Cold Regions Science and Technology 64 (2010) 104–109

Contents lists available at ScienceDirect

Cold Regions Science and Technology

journal homepage: www.elsevier.com/locate/coldregions

On stability sampling strategy at the slope scale

Jürg Schweizer *, Sascha Bellaire

WSL Institute for Snow and Avalanche Research SLF, Flüelastrasse 11, CH-7260 Davos, Dorf, Switzerland

ARTICLE INFO

Article history:
Received 5 December 2009
Accepted 27 February 2010

Keywords:
Snow cover
Snow sampling
Snow stability evaluation
Snow stability test
Snow avalanche release

ABSTRACT

Snow slope stability evaluation is often based on a single test location within a slope. However, we know that snow cover properties and stability may vary at the slope scale. Reliably estimating the slope-scale variability requires many samples, ideally more than 100. As this is unpractical, it has been proposed to perform at least two tests — about 10 m apart — on a given slope. In addition, if small column stability tests are used (such as the compression test), it seems reasonable to perform two tests at each of the two locations. Differences between the two tests at one location allow one to assess the small (or pit-) scale variability (and/or the test uncertainty), whereas differences between the pairs at different locations may hint at the slope-scale variability. We analyzed 22 small slopes each with four pairs of stability tests. In 61-75% of the cases the two stability tests at a specific location provided consistent results, depending whether we focused on the CT score or the fracture character (which was less variable). Comparing the different sampling locations on a given slope (~10–15 m apart) showed that at the slope-scale the differences between sampling locations (59–75%) were similar to the differences found at the pit-scale. Rather stable slopes tended to have more pit-scale variation than rather unstable slopes. Based on our analysis, we suggest an interpretation scheme and an adjusted sampling procedure. In particular, a second pit on a slope seems only necessary if the first pit does not indicate instability. In all other cases, a second pit can reduce the number of false-stable predictions.

1. Introduction

For assessing avalanche risk in backcountry operations and for public forecasting of the regional avalanche danger information on snowpack instability is of crucial importance. Snowpack instability manifests itself by recent avalanche activity, whumpfs or shooting cracks. In the absence of these obvious signs of instability manual observations of snow stratigraphy combined with stability tests are presently the method of choice to seek instability data. The problem with these measurements is twofold: (1) Their availability is limited since the measurements are time consuming and in addition may expose the field crew to an undesirably elevated level of risk. Consequently, these snow stability data are available only with low resolution in space and time. In the future, snow micro-penetrometer measurements may speed up sampling — provided stability can be derived from the SMP signal (Bellaire et al., 2009; Pielmeier and Marshall, 2009) and simulated snow stratigraphy data may complement manual observations (e.g. Schirmer et al., 2010-this issue). (2) The validity of the measurements is limited, since test results need to be extrapolated to the surrounding terrain. This limitation follows from the fact that the mountain snowpack is inherently variable at various scales (e.g. Schweizer et al., 2008). The uncertainty resulting from the extrapolation is difficult to quantify, but depends on the specific snow conditions. In general, the snowpack variability is often not such that observations are useless; they may at least reveal relevant weak layer and slab properties, and whether they may interact in a critical manner; stratigraphy tends to be spatially more uniform than stability test results. On average, point stability observations performed by very experienced forecasters are fairly reliable. Schweizer and Jamieson (2010) estimated the error rate to about 5–10%.

Obviously, we are not simply interested in the snowpack stability at a specific location, but rather would like to know whether an avalanche may release (or may be triggered) on slopes similar to the sampling site. The slope where the measurements are taken is typically safe as otherwise the safety of the field crew would be compromised. Due to spatial stability variations, point stability (which we observe) may not be the same as slope stability (which we wish to know). Variations in point stability at the slope scale can affect release probability as is sometimes exemplified when an avalanche is triggered by a skier who was not the first to enter the slope. This prompted various attempts to relate avalanche release probability to variability at the slope scale — at least conceptually (e.g. Kronholm and Schweizer, 2003). However, reliably measuring slope scale variability is too time consuming as a large number of measurements is required — even if using, for example, a snow micro-penetrometer. If the measurements should be amenable to a geostatistical analysis and the analysis has to yield accurate estimates,

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

0165-232X/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.coldregions.2010.02.013

the number of measurements required seems to be on the order of 50–100 (Webster and Oliver, 2007, p. 125). Obviously, such slope scale measurements of weak layer and slab properties can only be collected on safe slopes, either not steep enough to slide or at times of generally fair (or even higher) stability. Measurements at the slope scale (e.g. Kronholm et al., 2004) have revealed that the correlation length was often on the order of ≤10 m.

