Modélisation de la qualité de l'air

Frédéric Schindler

Année 2012/2013

Rapport de stage

M2 Mathématiques appliquées mention statistique
Modélisation de la qualité de l'air

Schindler Frédéric

Remerciements

Mon aventure a commencé à Bordeaux où se trouve AIRAQ dirigée par Mr Bourquin. J'y ai rencontré une équipe très active et chaleureuse. Mon encadrement y a été assuré par Pierre-Yves Guernion et Benoit Duval que je remercie pour leur disponibilité et leur patiente à mon égard. Ces deux mois m'ont permis de découvrir la qualité de l'air et de poser les fondements du stage que j'allais effectuer.

Quatre heures de train plus tard (au lieu de trois, remerciements à la SNCF) me voilà arrivé à Verneuil-en-Halatte. L'INERIS a été le second théâtre où j'ai eu la chance de me produire. J'y ais été accueilli par Laure Malherbe et Frédérik Meleux. Anthony Ung s'est ensuite présenté à moi et m'a accompagné pendant les quatre mois dans l'accomplissement du stage. Ces trois rencontres ont été très intéressantes et riches en enseignements, je tiens à les remercier pour tout ce qu'ils m'ont apporté. Je prolonge ces remerciements à tout le reste de l'équipe qui m'a permis une intégration sympathique et plus que facile, ainsi qu'à mes camarades stagiaires dont j'ai particulièrement apprécié la présence. Un grand merci à Bertrand Bessagnet pour m'avoir intégré dans son équipe.

De retour dans mon Alsace natale, il me reste à remercier Mr Poulin pour sa disponibilité et ses conseils et Mme Guillou qui a donné au master sa belle réputation. Je termine en remerciant tous les professeurs qui ont participé à ma formation durant mes années universitaires.

Introduction

Le rapport de stage a été rédigé durant l'année universitaire 2012-2013 dans le cadre du master 2 de mathématiques appliquées mention statistique. L'objectif du stage est d'améliorer les prévisions de qualité de l'air en région Aquitaine.

Les deux systèmes de modélisation d'AIRAQ puis de l'INERIS sont exposés, analysés et commentés. S'en suit l'argumentation des pistes d'amélioration de ces systèmes qui ont été explorées et un exemple de la modélisation sur des données.

La première partie du rapport permet d'appréhender plus facilement le système de surveillance de la qualité de l'air qui est actuellement en place à l'échelle nationale. Elle permet de comprendre le contexte particulier du stage et d'introduire les problèmes qui ont menés au sujet du stage.

Le sujet est alors exposé en une seconde partie, l'objectif du stage et la démarche de ce qui a été fait y sont décris plus en détail.

Les méthodes de modélisations utilisées à AIRAQ et à l'INERIS et leurs performances sont évaluées dans la troisième partie. Cela nous permet d'avoir des résultats d'évaluation de référence et de tirer des premières conclusions utiles pour la suite du stage. Cette partie décrit aussi le protocole d'évaluation des modèles qui a été utilisé.

Enfin, une dernière partie expose les pistes qui ont été étudiées, quelques résultats et les conclusions qui en découlent.

Table des matières

I. La surveillance de la qualité de l'air

1. Le système de surveillance

L'Etat délègue sa mission de surveillance de la qualité de l'air en France à des associations agréées par le ministère. L'ensemble de ces associations forme la fédération Atmo. La surveillance se fait en référence à la loi sur l'Air et l'Utilisation Rationnelle de l'Energie (**LAURE**) du 30 décembre 1996.

LAURE :*"L'objectif est la mise en œuvre du **droit reconnu à chacun à respirer un air qui ne nuise pas à sa santé**. Cette action d'intérêt général consiste à prévenir, à surveiller, à réduire ou à supprimer les pollutions atmosphériques, à préserver la qualité de l'air et, à ces fins, à économiser et à utiliser rationnellement l'énergie."*

Loi n° 96-1236 du 30 décembre 1996 sur l'air et l'utilisation rationnelle de l'énergie

i. AIRAQ

Il existe **26** Associations Agréées de Surveillance de la Qualité de l´Air (**AASQA**) en France. **AIRAQ est l'Association de surveillance de la qualité de l'air en Aquitaine.**

Les missions d'AIRAQ peuvent se résumer en 3 points :

- **Mesurer** les concentrations en polluants dans l'air, surveiller en permanence la qualité de l'air conformément à la réglementation. Les polluants mesurés sont ceux pour lesquels des effets sur la santé ou sur l'environnement ont été établis ou sont pressentis.
- **Exploiter** les données :
  - Validation, études, archivage
  - **Modélisation de la qualité de l'air dans le temps et l'espace.**
- **Informer** en permanence le grand public, les médias, Institutionnels et industriels.

Sur les 14 sites sous surveillance continue (Bordeaux, Pau, Bayonne, Périgueux, Agen, Arcachon, Marmande, Mont de Marsan, Dax, Ambès, Lacq, Tartas, Iraty et Le Temple) sont réparties une ou plusieurs stations de mesures fixes dans lesquelles se trouvent un ou plusieurs analyseurs fonctionnant en automatique et mesurant des polluants spécifiques.

ii. INERIS

Le LCSQA(Laboratoire Central de Surveillance de la Qualité de l'Air) est un groupement d'intérêt scientifique composé de 3 organismes :

- EMD: Ecole des Mines de Douai
- LNE: Laboratoire National de métrologie et d'Essais
- **INERIS: Institut National de l'Environnement Industriel et des Risques**

Il assure une mission d'appui scientifique et technique auprès du ministère chargé de l'environnement et des AASQA. Depuis 2011, il est responsable de la coordination technique de la surveillance de la qualité de l'air en France.

iii. L'indice Atmo

L'indice Atmo caractérise la qualité de l'air quotidienne d'une agglomération de plus de 100.000 habitants sur une échelle qui va de 1 (indice très bon) à 10 (indice très mauvais). Cet indice sans unité sert à communiquer sur la qualité de l'air de manière plus intuitive qu'avec des données de concentration par polluant. Pour une zone de moins de 100.000 habitants on parlera d'indice de la qualité de l'air simplifié (IQA).

Cet indice ne permet pas de mettre en évidence des phénomènes localisés de pollution mais une pollution globale de fond. Il tient compte des niveaux de **dioxyde de soufre** (SO_2), de **dioxyde d'azote** (NO_2), d'**ozone** (O_3) et des **particules en suspension** (PM).

Des sous-indices sont construits à partir de ces 4 indicateurs de pollution (tableau 1).

Tableau 1 : Journal officiel de la république française (données en µg/m^3)

Indice	SO2	NO2	O3	PM10
1	0 à 39	0 à 29	0 à 29	0 à 6
2	40 - 79	30 - 54	30 - 54	07 - 13
3	80 - 119	55 - 84	55 - 79	14 - 20
4	120 - 159	85 - 109	80 - 104	21 - 27
5	160 - 199	110 - 134	105 - 129	28 - 34
6	200 - 249	135 - 164	130 - 149	35 - 41
7	250 - 299	165 - 199	150 - 179	42 - 49
8	300 - 399	200 - 274	180 - 209	50 - 64
9	400 - 499	275 - 399	210 - 239	65 - 79
10	>=500	>=400	>=240	>=80

L'indice Atmo communiqué est le plus élevé des 4 sous-indices.

Les données de base pour le calcul journalier de chaque sous-indice sont :

- la **moyenne des concentrations maximales horaires** observées pour le dioxyde de soufre (SO_2), le dioxyde d'azote (NO_2) et l'ozone (O_3)
- la **moyenne des concentrations journalières** observées pour les particules en suspensions de diamètre médian inférieur à 10µm (PM_{10})

iv. Contexte réglementaire

En matière de pollution atmosphérique, il n'existe pas de seuil en deçà duquel les polluants sont sans effet pour la santé. Certaines personnes sont affectées par des niveaux très bas.

Face à ces enjeux sanitaires, les pouvoirs publics définissent des niveaux de pollution au-delà desquels des actions temporaires ou permanentes de réduction des émissions sont mises en œuvre. Il s'agit des **seuils d'alerte**.

La définition française officielle du seuil d'alerte est : « *niveau de concentration de substances polluantes dans l'atmosphère, fixé sur la base de connaissances scientifiques, au delà duquel une exposition de courte durée présente un risque pour la santé humaine ou de dégradation de l'environnement et à partir duquel des mesures d'urgences doivent être prises* » (*Loi sur l'air* du 30 décembre 1996).

Dans un Arrêté préfectoral relatif à la « *procédure d'information, de recommandation et d'alerte* », en cas de pollution atmosphérique, le **seuil d'alerte** est qualifié de *deuxième niveau* de pollution (le premier étant le *niveau « d'information et de recommandations», au-delà duquel une information à destination des groupes les plus sensibles de la population doit être émise)*.

La valeur limite est définie, par polluant, comme le *« niveau maximal de concentration de ce polluant dans l'atmosphère, fixé sur la base des connaissances scientifiques, dans le but d'éviter, de prévenir ou de réduire les effets nocifs de ces substances pour la santé humaine ou pour l'environnement »*.

Les seuils (tableau 2) sont fixés par les textes réglementaires.

Tableau 2 : Seuils d'information et de recommandation et seuils d'alerte (Décret n° 2010-1250 du 21 octobre 2010 relatif à la qualité de l'air)

Polluant	Seuil d'information et de recommandation (µg/m^3)	Seuil d'alerte (µg/m^3)
Dioxyde d'azote (NO_2)	200 (moyenne horaire)	400 (moyenne horaire) pendant 3 heures consécutives
Dioxyde de soufre (SO_2)	300 (moyenne horaire)	500 (moyenne horaire) pendant 3 heures consécutives
Particules en suspension (PM_{10})	50 (moyenne journalière)	80 (moyenne journalière)
Ozone (O_3)	180 (moyenne horaire)	240 (moyenne horaire)

On dit qu'il y a un **pic de pollution** dès que l'un de ces seuils de court terme est dépassé.

Depuis 2007 les AASQA ont la possibilité de déclencher une alerte (lorsque la concentration d'un polluant dépasse un seuil règlementé) **sur prévision** lors d'une trop forte concentration de polluant prévue à J+1**.** A ce jour cependant, les déclenchements se font majoritairement sur constat, en se fondant sur l'expérience des salariés d'AIRAQ.

A partir du **31/10/13, conformément à un arrêté qui devrait entrer en vigueur à cette date,** le déclenchement d'une alerte **sur prévision** devra être privilégié, en fonction de critères liés à l'étendue du territoire et au nombre d'habitants potentiellement exposés à ce dépassement. Cette prévision nécessite l'utilisation de modèles numériques.

2. Le système de prévision

i. La prévision de la qualité de l'air à l'échelle nationale

Le système de prévision **PREV'AIR** génère et diffuse quotidiennement des **prévisions** (à J+0, J+1, et J+2) et des cartographies des concentrations de polluants (O_3 , NO_2 , PM_{10} et PM_{25}) dans l'air. L'ensemble est disponible sur le site internet www.prevair.org . L'INERIS assure la mise en œuvre et la maintenance du système PREV'AIR pour faire des prévisions de qualité de l'air à l'échelle nationale.

Le système repose notamment sur le modèle déterministe **CHIMERE,** développé par l'institut Pierre-Simon Laplace (CNRS) et par l'INERIS. A la différence des modèles statistiques qui s'appuient sur un historique de données pour effectuer des prévisions, les **modèles déterministes** prennent en compte les émissions de gaz et de particules (liées ou non à l'activité humaine), les conditions naturelles (météorologiques, relief, type de sol,…) et les mécanismes de génération et de destruction des polluants pour réaliser des cartes de prévision.

Les modèles utilisés dans le système PREV'AIR sont développés par les partenaires du projet:

- L'Institut Pierre-Simon Laplace (IPSL) et l'**INERIS** pour le **modèle CHIMERE** (et sous-modèles: **AFM**, ASAFM, CFM, AWM, CWM,…). **Le rapport traite le modèle AFM.**
- **Météo France** pour le modèle **MOCAGE**
-

ii. Limites des modèles de prévision

Les prévisions nationales de PREV'AIR **n'intègrent pas pleinement les caractéristiques régionales**. Il en résulte certaines difficultés dans la prévision des concentrations les plus élevées, comme c'est le cas pour les particules PM_{10} en Aquitaine.

Afin de pallier ces difficultés et affiner les prévisions sur la France, l'INERIS a mis en place une **procédure de post-traitement des résultats de modélisation déterministe appelée « adaptation statistique » (ASAFM : Adaptation Statistique du modèle AFM).** Cette procédure consiste à prévoir les concentions en chaque site de mesure par modélisation statistique et à utiliser ces prévisions locales pour corriger spatialement les cartes de prévision issues de CHIMERE.

Toutefois cette correction ne suffit pas à anticiper l'ensemble des pics de pollution et à évaluer précisément les dépassements de seuil réglementaire, comme AIRAQ a pu le constater pour les PM_{10} en Aquitaine.

Parallèlement aux travaux conduits par l'INERIS, AIRAQ a développé **ses propres modèles statistiques de prévision (nommés Previ_ajust)** sur huit agglomérations d'Aquitaine. La concentration horaire brute fournie par PREV'AIR (résultats non corrigés du modèle déterministe CHIMERE France) est utilisée comme variable explicative. L'usage complémentaire de variables explicatives locales permet d'améliorer les prévisions en accentuant les effets météorologiques spécifiques à la région. Malgré cela, ces modèles **ne sont pas encore suffisamment performants** lors de la **prévision de pic de pollution** notamment pour les PM_{10}.

II. Présentation du sujet

1. Objectifs du stage

Le stage s'est divisé en deux parties, la première s'est déroulée à AIRAQ du 20 mars au 17 mai 2013 (2 mois) ; la seconde s'est déroulée à l'INERIS, du 21 mai au 17 septembre 2013 (4 mois). Ces deux parties servent un objectif commun: **définir une méthode plus performante de prévision de la qualité de l'air**. Plus précisément, à partir des travaux réalisés par AIRAQ et l'INERIS, l'objectif a été de définir les améliorations à apporter dans PREV'AIR pour une prévision plus efficace des concentrations de PM_{10} sur l'Aquitaine et, à terme, sur d'autres régions françaises.

