the underlying correlation coefficient as close as possible. For the other two scoring matrices the differences are slightly greater. This points out the similarity between S_P and S_r for symmetric multicategorical tables as has already been noted in the 2×2 case for moderate biases and probabilities of the event that were not too small.

5.2. Real-world example

Starting a few years ago the National Precipitation Verification Unit (NPVU) of the United States National Weather Service (NWS) regularly disseminates various statistics on verification of quantitative precipitation forecasts (QPF) from the River Forecasts Centers. The original multiple-categories contingency tables are available as well. Here we take the annual summary of QPF for day 1 (0–24 hour period), the contiguous US for the year 2005. The table is available at the NPVU web pages, http://www.hpc.ncep.noaa.gov/npvu/qpfv/. In its original form this 6×6 table counts for more than 11 million forecast-observation pairs categorized according the thresholds of 0.01, 0.10, 0.25, 0.50 and 1.00 inch. Let us mention that we do not advocate the agregation of data from different climatological regions into a single contingency table, but here we use the table to give some illustrative examples. Table 4 presents the data as relative frequencies given in percents.

As a first step, the thresholds between categories are calculated from marginal frequencies by inverting the 1-dimensional standard normal c.d.f. In this way we get the normalized thresholds, $z_{Oi} = 1.12, 1.67, 2.03, 2.41, 2.93$, and $z_{Fi} = 0.85, 1.47, 1.99, 2.49, 3.08$, for $i = 1, \ldots, 5$. Then the PCC of the table is estimated by the ML and the MCS method, both conditioned to the above thresholds. In ML (MCS) method the underlying function is maximized (minimized) by simple bisection algorithm, using the fact that joint probabilities of any standard BND are uniquely determined by a complete set of corresponding marginal frequencies and the correlation coefficient. The ML method gives 0.795, while the MCS one gives 0.782. The greatest difference between the theoretical and the observed joint frequencies is 0.35 percent for the ML method (Table 5), and 0.44 percent for the MCS one (table not shown). The sum of absolute differences over all cells is close to 1.8 percent for ML and 2.4 percent for the MCS method. Thus, the whole table may be reconstructed

Table 4. Annual summary of quantitative precipitation forecasts for day 1, the contiguous US, for the year 2005 from the NPVU of the NWS, in percents. The thresholds dividing the categories C1 – C6 are 0.01, 0.1, 0.25, 0.5 and 1 inch of precipitation, respectivelly.

	Observation						
Forec.	C1	C2	C3	C4	C5	C6	Σ
C1	76.96	2.76	0.40	0.13	0.05	0.01	80.31
C2	7.79	3.53	0.87	0.28	0.09	0.02	12.57
C3	1.67	1.62	0.90	0.41	0.15	0.03	4.77
C4	0.32	0.43	0.42	0.34	0.17	0.04	1.71
C5	0.05	0.08	0.10	0.13	0.13	0.05	0.54
C6	0.01	0.01	0.01	0.02	0.03	0.03	0.10
Σ	86.79	8.41	2.70	1.31	0.62	0.17	100.00

Table 5. Differences expressed in percents between original contingency table (Table 4) and the corresponding table of joint probabilities of BND (having the same marginal probabilities and the correlation coefficient equal to 0.795, obtained by the ML method).

	Observation					
Forec.	C1	C2	C3	C4	C5	C6
C1	0.214	–0.255	–0.040	0.037	0.034	0.009
C2	–0.263	0.355	–0.059	–0.050	0.007	0.011
C3	–0.037	–0.010	0.093	–0.021	–0.028	0.004
C4	0.055	–0.081	0.020	0.032	–0.016	–0.010
C5	0.025	–0.011	–0.014	0.005	0.005	–0.010
C6	0.005	0.002	–0.000	–0.002	–0.002	–0.003

almost perfectly using the five marginal frequencies of observations, five marginal frequencies of forecasts and the correlation coefficient. Alternatively, biases with frequencies of events may be plotted vs. respective thresholds (Fig. 5). By adding a single extra number, the TCC, all the essential information contained in the table is conveniently displayed. In general, if we are given the absolute frequencies, a $K \times K$ table could be reduced to $2K$ data (the TCC,

