Serge GUILLAS
Department of Statistical Science, University College London

Trends in stratospheric ozone profiles using functional mixed models

février 2016

Trends in stratospheric ozone profiles using functional mixed models

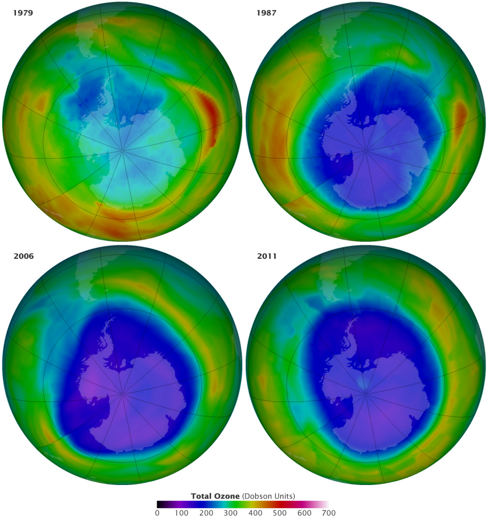
Serge Guillas

University College London
Department of Statistical Science

Park, A., Guillas, S., Petropavlovskikh, I., *Atmospheric Chemistry & Physics*, 13, 11473-11501, 2013

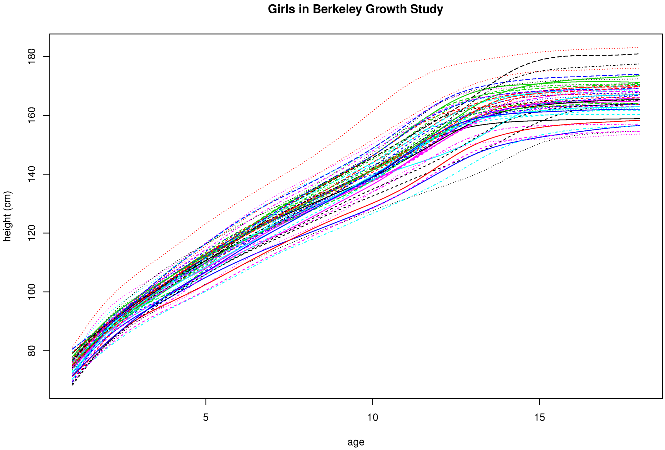
Outline

- Functional Data Analysis (FDA)
- Functional Principal Components Analysis (FPCA)
- mixed models for ozone trends



Functional Data Analysis (FDA)

- Branch of statistics dealing with analysis of data in functional forms such as *curves* or *images*.
- Functional data are intrinsically *infinite dimensional* and exhibit *high level of correlation* (Ramsay and Silverman, 2005).



FDA

- Two school of thoughts
  1. smoothing school
     - Consider each sample as a *smooth function*.
     - Conversion of discrete data into smooth functions using various approaches, e.g. *B-spline basis expansion, bivariate splines* (Guillas and Lai, 2010).
  2. stochastic school
     - View each sample as a set of *stochastic process*:
       $$X = \{X_t, t \in [0, T]\}.$$
- Design of functional data
  1. dense and regular grid
  2. sparse and irregular grid (longitudinal studies)

Notations

- Function space:

$$X \in \mathcal{L}^2[0,T] := \left\{ X, \int_0^T |X(t)|^2 \, dt < \infty \right\}$$

- Inner product:

$$f, g \in \mathcal{L}^2[0,T], \quad < f, g > := \int_0^T f(t) g(t) dt$$

- Norm:

$$\|f\|^2 := < f, f >$$

Notations: generalization of covariance matrix

- For $X \in \mathcal{L}^2[0,T]$ with $\mathsf{E}_X(t) = 0$ and $t \in [0,T]$, denote the covariance function by C_X:

$$\mathsf{C}_X(s,t) = \mathsf{E}[X(s)X(t)],$$

and compute its estimate as

$$\hat{\mathsf{C}}_X(s,t) = \frac{1}{n}\sum_{i=1}^{n} x_i(s)x_i(t).$$

- When $\int_0^T \int_0^T \mathsf{C}_X^2(s,t)dsdt < \infty$, write a covariance operator Γ_X of $X \in \mathcal{L}^2$, which maps $f(t)$ to $f(s)$ as

$$(\Gamma_X f)(s) = \int_0^T \mathsf{C}_X(s,t)f(t)dt.$$

Representation of functional data

- Let $x_i(t)$ be the ith underlying true function, observed at finite and dense grid points, $\{t_j, j = 1, .., N\}$. In practice, the observations y_{ij} include measurement errors:

