Conditional quantiles with functional covariates: an application to Ozone pollution forecasting

Hervé Cardot, Christophe Crambes & Pascal Sarda

Compstat - Prague

August 2004

Presentation of the data (1)

Data (ORAMIP) :

Presentation of the data (1)

Data (ORAMIP) :

- 9 variables : $NO, N2, O3, WD, WS, \ldots$ (hourly measurements)

Presentation of the data (1)

Data (ORAMIP) :

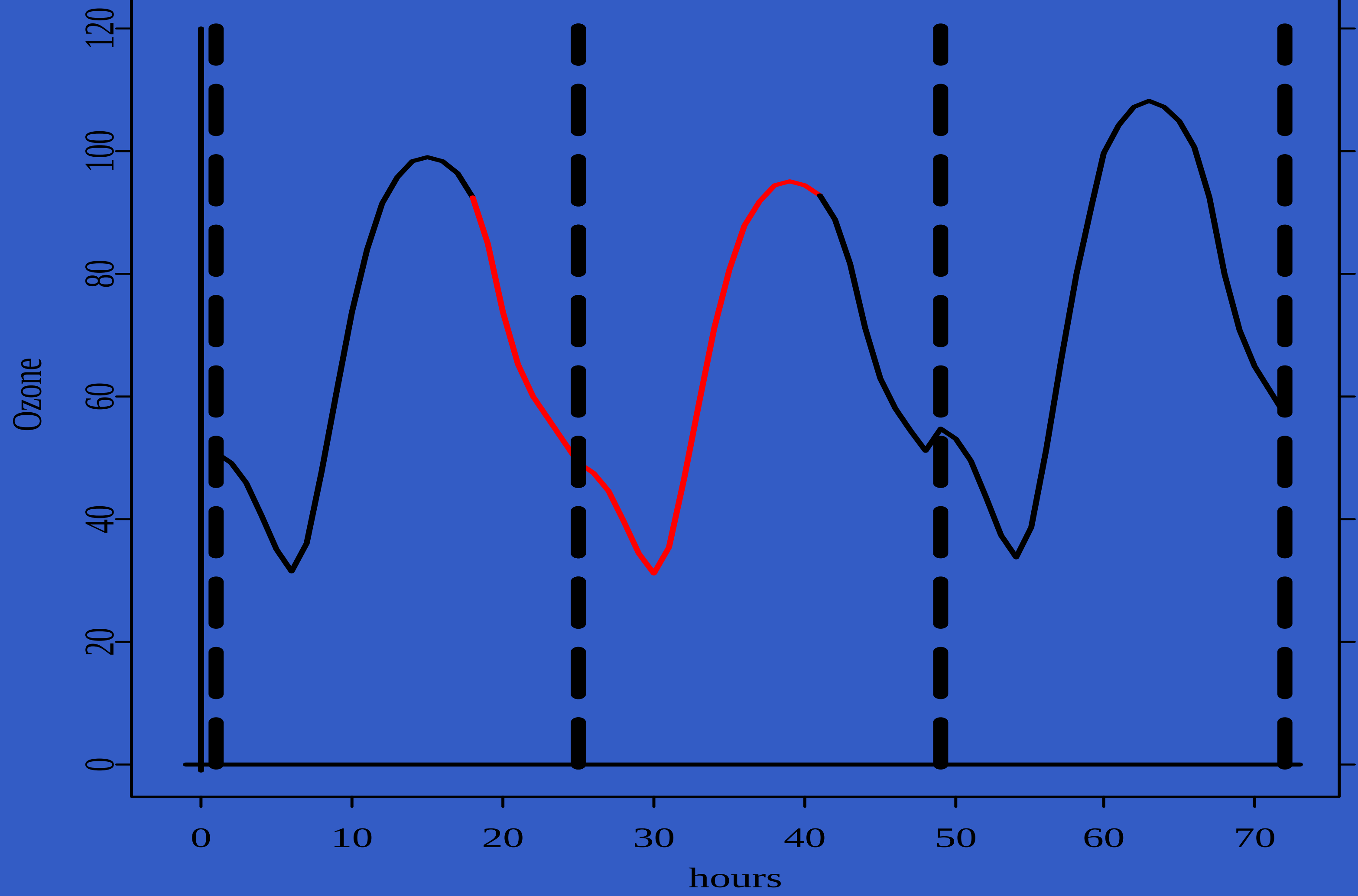
- 9 variables : $NO, N2, O3, WD, WS, \ldots$ (hourly measurements)
- 6 stations

