AUTOMATED NUMERICAL PREDICTION USING ELECTRONIC METEOROLOGICAL AND MANUAL SNOWPACK DATA

P. D. Cordy
University of British Columbia, Vancouver, BC

D. M. McClung
University of British Columbia, Vancouver, BC

C. J. Hawkins
University of British Columbia, Vancouver, BC

T. Weick
British Columbia Ministry of Transportation, Victoria, BC

ABSTRACT

Nearest neighbour algorithms using manual observation data can provide useful and accurate predictions of avalanche activity (McClung and Tweedy 1994, Floyer and McClung 2003, Roeger et al. 2003a, Zeidler and Jamieson 2004, Purves 2003). Here, a system is proposed that will use electronic data from automated weather stations in two distinctly different avalanche prone transportation corridors: Kootenay Pass and Bear Pass in British Columbia, Canada. The goal is to create a flexible, modular framework for numerical avalanche prediction using nearest neighbours that is automated, scalable, and that can be easily applied to different forecast operations. In addition to now-casts of avalanche probability, the program will provide advanced forecasts based on numerical or human meteorological forecasts (Roeger et al. 2003a). Furthermore, two methods of incorporating snowpack information into the avalanche predictions are outlined. The first is a simple threshold sum method similar to the one proposed by Schweizer and Jamieson (2003), and the second employs a data mining algorithm called MART (multiple additive regression trees). Probabilities generated by each algorithm will be combined using a Bayesian framework (McClung and Tweedy 1994).

KEYWORDS: Numerical Avalanche Prediction, Nearest Neighbours, MART, highways

1 INTRODUCTION

The primary goal of this research is to implement a lasting solution for numerical avalanche prediction in highways corridors in British Columbia, Canada. Previous work (McClung and Tweedy 1994, Floyer and McClung 2003, Roeger et al. 2003a) focused on manual data and was difficult to integrate into the British Columbia Ministry of Transportation's (hereafter: MoT) computer systems and forecaster protocols. Forecasters had to manually enter data into the prediction algorithm and the software soon became obsolete as the MoT updated its computing infrastructure.

The current research aims to reintroduce numerical prediction for maximum flexibility, functionality, longevity, and minimal effort on the part of forecasting personnel. Meteorological data will be fed automatically to the prediction algorithm, prediction outputs will be scalable and customizable for transplantation to any avalanche corridor for which sufficient data exists. The software will be implemented in Java so that it can be integrated into the MoT computer systems and so that it can be updated along with changes in MoT computer infrastructure. Nearest neighbour analysis is the chosen prediction algorithm due to its simplicity and success in many

Paul Cordy, UBC Geography,
1984 West Mall, Vancouver, BC
V6T 1Z2. fax: 604-822-6150
Email: paulcordy@gmail.com

studies (McClung and Tweedy 1994, Floyer and McClung 2003, Roeger et al. 2003a, Zeidler and Jamieson 2004, Purves 2003).

New challenges and opportunities arise in the shift from manually collected observations to an electronic weather station network. The proposed system makes use of hourly information and can be supplemented with manual data. Forecasters will also be able to automatically or manually input weather forecast information to extend the range of avalanche predictions into the future.

Solutions are also presented here for:

- missing values in the historic and current data.
- problems of disparity between the number of avalanche days versus non-avalanche days in the data,
- generation of hourly interval memory variables,
- mismatched observation intervals among electronic, manual, and snowpack data.

Finally, a method is proposed for integration of snowpack information into the numerical prediction system.

2 FUNCTIONAL DESIGN

The default prediction algorithm is a nearest neighbours analysis, but the program has been designed to allow substitution of other classification algorithms, and future versions will include a choice of algorithms such as Multiple Additive Regression Trees (MART) and a threshold sum for snowpack information.

