Doubly Cyclic Smoothing Splines and Analysis of Seasonal Daily Pattern of CO2 Concentration in Antarctica

Mihoko Minami

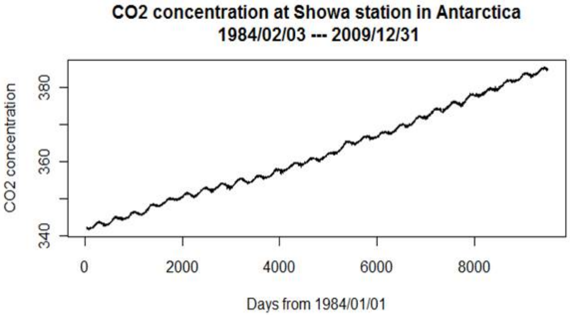
Keio University, Japan

August 15, 2016

Joint work with Ryo Kiguchi

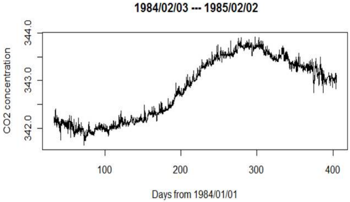
CO2 Data were provided by National Institute of Polar Research

Hourly observations of CO2 concentration at Syowa station in Antarctica (1984/ 2/ 3 - 2009/12/31)

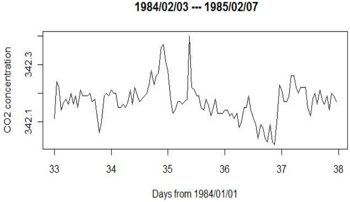


25+ years

(ppm)



1year



5days

Hourly observations of CO2 concentration at Syowa station in Antarctica

The data clearly show

- strong temporal trend
- strong seasonal variation

We are also interested to know if there is

- **a daily pattern ?** If so, does it **vary seasonally** ?
- an effect due to wind speed? If so, does it vary seasonally ?
- an effect due to wind direction?

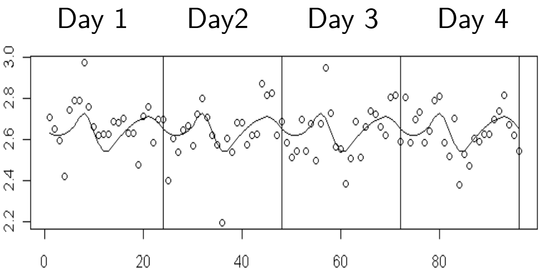
In this talk, we introduce the method to analyze a seasonal daily pattern, **doubly cyclic smoothing splines**, and show the results of analysis that give answers to the above questions.

Outline

1. Cyclic cubic smoothing splines

Cyclic cubic smoothing splines

- **The cyclic cubic smoothing spline** is a smoothing method to estimate periodic variation such as daily or annual pattern of time series observations.



- It fits **a cyclic cubic spline function** which is a periodic piece-wise cubic function with continuity up to the second derivative.

A cyclic cubic spline function

A cyclic cubic spline function $g(t)$ is

- **periodic** When the period is T,

$$g(t + kT) = g(t) \quad \text{for} \quad k = 0, \pm 1, \pm 2, \cdots$$

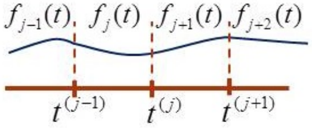
- **piece-wise cubic polynomial**
  Given knots $t^{(0)} < t^{(1)} < \cdots < t^{(K-1)} < t^{(K)}$ with $t^{(K)} - t^{(0)} = T$,

$$g(t) = f_j(t) \quad \text{for} \quad t \in [t^{(j-1)}, t^{(j)}),\ j = 1, 2, \cdots, K$$

  where $f_j(t)$s are cubic polynomial functions.

