Simultaneous Treatment of Random and Systematic Errors in the Historical Radiosonde Temperature Archive

Josh Browning

November 5, 2014

Table of contents

- Problem
  - Radiosonde Data
  - Random Errors
  - Systematic Errors
- Methods
  - Random Error Detection
  - Systematic Error Correction (Homogenization)
  - Sequencing
- Simulation Design
  - Modeling Data
  - Performance Metrics
- References

Project Overview

Problem: Random and systematic errors exist in the radiosonde temperature archive, and no studies have addressed how to handle both types of errors simultaneously.
Proposed Solution: I propose studying the effect of applying various sequences of quality control algorithms on simulated data which is designed to have the same structure as true radiosonde data.

Radiosonde Data

Figure: Sample Radiosonde

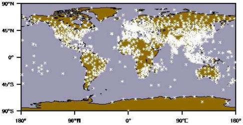

Figure: Global Radiosonde Network

- Small instruments are suspended below 2m hydrogen or helium balloons.
- They measure atmospheric variables such as temperature, wind speed, etc.
- Balloons are launched twice daily at roughly 00 UTC and 12 UTC, and at 700 sites worldwide.
- More than 1,300 additional launch sites exist in the historical archive.
- Considering the number of launch sites, multiple observations for different pressure levels, and the total number of launches, there are between 50 and 90 million historical soundings.

Problems with the data: Random Errors

Random errors can occur because of

1. Faulty data transmission
2. Sporadic instrumentation problems
3. Keystroke entries
4. Errors in data management
5. Many other reasons

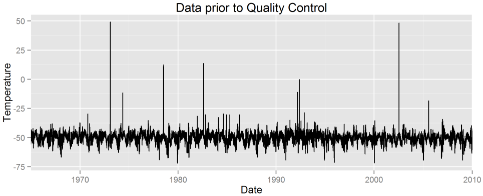

Figure: Data collected from Station 70219 in Bethel, Alaska.

Problems with the data: Systematic Errors

Systematic errors can occur because of

1. Station location changes
2. Urbanization of the area surrounding the station
3. Changes in instrumentation
4. Many other reasons

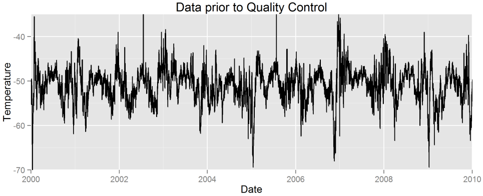

Figure: Data collected from Station 70219 in Bethel, Alaska.

Random Error Detection Methods

- The simplest way to detect random errors is to estimate the mean and standard deviation of the data and then to label observations as random errors if they are more than k standard deviations from the mean (where k may be 5 or so).
- In Anderson et al. (2014), a better alternative is proposed which uses robust estimates of the mean and standard deviation (more details on next slide).
- I propose using these techniques in my simulation study.

More on Robust Mean and Standard Deviation

1. First, the estimates of the mean, $\hat{\mu}$, and standard deviation, $\hat{\sigma}$, are initialized to

   $$\begin{aligned} \hat{\mu} &= \text{median}(\mathbf{x}) \\ \hat{\sigma} &= \text{MAD}(\mathbf{x}), \end{aligned}$$

   where **x** is a vector of the data, and *MAD* is the median absolute deviation, defined as

   $$MAD = \text{median}(|x_i - \text{median}(\mathbf{x})|).$$

2. Then, Winsorized values, y_i, are computed. These are defined as

   $$y_i = \begin{cases} \hat{\mu} - k\hat{\sigma} & : x_i \leq \hat{\mu} - k\hat{\sigma} \\ x_i & : \hat{\mu} - k\hat{\sigma} < x_i \leq \hat{\mu} + k\hat{\sigma} \\ \hat{\mu} + k\hat{\sigma} & : x_i > \hat{\mu} + k\hat{\sigma} \end{cases}$$

3. Updated estimates of $\hat{\mu}$ and $\hat{\sigma}$ are computed as the mean of **y** and the standard deviation of **y**, respectively.
4. Steps 2 and 3 are repeated until $\hat{\mu}$ changes by less than $10^{-6}\hat{\sigma}$.

Note: In step 3, one may choose to estimate $\hat{\sigma}_L$ and $\hat{\sigma}_R$ using only observations to the left and right, respectively, of the mean. This allows for improved outlier detection with skewed data.

Systematic Error Correction (Homogenization)

- Many homogenization methods have been proposed in the climate literature: Eskridge et al. (1995); Haimberger (2007); Lanzante (1996); Lanzante et al. (2003); Venema et al. (2012). However, these methods are not generally robust to outliers.
- Modern homogenization methods are also presented in the statistical literature Killick et al. (2012); Scott and Knott (1974).
- I propose comparing these methods and developing a homogenization method that is robust to outliers.

