Journa lof Glaciology, Vol. 22, No. 86, 1979

USING LINEAR DISCRIMINANT ANALYSIS TO CLASSIFY SNOWFALL SITUATIONS INTO AVALANCHING AND NON-AVALANCHING ONES*

By N. F. Drozdovskaya

(Sredneaziatskiy Regional'nyy Naucho-Issledovatel'skiy Gidrometeorologicheskiy Institut, Tashkent, U.S.S.R.)

Abstract. The existing methods of predicting avalanche danger often do not meet users' demands because of the empiric character of the insufficient volume of information used. In such forecasts the contribution of each individual parameter into the prognostic information is unknown, and this is very important when studying such an event as avalanche formation, which is conditioned by a complex interaction of numerous factors, including snow accumulation, snow thickness, and the conditions of its development. It is obvious that such problems can be successfully solved by statistical methods, and that explains the growing interest in numerical methods of avalanche forecasting. Problems of multi-dimensional observations arises in many scientific fields. The method suited for this problem is discriminant analysis, the purpose of which is to divide a multi-dimensional observation vector into predetermined classes.

This study considers the prognostic (diagnostic) problems of fresh-snow avalanches released during snowfall or in the two days after it has ceased. The theoretical basis is a complex of statistical methods: correlation and dispersion analysis, "sifting" for the choice of predictors' informative groups, construction of linear parametric discriminant functions, predictions based on training sample, and verification of discriminant functions based on independent material.

The archive used in the study consisted of 500 avalanching cases and 1 300 non-avalanching ones. All situations were grouped according to geomorphological characteristics. Each situation is described by eight meteorological characteristics. The results of classification of snowfall situations into avalanching and non-avalanching ones are as follows: reliability of p is from 75% to 91%, H from 0.15 to 0.51; based on independent material the reliability of p is from 63% to 85%, H from 0.10 to 0.56.

Résumé. *Utilisation de l'analyse discriminante linéaire pour classer les épisodes de chutes de neige en situations dangereuses ou non dangereuses du point de vue avalanches.* Les méthodes existantes pour la prévision du danger d'avalanche ne satisfont pas toujours à la demande des usagers en raison du caractère empirique du volume insuffisant d'informations utilisées. Dans les prévisions de l'espèce, la contribution de chaque paramètre particulier au pronostic est inconnue alors qu'elle est très importante dans l'étude d'évènements comme la formation d'avalanches régie par des interactions complexes entre de nombreux facteurs, chutes de neige, épaisseur du manteau, conditions de son développement. Il est clair que de tels problèmes peuvent être résolus avec succès par des méthodes statistiques, ce qui explique un intérêt croissant pour les méthodes numériques de prévision d'avalanches. Des problèmes d'observations multidimensionnelles se posent dans beaucoup de disciplines scientifiques. La méthode convenable pour de tels problèmes est l'analyse discriminante dont l'objectif est de diviser le vecteur observations multidimensionnelles en classes préétablies.

Cette étude considère les problèmes de pronostic (diagnostic) des avalanches de neige fraîche se produisant pendant une chute de neige ou dans les deux jours qui suivent. La base théorique est une réunion de méthodes statistiques: analyses de corrélation et de dispersion, "test" pour le choix des groupes d'informations prédicteurs, construction de fonctions discriminantes de paramètres linéaires, prévisions basées sur des essais sur échantillons, et vérification des fonctions discriminantes sur la base de matériels indépendants.

La population utilisée dans cette étude comprenait 500 cas d'avalanches et 1 300 cas de non-avalanches. Toutes les situations ont été groupées selon leurs caractéristiques géomorphologiques. Chaque situation est décrite par huit caractéristiques météorologiques. Les résultats de la classification des situations de chute de neige en avalancheuses et non-avalancheuses sont les suivants: la fiabilité de p va de 75% à 91%, celle de H de 0,15 à 0,51; à partir de données indépendantes, la fiabilité de p est de 63% à 85%, celle de H de 0,10 à 0,56.