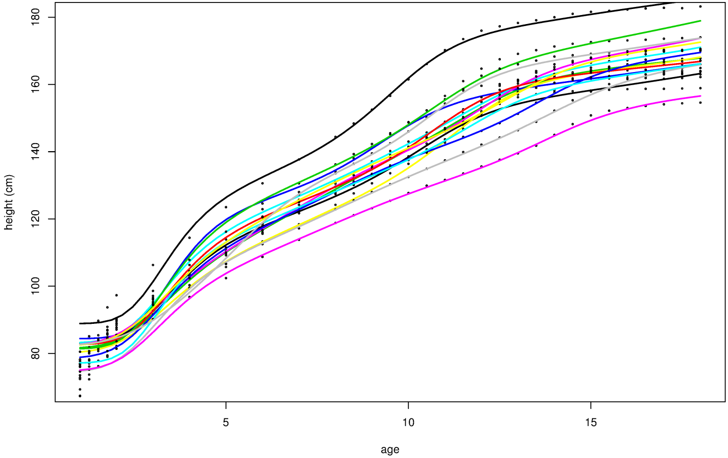
$$y_{ij} = x_i(t_j) + \epsilon_{ij}, \quad \forall t_j \in [0, T].$$

- Expand $x_i(t)$ in B-spline basis $\phi_k(t)$:

$$x_i(t) \ = \ \sum_{k=1}^{K} c_{ik}\phi_k(t),$$

  where c_{ik} are the associated B-splines coefficients.
- *Splines* are piecewise polynomials with the polynomial pieces joining at *knots*. To define spline basis system, we have to decide
  1. degree of polynomials
  2. location and the number of knots, or functions via K. K can control the level of smoothing.

Representation of functional data



Estimation of c_{ik} based on least squares fitting criterion

- In *regression splines*, selection of K plays a crucial role:

$$\min_{\mathbf{c}_i} \sum_{j=1}^{N} [y_{ij} - \sum_{k=1}^{K} c_{ik}\phi_k(t_j)]^2,$$

  but is computationally expensive.

- In *penalized splines*, K is chosen to be large enough to capture the maximum complexity, but the use of penalization controls the excessive variations:

$$\min_{\mathbf{c}_i} \left(\|\mathbf{y}_i - \mathbf{\Phi}\mathbf{c}_i\|^2 + \lambda \mathbf{c}_i^T \mathcal{P}^T \mathcal{P} \mathbf{c}_i \right),$$

  where $\mathbf{y}_i = [y_{i1}, .., y_{iN}]^T$, $\mathbf{\Phi}_{jk} = \phi_k(t_j)$ and $\mathbf{c}_i = [c_{i1}, .., c_{iK}]^T$. $\mathcal{P}^T\mathcal{P}$ is penalty matrix, measuring the roughness of $x_i(t)$.

Principal Component Analysis (PCA): multivariate

Transforms $X_1, \ldots, X_p$ into linearly uncorrelated random variables: to select the first few modes of variability and maximize variance explained.
First PC, $\mathbf{z}_1$, is solution of

$$\max_{\|\mathbf{z}_1\|=1} \; \mathbf{z}_1^T \mathbf{\Sigma}_X \mathbf{z}_1,$$

where $\mathbf{\Sigma}_X$ is the sample covariance matrix.
Subsequent PCs, $\mathbf{z}_2, \ldots, \mathbf{z}_p$, obtained by solving maximization above under orthogonality constraint: $\mathbf{z}_k^T \mathbf{z}_l = 0$, $\forall k \neq l$.
It is related to eigenvalue decomposition as

$$\mathbf{\Sigma}_X \mathbf{z}_j = \rho_j \mathbf{z}_j,$$

where ρ_j are eigenvalues and $\mathbf{z}_j$ are eigenvectors.

Functional Principal Component Analysis (FPCA)

Similar idea as for multivariate random vectors:

- Expansion of $X \in \mathcal{L}^2[0, T]$, in terms of eigen-functions of Γ_X.
- Tool for dimension reduction, an essential step for FDA.
- Functional Principal Components (FPCs) are often interpreted as *major modes of variation*.

