Annals of Glaciology 26 1998

Inferential statistics to verify prediction models

R. Bolognesi
Swiss Federal Institute for Snow and Avalanche Research, Antenne Valais, CH-1951 Sion, Switzerland

ABSTRACT. Models are powerful tools if their outputs are relevant!

Therefore, knowing the reliability of models is essential for people who wish to use them, as well as for researchers who attempt to improve them. Whatever the nature of the model output, objective evaluation consists of comparing predicted or calculated events with observed events.

Such comparison can only focus on available samples of observed events. Obviously, the results depend on the choice of the sample. However, inferential statistics enable one to extend results obtained from a random sample to general use.

An unbiased method of testing boolean avalanche-prediction models is suggested: the validity of this type of model should be characterized by the probability that the proportion of correct forecasts is within a given confidence interval. This interval is calculated from the sample size, according to the Gaussian table.

This unrestricted principle can be used to prove all kinds of static models, if ever their outputs are verifiable, enabling one to calculate the ratios of correct forecasts as well as the ratios of well-predicted events and can also be extended to verify probabilistic predictions.

INTRODUCTION

The development of avalanche-forecast systems keeps a great number of snow scientists busy. And for a very good reason: these systems promise to be invaluable as support tools. They use various models which emulate reality. Although they are becoming more and more sophisticated, these models are usually only very simplified reflections of reality, either because reality is only partly understood or describable or because the calculation tools (theoretical or technical) cannot deal with more complex representations.

Therefore, the following question must be answered: does the model give an appropriate reflection of observed reality?

It is obvious that this question cannot be avoided if the model is to become an operational tool.

In this event, the model must be checked with the utmost rigour.

But how can this verification be done?

In order to guarantee objective checking, the calculated values of the model must be compared to those actually measured. This test cannot be exhaustive and consequently cannot be a full inventory, as all possible cases of reality are not known (which is why a model is needed!). The validity of the model must therefore be established from verification of a restricted number of its assertions.

Thus, *verification of a model is a survey based on a sample poll*, that is to say a problem of inferential statistics, a technique which has as its aim establishing the likely characteristics of a particular group by observing only a sample of that group.

PRINCIPLES

Here, we shall deal with the verification of boolean forecasts which predict either the occurrence or the non-occurrence of an event (the developed principles can be extended to other types of forecast).

The verification of a model providing boolean forecasts consists in estimating the proportion p of correct forecasts which can be expected from the model, based on the proportion p′ of correct forecasts which is calculated using an available sample S′ of size n.

Thus, p is the proportion of exact forecasts which could be calculated if absolutely all the forecasts that the model can ever provide could be verified. Therefore, p is an objective quantifier of the reliability of the model.

Although p cannot be calculated, an approximation can be given.

The sample is a finite set of forecasts which can be verified. It may appear as a collection of pairs (predicted event/observed event). Provided that the composition of this sample is completely left to chance, it may be considered a random sample of the infinite number of predictions that can be made (Fig. 1).

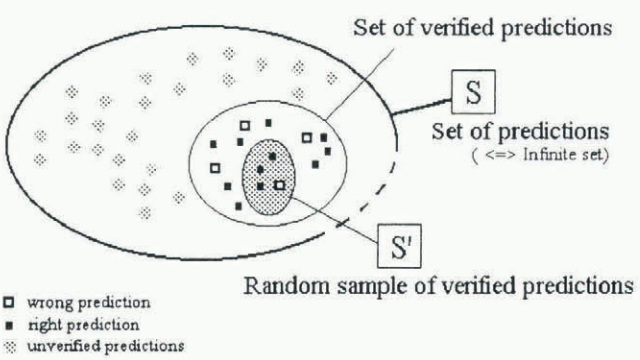


Fig. 1. The sample S′ used for the verification of a model is a sub-set of the infinite set S of past and future forecasts. To evaluate a model, it is possible to use either all verified forecasts at our disposal, assuming that the corresponding situations have randomly occurred, or only part of the verified forecasts, in which case this part must be a random sample.

https://doi.org/10.3189/1998AoG26-1-367-369 Published online by Cambridge University Press

As by definition, the probability of any forecast being exact is p, and as the verification of each forecast is a test which is independent of the $(n-1)$ others and applied according to one of two modes, the number x of exact forecasts can be considered a binomial random variable with mean np and variance $np\,(1-p)$.

