INSA – MS ESD Année 2019-2020
ROUFFIAC Jean-Eudes

STAT2 - Regression penalisée

Rapport de TP n°4

Titre : *Modèle Linéaire Gaussien Multiple*
Prévision de la pollution de l'air par les particules

1 Introduction

L'objectif de ce TP est de manipuler les instructions de base permettant de faire une régression linéaire multiple et d'expérimenter les méthodes de sélection de variables ainsi que de validation de modèle. C'est dans cette optique que nous cherchons à connaître l'influence de certaines variables sur la pollution par les particules fines PM10. Pour ce faire, le contexte suivant est donné : on s'intéresse à une station de mesures d'Atmo Normandie située dans l'agglomération du Havre sur la période 2007-2015. Il s'agit de la station HRI, station de fond urbain. Les données sont de deux types : des concentrations de polluants (PM10, NO, NO2, O3 et SO2) et des mesures de paramètres météorologiques (température, pression atmosphérique, humidité relative, vitesse et direction de vent, gradient température, pluviométrie).

Afin d'anticiper des éventuels dépassements de seuil, on souhaite modéliser la concentration journalière moyenne en PM10 à l'aide d'un modèle linéaire utilisant tout ou partie des variables explicatives disponibles.

2 Etude statistique

2.1 Statistique descriptive

A l'aide de la commande summary de R, nous pouvons avoir des informations sur les variables à notre disposition. Nous pouvons alors mettre en évidence deux choses :

- Il y a un nombre conséquent de variables manquantes, qu'il faudra donc enlever avant de construire notre modèle
- En analysant les valeurs minimales, maximales et moyennes, on s'aperçoit que les variables ont des ordres de grandeurs différents. Pour améliorer les performances du modèle, il sera alors grandement recommandé de centrer et réduire les données.

2.2 Corrélation

Avant d'introduire quelconque modèle, il est nécessaire de s'intéresser aux possibles corrélations entre les variables explicatives du jeu de données, pour ne garder que des variables explicatives non corrélées.
Dans un premier temps, on trace le graphique représentant la corrélation des variables deux à deux :

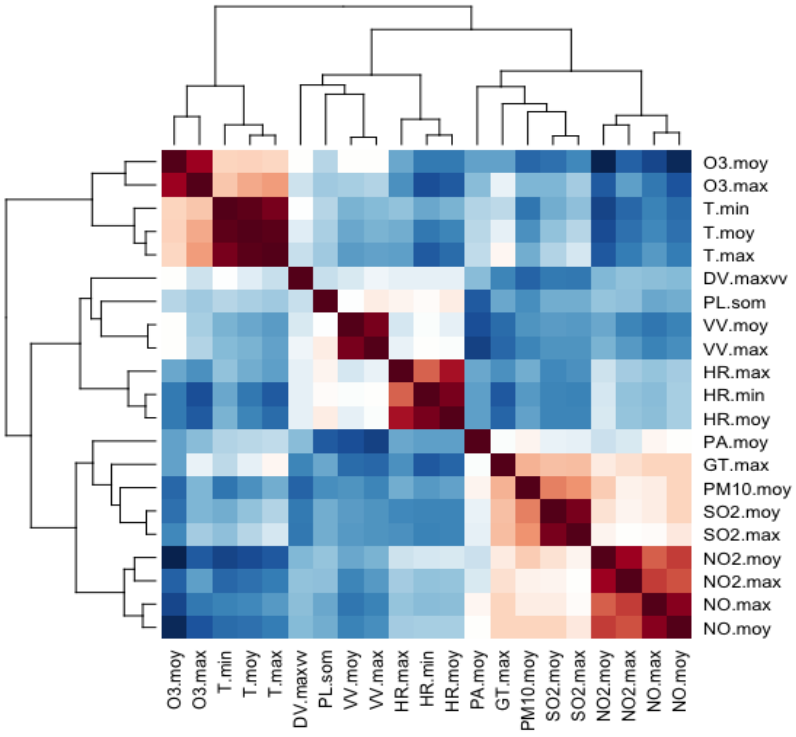

FIGURE 1 – Matrice de corrélation

Une forte corrélation sera représentée par du rouge foncé, une faible corrélation par du bleu. Ainsi nous pouvons voir à l'oeil nu sur la figure ci dessus que les variables T.max-T.moy, T.min-T.moy, S02.max-SO2.moy, T.min-T.max, HR.min-HR.moy, VV.max-VV.moy

