The outlier detection problem for radiosondes

Liangliang Wang
University of British Columbia
Email: l.wang@stat.ubc.ca

August 13, 2008
Outline

Description of the data

Proposed methods
Outline

Description of the data

Proposed methods
Radiosonde balloons
(photo source: US National Weather Service)
Radiosonde

- A radiosonde consists of instruments that are launched from the surface by balloon and carried through the atmosphere into the stratosphere.
- Temperature, water vapor, wind speed and wind direction and pressure are measured at different heights above the surface.

Data from NCAR data support section (DSS)

- There are 40-60 million unique soundings distributed over 1500 locations and over the period 1920-2007.
### Fields of the original data

**Station ID:** 1001 ; (1, 161, 392 observations)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unit</th>
<th>Missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 station id</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>2 year</td>
<td>4 digit</td>
<td>no</td>
</tr>
<tr>
<td>3 month</td>
<td>2 digit</td>
<td>no</td>
</tr>
<tr>
<td>4 day</td>
<td>2 digit</td>
<td>no</td>
</tr>
<tr>
<td>5 hour</td>
<td>2 digit</td>
<td>no</td>
</tr>
<tr>
<td>6 pressure</td>
<td>hP</td>
<td>no</td>
</tr>
<tr>
<td>7 Geopotential</td>
<td>meters</td>
<td>398,274 (34.29%)</td>
</tr>
<tr>
<td>8 <strong>Temperature</strong></td>
<td>degrees K and tenths</td>
<td>326,170 (28.08%)</td>
</tr>
<tr>
<td>9 Dew point</td>
<td>degrees K and tenths</td>
<td>503,847 (43.38%)</td>
</tr>
<tr>
<td>10 Wind Direction</td>
<td>degrees</td>
<td>441,379 (38.00%)</td>
</tr>
<tr>
<td>11 Wind Speed</td>
<td>m/s and tenths</td>
<td>441,286 (38.00%)</td>
</tr>
</tbody>
</table>
The outlier detection problem for radiosondes

Description of the data

Temperature vs. pressure of the first 40 unique time points
The outlier detection problem for radiosondes

Description of the data

- Use year and fraction of year:
  - e.g., September 1, 1957, Hour 0 → 1957.668.
- There are 28,788 unique time points;
  - Typically 1 or 2 radiosonde every day
- The range of pressure is 0 - 1,106 hP
- The range of temperature is 173.3 - 401.1 K
How the shape of the radiosonde changes with time (Year 1958)
The temperature cycle
Goals

For DSS

- to assemble a single consistent data base for all available radiosonde measurements.

For statisticians

- to determine objective ways of detecting unusual observations that can be due to systematic biases or random problems.
Outline

Description of the data

Proposed methods
Proposed methods

- Robust principal components analysis
  - Median-centered spherical PCA (Locantore, Marron, Simpson, Tripoli, Zhang, and Cohen 1999); (Gervini 2007)
  - Functional PCA through conditional expectation (PACE) (Yao, Muller, and Wang 2004)
- A 2-d (3-d) thin plate partial smoothing spline
  - As a project for this summer school
  - Using R packages: fields, ncdf
The outlier detection problem for radiosondes
modelling the discontinuity in the derivative for each radiosonde curve
Wahba’s method (1986)

Shiau, Wahba, and Johnson (1986)

Fix one curve: data \((p_j, x(p_j)), j = 1, \cdots, n,\)
- \(p_j\) is the pressure value,
- \(x(p_j)\) is the temperature value.