Consequently, the proportion p' equivalent to x/n is also a binomial random variable with mean np/n and variance $np\,(1-p)/n^2$, that is to say with mean p and variance $p\,(1-p)/n$.

If the value of n is large and if p' is not close to 0 or 1, the binomial distribution can be approximated by the normal distribution. With the help of the table of the cumulative normal distribution function, intervals containing p' with a certain probability, can be determined. These intervals are usually called "confidence intervals". For instance: $(p-2(p\,(1-p)/n)^{\frac{1}{2}}) < p' < (p+2(p\,(1-p)/n)^{\frac{1}{2}})$, with a probability called a "coefficient of confidence" which is equal to 0.95.

Considering that the maximal value of $p(1-p)$ is 0.25, we get : $(p-K/n^{\frac{1}{2}}) < p' < (p+K/n^{\frac{1}{2}})$, that is to say $(p'-K/n^{\frac{1}{2}}) < p < (p'+K/n^{\frac{1}{2}})$, where K is a real number depending on the chosen coefficient of confidence and where $K/n^{\frac{1}{2}}$ is the accuracy of estimation of p.

Therefore, it is possible to estimate p depending on n and p', without overestimating the value of p. For instance:

$$p = p' \pm 1/n^{\frac{1}{2}} \text{ with a minimum probability of 0.95,} \quad (1)$$

$$p = p' \pm 1.3/n^{\frac{1}{2}} \text{ with a minimum probability of 0.99.}$$

It is not surprising that the accuracy of the verification increases with the size of the sample (Fig. 2). Thus, for a proportion $p' = 50\%$ recorded from a sample of 100, it is possible to assert that the odds are at least 95 out of 100 that the reliability of the tested model lies between 40% and 60%. The same proportion p', recorded from a sample of 1000 would allow one to assert, with the same probability, that the reliability of the model is between 47% and 53% (rounded values). In fact, the accuracy varies in proportion to the square root of the sample size. Thus, in order to double the accuracy, the size of the sample must be multiplied by 4: accuracy turns out to be "expensive". Certainty is also expensive: to reach a probability of 99% instead of 95% for the preceding estimation, the sample size has to be increased to 1877 instead of 1000.

In order to carry out a verification which might be usable in making forecasts, it is necessary to know the proportion of exact forecasts which can be expected from the model for the two forms of the forecast (occurrence and

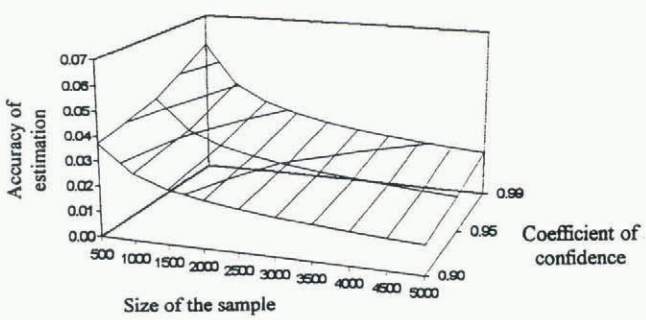
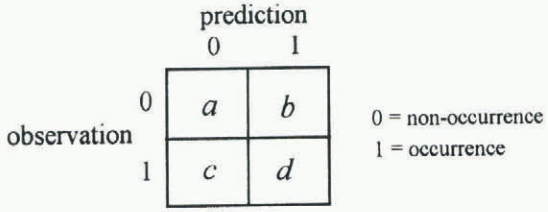


Fig. 2. Minimal accuracy of a model reliability estimation, depending on the size of the verification sample with various coefficients of confidence.



$$p_1 = (d/(b+d)) \pm K/(b+d)^{\frac{1}{2}}$$

$$p_2 = (a/(a+c)) \pm K/(a+c)^{\frac{1}{2}}$$

Fig. 3. Verification of a model for practical use. p_1 is the proportion of correct occurrence forecasts and p_2 is the proportion of correct non-occurrence forecasts which can be expected from the model.

non-occurrence of the event). So it is necessary to evaluate the proportions of exact forecasts p_1 and p_2 from the proportions of exact forecasts p_1' and p_2', calculated respectively from a sample of occurrence and non-occurrence predictions (Fig. 3). These proportions p_1 and p_2 will show the model user how much he can trust a given forecast.

