Statistics 860 Lecture 12. Smoothing Spline ANOVA: Second Variational Problem.

refs: Wahba book Ch 10 must read , Gu book.
Will talk about the second variational problem.
Software for SSANOVA is in R, see `gss, fields, assist`

Examples for today:
gu.wahba.tps.93.pdf - thin plate spline for lake latitude and longitude, cubic spline for calcium content

fing.pdf - "fingerprint" method for detection of global warming, a spline on the sphere for global latitude and longitude, , 30-vector for time, 30 years. Splines on the sphere are in `sphspl.pdf`

wahba.wang.gu.95.pdf progression of diabetic retinopathy See also lin.wahba.zhang.gao.klein.klein.2000.pdf

C. Gu and G. Wahba. Semiparametric analysis of variance with tensor product thin plate splines. J. Royal Statistical Soc. Ser. B, 55:353-368, 1993.

Chiang, A., Wahba, G., Tribbia, J., and Johnson, D. R. ” Quantitative Study of Smoothing Spline-ANOVA Based Fingerprint Methods for Attribution of Global Warming ” TR 1010, July 1999.

G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. Ann. Statist., 23:1865-1895, 1995.

First Variational Problem:

$$\mathcal{H}^\alpha = [1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)}$$

$$\mathcal{H} = \prod_{\alpha=1}^{d} \mathcal{H}^\alpha = \prod_{\alpha=1}^{d} [[1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)}]$$

$$= [1] \oplus \sum_\alpha \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} \mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)} + \cdots$$

Two factor interaction model:
$\mathcal{H} = [1] \oplus \sum_\alpha \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} \mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}$

Let $R^{(\alpha)}$ be the RK for $\mathcal{H}^{(\alpha)}$. Since all of these subspaces are orthogonal, the RK for $\mathcal{H}$ is:

$$R(s,t) = 1 + \sum_\alpha R^{(\alpha)}(s_\alpha, t_\alpha) + \sum_{\alpha<\beta} R^{(\alpha)}(s_\alpha, t_\alpha) R^{(\beta)}(s_\beta, t_\beta)$$

Second ANOVA Variational Problem:

$$\mathcal{H}^{\alpha} = [1]^{(\alpha)} \oplus \mathcal{H}_{\pi}^{(\alpha)} \oplus \mathcal{H}_{s}^{(\alpha)}$$

"π"="parametric",s="Smooth"
$\mathcal{H}_{\pi}^{(\alpha)}$ is spanned by $\{\phi_1^{(\alpha)}, \cdots, \phi_{M-1}^{(\alpha)}\}$ an orthogonal basis in $\mathcal{H}_{\pi}^{(\alpha)}$.
$\{\phi_{\nu}^{(\alpha)}\}$ span the null space of the penalty functional that we want to impose on $f^{(\alpha)}$–the main effects.

For $f \in \mathcal{H}^{\alpha}$,

$$P_{\{1^{(\alpha)}\}} f = \int_{\mathcal{T}^{(\alpha)}} f(z_\alpha) d\mu_\alpha(z_\alpha) = [\mathcal{E}_\alpha f] 1^{(\alpha)}$$

Define the inner product in $\mathcal{H}_\pi^{(\alpha)}$
as

$$\langle \phi_\mu^{(\alpha)}, \phi_\nu^{(\alpha)} \rangle = \int \phi_\mu^{(\alpha)} \phi_\nu^{(\alpha)} d\mu_\alpha$$

choose the $\phi_\nu^{(\alpha)}$ to be orthonormal.
In dealing with a single variable, the norm in $\mathcal{H}_\pi^{(\alpha)}$ is irrelevant, but it will affect the interaction term, as we shall see.
Define the orthogonal projector from $\mathcal{H}^{(\alpha)}$ onto $\mathcal{H}_\pi^{\alpha}$ as

$$P_\pi^{(\alpha)} f = \sum_{\nu=1}^{M-1} \phi_\nu^{(\alpha)} \int \phi_\nu^{(\alpha)}(z_\alpha) f(z_\alpha) d\mu_\alpha$$

$$\mathcal{H}^{(\alpha)} = \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)}$$
$$\mathcal{H}_\pi^{(\alpha)} = P_\pi^{(\alpha)}(\mathcal{H}^{(\alpha)})$$

