9.6 SPATIAL REGRESSION AS A TECHNIQUE FOR ASSESSING THE QUALITY OF TEMPERATURE DATA

Nathaniel B. Guttman *
NOAA National Climatic Data Center, Asheville, North Carolina

1. INTRODUCTION

Quality assurance procedures have been applied by the National Climatic Data Center (NCDC) (Guttman and Quayle 1990) in a mix of manual and automatic checks to assess the validity of weather data from the cooperative climatological stations. General testing approaches such as using threshold and step change criteria were designed for the single station review of data to detect potential outliers (Wade 1987; Meek and Hatfield 1994; Eischeid et al. 1995).

Recently, the use of multiple stations in quality assurance procedures has proven useful. Spatial tests compare a station's data against the data from neighboring stations (Wade 1987; Gandin 1988; Eischeid et al. 1995, Hubbard 2001). They involve the use of neighboring stations to make an estimate of the measurement at the station of interest. This estimate can be formed by weighting according to distance separating the locations (Guttman, 1988; Wade, 1987), or through other statistical approaches (e.g., multiple regression, (Eicheid et al., 1995) and linear regression (Hubbard et al. 2005).

The spatial regression test (SRT) described by Hubbard et al. (2005) and used at the High Plains Regional Climate Center (HPRCC) does not assign the largest weight to the nearest neighbor but, instead, assigns weights according to the root mean square error (RMSE) between the station of interest and each of the neighboring stations. Research has demonstrated excellent performance of the spatial regression test in identifying seeded errors (Hubbard et al. 2005). The SRT approach has been found in a previous study (You et al. 2004) to be more accurate than the inverse distance weighting (IDW) approach for the maximum air temperature (Tmax) and the minimum air temperature (Tmin). It was found that the RMSE was smaller for SRT estimates than for IDW estimates in all areas including the coastal and mountainous regions. Both the spatial regression and inverse distance methods were found to perform relatively poorer when the weather stations are sparsely distributed (You et al. 2004). The success of the spatial regression approach is in part due to its ability to implicitly resolve the systematic differences caused by temperature lapse rate with elevation; these differences are not accounted for in the IDW method.

* *Corresponding author address:* Nathaniel B. Guttman, National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801; e-mail: ned.guttman@noaa.gov.

2. CURRENT NCDC QUALITY ASSESSMENT

The NCDC quality assessment is based on accepting all observed data that are plausible. There are five steps in the evaluation of temperature data. Because of the volume of data that are processed as well as requirements to provide quality assessed digital data to customers in near real-time, a goal of the approach is to automate as much evaluation as possible. The operational processing consists of four steps:

a. Pre-edit – Input data records are checked for format and coding errors. Improper station identifiers, invalid characters, duplications, values that are not in a valid range, unexpected data, and other similar problems are identified and corrected if possible. If it is not possible to correct these errors, then a datum is labeled as missing.

b. Climate Division Consistency – Departures of a station's data from the monthly average of the data are calculated for all stations within a climatic division (see Guttman and Quayle 1996 for a description and history of the 344 climatic divisions in the contiguous U.S.). The average departure for each day is then calculated. A datum is flagged for further review if the departure for a given station and day differs from the divisional average for the day by more than $\pm$ 10F. For a given day, temperature means and variances are estimated from all the divisional data that have not been flagged for further review. Any flagged data that exceed $\pm$ 3 standard deviations from the mean for the day are then flagged as suspect. Validators also compare the divisional data to the top 10 and bottom 10 observed extremes for the State. This comparison is intended to identify gross keying errors and anomalous extreme values, and is performed both on the observed data and on the replacement values.

c. Consistency – This check insures that maximum, minimum, and observation time temperatures are internally consistent. Physically impossible relationships, such as the minimum temperature for a day being greater than the maximum temperature for the same day, are flagged as suspect. Often, these errors result from incorrect dates that are assigned to an observation (sometimes called "date shifting"); if possible, the flagged data are corrected.

d. TempVal – This spatial check uses grid fields derived from ASOS (Automated Surface Observation System)/ AWOS (Automated Weather Observing

System) hourly and daily temperature values as a "ground truth" to quality assure the Cooperative Network daily temperature data (Angel et al. 2003). Note that the previously described steps are only applied to the cooperative data; this step compares the cooperative data to an independent data network.