APPLICATION

The following example is given to show the use of inferential statistics in verifying predictions. A very simple model has been outlined which is expected to provide forecasts from cursory data. This model predicts whether at least one avalanche will occur in a given area during a given day. The model is based on the fact that most avalanches occur either during snowfalls, when snow is melting or after heavy snow transport by wind. This suggests that the occurrence of avalanches is a function of the thickness of the top layer of non-cohesive snow, on one hand, and of the intensity of the snow drift, on the other hand, which can be translated into the following formula:

$$A = (Ps > \alpha)\mathbf{V}(FFt > \beta)$$

with A, occurrence of at least one avalanche during the next hours; Ps, sinking of the first ram-penetrometer tube; FFt, amount of snow caught by the driftometer (Bolognesi, 1997) in 24 hours; α, β reals.

This simple model will now be evaluated according to the principles laid down previously. We have at our disposal a number of cases for which the forecasts can be produced (the necessary data being available) and verified (the occurrence like the non-occurrence of avalanches being established by the success or failure of daily trials to trigger avalanches in all of a given group of couloirs). Each of the forecasts established by the model will be verified in the following way: an occurrence forecast will be declared exact if at least one avalanche takes place either naturally or during trials, and a non-occurrence forecast will be classified as exact if no avalanche occurs in spite of release attempts.

In total, 278 such verified forecasts are available (this represents more than 5000 attempts to trigger avalanches!).

The proportion of exact forecasts which was calculated from a random sample of 100 forecasts, with $\alpha = 0.05$ and $\beta = 0$, is 0.8.

The hypothesis according to which the proportion p' of correct forecasts is a random variable approximately Gaussian is perfectly plausible, as shown by the distribution function of this variable (Fig. 4).

https://doi.org/10.3189/1998AoG26-1-367-369 Published online by Cambridge University Press

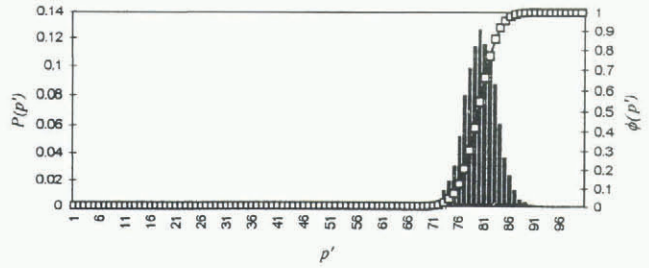


Fig. 4. Distribution function (diagram) and cumulative distribution function ϕ (curve) of the random variable $p'(P(p_i'))$ assumed equivalent to calculated frequencies of p_i').

So we can state, with a probability of at least 0.95, that the proportion of exact forecasts which we can hope for in applying the model proposed above, lies between 0.7 and 0.9 (see Equation (1)).

The proportions of exact forecasts calculated from a random sample of 100 occurrence forecasts and from a random sample of 100 non-occurrence forecasts are both nearly equal to 0.8. This information is vital when using the model for operational forecasts: for instance, the user knows that occurrence predictions have to be taken into account because they are often right.

DISCUSSION

The method of verification proposed here can be used to prove all types of models (which do not include any automatic learning procedure), providing that their forecasts or diagnoses are verifiable, that is to say, comparable to observed events or measured values. However some authors use other verification procedures if this requirement is not fulfilled (Bois and Obled, 1976; Giraud and others, 1994; Föhn and Schweizer, 1995).