Le travail a été orienté sur la **prévision des concentrations de PM_{10}.** A ce jour en effet, du fait des incertitudes sur leurs données d'entrée (les données d'émissions notamment) et de la multiplicité des processus impliqués dans la formation et le transport des particules, les modèles déterministes ne parviennent pas à reproduire précisément les niveaux et la variabilité des concentrations de PM_{10}. En période de fortes concentrations, une sous-estimation par les modèles est généralement constatée. Or compte tenu des enjeux sanitaires et des contraintes réglementaires (déclenchement de procédures d'information et d'alerte), une prévision fiable des PM_{10} est indispensable.

2. Démarche

La première partie du stage a eu pour objet de réaliser une **synthèse des études menées par AIRAQ** sur les modèles de prévision en région Aquitaine. Il s'agissait d'**identifier les variables permettant** d'expliquer les écarts entre les prévisions de concentration de polluant faites à l'échelle nationale et les mesures relevées sur les stations d'Aquitaine.

Cette analyse a été valorisée dans la seconde partie du stage, dans le cadre des travaux conduits par l'INERIS sur la prévision de la qualité de l'air. Une première étape a consisté à **comparer les méthodes de prévision utilisées respectivement par AIRAQ et l'INERIS et à déterminer les forces et faiblesses de chacune**. Le but était de tirer le meilleur de chaque méthode et de s'en inspirer pour la suite du stage.

L'étape suivante a été de déterminer les améliorations à apporter au modèle de prévision utilisé au niveau national en incorporant les informations apportées par le stage à AIRAQ. D'autres méthodes de modélisation peuvent être développées et évaluées si leurs apports semblent judicieux.

Tout au long du stage, des réunions de travail avec AIRAQ et l'INERIS ont été organisées permettant de poser les résultats et de définir un plan façonné en fonction des résultats obtenus.

3. Outils et sources d'information

La plupart des calculs ont été conduits avec le logiciel R (http://cran.r-project.org/). Une initiation à Linux (script Shell) m'a permis de manipuler avec plus de liberté les données mises à disposition par l'INERIS.

D'autre part, mon travail s'est appuyé sur une abondante documentation disponible sur les sites des AASQA, en particulier ceux d'AIRAQ (http://www.airaq.asso.fr/), du LCSQA (http://www.lcsqa.org/rapports) et de PREV'AIR (http://www.prevair.org).

J'ai enfin participé à une formation organisée par le LCSQA, en collaboration avec Frédéric Lavancier, maitre de conférences à l'université de Nantes. Cette formation était ouverte aux AASQA qui désiraient utiliser des méthodes statistiques un peu plus poussées dans leurs études. Les conseils de Frédéric Lavancier sur l'avancement de mon stage m'ont permis de diriger la suite de mon travail portant sur les outils de modélisation.

III. La Modélisation à AIRAQ et INERIS

Comme il a été expliqué en II.2, AIRAQ et l'INERIS ont développé des procédures d'adaptation statistique destinées à améliorer localement les prévisions fournies par le modèle déterministe CHIMERE.

En Aquitaine, AIRAQ a mis en place des modèles statistiques qui, à partir des données de CHIMERE récupérées chaque matin sur le site de PREV'AIR et de variables locales (mesures de la veille, prévisions météorologiques), fournissent des prévisions moyennes par agglomération.

A l'échelle nationale, l'INERIS a construit des modèles statistiques qui, à partir des données de CHIMERE, des mesures de la veille et de prévisions météorologiques de grande échelle (données de modèles météorologiques), fournissent des prévisions par station. Ces dernières sont ensuite exploitées selon une approche géostatistique pour corriger la carte de prévision issue de CHIMERE.

1. L'adaptation statistique chez AIRAQ

AIRAQ a développé un modèle d'adaptation statistique sur chacune des huit agglomérations de la région Aquitaine (Agen, Arcachon, BAB, Bordeaux, Dax, Lacq, Pau et Périgueux). Ce modèle est nommé Previ_ajust.

Chaque agglomération dispose d'une ou plusieurs stations de mesures réparties à des endroits stratégiques afin évaluer l'exposition de la population à la qualité de l'air selon l'origine de la pollution (station de fond urbain, station rurale, station de proximité automobile, station industrielle). Seules les stations de fond urbain sont utilisées pour la création des modèles. Situées dans des zones densément peuplées à proximité directes des sources d'émissions des polluants, ces stations permettent d'évaluer le niveau d'exposition moyen de la population aux phénomènes de pollution atmosphérique dits de « fond » dans les centres urbains.

Pour mettre en place **un modèle par agglomération** AIRAQ a fait la moyenne spatiale des mesures relevées sur l'ensemble des stations urbaines de fond que contient l'agglomération. Cette modification permet d'obtenir une prévision à l'échelle de l'agglomération plutôt qu'à l'échelle de la station.

Chaque agglomération dispose de deux modèles, un pour la période estivale et un second pour la période hivernale. La concentration en PM_{10} est plus élevée en hiver où les polluants sont émis en plus grande quantité (chauffage au bois,…) et les conditions météorologiques davantage propices à leur accumulation (absence de vent, création d'une couche d'inversion, …). Un modèle basé uniquement sur une année complète sans distinction de période donnera moins de poids aux valeurs élevées. La prévision sera tirée vers des valeurs plus faibles du fait que la période estivale est en général plus "plate".

Les modèles actuellement en opérationnel ont été construits sur un historique de données qui s'étend du 01/06/2007 au 31/12/2010. Les 16 modèles (8 agglomérations, été/hiver) ont été créés indépendamment mais résultent tous d'une même procédure de modélisation.

AIRAQ a développé un modèle permettant d'expliquer non pas directement la concentration en PM_{10} mais l'**erreur de prévision du modèle CHIMERE pour le jour considéré (J+0)** (Concentration mesurée – Prévision PREV'AIR).

Les variables explicatives de cette erreur sont constituées de données météorologiques et de la concentration en PM_{10} mesurée la veille (J-1).

La **régression linéaire multiple** sert à la modélisation, ce qui suppose une relation linéaire entre les variables explicatives et la variable réponse. Il convient de s'assurer que cette hypothèse est légitime. La sélection de variable à partir des données météo a été faite par la méthode stepwise sous R. De plus, de nouvelles variables ont été créées en en combinant deux variables météo. Elles ont une cohérence et sont intuitivement compréhensibles. Il n'y pas eu de transformation de variables pour linéariser certaines relations. Aucun traitement particulier n'a été mis en place pour les événements rares (une dizaine de pics de pollution par année). Ce type de modélisation a l'avantage d'être aisément interprétable.

2. L'adaptation statistique chez INERIS

L'INERIS travaille au niveau national et a développé **un modèle pour chacune des stations françaises qui mesurent la pollution de fond.**

La variable expliquée est ici **la concentration de PM_{10}.** Les variables explicatives de cette concentration comprennent la prévision de CHIMERE, des variables météorologiques et la concentration en PM_{10} mesurée la veille.

Le nombre de stations étant très important (à peu près 400 sur l'ensemble du territoire), la procédure de modélisation a été automatisée par un script Bash. La méthode doit être la plus souple et généralisable possible. Comme à AIRAQ, la **régression linéaire multiple** a été choisie. Le lien linéaire entre les variables explicatives et la concentration en PM_{10} est supposé. Pour chaque station de mesure, une présélection de variables est réalisée en fonction du coefficient de corrélation ; la sélection finale des prédicteurs résulte d'une procédure de type stepwise. Ces développements reposent sur les travaux méthodologiques et les évaluations réalisés lors du projet CITEAIR2 (Honoré et al., 2012[1]).

[1] Honoré C., Ung A., Corbet L., Malherbe L., 2012. Good Practice Guide on Urban Air Quality Forecast. CITEAIRII project, http://www.citeair.eu/.

3. Les données

AIRAQ et l'INERIS disposent d'un historique de données très fourni. Ils sont constitués de mesures faites par les AASQA sur le terrain et de données météorologiques fournies principalement par Météo France.

Les **PM_{10}** sont des particules en suspension dans l'air dont le diamètre est inférieur à 10 micromètres (1 µm = 1×10^{-6} m), d'où leur nom anglais de *particulate matter* 10, ou PM 10 en abrégé.

Leur origine peut être naturelle (érosion, volcanisme…) ou anthropique (fumée, usure, etc.). Les PM_{10} proviennent en grande partie du trafic automobile (diesel) ou des chauffages au fioul, au bois, des industries et de l'agriculture. Dans toutes les modélisations réalisées, les données **de PM_{10}** considérées sont les **concentrations moyennes journalières de PM_{10}** (i.e. les moyennes journalières des 24 concentrations horaires) mesurées aux stations et modélisées par CHIMERE.

En hiver, une diminution de **température** (2 à 4°C) peut provoquer la formation d'une **couche d'inversion**. En situation normale, la température de l'air diminue avec l'altitude (environ 1°C tous les 100 mètres). Quand l'air chaud s'élève dans les couches supérieures plus froides, il entraîne avec lui les polluants qui sont ainsi dispersés verticalement (principe de la montgolfière). Les inversions de température sont des cas particuliers ; l'atmosphère, au lieu de se refroidir avec l'altitude, réchauffe jusqu'à un certain niveau appelé niveau d'inversion. A ce niveau se forme une couche d'air plus chaude qu'on appelle couche d'inversion. Les substances provenant des chauffages, des industries et du trafic automobile, s'accumulent sous la couche d'inversion qui forme un « couvercle » empêchant les polluants de se disperser, il n'y a plus de brassage vertical. Si le vent est faible, la concentration des polluants peut alors augmenter très rapidement.

La h**auteur de la couche limite** est la partie de l'atmosphère influencée par la surface terrestre. Si la couche limite se situe à une basse altitude alors les polluants seront plus concentrés dans l'air. La mesure de cette hauteur nécessite des outillages de mesure coûteux, les erreurs d'estimation de la hauteur de la couche limite ne sont pas évaluables.

Le **vent** intervient tant par sa direction pour orienter les panaches de fumées que par sa vitesse pour diluer et entraîner les émissions. Le vent est modélisé par 2 paramètres :

- Vitesse en m/s
- Direction modélisée sur axe x (Est / Ouest) et y (Nord / Sud)

L'**humidité** (ou humidité relative) influence la transformation des polluants primaires émis. Les précipitations entraînent au sol les polluants les plus lourds. Elles peuvent parfois accélérer la dissolution de certains polluants. Mais, globalement, les concentrations en polluants dans l'atmosphère diminuent nettement par temps de pluie notamment pour les poussières

Les deux variables (insolation et rayonnement global) décrite ci-dessous sont liées, il est possible d'exprimer l'insolation en fonction du rayonnement global et inversement. Pour cela il faut introduire les variables *irradiation extraterrestre* et durée du jour mais elles ne seront pas traitées dans ce rapport.

L'Insolation désigne le temps d'ensoleillement. Un faible ensoleillement engendre généralement amélioration de l'indice de qualité de l'air sauf si la température est constante et élevée. Un fort ensoleillement donne une dégradation dont l'importance est à corréler avec la hausse de la température. Le **Rayonnement global** est l'énergie rayonnante totale du soleil, qui atteint une surface horizontale à la surface de la Terre au cours d'une unité de temps précise.

AIRAQ se sert de ces variables pour en créer de nouvelles qui semblent avoir un effet sur la concentration du polluant concerné. Elles résultent de combinaison entre deux variables météo. Leurs présence est intuitivement compréhensible mais aucune information sur la découverte de ces variables n'a été renseignée.

Les variables météorologiques sont récupérées auprès de Météo France. Il existe deux type de données : les données des stations de mesure et les données de prévision correspondant aux sorties des modèles. AIRAQ et INERIS ne travaillent pas avec les mêmes modèles météorologiques.

- AIRAQ utilise prévisions du modèle ARPEGE avec résolution 0.1° (à échéance 48h).
- Le modèle météorologique utilisé diffère selon la version de CHIMERE. Officiellement PREV'AIR fonctionne avec le modèle météo GF5 - code MM5 (résolution 0.15°).

Le but ici n'est pas d'entrer dans les détails de ces modèles météo mais il est bon de mettre un nom sur les données utilisées.

Tableau 3 : Variables météorologiques quantitatives.

Variable explicative	Notation	Unitée
Concentration en PM10		µg/m^3
Concentration en PM10 à J-1		µg/m^3
Température	T	°C
Hauteur de la couche limite	hght	mètre
Vent - Vitesse	FF	m/s
- **Direction**	DD	
Humidité	U	%
Insolation	INS	Minute d'ensoleillement
Rayonnement global	GLO	Joule/cm^3

Tableau 4 : Variables météorologiques qualitatives.

Variables explicatives	Notation	Type
Jour de la semaine	Day…	Qualitatif à 7 modalités
Week-end	WE	Qualitatif à 2 modalités

4. Procédure de comparaison des modèles

On nomme **période d'apprentissage** l'intervalle de temps sur lequel le modèle est déterminé, il est choisi dans l'historique de données dont nous disposons. Dans l'objectif de prévision ce choix est très important car le modèle doit être préparé à toutes les éventualités. La période doit couvrir de préférence une année entière.

Pour comparer les qualités des modèles d'AIRAQ et de l'INERIS, il faut confronter leurs performances de prévision sur une même plage de données (**période de validation**). Les statistiques de score (voir annexe 1) sont utilisées pour évaluer ces performances.

Dans l'analyse qui suit les modèles de l'INERIS et d'AIRAQ sont comparés sur la ville de Bordeaux. Cette agglomération a été choisie du fait des pics de concentration survenus. L'année 2012 est particulièrement intéressante pour évaluer un modèle car les pics enregistrés ont des causes différentes :

- Les **polluants primaires** sont émis dans l'air directement par une source donnée (trafic routier, chauffage au bois, industries,…).
- Les **polluants secondaires** se forment lorsque d'autres polluants (primaires) réagissent dans l'atmosphère (pas émis directement).