semblent corrélées deux a deux. Via le logiciel R avec la fonction "rquery.cormat", on peut alors afficher la corrélation entre les variables citées précédemment :

- Corrélation(T.max,T.moy) = 0.98
- Corrélation(T.min,T.moy) = 0.97
- Corrélation(S02.max,SO2.moy) = 0.92
- Corrélation(T.min,T.max) = 0.92
- Corrélation(HR.min,HR.moy) = 0.92
- Corrélation(VV.max,VV.moy) = 0.92

Ces résultats indiquent un lien très important entre les variables T.max et T.moy ainsi que T.min et T.moy c'est-à-dire entre les températures maximales et moyennes, ansi que les températures minimales et moyennes. Ces corrélations sont logiques car si des températures augmentent, la moyenne augmente aussi et vice-versa.
Les autres coefficients de corrélations, bien qu'ils soient importants, ne permettent pas d'écarter pour le moment une variable.
Ainsi, après cette analyse, nous allons envisager de nous séparer de la variable T.max, T.moy ou bien encore T.min. Pour valider ce choix, nous utilisons le VIF (Variance Inflation Factor) pour étudier la multi-colinéarité. La fonction VIF sur R nous donne les résultats suivants :

Variable	SO2.max	O3.max	NO.max	NO2.max	SO2.moy	O3.moy	NO.moy
Valeur	7.228970	7.156059	6.652925	6.223753	7.567991	7.722383	7.062206

Variable	NO2.moy	T.min	T.max	T.moy	HR.min	HR.max	HR.moy
Valeur	7.178599	58.947907	79.807114	208.807318	10.657762	4.462666	17.647311

Variable	VV.max	VV.moy	PL.som	PA.moy	GT.max	DV.maxvv
Valeur	7.577536	7.519881	1.282650	1.417535	2.005820	1.333490

Une valeur supérieure à 20 indique qu'une variable est fortement liée avec les autres variables. C'est le cas de T.moy, T.max et T.min. Puisque T.moy est la variable pour laquelle l'indicateur VIF est le plus important (plus de 200), c'est cette variable qu'on préfère retirer du modèle par la suite. Nous avons réitérer l'étude de multicolinéarité avec l'indicateur VIF en enlevant la variable T.moy pour vérifier que la valeur du VIF des variables restantes est en dessous de 20.

2.3 Modèle statistique

Dans cette partie on souhaite expliquer la variable d'intérêt Y, la contentration moyenne en PM10, par une liaison linéaire à partir des différentes variables explicatives. On introduit donc le modèle linéaire gaussien multiple, c'est à dire que l'on suppose que les données (y_j) sont les réalisations des variables aléatoires (Y_j) liées aux (x_{ji}) avec $i \in \{1; 2; ...; 20\}$ et $j \in \{1; 2; ...; 2493\}$ par la relation :

$$Y_j = \mu + \sum_{i=1}^{20} \alpha_i x_{j,i} + \epsilon_j$$

où :

- μ et α_i sont des paramètres réels inconnus,
- les erreurs (ϵ_j) sont des variables aléatoires que l'on suppose indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$

L'étude précédente nous a permis d'exclure la variable X11 (T.moy). Ainsi nous reprenons le modèle ci-dessus mais avec $i \in \{1; 2; ...; 20\} - \{11\}$.

On découpe aléatoirement l'ensemble des données en deux jeux de données : un échantillon d'apprentissage et un échantillon test. L'échantillon d'apprentissage servira à construire le modèle, qui sera ensuite utilisé avec les données de l'échantillon test pour évaluer sa capacité à prédire la variable à expliquer.

