Prévision statistique de la qualité de l'air et d'épisodes de pollution atmosphérique en Corse

Wani W. Tamas

UNIVERSITE DE CORSE – PASQUALE PAOLI
ECOLE DOCTORALE ENVIRONNEMENT ET SOCIETE
UMR CNRS 6134 (SPE)

Thèse présentée pour l'obtention du grade de

DOCTEUR EN AUTOMATIQUE, SIGNAL, PRODUCTIQUE, ROBOTIQUE

Mention : Génie informatique, automatique et traitement du signal

Soutenue publiquement par

Wani Théo TAMAS

le 17 novembre 2015

Prévision statistique de la qualité de l'air et d'épisodes de pollution atmosphérique en Corse

---

Directeurs :

M. Gilles Notton, Dr-HDR, Université de Corse
M. Christophe Paoli, Dr-HDR, Université de Corse

Rapporteurs :

M. Matthias Beekmann, DR, Université Paris Est Créteil
M. Pierre-Charles Maria, PREM, Université de Nice Sophia Antipolis

Jury :

M. Matthias Beekmann, DR, Université Paris Est Créteil
M. Dominique Lambert, Dr-HDR, Université Toulouse III - Paul Sabatier
M. Pierre-Charles Maria, PREM, Université de Nice Sophia Antipolis
M. Jean-François Muzy, DR, Université de Corse
M. Gilles Notton, Dr-HDR, Université de Corse
M. Christophe Paoli, Dr-HDR, Université de Corse

UNIVERSITE DE CORSE – PASQUALE PAOLI
ECOLE DOCTORALE ENVIRONNEMENT ET SOCIETE
UMR CNRS 6134 (SPE)

Thèse présentée pour l'obtention du grade de

DOCTEUR EN AUTOMATIQUE, SIGNAL, PRODUCTIQUE, ROBOTIQUE

Mention : Génie informatique, automatique et traitement du signal

Soutenue publiquement par

Wani Théo TAMAS

le 17 novembre 2015

Prévision statistique de la qualité de l'air et d'épisodes de pollution atmosphérique en Corse

---

Directeurs :

M. Gilles Notton, Dr-HDR, Université de Corse
M. Christophe Paoli, Dr-HDR, Université de Corse

Rapporteurs :

M. Matthias Beekmann, DR, Université Paris Est Créteil
M. Pierre-Charles Maria, PREM, Université de Nice Sophia Antipolis

Jury :

M. Matthias Beekmann, DR, Université Paris Est Créteil
M. Dominique Lambert, Dr-HDR, Université Toulouse III - Paul Sabatier
M. Pierre-Charles Maria, PREM, Université de Nice Sophia Antipolis
M. Jean-François Muzy, DR, Université de Corse
M. Gilles Notton, Dr-HDR, Université de Corse
M. Christophe Paoli, Dr-HDR, Université de Corse

Remerciements

C'est une importante page de ma vie qui se tourne avec la fin de cette thèse. Ce fut une aventure pleine de bons moments, de découvertes et de dépaysements ; nombreux sont celles et ceux qui m'ont accompagné et que je souhaite remercier.

Je tiens tout d'abord à dire un grand merci à mes directeurs de thèse, Gilles Notton, Christophe Paoli, et bien sûr Marie-Laure Nivet qui m'a tout autant suivi et accompagné. Merci également à Cyril Voyant pour m'avoir aidé et conseillé pendant ces travaux. Votre très grande disponibilité tout au long de cette thèse (y compris depuis Istanbul) m'a été d'une grande aide. Merci de m'avoir accueilli en Corse et à l'Université, de m'avoir guidé pendant mon doctorat et pour vos nombreuses relectures depuis LaTeX. Nos discussions me manqueront (clarté ou pédagogie ?). La prochaine fois, on se lance dans le Deep Learning ! Et merci à Aurélia Balu, qui a travaillé avec nous en temps qu'alternante, pour sa collaboration et son amitié. Mon travail dans notre « Team Prévision » restera une très bonne expérience grâce à vous tous !

Je souhaite remercier les membres de mon jury de thèse pour avoir évalué mon travail. Merci particulièrement à Pierre-Charles Maria et à Matthias Beekmann pour avoir accepté d'être les rapporteurs de ma thèse. Merci à Dominique Lambert et à Jean-François Muzy pour avoir participé à mon jury. La soutenance de ma thèse restera un très bon souvenir.

Je tiens à remercier l'Agence d'Aménagement Durable, de Planification et d'Urbanisme de la Corse (AAUC), l'Agence de Développement Economique de la Corse (ADEC) et Qualitair Corse pour avoir financé ces travaux. Merci à l'ensemble des employés de Qualitair Corse pour m'avoir accueilli au sein de l'association pendant ces trois années. Merci à Jean-Luc Savelli, à Rosanna Casale, à Guillaume Grignion, à Mathieu Lion et Louise Declerck, à Florent Bordier, à Nicolas Bernardi, à Gabrielle Pochet et à Mateo Navarro. Grâce à vous j'ai découvert le monde de la surveillance de la qualité de l'air, et votre aide à été précieuse au bon déroulement de ma thèse. Merci également aux collègues des autres AASQA, notamment Morgan Jacquinot chez Air PACA. Merci à Météo-France pour la fourniture de données utilisées dans ces travaux, et particulièrement à Patrick Rébillout de Météo-France Ajaccio.

Je voudrais dire merci à tout le personnel du laboratoire Science Pour l'Environnement (UMR CNRS 6134) de l'Université de Corse. L'ensemble de l'équipe m'a permis de travailler sereinement. Je remercie tous ses membres,

enseignants-chercheurs, personnels techniques et administratifs. Merci particulièrement à mes amis du laboratoire ; merci à Gauthier Lapa et Damien Foures pour notre travail commun sur les différents aspects du brassage amateur, merci au Dr Tom Toulouse, à Aurélia encore une fois et à Damien Grandi (longue vie à DataSensia !), à Camille (Anton joue déjà à Carcassonne ?), Jean-Baptiste (Max, il s'appelle Max, c'est son nom !), Mohamed et Andreï qui m'ont permis d'excaver agréablement mes restes de physicien pour parler excitation des trous noirs, à Raphaël, Marie, Christelle, Romain, Lara, Hélène. Merci pour nos discussions scientifiques, brassicoles, nanardesques ou autre. Les pauses café et les sessions escalade/rando me manqueront.

Merci aussi aux amis de Corte, Violette, Cédric Choupi, Lilian, Tiphaine, Dr Cyril Berquier, Marion, Mathieu et Louise encore, Emilie et Vincent, Géraldine, Claire, Laura, Yohan, Julie Chaurand (bonne chance pour ta thèse !), Pauline, Mickael et Morgane (et Nina !), mes anciens colocs du début de ma vie en Corse, Paulin, Maxence et Vincent, merci à vous et à ceux que j'oublie pour le bon temps passé. Et un grand merci à Pascal et Henri du Cyrnéa, et à Jean-Hugues pour votre importante contribution la vie nocturne cortenaise !

Un grand merci à ma famille, sans qui je ne serais jamais arrivé où j'en suis aujourd'hui. Vous avez vu, je suis docteur ! Merci Maman, Papa et Johanna. Merci Denis, Magda et gros bisous à Gael et Milena ! Vous m'avez manqué pendant ces trois années qui n'ont pas été les plus faciles, mais le soleil revient toujours au fil des jours !

Je pense aussi à mes proches éloignés, mon cousin Gyban, Ryad, Benjy, Latigone, Mathilde, Marcus, Charlotte et les autres, vous m'avez bien manqués ! Merci aux amis alsaciens, particulièrement Marie et Célia mais la liste semble vraiment être trop longue pour figurer intégralement ici. A bientôt pour d'autres escapades au bout du monde !

Et bien sûr, un grand merci à toi ma merveilleuse Julie, pour m'avoir accompagné et supporté tout au long de cette thèse.

We can only see a short distance ahead,
but we can see plenty there that needs to be done.
Alan Turing

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

AASQA	Association Agréée de Surveillance de la Qualité de l'Air
AAUC	Agence d'Aménagement Durable, de Planification et d'Urbanisme de la Corse
ACP	Analyse en Composantes Principales
ADEC	Agence de Développement Economique de la Corse
ADEME	Agence De l'Environnement et de la Maîtrise de l'Energie
AG	Algorithmes Génétiques
AR	Auto-Régressif
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
AROME	Application de la Recherche à l'Opérationnel à Méso-Echelle
BASTER	BASe de données en TEmps Réel
BFGS	Algorithme de Broyden – Fletcher – Goldfarb – Shanno
BIC	Bayesian Information Criterion (critère d'information bayésien)
BP	Back Propagation (algorithme de rétropropagation de l'erreur)
C90	Centile 90, valeur dépassée par uniquement 10% des points de l'échantillon
CAH	Classification Ascendante Hiérarchique
CAPA	Communauté d'Agglomération du Pays Ajaccien
CART	Classification And Regression Tree (arbre de classification et de régression)
CFC	Chlorofluorocarbure
ChArMEx	Chemistry-Aerosol Mediterranean Experiment
Citeair	Common Information To European AIR
CLA	Couche Limite Atmosphérique
CMAQ	Community Air Quality Modelling System
CNRS	Centre National de la Recherche Scientifique
CORSiCA	Centre d'Observation Régional pour la Surveillance du Climat et de l'environnement Atmosphérique et océanographique en Méditerranée occidentale
COV	Composé Organique Volatil
COVNM	Composé Organique Volatile Non Méthanique
CTM	Chemical Transport Model (modèle de chimie-transport)
DG	algorithme de Descente de Gradient
DREAL	Direction Régionale de l'Environnement, de l'Aménagement et du Logement
DV	Direction du Vent (en degrés)

ECI	Epaisseur de la Couche d'Inversion (en m)
EMEP	European Monitoring and Evaluation Programme
FA	Forêt Aléatoire (random forest)
FB	Fractional Bias
FN	Faux Négatif
FP	Faux Positif
FTP	File Transfer Protocol
FV	Fractional Variance
GEO	Géopotentiel (en $m^2.s^{-2}$)
GES	Gaz à Effet de Serre
GNU GPL	GNU General Public Licence
GPU	Graphics Processing Unit, processeur graphique
GRIB	General Regularly-distributed Information in Binary form
HAP	Hydrocarbure Aromatique Polycyclique
HCL	Hauteur de la Couche Limite atmosphérique
HO_x	Famille des radicaux $OH\cdot$, $HO_2\cdot$ et peroxyles organiques $RO_2\cdot$
HR	Humidité Relative (en %)
IA	Intelligence Artificielle
IM	Information Mutuelle (en bits)
INERIS	Institut National de l'EnviRonnement Industriel et des risqueS
INS	Inventaire National Spatialisé
IPSL	Institut Pierre Simon Laplace
IQA	Indice de la Qualité de l'Air
k-nn	méthode des k plus proches voisins (k-nearest neighbor)
LAURE	Loi sur l'Air et l'Utilisation Rationnelle de l'Energie
LCSQA	Laboratoire Central de Surveillance de la Qualité de l'Air
LISA	Laboratoire Inter-universitaire des Systèmes Atmosphériques
LM	Algorithme de Levenberg – Marquardt
LMD	Laboratoire de Météorologie Dynamique
LNE	Laboratoire National de métrologie et d'Essais
MA	Moving Average (moyenne mobile)
MAE	Mean Absolute Error (erreur absolue moyenne)
MAPE	Mean Absolute Percentage Error (erreur absolue moyenne en pourcentage)
MBE	Mean Bias Error (erreur biaisée moyenne)
MEDDE	Ministère de l'Ecologie, du Développement Durable et de l'Energie
MLR	Multiple Linear Regression (régression linéaire multiple)
MM5	Modèle à Méso-échelle PSU/NCAR
MOCAGE	MOdèle de Chimie A Grande Echelle
MSE	Mean Squared Error (erreur quadratique moyenne)
NEB	Nébulosité (en %)
NNToolbox	Neural Network Toolbox
NO_x	Oxydes d'azote (NO et NO_2)
nRMSE	normalized Root Mean Squared Error (racine de l'erreur quadratique moyenne normalisée)
NWP	Numerical Weather Prediction (modèle de prévision numérique du temps)
P	Précipitations (en mm)

PA	Pression Atmosphérique (en hPa)
PACA	Région Provence-Alpes-Côte d'Azur
PAN	PéroxyAcétylNitrate
PIMENT	laboratoire Physique et Ingénierie Mathématique pour l'Energie, l'environnemeNt et le bâtimenT (PIMENT)
PM1	Particules de moins de 1 µm de diamètre aérodynamique
PM10	Particules de moins de 10 µm de diamètre aérodynamique
PM2.5	Particules de moins de 2,5 µm de diamètre aérodynamique
PMC	Perceptron MultiCouche (multilayer perceptron)
ppmv	Partie par million du volume
PSQA	Programme de Surveillance de la Qualité de l'Air
RBF	Radial Basis Function
RMSE	Root Mean Squared Error (racine de l'erreur quadratique moyenne)
RNA	Réseau de Neurones Artificiels
RS	Rayonnement Solaire (en $j.m^{-2}$)
RT	Rayonnement Thermique (en $j.m^{-2}$)
SACOI	SArdaigne-COrse-Italie
SARCO	SARdaigne-COrse
SARIMA	Seasonal AutoRegressive Integrated Moving Average
SARIMAX	Seasonal AutoRegressive Integrated Moving Average with eXogenous inputs
SCG	algorithme Scaled Conjugate Gradient
SOM	Self-Organizing Map (carte auto-organisatrice, ou carte de Kohonen)
SVM	Support Vector Machine (machine à vecteur de support)
TAC	Turbine A Combustion
TC	Température (en degrés Celsius)
TDNN	Time-Delay Neural Network
TFP	Taux de Faux Positif)
TK	Température (en degrés Kelvin)
TVP	Taux de Vrai Positif)
URL	Uniform Resource Locator (adresse web)
VN	Vrai Négatif
VP	Vrai Positif
VV	Vitesse du Vent (en $m.s^{-1}$)
WRF	Weather Research and Forecasting model
ZNI	Zone Non Interconnectée
ZPS	Zone Protection Spéciale
ZR	Zone Régionale
ZSC	Zone Spéciale de Conservation
ZUR	Zone Urbaine Régionale

Glossaire

advection	Transport par un champ vectoriel (comme le vent)
Air PACA	AASQA de la région PACA
Aria Base	Application de recherche pour la gestion des modèles prédictifs statistiques
Aria Web	Application d'aide à la décision pour la prévision quotidienne de la qualité de l'air
assimilation statistique	Technique permettant la prise en compte de mesures en temps réel par un modèle numérique déterministe
centile	Quatre-vingt dix-neuf valeurs permettant de diviser en cent parts égales l'échantillon ordonné d'une variable aléatoire
classification	classification automatique : séparation supervisée de données en plusieurs groupes
clustering	partitionnement automatique : séparation non-supervisée de données en plusieurs groupes (clusters)
courbe ROC	pour Receiver Operating Characteristic, courbe présentant les taux de vrai positif et de faux positif d'un modèle
data mining	domaine de l'extraction de connaissance à partir de grandes quantités de données
délai	Décalage temporel appliqué à une série temporelle faisant correspondre des valeurs passées à ses valeurs présentes
endogène	Variable dépendante (variable modélisée)
exogène	Variable indépendante (autre variable que celle modélisée)
krigeage	Méthode géostatistique d'interpolation spatiale
machine learning	domaine de l'apprentissage automatique
micro-échelle	Echelle spatiale locale (jusqu'à 10 km environ)
méso-échelle	Echelle spatiale se situant entre la micro-échelle et l'échelle synoptique (entre 10 km et 2000 km environ)
métaheuristique	Méthode d'optimisation adaptée aux problèmes d'optimisation difficiles, inspirées d'analogies physiques ou biologiques
parcimonie	Principe selon lequel l'optimisation des performances d'un modèle passe par l'utilisation du moins de paramètres possibles
pruning	Elagage (suppression automatique de paramètres superflus)
Qualitair Corse	AASQA de la région Corse
quartile	Trois valeurs permettant de diviser en quatre parts égales l'échantillon ordonné d'une variable aléatoire
recuit simulé	Métaheuristique inspirée du recuit, technique de chauffe en métallurgie permettant de réduire l'énergie d'un matériau

smog	Expression désignant les épisodes de pollution d'aspect brumeux, contraction de « smoke » et « fog » (fumée et brouillard en anglais)
synoptique	Echelle spatiale correspondant à des phénomènes comme les dépressions et anticyclones ($>$ 1000 km et jusqu'à l'échelle planétaire)
échéance	Date future pour laquelle une variable prévue par un modèle prévisionnel

Nomenclature

Dans l'ensemble du document, les vecteurs et matrices seront écris **en gras**.

$\bar{x}$	Valeur moyenne de x
$\hat{y}$	Variable en sortie de modèle, estimation ou prévision de y
$\mathbf{H}$	Matrice hessienne d'une fonction, regroupant ses dérivées partielles secondes
$\mathbf{I}$	Matrice identité
$\mathbf{J}$	Matrice jacobienne d'une fonction, regroupant ses dérivées partielles
∇	Opérateur Nabla, utilisé pour représenter une opération de gradient ou de divergence. $\nabla = \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial z}$
σ_x^2	Variance d'une variable x
d	Index of Agreement
y	Variable endogène, variable cible du modèle
GWh	Gigawatt-heure
hPa	hectopascal
MW	Mégawatt
R	Coefficient de corrélation de Pearson
U	Composante ouest-est du vent
V	Composante sud-nord du vent

Introduction générale

L'air que nous respirons est un bien extrêmement précieux et fragile. La pollution atmosphérique a des conséquences sur le climat, sur la couche stratosphérique d'ozone, mais aussi sur la qualité de l'air ambiant auquel nous sommes exposés en permanence. Les sources anthropiques de polluants dégradant cet air sont nombreuses, et les enjeux économiques qui s'y rattachent sont importants. Lors de pics de pollutions, épisodes au cours desquels les concentrations en polluants atteignent leurs plus fortes valeurs, l'impact sur la santé humaine et sur la biosphère est maximum.

En Corse, des pics de pollution ont régulièrement lieu. On observe des pics de pollution aux particules, souvent lors d'épisodes de transport de poussières du Sahara couplé à de fortes émissions locales et une météorologie défavorable à la dispersion du polluant. Des pics d'ozone sont également enregistrés, favorisés par les conditions estivales.

Il est cependant possible de combattre ces pics. Les émissions anthropiques sont très souvent impliquées lors de tels épisodes, et des mesures d'urgence peuvent être prises afin de les réduire à court terme, et d'éviter d'atteindre les plus hauts niveaux de polluants. L'impact de ce type de mesure (diminution de l'activité industrielle, réduction du trafic routier, etc.) est bénéfique pour la santé, et notamment pour celle des personnes les plus sensibles : les enfants, les personnes atteintes de troubles respiratoires, les personnes âgées.

La prise de ces mesures d'urgence doit être anticipée pour une meilleure efficacité. La nécessaire prévision des pics de pollution peut être réalisée à partir de plusieurs types de modèles. On peut construire, à partir de nos connaissances sur le fonctionnement atmosphérique, des modèles déterministes qui modélisent l'ensemble des phénomènes en lien avec l'évolution de la qualité de l'air. D'autre part, il est possible de construire des modèles statistiques capables d'apprendre à extrapoler les relations sous-jacentes existantes entre les variables qui décrivent l'état de l'atmosphère. Ces deux familles de modèles peuvent également être utilisées conjointement.

En Corse, la prévision de la qualité de l'air n'est pas actuellement satisfaisante et doit être améliorée. Elle est réalisée par Qualitair Corse, l'AASQA (Association Agréée de Surveillance de la Qualité de l'Air) de la Corse, agréée par le Ministère de l'Ecologie, du Développement Durable et de l'Energie (MEDDE). C'est ce besoin de disposer d'un outil adapté aux contraintes de Qualitair Corse, aux particularités géographiques et climatiques de l'île et aux nécessités de la prévision qui a conduit ces travaux de doctorat. Ils ont été rendus possibles par un partenariat fort entre l'association et l'Université de Corse, et par un financement conjoint de l'Agence de Développement Economique de la Corse (ADEC), de l'Agence d'Aménagement Durable, de Planification et d'Urbanisme de la Corse (AAUC) et de Qualitair Corse.

L'objectif de ce travail consiste donc à appréhender la problématique de la prévision de la qualité de l'air en Corse et d'étudier les méthodes prédictives aptes à prévoir les pics de pollution, particulièrement ceux dus aux particules et à l'ozone, avec la meilleure précision possible. A partir des modèles développés, un outil devra être fourni à Qualitair Corse, afin qu'il puisse être utilisé

quotidiennement par le personnel de l'association. Le modèle devra donc être parcimonieux en ressources informatiques.

Les modèles retenus dans ce travail sont basés sur les Réseaux de Neurones Artificiels (RNA). Ces modèles, inspirés à l'origine du fonctionnement des neurones du système nerveux central, font partie des méthodes de l'Intelligence Artificielle (IA) les plus utilisées. Leur apprentissage automatique permet des applications dans de nombreux domaines de la modélisation, et ils sont beaucoup utilisés en prévision de séries temporelles. Les données que nous utiliserons en entrée des RNA seront issues à la fois de mesures locales, et de résultats bruts du modèle déterministe CHIMERE et du modèle météorologique AROME de Météo-France. Le fait de pouvoir ainsi coupler les RNA avec d'autres modèles est un avantage qui permet d'améliorer les performances de prévision obtenues avec les différents modèles séparément. Le RNA devra adapter ses paramètres internes lors de son apprentissage automatique afin de représenter au mieux les relations sous-jacentes entre ces données et la variable à prévoir. Une fois paramétré, le modèle est prêt à être utilisé pour effectuer des prévisions.

Nous présenterons tout d'abord les principaux éléments qui gouvernent la qualité de l'air. Le devenir des polluants, leurs liens avec la météorologie et quelques informations sur les principaux polluants de la troposphère seront étudiés au premier chapitre.

Nous nous attellerons au second chapitre à identifier les différents modèles disponibles pour la prévision. Le fonctionnement des modèles déterministes et statistiques sera traité, et l'état de l'art précis de la prévision statistique de la qualité de l'air sera présenté. Ces informations nous conduiront à adopter un modèle de réseau de neurones, le Perceptron MultiCouche (PMC).

Au troisième chapitre, nous nous intéresserons à la qualité de l'air en Corse, afin d'identifier les conditions responsables des pics de pollution qui affectent l'île et de souligner les spécificités locales. Ce sera l'occasion de présenter les données qui seront utilisées par les modèles prévisionnels, données mesurées par Qualitair Corse, fournies par Météo-France ou provenant d'Air PACA, l'AASQA de la région PACA. Nous nous pencherons également sur les moyens actuels disponibles pour la prévision. La plate-forme de prévision AIRES utilisant le modèle CHIMERE et utilisée actuellement pour la prévision des concentrations en Corse sera présentée.

Au chapitre quatre, nous détaillerons le fonctionnement des PMC, modèles que nous avons choisis dans le cadre de cette thèse. Nous verrons comment les utiliser en tant que modèle prévisionnel. Leur architecture sera étudiée, les différents algorithmes d'apprentissage seront présentés ainsi que les différents prétraitements à appliquer aux données. La configuration des RNA est un problème complexe, dépendant de paramètres multiples. Plutôt que de rechercher un modèle prédictif universel, l'approche proposée consistera à appliquer une méthodologie permettant de trouver systématiquement la meilleure configuration de modèle, pour une problématique donnée. Le quatrième chapitre sera l'occasion de présenter cette méthodologie de construction de ces modèles.

Au chapitre cinq, nous nous intéresserons à plusieurs méthodes permettant d'améliorer les prévisions en se conformant au principe de parcimonie. Nous mènerons une étude sur la sélection de variables à l'aide de métaheuristiques (algorithmes génétiques, recuits simulés), et à l'aide d'Analyse en Composantes Principales (ACP). Après une étude complémentaire sur l'élagage des RNA, nous étudierons l'apport dû à l'utilisation des données issues de la plate-forme AIRES en tant que variables d'entrée des modèles neuronaux. Ces éléments seront illustrés par un exemple de construction de modèles prédictifs voués à la prévision opérationnelle. L'intérêt de notre méthodologie sera ensuite validé dans un contexte différent de la Corse, celui de la prévision d'ozone et de particules en région PACA (Provence-Alpes-Côte d'Azur). Les performances obtenues avec notre méthodologie seront comparées à celles du modèle de Forêt Aléatoire (FA) actuellement

en place à Air PACA.

Le chapitre six sera consacré à l'amélioration de cette prévision afin de renforcer la détection des pics de pollution, événements aux valeurs extrêmes et rares par nature, donc difficilement reconnus par les RNA qui sont plus aptes à reproduire des évènements fréquents et aux valeurs moyennes. Nous présenterons notre stratégie basée sur l'usage de modèles hybrides, combinant classifieurs et PMC, et mènerons une étude de sensibilité détaillée à l'aide de courbes ROC (Receiver Operating Characteristic) pour estimer l'amélioration de la détection des pics de pollution.

Enfin, nous présenterons au chapitre sept deux applications développées lors de ces travaux. La première, Aria Base, est une application de recherche développée sous Matlab qui permet de mener les expériences nécessaires à la mise en place du modèle prédictif. La seconde, Aria Web, est une application web d'aide à la décision développée sous Python, et destinée à être utilisée quotidiennement par les prévisionnistes de Qualitair Corse.

Nous conclurons ensuite ces travaux, fortement pluridisciplinaires, en présentant le bilan des travaux réalisés. On verra comment nous avons répondu à la problématique posée. Nous détaillerons également les perspectives qui ont émergé de ce travail de doctorat.

Chapitre 1

Pollution atmosphérique et qualité de l'air

La pollution atmosphérique est un problème majeur de notre temps. Depuis la révolution industrielle du XIX^{e} siècle, les émissions anthropiques de polluants atmosphériques ont fortement augmenté. Elles sont responsables de problèmes écologiques majeurs, qui sont désormais bien connus : le réchauffement climatique, le trou dans la couche d'ozone et la dégradation de la qualité de l'air que l'on respire.

Le réchauffement climatique d'origine humaine est un des problèmes environnementaux les plus connus du grand public. Ce phénomène est provoqué par l'accumulation de Gaz à Effet de Serre (GES) d'origine anthropique dans l'atmosphère. Naturellement présents dans l'atmosphère, les GES ont vu leurs niveaux fortement augmenter durant le siècle dernier. Les émissions de dioxyde de carbone (CO_2) et de méthane (CH_4) principalement, mais également d'autres composés organiques s'oxydant pour finir sous forme de CO_2, de protoxyde d'azote (N_2O) ou d'autres GES en sont responsables. L'effet de serre joue un rôle majeur sur le climat terrestre et son dérèglement a des conséquences considérables, que l'on commence à observer.

L'augmentation de la température moyenne provoque la fonte de glaciers et la montée du niveau de la mer, l'intensification des sécheresses et la progression des zones arides, la multiplication des phénomènes météorologiques extrêmes, etc. La figure 1.1 montre quelques observations allant dans ce sens. L'impact de tels changements climatiques sur nos sociétés est difficile à anticiper.

Le trou de la couche d'ozone (O_3) est un autre problème environnemental important, qui a été largement médiatisé. La couche d'ozone, naturellement présente au niveau de la stratosphère, protège la surface de rayonnement ultraviolet les plus énergétiques et les plus nocifs qu'elle absorbe. L'introduction dans l'atmosphère d'espèces chimiques stables telles que les CFC (Chlorofluorocarbure) va mener à des réactions catalytiques de destruction d'ozone dans la stratosphère, à l'origine du phénomène de déplétion d'ozone.

L'impact potentiel sur l'environnement de la destruction de l'ozone stratosphérique a mené à l'interdiction de l'utilisation de la plupart des espèces chimiques concernées avec le protocole de Montréal en 1987. Si les émissions des produits dangereux pour l'ozone stratosphérique ont fortement baissé depuis, la couche d'ozone reste aujourd'hui sous surveillance.

La dégradation de la qualité de l'air au niveau du sol est un autre problème, qui nous intéresse dans ces travaux de doctorat. Les polluants atmosphériques présents dans la troposphère ont des conséquences néfastes sur la biosphère. Ils sont très nombreux et ont des actions différentes

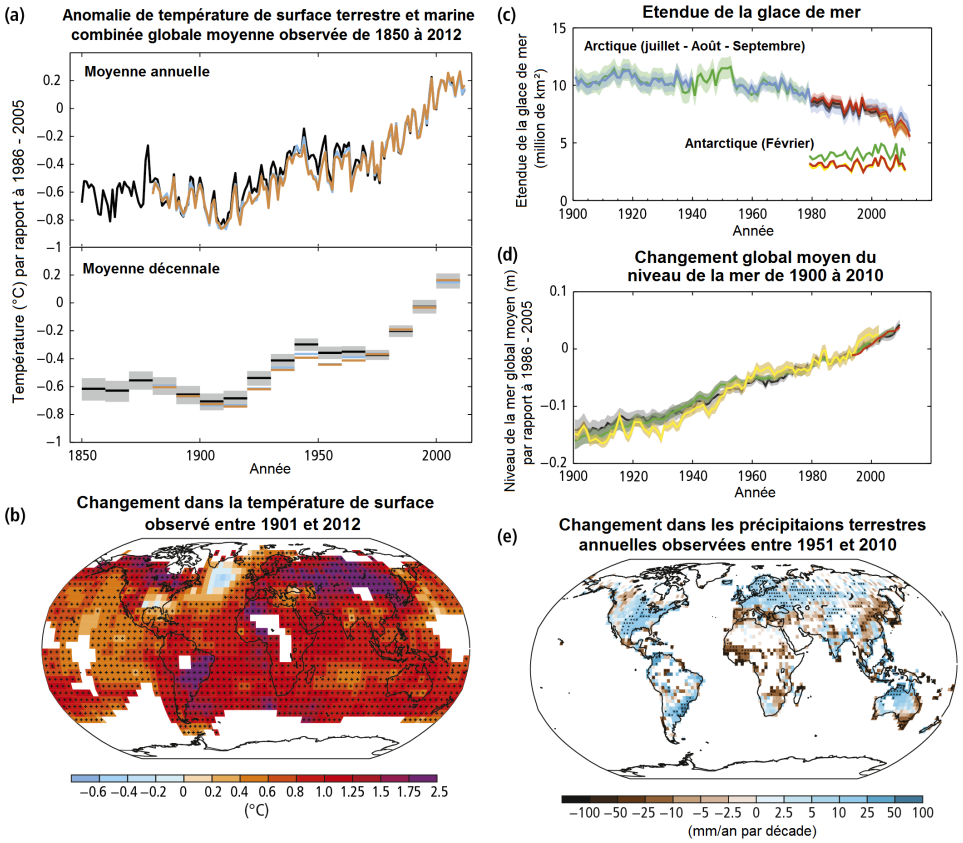


FIGURE 1.1 : Observations relatives au changement climatique illustrant les conséquences d'un réchauffement global (issues du rapport de l'ICPP, 2014).

sur les organismes vivants. La sensibilisation à la problématique de la qualité de l'air est récente, et les connaissances scientifiques sur la pollution atmosphérique le sont également.

Depuis l'invention de la machine à vapeur, la combustion de carburant a fortement augmenté. La qualité de l'air s'est dégradée dans les régions fortement industrialisées, et les phénomènes de smog (contraction de « smoke » et « fog ») sont devenus de plus en plus courants. Les émissions de gaz tels que le dioxyde de soufre (SO_2) ou les NO_x (oxydes d'azote, NO et NO_2) ont provoqué une acidification des gouttes d'eau des nuages et des particules menant à des pluies acides dégradant la végétation et le bâti.

Les épisodes exceptionnels de pollution de l'air des années 50 ont marqué les populations. En 1952, Londres a connu un smog meurtrier (photo en figure 1.2) qui provoquera le décès de milliers de personnes. On avance souvent un bilan de 4000 morts provoquées par ce smog, mais des études récentes montrent que ce nombre pourrait être bien plus élevé (Bell *et al.*, 2004). Cet événement, et bien d'autres épisodes similaires, seront à l'origine des efforts scientifiques pour mieux comprendre la pollution atmosphérique, ainsi que des efforts des gouvernements pour limiter cette pollution et instaurer une surveillance de la qualité de l'air. Des accidents industriels extrêmement graves comme la catastrophe de Bhopal en Inde en 1984 ou celle de la centrale nucléaire de Tchernobyl en 1986, ont également participé à la prise de conscience du fait que l'air que l'on respire est précieux et que sa qualité n'est pas inconditionnelle.

L'amélioration de la qualité de l'air est un défi pour nos sociétés, dont l'économie s'appuie largement sur des technologies polluantes. Les énergies renouvelables, proposées comme alternatives aux énergies fossiles fortement émettrices de polluants atmosphériques peinent encore à

FIGURE 1.2 : Photo de Londres pendant le grand smog de 1952.

trouver leur place dans le mix énergétique. Et certaines de ces énergies renouvelables, comme le bois-énergie, restent des sources de polluants. La volonté politique au niveau mondial d'une transition énergétique vers un modèle plus respectueux de l'environnement reste pour l'instant frileuse, et ce malgré le fait que la qualité de l'air prenne de plus en plus de place dans le débat public.

Désormais, la plupart des pays industrialisés ont adopté une politique plus ou moins ambitieuse vis-à-vis des problèmes posés par la pollution troposphérique. Ces politiques comprennent généralement un dispositif de surveillance de la qualité de l'air, qui en France est assurée par le réseau ATMO des AASQA (Associations Agréées de Surveillance de la Qualité de l'Air). Ce sont des associations régionales, qui assurent le suivi des concentrations en polluants, et le relayent auprès du public et des autorités. Les AASQA sont appuyées techniquement par le Laboratoire Central de Surveillance de la Qualité de l'Air (LCSQA), un groupement d'intérêt scientifique réunissant l'Institut National de l'EnviRonnement Industriel et des risqueS (INERIS), l'école des mines de Douai et le Laboratoire National de métrologie et d'Essais (LNE). Cet organisme permet de garantir la qualité de la surveillance et des informations produites par le réseau. C'est le Ministère de l'Ecologie, du Développement Durable et de l'Energie (MEDDE) qui donne leur agrément aux AASQA. Elles sont également chargées de mettre en place une prévision de la qualité de l'air, pour prévenir toutes conséquences négatives sur la santé. Qualitair Corse avec qui nous avons mené ces travaux est l'AASQA de la Corse.

Nous allons maintenant nous intéresser à l'atmosphère et à son fonctionnement, afin d'introduire quelques notions utiles à la compréhension de la problématique de la pollution atmosphérique. Nous verrons ensuite certains phénomènes météorologiques d'importance dans l'évolution des niveaux de polluants au niveau local. Nous nous focaliserons ensuite sur certains polluants de la troposphère qui nous intéressent particulièrement dans ces travaux ou qui ont une place importante dans les stratégies de surveillance de la qualité de l'air.

1.1 Introduction sur la qualité de l'air

L'atmosphère est composée, en volume, de 78.1% de diazote (N_2), de 20.9% de dioxygène (O_2) et de 0.93% d'argon (Ar), et d'autres gaz sous forme de traces. Elle est également constituée de vapeur d'eau (H_2O) dans des proportions variables allant de moins de 1% à 4%. Les gaz traces

les plus présents sont les espèces chimiquement stables, comme le CO_2 (400 ppmv), le CH_4 ou des gaz rares (néon, hélium, krypton). A cela s'ajoutent de très nombreuses espèces chimiques à de plus faibles concentrations, et à forte variabilité spatio-temporelle.

On parle de polluant pour désigner la part de ces éléments qui est indésirable, à cause des effets néfastes pour l'environnement et la santé humaine. Nous verrons quelles sont leurs origines, et leur devenir dans l'atmosphère.

1.1.1 Origines des polluants

Les polluants atmosphériques peuvent être des gaz ou des particules en suspension. On différencie en général les différents gaz par leur espèce chimique, mais les particules sont plus difficiles à classifier. On parle souvent d'aérosol, terme qui désigne le mélange formé par l'ensemble des particules et le gaz dans lequel elles sont en suspension. On parle aussi de PM (pour matière particulaire, Particulate Matter en anglais) ou simplement de particules.

Les polluants sont introduits dans l'atmosphère par leurs sources et en sont retirés par leurs puits. Certains de ces polluants, les polluants primaires, sont émis à la surface du globe (comme les particules dues à la combustion), et d'autres, les polluants secondaires, se forment par réaction chimique dans l'atmosphère (comme l'O_3). Les sources de polluants peuvent avoir une origine naturelle ou anthropique, et la part de chacune varie fortement en fonction du polluant concerné.

Les sources naturelles de polluants sont nombreuses, et peuvent concerner des polluants organiques ou inorganiques. Le volcanisme par exemple, est une source de composés soufrés, de CO_2 et de particules. Le vent provoque quant à lui la mise en suspension de poussières, d'aérosols marins, et même les éclairs sont des sources d'oxydes d'azote, apportant l'énergie nécessaire à la combinaison du diazote et du dioxygène en NO. Enfin, la végétation et la faune sont responsables de l'émission de nombreux gaz (Composés Organiques Volatils (COV), O_2, CO_2).

Les sources anthropiques sont également très diverses. La combustion de combustibles fossiles (industrie, transport, chauffage domestique, incinération, etc.) est responsable de l'émission de CO, CO_2, de COV, d'oxydes d'azote et de soufre, de métaux lourds, de particules. L'industrie émet en plus de nombreux produits spécifiques liés à chaque filière. La majeure partie des incendies peut être considérée comme d'origine humaine (Pechony et Shindell, 2010). Ce sont également d'importantes sources de particules, de COV, etc.

L'agriculture a une part importante dans la pollution anthropique. Une partie des pesticides et engrais utilisés par les agriculteurs se retrouve dans l'atmosphère. Les modes d'utilisation de ces produits jouent un rôle dans leur dispersion, leur épandage aérien par exemple est particulièrement néfaste à la qualité de l'air. Outre les émissions dues aux produits chimiques utilisés en agriculture, l'épandage de produits organiques comme le lisier ou les boues d'épuration implique également l'émanation de nombreux gaz. L'élevage est également un fort contributeur, même en ne comptant pas les émissions dues à la production agricole de la nourriture des animaux. La rumination des bovins d'élevage est responsable d'une importante part des émissions anthropiques de méthane, GES au pouvoir réchauffant très élevé.

Une fois émis par sa source, un polluant est transporté au sein de la masse d'air dans l'atmosphère. Ce transport se fait sous l'action du vent ou de dynamiques turbulentes. Le polluant finit par être retiré de l'atmosphère par son puits.

Chaque polluant a des puits plus ou moins importants. Il peut s'agir de dépôt sec, quand le polluant se dépose par contact, ou de dépôt humide, si le polluant est lessivé par la pluie. Le dépôt sec dépend des propriétés des surfaces sur lesquelles se déposent les polluants (la nature du

sol, de l'étendue d'eau, de la végétation, etc.). Le dépôt humide dépend de la présence de gouttes d'eau dans l'atmosphère (nuages), des propriétés de solubilité des polluants, des précipitations.

Le puits des espèces chimiques peut également correspondre à leur consommation lors d'une réaction chimique ou leur photolyse par le rayonnement solaire. Chaque espèce chimique a une durée de vie moyenne vis-à-vis de ces phénomènes, au bout de laquelle elle est détruite. Nous y reviendrons dans la suite de ce chapitre introductif. Tous ces processus sont très dépendants de la météorologie, qui joue un rôle majeur dans la disparition ou la stagnation des polluants.

1.1.2 Destin des polluants

Dans l'atmosphère, les réactions chimiques se font en phase gazeuse (ainsi qu'en phase aqueuse dans les gouttes d'eau des nuages). La chimie en phase gazeuse fait principalement intervenir des réactions radicalaires. Cette chimie est fortement dépendante du rayonnement solaire, qui apporte l'énergie nécessaire à la formation de radicaux libres (on parle de photochimie). Par exemple, l'ozone stratosphérique se forme grâce aux réactions photochimiques suivantes :

$$O_2 + h\nu \longrightarrow O\cdot \ + O\cdot \tag{1.1}$$

$$O_2 + O\cdot \ \longrightarrow O_3 \tag{1.2}$$

où $h\nu$ représente l'énergie apportée par le photon, ν étant sa fréquence et h la constante de Planck. $O\cdot$ est un radical libre. La réaction représentée à l'équation 1.1 décrit l'action du rayonnement, qui rompt la liaison covalente entre les deux atomes d'oxygène.

L'espèce chimique ainsi formée est appelée un radical libre, car elle possède un électron non apparié, habituellement représenté par un point (on note cependant que ce point est assez généralement omis par simplicité). Les radicaux ne respectent donc pas la règle de l'octet qui veut que les atomes soient entourés de huit électrons (contrairement aux ions qui se chargent électriquement en perdant ou captant le nombre d'électrons nécessaire au respect de cette règle). Cet électron non-apparié rend les radicaux libres très instables car extrêmement réactifs. Ces radicaux sont les réactifs principaux de l'atmosphère, et dirigent les réactions photochimiques. Le radical $OH\cdot$ est le plus important d'entre eux pour la chimie diurne. Les radicaux $HO_2\cdot$, $NO_3\cdot$, ainsi que l'ozone sont également très réactifs (bien plus que le dioxygène de l'air) et confèrent un caractère oxydant à la troposphère.

Le devenir des polluants atmosphériques dépend en grande partie de leurs réactions impliquant ces radicaux, ainsi que leur photolyse par le rayonnement solaire. La matière organique s'oxyde pour finir sous sa forme la plus oxydée, le CO_2. La lutte contre l'augmentation des concentrations en dioxyde de carbone ne peut donc pas concerner uniquement les émissions de CO_2, mais doit également passer par la gestion des émissions de ses précurseurs.

Les particules en suspension ont quant à elles des durées de vie qui dépendent de leurs propriétés physiques (masse, forme, etc.). Elles quittent l'atmosphère par dépôt sec ou humide, leurs propriétés de sédimentation ou leur solubilité sont donc importantes.

1.1.3 Transport des polluants

L'atmosphère se compose de plusieurs couches, qu'on peut identifier à partir du profil vertical de la température qu'on y observe (figure 1.3). Plusieurs phénomènes dirigent ce gradient. Il est tout d'abord dû au gradient de pression, lui-même imputable au poids de l'air. Le poids de ces

molécules d'air est compensé par la pression, menant à l'équilibre hydrostatique :

$$\frac{\partial P}{\partial z} = -\rho g \tag{1.3}$$

avec P la pression, z l'altitude, ρ la masse volumique de l'air et g l'accélération de la pesanteur. La pression est bien sûr modifiée localement suivant divers effets météorologiques. Cependant, ce gradient de pression permet de fixer une échelle d'altitude, qu'on peut alors exprimer en unité de pression (en hectopascal (hPa) le plus souvent) plutôt qu'en mètres. La loi des gaz parfaits combinée à l'équation 1.3 donne un gradient vertical négatif de la température T :

$$\frac{\partial T}{\partial z} = -\frac{g}{C_p} \tag{1.4}$$

avec C_p la capacité calorifique du gaz à pression constante.

La plus basse des couches est la troposphère, celle dans laquelle nous vivons. On y observe habituellement un gradient de température négatif. Au dessus d'elle se trouve la stratosphère, abritant la couche d'ozone. Le gradient de température y est positif, à cause du réchauffement provoqué par l'absorption du rayonnement solaire par l'ozone, dont la concentration est plus importante dans les hautes couches de la stratosphère. Au dessus on trouve la mésosphère, où la température diminue de nouveau avec l'altitude. Puis il y a la thermosphère où la température augmente, cette fois à cause de l'absorption de rayonnement par le dioxygène. L'exosphère se trouve au delà. Le gradient négatif de température de la troposphère permet la dispersion verticale des polluants. La stratosphère est beaucoup plus stable que la troposphère et les vents n'y soufflent quasiment pas verticalement.

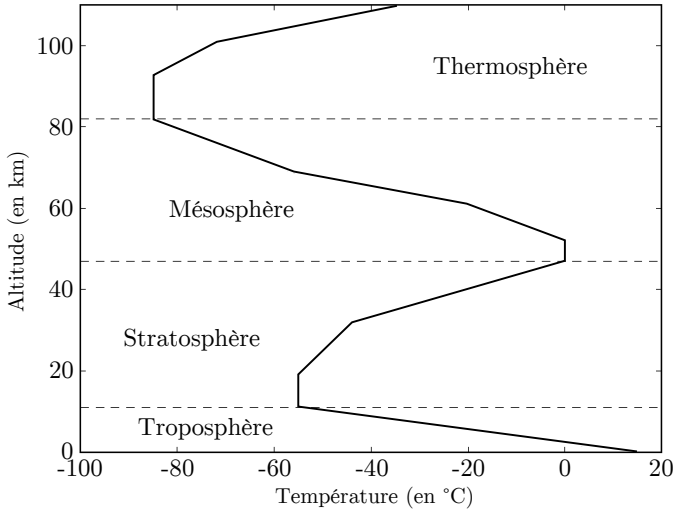


FIGURE 1.3 : Profil vertical de la température dans les différentes couches atmosphériques.

La troposphère dite « libre » est séparée du sol par une zone dotée d'un régime turbulent qu'on appelle la Couche Limite Atmosphérique (CLA). L'existence de la CLA est due à l'influence de la surface sur la dynamique des fluides de l'atmosphère. L'épaisseur de la troposphère est de l'ordre de la dizaine de kilomètres, celle de la CLA varie entre une centaine de mètres et quelques kilomètres.

Dans des conditions normales de stabilité, les polluants peuvent être transportés au sein de la couche limite sur des distances qui dépendent de leur durée de vie dans l'atmosphère. Les temps caractéristiques que mettent les polluants à passer d'une couche à l'autre sont donnés au tableau 1.1.

Transport	Temps caractéristique
Continental	1 semaine
Intercontinental	2 semaines
Hémisphérique	1 mois
Inter-hémisphérique	1 année
Couche limite atmosphérique	1 heure - 1 journée
Troposphère libre ($\simeq$ 5000 m)	1 semaine
Troposphère	1 mois
Échange troposphère vers stratosphère	de 5 à 10 ans
Échange stratosphère vers troposphère	de 1 à 2 ans

TABLEAU 1.1 : Temps caractéristiques du transport atmosphérique (issu de Sportisse, 2008).

Seules les espèces les plus stables peuvent donc être transportées jusqu'à la stratosphère. Pour cette raison d'ailleurs, les CFC ont été notamment remplacés par des espèces qui seraient tout aussi problématiques dans la couche d'ozone mais dont la durée de vie plus courte leur interdit de quitter la troposphère. Une masse d'air d'une semaine peut par contre être transportée à l'échelle continentale, ce qui permet une pollution longue distance pour les polluants ayant une durée de vie de cet ordre.

Si on s'intéresse à la qualité de l'air sur un site, elle dépend des sources de pollution locales, auxquelles s'ajoutent les polluants transportés et la pollution de fond. Les durées de vie de polluants, dont certaines sont indiquées en figure 1.4, doivent être prises en compte si l'on veut connaître leur origine potentielle. Les particules fines ou l'ozone ont par exemple des temps de vie leur permettant d'être transportés sur des milliers de kilomètres. Certains épisodes de pollution attribués à ces polluants peuvent donc avoir pour origine un épisode de transport.

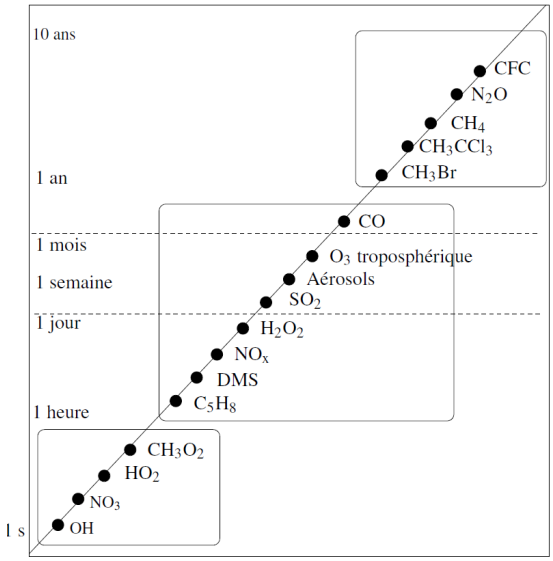


FIGURE 1.4 : Durées de vie et échelles à laquelle le transport est possible pour les principaux polluants (Bruno Sportisse).

1.2 Régimes météorologiques particuliers

Les émissions de polluants ne déterminent donc pas à elles seules l'état de l'air que nous respirons. La météorologie est, à ce propos, déterminante. Les conditions météorologiques peuvent favoriser la dispersion des polluants malgré une proximité avec des sources intenses, ou au

contraire causer une stagnation de l'air créant localement des épisodes de pollution alors que les émissions sont réduites. Lors d'épisodes de transport, des masses d'air âgées peuvent apporter de l'air pollué aux sources éloignées.

Les phénomènes à l'échelle synoptique (plusieurs milliers de kilomètres), comme ceux qui décrivent l'évolution des anticyclones et des dépressions, sont déterminants. Mais ce qui se passe à méso-échelle (entre la dizaine et le millier de kilomètres) ou micro-échelle (de l'ordre du kilomètre) l'est également, et plusieurs phénomènes de ces échelles ont la particularité d'être spécifique à leur localité. La dynamique locale des polluants peut être en partie déterminée par la configuration du site en question. Les mesures de polluants étant réalisées en station fixes, il est utile de connaitre les effets pouvant intervenir localement dans la concentration mesurée. Nous allons voir quelques phénomènes météorologiques locaux typiquement associés à certaines dynamiques de la qualité de l'air.

1.2.1 Brises thermiques

Les brises sont des régimes de vents associés à des différences spatiales de températures, qui sont en général dues à la configuration du terrain. Le principe d'une brise est le suivant : quand une zone au niveau du sol est plus froide qu'une zone voisine, une cellule de convection apparaît. L'air chaud s'élève, et provoque une aspiration de l'air plus froid, ce qui crée un courant d'air de la zone froide vers la zone chaude, la brise. Il existe plusieurs types de brises, en fonction du phénomène à l'origine du gradient de température.

La brise de mer prend place sur le littoral ou au bord des lacs suffisamment grands. L'eau a une inertie thermique plus élevée que le sol. En conséquence, quand le soleil se lève et chauffe la surface, le sol se réchauffe plus vite que l'eau. Il se crée alors une brise de mer, un vent qui souffle de la mer vers le sol. De nuit, le sol se refroidit plus vite que l'eau qui conserve mieux la chaleur. On observe une brise dans le sens inverse, un vent soufflant des terres vers le large qu'on appelle la brise de terre.

Le phénomène de brise de mer est problématique pour la qualité de l'air des villes situées sur le littoral, ce qui sera le cas des deux villes étudiées dans ce travail. En effet, les polluants émis par les sources urbaines ne se dispersent pas en direction du large, et sont maintenus sur la côte.

Dans les zones urbanisées peut se créer un phénomène d'îlot de chaleur urbain. Le béton accumule mieux la chaleur que le sol des zones rurales, et de plus les villes sont chauffées par les activités anthropiques, notamment le chauffage domestique. En ville, on mesure donc en général des températures plus élevées que celles mesurées en zones rurales, de quelques degrés. Cette différence est suffisante pour créer une brise de campagne, qui souffle depuis la périphérie vers le centre ville. Ce type de brise défavorise également la dispersion des polluants émis en zone urbaine.

Il existe des régimes de brises qui ne sont pas liés à des différences d'inertie thermique du sol mais à des différences d'ensoleillement pouvant intervenir dans des zones encaissées. De part sa trajectoire apparente dans le ciel, le soleil éclaire d'abord les cimes et le haut des montagnes ; le matin, une brise de pente peut souffler du bas vers le haut des montagnes. Une vallée entourée de pentes peut subir un appel d'air global, provoqué par l'ensemble des brises de pente, qui va induire une brise de vallée avec un vent soufflant depuis le bas de vallée vers le haut. Au cour de la journée, des brises vont naître sur les versants ensoleillés, et se dissiper lorsqu'ils se retrouvent à l'ombre. La nuit, l'air froid glisse des pentes vers les fonds de vallée.

Les masses d'air portées par le vent peuvent également être amenées à s'élever à cause

de l'obstacle que présente la montagne, sans intervention de brise. L'air qui monte peut être amené à condenser, car avec l'altitude la pression diminue, et ainsi l'air peut atteindre son point de rosée. Arrivé en crête, l'air s'est réchauffé et a perdu de l'humidité. Une fois la crête dépassée, l'air redescend de l'autre côté du relief, parfois par un fort vent chaud et sec. L'effet de Foehn bien connu dans certaines régions montagneuse suis cette dynamique. En Corse, ce type de phénomène explique la présence de certains nuages qu'on observe accrochés aux parois montagneuses.

Ces phénomènes influent largement sur la météorologie des zones montagneuses. Ils ont aussi un impact sur la qualité de l'air, participant au brassage de l'air.

1.2.2 Stabilité de la couche limite

Pendant la journée, le réchauffement du sol est à l'origine de turbulences qui forment une couche où prennent place des échanges verticaux, la Couche Limite Atmosphérique (CLA). Une couche limite très turbulente sera instable, ce qui favorisera la dispersion des polluants. La turbulence peut être due à des effets thermiques comme l'échauffement du sol ou la condensation des nuages, ou cinétiques (cisaillement de vents contraires, obstacles physiques au vent, etc.).

Les turbulences liées à l'ensoleillement de la surface établissent la CLA le matin et pendant la journée. Au coucher du soleil, ces turbulences diminuent avec le rayonnement du sol. Ce dernier se refroidit rapidement et une inversion thermique apparait : on observe un gradient positif de la température.

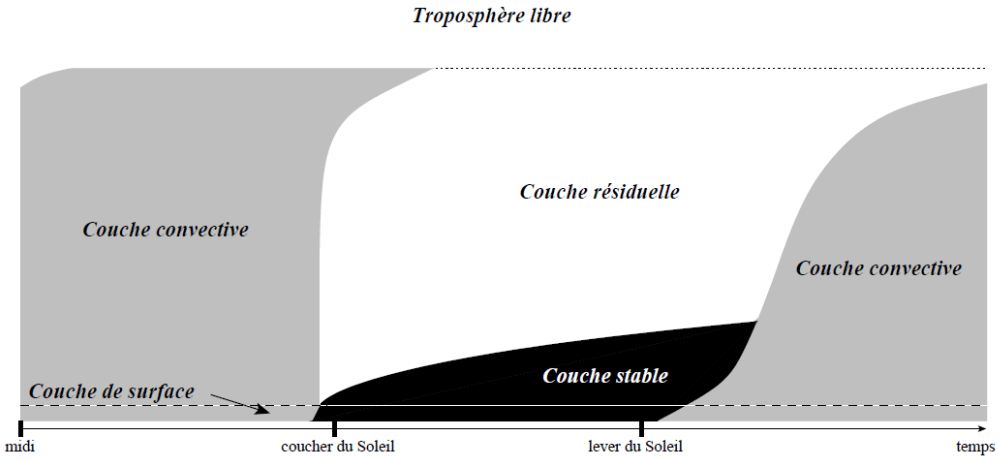


FIGURE 1.5 : Profil journalier de la couche limite atmosphérique (issu de Stull, 1988).

Un tel profil de température provoque une stabilité d'atmosphère. L'air plus chaud en altitude étant plus léger, il empêche l'air plus froid de monter. Il y a donc formation d'une couche stable, au dessus de laquelle se trouve le résidu de la couche limite diurne. Cette dernière n'est plus turbulente de nuit, on l'appelle la couche résiduelle. Au lever du jour, une nouvelle couche limite se met en place (figure 1.5).

Quand les polluants sont émis dans la couche limite, ils y sont transportés horizontalement par advection (transport par un champ vectoriel, comme le vent), et verticalement par turbulence. Les gaz à température relativement élevée (émis par combustion, par exemple) vont former des panaches qui s'élèvent par convection. En cas d'inversion thermique, quand le gradient de température est positif sur une certaine distance, les panaches sont bloqués. C'est pourquoi les inversions thermiques provoquent la stagnation des polluants. Les inversions thermiques diurnes

peuvent survenir et sont l'une des situations les plus favorables aux pics de pollution. Outre les cas d'inversions thermiques, le gradient thermique de la CLA peut conduire à des régimes plus ou moins stables. La Hauteur de la Couche Limite (HCL) peut être un bon indice de l'intensité des turbulences.

En terrain montagneux, les régimes de brise peuvent perturber la stabilité nocturne de la CLA et contribuer à son brassage. En crête ou en sommet de montagne, la situation topographique ne permet pas l'établissement d'une telle couche. C'est le contraire en plaine, où l'évolution temporelle de la CLA se conforme le plus au profil typique de la figure 1.5.