$\mathcal{H}_\pi^{(\alpha)} \perp \mathcal{H}_s^{(\alpha)}$ with the norm defined by

$$\|f^{(\alpha)}\|^2 = \|P_\pi^{(\alpha)} f^{(\alpha)}\|^2_{\mathcal{H}_\pi^\alpha} + \|(I - P_\pi^{(\alpha)}) f^{(\alpha)}\|^2_{\mathcal{H}_s^\alpha}$$

The RK for $\mathcal{H}_s^{(\alpha)}$ is

$$(I - P_{\pi(s_\alpha)}^{(\alpha)})(I - P_{\pi(t_\alpha)}^{(\alpha)}) R^{(\alpha)}(s_\alpha, t_\alpha) = R_s^{(\alpha)}(s_\alpha, t_\alpha)$$

The RK for $\mathcal{H}_\pi^\alpha$ is $\sum_{\nu=1}^M \phi_\nu^{(\alpha)}(s_\alpha)\phi_\nu^{(\alpha)}(t_\alpha)$

ANOVA Decomposition For the Second Variational Problem:

$$\prod_{\alpha=1}^{d} \mathcal{H}^{\alpha} = \prod_{\alpha=1}^{d} \{[1^{(\alpha)}] \oplus \mathcal{H}_{\pi}^{(\alpha)} \oplus \mathcal{H}_{s}^{(\alpha)}\}$$

In d-dimensions there are a maximum of 3^d subspaces, d=2, 9 subspaces

$$\begin{pmatrix} [1^{(1)}] \otimes [1^{(2)}] & \vdots & [1^{(1)}] \otimes \mathcal{H}_{\pi}^{(2)} & \vdots & [1^{(1)}] \otimes \mathcal{H}_{s}^{(2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathcal{H}_{\pi}^{(1)} \otimes [1^{(2)}] & \vdots & \mathcal{H}_{\pi}^{(1)} \otimes \mathcal{H}_{\pi}^{(2)} & \vdots & \mathcal{H}_{\pi}^{(1)} \otimes \mathcal{H}_{s}^{(2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathcal{H}_{s}^{(1)} \otimes [1^{(2)}] & \vdots & \mathcal{H}_{s}^{(1)} \otimes \mathcal{H}_{\pi}^{(2)} & \vdots & \mathcal{H}_{s}^{(1)} \otimes \mathcal{H}_{s}^{(2)} \end{pmatrix}$$

“Parametric” part (finite dimensional) is the 11, 12, 21 and 22 elements of this array.

$$\mathcal{H}_\pi^{(1)} \otimes \mathcal{H}_\pi^{(2)} : \quad \{\phi_\nu^{(1)}(t_1)\phi_\mu^{(2)}(t_2)\}_{\nu=1}^{M_1-1}\,{}_{\mu=1}^{M_2-1}$$

$$\mathcal{H}_\pi^{(1)} \otimes \mathcal{H}_s^{(2)} : \quad \{\phi_\nu^{(1)}(t_1)f^{(2)}(t_2)\}_{\nu=1}^{M_1-1}$$

where

$$\mathcal{E}_2 f^{(2)} = 0$$

$$P_\pi^{(2)} f^{(2)} = 0$$

$$\mathcal{H}_s^{(1)} \otimes \mathcal{H}_s^{(2)} : \quad f(t_1, t_2), \quad f \in \mathcal{H}$$

$$\mathcal{E}_{1(t_1)} f(t_1, t_2) = 0$$

$$\mathcal{E}_{2(t_2)} f(t_1, t_2) = 0$$

$$P_{\pi(t_1)}^{(1)} P_{\pi(t_2)}^{(2)} f(t_1, t_2) = 0$$

$$\mathcal{H} = \mathcal{H}_0 \oplus \sum \mathcal{H}^{\beta}$$

$\mathcal{H}_0$: all the parametric subspaces.

