Geophysical Journal International

Geophys. J. Int. (2011) **184,** 1445–1454 doi: 10.1111/j.1365-246X.2011.04930.x

Gambling score in earthquake prediction analysis

G. Molchan[1,2] and L. Romashkova[1,2]

[1] *International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Science, Moscow, Russia.*
E-mail: molchan@mitp.ru
[2] *The Abdus Salam International Centre for Theoretical Physics, SAND Group, Trieste, Italy*

Accepted 2010 December 17. Received 2010 November 29; in original form 2010 June 6

SUMMARY

The number of successes and the space–time alarm rate are commonly used to characterize the strength of an earthquake prediction method and the significance of prediction results. It has been recently suggested to use a new characteristic to evaluate the forecaster's skill, the gambling score (GS), which incorporates the difficulty of guessing each target event by using different weights for different alarms. We expand parametrization of the GS and use the M8 prediction algorithm to illustrate difficulties of the new approach in the analysis of the prediction significance. We show that the level of significance strongly depends (1) on the choice of alarm weights, (2) on the partitioning of the entire alarm volume into component parts and (3) on the accuracy of the spatial rate measure of target events. These tools are at the disposal of the researcher and can affect the significance estimate. Formally, all reasonable GSs discussed here corroborate that the M8 method is non-trivial in the prediction of $8.0 \le M < 8.5$ events because the point estimates of the significance are in the range 0.5–5 per cent. However, the conservative estimate 3.7 per cent based on the number of successes seems preferable owing to two circumstances: (1) it is based on relative values of the spatial rate and hence is more stable and (2) the statistic of successes enables us to construct analytically an upper estimate of the significance taking into account the uncertainty of the spatial rate measure.

Key words: Earthquake interaction, forecasting and prediction; Seismicity and tectonics; Statistical seismology.

1 INTRODUCTION

Earthquake prediction of the yes/no type usually deals with two characteristics: the rate of failures-to-predict, n, and the normalized measure of space–time alarm, τ (see e.g. Molchan 2003). In these terms one can characterize the strength of a prediction method and the significance of prediction results. In particular, the quantity $H = 1-(n+\tau)$ is the expected fraction of non-randomly predicted target events and therefore is a good candidate to characterize the strength of any prediction method at the research stage. The (n, τ) vector as an integral characteristic may mask inefficiency of a method in some parts of the space–time monitoring area. This weakness is somewhat compensated by relaxed requirements on the accuracy of the rate measure of target events in a subarea dg, $\lambda(dg)$, as well as by the ease with which the uncertainty of $\lambda(dg)$ can be incorporated in significance estimation (Molchan 2010; Molchan & Romashkova 2010).

The problem of choosing a suitable quantity (score) to characterize the strength of a prediction method is treated in an extensive literature (see e.g. Molchan 1997; Joliffe & Stephenson 2003). This is quite natural, because the choice of the score depends on the applications and prediction goals in mind. Even the simplest models of the score, $R = 1 - n - c\tau$, interesting as these are at the research stage with any positive *c,* lead to an infinity of optimal prediction strategies (Molchan 1997).

Recently Zhuang (2010) and Zechar & Zhuang (2010) suggested the so-called gambling score (GS), in which the forecaster is rewarded or punished for success or failure in prediction according to the risks that he has taken. In contrast to the other scores, the GS takes into account the predictions of two types: appearance and non-appearance of target events in a time–space subvolume of interest. Two types of alarms take place, for example, in the prediction of a subsequent strong earthquake (Vorobieva 1999). Usually forecaster suggests the prediction of target event. However in case the monitoring zone is fixed, the alarm zone complement can be considered by researcher as alarm of the alternative type.

Zhuang (2010) discussed a theoretical possibility of using the GS for comparison of any two forecasts. At present there is only one application of the GS approach to practical problems. This is an analysis of the Reverse Tracing of Precursors (RTP) prediction method (see e.g. Keilis-Borok *et al.* 2004) by Zechar & Zhuang (2010). This analysis includes (1) a quite uncommon methodology of comparison of the prediction results with the random guessing; (2) an example of GS parametrization: all RTP alarms are weighted differently so that the alarms with smaller prior probabilities of target events are considered more valuable. Under these conditions the requirement on the local accuracy of $\lambda(dg)$ becomes more stringent. Because the weighting of alarms is not unique and may be arranged in a variety of ways, any parametrization of the GS needs a justification.

In this study we discuss the GS approach as applied to the RTP method and consider a new application of the GS, namely, the significance estimation of prediction results. To illustrate how the GS works in a significance problem we use the M8 algorithm by Keilis-Borok & Kossobokov (1990), which is designed for the prediction of $M \geq 8.0$ events worldwide. The M8 algorithm was examined recently in our paper (Molchan & Romashkova 2010). This considerably simplifies our task of a preliminary analysis of data and predictions, so that we can concentrate on the methodology of the GS approach. In addition, we are in a position to compare the significance estimates based on the GS with those based on the conventional (n, τ) approach. To make reading more convenient, we give a list of essential notation in the Appendix.

2 THE GS APPROACH

2.1 The GS *R*

Let $Y = (y_1, \ldots, y_N)$ be a random binary sequence with the probabilities of outcome of the i-th component equal to

$$P(y_i = 1) = p_i,\ P(y_i = 0) = 1 - p_i. \tag{1}$$

Here, $y_i = 1$ may be interpreted as the occurrence of at least one target event in a space–time domain $A_i = G_i \times T_i$, and $y_i = 0$ as no event in A_i.

Let $X = (x_1, \ldots, x_N)$ be another binary vector to be treated as a prediction of Y: $x_i = 1$ means an alarm, that is, the 'positive prediction' of a target event in A_i, while $x_i = 0$ means no alarm, that is, the 'negative prediction' of a target event in A_i.

To be able to specify the quality of a prediction or the proximity of X and Y, an appropriate measure (score) $R(X, Y)$ is used. The choice of R may depend on a variety of factors like research and applied goals in the study of a prediction method, the requirements on the stability of the distribution of R given a reference seismicity model, etc.

The GS approach is of interest as being a tool for constructing meaningful models of R. This approach treats R as the net gain of the forecaster in a sequence of N trials. In the i-th trial the forecaster may stake $b_+(p_i)$ on the outcome $y_i = 1$ or $b_-(p_i)$ on the outcome $y_i = 0$. In the case of success the forecaster gets $a_+(p_i)$ or $a_-(p_i)$, respectively. The net gain is

$$R(X, Y) = \sum_{i \geq 1} [a_+(p_i) x_i y_i - b_+(p_i) x_i \bar{y}_i - b_-(p_i) \bar{x}_i y_i + a_-(p_i) \bar{x}_i \bar{y}_i], \tag{2}$$

where we denoted $\bar{c} = 1 - c$. To realize a fair game we suppose that the forecaster's expected gain under (1) in each trial is zero, that is,

$$E_Y[a_+(p_i) y_i - b_+(p_i) \bar{y}_i] = 0 = E_Y[a_-(p_i) \bar{y}_i - b_-(p_i) y_i],$$

where E_Y denotes the expectation with respect to the distribution of Y. This assumption restricts the number of unknown functions $\{a_\pm, b_\pm\}$ to two, because it requires

$$a_+(p)/b_+(p) = \bar{p}/p,\ a_-(p)/b_-(p) = p/\bar{p}. \tag{3}$$

As a result, R transforms to

$$R(X, Y) = \sum_{i \geq 1} w(p_i)(x_i - p_i)(y_i - p_i) + \sum_{i \geq 1} v(p_i)(y_i - p_i),$$

where

$$w(p) = b_+(p)/p + b_-(p)/\bar{p},\ v(p) = b_+(p) - b_-(p).$$

To further restrict the number of unknown functions, we shall interpret $a_+(p)$ and $a_-(p)$ as one and the same measure of the complexity in guessing the random events $\{y = 1\}$ and $\{y = 0\}$, respectively. If the measure depends on the probability of the event only, then according to (1) we have

$$a_-(p) = a_+(1 - p). \tag{4}$$

By (3) and (4), we get an analogous relation for b_+ and b_-:

$$b_-(p) = b_+(1 - p). \tag{5}$$

The corollary is that $w(p)$ is symmetric: $w(p) = w(1 - p)$, while v is antisymmetric: $v(p) = -v(1 - p)$.

