Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves

Piotr Kokoszka
Department of Statistics, Colorado State University

Joint work with

Oleksandr “Sasha” Gromenko
Department of Mathematics and Statistics, Utah State University

Jan Sojka
Department of Physics, Utah State University
Electron density profile

Height, km

EXOSPHERE

THERMOSPHERE

MESOSPHERE

STRATOSPHERE

TROPOSPHERE

Electron density, cm$^{-3}$

Ionospheric cooling trend
Figure: foF2 records in the northern hemisphere.
Figure: Locations of the 85 ionosonde stations used in this study, black discs. The two circles in northern Canada represent stations located in the auroral zone (dashed line), which were not used.
Let \( X(s; t) \) be a value of foF2 at time \( t \in [0, 1] \) and location \( s \). We postulate the functional spatio–temporal model

\[
X(s; t) = \mu(t) + \varepsilon(s; t) + \theta(s; t)
\]

where \( E\varepsilon(s_k, \cdot) = 0 \), and \( E\theta(s; t) = 0 \)

**Question:** Does the mean function \( \mu \) contain a linear trend?
**Assumption 1:** $\varepsilon(s_k, \cdot)$ is an element of a strictly stationary isotropic zero mean spatial random field taking values in the space $L^2$. This term admits the Karhunen-Loève expansion:

$$\varepsilon(s; t) = \sum_{j=1}^{\infty} \xi_j(s)v_j(t),$$

where $v_j(t)$ are the functional principal components (FPC’s), and $\xi_j(s)$ are the corresponding scores, $\xi_j(s) = \langle X(s) - \mu, v_j \rangle$.

**Assumption 2:** $E[\xi_j(s_k)\xi_{j'}(s_{\ell})] = \delta_{jj'}\Sigma_j(s_k, s_{\ell})$.

**Assumption 3:** $\theta(s; t)$ are iid random variables independent of $\varepsilon(s; t)$, $\text{Var}(\theta(s; t)) = \sigma_\theta^2$
Justification of stationarity and isotropy

GAIM, day 74, 2003

Kriging, day 74, 2003

Ionospheric cooling trend
Justification of independence of scores

\[ \xi_1 - \xi_2 \]

\[ \xi_1 - \xi_3 \]

\[ \xi_2 - \xi_3 \]
Estimation of the mean

For complete records:

\[ \hat{\mu}(t) = \sum_{k=1}^{N} w_k X(s_k; t), \quad \sum_{k=1}^{N} w_k = 1, \]


For incomplete records: local linear indexed regression (LLIR)

\[
(\hat{m}_0(t), \hat{m}_1(t)) = \arg \min_{m_0, m_1} \sum_{i=1}^{T} \kappa_{\mu} \left( \frac{t - t_i}{h_\mu} \right) \left\{ \sum_{k=1}^{N_i} w_k(t_i) X(s_k; t_i) - m_0 - m_1(t - t_i) \right\}^2,
\]

where

- \( \hat{m}_0(\cdot) \) – estimate of the mean function \( \mu(\cdot) \)
- \( T \) – total number of temporal observations in a complete record
- \( N_i \) – number of available observations at time \( t_i \)
- The weights \( w_k(t_i) \) account for the spatial correlation
Estimation of the weights

Define a functional variogram:

\[ 2\gamma(d_{k\ell}) = E \left\{ \int (X(s_k; t) - X(s_\ell; t))^2 dt \right\}. \]

For incomplete records numerical integration can be a source of a severe bias especially for short records. And thus should be avoided!

**Step 1:** Obtain the preliminary estimator

\[ 2\tilde{\gamma}(d_{k\ell}) = \frac{1}{p_{k\ell}} \sum_{P(d_{k\ell})} (X(s_k; t_i) - X(s_\ell; t_i))^2, \]

where \( P(d_{k\ell}) = \{(s_k, s_\ell) : \|s_k - s_\ell\| = d_{k\ell}\} \) and \( p_{k\ell} \) is the cardinality of \( P(d_{k\ell}) \). The points with \( d = 0 \) are not included.

