EVALUATING NEAREST NEIGHBOURS IN AVALANCHE FORECASTING - A QUALITATIVE APPROACH TO ASSESSING INFORMATION CONTENT

Ross S. Purves
Department of Geography
University of Zurich,
Winterthurerstrasse 190
Zurich
Switzerland
e-mail: rsp@geo.unizh.ch

Joachim Heierli
Swiss Federal Institute for Snow and Avalanche Research (SLF)
Flueelastrasse 11
Davos Dorf
Switzerland
e-mail: heierli@slf.ch

Abstract

Nearest neighbours (NN) approaches are a statistically-based pattern classification technique used extensively for computer-assisted decision making in tasks such as ski area and winter road management and backcountry avalanche forecasting. The essential hypothesis behind NN assumes that the available data (normally current meteorological and snowpack data) usefully select similar avalanche conditions from the past which are thus of interest to the forecaster. Most evaluations of NN rely on verification schemes based on avalanche events and delivering summary statistics of performance in the form of contingency tables. We argue that such approaches, whilst useful, oversimplify NN's information output, and present a complementary approach to verification.

Over two winters qualitative information was reported by forecasters, describing meteorological conditions, snowpack conditions, snow stability and related information. These information were entered in a web log, where forecasters could quickly and easily enter their assessment of the current situation.

Using these data we have reevaluated the quality of information delivered by NN, by comparing NN forecasts with randomly generated forecasts. Our results suggest that, although the performance of NN may appear good when measured by summary statistics, the usefulness of the information presented to the forecaster may often be low with the events selected by NN as 'examples' not corresponding to the general avalanche situation - e.g. NN selects single, skier triggered avalanches though on the forecast day large natural avalanches occurred.

These findings demonstrate that, firstly, our previous evaluation approaches insufficiently described the information content that we wished to evaluate and secondly, suggest weaknesses in the NN approach. While NN demonstrates considerable skill in selecting similar weather patterns, it has considerably less skill in selecting similar snowpack and stability patterns.

Finally, forecasters reported that filling in web logs was itself a useful process and led to more self reflection in their decision making process.

Keywords: Avalanche forecasting; validation; nearest-neighbours.

1. INTRODUCTION AND MOTIVATION

Nearest neighbours (NN) pattern classification is a popular tool in avalanche forecasting, with a long history of use in many countries for a range of forecasting purposes including forecasts used by ski patrol services to protect ski areas, by authorities responsible for safety of villages and roads and forecasting services providing backcountry forecasts (e.g. Bolognesi (1994); Buser (1983; 1989); Gassner et al. (2000); Mérindol et al. (2002); McCollister et al. (2002); Purves et al. (2003); Zeidler and Jamieson (2004)). The underlying principle of NN assumes that by choosing parameters which describe conditions on given days we can select similar days in the past. Typically, NN databases contain data describing weather and snowpack conditions for every day and the avalanche events associated with each of those days. In Canada, Zeidler and Jamieson (2004) showed that for a snowpack where persistent weak layers play an important role in avalanche events the addition of more information parameterising stability through

a daily skier instability index improved NN performance.

A key, and often neglected, element of forecasting involves model verification and validation. Verification and validation are defined by Rykiel (1996) as follows:

"Verification is a demonstration that the modelling formalism is correct… Validation is a demonstration that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model."

In practice the terms verification and validation are often used synonymously, but it is important to note that any forecasting tool must first be verified – in other words we must ensure that we implemented our tool *correctly* before it can be validated. However, it is only by validating a tool that its use with respect to its intended application can be usefully assessed.

In previous work (Heierli et al., 2004) we argued that NN output could be interpreted in a range of ways, namely:

- categorical forecasts;
- probability forecasts; and
- descriptive forecasts.

Categorical forecasts are perhaps the most typical interpretation of an NN tool. Here, the most similar days to a forecast day are selected, and if more than some threshold number of similar days had avalanche events, the forecast day itself is considered to be an avalanche day. Categorical forecasts can be validated through the production of a contingency table and associated forecast accuracy and skill measures (Doswell et al., 1990).