When dealing with the prediction of large events, the typical situation involves a small p, $p < 1/2$. To stimulate the prediction of rare events ($p < 1/2$), the natural requirement is $b_+(p) \leq b_-(p)$ for the stakes and $a_+(p) \geq a_-(p)$ for the gains. The simplest choice

$$b_+(p) = b_-(p), \quad 0 < p \leq 1/2 \tag{6}$$

is sufficient to satisfy the requirement on the gains. By (3) and (6), one has

$$a_+(p)/a_-(p) = (p^{-1} - 1)^2 \geq 1, \quad p \leq 1/2.$$

If $a_+(p)$ as a measure of complexity decreases on (0, 1), it follows that the requirement on the gains is satisfied without (6): $a_+(p) \geq a_+(1 - p) = a_-(p)$ when $p \leq 1/2$.

The requirement (6) is convenient, because $v(p) = 0$, so R is

$$R(X, Y) = \sum_{i \geq 1} w_i (x_i - p_i)(y_i - p_i), \quad w_i = w(p_i), \tag{7}$$

where $w(p)$ specifies all basis functions $\{a_\pm, b_\pm\}$:

$$w(p) = b_\pm(p)[p(1 - p)]^{-1} = a_+(p)(1 - p)^{-2} = a_-(p)p^{-2}. \tag{8}$$

Zechar & Zhuang (2010) use the score (2, 3) with $b_\pm(p) = 1$. In virtue of (8) the score is consistent with (7) when

$$w(p) = [p(1 - p)]^{-1}. \tag{9}$$

The representation (7) is of interest since it can be treated in purely geometric terms, that is, as a weighted correlation of the vectors $X - P$ and $Y - P$, where $P = (p_1, \ldots, p_N)$ is the mean of Y assuming (1). The weight w_i may reflect the degree of importance of the i-th prediction or the complexity involved in guessing the result $y_i = 1$ when $p \leq 1/2$. Indeed, using (8) one has

$$a_+(p) \leq w(p) \leq 4a_+(p), \quad 0 \leq p \leq 1/2,$$

that is, $a_+(p)$ and $w(p)$ are roughly equivalent. Note that the terms in R have a twofold representation:

$$w_i(x_i - p_i)(y_i - p_i) = w_i(\bar{x}_i - \bar{p}_i)(\bar{y}_i - \bar{p}_i).$$

Consequently, the $\{w_i\}$ in (7) should be interpreted in the same manner with respect to the prediction of the sequences $\{y_i\}$ and $\{\bar{y}_i\}$.

2.2 Models of the GS R

The weight (9) too heavily emphasizes the complexity of guessing rare events. An alternative is the following parametric family of $w(p)$:

$$w_\beta(p) = [4p(1 - p)]^{-\beta}, \quad \beta \geq 0, \tag{10}$$

which includes (9) ($\beta = 1$) and $w(p) = 1$ ($\beta = 0$). The family is normalized by the requirement $w_\beta(1/2) = 1$; it possesses the monotone property

$$w_\beta(p) \le w_{\beta'}(p), \quad 0 \le \beta \le \beta'.$$

Due to (8), this monotone property holds true for the bets $b_\pm(p)$ and for the return functions $a_\pm(p)$.

The weight

$$\tilde{w}_\beta(p) = 1 - \beta \ln[4p(1-p)] \qquad (11)$$

for any $\beta \ge 0$ lies between $w_\beta(p)$ and $w_0(p)$, that is, $w_0 \le \tilde{w}_\beta \le w_\beta$.

When p is small, the dominant part of $\tilde{w}_\beta$ is proportional to $\ln(1/p)$, that is, to the Shannon information, which resides in an event occurring with probability p. This is the reason why (11) may be interesting for the analysis of prediction results as well.

For the family (10) one has

$$a_+(p) = cp^{-\beta}(1-p)^{2-\beta}, \quad b_+(p) = c\,[p(1-p)]^{1-\beta}.$$

The function $a_+(p)$ as a complexity measure must decrease on (0,1); therefore one has $0 \le \beta \le 2$. The behaviour of the bet functions $b_+(p)$, $0 < p < 1/2$ is twofold: $b_+(p)$ increases for $0 \le \beta < 1$ and decreases for $1 < \beta < 2$. In the first case we stimulate the prediction of rare events, while in the second we realize the following game principle: the higher the return ratio [see (3)] the higher the bets. We will show that the models $w_\beta(p)$ with small β lead to a more stable statistical analysis of predictions.

Note that R_w with $w = w_{1/2}(p)$ and $w = w_1(p)$ admits of another statistical treatment. When $\beta = 1/2$, the weights w_i normalize the stochastic terms $(y_i - p_i)$ of R because the quantities

$$w_i(y_i - p_i) = (y_i - p_i)/\sqrt{p_i(1-p_i)} := y_i^{\text{norm}}$$

have zero means and unit variances under condition (1). For the same reason, when $\beta = 1$ and X, Y have the distribution (1), the score $R(X, Y)$ is the correlation of the normalized vectors $\{x_i^{\text{norm}}\}$ and $\{y_i^{\text{norm}}\}$, where $x_i^{\text{norm}} = (x_i - p_i)/\sqrt{p_i(1-p_i)}$. More generally, by (8) one has

$$R_w(x, y) = \sum_{i \ge 1} b(p_i) x_i^{\text{norm}} y_i^{\text{norm}}, \quad b(p) = b_\pm(p). \qquad (12)$$

The representations (7) and (12) are of interest as regards the question of which model of $w(p)$ should be considered 'natural'. According to the information arguments, this may be $\tilde{w}_\beta(p)$, while the statistical interpretation of (12) leads to the $w_1(p)$ for which $b(p) = 1$ in (12). This indeterminacy cannot be removed without additional argumentation.

The GS approach is far from being the only method for choosing the appropriate measure to assess the performance of a prediction algorithm. As an illustration we consider an example that is conceptually similar to the 'entropy score' (see Vere-Jones 1998).

The likelihood of y_i, as well as that of $\bar{y}_i = 1 - y_i$, has the form

$$l_i = y_i \log p_i + (1 - y_i)\log(1 - p_i).$$

The quantity $(-l_i)$ has the meaning of the amount of information the observer gets from the i-th prediction experiment. The goal of the forecaster may be formulated as follows: the sum $\sum_i (-l_i)$ should be maximized for those alarm zones where target events are to be expected and minimized where such events are not expected. Hence the goal function may be defined as

$$R_{LH}(X, Y) = \sum_{i \ge 1} (x_i - \bar{x}_i)(-l_i) = \sum_{i \ge 1} (x_i - 1/2)(y_i - 1/2) \times 2\ln\left(p_i^{-1} - 1\right) + c, \qquad (13)$$

where c is independent of $\{y_i\}$. Comparison of (7) with the right-hand side of (13) reveals differences in the centring of the X vector and in the evenness of the weight functions. Nevertheless, both of these goal functions can be used in prediction analysis. To make this clear, suppose that $p_i = p < 1/2$ for alarms of both types. This case corresponds to the H_0 hypothesis in the prediction of a subsequent strong earthquake by the SSE method (Vorobieva 1999). One has

$$R_{LH} = (\nu_{++} - 0.5N_y) \times 2\ln(p^{-1} - 1) + c$$

and

$$R_w = (\nu_{++} - pN_y)w(p) + c_1,$$

where ν_{++} is the number of successful positive alarms, $N_y = \#\{y_i = 1\}$ is the total number of alarm zones with target events, and c, c_1 are functions of $\{x_i\}$. Now the structural similarity of the scores is obvious.

If N_y and $\{x_i\}$ are fixed, the scores depend on ν_{++} only. Therefore the analysis of prediction results in the conditional situation will rely on the conventional statistic of successes. Note that in this case the distribution of ν_{++} does not depend on p and is the hypergeometric distribution with the parameters (N_+, N_y).

$$P(\nu_{++} = k) = C_{N_+}^{k} C_{N-N_+}^{N_y - k} / C_N^{N_y},$$

where N_+ is total number of positive alarms and C_a^b are the binomial coefficients. It is an ideal situation for analysis of the prediction results when distribution of the key statistic does not depend on unknown parameters. This is not the case in the generic situation.