**Step 2:** Fit \( \tilde{\gamma}(d_{k\ell}) \) to some valid parametric model

\[ \gamma(d) = (\sigma^2 - \sigma_\nu^2)(1 - \exp(-d^2/\rho^2)) + \sigma_\nu^2 1_{(0,\infty)}(d). \]

**Step 3:** Calculate the weights

\[ w(t_i) = \Sigma(t_i)^{-1}1/(1^T \Sigma(t_i)^{-1}1), \]

where \( \Sigma(t_i) \) is the \( N_i \times N_i \) covariance matrix for available observations at time \( t_i \).
Figure: Estimation of the weights for incomplete records. Left panel: Gray dots represent all available squared differences \((X(s_k; t_i) - X(s_\ell; t_i))^2\), \(1 \leq i \leq T\); black dots represent squared differences \((X(s_k; t_i) - X(s_\ell; t_i))^2\), for some fixed \(t_i\). Dashed lines separate regions \(P(d_{k\ell})\). Right panel: The thin line shows the estimated variogram, the bold line represents the fitted Gaussian variogram.
Estimation of the covariance function

Define

\[ c(t, t') = \text{Cov}(X(t), X(t')). \]

For complete records the estimation procedure is proposed in Gromenko, et al. (2012). But it does not work for incomplete records...

**Step 1:** Obtain the preliminary estimator \( \tilde{c}(t_i, t_{i'}) \). For fixed \( i \) and \( i' \) define the spatial scalar field \( \psi(s) = [X(s; t_i) - \hat{\mu}(t_i)] [X(s; t_{i'}) - \hat{\mu}(t_{i'})] \). Then

\[ \tilde{c}(t_i, t_{i'}) = E\psi(s) \]

This step is computationally very expensive!

**Step 2:** Smoothing

\[ \hat{u} = \arg \min_u \sum_{1 \leq i \neq i' \leq T} \kappa_c \left( \frac{t - t_i}{h_c}, \frac{t' - t_{i'}}{h_c} \right) \left\{ \tilde{c}(t_i, t_{i'}) - f(u, t, t', t_i, t_{i'}) \right\}^2, \]

where

- \( f(u, t, t', t_i, t_{i'}) = u_0 + u_1(t - t_i) + u_2(t' - t_{i'}), \ u = [u_0, u_1, u_2]^T \)
- \( \hat{u}_0(t, t') - \) estimate of the covariance function \( c(t, t') \)

In the presence of measurement error, the diagonal elements are contaminated by the noise variance \( \sigma_\theta^2 \) and should not be included as input for the smoothing step, Yao, et al. (2005).
Assume that the mean function $\mu(t)$ is a linear combination of $q$ known complete functions, so that

$$X(s; t) = \sum_{i=1}^{q} \beta_i z_i(s; t) + \sum_{j=1}^{\infty} \zeta_j(s) v_j(t) + \theta(s; t).$$

For the ionosonde data we select:

- $z_1(t)$ – constant
- $z_2(t) = t$ – the linear trend parameter
- $z_3(t)$ – the observed solar radio flux (SRF) measured in W/m$^2$Hz
- $z_4(s; t) = \sin I(s; t) \cos I(s; t)$, $I(s; t)$ – inclination of the Earth’s magnetic field

**Question:** Is the coefficient $\beta_2$ statistically significant?
Estimation of the regression coefficients

If the responses $X(s_k)$ are fully observed, Gromenko and Kokoszka (2013):

$$\hat{\beta} = \arg \min_{\beta} \left\| \sum_{k=1}^{N} w_k \left\{ X(s_k) - \sum_{i=1}^{q} z_i(s_k) \beta_i \right\} \right\|^2, \text{ subject to } \sum_{k=1}^{N} w_k = 1.$$ 

This leads to the solution

$$\hat{\beta} = Q^{-1} \langle z, w^T X \rangle = Q^{-1} \langle z, \hat{\mu} \rangle,$$

where

- $z = [z_{w1}, \ldots, z_{wq}]^T$
- The quantity $\langle z, \hat{\mu} \rangle$ is the $q \times 1$ vector with the $i$th entry $\langle z_{wi}, \hat{\mu} \rangle$
- $Q = [\langle z_{wi}, z_{wi'} \rangle, 1 \leq i, i' \leq q]$
- $z_{wi}(t) = \sum_{k=1}^{N} w_k z_i(t; s_k)$

For incomplete records estimate the mean function using (LLIR) and plug it into the equation above.
Significance of the regression coefficients

\[ \text{P-value} = 2\Phi(\beta_i / \sqrt{\text{Var}[\hat{\beta}_i]}). \]

Exact formula for the variance: 
\[ \text{Var}[\hat{\beta}] = Q^{-1}E \left[ (z, \hat{\mu} - \mu) (z, \hat{\mu} - \mu)^T \right] Q^{-1}, \]
where the middle term is a \( q \times q \) matrix whose \((i, j)\) element is
\[ \int \int z_{wi}(t)z_{wj}(t') \text{Cov}(\hat{\mu})(t, t')dt dt'. \]

The formula for \( \text{Cov}(\hat{\mu})(t, t') \) is derived in Gromenko, et al. (2013).