Model: \(x(p_j) = f(p_j) + \epsilon_j, j = 1, 2, \cdots, n.\) In a partial spline \(f\)
is modeled as

\[
f(p) = g(p) + \theta|p - p^*|
\]

- \(p^*\) = point of discontinuity in derivative (tropopause).
- Tropopause is at a known pressure value.
A partial spline estimate of $f$ is obtained by minimizing

$$\frac{1}{n} \sum_{j=1}^{n} \{ x(p_j) - g(p_j) - \theta |p_j - p^*| \}^2 + \lambda \int_{-\infty}^{\infty} [g^{(2)}(p)]^2 dp.$$ 

Partial spline models can be fitted using the `ssanova` in the R package `gss` through the specification of an optional argument `partial` (Gu 2002).
The outlier detection problem for radiosondes
modelling the discontinuity in the derivative for each radiosonde curve
Wahba’s method (1986)

1958.2137
A parametric model for one curve

Two connected parabolas are used to fit each radiosonde curve:

\[
x(p) = (\beta_0 + \beta_1 p + \beta_2 p^2) \cdot 1(p \leq p^*) + (\alpha_0 + \alpha_1 p + \alpha_2 p^2) \cdot 1(p > p^*),
\]

\[
\alpha_0 = \beta_0 + \beta_1 p^* + \beta_2 p^{*2} - \alpha_1 p^* - \alpha_2 p^{*2}.
\]

where

- \( p \) is the pressure level;
- \( x(p) \) is the corresponding temperature;
- \( p^* \) is the change point (tropopause) of the curve; the two parabolas are connected at \( p^* \);
- \( \beta_0, \beta_1, \beta_2, \alpha_0, \alpha_1, \alpha_2 \) are parameters;
- \( 1 \) is the indicator function.
How to choose the tropopause?

- I looked at the difference of two successive ratios of temperature to pressure:

\[
\begin{align*}
    r_1(p_j) &= \frac{x(p_{j+1}) - x(p_j)}{p_{j+1} - p_j} \\
    r_2(p_j) &= \frac{x(p_j) - x(p_{j-1})}{p_j - p_{j-1}} \\
    \Delta r(p_j) &= r_1(p_j) - r_2(p_j)
\end{align*}
\]

- Better methods?
The outlier detection problem for radiosondes
- modelling the discontinuity in the derivative for each radiosonde curve
- A parametric model for one curve

One good fit
The outlier detection problem for radiosondes
- modelling the discontinuity in the derivative for each radiosonde curve
- A parametric model for one curve

One bad fit
The general idea

- estimate the mean function $\mu(p)$;
- estimate the covariance function $G(s, p)$;
- functional PCA
  - The spherical principal components: see Locantore et al. (1999); Gervini (2007)
  - PACE: see Yao, Muller, and Wang (2004)
- use the first $K$ PCs to approximate curves
- outlier detection.
PCA is used to approximate curves using few parameters.

\[ \hat{x}_i(p) = \hat{\mu}(p) + \sum_{j=1}^{K} \xi_{ij} \phi_j(p). \]

- \( p \): the pressure value
- \( x_i(p) \): the temperature value at \( p \)
- \( \xi_{ij} \): principal components cores
- \( \phi_j(p) \): principal component function
Several ways:

- Plot pc scores
- $L^2$ type error

\[
\text{ERROR1} = \frac{\sum_{j=1}^{n_i} (\hat{x}_i(p_{ij}) - x_i(p_{ij}))^2}{n_i}
\]

\[
\text{ERROR2} = \frac{\sum_{j=1}^{n_i} (\hat{x}_i(p_{ij}) - \hat{\mu}(p_{ij}))^2}{n_i},
\]

where $\hat{\mu}$ is the estimated mean mean curve.

- Correlation
  - CORR1 = $\text{corr}(\hat{x}_i, x_i)$
  - CORR2 = $\text{corr}(\hat{x}_i, \hat{\mu})$
Some challenges

- Some PC methods require curves measured at common points; e.g., Gervini (2007)
- Some radiosondes are “short”
Median-centered spherical PCA

- Locantore, Marron, Simpson, Tripoli, Zhang, and Cohen (1999); Gervini (2007)
- The functional median $\tilde{\mu}(p)$
- The spherical principal components
  - $X(p)$ is projected onto the sphere:
    $\tilde{X}(p) = \frac{X(p) - \tilde{\mu}(p)}{\|X(p) - \tilde{\mu}(p)\|}$.
- The spherical covariance function
  $\tilde{G}(s, p) = \text{cov}(\tilde{X}(s), \tilde{X}(p))$,
  is used in the functional eigen-analysis.
The outlier detection problem for radiosondes