In a probability forecast the number of neighbours with events is assumed to be proportional to the probability of an event. This approach may be validated by distributions-oriented verification (Murphy and Winkler, 1986) producing measures such as reliability and resolution.

A descriptive forecast is one where a detailed list of events and associated observations are provided to the forecaster as an *aide memoire* and used as part of the decision making process. Many authors (e.g. Purves et al., 2003) have argued that such an interpretation of NN is the most useful to forecasters as it provides a way of integrating the output of NN into the overall decision making process, rather than simply providing a binary or probabilistic forecast for events.

However, validating descriptive interpretations of NN requires that a qualitative assessment of the information content of the neighbours and associated events be made. In a previous paper we made a first attempt at such a validation, asking an experienced forecaster to critically rate the usefulness of NN forecasts at the end of a winter (Heierli et al., 2004). This initial experiment suggested that, at least some of the time, NN provided useful information. Importantly, we define useful as the provision of both *correct* and *additional information* to the forecaster which is therefore a useful aid in the overall decision making process. In order to deliver correct information to the forecaster, we must first validate NN either categorically or probabilistically – a successful validation using these measures is a precondition to the provision of useful information.

In a descriptive interpretation of the results of NN the forecaster must first compare the similarity of the days returned and then interpret the conditions and events on the similar days with respect to likely conditions on the forecast day. Thus, if we wish to validate NN with respect to descriptive forecasting, we can pose two separate but related questions:

- How similar are the neighbours returned in terms of weather, snowpack and stability?
- Does the number of neighbours returned with avalanche events combine to suggest greater instability or avalanche hazard (as opposed to a greater probability of a single event)?

In the rest of this paper we describe an approach to qualitatively assessing the usefulness of NN by collecting additional data describing weather, snowpack, instability and mitigation measures from avalanche forecasters over a winter. These data are then used in starting to explore the first question posed above, before discussing the implications of the results for the use of NN in particular and forecast validation in general.

2. STUDY AREA

The data described in this paper were collected at the Parsenn ski area, in Davos Switzerland. The ski patrollers at Parsenn are responsible for avalanche control in an area of some 25km^2, within which a total of some 100 kms of ski pistes are available for guests. Numerous avalanche paths of all aspects and sizes are found within the ski area and are controlled through use of explosives and other mitigation measures.

Within the Parsenn ski area, the Swiss Federal Institute for Snow and Avalanches have maintained a measurement site for many years, meaning that this ski area has one of the longest and highest quality data sets of meteorological and snowpack measurements available. These data, together with information on natural, accidental and avalanches triggered by mitigation measures are all used in a version of NN configured for the Parsenn.

Figure 1 shows accuracy measures for categorical forecasts at the Parsenn over the last 37 years with a range of decision thresholds for 10 nearest neighbours. The experiments described in this paper were carried out with only two years data, and accuracy measures for these two years are shown in Figure 2.

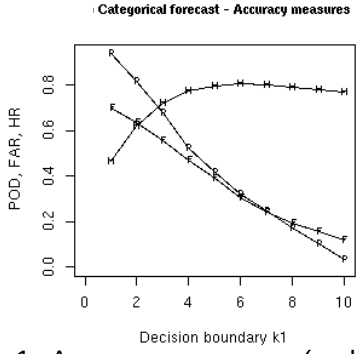


Figure 1: Accuracy measures (probability of detection, false alarm rate and hit rate) for the Parsenn over 37 years – decision boundary k1 is the threshold number of neighbours with an event considered indicative of an event on the target day

An important prerequisite to the qualitative evaluation described in the following sections was that NN for this area be validated first with respect to its categorical performance. Figures 1 and 2 show that NN for Parsenn has very similar properties in terms of probability of detection, false alarm ratio and hit rate for both datasets, suggesting that this prerequisite has been met.