$$1, \{\phi_{\mu}^{(1)}\}, \{\phi_{\nu}^{(2)}\}, \{\phi_{\nu}^{(1)}\phi_{\mu}^{(2)}\}$$

are $1 + (M_1 - 1) + (M_2 - 1) + (M_1 - 1)(M_2 - 1)$ elements.
$d = 2, \beta = 1, 2, 3, 4, 5$
are 5 nonparametric subspaces

$$\left(\begin{array}{c|c|c} & & \text{X} \\ \hline & & \text{X} \\ \hline \text{X} & \text{X} & \text{X} \end{array}\right)$$

to find f to minimize

$$\sum (y_i - L_i f)^2 + \sum_{\beta=1}^{5} \lambda_{\beta} \|P_{\beta} f\|^2 \qquad \text{for d=2}$$

if both M_1 and M_2 are > 1.

The RK's for the 5 nonparametric subspaces will be

$R_s^{(1)}(s_1, t_1)$, $R_s^{(2)}(s_2, t_2)$, $R_\pi^{(1)}(s_1, t_1)R_s^{(2)}(s_2, t_2)$,

$R_s^{(1)}(s_1, t_1)R_\pi^{(2)}(s_2, t_2)$ and $R_s^{(1)}(s_1, t_1)R_s^{(2)}(s_2, t_2)$.

Lemma

Let $\mathcal{H}_1 = \sum_{\beta=1}^{p} \oplus \mathcal{H}^\beta$, where the $\mathcal{H}^\beta$ are orthogonal subspaces of $\mathcal{H}_1$. If $f \in \mathcal{H}_1$, then

$$\|f\|^2_{\mathcal{H}_1} = \sum_{\beta=1}^{p} \|P_\beta f\|^2_{\mathcal{H}^\beta}$$

and the RK for $\mathcal{H}_1$ is $\sum_{\beta=1}^{p} R^\beta(s,t)$ where R^β is the RK for $\mathcal{H}^\beta$. Given $\theta_1, \cdots, \theta_p > 0$, then we may define another norm on $\mathcal{H}_1$ by

$$\|f\|^2_{\theta\mathcal{H}_1} = \sum_{\beta=1}^{p} \frac{1}{\theta_\beta} \|P_\beta f\|^2_{\mathcal{H}^\beta} = \sum_{\beta=1}^{p} \lambda_\beta \|P_\beta f\|^2_{\mathcal{H}^\beta}$$

and the RK for this norm is

$$\sum_{\beta=1}^{p} \theta_\beta R^\beta(s,t)$$

$$\boxed{\begin{array}{rcl} Td + \Sigma^\theta c & = & y \\ T'c & = & 0 \end{array}}$$

$$\Sigma^\theta = \theta_1 \Sigma_1 + \cdots + \theta_p \Sigma_p$$

all of the original formulas hold with Σ replaced by Σ^θ.

$$A(\lambda) = A(\lambda, \theta) = A(\lambda_1, \cdots, \lambda_p)$$

where $\lambda_\beta = \lambda \theta_\beta^{-1}$.

To make this unique, must put a constraint on $\theta_1, \cdots, \theta_p$. For example, $\sum_{j=1}^p log\theta = 0$.

RKPACK, gss in R. gcv.gml.pdf.

From `gu:wahba:tps.93.pdf`
Lake acidity in the Blue Ridge Mountains

$$y_i = \mu + f_1(t_1) + f_2(t_2) + f_{12}(t_1, t_2)$$

y_i is lake acidity (pH) in lake i
$t_1(i)$ is calcium content of lake $i(log_{10}mg/L)$
$t_2(i)$ is (centered latitude, longitude) $(x_1(i), x_2(i))$
of lake i
$f(t_1)$ is a cubic spline
$f(t_2)$ is a thin plate spline

Averaging operators for both calcium content and lake acidity are the marginal design measures:

$$\mathcal{E}_\alpha(f) = \frac{1}{n}\sum_{i=1}^{n} f(t_\alpha(i)), \alpha = 1, 2.$$

Unpenalized terms other than the constant function on the plot region are a linear function in calcium content and two linear functions in (latitude,longitude), $\phi_1^{((1)}(t_1), \phi_1^{(2)}(t_2)$ and $\phi_2^{(2)}(t_2)$.