Zusammenfassung. *Klassifizierung von Hängen hinsichtlich ihrer Lawinengefahr während eines Schneefalls mit Hilfe der linearen Unterscheidungsanalyse.* Infolge ihres empirischen Charakters oder wegen zu geringen Informationsvolumens werden die verfügbaren Methoden zur Vorhersage von Lawinengefahr oft den Anforderungen der Interessenten nicht gerecht. Bei Vorhersagen dieser Art ist der Anteil jedes einzelnen Parameters an der prognostischen Information unbekannt; dies ist von grosser Bedeutung für das Studium von Ereignissen wie der Lawinenbildung, die vom komplexen Zusammenwirken vieler Faktoren wie Akkumulation, Zusammensetzung und Metamorphose des Schnees abhängt. Offensichtlich können solche Probleme erfolgreich mit statistischen Methoden gelöst werden, woraus sich das wachsende Interesse der letzten Zeit an Rechenmethoden in der Lawinenvorhersage erklärt. Im vorliegenden Fall handelt es sich um die Unterscheidungsanalyse zur Aufspaltung eines vieldimensionalen Beobachtungsvektors in vorbestimmte Klassen. Diese Untersuchung gilt dem Problem der Vorhersage von Neuschneelawinen, die während oder bis zu zwei

* This paper was accepted for the Symposium on Applied Glaciology, Cambridge, September 1976, but not presented because of the absence of the author.

https://doi.org/10.3189/S0022143000014106 Published online by Cambridge University Press

Tagen nach Ende eines Schneefalles abgehen. Die theoretische Grundlage ist ein Komplex von statistischen Methoden: Korrelations- und Dispersionsanalyse; Filterung zur Auswahl der zur Vorhersage geeigneten Informationsgruppen; Aufbau linearer Parameter der Unterscheidungsfunktionen; Vorhersage auf der Basis von Lehrstichproben und Erprobung der Unterscheidungsfunktionen mit unabhängigem Material.

Das benutzte Archiv enthält 500 Fälle mit und 1 300 Fälle ohne Lawinen. Alle Hänge sind nach geomorphologischen Kennzeichen eingruppiert und nach 8 meteorologischen Verhältnissen beschrieben. Die Ergebnisse der Hangklassifizierung hinsichtlich ihrer Lawinengefahr sind die folgenden.

Die Zuverlässigkeit p liegt zwischen 75 und 91%; H ist 0,15–0,51; mit unabhängigem Material liegt p zwischen 63 und 85%; H ist 0,10–0,56.

The aim of discriminant analysis is to divide X vector-predictors into classes by an optimal method according to the prescribed predictant classes; the discriminant function makes it possible to determine the class of any $\mathbf{X}$ that is not included into the archive.

The simplest solution of the problem concerning the division of the object into classes is obtained with normally-distributed continuous predictors, identical covariant matrices within both classes of the object, and with inequality of mathematical expectations within the classes. In this case the problem is solved by means of a linear parameter discriminant function $U(X)$ that may be considered as some one-dimensional variable with normal distribution. The equality of the covariant matrices of different classes is a rather strong limitation, but it essentially simplifies the computational technique.

Non-significance of the differences between them is defined empirically, and then one matrix is constructed by averaging (taking the weights into consideration) the same elements. Another limitation is connected with the deviation from normal distribution that is conventional for some parameters.

Discriminant analysis methods are widely used in hydrometeorology (Gruza, 1967; Suzuki, 1969; Bois and Obled, 1973). Its application is simplest for forecasting of phenomena which assume alternative formulation. In this case the linear discriminant function at the limit of classes is:

$$U(X) = X'A - \frac{1}{2}(M_1+M_2)'A - \ln\frac{\Pi_2 C_2}{\Pi_1 C_1},$$

where M_1 and M_2 are vectors of mathematical expectations within classes Φ_1 and Φ_2 respectively; Π_1 and Π_2 are *a priori* probabilities of classes Φ_1 and Φ_2; C_1 and C_2 are values of erroneous classification of the first and the second kind, their relation is assigned according to *a priori* probabilities of classes and valuability of correct prediction of each class; C_1 is the value of the erroneous classification of the first kind which means that the $\mathbf{X}$ vector belonging to the first class is referred to the second one; C_2 is the value of the erroneous classification of the second kind which means that the $\mathbf{X}$ vector belonging to the second class is referred to the first one. These errors arise due to the fact that the probability density of different classes is in some way re-covered by each other. By means of the C_2/C_1 correlation it is possible to improve results of correct identification of one class by deterioration of results of identification of the other class. The prime is the transposition sign and A is a vector of linear coefficients

$$A = S^{-1}(M_1 - M_2),$$

where S denotes the mean covariant matrix.