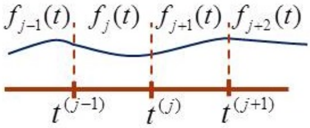
  That is, for $t \in [t^{(0)}, t^{(K)}]$, it can be expressed as

$$g(t) = \sum_{j=1}^{K} I_{[t^{(j-1)}, t^{(j)})} f_j(t)$$



A cyclic cubic spline function

$$g(t) = \sum_{j=1}^{K} I_{[t^{(j-1)}, t^{(j)})} f_j(t)$$



- is also **continuous up to the second derivative**
  For $j = 1, 2, \cdots, K-1$,

$$f_j(t^{(j)}) = f_{j+1}(t^{(j)}),\ f_j'(t^{(j)}) = f_{j+1}'(t^{(j)}),\ f_j''(t^{(j)}) = f_{j+1}''(t^{(j)})$$

- **The values at the both endpoints $t^{(0)}$ and $t^{(K)}$ are equal up to the second derivative.**

$$f_1(t^{(0)}) = f_K(t^{(K)}),\ f_1'(t^{(0)}) = f_K'(t^{(K)}),\ f_1''(t^{(0)}) = f_K''(t^{(K)})$$

Cyclic cubic smoothing splines

A cyclic cubic spline function is flexible.
To avoid overfitting, we impose a roughness penalty.

In a most simplified case, the model and object function are defined as:

Model

$$y_i = g(t_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),\ i = 1, \cdots, n,\ i.i.d.$$

where $g(t)$ is a cyclic cubic spline function.

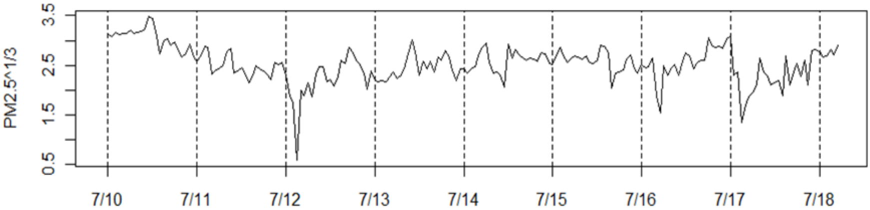
Penalized squared errors

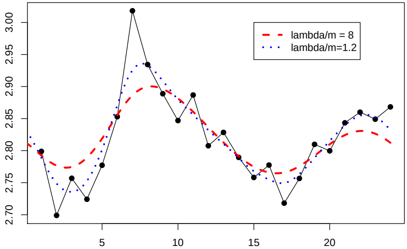
$$\Omega_\lambda(g) = \sum_{i=1}^{n} \{y_i - g(t_i)\}^2 + \lambda \int_{t^{(0)}}^{t^{(K)}} g''(t)^2 dt,\ \lambda > 0 \qquad (1)$$

λ is called a smoothing parameter.

Example: Daily pattern of PM2.5 $^{1/3}$

Hourly observations of PM2.5 $^{1/3}$ in air for 28 days at Fukuoka





In the left plot,

- black dots with solid lines depict hourly averages of PM2.5 $^{1/3}$
- the red dashed curve is the fitted cyclic cubic spline function with $\lambda = 8 \times 28$
- the green dotted curve is the fitted cyclic cubic spline function with $\lambda = 1.2 \times 28$

Smoothing mechanism of cyclic cubic smoothing splines

In the following, we assume knots are evenly spaced with

$$t^{(j)} - t^{(j-1)} = h \quad \text{for} \quad j = 1, 2, \cdots, K$$

Function value parameterization

We employ the function value parameterization to express cyclic cubic functions. For $j = 1, 2, \cdots, K$, let