2.3 The GS and the significance of prediction results

Let $\hat{X}$ and $\hat{Y}$ be samples of X and Y, respectively. The quantity $R(\hat{X}, \hat{Y})$ can be used to estimate the significance of prediction results. First, one has to specify the distribution of Y. We assume that the components of Y are independent and their distribution follows (1) (the H_0 hypothesis). The greater the value of $R(X, Y)$, the better is the prediction method. Therefore, the probability

$$\alpha_y = P_y(R(\hat{X}, Y) \ge R(\hat{X}, \hat{Y})), \qquad (14)$$

where P_y is the distribution of Y under H_0, provides the observed significance level for the prediction results.

Obviously, one has

$$\alpha_y = P\{\xi_R \ge \hat{\xi}_R\},$$

where ξ_R is a linear function of $\{y_i\}$, while $\hat{\xi}_R$ is the observed value of ξ_R.

To be more specific,

$$\xi_R = \sum_{i \ge 1} c_i y_i, \qquad (15)$$

where $c_i = (\hat{x}_i - p_i)w(p_i)$ for $R = R_w$, and $c_i = (2\hat{x}_i - 1)\ln[(1 - p_i)/p_i]$ for $R = R_{LH}$. The significance level α_y can be found from the distribution of ξ_R. Under H_0 this is the convolution of distributions $\{F_i\}$ of the type

$$\dot{F}_i(u) = \delta(u)(1 - p_i) + \delta(u - c_i)p_i,$$

where $\delta(\cdot)$ is the delta function. The convolution operation is convenient for numerical computations and for checking the accuracy of the distribution.

If the scatter of the $\{c_i\}$ is not too large (the case of R_w with $p_i > p_0 > 0$), the significance of the R statistic can be roughly inferred from large values of the normalized quantity ξ_R, i.e.

$$\xi_R^{\text{norm}} = (\xi_R - m_R)/\sigma_R, \qquad (16)$$

where m_R and σ_R^2 are the mean and variance of ξ_R.

$$m_R = \sum_{i\geq 1} c_i p_i, \quad \sigma_R^2 = \sum_{i\geq 1} c_i^2 p_i(1-p_i). \tag{17}$$

The estimation of α_y based on R needs two general comments.

(1) α_y may strongly depend on the choice of the $\{w_i\}$. Therefore a serious argumentation for the choice of R and, in particular, of $\{w_i\}$, is required.

(2) Usually the parameters $\{p_i\}$ for large events are small and their estimates are inaccurate, hence interval estimates of α_y are required. The problem is difficult for analytical solution, because the $\{p_i\}$ are involved in the distribution of ξ_R (see 15) both through the distribution of Y and through the coefficients $\{c_i\}$.

3 THE GS AND COMPARISON OF PREDICTION METHODS

Zechar & Zhuang (2010), to be referred hereinafter as [ZZ], use the GS R to compare a prediction method of interest X with the random guessing Z. This point should be discussed.

Let $A_i = G_i \times T_i$, $i = 1, \ldots, N$ be alarms of X. If the target events occurrence is Poissonian and stationary, the p parameters of the H_0 hypothesis can be specified as follows:

$$p_i = P\{y_i = 1\} = 1 - \exp(-\lambda_i T_i), \tag{18}$$

where λ_i is the rate of target events in G_i. In what follows the quantities (18) are referred as 'reference probabilities'. By definition, the distribution P_z of a random strategy Z is identical with the distribution of Y given by (18). To analyse the prediction performance, [ZZ] consider the random variable $R_w(Z, \hat{Y})$ instead of the $R(\hat{X}, Y)$ discussed above. The analogue of the significance (14) in this case is the following quantity:

$$\alpha_z = P_z\{R_w(Z, \hat{Y}) \geq R_w(\hat{X}, \hat{Y})\}. \tag{19}$$

If α_z exceeds a nominal level α_0, a practical conclusion may sound as follows: the method X looks no better than random guessing, that is, $X \prec_R Z$ or, which is more accurate, the GS R does not detect the preference of $\hat{X}$ compared with its randomized version Z (recall that Z and $\hat{X}$ have the same alarm zones). The conclusion is related to the observed seismicity $\hat{Y}$ only. This point is quite unusual for statistical analysis.

If we cancel the condition $Y = \hat{Y}$, then we arrive to a new characteristic of significance type.

$$\alpha_{zy} = P_{zy}\{R_w(Z, Y) \geq R_w(\hat{X}, \hat{Y})\}. \tag{20}$$

Here Z and Y under H_0 are random and independent, that is, their probability measure P_{zy} is the product measure of P_z and P_y. By (12), $R_w(Z, Y)$ under H_0 has zero mean and the variance $\sigma_{zy}^2 = \sum_{i\geq 1} b_i^2$. Hence, a rough conclusion like $X \prec_R Z$ which is related now to an arbitrary sample of Y may follow from the simple relation

$$R_w(\hat{X}, \hat{Y}) < k\sigma_{zy},$$

where $k \approx 2$. For the [ZZ] model of R one has $b_i = 1$, therefore σ_{zy}^2 is equal to the total number of alarms N.

It is important that the quantities α_y, α_z and α_{zy} are not necessarily correlated when different models of R_w are considered.

Indeed, let S be the space of vectors (Z, Y) with the product measure $P_{zy} = P_z \circ P_y$. Suppose $R_w(\hat{X}, \hat{Y}) = c$ and denote by U_c the subset of S where $R_w \geq c$. Then α_{zy} is the P_{zy} measure of S, while α_y and α_z are the conditional measures of sections of S given by the relations $Z = \hat{X}$ and $Y = \hat{Y}$, respectively. We illustrate our statement by an example of this construction.

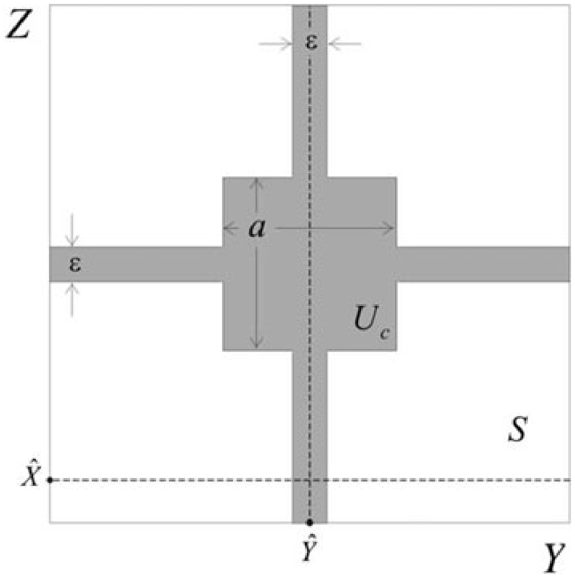

Figure 1. Example illustrating the relationship between α_{zy}, α_y and α_z (see Section 3 for details).

Let $S = [0, 1]^2$ be a square with the uniform measure. Suppose that U_c is a cross-like centrally symmetric figure as that shown in Fig. 1. The set U_c depends on two parameters, ε and a, where ε is small and $\varepsilon < a < 1$. Then one has $\alpha_{zy} = a^2 + 2(1-a)\varepsilon$ and α_y can assume the values 1, a or ε. Due to the symmetry of U_c the same is valid for α_z. It is clear that the parameters (ε, a) allow us to generate a vector $(\alpha_{z,y}, \alpha_y, \alpha_z)$ in which any pre-assigned components are small while the others are not. Fig. 1 shows an example with $\alpha_y = \varepsilon$, $\alpha_z = 1$, and therefore $\alpha_y < \alpha_{zy} < \alpha_z$.

Thus in the general situation, 'the conclusion like $X \prec_R Z$ does not mean that the method X is trivial', because α_y may be small.

[ZZ] apply their method of the comparison to the RTP algorithm (Shebalin *et al.* 2004, 2006). In this application

(i) the positive RTP alarms $A_i = G_i \times T_i$ are considered only, therefore $\hat{X} = (1, \ldots, 1)$;

(ii) the space–time alarm zones of the random strategy $Z = (z_1, \ldots, z_N)$ coincide with the RTP alarms: $T_i \approx 9$ months, the areas G_i do not have regular shapes, they represent the union of standard local areas determined by current and past seismicity, that is, the alarm zones of Z are fixed and not involved in the simulation. Therefore the solution $z_i = 1$ means only that the i-th RTP alarm A_i remains in force, while the solution $z_i = 0$ converts the positive i-th alarm into a negative one.