As it is not practical for operational purposes to completely sample a slope, the question arises how a single point stability measurement should be supplemented so that slope stability (or avalanche release probability) can more reliably be extrapolated. Accordingly, it has been proposed (e.g. Birkeland and Chabot, 2006) to do a second observation at a representative site beyond the correlation length from the first test and choosing the least stable of the two test results. As the correlation length is unknown, at least about 10 m have been proposed as the distance between two independent tests (Jamieson and Johnston, 1993; Schweizer et al., 2008). Furthermore, performing two tests side by side at one location can decrease the uncertainty of test results and may indicate the small scale variability (0.1–1 m).

Independent estimates suggest that the critical size for a self-propagating fracture is on the order of 1–10 m (Schweizer et al., 2003). The lower range is probably more relevant for skier-triggered avalanches, i.e. the critical size may be on the order of the slab thickness (McClung and Schweizer, 2006). We assume that a fracture can propagate (and avalanche release is possible) if at one of the two test locations (~10 m apart) test results suggest that initiation and propagation is possible. On the other hand, significant variations at the scale of 0.1–1 m rather indicate conditions unfavorable for fracture propagation.

The aim of the present study is to assess whether performing two pairs of tests about 10 m apart improves our ability to predict snow slope stability, and if so, to suggest a procedure for sampling.

2. Data

Our study area is located in the Eastern Swiss Alps near Davos (9°47.5′ E, 46°48′ N). We collected compression test (CT) results (Jamieson, 1999) on 22 slopes located above tree line, i.e. most slopes were not sheltered. The mean elevation of the test slopes was about 2400 masl and the mean slope angle was 25°. About half of the slopes had northerly, the other half mainly south-westerly aspect. On each slope, four pairs of CTs were performed. Pit locations where two adjacent tests were done, were about 10-18 m apart (mean distance: 14 m) (Fig. 1). These data were collected during the winters 2006–2007, 2007–2008 and 2008–2009 at 22 days in the course of a spatial variability study (Bellaire and Schweizer, 2008). On each slope, we also observed snow stratigraphy so that failures in CTs could be associated with a specific layer boundary. Occasionally, the stratigraphy was too variable and the failure depth found with the CT could not be related to a corresponding depth in the manual snow profile; these CT results were still recorded and subsequently analyzed.

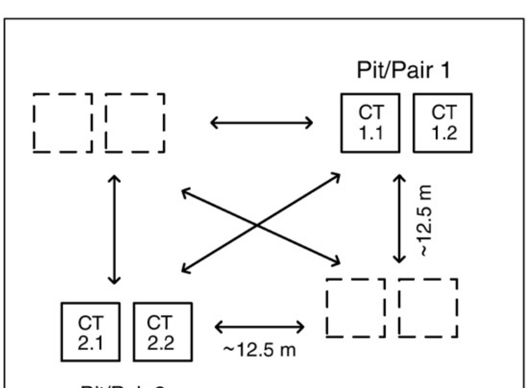

Fig. 1. Measurement set-up. At four locations on a slope, about 10–18 m apart, 2 compression tests (CT) were performed in one pit.

The primary grain type in most failure layers (73%) belonged to the group of persistent grain types (FC: faceted crystals, FCxr: rounding faceted crystals, DH: depth hoar and SH: surface hoar) (Fierz et al., 2009). Profiles were classified according to Schweizer and Wiesinger (2001). About two thirds of the profiles had well consolidated basal layers, whereas the others had weak basal layers (Table 1).

3. Methods

Compression tests were performed according to Jamieson (1999). The loading step at failure, failure depth and fracture character (van Herwijnen and Jamieson, 2007) were recorded. On 18 out of 22 slopes multiple failures occurred, i.e. the CT indicated several potential weak layers.

In the centre of the four pits with each a pair of CTs, a manual snow profile was observed complemented with a rutschblock (RB) test and another pair of CTs.

For assessing the variability within pairs (or pits) as well as between pairs (or pits) we considered (i) the CT score and the failure depth, and (ii) the fracture character and the failure depth.