- β_j be the function value of $g(t)$ at $t^{(j)}$, that is, $\beta_j = g(t^{(j)})$, and
- $b_j(t)$ be the corresponding cyclic cubic spline basis function with $b_j(t^{(i)}) = \delta_{ij}$ for $i = 1, 2, \cdots, K$,

so that a cyclic cubic spline function $g(t)$ can be expressed as

$$g(t) = \sum_{j=1}^{K} \beta_j b_j(t) \tag{2}$$

Penalty term with function value parameterization

The penalty term can be expressed as

$$\int_{t^{(0)}}^{t^{(K)}} g''(t)^2 dx = \boldsymbol{\beta}^T D^T B^{-1} D \boldsymbol{\beta} \tag{3}$$

where

- $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_K)^T$
- B and D are cyclic band matrices,

$$B = \frac{h}{6} G(4, 1) \text{ and } D = \frac{1}{h} G(-2, 1)$$

  where $G(a, b)$ denotes a cyclic band matrix

$$G(a,b) = \begin{pmatrix} a & b & & & b \\ b & a & b & & \\ & \ddots & \ddots & \ddots & \\ & & b & a & b \\ b & & & b & a \end{pmatrix}$$

Least penalized squared error estimate $\hat{\boldsymbol{\beta}}$

Suppose now that

- all observations were made at knots
- at each knot, m observations were obtained.

Let $\overline{y}$ denote the sample average vector at the knots. Then, we have

$$\sum_{i=1}^{n} \{y_i - g(t_i)\}^2 = \|\boldsymbol{y} - \mathbf{1}_m \otimes \overline{\boldsymbol{y}}\|^2 + m\|\overline{\boldsymbol{y}} - \boldsymbol{\beta}\|^2$$

so that the minimization of penalized squared errors is equivalent to the minimization of

$$S(\boldsymbol{\beta}) = m\|\overline{\boldsymbol{y}} - \boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^T D^T B^{-1} D\boldsymbol{\beta} \tag{4}$$

Least penalized squared error estimate

$$\hat{\boldsymbol{\beta}} = H\overline{\boldsymbol{y}} \quad \text{where} \quad H = \left(I_K + \frac{\lambda}{m} D^T B^{-1} D\right)^{-1}$$

Eigenvalues and eigenvectors of $G(a, b)$

For the even number of knots ($K = 2q$), eigenvalues of a cyclic band matrix $G(a, b)$ with $b > 0$ are in descending order

$$l_1 = a + 2b, \quad l_{2j} = l_{2j+1} = a + 2b\cos\frac{2\pi j}{k}, \quad l_{2q} = a - 2b$$
$$(j = 1, \cdots, q-1)$$

and the corresponding eigenvectors are

$$\boldsymbol{u}_1 = \frac{1}{\sqrt{k}}\left(1, 1, \cdots, 1, 1\right)^T,$$

$$\boldsymbol{u}_{2j} = \sqrt{\frac{2}{k}}\begin{pmatrix} \cos\left(2\pi j\frac{1}{k}\right) \\ \vdots \\ \cos\left(2\pi j\frac{i}{k}\right) \\ \vdots \\ \cos\left(2\pi j\right) \end{pmatrix}, \boldsymbol{u}_{2j+1} = \sqrt{\frac{2}{k}}\begin{pmatrix} \sin\left(2\pi j\frac{1}{k}\right) \\ \vdots \\ \sin\left(2\pi j\frac{i}{k}\right) \\ \vdots \\ \sin\left(2\pi j\right) \end{pmatrix}, j = 1, 2, \cdots, q-1$$

$$\boldsymbol{u}_{2q} = \frac{1}{\sqrt{k}}\left(1, -1, \cdots, 1, -1\right)^T.$$

Eigenvalues and eigenvectors of influence matrix H

Recall that the estimate $\hat{\boldsymbol{\beta}}$ of the function values at knots is given by

$$\hat{\boldsymbol{\beta}} = H\overline{\boldsymbol{y}} \quad \text{where} \quad \left(I_K + \frac{\lambda}{m} D^T B^{-1} D\right)^{-1}$$

Matrices D, B and I_K share the same eigenvectors, so does H.