Maintenant que nous avons vu les phénomènes qui peuvent localement influencer leurs concentrations, nous allons nous pencher sur les polluants qui nous intéressent, en précisant leurs sources et leurs puits, leur dynamique et les effets qu'ils ont sur la santé et l'environnement.

1.3 Principaux polluants

Nombreuses sont les espèces chimiques dangereuses pour l'environnement et la santé présentes dans la troposphère. Afin de surveiller efficacement la qualité de l'air, certains polluants sont de meilleurs candidats pour être suivis de manière automatique. Il est nécessaire de se focaliser sur ceux qui sont à la fois représentatifs de l'état général de l'air et mesurables automatiquement à une fréquence satisfaisante et à des coûts envisageables. De plus, une espèce elle-même connue pour ses effets néfastes fera une meilleure candidate.

Quatre polluants sont ainsi au centre de la surveillance, les particules en suspension, l'ozone, le dioxyde d'azote et le dioxyde de soufre. En France, l'Indice de la Qualité de l'Air (IQA) est calculé à partir de ces quatre polluants afin de synthétiser auprès du grand public l'information sur la qualité de l'air en temps réel. Nous allons passer ces polluants en revue et en décrire la dynamique et les effets. Nous verrons ensuite rapidement quels autres polluants font désormais également partie du dispositif de surveillance.

1.3.1 Particules en suspension

Les particules en suspension peuvent être liquides, solides ou mixtes avec un noyau solide entouré de liquide. Elles sont soit constituées d'un seul élément, soit regroupent plusieurs espèces chimiques. Elles peuvent être issues de sources primaires ou secondaires. Elles sont naturellement présentes dans l'atmosphère, et les activités humaines augmentent leurs niveaux.

Il en existe une grande diversité. Dans les particules primaires, il y a des particules minérales soulevées par les vents, comme les poussières ou les sels marins. Les combustions incomplètes produisent des particules de suie, composée de carbone élémentaire (souvent dénommé « black carbon ») et de carbone organique. Les cendres issues de combustion ou d'activités volcaniques forment également des particules. Une photographie de retombée d'un panache de ferry chargé en particules est présentée en figure 1.6. La biosphère émet aussi des particules, comme les pollens.

Les particules secondaires se forment quand à elles dans l'atmosphère, d'abord par nucléation, regroupement de molécules de gaz (COV, H_2SO_4, NH_3, etc.). Les particules ainsi formées sont les plus petites. Elles s'agrègent par coagulation et grossissent. Les gaz précurseurs d'aérosols secondaires sont nombreux, émis par la végétation comme par les activités humaines.

Les aérosols sont ainsi composés d'une large variété de particules. Leur forme diffère gran-

FIGURE 1.6 : Panache d'un ferry chargé en particules retombant sur Bastia (Crédit photo : Qualitair Corse).

dement (particules sphériques, en forme de filament, etc.), et leur propriétés chimiques (et leur toxicité) varient. Il est difficile d'analyser en détail un aérosol, on regroupe donc souvent les particules en fonction de leur propriétés aérodynamiques. Le diamètre aérodynamique d'une particule est le diamètre qu'aurait une particule sphérique de même vitesse de sédimentation. Cela permet d'utiliser une méthode de mesure des particules basée sur ce type de propriété.

Le terme PM10 par exemple désigne les particules au diamètre aérodynamique de moins de 10 µm. Ce sont les PM10 dont la mesure automatique s'est développée en premier, avant que l'on suive également les particules de moins de 2,5 µm de diamètre (PM2.5), puis celles de moins de 1 µm de diamètre (PM1). On note que le terme « particules fines », souvent employé pour désigner l'ensemble des particules en suspension, est utilisé par la communauté scientifique pour désigner spécifiquement les PM2.5. De même, on parle de particules grossières pour désigner les PM10 de plus de 2,5 µm de diamètre. Les processus gouvernant l'évolution des particules sont présentés en figure 1.7.

Les particules jouent un rôle important dans le fonctionnement atmosphérique. En effet, ce sont des particules qui font office de noyaux de condensation autour desquels se forment les gouttes d'eau des nuages. Dans un premier temps, une faible pollution favorise donc les temps couverts et pluvieux. Mais dans les atmosphères très chargées en particules, comme en milieu urbain pollué, le grand nombre de noyaux de condensation provoque un plus grand nombre de gouttelettes pour la même quantité d'eau. Ces gouttes seront donc plus fines et précipiteront moins, la pluie sera plus rare pour la même quantité d'eau.

Les particules ont également un impact direct sur le bilan radiatif terrestre. Leur grande taille (par rapport aux molécules de gaz) leur donne des propriétés optiques différentes de celles des molécules d'air. Les diamètres des molécules provoquent une diffusion du rayonnement incident par diffusion de Rayleigh, phénomène à l'origine de la couleur bleue du ciel. Les particules diffusent la lumière par diffusion de Mie, qui donne par exemple leur couleur blanche aux nuages. Le forçage radiatif induit par les particules leur donne un rôle important dans le changement climatique.

Les particules ont un effet néfaste sur la santé humaine. Leur petite taille leur permet de pénétrer profondément les voies respiratoires, provoquant des inflammations et l'aggravation

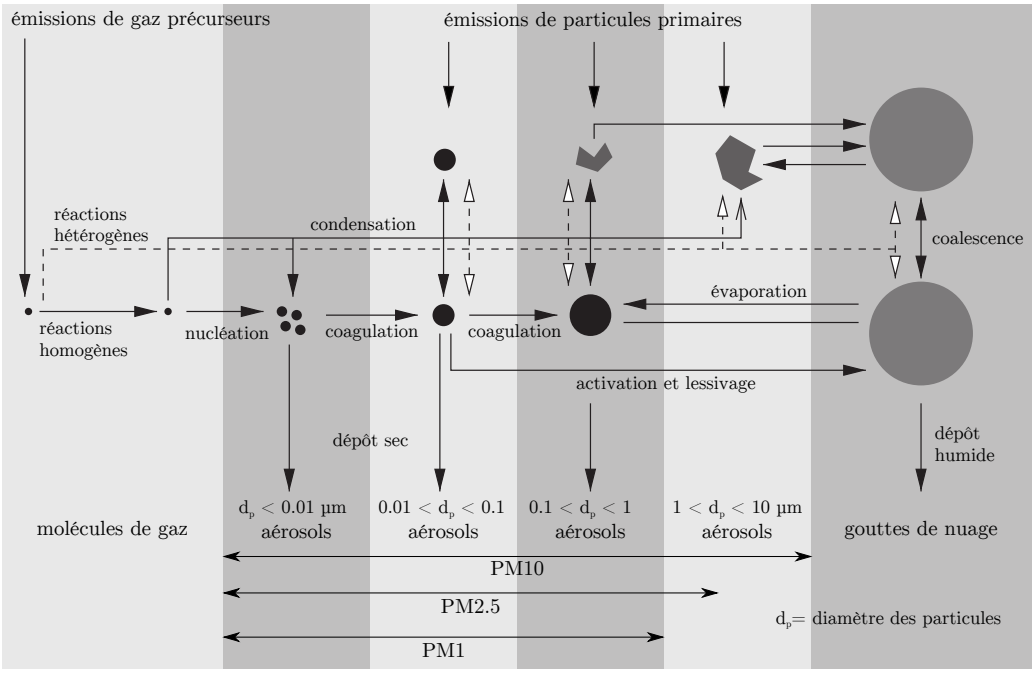


FIGURE 1.7 : Processus microphysiques des aérosols (issu de Raes *et al.*, 2000).

d'éventuels problèmes respiratoires ou cardiaques (Kappos *et al.*, 2004). Etant donnée l'hétérogénéité de la composition des particules, leur toxicité est difficile à envisager dans l'absolu. Elle dépendra des composés chimiques qui les composent, et de leur taille, leur permettant d'atteindre différents organes du corps (Englert, 2004). Pour cette raison, les particules plus fines sont considérées comme plus dangereuses que les particules grossières. L'inhalation de particules augmente également les probabilités de développer un cancer du poumon (Knaapen *et al.*, 2004). Les particules en suspension pourraient ainsi provoquer le décès de 42000 personnes par an en France (CBA, 2005).

1.3.2 Ozone

L'ozone est une molécule contenant trois atomes d'oxygène (O_3). C'est un polluant secondaire, sa formation dépend d'un cycle photochimique complexe. La couche d'ozone présente dans la stratosphère nous protège du rayonnement ultra-violet le plus énergétique du soleil. La formation de l'ozone stratosphérique est propre à cette couche de l'atmosphère (équation 1.1) et diffère du cycle troposphérique. Par opposition au rôle protecteur de l'ozone stratosphérique, l'ozone présent dans la troposphère, qui est néfaste pour l'environnement et pour l'humain, est parfois qualifié de « mauvais ozone ». Bien que ce ne soit pas une espèce radicalaire, c'est l'un des oxydants majeurs de la troposphère (avec le radical $OH\cdot$, NO_2, et le radical $NO_3\cdot$ pour la chimie nocturne). Il participe par exemple à la formation d'aérosols secondaires, par l'oxydation de COV comme l'isoprène émis par la végétation. De plus, l'O_3 est un GES.

Le cycle troposphérique de formation d'ozone fait intervenir un grand nombre d'espèces chimiques. La représentation simplifiée de ce cycle est présentée en figure 1.8. La photolyse de NO_2 libère un radical oxygène qui réagit avec le dioxygène de l'air et assure la formation nette d'ozone. Le cycle catalytique des NO_x dépend lui-même d'un cycle faisant intervenir de nombreuses espèces. Ce cycle est initié par l'introduction de radicaux $OH\cdot$ et $HO_2\cdot$, par des réactions de photolyse (entourées en vert sur la figure). L'oxydation par $OH\cdot$ de COV donne des radicaux $RO_2\cdot$ qui permettent l'oxydation de NO en NO_2, oxydation également assurée par le

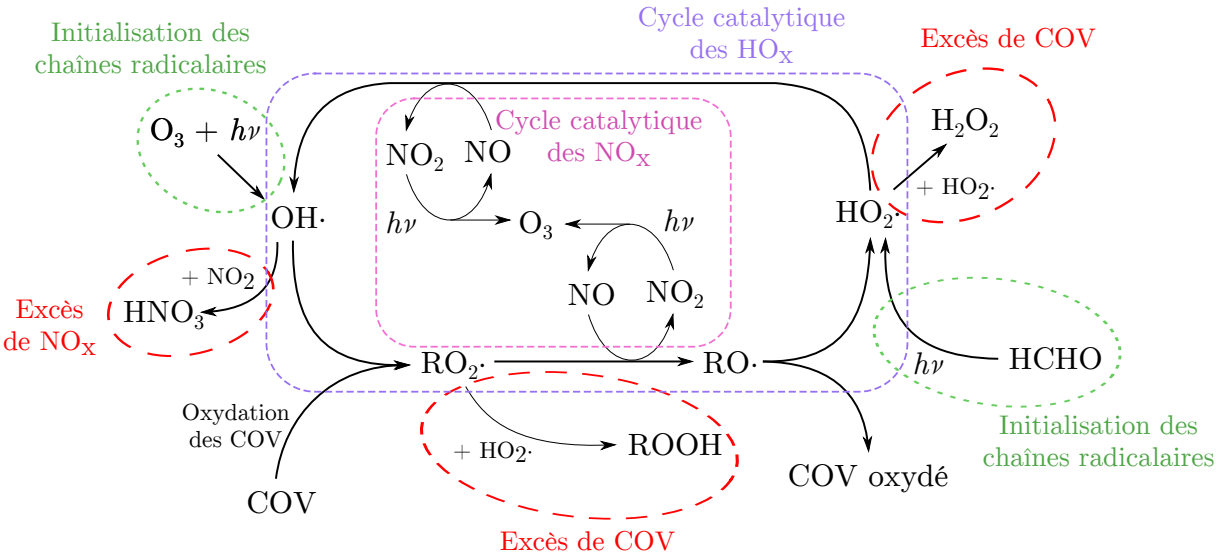


FIGURE 1.8 : Schéma simplifié du cycle troposphérique de formation d'ozone.

radical HO_2·. La formation d'ozone dépend du maintien de ce cycle des HO_x (les radicaux OH·, HO_2· et peroxyles organiques RO_2·), qui repose sur l'équilibre entre les réactions d'initiation et les réactions de terminaisons.

Les réactions de terminaison de ce cycle sont celles qui éliminent les radicaux. Certaines font intervenir des NO_x, d'autres des radicaux HO_2· dont la présence dépend des niveaux de COV. Si l'on essaie de simplifier ce cycle, on peut dire qu'un équilibre entre les niveaux de NO_x et de COV est nécessaire. En cas d'excès relatif de NO_x ou de COV, la formation d'ozone diminue. En plus de sa réaction de terminaison menant à la formation d'acide nitrique, les NO_x peuvent directement détruire l'ozone par la réaction de titration :

$$O_3 + NO \longrightarrow O_2 + NO_2 \tag{1.5}$$

A proximité des sources de NO_x, comme par exemple au niveau d'axes routiers importants, les niveaux d'O_3 sont donc très bas. Pour ces raisons, les concentrations les plus fortes en ozone se trouvent en général en milieu périurbain ou rural, plus loin des sources de NO_x et plus proches de celles de COV (végétation), là où le rapport des concentrations COV/NO_x est optimal.

La formation d'O_3 a lieu le jour, par la réaction suivante :

$$O_2 + NO_2 + h\nu \longrightarrow NO + O_3 \tag{1.6}$$

La nuit, l'ozone proche du sol est piégé dans la couche limite basse due à l'inversion nocturne. Les concentrations diminuent, principalement sous l'effet du dépôt sec. Dans la couche résiduelle cependant, les niveaux d'ozones restent élevés (Académie des sciences, 1993). La production diurne d'O_3 et son dépôt nocturne lui confère un profil journalier de concentration en cloche. De par la forte dépendance au rayonnement de sa formation, les concentrations d'ozone sont plus élevées en été qu'en hiver. En terrain montagneux, le léger brassage nocturne limite la diminution des concentrations d'O_3 la nuit. En crête, on peut parfois observer des profils journaliers avec des concentrations constantes le jour et la nuit. Quand on a une couche nocturne stable, il est possible d'observer un pic matinal d'ozone, dû au mélange avec la couche résiduelle que provoque l'établissement de la couche limite.

L'ozone à un effet néfaste sur la végétation. Oxydant puissant, il réagit avec les composants de la surface des cellules végétales, provoquant des dégâts souvent visibles sur les feuilles où

FIGURE 1.9 : Effet de l'ozone sur l'armoise (base de données des dommages de l'ozone sur les plantes).

des nécroses peuvent apparaitre (comme le montre la photographie en figure 1.9). L'activité de photosynthèse est réduite, et les plantes vont avoir besoin de consommer leur sucre (par respiration) pour réparer les tissus détruits par l'ozone. De plus la sénescence des feuilles est accélérée. Certaines espèces, comme le trèfle ou le tabac, sont particulièrement sensibles à l'ozone, même en faible concentration. Elles peuvent être utilisées en biosurveillance, leur état indiquant la présence d'ozone. L'impact écologique de l'ozone se traduit en impact économique, via la diminution induite des rendements en agriculture.

Chez l'humain, l'ozone peut pénétrer les voies respiratoires les plus fines, provoquant toux et crises d'asthme. Il a également été observé que l'exposition à de fortes concentrations d'ozone provoquait une hausse de la mortalité et des hospitalisations. On estime le nombre de morts prématurés dus à l'ozone à près de 3000 par an en France (CBA, 2005).

Le caractère fortement oxydant de l'O_3 et son rôle dans la photochimie troposphérique est également problématique, puisqu'elle participe à la formation de polluants secondaires, comme des particules fines. En surveillance de la qualité de l'air, le suivi des niveaux d'O_3 permet à la fois de surveiller un polluant dangereux et d'avoir une information sur l'état du réacteur photochimique qu'est la troposphère.

1.3.3 Oxydes d'azote

Les NO_x (oxydes d'azote, NO et NO_2) sont des polluants émis principalement lors de combustions à haute température. NO et NO_2 sont également de puissants oxydants. Bien qu'ils soient rarement écrits avec le point symbolisant l'électron non-apparié des espèce radicalaires, tous deux sont des radicaux libres. Leur durée de vie est tout de même plus longue que la plupart des radicaux. En réalité, les combustions produisent principalement du NO. Celui-ci est rapidement oxydé en NO_2. Le rapport NO/NO_2 peut ainsi permettre d'estimer l'âge de la masse d'air dans laquelle il est mesuré.

Les sources de NO_x sont variables (voir tableau 1.2). La combustion peut libérer de l'azote

Emissions dans l'atmosphère (en Tg)		
Sources anthropiques	**NO_x**	**NH_3**
Combustion de carburants fossiles et industrie	28.3	0.5
Agriculture	3.7	30.4
Combustion de biomasse et de biocarburants	5.5	9.2
Total anthropique	**37.5**	**40.1**
Sources naturelles		
Sols sous la végétation naturelle	7.3 (5–8)	2.4 (1–10)
Océans	—	8.2 (3.6)
Eclairs	4 (3–5)	—
Total naturel	**11.3**	**10.6**
Sources totales	**48.8**	**50.7**

TABLEAU 1.2 : Sources naturelles et anthropiques de NO_x et de NH_3 (issu du rapport de l'ICPP, 2014).

présent dans le combustible, mais elle apporte surtout l'énergie nécessaire pour briser le lien covalent du diazote de l'air, qui s'oxyde pour donner NO. La foudre a le même effet, et est responsable de la formation d'une quantité de NO non-négligeable, injectant des NO_x nécessaires à la formation d'O_3 parfois loin des sources fixes, là où la formation est limitée par les NO_x. Des NO_x sont également émis par les sols et par les pratiques agricoles.

Avec des temps de vie de l'ordre de la journée, on trouve principalement des NO_x à proximité de leurs sources. L'un des principaux puits de NO_2 vient de son oxydation par $OH\cdot$:

$$NO_2 + OH\cdot \longrightarrow HNO_3 \tag{1.7}$$

donnant de l'acide nitrique stable. Ce dernier contribue au phénomène de pluies acides. Le dépôt de NO_2 induit une acidification des milieux et participe à l'eutrophisation des sols. Les NO_x sont également toxiques pour l'être humain, particulièrement NO_2. Il irrite les bronches et provoque des troubles respiratoires voire des œdèmes pulmonaires.

Les NO_x sont surveillés, souvent via le NO_2 qui est le plus toxique. Ce polluant est un traceur des combustions, indicateur de l'intensité des émissions primaires telles que le trafic routier ou les rejets industriels. Ces niveaux sont souvent corrélés à ceux des PM10 également émis par combustion lors de pics de trafic ou d'intenses émissions industrielles.

1.3.4 Dioxyde de soufre

Le dioxyde de soufre (SO_2) est un gaz principalement émis par les combustions utilisant des combustibles soufrés (charbon, coke, fuel, gasoil, etc.). Certaines industries produisent également du SO_2 (chimie, papier, raffinage de pétrole, etc.). Il existe également des sources naturelles de dioxyde de soufre, comme le volcanisme par exemple, moins importantes que les sources anthropiques. C'est un polluant emblématique de l'ère industrielle, qui a vu ses niveaux baisser de manière importante en France ces dernières décennies (voir figure 1.10), grâce à des réglementations allant en ce sens.

Ce gaz est très soluble. Il se solubilise dans les gouttes d'eau des nuages et est responsable des pluies acides, avec les NO_x et NH_3. Chez l'humain, le SO_2 affecte l'appareil respiratoire, aggrave l'asthme et provoque également des irritations oculaires. A de fortes concentrations,

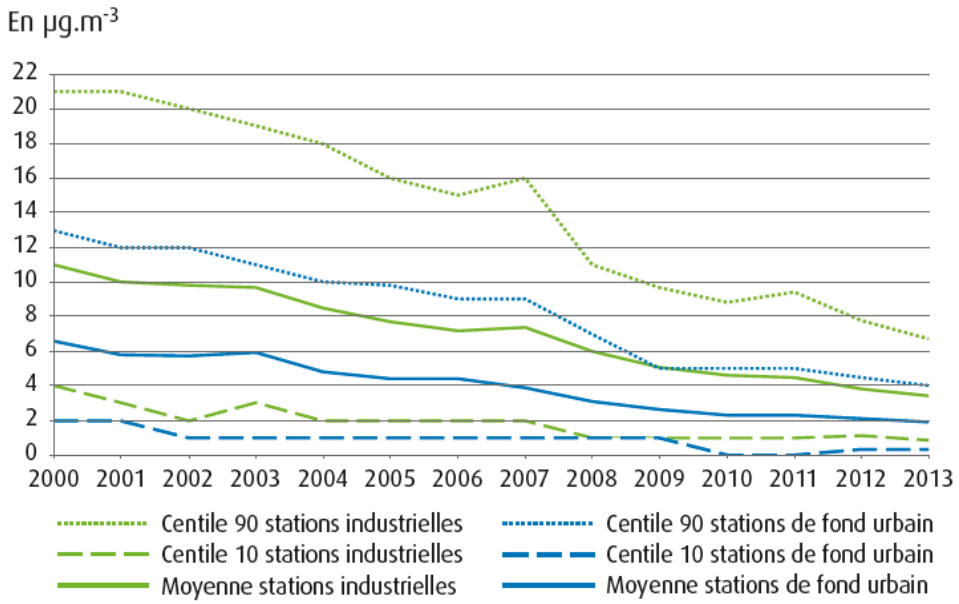


FIGURE 1.10 : Evolution des concentrations annuelles de SO_2 en France entre 2000 et 2013 (source : MEDDE, Bilan de la qualité de l'air en France en 2013).

il peut provoquer le décès de personnes atteintes de pathologies cardiovasculaires. Il est par exemple le principal responsable des décès liés au grand smog de Londres de 1952.

Le SO_2 sert de traceur d'activités industrielles. Il fait toujours partie à ce titre des polluants les plus surveillés, bien que ces niveaux ne soient plus problématiques en France. Il reste malgré tout un véritable problème dans les pays ayant massivement recours au charbon pour la production énergétique, comme la Chine par exemple.

1.3.5 Autres polluants

Parmi les autres polluants majeurs, on trouve la famille des COV. Elle regroupe tous les composés organiques (composés avec au moins un atome de carbone, à l'exception des oxydes de carbone CO et CO_2) qui ont une pression de vapeur saturante suffisante pour se volatiliser dans l'air. Le méthane, puissant GES particulièrement stable, est généralement mis à l'écart de cette catégorie (on parle alors de COVNM, Composés Organiques Volatiles Non Méthaniques). Les COV forment donc un ensemble très vaste de composés, aux sources, aux puits, aux effets sur l'environnement et sur la santé variables.

La majeure partie des COV est d'origine naturelle. La végétation en émet un grand nombre, dont les plus courants sont l'isoprène, les terpènes, etc. Les zones géologiques peuvent émettre des hydrocarbures. Le méthane (souvent exclu des COV) est lui émis par les animaux et les zones humides. Les sources anthropiques de COV sont principalement industrielles, et particulièrement liées à l'usage de solvants ou aux filières liées aux hydrocarbures.

Leurs effets dépendent de leur nature. On l'a vu, les COV jouent un rôle crucial dans le cycle de formation troposphérique de l'O_3. Leur oxydation mène également à la formation de particules secondaires. Certains d'entre eux sont particulièrement dangereux pour la santé, comme par exemple le benzène. Cet hydrocarbure de formule brute C_6H_6 forme un cycle aromatique. C'est une espèce cancérigène et génotoxique principalement émis lors de combustions incomplètes, mais aussi par le volcanisme. La famille des Hydrocarbures Aromatiques Polycycliques (HAP) est également devenue prioritaire dans les stratégies de surveillance. La plupart de ses représentants

sont toxiques, et un grand nombre d'entre eux est également mutagène ou cancérigène. Leurs sources sont sensiblement les mêmes que celles du benzène.

Le monoxyde de carbone est également un polluant produit lors de combustions incomplètes. Il est surtout connu pour ses effets néfastes en qualité de l'air intérieur, pour les accidents domestiques mortels qu'il peut provoquer notamment lors de mauvaises utilisations de systèmes de chauffage défaillants ou mal entretenus. Il est toxique car il se fixe sur l'hémoglobine en substitution du dioxygène, ce qui entraine de nombreux symptômes et peut mener au décès. On le trouve également dans l'atmosphère, à proximité du trafic, et ses niveaux sont désormais surveillés.

Parmi les polluants nouvellement surveillés, on note également la présence des métaux lourds. Ce terme est en réalité utilisé pour désigner les éléments traces, qu'ils soient considérés comme lourds ou non, et métalliques ou non. Les plus surveillés sont le plomb (Pb), l'arsenic (As), le cadmium (Cd) et le nickel (Ni). Les principales sources de métaux lourds sont industrielles (métallurgie) et liées à l'usage de combustibles dans lesquels ils sont présents. Avant les années 2000, le plomb était présent dans les carburants automobiles mais en est désormais banni et remplacé par des substituts.

Les métaux lourds s'accumulent dans l'organisme, et ont des effets néfastes à court et long terme. La plupart d'entre eux (l'arsenic, le cadmium et le nickel) sont cancérigènes. Le plomb provoque saturnisme et troubles neurologiques, et sa toxicité est doublée de forts soupçons sur son potentiel cancérigène (Hervé-Bazin, 2004).

1.4 Conclusion

De nombreux polluants, d'origine naturelle ou anthropique, contribuent à la dégradation de la qualité de l'air que nous respirons. La plupart sont liés à des processus de combustion, ainsi qu'à l'usage de produits chimiques dans l'industrie ou l'agriculture. Leurs niveaux peuvent conduire à des smogs dont les conséquences sont néfastes pour la santé humaine et pour l'environnement. Parmi eux, l'ozone, les particules en suspension, les oxydes d'azote se distinguent et sont surveillés depuis les dernières décennies. Les progrès techniques permettent désormais le suivis de « nouveaux polluants » : les métaux lourds, le benzène, les HAP, les particules ultrafines. Cependant, les PM10 et les NO_x sont de bons traceurs d'activités émettrices de polluants, l'ozone un bon indicateur de l'activité photochimique.

Le devenir des polluants dans l'atmosphère dépend de nombreux paramètres. Leur durée de vie dépend de leur réactivité vis-à-vis des oxydants de l'atmosphère et de leur photolyse, mais aussi de leur solubilité, de leur capacité de dispersion, de sédimentation. Ils peuvent être advectés par le vent, lessivés par l'eau des nuages. Tous ces phénomènes sont étroitement liés à la météorologie. Des régimes locaux de vents et de stabilité de couche limite sont importants dans l'évolution de la qualité de l'air.

Les mécanismes qui gouvernent le devenir des polluants permettent de modéliser l'évolution de leurs concentrations. Ces mécanismes peuvent être eux-mêmes modélisés (modèle déterministe), ou un modèle statistique peut les identifier et reproduire leurs effets grâce à l'analyse de données de pollution et de données météorologiques. Nous verrons dans le prochain chapitre comment à partir de ces connaissances sont construits les différents modèles permettant de prévoir l'évolution des concentrations de ces polluants. Ces prévisions permettent d'anticiper les pics de pollution et de mobiliser des moyens d'urgence pour éviter d'exposer les populations à un air qui soit néfaste pour leur santé.

Chapitre 2

La prévision de la qualité de l'air

La qualité de l'air est liée à l'émission naturelle ou anthropique de composés dans l'atmosphère, et à l'évolution de ces composés au gré des conditions météorologiques et de leur réactivité. Mais comment suivre cette évolution ? Nous allons voir dans ce chapitre que deux paradigmes existent.

Les modèles déterministes calculent cette évolution, sur la base de nos connaissances des phénomènes impliqués. Les modèles statistiques eux apprennent à utiliser les relations qu'ils identifient entre les différentes variables qui décrivent ces phénomènes. Ces deux familles de modèles ont un fonctionnement différent, que l'on abordera dans ce chapitre.

Nous verrons pourquoi nous nous sommes penchés sur les modèles statistiques, et plus particulièrement sur les réseaux de neurones artificiels, pour élaborer un modèle prévisionnel pour la Corse. Ce type de modèle demande peu de ressources informatiques et est facilement utilisable par une petite structure comme Qualitair Corse, l'AASQA de l'île. Ils offrent des performances intéressantes, et peuvent être utilisés conjointement avec d'autres modèles.

Nous discuterons ici de la manière dont on évalue les résultats de ces modèles. Puis nous ferons l'état de l'art de la prévision neuronale de la qualité de l'air depuis les années 2000 (voir notamment le tableau 2.1 page 47), afin d'identifier les modèles correspondant le plus à notre problématique et à nos besoins. Le Perceptron MultiCouche (PMC), un réseau de neurones à couches cachées fait partie des outils souvent adoptés en prévision. Nous verrons ses avantages et ses inconvénients en tant que modèle de prévision de la qualité de l'air en Corse.

2.1 Objectif de la prévision de la qualité de l'air

La prévision de la qualité de l'air est une problématique importante de la surveillance. De nos jours, elle est réalisée à l'aide de différents outils informatiques et avec des objectifs de modélisation différents. Tout d'abord, il est important de différencier les modèles climatiques des modèles de prévision de la qualité de l'air.

L'objectif des premiers est de prévoir l'évolution du climat, qui dépend grandement de la pollution atmosphérique anthropique. L'état de l'ensemble de l'atmosphère et ses interactions avec les océans et les continents sont modélisés avec des échelles de temps de l'ordre du siècle, en utilisant différents scénarii d'évolution des émissions anthropiques, qui dépendent fortement de modèles économiques et sont difficiles à prévoir. Ces prévisions doivent permettre d'orienter nos politiques à long terme afin de prendre en compte notre impact sur le climat, notamment vis-à-vis des émissions de GES provoquant un réchauffement climatique global d'origine humaine.

Nous n'aborderons pas ce type de modèle dans les travaux présentés ici.

Les modèles de prévision de la qualité de l'air, qui nous concernent, ont des horizons de prévision plus courts, de quelques heures à plusieurs jours. Leur objectif est double :

- Ils permettent en premier lieu de valider les connaissances théoriques sur les mécanismes atmosphériques qui gouvernent l'évolution de la qualité de l'air, et sont par là nécessaires à la recherche.
- Ils aident les autorités à surveiller l'état de l'air auquel la population et les milieux naturels sont exposés, à des fins de protection de la santé publique et de l'environnement.

Les modèles prédictifs permettent de fournir à temps aux autorités l'information nécessaire à la prise de mesures ponctuelles pour, à court terme, limiter ou atténuer les pics de pollution. Ces mesures sont principalement des limitations d'émissions de polluants via des limitations de trafic (automobile, maritime, aéroportuaire), d'activités industrielles (production de biens, d'énergie...) ou domestiques (notamment le chauffage). Elles peuvent également concerner l'exposition de la population, par des annulations d'événements publics d'extérieur par exemple, ce qui reste plus rare. La prévision de l'imminence d'un pic est également communiquée au public.

En France, ce sont principalement les préfectures qui mettent en place ce type de mesures au travers d'arrêtés préfectoraux, ou les mairies par arrêté municipal. Des consignes de santé en cas de pic sont diffusées afin que la population sache comment adapter ses activités pour limiter l'effet des fortes concentrations de polluants sur la santé. Pour limiter leur exposition, les personnes sensibles (enfants, personnes âgées ou atteintes de certaines maladies chroniques) sont invitées à rester à l'intérieur, et les activités sportives d'extérieurs doivent être limitées.

Les modèles de prévision permettent également un suivi permanent pouvant orienter les politiques publiques en matière d'aménagement territorial afin de prendre en compte la qualité de l'air et de l'améliorer. Les nouveaux aménagements sont de plus en plus pensés pour ne pas favoriser localement les fortes concentrations en polluant. Les modèles prédictifs peuvent être un outil efficace pour évaluer différents scenarii d'aménagement.

Historiquement, la prévision a commencé dans les années 60, aux débuts de la surveillance de la qualité de l'air qui a fait suite aux épisodes de pollution désastreux des années 50. Les premières prévisions étaient, en fait, des prévisions météorologiques de conditions défavorables à la qualité de l'air via des modèles de prévision du temps (Niemeyer, 1960). Les conditions menant à la stagnation des polluants étaient identifiées comme potentiellement dangereuses pour la qualité de l'air.

Dans les années 70, des modèles de prévision de la qualité de l'air à proprement parler sont apparus. On en distingue particulièrement deux types : les modèles statistiques, qui prévoient les valeurs futures d'une série temporelle à partir de l'étude d'un historique de données, et les modèles mécanistes ou déterministes qui simulent numériquement des modèles de connaissances de l'atmosphère. Ils sont similaires aux NWP (Numerical Weather Prediction, modèles de prévision météorologique) mais qui intègrent la prévision des concentrations en polluants atmosphériques.

2.2 Modèles déterministes

Avec l'évolution des connaissances sur les processus de transport des polluants, de leurs sources, de leurs puits et de processus physico-chimiques atmosphériques, des modèles de prévision de la qualité de l'air dit « mécanistes » ou « déterministes » sont apparus. Il en existe plusieurs types, différents en fonction de l'échelle géographique. Parmi eux, les modèles de chi-

mie transport (Chemical Transport Model (CTM) en anglais) couvrent des domaines spatiaux allant de l'échelle locale à l'échelle mondiale. Leur évolution depuis les premières prévisions de potentiel de pollution dans les années 60, déduites de prévisions météorologiques, aux modèles actuels est décrite par Zhang *et al.* (2012*a*).

Il existe deux grandes catégories de CTM (Jacob, 1999), les modèles eulériens (à système de coordonnées fixes) et les modèles lagrangiens (à système de coordonnées mobiles). Les premiers discrétisent l'espace en un ensemble de « boîtes », unités de volume au sein desquelles sont simulés les sources et les puits des polluants gérés et les réactions chimiques, et entre lesquelles le transport des polluants est simulé, à chaque pas de temps. On les appelle parfois modèles 3D en référence à cette représentation.

Les modèles lagrangiens n'ont pas de maillage fixe. Leur référentiel suit l'évolution spatiale d'une masse d'air, le long de sa trajectoire, en suivant une densité de probabilité de transition afin de prendre en compte l'aspect stochastique du transport. Cette approche est adaptée à des modèles de dispersion, étudiant l'évolution d'un panache pour une étude locale. Il existe également des modèles hybrides, comme le modèle MOCAGE (MOdèle de Chimie A Grande Echelle) de Météo-France (Peuch *et al.*, 2009), modèle eulérien qui inclue des modules lagrangiens pour un suivi d'événements exceptionnels plus précis.

On trouve également des modèles de dispersion plus simples, appelés parfois modèles gaussiens, adaptés à la modélisation de la petite échelle (échelle de la rue), proche des sources primaires, jusqu'à la dispersion de panaches à l'échelle continentale. Ce sont des modèles basés sur la mécanique des fluides qui modélisent uniquement la dispersion des polluants, et non leur transformation chimique. Ils prennent en compte le comportement des gaz et les principales variables météorologiques (vent, pluie...) fournies par d'autres modèles. Ce type de modèle est parfois utilisé pour assurer le suivi d'épisodes de pollution après un accident industriel (accident nucléaire par exemple). Certains CTM sont hybridés avec des modèles gaussiens, comme par exemple le modèle Polyphemus (Mallet *et al.*, 2007).

Les modèles eulériens ont des domaines spatiaux de taille différente. Certains domaines recouvrent toute la planète, on les appelle modèles globaux. Il existe également des modèles à micro-échelle (de l'ordre du kilomètre), des modèles à méso-échelle (entre la dizaine et le millier de kilomètres) ou des modèles à échelle synoptique (plusieurs milliers de kilomètres). Ils ont besoin de données provenant d'autres modèles recouvrant leur domaine pour pouvoir forcer les conditions aux limites de leur propre domaine. Le forçage consiste à assigner des valeurs venant de l'extérieur du modèle afin de lui apporter les données manquantes. Il est nécessaire pour prendre en compte les interactions avec l'extérieur du domaine (quand il ne s'agit pas d'un modèle global) mais également pour donner des conditions initiales à partir desquelles commencer la modélisation.

A partir de ces données, le modèle va calculer à chaque pas de temps les champs de concentration de chaque espèce chimique en chaque point de son domaine. Il utilise les équations de continuité, équations différentielles du premier ordre qui relient l'évolution spatio-temporelle de la concentration de chaque espèce à ses sources, ses puits et à son transport (Jacob, 1999). Le transport des polluant correspond à l'advection par le champ de vent. L'équation de continuité dans chaque unité de volume prend la forme :

$$\begin{aligned} \frac{\partial n}{\partial t} &= -\nabla \cdot \mathbf{F} + S - P \\ \frac{\partial n}{\partial t} &= -\nabla \cdot (n\mathbf{U}) + S - P \end{aligned} \tag{2.1}$$

avec n la concentration de l'espèce chimique en molécules.cm^{-3}, $\mathbf{F}$ le flux atmosphérique en molécules.cm^{-2}.s^{-1}, $\mathbf{U}$ le champ de vent local en cm.s^{-1}, S l'apport dû aux sources et P les pertes dues aux puits, tout deux en molécules.cm^{-3}.s^{-1}. Notons que dans ces travaux, nous notons les vecteurs et les matrices en gras.

Chaque unité de volume prend la forme d'une boîte élémentaire de côté ∂x, ∂y et ∂z. L'opérateur ∇ est utilisé afin de représenter la divergence d'un vecteur. On a de manière générale :

$$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \tag{2.2}$$

pour un vecteur $\mathbf{A}$ et ses composantes A_x, A_y et A_z respectivement selon x, y et z.

Le flux atmosphérique $F = n\mathbf{U}$ décrit le nombre de molécules qui entrent ou sortent par la face d'une « boîte » en fonction du temps sous l'action du champ de vent. Ainsi, $\partial F_x/\partial x$ représente le flux du polluant hors de cette boîte selon x. Dans l'équation 2.1, le terme $-\nabla \cdot \mathbf{F}$ correspond donc au bilan des gains et des pertes de molécules dus à l'advection au sein de la boîte.

Quant aux sources S et aux puits P, ils représentent l'ensemble des phénomènes apportant ou retirant des quantités de polluants au sein de nos boîtes. Les sources comprennent par exemple les émissions prises en charge par le cadastre, la création de polluant secondaire (photochimie atmosphérique, phénomènes comme les éclairs, etc.). Les puits regroupent par exemple les phénomènes de dépôt (sec et humide), les pertes dues à la photochimie atmosphérique, etc. Nous n'entrerons pas ici dans le détail de cette gestion, variable selon les CTM.

Le flux $\mathbf{F}$ n'est en fait pas de nature purement advective mais comporte une composante turbulente. On sépare les composantes advective et turbulente du flux, respectivement $\mathbf{F}_A$ et $\mathbf{F}_T$. La dispersion d'un panache due aux turbulences peut être décrite comme une décroissance gaussienne des concentrations autour de l'axe de trajectoire principale. $\mathbf{F}_T$ peut alors être décrit en utilisant la loi de diffusion de Fick. On note que la diffusion moléculaire ne contribue pas significativement au transport, mais le flux turbulent se comporte de la même manière. On a :

$$\overline{\mathbf{F}}_T = -\mathbf{K}\ n_a\ \nabla \cdot \overline{\mathbf{r}} \tag{2.3}$$

avec $\overline{\mathbf{F}}_T$ le flux turbulent moyenné dans le temps, n_a la concentration des molécules d'air (N_2 et O_2), $\overline{\mathbf{r}}$ le rapport de mélange de l'espèce chimique moyenné dans le temps et $\mathbf{K}$ la matrice des coefficients de diffusion turbulente selon x, y et z. Cette dernière est définie empiriquement par des mesures de $\overline{\mathbf{F}}_T$ et de $\nabla \cdot \overline{\mathbf{r}}$. L'équation de continuité 2.1 devient alors :

$$\frac{\partial \overline{n}}{\partial t} = -\nabla \cdot (\overline{n}\overline{\mathbf{U}}) + \nabla \cdot (\mathbf{K}\ n_a\ \nabla \cdot \overline{\mathbf{r}}) + \overline{S} - \overline{P} \tag{2.4}$$

où $-\nabla \cdot (\overline{n}\overline{\mathbf{U}})$ correspond à $\overline{\mathbf{F}}_A$, la composante advective du flux moyennée dans le temps. Le CTM peut ainsi calculer numériquement la concentration de chaque polluant en chaque point de la grille à chaque pas de temps. Les variables météorologiques nécessaires peuvent soit être gérées par un module météorologie du CTM (« on-line »), soit être fournies par un modèle météorologique externe (« off-line ») couplé au CTM. De même, certains CTM sont pourvus de modules permettant de gérer l'aérosol.

En France, le modèle eulérien CHIMERE (Menut *et al.*, 2013, description du module aérosol : Bessagnet *et al.*, 2004, 2008) a été développé par l'INERIS et deux laboratoires de l'Institut Pierre Simon Laplace (IPSL) : le Laboratoire Inter-universitaire des Systèmes Atmosphériques (LISA) et le Laboratoire de Météorologie Dynamique (LMD). Il est disponible en

ligne (http://www.lmd.polytechnique.fr/chimere/), distribué sous licence de logiciel libre GNU General Public Licence (GNU GLP). Son développement est continu depuis sa création, et il est décliné en plusieurs versions.

La plate-forme de prévision PREV'AIR (Rouil *et al.*, 2009) regroupe les modèles CHIMERE et MOCAGE. Ses prévisions sont disponibles au public (http://www2.prevair.org/). Cette plate-forme est le fruit d'une collaboration entre l'INERIS, le LCSQA, Météo-France et le Centre National de la Recherche Scientifique (CNRS). PREV'AIR fournit des prévisions sur le domaine national et européen (les prévisions sur des domaines couvrant les DOM-TOM sont prévues) avec une résolution de 0.15° × 0.1° soit à peu près 17 km × 11 km (voir figure 2.1). Les résultats de PREV'AIR sont utilisés par les AASQA, souvent en complément de leurs propres moyens de prévision.

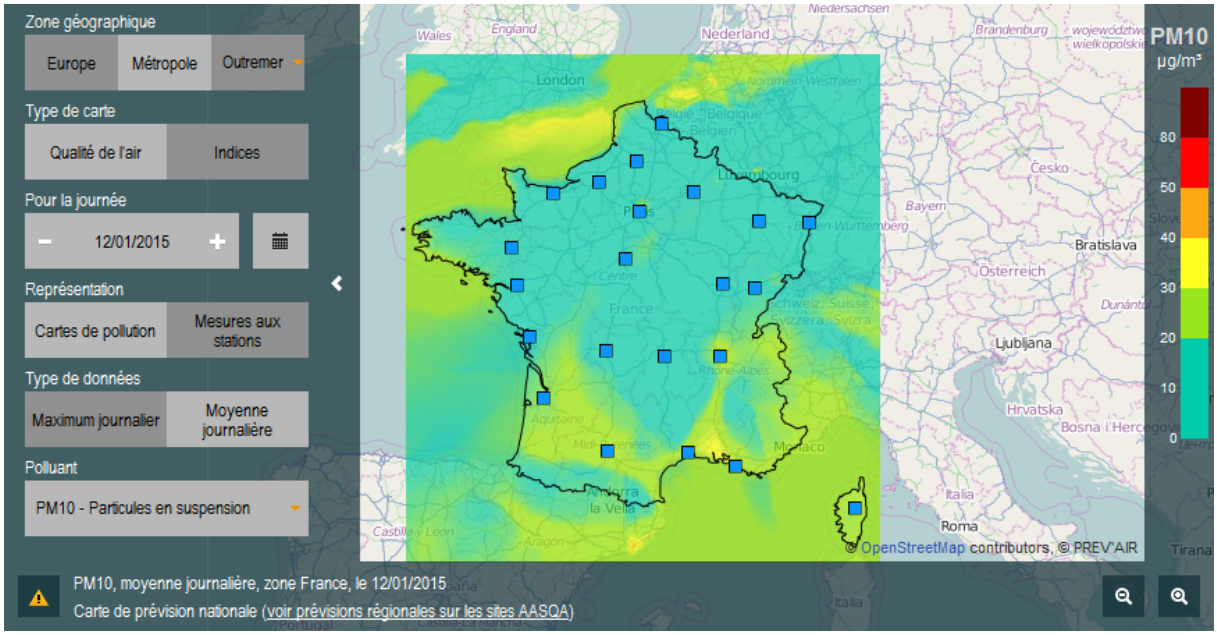


FIGURE 2.1 : Exemple de carte de prévision fournie par la plate-forme PREV'AIR.

Le développement interne de modèle prédictif est en effet commun à plusieurs AASQA, et leur permet d'avoir des outils spécialisés sur les domaines de compétence des associations (domaine du modèle coïncidant avec la région surveillée, inventaire régional des émissions, assimilation des mesures de la région...). PREV'AIR, quant à lui, utilise l'inventaire européen EMEP (European Monitoring and Evaluation Programme) et l'INS (Inventaire National Spatialisé) et assimile certaines données du réseau des AASQA. Il est utilisé à l'échelle nationale, et peut être moins précis localement que des modèles plus spécialisés géographiquement.

Par exemple Air PACA, l'AASQA de la région PACA, gère la plate-forme AIRES (le nom provient du mot « air » en occitan) qui utilise également CHIMERE, avec une météorologie fournie par le modèle MM5 (Modèle à Méso-échelle PSU/NCAR) et un cadastre régional des émissions à une résolution de 4 km × 4 km (prévisions en ligne http://www.aires-mediterranee.org/). Cette plate-forme fournit également des prévisions sur le domaine de la Corse (voir figure 2.2).

La situation géographique et la topographie de la Corse rendent l'utilisation de CTM plus délicate que dans la plupart des régions. En effet, le relief est très montagneux, ce qui demande une grande résolution spatiale pour modéliser correctement la topographie de l'île. Des régimes de brise de montagne prennent forme et ont un fort impact sur la qualité de l'air dans l'île, tout comme les phénomènes d'inversion thermique favorisés par les reliefs des vallées. De plus, bien

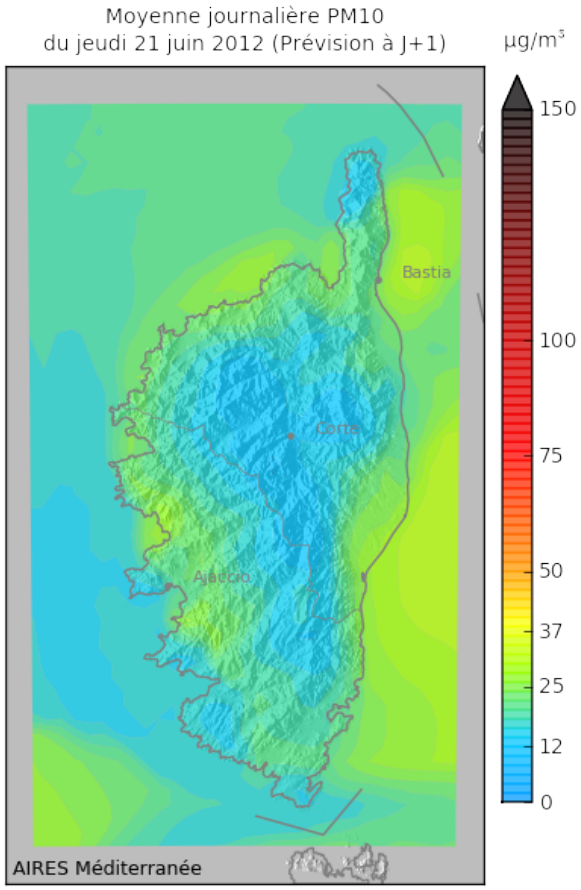


FIGURE 2.2 : Exemple de carte de prévision à méso-échelle fournie par la plate-forme AIRES.

que ce travail soit actuellement en cours à Qualitair Corse Corse (prévu pour 2015), il n'y a pas d'inventaire régional des émissions disponible pour réaliser un cadastre précis afin d'alimenter un CTM. Pour finir, la situation géographique de la Corse par rapport au continent français place souvent l'île en bordure du domaine de modélisation à l'échelle nationale, comme c'est le cas avec PREV'AIR par exemple, (voir figure 2.1). Cette proximité avec la bordure où les concentrations en polluants sont forcées à partir de modèles plus globaux ou de mesures de pollution de fond détériore la qualité des prévisions sur la Corse. PREV'AIR et AIRES n'en restent pas moins des outils très utiles pour la prévision opérationnelle de la qualité de l'air effectuée quotidiennement par Qualitair Corse.

Le modèle SKIRON (Kallos *et al.*, 1997) est également utilisé, notamment pour sa prévision des poussières sahariennes (http://forecast.uoa.gr/dustindx.php). Ces événements expliquent la majeure partie des épisodes de pollution aux particules en Corse mais aussi dans le sud de l'Italie (Pederzoli *et al.*, 2010) et dans le sud et l'est de l'Espagne (Rodriguez *et al.*, 2001), d'où l'utilité de cet outil. La figure 2.3 présente les résultats de prévision du modèle SKIRON lors d'un épisode de transport de poussières sahariennes. De tels épisodes de transport de poussières du Sahara ont également fréquemment lieux dans les Caraïbes (Petit, 2005). SKIRON a l'avantage de se focaliser sur le Sahara, ce que ne font pas PREV'AIR et AIRES puisque leur domaine est régional, national voire européen. Le forçage aux limites de tels domaines n'est pas suffisant pour prendre en compte correctement le transport de poussières du Sahara.

Au-delà des problématiques de domaine, les particules ont une dynamique complexe, et sont plus difficiles à modéliser pour les CTM que les polluants gazeux. L'aérosol a tendance à

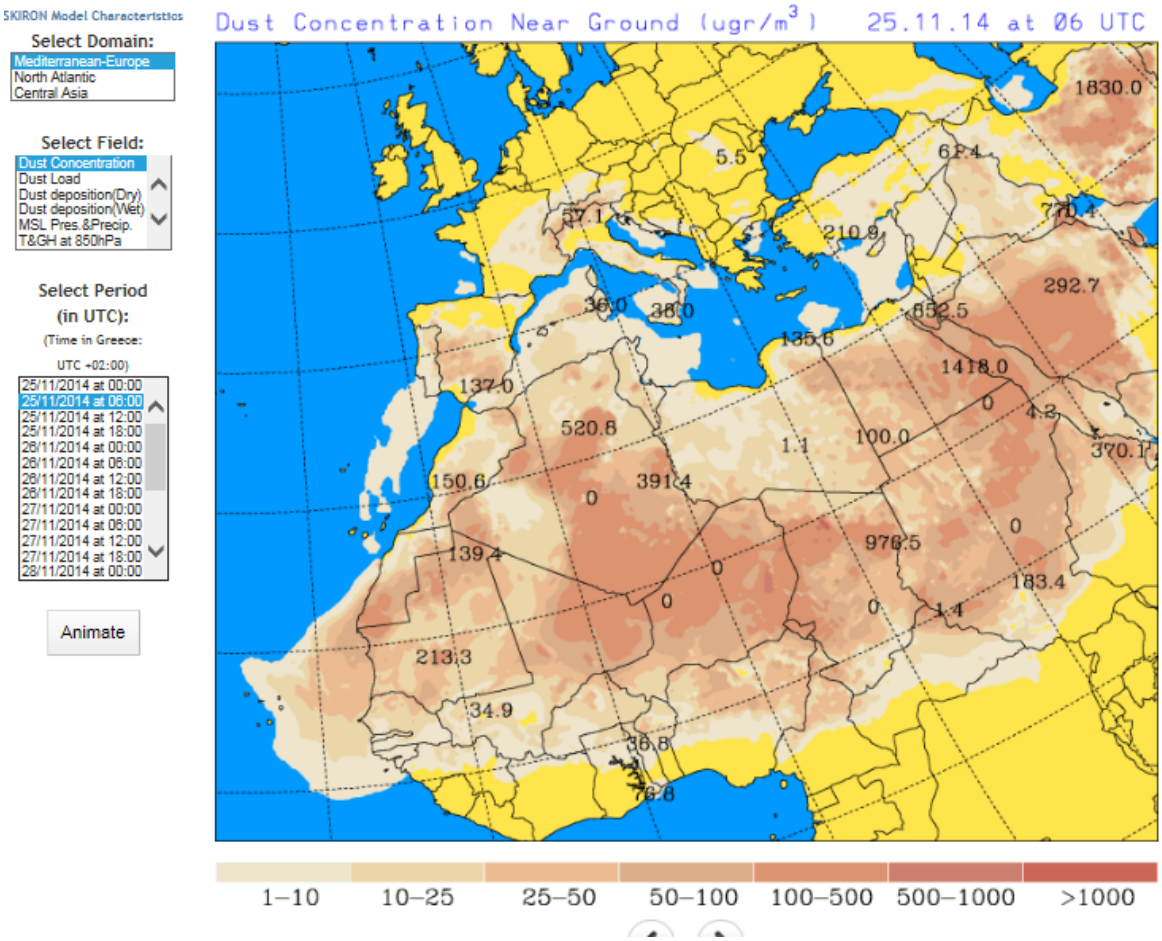


FIGURE 2.3 : Exemple de carte de prévision fournie par la plate-forme SKIRON montrant un épisode de transport de poussières sahariennes à l'échelle synoptique.

être sous-estimé, soit à cause d'une sous-estimation des émissions primaires, soit à cause des imprécisions sur les émissions de précurseurs et sur la formation d'aérosol secondaire (McKeen *et al.*, 2007; Yu *et al.*, 2008). Les recherches actuelles liées à la modélisation déterministe tentent de pallier ces défauts notamment grâce à des approches statistiques, par de la modélisation d'ensemble ou de l'assimilation statistique (Zhang *et al.*, 2012*b*).

La modélisation d'ensemble, déjà répandue en météorologie, permet d'apporter une notion d'incertitude à une sortie de modèle. Plusieurs résultats de CTM avec des données d'entrée légèrement différentes sont comparés, permettant de donner plus d'importance à un scénario qui ressort souvent et de le considérer comme plus probable. Une étude récente a ainsi permis d'évaluer la précision de PREV'AIR (Debry et Mallet, 2014). Plusieurs modèles différents peuvent être utilisés dans cette optique. Cela permet de palier un manque de données disponibles pour forcer les conditions initiales des modèles. Cette méthodologie nécessite logiquement plus de temps de calcul qu'une modélisation simple.

L'assimilation statistique (technique permettant la prise en compte de mesures en temps réel par un modèle numérique déterministe) fait partie des efforts de recherches actuels menés sur les CTM. Les mesures vont forcer en temps réel les calculs du CTM, qui va gagner en précision. Ces enjeux actuels de la modélisation déterministe sont revus par Zhang *et al.* (2012*b*). Dix-huit modèles utilisés en Europe pour la prévision de la qualité de l'air sont présentés par Kukkonen *et al.* (2012), qui soulignent l'intérêt de la modélisation d'ensemble et de l'assimilation statistique.

Les calculs effectués par un CTM pour réaliser une prévision nécessitent beaucoup de ressources informatiques. Typiquement, un supercalculateur est nécessaire pour faire fonctionner un CTM, et toutes les AASQA n'ont pas les moyens humains et financiers pour gérer ce matériel, ce qui limite le développement de modèles régionaux et renforce l'intérêt de l'échange de données inter-AASQA ainsi que des plates-formes de prévision disponibles en ligne comme celles citées précédemment.

Des systèmes de centralisation de données existent, afin de permettre l'assimilation de données au système PREV'AIR. Ainsi, l'ADEME (Agence De l'Environnement et de la Maîtrise de l'Energie) a développé BASTER (BASe de données en TEmps Réel), une base de données récupérant des données mesurées par des stations fixes de chaque AASQA afin que l'INERIS y ait accès quasiment en temps réel. Cependant, toutes les stations ne sont pas prises en compte pour l'assimilation. Un échange de données automatique a également été mis en place entre Qualitair Corse et Air PACA afin de permettre l'assimilation de données pour les prévisions d'AIRES qui couvrent le domaine corse.

2.3 Modèles statistiques

La modélisation statistique suit un tout autre paradigme. Le but d'un modèle statistique est de modéliser au mieux un phénomène, non pas à partir des connaissances que l'on en a, mais à partir des informations qui peuvent être tirées d'un jeu de données qui le décrit. Ainsi, il n'est pas nécessaire de formaliser le fonctionnement du phénomène étudié et on ne subit pas de biais dus à des imprécisions sur nos modèles de connaissances. Par contre, il est nécessaire de disposer d'un jeu de données qui permette au modèle de représenter le phénomène étudié de manière satisfaisante, et le modèle doit être capable d'exploiter l'information présente dans ce jeu de données. De plus, l'imprécision des données utilisées se répercutera sur les prévisions. Les modèles statistiques occupent une part importante des modèles utilisés pour la prévision de la qualité de l'air (Ionescu, 2013).