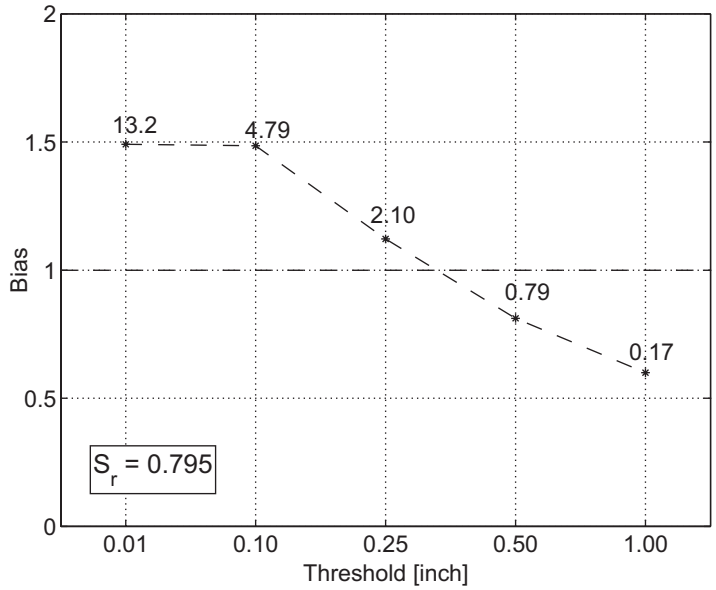

Figure 5. Biases vs. observation-threshold values for Table 4, labeled with the relative frequencies of exceeding the respective thresholds (P_O, in percents). The polychoric correlation coefficient, displayed in lower left corner, expresses the amount of association. It is the only additional peace of information required to essentialy reconstruct the original table.

$2K$–2 marginal frequencies, and the total number of observations), the polychoric correlation coefficient being the only non-trivial one. By measuring solely the association in the contingency table it should enable comparison between tables of different sizes and/or marginal frequencies. Moreover, by examining the differences between the original contingency table and the corresponding theoretical one, it might be possible to address some specific features of a particular forecasting system.

The standard errors based on the very large total number of inputs (N) in this example are unrealistically small, so we do not report them. This points to the problem of estimating the number of degrees of freedom, which is certainly smaller than N. Indeed, the table is comprised from numerical-model results and corresponding observations, and values for nearby grid elements are certainly not indepenent. However, let us recall that standard errors are getting lower as the number of categories increases (Olsson, 1979).

The present example may be used to examine how the various scores change with the probability (P_O) of the event. To that end we partitioned Table 4 at its thresholds and calculated the previously examined scores (Section 2) for each of the five 2×2 contingency tables obtained. Results are shown on Fig. 6. The decreasing sequence of P_O values corresponding to respective thresholds is indicated at the upper axis. It is readily seen that