Homogenization Methods: SNHT

This algorithm is commonly used in the climate literature. The algorithm works as follows:

1. For each observation, two means are computed: one for the N days prior to observation i, $\bar{X}_{L,i}$, and one for the N days following, $\bar{X}_{R,i}$
2. Then, the test statistic

$$T_i = \frac{N}{s_i}\left((\bar{X}_{L,i} - \bar{X}_i)^2 + (\bar{X}_{R,i} - \bar{X}_i)^2\right), \tag{1}$$

   is computed where $\bar{X}_i$ is the mean of $\bar{X}_{L,i}$ and $\bar{X}_{R,i}$, and s_i is the estimated standard deviation over the N days prior and N days following observation i.
3. If the largest T_i exceeds some threshold at time $i = i^*$, I conclude that a change point occurred at time i^*, and I adjust all observations after time i^* by $\bar{X}_{L,i^*} - \bar{X}_{R,i^*}$. A threshold of 100 is recommended in Haimberger (2007).
4. Repeat steps 1-3 until no test statistic exceeds the threshold.

Homogenization Methods: Robust SNHT

- The previous algorithm is not robust to random errors, as the influence function for the sample mean is unbounded.
- The Winsorized estimator of center and scale discussed previously is robust against random errors.
- Thus, to create a robust SNHT statistic, I propose replacing the means and standard deviation in Equation (1) with these estimators.

Homogenization Methods: BinSeg

- Several homogenization methods work by optimizing a cost function:
$$\sum_{i=1}^{m+1}[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i})] + \beta f(m), \tag{2}$$
where τ_i is the ith change point; m is the number of change points; $\mathcal{C}$ is a cost function; $y_{(\tau_{i-1}+1):\tau_i}$ is the observed data between the $(i-1)$ and ith change point; and $\beta f(m)$ is a penalty term on the number of change points, see Killick et al. (2012).
- Often, $\mathcal{C}$ is chosen to be twice the negative log likelihood, and $f(\cdot)$ is linear.
- Binary Segmentation (BinSeg) uses a greedy algorithm: at each step, a changepoint is selected which minimizes the cost function. The data is then partitioned in two, and optimization continues iteratively.

Homogenization Methods: PELT

- Pruned Exact Linear Time (PELT) is another algorithm for optimizing Equation (2), but it computes the exact minimum.
- It proceeds recursively as follows: first, the optimal number and location of change points is determined for observations 1 and 2 only. The optimal number and location of change points for the first three observations is then determined using this information, and more generally the optimal number and location of change points for the first $k+1$ observations is determined by considering the optimal configurations for the first $2, 3, \ldots, k$ observations.
- PELT is computationally efficient, and is implemented in the `changepoint` package in R Killick and Eckley (2014).

Quality Control Algorithm Sequencing

- I am also interested in understanding the effect that the sequence of the quality control algorithms has on the overall quality control procedure.
- Let “Ran” denote the random error detection algorithm and “Sys” the systematic error correction algorithm. Then, I propose investigating the following alternatives:
  - Ran→Sys
  - Sys→Ran
  - Ran→Sys→Ran
  - Sys→Ran→Sys

Overall Simulation Design

To understand the effect of these various algorithms/sequencings on the final quality controlled dataset, I propose the following simulation design:

1. Using observed radiosonde data, develop a data generating mechanism with which to simulate data from.
2. Simulate a dataset using the model from step 1, and contaminate it with systematic and random errors.
3. Apply the quality control algorithms described above to the simulated dataset, and determine
   - The best homogenization algorithms for sub-daily data in the presence of random errors
   - The best sequencing in the presence of both random and systematic errors
4. Repeat steps 2 and 3 1,000 times for each of 10 different radiosonde stations and 3 pressure levels.

Modeling Radiosonde Data- Part 1

1. In order to capture seasonal and hourly trends, I plan on fitting a Generalized Additive Model (GAM) to the radiosonde temperature data.
2. GAMs are flexible, non-parametric models that allow the response variable to be a linear combination of smoothed functions of the input variables, see Hastie and Tibshirani (1990).
3. The particular model I'll fit is

$$t_i = \beta_0 + s_1(h_i) + s_2(d_i) + \beta_1 y_i + \epsilon_i, \tag{3}$$

   where t_i is the temperature at a given station and pressure level; h_i, d_i and y_i are the hour, day, and year of the i-th observation, respectively; β_0 is the intercept; β_1 is the coefficient for the long term trend; and $s_1(\cdot)$ and $s_2(\cdot)$ are cubic regression splines.

