Available online at www.sciencedirect.com
ScienceDirect
Cold Regions Science and Technology 47 (2007) 50–59
cold regions science and technology
www.elsevier.com/locate/coldregions

A threshold sum approach to stability evaluation of manual snow profiles

Jürg Schweizer [a,*], J. Bruce Jamieson [b]

[a] WSL Swiss Federal Institute for Snow and Avalanche Research SLF, Flüelastrasse 11, CH-7260 Davos Dorf, Switzerland
[b] Department of Civil Engineering, Department of Geology and Geophysics, University of Calgary, Calgary, Alberta, Canada

Abstract

Snow profile interpretation has developed in the last few years from being based on experience into a semi-quantitative scientific method. Emphasizing structural rather than mechanical instability, threshold values were developed for key parameters such as weak layer grain size and hardness, and differences in grain size and hardness between layers. Despite promising attempts so far it has not been shown that this method works to quantitatively interpret snow profiles, in particular if the principal weakness is unknown. Our aim was to provide an easy and robust method based on a threshold sum approach to assess snowpack stability based on layer properties. Second, we investigated whether that method was also suited to find the principal weakness (in case it is unknown) and assess the probability for a skier-triggered avalanche on this weakness. Our data set consisted of 500 manual snow profiles observed over 16 years on skier-tested and skier-triggered avalanche slopes from both western Canada and Switzerland. A weighted threshold sum with the failure layer depth as independent variable scored highest (77% for the learning data set, 65% for the test data set). Detection of potentially critical layers proved to be less successful, in particular for the Swiss profiles. If the principal weakness was unknown, the stability classification for the potentially critical layers agreed with the observed stability for the Swiss profiles in about 53% and for the Canadian profiles in about 62% of the cases. The results emphasize that stability assessment should include – besides stability tests that help locate the principal weakness – analysis of snow layer properties, in particular grain size, type and hardness. The proposed threshold sum considering seven variables is well suited for profile analysis of manual profiles by practitioners. Stability classification of snow profiles simulated by snow cover models such as SNOWPACK will need further adaptation, in particular for application in transitional snow climates.

Keywords: Snow stability; Stability evaluation; Avalanche forecasting; Skier triggering; Snowpack stratigraphy; Snow profile

1. Introduction

Snowpack stability evaluation for avalanche forecasting relies on weather data, snowpack data and avalanche observations. Snowpack data in the form of snow profiles and stability tests are the crucial information in the absence of avalanche occurrence data to derive snow stability (Schweizer et al., 2003). Stability tests are powerful, but occasionally give misleading results, i.e. false-stable predictions. Also, stability test results seem to be more susceptible to spatial variations of snowpack properties than e.g. layer characteristics such as grain type and size (Kronholm, 2004).

Consequently, McCammon and Schweizer (2002) proposed to augment information on mechanical instability such as the shear strength or stability test scores with data on structural instability such as grain type and

* Corresponding author. Tel.: +41 81 4170164; fax: +41 81 4170110.
E-mail address: schweizer@slf.ch (J. Schweizer).

0165-232X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.coldregions.2006.08.011

size, or hardness difference across a potential failure interface. Structural instability was defined as the tendency of the surrounding snowpack to concentrate shear stresses at the weak layer or interface and to propagate a shear fracture along that layer or interface. They showed that, while no single parameter was a reliable predictor of instability, a simple count of the variables that were in a critical range (threshold sum) provided an approximate indicator of unstable conditions. No comparison to stable profiles was given and it is not clear whether the threshold sum can discriminate between stable and unstable conditions.

Based on a comparison of snow profiles from skier triggered avalanches with profiles from skier-tested slopes that did not release Schweizer and Jamieson (2003) showed that there are significant variables to predict instability and proposed corresponding critical ranges for each variable. Besides the score from a mechanical test (rutschblock), they found the following snow stratification variables to be indicative of snowpack instability: difference in grain size across the failure interface, failure layer grain size, difference in hardness across the failure interface and failure layer hardness. However, the multivariate classification tree they proposed was difficult to apply for operational forecasting and they did not provide any verification of their findings. In addition, their whole analysis was based on the assumption that the critical failure layer was known, i.e. a mechanical test was required to identify the critical weakness. This restriction hinders some applications, in particular, the application of their results to simulated snow cover profiles.

The aim of the present study was to combine the approaches by McCammon and Schweizer (2002) and Schweizer and Jamieson (2003) to (1) provide a robust and easy to use method to assess the probability of skier triggering from snow layer properties at the failure interface, and (2) demonstrate that the method can also be used to find potential failure layers when the location of the critical failure layer is unknown, or to identify additional weaknesses that did not show up in the stability test, or to apply the method to snow profiles simulated by a snow cover model such as SNOWPACK (Lehning et al., 1999).