As a result, α_z shows how efficient the random mechanism of the alarm cancellation could be: it is ineffective if $\alpha_z < \alpha_0$ and non-trivial otherwise. According to [ZZ], the RTP alarms admit of different options and interpretations. The so-called 'loose' interpretation involves six successful alarms and gives $\alpha_z \approx 0.0001$ for the R_w model (9). In contrast to this, the more 'strict' option involving two successes only gives $\alpha_z \approx 0.94$. Unfortunately, so divergent estimates of α_z can tell us nothing about the expediency of the random cancellation of positive RTP alarms. On the other hand, either of the α_z estimates is irrelevant to the significance of the RTP prediction results.

It is our opinion that the statistical analysis of any prediction method with few target events and a short monitoring period is premature (this is the case of RTP). For this reason we will analyse the significance α_y for the M8 algorithm prediction results. For the sake of simplicity we use the notation α for the quantity α_y.

4 SIGNIFICANCE ANALYSIS OF THE M8: THE CONVENTIONAL APPROACH

Descriptions of the M8 prediction algorithm can be found in Keilis-Borok & Kossobokov (1990), Kossobokov *et al.* (1999), Kossobokov & Shebalin (2003). A statistical analysis of the M8 prediction results was considered recently by Molchan & Romashkova (2010) using conventional methods. This enables us to provide a comparative analysis of the GS and conventional approaches based on a standardized use of earthquake catalogues.

We consider the M8 prediction results for $M \geq 8.0$ events. This prediction has been conducted since 1985, but we will only discuss the period of forward testing of the M8 algorithm, that is, 1992–2009 (see http://www.mitp.ru/en/predlist.html). The monitoring space G consists of a set of overlapping circles B_R of radius $R = 668$ km located along the Circum–Pacific and Alpine–Himalayan belts. Any circle is permanently in an alarm/non-alarm state. The states are revised for all circles simultaneously at intervals of 6 months, $\Delta t = 0.5$ yr. The union of the circles that have been in a state of alarm during 6 months Δt_i will be treated (in the GS approach) as a single domain of positive alarm, $A_i^+ = G_i \times \Delta t_i$; the domain $A_i^- = G_i^c \times \Delta t_i$ where G_i^c is the complement of G_i in G will be considered as a negative alarm.

Table 1 summarizes some results from our analysis of the M8 algorithm for predicting $M \geq 8.0$ events (see Molchan & Romashkova 2010). Some comment is in order.

The similarity principle, which is the basis of the M8 algorithm, requires specifying the magnitude range of the target $M \geq 8.0$ events. We consider two options: $8.0 \leq M < 8.5$ and $8.0 \leq M < 8.7$; of these two, the former is the more logical, since there is a special version of the M8 algorithm for predicting the $7.5 \leq M < 8.0$ events and this version uses the same similarity principle.

The prediction results for the monitoring period $T = 1992-2009$ involve the following:

(i) The number of target events N_e and the number of predicted events ν_e^+. For the option $8.0 \leq M < 8.5$ one has $\nu_e^+/N_e = 10/18$;

(ii) A normalized measure of space–time alarm τ,

$$\tau = \lambda(A^+)/\lambda(G \times T), \tag{21}$$

where $\lambda(A)$ is the expected number of target events in the space–time domain A, A^+ is the union of all positive alarms $\{A_i^+\}$, and $G \times T$ is the entire space–time prediction volume. We discuss two estimates for τ: $\hat{\tau} = 32.5$ per cent and $\tilde{\tau} = 35.4$ per cent.

Because the target events are few, the estimate of $\lambda(\cdot)$ is based on the hypothesis that the $M \geq M_-$ seismicity is stationary and on the regional Gutenberg–Richter (G–R) relations for $M_- \leq M \leq 8.0$ (see for details Molchan & Romashkova 2010). As a result, we find for τ a point estimate $\hat{\tau}$ and an upper estimate $\tilde{\tau}$ with confidence level 99 per cent. Both of these estimates are stable with respect to magnitude type (M_w or Ms) and to the threshold M_-above which all events are reported completely.

The choice of M_- is influenced by two opposite tendencies: with increasing M_- the G–R law hypothesis becomes more likely; at the same time the amount of available data $N_\lambda = \#\{M \geq M_-\}$ decreases, thereby making the uncertainty of τ larger. Our analysis (see Molchan & Romashkova 2010) shows that the following restriction on N_λ is reasonable:

$$N_\lambda/\text{area}\, G > 100/\text{area}\, B_R, \tag{22}$$

where B_R is the space unit of M8 alarms, that is, a circle of radius R.

We use M_w as the magnitude most consistent with the G–R law for the range $(M_-, 8.0)$. The resulting estimates $\hat{\tau} = 32.5$ per cent and $\tilde{\tau} = 35.4$ per cent are obtained with $M_- = 5.5$ and $N_\lambda = 8500$. The data are from the Centroid Moment Tensor catalogue, 1977–2004 (Ekstrom *et al.* 2005).

The significance of prediction results is based on the conditional distribution of ν_e^+ given N_e. Under the Poisson hypothesis for the target events, ν_e^+ has the binomial distribution with parameters (N_e, τ), which gives the significance level

$$\alpha = P\left\{\nu_e^+ \geq \hat{\nu}_e^+ | N_e = \hat{N}_e\right\},$$

where $\hat{\nu}_e^+$, $\hat{N}_e$ are samples of ν_e^+, N_e. The estimate $\tau = \hat{\tau}$ leads to the point estimate $\alpha = \hat{\alpha}$, while $\tilde{\tau}$ leads to the upper estimate $\alpha \leq \tilde{\alpha}$. For the [8.0, 8.5) option one has $\hat{\alpha} = 3.7$ per cent and $\tilde{\alpha} = 6.4$ per cent.

Since $\hat{\nu}_e^+$ and $\hat{N}_e$ are small for the forward M8 monitoring, the upper estimate $\tilde{\alpha}$ is unstable over time. In fact, a next target event in the monitoring zone will modify the pair $(\hat{\nu}_e^+, \hat{N}_e)$ to become $(\hat{\nu}_e^+, \hat{N}_e + 1)$ or $(\hat{\nu}_e^+ + 1, \hat{N}_e + 1)$. As a result, the estimate $\tilde{\alpha} = 6.4$ per cent will become 9.4 per cent or 3.8 per cent, respectively.

More information on the issues here discussed can be found in Molchan & Romashkova (2010).

5 SIGNIFICANCE ANALYSIS OF THE M8: GS APPROACH

5.1 General remarks

For the monitoring period $T = 1992 - 2009$ we have $N_+ = 36$ positive alarms $\{A_i^+\}$ and the same number $N_- = 36$ of negative ones $\{A_i^-\}$, each lasting 6 months. Almost all alarms are not simply connected in space; this circumstance will be discussed later. A positive alarm A_i^+ is treated as successful if it contains at least one target event; a negative alarm A_i^- is treated as successful if it does not contain target events. For this reason in the GS approach the number of successful positive alarms ν_{++} and the total number of alarms that cover target events, N_y, are not necessarily equal to the

Table 1. Significance level α for M8 prediction results based on the number of predicted events ν_e^+.

Period, T	Target events	ν_e^+/N_e	$\hat{\tau}$	$\hat{\alpha} \cdot 100$ per cent	$\tilde{\tau}$	$\tilde{\alpha} \cdot 100$ per cent	$\hat{\alpha}_{PS} \cdot 100$ per cent
1992–2009	$8.0 \leq M < 8.5$	10/.18	0.325	3.7	0.354	6.4	3.2
1992–2009	$8.0 \leq M < 8.7$	11/21	0.325	4.7	0.354	8.3	4.0

Notes: N_e is the number of target events; $\hat{\tau}$ and $\tilde{\tau}$ are a point estimate and an upper estimate, respectively, for the normalized measure of space–time alarm τ; $\hat{\alpha}$ and $\tilde{\alpha}$ are a point estimate and an upper estimate of α, respectively; $\hat{\alpha}_{PS}$ is a point estimate of α based on $Ms \geq 8.0$ earthquakes from the catalogue by Pacheco and Sykes (1992), $\#\{Ms \geq 8.0\} = 63$.

