ASSESSING THE PROBABILITY OF SKIER TRIGGERING FROM SNOW LAYER PROPERTIES

Jürg Schweizer[1], Charles Fierz[1] and J. Bruce Jamieson[2]

[1] Swiss Federal Institute for Snow and Avalanche Research SLF, Davos, Switzerland

[2] Department of Civil Engineering, Department of Geology and Geophysics, University of Calgary, 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 the threshold sum approach to assess snowpack stability based on layer properties. Second, we investigated whether that method is 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 consists 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 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. 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

1. INTRODUCTION

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. 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 as grain type and size (e.g. Kronholm, 2004).

McCammon and Schweizer (2002) proposed to complement information on mechanical instability such as the shear strength or stability test scores with data on structural instability such as grain type and 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. It was 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 RB score they found the following snow stratification variables to be indicative of snowpack instability: difference in

Corresponding author address:
Jürg Schweizer, Flüelastrasse 11,
CH-7260 Davos Dorf, Switzerland
phone: +41 81 4170111, fax: +41 81 4170110,
e-mail: schweizer@slf.ch

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 localize the critical weakness. This restriction hinders e.g. the application of their results to simulated snow cover profiles.

The aim of the present study is 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, 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 e.g. 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 at 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.

3. METHODS

As classifiers the five variables were used that showed very high significance 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 completed 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 that the probability of skier triggering strongly decreases with increasing slab thickness (Schweizer and Camponovo, 2001; Schweizer and Jamieson, 2001).

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 (4F), One-finger (1F), Pencil (P) and Knife (K). 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). In the case of the profiles from skier-triggered avalanches rutschblock tests were often not available. Occasionally, a compression test (Jamieson, 1999) was performed instead. Compression test scores were converted into comparable rutschblock scores.

Differences in 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 failure layer. If there was no difference in hardness, the layer with larger grain size was considered as the failure layer, and if there was no difference at all the lower layer was chosen as failure layer. If the failure interface was not reported, but the failure layer was known, first the difference in hardness and second the difference in grain size were considered to chose either the layer above or below the failure layer as the adjacent layer.

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). For each of the seven variables or classifiers a binary threshold function was determined. The outcomes (0 or 1 for each

classifier) were then combined by 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 threshold sum should relate to increasing instability. By applying the classification tree method to the threshold sum a split value was determined with respect to stable or unstable. As failure layer depth is not the same type of classifier as the other variables it was also attempted to not include the failure layer depth into the threshold sum, but into the final assessment as second independent variable besides the threshold sum. This approach, based on the proposal by McCammon and Schweizer (2002), is comparable in the unweighted case with 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, and it is first of all simply applicable by practitioners. Also it gives a range of instability (e.g. 1 to 7) which allows for indication of quasi- or transitional stability.

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 rate: FAR, and the true skill score (or so-called Hanssen and Kuipers discriminant: POD-FAR) (Purves et al., 2003; Wilks, 1995).

With the definitions used in contingency tables (Table 2) the measures are calculated as follows:

Accuracy or PFC = $\frac{a+d}{n}$

Unweighted average acc. = $0.5\left(\frac{a}{a+c}+\frac{d}{b+d}\right)$

Probability of detection POD = $\frac{d}{b+d}$

False alarm rate FAR = $\frac{c}{a+c}$

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

		observed stable	unstable
forecasted	stable	a : correct stables	b : misses
	unstable	c : false alarms	d : hits

True skill score POD – FAR = $\frac{d}{b+d} - \frac{c}{a+c}$

The accuracy measures the overall success of a model (correct classification of non-events and events). 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. Misses or false-stable predictions are given by 1-POD.

4. RESULTS

We first report on stability classification by the threshold sum method and then on detection of critical failure layers. The stability classification was done for the combined Swiss-Canadian sample, whereas layer section 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 were the ones given by Schweizer and Jamieson (2003). For the failure layer grain type the critical range was defined as persistent and for the failure layer depth the critical range was chosen arbitrarily based on the 5th percentiles and the 95th percentiles (middle 90%).

A univariate analysis showed that the RB score was the classifier with the highest accuracy and 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 single independent variable suggested a threshold sum of 5 as split value, i.e. < 5 stable, ≥ 5 unstable. As can be seen from Figure 1 the threshold sum seems to discriminate quite clearly between stable and unstable (non-parametric Mann-Whitney U-Test, p < 0.001). The accuracy measures of this model are given in Table 4. 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 POD and the true skill score decreased.

Alternatively, the failure layer depth can be considered as second independent variable besides the threshold sum in the final classification tree. This revealed first of all a split

Table 3: Critical ranges of variables and weights

Variable or classifier	Threshold, critical range	Weight
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

value of 4 for the threshold sum, but then the tree suggested to classify the cases with threshold sum ≥ 4 as unstable only if the failure layer depth was either ≥ 24 cm or ≤ 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.

As not all variables had the same classification power, 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 the coefficients, we derived the weights as 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, 3.5 to 6.5 mostly unstable conditions with a transitional range of 3.5 to 4. 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 stable, ≥ 3 unstable). The band of transitional stability

Table 4: Classification accuracy of different models for profiles with known failure layer. Scores are given for learning data set and below in brackets for test data set.