2.1 *The Nearest Neighbour model*

The nearest neighbour algorithm is still among the best classification algorithms available (Hastie et al 2001). Nearest neighbours analysis ranks the historic data in terms of similarity to meteorological conditions (predictors) in the current datum. The dominant class among the nearest historic data is the predicted class of the current datum. Similarity is measured by Euclidian distance metric in predictor space in which each predictor is represented along an axis in this space. By adjusting the relative weights of each predictor using a genetic algorithm (Purves 2003), predictive accuracy can be optimized with respect to the Hanssen-Kuipers fitness metric (Roeger et al. 2003b). The predicted avalanche probability is given by the proportion of nearest neighbours associated with avalanches to the total number of nearest neigbours (k). The default k is 30 nearest neighbours. When optimizing the genetic algorithm, k=6 (20% probability) is the "threshold k" that classifies a day as an avalanche day (McClung and Tweedy 1994).

A nearest neighbour is considered to be associated with avalanches if avalanches occurred in a twelve hour period following the prediction time on that day. Future versions of the program will allow the forecaster to vary this period such that nearest neighbours are associated with avalanches that occur before the prediction time or more than 12 hours afterward.

All available information about the nearest neighbours, including avalanche activity, forecaster generated avalanche hazard forecast levels, and meteorological conditions will be displayed to the forecaster.

2.2 *Customization and scaling*

Avalanche occurrence data can be filtered by type (natural, explosive triggered, wet or dry) and size, so that the algorithm predicts only the type and size of avalanches that are of interest. The user can also select which of the available predictor variables are to be used by the prediction algorithm. If manual observation predictor variables are used, the program will prompt the user to input the necessary data and these values will be used hourly until changed by the user on or before the next standard observation. Manual data are taken near electronic standard observation times (0600 and 1800) instead of hourly. Interval manual data will be interpolated simply by using the most recent values until the next standard observation.

Future versions will also enable forecasters to vary most other aspects of the prediction algorithm, including;

- predictor variable lags (such as cumulative snowfall for 24 or 48 hours)
- k and threshold k
- intervals for which maximum and minimum temperatures are logged.

- filtering avalanche occurrences by path number.

Thus the program is scalable; the algorithm can make a prediction for the entire region using all occurrences in all paths, or predictions can be made for sub-regions by considering only avalanche occurrences from select paths.

2.3 *Hourly prediction intervals*

The nearest neighbour algorithm is run every hour. The time when the algorithm computes a prediction is referred to as the prediction time. The prediction is based on the nearest data out of all of the historic data from the same hour. For example, if the prediction time is 1400 hours the nearest neighbour prediction algorithm will only reference the data taken at 1400 hours on each of the historic days in the database. This procedure reduces computation time and ensures that observation and prediction data are well matched.

2.4 *True forecasts*

The program can also use weather forecasts (Roeger et al. 2003a) instead of current observation values given by the automated weather stations. There will be the choice of manual and numerical forecast inputs. Numerical forecast inputs will be automatically queried from UBC Atmospheric Science Kalman-corrected predictions (Roeger et al. 2003b) at regular intervals specified by the forecaster. If the forecaster chooses to use local non-numerical weather forecasts, the manual forecast input window will allow them to enter values taken or estimated from local weather forecast service providers. These values are then used by the prediction algorithm to make a one-time only forecast.

2.5 *Missing values*

Occasionally there are missing values in the historic data, or an electronic sensor is not recording observations with confidence. The nearest neighbour algorithm will therefore only reference rows of historical data that have valid values for all of the predictors that the algorithm requires; the program will skip over rows with missing data. In the event that there is no current observation for a given predictor at prediction time, then the nearest neighbour algorithm will only calculate distance in predictor space based on the available values.

Future versions of the program will include observation estimator algorithms in which linear regression estimates the value that would be given by an inoperative electronic sensor using the value of a nearby operational sensor and the statistical relationship between the two. Another, perhaps more accurate, solution is to use the UBC atmospheric science Kalman-corrected prediction for the missing value at the inoperative sensor.