On calcule alors les coefficients estimés du modèle en se servant de l'échantillon d'apprentissage :

- μ = -1.571620e+02
- α_1 = 8.068645e-03
- α_2 = 1.219873e-01
- α_3 = -1.414175e-02
- α_4 = -1.096986e-01
- α_5 = 4.860253e-01
- α_6 = -9.830267e-02
- α_7 = 6.039496e-02
- α_8 = 2.910277e-01
- α_9 = -9.582354e-01
- α_{10} = 4.382731e-01
- α_{12} = -3.267914e-02
- α_{13} = 3.202387e-02
- α_{14} = 1.315495e-01
- α_{15} = -4.553983e-01
- α_{16} = 6.656842e-01
- α_{17} = -2.646592e-01
- α_{18} = 1.638161e-01
- α_{19} = 1.893967e+00
- α_{20} = -1.293130e-02

On va maintenant étudier la qualité de l'ajustement linéaire des données. On effectue tout d'abord un test de non régression pour savoir si les 19 variables explicatives ont une réelle influence sur Y.
On teste donc l'hypothèse nulle H_0 :" $\alpha_1 = \alpha_2 = ... = \alpha_{20} = 0$"
contre l'hypothèse alternative H_1 :" $\exists i \in \{1; 2; ...; 20\} - \{11\}$ tel que $\alpha_i \neq 0$" .

Le logicel R nous donne une p-value de $2.2.10^{-16}$, soit une p-value inférieure à 5%. On rejette donc l'hypothèse nulle au risque 5% et on en déduit qu'au moins une des variables explicatives a une influence significative sur la concentration moyenne.
De plus, la part de variance expliquée par le modèle de 52.95% et celle ajustée de 52.44%, nous montre que la qualité de la régression est moyenne. L'analyse des résidus montre qu'ils ne sont pas centrés. Enfin, on observe que pour certains coefficients, la p-value est supérieure à 5% et que donc ils n'ont pas d'influence significative sur la variable à prédire.

On veut maintenant ne garder dans le modèle que les régresseurs qui ont une influence réelle (significative) sur la variable à expliquer. Nous utiliserons pour cela la méthode de sélection de variables pas à pas descendante (backward regression).

Dans cette méthode, on part du modèle complet (les 19 variables) et à chaque étape on élimine la variable la moins significative au risque 5%. Pour ce faire on calcule les p-values associées à chaque variable et on cherche la plus grande de ces p-values qui est supérieure à 5%. Si elle existe, on retire alors cette variable du modèle. On réitère ce schéma jusqu'à ne plus pouvoir éliminer de variables. En appliquant cette méthode, on supprime successivement les variables suivantes :

- HR.min de p-value 0.818062
- HR.moy de p-value 0.748551

- NO.max de p-value 0.118567
- SO2.max de p-value 0.117025

On arrive donc à un nouveau modèle gardant 15 variables explicatives. On peut alors expliquer 52.41% de la variance. On a donc conservé le même pourcentage de variance expliquée que le modèle initial en enlevant 4 variables explicatives. Les performances entre les deux modèles sont donc sensiblement les mêmes. En effectuant un test de Fischer, on obtient une p-value de plus de 5%, ce qui permet donc d'accepter H_0 au risque 5% et donc d'affirmer au risque 5% que les variables enlevées du modèle ont bien raison d'être enlever.