2. Data

We used snow profile data from the Columbia Mountains of western Canada and the Swiss Alps collected during the winters of 1988–89 to 2003–04. About half of the profiles were taken near the fracture line of or on slopes adjacent to skier-triggered avalanches; these were called “unstable” profiles. The other half were so-called “stable” profiles observed on slopes that were skied but no avalanche was released. We split the data set into a learning data set of 424 cases, the same as used by Schweizer and Jamieson (2003), and a test data set of 109 profiles as shown in Table 1. The test data set included primarily profiles from the winters 2002–03 and 2003–04. Many of the stable profiles in the test data set had relatively poor stability due to targeted sampling (McClung, 2002) compared to the learning data set.

However, comparing the two samples for all seven variables showed that there were no statistically significant differences between the samples except for the variable failure layer hardness ($p=0.02$). Failure layers were slightly softer in the test sample (mean=1.5, median=1) than in the learning sample (mean=1.7, median=1–2). The suspected bias due to targeted sampling was in part confirmed by a slightly lower RB score in the test sample (mean=3.9, median=4) than in the learning sample (mean=4.3, median=4). However, the difference was statistically not significant ($p=0.06$).

For cases with missing data, values were not imputed so that for a multivariate analysis the learning data set reduced to 296 cases without any missing data. There were no missing values in the test data set. Overall, there were 230 stable cases and 175 unstable cases with no missing data.

3. Methods

Five variables were analyzed that showed very high significance as classifiers in the analysis by Schweizer and Jamieson (2003): Rutschblock (RB) score, failure layer (FL) grain size, failure layer hardness and differences in grain size and hardness across the failure interface. These were supplemented with failure layer grain type which also was highly significant in their analysis and failure layer depth. Failure layer depth was introduced to take into account the fact that the probability of skier triggering strongly decreases with increasing slab thickness (Schweizer and Camponovo, 2001; Schweizer and Jamieson, 2001). For shallow weak layers, in the range of the penetration depth, the probability of triggering is also decreasing.

Table 1
Characteristics of snow profile data sets used for model development and testing (number of profiles)

Data set	Country	Stable	Unstable
Learning	Canada	99	117
	Switzerland	105	103
Test	Canada	38	16
	Switzerland	30	25

The above variables are standard snowpack observations and described in Colbeck et al. (1990) and CAA (2002). For failure layer grain size the average grain size (in mm) was used. Failure layer hardness was analyzed using the hand hardness index from 1 to 6 for Fist (F), Four-finger (4 F), One-finger (1 F), Pencil (P) Knife (K) and Ice (I). Intermediate values were allowed, e.g. 2–3, or 2–.

For failure layer grain type primary and secondary grain type were considered for classification into either non-persistent or persistent as proposed by Jamieson and Johnston (1995). Rutschblock tests were performed as described in Schweizer (2002). For profiles from skier-triggered avalanches, rutschblock test results were not always available. Occasionally, a compression test (Jamieson, 1999) was performed instead of a rutschblock test. The compression test score was converted into a comparable rutschblock score.

Differences in average grain size and hardness were calculated between the failure layer and the adjacent layer, i.e. across the failure interface. If the location of the interface was recorded the layer with the lower hardness index was chosen as the possible failure layer. If there was no difference in hardness, the layer with larger grain size was chosen, and if there was no difference at all the lower layer was chosen. If the failure interface was not reported, but the failure layer was known, the failure interface was defined as the boundary with the larger difference in grain size or hardness to the adjacent layer (Schweizer and Jamieson, 2003).

The non-parametric Mann–Whitney *U*-Test was used to contrast variables from the stable/unstable and critical/non-critical data sets. Observed differences were judged to be statistically significant where the level of significance is $p<0.05$.

For each variable stable and unstable data were contrasted to find a split or threshold value that predicts whether the case under consideration belongs into the stable or the unstable category. To find the threshold value the classification tree method was used (Breiman et al., 1998). This corresponds to a binary threshold function for each of the seven variables or classifiers. The outcomes (0 or 1 for each classifier) were then combined by using a simple or weighted sum. This provided a value, also called threshold sum, between e.g. 0 and 7 for the case of unweighted summing. Increasing values of the threshold sum should relate to increasing instability. By applying the classification tree method to the threshold sum a split value could be determined with respect to stable or unstable states. As intermediate values of the failure layer depth are assumed to be most critical (due to the characteristics of the skier impact), there is probably a lower and an upper threshold for instability. Therefore, we also attempted not including failure layer depth in the threshold sum, but in the final classification tree as a second independent variable besides the threshold sum. The above described approach, based on the proposal by McCammon and Schweizer (2002), is comparable in the unweighted case to simply ticking boxes and counting the number of ticks. This is known to be a robust method of combining multiple classifiers that often outperforms more sophisticated expert systems (Kittler et al., 1998), and could be easily used by practitioners. Also it gives a range of instability (e.g. 1 to 7) which allows for indication of quasi-stability.