2.6 *Data bias and forecaster prior*

Avalanche days are much less common than non-avalanche days and this can bias the classification results toward non-avalanche days if the classification algorithm assumes parity between classes. In order to ensure meaningful prediction probabilities, Bayes rule (posterior probability α likelihood x prior) is used to account for this bias (Hastie et al. 2001). The likelihood is given by the nearest neighbours output, and the prior is a function of the ratio of avalanche days to non-avalanche days. Then, in the second round of inference, the forecaster can set their own prior belief as to the probability of avalanching as in McClung and Tweedy (1994).

2.7 *Memory variables*

Early iterations of the predictive system will use standard observation values for maximum and minimum air temperature, as well as cumulative (lagged) snowfall from the past two standard observations.

In future versions of the program, extra columns will be added to the database that log the maximum and minimum air temperature during the 12 hours preceding the observation time (hereafter referred to as “logs”), and that calculate the cumulative snowfall during the 24 hours preceding the observation (hereafter referred to as “lags”). When using the additional lag and log variable columns, the observation period shifts hourly in unison with the other

variables instead of shifting discretely by 12 hour jumps for variables taken at standard observations.

Lagged and logged variables will have real time filters that sum the lagged value or log the maximum or minimum of a value of these predictors from current observations.

3 FUTURE EXPANSION

In order to improve the accuracy of this prediction system, methods are presented here for inclusion of snowpack structure information.

Snowpack models will make predictions whenever the forecasters dig a snow pit and enter this information into the program. The relative contribution of these predictions will decrease with time as the snowpack inevitably changes from the observed state. Output probabilities from snowpack models will be combined using the Bayesian framework discussed above; a snowpack prior will be combined with the likelihood given by the bias prior adjusted nearest neighbour output. As before, the final round of inference merges the combined numerical model likelihood with the forecaster's prior. Early versions of the program will include a threshold sum model, and Multiple Additive Regression Trees will be offered later.

3.1 *The threshold sum model*

This model is based on the threshold sum method of Schweizer and Jamieson (2003). It is a simple prediction algorithm in which one chooses a critical threshold and a weight for each variable. The critical threshold is the value above (or below) which a variable is expected to significantly contribute to avalanche instability. The weight reflects the relative importance of a given variable in predicting instability. The threshold sum score is calculated by summing the weights of all the variables that have exceeded their threshold (or failed to exceed their threshold for variables where low values increase avalanche instability). Initially the ratio of threshold sum score to maximum possible score will give the probability of avalanching given the snowpack information. This value might have to be adjusted by a bias prior such as in the nearest neighbours analysis, but in contrast to the nearest neighbours bias prior, the snowpack bias prior might have to be estimated by the forecaster.

3.2 *Multiple Additive Regression Trees (MART)*

MART is a boosted version of a Classification and Regression Tree (CART) (Hastie et al. 2001). CARTs are optimized by a recursive binary partitioning of the predictor space such that each partition, or "rule" is made on a threshold of a single variable such that the data is separated into two classes with the least number of misclassified points. Thus a new datum filters down the rule branches until reaching a terminal node where the dominant class in that node dictates the class of the new datum.

Tree methods for classification are attractive for several reasons: they require virtually no data preprocessing, they are insensitive to missing data, they naturally perform feature selection, and data can be categorical, ordinal, numerical or a mixture of types with different distributions.

The MART algorithm generates a forest of trees of limited size and the predictions from all trees are averaged. Tree classifications are boosted in the following way: trees are built iteratively, and misclassified points from earlier trees carry heavier penalty for misclassification in successive trees. This way most of the variance is accounted for by early trees, and later trees capture outliers.

Since MART is normally used only as a classifier and not to generate probabilities, the program will output the model's verification accuracy as the probability of avalanching. Alternatively, the ratio of avalanche days to total days in the terminal nodes reached by the current datum can provide a probability of avalanching, but this method may be unstable.

