Projet de Data Mining

Prévision des niveaux de polluants dans les rues des métropoles

Clément Berenfeld & Hassan Saber

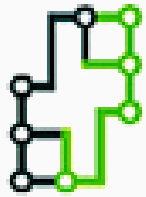
15 mars 2018

Sous la direction de Yannig Goude

Table des matières

Présentation des données

Description des données



Challenge data

plume LABS

Les jeux de données (entraînement et test) sont constitués des observations de :

- **28 stations**
- **6 villes** (inconnues)
- **18 meta-variables**
- **3 polluants** (NO2, PM10 et PM2.5)
- fréquence **horaire**, pendant **un an et demi**
- relevés **météorologiques**

→ **Critère** : RMSE

Visualisation

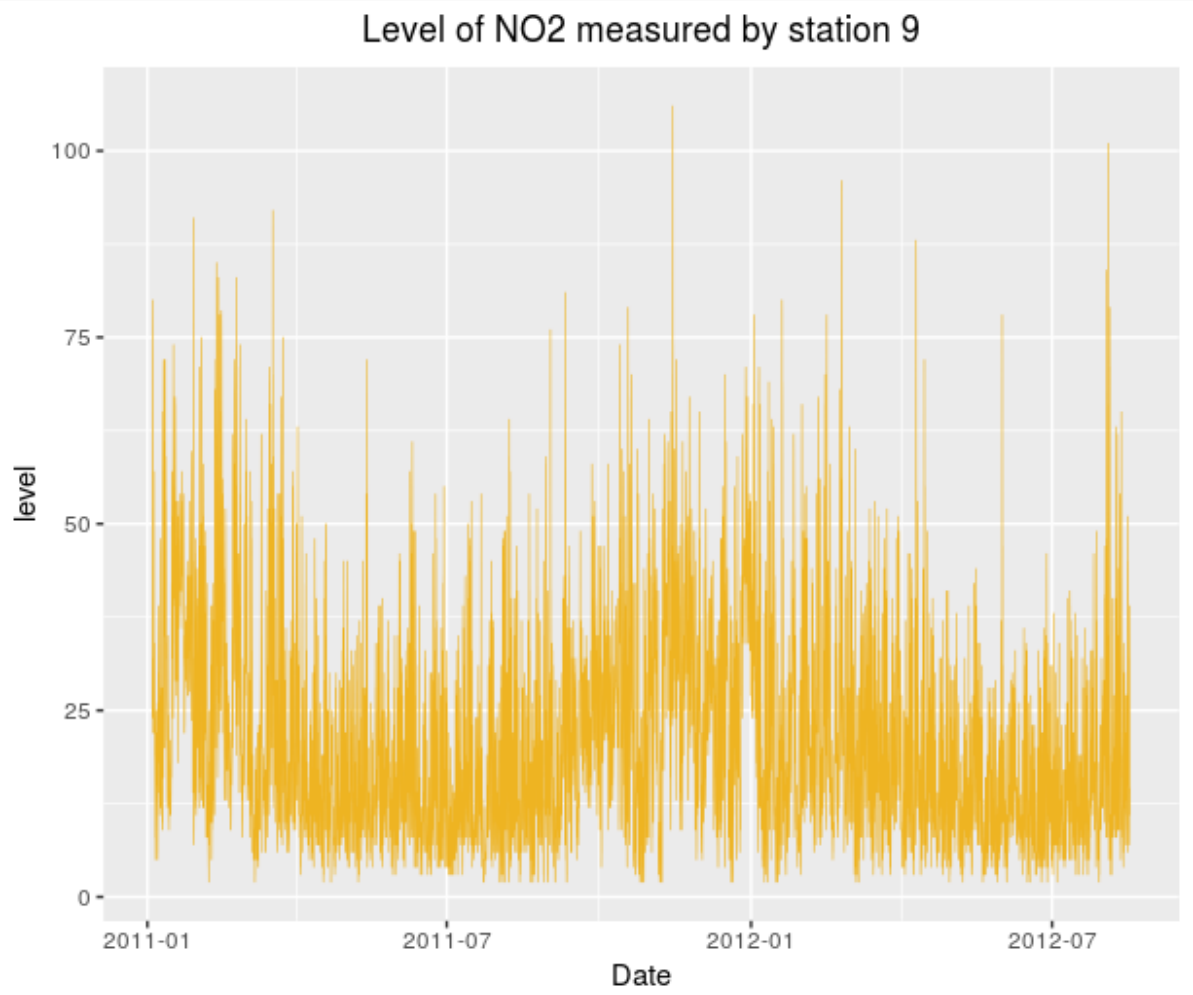


Figure 1 – Enregistrement du niveau de NO2 de la station 9

Saisonnalités

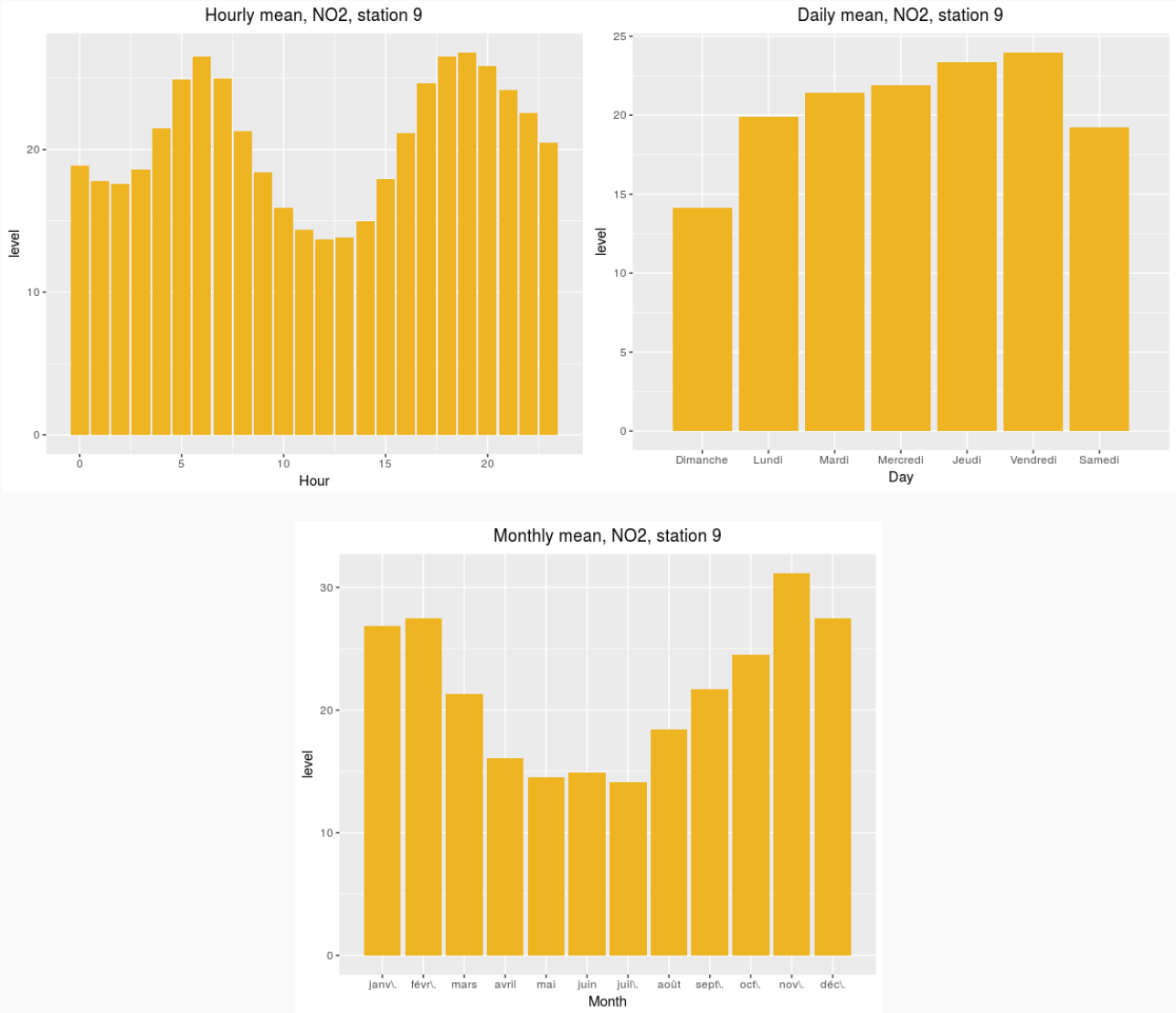
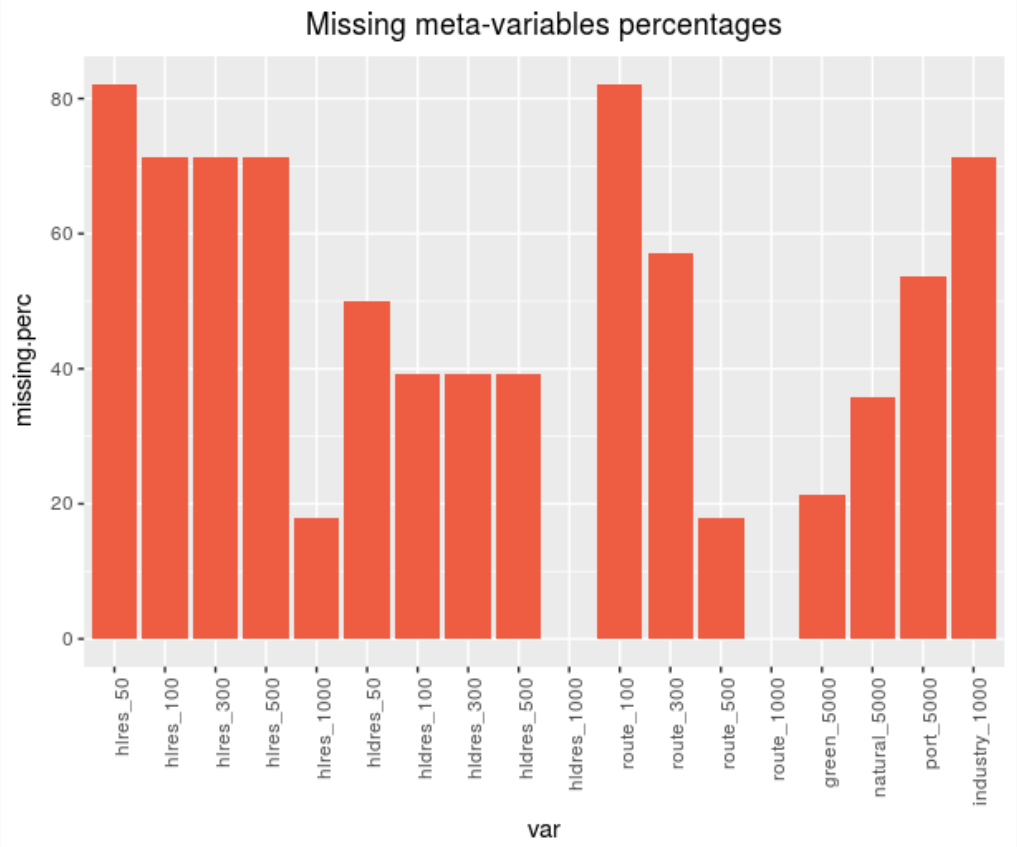


Figure 2 – Les différentes saisonnalités pour le NO2, station 9.

Complétion des données manquantes

Complétion



- Tous les relevés météo / polluants sont présents
- Mais un grand nombre de méta-données absentes ($> 45\%$)