Whether the concrete realization of the $\mathbf{X}$ vector belongs to classes Φ_1 and Φ_2 is determined by the following decisive rule: if $U(\mathbf{X}) \geqslant R$, then class Φ_1 is forecast (i.e. avalanching is expected); if $U(\mathbf{X}) < R$, then class Φ_2 is forecast (i.e. avalanching is not expected).

R denotes the threshold value of the discriminant function

$$R = \frac{1}{2}(M_1+M_2)'A - \ln\frac{\Pi_2 C_2}{\Pi_1 C_1}.$$

The purpose of the study was to classify the situations into avalanching and non-avalanching ones during snowfall (zero earliness forecast or diagnosis of the fresh-snow avalanches which

https://doi.org/10.3189/S0022143000014106 Published online by Cambridge University Press

slipped during snowfall or within the next two days when there was no rise in temperature). The concept "zero earliness forecast" can be explained as follows. Let us analyse the object which is described by observing parameters $y = (y_1, y_2, \ldots, y_m)$; these observations can be referred to the different moments of time t_{y_j}, $j = 1, \ldots, m$. Suppose that we have clarified the physical nature of this object and can describe it by another one $X = (x_1, x_2, \ldots, x_n)$, i.e. the characteristics of the X objects are also referred to different moments of time t_{x_i}, $i = 1, \ldots, n$. If $\max_i t_{x_i} < \min_j t_{y_j}$ then the analysis can be regarded by its character as prognostic.

Otherwise extrapolation in time and, consequently, the prediction is not possible. Let us assume that when $\max_i t_{x_i} < \min_j t_{y_j}$ the time interval $\tau = \min_j t_{y_j} - \max_i t_{x_i}$ indicates the forecast earliness, the moment of time $\max_i t_{x_i}$ represents the moment of forecasting and the time interval $\tau_y = \max_j t_{y_j} - \min_j t_{y_j}$ indicates the forecasting period or period of forecasting efficiency.

The case when the requirement $\max_i t_{x_i} < \min_j t_{y_i}$ is not realized, i.e. $\tau < 0$, is also interesting, and generally it is in accordance with the problem of statistical analysis or diagnosis. Sometimes it is called "zero earliness forecast".

Investigations were carried out on the basis of observed data obtained from snow-avalanche stations in Central Asia. For the analysis samples consisting of 500 cases of avalanching (class Φ_1) and 1 300 cases of non-avalanching (class Φ_2) were used. All data were differentiated into two altitude zones (higher and lower than 3 000 m) and into three sectors in each altitude zone. The first sector united cases of situations observed on slopes of southern aspect (S., S.E., and S.W. expositions), the second one covered cases with western and eastern aspects, and the third sector united cases observed with northern aspects (N., N.E., and N.W. expositions). The values of the vector-predictor in each zone differs in the variable quantity X_8 (the sum of the total daily radiation during the period of snowfall, which depends on the slope aspect). The whole totality of avalanches was considered as one case, if some avalanches were observed in similar height conditions and orographic aspect.

A vector whose components are analogue data on meteorological conditions in which the formation of a fresh snow layer occurs and which have decisive influence on avalanche formation in fresh snow, is taken as a predicting vector (predictor) $X = (x_1, x_2, \ldots, x_k)$, $k = 1, \ldots, 8$. The variables x_1, x_2, x_3 are introduced into the prediction scheme to characterize indirectly the state of the old-snow surface at the beginning of snowfall. Table I gives the content of the predictor.