In order to initiate MART analysis on snowpack data, relevant information must be mined from snow profile data files and tabled in a database. This process will be different for each geographical region Once the new snowpack variables are imported into the model, they are associated through their date stamp to the appropriate

avalanche occurrences from a relevant response period surrounding the snowpack observation. A reasonable approach is that the window should extend a few days after of the observation date and only one day previous. The user will, of course, have the freedom to choose whatever response period they feel is appropriate.

Unlike the nearest neighbour algorithm, once MART has been optimized there is no need to reference the historic data. One need only input the new snowpack structure values (and/or data from other variables upon which the MART module was optimized) and the algorithm will provide a prediction. The goal is to automate this process by standardizing the way forecasters enter data into a snow profile software package such that a single data mining algorithm can extract the data.

4 CONCLUSIONS

Methods presented here aim to largely automate the process of numerical avalanche prediction such that the algorithms merge seamlessly with the MoT database and protocols. Nearest neighbours analysis forms the core of the predictive system, and the forecaster has many choices as to the algorithm parameters, predictors, memory variables, occurrence data filters, and eventually the choice of prediction algorithm. Bayesian inference is used to combine probabilities given by different models and the forecaster's prior knowledge, as well as remove class size disparity bias. Avalanche predictions can be extended into the future using human or numerical weather forecasts. Prediction accuracy may be improved by the inclusion of snowpack structure information either by a simple threshold sum algorithm, for which no data mining is required, or by MART, which requires a historic snow profile database. The avalanche prediction system proposed here is designed to grow along with the evolving database, electronic weather station and computer infrastructure at the British Columbia Ministry of Transportation.

5 REFERENCES

Floyer, J. Statistical avalanche forecasting using meteorological data from Bear Pass, British Columbia, Canada. University of British Columbia. Dept. of Geography. Thesis. M.Sc., 2003.

Floyer, J. and D.M. McClung, 2003. Numerical avalanche prediction in Bear Pass, British Columbia, Canada, *Cold Regions Science and Technology.* 37, 333-342.

Haegli, P. and D.M. McClung. 2000. A new perspective on computer assisted avalanche forecasting: scale and scale issues. *Proceedings of the International Snow Science Workshop 2000*. Bozeman, MT, American Avalanche Association. 66-73.

Hastie, T., Tibshirani, R., Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag. New York.

McClung, D.M. 2002. The elements of applied avalanche forecasting Part II: the physical issues and the rules of applied avalanche forecasting. *Natural Hazards*. 25, 131-146.

McClung, D.M. and P. Schaerer. 1993. *The Avalanche handbook*. The Mountaineers, Seattle.

McClung, D.M., Tweedy, J., 1994. Numerical avalanche prediction: Kootenay Pass, British Columbia. *Journal of Glaciology*. 40, 350-358.

Purves, R., K. Morrison, G. Moss and B. Wright. 2003. Nearest Neighbours for avalanche forecasting in Scotland – development, verification and optimization of a model. *Cold Regions Science and Technology*. 37, 343-355.

Roeger, C., McClung, D. and R. Stull, 2003a. Verified combination of numerical weather and avalanche prediction models at Kootenay Pass, British Coloumbia, Canada, *Annals of Glaciology.* 38, 334-346.

Roeger, C., R. Stull, D. McClung, J. Hacker, Xingxiu, Deng,and H. Modzeleski, 2003b. Verification of fine-grid numerical weather forecasts in mountainous terrain weather for application to avalanche forecasting. *Weather and Forecasting*. 18, 1140 - 1160.

Schweizer, J. and B. Jamieson. 2003. A threshold sum approach to stability evaluation of manual snow profiles. *Cold Regions Science and Technology*. 37, 233-241.

Zeidler, A. and J.B. Jamieson. 2004. A Nearest neighbour model for forecasting skier-triggered avalanches on persistent weak layers in the Columbia Mountains. *Annals of Glaciology*. 38, 166-172.