Complétion

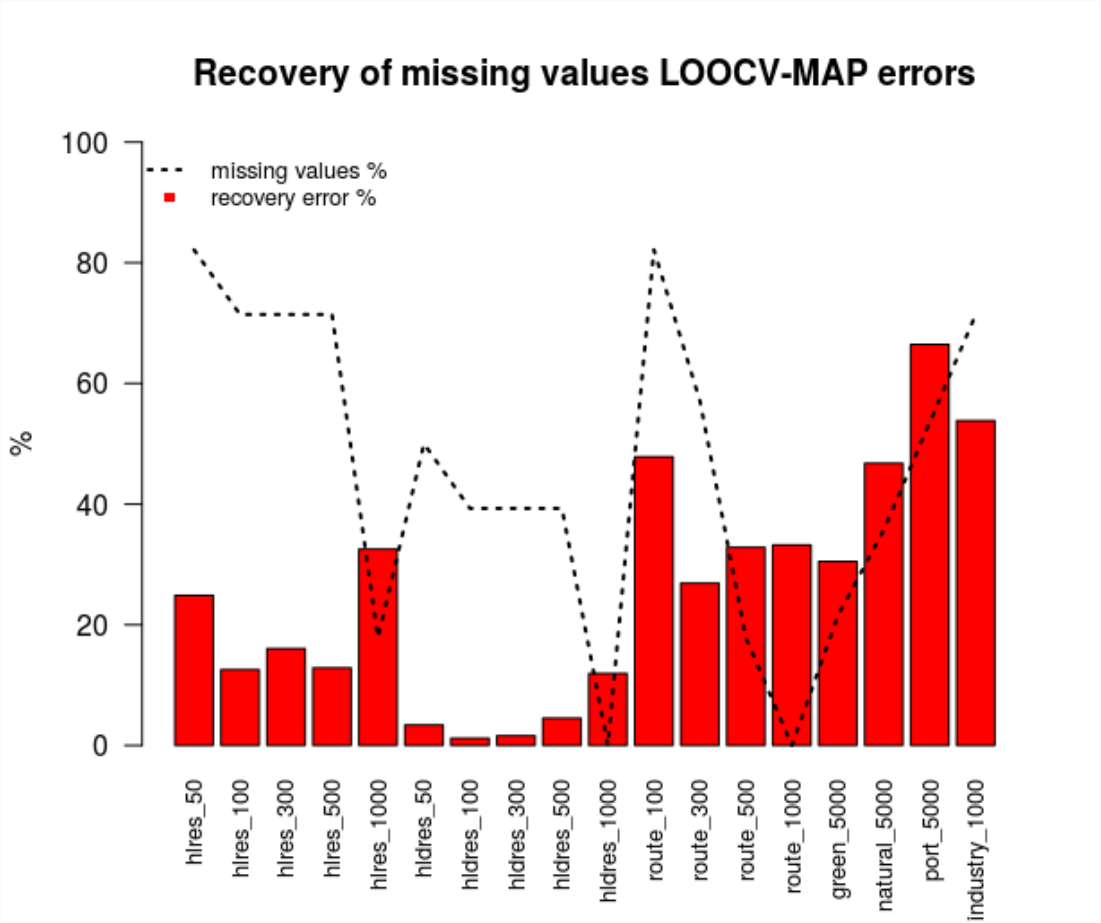


Figure 3 – Erreur de CV pour la reconstruction effectuée avec `missRanger`.

Prediction à long terme

Prediction à long terme

Modèles GAM uni-stations

Modèles GAM

Cadre

- Station fixée, polluant fixé (donc méta-variables fixées)
- Prédire le taux de pollution `level` à partir des autres variables

Modèle :

Ajuster un modèle GAM en intégrant séquentiellement les variables.
$\rightarrow$ Gain en **complexité**

Modèles GAM

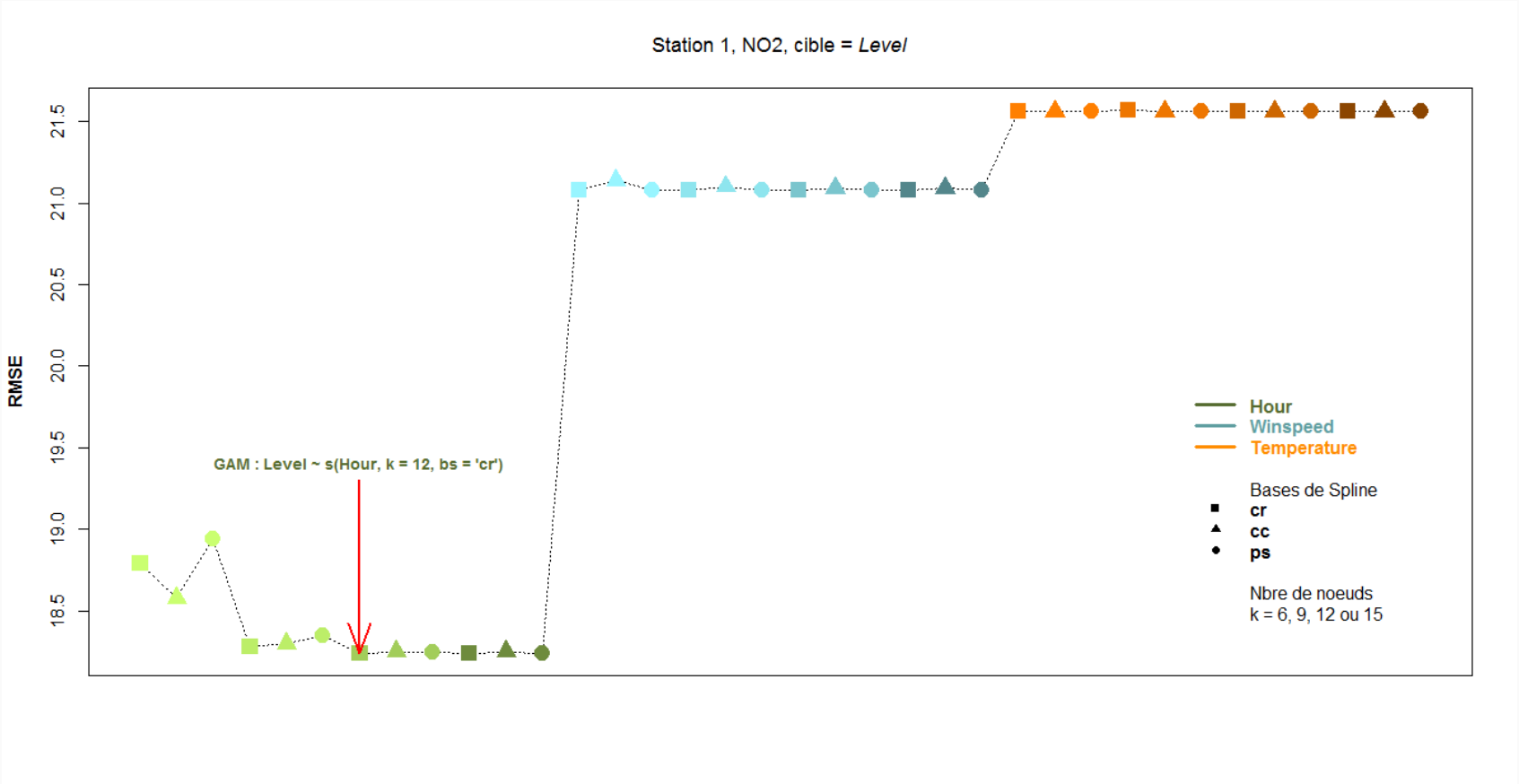


Figure 4 – Sélection et paramétrage de la première variable du modèle GAM pour la station 1