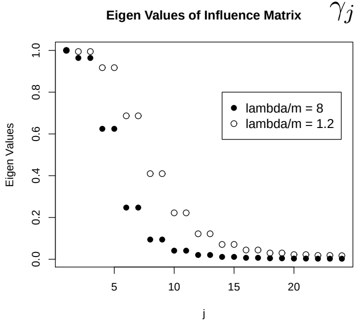
The eigenvalues of the influence matrix H are given in descending order by

$$\gamma_1 = 1, \quad \gamma_{2j} = \gamma_{2j+1} = \left(1 + \frac{\lambda}{m} \cdot \frac{12}{h^3} \cdot \frac{\left(1 - \cos\frac{2\pi j}{k}\right)^2}{\left(2 + \cos\frac{2\pi j}{k}\right)}\right)^{-1} \quad j = 1, \cdots, q-1,$$

$$\text{and} \quad \gamma_{2q} = \left(1 + \frac{\lambda}{m} \cdot \frac{48}{h^3}\right)^{-1}$$

Eigenvalues and eigenvectors of influence matrix H

γ_j

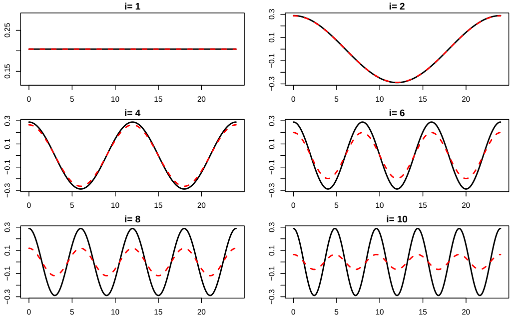


$$\hat{\boldsymbol{\beta}} = \left(I_k + \frac{\lambda}{m} D^T B^{-1} D\right)^{-1} \overline{\boldsymbol{y}}$$

$$= \sum_{j=1}^{k} \gamma_j (\boldsymbol{u}_j, \overline{\boldsymbol{y}})\, \boldsymbol{u}_j$$

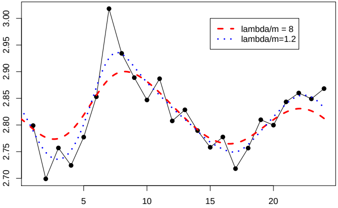
where γ_js are eigenvalues and $\boldsymbol{u}_j$ are eigenvectors of H.

$\boldsymbol{u}_j$



- The black solid curves are cyclic cubic spline basis functions corresponding to $\boldsymbol{u}_j$.
- The red dashed curves are cyclic cubic spline basis functions multiplied by γ_j $(=\gamma_j \boldsymbol{u}_j)$.

Smoothing mechanism of cyclic cubic smoothing splines



$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\sum_{j=1}^{k} \gamma_j \boldsymbol{u}_j \boldsymbol{u}_j^T\right) \overline{\boldsymbol{y}} \\ &= \sum_{j=1}^{k} \gamma_j (\boldsymbol{u}_j, \overline{\boldsymbol{y}})\ \boldsymbol{u}_j \\ \overline{\boldsymbol{y}} &= \sum_{j=1}^{k} (\boldsymbol{u}_j, \overline{\boldsymbol{y}})\ \boldsymbol{u}_j\end{aligned}$$

The smoothing mechanism can be understood as follows

- It decomposes the average observation vector $\overline{\boldsymbol{y}}$ into
  - the constant component (overall mean), and
  - $\sin$ and $\cos$ components with frequencies 1 to $q(= m/2)$.
- $\sin$ and $\cos$ components are shrunk. The higher the frequency is, the more the component is shrunk.
- The overall mean and shrunk components are summed up to produce $\hat{\boldsymbol{\beta}}$

2. A tensor product method:

an extenstion to a multivariate smoothing method

A tensor product method : an extenstion to a bivariate smoothing method

Suppose we have

basis functions for a function space Ω_1 : $\quad a_1(s), a_2(s), \cdots, a_{K_1}(s)$

basis functions for a function space Ω_2 : $\quad b_1(t),\ b_2(t),\ \cdots, b_{K_2}(t).$