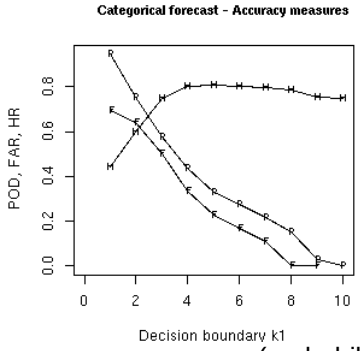


Figure 2: Accuracy measures (probability of detection, false alarm rate and hit rate) for the Parsenn over two years of data used in this study – decision boundary k1 is the threshold number of neighbours with an event considered indicative of an event on the target day

3. METHODOLOGY

3.1 Data collection

As discussed above, the aim of this work was to validate an NN tool with respect to the descriptive information provided. Since the descriptive information itself is integrated by the forecaster as a part of their overall decision making process, we decided to ask forecasters to write textual summaries of observations and actions relevant to avalanche activity and mitigation. To do this, we set up a *blog*[1] in which forecasters could post a description under a number of relevant categories, namely:

1. weather situation and recent changes;
2. snowpack development;
3. snowpack instabilities and avalanche events;
4. decisions taken to improve safety on ski pistes; and
5. problems in assessing or deal with avalanche hazard, for example through a lack of personnel or bad weather.

[1] Blogs or weblogs are websites where authors post information in the form of a diary, often subdividing the information by category.

To reduce the load on the forecasters writing the posts, it was agreed from the outset that posts would only be written on days of interest in terms of avalanche activity – in other words on days where little or no change in occurred to a stable snowpack the forecasters did not post information. This means that the data collected are conditioned by the forecasters, and that we only have a record of days considered *a priori* to be of interest by forecasters. Since the forecasters generally posted information to the blog after being out in the field, this assumption does not mean that unexpected events are not recorded.

In order to speed writing of posts a set of commonly used keywords were provided and could be inserted by the users through a single mouse click. The blog was designed in order that information could only be posted about a day on the day itself, thus preventing *post hoc* reassessment of conditions.

We used a modified version of Wordpress (www.wordpress.org), a popular blogging tool, to allow the forecasters to post descriptions. The descriptions themselves were stored in a MySql database which allowed easy access to data for the processing tasks described in the next section.

3.2 Measuring post similarity

The first question posed in the introduction asked *"How similar are the neighbours returned in terms of weather, snowpack and stability?*" To answer this question we wished to compare, for all the target days X for which we had a post, the posts which existed for the nearest neighbours $Y_{k,\ k=1...10}$.

The version of NN used every day at Parsenn has a total of 37 years worth of data. As datasets increase in size, the chances of neighbours being selected from days with posts decrease and so we configured a version of NN using only two winters, where a winter considered to be similar in character to the winter with posts was added to the dataset. As shown in Figures 1 and 2 NN performed reasonably with this reduced dataset.

Having created a dataset, we then performed a forecast for every target day X for which at least one post existed. For each of these target days X we then further selected only those days on which one or more of the $Y_{k,\ k=1...10}$ had at least one post.

We then compared the posts on each target day X for the three categories, weather, snowpack and stability with those for the neighbours $Y_{k,\ k=1...10}$ using the following ordinal rating scale:

- Rating 0: posts do not match and contain either contrary or disjoint information. At forecast time the neighbouring post is an unhelpful or even misleading example with no or negative value.
- Rating 1: the posts match partially, some features described, some not. At forecast time, the neighbouring post is of limited help or value.
- Rating 2: the two posts match well and describe broadly the same situation. At forecast time, the neighbour's post is a helpful, valuable example.

Table 2 shows a typical set of posts for a target day (16.02.2005) and three of its ten neighbours which had posts, together with example similarity ratings. The posts for the neighbours on both 24.01.2005 and 21.02.2005 match well (2), since in both cases the main salient features of the description (north winds and cold) are similar to the target day, and the light snow reported is considered unlikely to have been an important component in the description of the conditions. For the neighbour on 03.02.2005, because snow is reported but *not* categorised as light we assume that the new snow may have been an important feature of this day and thus, whilst some features are similar (e.g. northerly winds and cold) it is only a partial match (1).

