Automatic data quality control using spatial interpolation with statistical methods

C. Lussana (1), M. Ranci (1), M. R. Salvati (1), F. Uboldi (2)
(1) Servizio Meteorologico Regionale - ARPA Lombardia, Milano, Italy, (2) consultant, Novate Milanese (MI), Italy

Lombardia’s regional meteorological network

- High spatial resolution
- High temporal resolution
- Automatic stations
- Inhomogeneous data quality
- Complex topography and land use
- High representativity error

Quality Assurance System at Data Processing Center

DATABASE

Every day about 27’000 hourly measures from the local subnetworks enter the meteorological database.

Each measure has a quality flag according to collection / transmission / reception / temporal aggregation status

<table>
<thead>
<tr>
<th>Quality Flag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>complete and plausible</td>
</tr>
<tr>
<td>1</td>
<td>incomplete and plausible</td>
</tr>
<tr>
<td>2</td>
<td>not plausible</td>
</tr>
<tr>
<td>3</td>
<td>missing</td>
</tr>
</tbody>
</table>

Quasi-real time procedures:

- Plausibility check (as defined in 1)
  setting of thresholds from the results of long-term series analysis
- Temporal continuity checks (as defined in 1)
  - Step check
  - Persistence check
    setting of thresholds from QA managers experience
- Spatial continuity check (based on Qi described in 2)
  so far implemented only for temperature (see below)

Automatic QC output is published every night on internal WEB server and sent via e-mail to the QA manager

The final human judgment is supported by:

- review of the suspicious observations listed in the automatic QC output
- information from other investigative tools, such as data visualizers
- The QA manager merges the available information and daily sets the new quality flags in the database

365 stations from 9 local subnetworks set up in the past for different purposes:

- 310 thermometers
- 331 rain gauges
- 140 anemometers
- 115 barometers
- 95 global radiometers
- 28 net radiometers

Spatial Continuity check

Cross-Validation (CV) \( CV_m = \frac{y_m^{CV_a} - y_m^{CV_o}}{\sigma_m} \)

The analysis value \( y_m^{CV_a} \) is computed leaving the observation \( y_m^{CV_o} \) out.

In the left puddleplot, the circle radius is proportional to \( CV_m \): \( CV_m > y_m^{CV_o} > y_m^{CV_a} > y_m^{CV_o} \)

In the right puddleplot the \( CV_m \) is normalized respect to its root-mean-square for each station.

The difference is related to the analysis error of that area.

Data Quality Control test:

\( CV_m > T \) expected value

The test is computed using a station dependent threshold. The constant value \( T \) depends on the network characteristics (prior probability of gross error).

In the left graph, the hyperbole represents the threshold for the observation pointed by the arrow.

In the right graph, quantities \( (y_m^{CV_a} - y_m^{CV_o}) \) and \( (y_m^{CV_o} - y_m^{CV_a}) \) are normalized with respect to their RMS for each station. In this case the hyperbole represents the threshold for all the stations. The observation pointed by the arrow is the only one that does not pass the test in this case.

Conclusions

Plausible value and temporal continuity checks are engeneered for most observations of real meteorological phenomena to pass the various tests, moreover thresholds are set the same in the whole region for stations placed in different landscapes and at different heights. As a result, in some cases these tests are not stringent enough to catch malfunctioning sensors.

Very often, spatial continuity check successfully filters outliers, spikes and short persistences in the observations that are not detected by other checks.

Further developments

An effort is under way to better describe the characteristics of the region, since a better knowledge of the network enables a better tuning of the spatial continuity check parameters.

Work is in progress to extend the spatial continuity check to other meteorological variables.

Work is under way to implement more stringent QC checks in which the plausibility of each observation is defined by the comparison with some statistical parameters associated to the long time series of the corresponding sensor.

References

2) Uboldi, F., M.R. Salvati, C. Lussana, 'Spatial analysis of observations from high resolution automatic meteorological networks - poster EMS 2007

CORRESPONDING AUTHOR: Cristian Lussana c.lussana@arpalombardia.it