Sur l'année 2012 certains pics de PM10 ont été enregistrés à cause de polluant primaires et d'autres à cause des polluants secondaires. Le modèle doit être capable de prévoir ces deux familles de pics.

i. Graphique temporel :

Une première approche simple permet de se faire une idée sur les qualités d'un modèle, il suffit de tracer les séries temporelles.

Le graphique temporel (figure 1) regroupe pour l'année 2012 les mesures (en noir), le modèle d'adaptation statistique développé par l'INERIS (nommé ASAFM, en rouge) le modèle utilisé à AIRAQ (nommé Previ_ajust, en bleu). Les dates où des pics sont enregistrés sont indiquées sur le graphique. La ligne horizontale rouge marque le seuil de 50 µg/m^3 au-delà duquel la journée est classée en pic de pollution.

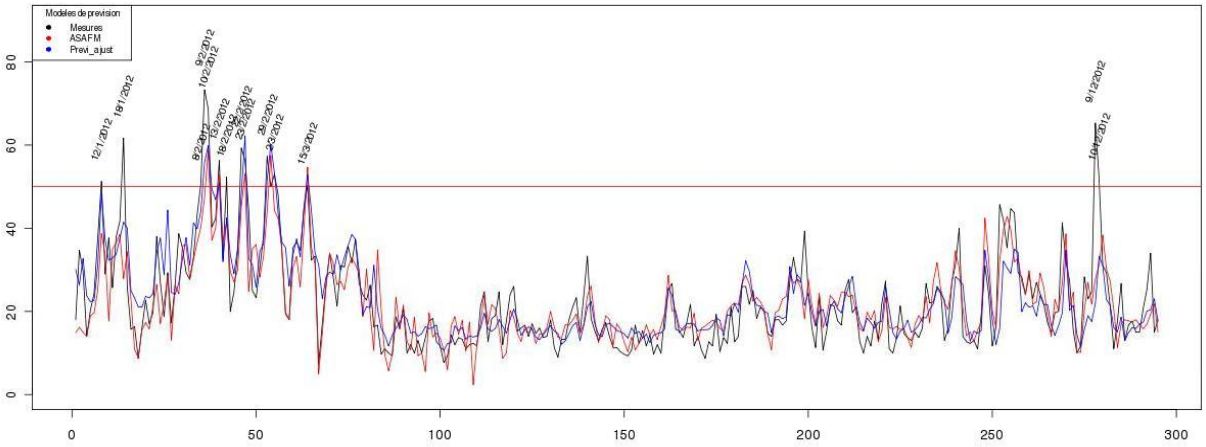


Figure 1 : Séries temporelle de la concentration en PM10 et des modèles de prévisions (ASAFM pour l'INERIS et Previ_ajust pour AIRAQ) - année 2012

Il est difficile de dégager un modèle plus performant que l'autre. Le pic du 18/01/2012 a été totalement ignoré par les modèles. Les deux pics consécutifs du 9 et 10 décembre 2012 ont été totalement sous-estimés. Cependant, aucun pic n'est prévu à tort.

Le modèle développé à l'INERIS travaille beaucoup avec les données mesurées à J-1, son attitude après la prévision d'un pic est importante à regarder. Les pics du 02/03/2012 et du 15/03/2012 ont été prévu mais s'en suit une forte sous-estimation alors que la mesure se situe dans la moyenne de l'année.

Les performances des deux modèles sont semblables sur la faible période (estivale).

ii. Graphique croisé :

Il s'agit de croiser les mesures relevées (ordonnée) avec les estimations du modèle qui ont été simulées (abscisse) sur la même période. Ce graphique permet de visualiser rapidement :

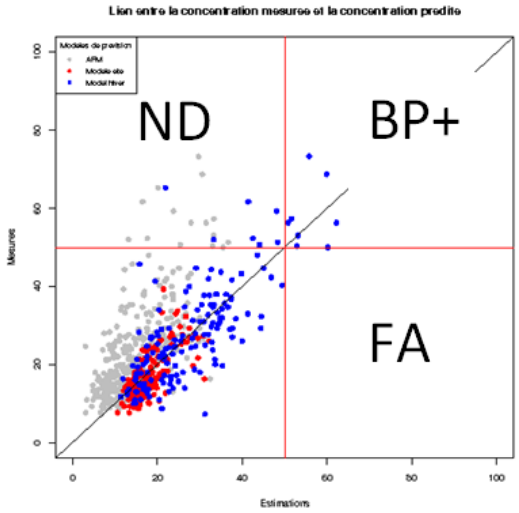
- La performance globale du modèle.
- Une tendance à la surestimation (droite de la diagonale) ou à la sous-estimation (gauche de la diagonale).
- La qualité de prévision du modèle sur les jours de pic.

La répartition des points donne l'intuition de :

- **BP+** (Bonne Prévision) : part de bonne prévision au-delà d'un seuil
- **BP-** : part de bonne prévision en-dessous d'un seuil
- **ND** (Non-Dépassement) : part de mauvaise prévision en-dessous d'un seuil
- **FA** (Fausse alerte) : part de mauvaise prévision au-delà d'un seuil

Une **bonne prévision** signifie que le modèle a prévu un dépassement ou un non-dépassement de seuil qui a été validé avec les mesures. Une **mauvaise prévision** indique que le modèle a prévu un dépassement ou un non-dépassement du seuil à tort.

Sur le graphique en exemple le seuil a été placé à 50µg/m^3 (droites verticale et horizontale en rouge) correspondant au seuil d'information et de recommandation pour les PM10.



Une mesure est d'autant mieux prévue que le point correspondant est proche de la diagonale.

Le graphique utilisé dans l'exemple représente en gris claire les prévisions du modèle CHIMERE (AFM), sur l'année 2012, qu'AIRAQ et INERIS veulent améliorer. Ensuite nous avons représenté les modèles été et le modèle "hiver" développés à AIRAQ, comparé aux mesures (respectivement été et hiver). On compare les modèles en représentant les nuages sur le même graphique croisé (figure 2).

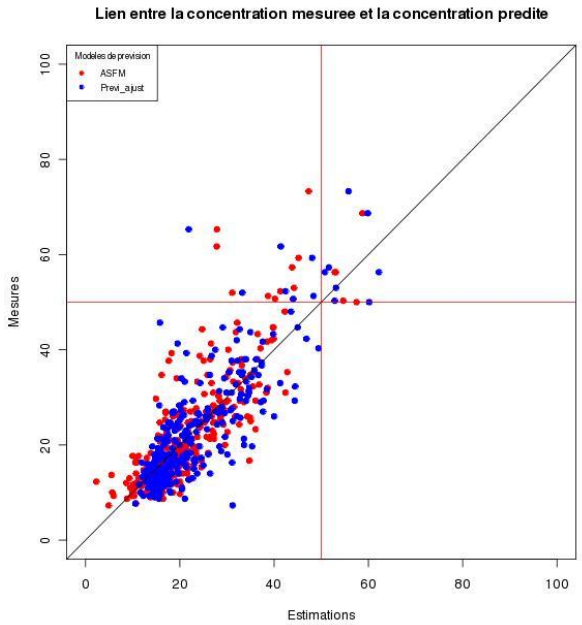


Figure 2 : Graphique croisé du modèle ASAFM (INERIS) et Prévi_ajust (AIRAQ) – année 2012

La plupart des mesures se situe entre 10 et 30 µg/m^3, l'adéquation globale du modèle est évaluée à partir de cet ensemble de données. La différence entre les modèles ASAFM et Previ_ajust n'est pas flagrante. Le modèle Previ_ajust a tendance à surestimer les faibles valeurs.

Previ_ajust est plus performant pour ce qu'il est du nombre de bonnes prévisions, on en dénombre 8 contre 5 pour le modèle ASAFM. C'est-à-dire que le modèle Previ_ajust a prévu 7 dépassements de seuil en concordance avec les mesures, mais 7 dépassements n'ont pas été détectés. L'étude des pics se fait sans prendre compte de l'écart entre les mesures et les prévisions.

Pour seulement 14 valeurs il est plus simple de regarder directement les données chiffrées enregistrées les jours de pic. On regarde en parallèle les mesures et les prévisions qui ont été faites.

Tableau 5 : Mesures et prévisions du modèle Prévi_ajust les jours de pic

Jour	mois	année	mesure	estimation
12	**1**	**2012**	51.3	48.4
18	**1**	**2012**	61.7	41.4
8	**2**	**2012**	50.7	44.1
9	**2**	**2012**	73.3	55.8
10	**2**	**2012**	68.7	59.9
13	**2**	**2012**	56.3	50.8
18	**2**	**2012**	52.3	42.5
22	**2**	**2012**	59.3	48.1
23	**2**	**2012**	56.3	62.2
29	**2**	**2012**	57.3	51.6
2	**3**	**2012**	53	53.1
15	**3**	**2012**	50.3	52.9
9	**12**	**2012**	65.3	21.9
10	**12**	**2012**	52	33.3

On remarque que certains dépassements non prévu sont comptabilisé alors que les prévisions n'étaient que de 3 µg/m^3 inférieures aux mesures. Les dépassements de décembre ont été totalement sous-estimés alors que la mesure est 15µg/m^3 au-dessus du seuil. Lors de la période du 8, 9 et 10 février le modèle n'a pas détecté le premier pic mais s'ajuste et est plus performant les deux jours suivants.

iii. Etude quantitative :

Les indices de score permettent de quantifier la distance globale entre le modèle et les mesures (voir Annexe 2).

Tableau 6 : Statistique de score pour les modèles ASAFM et Prévi_ajust évalués sur 2012.

Indice	ASAFM	Previ_ajust
Biais	0.54	-0.37
Ecart absolu moyen	4.81	4.83
Ecart quadratique moyen (RMSE)	6.85	6.86
Ecart quadratique centré moyen (RcMSE)	6.83	6.85
Correlation Pearson	0.82	0.82
Variabilite	1.26	1.2

Ces données ne renseignent pas sur la qualité de prévision des pics.

Le biais est négligeable pour les deux modèles, on retrouve la légère tendance à la surestimation (biais positif) du modèle Previ_ajust qu'on avait identifié. La surestimation se situe surtout sur les faibles valeurs, une sous-estimation des valeurs plus élevée peut faire balancer le biais vers 0.

L'écart absolu moyen (EAM) indique l'écart moyen entre les mesures et les prévisions. Les modèles ont un EAM semblable, la qualité est moyenne (voir annexe 1). Le RMSE montre que les modèles sont relativement dispersés autour des mesures, de l'ordre de 6.85µg/m^3 en moyenne, le RcMSE reste en accord avec ce constat. La corrélation est relativement élevée, les modèles ont globalement la même attitude que les mesures sur l'année 2012.

5. Etude des variables constitutives des modèles

En première partie nous exposons les modèles dans un tableau, en énumérant les variables explicatives présentes, les coefficients associés, et une « échelle approximative » qui permet de se faire une idée sur la plage de valeurs couverte par la variable. On se sert de l'historique de données sur la station 31002 pour se faire une idée approximative des valeurs prises par la variable. Ces bornes associées aux coefficients déterminés par la régression linéaire permettent d'évaluer la participation de la variable sur la concentration en PM_{10}. Pour chaque variable on note donc le min, le max, et le $3^{ème}$ quartile. Lorsque la variable est un polluant nous nous référons au tableau 1 (page 7), la valeur associée à l'indice 7 pour le polluant considéré remplacera le $3^{ème}$ quartile utilisé pour les variables météorologiques.

Lors du stage la modélisation mise en place a été testée sur les trois stations de Bordeaux, nous allons donc nous concentrer sur celles-ci. L'INERIS a développé un modèle par station tandis qu'AIRAQ a travaillé sur un modèle pour l'ensemble de l'agglomération bordelaise. Les trois premiers modèles exposés sont ceux utilisés par INERIS, le modèle d'AIRAQ sera ensuite examiné en détail.

La nomination des variables explicatives est décrite dans le tableau. L'explication des indices associés aux facteurs explicatifs est décrite dans le tableau 7.

Tableau 7 : Notation utilisée à l'INERIS

Notation	B		A
Indice	J-1	J+0	J+1
Définition	Hier	Aujourd'hui	Demain

Les prévisions à J+0 sont faites tôt le matin et sont logiquement plus fiables que les prévisions à J+1.

Il faut bien distinguer dans les modèles les données à J-1 qui sont des mesures (notées **mes**) et les prévisions (notées **mod**) à J+0 qui sortent de modèles qui ont leurs propres marges d'erreur.

i. Station 31001 :

Tableau 8 : Modèle de l'INERIS sur la station 31001

Facteur	Coefficient		Echelle approximative
(Intercept)	8.132	Constante	
min_PM10_mesB	0.455	Minimum du PM_{10} mesuré à J-1	0 à 80, indice 7 = 40
min_NO2_mod	1.037	Minimum des prévisions de NO_2 à J+0	0 à 400, indice 7 = 160
moy_PM10_mod	0.734	Moyenne des prévisions de PM_{10} à J+0	
moy_PM25_mod	-0.609	Moyenne des prévisions de PM_{25} à J+0	0 à 80, indice 7 = 40
moy_PM10_modB	-0.645	Moyenne des simulations AFM de PM_{10} pour J-1	
moy_PM25_modB	1.085	Moyenne des simulations de PM_{25} pour J-1	
moy_hght_mod	-0.003	Hauteur couche limite estimée à J+0	20 à 3500, Q3=700
min_tem2_mod	-0.152	Température min estimée à J+0	-5 à 15
dayMonday	-0.55	Jour de la semaine	
daySaturday	-1.859		
daySunday	-1.417		
dayThursday	0.194		
dayTuesday	1.184		
dayWednesday	1.537		

Dans la constitution des modèles l'INERIS (Tableau) garde les facteurs qui ont une corrélation d'au moins 0.3 avec la concentration en PM_{10}. Les facteurs météorologiques ont aussi un effet sur les PM_{10}, les NO_2 et l'O_3 qui peuvent expliquer la présence des autres polluants dans le modèle. Les PM_{25} sont les particules d'au plus 2.5µg de diamètre. On comprend que PM_{25} et PM_{10} soient corrélés, d'où la présence de ce facteur dans le modèle. Ce point nous laisse inquiet quand à un problème colinéarité entre les variables explicatives du modèle.