Post date	Category	Post content	Rating
16.02.2005	Weather	*North winds, over night no precipitation, cold, -18°C, changeable, thin cloud*	N/A
16.02.2005	Snowpack	*New snow drifting and forming slab on nearly all aspects*	N/A
16.02.2005	Stability	*Stable under 2000m, unstable over 2000m but on 15.2 little success with test releases*	N/A
24.01.2005	Weather	*Today, north winds, cold, -16°C, light snowfall, very cloudy*	2
03.02.2005	Weather	*Northwest winds, cold, snowfall, cloud above 2200-3000m*	1
21.02.2005	Weather	*Yesterday and today north winds, cold, light snow fall*	2
24.01.2005	Snowpack	*In places with lots of snow the snowpack is already somewhat stablised and is becoming more stable.*	0
03.02.2005	Snowpack	*Drifted snow collecting, near ridges of northwest to northeast aspects slopes are scoured*	1
21.02.2005	Snowpack	*Last week's new snow is very well bonded with the old snowpack*	0
24.01.2005	Stability	*- When moving from deep to shallow snow the snowpack is still very unstable. Yesterday a piste machine released an avalanche remotely. North facing slopes still very unstable.*	0
03.02.2005	Stability	*Stable until 2200. Above 2200m unstable, north slopes in gullies and hollows scoured*	1
21.02.2005	Stability	*Unstable on slopes where the drifted snow has not been disturbed and when moving between deep and shallow snow. Generally stable.*	1

Table 2: Example posts and their ratings (translated from the German)

In order to assess whether NN finds useful matches we must not only compare the similarity of posts, but also show that NN has more skill in selecting neighbours than would be the case for a random selected set of neighbours. Furthermore, we must also control for the subjectivity of comparisons made by the annotators and biases introduced by the annotation process.

To measure the importance of these biases and the influence of subjectivity two annotators blindly rated the similarity of the same set of posts for both random and NN selected neighbours independently.

4. RESULTS

Each comparison of a neighbour with the target day was made by two annotators. Figure 3 shows inter-annotator agreement for all judgements made by annotators A and B.

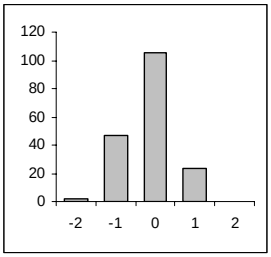
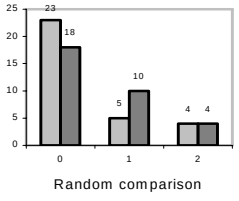
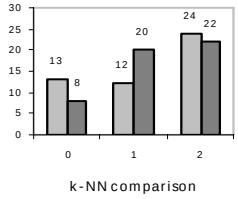


Figure 3: Count of differences in annotator judgements – value of 0 indicates complete agreement while negative values indicate that annotator B considered agreement better than annotator A.

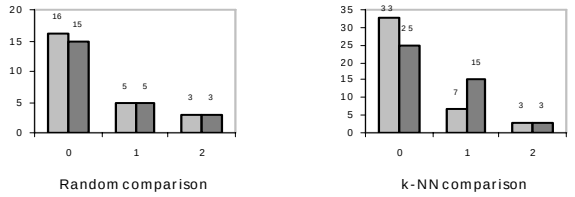
Figures 4-6 show the results of comparisons of neighbours generated randomly and by NXD for weather, snowpack and stability descriptions. The figures show frequency distributions for both annotators which were categorised as described in Section 3.2.





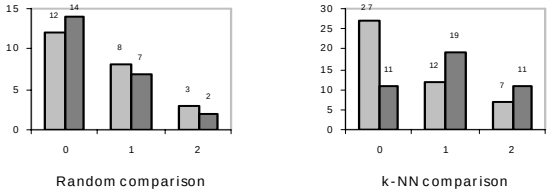
(a) (b)

Figure 4: Frequency distributions (for two independent raters) of comparisons of random (a) and k-NN (b) selections of nearest neighbours for weather similarity based on ordinal rating scale described in Section 3.2



(a) (b)

Figure 5: Frequency distributions (for two independent raters) of comparisons of random (a) and k-NN (b) selections of nearest neighbours for snowpack similarity based on ordinal rating scale described in Section 3.2



(a) (b)

Figure 6: Frequency distributions (for two independent raters) of comparisons of random (a) and k-NN (b) selections of nearest neighbours for stability similarity based on ordinal rating scale described in Section 3.2

Table 3 shows a summary of these results, where good and partial matches are merged into a single category.