Table 2
Contingency table (total of cases: $N=a+b+c+d$)

		Observed	
		Stable	Unstable
Forecasted	Stable	*a*: correct stable	*b*: misses
	Unstable	*c*: false alarms	*d*: hits

Table 3
Critical ranges of variables and weights

Variable or classifier	Threshold, critical range	Weights
RB score	<4	2
Difference in grain size (mm)	≥0.75	1
Failure layer grain size (mm)	≥1.25	1
Difference in hardness	≥1.7	1
Failure layer hardness	≤1.3	0.5
Failure layer grain type	Persistent	0.5
Slab thickness or failure layer depth (cm)	18–94	0.5

For failure layer detection, besides the classification tree methodology, discriminant analysis was used for the multivariate analysis. The discriminant analysis provides linear functions of the variables that best separate cases into the two groups of stable and unstable profiles. The variables in the linear function were selected by either a forward or a backward stepwise method. In forward stepping, at each step the variable is included into the model that contributes most to the separation of the groups, whereas backward stepping begins with all variables and subsequently at each step the variable is excluded that is least useful (Flury and Riedwyl, 1988).

To describe the performance of the different models the following measures for categorical forecasts were used: accuracy (or perfect forecast or hit rate, or probability correct forecast (PFC)), the unweighted average accuracy, the probability of detection (POD), the false alarm ratio (FAR), and the true skill score (or so-called

Hanssen and Kuipers discriminant (HK)) (Purves et al., 2003; Wilks, 1995). With the definitions used in contingency tables (Table 2) the measures are calculated as follows:

$$\text{Accuracy or PFC} = \frac{a+d}{N} \quad (1)$$

$$\text{Unweighted average accuracy} = 0.5\left(\frac{a}{a+c}+\frac{d}{b+d}\right) \quad (2)$$

$$\text{Probability of detection : POD} = \frac{d}{b+d} \quad (3)$$

$$\text{False alarm ratio : FAR} = \frac{c}{c+d} \quad (4)$$

$$\text{True skill score : HK} = \frac{d}{b+d}-\frac{c}{a+c} \quad (5)$$

The accuracy measures the overall success of a model. The unweighted average accuracy accounts better for rare events than the accuracy. The true skill score is a measure of the forecast success at discriminating between stable and unstable cases correctly. The probability of false-stable predictions is defined as 1 − POD. Although snow profiles are only one factor considered in avalanche forecasting, false-stable predictions can have more serious consequences than false alarms.

4. Results

We first report on stability classification by the threshold sum method and then on detection of potentially critical failure layers. The stability classification analysis was performed for the combined Swiss–Canadian sample, whereas the layer selection analysis was done separately.

4.1. Stability classification

Table 3 shows the critical ranges that were used for the initial classification by the unweighted threshold sum. For the first five variables the critical ranges or threshold values are the ones given by Schweizer and Jamieson (2003). For the failure layer grain type the critical range was defined as persistent (surface hoar, depth hoar or faceted crystals) and for the failure layer depth the critical range was chosen arbitrarily based on the 5th percentiles and the 95th percentiles (middle 90%).

The univariate classification power of the seven variables is summarized in Table 4. It shows that the rutschblock (RB) score was the classifier with the highest accuracy and it best discriminated between stable and unstable cases. The second and third best classifiers considering the true skill score were the difference in grain size and the difference in hardness across the failure interface.

A classification tree with the unweighted threshold sum as a single independent variable suggested a threshold sum of 5 as split value, i.e. <5 mostly stable, ≥ 5 mostly unstable. As can be seen from Fig. 1 the threshold sum seems to discriminate quite clearly between stable and unstable (non-parametric Mann–Whitney U-Test, $p<0.001$). A threshold sum of 4 can be attributed to a transitional range between mostly stable and mostly unstable conditions. The accuracy measures of this model are given in Table 5. If the failure layer depth was omitted the different accuracy scores were only slightly different. However, if the rutschblock score was not considered, particularly the probability of detection and the true skill score decreased.

Alternatively, the failure layer depth can be considered as a second independent variable besides the threshold sum in the final classification tree. This revealed first of all a split value of 4 for the threshold sum, but then the tree suggested to classify the cases with a threshold sum ≥ 4 as unstable only if the failure layer depth

Table 4
Univariate classification power for the learning data set

Variable	Number of cases, N	Accuracy (%)	Unweighted average accuracy (%)	Probability of detection (POD) (%)	False alarm ratio (FAR) (%)	True skill score HK (%)
RB score	369	67.8	66.7	56.0	33.6	33.0
Difference in grain size (mm)	356	65.8	65.5	71.9	34.5	31.0
Failure layer grain size (mm)	421	65.6	65.0	71.8	36.2	23.8
Difference in hardness	401	63.6	64.0	52.7	30.6	28.0
Failure layer hardness	401	63.1	63.3	58.0	33.7	26.6
Failure layer grain type	424	58.3	57.8	70.9	42.0	15.5
Slab thickness or failure layer depth (cm)	424	51.9	51.4	90.5	48.2	0.7