Observons maintenant sur le graphique ci dessous la concentration moyenne en PM10 estimée par le modèle reslm2 en fonction de la concentration réelle observée :

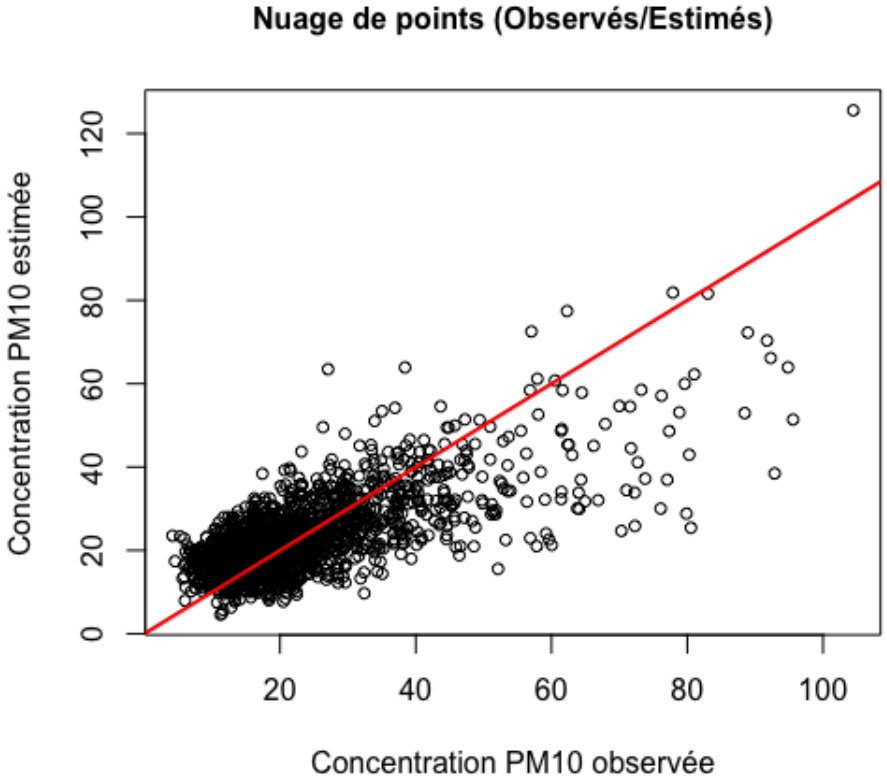

FIGURE 2 – Nuage de points des valeurs estimées en fonction des valeurs observées.

Le nuage de points observés/estimés montre la qualité de la régression, c'est à dire moyenne. On remarque que de nombreux points sont assez mal estimés, notamment pour de fortes concentrations de PM10. Le modèle a d'ailleurs tendance à estimer à la baisse lorsque la concentration moyenne de PM10 est élevée et inversement, le modèle sur-estime pour de faibles concentrations.

Nous étudions maintenant le graphe des résidus. Avec le logiciel R, nous traçons le graphe des résidus et le graphe des résidus studentisés.

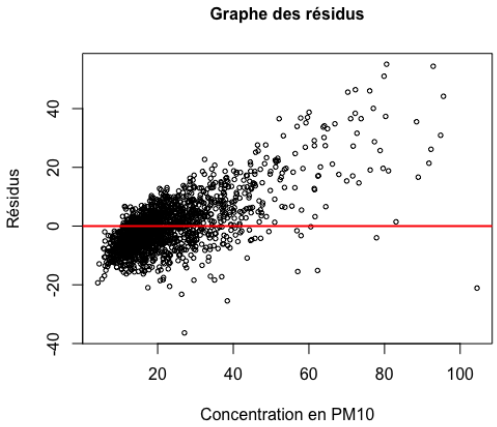

FIGURE 3 – Graphe des résidus

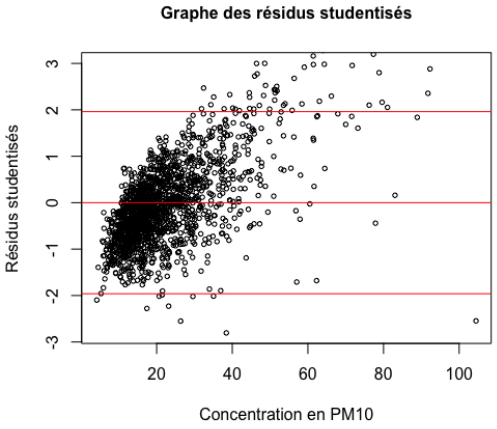

FIGURE 4 – Graphe des résidus studentisés (seuils à 95%)