Within a pair, we considered the two CTs as similar if (1) the difference in CT score was ≤2 for intermediate scores between 11 and 20 and the same weak layer failed (for lower (≤ 10) or higher scores (≥ 20), we considered a difference of ≤4 as sufficient for similarity), or (2) if the tests had similar fracture character and the same weak layer failed. Similar fracture character meant that the fracture character was (a) either sudden planar (SP) or sudden collapse (SC), or (b) one of the following three types: progressive collapse (PC), resistant planar (RP), or break (B) (van Herwijnen and Jamieson, 2007). The higher tolerance for similarity for scores ≤ 10 and scores ≥ 20 is based on the fact that the variations at these scores are often larger, i.e. more incidental and depending on the operator, but the interpretation is in fact unambiguous: either very weak or strong. Cases where the larger tolerance was applied were rare (5).

For comparing the CT score between pairs (or pits) we considered the mean score. Again, a difference of ≤2 was required for intermediate scores, and ≤4 for low or high scores. When checking whether the same failure layer existed in a pair 10-15 m apart, it was sufficient for similarity that the same weak layer failed in at least one of the two tests in the second pit. In other words, we simply wanted to know whether the weak layer also existed 10-15 m away from the first sampling location. Two pits could be similar, even if the CTs within a pit were different — and of course vice versa. For example, the first pair had scores of 13 and 16, and the second of 12 and 15. In that case, the pairs were each different, but the two pairs were similar. On the other hand, if the first pair had scores of 11 and 12, and the second of 15 and 16, the pairs were each similar, but the two pairs were different. For the fracture character we used the same criterion as described above for the within-pair comparison. In addition, if fracture characters were dissimilar each in both pits, the two pits were similar.

We rated each pit in terms of stability. If the mean CT score for the first sudden fracture was ≤13, we estimated the stability as 'poor', for a mean CT score ≥ 20 as 'good', and 'fair' in the other cases. This rating is based on a comparison of RB results with CT results primarily from the Columbia Mountains of western Canada (Schweizer and Jamieson, 2003); it is known to rather underestimate stability, i.e. yield a relatively high number of false–unstable results (e.g. Winkler and Schweizer, 2009).

In addition, at the end of the day stability was rated based on the profile and RB test results and on the presence or absence of any signs of instability in the area. Signs of instability were recent avalanching, whumpfs and shooting cracks (Jamieson et al., 2009). Provided the test slope was representative, this stability estimate should describe

Table 1
Characteristics of the 22 study slopes. Weak layer grain type is given according to the ICSSG (Fierz et al., 2009). Profile type refers to the classification of snow stratigraphy according to Schweizer and Wiesinger (2001). Stability is an estimate of local slope stability based on Rutschblock results, profile interpretation as well as existence of signs of instability. The CT score is the mean score of the four pairs on a slope for the primary weak layer. Number of pairs without variability indicates number of pairs (out of 4) with two similar test results; number of similar pairs on the slope shows the number of test pairs (or pits) (out of 6) that had similar test results — shown separately for the score (CT) and the fracture character (FC).

No	Date	WL grain type	Profile type	Stability	CT score	Number of pairs without variability out of 4 pairs (similar at pit-scale)		Number of similar pairs at the slope out of 6 pairs (similar at slope-scale)	
						CT	FC	CT	FC
1	2 Feb 2007	RG	3	Good	19	2	2	3	6
2	16 Feb 2007	FC	7	Fair	12	3	3	3	3
3	6 Mar 2007	DH	4	Good	13	3	1	6	2
4	8 Mar 2007	DF	4	Fair	11.5	3	3	5	5
5	15 Mar 2007	DH	3	Poor	14	3	4	3	6
6	10 Jan 2008	FC	7	Poor	11	2	4	1	6
7	17 Jan 2008	RG	4	Fair	11.5	3	2	6	6
8	23 Jan 2008	DH	6	Poor	12	3	4	4	6
9	31 Jan 2008	FCxr	6	Good	13.5	3	4	6	6
10	7 Feb 2008	DH	1	Fair	13	3	4	4	6
11	15 Feb 2008	FC	6	Good	17	2	4	3	3
12	19 Feb 2008	FCxr	7	Good	15	2	4	3	5
13	6 Mar 2008	DH	7	Good	16	1	2	2	2
14	18 Mar 2008	RG	6	Good	21	3	3	2	3
15	9 Jan 2009	FC	6	Good	19	3	2	2	3
16	14 Jan 2009	RG	6	Good	14	2	2	2	5
17	29 Jan 2009	DH	6	Fair	11.5	2	3	4	6
18	30 Jan 2009	RG	7	Fair	13	2	3	4	3
19	5 Feb 2009	SH	7	Good	21	4	4	6	6
20	19 Feb 2009	DH	7	Good	15	2	3	5	4
21	26 Feb 2009	FCxr	7	Good	19	3	4	4	6
22	17 Mar 2009	DH	3	Good	18	0	1	0	0
					Total	54	66	78	98

the slope-scale stability for similar slopes in the surroundings. The stability was rated as 'Poor' if either the profile (incl. the RB) was rated as 'very poor' or 'poor' (according to Schweizer and Wiesinger, 2001) and/or signs of instability were present that day, as 'Fair' if the profile (incl. the RB) was rated as 'fair' and signs of instability were observed, and 'Good' in the other cases.