	Random selection				k-NN selection			
Match quality	**Poor**		**Good or partial**		**Poor**		**Good or partial**	
Annotator	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Weather	72%	56%	28%	44%	27%	16%	73%	84%
Snowpack	67%	65%	33%	35%	77%	58%	23%	42%
Stability	52%	61%	48%	39%	59%	27%	41%	73%

Table 3: Summary of judgements for weather, snowpack and stability

5. DISCUSSION

A key task in this work concerned measuring the qualitative similarity of text snippets. Such comparisons are by their nature subjective, and an important first step in assessing the validity of the method is consideration of the reliability and consistency of the judgements made. Such tasks are common in other research fields such as information retrieval, where assessing the relevance of documents to a query is a common, subjective task (Voorhees, 2000). Figure 3 shows the inter-annotator agreement for our study. For 59% of the judgements, both annotators agreed exactly on a three point scale. If judgements were random, we would expect inter-annotator agreement of around 33%, which suggests that our annotation is of reasonable quality.

Figure 4 shows the distribution of judgements for descriptions of weather conditions randomly selected from the database, and identified by using NN. In the random case the distribution of judgements shows a clear peak at Rating 0, in other words most of the descriptions retrieved are not considered similar to the target day. On the other hand, the distribution for NN shows a peak at Rating 2, suggesting that NN has skill in selecting neighbours with weather conditions similar to the forecast day. Given that most implementations of NN are based on the use of weather data it is reassuring that NN appears well able to identify days considered either very similar (47%) or partially similar (32%) to the target day.

In contrast to NN's skill in selecting similar days in terms of weather, Figure 5 shows that NN appears to have little more skill than random selection in identifying days with similar snowpacks, despite the inclusion of some variables describing current snowpack conditions in NN. There are a number of possible reasons for this disparity. Firstly, the snowpack data used in NN may not reflect the descriptions given by our forecasters to the snowpack. Secondly, NN pattern matches days with similar weather conditions to the target day and a small number of previous days. This implies that recent weather conditions are often not enough to successfully discriminate between likely snowpack conditions resulting from the long term development of a snowpack over a winter.

Finally, Figure 6 shows that our results analysing NN's skills in identifying days with similar stability are somewhat ambiguous. One annotator considered NN to deliver matched or partially matched information only in around 42% of cases, whilst the other annotator considered around 73% of cases to have either a partial or

good match. This result suggests, firstly, that comparing stability descriptions is more difficult than comparing weather or snowpack descriptions and secondly that NN's skill in identifying similar stability conditions is mixed. Given that the results for snowpack indicated that NN had little skill in identifying days with similar snowpack conditions, this implies in turn that a family of avalanche events directly related to weather conditions (e.g. avalanches following a rapid thaw or heavy snowfall) are more likely to be selected by NN.

In discussing the results of these initial experiments it is important to sound a note of caution. These experiments are for a single area (the Parsenn), with a single winter's worth of descriptions and the quality of matches has been annotated only by two annotators (this paper's authors). However, our initial experiments suggest that such a qualitative method has considerable potential as a tool for the evaluation of computer-assisted avalanche forecasting, in this case through the use of NN.

One other point is worthy of discussion. In initiating this work we were mindful that those responsible for avalanche security are generally extremely busy, especially in conditions that might be deemed "interesting" – in other words those where blog entries have the most value to us. With this in mind we were unsure as to whether sufficient entries would be made to allow this work to be carried out. In fact, the Parsenn team reported that filling in the blog was a useful daily exercise in considering avalanche conditions and have asked how the information collected might be made available as a further, descriptive, output of NN.

6. CONCLUSIONS AND FURTHER WORK

This paper has presented a methodology for the validation of a particular family of forecast tools commonly used in avalanche forecasting, namely nearest neighbours. At the outset of the paper we set out a number of possible interpretations of the output of NN, two of which we argued methods already existed to validate (namely categorical and probabilistic interpretations) and, one – descriptive interpretation – for which no method, to our knowledge, existed.