On peut observer sur le graphe des résidus que les points sont répartis de manière équitable autour de 0 pour de faible/moyenne concentration, mais les résidus pour les fortes concentrations sont élevés (+ de 40 pour certains résidus). Nous observons une augmentation de l'amplitude des résidus en fonction de la concentration moyenne de PM10, l'hypothèse d'homoscedasticité n'est donc pas vérifiée (confirmé par le test de Breusch-Pagan qui donne une p-value plus petite que 2.2e-16) . Le graphe des résidus studentisés permet de s'interroger sur le caractère aberrant ou non de certaines données. En effet, on observe que les résidus pour le certaines données sont en dehors de bande de confiance à 95%.

Nous testons maintenant la normalité des résidus par le biais du test de Shapiro-Wilk. Nous obtenons une p-value très inférieure à 5% (p-value $<$ 2.2e-16). Ainsi on rejette au risque de 5% l'hypothèse que les résidus suivent une loi normale. On peut vérifier ce résultat visuellement avec l'histogramme des résidus ci-dessous ainsi que le QQplot :

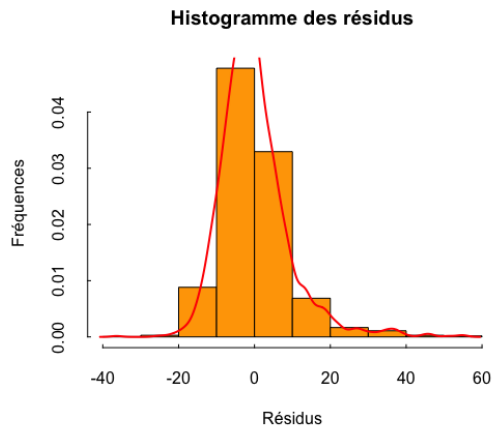

FIGURE 5 – Histogramme des résidus

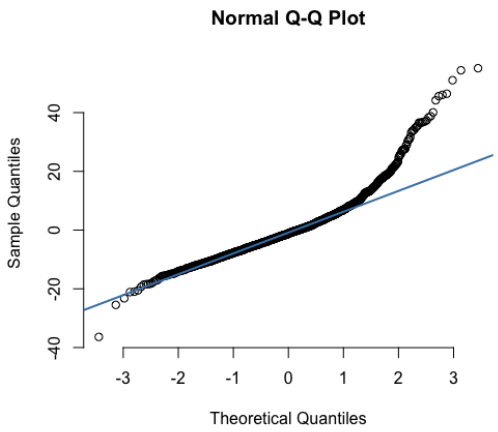

FIGURE 6 – Normal Q plot

Enfin en effectuant le test de Durbin-Watson, le logiciel R donne une p-value égale à 0 et donc inférieure à 5%. Ceci nous permet de rejeter l'hypothèse de non corrélation des résidus et donc de considérer que les résidus sont corrélés.
En conclusion, au vu de ces dernières observations nous pouvons dire que l'ajustement n'est pas de bonne qualité et qu'il faudrait donc mieux s'orienter vers un autre modèle.

Nous allons maintenant tester la commande *step* appliquée au modèle initial (reslm1). Cette commande permet d'éliminer les variables explicatives pas à pas suivant le critère d'information Akaike ou AIC. Avec cette méthode, seulement deux variables sont élminées : HR.min et HR.moy. On obtient un pourcentage de variance expliquée très légèrement meilleure que les deux précédents modèles (52.49%). Le tracé des différents graphiques montre les mêmes caractéristiques que le modèle reslm2.