Le plus souvent, les modèles statistiques prévisionnels travaillent avec des séries temporelles. Une série temporelle (parfois appelée « chronique ») est un échantillon dont les valeurs sont organisées dans le temps, selon un pas de temps la plupart du temps fixe. La série temporelle « cible » $y(t)$ est celle que l'on veut prévoir. La sortie du modèle est donc notée $\hat{y}(t)$. En entrée du modèle, l'utilisateur fournit une ou plusieurs séries temporelles $\mathbf{x}(t)$ (les « entrées », ou « inputs », qu'on peut réunir en une matrice $\mathbf{x}$). Il peut s'agir uniquement des éléments de la série temporelle $y(t)$, la variable appelée endogène, dans le cas d'un modèle univarié, mais également d'autres séries temporelles qu'on appelle variables exogènes dans le cas d'un modèle multivarié. Dans tous les cas, la sortie du modèle est nécessairement décalée dans le temps par rapport aux entrées, d'un certain « horizon » h. Un modèle prédictif statistique prend ainsi de manière générale la forme :

$$y(t+h) = \hat{y}(t+h) + e = f(\mathbf{p}, \mathbf{x}(t)) + e \tag{2.5}$$

où y est la variable à prévoir, $\hat{y}$ représente la prévision par le modèle de cette variable, h l'horizon de cette prévision, e un terme d'erreur, f représente le modèle, $\mathbf{p}$ ses paramètres et $\mathbf{x}(t)$ ses variables d'entrée.

A partir des données fournies, les modèles statistiques définissent leurs paramètres $\mathbf{p}$ grâce à différentes méthodes ou algorithmes, avant d'être prêts à fonctionner. Cette phase fait partie du domaine de l'apprentissage automatique, ou « machine learning » en anglais. Cet apprentissage prépare automatiquement le modèle, qui produit ensuite ses résultats en fonction des données

d'entrée qu'on lui fournit, sans que l'on puisse forcément interpréter son fonctionnement. Pour cette raison, certains modèles statistiques complexes sont parfois qualifiés de « boîtes noires » ou « black boxes », quand on ne peut pas les interpréter au delà de leur forme décrite dans l'équation 2.5. Ils ne sont pas adaptés à toutes les problématiques de prévision. A l'inverse, il est parfois possible d'interpréter les paramètres **p** de certains modèles plus simples. On peut alors comprendre le cheminement entre l'entrée et la sortie, qui peut révéler certains phénomènes du problème étudié.

Les données fournies à cette famille de modèle, en plus d'être suffisamment explicatives, doivent également correspondre à un échantillon suffisamment large pour receler un nombre d'exemples suffisant à un apprentissage du modèle, pour que celui-ci ait une bonne capacité de généralisation et soit robuste face à toutes les situations. Dans le contexte de la prévision de la qualité de l'air par exemple, il faut que toutes les situations particulières menant à des événements typiques de pollution soient suffisamment présentes dans les données d'apprentissage pour que le modèle, une fois paramétré, puisse les modéliser correctement. On peut ici risquer une analogie avec l'apprentissage des humains, qui vont essayer de prévoir quelque chose en se rappelant leur expérience du phénomène en question, qu'ils doivent alors connaître suffisamment bien.

C'est d'ailleurs par analogie avec l'être humain qu'une partie des modèles statistiques est rattachée à la famille de l'Intelligence Artificielle (IA), soit parce que ces modèles assurent une tâche qu'on assimile à une fonction du cerveau humain (perception, apprentissage...), soit parce que la nature de ces modèles est directement inspirée des connaissances sur le fonctionnement du cerveau humain (comme c'est le cas par exemple des Réseau de Neurones Artificiels (RNA)). L'appellation « intelligence artificielle » appliquée à un modèle peut parfois être mal interprétée, et donner l'impression qu'il s'agit d'une tentative de se rapprocher d'une intelligence humaine, ce qui n'est quasiment jamais le cas. Les RNA, par exemple, bien qu'inspirés à l'époque de leur création par des connaissances en neurobiologie (McCulloch et Pitts, 1943), restent plus proches du fonctionnement de modèles de Box-Jenkins (voir section 2.3.2) que de celui d'un système nerveux, s'en rapprocher n'étant d'ailleurs pas l'objectif de leur utilisation.

A la fin des années 90, les modèles statistiques étaient majoritairement utilisés dans le monde pour la prévision d'ozone grâce à leurs avantages en termes de simplicité et de performances (Fromage et Gilibert, 1997). Par la suite, les progrès en termes de connaissances atmosphériques et l'évolution de l'informatique ont favorisé l'essor des modèles déterministes. Mais les modèles statistiques ont continué à se développer et à être appliqués à la prévision de la qualité de l'air. Ils obéissent à un paradigme différent des modèles déterministes et ainsi les deux familles se complètent plus qu'elles ne se concurrencent.

Nous présentons ici quelques types de modèles statistiques utilisés en prévision de série temporelle, et notamment en prévision de la qualité de l'air. Nous commencerons par les modèles naïfs, les plus simples. Nous nous intéresserons ensuite à deux familles faisant référence en prévision de série temporelle : les modèles de Box-Jenkins et les arbres de décision. Nous verrons pour finir pourquoi les RNA ont fini par s'imposer et comment ils ont été appliqués à la prévision de la qualité de l'air.

2.3.1 Modèles naïfs

Les modèles qu'on appelle « naïfs » sont des modèles très simples. En prévision, il s'agit des modèles de persistance, dont le principe est de considérer que la valeur prédite est égale à la valeur actuelle (équation 2.6). Ce type de prévision peut paraître simpliste mais dans certains

cas il s'avère utile, par exemple quand la variable à prédire à une composante périodique de la même durée que l'horizon de prévision. On a :

$$\begin{aligned} \hat{y}(t+h) &= y(t) \\ y(t+h) &= y(t) + e(t) \end{aligned} \tag{2.6}$$

avec y la variable d'entrée (valeur actuelle de la variable à prédire), $\hat{y}$ la prédiction du modèle, t le temps, h l'horizon du modèle et $e(t)$ un terme d'erreur.

On peut aussi utiliser la persistance en prenant en compte un modèle de connaissance qui décrirait le profil de l'évolution de la variable à prédire. A partir de l'équation 2.6 on obtient :

$$y(t+h) = y(t)\frac{f(t+h)}{f(t)} + e(t) \tag{2.7}$$

avec f la fonction de t décrivant l'évolution de la variable d'après le modèle de connaissance.

On remarque qu'il ne s'agit pas réellement de modèles statistiques puisque les modèles de persistance n'utilisent pas d'historique de données. Dans leur philosophie, ce sont en réalité des modèles déterministes représentant une mécanique très simple. Ils sont décrits ici en raison de leur ressemblance avec les modèles « à horizon », tels que décrits dans l'équation 2.5. Ce type de modèle est surtout utilisé pour évaluer d'autres modèles prédictifs, le fait d'obtenir des performances qui ne seraient pas à la hauteur de celles d'un modèle de persistance étant rédhibitoire pour lesdits modèles. On peut également les utiliser si l'on ne dispose d'aucun historique de données.

2.3.2 Modèles linéaires

Les modèles de cette famille ont été fréquemment utilisés en prévision de série temporelle, notamment en qualité de l'air. Ces modèles tentent de capter les liens linéaires qui existent entre la variable à prédire et la ou les variables d'entrées, endogènes ou exogènes.

Un modèle régressif simple appliqué à la prévision de série temporelle à l'horizon h prend la forme :

$$y(t+h) = \alpha + \boldsymbol{\beta}\,\mathbf{x}(t) + e(t) \tag{2.8}$$

avec $y(t+h)$ la sortie du modèle, $\mathbf{x}(t)$ le vecteur des séries temporelles en entrée, α et $\boldsymbol{\beta}$ les coefficients de la régression et $e(t)$ un terme d'erreur. Si les séries temporelles $\mathbf{x}(t)$ comprennent des variables exogènes, on parle de régression linéaire multiple (souvent nommé Multiple Linear Regression (MLR) de son nom anglais). Il peut s'agir d'une même variable qu'on utilise plusieurs fois en lui appliquant des délais différents, c'est-à-dire qu'on utilise plusieurs séries temporelles qui représentent ses valeurs passées. Dans ce cas on parle de modèle AR (Auto-Régressif) (voir équation 2.9).

Ce type de modèle peut permettre d'obtenir de bons résultats. Stadlober *et al.* (2008) ont appliqué la régression linéaire multiple à la prévision de PM10 dans des villes des Alpes italiennes et autrichiennes. D'autres études utilisant les MLR sont citées dans le tableau 2.1.

La famille des modèles de Box-Jenkins (ARMA, ARIMA, SARIMA, SARIMAX, etc.) est basée sur les modèles AR et MA (de l'anglais Moving Average pour moyenne mobile). Ces modèles ont été fréquemment utilisés en prévision de la qualité de l'air. Ils ont été popularisés

par Box et Jenkins (Box *et al.*, 1994), qui ont développé une méthodologie afin de les paramétrer. Ils sont ici brièvement décrits.

Un modèle autorégressif d'ordre p (noté $AR(p)$) modélise une série temporelle $x(t)$ comme étant une combinaison linéaire de ses p précédentes valeurs et d'un terme d'erreur $e(t)$. Il s'agit donc d'une régression linéaire classique, appliquée à la série temporelle à prédire avec plusieurs délais. Selon un modèle $AR(p)$ on a :

$$y(t) = \sum_{i=1}^{p} \phi_i \, y(t-i) + e(t) \tag{2.9}$$

avec p l'ordre du modèle, ϕ_i les coefficients de la régression et $e(t)$ un terme d'erreur. On remarque que bien qu'aucun horizon h n'apparaisse, il s'agit bien de prévision puisque $y(t)$ n'est fonction que de ses valeurs précédentes. L'horizon de la prévision est égal à 1.

Pour décrire cette famille de modèles, il est utile d'introduire l'opérateur retard, ou opérateur délai noté L (parfois noté B) :

$$Ly(t) = y(t-1) \tag{2.10}$$

Plus généralement, noter L à la puissance n correspond à retarder n fois la variable y :

$$L^n y(t) = y(t-n) \tag{2.11}$$

En utilisant l'opérateur L, l'équation 2.9 devient :

$$y(t) = \sum_{i=1}^{p} \phi_i \, L^i y(t) + e(t) \tag{2.12}$$

D'un autre côté, un modèle moyenne mobile d'ordre q (noté $MA(q)$) modélise une série temporelle $y(t)$ comme une fonction linéaire des q valeurs d'une fonction de bruit blanc $e(t)$. L'évolution de la série temporelle est considérée comme une fluctuation autour de sa valeur moyenne, nulle si la série temporelle est centrée. Un modèle $MA(q)$ représente $y(t)$ suivant la relation :

$$\begin{aligned} y(t) &= \sum_{j=0}^{q} \theta_j \, e(t-j) \\ y(t) &= \sum_{j=0}^{q} \theta_j \, L^j e(t) \end{aligned} \tag{2.13}$$

avec q l'ordre du modèle, θ_j les coefficients de la régression et $e(t)$ un terme d'erreur.

En combinant les modèles $AR(p)$ et $MA(q)$, on obtient un modèle $ARMA(p,q)$ qui suit l'équation suivante :

$$y(t) - \left(\sum_{i=1}^{p} \phi_i \, L^i y(t)\right) = e(t) + \left(\sum_{j=1}^{q} \theta_j \, L^j e(t)\right)$$
$$\left(1 - \sum_{i=1}^{p} \phi_i \, L^i\right) y(t) = \left(1 + \sum_{j=1}^{q} \theta_j \, L^j\right) e(t) \tag{2.14}$$

Ces modèles, AR, MA et ARMA, décrivent des séries temporelles stationnaires. Une série temporelle est stationnaire si ses propriétés statistiques ne changent pas au cours du temps. Elle ne doit donc pas avoir de tendance, ni être périodique, et sa variance ne doit pas évoluer. Si ce n'est pas le cas comme avec la plupart des séries temporelles issues de données réelles, il est dans un premier temps nécessaire de les rendre stationnaires.

Plusieurs méthodes sont utilisables, l'une d'entre elle, permettant d'éliminer une tendance présente dans la série temporelle, a été intégrée au modèle ARMA (AutoRegressive Moving Average) pour donner le modèle ARIMA (pour AutoRegressive Integrated Moving Average), il s'agit d'une méthode de différenciation. Une tendance de la série temporelle se traduit par une différence constante entre les points successifs de la série. Une différenciation d'ordre 1, qui consiste à utiliser $y(t) - y(t-1)$ à la place de $y(t)$, peut alors supprimer cette différence et supprimer la tendance. On a :

$$\begin{aligned} y(t) - y(t-1) &= y(t) - L\, y(t) \\ &= (1 - L)\, y(t) \end{aligned} \tag{2.15}$$

Si la série temporelle présente une évolution non-stationnaire (à l'inverse d'une simple tendance) alors l'utilisation d'une différenciation d'ordre supérieur à 1 peut être utile. Le modèle ARIMA utilise $y_d(t)$, terme correspondant à une différenciation d'ordre d de la variable $y(t)$, d devant être déterminé de manière à obtenir une série stationnaire.

$$y_d(t) = (1 - L)^d \, y(t) \tag{2.16}$$

A partir de 2.14 et 2.16 on obtient le modèle $ARIMA(p, d, q)$:

$$\left(1 - \sum_{i=1}^{p} \phi_i \, L^i\right) (1 - L)^d \, y(t) = \left(1 + \sum_{j=1}^{q} \theta_j \, L^j\right) e(t) \tag{2.17}$$

En plus de la tendance, les modèles SARIMA (pour Seasonal AutoRegressive Integrated Moving Average) permettent d'éliminer la périodicité, ou saisonnalité, que présente une série temporelle. Si une série temporelle présente une saisonnalité s (par exemple $s = 12$ pour une série temporelle mensuelle à composante saisonnière) alors on peut la rendre stationnaire en la différenciant de s pas de temps. On a :

$$y_s(t) = (1 - L^s)^D (1 - L)^n \, y(t) \tag{2.18}$$

qui suit un modèle $ARIMA(P, D, Q)$:

$$\left(1 - \sum_{m=1}^{P} \Phi_m \, L^m\right) (1 - L)^D \, y_s(t) = \left(1 + \sum_{n=1}^{Q} \Theta_n \, L^n\right) e(t) \tag{2.19}$$

Le modèle $SARIMA(p,d,q)(P,D,Q)$ est construit ainsi :

$$\left(1-\sum_{m=1}^{P}\Phi_m\,L^{m\times s}\right)(1-L^s)^D\left(1-\sum_{i=1}^{p}\phi_i\,L^i\right)(1-L)^d\,y(t)=\left(1+\sum_{n=1}^{Q}\Theta_n\,L^{n\times s}\right)\left(1+\sum_{j=1}^{q}\theta_j\,L^j\right)e(t) \tag{2.20}$$

Intégrant le prétraitement des séries temporelles saisonnières et pourvues d'une tendance, le modèle SARIMA est un outil puissant, très utilisé en prévision.

On peut ajouter des variables exogènes à un modèle SARIMA, via différentes méthodes. Ce type de modèle, appelé SARIMAX (pour Seasonal AutoRegressive Integrated Moving Average with eXogenous inputs), peut par exemple prendre la forme :

$$\begin{aligned}\left(1-\sum_{m=1}^{P}\Phi_m\,L^{m\times s}\right)(1-L^s)^D\left(1-\sum_{i=1}^{p}\phi_i\,L^i\right)(1-L)^d\,y(t)=\\ \left(1+\sum_{n=1}^{Q}\Theta_n\,L^{n\times s}\right)\left(1+\sum_{j=1}^{q}\theta_j\,L^j\right)e(t)+\sum_{k=1}^{r}\eta_k\,L^k x(t)\end{aligned} \tag{2.21}$$

avec $x(t)$ une variable exogène et η ses coefficients régressifs.

L'utilisation des modèles de Box-Jenkins appliqués à la prévision de la qualité de l'air date des années 70. McCollister et Wilson (1975) ont appliqué des SARIMA à la prévision d'ozone à $h+24$ et $j+1$. Leurs modèles n'utilisaient qu'un délai appliqué aux séries temporelles à prévoir ($p=1$, fixé après avoir testé d'autres modèles), et les résultats étaient meilleurs que ceux obtenus avec des modèles de persistance. Cette famille de modèles a par la suite été fréquemment utilisée en modélisation statistique de la qualité de l'air jusque dans les années 2000 (Wolff et Lioy, 1978; Robeson et Steyn, 1990; Shi et Harrison, 1997; Slini *et al.*, 2002; Wang et Lu, 2006; Goyal *et al.*, 2006). Ils ont l'avantage d'être relativement lisibles. Pour les plus simples d'entre eux, il est possible d'interpréter la valeur des coefficients régressifs, ce qui peut aider à valider un modèle, ou peut permettre de souligner l'importance de certaines variables dans les processus qui dirigent l'évolution de la qualité de l'air. Ils sont par contre incapables de modéliser les relations non-linéaires qui peuvent exister entre les variables prédictives et la variable objective. Milionis et Davies (1994) ont passé en revue l'utilisation des modèles régressifs et des modèles stochastiques pour la modélisation de la qualité de l'air. Ils constatent le défaut des modèles régressifs par rapport à la non-stationnarité des séries temporelles de concentration de polluants atmosphériques et suggèrent l'hybridation de ces modèles pour améliorer leurs performances.

Ainsi, pour correctement prévoir le comportement non-linéaire des séries temporelles de concentration en NO_2, Chelani et Devotta (2006) utilisent un ARIMA couplé avec un modèle non-linéaire. La combinaison des deux modèles fonctionne mieux que chaque modèle séparé.

La prévision des particules fines est plus difficile que celle de l'ozone, à cause du comportement complexe de ce polluant. Les études de prévision des particules sont apparues plus tard que celle des polluants gazeux. Certaines utilisaient également des modèles régressifs, mais systématiquement couplés à d'autres modèles. Díaz-Robles *et al.* (2008) ont utilisé un ARIMA hybridé avec un RNA avec de très bons résultats (mais dans une situation particulière pour laquelle les hautes valeurs de PM10 suivent un cycle journalier très précis).

Plus récemment, Poggi et Portier (2011) ont développé une méthodologie utilisant des modèles régressifs pour prévoir les concentrations en PM10 à partir des données d'Air Normand, l'AASQA de Normandie. Les modèles sont utilisés après une phase de clustering (partitionnement automatique par apprentissage non-supervisé) qui identifie plusieurs groupes dans les

données. En fonction de la probabilité d'appartenir à un groupe, l'un des modèles régressifs est utilisé. Les résultats obtenus sont bons et soulignent l'intérêt du clustering pour la modélisation statistique.

Les modèles statistiques à horizon ont de manière générale l'avantage (et l'inconvénient) d'être spécialisés localement car ils sont configurés à partir de données issues des mesures localisées à l'endroit des prévisions, alors que les CTM ont des domaines plus larges. Comparés à des modèles de dispersion, des modèles SARIMA se sont révélés plus précis pour prévoir les concentrations de polluants gazeux à des horizons courts (Polydoras *et al.*, 1998).

Les modèles régressifs sont donc adaptés à la modélisation de la qualité de l'air appliquée à la prévision. Ils ont été largement utilisés jusqu'aux années 2000, et restent encore d'actualité quand ils sont hybridés avec d'autres modèles. Ils ont cependant le défaut de nécessiter des pré-requis de stationnarité de la part des séries temporelles modélisées.

2.3.3 Arbres de décision

Les arbres de décision sont des modèles utilisés en prévision de séries temporelles et dans d'autres domaines comme l'aide à la décision ou le data mining (domaine de l'extraction de connaissance à partir de grandes quantités de données). Un arbre de décision consiste en une succession de choix menant vers d'autres choix jusqu'à des décisions finales. Le processus peut se visualiser sous la forme d'un arbre, chaque choix étant représenté par une séparation en plusieurs « branches », et chaque décision par une « feuille ». Un arbre de décision est un modèle particulièrement lisible et interprétable, sa représentation permet de suivre le cheminement des choix (voir figure 2.4). On parle d'arbre de classification lorsque la décision finale correspond à un choix entre plusieurs valeurs qualitatives ou semi-qualitatives, c'est-à-dire à l'assignation à une classe. Dans le cas d'une décision correspondant à l'adoption d'une valeur d'une variable continue, on parle d'arbre de régression. Un arbre de l'un de ces types est indistinctement désigné par le terme CART (Classification And Regression Tree).

Les CART sont capables de modéliser les liens non-linéaires entres leurs entrées et sorties à l'issue d'un apprentissage automatique, ce qui les avantage dans le domaine de la prévision de la qualité de l'air. Cependant, leurs performances ne surpassent pas nécessairement les modèles régressifs classiques, comme a pu le constater Ryan (1995) pour la prévision de la concentration maximale d'ozone.

Gardner et Dorling (2000) ont comparé modèles régressifs, CART et RNA pour la prévision d'ozone. De nouveau, les CART sont surpassés, cette fois par les modèles neuronaux, mais ont l'avantage de la lisibilité par rapport à ces derniers. Cet avantage ne doit pas être sous estimé quand il s'agit de prévision. Même si le résultat est moins précis, l'analyse de l'arbre permet en effet aux auteurs de remonter aux mécanismes physiques derrière l'évolution des concentrations d'ozone.

Plus tard, c'est à la prévision des PM10 que sont appliqués les CART par Slini *et al.* (2006) avec des résultats prometteurs et comparables à ceux obtenus par des RNA.

Il existe une famille de modèles basés sur des arbres qui utilisent le principe de la modélisation d'ensemble : les Forêts Aléatoires (FA, random forests en Anglais). Les modèles FA ont été introduits par Breiman (2001*a*). Ce modèle est constitué d'un ensemble d'arbres de décision, dont le résultat moyen constitue la sortie de modèle. Chaque « arbre » de la « forêt » est entrainé avec un jeu de données d'apprentissage différent. Chacun de ces jeux est un sous-échantillon de l'ensemble des données d'apprentissage, constitué aléatoirement en utilisant la méthode de « bagging » (terme qui provient de la contraction de « bootstrap aggregating »).

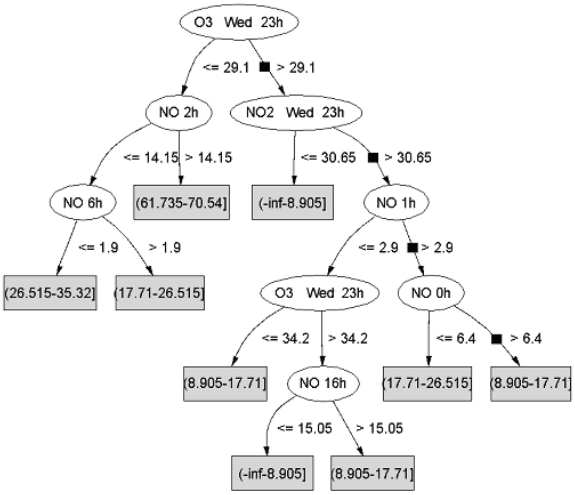


FIGURE 2.4 : Détail de la représentation d'un arbre de classification modélisant la concentration en NO à partir de variables exogènes provenant de Juhos *et al.* (2003).

Le bagging (Breiman, 1996) consiste à entrainer les différents modèles sur des échantillons bootstraps et à moyenner leurs résultats, ce qui permet d'augmenter la précision. Un désavantage de l'utilisation du bagging avec des arbres de décision est la perte de la lisibilité du modèle propre aux arbres. Chaque arbre de la forêt utilise des variables différentes, choisies aléatoirement. L'importance de chaque variable pour la prévision peut donc être quantifiée *a posteriori*. Il est possible d'utiliser des variables dépourvues de lien statistique avec la cible, leur importance dans la prévision sera amoindrie par le modèle.

Ces FA fonctionnent de manière opérationnelle pour la prévision de la qualité de l'air effectuée par Air PACA. Il seront utilisés à la section 5.4 (page 5.4) dans une étude comparative basées sur les données de PACA (Provence-Alpes-Côte d'Azur).

2.3.4 Réseaux de neurones artificiels

Les RNA sont une famille de modèles statistiques inspirés du fonctionnement des neurones biologiques et classés dans l'IA. Leur fonctionnement est plus largement détaillé au chapitre 4 (page 84). Ils ont plusieurs applications, comme le clustering, la classification (séparation automatique des données par apprentissage supervisé) ou la prévision de séries temporelles. Cette dernière application concerne de nombreux domaines des sciences humaines et sociales comme des sciences dites « exactes ».

Les RNA sont formés de neurones artificiels interconnectés. Chaque neurone reçoit des données, les traite et fournit une sortie, qui devient l'entrée des neurones suivants, jusqu'aux neurones de sortie du modèle. A partir du modèle de neurone formel proposé par McCulloch et Pitts (1943), les premiers modèles numériques n'ont d'abord comporté qu'un seul neurone.

Les neurones possèdent plusieurs entrées, sont dotés de paramètres et d'une fonction de transfert. Leurs paramètres sont appelés poids et biais (voir figure 2.5). Les poids sont des coefficients spécifiques à chaque entrée du neurone, par lesquels sont multiplié les données d'entrée. La somme de toutes les entrées ainsi pondérées est calculée, et on lui ajoute le biais du neurone. Cette valeur devient alors l'argument de la fonction de transfert du neurone, dont le résultat constitue la sortie du neurone. L'apprentissage automatique est la phase clé de la configuration d'un RNA lors de laquelle sont fixés les poids et biais du réseau, afin de s'adapter aux données

qui doivent être modélisées. Cet apprentissage, ou entrainement, est assuré par un algorithme d'apprentissage, le véritable outil de la modélisation neuronale, qui s'inscrit dans la discipline du machine learning (le domaine de l'apprentissage automatique) et de l'IA.

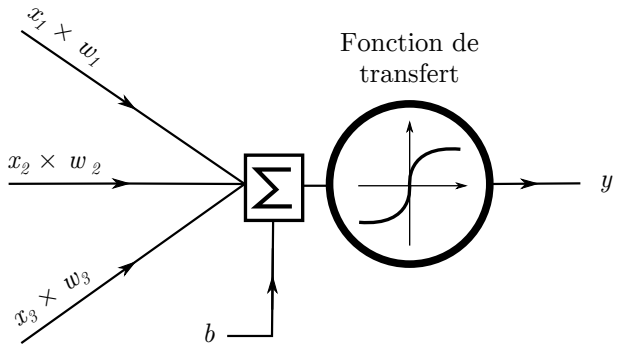


FIGURE 2.5 : Représentation d'un neurone formel recevant trois entrées x_i et produisant une sortie y avec des poids w_i, un biais b et une fonction de transfert f tels que $y = f(b + \sum_{i=1}^{3} w_i x_i)$.

Il existe plusieurs types de RNA, adaptés à différents problèmes. Leurs différences sont dues à la nature de leurs neurones, ainsi qu'à la manière dont ils sont interconnectés. Parmi ces architectures, le Perceptron proposé par Rosenblatt (1958) ne comporte qu'un neurone. Il permet une amélioration par rapport aux modèles linéaires. Le fait d'utiliser une fonction de transfert logique (fonction de Heaviside) donne au Perceptron les capacités d'un classifieur linéaire.

Le PMC (en anglais « MultiLayer Perceptron », ou MLP) permet, grâce à l'ajout de couches supplémentaires de neurones, de modéliser des fonctions non-linéaires. Entre l'entrée du réseau et les neurones de sortie, une ou plusieurs couches de neurones, qu'on appelle couches cachées, s'interposent (voir figure 2.6). Les capacités du PMC relancent l'intérêt de la communauté scientifique vis-à-vis des RNA. Son apprentissage est possible grâce à l'algorithme de rétropropagation de l'erreur, Back Propagation (BP) en anglais (Rumelhart *et al.*, 1986). D'autres algorithmes sont utilisés pour l'entrainer, comme l'algorithme de Broyden – Fletcher – Goldfarb – Shanno (BFGS) des noms de ses auteurs (voir Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), l'algorithme SCG (Scaled Conjugate Gradient) de Møller (1993), ou l'algorithme de Levenberg – Marquardt (LM) découvert par Levenberg (1944) puis indépendamment développé par Marquardt (1963). On trouvera plus de détails sur les algorithmes d'apprentissage à la section 4.3 page 90.

Le PMC présenté sur la figure 2.6 possède une couche cachée et une couche de sortie. Il possède trois variables d'entrée. Certains auteurs interprètent les variables d'entrée comme étant des « neurones d'entrée ». Le réseau présenté ici aurait ainsi donc trois couches de neurones. Cette appellation nous semble imprécise et nous distinguerons clairement les variables d'entrée et de sortie, des neurones, unités de calcul possédant des paramètres et une fonction de transfert. Le nombre de neurones de la couche de sortie définit le nombre de variables de sortie du modèle, ici trois.

Les réseaux récurrents sont un autre type important de RNA. Ils peuvent être semblables au PMC mais ont la particularité d'avoir des neurones qui propagent leurs sorties vers leurs propres entrées, ou vers celles de couches précédentes. Dans le cas de la modélisation d'une série temporelle, ce bouclage permet à un neurone de traiter en même temps l'information de deux pas de temps successifs. Cela induit une capacité de mémorisation pour le neurone, puisqu'il a encore accès à une information du temps précédant lorsqu'il traite les données du temps présent. Les neurones ont ainsi la possibilité de travailler directement sur les aspects temporels des phénomènes à modéliser. Ces réseaux, qu'on appelle réseaux bouclés, ou réseaux récursifs,

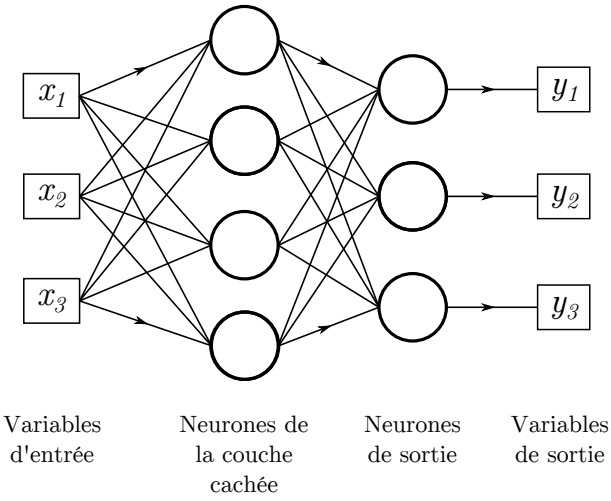


FIGURE 2.6 : Représentation d'un PMC à une couche cachée.

ou encore réseaux récurrents, sont particulièrement associés à la prévision de séries temporelles. Les réseaux non-bouclés sont appelés « feedforward » (« alimentés vers l'avant »).

Cependant, il est possible pour un RNA « feedforward » comme le PMC d'appréhender les phénomènes temporels à la manière des réseaux récurrents, et ce par un choix de délai (ou « lag » en anglais) à appliquer aux variables d'entrées. Il suffit en effet de fournir une même série temporelle à laquelle on applique différents délais au PMC pour que ses neurones aient également directement accès aux informations de différents pas de temps. Plusieurs séries temporelles décalées dans le temps sont ainsi formées à partir de la variable et constituent autant d'entrées du réseau. Certains réseaux utilisent en entrée uniquement différents délais de la variable endogène, à la manière des modèles AR. On les nomme Time-Delay Neural Network (TDNN). Une approche multivariée consiste à utiliser avec un PMC des variables endogènes et exogènes avec différents délais, choisis selon des méthodes de sélection de variables (voir section 5.1, page 113).

On peut noter également l'existence des réseaux « en cascade » dans le domaine de la prédiction, introduits par Fahlman et Lebiere (1990). Le principe est d'utiliser plusieurs RNA, chacun entrainé pour effectuer une prévision au même horizon (par exemple $h+1$). Seulement, on les fait fonctionner en cascade, la prévision du premier réseau étant utilisée en tant que données d'entrée par le second, et ainsi jusqu'au dernier. L'architecture peut être simplifiée à partir de ce principe pour ne former qu'un seul réseau de neurones à plusieurs couches cachées, recevant toutes des données d'entrée en plus des sorties des couches cachées précédentes.

Le lecteur intéressé par la famille des modèles neuronaux et leurs applications pourra lire utilement et en français le livre de Dreyfus (2004). Concernant l'apprentissage automatique, on note la présence de l'ouvrage de Cornuéjols *et al.* (2002), toujours en français. Tous les aspects des RNA utiles à nos travaux seront plus largement développés dans les chapitres 4 et 5, où ils seront mis en valeurs par nos résultats.

Les réseaux de neurones sont inspirés de modèles biologiques et se rattachent au machine learning. En cela, ils sont très différents d'autres modèles statistiques comme les modèles de Box-Jenkins, qui reposent uniquement sur des propriétés statistiques des séries temporelles. Les différences dans leurs origines et leur fondement classent les premiers dans l'IA et les seconds dans la statistique. Leurs différences ont un sens, mais parfois plus épistémologique que pratique. Dans la pratique, ces modèles peuvent être très similaires selon leur configuration, ils peuvent utiliser les mêmes jeux de variables, parfois les mêmes algorithmes d'apprentissage. On peux par

exemple construire un RNA à fonction de transfert linéaire qui reproduise un ARIMA.

De manière générale, les réseaux neuronaux ne sont pas systématiquement préférés en prévision. La revue de Zhang *et al.* (1998) concerne l'utilisation des RNA en prévision dans tous les domaines. Un avantage des RNA est qu'ils sont applicables à tous les problèmes, sans besoin d'*a priori* sur les variables. Il est par contre fastidieux de les configurer correctement. Le nombre de neurones peut être trop élevé, ce qui risque de sur-spécialiser les modèles sur leurs données d'apprentissage, les rendant incapables ensuite de correctement prédire avec de nouvelles données. Ce sur-apprentissage, qu'on appelle l'« overlearning », doit être géré.

En 1989, Hornik *et al.* apportent la preuve que les PMC sont capables de modéliser avec précision toute fonction, à condition qu'elle soit « lisse », c'est-à-dire infiniment dérivable. Cette propriété renforce l'intérêt du PMC et son utilisation en prévision se généralise. On a là un outil qui se différencie clairement de ses concurrents comme les modèles de Box-Jenkins, qui en théorie ne sont capables de modéliser que des séries temporelles parfaitement stationnaires (même si certaines versions intègrent la suppression de la tendance et de la saisonnalité). Le travail sur les séries temporelles peut sembler moins lourd avec les PMC, mais d'autres problèmes leur sont propres. L'apprentissage est délicat, puisque les valeurs des paramètres du réseau doivent être optimisées pour la prévision tout en évitant le sur-apprentissage. Cette problématique sera abordée en détail à la section 4.3, page 90. Le prétraitement des données facilite cet apprentissage. C'est une étape nécessaire que nous verrons à la section 4.4, page 98.

Avec les RNA, il est important de respecter le principe de parcimonie. C'est un principe en modélisation statistique qui veut que le meilleur modèle soit celui qui comporte le moins de paramètres possibles. Un modèle parcimonieux a de meilleures capacité de généralisation, alors qu'un trop grand nombre de paramètres a tendance à sur-spécialiser les modèles. Pour cela, il peut être nécessaire de réduire le nombre de variables d'entrée, en évitant d'utiliser des variables dont l'information est redondante, ce qui peut réduire les performances du modèle. Il est donc nécessaire d'adopter une démarche de sélection de variable efficace, ce que nous examinerons à la section 5.1, page 113. D'autres méthodes, comme les techniques de pruning (élagage), peuvent aider à rester parcimonieux, en supprimant les paramètres du RNA les moins utiles. On abordera le pruning à la section 5.2, page 127.

La précision des modèles statistiques est mesurée en comparant les prévisions aux mesures qui ont ensuite été effectuées. Cette démarche est importante et permet de statuer sur la qualité des modèles. Nous allons donc la présenter, avant de passer en revue l'état de l'art de la prévision de la qualité de l'air à l'aide de RNA à la section 2.5 page 46.

2.4 Evaluation

L'évaluation a pour but de mesurer la capacité d'un modèle à effectuer de bonnes prévisions, de manière objective. Cette tâche est délicate, car la définition d'une « bonne » capacité de prévision n'existe pas. En effet, l'erreur qu'un modèle commet n'est pas la même à chacune de ses prévisions, mais on peut étudier son comportement moyen, avec des approches variables.

Avec des modèles à apprentissage automatique, pour lesquels les relations entre entrées et sorties sont difficilement interprétables, une approche de propagation des incertitudes à partir des erreurs de mesure n'est pas adaptée. Pour évaluer les modèles, on compare point par point leur résultats avec la variable mesurée qu'ils sont censés modéliser.

Nous définirons plusieurs indices ou critères de précision, qui sont autant d'outils différents permettant de quantifier la précision des modèles et qui sont largement utilisés en prévision sta-

tistique. L'usage de ces indices est important pour pouvoir comparer les travaux de différentes équipes. Nous nous appuierons également sur des analyses de sensibilité qui évalueront les capacités des modèles à correctement prévoir les pics de pollution. Des outils comme les courbe ROC (pour Receiver Operating Characteristic) sont moins utilisés dans la littérature que les indices classiques mais sont particulièrement adaptés à notre problématique. Nous présenterons également les représentations graphiques des observations et des prévisions que nous utilisons afin de visualiser la justesse des modèles.

Tous ces indices et ces représentations sont générés à partir de données n'ayant pas servi pour l'apprentissage du modèle. Un tel « test à l'aveugle » garantit donc une évaluation correcte, qui prend en compte les capacités de généralisation du modèle. Cela nous place dans les conditions rencontrées lors d'une utilisation opérationnelle, où les données utilisées sont nouvelles.

Nous commencerons par passer en revue les différents scores d'erreur qui sont utilisés, avant de présenter les représentations graphiques qu'on peut utiliser.

2.4.1 Indices d'erreur

Un indice d'erreur sert à quantifier les différences qui existent entre les mesures de la variable à prévoir et sa prévision par le modèle. Concrètement, il s'agit de comparer deux séries temporelles, la série observée y et celle prédite $\hat{y}$. L'« observée » étant issue d'une mesure effectuée en station, elle est elle-même entachée d'erreur. L'évaluation de nos modèles ne portera pas exactement sur les différences entre leurs prévisions et la réalité, mais avec un aperçu de cette réalité. Etant donné que les modèles sont entrainés à prévoir une série temporelle mesurée, il est plus juste de les considérer comme des modèles de prévision de la mesure de la qualité de l'air plutôt que comme des modèles de prévision de la qualité de l'air.

Il existe plusieurs indices utilisés pour quantifier les différences entre l' « observé » et le « prédit ». Certains ne mesurent que l'erreur systématique (biais) commis par le modèle, d'autre intègrent également l'erreur aléatoire (bruit).

La Mean Bias Error (MBE) représente le biais moyen des prévisions :

$$MBE = \frac{1}{n}\sum_{i}^{n}(\hat{y}(i) - y(i)) \tag{2.22}$$

avec $\hat{y}$ la variable prédite, en sortie du modèle, y la variable observée, cible du modèle, n le nombre de ces prévisions et observations et i leur indice.

Les prévisions vont tantôt sous-estimer les observations, tantôt les surestimer. La MBE ne permet donc pas d'évaluer la précision d'un modèle car elle ne mesure que l'erreur systématique et les erreurs peuvent se compenser. On l'utilise pour se rendre compte à quel point un modèle a tendance à sur ou sous-estimer les valeurs.

La Mean Absolute Error (MAE) utilise une valeur absolue de l'écart entre l'observé et le prédit, mesurant ainsi l'erreur systématique et l'erreur aléatoire. Elle peut donc être utilisée pour évaluer la précision d'un modèle. On a :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}(i) - y(i)| \tag{2.23}$$

L'erreur quadratique moyenne (MSE pour Mean Squared Error) utilise le carré de l'écart entre observé et prédit plutôt que la valeurs absolue. L'indice pénalise donc plus les gros écarts

entre observé et prédit, ce qui est préférable en prévision de la qualité de l'air puisque les pics de pollution sont particulièrement importants à prévoir. On a :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}(i) - y(i))^2 \tag{2.24}$$

La Mean Squared Error (MSE) est l'erreur qui est minimisée lors de l'apprentissage automatique par les algorithmes que nous avons utilisés. Cet indice est donc important car il montre la performance de cet apprentissage, même si d'autres indices peuvent être préférables pour juger les modèles. L'utilisation de la MAE est également possible pour l'apprentissage, mais l'usage de la MSE a l'avantage d'être plus sensible à l'erreur sur les valeurs extrêmes, que l'on cherche à prévoir le mieux possible. On préfère présenter des résultats en donnant sa racine carrée, la Root Mean Squared Error (RMSE), qui contient la même information mais a la même dimension que la variable prédite.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}(i) - y(i))^2} \tag{2.25}$$

La RMSE est peut être l'indice le plus apprécié, à la fois pour sa lisibilité et pour sa bonne estimation de la précision. La normalized Root Mean Squared Error (nRMSE) est une version normalisée de la RMSE. Cette normalisation permet d'exprimer l'erreur en pourcentage, mais peut se calculer de différentes manières (pas toujours précisées par les auteurs). On peut par exemple diviser la RMSE par la valeur moyenne de la variable observée, par son écart-type ou par l'écart entre sa valeur minimale et sa valeur maximale. Nous adopterons la normalisation par la moyenne afin de rapporter l'erreur à la situation « normale » :

$$nRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}(i) - y(i))^2}}{\bar{y}} = \frac{RMSE}{\bar{y}} \tag{2.26}$$

avec $\bar{y}$ la moyenne algébrique de y.

La Mean Absolute Percentage Error (MAPE) est proche de la MAE, mais chaque écart entre l'observé et le prédit est divisé par l'observé, afin de ne considérer que des écarts relatifs. On a :

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}(i) - y(i)}{y(i)}\right| \tag{2.27}$$

Cet indice présente par contre le sérieux désavantage d'être instable quand $y(i)$ s'approche de zéro, et de ne pas être défini pour cette valeur.

Le Fractional Bias (FB) représente de manière adimensionnée le biais du modèle.

$$FB = 2\,\frac{\bar{y} - \bar{\hat{y}}}{\bar{y} + \bar{\hat{y}}} \tag{2.28}$$

Cet indice permet de représenter le biais de manière normalisée et bornée entre -2 et 2. Une valeur négative indique une surestimation lors de la prévision, et une valeur positive une sous-estimation, tandis que l'absence d'erreur systématique donnera une valeur nulle. La Fractional Variance (FV) est calculée de manière similaire pour représenter la différence entre les variances de l'observé et du prédit :

$$FV = 2\,\frac{\sigma_y^2 - \sigma_{\hat{y}}^2}{\sigma_y^2 + \sigma_{\hat{y}}^2} \tag{2.29}$$

avec σ_y^2 la variance de la variable observée et $\sigma_{\hat{y}}^2$ la variance de la variable prédite. La FV est également bornée entre -2 et 2. Ses valeurs négatives indiquent une plus forte variance de la variable prédite et inversement, une valeur nulle démontrant une conservation de la variance.

Le coefficient de corrélation, ou coefficient de Pearson (noté R) est parfois utilisé pour évaluer les modèles. Il s'exprime ainsi :

$$R = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.30}$$

R varie entre 1 pour des variables parfaitement corrélées et -1 pour des variables parfaitement anti-corrélées. On utilise également sa valeur au carré, R^2 variant entre 0 et 1. Une valeur nulle de R indique des variables totalement décorrélées. Mais la corrélation entre $\hat{y}$ et y n'indique pas forcement la précision du modèle mais uniquement une relation linéaire entre les deux variables. R n'est donc pas préféré pour mener une évaluation, mais beaucoup d'études se basent tout de même sur cet indice.

L'évaluation peut se faire avec l'indice d'agrément (« index of agreement » en anglais), indice proposé par Willmott (1982) et noté d.

$$d = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} \tag{2.31}$$

L'indice d'agrément d a l'avantage d'être normalisé et borné entre 0 pour la pire prévision et 1 pour une prévision parfaite, d'où une certaine lisibilité et une possibilité de comparer des modèles prédictifs appliqués à différents polluants. Le dénominateur représente une erreur potentielle et est utilisé pour normaliser le numérateur qui correspond à une erreur quadratique. Cette fraction est nulle pour une prévision parfaite et se rapproche de 1 quand la précision diminue. Une telle normalisation a plus de sens que celles qui peuvent être utilisées pour le calcul de la nRMSE. En effet, la nRMSE devient instable quand le dénominateur est proche de 0, et n'est pas bornée. Cet indice est de plus en plus utilisé depuis qu'il a été proposé, remplaçant petit à petit R dans un premier temps, puis finalement la RMSE qui s'exprime dans l'unité de la variable y.

Quand il s'agira de comparer des modèles appliqués à des problèmes différents, comme par exemple des modèles de prévision de polluants différents, nous choisirons d'utiliser d. La RMSE et la MAE seront également favorisées pour leur lisibilité. Cependant, il est important de systématiquement fournir tous ces indices, afin de permettre une comparaison, même imparfaite, avec d'autres études utilisant des métriques différentes. Ils seront donc consignés quand l'expérience présente un intérêt majeur.

On verra que quand on s'intéresse à une configuration de PMC, il est nécessaire de l'expérimenter plusieurs fois de suite, chaque entrainement et évaluation donnant des résultats légèrement différents. En effet, l'initialisation des paramètres des PMC comporte souvent une part d'aléatoire (systématiquement avec l'approche que nous avons utilisé, voir section 4.3.2 page 95), ce qui explique des variations de scores pour une même configuration. Chacune sera donc testée plusieurs fois (une dizaine de fois, sauf indications autres), et les scores seront présentés sous forme de « boîtes à moustaches » (exemple en figure 2.7).

Pour décrire ces graphiques, rappelons la notion de « quantile » d'un échantillon d'une variable aléatoire. Les quantiles sont les valeurs qui permettent de diviser un échantillon ordonné en plusieurs échantillons de même taille. L'échantillon étant ordonné, la valeur des quantiles permet de se rendre compte de la rareté relative des valeurs que prend la variable. Si l'on divise l'échantillon en quatre, les quantiles sont appelés des quartiles. S'il est divisé en cent, on parle

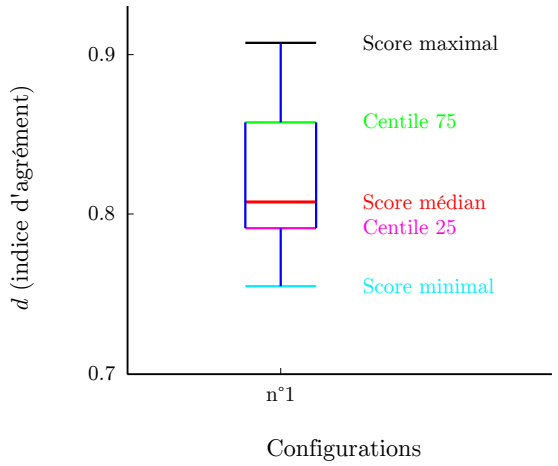


FIGURE 2.7 : Exemple de boîte à moustache indiquant l'indice d'agrément obtenu pour une configuration de modèle prévisionnel.

de centiles. Ainsi, un échantillon peut être divisé en quatre. Les trois valeurs qui permettent cette séparation sont appelés quartiles. Le premier indique la valeurs en dessous de laquelle on trouve 25 % des valeurs ; il est équivalent au 25$^{\text{ème}}$ centile. Le second quartile correspond à la médiane de l'échantillon, et le troisième équivaut au 75$^{\text{ème}}$ centile.

Sur nos boîtes à moustaches, les extrémités de la boîte représentent les premier et troisième quartiles du score présenté, et la médiane est représentée à l'intérieur de la boîte. Les deux branches indiquent les valeurs maximale et minimale obtenues.

2.4.2 Représentations graphiques des prévisions

Une représentation graphique des prévisions des modèles et des valeurs observées est une bonne manière de se rendre compte du comportement du modèle. La figure 2.8 montre ce type de représentation qu'on qualifiera de « courbe observé/prédit ». Le principal intérêt de ce type de représentation est sa lisibilité, puisqu'elle permet de voir clairement la manière dont chaque point mesuré est prédit.

Le principal problème de cette représentation est qu'on ne peut représenter les séries temporelles que pour un certain nombre de points consécutifs (quelques jours pour des séries temporelles horaires par exemple), sans quoi la figure devient illisible. Cette représentation est judicieuse pour illustrer un exemple particulier, mais il est délicat de s'en servir pour juger les capacités d'un modèle. On l'utilisera donc pour illustrer une discussion autour d'un modèle.

Une autre représentation très utilisée est le nuage de points, « scatter plot » en anglais. Il s'agit de représenter chaque prévision en fonction de la valeur observée, formant ainsi un nuage de points. On peut alors se rendre compte de plusieurs choses.

Tout d'abord, on trace habituellement la droite correspondant à observé = prédit. Un modèle parfait ne produit que des points sur cette droite, et on peut se rendre compte visuellement de l'écart moyen entre prévision et mesure, et ce pour toutes les concentrations. On peut également se rendre compte du biais du modèle, en se représentant le centre de gravité des points et en le comparant avec l'axe observé = prédit.

On peut choisir un seuil que l'on représente par des droites horizontale et verticale, qui divise le graphique en quatre. Cela aide à se rendre compte de la manière dont les dépassements de ce seuil sont prévus. Nous tracerons sur nos nuages de points de telles droites, et en l'absence de

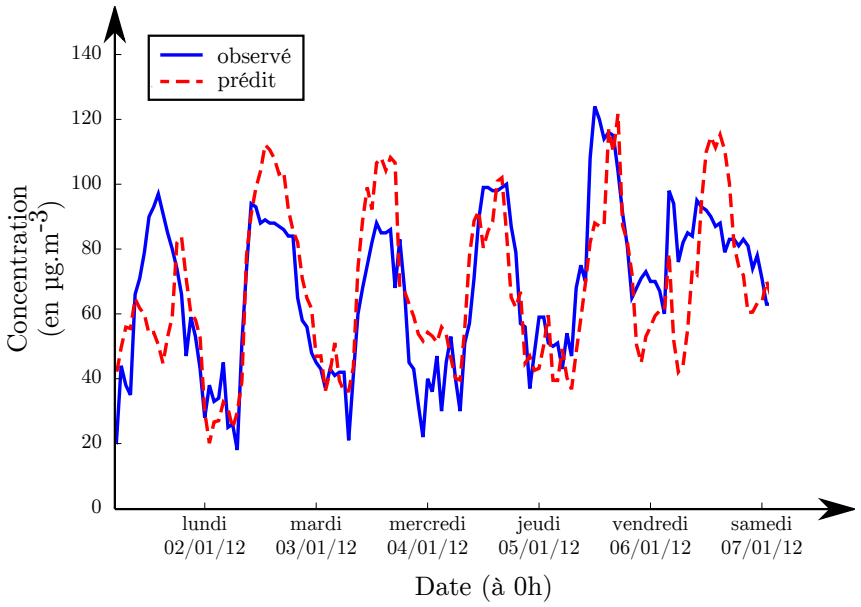


FIGURE 2.8 : Exemple de courbe « observé/prédit » d'un modèle de prévision d'O_3 à Canetto à l'horizon h + 24.

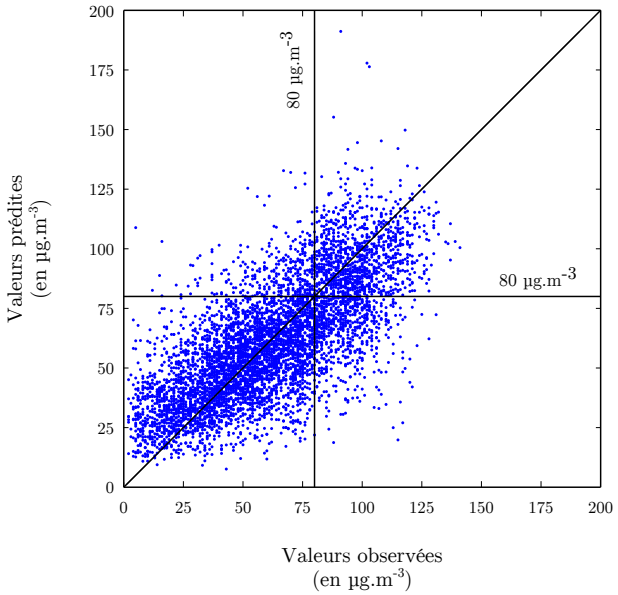


FIGURE 2.9 : Exemple de courbe de dispersion d'un modèle de prévision d'O_3 à Canetto à l'horizon h + 24, avec le centile 90 égal à 80 $\mu g.m^{-3}$ indiqué.

seuil spécifié sur les figures, ces droites correspondront au centile 90 (noté C90, 90% des données de l'échantillon sont inférieures à la valeur du C90) des concentrations observées. Cela permet de visualiser les points observés et prédits correspondant aux plus hautes valeurs rencontrées dans le jeu de données.

Lorsqu'on voudra évaluer en détail un modèle, nous réaliserons en plus une étude de sensibilité plus poussée, notamment grâce à l'usage de courbe ROC, que nous présentons maintenant.

2.4.3 Analyse de sensibilité

La prévision de la qualité de l'air est majoritairement utilisée de manière opérationnelle avec l'objectif de prévoir l'imminence de pics de pollution. Il est utile de considérer les modèles prévisionnels comme des classifieurs, prévoyant soit une situation normale, soit un dépassement de seuil. En plus de l'évaluation de la précision moyenne sur tout le jeu de test basée sur des scores comme la RMSE, il est utile de mener une étude de sensibilité afin de savoir si le modèle a de bonnes capacités de prévision de pics.

Pour cela, il est nécessaire d'adopter un seuil à partir duquel on interprète la prévision comme celle d'un dépassement, et en deçà duquel on considère qu'on est en situation normale. Il est par exemple possible d'adopter les seuils en vigueur en France. Une fois le seuil défini, on peut croiser les valeurs observées et les valeurs prédites pour comptabiliser les bonnes et mauvaises prévisions, qu'on peut par exemple représenter en traçant une matrice de contingence du modèle (exemple en figure 2.10).

		Observés négatifs	Observés positifs
		5975	652
Prédits positifs	670	Faux positifs 323 Erreur type I Erreur α	Vrais positifs 347
Prédits négatifs	5957	Vrais négatifs 5652	Faux négatifs 305 Erreur type II Erreur β

FIGURE 2.10 : Exemple de matrice de contingence pour l'évaluation d'un modèle de prévision des moyennes sur 24h glissantes de concentration de PM10 à Canetto, pour un seuil de 28 $\mu g.m^{-3}$.

Un « Vrai Négatif » (VN) correspond à la prévision juste d'une situation normale, et un « Vrai Positif » (VP) à la prévision juste d'un pic. Un « Faux Positif » (FP) correspond à une erreur de type I ou erreur de type α, c'est-à-dire à une fausse alerte, une prévision de pic qui n'a finalement pas eu lieu. Un « Faux Négatif » (FN) correspond à une erreur de type II ou erreur de type β, c'est à dire à un pic qui n'a pas été prévu.

A partir des informations représentées dans les matrices de contingence, on peut calculer plusieurs indices. On s'intéressera au Taux de Vrai Positif (TVP) (également appelé « sensibilité », ou « true positive rate » en anglais) qui correspond au rapport entre les vrais positifs et l'ensemble des valeurs observées positives, ainsi qu'au Taux de Faux Positif (FPR) (« false positive rate » en anglais), rapport entre les faux positifs et l'ensemble des valeurs observées négatives. On a :

$$TVP = \frac{VP}{VP + FN} \tag{2.32}$$

$$FPR = 1 - \frac{VN}{VN + FP} \tag{2.33}$$

Grâce à ces deux indices, on a d'importantes informations sur le comportement d'un modèle.

Le TVP donne la précision de détection des situations de dépassement de seuil, et le TFP celle des situations normales.

Le principal défaut de cette approche est qu'elle est dépendante d'un seuil de concentration, qui doit être fixé assez arbitrairement. En effet, les seuils légaux en vigueurs sont issus de compromis de la part des pouvoirs publics entre d'un côté les connaissances scientifiques sur les effets néfastes des polluants sur la santé humaine et de l'autre la réalité des concentrations observées et des contraintes industrielles. Adopter un seuil en particulier ne semble donc pas le meilleur moyen d'évaluer un modèle.

Les courbes ROC (présentées par Fawcett, 2006) peuvent être utilisées pour s'affranchir du choix d'un seuil. Ces courbes représentent le TVP d'un modèle en fonction de son TFP. Pour un seuil donné, cela correspond à un point. Mais pour tous les seuils possibles entre 0 et un seuil maximum, cela forme une courbe ROC.

Un exemple en est donné en figure 2.11. On y voit un nuage de points ainsi qu'une courbe ROC, qui correspondent à un modèle de prévision de concentration de PM10 à Canetto, celui dont la matrice de contingence est donnée en figure 2.10. Sur le nuage de points, les seuils indiqués représentent la valeur du C90 des concentrations observées, 28 $\mu g.m^{-3}$. Pour ce seuil, les informations données par la matrice de contingence (voir figure 2.10) sont reportées. Le nombre de 347 VP par exemple correspond au nombre de points dans le cadre en haut à droite, dans lequel les points observés et prédits dépassent touts les deux les 28 $\mu g.m^{-3}$. Un nuage de points peut donc être consulté comme une matrice de contingence.