Functional PCA through conditional expectation (PACE)

The PACE model

See Yao, Muller, and Wang (2004)

\[ Y_{ij} = X_i(p_{ij}) + \epsilon_{ij} = \mu(p_{ij}) + \sum_{k=1}^{\infty} \xi_k \phi_k(p_{ij}) + \epsilon_{ij}, \quad p_{ij} \in \mathcal{T}, \]

- \( Y_{ij} \): observation for the \( i \)th subject at the pressure value \( p_{ij}, i = 1, \cdots, n, j = 1, \cdots, N_i; \)
- Measurement error \( \epsilon_{ij} \sim N(0, \sigma^2); \)
- Covariance function \( G(s, p) = \text{cov}(X(s), X(p)) \)
The outlier detection problem for radiosondes

Functional PCA through conditional expectation (PACE)

- The mean function $\mu(p)$ is estimated based on the pooled data from all individuals by a local linear smoother;
- The covariance surface $G(s, p)$ is estimated via the local linear surface smoother using the "raw" covariance

$$G_i(p_{ij}, p_{il}) = (Y_{ij} - \hat{\mu}(p_{ij}))(Y_{il} - \hat{\mu}(p_{il})), j \neq l$$

- The variance of the measurement errors, $\sigma^2$,

$$\hat{\sigma}^2 = \frac{1}{|T|} \int_T \{ \hat{V}(p) - \tilde{G}(p) \} dp,$$

where $\hat{V}(p)$ is the estimate for $\{ G(p, p) + \sigma^2 \}$, and $\tilde{G}(p)$ is the estimate for $\{ G(p, p) \}$. 
Eigenfunctions $\hat{\phi}_k$ and eigenvalues $\hat{\lambda}_k$ are estimated by solving the eigenequations as follows,

$$\int_L \hat{G}(s, p) \hat{\phi}_k(p) dp = \hat{\lambda}_k \hat{\phi}_k(p),$$

where $\hat{G}(s, p)$ is the smoothed covariance surface.

Estimates for the FPC scores $\xi_{ik}$:

$$\tilde{\xi}_{ik} = E[\xi_{ik} | Y_{i1}, \cdots, Y_{iN_i}].$$
Prediction for individual curves

- The curve $X_i(p)$ for the $i$-th subject is approximated with the first $K$ eigenfunctions:

$$\hat{X}_i^K(p) = \hat{\mu}(p) + \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\phi}_k(p).$$

- Note: PACE has no trouble with short curves or non-common pressure values
Curves from the 941st time point to the 970th time point with short curves
The outlier detection problem for radiosondes

Some illustrations

PACE

Predicted curves via PACE (in black solid lines)
The outlier detection problem for radiosondes

Some illustrations

The first 5 PCs via PACE
The mean curve and the effects of adding and subtracting a suitable multiple of each PC via PACE
A toy data set

two types of outliers: different curve shape; position shift.

The blue lines represent 28 curves from June 1980. The red lines are two curves from December 1980.
The outlier detection problem for radiosondes

- Some illustrations
- A toy data set

Pairs of PC scores via PACE

![Pairs of PC scores via PACE](image)
Outliers by pairs of PC scores via PACE

- picked up some curves with spikes;
- picked up curves with a certain shift.
The outlier detection problem for radiosondes

- Some illustrations
- A toy data set

ERROR1 via PACE
Outliers by ERROR1 via PACE

- picked up all curves with spikes;
- didn’t pick up curves with a certain shift.
1-CORR1 vs. ERROR1 via PACE
Outliers by 1-CORR1 vs. ERROR1 via PACE

- picked up all curves with spikes;
- didn’t pick up curves with a certain shift.
- be consistent with the results using ERROR1 only.
1-CORR2 vs. ERROR2 via PACE
Outliers by 1-CORR2 vs. ERROR2 via PACE

- picked up some curves with spikes;
- picked up curves with a certain shift.
Reference