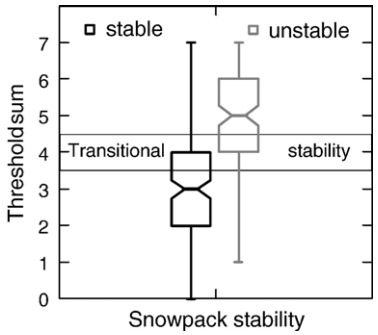

Fig. 1. Unweighted threshold sum with 7 variables for stable and unstable samples of snow profiles (N=296). The band of transitional stability indicates the threshold sum values for which about 50% of the cases are classified either as stable or unstable. Stable data is given on the left and unstable on the right. Boxes span the interquartile range from 1st to 3rd quartile with a horizontal line showing the median. Notches at the median indicate the confidence interval (p<0.05). Whiskers show the range of observed values that fall within 1.5 times the interquartile range above and below the interquartile range.

was either ≥24 cm or ≤94 cm. We interpret this as the approximate range in this data set for which the slabs in this data set had sufficient stiffness for fracture propagation (≥24 cm) and were not too thick for skiers to reliably initiate fractures (≤94 cm). Cases with threshold sum ≥4 that did not fall in this range of failure layer depth were classified as stable. This slightly improved the accuracy scores.

Table 4 shows that not all variables had the same classification power. Accordingly, weighting the classifiers seemed appropriate. We did not try to optimize the weights, but have chosen them such that the method remains simple and easily applicable for practitioners. A discriminant analysis provided a priori values in the form of the coefficients of the canonical discriminant function, from which we derived the weights given in Table 3.

The weighted threshold sum including all 7 classifiers improved the true skill score by about 4%. The split value given by the classification tree to discriminate between stable and unstable profiles was 3.5, i.e. a threshold sum of 0 to 3 indicated mostly stable conditions (good stability), 3.5 to 6.5 mostly unstable conditions (poor stability) with a transitional range of 3.5 to 4 (fair stability). In this range about 50% of the profiles were each rated as stable and 50% as unstable. Finally, the two models that showed the best performance in terms of true skill score were combined: weighted threshold sum with 6 variables and the failure layer depth as a second independent variable in the stability assessment with the classification tree. The split value suggested by the tree method was 3 (<3 mostly stable, ≥3 mostly unstable). The band of transitional stability ranged from 3 to 3.5. The classification tree (Fig. 2) showed that independent of failure layer depth, threshold sums of 0 to 2.5 indicated mostly stable conditions, and of 4.5 to 6

Table 5
Classification accuracy of different models for profiles with known failure layer

Model	Critical range	Accuracy (%)	Probability of detection (POD) (%)	False alarm ratio (FAR) (%)	True skill score HK (%)
Unweighted threshold sum (7 variables)	≥5	72 (64.2)	61.2 (58.6)	27.4 (47.8)	42.1 (26.2)
Unweighted threshold sum (6 variables, without FL depth)	≥4	71.6 (64.2)	61.9 (58.6)	28.4 (47.8)	41.5 (26.2)
Unweighted threshold sum (5 variables, without RB score and FL depth)	≥4	67.5 (66.1)	52.3 (48.8)	22.0 (44.4)	36.3 (25.3)
Unweighted threshold sum (6 variables) plus FL depth	≥4 and 24–93 cm	74.7 (66.1)	59.0 (58.5)	20.2 (41.7)	46.6 (29.1)
Weighted threshold sum (7 variables)	≥3.5	73.7 (63.3)	71.6 (73.2)	29.2 (49.2)	46.3 (30.5)
Weighted threshold sum (6 variables) plus FL depth	≥4.5 or 3–4 and 34–78 cm	77.0 (65.1)	64.9 (63.4)	19.4 (46.9)	52 (29.6)
Unweighted threshold sum (6 variables without RB score)	≥5	66.7 (66.1)	49.4 (48.8)	20.9 (44.4)	35.2 (25.3)
Unweighted threshold sum (6 variables without RB score)	≥4	67.2 (63.3)	73.3 (73.2)	33.2 (49.2)	33.8 (30.5)
Weighted threshold sum (6 variables without RB score)	≥3.5	67.4 (66.1)	52.3 (48.8)	22.0 (44.4)	36.2 (25.3)
Weighted threshold sum (6 variables without RB score)	≥3	68.9 (64.2)	70.5 (61.0)	29.9 (47.9)	37.7 (27.2)

Scores are given for learning data set and below in parentheses for test data set. The first 6 models were used to classify profiles with known failure layer. The last 4 models were subsequently used to classify the potentially critical layers determined with the layer selection routine.

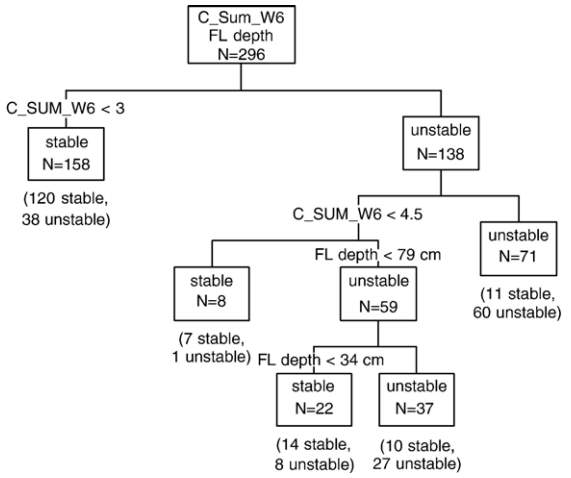

Fig. 2. Classification tree for model with weighted threshold sum (6 variables) and failure layer depth as a second independent variable (for learning data set: $N=296$).