A key question in validating the descriptive output of NN was set out as follows:

- How similar are the neighbours returned in terms of weather, snowpack and stability?

We presented a method which examined this similarity in terms of descriptions collected throughout a winter by the avalanche security team of the Parsenn ski resort in Davos, Switzerland. These descriptions were compared to days selected by NN and those randomly selected from a database, to test the hypothesis that NN should have more skill than random selection of neighbours.

Our initial results show that, while this is the case for neighbours in terms of descriptions of the prevailing weather, NN has poor skill in selecting days with similar snowpacks and, at best, ambiguous skill in terms of stability descriptions. With respect to the skill of NN in selecting days with similar stability descriptions we note that inter-annotator agreement for these descriptions was considerably poorer than for both weather and snowpack descriptions.

Importantly, the method appears both to have potential in exploring the descriptive quality of NN and to be an acceptable and useful mode of data collection for those responsible for avalanche security.

In the introduction we set out a second important question for analysis with respect to the descriptive output of NN:

- Does the number of neighbours returned with avalanche events combine to suggest greater instability or avalanche hazard (as opposed to a greater probability of a single event)?

In order to assess this question we intend to compare descriptions for target days with the number of nearest neighbours retrieved (e.g. whether the target day's description suggests many avalanches on days with many neighbours with avalanche events). We have also collected a further winter's data from the Parsenn and in a forthcoming paper will present results of the analysis of this larger dataset.

ACKNOWLEDGEMENTS

Many thanks to Hampi Amacker, Gian Darms, Walter Düesel, Romano Pajarola and Graham Moss for recording the blog. We are grateful to Felix Hebeler who adapted the blog interface for the special needs of this project and to Roli Meister who helped us by carefully selecting the most similar winter to the test winter in the last 15 years.

REFERENCES

Bolognesi, R., 1994: Local avalanche forecasting in Switzerland: strategy and tools, *Proceedings, International Snow Science Workshop, Snowbird, UT, USA*, 463– 472.

Buser, O., 1983: Avalanche forecast with the method of nearest neighbours: An interactive approach. *Cold Reg. Sci. Technol.,* **8**, 155-163.

Buser, O., 1989: Two years experience of operational avalanche forecasting using the nearest neighbours method. *An.Glaciol.,* **13**, 31-34.

Doswell, C.A., Davies-Jones, R., and Keller, D.L., 1990: On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables. *Weather and Forecasting,* **5**, 576-585.

Gassner, M., Birkeland, K., Etter, H.J. and Leonard, T., 2000: NXD2000: An improved avalanche forecasting program based upon the nearest neighbour method. *Proceedings, ISSW, Big Sky, Montana, USA,* 52-59.

Heierli, J., Purves, R.S., Felber, A. and Kowalski, J., 2004, Verification of nearest neighbours interpretations in avalanche forecasting. *An. Glaciol.*, **38**, 84-88.

McCollister C., Bikeland, K., Hansen, K., Aspinall, R., Comey, R., 2002: A probabilistic technique for exploring multi- scale spatial patterns in historical avalanche data by combining GIS and meteorological nearest neighbors with an example from the Jackson Hole Ski Area, Wyoming. *Proceedings*, *ISSW*, *Penticton, B.C., Canada*.

Mérindol, L., Guyomarc'h, G., and Giraud, G., 2002: A French Local Tool for avalanche hazard forecasting : Astral, current state and new developments. *Proceedings*, *ISSW*, *Penticton, B.C., Canada*.

Murphy, A.H., and Winkler, R.L., 1986: A general framework for forecast verification. *Mon. Wea. Rev.,* **115**, 1330-1338.

Purves, R.S., Morrison, K., 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**, 343-355.

Rykiel E.J., 1996: Testing ecological models: the meaning of validation. *Ecological Modelling*, **90**, 3, 229-244.

Voorhees, E.M., 2000: Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, **36**, 697–716.

Zeidler, A., and Jamieson, B., 2004: A nearest-neighbour model for forecasting skier-triggered dry-slab avalanches on persistent weak layers in the Columbia Mountains, Canada. *An. Glaciol.*, **38**, 166-172.