When comparing variables from two groups, the non-parametric Mann–Whitney *U*-Test was used. To check for differences in the relative frequency of two categories, the two proportion *Z*-test was applied (Spiegel and Stephens, 1999). A level of significance $p = 0.05$ was chosen to decide whether the observed differences were statistically significant.

4. Results

On each of the 22 slopes, we evaluated the within-pair variability (pit-scale) for the four locations, and the between-pairs variability (slope-scale) for the six combinations existing between the four pairs (Table 1).

4.1. Pit-scale variability

Considering the CT score and the fracture depth, within-pair similarity was found in 54 cases (61%) (Table 1). In the remainder of cases ($N = 34$, 39%), the reason for within-pair dissimilarity was as follows: In about half of these cases the difference in score was too large, in about one third of the cases the layers that fractured were different, and in the remaining cases the score and the fracture layers were dissimilar. The median score was higher (16) in pits with different scores than for pits with similar scores (13). However, the difference was statistically not significant (*U*-test, $p = 0.22$). Remarkably, within-pair variability existed on all except one slope. On 19 out of 22 slopes, either three or two pits were similar (Table 1) out of four possible pits.

The fracture character was less variable and agreed in 66 out of the 88 pairs (75%) (Table 1). Sudden failures were most often observed. The reasons for dissimilarity were different failure layers (32%), different fracture characters but the same failure layers (25%), and different failure layers as well as fracture characters (43%). Again the median score was slightly higher for pits with different fracture characters than for pits with similar characters (not significant, $p = 0.21$). Within-pair variability on a given slope was much less frequent when fracture character was considered rather than CT score. On 15 out 22 slopes, in either four or three pits (out of four) a similar fracture character was observed (Table 1).

When the CT score (and/or the failure layer) differed within a pit, in more than half of those cases (53%) the fracture character was also different. Very rarely (<5%) was a pair rated as similar based on the CT score, but different based on the fracture character.

Table 2
Stability rating of test pairs (based on CT score) indicating the difference between ratings on a given slope.

Stability			
2nd pair of tests	1st pair of tests		
	Poor ($N = 37$)	Fair ($N = 37$)	Good ($N = 14$)
Poor	68%	32%	0%
Fair	32%	54%	38%
Good	0%	14%	62%

4.2. Slope-scale variability

At the slope scale, the test results in a pit (10–15 m apart) were judged as similar in 78 out of 132 cases (59%) (Table 1), if the CT score (and the failure layer) was considered. For the CT fracture character, a significantly ($p = 0.01$) higher agreement was found (74%). Considering individual slopes (Table 1), there was one case where none of the pairs were similar. On the other hand, all four pits were similar on

Table 3
Variability situations, their frequency separately for CT score (CT) and CT fracture character (FC) and a possible interpretation (see text for explanations).

Situation	Frequency		Variability	Initiation	Propagation	Stability interpretation (depending on CT score) (sudden fracture)
	CT	FC				
ssS	35	65	No (or few)	Possible	Likely	'poor', 'fair' or 'good'
ssD	12	9	Slope scale	Possible	Likely	'poor', 'fair' or 'good'
ddS	7	5	Small scale (consistent)	Rather unlikely	Rather unlikely	'fair' or 'good'
ddD	10	6	Small scale and slope scale	Rather unlikely	Rather unlikely	'fair' or 'good'
sdS	36	28	Some small scale	Possible	Likely	'poor', 'fair' or 'good'
sdD	32	19	Some small scale and slope scale	Possible	Possible	'poor', 'fair' or 'good'

10 slopes if the fracture character was considered, but only on 4 slopes if the CT score was considered as criterion for similarity.