Dans le cadre de la prévision, la variable min_NO2_mod sera estimée à partir d'un autre modèle de prévision. Il y a alors un terme d'erreur associé qui peut avoir une grande importance si l'erreur est élevée. Le coefficient associé à cette variable est de 1.037 ce qui signifie que si le modèle se trompe de 10µg/m^3 pour sa prévision de NO_2, alors la concentration en PM10 sera immédiatement déviée de 10.37 µg/m^3.

L'interprétation de la différence entre les **jours de la semaine** est assez délicate. La modalité prise comme référence est ici le vendredi (Friday). Les coefficients associés aux autres jours de semaine s'interprètent par rapport à cette modalité : par exemple le jeudi, un surplus de 0,194 µg/m^3 est ajouté au modèle alors que le samedi une concentration de 1,859 est retirée.

Le coefficient associé à la **température** est négatif, ce qui est accord avec les observations. Les pics de PM10 sont observés lors de faibles températures. La **hauteur de la couche limite** a une place minime dans la prévision.

ii. Station 31007 :

Tableau 9 : Modèle de l'INERIS sur la station 31007

Facteur	Coefficient		Echelle approximative
(Intercept)	4.681	Constante	
min_PM10_mesB	0.487	Minimum des PM_{10} mesurés à J-1	0 à 80, indice 7 = 40
min_NO2_mod	1.44	Minimum des prévisions de NO_2 à J+0	0 à 400, indice 7 = 160
moy_PM10_mod	0.487	Moyenne des prévisions de PM_{10} à J+0	
moy_PM10_modB	-0.489	Moyenne des simulations de PM_{10} à J-1	
moy_PM25_modB	0.771	Moyenne des simulations de PM_{25} à J-1	0 à 80, indice 7 = 40
typedayWE	-1.919	Week-end	

On remarque dans le tableau que la variable **concentration en NO_2 prévue par le modèle** est toujours présente. Il y a au moins 0.3 de corrélation entre la concentration en PM10 et la concentration en NO_2 (voir méthode de construction du modèle). Cette valeur de corrélation est peut-être due à des facteurs qui influent sur la concentration en PM10 et conjointement sur la concentration en NO_2.

En week-end la prévision des concentrations en PM_{10} est amputée de 2µg/m^3.

Aucun facteur météorologique n'est présent dans ce modèle.

iii. Station 31002 :

Tableau 10 : Modèle de l'INERIS sur la station 31002

Facteur	Coefficient		Echelle approximative
(Intercept)	6.423	Constante	
moy_PM10_mesB	-6.966	Moyenne des mesures en PM_{10} à J-1	0 à 80, indice 7 = 40
min_NO2_mod	0.568	Minimum des prévisions de NO_2 à J+0	0 à 400, indice 7 = 160
min_O3_mod	0.045	Minimum des prévisions de O_3 à J+0	0 à 240, indice 7 = 150
moy_PM10_mod	0.305	Moyenne des prévisions de PM_{10} (AFM) à J+0	
moy_PM25_mod	0.421	Moyenne des prévisions de PM_{25} à J+0	0 à 80, indice 7 = 40
moy_PM10_modB	6.974	PM_{10} moyen simulé à J-1	
moy_PM25_modB	0.527	PM_{25} moyen simulé à J-1	
moy_hght_mod	-0.005	Hauteur couche limite simulée à J-1	20 à 3500, Q3=700
moy_PM10_erreur_modB	-7.521	Erreur de la veille de prévision du PM_{10} moyen	
typedayWE	-1.636	Week-end	

Les variables explicatives (Tableau 10) sont pour beaucoup des variables estimées par des modèles. Ces modèles ont leur propre erreur de prévision qui est alors intégré dans le modèle qui a été créé.

moy_PM10_mesB et **moy_PM10_modB** on quasiment le même facteur en moyenne absolue (|-6.966| et 6.974 que nous pouvons arrondir à 6.97). Dans la formulation du modèle on obtient :

6.97*(moy_PM10_modB – moy_PM10_mesB) = 6.97*(erreur de prévision à J-1). On retrouve le facteur **erreur de prévision du modèle à J-1** avec un coefficient de -7.521.

La **hauteur de la couche limite** a un rôle important, une hauteur élevée implique une concentration en PM10 plus faible. Le signe du coefficient est en accord avec les observations.

Il y a beaucoup de facteurs (4 sur 10) qui font référence à la **situation à J-1**. En regardant les graphiques temporels des variables durant les J-1 de chaque pic (sur les 24h précédant le pic), il n'est pas possible de dégager une journée type « pré-pic ». Le fait de garder les variables basées sur J-1 peut donc altérer la prévision des pics.

Dans les variables explicatives on trouve le NO_2 dont le lien avec les PM_{10} est difficile à interpréter. L'évolution conjointe des concentrations peut-être due à des variables explicatives communes pour les PM_{10}, NO_2 et O_3.

iv. Modèle AIRAQ :

Tableau 11 : Modèle Ete de AIRAQ pour la ville de Bordeaux

Facteur	Coefficient		Echelle approximative
(Intercept)	4.86	Constante	
min_T_mod	0.21	Minimum des prévi de température à J+0	-5 à 13, Q3=7
moy_PM10_mesB (jusqu'à 16h)	2.49	Moyenne des mesures de PM10 à J-1	
moy_PM10_mod	-0.65	Moyenne des prévi de PM10 (AFM) à J+0	0 à 80, indice 7 = 40
moy_Vy_mod	-0.313	Direction moyenne de l'origine du vent prédit à J+0	-6 à 10, Q3=1.5
PM10_mesB * PM10_mod	0.00797		80 à 1900, Q3=300

Tableau 12 : Modèle Hiver de AIRAQ pour la ville de Bordeaux

Facteur	Coefficient		Echelle approximative
(Intercept)	13.25	Constante	
moy_VV_mod	-1.37	Moyenne des prévi de force de vent à J+0	0 à 8, Q3=4.5
moy_PM10_mesB	0.25	Moyenne des mesures de PM10 à J-1	0 à 40, indice 7 = 40
Moy_PM10_mod	-0.398	Moyenne des prévi de PM10 (AFM) à J+0	
PM10_mesB * ampT_mod	0.016		0 à 1300, Q3=250
PM10_mod * min_T_mod	0.0278		10 à 230, Q3=80

Commentaires :

Ces modèles ont été créés pour modéliser l'erreur de prévision du modèle AFM. Il s'agit d'une régression linéaire sur Y=concentration mesurée – prévision de CHIMERE.

- **Le modèle CHIMERE sous-estime nettement les concentrations en PM10, donc Y est positif.**
- **On cherche à avoir un Y le plus petit possible**

Ces modèles sont très intéressants pour avoir une idée des facteurs météorologiques spécifiques à la région Aquitaine. Le modèle hiver tourne durant la période où les pics de PM10 sont enregistrés, et donc où l'erreur de prévision est plus importante.

Il y a un modèle par saison, **l'effet temporel** est donc présent.

Les variables composés (interactions d'ordre 2) qu'on retrouve dans les modèles ont été considérés comme des variables à part entière lors de la création des modèles. Les facteurs simples composants ces éléments ne sont pas inclus dans les modèles.

Pour faire des prévisions à J+0, on se sert des mesures à J-1 jusqu'à 16h. Le fait de prendre en compte la mesure à J-1 inhibe la prévision des pics qui sont des événements ponctuels. Le fait de travailler avec de telles données nous oblige à surveiller les prévisions faites suivant une période de pics.

Le facteur «**température min** » est présent dans le modèle été. Les pics sont enregistrés lors de températures faibles. Le coefficient de ce facteur est négatif, une température négative provoque une hausse de l'écart. L'erreur de prévision devient moins importante quand la température augmente.

Dans les deux modèles la variable **moy_PM10_mesB** a un coefficient positif, il y a donc un problème de prévision plus important quand on se situe dans une période de forte concentration en PM10. Un coefficient négatif est associé au facteur **Moy_PM10_mod**, ce qui veut dire que quand la prévision de CHIMERE est grande l'écart avec les observations diminue. En pratique les pics sont très mal prévus. Le faible nombre d'événement de la sorte ne permet pas de changer le signe du coefficient en question. Ce dernier point nous montre encore une fois que le modèle n'est **pas apte à prévoir les pics** de concentration en PM10.

Il reste intrigant de voir un coefficient négatif pour la variable « **vitesse du vent** » (modèle hiver), l'erreur de prévision semble être moins importante quand il y a beaucoup de vent (sans prendre en compte sa direction).

v. Bilan

L'INERIS a développé un script R qui, une fois lancé, sortira un modèle de régression linéaire multiple par station. Cette méthode apporte un gain de temps conséquent, mais une étude spécifique faite sur chaque station offrirait des résultats plus performants. Les modèles ainsi créés sont très différents sur les trois stations de fond traitées à Bordeaux.

La présence de certains facteurs explicatifs semble étonnante comme NO_2 et O_3. Ces deux facteurs ne permettent à priori pas d'expliquer la concentration en PM_{10} mais ils sont présents dans les modèles. Dans le cadre de la prévision ces variables sont estimées à partir d'autres modèles, les estimations faites comportent leur propre terme d'erreur. Le travail avec J-1 permet d'utiliser des mesures fiables mais la prévision des événements ponctuels peut en pâtir. Pire, la présence de ces variables peut apporter de la multicolinéarité dans le modèle comme nous le verrons par la suite.

Les jours de la semaine (ou une variable binaire "week-end ") servent de variable temporelle pour l'INERIS, leur contribution sur la concentration en PM_{10} reste faible (de l'ordre de 1 ou 2µg/m^3). Le graphique des séries temporelles sur 2012 nous laisse imaginer une tendance été et hiver, la variabilité de la concentration en PM_{10} semble beaucoup plus élevée durant la saison froide.

IV. Travail d'amélioration

AIRAQ et INERIS travaillent tous les deux avec des modèles linéaires. Ces modèles ne sont pas assez performants, notamment pour la prévision de pic de pollution. Le principe de régression linéaire repose sur la minimisation de l'estimateur des moindres carrés. Le modèle linéaire s'obtient en calculant les coefficients du modèle qui permettent de minimiser cette statistique, l'objectif de la méthode est d'avoir le meilleur modèle "en moyenne".

Sur une l'année 2012 on dénombre 14 pics, ces événements sont ponctuels et très peu représentés. Isoler les jours de pics permet de faire un modèle en éliminant la masse qui ampute leurs prévisions. Le modèle alors créé sur cette nouvelle base de données devra être utilisé sous certaines conditions à définir. Pour mettre en place un tel système il faut définir la frontière à partir de laquelle utiliser ce nouveau modèle, et imaginer un protocole d'utilisation qui soit le meilleur possible afin de ne pas lancer le mauvais modèle de prévision au mauvais moment.

Dans le cas où il existerait une journée type "pic de pollution" définie selon l'état de certains facteurs météo, la méthode CART est une alternative non linéaire à la régression linéaire multiple. Il s'agit d'une méthode de modélisation par arbre binaire (voir annexe 4).

Pour définir la meilleure méthode de modélisation il faut les mettre en place sur une base de données de référence. Les modèles bruts sont alors couplés à une nouvelle base de données pour faire des essais de prévisions.

L'essai des modèles sur un cas concret permet aussi d'évaluer les qualités et défaillances de la méthode, le but étant de tirer des leçons de toutes ces mises en pratiques.

1. La station d'étude

Le choix de la station est très important pour tester les outils de modélisation. Il est préférable de travailler sur une station où les mesures sont les plus variables. En effet la prévision des pics de pollution doit être analysée par la suite, il faut donc que les données comptent un nombre de pics suffisamment grand.

La ville de Bordeaux dénombre 3 stations urbaines de fond (31001: Grand-Parc, 31002: Talence et 31007: Bassens).Les graphiques (figure 3) présentent l'évolution de la concentration en PM_{10} pendant le $2^{ème}$ semestre (voir encadré bleu page 29) de 2012. On indique en rouge sur le graphique le seuil de pic (50 µg/m^3) et un second seuil informatif placé à 35µg/m^3 (en bleu).

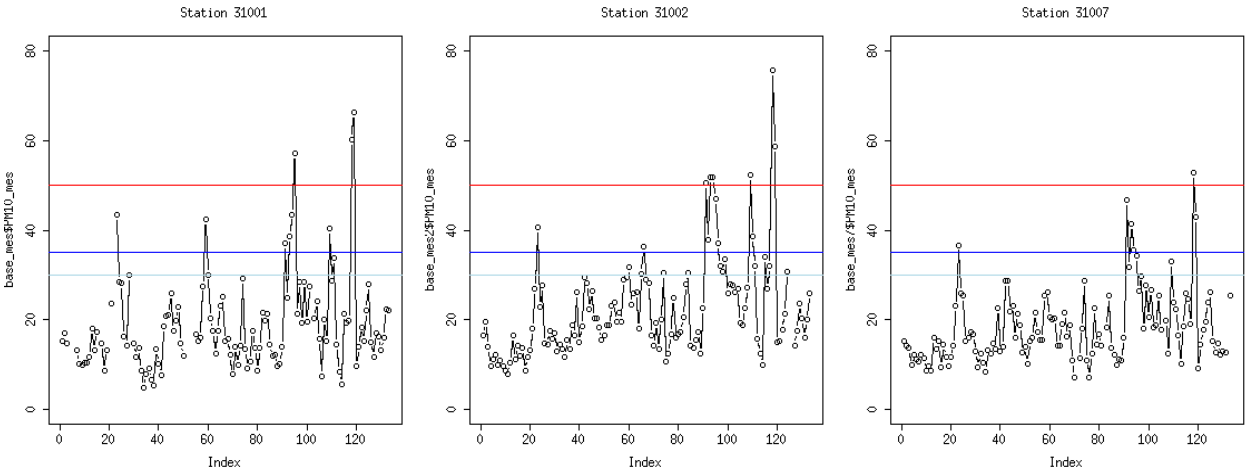


Figure 3 : séries chronologiques des mesures de 2012 sur les trois stations de Bordeaux.