For the cubic spline term
$\phi_1^{(1)}(t_1) = t_1 - \frac{1}{n}\sum_{i=1}^n t_1(i)$.
For the thin plate spline with $m = 2, d = 2$, let
$\psi_1^{(2)} = x_1 - \frac{1}{n}\sum_{i=1}^n x_1(i)$
$\psi_2^{(2)} = x_2 - \frac{1}{n}\sum_{i=1}^n x_2(i)$
and obtain an orthogonal pair
$\phi_1^{(2)}, \phi_2^{((2)}$
(satisfies $\frac{1}{n}\sum_{i=1}^n \phi_1^{(2)}(t_2(i))\phi_2^{(2)}(t_2(i)) = 0$).

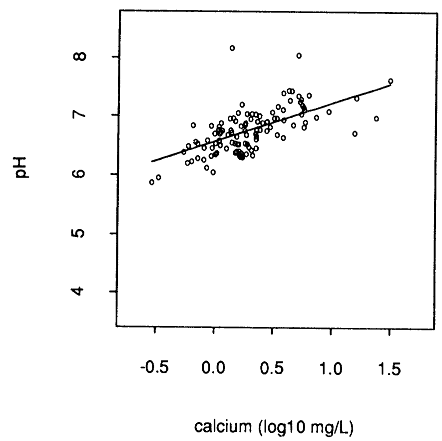


Fig. 2. Calcium main effect for the Blue Ridge model

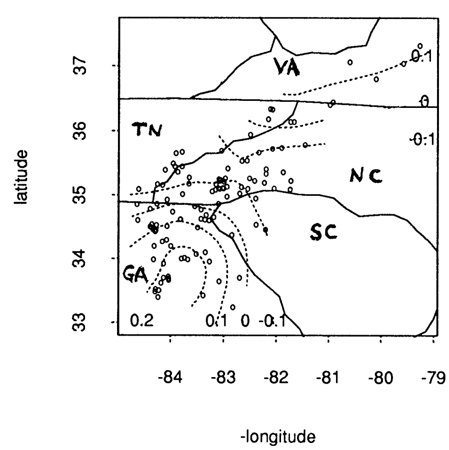


Fig. 3. Geography main effect for the Blue Ridge model

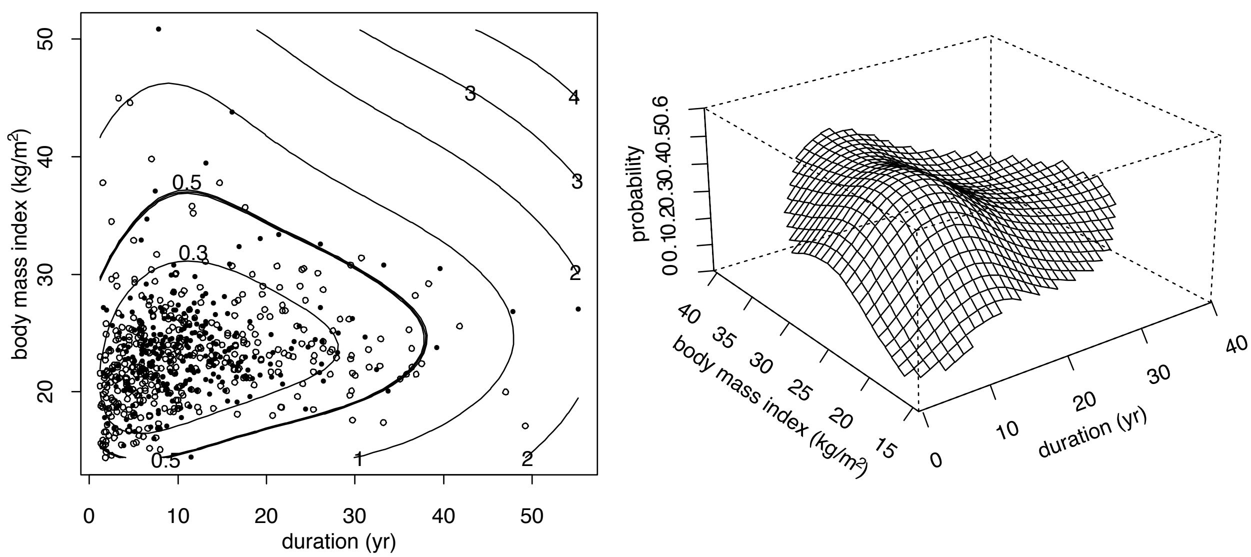
From `wahba:wang:gu:95.pdf`
Wisconsin Epidemiological Study of Diabetic Retinopathy
$n = 891$. Younger onset diabetics.
$y =$ four year progression of diabetic retinopathy, $1 = yes$, $0 = no$. Model variables:

1. `dur`: duration of diabeties at baseline
2. `gly`: glycosylated hemoglobin, a measure of hyperglycemia, %
3. `bmi`: body mass index-weight in kg /(height in m)^2

The model

$$f(\texttt{dur}, \texttt{gly}, \texttt{bmi}) = \mu + f_1(\texttt{dur}) + a_2 \cdot \texttt{gly}$$

$$+ f_3(\texttt{bmi}) + f_{13}(\texttt{dur}, \texttt{bmi})$$



Left: data and contours of constant posterior standard deviation. Right: estimated probability of progression as a function of duration and body mass index for glycosylated hemoglobin fixed at its median.

Time and Space Models on the Globe

Here $t = (t_1, t_2) = (x, P)$ where x is year, and P is (latitude, longitude). The RKHS of historical global temperature functions that was used in Chiang, Wahba, Johnson and Tribbia (1999) is