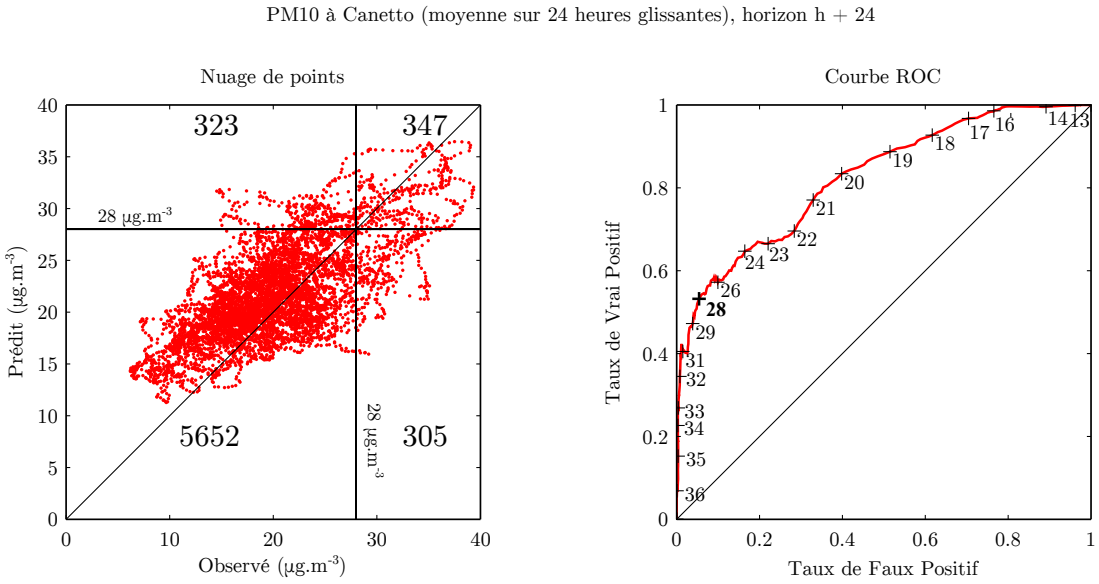


FIGURE 2.11 : Exemple d'un nuage de points et de la courbe ROC correspondants à un modèle de prévision des moyennes sur 24h glissantes de concentrations de PM10 à Canetto.

On remarque que le nuage de points dessine parfois des motifs de courbes. Leur présence est due à l'usage de moyennes sur 24 heures glissantes, avec lesquelles l'évolution des concentrations est lissée. Certains épisodes peuvent ainsi être identifiés sur le nuage de points.

A droite de la figure 2.11 est présentée la courbe ROC. Chaque seuil y est représenté, sauf quand cela rend l'affichage confus. Ainsi, si on s'arrête au point correspondant à un seuil de 28 $\mu g.m^{-3}$ (en gras sur la courbe), on retrouve un TVP de 0.53, correspondant aux 347 VP pour 305 FN que l'on retrouve sur le nuage de points. Le TFP de 0.05 correspond aux 5652 VN pour 323 FP.

Pour les plus petites valeurs de seuil, la courbe se rapproche du point TPR = 1 et FPR = 1. En effet, pour de trop petits seuils, on n'observe quasiment que des dépassements, correctement prédits. Et les rares valeurs observées sous le seuil ont peu de chance d'être prédites ainsi. Pour les seuils trop élevés par rapport aux valeurs observées présentes dans le jeu de test, c'est l'inverse. Quasiment toutes les valeurs observées sont sous le seuil et on a peu de fausses alertes, d'où un TFP tendant vers 0. Il devient alors difficile au modèle de correctement prévoir les rares dépassements, d'où le TVP qui tend également vers 0. Un tel profil est typique des courbes ROC. La droite noire correspondant à FPR = TPR indique la courbe qu'aurait un modèle aux prévisions aléatoires. La courbe d'un bon modèle doit se trouver bien au dessus de cette droite. Un modèle parfait n'aurait que des points en haut à gauche du diagramme, tandis que si un modèle fournit un point sous la droite, c'est qu'il vaut mieux faire confiance à l'inverse de ses prévisions.

La courbe ROC d'un bon modèle prendra donc la forme d'un « arc » longeant fortement le côté gauche et le haut du cadre. L'aire sous la courbe ROC sera donc maximale pour un bon modèle. Cette aire sous la courbe est parfois utilisée pour évaluer les modèles, ce qui donne une information de précision globale, indépendante de tout seuil.

Nous préférerons nous concentrer sur les seuils indiqués sur la courbe. Leur position indique la capacité du modèle à prévoir leurs dépassements. Pour la précision globale du modèle, nous préférerons les indices d'erreur présentés précédemment à la section 2.4.1.

2.5 Etat de l'art sur la prévision de la qualité de l'air avec des RNA

Arrivés dans le domaine de la prévision de la qualité de l'air après les modèles de Box-Jenkins, ils ont été comparés à ces derniers (Yi et Prybutok, 1996; Comrie, 1997). Les auteurs notent que les modèles neuronaux utilisés n'ont pas besoin de pré-requis vis-à-vis des données utilisées. Ils obtiennent de meilleures performances avec les RNA pour la prévision de concentration d'ozone.

L'utilisation des RNA en prévision de la qualité de l'air a été passée en revue par Gardner et Dorling (1998) au travers de l'usage du PMC. A partir de cette époque, le PMC sera le principal type de modèle neuronal utilisé en prévision de concentrations en polluants atmosphériques. Le tableau 2.1 synthétise les principales applications de réseaux neuronaux à la prévision de la qualité de l'air depuis le début des années 2000, par ordre d'années de publication.

TABLEAU 2.1 : Revue des études sur la prévision de qualité de l'air avec des modèles neuronaux.

Publication	Modèles prédictifs (algo. d'apprentissage)	Polluants prédits	Horizons de prédiction	Données d'entrée	Remarques et observations des auteurs
Gardner et Dorling (1999)	PMC(SCG)	NO_2, NO_x	h+0, h+1, h+24	endo, proj. met, IT	Le PMC admet la substitution de variables explicatives par d'autres
Gardner et Dorling (2000)	**PMC**(SCG), CART, LIN	O_3	h+0	endo, met, IT	Grâce à sa non-linéarité, le PMC est plus apte à modéliser O_3 que LIN
Kolehmainen *et al.* (2000)	PMC	PM10, NO_2, CO	j+1	endo, met, IT	Prévision des gaz plus efficace que celle des PM
Kao et Huang (2000)	**PMC**(BP), ARIMA	O_3, SO_2	h+1, h+24	endo	PMC meilleur qu'ARIMA pour le h+24, et de manière moins tranchée pour le h+1
Kolehmainen *et al.* (2001)	**PMC**(LM), SOM	NO_2	j+1	endo, met, IT	Difficultés à prévoir les valeurs extrêmes
Perez et Reyes (2001)	PMC(BP), **P**, pers	PM2.5	h+1~24	endo	Avec uniquement des lags de l'endogène en entrée, P surpasse pers jusqu'à h+6
Abdul-Wahab et Al-Alawi (2002)	PMC(BP)	O_3	h+0	poll, met	Les meilleurs variables d'entrée sont les précurseurs chimiques d'O_3
Perez et Reyes (2002)	**PMC**(Δrule), P	max24hma(PM10)	h+30	endo, pred. met	L'importance des variables utilisées est plus grande que celle du choix du modèle
Viotti *et al.* (2002)	PMC(BP)	O_3, NO_2, NO_x, CO, benzène	h+1, h+3, h+24, h+48	endo, met	Le choix des variables d'entrée doit être fonction des polluants cibles et de l'horizon
Juhos *et al.* (2003)	PMC(BP), CART	NO_2, NO	h+1~24	endo, poll, met	La lisibilité des CART utile, mais leurs classes peuvent couvrir de larges gammes de concentration
Kukkonen *et al.* (2003)	**PMC**(DG), LIN, det	PM10, NO_2	h+24	endo, met, trafic	La non-linéarité des RNA est nécessaire pour la prévision de polluants
Wang *et al.* (2003)	RBF, **SVM**	RSP	h+72	endo, poll, met	Avantage des SVM sur les RBF
Niska *et al.* (2004)	PMC(SCG)	NO_2	h+24	endo, poll, proj. met	Les AG sont utiles en sélection de variables, mais sont pénalisés par des temps de calcul très longs

Publication	Modèles prédictifs (algo. d'apprentissage)	Polluants prédits	Horizons de prédiction	Données d'entrée	Remarques et observations des auteurs
Jiang *et al.* (2004)	PMC(BP)	index	j+1	pred. met, IT	Plus efficace de prévoir les concentrations pour calculer l'index que de le prévoir directement. Une couche cachée suffit
Ordieres *et al.* (2005)	PMC, **RBF**, LIN, pers	PM2.5	j+0 (à 8h)	endo, met	Le RBF surpasse le PMC
Hooyberghs *et al.* (2005)	PMC(BP)	PM10	j+0~2	endo, pred. met, IT	Variables météo plus importantes que celles d'émissions. Parmi elles, HCL l'est particulièrement
Corani (2005)	**PMC**(LM), LL	PM10, max8hma(O_3)	j+0 (à 9h)	endo, poll, met	La désaisonnalisation des variables offre un gain de précision
Niska *et al.* (2005)	PMC(BP)	PM2.5, NO_2	h+24	endo, poll, pred. met, IT	L'utilisation de sorties de NWP en entrée améliore les prévisions. Les pics de pollution restent durs à prévoir
Agirre-Basurko *et al.* (2006)	**PMC**(BP), MLR, pers	O_3, NO_2	h+1~8	endo, poll, met, trafic	Le modèle le plus précis est PMC, sauf pour les horizons courts (h+2 et h+3)
Slini *et al.* (2006)	PMC(LM), **CART**, MLR	PM10	j+1	endo, met	Le meilleur modèle dépends de la métrique choisie pour l'évaluation
Perez et Reyes (2006)	**PMC**(BP), MLR, pers	max24hma(PM10)	j+1	endo, met, pred. met	Le choix qui revêt le plus d'importance est celui des variables d'entrée
Lu *et al.* (2006)	PMC(SCG)	max(O_3)	j+0	endo, met	Intérêt du clustering des données par SOM avant l'entrainement d'un PMC spécifique à chaque cluster
Sousa *et al.* (2007)	**PMC**, MLR	O_3	h+24	endo, poll, met	L'ACP en prétraitement augmente la précision des modèles et réduit le nombre de variable d'entrées
Brunelli *et al.* (2007)	PMC, **rPMC**	max(PM10, O_3, NO_2, SO_2, CO)	j+2	met	PMC et rPMC comparés uniquement pour la prévision de SO_2

Publication	Modèles prédictifs (algo. d'apprentissage)	Polluants prédits	Horizons de prédiction	Données d'entrée	Remarques et observations des auteurs
Dutot *et al.* (2007)	**PMC**(LM), MLR, det	max(O_3)	j+1	endo, pred. met	Le PMC et le CTM CHIMERE ont des résultats proches. Ceux de MLR sont également assez bons
Ibarra-Berastegi *et al.* (2008)	PMC, RBF, GRNN, MLR, pers	O_3, NO_2, NO, SO_2, CO	h+1~8	endo, poll, met	Aucun des modèles ne se démarque clairement
Gautam *et al.* (2008)	PMC(BP)	O_3	h+1	endo	Résultats améliorés par une recherche de poids par NN lors du test
Kurt *et al.* (2008)	PMC(BP)	PM10, SO_2, CO	j+1~3	met, IT	La prévision à j+1 répétée 3 fois en utilisant ses résultats fonctionne mieux que la prévision j+3 directe
Díaz-Robles *et al.* (2008)	PMC(LM), MLR, ARIMA	PM10	j+1	endo, met	Un modèle hybride ARIMA-PMC est meilleur que chaque modèle utilisé séparément
Perez et Salini (2008)	PMC(BP), MLR, HCA	max24hma(PM2.5)	j+1	endo, met, pred. met	Le modèle de clustering HCA n'est pas concerné par les mauvaises prévisions de valeurs extrêmes
Coman *et al.* (2008)	**PMC**(SCG, BFGS), cPMC	O_3, NO_2	h+24	endo, met, IT	Le modèle en cascade est (à peine) moins bon que le PMC
Cai *et al.* (2009)	**PMC**(BP), MLR, disp	PM10, O_3, NO, CO	h+14	endo, met, IT, trafic	Le PMC peut être plus précis qu'un modèle dispersif
Hrust *et al.* (2009)	PMC(BP)	PM10, O_3, NO_2, CO	h+1~17	endo, met, IT	La prévision est améliorée par l'utilisation de variables moyennées sur des pas de temps différents
Paschalidou *et al.* (2011)	**PMC**(BFGS), RBF, MLR	PM10	h+24	endo, met, IT	Le PMC a été capable de prévoir efficacement les épisodes de transport de poussières sahariennes à Chypre
Sfetsos et Vlachogiannis (2010)	PMC(LM), MLR	PM10	h+1~24, j+1	endo, met	Le calcul des valeurs journalières à partir de prévision horaires est meilleur que leur prévision directe

Publication	Modèles prédictifs (algo. d'apprentissage)	Polluants prédits	Horizons de prédiction	Données d'entrée	Remarques et observations des auteurs
Carnevale *et al.* (2011)	PMC(LM)	PM10, max(O_3)	j+1~3 à 12h	endo, pred. met, poll	La prévision localisée aux stations peut être efficacement extrapolée par krigeage pour obtenir des cartes de prévision
Zainuddin et Pauline (2011)	WTNN	NO_2, NO, NO_x	h+1	endo, poll, met (exclu)	La précision des WTNN est améliorée par une sélection des fonctions (wavelet) utilisées comme fonction d'activation
Shekarrizfard *et al.* (2012)	PMC(LM, SCG)	PM10	j+1	proj. met	Le débruitage des donnée par WT permet de rendre les prévisions plus robustes
Fernando *et al.* (2012)	**PMC**(SCG), det	PM10	h+24	endo, met, IT	Le PMC est plus précis que le CTM CMAQ
Perez (2012)	PMC, NN	max24hma(PM10)	j+1	endo, met, pred. met	La prise en compte des deux modèles permet une prévision plus efficace
de Mattos Neto *et al.* (2014)	PMC	PM10, PM2.5	j+1	endo	La prise en compte en post-traitement de la composante « marche aléatoire » des PM amélioire les prévisions
He *et al.* (2014)	PMC(BP)	PM10, PM1	prochain feu vert	endo, met, trafic	L'ACP en prétraitement améliore les prévisions, qui sont meilleures pour les PM1 que pour les PM10
Mishra et Goyal (2015)	**PMC**, MLR	NO_2	h+1	met, poll, IT	Le PMC profite de l'ACP en prétraitement et surpasse la MLR

Quand un modèle s'avère plus efficace que les autres selon les auteurs, il est écris **en gras**.

Abréviations - Modèles : P : Perceptron, SOM : Self-Organizing Map, HCA : Hybrid Clustering Algorithm, LIN : modèle linéaire, RBF : Radial Basis Function (RNA), WTNN : Wavelet Transform Neural Network, SVM : Support Vector Machines, rPMC : PMC récursif, cPMC : PMC en cascade, NN : Nearest Neighbors, LL : Lazy Learning (modèle linéaire), pers : modèles de persistance, det : modèle déterministe (CTM), disp : modèles de dispersion

Algorithmes d'apprentissages : BP : BackPropagation (rétropropagation), SCG : Scaled Conjugate Gradient, LM : Levenberg-Marquardt, DG : Descente de Gradient, BFGS : Broyden–Fletcher–Goldfarb–Shanno

Horizons : h indique la prévision d'une quantité horaire, j celle d'une quantité journalière. x~y indique tous les horizons de x à y. Les prévision à $j+0$ sont généralement effectuées le matin pour la journée même, l'heure à laquelle elle sont faite peut être indiqué entre parenthèses. ppfv indique uneprévision

pour la prochaine période de feu vert à un feu tricolore

Polluants prédits : max indique le maximum des concentrations, $x$hma indique une moyenne glissante de x heures. index indique une prévision d'un index de la qualité de l'air plutôt que d'un polluant. RSP : Particules respirables.

Variables d'entrée : endo : variable endogène, poll : concentrations de polluants, met : mesures météorologiques, pred. met : prévision météorologiques, proj. met : mesures météorologiques utilisées à l'avance, comme des prévisions, IT : Indices Temporels, trafic : données décrivant le trafic routier, HCL : Hauteur de la Couche Limite atmosphérique.

Autres : ACP : Analyse en Composante Principale, AG : Algorithmes Génétiques, WT : Wavelet Transform, NWP : Numerical Weather Prediction.

Plusieurs enseignements ressortent clairement de l'ensemble de ces études, dont nous allons discuter. Remarquons tout d'abord qu'il est très délicat de comparer des scores de modèles prévisionnels provenant de plusieurs études. Il est nécessaire que les horizons de prévision soient les mêmes, ainsi que les polluants étudiés et leur résolution temporelle (données horaires, journalières, etc.). Certains auteurs prévoient des moyennes des concentrations, d'autres leurs valeurs maximales ou des indices de qualité de l'air calculés à partir de ces concentrations.

De plus, les métriques utilisées pour évaluer la précision des modèles varient. Si l'on retrouve souvent certains indices (coefficient de corrélation, erreur quadratique moyenne...) vus à la section 2.4 (page 38), ils ne sont pas toujours adaptés à la comparaison de performances, et tous ne sont pas toujours utilisés par les auteurs. De plus, la période de temps consacrée à l'évaluation des modèles, c'est-à-dire le jeu de test, est loin d'être similaire entre plusieurs études. Certaines équipes évaluent leurs modèles sur une ou plusieurs années, d'autres sur à peine quelques semaines ce qui implique un biais saisonnal. D'autres constituent des jeux de test aléatoirement ou font de la validation croisée. Ces raisons nous ont conduit à ne pas indiquer dans le tableau 2.1 les indices d'erreur obtenus, ce qui pour nous n'aurait pas eu de sens dans ces conditions. Nous préférerons comparer les performances des différents modèles étudiés au sein d'une même étude.

La première conclusion qui est partagée par ces études est que les RNA paraissent tout à fait adaptés à la prévision de la qualité de l'air à moyen terme (entre $h+6$ et $j+3$). En dessous de $h+6$, ils peuvent perdre en précision face à des modèles plus simples comme des modèles naïfs ou linéaires, ou alors ne plus offrir de gain de précision tout en étant plus complexes. Mais ces horizons n'ont pas vraiment d'intérêt pour la prévision opérationnelle de la qualité de l'air.

Quand on compare les PMC aux modèles déterministes pour la prévision d'ozone ou de particules, ils sont effectivement plus précis (Dutot *et al.*, 2007; Fernando *et al.*, 2012), même s'ils ne fournissent que des prévisions localisées aux stations de mesure. Le PMC apparait également de manière générale plus approprié à la prévision des polluants que les modèles linéaires (Gardner et Dorling, 2000; Kao et Huang, 2000; Perez et Reyes, 2002; Kukkonen *et al.*, 2003; Corani, 2005; Agirre-Basurko *et al.*, 2006; Sousa *et al.*, 2007; Dutot *et al.*, 2007; Cai *et al.*, 2009; Paschalidou *et al.*, 2011; Mishra et Goyal, 2015).

Deuxièmement, l'élément le plus important quand on s'attaque à un problème de modélisation concerne clairement les données que l'on utilise. Les variables généralement employées sont des mesures (de pollution ou de variables météorologiques), des sorties de modèles prédictifs (NWP voire CTM), des indices temporels (qui servent à apporter aux modèles d'importantes informations sur le contexte de la prévision, les activités humaines et naturelles étant régulées par des cycles temporels précis, annuels, hebdomadaires, journaliers, etc.) et parfois des mesures relatives au trafic routier. Le choix des variables d'entrée apparait comme plus important que le choix du type de modèle. Les données définissent ensuite la manière dont on va construire le modèle. Le principe de parcimonie est rappelé par nombre d'auteurs comme étant incontournable.

Ainsi, le domaine de la sélection de variables est abordé par de nombreux auteurs. Ce domaine (« feature selection » en anglais) regroupe les méthodes qui permettent de ne retenir d'un jeu de variables que celles qui sont les plus utiles pour le modèle, optimisant ainsi ses performances. Des techniques apparaissent régulièrement comme l'utilisation de l'Analyse en Composantes Principales (ACP), de méthodes heuristiques comme les algorithmes génétiques (Niska *et al.*, 2004; Ibarra-Berastegi *et al.*, 2008), le recuit simulé (« simulated annealing » en anglais, Juhos *et al.*, 2003), l'utilisation de la notion d'Information Mutuelle (IM) entre les variables (Perez et Reyes, 2001), ou des critères bayésiens comme le BIC (Bayesian Information Criterion).

Le prétraitement des données d'entrée ne peut pas être négligé. De nombreuses transforma-

tions sont nécessaires pour que le modèle puisse comprendre et traiter le jeu de données, et des gains de précision remarquables sont obtenus lors d'études dédiées aux différents prétraitements (Corani, 2005; Hrust *et al.*, 2009; Shekarrizfard *et al.*, 2012).

Au final, le fait d'être capable d'améliorer les performances de ce type de modèle par la gestion des variables utilisées permet d'une certaine manière d'obtenir des informations sur le fonctionnement aérologique local. On peut ainsi confirmer l'importance de certains paramètres. On peut à l'inverse, grâce à des connaissances sur le fonctionnement atmosphérique, apporter une expertise lors de la sélection de variables et de la configuration du modèle, pour tenter de mieux prendre en compte un phénomène connu. Nos modèles « boîtes noires » deviennent, en quelque sorte, des « boîtes grises » puisqu'ils peuvent permettre d'avoir une réflexion sur les mécanismes de l'atmosphère.

L'hybridation des modèles semble être le troisième point d'importance. Plusieurs modèles peuvent être utilisés conjointement, comme par exemple des modèles prédictifs différents Díaz-Robles *et al.* (2008); Perez (2012). L'hybridation de modèle de clustering avec des modèles prédictifs semble également prometteuse (Kolehmainen *et al.*, 2000; Lu *et al.*, 2006; Poggi et Portier, 2011). Quand on utilise en entrée des données issues de modèles déterministes, on peut également voir cela comme une hybridation de modèle. Le « post-traitement » de modèle déterministe donne des résultats intéressants (Hooyberghs *et al.*, 2005; Niska *et al.*, 2005; Dutot *et al.*, 2007; Carnevale *et al.*, 2011 ; les différents modèles de Perez *et al.*).

D'autres améliorations de performances d'un PMC sont possibles, cette fois dans l'optimisation de la configuration du modèle. Par exemple, des techniques de pruning (Corani, 2005) permettent d'obtenir des modèles plus parcimonieux en supprimant les paramètres (voir des neurones ou même des variables d'entrées) les moins utiles. Des schémas de régulation bayésienne peuvent également être utilisés pour optimiser automatiquement ce nombre de paramètres (Lauret *et al.*, 2008).

De nombreux auteurs constatent par contre qu'il est difficile de prévoir les évènements extrêmes avec les modèles statistiques en général. Les algorithmes d'apprentissage sont faits pour minimiser l'erreur moyenne commise lors des prévisions. Les situations les plus courantes sont ainsi favorisées par rapport aux évènements rares de part leur abondance dans les jeux de données d'entrainement. L'utilisation d'indices d'erreur calculés sur tout le jeu de test pour évaluer des modèles peut être problématique car ces indices ne se focalisent pas spécialement sur la précision lors d'évènements de fortes concentrations, les plus importantes à prévoir (la problématique de l'évaluation a été abordée à la section 2.4, page 38).

Les études passées en revue font ressortir le fait qu'aucune méthode générique ne semble adaptée à tous les problèmes de prévision de qualité de l'air. Les différences entre les contextes d'études (nature, origine géographique et qualité des données, horizons de prévision) impliquent un travail approfondi de configuration des modèles préalable à toute expérimentation avec ces méthodes. A partir d'un modèle de RNA (souvent de PMC) qui sert de base à la prévision, les auteurs apportent tous des modifications, des enrichissements différents à la procédure de prévision jusqu'à obtenir des scores satisfaisants avec leurs données. Cet empirisme fait partie des critiques générales contre les RNA, mais n'empêche nullement l'obtention de très bons résultats. Cela souligne l'importance des données utilisées, qui doivent gouverner la construction d'un bon modèle prévisionnel.

Des réseaux de neurones comme le PMC peuvent également être utilisés pour des problèmes différents de la prévision localisée à court terme. Une application intéressante de ce type de modèle est donnée par Pfeiffer *et al.* (2009), avec une étude sur la modélisation de valeurs annuelles de NO_2 à Chypre. Les données utilisées sont des données d'inventaire, et les mesures

de NO_2 qui servent de cible ont été obtenues par campagne de tubes passifs. Ce type de donnée est adapté au pas de temps mensuel ou annuel, et cette étude montre que là aussi les RNA peuvent être utilisables.

Outre ce cas au pas de temps annuel, il existe des applications des PMC à la prévision journalière (localisée aux stations, donc) extrapolées par krigeage (méthode d'interpolation spatiale) afin de fournir des cartes de prévision couvrant une ville (Milan en l'occurrence, dans les travaux de Carnevale *et al.*, 2011). Ce type d'approche demande par contre des données rarement disponibles, avec un bon maillage de stations automatiques. Elle est également suggérée par Fernando *et al.* (2012), après une étude utilisant le PMC à Phoenix (Arizona, USA) comparé au modèle déterministe CMAQ (Community Air Quality Modelling System, `https://www.cmascenter.org/cmaq/`). Les réseaux neuronaux surpassant le modèle déterministe, les auteurs envisagent cette solution moins couteuse en calcul.

On peut enfin citer un autre type de modèle, les Support Vector Machine (SVM), des classifieurs non-linéaires. Leur principe est d'assigner les données d'entrée à leur classe en les transposant dans un espace de haute dimension séparé par un hyperplan délimitant les classes. Leur forme actuelle a été proposée par Cortes et Vapnik (1995) et ils sont désormais utilisés dans de nombreux problèmes de régression et de classification. Un exemple d'application des SVM à la prévision de la qualité de l'air est donné par Suárez Sánchez *et al.* (2011), pour de la prévision mensuelle, où les SVM surpassent alors le PMC. L'étude de Wang *et al.* (2003) applique des SVM à de la prévision de particules à un horizon de 72h, qui surpassent les RNA utilisés (Radial Basis Function (RBF)). L'application des SVM à la prévision de la qualité de l'air pourrait donner des résultats intéressants même si peu d'études y ont encore été dédiées.

2.6 Conclusion

La prévision de la qualité de l'air peut utiliser plusieurs types de modèles. Nous avons vu le fonctionnement et les avantages et inconvénients des CTM, des modèles statistiques et parmi eux particulièrement des RNA. Les CTM, basés sur les connaissances du fonctionnement atmosphérique, permettent une prévision à l'échelle de leur domaine, qui peut recouvrir toute la Corse. Au delà de la prévision, ils permettent de valider les modèles de connaissance sur lesquels ils sont basés et permettent de les améliorer. Mais leur usage est lourd. Un supercalculateur est nécessaire à leur fonctionnement, trop couteux pour une petite AASQA comme Qualitair Corse. De plus, des prévisions de CTM comme CHIMERE ou Skiron, administrés par d'autres organismes qui en diffusent les résultats, sont déjà utilisés par Qualitair Corse.

Les réseaux de neurones offrent des performances élevées appliqués à la prévision de la qualité de l'air. Leur non-linéarité leur permet de représenter efficacement les relations qui existent entre les variables décrivant l'état de l'atmosphère. Ils sont adaptés à l'usage d'une AASQA, ne nécessitant que peu de ressources de calcul pour leur entrainement (un ordinateur classique est suffisant). Ils peuvent utiliser des données issues de CTM et ainsi prendre en compte leur prévision.

Parmi eux, notre état de l'art a permis de retenir particulièrement le PMC, pour ces capacités d'approximateur universel et ses bons résultats dans notre domaine. Il s'agit d'une RNA « feedforward » à au moins une couche cachée.

On peut résumer les attributs du PMC :

Avantages :

– Pas de conditions de stationnarité par rapport aux variables utilisées

- Capacité à modéliser les relations non-linéaires entre les variables d'entrée et de sortie
- Performances de prévision en général supérieures aux modèles concurrents
- Faibles ressources informatiques et faible temps de calculs nécessaires
- Possibilités d'hybridation avec d'autres modèles

Inconvénients :

- L'aspect « black box » limitant la compréhension du modèle et la possibilité de s'en servir pour étudier le fonctionnement atmosphérique
- La nécessité de fixer empiriquement un grand nombre de paramètres en fonction du problème étudié (architecture du réseau, nombre de couches, de neurones, choix de l'algorithme d'apprentissage, etc.)
- La nécessité de disposer de larges échantillons pour l'apprentissage
- Différents prétraitements doivent être appliqués aux données pour obtenir des performances correctes

Dans ces travaux, nous abordons la prévision de la qualité de l'air principalement via l'utilisation de RNA, et notamment du PMC. Notre travail a pour but de développer des modèles précis, qui permettent d'anticiper au mieux les pics de pollutions, et qui soient utilisables de manière opérationnelle par une AASQA en respectant ses contraintes (contraintes matérielles, humaines, légales). Nous nous penchons principalement sur les particules et l'ozone, polluants les plus problématiques en Corse.

Les horizons de prévision sont de l'ordre de la journée (de quelques heures à plusieurs jours). Nous tentons d'avoir une approche « boîte grise » de la prévision par RNA, en travaillant sur les données, et en essayant d'améliorer les performances de nos modèles en leur simplifiant les problèmes que nous leurs soumettons à partir de connaissances sur les mécanismes de l'évolution de la qualité de l'air.

Chapitre 3

La qualité de l'air en Corse

La Corse est la plus petite mais la plus montagneuse des trois grandes îles situées en Méditerranée occidentale, les deux autres étant la Sardaigne et la Sicile. Elle est bordée à l'ouest par le bassin provençal, à l'est par la mer Tyrrhénienne et au nord par la mer Ligurienne. Au sud se situe l'île de la Sardaigne, à l'est et au nord la côte italienne et au nord-ouest les côtes françaises et la région PACA (voir figure 3.1).

FIGURE 3.1 : Situation de la Corse en mer Méditerranée (Open Street Map).

L'île est traversée par plusieurs chaînes de montagnes. Les plus importantes sont le massif du Monte Cintu, le massif du Monte Rotondo, le massif du Monte Renoso et le massif du Monte Incudine. Le sommet du Monte Cintu est le point culminant de l'île, à 2706 m d'altitude. Le relief de l'île est particulièrement prononcé, avec une altitude moyenne de 568 m pour une superficie de 8 680 km^2. La photographie en figure 3.2, prise depuis la station spatiale internationale (ISS), illustre l'aspect montagneux de l'île.

La Corse entre naturellement dans la zone climatique Méditerranéenne, avec des affinités subtropicales et tempérées selon la saison. L'importance de son relief introduit des contrastes marquants puisque le climat évolue de thermo-méditerranéen en climat alpin en seulement quelques kilomètres. On peut voir sur les diagrammes ombrothermiques de la figure 3.3 les différences

FIGURE 3.2 : Vue de la Corse depuis l'ISS (Credits : Terry W. Virts).

climatiques entre des villes côtières comme Bastia ou Ajaccio, et des villes au centre de l'île (Corte) ou à plus de 1000 m d'altitude (Bocognano).

La Corse est une région française au statut de Collectivité Territoriale, la seule de France métropolitaine, ce qui lui confère des pouvoirs spécifiques plus étendus que ceux des autres régions. Elle se divise en deux départements, Corse-du-Sud et Haute-Corse. La délimitation des deux départements suit la chaîne montagneuse qui sépare l'île en deux territoires : le Cismonte au nord (l'En-Deçà-des-Monts) correspond au département de la Haute-Corse (2B), et le Pumonti (l'Au-Delà des Monts) correspond au département de la Corse-du-Sud (2A). En plus du réseau routier littoral, le transport routier et ferroviaire entre les départements est possible grâce aux cols de Vizzavona, Vergio, Verde et Bavella.

C'est une île peu peuplée, comptant 322120 habitants en début 2013 pour une densité de population de 37.1 habitants.km^{-2} (contre 117 habitants.km^{-2} pour la France métropolitaine) selon l'INSEE (www.insee.fr). La majorité de la population vit sur le littoral, notamment autour des capitales des deux départements que sont les villes côtières d'Ajaccio en Corse-du-Sud et de Bastia en Haute-Corse. Au niveau des villes, Ajaccio intramuros compte 65000 citoyens contre 44165 pour Bastia, puis viennent Porto-Vecchio avec 11309 administrés, Borgo (7644), Biguglia (6934), Corte (6829), Calvi (5486), Furiani (5283), Lucciana (4246) et Ghisonaccia (3738). Ces dix communes totalisent à elles seules plus de la moitié des personnes qui résident en Corse. Les territoires ruraux sont beaucoup moins densément peuplés, notamment à cause du relief et malgré la présence de très nombreux villages toujours habités (360 communes au total sur l'île).

L'économie de l'île est dominée par le tourisme, avant d'autres secteurs d'activité comme le bâtiment. L'agriculture et l'industrie y sont faibles, et se concentrent autour du littoral. La plaine orientale est le territoire le plus propice à l'agriculture, et regroupe des productions fruitières, viticoles et maraichères. Le vignoble corse se développe sur différentes zones viticoles réparties sur tout le territoire. L'élevage extensif d'ovins et plus récemment de bovins est également présent sur l'île.

Il y a donc relativement peu de sources de pollution anthropique sur l'île. En ce qui concerne les sources industrielles, on note la présence de deux centrales thermiques de production d'électricité gérées par EDF. Du fait de son insularité, la Corse est en effet une ZNI (Zone Non Interconnectée) et de ce fait, elle est confrontée à de nombreux problèmes d'approvisionnement

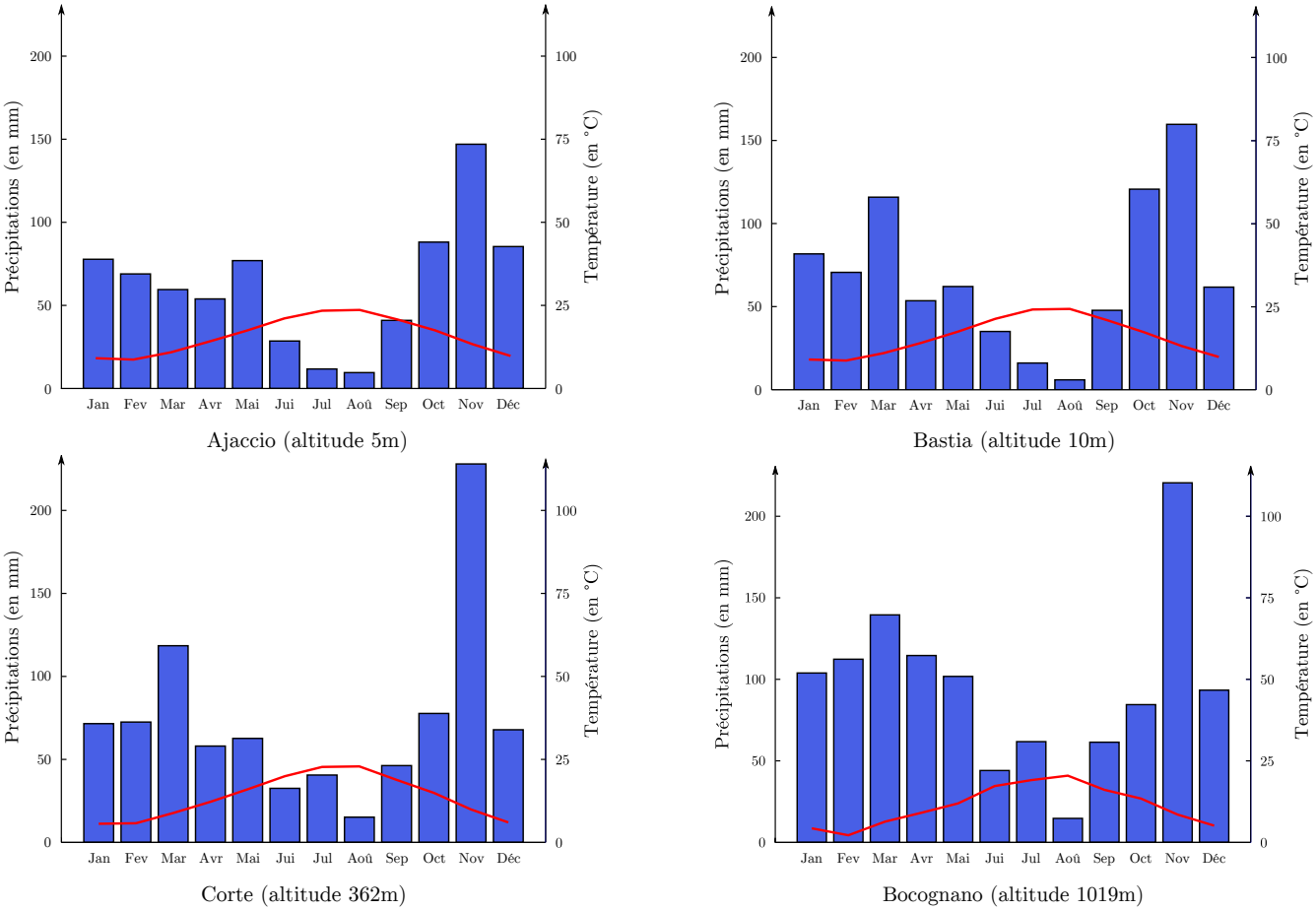


FIGURE 3.3 : Diagrammes ombrothermiques représentant les histogrammes des précipitations et les courbes des températures moyennes mensuelles à Ajaccio, Bastia, Corte et Bocognano .

en énergie électrique. L'utilisation de moteurs à combustion interne est prédominante dans la totalité des îles françaises (Notton, 2015), et est une source importante de polluants atmosphériques.

La centrale du Vazzio est située dans la ville d'Ajaccio et celle de Lucciana au sud de Bastia. La première, d'une puissance totale de 132 MW, utilise un fuel lourd à très basse teneur en soufre, et est équipée d'un réducteur catalytique. La centrale de Lucciana, d'une puissance de 54.5 MW jusque fin 2013, utilisait également du fuel lourd. La nouvelle centrale dont la mise en service a commencé début 2014 a une puissance de 112 MW et fonctionne au fuel léger. Cette dernière est également équipée de quatre TAC (Turbines A Combustion) d'une puissance totale de 125 MW, utilisées pour l'écrêtage des pointes de consommation.

Il convient de préciser une différence importante au niveau de l'emplacement des centrales thermiques du Vazzio et de Lucciana. La centrale du Vazzio est située à très grande proximité de la ville d'Ajaccio alors que la centrale de Lucciana est située à plusieurs kilomètres du centre ville de Bastia. Les emplacements des centrales thermiques et des ports sont présentés sur la figure 3.11 pour Ajaccio et les figures 3.13 et 3.14 pour Bastia.

La Corse est partiellement interconnectée avec l'Italie, grâce à deux câbles sous-marins :

– l'interconnexion SACOI (pour SArdaigne-COrse-Italie) entre Bastia et l'Italie (de 50 MW, courant continu)
– l'interconnexion SARCO (pour SARdaigne-COrse) entre Bonifacio au sud de la Corse et

la Sardaigne (de 100 MW, courant alternatif)

Les énergies renouvelables sont présentes sur l'île, notamment via un réseau de barrages hydrauliques. Plusieurs centrales photovoltaïques sont également présentes, ainsi que des fermes éoliennes. La figure 3.4 illustre la répartition des moyens de production d'électricité en Corse en 2014.

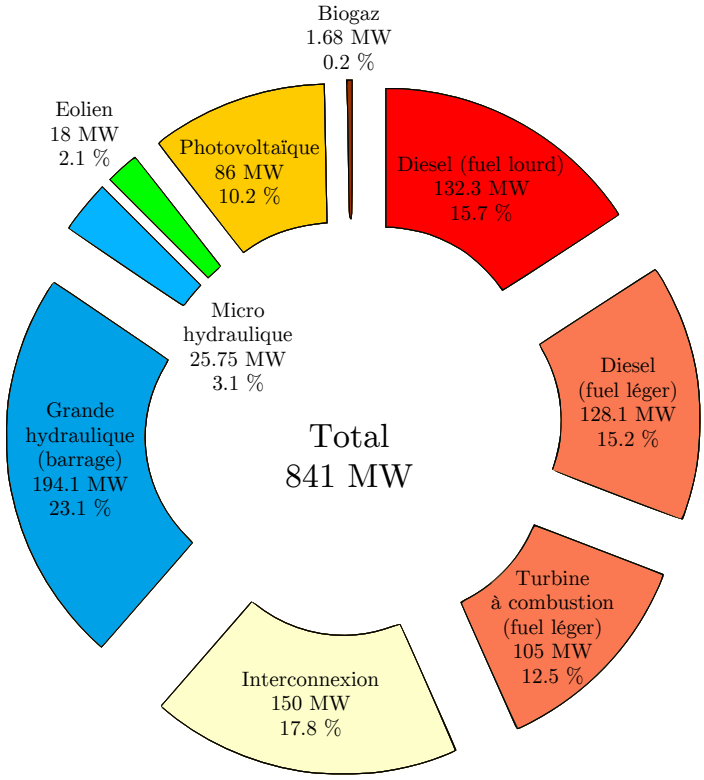


FIGURE 3.4 : Répartition des moyens de production électrique en puissance installée en 2014 (source EDF).

La part des différentes énergies dans la production d'électricité varie d'une année sur l'autre en particulier selon la météorologie car la majeure partie de la consommation est sujette aux variations saisonnières de température ; elle dépend également de la pluviométrie qui intervient sur la part plus ou moins importante de l'énergie hydraulique.

La répartition de la production d'électricité pour 2012 et 2013 est présentée sur la figure 3.5. En 2012, la production d'électricité a été de 2197 GWh avec une pointe de 530 MW, la part des énergies renouvelables représentait 21.6 % . En 2013, elle a été de 2235 GWh avec une pointe de 495 MW et une part d'énergies renouvelables de 33.2 %. La part de l'hydraulique de 598 GWh en 2013 a atteint un niveau jamais égalé, en hausse de 78 % par rapport à 2012 et qui est dû d'une part à une bonne pluviométrie et d'autre part à la mise en service d'une nouvelle centrale hydraulique de 55 MW sur le Rizzanese.

Outre la production d'énergie, on note comme sources industrielles d'émission de polluants atmosphériques les carrières en périphérie de Bastia et d'Ajaccio émettrices de particules. Un faible tissu industriel est présent. Quelques chaudières bois sont utilisées et sont émettrices de particules. Le brûlage de déchets verts, bien qu'interdit, est également une source très importante de pollution en Corse. Ce brûlage est en partie liée au besoin de démaquisage autour des habitations. Une campagne de sensibilisation sur ce sujet est actuellement en cours en Corse auprès de la population.

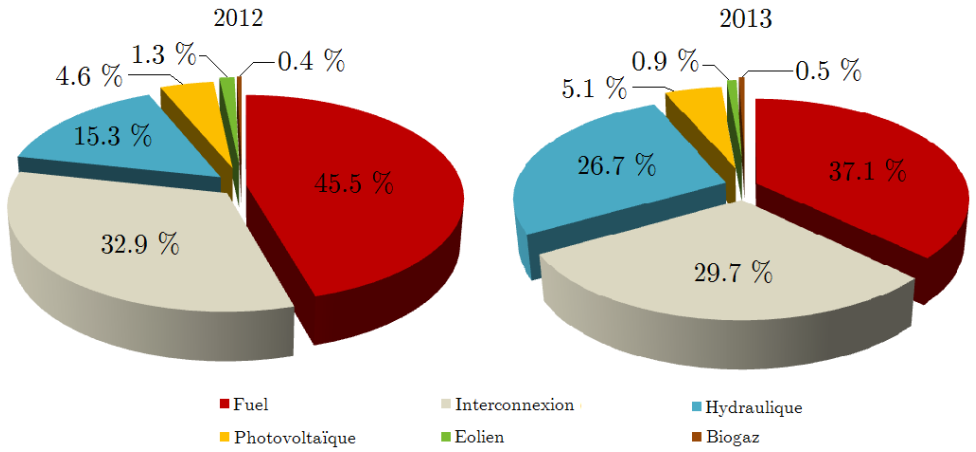


FIGURE 3.5 : Répartition de la production d'électricité en 2012 et 2013.

Au niveau du transport, les véhicules personnels sont très utilisés en Corse. Ceci est dû au manque de transport en commun y compris dans les centres urbains. On compte 612 véhicules pour 1000 habitants en Corse, pour 509 véhicules pour 1000 habitants en France. (Source : Ministère du transport, INSEE). Au classement du nombre de véhicules par habitant, les deux départements corses sont aux deux premières places. Le train qui relie Ajaccio à Bastia et Calvi fonctionne grâce à un moteur diesel.

L'interconnexion de la Corse avec le continent et la Sardaigne est assurée par six ports. Quatre aéroports civils sont également présents sur l'île. La pollution liée au transport est fortement influencée par l'affluence touristique ; entre 2.5 et 3 millions de touristes par an viennent en Corse, répartis surtout entre Juin et Septembre. Les émissions résidentielles comme celles causées par les climatisations sont également décuplées par le tourisme. Si le trafic aérien impacte la qualité de l'air, celui des bateaux (transport bord à bord et croisières) a une influence très importante car les deux ports principaux, à Bastia et Ajaccio, sont situés en centre ville des deux agglomérations (le port de Bastia est d'ailleurs concerné par un important projet d'extension). Les cheminées des navires sont à hauteur d'habitation et les moteurs des ferrys restent la plupart du temps allumés lors de leurs passages à quai.

Même si la Corse est peu industrialisée, ces différentes activités impactent la qualité de l'air de manière importante, et ce d'autant plus que les conditions météorologiques sont souvent propices à l'augmentation de leur concentration, comme nous le verrons ultérieurement.

La qualité de l'air est surveillée par Qualitair Corse, l'AASQA chargée de mesurer, informer et alerter sur l'état de l'air pour la région Corse. Il s'agit d'une association loi 1901, comme toutes les AASQA. Son siège est basé à Corte. Nous présenterons dans la suite de ce chapitre cet organisme de surveillance et son réseau de mesure, avant de nous intéresser aux données issues des mesures automatiques que nous avons utilisées dans ces travaux. Nous poursuivrons par une présentation des centres Météo-France, de leurs mesures et des sorties de modèles numériques mises à notre disposition par cet organisme.

3.1 Qualitair Corse

3.1.1 Présentation de l'AASQA

La création des AASQA a été rendu obligatoire par la Loi n°96-1236 du 30 décembre 1996 sur l'Air et l'Utilisation Rationnelle de l'Energie (LAURE) et par ses différents décrets d'appli-

cation. Cette loi, qui pose comme objectif fondamental « la mise en œuvre du droit reconnu à chacun à respirer un air qui ne nuise pas à sa santé », s'articule autour de trois grands axes :

- la surveillance et l'information
- l'élaboration d'outils de planification
- la mise en place de mesures techniques, de dispositions fiscales et financières, de contrôles et sanctions

Cette loi stipule la mise en place progressive d'un dispositif de surveillance de la qualité de l'air, devant être étendu à l'ensemble du territoire national au 1er janvier 2000. Cette surveillance est déléguée à des organismes agréés, les AASQA, associant notamment l'état, les collectivités territoriales, les industriels contribuant aux émissions de polluants, des associations de consommateurs ou de protection de l'environnement et des représentants des professions de santé ou personnes qualifiées. Au-delà de la mission de surveillance, les organismes agréés concourent à l'exercice du « droit à l'information sur la qualité de l'air ... reconnu à chacun sur l'ensemble du territoire » (article 4 de la loi LAURE).

Sous l'impulsion de cette loi, Qualitair Corse a été créée le 17 octobre 2003, il s'agit de l'avant-dernière AASQA créée en France. L'association fait partie de la Fédération ATMO qui regroupe l'ensemble des AASQA (figure 3.6). On peut trouver des informations utiles sur Qualitair Corse sur son site internet : www.qualitaircorse.org.

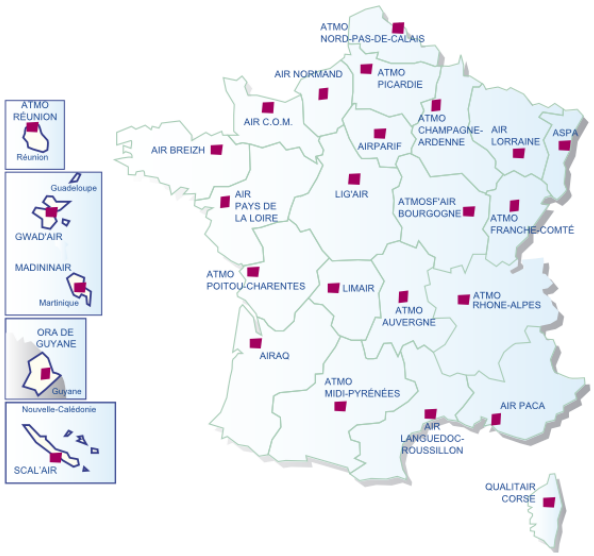


FIGURE 3.6 : Réseau des AASQA de la Fédération ATMO.

Les AASQA sont appuyées techniquement par le LCSQA qui garantie la qualité de la surveillance. Ce sont obligatoirement des associations à but non-lucratif, dont les membres sont des représentants de quatre collèges :

- collège de l'état (préfectures, Direction Régionale de l'Environnement, de l'Aménagement et du Logement (DREAL), etc.)
- collège des collectivités locales (intercommunalités, conseils généraux, etc.)
- collège des émetteurs de pollution atmosphérique (industriels, chambres de commerce et d'industrie, etc.)
- collège de personnalités qualifiées (associations de protection de l'environnement, chercheurs, professionnels de la santé, etc.)

Cette composition atypique regroupant aussi bien les émetteurs de pollution que les défenseurs de l'environnement se veut une garantie de transparence et de crédibilité des informations diffusées. Les missions de Qualitair Corse sont les suivantes :

- Surveiller la qualité de l'air, à l'aide de stations fixes et de campagnes de mesures
- Etudier les données afin dévaluer la qualité de l'air sur le territoire
- Prévoir la qualité de l'air afin d'anticiper les pics de pollution
- Informer le public et les autorités de l'état de l'air, notamment en cas de pic de pollution
- Conseiller les décideurs à l'occasion de projets d'aménagement afin d'assurer la prise en compte de la qualité de l'air

Certains polluants sont particulièrement suivis par les AASQA, pour leur représentativité de l'état de l'air, leur dangerosité mais aussi leur mesurabilité. Il s'agit de l'ozone, des PM10, du dioxyde d'azote et du dioxyde de soufre qui ont été introduits à la section 1.3 (page 13). Ces polluants sont réglementés, c'est-à-dire que leur suivi est obligatoire et que leurs concentrations ne doivent pas dépasser certains seuils.

Il existe deux seuils distincts de concentration pour chacun de ces polluants amenant une réaction particulière. Le premier est le seuil « d'information et de recommandation » et le second le seuil « d'alerte ». Ils sont spécifiés pour chaque polluant dans le tableau 3.1. Atteindre le premier seuil déclenche des actions d'information du public, des maires, des établissements de santé et des médias et la diffusion de recommandations médicales et de limitation d'émissions en polluants. En plus de ces actions, le dépassement du seuil d'alerte implique lui des actions de restriction ou de suspension de certaines activités polluantes.

Tableau 3.1 : Valeurs des concentrations correspondant aux seuils d'information et d'alerte en 2015 pour l'O_3, les PM10, le NO_2 et le SO_2.

Polluant	Moyenne	Seuil d'information ($\mu g.m^{-3}$)	Seuil d'alerte ($\mu g.m^{-3}$)
O_3	Horaire	180	240
PM10	Journalière	50	80
NO_2	Horaire	200	400
SO_2	Horaire	300	500

En cas de dépassement de ces seuils, Qualitair Corse communique envers le public et les autorités légales (préfecture et DREAL). La préfecture doit être à même de prendre des mesures temporaires adéquates afin de limiter les émissions de polluants lors du pic anticipé et d'en diminuer l'ampleur. Ces mesures peuvent concerner les industriels, avec des réductions des activités polluantes, les collectivités, avec des gratuités dans les transports en commun par exemple, ou l'ensemble de la population, avec par exemple des restrictions d'utilisation des véhicules personnels.

L'IQA est calculé à partir des concentrations de ces polluants. C'est un indice qui va de 1 pour une très bonne qualité de l'air à 10 pour une qualité de l'air très mauvaise. Il est représenté à la figure 3.7. Les concentrations de chacun des polluants de l'IQA permettent de calculer leur sous-indice. Le sous-indice d'un polluant égal à huit correspond à son seuil d'information, et à son seuil d'alerte quand il est égal à dix. L'IQA est égal à son sous-indice le plus élevé.

Ces polluants réglementaires sont mesurés grâce à différents analyseurs présentés plus bas (section 3.1.3), abrités dans des stations de mesures. Ces stations leur apportent les conditions nécessaires à leur fonctionnement (abri de la pluie, régulation de la température interne, humi-

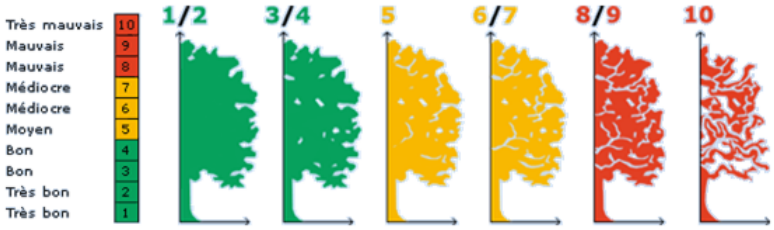


FIGURE 3.7 : Représentation de l'Indice de la Qualité de l'Air (IQA) émis par Qualitair Corse.

dité, etc.) et limite leur accès aux seuls techniciens de l'AASQA, qui assurent leur maintenance et leur étalonnage régulièrement. Elles peuvent être fixes ou mobiles pour permettre des campagnes temporaires de mesure.

Le réseau de stations fixes permet le suivi automatique et régulier de ces polluants. Le placement de ces stations fixes correspond à des critères bien précis, expliqués plus loin. Il fait suite à une étude locale qui a déterminé le lieu le plus propice pour capter les niveaux de polluants les plus élevés. Pour cela, des stations mobiles sont utilisées, ainsi que des préleveurs passifs afin de produire une cartographie de pollution qui aide à identifier la position optimale de la future station.

Il existe plusieurs types de station fixe, résumés au tableau 3.2, qui déterminent le type de zone dans laquelle leurs mesures sont représentatives. Les stations « urbaines » sont représentatives de la qualité de l'air respiré en ville. Plus loin des centres-villes, les stations « périurbaines » mesurent les polluants dans des zones représentatives des banlieues et zones moins densément urbanisées. Les stations « rurales » sont situées loin des sources, et représentent la pollution de fond et des régions reculées. Elle permettent également de connaître l'impact de la pollution sur les écosystèmes. Ces trois types de stations sont considérées comme des stations de « fond », représentatives de larges zones.

TABLEAU 3.2 : Typologie des stations fixes de surveillance de la qualité de l'air.

Typologie	Définition	Zone
Urbaine	Exposition moyenne des populations	100 m - 1 km
Périurbaine	Exposition en périphérie des villes	1 - 5 km
Rurale	Evaluation de l'impact sur les écosystèmes	5 - 25 km
Industrielle	Influence de sources industrielles	100 m - 5 km
Trafic	Exposition maximale en proximité	1 - 5 m

Les stations « trafic » sont situées à la proximité immédiate des axes routiers les plus importants, afin de représenter les concentrations maximales auxquelles sont exposés piétons et usagers de la route (automobilistes, cyclistes, usagers de transports en commun, etc.). Enfin les stations « industrielles » sont elles dévolues à la surveillance des émissions dues aux acteurs industriels. Ces deux types de stations sont considérées comme des stations de « proximité » étant donnée l'origine des polluants auxquels elles sont dédiées.

Jusqu'en 2014, c'était principalement la consultation de ces mesures qui servaient à déclencher les procédures liées aux dépassements de seuil, sur constatation du dépassement. Le déclenchement des alertes et des mesures à prendre est désormais régi par l'arrêté du 26 mars 2014 (Journal Officiel du 29 mars 2014), présenté en annexe A, page 196. Le changement de procédure est particulièrement important car ce n'est plus la constatation du dépassement qui fait office de déclencheur, mais la prévision de ce dépassement. Cette évolution améliore les pos-

sibilités de réaction face à un pic de pollution, elle favorise les mesures d'urgence demandant un certain temps de mise en place.