Each pit was rated in terms of stability; results are compiled in Table 2. From the 88 pairs of tests, 42% were rated as 'poor', 42% as 'fair', and the remaining 16% were rated as 'good'. In terms of stability variation across the slope, it was found that the stability rating between pits agreed in 61% of the cases. In the other 39% of the cases, one of the pits was either rated as 'poor' and the other as 'fair', or alternatively as 'fair' and 'good'. No combination with a pit rated as 'poor' and the other on the same slope as 'good' occurred (which would have indicated large slope-scale variation). However, only on 6 out of the 22 slopes all four stability ratings agreed (3 'poor', 2 'fair', and 1 'good'). On most slopes (59%), one pair of tests in a pit was rated either lower or higher than the other three pairs in the neighboring pits. If we randomly choose one of the four pits as the first (and repeat this procedure for all four locations), and wonder what stability a second pair of tests in a pit 10–15 m apart would have revealed, Table 2 shows that with an initial rating of 'poor', there is a good chance (68%) that in the second pit the pair of tests will also indicate rather unstable conditions. This consistency is less pronounced if the initial rating was 'fair' or 'good'. In those cases, there was a considerable chance that the second pair of tests indicated a stability lower than found in the first pit. In other words, the sampling at the location of the first pit overestimated stability — potentially resulting in false–stable prediction.

4.3. Pit and slope-scale variability

If the pit-scale and slope-scale variability are jointly considered, i.e. the differences within as well as between pairs, six different situations can be identified (Table 3): (1) no within- and no between-pairs differences: ssS; (2) no within-pair differences, but the pairs are different: ssD; (3) differences within pairs, but no difference between the pairs: ddS; (4) differences within and between pairs: ddD; (5) one of the pairs shows within-pair variability, but the pairs are similar: sdS; (6) one of the pairs shows within-pair variability and the pairs are different: sdD.

The first situation indicating very similar test results at the pit — as well as the slope-scale (ssS), was found in almost half of the cases (48%) if the fracture character was considered, but only in 27% of the cases for the CT score. In about two thirds of these cases at least one of the pairs was rated as 'poor'. On the other hand, very different results, different at the pit-scale as well as at the slope-scale (ddD), occurred in only 10 cases for the CT score and in 6 cases for the fracture character. In about 90% of these cases (ddD) at least one pair was rated as 'fair'. In other words, overall, in $<10\%$ of all cases variability existed at the pit — as well as the slope-scale and in these cases test results indicated rather stable conditions. In many cases, however, 'some' variability (ssD, ddS, sdS, and sdD) was found.

If we compare the joint pit- and slope-scale variability to the slope-scale stability estimate as given in Table 1, conditions with different CT results within pairs (ddD and ddS) were found more often on slopes with a stability estimate of 'Good' ($p = 0.05$). No other trends were found if considering the CT score. However, if considering the fracture character, the situation with neither within- nor between-pairs differences (ssS) was more frequently found on 'Poor' slopes ($p = 0.02$). In fact, on the three slopes with a stability estimate of 'Poor', only this situation with little variability (ssS) occurred. Also, on 'Good' slopes clearly more within-pit variation (ddS or ddD) was found ($p = 0.03$). In other words, there was a trend towards increasing variability with increasing slope stability.

4.4. Slope-scale stability

We compared our stability rating based on the CT results to the stability estimate given in Table 1 that is based in part on an expert assessment of the stability for similar slopes in the surroundings of the sampling site (Table 4). Pits with CT test results rated as 'poor' (pit-scale) were mostly found on slopes that were judged as 'Fair'. Similarly, pits with test results rated as 'fair' were mostly found on slopes for which stability was estimated to be 'Good'; all pits with tests rated as 'good' were found on slopes with a stability estimate of 'Good'.

On the other hand, considering whether the tests were able to recognize 'poor' conditions, the pits on the three slopes estimated as 'Poor' according to Table 1, were all rated as either 'poor' (67%) or 'fair' (33%). On slopes with a stability estimate of 'Fair' or 'Good', most pits were rated as 'poor' or 'fair', respectively. On 'Poor' and 'Fair' slopes at least two out of four pits on a slope were rated as 'poor'. On slopes with a stability estimate of 'Good', the variability was relatively large: There were slopes with three out of four pits rated as 'poor' and at the other end of the spectrum slopes with all four pits rated as 'good'.

If only the agreement between the pit-scale stability ratings on a specific slope was considered, there was no difference between slopes estimated as either 'Poor', 'Fair' or 'Good': on most slopes at least three pits had the same stability rating. In other words — if the pit stability was considered — the amount of slope-scale variation was not related to slope stability.