Modèles GAM

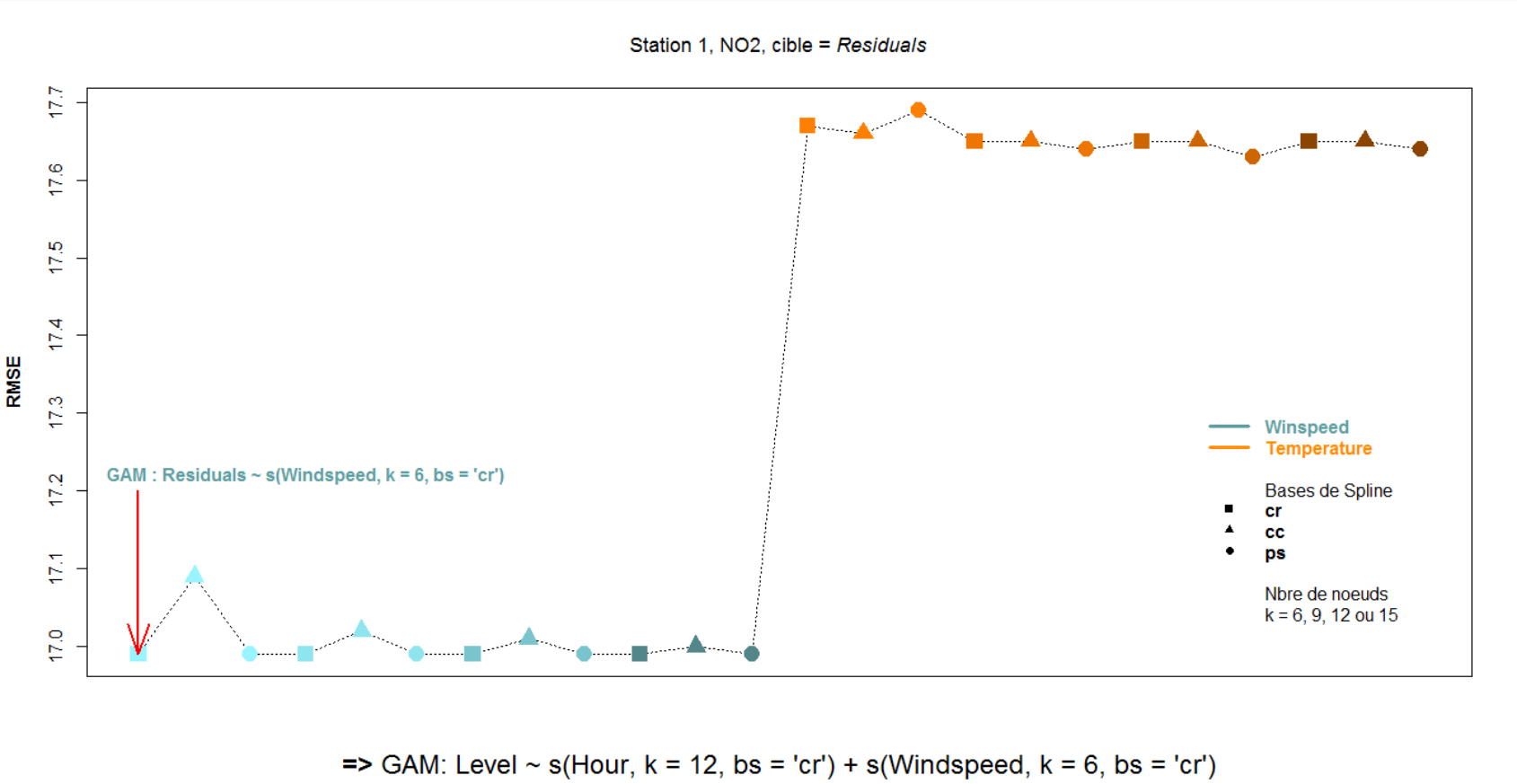


Figure 5 – Sélection et paramétrage de la deuxième variable du modèle GAM pour la station 1