L'article 2 de cet arrêté précise par exemple qu'un épisode de pollution peut se caractériser ainsi : « pour les départements de moins de 500 000 habitants, lorsqu'au moins une population de 50 000 habitants au total dans le département est concernée par un dépassement de seuils d'ozone, de dioxyde d'azote et/ou de particules PM10 estimé par modélisation en situation de fond », ou « dès lors qu'une surface d'au moins 100 km^2 au total dans une région est concernée par un dépassement de seuils d'ozone, de dioxyde d'azote et/ou de particules PM10 estimé par modélisation en situation de fond ». La notion de « persistance d'un épisode de pollution aux particules PM10 » est introduite ; elle concerne les cas où un dépassement du seuil d'information a été constaté deux jours de suite, et que ce dépassement est également prévu pour le lendemain. Le reclassement de l'épisode de pollution aux particules en épisode de persistance lui confère le même niveau de réaction qu'un épisode d'« alerte ».

Les AASQA doivent donc disposer de moyens de prévision efficaces, ce qui souligne l'intérêt de ces travaux de doctorat. On a vu au chapitre 2 que plusieurs types de modèles peuvent réaliser ces prévisions. Des modèles déterministes (section 2.2, page 22) peuvent être utilisés, comme le modèle AIRES déployé chez Air PACA qui couvre le domaine corse, mais aussi comme le modèle national Prév'air ou le modèle Skyron. On a aussi vu que l'absence d'inventaire des émissions corses, ainsi que l'absence d'assimilation statistique des données issues du réseau de Qualitair Corse nuisent à la qualité des prévisions d'AIRES. Notons qu'un inventaire des émissions est en cours de finalisation à Qualitair Corse, qui a de grandes chances d'améliorer les prévisions d'AIRES sur l'île. La section 3.3 sera consacrée à l'évaluation de la précision d'AIRES en Corse. On s'y rendra compte que cette précision n'est pour l'heure pas satisfaisante.

Les cartes de prévision fournies par ces modèles sont consultées par le personnel d'astreinte chargé de réaliser les prévisions à Qualitair Corse. Mais en l'absence de tout autre outil qu'AIRES et les différentes plates-formes de prévision disponibles, Qualitair Corse avait besoin d'un outil propre adapté à la prévision de la qualité de l'air sur l'île. C'est cette situation qui a donné lieu à ce travail de doctorat, qui a pour but de réaliser un tel outil pour le mettre à disposition de l'AASQA. Le chapitre 7 présentera les outils développés en ce sens, les applications « Aria Base » et « Aria Web ».

Le suivi et la prévision de ces polluants réglementaires sont au cœur de l'activité des AASQA. Mais Qualitair Corse mène d'autres activités. D'autres polluants sont également mesurés (métaux lourds, HAP, etc.) qui sont également réglementés mais dont les seuils de concentration concernent les valeurs moyennes annuelles. Des campagnes de mesures sont régulièrement menées afin d'étudier la qualité de l'air dans les zones dépourvues de stations fixes (grandes villes autre que Bastia et Ajaccio, zones industrielles, etc.). Elles permettent d'apporter des connaissances sur la qualité de l'air de l'ensemble de la Corse, ainsi que de valider ou faire évoluer la stratégie de surveillance de l'AASQA.

3.1.2 Réseau de surveillance

La surveillance de la qualité de l'air s'effectue suivant une stratégie, développée dans un Programme de Surveillance de la Qualité de l'Air (PSQA) renouvelé tous les cinq ans. Ce plan découpe le territoire corse en plusieurs zones. Quand une valeur réglementaire est dépassée au sein d'une zone, c'est l'ensemble de cette zone qui est considéré comme en dépassement.

Ces zones sont au nombre de trois. La ZUR (Zone Urbaine Régionale) d'Ajaccio recouvre le golfe d'Ajaccio et ses environs et remonte dans la vallée de la Gravona. La ZUR de Bastia s'étire

au nord de la ville dans le Cap Corse jusqu'à Erbalunga, et descend dans la banlieue de Bastia autour de l'étang de Biguglia, jusqu'à Casamozza et jusqu'à Monte en Castagniccia. Le reste de la Corse forme la ZR (Zone Régionale) corse. Le réseau de stations fixes est constitué de neuf stations, réparties sur le territoire et particulièrement dans les ZUR d'Ajaccio et de Bastia. Les zones ainsi que la position des stations fixes sont indiquées en figure 3.8. Le tableau 3.3 présente le parc de stations fixes de Qualitair Corse.

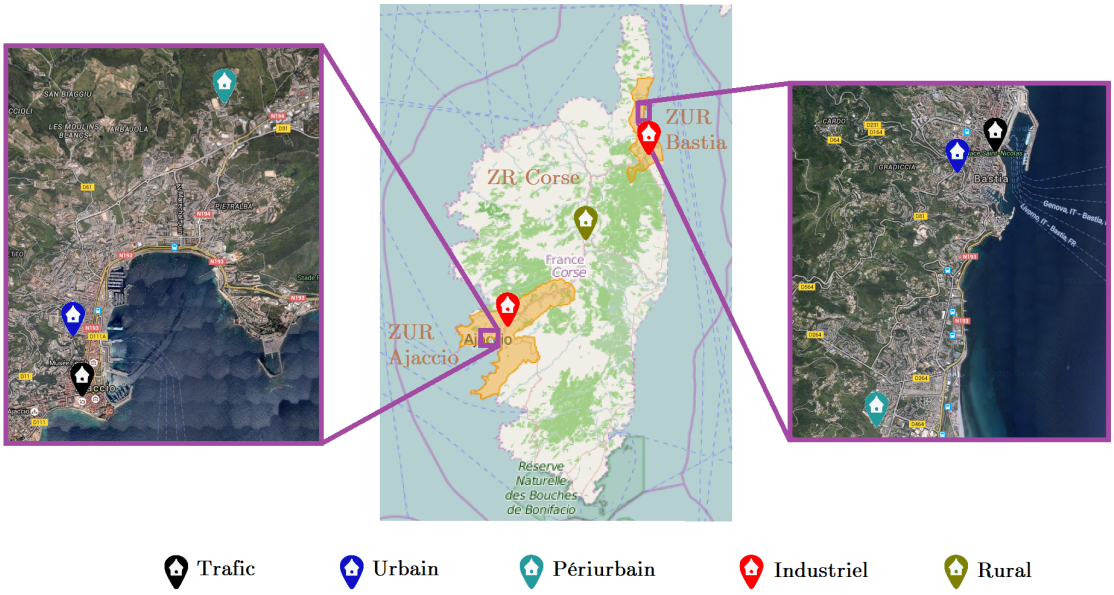


FIGURE 3.8 : Position des stations fixes de Qualitair Corse.

Les premières mesures fixes ont donc débuté en 2006, à la station de Canetto. Chaque ZUR est équipée d'une station « trafic », d'une station « urbaine », d'une station « périurbaine » et d'une station « industrielle » aux alentours de chaque centrale thermique. La ZR recouvre la majorité du territoire mais est beaucoup moins densément peuplée et les sources de polluants y sont moindres qu'autour des deux grandes villes. La seule station fixe de la ZR est la station « rurale » de Venaco, dans la commune éponyme en centre-Corse.

Outre le suivi des PM10, de l'ozone, du NO_2 et du SO_2, certaines stations de l'AASQA effectuent également des mesures de particules fines (PM2.5) et de CO.

Au delà de son réseau de station fixe, Qualitair Corse utilise des stations mobiles (photographie en figure 3.9) et des préleveurs passifs (tubes passifs) afin de réaliser des campagnes de mesures sur l'intégralité du territoire corse. Les préleveurs passifs permettent de mesurer sur filtre les quantités intégrées de polluant sur toute leur période d'exposition. Ils sont utilisés pour suivre l'O_3, le NO_2 et le benzène. Les campagnes temporaires permettent d'évaluer la stratégie de surveillance de la qualité de l'air en apportant des informations spatialisées sur les concentrations, et de vérifier quelles zones enregistrent les plus fortes concentrations. Elles peuvent motiver des évolutions du PSQA. Elles permettent aussi d'élargir les connaissances sur la dynamique locale des polluants, en étudiant des zones où aucune mesure fixe n'est prévue.

Des préleveurs actifs sont également utilisés. Contrairement aux tubes passifs, un débit contrôlé d'air augmente leurs capacités de prélèvement, leur permettant de mesurer des quantités de polluants beaucoup moins concentrés. Ils sont utilisés pour la mesure de « nouveaux polluants », espèces dangereuses qui ont été réglementées plus récemment que l'ozone, les particules les oxydes d'azote et de soufre : il s'agit de « métaux lourds », de COV comme le benzène et de HAP, présentés à la section 1.3.5 page 19.

Tableau 3.3 : Réseau de stations fixes de surveillance de la qualité de l'air en Corse.

Ville	Station	Type	Altitude	Polluants mesurés	Début	Fin
Ajaccio	Canetto	Urbain	39 m	O_3, PM10, PM2.5, NO_x, SO_2	2006	-
	Sposata	Périurbain	60 m	O_3, NO_x	2007	-
	Diamant	Trafic	12 m	PM10, NO_x	2008	-
Sarrola-Carcopino	Piataniccia	Industriel	30 m	O_3, PM10, NO_x, SO_2, CO	2006	-
Grosseto-Prugna	Porticcio	Périurbain	4 m	NO_x	2007	2011
Bastia	Giraud	Urbain	60 m	O_3, PM10, NO_x, SO_2	2006	-
	Montesoro	Périurbain	47 m	O_3, PM10, PM2.5, NO_x	2007	-
	Saint Nicolas	Trafic	5 m	PM10, NO_x	2008	-
Lucciana	La Marana	Industriel	15 m	O_3, PM10, NO_x, SO_2	2007	-
Venaco	Venaco	Rural	653 m	O_3, PM10, PM2.5, NO_x	2011	-

Qualitair Corse utilise un préleveur bas débit Partisol, un préleveur de type Leckel à moyen débit, un préleveur DA80 à haut débit et un préleveur Sypac. Ils permettent de prélever sur filtre des quantités de polluant dépendant de la durée de prélèvement, les filtres étant analysés *a posteriori* en laboratoire pour déterminer les concentrations atmosphériques moyennes correspondantes. On peut ainsi étudier les concentrations de métaux lourds, de benzène, de HAP, de particules mais aussi de pesticides.

Qualitair Corse a participé à l'élaboration d'un programme stratégique de surveillance des pollens. Elément fécondant mâle des plantes à fleur et allergène important, les pollens forment des grains de quelques dizaines de micromètres de diamètre. Le pollen n'est pas un polluant et n'est pas règlementé, cependant il pose un problème de santé vis-à-vis des personnes sensibles. Afin de suivre les pics de pollen, Qualitair Corse a évalué la possibilité de mise en place de « pollinariums sentinelles », espaces où les espèces régionales aux pollens allergisants sont réunies. Leur observation quotidienne permet de détecter le début et la fin d'émission de pollen et de transmettre l'information aux personnes allergiques, qui doivent commencer leurs traitements avant le pic de pollen.

Même s'il ne fait pas partie du réseau de Qualitair Corse, on citera un autre site abritant des mesures fixes en Corse. Il s'agit du Centre d'Observation Régional pour la Surveillance du Climat et de l'environnement Atmosphérique et océanographique en Méditerranée occidentale (CORSiCA) situé à Ersa, tout au nord du Cap Corse, la péninsule au nord de l'île. Ce super-site regroupe un grand nombre d'équipements de mesure faisant partie de la campagne ChArMEx (Chemistry-Aerosol Mediterranean Experiment) (Dulac *et al.*, 2013) à laquelle Qualitair Corse a participé. Il est situé en un lieu très reculé et a pour but de mesurer la pollution de fond dans le bassin méditerranéen. Son emplacement le situe dans une zone où l'on peut observer les épisodes de transport de masses d'air polluées depuis le continent (notamment depuis les

FIGURE 3.9 : Station mobile de Qualitair Corse.

Bouches-du-Rhône et la vallée du Pô, zones particulièrement industrialisées), tout comme les épisodes de transport de poussières sahariennes.

Il accueille des équipements permettant l'étude des aérosols (Nicolas *et al.*, 2013), notamment un TEOM-FDMS 1405 mesurant les PM1. L'ozone y est également mesuré parmi d'autres gaz (Pichon *et al.*, 2013). Les mesures d'O_3 et de PM1 du site CORSiCA sont très intéressantes pour Qualitair Corse, notamment mise en relation avec les mesures effectuées à Bastia pour se rendre compte des niveaux de fond sur les crêtes du Cap. En temps que données exogènes en entrée de RNA, elles apportent également de l'information sur les niveaux de fond et les arrivées de masses d'air d'Italie ou du sud de la France aux modèles prévisionnels.

Après cet aparté sur le super-site CORSiCA, nous allons à présent nous intéresser aux appareils de mesure de concentrations utilisés pour chaque polluant dans les stations fixes, ainsi qu'à la communication et à la validation des données produites.

3.1.3 Instrumentation

Les stations fixes du réseau enregistrent les concentrations en polluants de manière automatique et les transmettent à Qualitair Corse en temps réel. Ces données sont ensuite validées par l'équipe technique de l'AASQA deux fois par jour, afin d'assurer leur représentativité. La validation technique permet d'identifier les dysfonctionnements matériels, et la validation environnementale a pour but de déterminer les situations où la mesure, bien que juste, n'est plus caractéristique de l'état de l'atmosphère dans sa zone de représentativité. Cela permet d'éviter qu'un phénomène très localisé soit interprété comme effectif sur l'ensemble d'une ZUR ou de la ZR. Les données invalidées sont mises à part des jeux de données valides (mais conservées). Les données qui ont été utilisées par nos modèles sont les données validées.

Tous les deux jours, une vérification automatique de la dérive des appareils est menée. Cette

opération se fait grâce à l'injection de gaz étalons de concentration connue.

Intéressons-nous au parc analytique de Qualitair Corse dévolu aux mesures automatiques de PM10, O_3, NO_2 et SO_2. Les mesures d'ozone sont réalisées grâce à l'analyseur Thermo modèle $49i$. Cet appareil est basé sur l'absorption des radiations UV de longueur d'onde 254 nm par l'ozone. Le degré d'absorption de la lumière UV est directement corrélé à la concentration de l'ozone. Les NO_x sont mesurés par des analyseurs Thermo $42i$ qui utilisent la chimiluminescence de la réaction

$$NO + O_3 \longrightarrow NO_2 + O_2 + h\nu \tag{3.1}$$

grâce à un ozoniseur fournissant le réactif à partir d'air sec. La concentration en NO peut ainsi être déterminée. Les échantillons d'air peuvent également passer par un convertisseur transformant le NO_2 en NO préalablement à la réaction avec l'ozone, afin de quantifier les NO_x. La concentration en NO_2 est obtenue par soustraction entre celles de NO_x et celle de NO. Le dioxyde de soufre est lui mesuré grâce à l'analyseur Thermo $43i$ par photométrie UV Pulsée. La désexcitation du SO_2 après absorption de rayonnement UV émet un rayonnement qui est mesuré et à partir duquel la concentration en dioxyde de soufre est déterminée.

Les particules en suspension sont suivies grâce à des TEOM 1400 équipés de modules FDMS 8500 et à la station Canetto d'un TEOM 1405-DF incluant le module FDMS. Le fonctionnement du TEOM 1400 suit les principes suivants : après une arrivée d'air par une tête de prélèvement ne laissant passer que les particules en dessous d'un certain diamètre aérodynamique, les échantillons d'air passent au travers d'un filtre retenant les particules. Ce filtre est pesé régulièrement, permettant de mesurer la masse des particules, et leur concentration est déduite grâce au débit d'air. La mesure de la masse s'effectue indirectement par la mesure de l'oscillation du filtre. Le problème du TEOM 1400 est que l'air doit être chauffé aux alentours de 50°C, ce qui a pour effet de volatiliser la fraction soluble des particules qui n'est donc pas pesée. Pour compenser cette perte, les mesures doivent être corrigées par un facteur empirique représentant la part de cette fraction. Les modules FDMS 8500 ont par la suite été ajoutés à ces analyseurs afin de prendre en compte cette fraction volatile de manière précise.

Le module FDMS fonctionne sur une base séquentielle. Pendant une première phase de 6 minutes, l'air prélevé est déshydraté dans une cartouche avant de rejoindre l'analyseur et d'y être pesé. Lors de la seconde phase de 6 minutes également, l'air prélevé est envoyé vers un piège à particules avant d'être renvoyé vers l'analyseur, exempt de particule. La fraction volatile des particules impactées sur le filtre lors de la précédente phase va se volatiliser pendant cette phase, induisant une perte de masse permettant de l'estimer.

Le TEOM 1405-DF intègre deux modules FDMS et deux microbalances, et permet de mesurer deux classes de particules simultanément (à Canetto les PM10 et les PM2.5). Le passage des TEOM 1400 avec correction des mesures aux TEOM équipé de FDMS a créé une certaine discontinuité dans les mesures. Les séries temporelles de particules que nous avons utilisées sont des jonctions des mesures corrigées et des mesures avec FDMS.

La station de Venaco abrite une jauge bêta (un BAM 1020) pour la mesure des particules. Cet appareil mesure la masse des particules accumulées sur une bande grâce à une source au ^{14}C émettant un rayonnement bêta, atténué par la présence des particules.

3.2 Présentation des données

Nous allons maintenant présenter les données elles-mêmes, issues des mesures de ces appareils. Elles permettent de se rendre compte de la dynamique des polluants, et sont celles qui ont

été utilisées par nos modèles. Ces données sont organisées en séries temporelles, chaque valeur étant rattachée à sa date de mesure, et donc ordonnée dans le temps (exemple en figure 3.10).

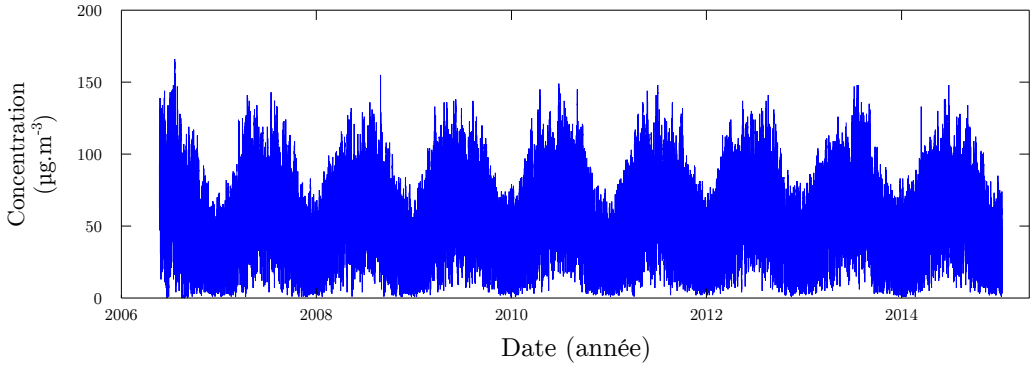


FIGURE 3.10 : Série temporelle de concentration d'ozone à Canetto.

Les mesures automatiques sont effectuées toutes les cinq minutes pour la plupart des appareils, puis transmises à Qualitair Corse sur une base quart-horaire. Pour ces travaux de thèse, nous avons décidé d'utiliser des séries temporelles horaires ; en effet, les déclenchements de procédures se font sur la base de données horaires (sauf concernant les particules pour lesquelles les seuils concernent la moyenne journalière du polluant). Avec ce pas de temps, il est possible de suivre la plupart des phénomènes ; de plus, cela nous permet de limiter le volume des données par rapport à l'usage d'une base quart-horaire, de diminuer les temps d'apprentissages des modèles sans perdre d'information utile. Ce pas de temps correspond également à celui des données des stations Météo-France qui nous sont transmises (voir section 3.2.6 page 78).

3.2.1 Données d'Ajaccio

A Ajaccio, trois stations fixes sont situées au sein de la ville (Canetto, Sposata et Diamant). La station industrielle Piataniccia se situe en périphérie, dans la vallée de la Gravona (voir figure 3.11). La station « trafic » Diamant est située sur la place éponyme, en plein centre ville, mais sera déplacée prochainement le long d'un axe routier où des campagnes mobiles ont permis d'identifier des concentrations potentiellement plus élevées en moyenne que sur la place du Diamant.

La station « urbaine » de Canetto est située sur le site de l'ancienne usine de traitement de l'eau (elle est à 400 m du port d'Ajaccio). Celle de Sposata ce situe en retrait de la route de Mezzavia.

Comme on peut le voir sur les roses des vents, la ville est sous l'influence du régime de brise de mer diurne et de brise de terre nocturne (voir section 1.2 page 10). Quand les conditions météorologiques permettent à ce régime de s'établir, les émissions de la centrale du Vazzio, qui utilise actuellement du fuel lourd comme carburant, sont donc transportées vers la station de Piataniccia. En situation de brise de terre, les stations du centre-ville peuvent également être sous l'influence du panache du Vazzio.

Les profils journaliers moyens des concentrations d'ozone, de dioxyde d'azote et de PM10 sont indiqués à la figure 3.12.

La station périurbaine de Sposata enregistre également des variables météorologiques, à savoir la Température en degrés Celsius (TC), la Pression Atmosphérique (PA), la Direction du Vent (DV), la Vitesse du Vent (VV), les Précipitations (P) et l'Humidité Relative (HR). La ville est marquée par un régime de brise de mer, que l'on retrouve sur les roses des vents.

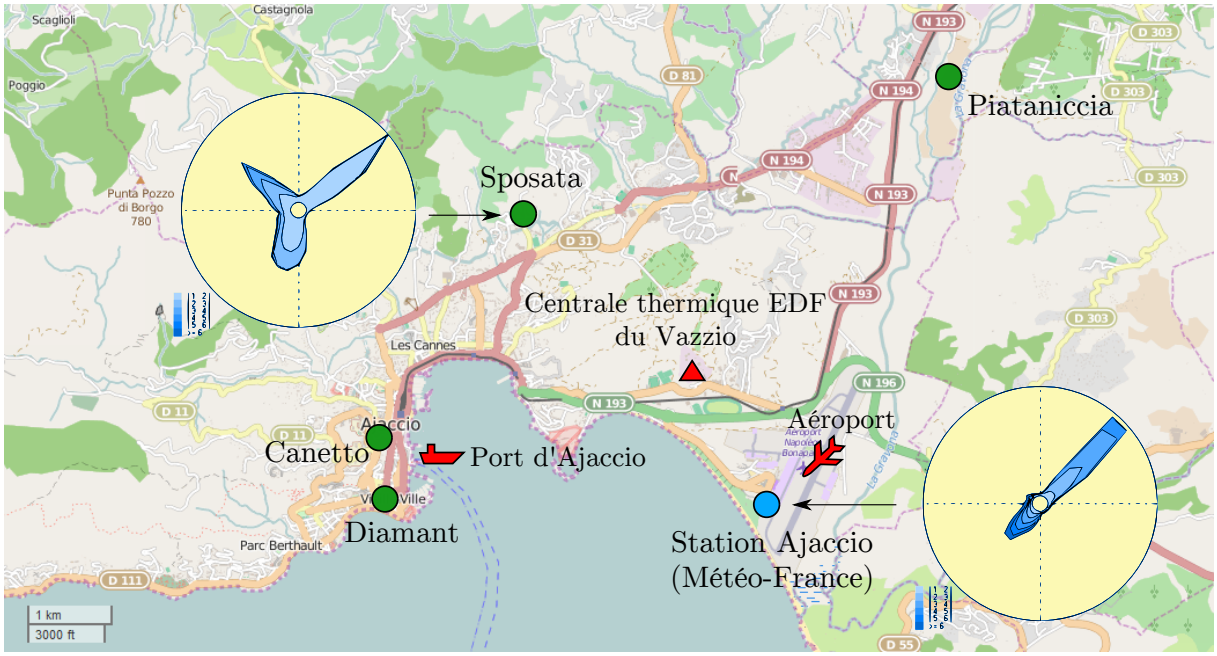


FIGURE 3.11 : Stations à Ajaccio avec les roses des vents (fond Open Street Map).

On peut voir sur le profil d'ozone (figure 3.12) que la dynamique varie en fonction des stations. On retrouve le profil en cloche typique du polluant, avec des dynamiques nocturnes différentes à Piataniccia, où les concentrations d'ozone redescendent bien, et à Canetto et Sposata où l'ozone redescend moins. On peut attribuer cette différence à la situation de Piataniccia, dans la vallée de la Gravona, vallée qui peut favoriser les situations de stabilité nocturne qui augmentent le dépôt sec (voir section 1.3.2 page 15). On retrouve sur les trois profils la diminution matinale des concentrations liée au pic de trafic routier émettant des NO_x qui titrent l'ozone (voir équation 1.5 page 16).

Les deux pics de trafic sont bien visibles sur les profils de NO_2, avec des concentrations plus fortes pour les sites plus exposés au trafic automobile (trafic, puis urbain, périurbain et industriel). On retrouve ces pics sur les profils de particules.

Aucun dépassement de seuil d'information ni *a fortiori* de seuil d'alerte n'a été enregistré dans la région ajaccienne pour l'ozone. Ce sont les particules qui posent problème. Les épisodes de pollution aux PM10 ont souvent lieu en conjonction avec un épisode de transport de poussières sahariennes, qui arrivent sur la Corse. Ces particules s'ajoutent à celles émises localement, notamment par la centrale thermique, le trafic routier et le chauffage urbain et peuvent mener à des dépassements. L'importance des épisodes de transport de poussières sahariennes sur les niveaux de particules dans l'ouest méditerranéen a récemment été souligné par Léon *et al.* (2015) et Di Biagio *et al.* (2015) dans le cadre de la campagne ChArMEx.

Au niveau local, une étude sur la caractérisation des particules fines a été menée par Qualitair Corse et Météo-France sur le territoire de la Communauté d'Agglomération du Pays Ajaccien (CAPA). Son rapport est disponible sur le site internet de Qualitair Corse. Cette étude montre l'importance des épisodes de transport lors des pics de pollution aux particules dans le pays ajaccien. Ces épisodes peuvent être liés aux poussières sahariennes, mais certains sont aussi dûs aux polluants émis dans la vallée du Pô en Italie lors de flux de nord-est.

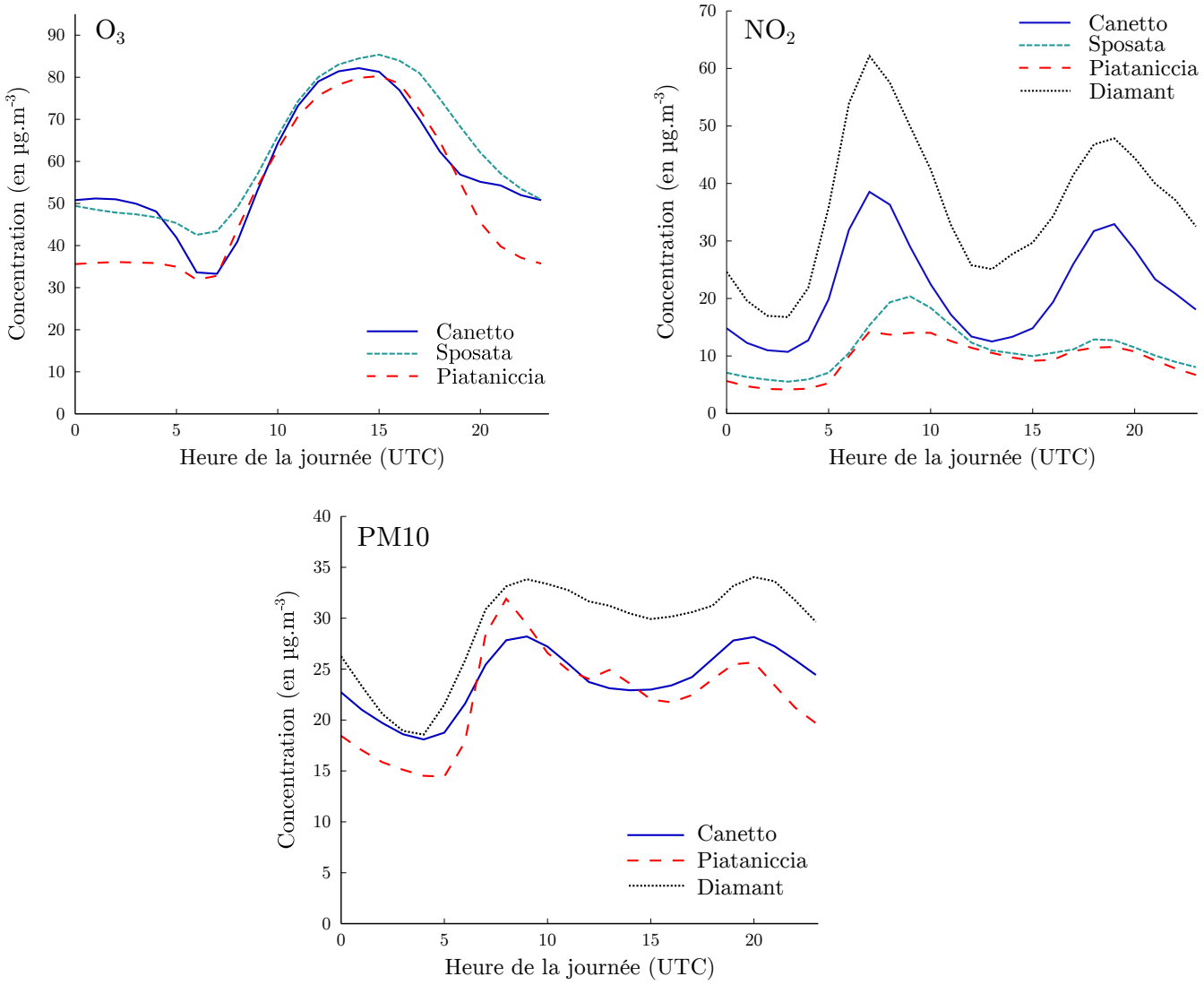


FIGURE 3.12 : Profils journaliers moyens des concentrations d'ozone, de dioxyde d'azote et de PM10 mesurées aux stations d'Ajaccio.

3.2.2 Données de Bastia

Le réseau de mesures à Bastia comprend les stations Giraud (urbaine), Montesoro (périurbaine) et Saint Nicolas (trafic). Cette dernière est située à la sortie nord du tunnel routier de Bastia, sous la place Saint Nicolas et à quelques dizaines de mètres du port de Bastia (voir la carte en figure 3.13). La station Giraud est située derrière le collège de même nom. La station Montesoro se situe plus au sud du centre-ville.

Plus au sud se trouve la station de surveillance industrielle de La Marana, à proximité de l'aéroport Poretta, d'une gravière et de la centrale thermique EDF de Lucciana (voir figure 3.14). La station est au sud de l'étang de Biguglia, qui est une réserve naturelle, lieu d'étape pour les oiseaux migrateurs et de nidification pour plusieurs espèces, et à ce titre classé Zone Protection Spéciale (ZPS) et Zone Spéciale de Conservation (ZSC) Natura 2000.

Les profils moyens des concentrations mesurées dans la région bastiaise ont été tracés en figure 3.15. Pour l'ozone, si l'on retrouve bien le profil en cloche à La Marana, ce n'est pas le cas pour les stations urbaine et périurbaine. Ces dernières sont situées dans la ville même de Bastia,

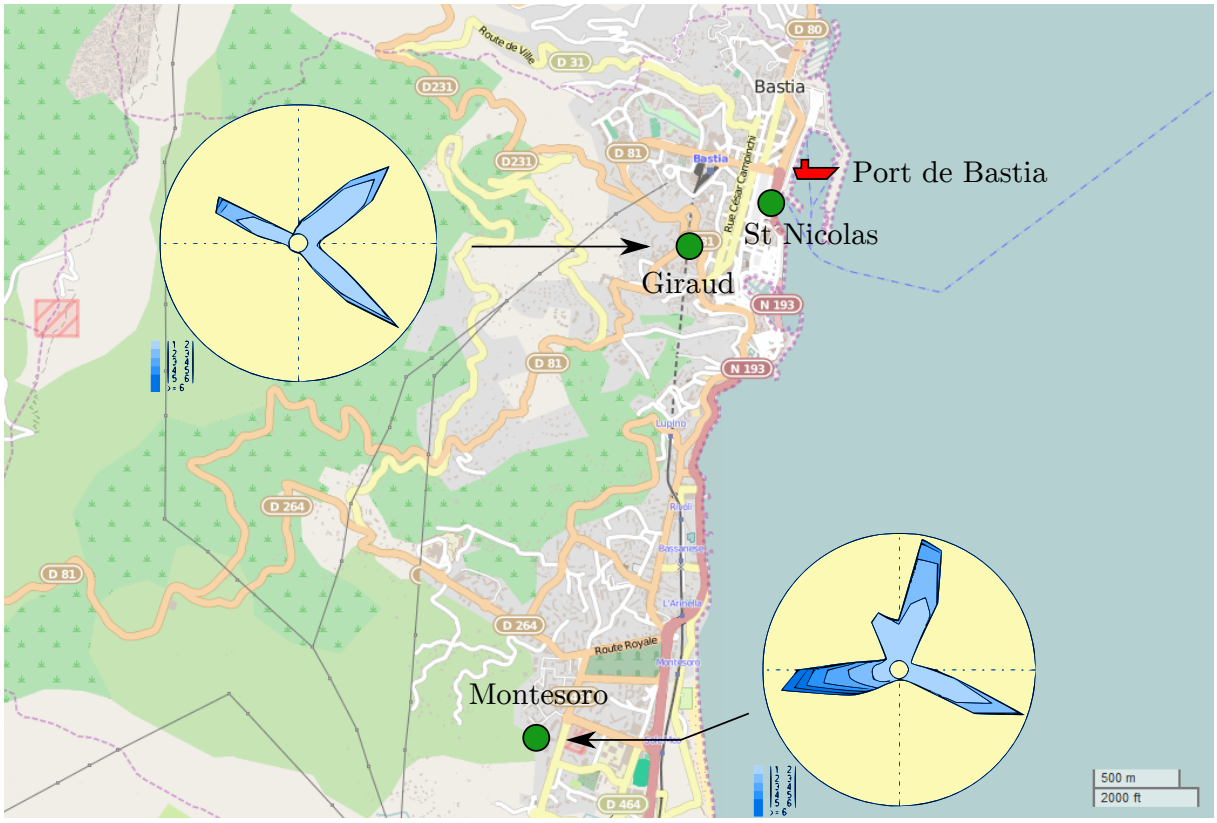


FIGURE 3.13 : Stations à Bastia avec les roses des vents (fond Open Street Map).

FIGURE 3.14 : Station industrielle à Lucciana et rose des vents (fond Open Street Map).

au pied de la chaîne de la Serra qui traverse la péninsule du Cap Corse. Cette situation peut empêcher la mise en place d'une couche stable nocturne, quand au coucher du soleil les brises de pente s'inversent et créent des déplacements d'air. Mélangés aux concentrations d'ozone plus fortes de la couche résiduelle en altitude, les niveaux restent élevés la nuit. On se rapproche alors des profils d'ozone qu'on trouve en crête. La Marana n'est pas concernée car la station est située plus au sud, en plaine. La nuit, le dépôt sec y fait baisser les concentrations et on retrouve un profil en cloche.

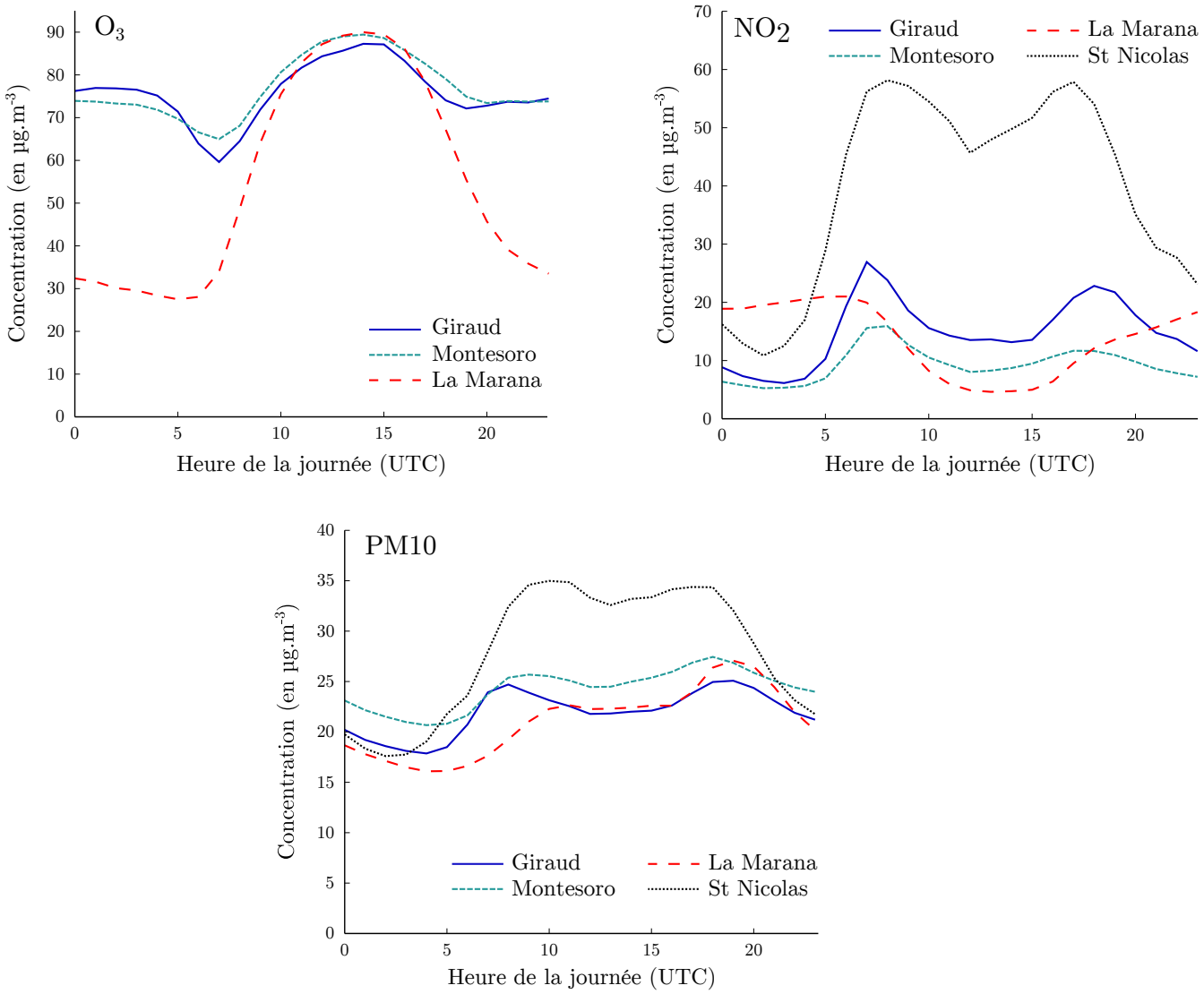


FIGURE 3.15 : Profils journaliers moyens des concentrations d'ozone, de dioxyde d'azote et de PM10 mesurées aux stations de Bastia.

On retrouve l'impact des émissions de NO_x liés au trafic sur les profils d'ozone et de NO_2. La station Saint Nicolas située en sortie du tunnel voit ses niveaux moins redescendre entre les deux pics que la station Diamant à Ajaccio. Sur les profils de NO_2 on retrouve des différences normales de niveaux entre les différents types de station, c'est-à-dire une amplitude des pics qui correspond à l'exposition aux sources de chaque type de station (voir tableau 3.2). Le profil à La Marana peut sembler atypique. Si l'on se réfère à la carte de la région, on remarque que cette station est placée sous le vent de la centrale thermique EDF de Lucciana quand le vent vient du sud-ouest, par exemple en situation de brise de terre durant la nuit. Quand la brise de

terre s'installe, on voit les concentrations de NO_2 augmenter car la station est influencée par les émissions de la centrale et celles du trafic routier sur la nationale 193. Quand la brise de mer s'installe, la concentration diminue car il y a peu de sources entre la station et le littoral. Cette tendance à recevoir le panache industriel plutôt la nuit que le jour est également visible dans une moindre mesure sur le profil de PM10.

Des données météorologiques sont enregistrées à la station Giraud (TC, PA, DV, VV, P et HR). Le matériel de mesure météorologique était préalablement installé à la station Montesoro. Ce déplacement a eu pour conséquence de diviser en deux le jeu de données météorologiques, ce qui pose problème pour l'apprentissage automatique, qui nécessite des séries temporelles les plus longues possibles. L'usage alternatif des données de Météo-France pose un problème d'éloignement, car elles sont mesurées au niveau de l'aéroport Poretta, à une quinzaine de kilomètres du centre-ville. Quand nous avons voulu utiliser en entrée de modèle les mesures météorologiques de Bastia, nous avons donc fusionné les séries temporelles météorologiques de la stations Giraud avec celles de la station Montesoro afin d'obtenir une série temporelle plus grande. Le plus souvent cependant, nous avons préféré l'usage des sorties de modèle de Météo-France, présentés à la section 3.2.6 page 78.

3.2.3 Données de Venaco

Site fixe le plus récent, la station rurale de Venaco se trouve au centre de la Corse (voir figure 3.8 page 65). Cette station fait partie du dispositif national de suivi de l'équivalence des mesures de particules fines mis en place par le LCSQA. Ce dispositif a permis de vérifier l'équivalence entre différentes méthodes de mesure des particules (jauges bêta) par rapport à la méthode de référence qu'est la gravimétrie (par pesée de filtre).

FIGURE 3.16 : Station rurale de Venaco.

Les profils moyens de concentrations sont présentés sur la figure 3.17. On peut voir que l'ozone a un profil très stable, lié à la position en crête de la station. Les niveaux de particules et de NO_2 sont bas car ils correspondent aux niveaux de fond régional. Les pics de trafic sont à peine perceptibles sur la courbe de NO_2.

Les séries temporelles des données enregistrées à la station de Venaco, inaugurée en 2011, sont plus courtes que celles des autres stations de Qualitair Corse. Ceci était particulièrement

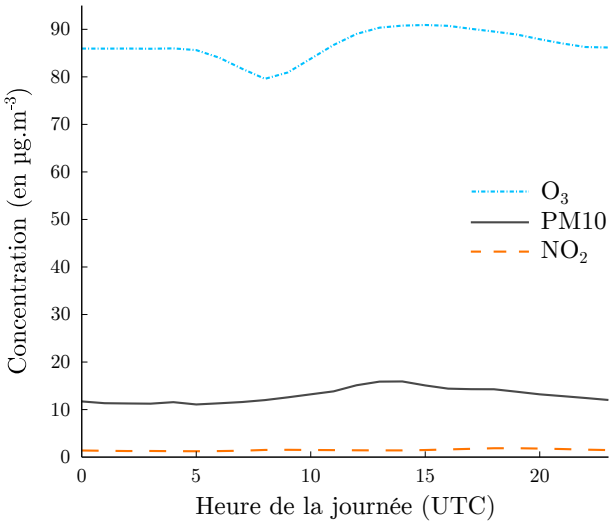


FIGURE 3.17 : Profils journaliers moyens des concentrations d'ozone, de dioxyde d'azote et de PM10 mesurées à la station rurale Venaco.

vrai au début de ces travaux de doctorat en 2012. De plus, des interruptions de mesures plus longues que sur les autres stations ont eu lieu à la station rurale. Ces éléments nous ont conduits à écarter la plupart du temps les données de Venaco pour ne pas réduire les jeux d'apprentissage de nos modèles.

A la fin de ces travaux, les données ont tout de même été suffisantes pour créer des modèles statistiques de prévision à Venaco. L'usage de ces données en temps que variables exogènes pour des modèles prédictifs à Bastia et Ajaccio permettrait de représenter les niveaux de fonds enregistrés sur la Corse. Cette perspective deviendra intéressante à l'avenir, avec l'augmentation de la taille des séries temporelles de Venaco.

3.2.4 Incertitudes de mesure

Qualitair Corse fait partie d'un groupe de travail inter-AASQA qui a pour but d'évaluer régulièrement les incertitudes de mesures des polluants. Ces mesures doivent respecter une incertitude relative maximale, de 15 % pour les gaz et de 25 % pour les particules, par rapport à une concentration donnée. Cette concentration doit être au voisinage de la valeur limite réglementaire, c'est-à-dire le seuil d'information et de recommandation du polluant (voir tableau 3.1 page 62), et ce afin de connaitre les incertitudes dans les conditions où des déclenchements de procédure peuvent avoir lieu.

L'incertitude de mesure est due à plusieurs sources différentes. Chaque source contribuant à l'incertitude est évalué séparément. Sont ainsi évaluées les incertitudes liées :

- aux réglages et étalonnages des appareils
- à l'analyseur lui-même
- à la ligne de prélèvement
- au système d'acquisition
- au milieu environnant
- à la présence d'espèces chimiques interférant lors des mesures

Le tableau 3.4 indique ces incertitudes, pour l'ozone, le NO_2 et le SO_2, calculées pour l'ensemble du parc analytique de Qualitair Corse. L'évaluation des incertitudes est en cours pour

les mesures de PM10, mais n'est pas encore disponible.

Tableau 3.4 : Incertitudes évaluées sur les mesures de gaz à Qualitair Corse.

Polluant	Valeur limite ($\mu g.m^{-3}$)	Incertitude 2012	Incertitude 2013
O_3	240	12.1 %	10.6 %
NO_2	211	13.4 %	12.9 %
SO_2	798	15.7 %	13.9 %

Les incertitudes respectent bien les recommandations, à l'exception des mesures de SO_2 de 2012 qui ont dépassé les 15 %. Elles se sont améliorées depuis et satisfont désormais les recommandations.

Il est important dans le cadre de ces travaux sur la prévision de garder en tête ces incertitudes. Les modèles à apprentissage que nous avons utilisés sont faits pour reproduire les interactions existant entre plusieurs variables. Comme pour toute mesure, les données qui représentent ces variables sont entachées d'erreur ce qui signifie que le modèle prédictif le plus performant possible restera condamné à reproduire cette incertitude.

Les modèles ne sont pas à proprement parler des modèles de prévision de concentration, mais bien des modèles de prévision des mesures de ces concentrations. Ce fait ne doit pas être oublié, notamment lors de l'évaluation des modèles qui est réalisée à partir de ces données, et quand il faut juger la précision obtenue.

3.2.5 Bilan

Nous avons présentés les données enregistrées dans chaque zone. Le tableau 3.5 donne quelques informations sur les données de l'ensemble des stations fixes que nous avons présentées. Plusieurs choses doivent en être retenues. Tout d'abord, les séries temporelles les plus longues sont celles des deux stations « urbaines », Canetto et Giraud. Etant de plus représentatives de la qualité de l'air respiré par la majorité de la population, ce sont les données de ces stations qui seront le plus utilisées pour la prévision.

Le second point d'importance concerne les valeurs. Les moyennes des concentrations sont relativement basses, comparées aux données d'autres AASQA en régions comprenant plus de sources anthropiques de pollution. Aucune des valeurs maximales d'ozone n'atteint le seuil d'information. Pour les PM10, cependant, la valeur maximale atteint 525 $\mu g.m^{-3}$ à Canetto et 692 $\mu g.m^{-3}$ à Piataniccia lors du plus important épisode de pollution enregistré sur l'île, ayant eu lieu fin novembre 2014. Bien qu'exceptionnel, cet épisode montre que ce polluant nécessite un travail de prévision afin d'anticiper ces pics. La relative rareté d'épisodes de pollution dans les données apporte cependant une difficulté pour l'apprentissage automatique.

A Ajaccio depuis le début des mesures, la moyenne sur 24 heures glissantes de la concentration en PM10 a dépassé les 50 $\mu g.m^{-3}$ à 25 occasions à Canetto, et 12 fois à Piataniccia (où les mesures de PM10 ont commencé en janvier 2013). Ces dépassements n'ont pas tous donné lieu à des déclenchements de procédure ; en effet, le seuil d'information était fixé à 80 $\mu g.m^{-3}$ (qui est actuellement le seuil d'alerte) jusqu'en 2012. De plus, la moyenne journalière de minuit à 23 heures a souvent été utilisée plutôt que la moyenne sur 24 heures glissantes, ce qui ne permet pas d'identifier tous les dépassements. Le seuil d'alerte a quant à lui été franchi une fois, lors de l'épisode de novembre 2014. Ce pic de pollution était lié à un épisode de transport de poussières sahariennes.

Tableau 3.5 : Statistiques sur les mesures de fond d'O_3, de PM10, de PM2.5 et de NO_2 jusqu'au 12/01/2015. Les moyennes, écarts-type et maximums sont fournis en $\mu g.m^{-3}$.

Station	Polluant	Début	Nombre de points	Moyenne	Ecart-type	Max
Canetto	O_3	23/05/06	74561	58.04	29.10	166
	PM10	09/03/07	64672	23.94	11.95	525
	NO_2	23/05/06	74432	21.33	16.23	128
Giraud	O_3	01/08/06	71688	75.79	23.28	164
	PM10	15/02/07	63642	21.92	10.81	149
	NO_2	01/08/06	70837	15.94	12.63	130
Sposata	O_3	10/03/07	67358	61.72	27.64	171
	NO_2	10/03/07	67073	11.11	9.37	99
Montesoro	O_3	06/08/07	63472	76.96	23.72	177
	PM10	06/08/07	21406	24.23	10.43	110
	PM2.5	23/11/07	53925	12.23	6.63	92
	NO_2	06/08/07	62885	9.29	8.31	95
Venaco	O_3	29/03/11	23142	86.66	18.24	165
	PM10	29/04/11	15714	12.99	9.78	287
	PM2.5	20/03/13	7195	8.75	4.44	37
	NO_2	01/01/13	17219	1.49	0.96	15

Sur Bastia, douze dépassements du seuil d'information ont eu lieu à Giraud, neuf à Montesoro et trois à La Marana, si l'on se réfère à la moyenne sur 24 heures glissantes des PM10. Il n'y a pas eu de déclenchement de procédure d'alerte, l'épisode de novembre 2014 qui a déclenché l'alerte en région ajaccienne n'ayant déclenché à Bastia que la procédure d'information et de recommandation. Le nombre de dépassements liés aux PM10 est plus faible qu'à Ajaccio, ce qui se traduit également par une concentration moyenne plus faible. Cette différence pourrait provenir de la position de Bastia, au nord de la Corse et plus éloignée du Sahara. La concentration en ozone est toujours restée en deçà du seuil d'information depuis que le polluant est mesuré, même si elle l'a souvent fortement approché. Les seules valeurs atteignant le seuil des 180 $\mu g.m^{-3}$ qui aient été enregistrées en Corse l'ont été par une station mobile, alors située sur le col qui surplombe Bastia, ou par le site CORSiCA au nord du Cap.

On se trouve donc dans une situation où les concentrations en polluants sont préoccupantes sur l'île, mais dépassent rarement les seuils réglementaires. Les particules posent le plus problème, et c'est sur ce polluant qu'il conviendra de se focaliser. L'ozone peut également être responsable de pics de pollution, mais les autres polluants (NO_2 et SO_2) ont peu de chances d'être impliqués dans des dépassements de seuils, du moins dans un avenir proche.

On développera des modèles prévisionnels essentiellement dédiés aux PM10 et à l'ozone, dans les stations de fond et particulièrement aux stations urbaines, les plus anciennes et représentatives de zones à la fois des plus polluées et abritant la plus grande part de la population corse.

3.2.6 Données Météo-France

Météo-France utilise vingt-six stations en Corse, réparties sur toute l'île. Les stations situées à l'aéroport Campo Dell'Oro d'Ajaccio et à proximité de l'aéroport Poretta à Bastia sont les plus proches des deux agglomérations où les mesures de polluants atmosphériques sont réalisées. Les données météorologiques issues de ces deux stations seront utilisées pour compléter les variables météorologiques des stations de Qualitair Corse. Ces données sont récupérées automatiquement par Qualitair Corse chaque jour.

En plus des données disponibles en stations fixes de Météo-France, nous avons accès aux sorties de modèles de Météo-France. Tout comme pour les données de stations, ces sorties de modèles sont communiquées automatiquement à Qualitair Corse dès qu'elles sont disponibles. Il s'agit des sorties des modèles ARPEGE et AROME, sur un domaine géographique recouvrant la Corse (figure 3.18). Nous avons ainsi accès à deux sorties par jour, celles calculées à minuit et celles calculées à midi. La récupération automatique de ces données leur permet d'être utilisées pour la prévision opérationnelle de la qualité de l'air, comme on le verra à la section 7.2 (page 173).

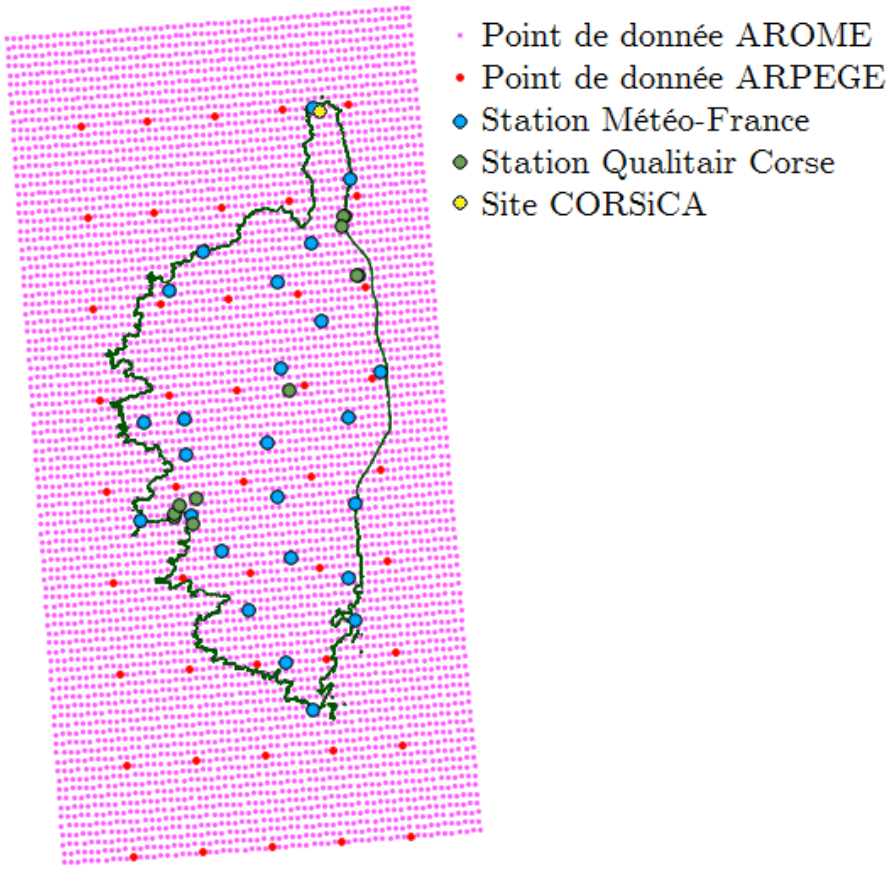


FIGURE 3.18 : Cartographie des données disponibles.

Le modèle AROME (Bouttier, 2007; Seity *et al.*, 2011) déployé par Météo-France en 2008 est un modèle non-hydrostatique avec une résolution spatiale de 2.5 km. Cette résolution permet d'améliorer sa prise en compte des phénomènes convectifs par rapport à son prédécesseur ALADIN, dont la résolution peut aller de 7 à 10 km. Une version d'AROME d'avril 2015 offre

même une résolution d'1.3 km. Le domaine d'ARPEGE couvre le monde entier à une résolution horizontale variable, autour de 10 km en France métropolitaine. Ce sont donc les sorties d'AROME qui seront utilisées dans ces travaux. Les variables d'AROME qui ont été rendues disponibles pour Qualitair Corse sont les suivantes :

- PA (en Pa)
- Géopotentiel (GEO) (en m^2s^{-2})
- Température en degrés Kelvin (TK)
- Composante ouest-est du vent U (en $m.s^{-1}$)
- Composante sud-nord du vent V (en $m.s^{-1}$)
- Composante verticale du vent (en $m.s^{-1}$)
- HR (en %)
- Nébulosité (NEB) totale (en %)
- NEB de l'étage inférieur (en %)
- NEB de l'étage moyen (en %)
- NEB de l'étage supérieur (en %)
- P totales (en $kg.m^{-2}$, regroupe les précipitations liquides et solides)
- Rayonnement Solaire (RS) (rayonnement de courtes longueurs d'onde, en $j.m^{-2}$)
- Rayonnement Thermique (RT) (rayonnement de grandes longueurs d'onde, en $j.m^{-2}$)
- Hauteur de la Couche Limite (HCL) (en m)

Ces variables sont calculées à plusieurs niveaux verticaux. Certaines n'ont pas d'« altitude » à proprement parler (comme HCL). D'autres sont disponibles au niveau du sol, ou recalculées au niveau de la mer (comme la pression). La plupart sont disponibles entre 10 m et 3000 m d'altitude, ainsi qu'à 800 et 700 hPa. Chaque fichier de sortie de modèle AROME contient les données modélisées pour l'heure de la prévision, ainsi qu'à plusieurs échéances de prévision : toutes les trois heures jusqu'à $h + 30$.