Modeling Radiosonde Data- Part 2

1. Typically the error term, ϵ_i, in the GAM model would be modeled as normal with some unknown variance, but the distribution of the error terms could be skewed or have heavier tails than a normal distribution.
2. Thus, I'll use a skew-t distribution for the errors of this model, which has 4 parameters, ξ, σ, α, and ν which are useful in controlling the first four moments of the distribution, see Azzalini and Capitanio (2003).

Modeling Radiosonde Data- Part 3

1. Additionally, I expect there to be temporal correlation in the error terms.
2. Since I have already included hourly and seasonal terms in the model, I expect most of this autocorrelation to be explained, and so an AR(1) time series model is sufficient to account for the remaining structure in the residuals.
3. For radiosonde data, observations are not equally spaced in time: Launches are scheduled globally at 0 and 12 UTC, but many deviations from this pattern are observed.
4. Thus, to estimate the lag-h autocorrelation, $\phi(h)$, in hours, I must use only those observations that are h time steps apart:

   $$\widehat{\phi}(h) = \frac{1}{|\mathcal{P}_h|} \sum_{(\widehat{\epsilon_i},\widehat{\epsilon_j})\in\mathcal{P}_h} \frac{(\widehat{\epsilon_i}-\bar{\epsilon_i})(\widehat{\epsilon_j}-\bar{\epsilon_j})}{\sqrt{s_{\epsilon_i}s_{\epsilon_j}}}, \tag{4}$$

   where $\mathcal{P}_h$ is the set of all pairs of residuals that are h hours apart (or within some window), and $\widehat{\epsilon_i}$ is the observed residual from Equation (3).
5. For an AR(1) model, I need only estimate $\phi(\cdot)$ at $h = 12$ hours, and I plan on using a window of 5% of 12 hours, or 0.6 hours.

Performance Metrics

- **True Positive Rate:** The proportion of simulated errors correctly detected by the quality control algorithm.
- **False Positive Rate:** The proportion of valid data points incorrectly identified as errors by the quality control algorithm.
- **Efficiency:** Let $\mathbf{x}$, $\mathbf{c}$, and $\mathbf{h}$ be the original, contaminated, and contaminated and homogenized time series, respectively and let the i-th observation be denoted by x_i, c_i, and h_i respectively. The Root Mean Square Error (RMSE) of $\mathbf{h}$ is then defined as follows:

  $$\mathrm{RMSE}(\mathbf{h}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h_i - x_i)^2}.$$

  Then, the efficiency of the homogenized series, where 1 means perfect skill, 0 means no improvement, and negative values indicate degradation is

  $$\mathrm{Eff}(\mathbf{h}) = \frac{\mathrm{RMSE}(\mathbf{c}) - \mathrm{RMSE}(\mathbf{h})}{\mathrm{RMSE}(\mathbf{c})}.$$

Summary of Proposed Study

For the defense, a summary of the simulation study will be given along with conclusions on the performance of the various homogenization algorithms and sequencings. Additionally, the recommended method will be applied to an actual dataset as a case study.

References

A. Anderson, J. M. Browning, J. Comeaux, A. S. Hering, and D. Nychka. A comparison of automated statistical quality control methods for error detection in historical radiosonde temperatures. *Submitted*, 2014.

A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 367–389, 2003.

P. Domonkos. Measuring performances of homogenization methods. *Quarterly Journal of the Hungarian Meteorological Service*, 117(1): 91–112, 2013.

R. E. Eskridge, O. A. Alduchov, I. V. Chernykh, Z. Panmao, A. C. Polansky, and S. R. Doty. A comprehensive aerological reference data set (CARDS): Rough and systematic errors. *Bulletin of the American Meteorological Society*, 76(10): 1759–1775, 1995.

L. Haimberger. Homogenization of radiosonde temperature time series using innovation statistics. *Journal of Climate*, 20(7): 1377–1403, 2007.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990.

R. Killick and I. A. Eckley. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3): 1–19, 2014. URL `http://www.jstatsoft.org/v58/i03/`.

R. Killick, P. Fearnhead, and I. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500): 1590–1598, 2012.

J. R. Lanzante. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *International Journal of Climatology*, 16 (11): 1197–1226, 1996.

J. R. Lanzante, S. A. Klein, and D. J. Seidel. Temporal homogenization of monthly radiosonde temperature data. part I: Methodology. *Journal of Climate*, 16(2): 224–240, 2003.

A. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974. doi: 10.2307/2529204.

V. K. Venema, O. Mestre, E. Aguilar, I. Auer, J. A. Guijarro, P. Domonkos, G. Vertacnik, T. Szentimrey, P. Stepanek, P. Zahradnicek, et al. Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8(1): 89–115, 2012.