La station 31002 est placée dans une zone résidentielle, donc les pics de PM10 y sont plus fréquents notamment à cause des émissions dues au chauffage au bois.

Tableau 13 : Dénombrement du nombre de valeurs dépassant certaines valeurs de référence

Stations	Pics (>50 µg/m^3)	>35 µg/m^3	>30 µg/m^3
31001	3	9	11
31002	6	12	23
31007	1	6	9

Les 6 pics enregistrés sur la station 31002 (tableau 8) sont décisifs dans le choix de la station qui sera utilisée pour faire nos études. De plus, des relevés ponctuels les jours de pics, en parallèle des analyseurs automatiques, permettent de différencier l'origine primaire ou secondaire des particules.

Il existe différentes versions du modèle CHIMERE, jusque là nous avons fonctionné avec FRA10, c'est-à-dire sur la zone France avec une précision de 10X15km. Une version plus précise nommée FRA05 sera prochainement mise en place. Il s'agit alors pour AIRAQ de ne plus corriger les sorties de FRA10, mais de commencer à travailler sur les sorties de FRA05.

Problème : l'historique des données pour la nouvelle version FRA05 ne commence qu'à partir de juillet 2012.

2. La sélection de variable

i. Importance du choix des variables

La sélection des variables est une étape importante en régression linéaire. Le modèle doit prendre en compte le plus d'information possible sans tomber dans la **surparamétrisation**. Un modèle qui contient trop de facteurs explicatifs ne sera pas généralisable à d'autres jeux de données. L'INERIS réutilise le modèle brut pour faire de la prévision, le piège serait de vouloir ajouter un maximum de variables explicatives dans le modèle statistique en se fiant à l'amélioration du R^2.

Il faut éviter les effets de **multicolinéarité** qui se produisent quand une variable explicative peut s'écrire comme combinaison linéaire des autres variables explicatives. Une telle variable doit être détectée pour éviter des problèmes numériques lors de certains calculs. L'inverse d'une matrice (pXq) ne peut être calculé que si elle est de rang p (ou régulière). Si elle est de rang inférieur, autrement dit s'il existe des relations linéaires entre ses colonnes, alors elle est singulière et **non inversible**. Plus simplement, la régression linéaire devient infaisable.

L'INERIS et AIRAQ développent des modèles en y intégrant des variables météo. Intuitivement, il ne serait pas étonnant d'y retrouver des problèmes de colinéarité (par exemple entre la température et les rayonnements solaires).

Un effet supposé de colinéarité entre deux variables peut être évalué par la valeur de corrélation de Pearson. Une valeur proche de 1 ou de -1 laisse supposer une liaison très forte entre les facteurs concernés.

Un moyen identifier les variables impliquées dans les multicolinéarités est d'effectuer des **régressions linéaires de chacune des variables en fonction des autres**. On calcule ensuite :

- Le **R^2** de chacun des modèles. Si le R^2 vaut 1, alors il existe une relation linéaire entre la variable dépendante du modèle (X_k) et les variables explicatives ($X_1,\dots,X_{k-1},X_{k+1},\dots,X_n$).
- La **tolérance** pour chacun des modèles. La tolérance vaut ($1-R^2$). Elle est utilisée comme un critère de filtrage des variables. Si une variable a une tolérance inférieure à un seuil fixé (la tolérance est calculée en prenant en compte les variables déjà utilisées dans le modèle), on ne la laisse pas entrer dans le modèle car sa contribution est négligeable et elle risquerait d'entraîner les problèmes numériques que nous avons soulignés plus tôt.
- Le **VIF**. Le VIF ou Variance Inflation Factor qui est égal à l'inverse de la tolérance. Des valeurs élevées de VIF indiquent donc la présence de multicolinéarité.

Les interactions d'ordre 2 servent à corriger un effet conjoint existant entre les variables concernées. Si deux variables sont fortement corrélées il ne s'agit pas d'entrer l'interaction entre ces variables dans le modèle, mais plutôt de rejeter la moins pertinente. Ajouter une variable d'interaction ne corrigera pas le problème de multicolinéarité. Ni AIRAQ et ni INERIS ne prennent en compte les interactions d'ordre 2 car le nombre de variables est déjà très élevé et les modélisateurs ont préférés ne pas sur-paramétrer le modèle.

Il existe plusieurs critères pour sélectionner des variables explicatives parmi k variables explicatives disponibles.

Le critère du R^2 est le plus connu mais pas le plus pertinent car il augmente avec l'introduction de nouvelles variables même si celles-ci sont peu corrélées avec la variable expliquée.

Il existe alors plusieurs alternatives (voir Annexe 2):

- le R^2 ajusté (R^2_{aj})
- le critère AIC
- le critère BIC
- autres…

Le logiciel R se sert du critère AIC par défaut dans sa fonction "step".

ii. Les méthodes de sélection

Il n'est pas raisonnable de se fier uniquement aux résultats statistiques fournis par un programme informatique. En effet, pour décider d'ajouter ou de supprimer une variable dans un modèle il faut conserver une part d'intuition et de logique.

Les méthodes de type **pas à pas** consistent à considérer d'abord un modèle faisant intervenir un certain nombre de variables explicatives. Puis elles procèdent par élimination ou ajout successif de variables.

- la méthode est **descendante** lorsqu'elle élimine des variables
- la méthode est **ascendante** lorsqu'elle ajoute des variables
- La méthode **stepwise** est une combinaison de ces deux méthodes.

Toutes ces procédures ne mènent pas forcément à la même solution quand elles sont appliquées au même problème.

Il faut rester vigilant par l'automatisation de ces méthodes sur les logiciels de statistique car le problème de multicolinéarité, comme nous l'avons souligné précédemment, n'est pas pris en compte dans le processus de sélection.

AIRAQ et INERIS ont utilisé une méthode simple de sélection de variables en se servant de l'indice AIC (par défaut sous R) avec la sélection ascendante.

L'INERIS est allé un peu plus loin dans la sélection des variables explicatives:

Pour chaque variable explicative sont calculées les sous-variables **min**, **max**, **moyenne**, **amplitude** (delta), **log** (cas de la couche limite) des différentes journées (données horaires). Ces sous-variables sont stockées, par la suite seulement une sera choisie pour chaque variable pour ne pas accentuer l'effet de multicolinéarité dans le modèle.

Par exemple : la température minimum journalière et la température maximum journalière sont très corrélés, il faut donc en choisir une des deux pour représenter la variable température.

La sélection des variables commence avec le calcul des corrélations (Pearson) entres les sous-variables (**min**, **max**,…) et la concentration du polluant étudié. Pour chaque variable on garde la sous-variable la plus corrélée avec la concentration en polluant est conservée. Si la valeur de corrélation est inférieure à 0.30 pour toutes les sous-variables, la variable concernée est alors totalement supprimée. Le seuil de 0.30 est une valeur arbitraire choisie empiriquement par l'INERIS lors de la mise en place et de l'évaluation de la méthodologie. La corrélation de Pearson calcule le lien de la relation linéaire entre deux variables, il n'est donc pas surprenant de s'y fier dans le cas d'une régression linéaire.

Pour une modélisation à la station la méthode de sélection "bestsubset " est plus intéressante car tous les modèles possibles sont comparés. Le nombre de variables explicatives est assez conséquent, cette méthode prendra plus de temps que celles ascendantes ou descendante. La sélection stepwise est un bon compromis entre efficacité et durée de mise en place. Le plus simple est de tester toutes ces méthodes sur une même base de données et de comparer les modèles ainsi créés, la méthode plus performante sera ainsi évaluée.

Le critère BIC favorise les modèles plus parcimonieux (qui ont le moins de variables possible). AIC a été introduit pour retenir des variables pertinentes lors de prévisions (voir annexe 1)

L'intuition des variables inclues dans le modèle est très importante. Les **matrices de corrélation** peuvent être tracées pour se faire une idée des variables qui semblent corrélées, et éventuellement imaginer des transformations pour linéariser le lien entre des variables explicatives et la variable réponse.

L'INERIS ne peut pas prêter une attention particulière à chaque modèle, il est alors plus simple de travailler avec la sélection stepwise et le critère AIC. Le modèle créé ne doit pas contenir trop de variables car l'objectif n'est pas que le modèle colle parfaitement sur la période d'apprentissage, mais bien de faire un modèle généralisable à des fins de prévision. Pour définir un modèle performant sur la région de Bordeaux il est intéressant de mettre en place et d'évaluer les différentes méthodes de sélection. La méthode par itération de tests de Student est longue à développer mais laisse plus de marge à l'intuition, on procède manuellement avec le logiciel R.

3. Modèle sur les grandes valeurs

Un modèle défini uniquement sur les grandes valeurs pose beaucoup de questions lors de sa mise en place :

- Il faut définir ce qu'est une "grande valeur".
- Combien de données au minimum doit-on avoir à disposition pour faire un "bon" modèle ?
- Comment mettre en place un tel modèle dans un objectif de prévision ?
- L'erreur faisant suite à l'utilisation du modèle au mauvais moment ne doit pas être négligée.

L'essai se fait sur la station 31002, avec les données de l'année 2012. Une valeur est supposée grande lorsqu'elle dépasse le seuil de **35µg/m^3**. Cette valeur correspond à l'indice Atmo 6 pour la concentration en PM_{10}.

Si ce seuil est fixe, la modélisation sera difficile à mettre en place à l'INERIS. En effet, dans certaines stations les valeurs de concentration en PM_{10} dépassent rarement ce seuil (avec présence de certains pics). La base de données ainsi définie ne sera pas assez fournie pour faire un modèle fiable. Le seuil définissant les grandes valeurs doit alors s'adapter aux données des stations traitées. Si le modèle est créé sur une année complète on peut choisir la valeur associé à un certain quantile. Il faudrait au moins 30 valeurs dépassant ce seuil. Si aucun dépassement n'est détecté sur l'année la question se pose sur l'utilité d'un tel modèle.

i. Mise en place du modèle

Un premier essai de cette méthode de modélisation se fait avec les données à la station 31002. Les données du modèle AFM dans sa version FRA05 ne sont disponibles qu'à partir de juillet 2012. Le protocole de modélisation sera posé sur les données dont nous disposons en 2012 et on se réserve les données de 2013 pour évaluer la qualité de prévision du modèle. Ce travail devra être refait dès lors que l'historique disponible sera plus conséquent. Le plus simple serait d'avoir une période d'apprentissage d'au moins une année complète et une période d'évaluation d'une année.

Le script de modélisation utilisé à INERIS est modifié pour permettre de faire cette modélisation. L'étape de sélection de variable (en calculant les sous-variables min, max,…) est conservée pour cette étape de modélisation.

Su ce jeu de données les grandes valeurs sont désigné par le **3ème quartile** (choix de l'utilisateur). Le 3ème quartile se situe à 27.46µg/m^3, il y a alors 33 journées supérieures à cette valeur sur le deuxième semestre de 2012.

La régression linéaire semble appropriée au jeu de données, la fonction scatter.smooth sous R permet de s'en faire une idée. Un exemple des graphiques obtenus avec les données de moyennes journalières est présenté par la figure 4.

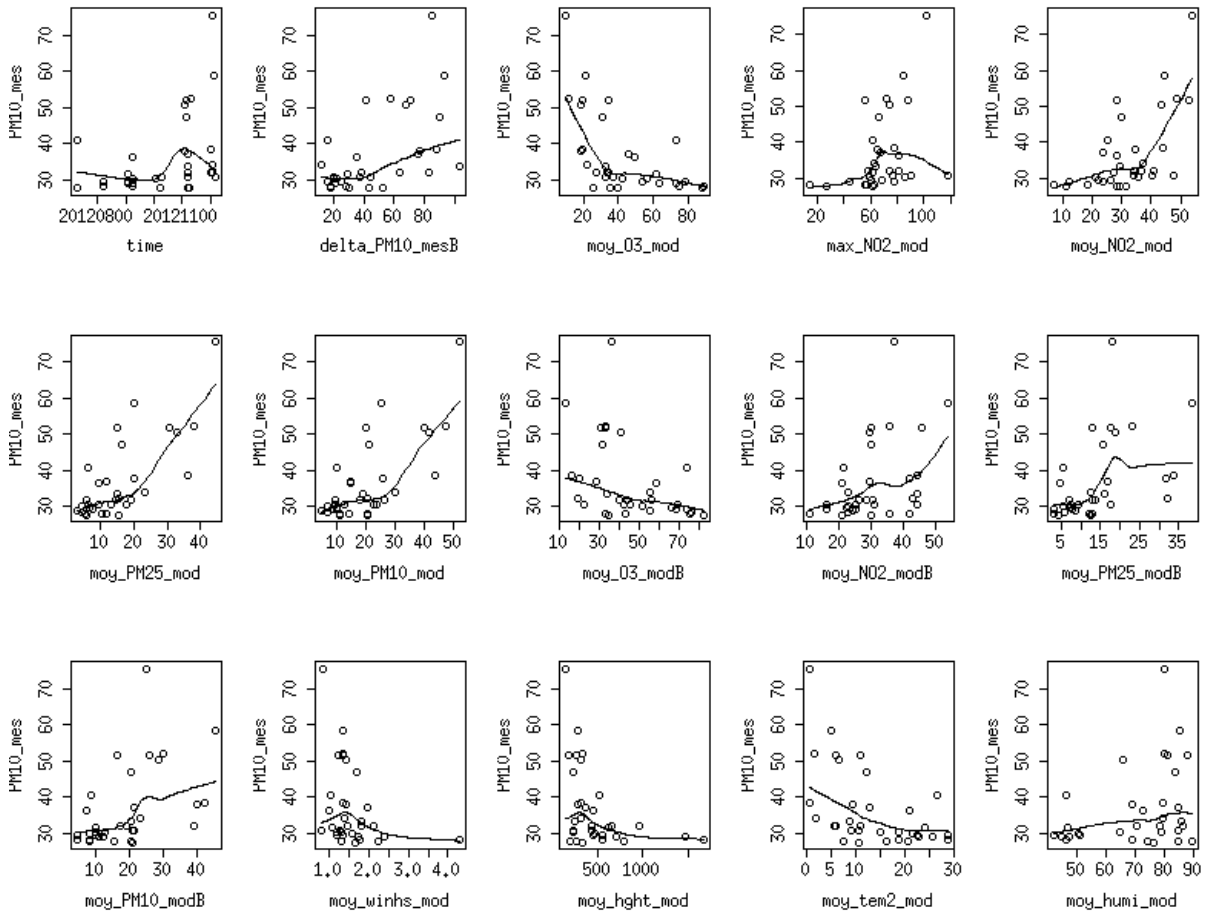


Figure 4 : Série de graphiques type scatter.smooth entre la variable concentration en PM_{10} et les variables météorologiques utilisées à l'INERIS

La relation de linéarité entre la concentration en PM_{10} et les variables explicatives est parfois floue mais l'hypothèse d'une relation linéaire semble convenir. Une transformation linéarisante pourrait être testée par la suite sur les variables qui peuvent apporter un doute.