Nous avons passé en revue les données dont nous disposons, venant de Météo-France ou de Qualitair Corse. Nous allons maintenant nous intéresser en particulier au modèle AIRES, utilisé jusqu'à présent par Qualitair Corse pour réaliser les prévisions de qualité de l'air.

3.3 La plate-forme AIRES en Corse

La plate-forme AIRES d'Air PACA fournit des prévisions sur un domaine incluant la Corse. Ces données sont utilisées par Qualitair Corse pour réaliser les prévisions du jour au lendemain à l'attention du public et des autorités. En plus des résultats de prévisions disponibles en ligne (http://www.aires-mediterranee.org/html/airesv3/), les sorties brutes du modèle sont envoyées à l'AASQA quotidiennement.

Ce modèle, basé sur CHIMERE et illustré en figure 3.19, est intéressant pour Qualitair Corse. L'AASQA à souvent l'occasion de travailler avec Air PACA du fait de la proximité des deux régions, ce qui facilite les échanges autour de l'utilisation d'AIRES et permet d'y apporter les améliorations nécessaires.

Faisant partie des principaux modèles de références consultés pour la prévision en Corse, il est important de se rendre compte objectivement de ses capacités. Nous allons évaluer ses capacités de prévision sur l'île, ce qui nous permettra de savoir s'il peut être directement utilisé pour assurer les prévisions.

Nous avons évalué les sorties de ce modèle, afin de savoir si l'on peut s'y fier actuellement pour la prévision en Corse. Cette évaluation s'est faite en comparant les sorties du modèle les

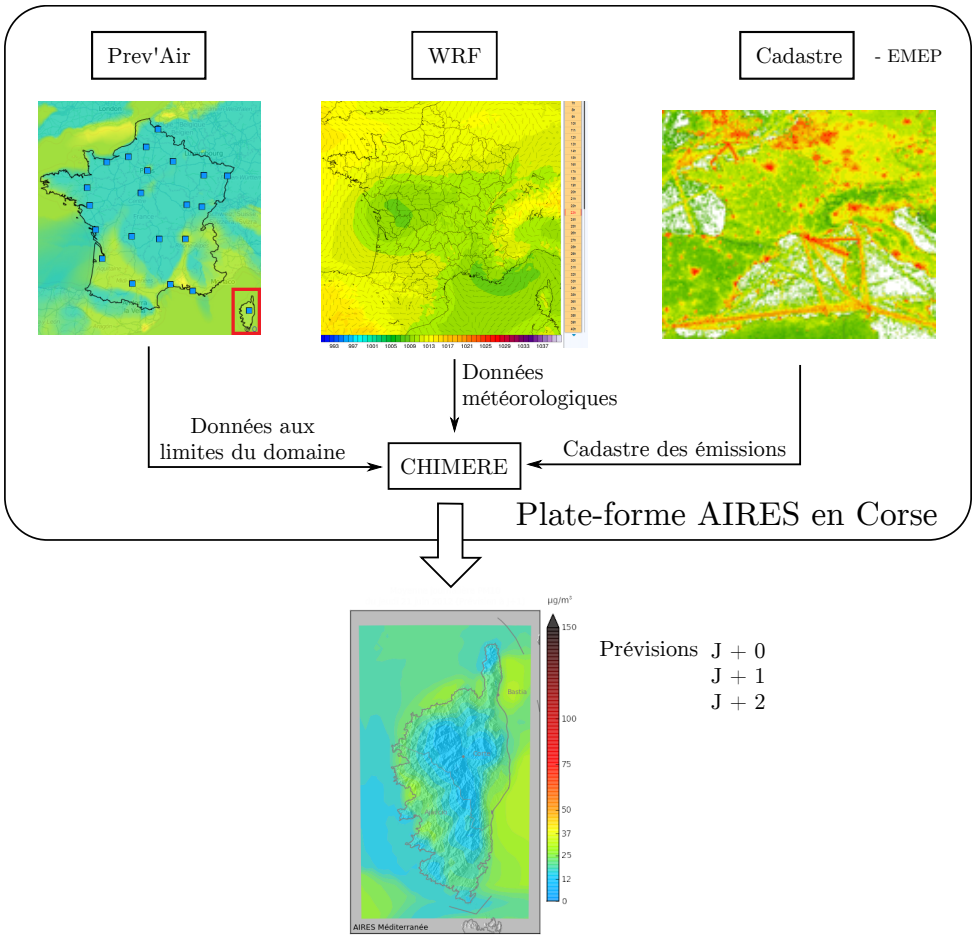


FIGURE 3.19 : Schéma du fonctionnement de la plate-forme AIRES en Corse.

plus proches d'une station avec les mesures réalisées à cette station (voir section 2.4 page 38). Nous avons pour cela utilisé les données des stations périurbaines pour évaluer les prévisions d'ozone, et des stations urbaines pour les particules, c'est-à-dire celles où les concentrations sont les plus élevées, ainsi que les données de la station rurale de Venaco pour les deux polluants. Les sorties de modèle utilisées correspondent aux données de 2011 à 2013. Pour l'ozone, la figure 3.20 montre les résultats à la station Sposata d'Ajaccio, à la station Montesoro de Bastia et à la station de Venaco.

Ces résultats sont moins bons que ceux obtenus par le même modèle en région PACA. AIRES a tendance à sous estimer les niveaux d'ozone. Quand on se penche sur les nuages de points obtenus aux trois stations, on se rend compte que cette sous-estimation est plus importante à Venaco, puis à Montesoro. Les scores de MBE et de FB vont dans le même sens. Ceci pourrait être dû aux imprécisions sur le relief, plus importantes en situation encaissée, comme c'est le cas à Bastia et surtout à Venaco.

Dans les trois cas, on peut dire que les résultats d'AIRES aux stations ne permettent pas sans plus d'analyse de prévoir les fortes valeurs d'ozones.

Les résultats pour les PM10 sont données en figure 3.21 à Canetto (Ajaccio), à Giraud (Bastia) et à Venaco.

Les résultats sont similaires à Bastia et Ajaccio, avec une RMSE proche de 15.7 $\mu g.m^{-3}$ et un indice d'agrément d autour de 0.48. La MBE négative indique encore une sous-estimation des valeurs, bien qu'on puisse voir sur les nuages de points certaines prévisions beaucoup plus

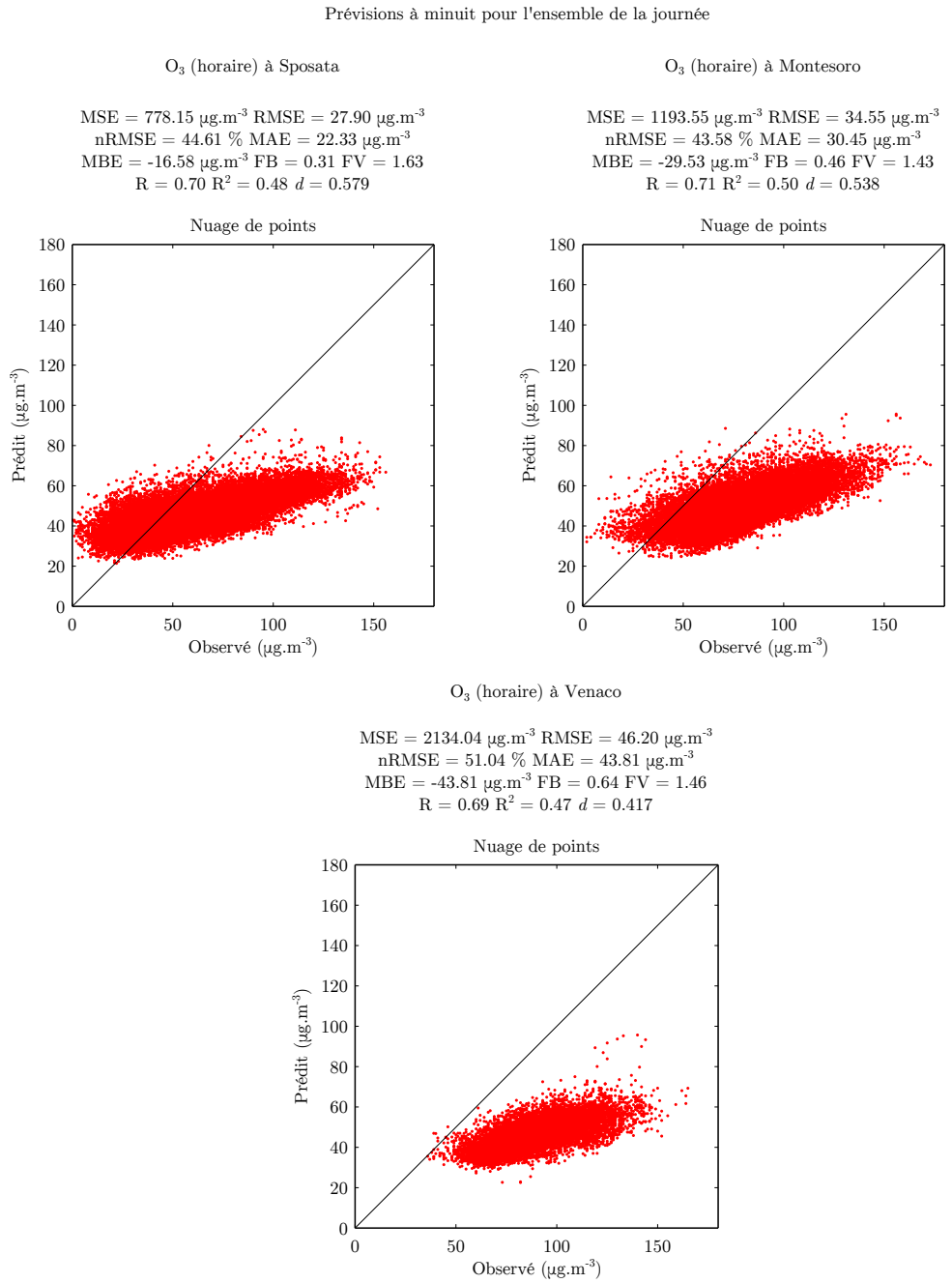


FIGURE 3.20 : Evaluation des prévisions d'ozone horaires brutes d'AIRES en Corse.

élevées que les observations. Il y a malheureusement beaucoup plus de fausses alertes ou d'alertes ratées que de dépassements correctement prévus pour le seuil d'information.

A Venaco les scores sont meilleurs. La RMSE atteint 12.32 $\mu g.m^{-3}$, ce qui peut en partie s'expliquer par des concentrations en particules plus faibles. Dans les trois cas, les dépassements du seuil d'information sont difficiles à prévoir.

Que ce soit pour les particules ou pour l'ozone, on peut voir que les performances d'AIRES en Corse sont faibles, ce qui relance l'intérêt de travailler sur un autre type de modèle prédictif. AIRES est un modèle de qualité, et ses performances sont en parties limitées par l'absence de cadastre régional des émissions, qui sera bientôt disponible. Enfin aucune correction statistique ne lui est appliquée en Corse. Nous envisageons donc d'utiliser ses sorties en entrées de nos

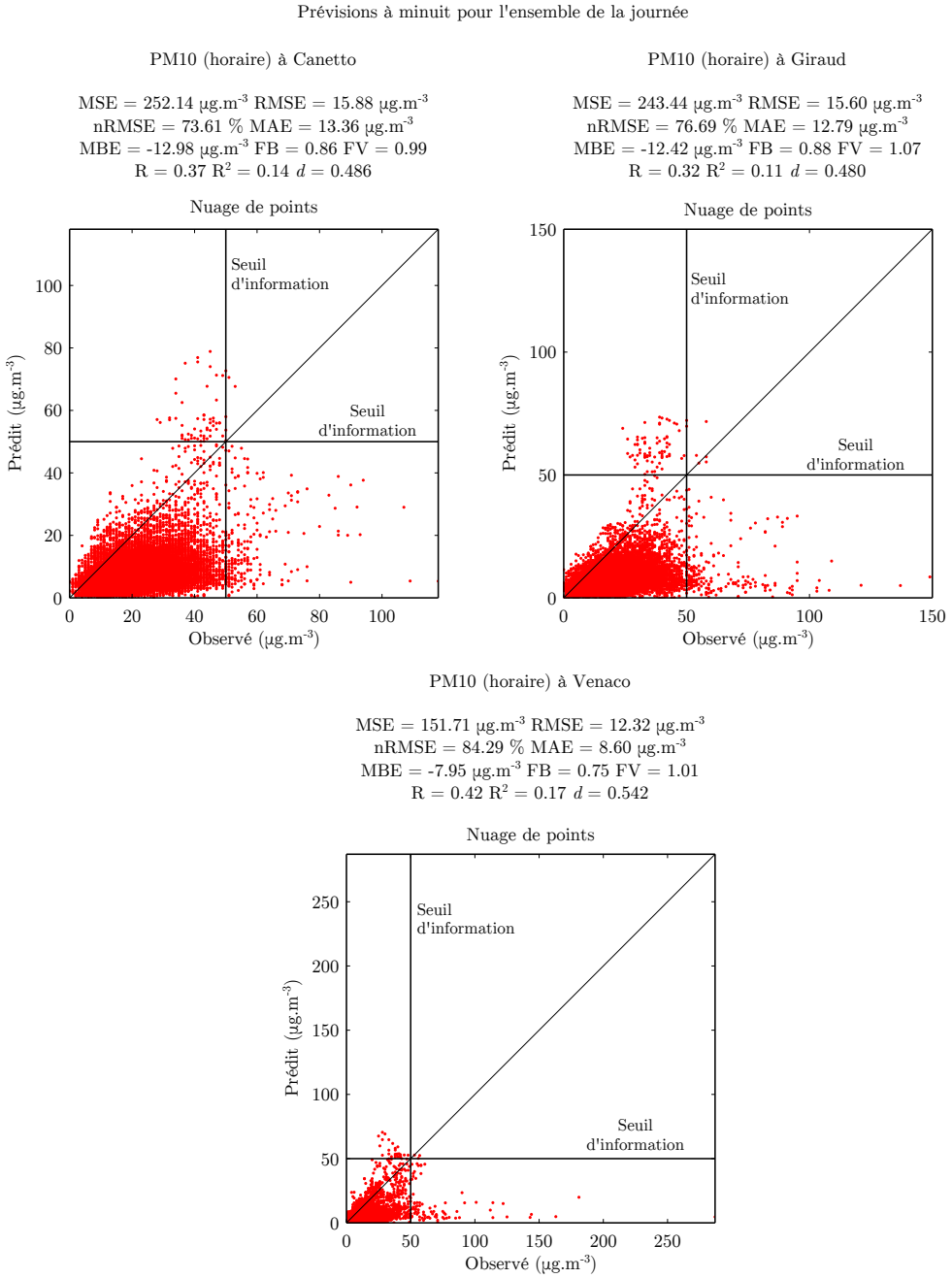


FIGURE 3.21 : Evaluation des prévisions de PM10 horaires brutes d'AIRES en Corse.

modèles, un point qui sera abordé à la section 5.3.1 (page 128).

Le manque de précision obtenue actuellement par le CTM justifie ces travaux. Afin que Qualitair Corse dispose d'un outil adapté, nous consacrerons nos effort à rechercher le meilleur modèle statistique de prévision. Nous développerons les applications présentées au chapitre 7, page 166, qui permettent de gérer ces modèles et de les utiliser de manière opérationnelle.

3.4 Conclusion

Dans ce chapitre, nous avons abordé la dynamique de la qualité de l'air en Corse, et nous nous sommes intéressés aux données qui la décrivent et que nous utiliserons dans nos travaux.

Nous avons donc à disposition pour notre étude sur la prévision plusieurs types de données : des mesures automatiques de concentrations en polluants fournies par Qualitair Corse, des mesures météorologiques provenant de l'AASQA et de Météo-France ainsi que des sorties du modèle AIRES.

Si l'on exclut les stations de proximité dévolues à la surveillance liée à certaines sources, à la fois locales et difficiles à prévoir avec des variables environnementales, à savoir les stations de type « trafic » et « industriel », il reste les stations « urbaines », « périurbaines » et « rurales ». C'est à ces stations qu'on essayera de prévoir les concentrations, dans le but de réaliser un modèle prédictif utile pour l'AASQA. Des informations sur les mesures de polluants nous intéressant sont présentées au tableau 3.5.

Les mesures de deux sites cependant ne pourront être systématiquement intégrées à nos jeux de données. Il s'agit du site rural de Venaco et du super-site CORSiCA lié à la campagne internationale ChArMEx. La raison est la taille des séries temporelles qui en proviennent, beaucoup plus petites que celles des sites urbains et périurbains. Leur usage aurait pour conséquence de réduire l'ensemble des jeux de données, affectant grandement l'apprentissage des réseaux neuronaux. Cependant, ces données seront très intéressantes pour apporter des informations sur les niveaux de fond d'ici quelques années.

L'usage de sorties de modèle AROME est une opportunité qui nous permettra d'utiliser des données météorologiques prédictives de qualité en entrée de modèle, un point qui est connu pour augmenter l'intérêt des réseaux de neurones en prévision de la qualité de l'air.

Nous avons mis en place une récupération automatique des données Météo-France citées, afin que les modèles que nous construirons soient utilisables de manière opérationnelle à Qualitair Corse, ce qui est le but de ces travaux. L'application « Aria Web », adaptée pour un fonctionnement sur un serveur tel que celui de Qualitair Corse a été développée pendant ces travaux. Elle gère la récupération des données et réalise les prévisions des modèles neuronaux entrainés. Elle sera présentée à la section 7.2.

Ces données doivent nous permettre de monter des modèles statistiques qui amélioreront la prévisibilité des concentrations. Les prévisions du modèle AIRES sont pour l'instant insuffisamment précises pour aider efficacement le personnel de Qualitair Corse à effectuer les prévisions quotidiennes.

Nous nous focaliserons particulièrement sur l'ozone et les particules, les deux polluants les plus problématiques en Corse. Les PM10 nous intéresseront particulièrement même s'ils sont plus difficiles à prévoir, que cela soit à l'aide de modèles déterministes ou à l'aide de modèles statistiques. Le dioxyde de soufre, très lié à l'utilisation de combustibles contenant du soufre, a des niveaux suffisamment bas en Corse pour que nous l'excluions complètement de nos travaux de prévision.

Chapitre 4

Méthodologie de prévision avec les modèles neuronaux

Les Réseau de Neurones Artificiels (RNA) sont une famille de modèles connexionnistes, inspirés des capacités des neurones biologiques. Leur champ d'action est large ; on les utilise en économétrie, en prévision, en reconnaissance de forme et dans bien d'autres domaines. Dans notre cas, ils sont utilisés pour la prévision en tant que modèle d'« ajustement de courbe » (« curve fitting »), c'est à dire qu'ils sont entrainés à reproduire une courbe de données, la sortie attendue, à partir d'autres données, les entrées. Leur apprentissage les amène à extrapoler les relations identifiées entre les variables d'entrée et de sortie. Ils se comportent en considérant que les mêmes conditions apporteront les mêmes conséquences.

Les RNA peuvent revêtir des formes différentes. Les Perceptrons MultiCouche (PMC) que nous avons utilisés en sont l'une des plus répandues. Dans ce chapitre, nous nous intéresserons tout d'abord à l'origine des RNA et à leur fonctionnement, à la section 4.1. Le PMC sera ensuite présenté plus en détail à la section 4.2 (page 87), où nous évoquerons son architecture et son fonctionnement.

Nous consacrerons ensuite la section 4.3 (page 90) à l'apprentissage automatique utilisé par ces modèles, et aux différents algorithmes utilisés pour cette étape. Cet apprentissage est crucial et les capacités des modèles prédictifs construits en dépendent. On verra les différents algorithmes candidats à cet apprentissage, leur initialisation et leur régulation, avant de les comparer pour sélectionner celui que l'on utilisera dans nos travaux.

Nous nous intéresserons à la section 4.4 (page 98) aux différents prétraitements qu'il convient d'utiliser pour préparer les données à l'apprentissage automatique et à la prévision opérationnelle.

Après avoir présenté ces différents aspects de l'usage du PMC, nous pourrons présenter à la section 4.5 (page 108) la méthodologie que nous avons adopté pour configurer les réseaux de neurones. Cette méthodologie nous permet d'optimiser les modèles prévisionnels aux problèmes qu'on leur soumet. Elle permet d'identifier les meilleures options de configuration des réseaux.

La majeure partie des expériences a été réalisée avec le logiciel Matlab (R2012a), et différentes fonctions de sa Neural Network Toolbox (NNToolbox). Afin de suivre notre méthodologie avec rigueur, nous avons utilisé une application que nous avons développée pour mener nos expérimentations (voir chapitre 7 page 166).

4.1 Modèles neuronaux

Les RNA sont des modèles numériques de calcul de la famille de l'Intelligence Artificielle (IA), constitués de plusieurs neurones formels interconnectés. Les neurones formels sont des automates inspirés de modèles de neurones biologiques du système nerveux central et traitent l'information qu'ils reçoivent en produisant un signal de sortie. Ces neurones artificiels prennent la forme de fonctions mathématiques à plusieurs variables. Les RNA, constitués d'un ensemble de ces neurones formels, suivent une approche connexionniste de la modélisation, c'est-à-dire qu'ils utilisent des unités de calcul simples et interconnectées pour modéliser des problèmes complexes.

Le computationnalisme était avant l'arrivée du connexionnisme la principale approche de l'IA. Selon ce paradigme, on peut considérer l'esprit comme un processus de traitement de l'information, c'est-à-dire un ensemble de règles définissant la manière d'interpréter les signaux extérieurs pour produire une réponse. Ces règles sont alors indépendantes du mécanisme permettant leur exécution. L'équivalent de la pensée humaine pourrait ainsi être créé grâce à un autre mécanisme que le corps humain et son cerveau, typiquement un ordinateur dont les algorithmes de fonctionnement seraient les mêmes que ceux de l'être humain.

Cette théorie suit une approche dite « top - down ». Pour schématiser ce point de vue, c'est en partant de ce qu'on pense être l'intelligence qu'on imagine un ensemble d'algorithmes équivalents, qui fonctionneraient quelle que soit la machine qui les exécute.

Le connexionnisme est un paradigme alternatif apparu dans les années 80 qui se différencie du computationnalisme. Selon une approche connexionniste, c'est en partant du fonctionnement même des mécanismes à l'origine de la pensée (le cerveau) que l'on peut développer une intelligence artificielle capable de produire l'équivalent de cette pensée. Cette approche suit donc un concept d'émergence, une approche « bottom – up ». Pour schématiser, c'est en partant de ce qu'on pense être le fonctionnement des éléments du cerveau qu'on développe un modèle équivalent capable d'agir comme un esprit humain. Ces deux paradigmes sont complémentaires et ne sont évidemment pas utilisés pour reproduire l'intelligence humaine mais offrent des résultats intéressants dans de nombreux domaines.

Les modèles connexionnistes que nous utiliserons, les RNA, font partie du domaine de l'apprentissage automatique, plus connu sous son appellation anglaise : le « machine learning ». La philosophie de l'apprentissage automatique réside dans l'optimisation des méthodes d'apprentissage elles-mêmes, rendant la machine capable de s'adapter à la problématique qu'on lui soumet. C'est cela qui place ces méthodes dans la famille de l'intelligence artificielle.

Cela peut créer une distance avec d'autres méthodes plus classique de la statistique, où l'on travaille surtout sur des modèles de données, et où c'est à partir des caractéristiques de ces données que l'on se permet de proposer des modèles prédictifs (par exemple, les modèles de la familles des ARMA (AutoRegressive Moving Average)). Les méthodes du « machine learning » et leur aspect « boîte noire » peuvent rebuter certains statisticiens pour leur opacité. Cependant, l'apprentissage automatique a toute sa place dans la statistique, puisqu'on parle de modèle entièrement basé sur les relations sous-jacentes qui existent entre des données. Breiman (2001*b*) a expliqué la différence entre ces deux points de vue et l'intérêt de travailler sur les algorithmes d'apprentissage propre au « machine learning ».

Les réseaux de neurones artificiels, souvent appelés simplement réseaux de neurones, ont été utilisés pour la première fois dans les travaux de McCulloch et Pitts (1943) afin de modéliser le comportement des neurones biologiques dont le cerveau est composé, bien avant l'apparition du courant connexionniste.

Les neurones biologiques sont des cellules très spécialisées et primordiales dans le fonctionnement du système nerveux, schématisés à la figure 4.1. Ils sont composés d'une partie centrale, le soma ou péricaryon, prolongée par les dendrites et l'axone. Les neurones sont interconnectés grâce aux synapses, zones de contact entre les axones et les dendrites.

La transmission nerveuse se fait par signaux électrochimiques. Les neurones reçoivent ces signaux sous forme de neurotransmetteurs chimiques au niveau des zones synaptiques. Des récepteurs sur les dendrites captent ces neurotransmetteurs et répondent en faisant varier le potentiel électrique de la membrane du neurone. Le potentiel postsynaptique correspond au signal unitaire produit au niveau d'une synapse. Au niveau du cône d'émergence de l'axone a lieu une sommation des potentiels postsynaptiques. Si cette somme dépasse le seuil d'excitabilité du neurone, alors le signal est propagé par l'axone vers d'autres synapses où seront libérés d'autres neurotransmetteurs vers d'autres dendrites.

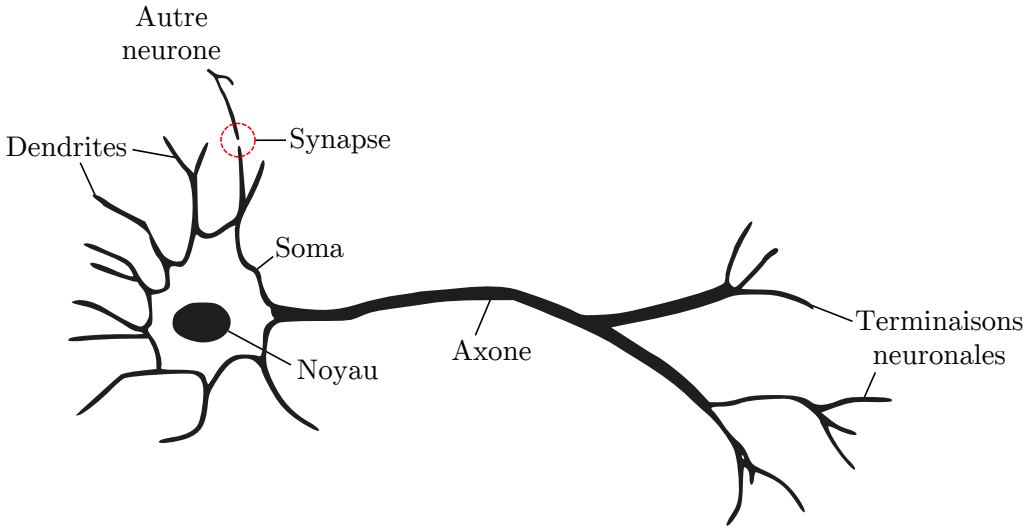


FIGURE 4.1 : Schéma d'un neurone biologique.

Ce seuil d'excitabilité donne à la transmission neuronale un caractère booléen, suivant une loi du « tout-ou-rien », qui à permis à McCulloch et Pitts d'établir un modèle logique de réseau de neurones. Les neurones formels sont l'équivalent mathématique des neurones biologiques.

Un neurone formel a plusieurs entrées et une sortie. Les entrées sont vues comme l'équivalent des dendrites. La valeur de chaque entrée est d'abord multipliée par un coefficient appelé poids et propre à chaque entrée. La somme de toutes les entrées ainsi pondérées est ensuite calculée et une valeur fixe, appelée biais et propre à chaque neurone, est éventuellement ajoutée à cette somme. La somme devient l'argument d'une fonction de transfert, reproduisant ainsi la sommation des potentiels électriques aux niveaux des cônes d'émergence des axones.

De nombreux types de fonctions sont utilisés comme fonction de transfert, notamment des fonctions logistiques. L'utilisation de telles fonctions, ou de fonction comme celle de Heaviside (échelon) permet de donner au neurone artificiel un équivalent des seuils d'excitabilité des neurones biologiques. Le résultat de ce calcul est la sortie du neurone. Elle est transmise aux neurones suivants, imitant le rôle de l'axone biologique.

Les neurones formels de McCulloch et Pitts avaient pour but d'étudier le fonctionnement des systèmes nerveux biologiques. C'est Hebb (1949) qui fournit un moyen d'utiliser ces modèles grâce à sa règle d'apprentissage ; la règle de Hebb. Cette règle stipule que plus les connexions entre les neurones biologiques sont utilisées, plus ces connexions deviennent efficaces. Cette règle

a permis de fixer les poids de neurones artificiels.

Le Perceptron est créé par Rosenblatt (1958) en reprenant les résultats de Hebb. Il s'agit d'un classifieur linéaire formé d'un unique neurone. A l'époque, d'autres réseaux de neurones linéaires ont été appliqués à des problèmes d'ingénierie comme le modèle ADALINE (Widrow et Hoff, 1960), mais les critiques émises par Minsky et Papert (1969) sur les limites du Perceptron et des classifieurs linéaires vont éloigner les chercheurs des RNA un certain temps. Ces critiques portent notamment sur l'incapacité des modèles neuronaux de l'époque de traiter certaines opérations, comme par exemple l'opérateur logique XOR (opérateur OU exclusif). Le PMC présenté par Rumelhart *et al.* (1986) et son apprentissage par « rétropropagation de l'erreur » ne possède plus ces défauts et est capable de traiter les problèmes non-linéaires. Cet apprentissage revêt un aspect particulièrement important dans l'usage de RNA.

La valeur des poids et les biais qui sont les paramètres des neurones est déterminée pendant la phase d'apprentissage automatique, qui peut être mise en œuvre grâce à différents algorithmes. On verra plusieurs algorithmes dont la rétropropagation de l'erreur à la section 4.3 page 90. Cet apprentissage peut être supervisé ou non-supervisé. Lors de l'apprentissage, l'utilisateur fournit à l'algorithme des données d'apprentissage qui serviront d'exemples au modèle. Si l'apprentissage est supervisé, ces données comprendront des données cibles indiquant le résultat auquel doit parvenir le RNA pour chaque cas présent dans les données d'apprentissage. L'apprentissage non-supervisé n'utilise pas de données cibles mais laisse le réseau adapter ses poids et biais selon d'autres méthodes dépendant moins de l'apriori de l'utilisateur sur le résultat.

Un RNA entrainé, c'est-à-dire ayant passé la phase d'apprentissage, peut être utilisé pour de nombreux problèmes. En effet, les caractéristiques d'un RNA sont très nombreuses (nombre et spécificités des neurones, topologie du réseau, algorithme d'apprentissage, etc.) et il existe de nombreux archétypes de RNA adaptés à de nombreuses tâches spécifiques. Ils sont notamment utilisés dans des domaines comme la reconnaissance de formes, le traitement du signal ou la modélisation de fonctions et la prédiction de séries temporelles, qui les placent à la fois dans la famille des méthodes d'intelligence artificielle et dans celle des modèles statistiques.

4.2 Perceptron Multicouche

Le Perceptron Multicouche est un type de RNA qui se caractérise par son architecture « feedforward » (c'est à dire non-bouclée), composée d'une succession d'au moins deux couches de neurones ; une couche cachée et une couche de sortie. Ce type de modèle est utilisé pour des problèmes de classification supervisée ou des problèmes de régression. Il est employé en prévision de série temporelle (Zhang *et al.*, 1998), et notamment en qualité de l'air. Le Perceptron Multi-Couche (PMC) est particulièrement populaire grâce à ses capacités lui permettant d'approximer toute fonction « lisse », c'est-à-dire infiniment dérivable (Hornik *et al.*, 1989).

Les variables d'entrées d'un PMC appliqué à la prévision sont des séries temporelles. Il peut s'agir de séries temporelles décrivant la variable endogène, ou des variables exogènes. Les valeurs de chaque séries temporelle qui sont fournies en entrées sont regroupées dans le vecteur $\mathbf{x}$. Il est fourni au PMC qui réalise la prévision $\hat{y}$.

Un schéma de PMC est donné à la figure 4.2. Nous y représentons les opérations gérées par un neurone caché : toutes les entrées sont multipliées par un poids spécifique, puis sont sommées en ajoutant un biais. Cette somme devient l'argument d'une fonction de transfert, ou fonction d'activation, dont le résultat devient la sortie du neurone.

La figure 4.2 présente un PMC à une couche cachée de m neurones et un neurone de sortie,

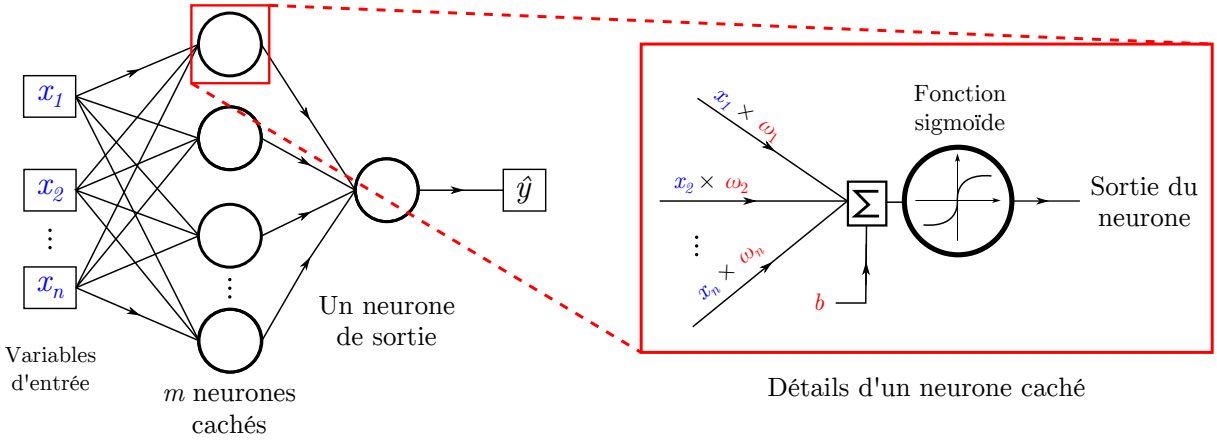


FIGURE 4.2 : Schéma d'un PMC avec mise en valeur d'un neurone, ses poids ω et son biais b.

ayant n variables d'entrée. Son équivalent mathématique est une fonction paramétrique f telle que :

$$f(\mathbf{x}, \omega, \mathbf{b}) = \sum_{j=1}^{m} \omega_{js} \, . \left(g(\sum_{i=1}^{n} \omega_{i,j} \, . \, x_i + b_j) \right) + b_s = \hat{y} \tag{4.1}$$

avec $\mathbf{x}$ le vecteur d'entrée contenant la valeur des n variables, ω le vecteur contenant les poids des neurones et $\mathbf{b}$ le vecteur contenant leurs biais. $\hat{y}$ est la sortie du modèle, b_j le biais du neurone caché j et $\omega_{i,j}$ son poids affecté à l'entrée x_i, g la fonction de transfert des neurones cachés, b_s le biais du neurone de sortie et $\omega_{j,s}$ son poids affecté à la sortie du neurone caché j.

La fonction de transfert g que nous avons utilisée pour les neurones cachés est une fonction tangente hyperbolique, représentée en figure 4.3 et de formule :

$$g(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{4.2}$$

Cette fonction représente une transition douce entre ses bornes, -1 et 1. L'avantage de son usage par rapport à une fonction échelon représentant une transition discontinue est qu'elle est infiniment dérivable. Sa dérivabilité est importante dans le cadre de l'apprentissage. Ce type de fonction de transfert pour les neurones cachés confère au PMC son aspect non-linéaire. Le PMC correspond donc à un modèle de régression non-linéaire des variables d'entrée.

L'utilisation d'une fonction linéaire pour l'ensemble des neurones aurait mené à un modèle régressif linéaire. On remarque dans l'équation 4.1 que la fonction du neurone de sortie n'est pas explicitée. En effet, nous utiliserons systématiquement une fonction de transfert linéaire pour le neurone de sortie, qui correspond à la multiplication par un facteur. On peut considérer que ce facteur est intégré dans les paramètres du neurone de sortie.

Utiliser un modèle non-linéaire est important, puisque les relations entre les concentrations d'ozone et les variables météorologiques sont non-linéaires (Gardner et Dorling, 2000; Schlink *et al.*, 2003). La dynamique des concentrations en particules et ses relations avec les autres variables sont également non-linéaires (Ionescu, 2013). Cela fait partie des raisons ayant menées au choix du PMC.

Pour réaliser un modèle prédictif à partir d'un PMC, il convient de décaler dans le temps les données d'apprentissage correspondant aux variables d'entrée et celles correspondant à la variable de sortie. Ce décalage correspondra à l'horizon de la prévision. Dans le cas de séries

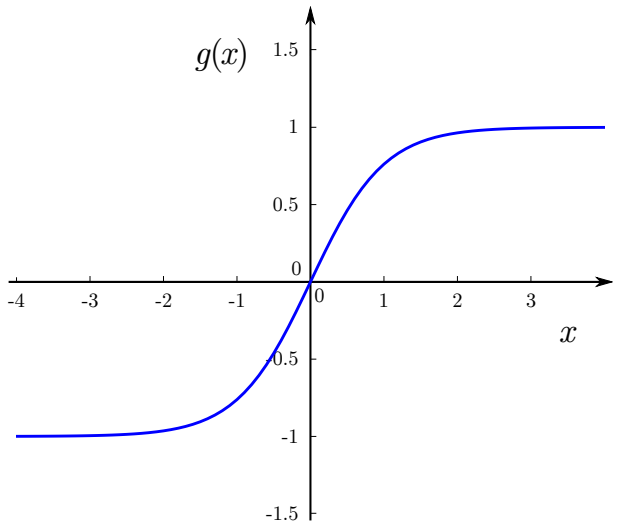


FIGURE 4.3 : Représentation de la fonction tangente hyperbolique $g(x)$.

temporelles horaires par exemple (une donnée chaque heure), on décale la série temporelle cible de manière à faire correspondre à chaque heure la valeur qui serait observée x heures plus tard. Ainsi, lors de l'apprentissage, le PMC s'entraine à prévoir les valeurs futures, à l'horizon $h + x$.

Ce décalage dans le futur est appliqué à la série temporelle cible. On peut également appliquer des décalages dans le passé aux variables d'entrée. L'utilisation de valeurs précédant les observations les plus récentes peut apporter de l'information utile pour la prévision au modèle. On dit qu'on applique un délai à la série temporelle. Par exemple, les réseaux de neurones réalisant des prévisions uniquement à l'aide de la variable endogène utilisent plusieurs entrées, chacune correspondant à un différent délai appliqué à la série temporelle endogène.

Une variable pourra donc être utilisée plusieurs fois en entrée d'un PMC, en lui appliquant un délai différent à chaque fois. Afin de préciser des délais, on considère quand on crée un modèle prédictif qu'il fonctionne à l'heure de référence h, pour prévoir à l'horizon $h+x$ (x étant l'horizon). On dira par exemple qu'on utilise une variable à $h-2$ pour dire qu'elle a subi un délai de deux heures.

Certaines variables sont des sorties d'un autre modèle prédictif, souvent des modèles météorologiques. On peut alors utiliser les prévisions de ces modèles, dont plusieurs échéances sont disponibles. On dira, par rapport à l'heure de référence, qu'on utilise la variable à $h+y$, y étant l'échéance utilisée. Cette thèse a pour but de proposer un modèle à l'usage opérationnel. Il sera donc constamment pris soin d'utiliser les sorties de modèles à des échéances qui sont disponibles à l'heure h en usage opérationnel.

La figure 4.4 montre un exemple de PMC avec cinq entrées : la concentration d'ozone sans délai et avec un délai d'une heure, la concentration en NO_2 avec des délais de une et deux heures et la température issue du modèle AROME à l'échéance $h+24$. Ce PMC prévoit la concentration d'ozone à l'horizon $h + 24$.

Les données utilisées par le PMC subissent en général plusieurs prétraitements avant d'être utilisées. Ces prétraitements, évoqués à la section 4.4, page 98, doivent rendre ces données plus intelligibles pour le réseau. Ils concernent autant les données d'entrées que les données cibles. Dans le cas d'un prétraitement de la cible, l'apprentissage du PMC entraine le réseau à prévoir une variable prétraitée. Un post-traitement est alors nécessaire pour retomber sur la variable que l'on veut prévoir.

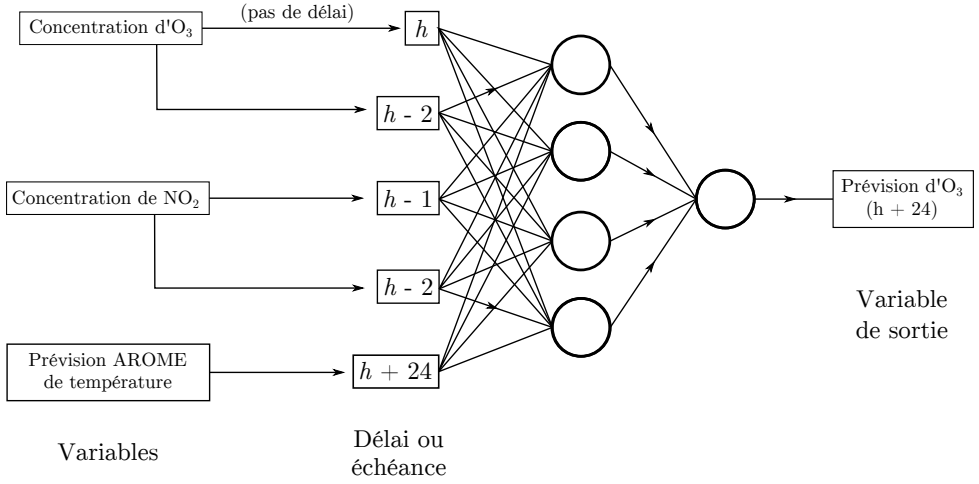


FIGURE 4.4 : PMC avec en entrée cinq variables pour la prévision d'O_3 à l'horizon $h + 24$.

Une fois les données préparées, l'apprentissage du réseau permet de le rendre opérationnel. Nous allons maintenant voir comment est mené cet apprentissage.

4.3 Apprentissage

L'apprentissage automatique sert à fixer la valeur de tous les paramètres du réseau, à savoir les poids et les biais de ses neurones. Pour cela, un algorithme d'apprentissage est utilisé pour optimiser les valeurs de ces paramètres, de manière à faire correspondre au mieux les sorties du modèle aux valeurs observées attendues.

Pour cela, un jeu de données d'apprentissage est utilisé par l'algorithme. Ce jeu regroupe plusieurs séries temporelles, qui décrivent les variables d'entrée et de sortie. Ainsi, on dispose d'exemples des sorties attendues qui permettent un apprentissage supervisé : Une série temporelle cible contenant les valeurs à prévoir par le réseau est fournie pour l'apprentissage (par opposition, les apprentissages non-supervisés n'utilisent aucune cible).

Les algorithmes que nous avons utilisés pour l'apprentissage supervisé sont des algorithmes « à direction de descente ». Ce type d'algorithme à pour objectif de trouver le minimum d'une fonction paramétrique, qui dans notre cas représente l'erreur quadratique moyenne (Mean Squared Error (MSE)) commise par le modèle, avec comme paramètres les poids et biais du réseau. Pour trouver ce minimum, ce type d'algorithme estime à chaque itération des informations concernant les dérivées de cette fonction. Dans l'espace des paramètres, une direction apparait alors comme la meilleure à suivre pour diminuer le plus efficacement la valeur de cette fonction. Les paramètres sont modifiés pour suivre cette direction, d'où le nom de cette famille d'algorithme.

Nous avons vu à la section 2.3.4 qu'en prévision de la qualité de l'air à l'aide de RNA, plusieurs algorithmes d'apprentissage étaient utilisés. Nous avons donc envisagé l'usage des principaux d'entre eux (voir tableau 2.1 page 47), l'algorithme de Descente de Gradient (DG), de Levenberg – Marquardt (LM), de Broyden – Fletcher – Goldfarb – Shanno (BFGS) et du SCG (Scaled Conjugate Gradient). Le principe de fonctionnement de ces algorithmes sera expliqué à la section suivante.

A la section 4.3.2 (page 95), nous aborderons l'initialisation des paramètres des PMC, à partir de laquelle l'algorithme cherche les meilleures valeurs pour les poids et biais. Nous aborderons

ensuite un point important à la section 4.3.3 (page 95) : le principe de parcimonie. Nous verrons comment l'apprentissage est régulé pour éviter le sur-apprentissage, un apprentissage « par cœur » des données d'apprentissage qui nuit aux capacités de généralisation du modèle et aux prévisions opérationnelles.

Enfin, après avoir présenté ces aspect de l'apprentissage et des algorithmes candidats, nous comparerons à la section 4.3.4 (page 97) les performances obtenues avec ces derniers, appliqués à notre problématique de prévision de la qualité de l'air en Corse.

4.3.1 Fonctionnement des algorithmes

Cette section est destinée à l'explication du fonctionnement des algorithmes de Descente de Gradient (DG) et de Levenberg – Marquardt (LM) en particulier. Les algorithmes BFGS et SCG sont décrits en annexe D, page 223.

Le rôle des algorithmes d'apprentissage est donc d'optimiser une fonction paramétrique à plusieurs variables f, qui correspond à la représentation mathématique du PMC. Les paramètres de cette fonction correspondent aux poids et biais du réseau, et les variables à ses entrées. Plusieurs approches sont possibles et ont donné naissance à des algorithmes différents.

Voyons comment est défini l'objectif de cette optimisation. Les m paramètres du modèle (n'importe quel modèle paramétrique mais dans notre cas un PMC) sont regroupés dans le vecteur $\mathbf{p}$ de taille m, les entrées dans la matrice $\mathbf{x}$ de taille $l \times n$ avec l le nombre de variables et n la taille de l'échantillon, et les sorties dans le vecteur $\hat{\mathbf{y}}$ de taille n. Le PMC correspond donc à la fonction :

$$f(\mathbf{x}_i, \mathbf{p}) = \hat{y}_i \tag{4.3}$$

pour 1 < i < n. Pendant l'apprentissage, l'algorithme utilisé doit trouver le vecteur de paramètres $\mathbf{p}$ qui optimise f afin d'en diminuer l'erreur. La différence entre les valeurs $\hat{\mathbf{y}}$ fournies par la fonction f et les valeurs observées $\mathbf{y}$ est donnée par r, le résidu de f :

$$r(\mathbf{x}_i, \mathbf{p}) = y_i - f(\mathbf{x}_i, \mathbf{p}) \tag{4.4}$$

On utilise en général la somme des carrés de r, qui permet de quantifier l'erreur commise par le modèle pour l'ensemble des n exemples fournis par les données d'apprentissage. Prenons

$$\begin{aligned} S(\mathbf{p}) &= \frac{1}{2}\sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i, \mathbf{p})\right)^2 \\ S(\mathbf{p}) &= \frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{p})^2 = \frac{1}{2}\|\mathbf{r}(\mathbf{p})\|^2 \end{aligned} \tag{4.5}$$

avec un facteur $\frac{1}{2}$ arbitraire mais qui simplifiera les équations futures.

C'est donc cette fonction S que les algorithmes à direction de descente doivent minimiser pour entrainer le réseau de neurones. Les algorithmes d'apprentissage que nous utilisons sont des algorithmes itératifs, qui à chaque itération modifient les paramètres $\mathbf{p}$ en recherchant le minimum de S. Ils se basent sur l'erreur S entre les prévisions et les observations, ainsi que sur des estimations des dérivées au premier et second ordre de cette erreur.

La fonction S a un vecteur de m paramètres $\mathbf{p}$ pour lequel elle est minimale, vecteur qui correspond à la configuration pour laquelle la fonction f est optimisée. La fonction S peut se représenter comme une surface dans un espace euclidien à m dimensions, m étant le nombre de

paramètres de la fonction f. A partir d'un point de cette surface correspondant au vecteur $\mathbf{p}$ initial, l'algorithme va parcourir cette surface en cherchant à chaque itération à « descendre » vers le minimum de f. L'utilisation de dérivées de la fonction d'erreur S permet de privilégier une direction dans l'espace des paramètres vers laquelle cette erreur semble diminuer.

Cette descente s'effectuera le long d'un vecteur qu'on notera $\mathbf{a}$ et qui devra être estimé à chaque itération de l'algorithme. Le vecteur de paramètres de l'itération suivante correspondra donc à $\mathbf{p} + \mathbf{a}$. Les itérations se poursuivront jusqu'à obtenir la précision désirée (S suffisamment petit), ou dans notre cas, jusqu'à l'interruption du processus par régulation en vue d'éviter le sur-apprentissage (voir section 4.3.3, page 95). Le vecteur $\mathbf{a}$ donnant la direction de descente à chaque itération, certains algorithmes adaptent le pas de descente noté α pour fixer l'amplitude de la descente.

L'algorithme de DG est le premier algorithme à avoir été utilisé pour entrainer un PMC. On le retrouve également sous le nom de rétro-propagation de l'erreur, ou Back Propagation (BP) (Rumelhart *et al.*, 1986). L'algorithme de DG calcule à chaque itération un nouveau vecteur de paramètres $\mathbf{p}_{k+1}$ grâce à un vecteur d'incrément pour obtenir $\mathbf{p}_{k+1} = \mathbf{p}_k + \mathbf{a}_k$. Le vecteur $\mathbf{a}_k$ est déterminé pour aller dans la direction inverse de celle du gradient de $S(\mathbf{p})$ afin de se rapprocher de son minimum, d'où le nom de la méthode.

Le gradient d'un champ scalaire est le vecteur pointant dans la direction de la plus forte augmentation du champ, d'amplitude égale au taux d'accroissement. Cela correspond pour une fonction à plusieurs variables au vecteur contenant ses dérivées partielles.

$$\nabla S(\mathbf{p}_k) = \begin{pmatrix} \dfrac{\partial S(\mathbf{p}_k)}{\partial p_1} \\ \cdots \\ \dfrac{\partial S(\mathbf{p}_k)}{\partial p_m} \end{pmatrix} \tag{4.6}$$

Introduisons $\mathbf{J}_k$ la matrice jacobienne de $\mathbf{r}$, c'est-à-dire la matrice de dimension $n \times m$ contenant les dérivées partielles de $\mathbf{r}$ à l'itération k :

$$\mathbf{J}_k = \begin{pmatrix} \nabla r(\mathbf{x_1}, \mathbf{p}_k)^\mathbf{T} \\ \nabla r(\mathbf{x_2}, \mathbf{p}_k)^\mathbf{T} \\ \vdots \\ \nabla r(\mathbf{x_n}, \mathbf{p}_k)^\mathbf{T} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial r(\mathbf{x}_1, \mathbf{p}_k)}{\partial p_1} & \dfrac{\partial r(\mathbf{x}_1, \mathbf{p}_k)}{\partial p_2} & \cdots & \dfrac{\partial r(\mathbf{x}_1, \mathbf{p}_k)}{\partial p_m} \\ \dfrac{\partial r(\mathbf{x}_2, \mathbf{p}_k)}{\partial p_1} & \dfrac{\partial r(\mathbf{x}_2, \mathbf{p}_k)}{\partial p_2} & \cdots & \dfrac{\partial r(\mathbf{x}_2, \mathbf{p}_k)}{\partial p_m} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial r(\mathbf{x}_n, \mathbf{p}_k)}{\partial p_1} & \dfrac{\partial r(\mathbf{x}_n, \mathbf{p}_k)}{\partial p_2} & \cdots & \dfrac{\partial r(\mathbf{x}_n, \mathbf{p}_k)}{\partial p_m} \end{pmatrix} \tag{4.7}$$

en utilisant l'exposant $^\mathbf{T}$ pour désigner la matrice transposée.

Chacune des lignes de $\mathbf{J}_k$ correspond au gradient de la fonction $r(\mathbf{x}_i, \mathbf{p}_k)$, le résidu de f pour le couple d'entrées $\mathbf{x}_i$ et de sortie y_i, avec les paramètres $\mathbf{p}$ de l'itération k.

Le gradient d'une erreur quadratique $S(\mathbf{p})$ (équation 4.5) s'écrit alors :

$$\nabla S(\mathbf{p}_k) = \mathbf{J}_k^\mathbf{T} \mathbf{r}(\mathbf{p_k}) \tag{4.8}$$

Le vecteur d'incrément $\mathbf{a}$ peut être pris égal à l'opposé du gradient, afin d'aller dans la direction de la minimisation de $S(\mathbf{p})$:

$$\mathbf{a}_k = -\nabla S(\mathbf{p}_k) = -\mathbf{J}_k^\mathbf{T} \mathbf{r}(\mathbf{p}_k) \tag{4.9}$$

Le vecteur d'incrément $\mathbf{a}$ correspond à la direction de plus forte descente au point $\mathbf{p}$ où il a été calculé. La surface de l'erreur $S(\mathbf{p})$ n'étant pas purement linéaire, le pas de descente doit être suffisamment petit pour rester dans la zone où le gradient reste valable. Plus l'on s'éloigne de $\mathbf{p}$, plus le gradient a des risques de ne plus correspondre à la direction de plus forte pente. A chaque itération, on choisit donc le nouveau vecteur de paramètre $\mathbf{p}_{k+1}$ tel que

$$\mathbf{p}_{k+1} = \mathbf{p}_k + \alpha_k \mathbf{a}_k \tag{4.10}$$

avec α_k un scalaire permettant de fixer l'amplitude de la descente. Il peut être déterminé grâce à une méthode de recherche linéaire le long de la direction donnée par $\mathbf{a}_k$. Le but de cette recherche est de trouver la valeur de α_k pour laquelle $S(\mathbf{p}_{k+1})$ est minimale, on peut chercher la valeur de α_k pour laquelle la dérivée de $S(\mathbf{p}_{k+1})$ est nulle :

$$\frac{\partial S(\mathbf{p}_{k+1})}{\partial \alpha_k} = \frac{\partial S(\mathbf{p}_k + \mathbf{a}_k \alpha_k)}{\partial \alpha_k} = 0 \tag{4.11}$$

Pour éviter ce calcul, de nombreuses méthodes permettent de fixer la valeur pour α_k, comme par exemple lui assigner l'amplitude du gradient.

Un des inconvénients de cet algorithme est qu'il converge très lentement, l'information du premier ordre contenu dans le gradient ne pouvant être extrapolée loin du point où elle est calculée. D'autres algorithmes plus complexes estiment des informations du second ordre, ce qui leur permet une meilleure convergence vers le minimum de S, parfois au prix de lourds calculs.

La méthode de Newton est à l'origine des algorithmes à direction de descente du second ordre (Dreyfus, 2008). Elle nous mènera à l'algorithme de Gauss-Newton qui permet d'appréhender l'algorithme de LM. L'approche de Newton consiste en l'utilisation d'un développement de Taylor du premier ordre de ∇S, allant jusqu'au terme linéaire (Kelley, 2003). Introduisons $\mathbf{H}$ la matrice hessienne, c'est-à-dire la matrice carrée regroupant les dérivées partielles secondes d'une fonction. C'est une matrice par définition symétrique. On peut représenter $\mathbf{H}_{r_i}$ la hessienne de $\mathbf{r}$ calculée pour un vecteur d'entrée $\mathbf{x}_i$ (la hessienne complète pour les n mesures étant de dimension $n \times m \times m$ et difficile à écrire) :

$$\mathbf{H}_{r_i} = \begin{pmatrix} \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_1^2} & \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_1 \partial p_2} & \cdots & \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_1 \partial p_m} \\ \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_2 \partial p_1} & \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_2^2} & \cdots & \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_2 \partial p_m} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_m \partial p_1} & \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_m \partial p_2} & \cdots & \dfrac{\partial^2 r(\mathbf{x}_i, \mathbf{p})}{\partial p_m^2} \end{pmatrix} \tag{4.12}$$

On dérive l'expression du gradient de S décrit en (4.8) et on obtient la hessienne de S :

$$\mathbf{H}_S = \mathbf{J}^\mathbf{T}\mathbf{J} + \sum_{i=1}^{n} r(\mathbf{x}_i)\mathbf{H}_{r_i} \tag{4.13}$$

Le développement de Taylor du premier ordre de ∇S prend donc la forme :

$$\begin{aligned} \nabla S(\mathbf{p} + \mathbf{a}) &= \nabla S(\mathbf{p}) + \mathbf{H}_S(\mathbf{p})\,\mathbf{a} \\ \nabla S(\mathbf{p} + \mathbf{a}) &= \mathbf{J}^\mathbf{T}\mathbf{r} + \left(\mathbf{J}^\mathbf{T}\mathbf{J} + \sum_{i=1}^{n} r(\mathbf{x}_i)\mathbf{H}_{r_i}\right)\mathbf{a} \end{aligned} \tag{4.14}$$

L'algorithme de Gauss-Newton utilise une approximation du calcul de $\mathbf{H}_S$ décrit dans la méthode de Newton. Dans le développement de Taylor de ∇S en $\mathbf{p}$ d'ordre 1 (équation 4.14), on néglige le terme comprenant la hessienne $\mathbf{H}_{r_i}$. On a alors :

$$\nabla S(\mathbf{p}+\mathbf{a}) = \mathbf{J^T r} + \mathbf{J^T J a} \tag{4.15}$$

L'annulation du gradient donne donc :

$$\begin{aligned} \mathbf{J^T r} + \mathbf{J^T J a} &= 0 \\ \mathbf{J^T J a} &= \mathbf{J^T r_{p_k}} \end{aligned} \tag{4.16}$$

Si on a $n > m$, c'est-à-dire si l'on dispose de plus d'observations au sein de notre échantillon de données qu'il y a de paramètres à la fonction f (ce qui sera toujours le cas), alors l'équation 4.16 correspond à un système de n équations à m inconnues, et a une solution permettant de trouver $\mathbf{a}$.