TABLE I. THE MEANING OF THE PREDICTORS USED

Predictor	*Meaning in classes Φ_1 and Φ_2*
x_1	The average air temperature three hours before the beginning of snowfall, in degrees
x_2	Air temperature at the moment of the beginning of snowfall, in degrees
x_3	Absolute air humidity at the moment of snowfall beginning, in mbar
x_4	The duration of the period from the beginning to the end of the snowfall for class Φ_2 and from the snowfall beginning up to the avalanching for class Φ_1, in hours
x_5	The sum of solid precipitation for the snowfall period for class Φ_2 and for the period from the beginning of snowfall up to avalanching for class Φ_1, in mm
x_6	The sum of average daily air temperature for the snowfall period for class Φ_2 and during the period from snowfall beginning up to avalanching for class Φ_1, in degrees
x_7	The sum of average daily absolute air humidity during the period from the beginning to the end of snowfall for class Φ_2 and for the period from snowfall beginning up to avalanching for class Φ_1, in mbar
x_8	The sum of daily total radiation during snowfall period for class Φ_2 and for the period of snowfall beginning up to avalanching for class Φ_1, in cal/cm² (or J/cm²) daily

https://doi.org/10.3189/S0022143000014106 Published online by Cambridge University Press

Discriminant analysis provides for two stages according to the scheme used: training and control testing. Training implies the first stage of the experiment when the training is carried out on the basis of predictor sampling with a known division into classes, i.e. the correlational analysis is realized, the covariance matrix is constructed, the predictors are sifting, the coefficient of the discriminant function and the indicator of divisibility (NR) are assessed. The stage of control testing includes the assessment of the reliability sample forecasts on dependent and independent material. The material is referred to as "dependent" if the training is carried out including it; "independent" material is material not included in the training stage, but used for testing of the discriminant functions constructed on the basis of dependent material.

Different statistical characteristics are considered within each stage. "Double 'Student's' *t*-criterion" determines divergence variability of $\bar{X}_{\Phi_1}$ and $\bar{X}_{\Phi_2}$, the arithmetic mean values of samples obtained from the results of independent measurements. This is one means of predictor informativity and serves as an assessment of the possibility of their application for classification of situations into classes Φ_1 and Φ_2 classes during snowfalls.

The calculations showed that considering error probability from 0.10 to 0.05 for predictors x_4–x_8, the difference in the means of samples of classes Φ_1 and Φ_2 is mainly significant; insignificant differences prevail for predictors x_1–x_3, this indicates their low informativity according to this criterion.

Informativity of different predictor groups was defined by a "sifting method" that includes the definition of a division index (NR):

$$(NR) = \frac{\bar{U}_1(X) - \bar{U}_2(X)}{\sqrt{D}},$$

where $\bar{U}_1(X)$ and $\bar{U}_2(X)$ are mean values of discriminant function within classes Φ_1 and Φ_2; $\bar{U}_1(X) - \bar{U}_2(X) = \Delta^2$ (Mahalanobis distance), and D is the dispersion of the linear discriminant function within both classes. The "sifting" method consists of the following procedures. First all the components of the X vector are inspected and a component is chosen that provides the best classification into classes Φ_1 and Φ_2 ((NR) value being taken as a criterion). Then $U(X)$ is calculated for the chosen component together with each other one. The pair which gives maximum (NR) value is taken as the optimal one. The optimal pair is used to choose an optimal triple, etc.

In Table II the results of "sifting" are presented for six samples, where the predictors are set in the order of their choice at "sifting". P_{T} in this table denotes the theoretical probability of correct classification (Gnedenko, 1965).

The data presented in Table II which show the degree of informativity increasing (through (NR)) as each predictor is added, allow us to conclude that the optimal informativity group (in the given complex) is obtained by means of variables characterizing the total value of precipitation, solar radiation, absolute air humidity, precipitation duration, and, in three of the six cases chosen, the sum of daily mean air temperatures. The order of these variables determined by their informativity rank was different in the two altitudinal zones.

In the altitude zone higher than 3 000 m, the classification of snowfall situations depends greatly upon precipitation value, while in the zone lower than 3 000 m it depends upon solar radiation. In the first zone during the third step of "sifting" the x_7 variable, the sum of daily mean absolute air humidity, was chosen; in the second zone this variable entered the optimal complex only in the fourth step.