Modèles GAM

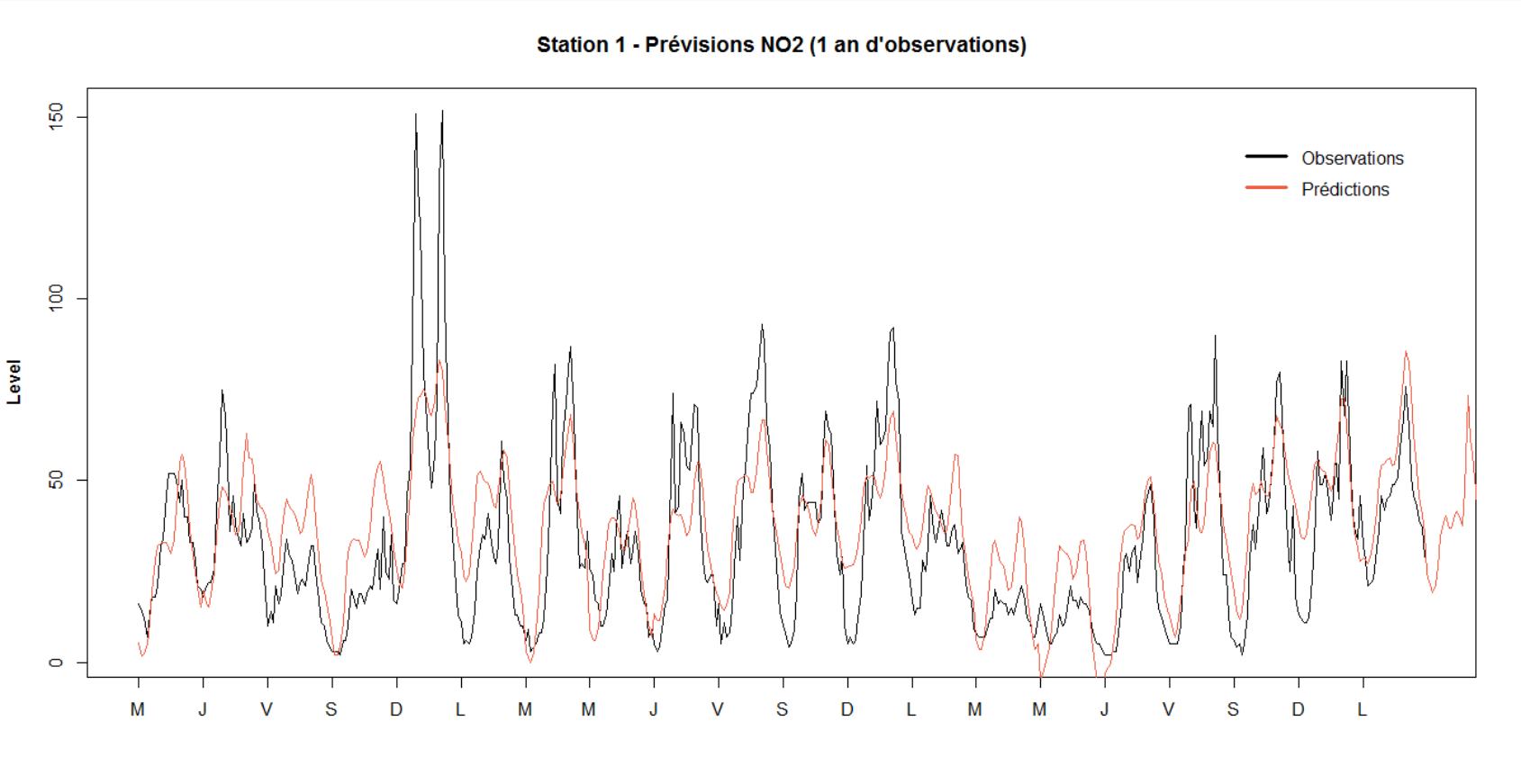


Figure 6 – Prédictions du taux de NO2 pour l'année 2012 à partir de l'année 2011

Modèles GAM

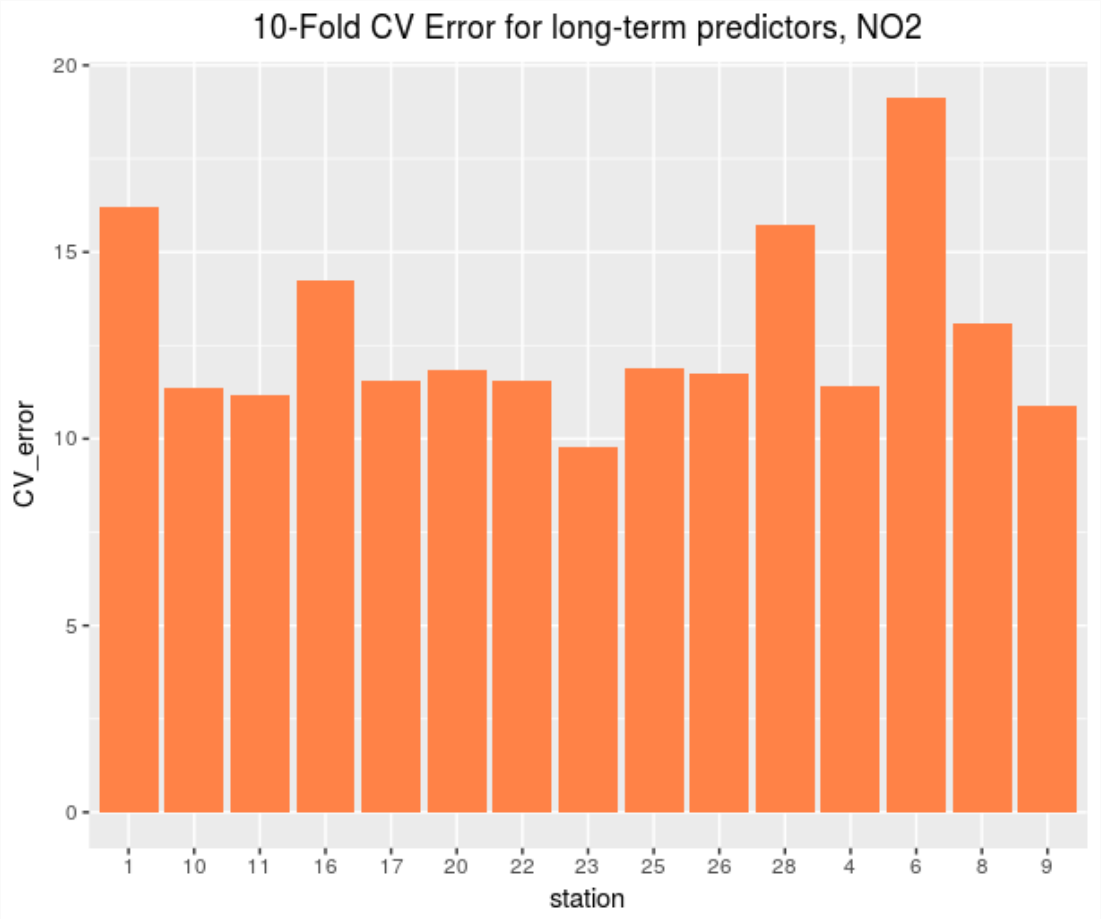


Figure 7 – Erreur CV 10-fold sur les modèles GAM pour le polluant NO2

Prediction à long terme

Agrégation des modèles uni-stations

Agrégation des modèles uni-stations

Notations :

- $\mathcal{S}$: l'ensemble des stations
- Ω : l'ensemble des variables
- $\mathcal{S}_{\text{train}} \subset \mathcal{S}$ les stations d'entraînement
- $\mathcal{S}_{\text{test}} \subset \mathcal{S}$ les stations tests

Modèle :

$$Y_i = f_s(X_i) + \epsilon_i$$

On dispose de $\{\widehat{f_s}\}_{s \in \mathcal{S}_{\text{train}}}$ grâce aux modèles GAM .

$\rightarrow$ On aimerait trouver $\{\widehat{f_u}\}_{u \in \mathcal{S}_{\text{test}}}$.

Agrégation des modèles uni-stations

Idée : Pour $u \in \mathcal{S}_{\text{test}}$:

$$\widehat{f}_u = \sum_{s \in \mathcal{S}_{\text{train}}} W_s(\mathcal{S}_{\text{train}}, u) \widehat{f}_s$$

avec

$$W_s(\mathcal{S}_{\text{train}}, u) = \frac{K_h(d_{\mathcal{S}}(s,u))}{\sum_{s' \in \mathcal{S}_{\text{train}}} K_h(d_{\mathcal{S}}(s',u))}$$

où K_h est un noyau et $d_{\mathcal{S}}$ une distance sur $\mathcal{S}$.

Agrégation des modèles uni-stations

Idée : Pour $u \in \mathcal{S}_{\text{test}}$:

$$\widehat{f}_u = \sum_{s \in \mathcal{S}_{\text{train}}} W_s(\mathcal{S}_{\text{train}}, u) \widehat{f}_s$$

avec

$$W_s(\mathcal{S}_{\text{train}}, u) = \frac{K_h(d_{\mathcal{S}}(s,u))}{\sum_{s' \in \mathcal{S}_{\text{train}}} K_h(d_{\mathcal{S}}(s',u))}$$

où K_h est un noyau et $d_{\mathcal{S}}$ une distance sur $\mathcal{S}$.

$\rightarrow$ Quelle distance choisir sur $\mathcal{S}$?