Nous allons maintenant comparer nos trois précédents modèles en apprentissage et en test en regardant la valeur du RMSE et MAE pour chaque modèle. Nous regroupons les résultats dans le tableau suivant :

Echantillon	**Apprentissage**			**Test**		
Modèle	reslm1	reslm2	reslm3	reslm1	reslm2	reslm3
RMSE	9.207177	9.220645	9.207593	9.744864	9.722622	9.747771
MAE	6.541772	6.544573	6.539635	7.022397	7.018797	7.020748

Nous pouvons remarquer que les résultats sont similaires. Nous observons une légère dégradation pour les 3 modèles lorsque l'on passe de l'échantillon d'apprentissage à l'échantillon test, mais cela est acceptable. Nous ne sommes donc pas en présence de sous ou sur-apprentissage. A la vue de ces résultats, nous allons choisir comme meilleur modèle le modèle 2, car c'est celui qui contient le moins de variables explicatives et qui donne le meilleur résultat sur l'échantillon test.

On nomme alors "reslm" le modèle 2. Évaluons ses performances en estimation sur l'échantillon d'apprentissage et en prévision sur l'échantillon test. Pour ce faire, nous allons

dans un premier temps calculé les indicateurs classiques de performance en estimation sur notre échantillon d'apprentissage et en prévision sur notre échantillon test. Les résultats sont représentés dans le tableau ci-dessous.

Erreurs	Estimation	Prévision
R	0.73	0.67
EV	0.53	0.45
MAE	6.54	7.02
RMSE	9.22	9.72

R représente la corrélation entre les valeurs observées et celles estimées/prévues. En estimation la valeur est de 0.73 contre 0.67 pour l'échantillon test en prévision. Ces valeurs sont assez basses, notons une légère dégradation des performances lorsque l'on passe de l'échantillon d'apprentissage à l'échantillon test. Le MAE (moyenne de l'erreur absolu) est à 6.54 en estimation et 7.02 en prédiction. Ces valeurs sont corrects sans plus. Il en est de même pour le RMSE qui est à 9.22 en estimation et 9.72 en prédiction.

On peut classer la concentration en PM10 en trois intervalles :

- Niveau 0 : [0,50] $\mu g/m^3$
- Niveau 1 : [50,80] $\mu g/m^3$
- Niveau 2 : [80,max] $\mu g/m^3$

Construisons alors les tableaux des dépassements et les performances en prévisions qui s'en suivent pour l'estimation de notre échantillon d'apprentissage et la prévision de notre échantillon test.

2*Tableau des dépassements	Estimation			Prévision		
	Niveau 0	Niveau 1	Niveau 2	Niveau 0	Niveau 1	Niveau 2
Niveau 0	1655	8	0	712	5	0
Niveau 1	54	15	1	21	4	0
Niveau 2	3	7	2	5	0	1

Performances	Estimation	Prévision
POD	0.3	0.16
FAR	0.24	0.5
TS	0.28	0.14
SI	0.3	0.15

Nous remarquons grâce aux tableaux de dépassements que le modèle à tendance à sous estimer pour les moyennes et fortes concentration. Pour ce qui est des performances en prévisions :

- POD : représente le taux de bonne détection des concentrations supérieures à 50 $\mu g/m^3$. Pour ce modèle nous avons un taux de 30% pour l'estimation des données d'apprentissage et 16% pour la prévision des données test. Ces résultats ne sont pas satisfaisants et encore une fois la dégradation des données est mise en évidence lors du passage de l'échantillon d'apprentissage à l'échantillon test ;
- FAR : correspond au taux de fausses alarmes. Un taux de 50% en prédiction n'est pas une valeur raisonnable ;
- TS : correspond au Threat Score. Ce taux à 14% n'est pas bon.

Regardons maintenant les graphiques représentant la concentration de PM10 estimée en fonction de celle observée et la concentration de PM10 prédite en fonction de la concentration observée.