mostly unstable conditions. In the intermediate range of 3 to 4 the failure layer depth was decisive. If the failure layer depth was ≥ 34 cm and <79 cm the profiles were rated as unstable, otherwise as stable. This model had the best performance of all models. Compared to the initial model with 7 unweighted variables the accuracy increased by 5% and the true skill score by 10%. However, the increase in the true skill score was in part due to a decreased false-alarm rate rather than an increase of the probability of detection, i.e. a decrease of the false-stable predictions. At 35%, the proportion of false-stable predictions was relatively high.

The performance of the test data set with 109 profiles was on average for the 10 models given in Table 5 only slightly poorer than for the learning set in regard to accuracy and probability of detection. However, the false alarm ratio was substantially higher. The true skill score decreased by about one-third from 41% to 28%.

McCammon and Schweizer (2002) pointed out that the threshold sum approach might have potential to avoid false-stable conditions. We considered all unstable profiles with rutschblock score ≥ 4 as potential cases of false-stable prediction ($N=60$). Applying the weighted threshold sum (6 variables without the RB score) revealed that 21 out of 60 cases (35%) with threshold sum ≥ 3.5 were rated as unstable. Another 20 cases had a threshold sum value of 3 which was considered as transitional. In total, when applying the weighted threshold sum with 6 variables and a threshold value of 3 (≥ 3: unstable), 67% of the potential false-stable predictions were recognized as rather unstable, in contrast to the RB score.

To compare the threshold sum approach to a pure classification tree analysis, we cross-validated the threshold sum approach model which had the best accuracy (weighted threshold sum with 6 variables and the failure layer depth as a second independent variable) and a direct classification tree using all 7 variables. We performed a 10-fold cross-validation, i.e. we randomly split the joint data set (learning and test: $N=405$) into 10 groups, trained the model with nine groups and tested the model on the excluded group, and after performing this ten times with different combinations of the training set, averaged the performance. The cross-validated average accuracy of these 10 runs was $68\pm5.5\%$ for the threshold sum model and $70\pm8.1\%$ for the classification tree. Hence, the decrease in classification accuracy of the simple threshold sum method compared to the direct classification tree using all seven variables was relatively minor and justified by the ease of application of the former method.

4.2. Failure layer detection

When searching for potentially critical failure layers, the stable and unstable data sets were combined and critical failure layers (identified by stability tests or by an avalanche) were contrasted with non-critical failure layers which included all other layers in the profiles.

Table 6
Contrasting critical and non-critical layers for each the Swiss and Canadian data set

Variable	Swiss				Canadian			
	Critical, $N=270$	Non-critical, $N=2000$	U-test, p	Critical range	Critical, $N=216$	Non-critical, $N=1833$	U-test, p	Critical range
Difference in grain size (mm)	0.5	0.375	<0.001	≥1.125	1.25	0.25	<0.001	≥0.7
Failure layer grain size (mm)	1.125	1.0	0.293	N/A	2	1.0	<0.001	≥1.6
Difference in hardness	1.5	1	<0.001	≥1.5	1.3	1	<0.001	≥1.2
Failure layer hardness	1 (F)	1.5 (F to 4F)	<0.001	≤1.5 (F to 4F)	2 (4F)	3 (1F)	<0.001	≤2.7
Failure layer grain type*	Persistent	Persistent	0.548	N/A	Persistent	Non-persistent	<0.001	Persistent
Failure layer depth (cm)	43	54	0.002	N/A	49	47.5	0.86	N/A

Medians are given, and in case of the grain type the mode (*) is shown. The critical range resulted from the first split of a univariate classification tree.

Table 7
Critical ranges for failure layer selection

Variable	Critical range	
	Switzerland	Canada
Failure layer grain size	≥1.125 mm	≥1.2 mm
Failure layer hardness	≤1.5 (F to 4F)	≤2.7 (1F–)
Difference in grain size	≥1.125 mm	≥0.7 mm
Difference in hardness	≥1.5	≥1.3
Failure layer grain type	Persistent	Persistent
Failure layer depth	13–89 cm	19–86 cm

This data set of layers was unbalanced with only about 11% critical layers compared to 89% non-critical layers (Table 6).

Initially, we tried to apply the same critical ranges given in Table 3 that were used for the stability assessment. With these ranges, for e.g. the Swiss profiles, about 35% of the critical layers in the stable profiles and about 64% in the unstable profiles were correctly classified. The scores for the test data set were similar: 37% and 63%, respectively. A potentially critical layer was considered correctly classified as critical if the threshold sum was maximal at either the upper or lower interface of the failure layer.