This method is designed to help to make use of a model by indicating the probability that a predicted event will occur; therefore this indication is of great interest to forecasters who have to know how much they can trust a given prediction. However, knowing the probabilities that events will be predicted is a prerequisite to evaluating whether a model is usable or not. These probabilities can be calculated according to the same principles. Let us imagine a first (realistic) practical case: 1450 predictions from an avalanche-forecasting model are compared to observed events; this comparison shows that 350 avalanches occurred, whereas 400 avalanches were predicted and that 50 unexpected avalanches occurred (Fig. 5a).

We can infer from this comparison that:

The probability of any prediction being exact is $0.90 \pm \epsilon_1$.

The probability of any occurrence prediction being exact is $0.75 \pm \epsilon_2$.

The probability of any non-occurrence prediction being exact is $0.95 \pm \epsilon_3$.

The probability of any avalanche being predicted is $0.86 \pm \epsilon_4$.

The probability of any non-occurrence being predicted is $0.91 \pm \epsilon_5$.

According to these results, we can consider that this model is of interest to forecasters.

Now, let us imagine a second practical case: we have to evaluate a model which invariably predicts that no avalanche will occur (Fig. 5b). Because avalanches are not frequent events (24% in our example), this model seems to be reliable: the probability of any prediction being correct is $0.76 \pm \epsilon$. But the probability of any occurrence prediction being exact cannot be calculated and the probability that at least one avalanche will be predicted is 0! Consequently we can consider that such a model is fanciful.

		pred. 0	pred. 1	Σ
obs.	0	1000	100	1100
	1	50	300	350
	Σ	1050	400	1450

a

		pred. 0	pred. 1	Σ
	0	1100	0	1100
	1	350	0	350
	Σ	1450	0	1450

b

Fig. 5. (a) Contingency table showing an imperfect but usable model. (b) Contingency table showing a fanciful and unusable model which yet gives a high ratio of correct predictions.

Finally, the method can be extended to verify probabilistic forecasts. In order to do this, we have to compare probabilities to frequencies of events. We would be able to claim that the model is perfect if, for m occurrences predicted with probability P, mP occurrences are observed when m tends towards infinity and this whatever may be the value of P.

So to make a reliability measure of a probabilistic prediction model, it will be sufficient to estimate the probable frequencies of the occurrence of the event (according to the principles presented in section 1) for the various classes of probabilities calculated by the model.

CONCLUSION

The verification of a prediction model is neither a simple task nor an insignificant operation capable of being reduced to a few improvised tests.

Verification requires consistent samples, the gathering of which may demand more time (and money) than the creation of the model itself. Therefore, the design of the verification procedures should be an imperative preliminary stage for all scientific modelling projects, for it is true that it is futile to create a model incapable of being verified.

ACKNOWLEDGEMENTS

Thanks are due to B. Corboz, O. Buser and F. Taillard for their helpful co-operation.

REFERENCES

Bois, P. and C. Obled. 1976. Prévision des avalanches par des méthodes statistiques. Aspects méthodologiques et opérationnels. (Application à la région de Davos, Suisse.) *Houille Blanche*, **31** (6–7), 509–531.

Bolognesi, R. 1997. The driftometer. *ISSW'96. International Snow Science Workshop, 6–10 October 1996, Banff, Alberta. Proceedings.* Revelstoke, B.C., Canadian Avalanche Association, 144–148.

Föhn, P. M. B. and J. Schweizer. 1995. Verification of avalanche danger with respect to avalanche forecasting. *In* Sivardière, F., *ed. Les apports de la recherche scientifique à la sécurité neige, glace et avalanche. Actes de Colloque, Chamonix 30 mai–3 juin 1995.* Grenoble, Association Nationale pour l'Étude de la Neige et des Avalanches (ANENA), 151–156.

Giraud, G., E. Brun, Y. Durand and E. Martin. 1994. Validation of objective models to simulate snow cover stratigraphy and avalanche risks. *ISSW'94. International Snow Science Workshop, 30 October–3 November 1994, Snowbird, Utah. Proceedings.* Snowbird, UT, P.O. Box 49, 509–517.

https://doi.org/10.3189/1998AoG26-1-367-369 Published online by Cambridge University Press