La matrice de corrélation est tracée (figure 5) pour déceler les éventuels problèmes de colinéarité entre les variables utilisées pour la modélisation.

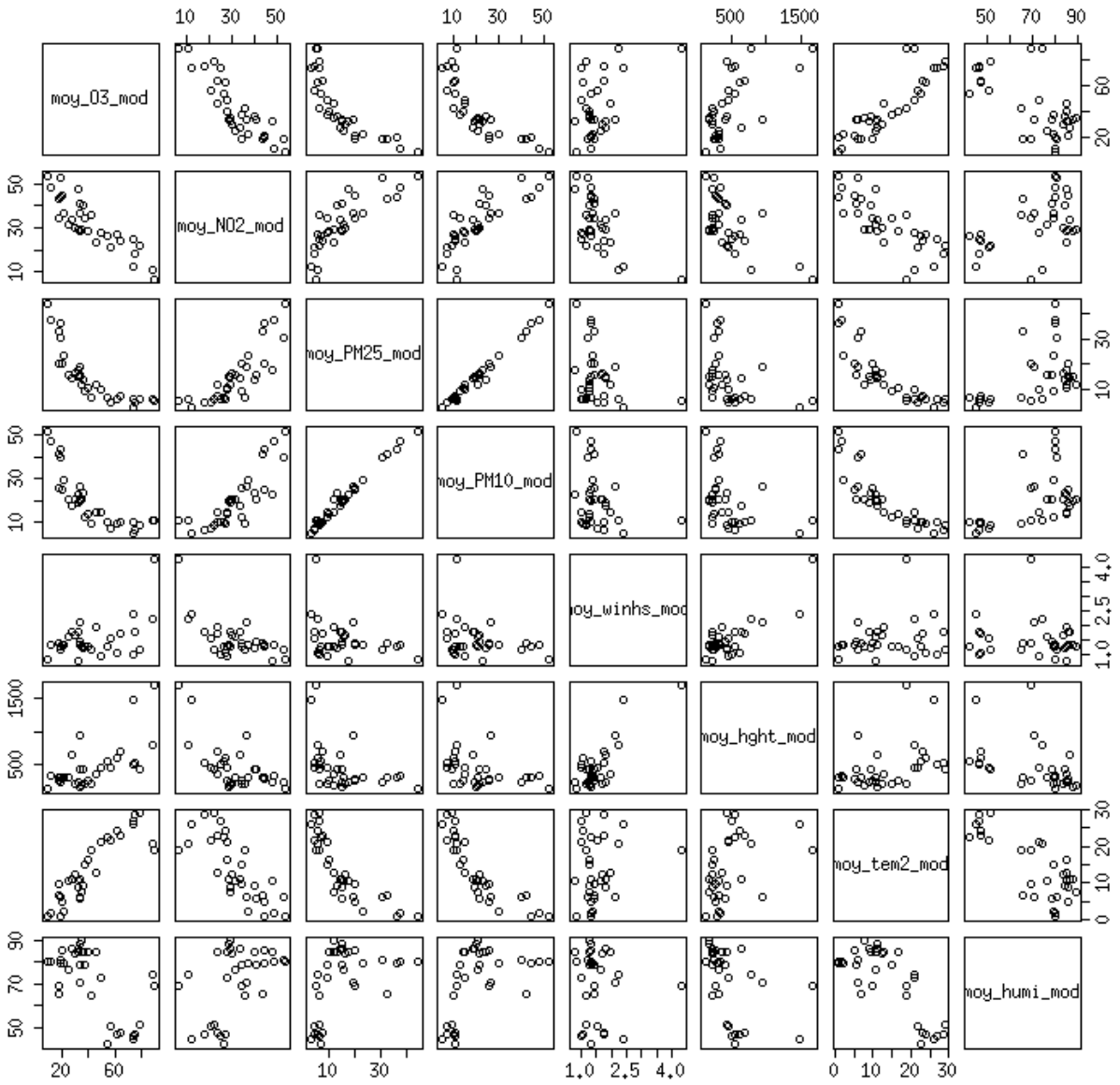


Figure 5 : Matrice de corrélation

moy_PM10_mod est une variable explicative qui représente les prévisions du modèle CHIMERE que nous voulons corriger. La corrélation entre cette variable et les autres variables de polluants est très forte.

La variable de température moy_tem2_mod semble corrélée à toutes les autres variables de polluants utilisées. Il semble également y avoir une relation entre la hauteur de la couche limite (hght) et la force du vent (winhs).

```
> eigen(t(x)%*%x)
$values
[1] 1.114903e+07 1.187535e+05 1.597914e+04 3.445659e+03 1.057395e+03
[6] 2.261890e+02 2.566906e+01 2.055957e+00
```

Les valeurs propres de cette matrice (X'X) sont éloignées de 0.

Les valeurs élevées du VIF (voir annexe 3) indiquent la présence de multicolinéarité. Ces valeurs sont inquiétantes si elles dépassent 5.

```
> vif(g1)
       delta_PM10_mesB             moy_O3_mod           min_NO2_mod
             18.163853              18.531080                5.919505
          moy_PM25_mod           moy_PM10_mod           moy_O3_modB
            193.022151             200.105683               15.581604
          moy_NO2_modB        delta_PM25_modB       delta_PM10_modB
             8.717890             220.746111              202.939585
       delta_winhs_mod       log_min_hght_mod          min_tem2_mod
             3.032814               2.290941               13.081656
          max_humi_mod delta_PM10_erreur_modB
             2.976370              17.266063
```

Il faut faire un tri parmi les variables explicatives car certaines valeurs de VIF sont très grandes. La présence de variables tels que O_3 et NO_2 a déjà été discutée dans la troisième partie du rapport, il serait judicieux de les mettre de coté pour la suite de l'exercice. En supprimant les supposées variables qui provoquent la multicolinéarité il nous reste les valeurs de vif:

```
vif(g4)
  moy_PM10_mod  delta_PM10_modB  delta_winhs_mod log_min_hght_mod
      3.915741          2.383605         1.497409          1.701207
  min_tem2_mod     max_humi_mod
      6.483271          1.798617
```

Le R^2 ajusté du modèle est passé de 0.63 pour le modèle à 14 variables explicative à **0.57** pour le modèle à 6 variables explicatives, mais ce dernier n'est plus soumis au problème de multicolinéarité. On prolonge la modélisation avec une sélection de variable type best subset avec le critère BIC (figure 6), la même procédure a été relancée en se servant du critère AIC et les résultats sont semblables.

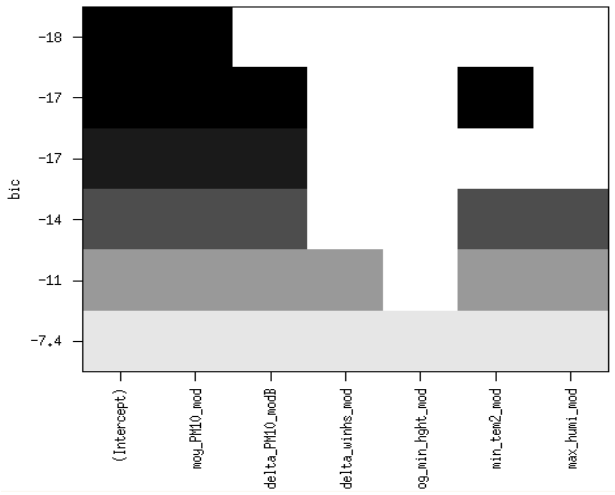


Figure 6 : Sélection de variables par méthode best subset

Selon le critère BIC, le meilleur modèle est celui ayant la plus petite valeur pour cette statistique, donc situé le plus haut dans la figure 6. Nous gardons le modèle à 4 variables explicatives : intercept, moy_PM10_mod, delta_PM10_modB, et min_tem2_mod.

La nécessité de transformer les variables explicatives peut être évaluée par des "ceres plot" ou les "partial residual plot".

ii. Validation du modèle

Le modèle est créé à partir d'une méthode de régression linéaire, il faut donc dresser l'étude des résidus studentisés. Ce point tient à évaluer l'indépendance, l'homoscédasticité et la normalité des résidus de manière graphique ou par des tests statistiques.

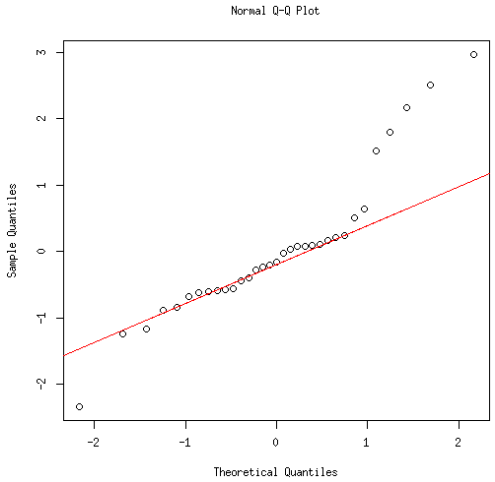


Figure 7 : QQplot des résidus du modèle

La queue de distribution observée dans la figure 7 s'écarte beaucoup de la droite rouge. L'hypothèse de normalité est rejetée mais elle peut être négligée. Le test de Shapiro nous conforte dans le rejet de l'hypothèse de normalité, à noter que le test est sensible aux nombres de valeurs dans les données d'étude :

```
        Shapiro-Wilk normality test

data:  as.matrix(res)
W = 0.8987, p-value = 0.004917
```

On rejette donc H_0 qui est l'hypothèse de normalité.

Pour l'étude de l'homoscédasticité des résidus on continue à travailler graphiquement (figure 8) en gardant en mémoire que le nombre de résidus relativement faible.

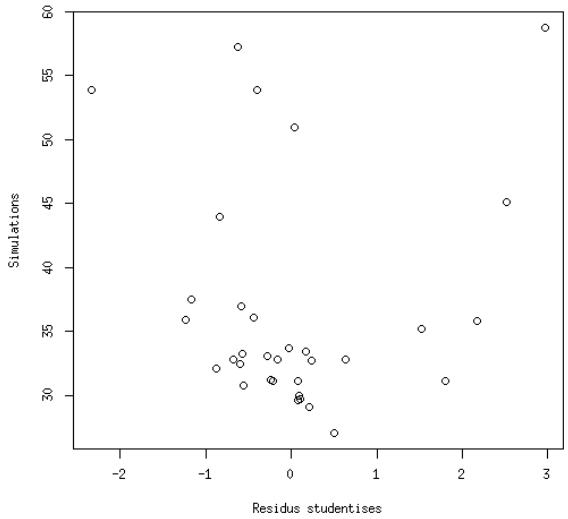


Figure 8 : Etude de l'homoscédasticité des résidus

Il n'y pas de structure qui se dégage du nuage de points, l'homoscédasticité semble être une hypothèse valide.

iii. Evaluation du modèle

En se servant de ce modèle pour faire des simulations sur la période d'acquisition on obtient le graphique (figure 9). Ce graphique représente seulement les journées dépassants le seuil que nous nous sommes fixés soit 27.46µg/m^3. La ligne horizontale en rouge indique la limite de 50µg/m^3 qui symbolise le seuil de dépassement pour les PM_{10}.

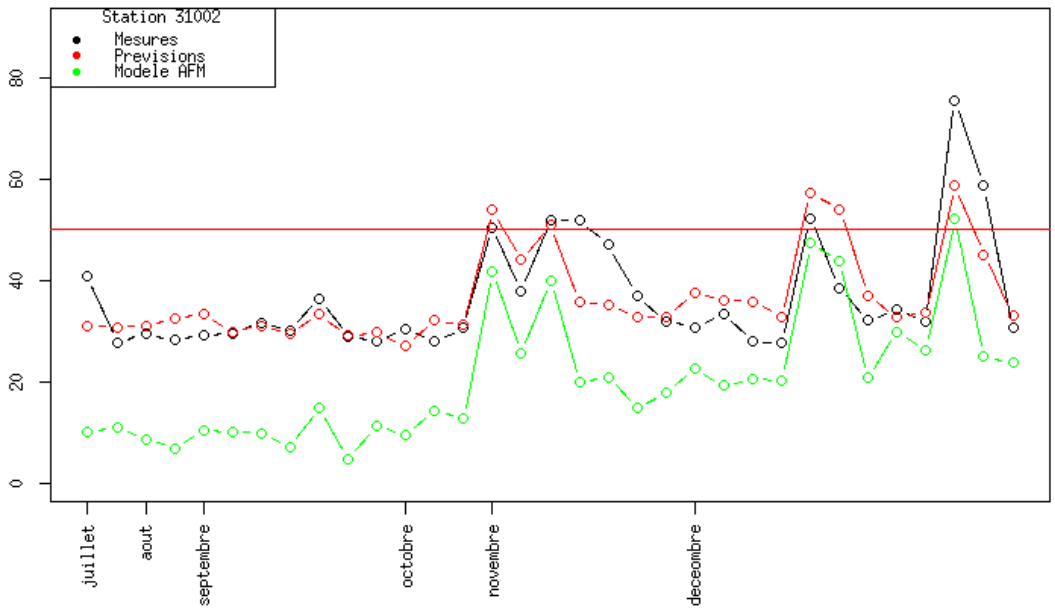


Figure 9 : Graphique temporel sur les jours de 2012 dépassants le seuil de 27.5µg/m^3

Sur les 6 pics enregistrés, seulement 4 sont bien prévus, et un dépassement de seuil a été annoncé à tort. La période de pics enregistrée en novembre est intéressante car le modèle a totalement sous estimé la période suivant le premier dépassement mesuré. Le modèle d'adaptation statistique colle globalement beaucoup mieux aux mesures que le modèle AFM. La même étude graphique doit être réalisée sur une période neutre, car le modèle d'adaptation a été construit sur la base des données de 2012.