Approximation:
\[ \text{Var}[\hat{\beta}] = Q^{-1}\Omega \text{Var}[\zeta w] \Omega^T Q^{-1} + \sigma^2 Q^{-1}w^T w, \]
where, assuming that the weights are known constants,
\[ \text{Var}[\zeta w] = \text{diag} \left( w^T \Sigma_1 w, \ldots, w^T \Sigma_p w \right). \]

The last expression is a \( p \times p \) diagonal matrix. The number \( p \) of the FPC’s is typically selected to capture about 85-90% of the variance.
Table: Trends and P-values for different bandwidths. “NM” denotes estimation without magnetic inclination, $z_4$, “M” denotes estimation when magnetic inclination is included. The number of the FPC’s, $p$, was chosen to obtain the cumulative variance closest to but greater than 85% (indicated as Final CV).

<table>
<thead>
<tr>
<th>$h_\mu$</th>
<th>$h_C$</th>
<th>Final CV,%</th>
<th>$p$</th>
<th>NM $\beta_2, 10^{-3}$MHz/Year</th>
<th>P-value</th>
<th>M $\beta_2, 10^{-3}$MHz/Year</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10</td>
<td>86.77</td>
<td>3</td>
<td>$-5.22 \pm 2.25$</td>
<td>0.020</td>
<td>$-4.91 \pm 2.24$</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>90.44</td>
<td>3</td>
<td>$-5.22 \pm 2.30$</td>
<td>0.023</td>
<td>$-4.91 \pm 2.29$</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>87.07</td>
<td>2</td>
<td>$-5.22 \pm 1.48$</td>
<td>$4.36 \cdot 10^{-4}$</td>
<td>$-4.91 \pm 1.47$</td>
<td>$8.62 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>88.75</td>
<td>4</td>
<td>$-5.18 \pm 2.15$</td>
<td>0.016</td>
<td>$-4.88 \pm 2.14$</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>87.57</td>
<td>3</td>
<td>$-5.18 \pm 2.11$</td>
<td>0.014</td>
<td>$-4.88 \pm 2.10$</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>86.50</td>
<td>2</td>
<td>$-5.18 \pm 1.83$</td>
<td>$4.67 \cdot 10^{-3}$</td>
<td>$-4.88 \pm 1.82$</td>
<td>$7.52 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>88.55</td>
<td>4</td>
<td>$-5.28 \pm 2.16$</td>
<td>0.014</td>
<td>$-4.99 \pm 2.15$</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>87.09</td>
<td>3</td>
<td>$-5.28 \pm 2.11$</td>
<td>0.012</td>
<td>$-4.99 \pm 2.11$</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>91.36</td>
<td>3</td>
<td>$-5.28 \pm 2.20$</td>
<td>0.016</td>
<td>$-4.99 \pm 2.19$</td>
<td>0.023</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>88.58</td>
<td>4</td>
<td>$-5.07 \pm 2.13$</td>
<td>0.017</td>
<td>$-4.79 \pm 2.12$</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>87.02</td>
<td>3</td>
<td>$-5.07 \pm 1.99$</td>
<td>0.011</td>
<td>$-4.79 \pm 1.98$</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>91.04</td>
<td>3</td>
<td>$-5.07 \pm 2.04$</td>
<td>0.013</td>
<td>$-4.79 \pm 2.03$</td>
<td>0.018</td>
</tr>
</tbody>
</table>
Figure: (a) Distribution of the trend parameter as a function of the interval length, (b) the number of stations as a function of the interval length. (c) distribution of the trend parameter as a function of the number of stations. Light gray - full region, dark gray - central 50 percent, dotted line - median, blue solid line - average. Bold black line is the final estimator.


Danke für die Aufmerksamkeit !