Table 6 summarizes the properties of critical and non-critical failure layers for Swiss and Canadian profiles. It shows that in general the differences between critical (failure) layers and non-critical layers were small, in particular for the Swiss profiles. Accordingly, for the Swiss profiles, the distributions for failure layer grain size and failure layer grain type were statistically not significantly different. Also, by applying a univariate classification tree to find split values for defining the critical range no split values were found for failure layer grain size and type, and failure layer depth. This was partly because the samples were very unbalanced. In the case of the Canadian profiles, the distributions for failure layer depth were statistically not significantly different and no split value was found by the univariate classification tree.

Applying discriminant analysis showed that, in the case of the Swiss data set, all four independent variables were chosen. None was rejected in either the forward or backward stepwise mode. The discriminant function had an overall success score of 61% (jackknifed classification). The success scores were 61% for the non-critical cases and 64% for the critical cases. In the discriminant functions failure layer hardness and difference in hardness had more weight than failure layer grain size and difference in grain size. This was in contrast to the above stability classification. As there were substantial differences between the Swiss and the Canadian samples the subsequent analysis was done separately.

4.2.1. Swiss data set

For the case of the Swiss profiles, a re-analysis was done for the variables failure layer grain size and failure layer depth. Comparing failure layer grain size to failure layer depth showed a statistically significant correlation between the two variables ($N=2270$, $R^2=0.08$, $p<0.001$) with larger grain sizes at larger depth. It was assumed that the large grain sizes usually existing in bottom layers had caused failure layer grain size to no longer be a significant variable (see above, Table 6).

Therefore, all cases with failure layer depth ≥100 cm were excluded from the analysis of finding a critical range. By restricting the sample, there was again a statistically significant difference (U-Test, $p=0.024$) in failure layer grain size when contrasting critical to non-critical layers. With the univariate classification tree method a new split value (≥1.125 mm) was found. With this split value 53.8% of all layers were correctly classified (54.3% of non-critical layers, 50.7% of critical layers). The univariate tree analysis for the failure layer depth resulted in a new critical range of 13 to 89 cm. Within this range, 89.6% of all critical layers were found, whereas only 38.8% of the non-critical layers were within this range. The new critical ranges are given in Table 7.

With the critical ranges of Table 7 and an unweighted threshold sum (6 variables) 42% of the critical failure layers in the stable profiles and 64% in the unstable profiles were recognized (Table 8). The unweighted average accuracy was 53%. The layer selection routine typically proposed many more (43%) layers as potentially critical layers than was observed. Weighting the

Table 8
Failure layer selection for Swiss (left) and Canadian (right) profiles

Data set	Layers	Observed failure layers	Potentially critical failure layers	Correctly classified failure layers (%)	Unweighted average accuracy (%)
Stable learning	1418/958	153/99	208/160	42/70	53/73
Unstable learning	1059/1092	132/117	199/174	64/77	
Stable test	404/336	38/38	72/55	53/58	62/70
Unstable test	289/135	38/16	64/27	71/81	

variables as in the case of stability classification did not improve detection results, but reduced the number of ties. The performance of the test data was better: 53% of the critical layers in the stable profiles, and 71% in the unstable profiles were correctly recognized.

Finally, we tried to assess stability discrimination for the potentially critical layers that were found with the layer selection procedure. For each of these layers the unweighted and weighted threshold sum (6 variables without the RB score) with the critical ranges as given in Table 3 were calculated. Based on the threshold value ≥ 5 for the unweighted sum and ≥ 3.5 for the weighted sum the potentially critical layers were classified into stable and unstable (Table 5). If in one profile the classification for different critical layers was different, the unfavourable case was considered, i.e. the profile was classified as unstable. Contingency tables for the learning data set showed that the probability of detection was poor: below 50% (Table 9). Consequently, we tried to improve the probability of detection (and to decrease the rate of false-stable predictions) by decreasing the threshold value to 4 for the unweighted sum and to 3 for the weighted sum. This adjustment improved the probability of detection but also increased the number of false alarms, so that the true skill score remained very poor. As in the case of the stability classification with known critical weakness (see above) the weighted threshold sum performed slightly better. The scores for the test data set were nearly always higher than for the learning data set.

4.2.2. Canadian data set

The thresholds of the 6 variables in Table 6 were adjusted to optimize the true skill score for discriminating between critical and non-critical layers in the learning sample. This yielded the same results for layer hardness, difference in grain size and grain type. However, the revised thresholds for grain size, difference in hardness were 1.2 mm and 1.3 (F+), respectively. Also, optimizing the true skill score for failure layer depth yielded lower and upper thresholds of 19 and 86 cm, respectively. While it is unclear whether the classification tree or optimizing the true skill score is preferable for selecting thresholds (Table 7), we chose to use the thresholds from the true skill score so that all thresholds, including those for failure layer depth were selected by a single method.

With the critical ranges of Table 7 and an unweighted threshold sum (6 variables) 70% of the critical failure layers in the stable profiles and 77% in the unstable profiles were recognized (Table 8). The unweighted average accuracy was 73%, suggesting that critical layers are more easily recognized in the Columbia Mountains of western Canada than in the Swiss Alps. The layer selection routine recognized 55% more layers as potentially critical layers than were observed; however, this is not surprising since only one critical layer per profile was selected as defined by the avalanche, rutschblock or compression test (layer with lowest score). Weighting the variables reduced the number of potentially critical layers but also reduced the unweighted average accuracy to 65%. The performance of the test data was less balanced. For stable profiles and unstable profiles, respectively, 58% and 81% of the critical layers were recognized.