Presentation of the data (1)

Data (ORAMIP) :

- 9 variables : $NO, N2, O3, WD, WS, \ldots$ (hourly measurements)
- 6 stations
- 4 years : $1997 - 2000$ (15^{th} May - 15^{th} Sept)

Presentation of the data (2)



Presentation of the data (3)

Presentation of the data (3)

- variable of interest : max of $O3$ every day:

$$Y = {}^t(Y_1, \ldots, Y_n)$$

Presentation of the data (3)

- variable of interest : max of $O3$ every day:

$$Y = {}^t(Y_1, \ldots, Y_n)$$

- covariates : $NO, N2, O3, DV$ or VV:

	18h	...	24h	1h	...	17h
day 0/day 1	$X_{1,1}$	...	...	...	...	$X_{1,24}$
$\vdots$	$\vdots$					$\vdots$
day $n-1$/day n	$X_{n,1}$	...	...	...	...	$X_{n,24}$

Presentation of the data (3)

- variable of interest : max of $O3$ every day:

$$Y = {}^t(Y_1, \dots, Y_n)$$

- covariates : $NO, N2, O3, DV$ or VV:

	18h	...	24h	1h	...	17h
day 0/day 1	$X_{1,1}$	...	...	...	...	$X_{1,24}$
$\vdots$	$\vdots$					$\vdots$
day $n-1$/day n	$X_{n,1}$	...	...	...	...	$X_{n,24}$

- $(X_i, Y_i)_{i=1,\dots,n}$ couples of random variables with $Y_i \in \mathbb{R}$ and $X_i \in L^2(I)$

Presentation of the data (3)

- variable of interest : max of $O3$ every day:

$$Y = {}^t(Y_1, \ldots, Y_n)$$

- covariates : $NO, N2, O3, DV$ or VV:

	18h	...	24h	1h	...	17h
day 0/day 1	$X_{1,1}$	...	...	...	...	$X_{1,24}$
$\vdots$	$\vdots$					$\vdots$
day $n-1$/day n	$X_{n,1}$	...	...	...	...	$X_{n,24}$

- $(X_i, Y_i)_{i=1,\ldots,n}$ couples of random variables with $Y_i \in \mathbb{R}$ and $X_i \in L^2(I)$
- X_i is known in $t_1, \ldots, t_p \in I$ (equispaced)

Definition of the conditional quantiles

Definition of the conditional quantiles

- $\alpha \in]0,1[$, $x \in L^2(I)$

Definition of the conditional quantiles

- $\alpha \in]0,1[$, $x \in L^2(I)$

- α conditional quantile :

$$P\left(Y \leq g_\alpha(X) | X = x\right) = \alpha$$

Definition of the conditional quantiles

- $\alpha \in]0,1[$, $x \in L^2(I)$

- α conditional quantile :

$$P\left(Y \leq g_\alpha(X) | X = x\right) = \alpha$$

- property :

$$g_\alpha(x) = \text{arg}\min_{a \in \mathbb{R}} \mathbb{E}\left(l_\alpha(Y - a) | X = x\right)$$

$$\text{with } l_\alpha(u) = |u| + (2\alpha - 1)u$$

Presentation of the model

Presentation of the model

- model (cf. Koenker and Bassett, 1978) :

$$g_\alpha(X) = c + \langle \Psi_\alpha, X \rangle = c + \int_I \Psi_\alpha(t) X(t)\, dt$$

Presentation of the model

- model (cf. Koenker and Bassett, 1978) :

$$g_\alpha(X) = c + \langle \Psi_\alpha, X \rangle = c + \int_I \Psi_\alpha(t) X(t)\, dt$$