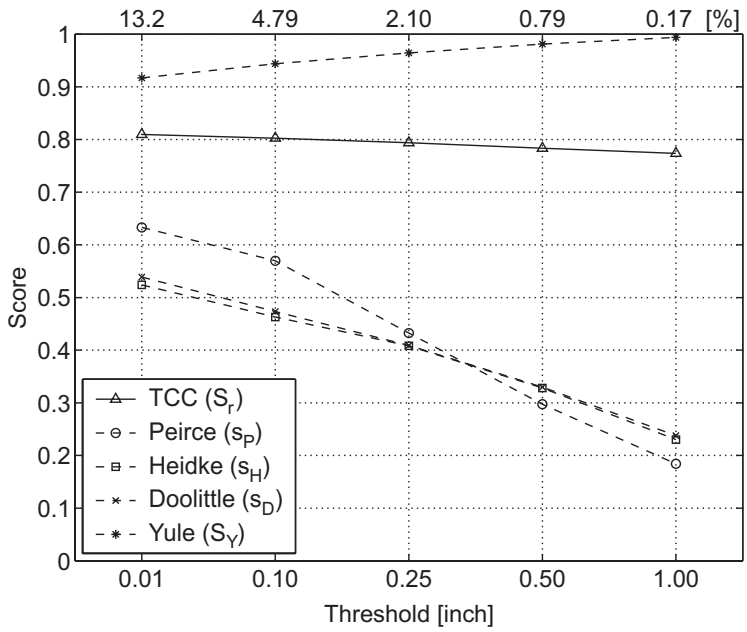

Figure 6. Various scores calculated for two-class tables obtained by partitioning Table 4 at its thresholds. The P_O values in the upper axes are the same as in Figure 5.

Peirce, Heidke and Doolittle scores decrease considerably, together with P_O, while at the same time Yule's score slowly increases. On the contrary, the TCC first remains nearly constant and then slowly decreases as the threshold of one inch, corresponding to a rather extreme event, is reached.

We may also try to examine the influence of bias on the considered scores by performing a sort of »hedging« to Table 4. Thus, we simulate the overforcasting of precipitation exceeding 1 inch by simply moving all the data from the middle four forecast categories into the last category. In this way we obtained a new 6×6 table having four middle rows filled with zeros, the first row common with Table 4 and the last row equal to the sum of all-but-the-first rows of Table 4. The overall PCC value obtained by the ML method now reads 0.798, while MCS method gives 0.782, which are almost the same as for the original table. The sum of absolute differences between the fitted and the observed table is now around 1 percent for the ML, and around 1.8 percent for the MCS method. The overall Peirce score (ESS) increases from 0.423 for the original table to 0.711 for the 'hedged' one.

Finally, in order to examine the trends in precipitation forecasts over last several years we calculated various parameters for observational thresholds of 0.01 and 1 inch (Table 6). For precipitation exceeding 0.01 inch ($P_O = 0.12$ on average) we see overall decrease of positive bias (overforecasting) that is particularly well noticable between years 2003 and 2004. The TCC (S_r) remains nearly constant over the period. A slight decrease of Peirce (s_P) and increase of Heidke score (s_H) between years 2003 and 2004 is primarily a consequence of the reduced bias (Fig. 2). For precipitation exceeding 1 inch ($P_O = 0.002$ on average) considerable improvement of negative bias (underforecasting) between years 2002 and 2003 is evident. The TCC exibits increase that is most noticeable between years 2003 and 2004 when it gets close to 0.8 thus approaching a level of association between observations and forecasts that already exists for the lower threshold. The increase of s_P, s_H, s_D scores over the

Table 6. Various parameters calculated for Table and analogous tables for years 2001–2004, for the two outermost categories.

	Threshold = 0.01 inch					Threshold = 1 inch				
Year	2001	2002	2003	2004	2005	2001	2002	2003	2004	2005
S_r	0.789	0.801	0.802	0.796	0.810	0.646	0.692	0.711	0.760	0.774
s_P	0.624	0.637	0.636	0.616	0.633	0.070	0.096	0.129	0.162	0.184
s_H	0.478	0.493	0.503	0.513	0.524	0.106	0.141	0.168	0.214	0.230
s_D	0.501	0.515	0.523	0.527	0.539	0.122	0.160	0.177	0.225	0.238
S_Y	0.911	0.918	0.915	0.907	0.917	0.985	0.988	0.990	0.993	0.994
Bias	1.697	1.671	1.614	1.477	1.491	0.332	0.362	0.532	0.516	0.600
$P_O(\%)$	0.109	0.111	0.123	0.138	0.132	0.002	0.003	0.002	0.002	0.002

years is a consequence of improved association. Their rather low values as compared to those for 0.01 inch threshold are due to small P_O. Yule's score, for low values of P_O, is weakly sensitive to bias as well as to S_r. Actually, for P_O = 0.002 the S_Y is greater then 0.9 for all $S_r > 0.4$ irrespectively of bias, and increases very slowly as association (S_r) increses (not shown). For P_O that small it is likely that Yule's score would be close to one. This is also a reason why Yule's cube is sometimes reported instead of the original value.