As with the Swiss profiles, we assessed the stability for the potentially critical failure layers that were found with the layer selection procedure. The results are shown in Table 9 for the unweighted sum with thresholds of 4 and 5 and weighted sum with thresholds 3 and 3.5. Contingency tables for the learning data set showed that the probability of detection ranged from 65% to 92%, and the false alarm ratios were quite high, ranging from 35% to 42% (Table 9). The true skill score

Table 9
Classification accuracy of different models for Swiss (left) and Canadian (right) profiles evaluated separately for the case when failure layer is unknown

Model	Critical range	Accuracy (%)	Probability of detection (POD) (%)	False alarm ratio (FAR) (%)	True skill score HK (%)
Unweighted threshold sum	≥5	47.1/62.0	40.8/65.0	53.8/35.0	−5.9/23.5
(6 variables without RB score)		(49.1/69.0)	(68.0/75.0)	(54.1/52.0)	(1.3/40.8)
Unweighted threshold sum	≥4	49.0/60.2	73.8/92.3	51.0/41.6	−1.5/14.5
(6 variables without RB score)		(49.1/35.2)	(88.0/81.3)	(53.2/71.1)	(4.7/−3)
Weighted threshold sum	≥3.5	47.1/60.6	48.5/70.9	53.3/38.1	−5.7/19.4
(6 variables without RB score)		(50.9/63.0)	(68.0/75.0)	(52.8/57.1)	(4.7/32.9)
Weighted threshold sum	≥3	52.9/59.7	72.8/88.9	48.3/41.6	6.1/14.1
(6 variables without RB score)		(49.1/40.7)	(76.0/81.3)	(53.7/69.0)	(2.7/5)

Stability classification is based on the potentially critical failure layers. Those were classified into stable or unstable and the stability rating compared with the observed stability rating of the profiles. Scores are given for learning data set and below in brackets for test data set. Swiss scores are given on the left, Canadian on the right.

was highest with the unweighted sum and a threshold of 5. However, the other models gave higher probabilities of detection but more false alarms, which might be acceptable in some circumstances given that the consequences for false-stable predictions are much greater than for false alarms. For the test data set, the probability of detection and true skill scores were better for the higher thresholds: 5 unweighted and 3.5 weighted, and much lower for the models with lower thresholds. We attribute the better results for the Canadian data compared to the Swiss data to the better recognition of critical weak layers (Table 8).

5. Conclusions

Introducing a simple threshold sum approach to discriminate between stable and unstable profiles proved to be successful for the case when the principal weakness was known. Seven variables were used to determine the threshold sum: Rutschblock score, failure layer grain size, failure layer hardness and differences in grain size and hardness across the failure interface, failure layer grain type and failure layer depth. The rutschblock score was the classifier with the highest skill score i.e. it best discriminated between stable and unstable cases, followed by difference in grain size and the difference in hardness across the failure interface. We did not consider fracture type (Schweizer, 2002), fracture character (van Herwijnen and Jamieson, 2006-this issue) or shear quality (Johnson and Birkeland, 2002) of the stability test since for most of the profiles these data were not available, but we recommend to consider it for stability assessment (Schweizer et al, in press).

A weighted threshold sum with the failure layer depth as independent variable scored highest (77% for the learning data set, 65% for the test data set, cross-validated accuracy: 68%). Detection of potential critical layers proved to be less successful, in particular for the Swiss profiles. If the principal weakness was unknown, the stability classification for the potentially critical layers agreed with the observed stability for the Swiss profiles in about 53% and for the Canadian profiles in about 62% of the cases. Potential failure layers were better recognized in unstable profiles than in stable profiles. In part, the limited accuracy of these results is inherent to point observations of the spatially variable snowpack since skiers can trigger slabs where snowpack properties are substantially different from the nearby profile site.

The results emphasize that stability assessment should include – besides stability tests that help locate the principal weakness and provide a first assessment (score and fracture type) – analysis of snow layer properties, in particular grain size, type and hardness of layers adjacent to the principal weakness. The proposed threshold sum considering seven variables is well suited for analysis of manual profiles by practitioners. The procedures for identifying critical layers and for assessing the stability of profiles may prove to be a useful training tool. Stability classification of snow profiles simulated by snow cover models such as SNOWPACK will need further adaptation (Schweizer et al., 2006), especially for application in areas with a primarily transitional snow climate (McClung and Schaerer, 1993) such as in the Swiss Alps.

Acknowledgements

We are grateful to numerous people who collected the profile data. We thank Charles Fierz for help with data analysis. For the Canadian contribution to this paper we acknowledge the financial support from the BC Helicopter and Snowcat Skiing Operators Association, the Natural Sciences and Engineering Research Council of Canada, Mike Wiegele Helicopter Skiing, Canada West Ski Areas Association and the Canadian Avalanche Association. We are grateful to the Editor Kelly Elder and to two anonymous reviewers who made valuable suggestions that helped to improve the paper.