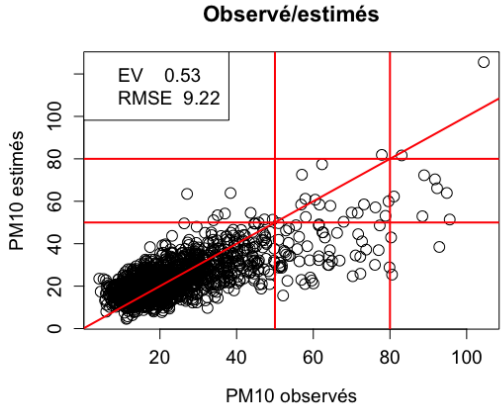

FIGURE 7 – Concentration PM10 estimée en fonction de observée

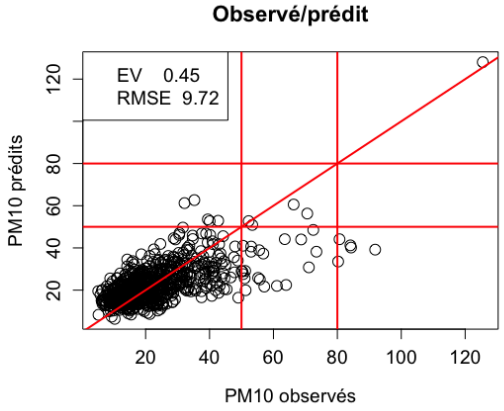

FIGURE 8 – Concentration PM10 prédite en fonction de observée

Nous voyons ainsi que le modèle a tendance à surestimer les faibles valeurs et à sous-estimer les moyennes et fortes concentrations. En conclusion, nous pouvons dire que ce modèle ne permet pas de prédire convenablement la concentration en PM10.

Nous gardons toujours ce modèle, et nous allons construire les intervalles de prévision à 95% pour les prévisions obtenues à partir de l'échantillon test. Pour évaluer la pertinence de ces intervalles de prévision à 95%, nous afficherons graphiquement ces intervalles en y ajoutant la vraie valeur de PM10.moy et sa prédiction.

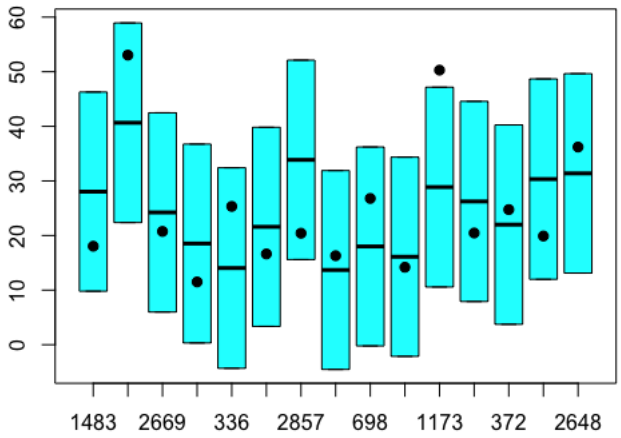

FIGURE 9 – Intervalles de prévision avec indication de la vraie valeur de PM10.moy et sa prédiction

Nous affichons aléatoirement 15 valeurs de l'échantillon test. En faisant plusieurs exécutions, on peut se rendre compte de la fiabilité des intervalles de prévision à 95% car peu de prévisions sont en dehors de l'intervalle. Cependant les intervalles ne sont pas précis. L'étendue des intervalles sont de l'ordre de 30 $\mu g/m^3$, ce qui est très grand, étant donné que dès 50 $\mu g/m^3$, la population doit être informée et des recommandations doivent être diffusées.

Modèles subsidiaires

Nous allons maintenant refaire la construction d'un modèle comme dans la partie précédente en apportant ces changements :

- On considère l'ensemble des variables explicatives disponibles dans le modèle (hormis la date)
- En utilisant des données centrées-réduites.

Nous expliquerons rapidement comment arriver au modèle final dans ces deux cas, puis nous afficherons les performances sous forme de tableau pour comparer avec le modèle obtenu dans la partie précédente.