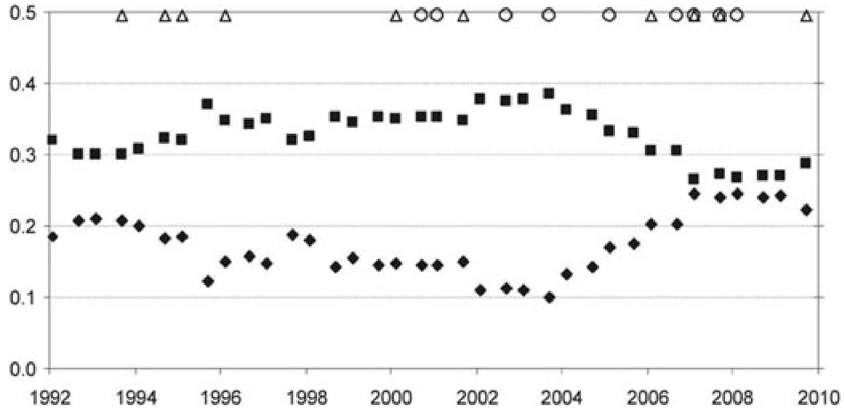

Figure 2. The reference probability of target event for time sequences of 6-month M8 alarms ordered in time: positive alarms (p^+, diamonds), and negative alarms (p^-, squares). Notation: Alarms that have covered target events are marked on the upper horizontal axis: successful positive alarms (triangles) and false negative alarms (circles).

number of predicted events ν_e^+ and the number of all target events N_e, respectively. In particular, for the option $8.0 \le M < 8.5$ we have $\nu_e^+/N_e = 10/18$ and $\nu_{++}/N_y = 10/17$.

The reference probabilities $p_i^\pm$ for the alarms $A_i^\pm$ are found from (18) using the rates of target events. The method for estimating the rates has been explained above. The alarms $A_i^\pm$ are ordered in time, so the plots $i \to p_i^\pm$ are shown as functions of time in Fig. 2.

Histograms for $\{p_i^+\}$ and $\{p_i^-\}$ (Fig. 3a) give an idea of the ranges of $p^\pm$: (0.1, 0.25) for p^+ and (0.26, 0.39) for p^-, as well as demonstrate typical values of these quantities. It is important for the subsequent argument that the reference probabilities for positive M8 alarms are bounded away from zero: $p_i^+ \ge 0.1$ for all alarms, and $p_i^+ \ge 0.15$ for successful ones.

We use the GSs R_{LH} and R_w with the following models of w:

$$w_0 \le \tilde{w}_{1/2} \le w_{1/2} \le w_1 \le w_{3/2}, \tag{23}$$

see (10) and (11). According to Section 2.3, the significance level α based on the GS statistic is

$$\alpha = P(\xi_R \ge \hat{\xi}_R), \tag{24}$$

where $\hat{\xi}_R$ is an observed value of ξ_R. Now ξ_R is $\xi_R^+ - \xi_R^-$, ξ_R^+ or $-\xi_R^-$ if all, positive or negative alarms are considered, respectively. Here

$$\xi_R^\pm = \sum_{i\ge1} c^\pm(p_i^\pm) y_i^\pm, \tag{25}$$

where $y_i^\pm = 1$, if $A_i^\pm$ contains a target event and $y_i^\pm = 0$ otherwise. The functions $c^\pm(p)$ are

$$R_w: \quad c^+(p) = w(p)(1-p), \quad c^-(p) = w(p)p \tag{26}$$

and

$$R_{LH}: \quad c^+(p) = c^-(p) = \ln[(1-p)/p]. \tag{27}$$

5.2 The relationship between α and $w(p)$: a theoretical analysis

The functions $c^\pm(p)$ for weights (23) are shown in Fig. 4. The $c^-(p)$ vary slowly for $\beta \le 1$ or $\beta \le 3/2$, $p^- \ge 0.05$, and are little sensitive to the model of w, especially in the range of $\{p_i^-\}$ for the M8 alarms. It follows that the statistic $\xi_{R_w}^-$ must be little sensitive to the choice of w.

The behaviour of $c^+(p)$ is essentially different. Near $p = 0$ we have a rapid decrease of $c^+(p)$ like $O(p^{-\beta})$ for $w = w_\beta$. Far away from $p = 0$ both the decrease of c^+ and its dependence on the w model are substantially weaker. Therefore the estimates of α may depend on the behaviour of w and on the distribution of $\{p_i^+\}$ in the range (0, 1/2). The following model situation illustrates this statement.

We begin by considering the case of positive alarms. For the monotone family of weights $w_\beta(p)$ the functions $c^+(p)$ increase with β (see Fig. 4). Hence the statistic $\xi_{R_w}^+ = \sum c^+(p_i^+)y_i^+$ increases with β as well because y_i are non-negative. Suppose there exist two non-intersecting intervals $I = (0, a)$ and $J = (b, 1)$ which cover the set $\{p_i^+\}$, and $c^+(b) << c^+(a)$ when the parameter β is close to β_0, say 1. Consider two possibilities. The first: the observed values are $\hat{y}_i^+ = 0$ for all p_i^+ from the interval I. Then the observed statistic $\{\hat{\xi}_{R_w}^+\}$ as the function of the parameter β is almost constant for all β nearby β_0 and therefore the probability of the event $\{\xi_{R_w}^+ \ge \hat{\xi}_{R_w}^+\}$ must increase with increasing β. Consequently, the significance of prediction results will be the worst for the model w_β with the greatest β.

The second possibility: $\hat{y}_{i0}^+ \ne 0$ for some $p_{i0}^+ < a$. For the sake of simplicity we assume that the interval I contains a single point, p_{i0}^+. Due to $c^+(b) \ll c^+(a)$, the event $\{\xi_{R_w}^+ \ge \hat{\xi}_{R_w}^+\}$ when the parameter β is close to β_0 can only occur if $y_{i0}^+ = 1$. In other words, the probability of this event will be

$$\alpha \approx P\left\{c^+(p_{i_0}^+)y_{i_0}^+ \ge c^+(p_{i_0}^+)\right\} = p_{i_0},$$

that is, when β is large, the significance of the R statistic is controlled by a single successful alarm with a very small p.

By Fig. 4, $\xi_{R_w}^-$ is weakly dependent on w for $\beta \le 1$ or $\beta \le 3/2$, $p^- \ge 0.05$. Therefore, under these conditions all considerations outlined earlier remain valid for the statistic $\xi_{R_w}^+ - \xi_{R_w}^-$ as well. It remains to verify our heuristic arguments by estimating α.

5.3 α estimates for M8 alarms

A rough idea of α is provided by the normalized quantity ξ_R, that is, ξ_R^{norm} (see 16). For example, the inequality $\xi_R^{\text{norm}} > 2$ argues in favour of the prediction results being significant. The exact significance estimates are based on the distribution of ξ_R under H_0 (see Section 2.3). The estimates of α and ξ_R^{norm} are summarized in Table 2. They are given for different options: two magnitude ranges of target events; three types of alarms: positive (+), negative (–) and all (+/–); six models of the GS, namely, R_{LH} and five models of R_w with w given by (23).