L'évaluation du modèle sur l'ensemble des données de 2013 est présentée par un graphique temporel (figure 10).

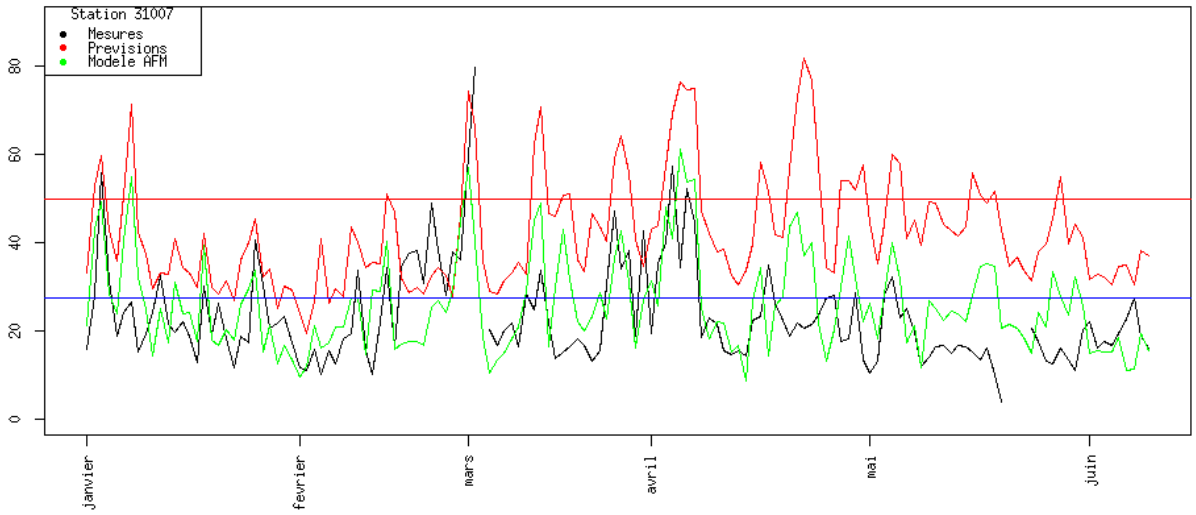


Figure 10 : Graphique temporel, évaluation du modèle sur 2013

Un point important à étudier est l'attitude du modèle sur les journées où la concentration enregistrée en PM_{10} est plus faible. Dans le cas où l'utilisateur se trompe de modèle il faut savoir quelle sera l'erreur de prévision engendrée.

L'utilisation du modèle au mauvais moment peut entrainer une lourde surestimation. La ligne bleue (figure 10) est tracée à 27.5µg/m^3 qui a été notre seuil lors de la création du modèle. Les mesures se situant au-dessus de ce seuil sont censée être bien représentées par le modèle que nous venons de créer (en rouge). Fin mars le modèle prévoit des dépassements de seuil à tort alors que les mesures se situent entre 30 et 50µg/m^3, le modèle à corriger (AFM en vert) a mieux fonctionné sur cette période. Les pics ont tous été prévu, mais beaucoup d'autres ont été annoncés à tort.

Si le lancement du modèle est décidé en fonction des prévisions du AFM, alors on s'expose à de grosses surestimations. On remarque entre mai et juin que le modèle AFM a prévu des concentrations en PM_{10} supérieures au seuil de 27.5µg/m^3 que nous nous étions fixé, alors que les mesures sont beaucoup plus faibles.

4. Autres modélisations envisagées

La régression non linéaire par la méthode CART est une piste envisagée. La méthode **Classification And Regression Tree** (CART) permet la construction d'un **arbre de décision** en déterminant une séquence de nœuds. Le modèle devra être mis en place sur un historique de données contenant au moins une année complète. L'écriture d'un script R détaillé sur la modélisation par CART peut être faite. L'outil sera testé sur un jeu de données plus fourni, comme par exemple les données de CHIMERE FRA10. La mise en place prochainement de CHIMERE FRA05 suggère de refaire la modélisation lorsque l'historique sera plus conséquent.

La durée du stage ne nous permet pas de mettre en place d'autres méthodes de régression non-linéaires.

Dans le cadre de prévision des pics les modèles linéaire "de base" ne sont clairement pas adaptés.

Conclusions

Le stage avait comme objectif d'améliorer les prévisions de qualité de l'air en région Aquitaine dans le système national PREV'AIR. Son déroulement particulier permet de tirer parti de ce qui est fait à AIRAQ et à l'INERIS.

AIRAQ et l'INERIS ont intégrés des modèles basés sur une méthode de régression linéaire. Les modèles existants ne sont pas assez performants pour faire de bonnes prévisions. Les dépassements de pic sont mal prévus, sur 15 dépassements enregistrés en 2012 seulement 7 sont détectés par les modèles. Ces événements sont ponctuels et difficilement prévisibles.

L'INERIS a automatisé la procédure de modélisation, la possible multicolinéarité entre les variables n'y est pas bien traité. L'exercice de modélisation fait dans ce rapport a bien mis en lumière ce problème lors de la modélisation sur les jours où la concentration dépasse un certain seuil. Les variables explicatives telles que la concentration en NO_2 ou en O_3 ont été retirées car elles amènent un problème de multicolinéarité dans le modèle. La mise en place de ce modèle conclu sur le rejet de cette méthode car sa manipulation est très dangereuse, et le en lui-même n'est pas très performants dans ses prévisions. La méthode CART semble intéressante pour ce type de données. La méthode a déjà été comprise, sa mise en place sera faite très prochainement.

On remarque que la concentration de PM_{10} est plus forte en hiver, cette tendance doit apparaître dans le modèle. Il faut évaluer quel facteur serait le plus intéressant entre "saison", "jour", "week-end",…

Un modèle de prévision performant n'a pas encore été mis au point mais certaines pistes ont été écartées, et d'autres peuvent voir le jour pendant la suite du stage.

Mes maîtres de stage m'ont laissé une totale liberté dans mes choix. Chaque décision pouvait être discutée et argumentée avec le personnel de l'équipe. Des réunions ponctuelles ont permis de faire le point sur les résultats obtenus et de définir des objectifs pour la suite du stage. Ce type de déroulement apporte une régularité dans le travail.

Les méthodes de modélisation ont été longuement discutées, notamment avec Frédéric Lavancier qui a pu proposer de nouvelles idées et rejeter certaines pistes que nous avions imaginés.

Le stage a été une très bonne expérience. J'espère pouvoir évoluer dans le même climat de confiance, et garder cette liberté de communication avec mes collègues pour mon avenir. Le domaine de la qualité de l'air est très intéressant et reste une très bonne piste que j'aimerais poursuivre à la suite de mon M2.

Tableaux et figures

Annexe 1 : La sélection de variables

i. Les méthodes basées sur un critère de sélection

Méthode descendante :

La procédure commence par effectuer une régression pour le modèle incluant toutes les k variables explicatives à disposition. En calculer la valeur du critère qui a été choisi. Oter du modèle de départ tour à tour chaque variable et en enregistrer la valeur du critère de comparaison choisi. Il suffit ensuite de sélectionner le modèle le plus performant au sens du critère, et de recommencer l'étape précédente. La sélection s'arrête lorsque le modèle ne peut plus être amélioré en retirant un des facteurs du modèle.

Méthode ascendante :

La démarche est la même que précédemment sauf que l'on part d'un modèle à 1 facteur en ajoutant itérativement les variables explicatives dans le modèle.

Lorsque le nombre de variables explicatives, noté k, à disposition n'est pas trop élevé, il est envisageable de considérer **tous les modèles possibles**. C'est une méthode fastidieuse et difficile à utiliser sans un ordinateur rapide. Il faut calculer toutes les régressions possibles impliquant un sous-ensemble des k variables explicatives à disposition, soit un total de 2^k régressions. Les modèles sont comparés en calculant une statistique, souvent le R^2 ou l'AIC, le meilleur modèle selon ce critère sera gardé.

ii. Les critères de sélection

Le **coefficient de détermination ajusté** tient compte du nombre de variables. En effet, le principal défaut du R^2 est de croître avec le nombre de variables explicatives. Or, on sait qu'un excès de variables produit des modèles peu robustes. C'est pourquoi on s'intéresse davantage à cet indicateur qu'au R^2.

R^2_{aj} n'augmente pas forcément lors de l'introduction de variables supplémentaires dans le modèle. L'utilisateur a alors la possibilité de comparer deux modèles n'ayant pas le même nombre de variables à l'aide du R^2_{aj} et choisir le modèle pour lequel cette statistique est la plus grande.

$$R^2_{aj} = 1 - \frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} - \frac{n-1}{n-p}$$

Le **critère AIC** représente donc un compromis entre le biais, diminuant avec le nombre de paramètres libres, et la parcimonie, volonté de décrire les données avec le plus petit nombre de paramètres possibles. Le meilleur modèle est celui possédant l'AIC le plus faible. Soit *L* la vraisemblance du modèle, k le nombre de paramètre du modèle et n le nombre d'observations:

$$AIC = -2\log(L) + 2k$$

On obtient dans le cas gaussien:

$$AIC = 2k + n\left[\ln\left(\frac{2\pi\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{n}\right) + 1\right]$$

Dans le cas où le nombre de paramètre est grand par rapport au nombre d'observations l'AIC corrigé est recommandé.

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

Le **critère BIC** est plus parcimonieux que le critère AIC puisqu'il pénalise plus le nombre de variables présentent dans le modèle.

$$BIC = -2\log(L) + 2\log(k)$$

AIC a été introduit pour retenir des variables pertinentes lors de prévisions, et le critère BIC vise la sélection de variables statistiquement significatives dans le modèle.

iii. Sélection par tests de Student

Le concept est semblable à ceux développés pour la sélection dite "automatique". Le choix du modèle ne se fera plus selon un critère choisi par l'utilisateur, mais en passant par des tests de significativité.

Méthode descendante :

Cette méthode consiste à calculer la régression en incluant toutes les k variables explicatives à disposition et à effectuer un t**est de Student** pour chacune des variables explicatives.

Deux cas se présentent :

- Les variables sont toutes significatives, le modèle est alors choisi. Nous arrêtons là notre analyse.
- La variable la moins significative est éliminée.

Cette procédure est réitérée jusqu'à ce que toutes les variables soient significatives.

Avantages et défauts de la méthode descendante:

- La méthode descendante est très satisfaisante pour l'utilisateur préférant avoir toutes les variables possibles afin de ne rien ignorer.
- C'est une procédure plus simple en terme d'interprétation que la sélection par critère AIC ou BIC.
- Problème : Il n'est plus possible de réintroduire une variable une fois qu'elle a été supprimée.

Méthode ascendante :

On commence par réaliser les k régressions possibles avec une seule variable explicative. Le test de Student est évalué pour chacune d'entre elles. On retient le modèle pour lequel la variable explicative est la plus significative.

On continue réaliser les k-1 régressions possibles avec une seule variable explicative. Le test de Student est évalué pour chacune d'entre elles. On retient le modèle pour lequel la variable explicative est la plus significative.

De même avec les (k-2) régressions possibles avec trois variables explicatives…

Le processus se termine lorsque nous ne pouvons plus introduire des variables significatives dans le modèle.

Avantages et défauts de la méthode ascendante:

- évite de travailler avec plus de variables que nécessaire.
- Problème : une variable introduite dans le modèle ne peut plus être éliminée.

Le problème de ces méthodes est l'irréversibilité de la sélection à chaque étape. Le modèle final peut alors contenir des variables non significatives. Pour résoudre ce problème il existe une méthode appelée **stepwise.**

La procédure stepwise :

Elle permet de réexaminer les tests de Student pour chaque variable explicative anciennement admise dans le modèle. On retire la variable la moins significative s'il y en a.

Le processus continue jusqu'à ce que plus aucune variable ne puisse être introduite ni retirée du modèle.

Dans la pratique, la procédure stepwise et la procédure descendante sont les plus utilisées.

Annexe 2 : Les indices de score

a. Erreur absolue moyenne (>0) en µg/m^3

L'erreur absolue indique la moyenne des différences entre les concentrations observées et calculées par le modèle.

$$EAM = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|$$

Cet indicateur est toujours positif. Il doit être le plus petit possible. Il renseigne sur l'amplitude moyenne de l'écart entre les concentrations observées et calculées par le modèle, indépendamment du signe de cet écart.

b. Biais en µg/m^3

Il s'agit de la moyenne des différences entre concentration observées et calculées.

$$Biais = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i) = \bar{x} - \bar{\hat{x}}$$

Il doit être le plus petit possible. Son signe donne une information sur la tendance générale du modèle :

- Biais négatif : surestimation
- Biais positif : sous-estimation

Il faut être vigilant en manipulant cette statistique car la formule ne tient pas compte du signe de la différence, ce qui veut dire que des écarts en positif et en négatif peuvent se compenser. Un biais petit ne signifie pas forcément que les données simulées coïncident avec les mesures.

c. Racine de l'erreur quad moyenne (RMSE) en µg/m^3

Le RMSE sur l'erreur moyenne du modèle. Un poids plus important est donné aux erreurs conséquentes.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}$$

d. Racine de l'erreur quad centrées moyenne (RcMSE)

La racine carrée de l'erreur quadratique moyenne centrée se calcule comme la RMSE, à ceci près que l'erreur est débiaisée avant d'être élevée au carré.

$$RcMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x}) - (\hat{x} - \bar{\hat{x}})\right]^2} = \sqrt{RMSE^2 - Biais^2}$$

e. Corrélation

La corrélation utilisée est celle de Pearson. On suppose implicitement une relation linéaire entre les variables concernées.

$$cor(x, \hat{x}) = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})(\hat{x_i} - \hat{x})\right]}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(\hat{x_i} - \bar{\hat{x}})^2}}$$

Le coefficient de corrélation mesure l'intensité du lien linéaire entre les variables. Il indique la propension des deux variables à évoluer dans le même sens ou, s'il est négatif, en sens contraire.

f. Variabilité

La variabilité est le rapport des écarts types des estimations et des observations.

$$variabilite(\hat{x}, x) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x_i} - \bar{\hat{x}})^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

Cette statistique indique si l'amplitude des concentrations prévues est inférieure ou supérieure à l'amplitude des concentrations observées.