If only two classes were considered ('poor' vs. fair-to-good') the agreement (unweighted average accuracy) between stability derived from the tests in one pit and the slope-scale estimate was 64%.

Table 4
Pit-scale stability based on CT results (Table 3) vs. stability estimate (Table 1).

Pit-scale stability	Stability estimate (slope-scale)		
	Poor	Fair	Good
poor	8	20	9
fair	4	4	29
good	0	0	14

5. Discussion

We investigated a relatively small area (~150 m^2) on a given slope. The reason for restricting ourselves to a small area is — apart from the advantage that sampling on the same slope is possible several times during the winter — based on the fact that most spatial variability studies that used geostatistics for analysis found correlation lengths ≤10 m (e.g. Kronholm, 2004). We can therefore assume that beyond a distance of about 10 m measurements will be un-correlated. In other words, sampling 10 m, 50 m or 100 m away from the first pit will probably not much change the results as measurements are in any case un-correlated — provided we are still sampling on the same slope.

For our analysis of spatial variability at the pit- and slope-scale, we relied on the compression test. The compression test is well established, quickly done, but also relatively prone to errors. In particular, the support is rather small and the CT tends to underestimate stability (Winkler and Schweizer, 2009). On the other hand, the chance of false-stable predictions tends to be small. Nevertheless, as we focus on relative changes of stability on a small test slope, the compression test is well suited for our analysis. The data collected in this study represent a wide variety of snowpack conditions — representative of a transitional snow climate (McClung and Schaerer, 2006).

We were not able to conduct sampling during conditions of 'very poor' stability. In only 7 out of 88 pits a CT score ≤10 was observed. On 9 out of 22 days we at least observed signs of instability on the slope we sampled or on adjacent slopes. Nevertheless, our dataset is somewhat biased toward higher levels of stability. However, this is not necessarily a flaw, since estimating slope-scale stability during periods of 'very poor' stability is not as challenging for skilled observers as estimating variations between 'poor' and 'good' stability. So, biasing this study towards the spectrum between 'poor' and 'good' stability has focused on the 'tricky' range in stability evaluation.

The amount of variability we found at the pit- and slope-scale is comparable to previous studies. For example, Landry et al. (2004) reported that 25–39% of pits dug on relatively "uniform" slopes were found to not be statistically representative of that slope. The agreement found between tests in a single pit and between two pits on the same slope (about 60–75%) reflects — among other things — the reliability of stability tests. Interestingly, the agreement between pits was not lower than the agreement within pits. Our agreement value is relatively low since we not simply considered two classes (stable vs. unstable).

At the pit- as well as at the slope-scale similar fracture character was more frequently found than similar CT score. This coincides with the common hypothesis that snowpack properties (or test results) related to fracture propagation should be less variable than properties related to fracture initiation (Schweizer et al., 2008). In fact, this was shown e.g. for the rutschblock test by Campbell and Jamieson (2007) and for the compression test by van Herwijnen et al. (2009).

If we perform two tests in one pit and then another pair of tests in a second pit on the same slope, obviously various degrees of variation can occur — six situations were described above. It is not fully clear how this information should be interpreted, in particular if test results differ at the small (or pit-) scale (~1 m) and/or the large (or slope-) scale (~10 m). In Table 3 an attempt is shown to derive slope stability based on two pits with each a pair of tests. We assume that small scale variability will hinder fracture initiation and in particular fracture propagation. Further assumptions are that (i) the stability rating of the pit with the lower score is decisive, and (ii) that small (i.e. pit-) scale variability favors stability. This would mean that in situations with pit-scale variability (ddS or ddD) the slope stability could be rated one level higher than the lower pit rating. This assumption is supported by our test results since pit-scale variability was more frequently found with 'Good' stability. In the other four situations the slope would simply be rated based on the pit with the lower scores (for the first sudden fracture). In other words, we assume that a fracture once initiated in one area of the slope may still spread across regardless whether there are more stable areas 10–15 m away. Of course, if no differences between pits are found propagation is even more likely — provided the slab is such that propagation is favored at all.

Obviously, performing two pairs of tests provides additional information. However, we suggest that it is not always necessary to dig a second pit on a slope — in particular not if we have already found an instability — as seeking instability is the purpose of sampling (McClung, 2002). Therefore, the following procedure is proposed. If at a sampling location on a slope two adjacent stability tests show similarly low scores and sudden fractures, no further sampling is required. However, if either the two scores are similar and indicate rather 'fair' or 'good' stability, or the two scores are different, a second pair of tests on the same slope about 10–15 m beyond the first sampling location can be useful. If at the second location similarly low scores are found, stability is expected to be rather 'poor'. If as well intermediate scores, dissimilar scores or a different weak layer are encountered the stability is at least 'fair', and in the case of consistent small scale variability maybe even 'good'. Based on our results (Table 2) digging a second pit would only be necessary in about 58% of the cases. In about one third of these cases the second pit would have indicated a lower stability than the first one.