The data for a cooperative site are compared to the grid estimates at the site. When the difference between a cooperative value and the estimated value is greater than ± (7F + gradient of the grid at the site), the cooperative datum is flagged as suspect. Note that the constant 7F is usually much greater than the gradient, so that the threshold is approximately a fixed value; the acceptance range for an observed datum is of the order of 15-20F.

The assessment methodology not only identifies suspect data, but also yields estimations of values that are likely to be correct. When the data are archived, the original observed values are always retained with the estimated values.

3. FUTURE NCDC QUALITY ASSESSMENT

Using a data seeding methodology, the NCDC and HPRCC compared the TempVal and SRT methods of detecting suspect data values. The comparison showed that the two techniques performed equally in detecting large Type II errors, but that SRT can detect more moderate and small errors than TempVal. TempVal has proven operationally useful in identifying date shifters (wrong date associated with a datum), observation time problems (wrong time, changes in observer's schedules that have not yet been officially recorded), and anomalous extremes. It was concluded that TempVal be retained as an assessment tool, and SRT be added to the NCDC processing system.

The SRT methodology is being incorporated into the NCDC processing system. Because the SRT software developed by the HPRCC could not be "plugged and played" in the NCDC system, new code had to be written. In order to insure that the code was written as intended, parallel testing is being conducted on January through May, 2005 data for the lower 48 states. The NCDC version of the software is being run at the NCDC, and the HPRCC version is being run at the HPRCC, and results are being compared for one-to-one correspondence.

The data for the test period that are flagged by the SRT will be evaluated for plausibility as well as compared to the data flagged by TempVal. The number of flagged values will also be evaluated in terms of the level of acceptance of wrong decisions and of the manual resources needed to review the suspect data. The results of the evaluations will be presented at the Conference.

4. REFERENCES

Angel, W.E., M.L. Urzen., S.A. Del Greco., and M.W. Bodosky, 2003: Automated validation for summary of the day temperature data (TempVal). *83rd AMS Annual Meeting, combined preprints CD-ROM, 9-13 February 2003, Long Beach CA, 19th Conference IIPS [International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology]*, American Meteorological Society, Boston, Mass., File 15.3, 4 pp. (February 2003).

Eischeid, J. K., C. B. Baker, T. Karl and H. F. Diaz, 1995: The quality control of long-term climatological data using objective data analysis. *J. Appl. Meteor.,* 34, 2787-2795.

Gandin, L. S., 1988: Complex quality control of meteorological observations. *Mon. Wea. Rev.*, 116, 1137-1156.

Guttman, N., C. Karl, T. Reek and V. Shuler, 1988: Measuring the performance of data validators. *Bull. Amer. Meteor. Soc.*, 69, 1448-1452.

Guttman, N. V. and R. G. Quayle, 1990: A review of cooperative temperature data validation. *J. Atmos. Oceanic Tech.*, 7, 334-339.

Guttman, N. V. and R. G. Quayle, 1996: A historical perspective of U.S. climate divisions. *Bull. Amer. Meteor. Soc.,* 77, 293-303.

Hubbard, K. G., 2001: Multiple station quality control procedures. *in* Automated Weather Stations for Applications in Agriculture and Water Resources Management. AGM-3 WMO/TD No. 1074, 248P.

Hubbard, K. G. , S. Goddard, W. D. Sorensen, N. Wells, and T. T. Osugi, 2005: Performance of Quality Assurance Procedures for an Applied Climate Information System, *J. Atmos. Oceanic Tech.,* 22,105-112.

Meek, D. W. and J. L. Hatfield, 1994: Data quality checking for single station meteorological databases. *Agric. Forest Meteor.*, 69, 85-109.

Wade, C. G., 1987: A quality control program for surface mesometeorological data. *J. Atmos. Oceanic Tech.*, 4, 435-453.

You, J, K. G. Hubbard., and S. Goddard, 2004: Comparison of Estimates from Spatial Regression and Inverse Distance Method, *J. Atmos. Oceanic Tech.,* submitted.