- we want to estimate the function $\Psi_\alpha \in L^2(I)$: spline estimation

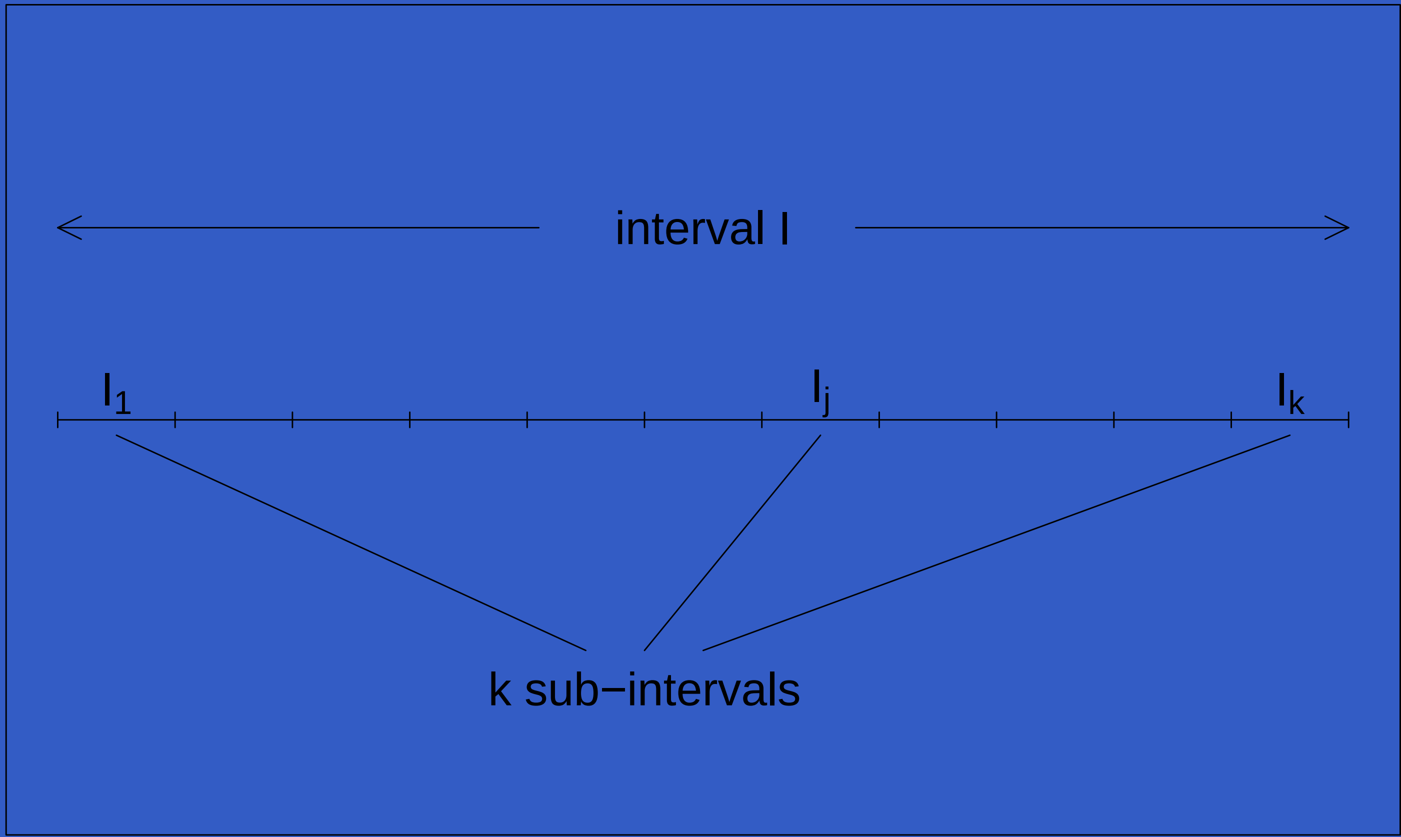
Spline estimation of Ψ_α

Spline estimation of Ψ_α

$k \in \mathbb{N}^\star$, $q \in \mathbb{N}$

Spline estimation of Ψ_α

$k \in \mathbb{N}^\star$, $q \in \mathbb{N}$



Spline estimation of Ψ_α

$k \in \mathbb{N}^\star$, $q \in \mathbb{N}$

$\mathbf{B_{k,q}} = {}^t(B_1, \ldots, B_{k+q})$ B-splines basis

Spline estimation of Ψ_α

$k \in \mathbb{N}^\star$, $q \in \mathbb{N}$

$\mathbf{B_{k,q}} = {}^t(B_1, \ldots, B_{k+q})$ B-splines basis

estimator : $\widehat{\Psi}_\alpha = {}^t\mathbf{B_{k,q}}\widehat{\boldsymbol{\theta}} = \sum_{j=1}^{k+q} \widehat{\theta}_j \; B_j$

Spline estimation of Ψ_α

$k \in \mathbb{N}^\star$, $q \in \mathbb{N}$

$\mathbf{B_{k,q}} = {}^t(B_1, \ldots, B_{k+q})$ B-splines basis

estimator : $\widehat{\Psi}_\alpha = {}^t\mathbf{B_{k,q}}\widehat{\boldsymbol{\theta}} = \sum_{j=1}^{k+q} \widehat{\theta}_j B_j$

Determination of $\widehat{c}$ and $\widehat{\theta}$

Determination of $\widehat{c}$ and $\widehat{\boldsymbol{\theta}}$

- $\widehat{\boldsymbol{\theta}}$ and $\widehat{c}$ solution of the minimisation problem :

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^{n} l_\alpha(Y_i - c - \langle {}^t\mathbf{B}_{\mathbf{k},\mathbf{q}} \boldsymbol{\theta}, X_i \rangle) + \rho \parallel ({}^t\mathbf{B}_{\mathbf{k},\mathbf{q}} \boldsymbol{\theta})^{(m)} \parallel^2 \right\}$$

Determination of $\widehat{c}$ and $\widehat{\theta}$