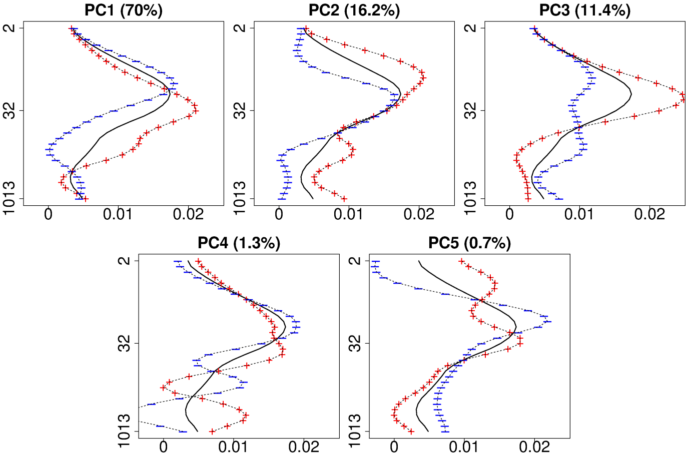


Figure: Mean function & variations induced by smoothed functional PCs

Ozone Trend Analysis

Objectives

- Reveal non-linear effects of time covariates and atmospheric forcings on ozone variations using penalized splines, where each effect is fitted as an additive smooth function.
- Remove the effects of atmospheric influences on ozone and obtain trend estimates more genuinely corresponding to the variations due to the changing emissions of ODSs and GHGs.
- Employ FDA approach to identify more precisely the covariates' effects that vary along different altitudes.
- Allow heteroscedasticity to account for the observed periodidicity in regression errors.

Description of ozone data

(Meiring, 2007)

- Umkehr daily ozone observations as functions of altitude (0-45km, layer 29-60) from 1978-2011 are investigated at Boulder (USA) and Arosa (Switzerland).
- Ozone observations are unequally spaced in time, so the daily records are averaged and monthly means are used.
- Remove observations of two volcanic periods, 1982-1983 (EI Chinchón) and 1991-1993 (Pinatubo).

Ozone as functional data

Denote y_{ij} the altitude-dependent monthly mean ozone at time i and layer j. Then, true but unknown ozone profile, $x_i(a_j)$, is

$$y_{ij} = x_i(a_j) + e_{ij}, \quad a_j \in [0, 45km], \quad e_{ij} \sim N(0, \sigma_y^2),$$

where e_{ij} are i.i.d. observational errors.

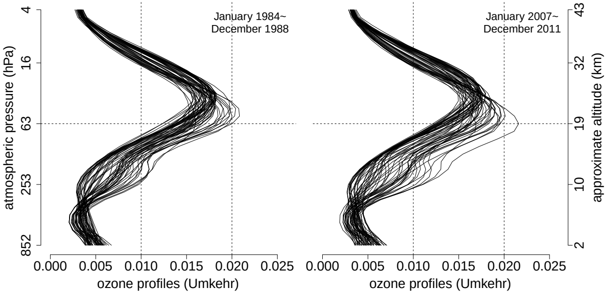


Figure: Boulder: Estimated $x_i(a)$ using smoothing splines.

Covariate data

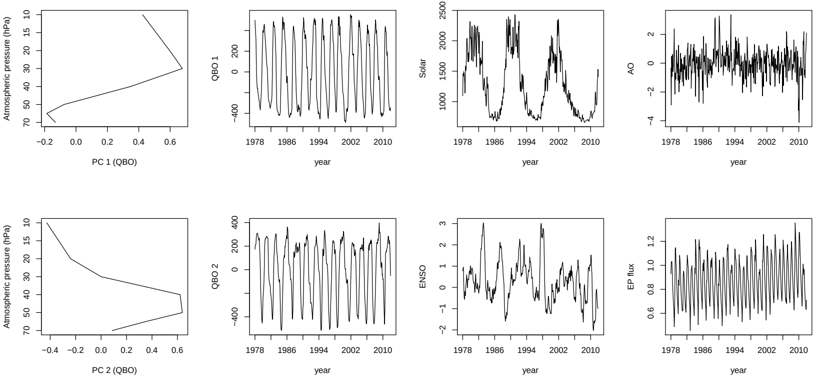


Figure: Atmospheric forcing as covariate data

Two alternative statistical models for trend analysis

- Conventional multivariate approach: fit regression separately for each altitude a_j with autoregressive noise δ_i (Miller et al., 2006)

$$y_{ij} = g_{1j}(m_i) + g_{2j}(yr_i) + \sum_{r=3}^{9} g_{rj}(z_{ri}) + \delta_{ij}.$$

Here, borrowing of information across altitudes is not possible.