6. Summary and Conclusions

The notions of TCC (PCC) for 2×2 ($K \times K$) contingency tables were recalled. The mild assumption on the existence of continuous latent variables is distinguished from the strong one that requires BND. However, the latter assumption is being applied not to latent variables themselves, but to transformed variables whose marginal distributions are exactly normal. Next, it was shown that, in practice, the transformation of latent variables into SND-s should be close to a linear one implying that TCC (PCC) should be approximation of the true correlation coefficient between latent variables. Apparent linearity of transformation explains the close agreement of 6×6 table from Section 5 with BND.

The TCC possesses a number of properties that are beneficial for any measure of association as it was discussed in Section 2. Although most of them are valid for the contingency table itself without any additional asumptions, from practical point of view it is the latent variables assumption that implies the most important property. Namely, the TCC does not depend on climatological probability (P_O), nor on bias (B). Thus, in practice, it may be expected that TCC and PCC would facilitate the comparison of forecasts over different climatological regions, and also be resistant to »hedging«.

Apparently, the TCC and PCC are measures of association only. They can not be used as measures of overall quality of forecasts, as neither measure can. Additional quantities are necessary to express the rest of information carried by the contingency table. It was shown that 2×2 table can be perfectly reconstructed by using the triplet (S_r, B, P_O), so the information contained in the table is naturally and clearly divided between association, bias and (climatological) probability of the event. It is the assumption on continuous latent variables that gives meaning to the assertion that the three quantities do not depend on each other. Similar division may be proposed for tables of higher order. In case of a $K \times K$ table with K–1 underlying events of exceeding respective thresholds, we have one correlation coefficient, namely the PCC, K–1 biases and K–1 marginal probabilities. Using these parameters and the BND we should be able to essentially reconstruct the original table. In practice, the quality of such a reconstruction should be assesed for each table separately. The gain should be threefold. First is the reduction in dimensionality from K^2

to $2K$ (if we add the total number of table elements as an additional parameter). Second, all parameters except the PCC are fairly simple with the meaning obvious even to non-meteorologists. Eventually, the table may be clearly expressed graphically as in Figure 5. Third, the differences between observed and theoretical contingency table may be further investigated using the distribution-oriented approach.

We analyzed the four common scores, namely Peirce, Heidke, Doolittle and Yule's measure, (the first three of them transfomed according to (10)), as functions of S_r, B and P_O. It appeared that all the four measures are close to each other and to S_r for moderate biases ($0.8 < B < 1.4$) and not-too-rare events ($P_O > 0.15$). This emphasizes the fact that they are, as many others, essentially measures of association. However, all four scores change considerably as climatological probability P_O is getting small. This is a serious deficiency of many common measures for forecast evaluation, since the values of a particular score can not be comapared to each other, not even within a relatively small geographical area, if the frequencies of the event within that area vary considerably (Reed, 1983). Moreover, different behavior of scores with respect to bias is found when they are applied to rare vs. non-rare events. In the latter case existence of bias is penalized. In the former it is favored. This is in accordance with the well known fact that usual measures like Peirce or Heidke did not prove suitable for rare events for which a special group of measures had been developed (*e.g.* critical success index, Gilbert's skill score). Regarding the properties of Yule's score we actually rediscovered some findings of Pearson and Heron (1913).

Summarizing, we believe that TCC (S_r) can be used as a measure of association that is less prone to the above mentioned deficiencies of scores. We may expect to face somewhat lower values of S_r for forecasts of rare events as opposed to forecasts of moderate-frequency events. However, this decrese should be an indication of difficulties related to the forecasting of rare events, rather than being a characteristic of S_r itself. Essentially S_r is an ordinary correlation coefficient, a concept that was introduced into statistics by Pearson a century ago. Over the years the intuition and feeling for its meaning and usage has been largely developed. Thus its values are much more comprehensible than the (usually lower) values of commonly used scores. Besides, it enables natural partition of information contained in a contingency table between association and bias, and eventual reduction of dimensionality in case of multicategorical forecasts.