$$\mathcal{H} = [[1^{(1)}] \oplus [\phi] \oplus \mathcal{H}_s^{(1)}] \otimes [[1^{(2)}] \oplus \mathcal{H}_s^{(2)}],$$

a collection of functions $f(x, P)$, on $\mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)}$ $=\{1, 2, ..., 30\} \otimes \mathcal{S}$, where $\mathcal{S}$ is the sphere, and ϕ is a function which averages to 0 on $\mathcal{T}^{(1)}$. $\mathcal{H}$ and f have the corresponding (six term) decompositions given next:

$$
\begin{array}{rcccccccc}
\mathcal{H} & = & [1] & \oplus & [\phi] & \oplus & [\mathcal{H}_s^{(1)}] & \oplus & [\mathcal{H}_s^{(2)}] \\
f(x,P) & = & C & + & d\phi(x) & + & f_1(x) & + & f_2(P) \\
 & = & mean & + & global & + & time & + & space \\
 & & & & time & & main & & main \\
 & & & & trend & & \textit{effect} & & \textit{effect}
\end{array}
$$

$$
\begin{array}{cccc}
\oplus & [[\phi] \otimes \mathcal{H}_s^{(2)}] & \oplus & [\mathcal{H}_s^{(1)} \otimes \mathcal{H}_s^{(2)}] \\
+ & \phi(x) f_{\phi,2}(P) & + & f_{12}(x,P) \\
+ & trend & + & space- \\
 & by\ space & & time \\
 & \textit{effect} & & interaction
\end{array}
$$

A sum of squares of second differences was applied to the time variable, and a spline on the sphere penalty (Wahba:1981,1982)) was applied to the space variable. For a cross country skier in the Midwest, as this author is, the results were very disappointing, in that they clearly showed a warming trend stretching from the Midwest towards Alaska (trend by space term) which was stronger than the global mean trend.