- $\widehat{\theta}$ and $\widehat{c}$ solution of the minimisation problem :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l_\alpha(Y_i - c - \langle {}^t\mathbf{B}_{\mathbf{k},\mathbf{q}}\boldsymbol{\theta}, X_i\rangle) + \rho \parallel ({}^t\mathbf{B}_{\mathbf{k},\mathbf{q}}\boldsymbol{\theta})^{(m)} \parallel^2 \right\}$$

empirical version of

$$\mathbb{E}\left(l_\alpha(Y - c - \langle s, X\rangle)\right)$$

Determination of $\widehat{c}$ and $\widehat{\theta}$

- $\widehat{\boldsymbol{\theta}}$ and $\widehat{c}$ solution of the minimisation problem :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l_\alpha(Y_i - c - \langle {}^t\mathbf{B_{k,q}}\boldsymbol{\theta}, X_i\rangle) + \rho \parallel ({}^t\mathbf{B_{k,q}}\boldsymbol{\theta})^{(m)} \parallel^2 \right\}$$

penalization

Determination of $\widehat{c}$ and $\widehat{\boldsymbol{\theta}}$

- $\widehat{\boldsymbol{\theta}}$ and $\widehat{c}$ solution of the minimisation problem :

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^{n} l_\alpha (Y_i - c - \langle {}^t\mathbf{B}_{\mathbf{k},\mathbf{q}} \boldsymbol{\theta}, X_i \rangle) + \rho \parallel ({}^t\mathbf{B}_{\mathbf{k},\mathbf{q}} \boldsymbol{\theta})^{(m)} \parallel^2 \right\}$$

- no explicit solution

Determination of $\widehat{c}$ and $\widehat{\theta}$

- $\widehat{\theta}$ and $\widehat{c}$ solution of the minimisation problem :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l_\alpha(Y_i - c - \langle {}^t\mathbf{B_{k,q}}\boldsymbol{\theta}, X_i\rangle) + \rho \parallel ({}^t\mathbf{B_{k,q}}\boldsymbol{\theta})^{(m)} \parallel^2 \right\}$$