At the boundary of this group the growth of theoretical forecast reliability essentially stops. This is rather important when defining the optimal group of predictors which is included into forecast, because introduction into the prediction scheme of features insignificantly affecting the classification quality, leads to the deterioration of the reliability of the forecast on independent material, although the forecast based on dependent material can be improved to some extent.

https://doi.org/10.3189/S0022143000014106 Published online by Cambridge University Press

TABLE II. COMPARATIVE INFORMATIVITY OF DIFFERENT PREDICTORS

Sample		*Predictor*	*(NR)*	P_T	*Sample*		*Predictor*	*(NR)*	P_T
Altitude zone I	Sector I	x_5	0.491		Altitude zone II	Sector I	x_8	1.00	
		x_8	0.877	0.666			x_5	1.58	0.785
		x_7	0.983	0.688			x_4	1.63	0.791
		x_6	1.129	0.716			x_1	1.70	0.802
		x_4	1.221	0.729			x_7	1.78	0.813
		x_3	1.226 7	0.729			x_6	1.80	0.816
		x_2	1.230 2	0.732			x_2	1.82	0.818
		x_1	1.256 7	0.735			x_3	1.83	0.818
	Sector II	x_5	0.558			Sector II	x_8	0.719	
		x_8	0.750	0.644			x_5	1.07	0.705
		x_7	1.049	0.698			x_4	1.209	0.726
		x_4	1.078	0.705			x_7	1.532	0.776
		x_3	1.091	0.709			x_6	1,846	0.821
		x_2	1.124 4	0.716			x_2	1.90	0.836
		x_1	1.131 6	0.716			x_3	1.98	0.838
		x_6	1.132 7	0.716			x_1	1.99	0.838
	Sector III	x_5	0.744			Sector III	x_8	0.413	
		x_8	1.004	0.691			x_5	0.579	0.614
		x_7	1.065	0.705			x_4	0.729	0.640
		x_4	1.144	0.715			x_7	0.951	0.681
		x_6	1.233 5	0.732			x_2	1.043	0.698
		x_3	1.235 2	0.732			x_3	1.044 2	0.698
		x_1	1.237 4	0.732			x_6	1.044 3	0.698
		x_2	1.238 1	0.732			x_1	1.044 5	0.698

The predictors which were not included into the optimal complex include the variables x_1, x_2, and x_3 indirectly characterizing the state of the old-snow surface before the snowfall began.

The next stage of the experiment consisted of making a forecast on the basis of the training sample, and for this purpose the discriminant function value was found, and the classification of situations into avalanching and non-avalanching ones with different values of C_1 and C_2 was made using the rule mentioned above. The best results for satisfactory classification of situations into the two classes were obtained with the values of the erroneous classification inversely proportional to the class of the appearance probability (except sector II in the altitude zone II, where the best result was obtained with a lower ratio C_1/C_2).

Table III shows the forecast estimation based on the training sample at various C_1/C_2 ratios and with either eight- and five-dimensional vector predictors for the sectors of the first altitude zone in order to estimate the effect of the dimension N_X of the vector-predictor on the forecast result (when $N_X = 5$, the predictors which were not introduced in Table II, were excluded).

The following symbols are used in Table III:

n the number of cases of coincident events,
(Π_1, Φ_1) avalanches predicted and observed,
(Π_1, Φ_2) avalanches predicted but not observed,
(Π_2, Φ_1) avalanches observed but not predicted,
(Π_2, Φ_2) avalanches not predicted and not observed,
p percentage of reliable forecasts,
H criterion of reliability used for estimation of prediction of rare events; this criterion varies from 0 for a random forecast to 1 for an ideal one (Bagrov, 1965).

This last criterion is more effective for assessing avalanche prediction, as is obvious, for instance, from the forecast result for sector II of altitude zone II, where at $p = 90.9\%$, H

https://doi.org/10.3189/S0022143000014106 Published online by Cambridge University Press

equals only 0.15%. The high percentage of forecast reliability with a low value of the H criterion in this case was due to successful prediction of class Φ_2 with a comparatively small number of cases in class Φ_1.

As a rule forecasts made by discriminant functions constructed on independent material proved to have less reliability, but quite comparable with the reliability of an avalanche forecast by any other empirical method, and testifies to the usefulness of the proposed method. The forecast results on independent samples are given in Table IV.