A tensor product method uses products of basis functions on $\Omega_1 \times \Omega_2$

$$a_i(s)b_j(t), \quad i = 1, 2, \cdots, K_1,\ j = 1, 2, \cdots, K_2$$

as its basis functions. Thus, a bivariate function for the tensor product method can be expressed as

$$f_{st}(s,t) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \beta_{ij} a_i(s) b_j(t) \tag{5}$$

Roughness penalty for the tensor product method

Roughness penality for a tensor product smoothing function is defined as

$$J(f_{st}) = \int_{\Omega_s \times \Omega_t} \lambda_s \left(\frac{\partial^2 f_{st}}{\partial s^2}\right)^2 + \lambda_t \left(\frac{\partial^2 f_{st}}{\partial t^2}\right)^2 ds\, dt.$$

When knots are evenly spaced, the penalty term can be approximated as

Penalty term for a tensor product smoothing function

$$J(f_{st}) \approx \lambda_s \boldsymbol{\beta}^T (S_s \otimes I_{K_t}) \boldsymbol{\beta} + \lambda_t \boldsymbol{\beta}^T (I_{K_s} \otimes S_t) \boldsymbol{\beta} \quad (6)$$

where $\boldsymbol{\beta}$ is a vector of appropriately rearranged function values at grids.

(Wood, 2006)

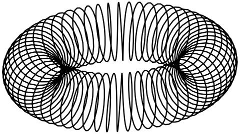
3. Doubly cyclic cubic smoothing splines

Doubly cyclic cubic smoothing splines

Doubly cyclic cubic smoothing splines are generated using a tensor product method with:

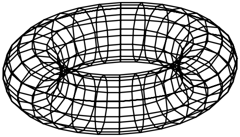
- basis functions for a function space with a yearly period : $f_1^a,\ f_2^a, \cdots,\ f_{K_a}^a$
- basis functions for a function space with a daily period : $f_1^d,\ f_2^d, \cdots,\ f_{K_d}^d$.

We start with a univariate function of time t, defined on a coil, that winds around a torus:



$$f(t) = \sum_{i=1}^{K_a} \sum_{j=1}^{K_d} \beta_{ij} f_i^a(t) f_j^d(t)$$

Then, to have a function that is smooth in two directions, we re-express this as:



$$\tilde{f}(s,t) = \sum_{i=1}^{K_a} \sum_{j=1}^{K_d} \beta_{ij} f_i^a(s) f_j^d(t)$$

and consider penalty to this function.

What does doubly cyclic cubic smoothing spline do?

When

- knots are evenly spaced, and
- the numbers of observations are equal for all knots,

then, original basis functions can be linearly transformed into

Oorthogonal basis functions

f_{ij}^{cos} : cyclic cubic spline function whose values at knots are equal to $\cos\left(2\pi t(ih_a + jh_d)\right)$ for $i = 0, \cdots q_a^*, j = 0, \pm 1, \cdots \pm q_d^*$

f_{ij}^{sin} : cyclic cubic spline function whose values at knots are equal to $\sin\left(2\pi t(ih_a + jh_d)\right)$ for $i = 0, \cdots q_a^*, j = 0, \pm 1, \cdots \pm q_d^*$

where $q_d^* \leq K_d/2 - 1$, $q_a^* \ll K_a/2 - 1$ and for $h_a = 1/K_a,\ h_d = 1/K_d$,

These are the eigenvectors of the influence matrix for doubly cyclic cubic smoothing splines.

What do doubly cyclic cubic smoothing splines do?