6. Conclusions

We explored the value of a simple sampling scheme that attempts to capture the small (~1 m) as well as the large (slope-) scale variability (~10–15 m). We analyzed the variations between two compression tests in a single pit and between two pits each with a pair of tests. Hence this is the first study — to our knowledge — that explicitly analyzes the value of a second pit on the same slope.

The present study has been performed on non-sheltered slopes above tree line. The area investigated on a given slope was relatively small (~150 m^2). The approach — suggested primarily in an operational setting — ignores the contribution of the public bulletin or quick field observations (Jamieson et al., 2009) to assessing slope stability.

At the pit- as well as at the slope-scale about 60% of the tests revealed similar results if the CT score and the failure depth, and about 75% if the fracture character and the failure depth, were considered. Hence, the fracture character was less variable than the CT score.

Rating the stability of pits based on the first sudden fracture revealed that in 61% of the cases the stability rating between two pits on a single slope agreed. No combination with a pair rated as 'poor' and the other on the same slope as 'good' occurred (which would have indicated large slope-scale variation).

Comparing the compression test results to a subjective local stability estimate confirmed that the CT underestimates stability — as has been pointed out by Winkler and Schweizer (2009). Variations at the pit-scale were more frequently found on slopes rated as 'Good', whereas as on slopes rated as 'Poor' similar conditions dominated. However, our dataset only contained three slopes with a stability estimate of 'Poor'. Nevertheless, our data indicate a trend to more pit-scale variations on slopes with rather stable conditions. Therefore, we suggest that with largely varying score, fracture character and/or failure depth at the pit-scale the stability can probably be rated higher — although it is presently not fully clear how the various degrees of variations (Table 3) that can be found when digging two pits and doing two tests in each pit should be interpreted. Table 3 should be considered as a preliminary proposal only.

Whereas two pits on a single slope obviously provide more information, we suggest that a second pit is not always necessary. If at a sampling location on a slope two stability tests show similarly low scores and sudden fractures, no further sampling is required. However, if either the two scores are similar and indicate rather 'fair' or 'good' stability, or the two scores are different, a second pair of tests on the same slope about 10 m away from the first sampling location can be useful. If at the second location low scores are found, stability is expected to be rather 'poor'. If as well intermediate scores, dissimilar scores or a different weak layer are encountered the stability is at least 'fair', and in the case of consistent small scale variability maybe even 'good'. For our dataset, proceeding to a second sampling location about 10 m apart would have been necessary in about 58% of the cases. In about two thirds of these cases, results in the second pit would have confirmed the findings at the first location. In other words, in about 20% of all cases test results in the first pit overestimated stability — which is obviously not desired. Accordingly, if no instability was found at the first sampling location, a second pair of measurements can clearly reduce the number of false-stable predictions. Due to the high sensitivity of the CT, false-stable predictions are in general relatively rare. This advantage does not exist if other tests are used. In that case, the proposed procedure needs to be re-visited.

Acknowledgements

We would like to thank Alec van Herwijnen, Christoph Mitterer, Sina Schneider, Michael Schirmer and Charles Fierz for their help with the field work. The manuscript was much improved as a result of reviews by two anonymous referees who provided very helpful and extremely constructive comments. We gratefully acknowledge financial support by the European Commission (FP6-STRep-NEST, project no. 043386: TRIGS).