Agrégation des modèles uni-stations

On choisit :

$$\forall s, s' \in \mathcal{S},\ \ d_{\mathcal{S}}(s, s') = \|f_s - f_{s'}\|_{L^2(\Omega)}$$

Estimation : Sur $\mathcal{S}_{\text{train}}$, par méthode MC

$$\forall s, s' \in \mathcal{S}_{\text{train}},\ \ \widehat{d}_{\mathcal{S}}(s, s') = \|\widehat{f}_s - \widehat{f}_{s'}\|_{L^2(\Omega)}$$

Extension : On fait l'hypothèse que

$$\forall s, s' \in \mathcal{S},\ \ d_{\mathcal{S}}(s, s') = \sum_{i=1}^{q} \lambda_i d_i(s_i, s'_i)$$

Agrégation des modèles uni-stations

Estimation des coefficients

On note $n = \mathrm{Card}(\mathcal{S}_{\text{train}})$, $N = n(n-1)/2$ et on pose :

$$X = (d_i(s_i, s_i'))_{(s,s)',i} \in \mathbb{R}^{N \times q}$$

$$Y = (\widehat{d}_{\mathcal{S}}(s, s'))_{(s,s')} \in \mathbb{R}^{N}$$

Et on résout ce problème quadratique grâce à la librarie `quadprog` :

$$\begin{aligned} &\text{minimiser} \quad \|Y - X\Lambda\|^2 \\ &\text{sous contraintes} \quad \Lambda \geq 0 \end{aligned}$$

Choix de h : Par validation croisée LOO sur $\mathcal{S}_{\text{train}}$.

Agrégation des modèles uni-stations

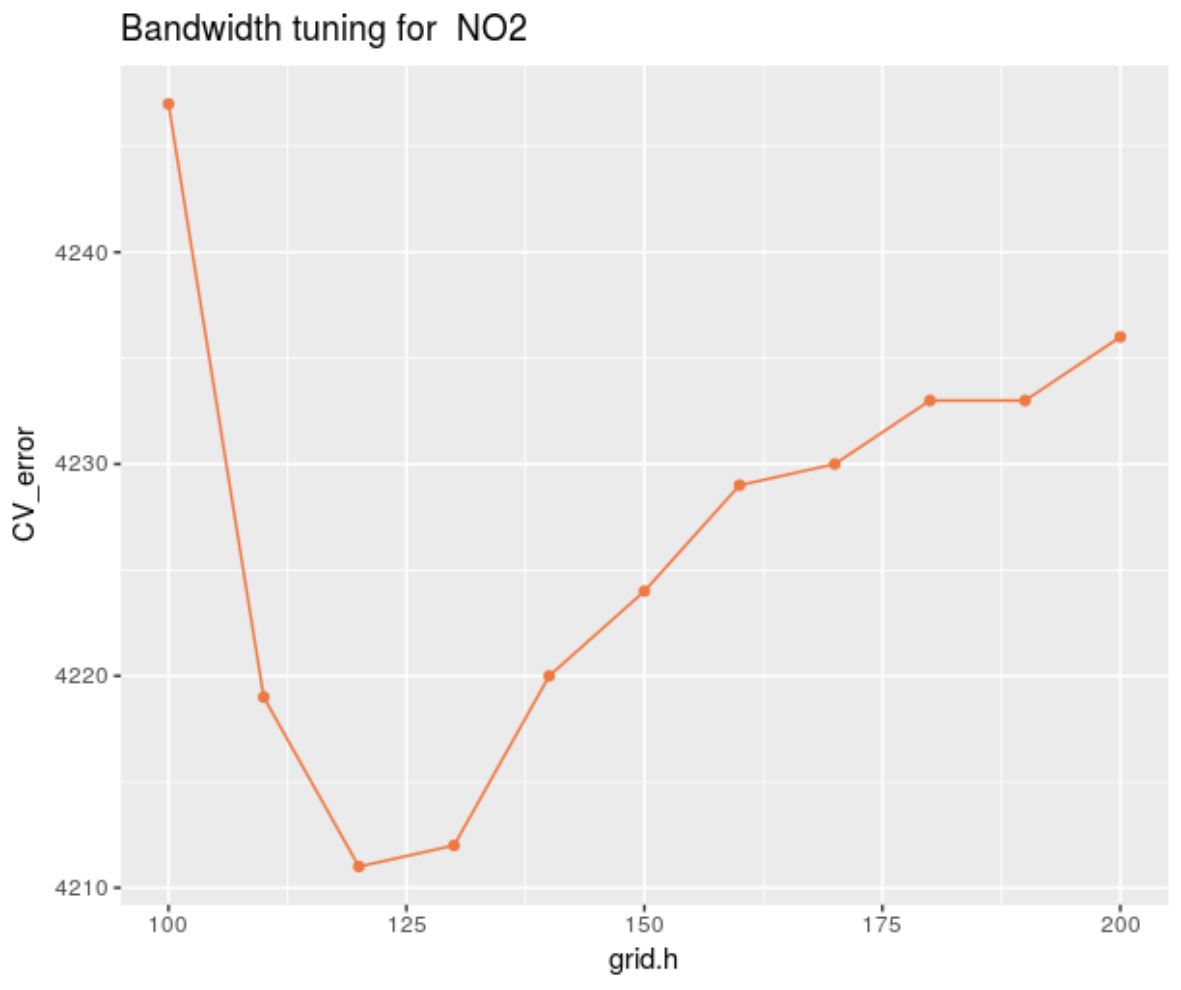


Figure 8 – Résultat du réglage du paramètre h pour NO2 (noyau triangle)