Estimated function

$$\hat{f}(t) = \sum_{i=0}^{q_a^*} \sum_{j=-q_d^*}^{q_d^*} \left(1 + \lambda_a \frac{(1-\cos 2\pi i h_a)^2}{2+\cos 2\pi i h_a} + \lambda_d \frac{(1-\cos 2\pi j h_d)^2}{2+\cos 2\pi j h_d}\right)^{-1} \times \left(\frac{< \boldsymbol{u}_{ij}^{cos}, \overline{\boldsymbol{y}} >}{|\boldsymbol{u}_{ij}^{cos}|^2} f_{ij}^{cos} + \frac{< \boldsymbol{u}_{ij}^{sin}, \overline{\boldsymbol{y}} >}{|\boldsymbol{u}_{ij}^{sin}|^2} f_{ij}^{sin} \right)$$

where $q_d^* \leq K_d/2 - 1$, $q_a^* \ll K_a/2 - 1$ and for $h_a = 1/K_a,\ h_d = 1/K_d$,

$\boldsymbol{u}_{ij}^{cos}$: vectors of values of $\cos(2\pi t(ih_a + jh_d))$ at knots

$\boldsymbol{u}_{ij}^{sin}$: vectors of values of $\sin(2\pi t(ih_a + jh_d))$ at knots

f_{ij}^{cos} : cyclic cubic spline function with $\cos(2\pi t(ih_a + jh_d))$ as values at knots

f_{ij}^{sin} : cyclic cubic spline function with $\sin(2\pi t(ih_a + jh_d))$ as values at knots

$\overline{\boldsymbol{y}}$: vector of the averages at knots

Wiggly components are shrunk more

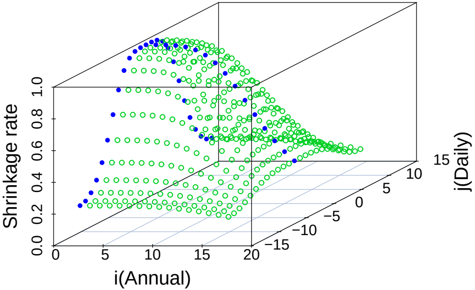
The doubly cyclic cubic smoothing spline shrinks the components of

- basis function with values at knots $\cos\left(2\pi t(ih_a \pm jh_d)\right)$, and
- basis function with values at knots $\sin\left(2\pi t(ih_a \pm jh_d)\right)$

by multiplying them by the shrinkage rate

$$\left(1 + \lambda_a \frac{(1-\cos 2\pi i h_a)^2}{2+\cos 2\pi i h_a} + \lambda_d \frac{(1-\cos 2\pi j h_d)^2}{2+\cos 2\pi j h_d}\right)^{-1}$$

and sum them up.



The larger i or j is, the more wiggly the basis function is in either direction.

The more wiggly the basis function is in either direction, the more its coefficient is shrunk.

4. Analysis of CO2 concentration at Syowa station in Antarctica

A model with temporal trend and seasonal daily pattern

We start with a linear additive model for CO2 concentration with temporal trend and seasonal daily pattern as explanatory terms.

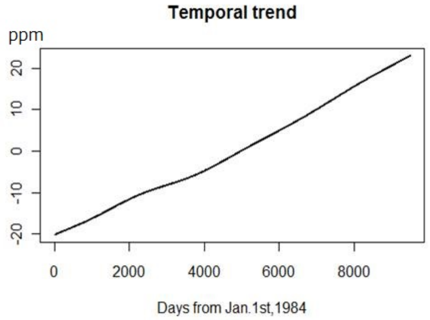
$$\text{Model 1:} \qquad Y = f_{\mathrm{tr}}(t) + f_{\mathrm{day,year}}(t) + \epsilon$$

where

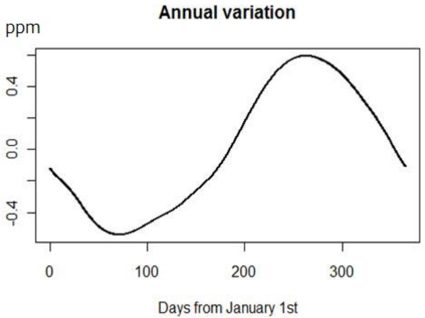
Y	: CO2 concentration
$f_{\mathrm{tr}}(t)$	: a cubic spline function of time t for temporal trend
$f_{\mathrm{day,year}}(t)$	: a doubly cyclic cubic spline function of time t with daily and annual cycles
ϵ	: random error with variance σ^2

We used R package *mgcv* by Simon Wood for analysis.