Le tableau ci-dessous établit une correspondance entre la qualité des résultats du modèle et les indicateurs statistiques précédemment décrits.

Tableau 1 : *Appréciation des indices de score (valeurs subjectives)*

Qualité des résultats	PM10	O3
	Biais	
Bonne	-2,5 à 2,5	-10 à 10
Moyenne	-4 à -2.5 / 2.5 à 4	-15 à -10 / 10 à 15
Mauvaise	<-4 ou >4	<-15 ou >15
	Ecart Absolu moyen	
Bonne	0 à 4	0 à 15
Moyenne	4 à 6,5	15 à 25
Mauvaise	>6,5	>25
	RMSE	
Bonne	0 à 5,5	0 à 20
Moyenne	5,5 à 8	20 à 30
Mauvaise	>8	>30
	RcMSE	
Bonne	0 à 4	0 à 15
Moyenne	4 à 6,5	15 à 25
Mauvaise	>6,5	>25
	Corrélation	
Bonne	>0,8	>0,8
Moyenne	0,6 à 0,8	0,6 à 0,8
Mauvaise	<0,6	<0,6
	Variabilité	
Bonne	0,8 à 1,2	0,8 à 1,2
Moyenne	0,6 à 0,8 / 1,2 à 1,4	0,6 à 0,8 / 1,2 à 1,4
Mauvaise	<0,6 ou >1,4	<0,6 ou >1,4

Annexe 3 : Méthode CART

i. Définitions

La méthode **Classification And Regression Tree** (CART) permet la construction d'un **arbre de décision** en déterminant une séquence de nœuds.

- Un **nœud** est défini par le choix conjoint d'une variable explicative et d'une division qui induit une partition en deux classes.
- Une **division** est elle-même définie par une valeur seuil de la variable quantitative sélectionnée ou un partage en deux groupes des modalités si la variable est qualitative.
- À la racine (nœud initial) correspond l'ensemble de l'échantillon ; la procédure est ensuite itérée sur chacun des sous-ensembles.

A un **attribut qualitatif** ayant *n* modalités, on peut associer autant de tests qu'il y a de partitions en deux classes. Dans le cas d'**attributs continus** il y a une infinité de tests envisageables. On découpe l'ensemble des valeurs possibles en segments, ce découpage peut être fait selon les connaissances de l'utilisateur ou de façon automatique.

Dans le cas **Y quantitative**, **à chaque nœud terminal est associée une valeur : la moyenne des observations associées à ce nœud terminal**. Dans le cas qualitatif, chaque nœud terminal est affecté à une classe T` de Y en considérant le mode conditionnel :

- celle la mieux représentée dans le nœud et il est ensuite facile de compter le nombre d'objets mal classés
- la classe a posteriori la plus probable au sens bayésien si des probabilités a priori sont connues

Remarques:

Le CART fabrique des arbres binaires (toujours deux branches par nœud). C'est un des algorithmes les plus performants et les plus répandus.

*Il existe un autre algorithme **C4.5**, ou encore son amélioration nommée **C5**, qui fabrique des arbres qui ne sont pas nécessairement binaires (0 à n branches par nœud).*

ii. Critères de séparation

Le critère de division repose sur la définition d'une **fonction d'hétérogénéité** (ou de **désordre**). L'objectif étant de partager les individus en groupes les plus homogènes **au sens de la variable à expliquer**. L'hétérogénéité d'un nœud se mesure par une fonction non négative qui doit être :

- nulle si, et seulement si, le nœud est homogène (tous les individus appartiennent à la même modalité ou prennent la même valeur de Y).
- Maximale lorsque les valeurs de Y sont équiprobables ou très dispersées.

Parmi toutes les divisions admissibles d'un nœud, l'algorithme retient celle qui minimise la somme des désordres des groupes d'individus qui découlent du nœud.

Il y a plusieurs fonctions d'hétérogénéité par type de variable à expliquer (qualitative ou quantitative). Le rapport ne traitera que de celles utilisées par défaut dans les logiciels.

Si **Y est qualitative** :

Soient Y la variable réponse à m motalités T_l pour l=1,…, m, n_k l'effectif du k-ième nœud (ou "classe"), on note la probabilité qu'un élément du k-ième nœud appartienne à la l-ieme classe :

$$p_{lk} = P(T_l|k),$$

avec: $$\sum_{l=1}^{m} p_{lk} = 1$$

Un nœud k de l'arbre représente un sous-ensemble de l'échantillon d'effectif:

$$n_k = \sum_{l=1}^{m} n_{lk}$$

Le désordre au sein du k-ième nœud est alors :

$$D_k = -2 \sum_{l=1}^{m} n_k p_{lk} \log(p_{lk})$$

Par convention :

$0 \log(0) = 0$

L'hétérogénéité de la partition peut se mesurer par:

$$D = \sum_{k=1}^{K} D_k = -2 \sum_{k=1}^{K} \sum_{l=1}^{m} n_k p_{lk} \log(p_{lk})$$

Cette quantité est positive ou nulle, elle est nulle si et seulement si les probabilités p_{lk} ne prennent que des valeurs nulles sauf une égale à 1 correspondant à l'absence de mélange.

Si on pose n_{lk} l'effectif observé de la l-ième classe dans le k-ième nœud, alors p_{lk} peut être estimé :

$$p_{lk} = \frac{n_{lk}}{n_k}$$

Les fonctions d'hétérogénéité sont estimées (on remplace p_{lk} par sa forme développée):

$$D_k = -2\sum_{l=1}^{m} n_{lk} \log\left(\frac{n_{lk}}{n_k}\right)$$

$$D = \sum_{k=1}^{K} D_k = -2\sum_{k=1}^{K}\sum_{l=1}^{m} n_{lk} \log\left(\frac{n_{lk}}{n_k}\right)$$

Dans le cas Y qualitative, il existe plusieurs fonctions d'hétérogénéité, ou de désordre d'un nœud. Ce peut être le nombre de mal classés, un critère défini à partir de la notion d'entropie ou encore à partir de la concentration de Gini. En pratique, il s'avère que le choix du critère importe moins que celui du niveau d'élagage.

L'indice de Gini par exemple est très utilisé en économie comme mesure des inégalités dans une population. L'indice est toujours compris entre 0 et 1, s'il est proche de 0, ceci signifie que les différences relatives sont en moyenne faibles par rapport à la moyenne: les inégalités dans la classe sont faibles. Si l'indice de Gini est proche de 1, alors au contraire il y a de fortes différences relatives en moyenne : les inégalités sont fortes. Le but de la méthode CART est d'avoir des classes homogènes à la sortie des points terminaux.

Si **Y est quantitative** :

Si on note n_j le nombre d'individus le j-ème classe et J le nombre de classes différentes, on peut alors estimer l'hétérogénéité de la classe j par:

$$D_j = \sum_{i=1}^{n_j} (y_{ij} - y_{.j})^2 \qquad y_{.j} = \sum_{i=1}^{n_j} y_{ij}$$

Et ainsi on mesure l'hétérogénéité du modèle avec:

$$D = \sum_{j=1}^{J}\sum_{i=1}^{n_j} (y_{ij} - y_{.j})^2$$

Il s'agit de l'inertie intra qui vaut D = 0 si et seulement si $y_{ij}=y_{.j}$ pour tout i et tout j.

On pose :

$$y_{..} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}$$

La différence d'hétérogénéité entre l'ensemble non partagé et l'ensemble partagé selon la partition J est :

$$\Delta = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-y_{.j})^2 - \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-y_{..})^2 = \sum_{j=1}^{J} n_j(y_{..}-y_{.j})^2$$

L'objectif, à chaque étape, est de maximiser Δ c'est-à-dire de trouver la variable induisant une partition en 2 classes associée à une inertie intraclasse minimale ou encore qui rend l'inertie interclasse la plus grande.

iii. L'algorithme CART

L'algorithme considéré nécessite :

- la définition du critère permettant de sélectionner la "meilleure" division parmi toutes celles admissibles pour les différentes variables.
- une règle permettant de décider qu'un nœud est terminal
- l'affectation de chaque nœud terminal à l'une des classes ou à une valeur de la variable à expliquer.

Le point 2 est le plus délicat. Il correspond encore à la recherche d'un modèle parcimonieux. Un arbre trop détaillé, associé à une sur-paramétrisation, est instable et donc probablement plus défaillant pour la prévision d'autres observations.

L'Algorithme peut s'expliquer assez simplement :
Pour chaque nœud on choisit la variable qui, par ses catégories, sépare le mieux les individus en fonction de la variable réponse. Pour faire ce choix il faut se référer à la fonction d'hétérogénéité appliquée à chaque groupe possible.

On réitère l'opération jusqu'à ce qu'elle ne soit plus possible ou plus souhaitable (il ne reste plus que des nœuds terminaux, plus de variable explicative,…).

iv. Avantages et inconvénients de la méthode CART

Avantages :

- Résultats explicites
- Règles de décisions simples
- Peu de perturbation des individus extrêmes (voir les pics de pollution)
- Ne requière pas d'hypothèses sur les distributions des variables et semble particulièrement adaptée au cas où les variables explicatives sont nombreuses.

Inconvénients :

- Utilisation des variables non simultanée mais séquentielle
- Nécessité d'un grand nombre d'individus : avoir au minimum 20 ou 30 individus par nœud
- Problèmes des arbres trop étoffés :
    - Complexité de l'arbre, trop de règles
    - Trop spécifique aux données d'apprentissage

v. Elagage

Le dernier problème de la méthode CART qui a été soulevé peut être atténué par une méthode dite d'**Elagage**.

La première stratégie utilisable pour éviter un sur-ajustement massif des arbres de décision consiste à proposer des critères d'arrêt lors de la phase d'expansion. C'est le principe du **pré-élagage**. Nous considérons par exemple qu'une segmentation n'est plus nécessaire lorsque le groupe est d'effectif trop faible, ou encore lorsque la pureté d'un sommet a atteint un niveau suffisant nous considérons qu'il n'est plus nécessaire de le segmenter.

La méthode CART peur amener à des arbres extrêmement raffinés et donc à des modèles de prévision très instables car fortement dépendants des échantillons qui ont permis leur estimation. On se trouve donc dans une situation de sur-ajustement à éviter au profit de modèles plus parcimonieux donc plus robuste au moment de la prévision. Cet objectif est obtenu par une **procédure d'élagage** (ou pruning) de l'arbre. Le principe de la démarche consiste à construire une suite emboîtée de sous-arbres de l'arbre obtenu avec la méthode CART par élagage successif puis à choisir, parmi cette suite, l'arbre optimal au sens d'un critère définit par l'utilisateur.

Objectif: supprimer les parties de l'arbre qui ne semblent pas performantes pour prédire la classe de nouveaux cas (remplacées par un nœud terminal (associe à la classe majoritaire).

Processus: généralement de type "bottom-up" (du bas vers le haut: des extrémités vers la racine), base sur une estimation du taux d'erreur de classification. Un arbre est élagué à un certain nœud si le taux d'erreur estimé a ce nœud est inferieur au taux d'erreur obtenu en considérant les sous-arbres terminaux.

Bibliographie

[1] Arrêté du 21 décembre 2011 modifiant l'arrêté du 22 juillet 2004 relatif aux indices de la qualité de l'air - JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE (Dernière modification : 31/12/2012). Disponible sur :
http://www.airaq.asso.fr/fileadmin/user_upload/fichiers/REGLEMENTATION/joe_20111231_0021.pdf

[2] Bernard Delyon. *Régression* le 21 mai 2013.
Disponible sur :
http://perso.univ-rennes1.fr/bernard.delyon/regression.pdf

[3] François Bavaud. *Modèles et données: une introduction à la statistique uni-, bi- et trivariée*. Editions L'Harmattan, 1999.

[4] Frédéric Bertrand. Cours diapositives : *choix du modèle*. Année 2010-2011.
Disponible sur :
http://www-irma.u-strasbg.fr/~fbertran/enseignement/Estimation_2010/Cours1-Anado.pdf

[5] Frédérik Meleux, Laure Malherbe, Anthony Ung. *Modélisation – traitements numériques* (Décembre 2010).

[6] Frédérik Meleux, Laure Malherbe, Anthony Ung. *Modélisation – traitements numériques* (Décembre 2011).

[7] G. Ghattas (Université de Méditerranée). *Importance des variables dans les méthodes CART.*
Disponible sur :
http://lumimath.univ-mrs.fr/~ghattas/mypapers/importance.pdf

[8] Honoré C., Ung A., Corbet L., Malherbe L., 2012. Good Practice Guide on Urban Air Quality Forecast. CITEAIRII project,disponible sur:
http://www.citeair.eu/.

[9] Loi n° 96-1236 du 30 décembre 1996 sur l'air et l'utilisation rationnelle de l'énergie (Dernière modification : 14 juin 2006). Disponible sur :
http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000005622536&dateTexte=20101202

[10] Plaquette Prévision de la qualité de l'air. AIRFOBEP, 2007.
Disponible sur :
http://www.airfobep.org/docs/Prevision_qualite_air_07.pdf

[11] Wikistat, auteur inconnu. *Arbres binaires de décision* (dernière modification le 02/07/2012).
Disponible sur :
http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-cart.pdf

[12] Yadolah Dodge. *Analyse de régression appliquée*. Dunod, 1999.

[13] Auteur non connu (Université de Lille). *Apprentissage automatique : les arbres de décision* (dernière modification le 03/07/2002).
Disponible sur :
http://www.grappa.univ-lille3.fr/polys/apprentissage/sortie004.html