Agrégation des modèles uni-stations

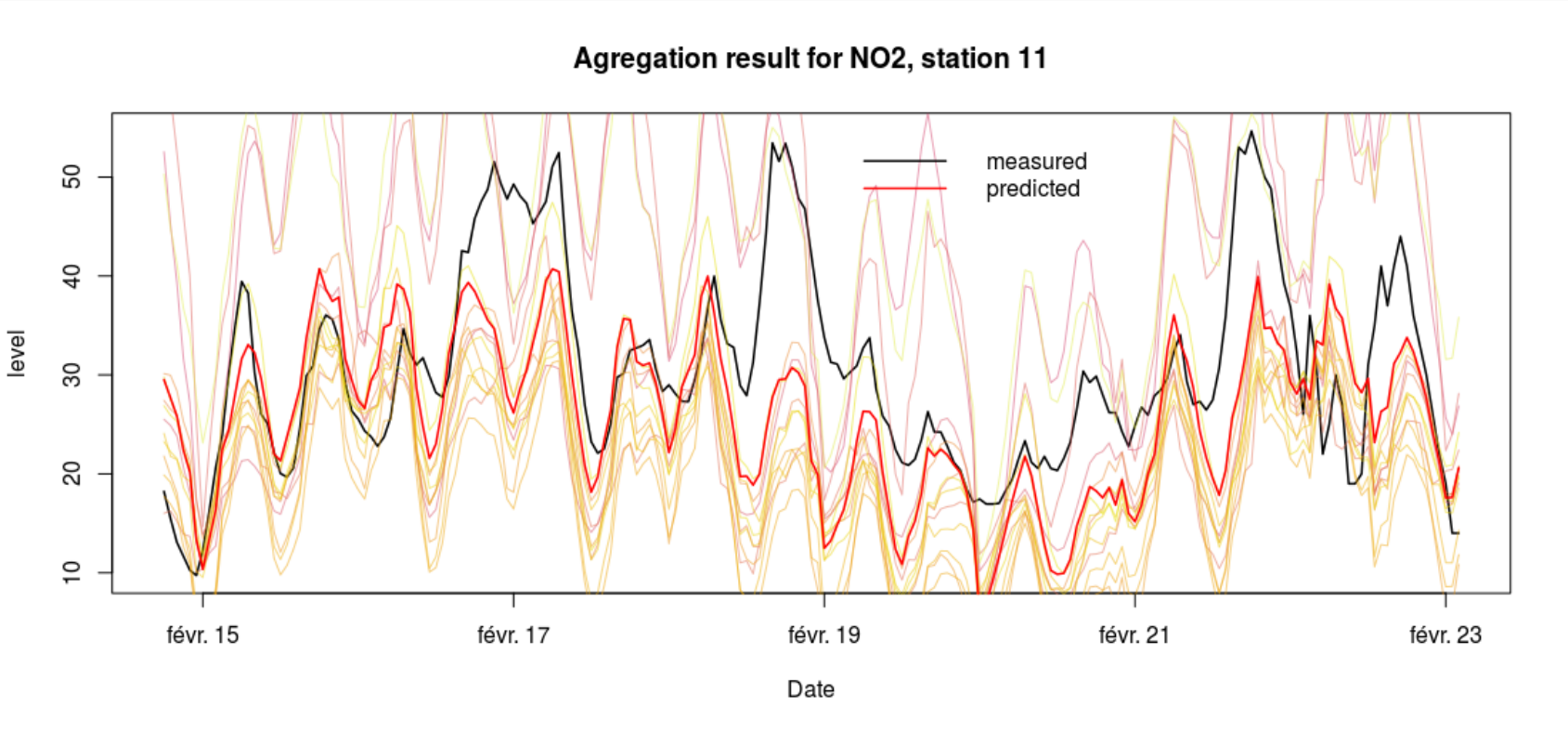


Figure 9 – Mélange des prédicteurs pour la station 11, avec la largeur de bande optimale. En noir, les vrais relevés, en rouge, les estimations agrégées des prédicteurs.

Agrégation des modèles uni-stations

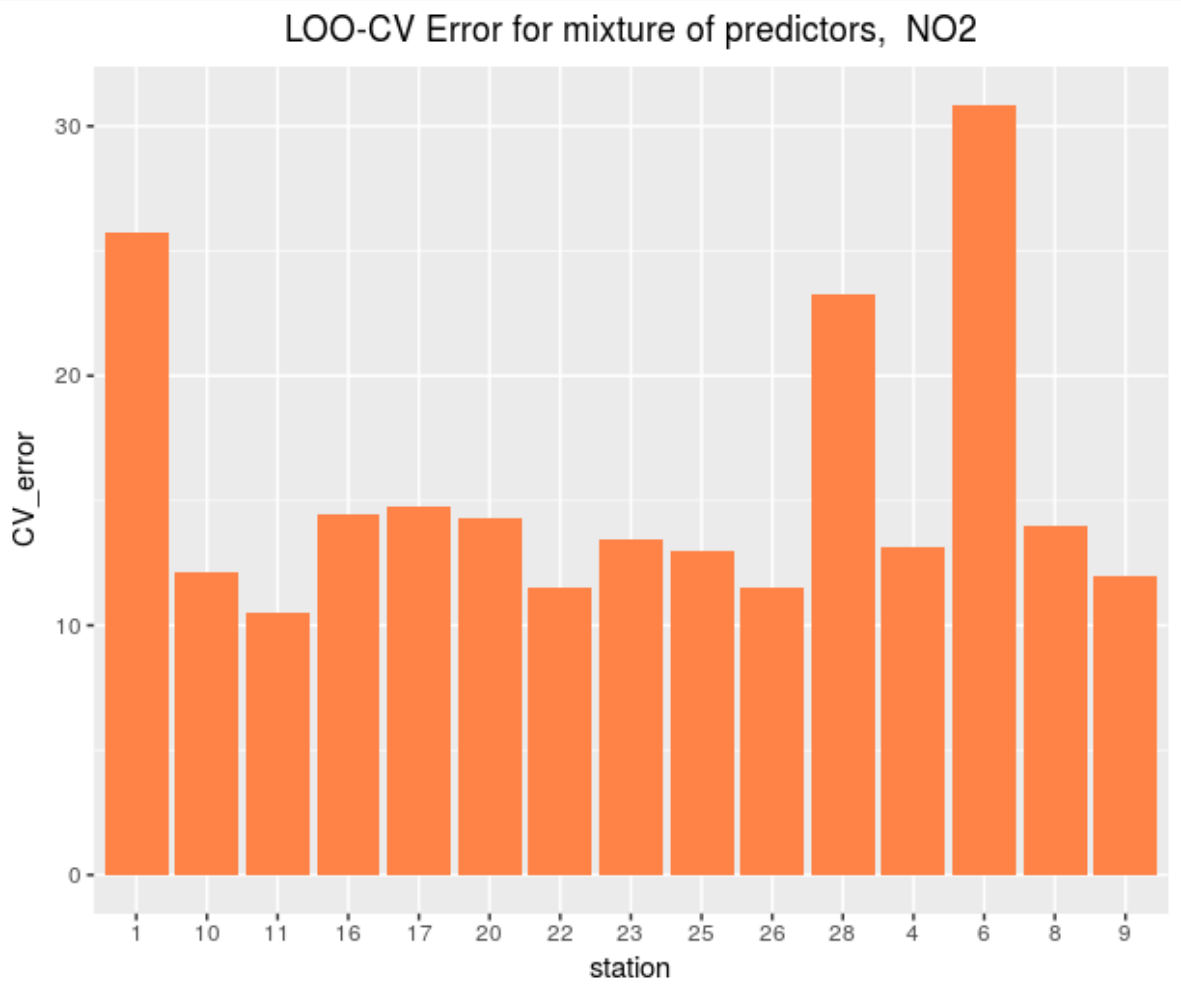


Figure 10 – Erreur d'estimation par CV-LOO pour le modèle de mélange

Prédiction à court terme

Modèles uni-stations

On rajoute un champ `level1` puis on entraîne les modèles GAM :