Temporal trend and annual variation

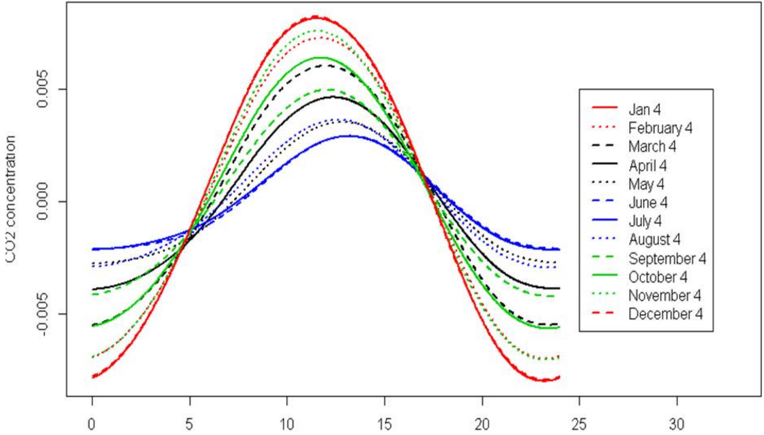


- Temporal trend is almost linear
- CO_2 concentration has increased 40ppm in 25 years



- The range of annual variation is 1.1ppm
- CO_2 concentration is low in summer and high in winter

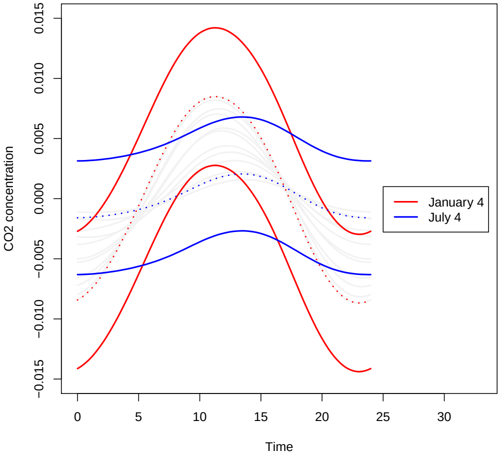
Seasonal variation of daily Pattern in CO2 concentration



Daily pattern of CO2 concentration has a seasonal variation
It has the largest daily variation (0.017ppm) in summer (January 4th)

Confidence interval curves for daily pattern

95% confidence interval curves for January 4th and July 4th.



Hourly variation is significant in summer (January 4th),

but not significant in winter (July 4th).

Effects of wind speed and direction

Wind speed might have an effect on CO2 concentration.
The effect of wind speed might depend on the wind direction.

Model 2: $$y = f_{\text{tr}}(t) + f_{\text{day,year}}(t) + f_{\text{ws,wd}}(s, d) + \epsilon$$

where

Y	: Co2 concentration
$f_{\text{tr}}(t)$	: a cubic spline function of time t for temporal trend
$f_{\text{day,year}}(t)$	: a doubly cyclic cubic spline function of time t with daily and annual cycles
$f_{\text{ws,wd}}(s, d)$	: tensor product of a cubic spline function of wind speed s and a cyclic spline function of wind direction d
ϵ	: random error with variance σ^2

Seasonal effect of wind speed

The effect of wind speed might differ by season.

Model 3: $$y = f_{\mathrm{tr}}(t) + f_{\mathrm{day,year}}(t) + f_{\mathrm{ws,year}}(s,t) + \epsilon$$

where

Y	: Co2 concentration
$f_{\mathrm{tr}}(t)$	: a cubic spline function of time t for temporal trend
$f_{\mathrm{day,year}}(t)$	: a doubly cyclic cubic spline function of time t with daily and annual cycles
$f_{\mathrm{ws,year}}(s,t)$	: tensor product of a cubic spline function of wind speed s and a cyclic cubic spline function with annual cycle of t
ϵ	: random error with variance σ^2