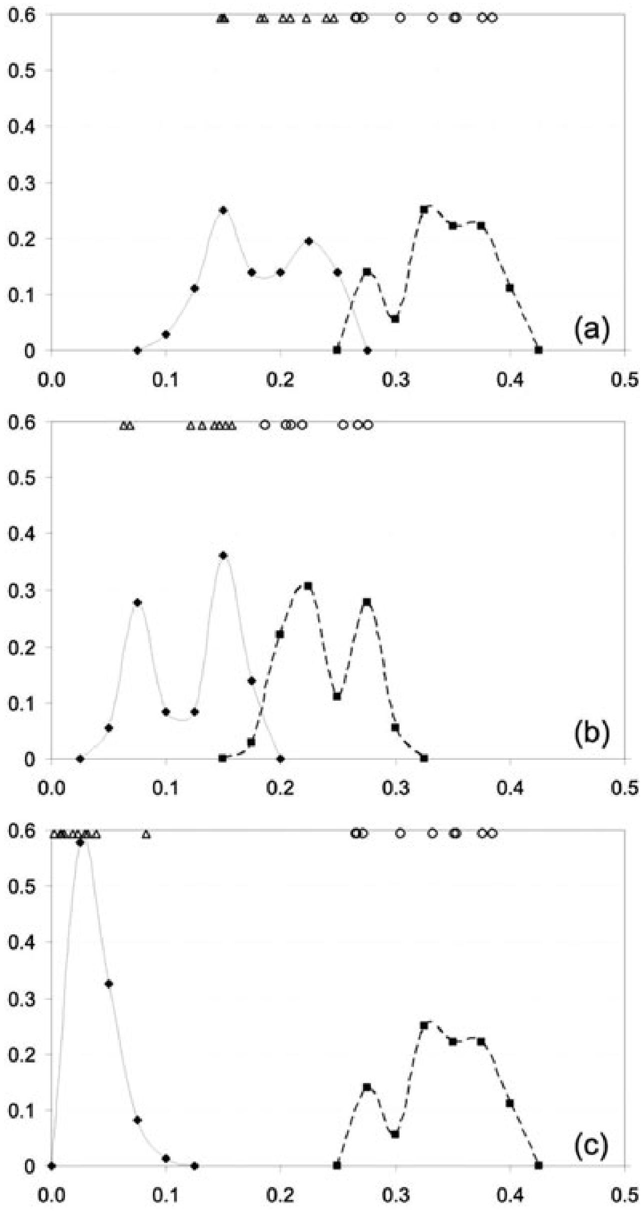

Figure 3. The normalized histograms of the reference probabilities $\{p_i\}$ for M8 alarms: positive alarms (diamonds) and negative alarms (squares). The estimates of $\{p_i\}$ are related to (a) 72 alarms and $M_w \geq 5.5$ (see Fig. 2), (b) 72 alarms and $Ms \geq 8.0$ events from the catalogue by Pacheco & Sykes (1992) and (c) 328 alarms and $M_w \geq 5.5$. Notation: the upper horizontal axes mark the reference probabilities for the alarms that cover target events: successful positive alarms (triangles) and false negative alarms (circles).

Conclusions from Table 2 are as follows:

(1) Like-sign M8 alarms are not significant: for the $8.0 \leq M < 8.5$ option and for any model of R, the (+) and (–) alarms give $\xi_R^{\rm norm} < 2$; $\alpha = 7 - 14$ per cent for positive M8 alarms. As was to be expected, $\xi_{R_w}^{\rm norm}$ for negative alarms is nearly independent of $w(p)$, $\xi_{R_w}^{\rm norm} \approx 1.8$.

(2) For all types of alarm, (+), (–) and (+/–), the significance of prediction results grows as w_β varies from $w_{3/2}$ to w_0, that is, $\xi_R^{\rm norm}$ increases while α decreases. This is in agreement with our theoretical analysis of the case where the positive alarm with the smallest reference probability p is false (see 5.2).

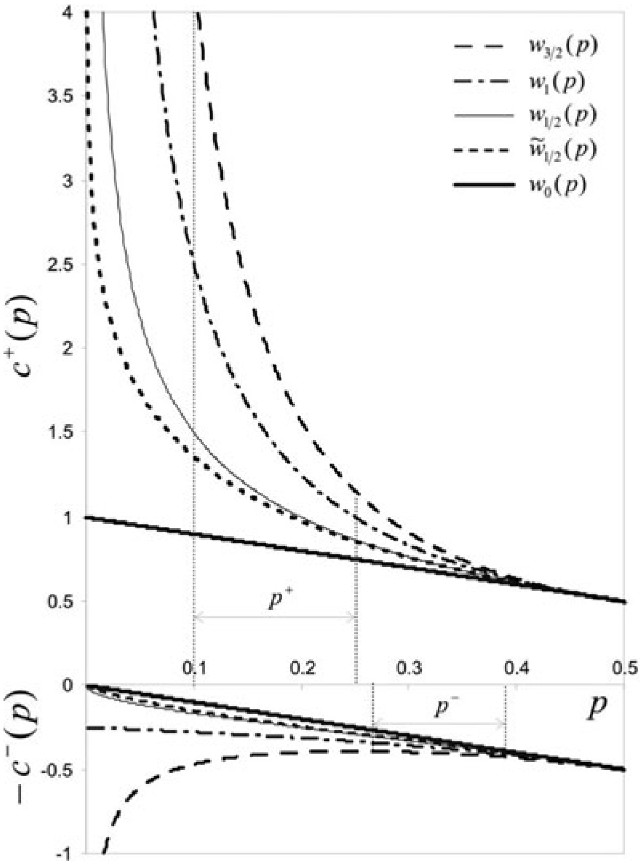

Figure 4. Functions $c^+(p)$ (top panel) and $-c^-(p)$ (bottom panel) for different models of the alarm weight function $w(p)$. Vertical dotted lines mark the range of the reference probabilities for M8 alarms: negative alarms (p^-, $N_- = 36$) and positive alarms (p^+, $N_+ = 36$).

(3) Of the two ΔM options, the lowest values of α are obtained for $8.0 \leq M < 8.5$. This fact supports the preliminary arguments in favour of the [8.0,8.5) option (see 5.1) and is in agreement with a similar conclusion in (Molchan & Romashkova 2010) based on the observed total number of M8 successes (see Table 1).

(4) The α for the $8.0 \leq M < 8.5$ option and $\beta \leq 1$ varies between 1.6 per cent ($w = w_0$) and 4.8 per cent ($w = w_1$), thus confirming that the M8 algorithm is non-trivial; the simple statistic ν_e^+ gives a similar point estimate, $\alpha = 3.7$ per cent (see Table 1).

5.3.1 *Rapidly changing weights $w(p)$*

We can see (Table 2) that the use of $w(p)$ rapidly changing near $p = 0$ leads to less favourable estimates of α. This can be explained by the loss of information when the weights of the alarms are essentially different in size, for example, the sequence of weights (1, 0, 0...) takes into account the first alarm only. Such effect we can observe for $\beta > 1$: the $w_{3/2}$ model results in $\alpha \geq 7.9$ per cent and thus casts doubt on the M8 method. However, for large β the relation between α and $w_\beta(p)$ is unstable. Consider the following numerical experiment.

Suppose i_0 is the ordinal number of the M8 positive successful alarm having the smallest reference probability p^+ among all successful ones. It is the alarm for the first 6 months of 2000, $p_{i_0}^+ = 0.15$ (see Fig. 2). Suppose the parameter has been revised to make $p_{i_0}^+ = 0.05$. This uniquely specifies $p_{i_0}^-$, since for any i

$$(1 - p_i^+)(1 - p_i^-) = \exp(-\Lambda \cdot T_i),$$

Table 2. Significance level α for M8 prediction results based on the gambling score: R_w and R_{LH}.

Score	$8.0 \le M < 8.5$, $\nu_{++}/N_y = 10/17$							$8.0 \le M < 8.7$, $\nu_{++}/N_y = 10/19$
	ξ_R^{norm}			$\hat{\alpha} \cdot 100$ per cent		$\hat{\alpha}_{PS} \cdot 100$ per cent		$\hat{\alpha} \cdot 100$ per cent
	(–)	(+)	(+/–)	(+)	(+/–)	(+)	(+/–)	(+/–)
(R_w, w_0)	1.85	1.60	2.26	7.3	1.6	0.3	0.2	3.3
$(R_w, \tilde{w}_{1/2})$	1.82	1.49	2.08	7.6	2.4	0.3	0.3	4.3
$(R_w, w_{1/2})$	1.81	1.46	2.02	8.0	2.7	0.4	0.4	4.7
(R_w, w_1)	1.78	1.29	1.75	10.5	4.8	0.9	0.8	7.0
$(R_w, w_{3/2})$	1.74	1.10	1.47	13.9	7.9	2.1	1.9	10.3
R_{LH}	1.37	1.41	1.90	8.7	3.4	0.4	0.6	5.9

Notes: Alarms: positive (+), negative (–) and all (+/–); ν_{++}/N_y is the number of successful positive alarms versus the total number of alarms that cover target events; $\hat{\alpha}$ is a point estimate of α; $\hat{\alpha}_{PS}$ is a point estimate of α based on the $Ms \ge 8.0$ earthquakes from the catalogue by Pacheco & Sykes (1992); ξ_R^{norm} is a normalized statistic related to α (see 16 and 17 in the text).