Model	Critical range	Accuracy (%)	Probability of detection POD (%)	False alarm rate FAR (%)	True skill score POD-FAR (%)
Unweighted threshold sum (7 variables)	≥ 5	72 (64.2)	61.2 (58.6)	19.1 (32.4)	42.1 (26.2)
Unweighted threshold sum (6 variables, without FL depth)	≥ 4	71.6 (64.2)	61.9 (58.6)	20.4 (32.4)	41.5 (26.2)
Unweighted threshold sum (5 variables, without RB score and FL depth)	≥ 4	67.5 (66.1)	52.3 (48.8)	16 (23.5)	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)	12.3 (29.4)	46.6 (29.1)
Weighted threshold sum (7 variables)	≥ 3.5	73.7 (63.3)	71.6 (73.2)	25.3 (42.6)	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)	13 (33.8)	52 (29.6)

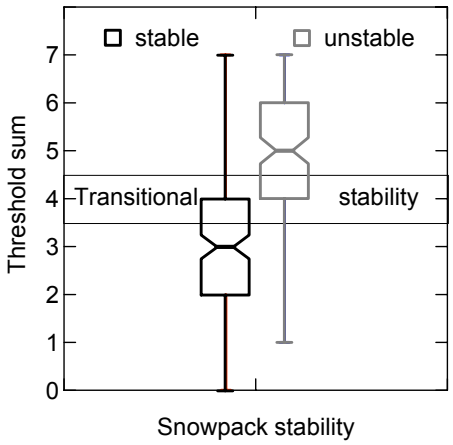

Figure 1: Unweighted threshold sum with 7 variables for stable and unstable sample of snow profiles. The band of transitional stability indicates the threshold sum values for which about 50% of the cases are either classified as stable or unstable. Stable data is given on the left and unstable on the 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.

ranged from 3 to 3.5. The classification tree showed that independent of failure layer depth, threshold sums of 0 to 2.5 indicated mostly stable conditions, and of 4.5 to 6 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 unfortunately due a decreased false alarm rate rather than an increase of the POD, i.e. a decrease of the false-stable predictions. With 35% the proportion of false-stable predictions was relatively high.

The performance of the test data set with 109 cases primarily from the winters 2002-03 and 2003-04 was on average for the 6 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 rate was substantially higher and accordingly 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 is 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 potentially unstable – in contrast to the RB score.

4.2 Failure layer detection

When searching for potentially critical failure layers, the stable and unstable datasets 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. This left the dataset unbalanced with only about 11% critical layers compared to 89% non-critical layers. 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. If several layers had the maximum score they were all considered as potentially critical.

Initially, we tried to apply the same critical ranges as given in Table 3 that were used for the stability assessment. However, as the differences

Table 5: 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

between critical (failure) layers and non-critical layers were small, in particular for the Swiss profiles, the POD was low. Re-analysis showed that for layer selection separate threshold values were needed (Table 5). With these critical ranges and an unweighted threshold sum (6 variables) 42% of the critical failure layers in the stable Swiss profiles and 64% in the unstable Swiss profiles were recognized. The layer selection routine typically proposed about 1.4 times more layers as potentially critical layers than were observed. Weighting the 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. The detection rate for the Canadian profiles was higher: 72% for the stable profiles and 74% for the unstable profiles.

Finally, stability for the potentially critical failure layers that were found with the layer selection procedure was assessed. For each of the potential failure layers the unweighted and weigthed threshold sum (6 variables without the RB score) were calculated. The critical ranges were ≥ 5 for the unweighted threshold sum and ≥ 3.5 for the weighted sum. If in one profile the classification for different layers was different, the unfavourable case was considered, i.e. the profile was classified as unstable. As in the case of the stability classification with known critical weakness (see above) the weighted threshold sum performed slightly better. However, for the Swiss profiles the accuracy was low, i.e. just about 53%, whereas it was higher for Canadian profiles: 62%. The scores for the control dataset were comparable and for the Swiss profiles nearly always higher than for the learning data set.

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. 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 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.

The results emphasize that stability assessment should include – besides stability tests that help to locate the principal weakness – analysis of snow layer properties, in particular grain size, type and hardness. Even when doing a hasty stability test observations of failure and adjacent layer properties can improve the assessment. 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 e.g. the snow cover model SNOWPACK will need further adaptation, in particular for application in transitional snow climates.

ACKNOWLEDGEMENTS

We are grateful to numerous people who collected the profile data. 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.

REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A. and 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. and 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.

Jamieson, J.B., 1999. The compression test - after 25 years. The Avalanche Review, 18(1): 10-12.

Jamieson, J.B. and 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.

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. and Zimmerli, M., 1999. Snowpack model calculations for avalanche warning based upon a network of wether and snow stations. Cold Reg. Sci. Technol., 30(1-3): 145-157.

McCammon, I. and Schweizer, J., 2002. A field method for identifying structural weaknesses in the snowpack. 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. 477-481.

Purves, R., Morrison, K.W., Moss, G. and 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): 343-335.

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

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

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

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

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