- no explicit solution
- algorithm : Iterative Reweighted Least Squares

Multiple conditional quantiles

Multiple conditional quantiles

- v covariates $X^1, \ldots, X^v$

Multiple conditional quantiles

- v covariates $X^1, \ldots, X^v$
- model :

$$\begin{aligned} & g_\alpha(X^1, \ldots, X^v) \\ = \; & c + \int_I \Psi_\alpha^1(t) X^1(t)\, dt + \ldots + \int_I \Psi_\alpha^v(t) X^v(t)\, dt \end{aligned}$$

Multiple conditional quantiles

- v covariates $X^1, \ldots, X^v$
- model :

$$
\begin{aligned}
& g_\alpha(X^1, \ldots, X^v) \\
= \ & c + \int_I \Psi_\alpha^1(t) X^1(t)\, dt + \ldots + \int_I \Psi_\alpha^v(t) X^v(t)\, dt
\end{aligned}
$$

- algorithm : *backfitting* + Iterative Reweighted Least Squares

Application to the pollution data

Application to the pollution data

- learning sample : $(X_{l_i}, Y_{l_i})_{i=1,\ldots,n_{learn}}$
- test sample : $(X_{t_i}, Y_{t_i})_{i=1,\ldots,n_{test}}$

Application to the pollution data

- learning sample : $(X_{l_i}, Y_{l_i})_{i=1,\ldots,n_{learn}}$
- test sample : $(X_{t_i}, Y_{t_i})_{i=1,\ldots,n_{test}}$
- number of knots : $k = 8$ (equispaced)
- degree of splines functions : $q = 3$
- order of derivation in the penalization : $m = 2$

Application to the pollution data

- learning sample : $(X_{l_i}, Y_{l_i})_{i=1,\dots,n_{learn}}$
- test sample : $(X_{t_i}, Y_{t_i})_{i=1,\dots,n_{test}}$
- number of knots : $k = 8$ (equispaced)
- degree of splines functions : $q = 3$
- order of derivation in the penalization : $m = 2$
- choice of ρ : Generalized Cross Validation

Quality criteria of the models

Quality criteria of the models

$$C_1 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \widehat{Y_{t_i}})^2}{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \overline{Y}_l)^2}$$

Quality criteria of the models

$$C_1 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \widehat{Y_{t_i}})^2}{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \overline{Y}_l)^2}$$

$$C_2 = \frac{1}{n_t}\sum_{i=1}^{n_t} \mid Y_{t_i} - \widehat{Y_{t_i}} \mid$$

Quality criteria of the models

$$C_1 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \widehat{Y_{t_i}})^2}{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \overline{Y}_l)^2}$$

$$C_2 = \frac{1}{n_t}\sum_{i=1}^{n_t} \mid Y_{t_i} - \widehat{Y_{t_i}} \mid$$

$$C_3 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t} l_\alpha(Y_{t_i} - \widehat{Y_{t_i}})}{\frac{1}{n_t}\sum_{i=1}^{n_t} l_\alpha(Y_{t_i} - q_\alpha(Y_l))}$$

Results (conditional median)