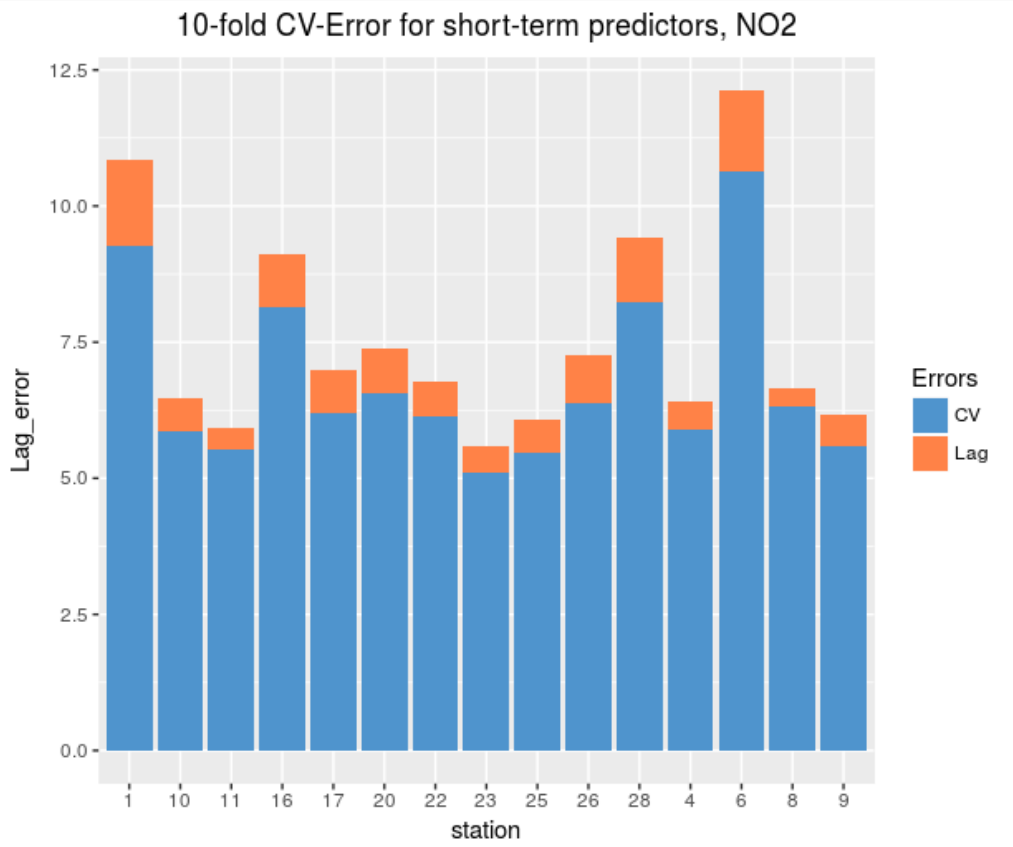
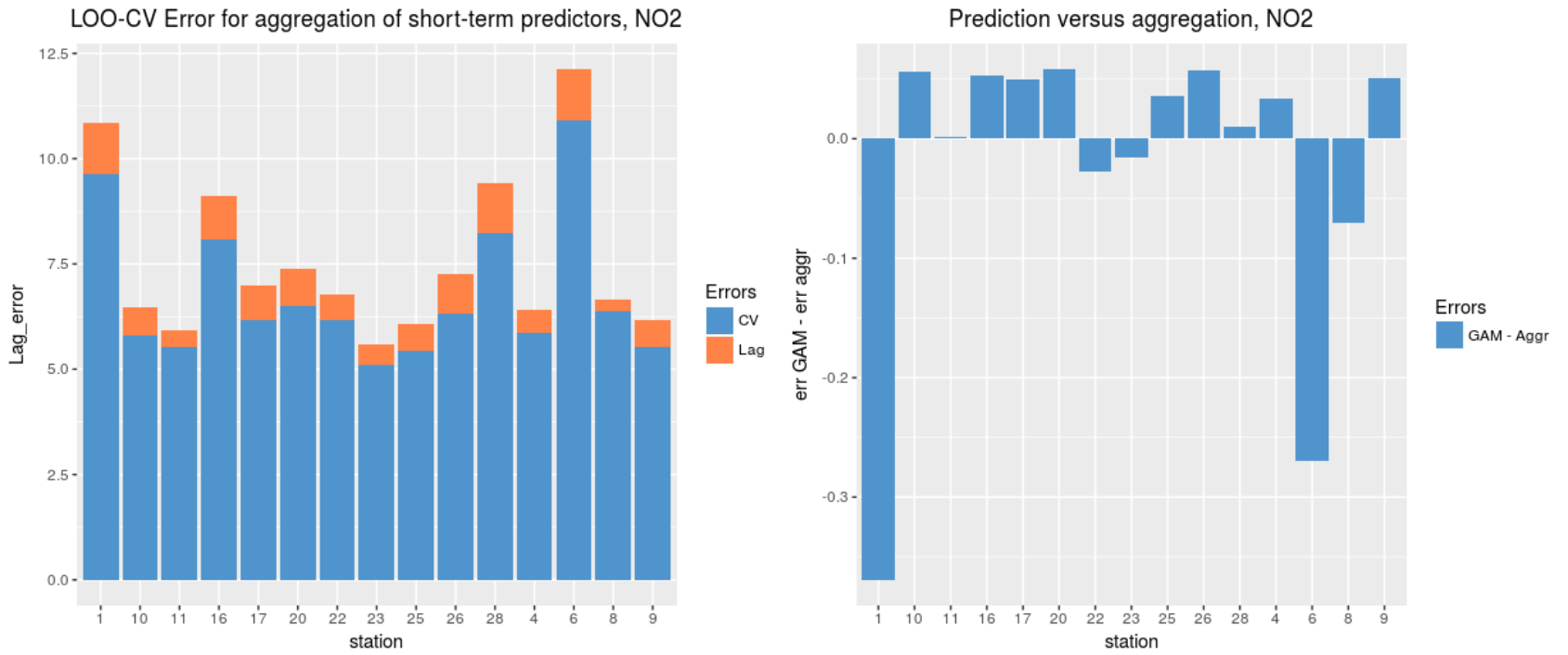


Figure 11 – Erreur d'estimation pour les prédicteurs court-terme

Agrégation des prédicteurs

Puis agrégation sequentielle des prédicteurs avec `opera` (modèle : `MLpol`)



Conclusion

Conclusion

Les différentes approches et leurs résultats sont résumés dans le tableau ci-dessous.

		RMSE	MAPE (%)
Prédiction court-terme	Experts (CV)	6.23	24.71
	Mélanges (CV)	6.18	24.8
	Challenge	433	-
Prédiction long-terme	Experts (CV)	12.73	63.99
	Mélanges (CV)	14.27	85.78
	Challenge	275	-
Forêts aléatoires	Challenge	493	-
	Benchmark	501	-
	1^{ers} du challenge	192	-


Merci pour votre attention !
Des questions ?