SNOW SENSOR DATA QUALITY INDEXING[1]

by

Clay Brandow[2] and Anthony Lourick[3]

INTRODUCTION

California's automatic snow sensing network produces volumes of snow water equivalent (SWEQ) data. Even when the data are reduced to daily values, the sensor network produces over 10,000 SWEQ values per snow season. With such a large and rapidly growing record, identifying which sensors and periods have good data--and which do not--is a problem. One solution is to index the quality of snow sensor data.

This paper describes a snow data quality index (SDQI) system devised for California's snow sensor network. Using a unique feature of the California Data Exchange Center (CDEC) data base, SDQI scores--indicating the relative quality of the data--can be calculated for each sensor by water year. An SDQI score of zero indicates no usable data. An SDQI score of 100 indicates perfect data, though a perfect score is more likely to indicate that the data has not been rigorously examined.

SDQI scores can be computer generated, using a program written for this purpose. The scores can be displayed in a table (two-dimensional matrix), with water years across the top and snow sensor names (mnemonics) down the side. Such tables make it easy to compare the data quality for specific snow sensors for specific periods. They can also reveal how the network is doing as a whole.

EDITING

A unique feature of CDEC's data base is that for each snow sensor two daily values are stored---a raw value and an edited value. Each snow sensor reports SWEQ at least once a day to the CDEC computer, during the snow season. The computer is programmed to pick one of those readings as the daily SWEQ value for that sensor. Those values are placed in both the raw and edited SWEQ data files. Values in the raw file remain unchanged. Values in the edited file are adjusted to reflect the actual SWEQ as closely as possible. Using the yet to be adjusted values in the edited file, the computer generates an initial "Daily Snow Sensor Report". This report is often referred to as CDEC's "Page 6".

Each morning, an assigned CDEC person looks at that initial report (Page 6) and edits the data contained in it, producing a revised report. Then he or she makes the Daily Snow Sensor Report--with edited data--available to the public on the CDEC computer. Simultaneously and automatically, the edited data from the revised Daily Snow Sensor Report replaces the raw data in the edited data file. The edited data file now contains edited values for that day. The raw data file remains unchanged, and is permanent record of the data received via telemetry.

At the end of the water year, the edited files are edited a second time, taking advantage of additional information and hindsight.

Since 1989, CDEC has been setup to handle snow sensor data in the manner just described. The data prior to 1989 went through a one-time edit to create an edited data file for each sensor.

SDQI takes advantage of the fact that CDEC has raw SWEQ data files and edited SWEQ data files, and that they can be compared, revealing how much editing the incoming (raw) data required.

---

[1]Presented at the Western Snow Conference, April 12-15, 1991, Junuea, Alaska.

[2] Hydrologist, California Department of Forestry and Fire Protection, 1416 Ninth St., Sacramento, CA. 95814

[3] Associate Programmer Analyst, California Department of Water Resources, 1416 Ninth St., Sacramento, CA. 95814

INDEXING

Most data contain errors. The number of defective values divided by the total number of values in the data set can be expressed as a percent defective (J.I. Naus 1975). A simple data quality index could be expressed as 100 minus the percent defective.

Snow sensor SWEQ data contain errors, but some of the defective data are useful in approximating actual values. SDQI is a weighted index which recognizes this point. The basic premise is data requiring less editing to reflect actual SWEQs are higher quality data.

A high weight is given to data that arrives at CDEC and are usable as is--requiring no adjustment during editing. Lesser weight is give to data that contain minor errors needing slight adjustment. A low weight is given to data that contain gross errors or are missing, requiring major adjustment or synthesizing. When there is not enough information to derive even synthetic SWEQ values, a weight of zero is given.

The Snow Data Quality Index (SDQI) program calculates SDQI scores using the following expression:

$$SDQI = \frac{1GD + 0.66FD + 0.33PD + 0.33SD + 0.0ND}{N} \times 100$$

where:

SDQI is the quality index score for a sensor for a water year expressed as a percentage. Assuming adequate editing, an SDQI score of 100 indicates perfect data. Zero equals no useable data for the water year.

GD is the number of days with good data: data where the adjusted value for that day is no more than 0.5 inches (1.3 cm) different from the raw value. Weight each GD day with a value of 1.

FD is the number of days with fair data: data where the adjusted value for that day is no more than 2 inches (5 cm) different from the raw value. Weight each FD day with a value of 0.66.

PD is the number of days with poor data: data where the adjusted value for that day is more than 2 inches (5 cm) different from the raw value. Weight each PD day with a value of 0.33.

SD is the number of days with synthesized data: data where an adjusted value exists, but no raw data value exists. (Note, missing values in the SWEQ data base are listed as -999.9) Weight each SD day with a value of 0.33.