- Full functional approach: fit functional regression in one setting as

$$x_i(a) = g_1(m_i, a) + g_2(yr_i, a) + \sum_{r=3}^{9} g_r(z_{ri}, a) + \delta_i(a).$$

Here, the covariate effects smoothly vary along altitudes.

Our 2-step functional approach via dimension reduction

1. dimension reduction (truncated FPCA):

$$x_i(a) \approx \sum_{l=1}^{d} \xi_{il}\zeta_l(a), \quad \xi_{il} = \int \zeta_l(a)x_i(a)da,$$

   where $\zeta_l(a)$ is the lth functional PC and ξ_{il} is its score. $d = 5$.

2. estimation of covariate effects: Additive Mixed Model

$$\xi_{il} = g_{1l}(m_i) + g_{2l}(yr_i) + \sum_{r=3}^{9} g_{rl}(z_{ir}) + \delta_{il}, \delta_{il} \sim N(0, \sigma_{il}^2),$$

$$\log(\sigma_{il}^2) = \delta_{1l}\sin(2\pi\tilde{m}_i) + \delta_{2l}\cos(2\pi\tilde{m}_i), \tilde{m}_i = \mathrm{m}_i/12.$$

   g_{rl} are fitted using penalized splines in mixed effects model framework (Wood, 2006).
   Observed annual pattern in errors modeled.

Estimation results: dimension reduction

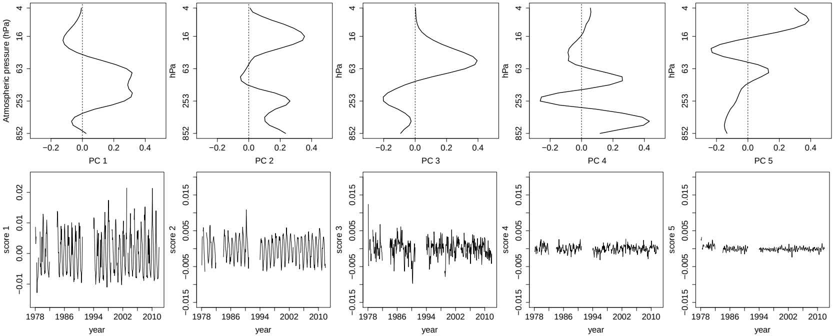


Figure: Smoothed functional PCs (first) and PC scores.

Estimation results: covariate effects

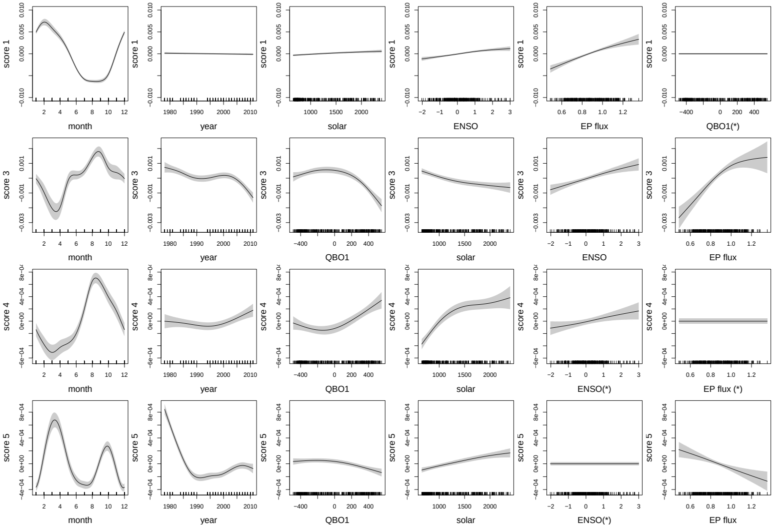


Figure: Boulder: Fitted smooth curves (solid line) and 95% Bayesian credible intervals (shaded areas) for selected scores and covariates.

Trend analysis

Recall the FPCA decomposition.
Using the estimated scores after other effects of covariates are removed, compute the trend at altitude a as

$$O_i(a) = \sum_{l=1}^{d} \zeta_l(a)\hat{g}_{2l}(yr_i), \quad (1)$$

where $O_i(a)$ is the estimated ozone trend at altitude a for year i and $\hat{g}_{2l}(yr_i)$ is lth fitted PC score of year term.