Model selection for CO2 concentration

We fitted the following three models and compared AIC

model	formula	AIC
model 1	$y = f_{\mathrm{tr}}(t) + f_{\mathrm{day,year}}(t) + \epsilon$	19942.5
model 2	$y = f_{\mathrm{tr}}(t) + f_{\mathrm{day,year}}(t) + f_{\mathrm{ws,wd}}(s,d) + \epsilon$	18725.6
model 3	$y = f_{\mathrm{tr}}(t) + f_{\mathrm{day,year}}(t) + f_{\mathrm{ws,year}}(s,t) + \epsilon$	16189.5

where

$f_{\mathrm{tr}}(t)$: a cubic spline function of time t

$f_{\mathrm{day,year}}(t)$: a doubly cyclic cubic spline function of time t with daily and annual cycles

$f_{\mathrm{ws,wd}}(s,d)$: tensor product of a cubic spline function of wind speed s and a cyclic spline function of wind direction d

$f_{\mathrm{ws,year}}(s,t)$: tensor product of a cubic spline function of wind speed s and a cyclic cubic spline function with annual cycle of t

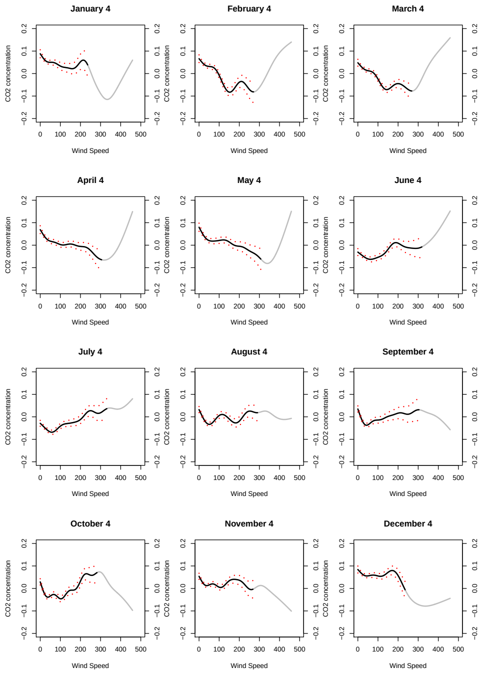
Seasonal change of wind speed effect

January - March

April - June

July - September

October - December



Conclusions

- We proposed the doubly cyclic cubic smoothing spline method.
- For a simple model, the eigenvalues and eigenvectors of the influence matrix can be explicitly expressed with the values of trigonometric functions with different frequencies.
- This expression shows that the more wiggly the basis function is, the more its coefficient is shrunk.
- We analyzed CO2 concentration at Syowa station in Antarctica using this method. CO2 concentration has a strong temporal trend and annual variation.
- Daily pattern of CO2 concentration has a seasonal variation. Hourly variation is significant in summer (January), but not significant in winter (July).
- The effect of wind speed also has annual variation.
- Flexible regression models using nonparametric smoothing methods enable us to analyze the data of interest more precisely.

Reference

1. Green, P.J. and Silverman, B.W. (1993) Nonparametric Regression and Generalized Linear Models: A roughness penalty approach, Chapman&Hall/CRC.
2. Wood, S.N. (2003) Thin plate regression splines. J. R. Statist. Soc. B 65, 95-114
3. Wood, S.N. (2006a) Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. Biometrics 62(4): 1025-1036
4. Wood, S.N. (2006b) Generalized Additive Models: An Introduction with R, Chapman Hall/CRC.
5. Kiguchi, R. and Minami, M. (2012) Cyclic Cubic Regression Spline Smoothing and Analysis of CO2 Data at Showa Station in Antarctica, Proceedings of International Biometric Conference 2012.
6. Kiguchi, R. (2014) Doubly Cyclic Smoothing Splines and Analysis of CO2 Data at Syowa Station in Antarctica, Master's thesis, Keio University.

Thank you for your attention!