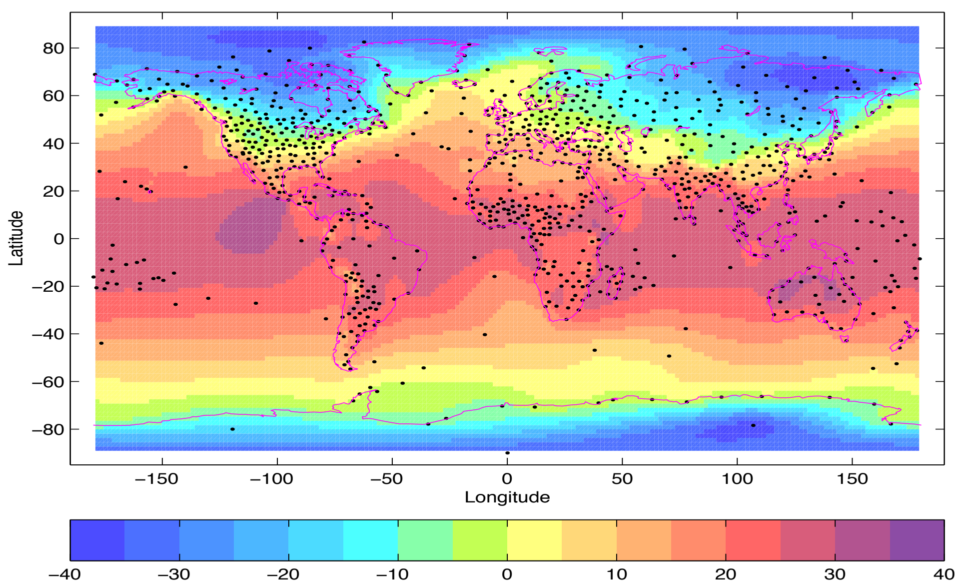


Figure 7: Mean of the historical average winter temperature (^{o}C), 1961-1990.