Table 3. Significance level α for the M8 predictions of $8.0 \le M < 8.5$ events (72 alarms) depending on the reference probability $p_{i_0}^+$ for a particular positive alarm, i_0.

Score		(R_w, w_0)	R_{LH}	(R_w, w_1)	$(R_w, w_{3/2})$
$\hat{\alpha} \cdot 100$ per cent	$p_{i_0}^+ = 0.15$	1.6	3.4	4.8	7.9
	$p_{i_0}^+ = 0.05$	1.2	1.8	1.0	0.8

where $\Lambda = \lambda(G)$ is the rate of target events in G. Suppose the other reference probabilities remain unchanged. Then one has $p_{i_0}^+ = \min_i p_i^+$. This case has been considered in Section 5.2 theoretically. As was to be expected, the α estimate based on w_β, is the smallest now for the model $w = w_{3/2}$, namely 0.8 per cent instead of the former 7.9 per cent (see Table 3). In other words, the original conclusion that the prediction results are not significant becomes its opposite.

At the same time, the α estimates based on w_β with small β or on the information type scores look stable, for example, for the case $w = w_0$ one has $\alpha = 1.2$ per cent (the hypothetical $p_{i_0}^+$) as against $\alpha = 1.6$ per cent (the original $p_{i_0}^+$).

5.3.2 Simply connected M8 alarms

We have considered one of the possible variants of subdivision of the M8 alarm volume into isolated parts using the half-year layers of the volume. In this case the numbers of positive and negative alarms are the same and equal to 36. Taking into account that these alarms are not simply connected in space, we can improve the situation considering the following alternative subdivision: it consists of all the half-year simply connected positive subalarms ($N_+ = 292$) plus the previous set of negative alarms ($N_- = 36$).

The additional splitting of the alarms can cause the following difficulties:

(1) The greater the number of alarms the more often the property of independence of target events in the alarm zones (H_0hypothesis) is used to estimate the significance of predictions. A substantial increase in the number of alarms can have a negative influence on the validity of the α. In our case we have 328 alarms versus 72 in the previous variant.

(2) The numbers $N_\lambda(A_i)$ of $\{M \ge M_-\}$ events used for estimating the reference probabilities have changed considerably: for the positive alarms one has $N_\lambda(A_i) = 15 - 1433$ versus $2412 - 4156$ in the previous variant. The diminution of $N_\lambda(A_i)$ affects the accuracy and values of the reference probabilities.

(3) The values of the reference probabilities has been displaced towards 0: the range of p for the positive alarms is 0.001–0.1 (Fig. 3c) versus the previous 0.1–0.3. (Fig. 3a). The same holds for the reference probabilities for successful positive alarms; the range of p is 0.001–0.08 versus the previous 0.15–0.25. For such a case our previous analysis predicts a possible instability of α.

Table 4 compares the α estimates for the cases of 72 and 328 alarms in the prediction of $8.0 \le M < 8.5$ events. We have a quite good stability of the α estimates for scores of information type, that is, for $(R_w, \tilde{w}_{1/2})$ and R_{LH}, and its instability for the weights w rapidly changing near $p = 0$. In particular, the model $w_{3/2}$ results in the most unfavourable αestimate for the case of 72 alarms and very good for 328 alarms, namely 7.9 per cent and 0.6 per cent. For the model w_0 the situation is reverse but more stable: α is 1.6 versus 3.7.

5.4 The GS with unreliable $\{p_i\}$

The reference probabilities for $M \ge 8.0$ are based on the $M_w \ge M_- = 5.5$ earthquakes. These data can include aftershocks in addition to main shocks, hence the estimates of $\lambda_i = \lambda(G_i)$ and $p_i^\pm$ may be overestimated. There is a practice of estimating $\{\lambda_i\}$ directly based on past target events, even when G_i contains a single event (see e.g. [ZZ]). The uncertainty of α for ν_e^+ has been investigated in relation to the quantity $N_\lambda = \#\{M \ge M_-\}$ (Molchan & Romashkova 2010). A similar analysis for the R statistic is more difficult. For this reason we repeat the analysis of the M8 predictions using for estimation of $\{\lambda_i\}$ all $Ms \ge 8.0$ earthquakes for the period 1900–1984 from the Global Catalogue by Pacheco & Sykes (1992) with $N_\lambda = 63$. Such estimates are strongly advocated by Marzocchi *et al.* (2003).

Because N_λ is too small, the estimates of $\{\lambda_i\}$ are highly unreliable. Nevertheless the point estimates of τ, τ_{PS} remain practically unchanged; in particular, for the $8.0 \le M < 8.5$ option one has $\hat{\tau}_{PS} = 0.32$ versus the original $\hat{\tau} = 0.33$ (Table 1). As a result, the point estimates of α based on ν_e^+ are stable too: $\hat{\alpha}_{PS} = 3.2$ per cent versus $\hat{\alpha} = 3.7$ per cent (Table 1). At the same time, the αestimates based on the R_w statistics have diminished considerably: for the $8.0 \le M < 8.5$ option one has $\hat{\alpha}_{PS} = 0.2$ per cent versus $\hat{\alpha} = 1.6$ per cent (Table 2) with $w = w_0$, and $\hat{\alpha}_{PS} = 0.8$ per cent versus $\hat{\alpha} = 4.8$ per cent (Table 2) with $w = w_1$. Even when positive alarms only are considered, the new α estimates are rather optimistic, for example, $\hat{\alpha}_{PS} = 0.9$ per cent versus $\hat{\alpha} = 10.5$ per cent (Table 2) with $w = w_1$ and the $8.0 \le M < 8.5$ option.

Table 4. Significance level α for the M8 predictions of $8.0 \leq M < 8.5$ events depending on the gambling score model and subdivision of the M8 alarm volume into parts.

Score		(R_w, w_0)	$(R_w, \tilde{w}_{1/2})$	$(R_w, w_{1/2})$	(R_w, w_1)	$(R_w, w_{3/2})$	R_{LH}
$\hat{\alpha} \cdot 100$ per cent	72 alarms	1.6	2.4	2.7	4.8	7.9	3.4
	328 alarms	3.7	2.1	0.5	0.5	0.6	2.0

These results reveal a fundamental difference between the conventional and the GS approach. The first operates with the conditional distribution of ν_e^+ and hence with the relative characteristic τ (see 21). In other words, a rescaling of the rate parameters, $\lambda_i \to \rho\lambda_i$, does not affect τ. This transformation of $\{\lambda_i\}$ arises when the magnitude range of target events, $M \geq M_0$, is shifted by $\delta M \approx -\lg \rho$. If M_0 is fixed, the shift may have been caused by the fact that a different catalogue or a different magnitude type is used. That is why the τ estimates for the M8 alarms weakly depend on the choice of magnitude type and on the magnitude range $M > M_-$ used to estimate $\{\lambda_i\}$ (see Kossobokov 2005; Molchan & Romashkova 2010).

In contrast to this, the GS approach uses absolute values of $\{\lambda_i\}$. Therefore, the estimates of α may be sensitive to the transformation $\lambda_i \to \rho\lambda_i$. Fig. 3(b) shows the histogram of $\{p_i\}$ estimates based on $Ms \geq 8.0$ events from the Pacheco & Sykes (1992) catalogue. We can see that the range of $\{p_i^+\}$ is (0.05, 0.2) (Fig. 3b) versus the previous (0.1, 0.25) (Fig. 3a). As a result, we have more optimistic estimates of α (see $\hat{\alpha}_{PS}$ in Table 2).

6 CONCLUSION AND DISCUSSION

(1) We discussed a version of the GS approach in which the results of prediction are considered with weights that depend on reference probabilities of the alarms. A fair scoring scheme helps to reduce (but not entirely remove) the number of unknown functions in the GS. For this reason a serious argumentation for the choice of the GS model is required in each case of its application.