C'est ce calcul que va réaliser l'algorithme à chaque itération. Le minimum ne sera pas atteint en une fois à cause du développement de Taylor du premier ordre (équation 4.14) ainsi que de l'approximation du terme de second ordre $\mathbf{H}_S$. On évite ainsi le calcul de $\mathbf{H}_{r_i}$, l'erreur commise se corrigeant itérativement. La jacobienne $\mathbf{J}$ doit être recalculée à chaque itération. Quand l'erreur quadratique moyenne est satisfaisante selon les critères de l'étude réalisée, le vecteur $\mathbf{p}$ de l'itération courante est conservé pour paramétrer la fonction f ainsi optimisée.

Plusieurs variantes de cet algorithme ont depuis été développées, dont la plus utilisée est l'algorithme de Levenberg-Marquardt.

L'algorithme de Levenberg-Marquardt, qui est également connu sous la dénomination anglaise « damped least-squares » (moindres carrés amortis) est une amélioration de l'algorithme de Gauss-Newton, très largement utilisée dans de nombreux domaines.

Levenberg a modifié l'algorithme de Gauss-Newton (Levenberg, 1944) en y introduisant un facteur d'amortissement λ. L'équation 4.16 devient :

$$(\mathbf{J^T J} + \lambda \mathbf{I})\mathbf{a} = \mathbf{J^T r_p} \tag{4.17}$$

avec $\mathbf{I}$ la matrice identité. Le facteur λ est positif et change à chaque itération. Une petite valeur de λ rapproche l'algorithme de celui de Gauss-Newton, tandis que si λ est grand, l'algorithme se rapprochera d'une descente de gradient. En effet, étant donnée la formule du gradient (équation 4.8), une valeur élevée de λ aura pour effet d'atténuer la direction favorisée par le terme $\mathbf{J^T J}$ au profit du membre de droite de l'équation 4.17 qui correspond à celle du gradient.

Si lors d'une itération l'erreur quadratique diminue avec λ petit, cela indique que le développement de Taylor de $S(\mathbf{p})$ de la méthode Gauss-Newton fonctionne et λ est diminué en conséquence afin de se rapprocher de la direction de Gauss-Newton, permettant la convergence rapide vers la solution et évitant la lenteur de la descente de gradient. Si au contraire l'erreur augmente, alors λ est augmenté de ce même facteur, puis l'itération est reprise, sa direction étant plus proche de celle du gradient. Chaque itération fait donc évoluer $\mathbf{p}$ dans une direction se trouvant entre celle du gradient et celle fournie par la méthode de Gauss-Newton.

Marquardt (1963), n'ayant alors pas connaissance des travaux de Levenberg, met au point un algorithme similaire où la matrice identité $\mathbf{I}$ est remplacée par la matrice contenant les éléments de la diagonale de $\mathbf{J^T J}$. On obtient ainsi

$$(\mathbf{J}_k^\mathbf{T}\mathbf{J}_k + \lambda\ diag(\mathbf{J}_k^\mathbf{T}\mathbf{J}_k))\mathbf{a}_k = \mathbf{J}_k^\mathbf{T}\mathbf{r}_k \tag{4.18}$$

la formule de l'algorithme de LM.

Les algorithmes BFGS et SCG sont deux autres algorithmes du second ordre que nous avons utilisés, et comparés au LM et à la DG. Le lecteur intéressé trouvera plus d'informations à leur sujet à l'annexe D (page 223). Nous allons aborder à la section suivante l'initialisation des neurones qui permet de donner une valeur à **p** avant la première itération de l'apprentissage.

4.3.2 Initialisation des paramètres

Les algorithmes à direction de descente nécessitent un point de départ dans l'espace des paramètres, d'où commencer pour converger vers le point minimisant l'erreur du modèle. Il est donc nécessaire d'initialiser le vecteur de paramètres **p**. Cette étape est importante car elle peut influencer les performances de l'apprentissage.

Cette initialisation est parfois réalisée en fixant tous les poids et biais à 0. Cependant, cette valeur n'a pas de raison particulière d'être sélectionnée, et son utilisation répétée (ou l'utilisation de toute autre constante) crée un biais. Une telle initialisation favorise la convergence vers un optimum local lors de l'apprentissage, une situation qui doit être évitée. Une autre approche consiste à initialiser aléatoirement les paramètres. D'autres méthodes que l'initialisation aléatoire pure, (mais qui gardent une composante aléatoire) ont montré de meilleurs résultats.

L'algorithme de Nguyen-Widrow par exemple utilise une initialisation aléatoire des paramètres, mais en distribuant la région active de chaque neurone uniformément sur l'espace de ses entrées (Nguyen et Widrow, 1990). Le fait d'utiliser une initialisation ayant une composante aléatoire implique que pour une configuration donnée, on n'aura pas le même modèle si on procède à plusieurs apprentissages.

C'est pour cette raison qu'il sera nécessaire de procéder à plusieurs apprentissages lorsqu'on veut expérimenter une configuration particulière. Cela alourdit le processus expérimental mais donne une précieuse information sur la robustesse des résultats obtenus pour une configuration donnée. Nous avons utilisé l'algorithme de Nguyen-Widrow pour initialiser nos PMC.

4.3.3 Parcimonie et régulation de l'apprentissage

Respecter le principe de parcimonie est important quand on utilise des modèles paramétriques tels que le PMC. Ce principe s'applique en utilisant le moins de paramètres possible, afin de limiter la capacité de spécialisation du modèle au profit de sa capacité de généralisation. Lors de l'apprentissage, l'algorithme va donc configurer les paramètres **p** du modèle afin de lui faire intégrer les relations existant entre les données d'entrée et la sortie désirée. Le jeu de données utilisé à cette fin s'appelle le jeu d'entrainement, ou le jeu d'apprentissage, voire le set d'apprentissage. Une fois entrainé, le modèle est prêt à être utilisé de manière opérationnelle. L'objectif de cet apprentissage est donc de rendre le modèle capable d'extrapoler les relations sous-jacentes entre entrées et sorties qu'il aura su capter pendant l'entrainement lors d'une future utilisation, en utilisant de nouvelles données d'entrées.

Il existe un risque que le modèle se sur-spécialise sur les données d'entrainement lors de l'apprentissage. En quelque sorte, le modèle peut apprendre « par cœur » les relations entre entrées et sortie du jeu d'apprentissage, et avoir une faible capacité de généralisation. Ce problème que l'on appelle le sur-apprentissage (« over-learning » en anglais) prend une importance d'autant plus grande que l'on a de paramètres (poids et biais) dans un modèle. Le nombres de poids et de biais dépend du nombre de neurones et de variables d'entrée, d'où l'importance de respecter le principe de parcimonie en limitant la taille du réseau et le nombre de variables. Un petit jeu d'entrainement (en taille d'échantillon) n'apportant pas assez d'exemples pour l'apprentissage

est également propice au sur-apprentissage.

Une bonne capacité de généralisation assure que le modèle obtiendra toujours une bonne précision lorsqu'il sera confronté à de nouvelles données, n'ayant pas été présentées lors de l'apprentissage. C'est évidement le cas lors d'une utilisation dite « on line », ou opérationnelle du modèle, quand on s'en sert pour réaliser les prévisions de routine. Un modèle avec une mauvaise capacité de généralisation, due à un sur-apprentissage, fera de mauvaises prévisions dans ce contexte.

Afin de respecter ce principe, il est nécessaire de limiter les variables d'entrée, qui impliquent un nombre proportionnel de poids et biais du réseau. Le nombre de neurones doit également être limité. Ainsi, il est nécessaire de trouver un compromis permettant d'avoir suffisamment de variables et de neurones tout en respectant le principe de parcimonie. Il existe plusieurs méthodes de sélection de variables (voir section 5.1, page 113) qui permettent de ne conserver que les meilleures d'entre elles pour la prévision et d'en limiter la redondance. Il existe également des méthodes de pruning (élagage) qui permettent de supprimer les paramètres superflus d'un RNA.

Il reste nécessaire d'éviter le sur-apprentissage en utilisant par exemple la technique largement répandue de l'« early stopping ». Cette technique utilise un second jeu de données pendant l'apprentissage, le jeu de validation. A chaque itération de l'algorithme d'apprentissage, la précision du modèle est calculée sur le jeu de validation (dont les données ne sont pas utilisées par l'algorithme d'apprentissage). Au fur et à mesure que l'algorithme améliore la précision sur le jeu d'apprentissage, l'erreur diminue également sur le jeu de validation. Cependant, quand le modèle commence à se sur-spécialiser sur les données du jeu d'apprentissage, l'erreur se met à augmenter sur le jeu de validation. L'entrainement est alors interrompu. Afin de s'assurer que l'on n'interrompt pas trop tôt l'apprentissage, ce dernier est interrompu quand l'erreur de validation recommence à augmenter pendant six itérations consécutives.

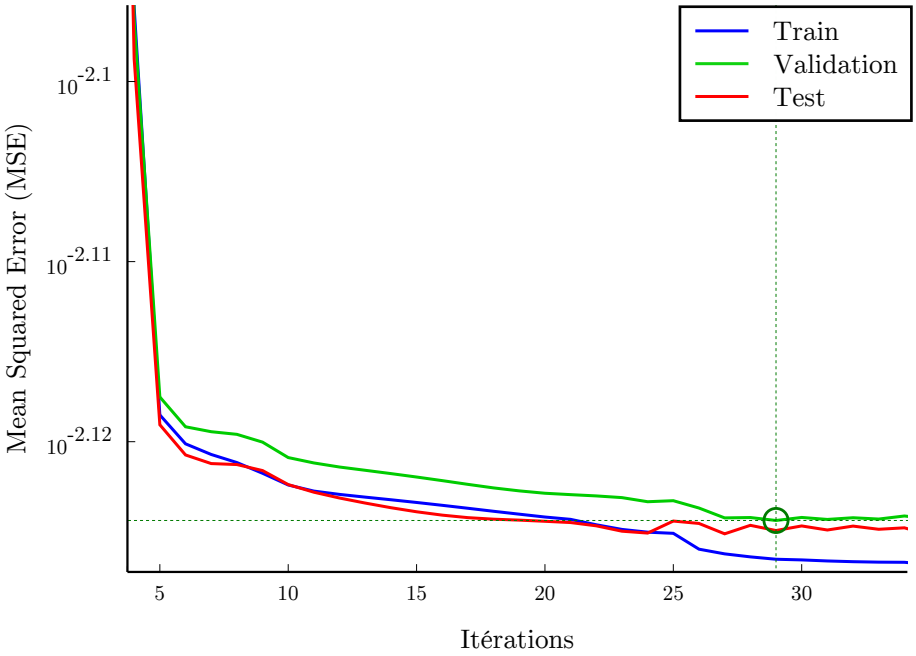


FIGURE 4.5 : Erreur lors de l'apprentissage sur les jeux d'apprentissage, de validation et de test.

Une fois que l'entrainement est arrêté par l'early-stopping, on peut évaluer le modèle en calculant sa précision à l'aveugle sur un troisième jeu, le jeu de test. On remarquera qu'il existe

une certaine confusion au niveau des termes « validation » et « test », qui sont parfois interverti. Le jeu utilisé pendant l'apprentissage pour le processus d'early-stopping est bien le jeu de validation, et celui utilisé pour l'évaluation finale est bien le jeu de test. Ce procédé amène donc à diviser le jeu de données dont on dispose en trois :

- Le jeu d'entrainement, utilisé par l'algorithme d'apprentissage
- Le jeu de validation, utilisé pour la régulation de l'apprentissage
- Le jeu de test, utilisé pour l'évaluation du modèle

La figure 4.5 montre la progression d'une phase d'apprentissage. On y voit l'erreur de test diminuer avec les itérations, cette erreur étant diminuée par l'algorithme. L'erreur calculée sur les deux autres jeux indépendants diminue également. A partir de la $29^{\text{ième}}$ itération, l'erreur calculée sur le jeu de validation arrête de diminuer, alors que la précision s'améliore toujours sur le jeu d'apprentissage. L'apprentissage est donc interrompu, ce qui évite que l'erreur sur le jeu de validation ainsi que sur le jeu de test ne recommence à augmenter.

Travailler avec des RNA nécessite donc de détenir un jeu de variables large. Cet échantillon devra être divisé en trois, en ayant tout de même un jeu d'entrainement suffisamment large et représentatif. Dans notre cas, jusqu'à huit ans de données de qualité de l'air et de données météorologiques ont pu être réunis selon les sites de mesures, permettant d'adopter cette approche de prévision.

4.3.4 Comparaison d'algorithmes d'apprentissage

Les algorithmes à direction de descente que l'on a abordé (DG, LM, BFGS, SCG) sont les plus utilisés dans la littérature (voir tableau 2.1 page 47). Il n'y a pas de consensus sur le choix optimal, même si le LM est peut-être le plus utilisé. Chacun de ces algorithmes a des caractéristiques intéressantes (capacité de convergence, coût en calcul de chaque itération, etc.).

Une étude préalable dont les résultats sont montrés sur la figure 4.6 permet de se rendre compte de l'intérêt de chaque algorithme dans notre cas. Nous avons procédé pour l'entrainement de 10 PMC à la prévision des concentrations horaires d'O_3 à Giraud à $h+24$, pour chacun de ces quatre algorithmes. La configuration utilisée pour ce test correspond à la configuration suivante : les données d'entrées sont les mesures d'O_3 et de NO_2 de la station, ainsi que des sorties du modèle AROME (pour Application de la Recherche à l'Opérationnel à Méso-Echelle) au même horizon (composantes U et V du vent à 10m et à 800hpa, température à 2m, la hauteur de la couche limite, le rayonnement global et la nébulosité). Ces données ont été centrées et réduites. L'initialisation des poids et biais du réseau a été réalisée à l'aide de l'algorithme de Nguyen-Widrow, l'expérience est menée dix fois pour chaque algorithme.

Sur la figure 4.6, le nombre d'itérations moyen sur les dix expériences est spécifié, sauf pour l'algorithme de descente de gradient. En effet, cet algorithme a systématiquement mené à des minimums locaux de précision, dont il n'a pas été capable de sortir. La précision a arrêté d'augmenter sans atteindre l'early stopping. La figure 4.6 montre la précision atteinte après mille itérations, l'indice d'agrément est alors faible et n'évolue plus.

Entre le LM, le SCG et le BFGS, c'est le LM qui obtient les meilleurs scores d'indice d'agrément. Ces scores sont également les plus stables, comparés à ceux des deux autres algorithmes. Enfin, sa convergence est la plus rapide, grâce à la méthode de descente variant entre la direction de Gauss-Newton et celle du gradient.

Pour ces raisons, on utilisera l'algorithme de LM dans ces travaux, initialisé par l'algorithme de Nguyen-Widrow.

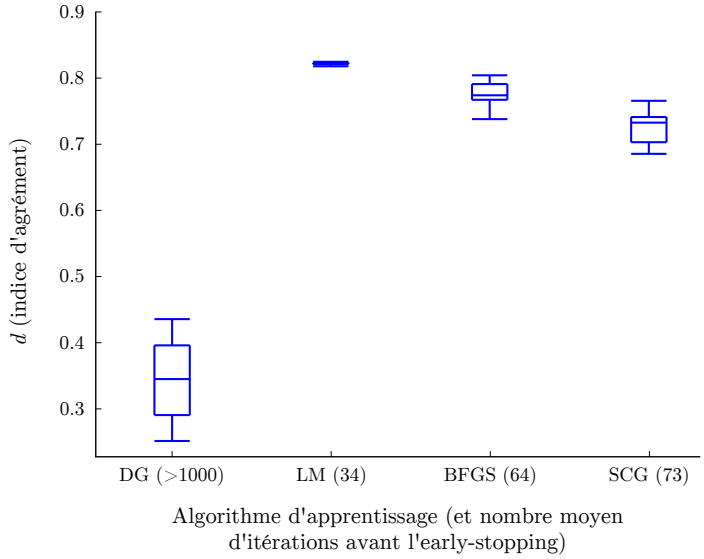


FIGURE 4.6 : Indices d'agrément obtenus par 10 modèles prédictifs d'O_3 à h + 24 à Giraud, en fonction de l'algorithme d'apprentissage utilisé. L'apprentissage par DG reste systématiquement piégé dans un minimum local, il est interrompu après 1000 itérations.

4.4 Traitement de données

Avant de commencer l'apprentissage d'un RNA, plusieurs prétraitements des données sont nécessaires. Certains d'entre eux sont obligatoires, d'autres sont optionnels. Les prétraitements doivent rendre les données intelligibles pour le réseau de neurones, afin de permettre et faciliter l'apprentissage. Le prétraitement de la variable cible y peut également impliquer un post-traitement de la sortie de modèle $\hat{y}$ afin de la faire correspondre à la variable initiale.

Plusieurs prétraitements doivent être envisagés en fonction du problème. Les prétraitements choisis font partie de la configuration du PMC qui est déterminée lors d'expérimentations, afin d'isoler la configuration optimale pour un problème donné.

Nous allons présenter les traitements que nous avons utilisés, et leur impact sur la précision des modèles. Nous verrons comment sont gérées les valeurs manquantes des jeux de données. Nous présenterons ensuite la transformation des variables circulaires comme la direction du vent, la stationnarisation des séries temporelles et leur normalisation. Enfin, nous montrerons les avantages apportés par une transformation des variables par Analyse en Composantes Principales (ACP).

4.4.1 Gestion des données manquantes

Les jeux de données réels peuvent comporter des valeurs manquantes. Ces « trous » sont dus à plusieurs facteurs : les périodes de maintenance des appareils de mesure, les dysfonctionnements ponctuels des mesures ou de la chaîne d'acquisition, les données invalidées (voir section 3.1 page 60).

Ces données manquantes sont problématiques. Sous Matlab, quand dans une série temporelle, un point est manquant, sa valeur est remplacée par l'indication NaN (pour « Not a Number » en anglais). C'est la forme spécifiée par la norme IEEE 734 pour la représentation des nombres à virgule flottante en binaire (IEEE Computer Society *et al.*, 2008). Aucun calcul mathématique n'est possible à partir d'un NaN, toute opération impliquant un NaN voit son résultat être

également NaN. Ainsi, quand une des variables d'un jeu de données est manquante à une certaine date, c'est l'ensemble des données à cette date formant un point qui ne peut plus être utilisé.

La première réaction face au problème des données manquantes est de supprimer tous les points pour lesquels l'une des variables présente un manque de données. C'est l'attitude que nous aurons en général vis-à-vis de ce problème.

Mais dans certaines situations, ce procédé implique une réduction drastique de la taille du jeu de données. Par exemple, si une série temporelle est utilisée plusieurs fois en entrée avec plusieurs délais, un point de donnée manquant de la série temporelle originale va se retrouver décalé dans le temps sur la série ayant subi le délai. Au final on se retrouve avec deux dates pour laquelle une des variables d'entrée présente une donnée manquante. Ceci peut grandement augmenter le nombre de valeurs manquantes. Quand on dispose de peu de données pour l'apprentissage, cela peut être problématique.

Il peut alors être bénéfique de remplacer les données manquantes de certaines variables par une estimation de leur valeur. Le procédé peut paraitre néfaste, car ces estimations risquent de biaiser l'apprentissage. L'évaluation d'un réseau de neurones entrainé avec et sans remplacement des données manquantes peut apporter la réponse.

Il existe plusieurs méthodes pour réaliser ce type de remplacement. Tout d'abord, on peut remplacer les variables manquantes par la valeur moyenne sur les autres années à la même date et heure, méthode qu'on appellera le remplacement par profil. On peut également utiliser la méthode plus couteuse en temps de calcul des k plus proches voisins (knn pour « k-nearest neighbors » en anglais). Quand un point d'une variable est manquant, on recherche les autres variables à la même date. A partir de ces autres variables, on identifie la situation la plus proche dans le jeu de données (proche par exemple au sens d'une distance euclidienne). La valeur correspondante de la variable manquante est alors adoptée pour remplacer le NaN.

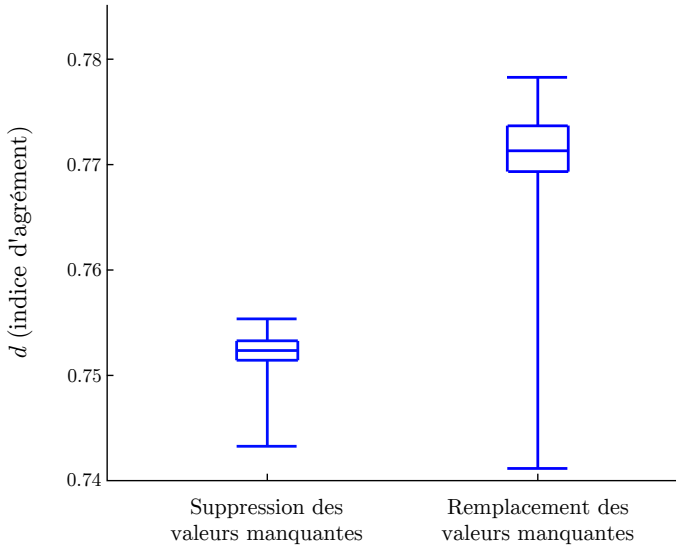


FIGURE 4.7 : Indices d'agrément obtenus par 10 modèles prédictifs d'O_3 à h + 24 à Giraud, avec et sans remplacement par profil des valeurs manquantes.

Ces méthodes sont à utiliser au cas par cas, quand nos jeux de données sont de petites tailles et qu'il faut éviter de supprimer trop de données manquantes. Il est nécessaire d'en vérifier la pertinence par une comparaison avec la simple suppression de tout point comportant une valeur manquante. La figure 4.7 montre un exemple de modèle prédictif bénéficiant positivement d'un remplacement par profil des valeurs manquantes.

4.4.2 Projection de variables circulaires

Certaines variables dites bornées ont la particularité d'avoir la valeur de leur borne supérieure qui correspond à la valeur de leur borne inférieure. On les appelle des variables circulaires. C'est par exemple le cas de la direction du vent. Exprimée en degrés, une direction de 0° est la même qu'une direction de 360°.

Pour un réseau de neurones, les valeurs extrêmes de telles variables sont difficiles à interpréter. Elles paraissent opposées alors qu'elles décrivent au contraire des états très proches. Ceci peut être corrigé d'une manière simple : en projetant une variable circulaire x sur les axes correspondant à son sinus et cosinus. On utilise donc deux variables $cos(x)$ et $sin(x)$ à la place de x, qui à elles deux apportent autant d'informations. On fera attention à l'unité originale de x, qui sera ramenée en radian au besoin. Les nouvelles variables ne présentent plus de discontinuité préjudiciable à leurs bornes.

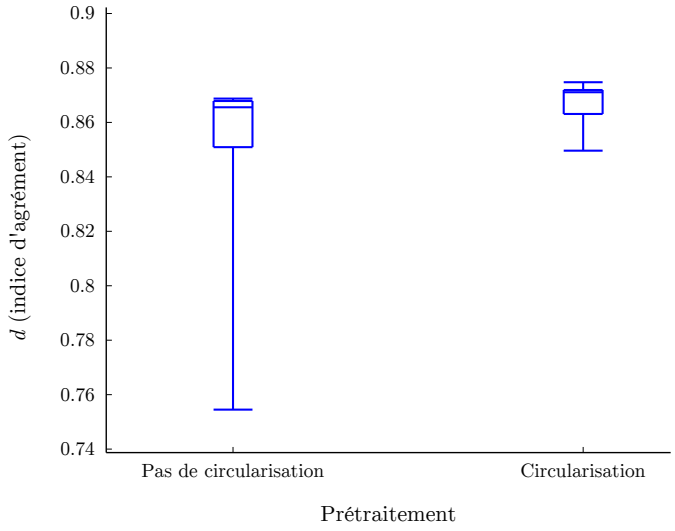


FIGURE 4.8 : Indices d'agrément obtenus par 10 modèles prédictifs d'O_3 à h + 24 à Venaco, avec et sans projection de la variable de direction du vent.

La figure 4.8 montre un exemple de l'impact de cette mesure sur les prévisions d'O_3 à Venaco, en utilisant les variables météorologiques mesurées au niveau de la station et les mesures de concentration d'O_3 et de NO_2 en données d'entrée. La direction du vent est incluse telle quelle puis projetée, avec un gain de précision et de stabilité dans les résultats.

Ce traitement sera systématiquement appliqué aux directions de vent mesurées en station (les sorties de modèles sont déjà présentées selon leurs composantes nord-sud et est-ouest) ainsi qu'aux variables temporelles comme l'heure de la journée ou le jour de l'année.

4.4.3 Stationnarisation

L'utilisation de RNA ne nécessite pas de satisfaire des conditions de stationnarité des séries temporelles, comme d'autres modèles, par exemple ceux de Box-Jenkins. Cependant, le fait d'avoir des séries temporelles stationnaires aide les réseaux de neurones à les modéliser correctement. En effet, si une série temporelle comprend une composante périodique ou une tendance identifiable (ou les deux), alors cela correspond à une information superflue puisque connue.

Si l'on rend la série temporelle stationnaire en la débarrassant de ses composantes périodiques, on simplifie la tâche du réseau de neurones, qui n'aura plus à tenir compte de ce type

d'information lors de l'apprentissage. Il aura plus de ressources disponibles pour se consacrer à la prévision de la composante non-périodique qui nous intéresse.

La plupart des séries temporelles que l'on utilise présentent une périodicité journalière et saisonnière, du fait d'une dépendance directe ou indirecte à l'ensoleillement. C'est le cas par exemple de l'O_3, sa formation s'inscrivant dans un cycle photochimique dont l'activité dépend directement du rayonnement solaire. Cette périodicité est visible si l'on trace la fonction d'autocorrélation de la série temporelle, c'est-à-dire le coefficient de corrélation R entre cette série temporelle et la même série décalée dans le temps.

Introduisons le concept d'Information Mutuelle (IM). L'IM est une notion issue de la théorie de l'information de Claude Shannon, et sert à quantifier la quantité d'information que partagent deux signaux. Elle est basée sur l'entropie de Shannon, notion qui rappelle par sa formule l'entropie thermodynamique, ou entropie de Boltzmann, d'où son nom. L'entropie thermodynamique représente le désordre d'un système à l'échelle microscopique, l'entropie de Shannon quant à elle représente le désordre au sein d'un signal, et indirectement sa prédictibilité (Shannon, 1948). Soit une variable x pouvant prendre n valeurs discrètes. On a

$$H(x) = -\sum_{i=1}^{n} P_i \, log_2(P_i) \tag{4.19}$$

avec $P_i = P(x = x_i)$ la probabilité que x prenne la valeur x_i. L'utilisation d'un logarithme de base 2 permet d'avoir une entropie exprimée en bits. Prenons l'exemple d'un tir à pile ou face. On utilise une pièce normale qui a autant de chance de tomber sur pile que sur face. Il y a donc $n = 2$ états possibles, tous deux dotés d'une probabilité d'occurrence $P_i = 0.5$. L'entropie est égale à un bit, ce qui veut dire que l'information moyenne fournie par un lancer de pièce correspond à un bit d'information (0 ou 1). Prenons maintenant une pièce truquée qui ne tombe que sur face. L'entropie d'un lancer est nulle, car le résultat d'un tel lancer ne fournit aucune information, le système étant parfaitement prévisible. Plus l'entropie d'un signal est élevée, plus son nombre de possibilités le rend imprédictible.

L'entropie conjointe entre deux variables x et y se calcule par

$$H(x, y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P_{i,j} \, log_2(P_{i,j}) \tag{4.20}$$

avec m le nombre de valeurs que peut prendre y, et $P_{i,j} = P(x = x_i, y = y_j)$ la probabilité que x prenne la valeur x_i quand y prend la valeur y_j. On peut alors calculer l'IM de x et y, telle que

$$IM(x, y) = H(x) + H(y) - H(x, y) \tag{4.21}$$

Cette information mutuelle s'exprime également en bits et correspond à la quantité d'information moyenne sur x que fournit une réalisation de y. L'IM est intéressante quand on étudie le lien entre plusieurs variables. Leur corrélation croisée est souvent utilisée pour étudier ces relations, mais ne concerne que les liens linéaires qui existent entre les variables, alors que l'IM quantifie leur interdépendance statistique, qu'il s'agisse de relation linéaire ou non. Nous l'utilisons ici pour détecter une saisonnalité dans les séries temporelles, mais elle sera surtout utile pour la sélection de variable (voir section 5.1.1 page 115).

La figure 4.9 montre cette autocorrélation ainsi que l'auto-information mutuelle. On voit bien la périodicité annuelle, représentée par un pic de corrélation pour 365 jours de décalage. Vers 180 jours de délai, les concentrations d'O_3 apparaissent anticorrélées, puisqu'il s'agit de la saison opposée. Au niveau de l'information mutuelle, l'anticorrélation apporte autant d'information que

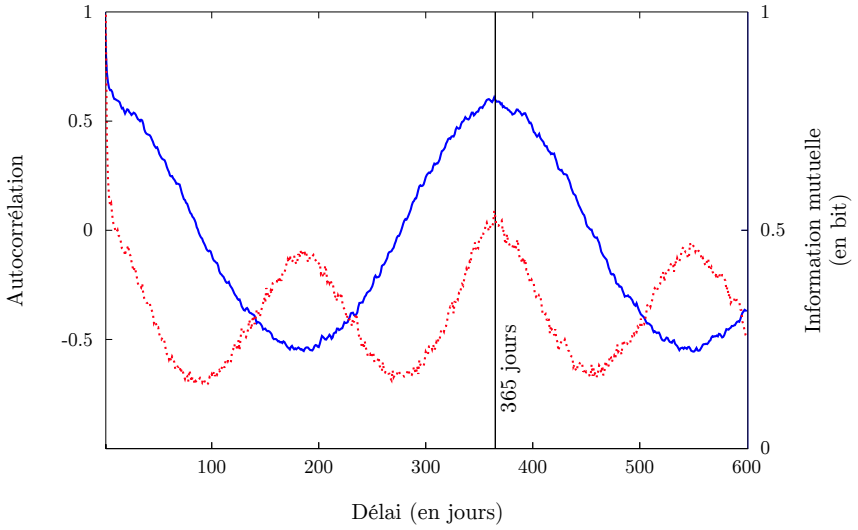


FIGURE 4.9 : Autocorrélation (en trait plein bleu) et information mutuelle (en pointillé rouge) entre la série temporelle journalière d'O_3 mesurée à Canetto et la même série temporelle décalée.

la corrélation puisqu'il s'agit d'une relation d'interdépendance, et on retrouve deux pics annuels dus à la saisonnalité de l'O_3. Si l'on traçait le même graphique pour des délais de l'ordre de la journée, on observerait le même type de pics pour la période de 24 heures.

Afin de rendre stationnaire une telle série temporelle, on peut utiliser la méthode suivante. A partir de notre historique de données, on calcule le profil journalier et le profil annuel de la série temporelle. Sur ces profils, on voit bien la périodicité. On peut alors créer une série temporelle correspondant au profil type de notre variable. On stationnarise la variable en lui soustrayant son profile type. La figure 4.10 montre l'autocorrélation et l'information mutuelle de la série temporelle d'O_3 ainsi stationnarisée avec elle-même.

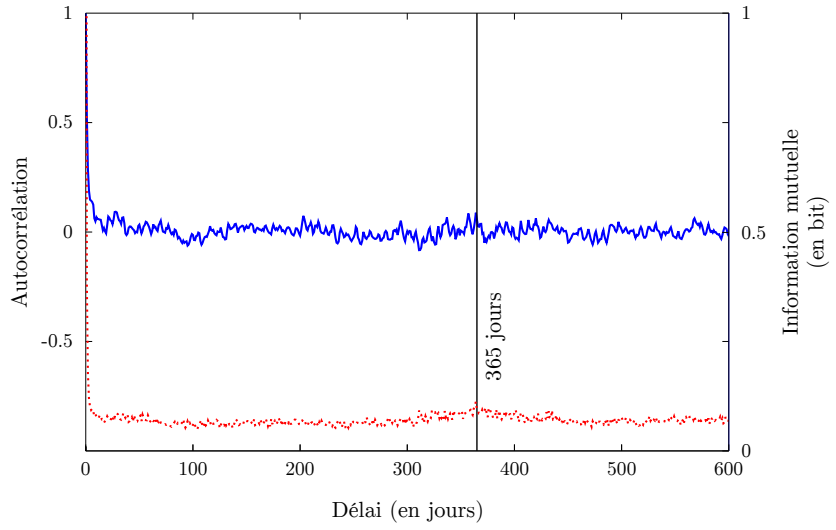


FIGURE 4.10 : Autocorrélation (en trait plein bleu) et information mutuelle (en pointillé rouge) entre la série temporelle journalière d'O_3 mesurée à Canetto stationnarisée et la même série temporelle décalée.

On ne voit plus apparaitre de pics indiquant une périodicité, ni pour l'autocorrélation, ni pour l'information mutuelle. Un léger résidu de pics annuels d'information mutuelle est perceptible.

Cette opération permet de désaisonnaliser les séries temporelles.

Les données d'entrée tout comme les données cibles peuvent être stationnarisées. Si l'on effectue un tel traitement sur la donnée cible, alors il convient de déstationnariser les sorties de modèle en post-traitement avant l'évaluation ou l'utilisation opérationnelle, afin de retrouver la prévision de la série temporelle originelle.

La figure 4.11 montre les résultats de modèles prédictifs d'O_3 à Giraud à l'horizon $h+24$. Les données utilisées en entrée sont les mesures d'O_3 et de NO_2 de la station, ainsi que des sorties du modèle AROME au même horizon (composantes U et V du vent à 10m et à 800hpa, Température en degrés Kelvin (TK) à 2m, Hauteur de la Couche Limite (HCL), Rayonnement Solaire (RS) et Nébulosité (NEB)). On présente tout d'abord les résultats du modèle sans stationnarisation, puis avec stationnarisation des entrées, puis avec stationnarisation des entrées et sorties. On remarque que c'est cette dernière configuration qui obtient les meilleurs scores.

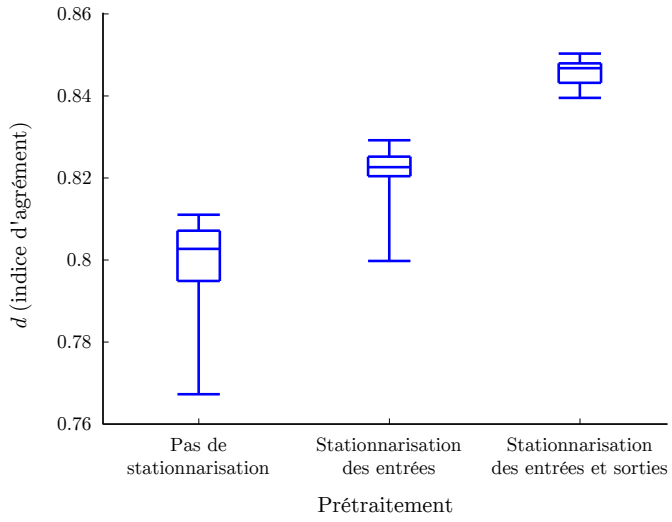


FIGURE 4.11 : Indices d'agrément obtenus par 10 modèles prédictifs d'O_3 à h + 24 à Giraud, en fonction des données qui ont été stationnarisées.

4.4.4 Normalisation

Certaines variables d'entrée peuvent prendre plus d'importance que d'autres à cause de leurs valeurs. L'unité dans laquelle elles sont exprimées peut créer un biais. En effet, quand les neurones somment leurs entrées, si l'une d'entre elle a des valeurs systématiquement plus élevées que les autres à cause de son unité ou de son ordre de grandeur, il va falloir que son poids soit configuré en conséquence pendant l'apprentissage et corrige cette sur-représentation. Les algorithmes sont capables de gérer ce biais en agissant sur les poids, mais une normalisation de toutes les entrées en prétraitement qui annule l'impact de l'unité sera plus efficace et simplifiera le problème d'optimisation que l'on soumet à l'algorithme.

Pour cette raison, les données d'entrée son habituellement normalisées. On peut par exemple ramener toutes leurs valeurs entre -1 et 1 en appliquant à chaque variable $\mathbf{x}$

$$\mathbf{x}_n = \frac{2\left(\mathbf{x} - min(\mathbf{x})\right)}{max(\mathbf{x}) - min(\mathbf{x})} - 1 \tag{4.22}$$

et en utilisant la variable normalisée $\mathbf{x}_n$ pour l'apprentissage et pour les prévisions.

Cependant, une telle normalisation, si elle supprime le biais dû à l'unité, n'est pas toujours adaptée. Dans le cas ou certaines variables ont de rares valeurs extrêmes, qu'elles soient dues à des erreurs de mesure ou pas, ce type de normalisation va pénaliser les valeurs proches de la moyenne, qui vont se retrouver proches de 0, peu signifiantes par rapport aux valeurs extrêmes et aux autres variables. Il vaut mieux dans ce cas centrer les variables autour de 0, pour qu'elles soient toutes centrées autour de la même valeur, puis les réduire en les divisant par leur variance (Dreyfus, 2004) :

$$\mathbf{x}_n = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sigma_x^2} \tag{4.23}$$

avec $\bar{\mathbf{x}}$ la moyenne de $\mathbf{x}$ et σ_x^2 sa variance.

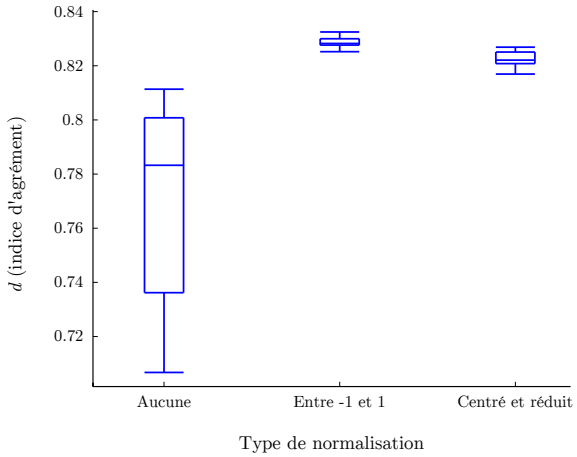


FIGURE 4.12 : Indices d'agrément obtenus par 10 modèles prédictifs d'O_3 à h + 24 à Giraud, en fonction du type de normalisation appliquée.

Ce simple prétraitement est très important. Il a un impact clair et systématique sur les scores des modèles. La figure 4.12 montre les scores calculés pour dix PMC prévoyant l'O_3 à Giraud à $h + 24$, sans normalisation et avec les deux procédés présentés ci-dessus. Ces deux prétraitements améliorent les scores du modèle, et les rendent plus stables. La normalisation entre -1 et 1 apparait plus adaptée pour cet exemple, cependant, centrer et réduire les variables peut s'avérer plus intéressant selon les cas, en fonction de la variance des variables utilisées.

Pour nos expériences, nous normaliserons systématiquement nos variables d'entrée et nos variables cibles en les centrant et les réduisant.

4.4.5 Analyse en Composantes Principales

L'ACP est une méthode d'analyse multivariée, à la fois géométrique et statistique. Elle permet d'obtenir à partir d'un jeu de variables partiellement corrélées entre elles de nouvelles variables, combinaisons linéaires des premières, et totalement décorrélées les unes des autres. Cette transformation se fait par projection des variables dans un nouveau repère, sur des axes qui sont les « composantes principales » des premières variables. Ces axes correspondent à de nouvelles variables qui n'ont peut-être plus de réalité physique mais plutôt une réalité conceptuelle, et qui expliquent au mieux la variance au sein des données.

L'ACP se fait en plusieurs étapes. Les variables sont réunies en une matrice $\mathbf{M}$ de n lignes et m colonnes, avec n égal à la taille des échantillons et m au nombre de variables. Afin de faire contribuer toutes les variables de la même manière à l'ACP, celles-ci sont habituellement centrées

et réduites. Dans le cas où l'on veut conserver la variance de chaque variable, pour éviter qu'une variable fortement entachée de bruit contribue autant que les autres à l'ACP par exemple, il arrive qu'on ne réduise pas les variables. Dans notre cas, nous les avons systématiquement normalisées.

On obtient la matrice $\mathbf{K}$ en multipliant $\mathbf{M}$ par sa transposée.

$$\mathbf{K} = \mathbf{M^T M} \tag{4.24}$$

$\mathbf{K}$ est la matrice de corrélation des variables de $\mathbf{M}$, ou la matrice de variance-covariance si les variables n'ont pas été réduites. On obtient les vecteurs propres $\mathbf{v}$ de $\mathbf{K}$ en la diagonalisant. On a :

$$\mathbf{K}\,\mathbf{v}_i = a_i\,\mathbf{v}_i \tag{4.25}$$

avec a_i la valeur propre associée au vecteur propre $\mathbf{v}_i$ et $1 < i < m$.

Les m composantes principales données par les vecteurs propres de $\mathbf{K}$, sont orthogonales entre elles, c'est-à-dire linéairement décorrélées. La valeur propre de chaque composante principale correspond à la variance observée sur l'axe correspondant. Chaque vecteur propre explique un certain pourcentage de la variance totale, et on utilise ce pourcentage pour les hiérarchiser du plus important au moins important.

On obtient $\mathbf{p}_i$ la projection des points sur la composante principale i par le produit :

$$\mathbf{p}_i^{\mathbf{T}} = \mathbf{v}_i^{\mathbf{T}}\mathbf{M}^{\mathbf{T}} \tag{4.26}$$

L'ACP a de nombreuses applications. En premier lieu, les graphiques des variables tels qu'on peut le voir sur la figure 4.13 montrent la contribution de chaque variable à chacune des composantes principales. Les corrélations entre les variables initiales apparaissent, permettant d'étudier leurs relations. Il est également possible d'étudier la manière dont chaque variable contribue à la dispersion des points du jeu de données. Plus une variable est responsable de cette dispersion, plus elle contribue aux axes de plus grande valeur propre, et inversement.

L'ACP permet également de réduire le nombre de variables du jeu de données. Puisque les vecteurs propres sont hiérarchisés en fonction de leur valeur propre, on obtient ainsi un critère de sélection en ne retenant que certaines variables transformées. On verra cette application de l'ACP pour la sélection de variables à la section 5.1.4, page 124.

La figure 4.13 est un exemple de graphique des variables, résultat d'une ACP sur des données mesurées à la station Giraud (concentrations d'O_3, de NO, de NO_2, de PM10, composantes est-ouest (U) et sud-nord (V) du vent) et des sorties de modèle AROME au point de coordonnées les plus proches de celle de la station : HCL, Epaisseur de la Couche d'Inversion (ECI) entre 0 et 1000 m.

Il faut être prudent avant de tirer des conclusions de ce graphique car la variance expliquée par l'ensemble de ces deux axes ne correspond qu'à 57.69% de la variance totale. Cependant, cette représentation peut aider à identifier des liens linéaires entre des variables et peut être pris en compte par l'utilisateur lors d'une étude préliminaire de sélection de variables.

On retrouve sur cette projection certaines relations entre les variables. Les concentrations de NO sont logiquement corrélées à celles de NO_2, NO étant principalement produit lors de réactions de combustion, et rapidement oxydé en NO_2. Ces concentrations sont anticorrélées avec celles d'O_3, la réaction avec NO étant un puits chimique de l'ozone. La hauteur de la couche limite s'oppose à l'épaisseur de la couche d'inversion thermique, puisque ces inversions s'apparentent à des « couvercles » limitant la hauteur de cette couche limite. Ces situations stables favorisent

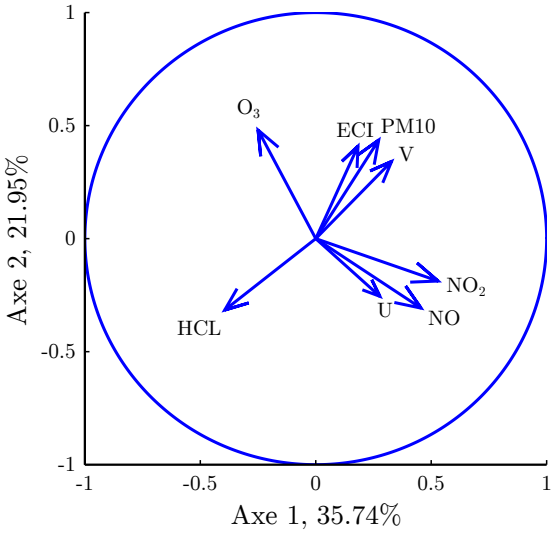


FIGURE 4.13 : Contribution de chaque variable aux 2 premières composantes principales, avec les pourcentages de variance expliquée indiqués pour chaque axe.

la stagnation des polluants et apparaissent très corrélées aux concentrations de PM10. Le vent du sud (au travers de la composante V) est corrélé aux concentrations en PM10, ce qui peut s'expliquer par le rôle important des épisodes de transport de poussières sahariennes sur les fortes concentrations de ce polluant en Corse. La composante U est corrélée sur les 2 axes à NO_2 et sur le second aux PM10, ce qui peut représenter l'apport de ces polluants sur la station depuis le centre-ville et le port en vent d'est. Le vent d'ouest, lui, apporte de l'ozone rural et d'altitude des crêtes du Cap Corse, où ses niveaux sont plus élevés qu'en milieu urbain. Ceci se dénote par l'anticorrélation de U et O_3 sur le graphique.

L'ACP apporte encore une autre possibilité de visualisation. En utilisant certaines composantes principales pour former les axes d'un repère, par exemple celles ayant la plus grande variance, l'ACP permet une visualisation en deux (ou trois) dimensions du nuage de point, per-

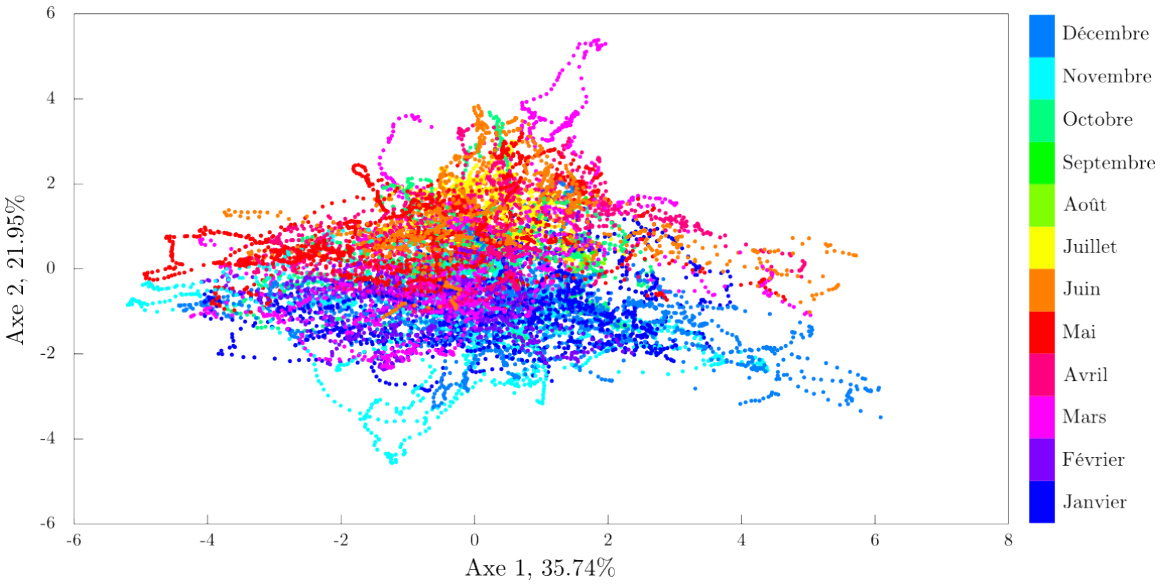


FIGURE 4.14 : Projection des individus sur les deux premiers axes de l'ACP, avec les pourcentages de variance expliquée indiqués pour chaque axe.

mettant d'étudier les relations entre les individus des échantillons, d'identifier des groupes de points ou des points atypiques.

La figure 4.14 montre la projection du nuage de points correspondant aux variables brutes sur les deux mêmes premiers axes de l'ACP. Ici, chaque point est colorisé suivant le mois correspondant. On remarque que la distinction entre été et hiver ressort clairement le long de l'axe 2. La période de l'année influence évidemment toutes les variables environnementales et météorologiques ayant servi à effectuer l'ACP. Etant donné l'importance majeure du cycle des saisons sur ces variables, il n'est pas surprenant qu'une des composantes principales y soit fortement corrélée. C'est un exemple d'ACP qui permet de mettre à jour l'influence d'un phénomène particulier (le cycle saisonnal) qui n'est pas directement représenté par une variable.

Le fait que les variables issues de l'ACP soient orthogonales entre elles a souvent un impact positif sur la précision des modèles prédictifs, par rapport à l'utilisation des variables initiales. En effet, leur décorrélation fait diminuer leur redondance, et le jeu de données est plus clair pour le réseau de neurones.

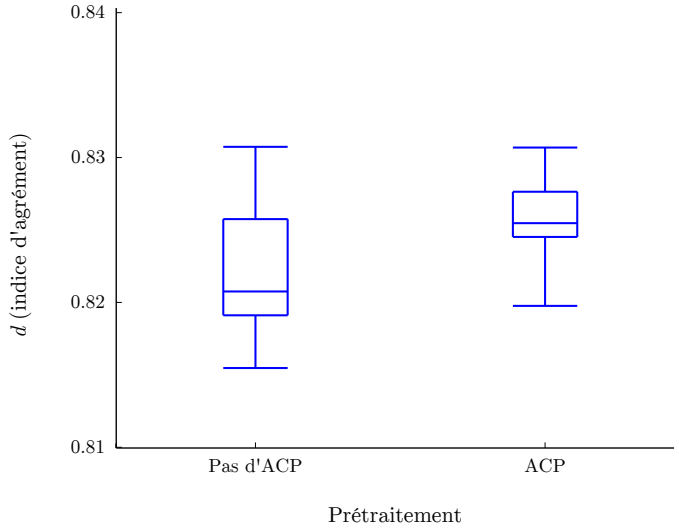


FIGURE 4.15 : Indices d'agrément de modèles de prévision des concentrations d'O_3 à La Marana, avec et sans utilisation des composantes principales à partir de variables centrées et réduites.

La figure 4.15 nous montre un exemple de bénéfice visible de l'utilisation de composantes principales (calculées à partir de données normalisées) sur la prévision.

4.4.6 Intérêt des prétraitements

Nous avons examiné les principales opérations de prétraitement qui sont utilisées dans la littérature, afin de rendre les données brutes intelligibles pour un RNA. Les prétraitements que l'on a vus sont autant d'outils permettant d'améliorer les capacités de nos modèles prévisionnels. Certains d'entre eux comme la normalisation des variables seront utilisés systématiquement, d'autres comme le remplacement des données manquantes le seront au cas par cas, avec évaluation attentive de l'impact sur les résultats.

Le choix des prétraitements à utiliser alourdit la méthodologie, puisqu'il faut souvent chercher leur meilleure configuration pour chaque problème. C'est l'un des défauts majeurs des RNA, qui demandent beaucoup de tests avant d'identifier la meilleure configuration. Cependant, le gain en précision est bien réel pour les modèles prédictifs appliqués à tous les polluants. La figure 4.16

montre un exemple de ce gain pour la prévision de particules à Canetto, sans aucun prétraitement et en utilisant stationnarisation, normalisation, projection des variables circulaires et ACP.

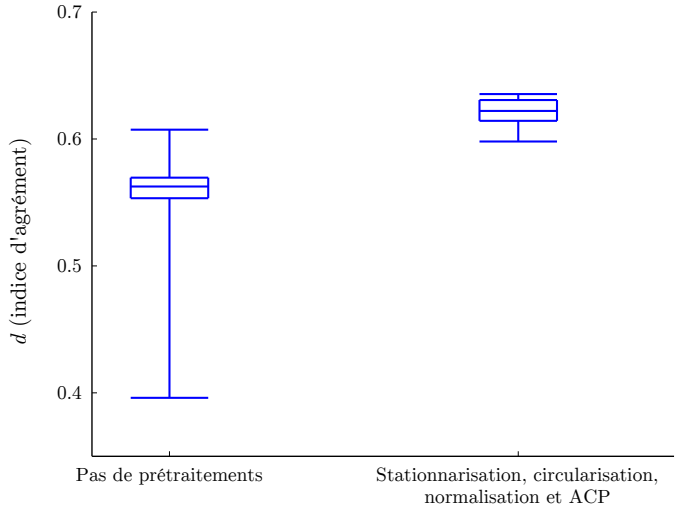


FIGURE 4.16 : Indices d'agrément de modèles de prévision des concentrations en PM10 à Canetto, avec et sans prétraitements.

4.5 Configuration des PMC et expérimentations

Maintenant qu'on a vu les étapes nécessaires à la construction d'un PMC prédictif, à savoir le prétraitement des entrées et l'apprentissage, nous allons nous pencher sur la méthode à adopter pour fixer efficacement la configuration du réseau. En effet, les PMC laissent à leur utilisateur plusieurs possibilités de configuration.

La méthodologie utilisée pour définir la configuration d'un modèle prédictif permet d'optimiser les prévisions. Cette méthodologie a pour but de fixer tous les points de configuration du modèle, qui peuvent être interdépendants.

La configuration d'un PMC correspond aux choix suivants :

- Variables d'entrée
- Délais et/ou échéances à utiliser
- Prétraitements et post-traitements à appliquer
- Architecture du PMC (nombre de couches et nombre de neurones)
- Algorithme d'apprentissage
- Division des données entre jeu d'apprentissage, de validation et de test

On note que le choix de l'algorithme d'apprentissage se portera désormais systématiquement vers l'algorithme de LM. Ce point ne fera plus partie dans notre cas des points de configuration à fixer.

Les autres points doivent être fixés pour chaque problème différent. La configuration optimale change pour la prévision de chaque polluant, à chaque station, et suite à l'évolution des jeux de données disponibles. Ces pour cela que la modélisation à l'aide de PMC nécessite beaucoup d'expérimentations. La recherche empirique de la configuration optimale est nécessaire, et fait partie des inconvénients des réseaux neuronaux.

L'évaluation à l'aide du jeu de test permet de départager les modèles. Elle est menée avec les

outils présentés précédemment, à la section 2.4 (page 38). Il est important de garder à l'esprit qu'en fonction des données disponibles, les jeux de test changent, ce qui se répercute sur les indices d'erreurs obtenus par les modèles. On ne compare donc des modèles entre eux que s'ils ont été évalués avec les mêmes jeux de test.

Avant de construire le modèle, on doit définir le problème. Il s'agit de définir le prédictand (la variable à prévoir), typiquement la concentration d'un polluant mesurée à une station fixe. L'horizon de la prévision est également choisi.

On procède ensuite par ordre pour fixer tous les détails de la configuration. Comme indiqué sur la figure 4.17, on commence par choisir les variables initiales qu'on veut utiliser en entrée du modèle. Ce choix s'appuie sur les connaissances des mécanismes gouvernant la qualité de l'air, et dépend fortement des données disponibles. Ensuite, une méthode de sélection de variables peut être choisie pour réduire la taille du jeu de variables d'entrée (on verra ce point plus en détail à la section 5.1 page 113).