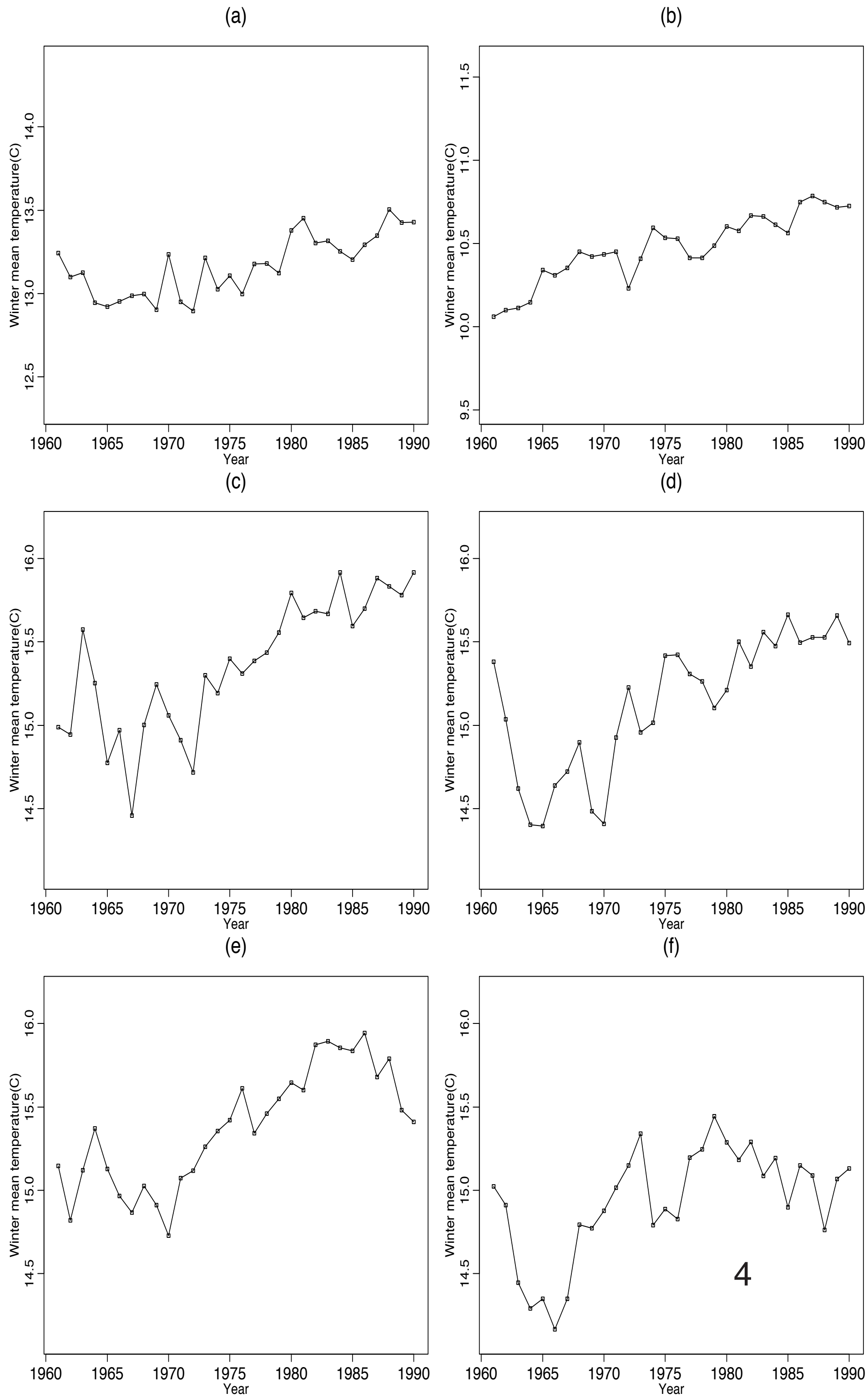


Figure 8: Yearly average winter temperatures (^oC): (a) Historical (b) GFDL forced (c)

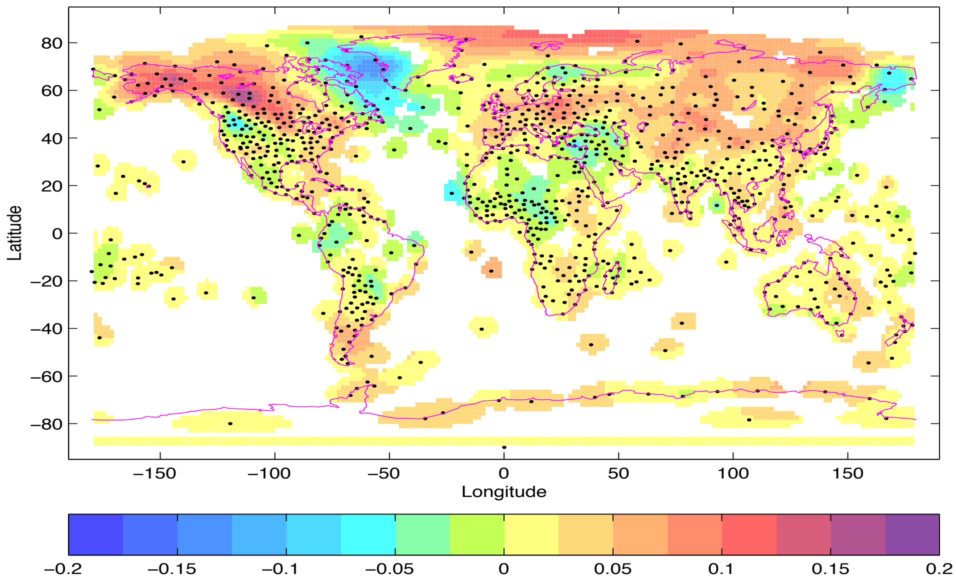


Figure 9: Linear trend of the historical average winter temperature (^{o}C/yr), 1961-1990.