(2) On the basis of the GS Zechar & Zhuang (2010) suggested a method for comparison of predictions with their randomized version. It looks like a comparison of the prediction method with random guess. Theoretical arguments show that any outcome of such comparison cannot exclude the possibility that the prediction results may be statistically significant.

(3) The problem of earthquake predictability is still debatable. For this reason we apply the GS approach to the significance analysis of results from prediction of $M \geq 8.0$ events by the M8 algorithm.

Theoretical considerations and straightforward estimates of the significance level α show a strong dependence of α on the GS weight function model and on the distribution of alarm reference probabilities near zero. Both of these factors may affect the estimate of α in either direction and this can be exploited by the researcher. At the same time, the distribution of the reference probabilities is affected by:

(i) the partitioning of the entire alarm space–time into subareas, that is, into individual alarms A_i;

(ii) the method used to estimate the rate of target events in the alarms, $\{\lambda_i\}$; this involves the choice of the catalogue and the type of magnitude, the choice of small events to extrapolate the recurrence of target events and so on.

(4) All estimates of α based on the reasonable weight functions, w_β, $\beta \leq 1$, show that the M8 algorithm is non-trivial for the prediction of $8.0 \leq M < 8.5$ events: 0.5 per cent $< \alpha <$ 5 per cent. This interval covers the estimate $\alpha = 3.7$ per cent, which was obtained in the conventional way using the number of predicted events, ν_e^+.

(5) The significance analysis of M8 results based on the statistic ν_e^+ has some advantages.

(i) There are analytical upper bounds for α that incorporate the number of data N_λ used for estimating the rate of target events $\{\lambda_i\}$.

(ii) Estimates of α are not affected when $\{\lambda_i\}$ are replaced with $\{\rho\lambda_i\}$. This makes α stable under the choice of the estimation method for $\{\lambda_i\}$.

(iii) Estimates of α do not require any partitioning of the alarm space. The more detailed is the partitioning, the more difficult it is to interpret α. For small alarm domains α may merely reflect the uncertainty in $\{\lambda_i\}$.

(6) The following quantities are important in the statistical analysis of a prediction method: the number of target events N_e for the monitoring period, the number of events N_λ used to estimate $\{\lambda_i\}$ and the rate of growth of the GS weight function $w(p)$ near $p = 0$. Small values of N_e, N_λ and high rates of growth of $w(p)$ near $p = 0$ destabilize the estimates of significance of prediction results. For this reason the rules that regulate the testing of prediction algorithms should include restrictions on the above quantities. In our analysis of the M8 algorithm we have $N_e \approx 20$, the restrictions (22) on N_λ, and we consider the statistics (ν_e^+, τ) and $(R_w, w = 1)$ as the most preferable.

ACKNOWLEDGMENTS

We are grateful to Dr. Zechar and an anonymous reviewer for useful comments. This work was partly supported by the Russian Foundation for Basic Research through grant 11-05-00342.

REFERENCES

Ekstrom, G., Dziewonski, A.M., Maternovskaya, N.N. & Nettles, M., 2005. Global seismicity of 2003: centroid-moment tensor solutions for 1087 earthquakes, *Phys. Earth planet. Inter.,* **148,** 327–351.

Joliffe, L.T. & Stephenson, D.B. (eds), 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* John Willey & Sons, Hoboken.

Keilis-Borok, V.I. & Kossobokov, V.G., 1990. Premonitory activation of earthquake flow: algorithm M8, *Phys. Earth planet.Inter.,* **61,** 73–83.

Keilis-Borok, V.I., Shebalin, P.N., Gabrielov, A. & Turcotte, D.L., 2004. Reverse traicing of short-term earthquake precursors, *Phys. Earth planet.Inter.,* **145,** 75–85.

Kossobokov, V.G., 2005. Earthquake prediction: principles, implementation, perspectives, *Comput. Seism.,* **36,** 3–175 (in Russian).

Kossobokov, V.G., Romashkova, L.L., Keilis-Borok, V.I. & Healy, J.H., 1999. Testing earthquake prediction algorithms: statistically significant real-time prediction of the largest earthquakes in the Circum-Pacific, 1992–1997, *Phys. Earth planet.Inter.,* **111**(3–4), 187–196.

Kossobokov, V.G. & Shebalin, P.N., 2003. Earthquake prediction, in *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction,* pp. 141–207, eds Keilis-Borok, V.I. & Soloviev, A.A., Springer Publishers, Heidelberg.

Marzocchi, W., Sandri, L. & Boschi, E., 2003. On the validation of earthquake-forecasting models: the case of pattern recognition algorithms, *Bull. seism. Soc. Am.,* **93**(5), 1994–2004.

Molchan, G.M., 1997. Earthquake prediction as a decision-making problem, *Pure appl. Geophys.,* **149,** 233–247.

Molchan, G. M., 2003. Earthquake prediction strategies: a theoretical analysis, in *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction,* pp. 209–237, eds Keilis-Borok, V.I. & Soloviev, A.A., Springer Publishers, Heidelberg.

Molchan, G.M., 2010. Space-time earthquake prediction: the error diagrams, *Pure appl. Geophys.,* **167,** 8–9, doi:10.1007/s00024-010-0087-z.

Molchan, G.M. & Romashkova, L.L., 2010. Earthquake prediction analysis based on empirical seismic rate: the M8 algorithm, *Geophys. J. Int*., **183,** 1525–1537.

Pacheco, J.F. & Sykes, L.R., 1992. Seismic moment catalog of large shallow earthquakes, 1990 to 1989, *Bull. seism. Soc. Am.,* **82**, 1306–1349.

Shebalin, P., Keilis-Borok, V., Gabrielov, A., Zaliapin, I. & Turcotte, D., 2006. Short-term earthquake prediction by reverse analysis of lithosphere dynamics, *Tectonophysics,* **413,** 63–75.

Shebalin, P., Keilis-Borok, V., Zaliapin, I., Uyeda, S., Nagao, T. & Tsybin, N., 2004. Advance short-term prediction of the large Tokachi-Oki earthquake, September 25, 2003, M = 8.1. A case history, *Earth planets Space,* **56,** 715–724.

Vere-Jones, D., 1998. Probability and information gain for earthquake forecasting, *Comput. seism.,* **30,** 248–263.

Vorobieva, I.A., 1999. Prediction of subsequent large earthquake, *Phys. Earth planet. Inter.,* **111**(N3–4), 197–206.

Zechar, J.D. & Zhuang, J., 2010. Risk and return: evaluating RTP earthquake predictions, *Geophys. J. Int.,* **182,** 1319–1326.

Zhuang, J., 2010. Gambling scores for earthquake forecasts and predictions, *Geophys. J. Int.,* **181,** 382–390, doi:10.1111/j.1365-246X.2010.04496.x.

APPENDIX: LIST OF ESSENTIAL NOTATION

R_w, GS with weight function $w(p)$, formula (7);
w_β, model of $w(p)$, formula (10);
R_{LH}, score of information type (formula (13);
$A_i^\pm$, positive (+) and negative (–) space–time alarms;
$N_\pm$, number of positive (+) and negative (–) alarms;
$N = N_+ + N_-$, total number of alarms;
N_y, number of alarms that cover the target events;
ν_{++}, number of successful positive alarms;
N_e, number of target events;
$\hat{N}_e$, observed value of N_e;
ν_e^+, number of predicted events;
$\hat{\nu}_e^+$, observed value of ν_e^+;
λ_i, rate of target events in the *i*-th alarm zone;
N_λ, total number of $\{M \geq M_-\}$ events used for estimation of target event rates;
τ, normalized measure of space–time alarms;
$\hat{\tau}$ and $\tilde{\tau}$, point estimate and upper estimate of τ;
α, significance of prediction results;
$\hat{\alpha}$ and $\tilde{\alpha}$, point estimate and upper estimate of α;
$\hat{\alpha}_{PS}$ and $\hat{\tau}_{PS}$, $\hat{\alpha}$ and $\hat{\tau}$ for the case of the Pacheco & Sykes (1992) catalogue;
p_i^+, reference probability of target event for positive alarm;
p_i^-, reference probability of target event for negative alarm.