TABLE III. ESTIMATION OF SUCCESS OF CLASSIFICATION OF SITUATIONS DURING SNOWFALLS INTO AVALANCHING AND NON-AVALANCHING ONES

			Ratio of successful and unsuccessful forecasts in classes Φ_1 and Φ_2		*Estimation of the forecast reliability*	
Sector	*Dimensions* N_x	*Value* C_1/C_2	$\frac{n(\Pi_1, \Phi_1)}{n(\Pi_1, \Phi_2)}$	$\frac{n(\Pi_2, \Phi_1)}{n(\Pi_2, \Phi_2)}$	p %	H
			Altitude zone I			
Sector I	8	3/1	$\frac{86}{36}$	$\frac{30}{304}$	80.2	0.50
	5	3/1	$\frac{90}{64}$	$\frac{26}{306}$	81.5	0.51
Sector II	8	3/1	$\frac{35}{45}$	$\frac{41}{325}$	80.7	0.34
	5	3/1	$\frac{32}{40}$	$\frac{44}{330}$	81.2	0.32
Sector III	8	5/1	$\frac{65}{69}$	$\frac{34}{432}$	82.8	0.47
	5	5/1	$\frac{66}{67}$	$\frac{33}{434}$	83.3	0.47
			Altitude zone II			
Sector I	8	10/1	$\frac{21}{37}$	$\frac{11}{316}$	87.6	0.43
Sector II	8	8/1	$\frac{7}{5}$	$\frac{57}{494}$	90.9	0.15
Sector III	8	4/1	$\frac{58}{58}$	$\frac{23}{189}$	75.3	0.42

TABLE IV. ESTIMATION OF SUCCESS OF CLASSIFICATION OF SITUATIONS DURING SNOWFALLS INTO AVALANCHING AND NON-AVALANCHING ONES BASED ON INDEPENDENT MATERIAL

	Altitude zone I				*Altitude zone II*			
			Estimation of forecast reliability				*Estimation of forecast reliability*	
Sector	*Dimension* N_x	*Value* C_1/C_2	p %	H	*Dimension* N_x	*Value* C_1/C_2	p %	H
Sector I	8	3/1	78.8	0.50	8	5/1	84.7	0.10
	5	3/1	80.0	0.56	5	5/1	76.4	0.22
Sector II	8	5/1	70.0	0.20	8	5/1	79.6	0.54
	5	5/1	72.0	0.33	5	5/1	79.8	0.56
Sector III	8	4/1	79.1	0.51	8	3/1	63.2	0.30
	5	4/1	80.6	0.53	5	3/1	62.1	0.28

The work is far from being completed; improvement of the given scheme implies drawing in a greater volume of material, using additional predictors, and transition to a forecast with an "earliness" which differs from zero.

MS. received 16 August 1976 and in revised form 4 July 1977

REFERENCES

Bagrov, N. A. 1965. Statisticheskiy analiz rezul'tatov ispytaniy nekotorykh sposobov prognoza [Statistical analysis of the investigation results obtained with some forecast techniques]. *Meteorologiya i Gidrologiya*, 1965, No. 8, p. 40–46.

Bois, P., *and* Obled, C. 1973. Vers un système operationel de prévision numérique des avalanches à partir de méthodes statistiques. *Hydrological Sciences Bulletin*, Vol. 18, No. 412, p. 419–29.

Gnedenko, B. V. 1965. *Kurs teorii veroyatnostey* [*Probability theory*]. Moscow, [Izdatel'stvo] "Nauka".

Gruza, G. V. 1967. O nekotorykh prakticheskikh priyemakh diskriminantnogo analiza [On some discriminant analysis practical techniques]. *Trudy SANIGMI*, Vyp. 29(44), p. 160–66.

Suzuki, E. 1969. Discrimination theory based on categorical variables and its application to meteorological problems. *Journal of the Meteorological Society of Japan*, Vol. 47, No. 3, p. 145–58.

Shmakova, V. S. 1969. Postroyeniye diskriminantnykh funktsiy dlya trekhdnevnogo prognoza minimal'noy temperatury [Discriminant functions structure for three-days range forecast of minimum temperature]. *Trudy SANIGMI*, Vyp. 40(55), p. 107–11.

Storm, R. 1965. *Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle.* Leipzig, Fachbuchverlag.

https://doi.org/10.3189/S0022143000014106 Published online by Cambridge University Press