Trend analysis

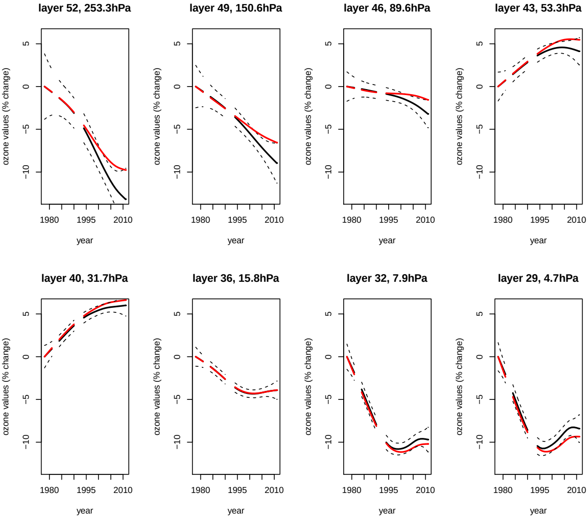


Figure: Arosa: Estimated trends (without EP flux: red)

Trend analysis

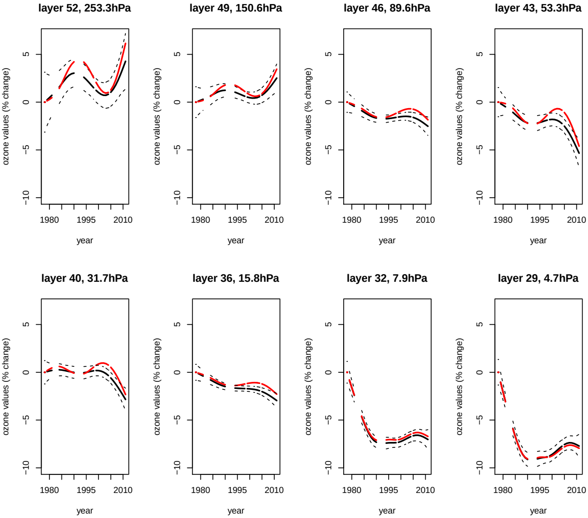


Figure: Boulder: Estimated trends (without EP flux: red)

Conclusions and future work

- Great variations in covariates's effects across altitudes are found: benefits of functional approach
- Our model can capture fine variations in the profiles such as semi-annual-oscillation.
- Using heteroscedastic error structure:
  - more accurate estimates of influences and trends
  - enhanced uncertainty quantification of the estimates (width of confidence intervals)
- To improve the fit we can include short-term-dynamical transport terms.
- Add more stations and incorporate latitudes to borrow strength across stations.

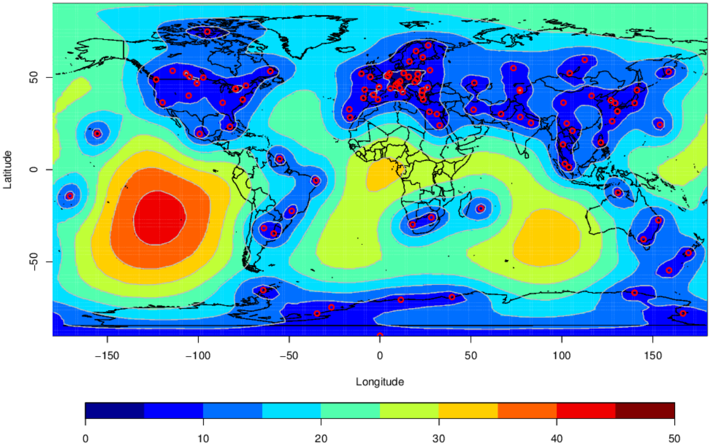


Figure: Uncertainties in estimates of global total column ozone (Chang, Guillas, Fioletov, AMT 2015)

References

Chang, K. L., S. Guillas, and V. E. Fioletov (2015). Spatial mapping of ground-based observations of total ozone. *Atmospheric Measurement Techniques* 8 : 3967-4009.

Guillas, S. and M. J. Lai (2010). Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics* 22, 477-497.

Meiring, W. (2007). Oscillations and time trends in stratospheric ozone levels: A functional data analysis approach. *Journal of the American Statistical Association* 102, 788-802

Miller, A., A. Cai, G. Tiao, D. Wuebbles, L. Flynn, S. Yang, E. Weatherhead, V. Fioletov, I. Petropavlovskikh, X. Meng, S. Guillas, R. Nagatani, and G. Reinsel (2006). Examination of ozonesonde data for trends and trend changes incorporating solar and arctic oscillation signals. *Journal of Geophysical Research* 111, D13305.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer Science Business Media Inc.

Wood, Simon. *Generalized additive models: an introduction with R*. CRC press, 2006.