Certainly, the value and usefulness of any particular score can not be decided from a single paper, nor can it be decided using heuristic and/or statistical arguments. The only way is systematic, long-term evaluation of scores within forecast practice, and this requires the complete contingency tables or equivalent information to be systematically reported. Unfortunately, very often this is not the case, preventing us to fully comprehend a large

number of presently used measures and see how they reflect the very slow but constant progress in weather forecasting.

Acknowledgement – This study was supported by the Croatian Ministry of Science, Education and Sports (grant 0119330).

References

Barnston, A. G. (1992): Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score, *Weather Forecast.*, **7**, 699–709.

Brooks, C. E. P. and Carruthers, N. (1953): *Handbook of Statistical Methods in Meteorology*. Her Majesty's Stationery Office, London, 412 pp.

Brown, M. B. (1977): Algorithm AS 116: The tetrachoric correlation and its asymptotic standard error, *Appl. Stat-J. Roy. St. C.,* **26**, 345–351.

Daan, H. (1984): *Scoring Rules in Forecast Verification*. Number 4 in PSMP Publication Series, WMO, 62 pp.

Doswell, C. A., III, Davies-Jones, R. P. and Keller, D. L. (1990): On summary measures of skill in rare event forecasting based on contingency tables, *Weather Forecast.*, **5**, 576–585.

Gandin, L. S. and Murphy, A. H. (1992): Equitable skill scores for categorical forecasts. *Mon. Weather Rev.*, **120**, 361–370.

Gerrity, J. P., Jr. (1992): A note on Gandin and Murphy's equitable skill score, *Mon. Weather Rev.*, **120**, 2709–2712.

Gringorten, I. I. (1971): Modeling conditional probability, *J. Appl. Meteorol.*, **10**, 646–657.

Gringorten, I. I. (1972): Conditional probability for an exact noncategorized initial condition, *Mon. Weather Rev.*, **100**, 796–798.

Hamdan, M. A. (1970): The equivalence of tetrachoric and maximum likelihood estimates of ρ in 2×2 tables. *Biometrika*, **57**, 212–215.

IMSL Library 1 (1975): International Mathematical and Statistical Libraries, Houston, Texas, 5 edition.

Johnson, N. L. and Kotz, S. (1972): *Distributions in Statistics. 4. Continuous Multivariate Distributions*. Wiley, New York, 333 pp.

Juras, J. (1982): Comparison of models for estimating the joint probability of a weather event, *J. Appl. Meteorol.*, **21**, 1926–1928.

Kendall, M. G. and Stuart, A. (1973): *The Advanced Theory of Statistics*, volume 2. Griffin, London, 723 pp.

Kruizinga, S., Blaauboer, D. and van Vliet., K. (1989): Statistical input for a fog forecasting system. *Preprints, 11th Conference on probability and statistics*, Amer. Meteor. Soc., Montrey, California, 84–87.

Martinson, E. O. and Hamdan, M. A. (1971): Maximum likelihood and some other asymptotically efficient estimators of correlation in two way tables. *J. Stat. Comput. Sim.*, **1**, 45–54.

Marzban, C. (1998): Scalar measures of performance in rare-event situations, *Weather Forecast.*, **13**, 753–763.

Marzban, C. and Lakshmanan, V. (1999): On the uniqueness of Gandin and Murphy's equitable performance measures, *Mon. Weather Rev.*, **127**, 1134–1136.

Murphy, A. H. (1995): The coefficient of correlation and determination as measures of performance in forecast verification, *Weather Forecast.*, **10**, 681–688.

Murphy, A. H. (1996): The Finley affair: A signal event in the history of forecast verification, *Weather Forecast.*, **11**, 3–20.

Murphy, A. H. and Epstein, S. (1989): Skill scores and correlation coefficients in model verification, *Mon. Weather Rev.*, **117**, 572–581.