References

Bellaire, S., Schweizer, J., 2008. Deriving spatial stability variations from penetration resistance measurements. In: Campbell, C., Conger, S., Haegeli, P. (Eds.), Proceedings ISSW 2008, International Snow Science Workshop, Whistler, Canada, 21–27 September, pp. 188–194.
Bellaire, S., Pielmeier, C., Schneebeli, M., Schweizer, J., 2009. Stability algorithm for snow micro-penetrometer measurements. J. Glaciol. 55 (193), 805–813.
Birkeland, K.W., Chabot, D., 2006. Minimizing "false stable" stability test results: why digging more snowpits is a good idea. In: Gleason, J.A. (Ed.), Proceedings ISSW 2006. International Snow Science Workshop, Telluride CO, U.S.A., 1–6 October, pp. 498–504.
Campbell, C., Jamieson, J.B., 2007. Spatial variability of slab stability and fracture characteristics within avalanche start zones. Cold Reg. Sci. Technol. 47 (1–2), 134–147.
Fierz, C., Armstrong, R.L., Durand, Y., Etchevers, P., Greene, E., McClung, D.M., Nishimura, K., Satyawali, P.K., Sokratov, S.A., 2009. The international classification for seasonal snow on the ground. HP-VII Technical Documents in Hydrology, 83. UNESCO-IHP, Paris, France. 90 pp.
Jamieson, J.B., 1999. The compression test — after 25 years. The Avalanche Review 18 (1), 10–12.
Jamieson, J.B., Johnston, C.D., 1993. Rutschblock precision, technique variations and limitations. J. Glaciol. 39 (133), 666–674.
Jamieson, B., Haegeli, P., Schweizer, J., 2009. Field observations for estimating the local avalanche danger in the Columbia Mountains of Canada. Cold Reg. Sci. Technol. 58 (1–2), 84–91.
Kronholm, K., Schweizer, J., 2003. Snow stability variation on small slopes. Cold Reg. Sci. Technol. 37 (3), 453–465.
Kronholm, K., Schneebeli, M., Schweizer, J., 2004. Spatial variability of micropenetration resistance in snow layers on a small slope. Ann. Glaciol. 38, 202–208.
Landry, C., Birkeland, K., Hansen, K., Borkowski, J., Brown, R., Aspinall, R., 2004. Variations in snow strength and stability on uniform slopes. Cold Reg. Sci. Technol. 39 (2–3), 205–218.
Pielmeier, C., Marshall, H.-P., 2009. Rutschblock-scale snowpack stability derived from multiple quality-controlled SnowMicroPen measurements. Cold Reg. Sci. Technol. 59 (2–3), 178–184.
McClung, D.M., 2002. The elements of applied forecasting — part I: the human issues. Nat. Hazards 26 (2), 111–129.
McClung, D.M., Schaerer, P., 2006. The Avalanche Handbook. The Mountaineers Books, Seattle WA, U.S.A. 342 pp.
McClung, D.M., Schweizer, J., 2006. Fracture toughness of dry snow slab avalanches from field measurements. J. Geophys. Res. 111 (F4), F04008. doi:10.1029/2005JF000403.
Schweizer, J., Jamieson, J.B., 2003. Snowpack properties for snow profile analysis. Cold Reg. Sci. Technol. 37 (3), 233–241.
Schweizer, J., Wiesinger, T., 2001. Snow profile interpretation for stability evaluation. Cold Reg. Sci. Technol. 33 (2–3), 179–188.
Schirmer, M., Schweizer, J., Lehning, M., 2010. Statistical evaluation of local to regional snowpack stability using simulated snow-cover data. Cold Reg. Sci. Technol. 64 (2), 110–118 (this issue).
Schweizer, J., Jamieson, J.B., Schneebeli, M., 2003. Snow avalanche formation. Rev. Geophys. 41 (4), 1016.
Schweizer, J., Kronholm, K., Jamieson, J.B., Birkeland, K.W., 2008. Review of spatial variability of snowpack properties and its importance for avalanche formation. Cold Reg. Sci. Technol. 51 (2–3), 253–272.
Schweizer, J., Jamieson, J.B., 2010. Snowpack tests for assessing snow-slope instability. Ann. Glaciol. 41 (54), 187–194.
Spiegel, M.R., Stephens, L.J., 1999. Schaum's Outline of Theory and Problems of Statistics. Schaum's outline series, McGraw-Hill, New York. 538pp.
van Herwijnen, A., Jamieson, J.B., 2007. Fracture character in compression tests. Cold Reg. Sci. Technol. 47 (1–2), 60–68.
van Herwijnen, A., Bellaire, S., Schweizer, J., 2009. Comparison of micro-structural snowpack parameters derived from penetration resistance measurements with fracture character observations from compression tests. Cold Reg. Sci. Technol. 59 (2–3), 193–201.
Webster, R., Oliver, M.A., 2007. Geostatistics for environmental scientists. Statistics in Practice, Wiley, Chichester, West Sussex, U.K.. 271pp.
Winkler, K., Schweizer, J., 2009. Comparison of snow stability tests: extended column test, rutschblock test and compression test. Cold Reg. Sci. Technol. 59 (2–3), 217–226.