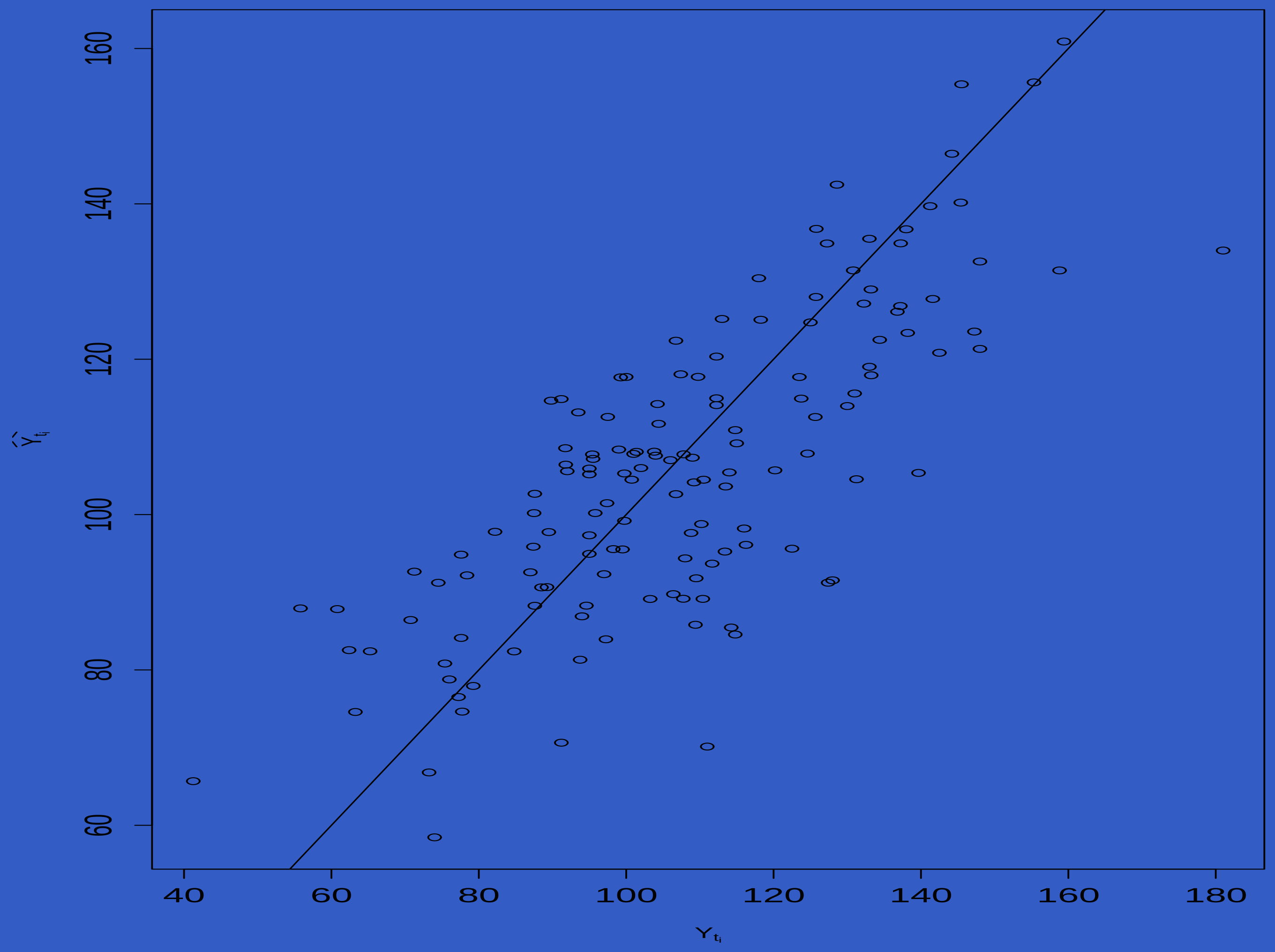
Results (conditional median)

Models	Variables	C_1	C_2	C_3
1 covariate	N2	0.814	16.916	0.906
	O3	**0.414**	**12.246**	**0.656**
	WS	0.802	16.836	0.902
2 covariates	O3, NO	0.413	11.997	0.643
	O3, N2	0.413	11.880	0.637
	O3, WS	0.414	12.004	0.635
3 covariates	O3, NO, N2	0.412	12.127	0.644
	O3, N2, WD	0.409	12.004	0.645
	O3, N2, WS	0.410	11.997	0.642
4 covariates	**O3, NO, N2, WS**	**0.400**	**11.718**	**0.634**
5 covariates	O3, NO, N2, WD, WS	0.401	11.750	0.639

Forecasting (conditional median)

Forecasting (conditional median)

Predicted maximum of Ozone versus measured maximum of Ozone (covariates : O3, NO, N2, WS) :



Conclusion

- satisfying predictions
- improvements : use of other covariates (temperature, . . .)
- outlook : model where X_i is not observed (we observe $W_i = X_i + \delta_i$)