National Bureau of Standards (1959): *Tables of Bivariate Normal Distribution Function and Related Functions*. Number 50 in NBS Appl. Math. Series, Govt. Printing Office, Washington, D. C., 258 pp.

Olsson, U. (1979): Maximum likelihood estimation of the polychoric correlation coefficient, *Psichometrika*, **44**, 443–460.

Panofsky, H. A. and Brier, G. W. (1958): *Some Applications of Statistics to Meteorology*. Pennsylvania State University, University Park, 224 pp.

Pearson, K. (1900): Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable, *Philos. Tr. R. Soc. S.–A*, **195**, 1–47.

Pearson, K. and Heron, D. (1913): On theories of association. *Biometrika*, **9**, 159–315.

Potts, J. M., Folland, C. K., Jolliffe, I. T. and Sexton, D. (1996): Revised »LEPS« scores for assessing climate model simulations and long-range forecasts. *J. Climate*, **9**, 34–53.

Reed, R. J. (1983): A note on the relationship between relative precipitation frequency and percent of correct forecasts. *B. Am. Meteorol. Soc.*, **64**, 148–149.

Rousseau, D. (1980): A new skill score for the evaluation of yes/no forecasts. *Proc. WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, WMO, Nice, France, 167–174.

Sheppard, W. F. (1898): On the application of the theory of error to cases of normal distributions and normal correlations, *Philos. Tr. R. Soc. S.–A*, **192**, 101–167.

Stansky, H. R., Wilson, L. J. and Burrows, W. R. (1989): *Survey of Common Verification Methods in Meteorology*. Number 358 in WMO/TD., WMO, Geneva, Switzerland, 114 pp.

Stephenson, D. B. (2000): Use of the »odds ratio« for diagnosing forecast skill, *Weather Forecast.*, **15**, 221–232.

Ward, M. N. and Folland, C. K. (1991): Prediction of seasonal rainfall in the north nordeste of Brazil using eigenvectors of sea-surface temperature, *Int. J. Climatol.*, **11**, 711–743.

Wilks, D. (1995): *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, 464 pp.

Woodcock, F. (1976): The evaluation of yes/no forecasts for scientific and administrative purposes, *Mon. Weather Rev.*, **104**, 1209–1214.

SAŽETAK

Primjena tetrahoričkog i polihoričkog koeficijenta korelacije u verifikaciji prognoza

Josip Juras i Zoran Pasarić

Tetrahorički (polihorički) koeficijent korelacije dobro je poznata mjera asocijacije u kontingencijskim tablicama veličine 2×2 ($K \times K$). Ove mjere počivaju na dvjema pretpostavkama: 1) U pozadini kontingencijske tablice nalaze se neprekidne latentne varijable, te 2) zajednička funkcija distribucije pripadnih standardnih normalnih devijata je bivarijantna normalna razdioba. Pokazano je da tetrahorički, odnosno polihorički koeficijent korelacije predstavlja procjenu Pearsonovog koeficijenta korelacije izmedju latentnih varijabli. Posljedično, ove mjere ne ovise o pristranosti, kao ni o marginalnim čestinama, što rezultira rasčlambom informacije sadržane u kontingencijskoj tablici na

tri dijela. Prvi se odnosi na povezanost, drugi na pristranost, a treći daje čestinu razmatrane pojave. Korištenjem dobivenog rastava analizirana je ovisnost drugih verifikacijskih mjera o pristranosti i o marginalnim čestinama. Rezultati su prirodno prošireni na tablice oblika $K \times K$, pri čemu se dimenzija problema smanjuje s K^2 na 2K. Teorija je primjenjena u analizi tablica veličine 6×6 koje opisuju kvantitativne prognoze oborine.

Ključne riječi: tetrahorički koeficijent korelacije, tablica kontingencije, ocjena prognoza

Corresponding author's address: Dr. Zoran Pasarić, Department of Geophysics, Faculty of Science, University of Zagreb, 10000 Zagreb, Horvatovac b.b., Croatia, tel: +385 1 4605 922, e-mail: pasaric@irb.hr.