ND is the number of days (during the snow season) with no data: no data exists for that day and none could be synthesized, due to a lack of information (ie. no applicable control measurements, no opportunity to interpolate, etc.). Weight each ND day with a value of zero.

N is the total number of days with snow on the ground. That is to say the adjusted snow water content value for that day is greater then zero. For truncated records, those water years where a sensor record does not begin and end with zero, N defaults to a number set by the program operator. In the example presented below, the default N was set at 150 days.

The program can be run for all the sensors or a selected few. It can be run for the whole record or for a specified period of water years. The output for 1965 to the present for all the sensors consumes about six wide pages of computer printout. Too big to include here.

Here is a sample output for two sensors:

Station ID	Average SDQI	Water Year '80	'81	'82	'83	'84	'85	'86	'87	'88	'89	'90
GIN (Gin Flat)	85	*	*	84	88	97	81	73	89	91	96	70
PDS (Paradise)	54	^	85	30	83	0	85	58	6	80	56	59

* The record for the Gin Flat snow sensor starts in WY 1982.
^ The record for Paradise snow sensor starts in WY 1981.

For an up-to-date SDQI run for any or all of the sensors contact either of the authors.

LIMITATIONS

Since SDQI scores are based on a comparison of the raw SWEQ data to the edited SWEQ data, SDQI scores are only as good as the data editing. If new data is not rigorously edited, the SDQI scores will be too high, and chronic problems are more likely to go undetected and unresolved.

Also, in order to make valid SDQI score comparisons, and more importantly to build a good data set, the data editing protocol must be consistent from year to year and from sensor to sensor. Consistently collecting and using SWEQ field control measurements is particularly important to good data editing.

Finally, SDQI is a relative index of data quality and lumps many different problems together. To see why a sensor is performing poorly, a closer look at the data and the sensor maintenance record is needed. For the future, development of more remote diagnostic tools and techniques would be extremely useful in the effort to keep data quality high at a reasonable cost.

USES

There are two basic concerns in snow sensor data quality. First, how can we make best use of a historical data set that has problems? Second, how can we make new data meet or exceed certain minimum objectives and standards?

SDQI can help. First, SDQI can be used as a tool by hydrologists to avoid poor quality and missing data when selecting sensors and calibration periods for new water-supply forecasting procedures. Other users of the data set can use SDQI for similar purposes. Second, SDQI can be used as a tool by CDEC managers to identify trends and chronic problems in the performance of specific sensors and basin networks.

CDEC needs to officially adopt some snow sensor data quality objectives (DQOs) and standards. The component parts of the SDQI expression could provide the basis for several DQOs. For example, a DQO stating that 80% of the daily SWEQ values should fall into the good data (GD) classification (standard) could easily be audited using SDQI with a slight modification. Likewise, a DQO prescribing that 95% of a sensor's snow season days have useable SWEQ values--synthesized data (SD) or better classification (standard)--could be easily tracked. Another DQO could be adopted addressing length of periods with no data. These are just a few rough suggestions. DQOs actually adopted by CDEC should reflect the data quality needs of the data users. Consequently, data users should help write the official DQOs for the network.

To use the SDQI computer program to track DQOs would require writing some new output routines, a fairly straight forward task. Using SDQI derivatives to audit data quality could be a very useful tool for annual management reviews of sensor and network performance.

CONCLUSION

CDEC is striving to produce data of sufficient quality to meet the needs of its users. In a small way, Snow Data Quality Indexing may be able to help in this quest.

REFERENCES

Blacker, Stanley M. and Brantly, Eugene P. (1984) Data Quality Objectives and the EPA Quality Assurance Program, Proceedings of the APCA Conference on Quality Assurance in Air Pollution Measurements, Boulder, CO., p. 261-278.

Burnash, Robert J. C. and Ferral, Larry R. (1979) Contamination in the Hydrologic Data Base, Proceedings of the Engineering Conference on Improved Hydrologic Forecasting: Why and How, ASCE, Asilomar, CA., pp. 303-311.

Johnson, Ted R. and Penkala, Stanley J. (1985) Quality Assurance in Air Pollution Measurements: Summary of an APCA Specialty Conference. Journal of the Air Pollution Control Association. 35(4):337-344.

Naus, Joseph I. (1975) Data Quality Control and Editing, pub. Marcel Dekker, Inc., New York, New York, pp. 73-78.

Mosely, M. P. and McKerchar, A. I. (1989) Quality Assurance Programme for Hydrometric Data in New Zealand. Hydrological Sciences Journal. 34(2):185-202.

Suits, R. D. (1985) California Snow Sensor Performance. Masters Thesis, University of California, Davis, 142 pp.