References

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1998. Classification and Regression Trees. CRC Press, Boca Raton, U.S.A. 368 pp.

CAA, 2002. Observation Guidelines and Recording Standards for Weather, Snowpack and Avalanches. Canadian Avalanche Association (CAA), Revelstoke BC, Canada. 78 pp.

Colbeck, S.C., Akitaya, E., Armstrong, R., Gubler, H., Lafeuille, J., Lied, K., McClung, D., Morris, E., 1990. The International Classification of Seasonal Snow on the Ground. International Commission on Snow and Ice (ICSI). International Association of Scientific Hydrology, Wallingford, Oxon, U.K. 23 pp.

Flury, B., Riedwyl, H., 1988. Multivariate Statistics: A Practical Approach. Chapman and Hall, London, U.K. 296 pp.

Jamieson, J.B., 1999. The compression test – after 25 years. Avalanche Rev. 18 (1), 10–12.

Jamieson, J.B., Johnston, C.D., 1995. Monitoring a shear frame stability index and skier-triggered slab avalanches involving persistent snowpack weaknesses. Proceedings International Snow Science Workshop, Snowbird, Utah, U.S.A., 30 October–3 November 1994. ISSW 1994 Organizing Committee, Snowbird UT, U.S.A., pp. 14–21.

Johnson, R.F., Birkeland, K.W., 2002. Integrating shear quality into stability test results. In: J.R. Stevens (Editor), Proceedings ISSW 2002, International Snow Science Workshop, Penticton BC, Canada, 29 September–4 October 2002. International Snow Science Workshop Canada Inc., BC Ministry of Transportation, Snow Avalanche Programs, Victoria BC, Canada, pp. 508–513.

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20 (3), 226–239.

Kronholm, K., 2004. Spatial variability of snow mechanical properties with regard to avalanche formation. Ph.D. Thesis, University of Zurich, Zurich, Switzerland, 192 pp.

Lehning, M., Bartelt, P., Brown, R.L., Russi, T., Stöckli, U., Zimmerli, M., 1999. Snowpack model calculations for avalanche warning based upon a network of weather and snow stations. Cold Reg. Sci. Technol. 30 (1–3), 145–157.

McCammon, I., Schweizer, J., 2002. A field method for identifying structural weaknesses in the snowpack. In: Stevens, J.R. (Ed.), Proceedings ISSW 2002. International Snow Science Workshop, Penticton BC, Canada, 29 September–4 October 2002. International Snow Science Workshop Canada Inc., BC Ministry of Transportation, Snow Avalanche Programs, Victoria BC, Canada, pp. 477–481.

McClung, D.M., 2002. The elements of applied avalanche forecasting: Part I. The human issues. Nat. Hazards 26 (2), 111–129.

McClung, D.M., Schaerer, P., 1993. The Avalanche Handbook. The Mountaineers, Seattle WA, U.S.A. 271 pp.

Purves, R., Morrison, K.W., Moss, G., Wright, D.S.B., 2003. Nearest neighbours for avalanche forecasting in Scotland – development, verification and optimisation of a model. Cold Reg. Sci. Technol. 37 (3), 343–355.

Schweizer, J., 2002. The Rutschblock test – Procedure and application in Switzerland. Avalanche Rev., 20 (5): 1, 14–15.

Schweizer, J., Camponovo, C., 2001. The skier's zone of influence in triggering slab avalanches. Ann. Glaciol. 32, 314–320.

Schweizer, J., Jamieson, J.B., 2001. Snow cover properties for skier triggering of avalanches. Cold Reg. Sci. Technol. 33 (2–3), 207–221.

Schweizer, J., Jamieson, J.B., 2003. Snowpack properties for snow profile interpretation. Cold Reg. Sci. Technol. 37 (3), 233–241.

Schweizer, J., Jamieson, J.B., Schneebeli, M., 2003. Snow avalanche formation. Rev. Geophys. 41 (4), 1016. doi:10.1029/2002RG000123.

Schweizer, J., Bellaire, S., Fierz, C., Lehning, M., Pielmeier, C., 2006. Evaluating and improving the stability predicitons of the snow cover model SNOWPACK. Cold Reg. Sci. Technol. 46 (1), 52–59.

Schweizer, J., McCammon, I., Jamieson, J.B., in press. Snow slope stability evaluation using concepts of fracture mechanics. In: Gleason, J.A. (Ed.), Proceedings ISSW 2006. International Snow Science Workshop, Telluride CO, U.S.A., 1-6 October 2006.

van Herwijnen, A., Jamieson, J.B. 2007. Fracture character in compression tests. Cold Reg. Sci. Technol. 47, 64–72 (this issue). doi:10.1016/j.coldregions.2006.08.016.

Wilks, D.S., 1995. Statistical methods in the atmospheric sciences: an introduction. International Geophysics, vol. 59. Academic Press, San Diego CA, U.S.A. 467 pp.