On considère l'ensemble des variables explicatives (hormis la date)

Après avoir effectué successivement le VIF, nous excluons les variables suivantes : T.moy, DV.maxvv.fact. On applique ensuite la fonction lm de R sur nos données. En regardant le summary, on se rend compte que la fonction lm a converti automatiquement les variables qualitatives en n variables numériques avec n le nombre de modalités de la variables qualitative. Le modèle comporte alors 32 variables explicatives, pour une variance expliquée de 56.83%, ce qui est mieux que notre modèle précédent.

On applique alors la commande *step* au modèle, ce qui permet de supprimer du modèle les variables explicatives jugées non influentes. Les variables jour, HR.min, S02.max, HR.moy et VV.max sont supprimées du modèle successivement. La variable qualitative indiquant la direction du vent a donc une influence sur la concentration en PM10, et elle permet d'augmenter la variance expliquée du modèle par rapport au modèle de la partie précédente (56.91% contre 52.41%). Regroupons les indicateurs de performances du modèle et de celui de la partie précédente afin de les comparer (on appelle modèle 1 le modèle de la partie précédente et modèle 2 le modèle de cette partie) :

	Modèle 1		**Modèle 2**	
Performances	**Estimation**	**Prévision**	**Estimation**	**Prévision**
R	0.73	0.67	0.76	0.72
EV	0.53	0.45	0.57	0.52
MAE	6.54	7.02	6.14	6.38
RMSE	9.22	9.72	8.76	9.02
POD	0.3	0.16	0.32	0.16
FAR	0.24	0.50	0.24	0.44

On peut donc voir que les résultats sont meilleurs que pour le premier modèle. Le modèle reste en revanche moyen car le taux de fausse alarme est à 44% en prévision et le taux de bonne détection des concentrations de PM10 supérieures à 50 $\mu g/m^3$ est de seulement 16% en prévision. Le pourcentage de variance expliquée est quant à lui de 52%, ce qui est faible. Ce modèle est donc légèrement meilleur que le modèle sans prendre en compte les variables explicatives. De plus, on a pu voir que le logiciel R transforme automatiquement les variables qualitatives en quantitatives, comme expliqué précédemment.

Les données sont centrées-réduites

Regardons maintenant les résultats obtenus lorsque l'on centre et réduit les données. En reproduisant la même étude statistique que pour les autres modèles (VIF + réduction de variables), nous obtenons comme meilleur modèle le modèle avec les 17 variables explicatives suivantes : NO.max, SO2.max, VV.max, NO.moy, VV.moy, T.max , HR.max, PL.so, DV.maxvv , O3.moy, NO2.max , GT.max, SO2.moy , NO2.moy , O3.max, T.min , PA.moy. On appelle ce modèle le modèle 3. Comparons alors avec les 2 modèles précédents.

	Modèle 1		**Modèle 2**		**Modèle 3**	
Perf	**Estimation**	**Prévision**	**Estimation**	**Prévision**	**Estimation**	**Prévision**
R	0.73	0.67	0.76	0.72	0.73	0.67
EV	0.53	0.45	0.57	0.52	0.53	0.44
MAE	6.54	7.02	6.14	6.38	6.54	7.02
RMSE	9.22	9.72	8.76	9.02	9.21	9.75
POD	0.3	0.16	0.32	0.16	0.33	0.16
FAR	0.24	0.50	0.24	0.44	0.29	0.50

Nous obtenons alors des résultats très similaires au premier modèle, c'est à dire au modèle sans les variables qualitatives et sans avoir centré et réduit les données. Ce résultat

est surprenant car les variables ont des données d'ordres de grandeurs différents pouvant engendrer des instabilités dans la modélisation.

3 Conclusion

Pour conclure, on peut dire que les modèles que nous avons mis en place durant ce TP ne sont pas efficaces pour prédire correctement la concentration en PM10. Nous avons pu voir que le taux de fausse alerte est grand et le taux de bonne prédiction de concentrations supérieures à 50 $\mu g/m^3$ est faible. Nos modèles sous estiment fortement les hautes concentrations.

Il serait donc préférable d'utiliser un autre modèle comme les forêts aléatoires.