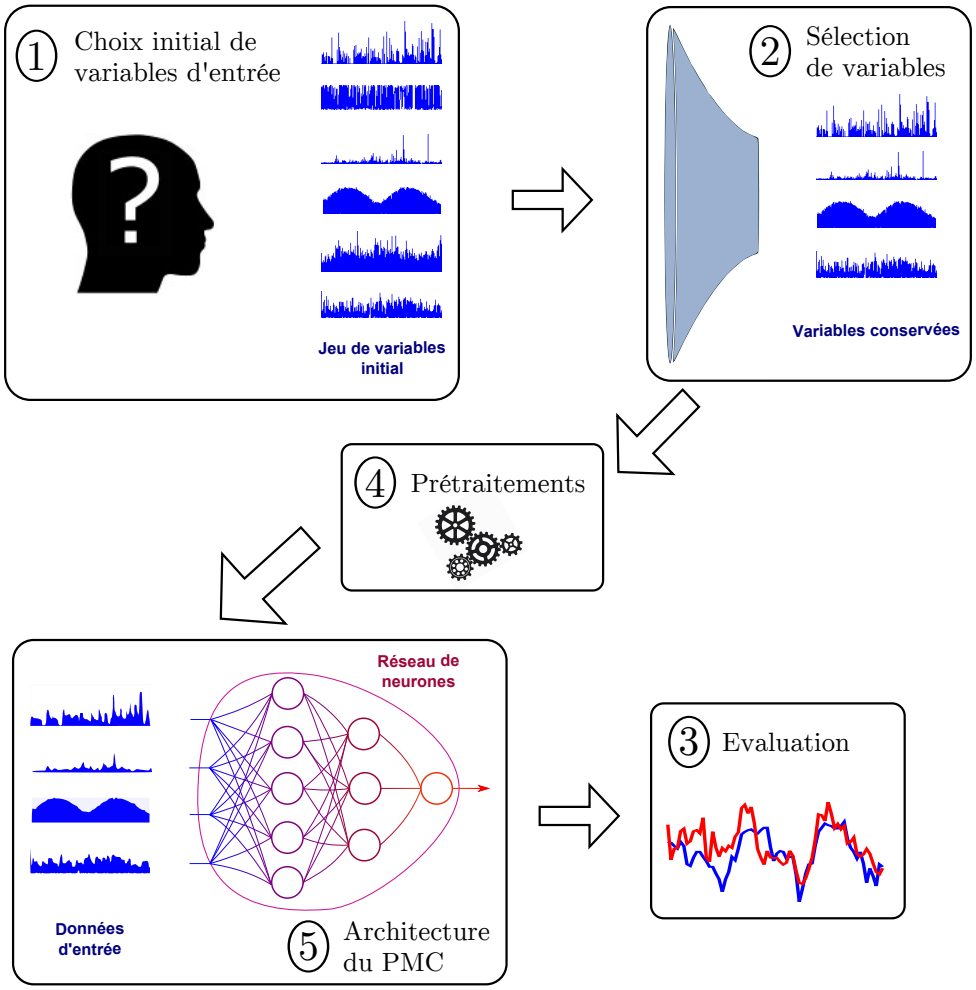


FIGURE 4.17 : Options de configuration d'un modèle à fixer, avec indiqué l'ordre dans lequel les fixer. Les flèches indiquent l'ordre d'exécution de ces étapes.

En fonction des séries temporelles sélectionnées, il est important d'adapter l'évaluation, en choisissant la taille et les données du jeu de test, et par la même occasion des jeux d'apprentissage et de validation. On peut alors choisir les prétraitements à appliquer. La gestion des données manquantes dépendra de leur quantité dans les différents jeux de données. S'il n'y en a pas un

nombre significativement élevé, les données manquantes sont simplement supprimées.

Enfin, l'architecture du PMC peut être fixée, le nombre de neurones et de couches pouvant être choisis en fonction du nombre de variables d'entrée et de la taille du jeu d'apprentissage. Pour éviter le sur-apprentissage, on limitera le nombre de neurones si le jeu d'apprentissage est petit.

Tous ces points forment une configuration, qui doit être jugée en fonction des résultats de l'évaluation du modèle. La méthodologie pour isoler la meilleure configuration est la suivante :

- Chaque configuration envisagée est pré-configurée
- L'ensemble des modèles pré-configurés sont entrainés et évalués plusieurs fois
- Les résultats sont analysés et mènent vers d'autres expérimentations de configuration, jusqu'à l'obtention de résultats stables

Chaque expérience permet d'évaluer une configuration. Elles doivent être répétées un certain nombre de fois afin de juger la stabilité des résultats. En effet, l'initialisation aléatoire des paramètres des neurones par l'algorithme de Nguyen-Widrow crée des différences de précision pour une même configuration testée plusieurs fois. Typiquement, une même configuration est testée dix fois. On mène donc un grand nombre d'expériences pour converger finalement vers la meilleure configuration.

Afin de pouvoir mener les expérimentations nécessaires, nous avons développé (sous Matlab) et utilisé l'outil Aria Base. Cette application est introduite à la section 7.1 page 167, et ses fonctionnalités détaillées sont présentées à l'annexe C page 211. L'outil permet de fixer en détail l'ensemble des configurations à examiner. Ensuite l'expérimentation, qui correspond à l'ensemble des prétraitements, entrainement et évaluation de tous ces modèles, est menée automatiquement. Tous les résultats de ces expériences sont ensuite archivés.

C'est avec cette application et en suivant cette méthodologie que nous avons construit les modèles prévisionnels, ceux présentés dans ces travaux tout comme ceux destinés à l'usage opérationnel à Qualitair Corse.

4.6 Conclusion

Lors de ce chapitre, nous avons expliqué les principes de l'usage des PMC pour la prévision statistique de la qualité de l'air, après avoir évoqué l'origine de cette famille de modèle. Nous nous somme familiarisé avec PMC et son apprentissage automatique, permettant de fixer les paramètres de ses neurones.

Nous avons envisagé plusieurs algorithmes d'apprentissage pour réaliser cette étape. La Descente de Gradient (DG), l'algorithme de Levenberg – Marquardt (LM), de Broyden – Fletcher – Goldfarb – Shanno (BFGS) et du SCG (Scaled Conjugate Gradient) ont été considérés car l'état de l'art sur la prévision de la qualité de l'air à l'aide de RNA recèle plusieurs exemples de leur usage. Après avoir présenté les bases théoriques des deux premiers, l'expérimentation a montré que pour la prévision de concentration en polluant en Corse avec nos données, l'algorithme de LM est le meilleur candidat, pour la précision et la robustesse de ses résultats.

Pour préparer les données à leur usage par cet algorithme, nous avons passé en revue les prétraitements nécessaires. Ces prétraitements permettent de présenter des données dénuées de leurs composantes saisonnières, normalisées, les données circulaires sont projetées. On peut éventuellement remplacer les valeurs manquantes. L'ACP enfin permet de présenter un jeu de données sans relation linéaire entre les variables.

On a pu remarquer l'intérêt de chacune de ces opérations sur les données. Les gains de précision apportés par ces prétraitements ont été illustrés à l'aide de modèles construits pour la prévision de concentration.

Enfin, nous avons présenté notre méthodologie afin d'identifier les meilleures configurations possibles lorsque l'on souhaite construire un modèle prévisionnel. Un grand nombre d'expérimentations est alors nécessaire, qui sont menées grâce à un outil que nous avons développé.

La méthode de travail avec les PMC qui a été présentée permet de faire fonctionner les modèles prévisionnels. Disons que cela permet leur fonctionnement basique. A partir de là, des innovations sont possibles pour perfectionner la prévision. Nous allons désormais nous consacrer à des expériences ayant pour but d'identifier diverses méthodes permettant d'améliorer les performances de nos modèles prévisionnels.

Chapitre 5

Améliorations des performances de prévision de la qualité de l'air

Nous avons vu au chapitre précédent comment construire un modèle prévisionnel basé sur le PMC. La méthodologie permettant d'identifier la configuration optimale pour chaque problème a été présentée. Nous allons maintenant nous intéresser à des méthodes que nous avons développées lors de ces travaux et qui permettent d'améliorer les performances des prévisions.

Nous nous intéresserons tout d'abord au domaine de la sélection de variables (« feature selection » en anglais), appréhendé comme un problème d'optimisation, notamment à l'aide de métaheuristiques. Pour être en conformité avec le principe de parcimonie, on utilisera des notions de la théorie de l'information et des connaissances en aérologie pour exclure à bon escient les variables les moins pertinentes. Nous observerons ensuite l'amélioration apportée par ces traitements préalables sur les performances de la prédiction (section 5.1, page 113).

Nous avons mis au point une méthode d'élagage permettant de supprimer les poids et biais superflus des PMC. Ce processus utilise l'algorithme de LM pour identifier les paramètres neuronaux les moins actifs, afin de les supprimer avant l'apprentissage habituel. Elle sera discutée à la section 5.2 de ce chapitre (page 127).

Après avoir présenté ces méthodes qui sont applicables à toute étude sur la prévision à l'aide de PMC, nous nous concentrerons de nouveau sur la problématique corse de prévision de la qualité de l'air (section 5.3, page 128). Dans ce cadre, l'usage de certaines variables, les sorties de modèles issues de la plate-forme AIRES, sera tout d'abord discuté. Nous pourrons ensuite présenter les modèles prévisionnels construits en appliquant notre méthodologie et en utilisant tous les outils présentés jusqu'ici. Ces modèles de prévision de concentration de PM10 et d'ozone à l'horizon $h + 24$ pour les villes de Bastia et d'Ajaccio nous permettront d'évaluer l'intérêt de notre méthode.

Nous avons ensuite voulu valider la robustesse de cette méthodologie, en nous adaptant à une situation de prévision différente de celle de Corse. Nous avons travaillé la prévision telle qu'elle est menée en région PACA, à partir des sorties du modèle AIRES et des mesures effectuées dans toutes les Bouches-du-Rhône. Cette étude, présentée à la section 5.4 page 137, nous a permis à la fois de nous confronter à une situation plus polluée qu'en Corse, à la fois de comparer nos modèles à ceux utilisés chez Air PACA, l'AASQA de la région PACA : les Forêts Aléatoires (FA, random forests en Anglais). Elle permettra de confirmer l'intérêt des PMC pour la prévision de la qualité de l'air, avec des variables d'entrée issues de mesure comme issues de CTM.

Une conclusion sera enfin apportée à ce chapitre à la section 5.5 (page 145).

5.1 Sélection de variables

Afin de respecter le principe de parcimonie, il est nécessaire d'avoir une méthode de sélection de variables (« feature selection » en anglais). En effet, beaucoup de variables exogènes potentielles sont en général disponibles. Nous avons eu accès à tous les polluants mesurés aux stations de surveillance de la qualité de l'air, ainsi qu'aux variables météorologiques quand elles y sont mesurées. On peut ajouter à ces données les mesures des stations de Météo-France les plus proches, qui sont souvent représentatives des conditions météorologiques observées à la station de Qualitair Corse. A cela se sont ajoutées les sorties du modèle AROME, ainsi que parfois des indices temporels.

En plus du choix des variables à utiliser, il faut également envisager d'appliquer plusieurs délais aux séries temporelles. Par exemple, si l'on utilise la concentration en NO_2 pour un modèle à l'horizon $h + 24$, on peut utiliser les valeurs de NO_2 à h, mais aussi à $h - 1$, $h - 2$, etc. Quand les variables sont des prévisions issues d'autres modèles comme par exemple AROME, plusieurs échéances de prévision peuvent de la même manière être sélectionnées. Dans ce cas, on prendra bien soin de s'assurer que l'échéance est bien disponible en situation opérationnelle au moment d'effectuer la prévision. Le fait de pouvoir utiliser plusieurs occurrences d'une série temporelle décalées dans le temps complexifie la sélection de variables.

La première méthode de sélection de variables est l'intuition de l'utilisateur. A partir de nos connaissances en aérologie, on estime quelles sont les variables les plus pertinentes. Par exemple, s'il s'agit de prévision des concentrations en particules à Ajaccio, on imagine bien que des indices temporels permettant d'apporter une information sur les activités humaines émettrices de particules sont intéressants comme variable d'entrée. Le jour de la semaine indique les jours de repos où moins de véhicules circulent, les heures de la journée permettent d'identifier les heures de pic de trafic routier. La hauteur de la couche limite, la direction et la force du vent ainsi que les précipitations le sont aussi pour leur impact sur la dispersion des particules et leur dépôt humide. Les composantes du vent en altitude fournies par AROME semblent également intéressantes, pour identifier les situations de vent sud pouvant apporter des poussières sahariennes. Les autres polluants de la station semblent également utiles.

La seconde méthode est l'utilisation d'outils objectifs qui servent de critères d'évaluation. Cela permet une approche moins subjective et pouvant gérer plus de paramètres, comme par exemple le choix des délais à appliquer aux variables à utiliser. L'adoption d'un critère d'évaluation des variables permet de quantifier leur pertinence afin de pouvoir les classer. Par exemple, on peut utiliser comme critère l'IM entre les variables à t et la cible à $t + x$. Il est également possible d'utiliser la corrélation entre les variables et la cible, mais cette corrélation ne prend en compte que les relations linéaires alors qu'on sait qu'il existe des liens non-linéaires entre nos séries temporelles cibles et les variables exogènes. C'est même une des raisons du choix des PMC comme modèle prédictif.

Ce type de critère peut être utilisé par l'utilisateur pour sélectionner les variables une par une. Mais on peut également utiliser des critères permettant d'évaluer l'ensemble d'un jeu de variables d'entrée, ce qui permet d'automatiser la sélection de variables à l'aide de méthodes de recherche.

La troisième méthode consiste en l'utilisation de métaheuristiques (méthodes d'optimisation adaptées aux problèmes d'optimisation difficiles, inspirées d'analogies physiques ou biologiques) utilisant ces critères d'évaluation. Les métaheuristiques sont des algorithmes d'optimisation qu'on peut utiliser quand on ne dispose d'aucune information sur le système que l'on veut optimiser, mais qu'on peut tout de même l'évaluer grâce à un critère. L'ouvrage de Dréo *et al.*

(2003) en décrit les principaux algorithmes. La fonction à optimiser (fonction objectif) a comme arguments des variables binaires indiquant l'utilisation ou non de chaque variable envisagée, et un critère d'évaluation lui permet de fournir un résultat pour les variables en argument. Les métaheuristiques sont des algorithmes itératifs, dont le but est de converger vers l'extremum global (maximum ou minimum en fonction du critère utilisé). Leur objectif est donc le même que celui des algorithmes d'apprentissage à direction de descente (vus à la section 4.3), mais leur fonctionnement est différent. En effet, contrairement à l'apprentissage d'un RNA dont les paramètres sont continus et l'erreur dérivable, on ne peut dériver notre fonction objectif ni estimer de meilleure direction de descente. Nous utiliserons deux métaheuristiques, les Algorithmes Génétiques (AG) et le recuit simulé.

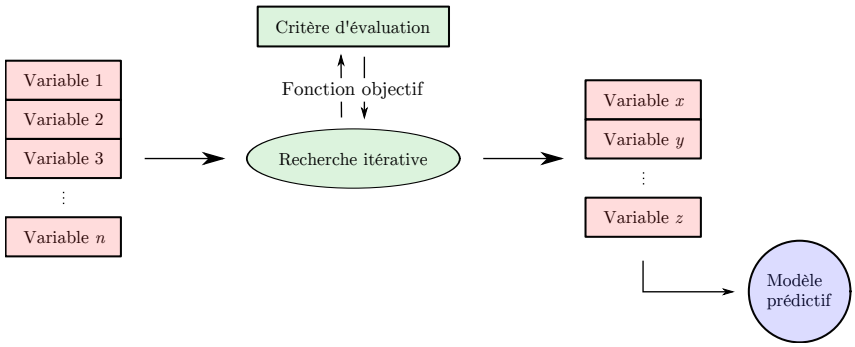


FIGURE 5.1 : Sélection de variables par approche « filter ».

Deux types d'approches concernant le critère d'évaluation des métaheuristiques se distinguent, les approches dites « filter » et « wrapper », que nous avons toutes deux expérimentées. L'approche par filtrage (filter en anglais) a été la première utilisée, car elle est moins coûteuse en ressources informatiques. Il s'agit d'utiliser un critère d'évaluation indépendant du modèle prédictif pour juger de la pertinence des variables. On élabore donc le jeu de variables d'entrée par rapport à ce critère, puis on entraine notre modèle prédictif avec les variables sélectionnées (procédé illustré en figure 5.1).

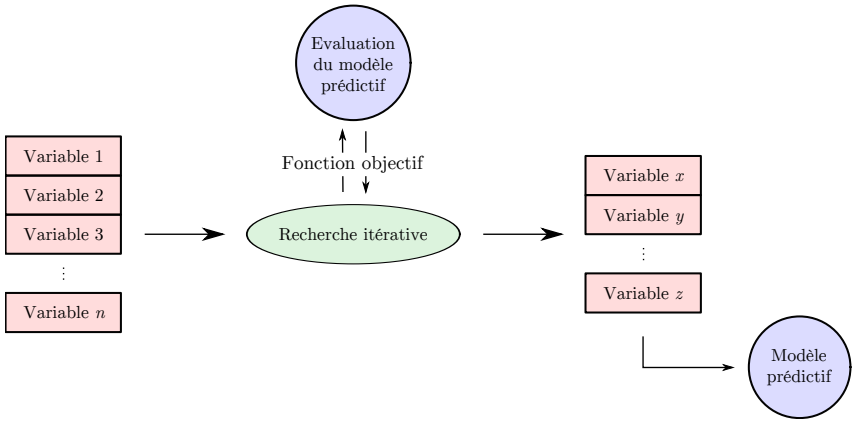


FIGURE 5.2 : Sélection de variables par approche « wrapper ».

L'approche « wrapper » à été décrite par Kohavi et John (1997). Elle utilise le modèle prédictif lui-même pour sélectionner les variables. La précision obtenue par le modèle prédictif sert de critère d'évaluation, ce qui assure que les variables lui sont adaptées (voir figure 5.2). C'est un avantage par rapport au filtrage car un critère indépendant, même pertinent, ne peut pas exactement représenter l'intérêt d'un jeu de variables pour le modèle prédictif. Il est donc

nécessaire, dans notre cas, d'entrainer et d'évaluer un PMC pour évaluer un jeu de variables, la précision du PMC faisant office de critère d'évaluation. Cette méthode est souvent plus couteuse en temps que le filtrage.

Nous allons voir comment une étude des variables à l'aide de leur Information Mutuelle (IM) peut permettre à l'utilisateur de retenir certains délais à appliquer aux variables. Nous verrons ensuite à la section 5.1.2 (page 117) la sélection de variable par Algorithme Génétique (AG), puis à la section 5.1.3 (page 122) par algorithme de recuit simulé. Ces deux métaheuristiques nous permettront d'utiliser une approche par filtrage (avec les AG) et une approche « wrapper » (avec le recuit simulé).

Enfin, nous nous pencherons sur l'utilisation des capacités de l'ACP afin de réduire la taille du jeu de variables à la section 5.1.4 (page 124). L'ACP n'est pas à proprement parler une méthode de sélection de variables, puisque l'objectif de son usage n'est pas d'identifier certaines variables en fonction de leur intérêt pour la prévision. Cependant, on a vu que les composantes principales issues de la transformation étaient hiérarchisées, ce qui rend possible un choix et permet *de facto* d'utiliser l'ACP comme méthode de sélection de variables.

Nous conclurons enfin à la section 5.1.5 (page 125) sur l'intérêt de ces différentes méthodes pour notre problématique.

5.1.1 Sélection de variables par information mutuelle

Nous avons précédemment vu que l'IM représente la quantité d'information qui est partagée par deux variables. Cette notion permet de départager efficacement les variables, apportant un critère d'évaluation rendant possible une sélection de variables par des outils objectifs. Elle est souvent utilisée à ces fins (Frénay *et al.*, 2013), notamment en prévision de la qualité de l'air (Perez et Reyes, 2001). En sélection de variables, l'IM peut avantageusement remplacer le coefficient de corrélation, afin de quantifier les relations non-linéaires entre les variables. L'IM peut s'interpréter comme le degré de prédictibilité d'une variable connaissant la seconde (Li, 1990).

Nous avons introduit l'IM à la section 4.4.3 page 101. Pour rappel :

$$IM(x,y) = H(x) + H(y) - H(x,y) \tag{4.21}$$

Il existe plusieurs méthodes pour calculer numériquement l'IM entre deux variables. Elles passent par les calculs de l'entropie des deux variables $H(x)$ et $H(y)$ et de celui de leur entropie conjointe $H(x,y)$. Pour rappel :

$$H(x) = -\sum_{i=1}^{n} P_i \, log_2(P_i) \tag{4.19}$$

$$H(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{i,j} \, log_2(P_{i,j}) \tag{4.20}$$

Le calcul de l'IM implique donc d'estimer les densités de probabilité des variables. La méthode la plus simple est la méthode par histogramme, qui revient à discrétiser la densité de probabilité pour pouvoir l'estimer numériquement. Cette méthode nécessite de fixer la largeur des barres de l'histogramme. Pour cela nous avons utilisé la méthode de Sturges (Sturges, 1926; Legg *et al.*, 2007).

D'autres méthodes existent pour estimer l'entropie. On note tout d'abord la méthode par kernel, c'est-à-dire utilisant une fonction de pondération (appelée kernel ou en français noyau) pour le calcul des densités de probabilité. Il s'agit d'une amélioration de la méthode par histogramme. Les « bâtons » des histogrammes sont remplacés par des kernels (par exemple des fonctions gaussiennes), qui permettent de lisser le calcul des probabilités (Moon *et al.*, 1995). Le calcul des densités de probabilité est plus long car ces fonctions kernels ont une forme plus complexe que les simples bâtons d'histogramme. Nous avons dû en abandonner l'utilisation pour la sélection de variables à cause de temps de calcul trop élevés. On note également la méthode par k plus proches voisins (k-nn, de l'anglais k-nearest neighbor) basée sur le calcul des distances moyennes de chaque point avec ses plus proches voisins (Kraskov *et al.*, 2004). Une comparaison de ces méthodes est effectuée dans la thèse de Master de Hao (2005).

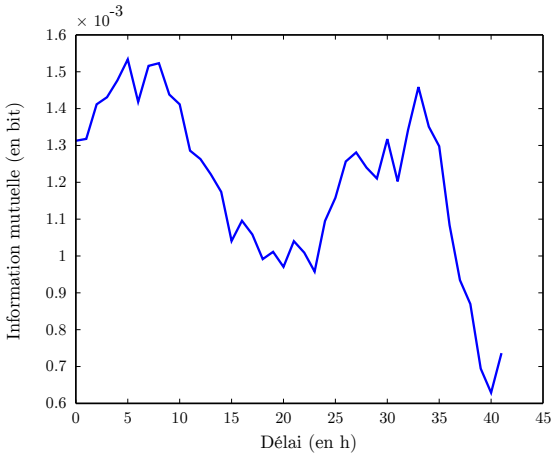


FIGURE 5.3 : Information mutuelle entre la série temporelle de PM10 à Canetto à $h+24$ et celle d'O_3 en fonction du délai appliqué à cette dernière (station Sposata).

Notre première approche utilisant l'IM a été de s'en servir pour sélectionner les délais pour les variables d'entrée. Pour chaque variable d'entrée, son IM avec la cible (série temporelle décalée dans le temps pour correspondre à l'horizon) est calculée, et ce pour tous les lags possibles entre 0 et 48. Ce procédé permet d'identifier le délai optimal pour maximiser l'information fournie sur la cible par la variable d'entrée. On s'imagine qu'en général, c'est l'observation la plus proche de l'horizon de prévision qu'il faut choisir, ce qui correspond à la variable non-décalée. Mais parfois, une variable partage plus d'information avec la cible quelques heures avant la dernière observation, comme on peut le voir sur la figure 5.3. On y voit qu'il vaut mieux utiliser la variable décalée de cinq heures si l'on veut maximiser l'IM avec la cible. On remarque un second pic d'IM vers 29 heures de délai, c'est-à-dire 24 heures après le premier pic qui révèle la périodicité de l'ozone.

La sélection de délais en choisissant les pics d'information mutuelle avec la cible permet d'améliorer les résultats (voir un exemple sur la figure 5.4). C'est un procédé assez simple, que nous avons utilisé dans le modèle de prévision d'ozone présenté lors du congrès EENVIRO 2013 (Tamas *et al.*, 2014). Cependant cette approche est limitée par son aspect univarié. Si l'IM entre les variables et la cible est optimisée, la redondance d'information entre les variables n'est pas gérée. Nous avons donc également utilisé des méthodes différentes qui se basent cette fois sur l'usage de métaheuristiques (« filter » et « wrapper ») et que nous allons maintenant présenter.

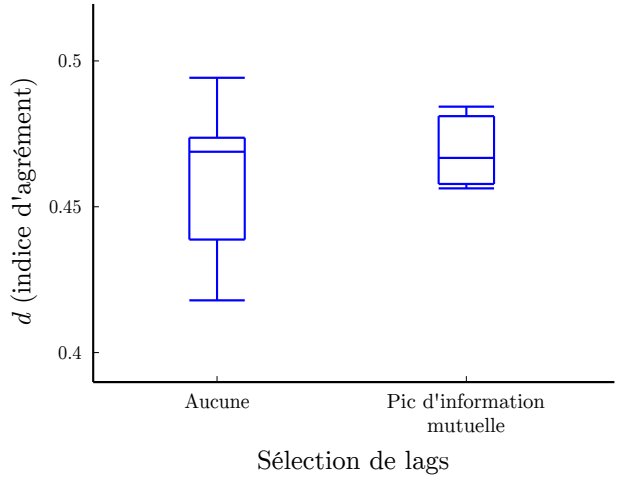


FIGURE 5.4 : Indices d'agrément obtenus par 10 modèles prédictifs de PM10 à h + 24 à Canetto, sans et avec sélection de lags en retenant les pics d'information mutuelle.

5.1.2 Algorithmes génétiques

Les AG sont des métaheuristiques utilisées en sélection de variables. Ils s'inspirent de processus naturels guidant l'évolution génétique des espèces. Ils ont été introduits par Holland (1975) et sont utilisés en combinaison avec les RNA avec succès, en sélection de variables mais également en gestion des poids et biais, d'architecture ou en sélection d'algorithme d'apprentissage (Yao, 1999). Ils ont été appliqués à la sélection de variables pour la prévision de la qualité de l'air avec une amélioration des résultats notables (Niska *et al.*, 2004; Ibarra-Berastegi *et al.*, 2008).

Leur principe est de faire évoluer une population en effectuant une sélection de ses meilleurs individus. La population est représentée par un ensemble de « chromosomes » qui représentent chaque individu. Ces individus correspondent à une configuration, que la métaheuristique doit optimiser. Il s'agit donc dans notre cas d'un choix de variables d'entrée uniquement. Les « chromosomes » qui représentent ces individus sont une analogie avec les chromosomes composés d'ADN et portant les gènes des organismes vivants. Nos chromosomes prennent la forme de vecteurs composés de nombres binaires de la taille du nombre initial de variables. Chacun de ces nombres binaires, qu'on assimile à des gènes, représente la présence ou l'absence d'une variable dans le jeu de données final, comme l'indique la figure 5.5. Un individu correspond donc à un choix de sélection de variables.

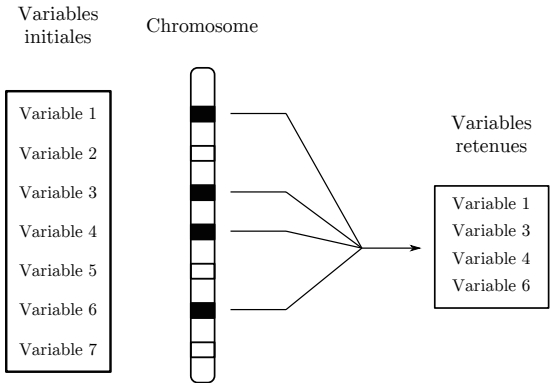


FIGURE 5.5 : Représentation d'un chromosome utilisé par les AG pour représenter un choix de sélection de variables.

Les chromosomes évoluent à chaque itération, et les meilleurs d'entre eux selon la fonction objectif sont sélectionnés et conservés pour l'itération suivante. L'évolution se fait par analogie à deux processus importants en génétique, le croisement ou enjambement (« crossover » en anglais) et la mutation.

Dans la nature, les croisements ont lieu pendant la méiose, étape de la production de gamètes lors de laquelle les chromosomes se mélangent entre eux. Il consiste en une recombinaison de deux chromosomes, qui se chevauchent et s'échangent des fragments, participant ainsi au brassage génétique. La mutation correspond à une modification spontanée et aléatoire de l'information génétique. Dans le cas d'AG, elle est représentée par la transposition aléatoire d'un élément binaire d'un chromosome, échangeant la présence de la variable concernée par son absence du jeu de données final, ou inversement.

Ces deux processus assurent le brassage génétique au sein de la population. Un nombre défini de chromosomes joue le rôle de « parent », c'est-à-dire que le produit de leur croisement donnera de nouveaux chromosomes « enfants » qui s'ajouteront à la population. Le même nombre de chromosomes en sera éliminé à chaque itération afin d'en maintenir l'effectif, mimant ainsi la sélection naturelle. A chaque itération, chaque chromosome correspondant à un ensemble de variables est évalué afin d'identifier les moins adaptés et les supprimer. Cette évaluation identifie également le meilleur candidat de l'itération.

A chaque itération, il est donc nécessaire d'évaluer l'ensemble des chromosomes. Cette contrainte ne favorise pas l'approche « wrapper ». En effet, à chaque itération il serait nécessaire d'entrainer et d'évaluer un grand nombre de PMC ce qui implique de longs temps d'exécution. De plus, une initialisation des paramètres du PMC avant apprentissage utilise une méthode avec une composante aléatoire (algorithme de Nguyen-Widrow), ce qui implique que les performances peuvent varier entre deux modèles ayant les mêmes variables. Ceci limite la pertinence d'une sélection d'individus basée sur la précision d'un seul PMC. Fixer d'avance les paramètres des neurones avec cet algorithme et les conserver pour toute l'expérience n'aurait pas plus de sens, puisque les différents chromosomes ont un nombre de variables différent ce qui modifie l'architecture des PMC correspondants.

Nous avons donc adopté une approche par filtrage avec les AG. Comme critère d'évaluation (noté C), nous avons utilisé la fonction basée sur l'information mutuelle proposée par Peng *et al.* (2005). Ce critère maximise l'IM entre chaque variable du jeu et la cible, tout en pénalisant l'IM entre les variables d'entrée. Elle minimise donc la redondance entre les variables et maximise leur pertinence vis-à-vis de la cible. On a :

$$C = \frac{1}{l} \sum_{\forall i} \left(IM(\mathbf{x}_i, \mathbf{y}) - \frac{1}{l-1} \sum_{j \neq i} IM(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{5.1}$$

avec l le nombre de variables, $\mathbf{x}_i$ une variable, $\mathbf{y}$ la variable cible. Ce critère a été utilisé avec des AG pour de la sélection de variables de RNA en prédiction de séries temporelles par Khazaee *et al.* (2008).

La taille de la population, le nombre de parents au sein de celle-ci et le taux de mutation sont des éléments clés du bon déroulement du processus de sélection. A chaque itération, chaque chromosome a une probabilité de subir une mutation, qui intervient sur l'un de ses gènes choisi aléatoirement. Puis chaque parent effectue un croisement avec l'un des chromosomes choisi au hasard, en pondérant sa probabilité d'être sélectionné par sa fonction C. Le croisement s'effectue entre deux gènes, choisis aléatoirement. La population doit être assez large pour permettre une certaine diversité, qui évite de tomber dans un maximum local de C, sachant que la durée d'une itération augmente avec la taille de la population. Le nombre de parents ainsi que le taux

de mutation favorisent le brassage. Ce brassage doit être assez grand pour éviter les maximums locaux, mais pas trop pour tout de même converger vers un optimal de C. Les opérations menées par l'algorithme génétique sont illustrées sur la figure 5.6.

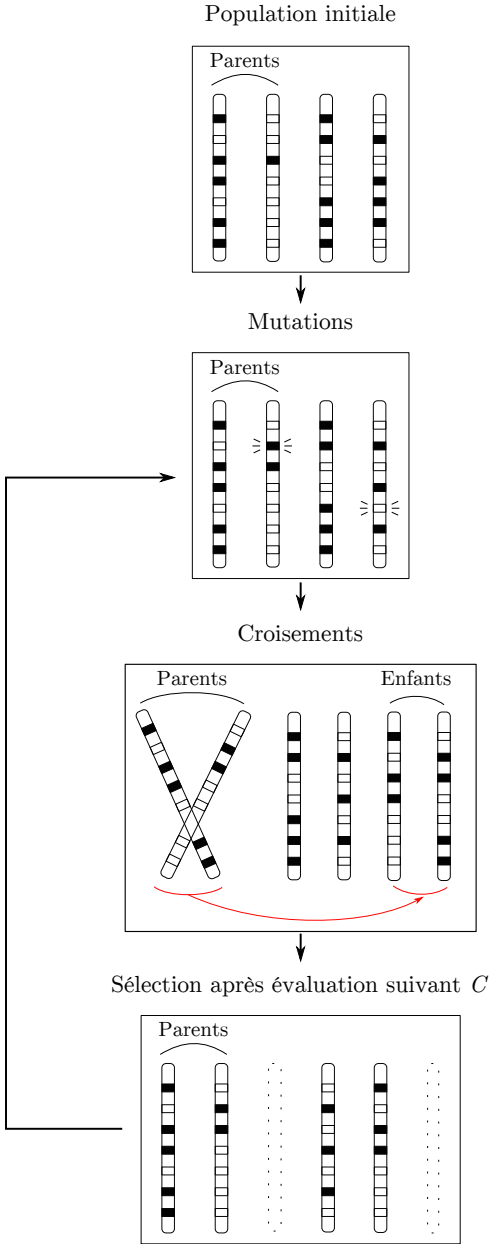


FIGURE 5.6 : Illustration d'un algorithme génétique.

Nous avons mené plusieurs expériences avec ces AG pour la sélection de variables et de délais, la principale difficulté étant de trouver les paramètres optimaux (taille de la population, nombre de parents, taux de croisement, taux de mutation). Fixer ce type de paramètre est une des difficultés majeures des métaheuristiques en général (Dréo *et al.*, 2003) et demande toujours une part d'empirisme.

Nous avons appliqué un AG à la sélection de variables et de délais pour un modèle prédictif de concentration de PM10 à Canetto, à l'horizon $h+24$. Les variables mesurées à choisir étaient les variables mesurés à Canetto (PM10, O_3, NO_2) et à Sposata (O_3, NO_2, Vitesse du Vent (VV), Direction du Vent (DV), Précipitations (P), Température en degrés Celsius (TC)) ainsi que les

sorties de modèle AROME au point le plus proche de la station (Géopotentiel (GEO) à 800 hPa, RS, Rayonnement Thermique (RT), Humidité Relative (HR) à 2 m, P, Température en degrés Kelvin (TK) à 2m, Epaisseur de la Couche d'Inversion (ECI) entre 0 et 1000 m, U et V à 10 m, U et V à 800 hPa). Pour limiter la complexité du problème, nous avons soumis un nombre restreint de délais. Pour les variables mesurées, les délais de 0 et 5 heures ont été proposés. Et pour les sorties de modèle, les échéances à $h+10$, $h+20$ et $h+24$ ont été soumises. Ceci correspond à un jeu initial de 51 variables, qui offre plus de 2.10^{15} possibilités de sélection de variables d'entrée.

Notre première approche (l'expérience AG **a**) a été d'utiliser une population limitée (30 individus) pour tester plus rapidement l'impact des autres paramètres. L'algorithme était utilisé sur une trentaine d'itérations. En effet, sur de si petites populations, l'algorithme converge rapidement vers un minimum local. La sélection des meilleurs individus est trop rapide par rapport aux mécanismes de brassage génétique. Les croisements effectués à partir des meilleurs individus mènent vite à une certaine « consanguinité » dans la population, et tous les chromosomes sont rapidement identiques ou très similaires. Mais l'avantage de cette approche est de pouvoir reprendre l'expérience à partir du meilleur résultat obtenu, en lançant de nouveau l'algorithme avec des paramètres différents mais en incluant dans la population le meilleur chromosome obtenu. Les autres individus sont renouvelés et assurent un certain brassage entre les expériences.

Répété plusieurs fois, ce procédé nous a permis de converger vers une valeur de C de plus en plus grande (voir figure 5.7, **a**). Cela a surtout permis d'explorer les paramètres autres que la taille de la population afin de lancer l'expérience suivante (AG **b**). On utilise maintenant une population d'une centaine d'individus. Le taux de mutation est fixé empiriquement à 15 %, et les 15 meilleurs individus à chaque itération font office de parents. Cette plus large population permet à l'algorithme de converger correctement (voir figure 5.7, **b**).

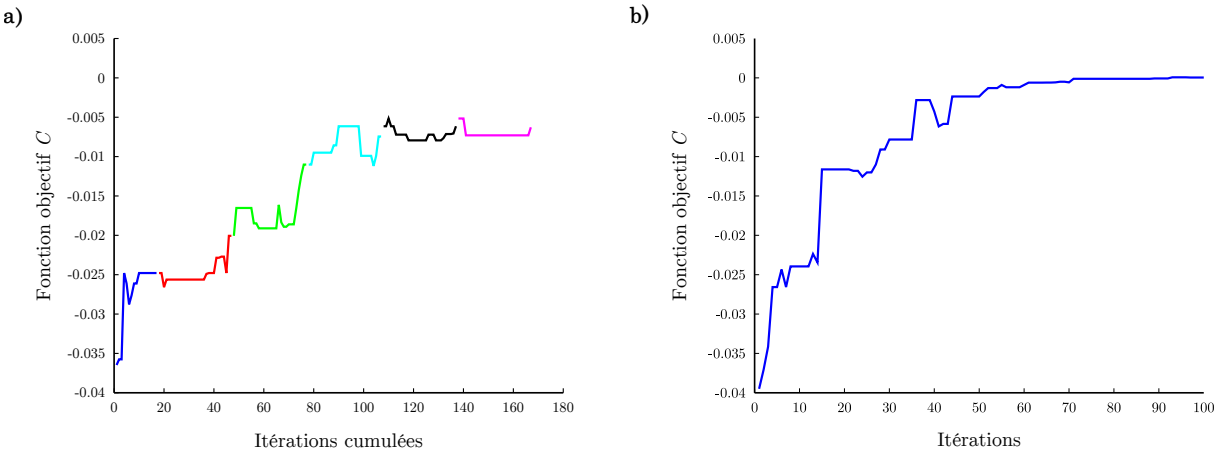


FIGURE 5.7 : Résultats de l'optimisation de C par algorithmes génétiques **a)** Plusieurs expériences reprenant le meilleur chromosome de l'expérience précédente **b)** expérience unique sur une population de 100 chromosomes.

Les variables qui correspondent à la suite d'expériences **a** sont les suivantes ; Les mesures à Canetto : PM10 à h et $h-5$ et NO_2 à h et $h-5$, les mesures à Sposata : P à h et $h-5$, ainsi que les sorties de modèle AROME suivantes : RT à $h+20$, ECI entre 0 et 1000 m à $h+10$, P à $h+10$, $h+20$, et $h+24$, et V à 800 hPa à $h+24$. Celle de l'expérience **b** sont beaucoup moins nombreuses, il s'agit des variables suivantes : PM10 à h à Canetto, P à $h-5$ à Sposata ainsi que les sorties de modèle P à $h+10$ et V à 800 hPa à $h+24$. Cette sélection drastique indique que pour optimiser C, l'AG a surtout supprimé des variables afin de limiter la redondance entre

celles-ci.

Nous avons construit les modèles prédictifs utilisant les variables sélectionnées par ces deux expériences, ainsi qu'un modèle témoin contenant les variables mesurées sans délai et les sorties de modèle à l'échéance $h + 24$. Les résultats sont indiqués sur la figure 5.8.

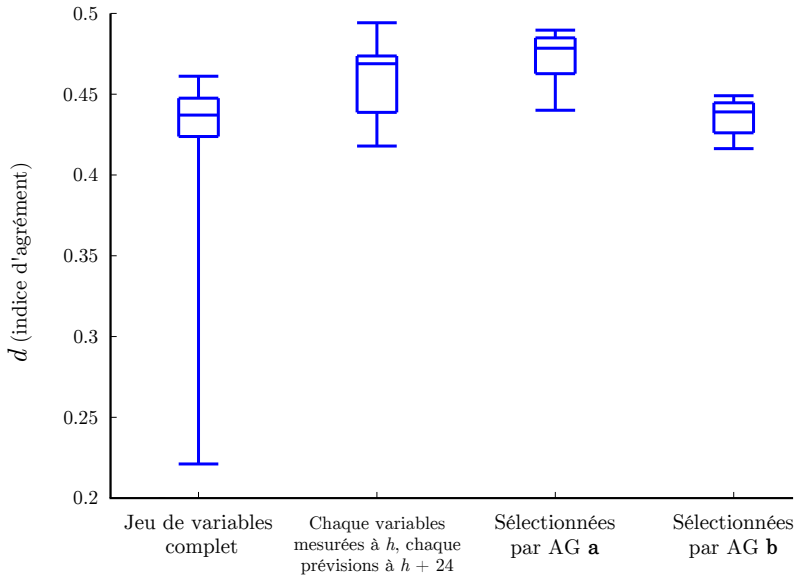


FIGURE 5.8 : Indices d'agrément obtenus par les modèles prédictifs de PM10 à Canetto à $h+24$ pour plusieurs sélections de variables.

On se rend compte que la meilleure sélection de variables par rapport à la fonction objectif C n'est pas la meilleure en termes de résultats de modèle prédictif. Elle est surpassée par le jeu de variables obtenu par AG **a**, dont la fonction C n'a pas été autant optimisée. Cela peut s'expliquer par le peu de variables retenues par AG **b**, et illustre l'imperfection de la fonction C. Elle ne convient pas parfaitement à notre problème, même si elle permet dans un premier temps d'optimiser les résultats, comme le montrent les résultats obtenus avec AG **a**.

Cette fonction C minimise la redondance entre les variables et maximise leur pertinence vis-à-vis de la cible. Nous avons tenté de l'améliorer en diminuant la minimisation de redondance grâce à un facteur k, pour obtenir la fonction C' :

$$C' = \frac{1}{l} \sum_{\forall i} \left(IM(\mathbf{x}_i, \mathbf{y}) - k \frac{1}{l-1} \sum_{j \neq i} IM(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{5.2}$$

Adopter une valeur de k inférieure à 1 permet d'accepter plus de redondance d'information entre les variables et d'avoir des jeux de variables plus larges. Cependant cette démarche apporte de l'empirisme supplémentaire.

On pourrait mettre en place une régulation des AG, qui à chaque itération vérifierait les performances obtenues par un PMC utilisant le jeu de variables courant (dans la même philosophie que la régulation de l'apprentissage des RNA). On passerait alors à une approche quasiment « wrapper » puisque nécessitant l'utilisation du modèle prédictif lui-même. Le problème de l'initialisation aléatoire des paramètres se poserait également. On a préféré passer à une réelle approche « wrapper », en changeant de méthode de recherche et en utilisant une autre méta-heuristique.

5.1.3 Recuits simulés

Le recuit simulé (« simulated annealing » en anglais) est une métaheuristique inspirée du domaine de la métallurgie. Il a été proposé en tant que méthode d'optimisation par Kirkpatrick *et al.* (1983) puis indépendamment par Cerny (1985). Les propriétés thermodynamiques du procédé de recuit sont utilisées à des fins d'optimisation.

En métallurgie, le recuit est une pratique qui permet, grâce à un cycle de chauffe et de refroidissement contrôlés, de limiter les imperfections microscopiques au sein de la pièce travaillée. Les métaux ont une structure cristalline. La structure formée peut avoir des défauts, qui se traduisent par une mauvaise position des atomes qui la composent. Lors d'un recuit, la montée en température permet d'apporter aux atomes l'énergie nécessaire pour se diffuser au sein du solide. Lors du refroidissement, ils vont retrouver une position finale. La manière dont ce refroidissement est effectué influence la nature finale solide. Un bon schéma de refroidissement permettra au solide d'atteindre un état stable, qui correspond à un minimum de son énergie. Un tel procédé permet de rendre les métaux plus malléables, plus ductiles et facilite leur travail. C'est l'inverse de la méthode de trempe, où l'on refroidit brutalement la pièce travaillée pour obtenir un métal dur et cassant.

Le recuit simulé est inspiré du recuit métallurgique, en particulier de son schéma de refroidissement. L'analogie est faite avec les atomes dont la température permet d'abord de se déplacer dans le cristal avant de trouver une position diminuant l'énergie interne du matériau.

Au niveau algorithmique, le recuit simulé peut être vu comme une évolution de l'algorithme de « hill climbing ». Cet algorithme d'optimisation consiste à modifier à chaque itération un élément de la configuration et à l'évaluer à nouveau. En cas d'amélioration du résultat de la fonction objectif, cette nouvelle configuration est conservée, sinon l'algorithme réutilise la configuration précédente pour sa prochaine itération. Il s'agit donc d'une méthode de recherche locale, qui converge vers un extremum local de la fonction objectif.

Le recuit simulé reprend la méthode du « hill climbing », en y introduisant le principe du schéma de refroidissement. L'algorithme utilise un vecteur binaire représentant la configuration à optimiser (semblable à un chromosome d'AG, voir figure 5.5 page 117). Un paramètre qu'on appelle par analogie la température est initialisé avec une certaine valeur, et va diminuer au cours des itérations selon un schéma de refroidissement. A chaque itération, l'algorithme va intervertir un élément du chromosome aléatoirement, comme pour une mutation d'AG. La différence avec le hill climbing vient du fait qu'en cas de non-amélioration de la fonction objectif, l'algorithme a une chance de tout de même retenir le nouveau chromosome. Cette probabilité dépend de la température. Ainsi, au début de l'expérience l'algorithme a de grandes chances de retenir les configurations qui n'optimisent pas la fonction objectif, mais au fil des itérations il se comporte de plus en plus comme un algorithme de hill climbing, convergeant vers l'optimum local. Le recuit simulé permet ainsi d'éviter efficacement la convergence vers un extremum local de la fonction objectif (Bertsimas et Tsitsiklis, 1993).

Par rapport aux AG qui travaillent sur une population large de chromosomes, le recuit simulé n'effectue qu'une évaluation par itération. Ceci nous a permis d'adopter une approche « wrapper » utilisant un PMC prédictif dont l'évaluation fait office de critère d'évaluation pour la métaheuristique. La fonction objectif qu'il faut optimiser est l'équivalent de l'énergie d'un matériau qui doit être minimisée par le recuit « physique ». On la notera donc E. Nous avons utilisé l'inverse de l'indice d'agrément d obtenu par le modèle comme fonction objectif E. Le choix de considérer l'inverse de l'indice d'agrément s'explique par le fait que l'on cherche à maximiser d plutôt que de le minimiser.

Nous utiliserons également une variable T, analogue à la température d'un recuit, qui diminue au cours des itérations. Cette variable température agit sur la probabilité de conserver une configuration n'améliorant pas la précision du PMC, suivant la règle de Metropolis (Metropolis *et al.*, 1953). Cette probabilité est donnée par :

$$\begin{aligned} P &= e^{\frac{-(E_k - E_{k-1})}{T_k}} && \text{si } E_k > E_{k-1} \\ P &= 1 && \text{si } E_k \leqslant E_{k-1} \end{aligned} \tag{5.3}$$

avec E_k et T_k respectivement l'énergie et la température à l'itération k. La température diminuant au cours de l'expérience, les configurations ne diminuant pas E ont de moins en moins de chances d'être conservées.

Les paramètres du recuit simulé à fixer sont la température initiale T_0 ainsi que le schéma de refroidissement, c'est à dire la règle encadrant la diminution de T au cours des itérations. Cette diminution doit être assez lente pour que l'algorithme ait de bonnes propriétés de convergence. Plusieurs schémas sont classiquement utilisés, comme une diminution linéaire ($T_k = T_0 - \alpha\, k$ avec $\alpha > 0$), ou une diminution en créneaux ($T_{k+1} = T_k - \alpha$ à un nombre régulier d'itérations). La température initiale peut être fixée de manière à obtenir un certain ratio d'acceptation en début d'algorithme (Ben-Ameur, 2004) mais une certaine correction empirique reste nécessaire. Le schéma de refroidissement sera lui aussi choisi après plusieurs essais, et on adoptera un schéma de décroissance exponentielle :

$$T_k = T_0\, e^{-\alpha k} \tag{5.4}$$

avec α une constante strictement positive. Afin de s'assurer que la température ne baisse ni trop vite ni trop lentement, on fixe empiriquement α en fonction du nombre d'itérations que l'on prévoit.

Ainsi, on a soumis le même problème de sélection de variables que celui traité avec les AG. Le jeu initial de variables d'entrée comprend donc les variables mesurées à Canetto (PM10, O_3, NO_2) et à Sposata (O_3, NO_2, VV, DV, P, TC) à h et $h-5$, et les sorties du modèle AROME au point le plus proche de la station (GEO à 800 hPa, RS, RT, HR à 2 m, P, TK à 2m, ECI entre 0 et 1000 m, U et V à 10 m, U et V à 800 hPa) aux échéances $h+10$, $h+20$ et $h+24$. La figure 5.9 montre l'évolution de E le long de l'expérience retenue.

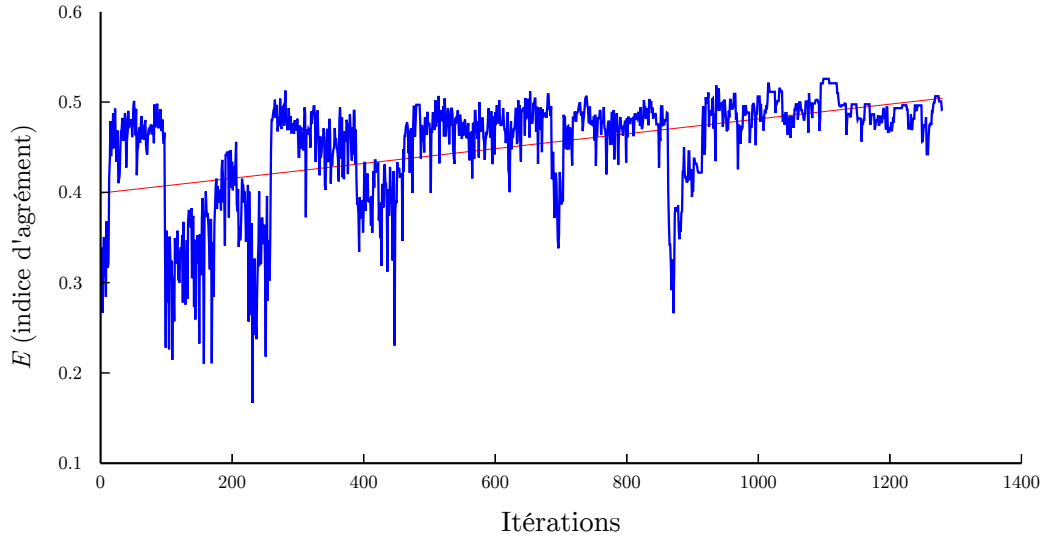


FIGURE 5.9 : Evolution de l'énergie E lors de l'expérience de recuit simulé, avec en rouge la droite de moindre carré correspondante illustrant la convergence.

On se rend compte que E converge au fil des itérations vers des valeurs maximales, d'abord chaotiquement puis de manière plus stable. On remarque vers la fin de l'expérience que l'algorithme accepte beaucoup moins les configurations diminuant les performances. Nous avons construit un modèle prédictif à partir des variables correspondant à la configuration la plus optimisée lors du recuit. Après cette optimisation, elle regroupait les PM10 mesurées à Canetto et l'O_3 à Sposata, sans aucuns lags, ainsi que les sorties de modèle suivantes : GEO à 800 hPa aux échéances $h+10$ et $h+24$, P à $h+20$ et $h+24$, ECI entre 0 et 1000 m à $h+24$, TK à 2 m à $h+10$, $h+20$ et $h+24$, U à 10 m à $h+10$, V à 10 m à $h+24$, U à 800 hPa à $h+10$ et à $h+20$, et V à 800 hPa à $h+10$ et $h+24$.

Les résultats du modèle avec ces variables sont montrés en figure 5.10. Ils y sont comparés à ceux obtenus avec le jeu de variables initial, ainsi qu'aux résultats obtenus pour ce même jeu, mais sans délai pour les mesure, et avec les échéances $h+24$ pour les sorties de modèle.

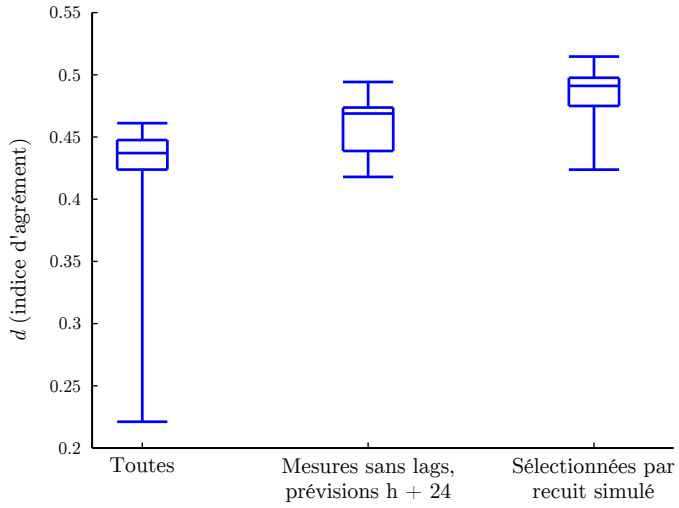


FIGURE 5.10 : Indices d'agrément obtenus avec le modèle prédictif de PM10 à Canetto à $h+24$, pour différentes sélections de variables d'entrée incluant le recuit simulé.

Les résultats du modèle avec recuit simulé sont meilleurs que ceux des deux configurations de base, l'un comprenant juste une occurrence de chaque variable, l'autre toutes les variables avec les délais et échéances envisagés. Le procédé du recuit simulé nécessite des paramétrages empiriques mais permet efficacement de gérer le choix des variables.

5.1.4 ACP en sélection de variables

L'ACP, qui fait déjà partie de nos outils de prétraitement (section 4.4.5 page 104) permet également la réduction du nombre de variables. L'ACP est une transformation du nuage de points que forment les variables, qui consiste en une projection sur ses vecteurs propres. Le calcul des vecteurs propres et de leur valeur propre respective permet de quantifier la participation de chacune des variables à chaque vecteur. En prétraitement, l'ensemble des variables est donc transformé en l'ensemble des composantes principales, qui sont hiérarchisées par leurs valeurs propres. A des fins de réduction du nombre de variables, il est possible de ne conserver qu'un certain nombre des composantes principales. La hiérarchisation par la valeur propre offre alors un moyen de sélection. On peut par exemple conserver les composantes jusqu'à expliquer un certain pourcentage de la variance, et « oublier » les autres. On obtient ainsi un jeu de variables plus petit, mais porteur de sens.

On peut voir ce procédé comme une sélection des composantes principales de type « filter », le critère d'évaluation étant la valeur propre des composantes. Cette sélection de variable ne comporte pas de choix aléatoire contrairement aux approches utilisant des heuristiques. Elle a l'avantage de la rapidité, puisque si l'on utilise déjà l'ACP en prétraitement, les valeurs propres des composantes principales sont déjà calculées et il ne reste plus qu'à se défausser des composantes à exclure.

Cette méthode laisse un choix à l'utilisateur, le nombre de composantes à conserver. Il peut être plus simple de raisonner en pourcentage de la valeur propre totale conservée. A partir de la somme de toute les valeurs propres, on conserve toutes les composantes principales jusqu'à ce qu'elles réunissent un certain pourcentage de cette valeur propre totale. La figure 5.11 montre les résultats des modèles prédictifs ainsi construits, en conservant les composantes pour plusieurs pourcentages de valeur propre, entre 100 % (pas de suppression de composantes) et 45%.

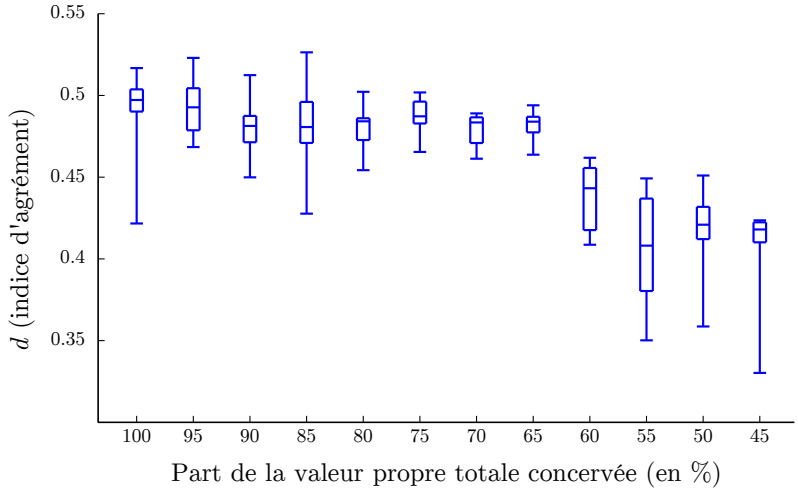


FIGURE 5.11 : Indices d'agrément obtenus avec le modèle prédictif de PM10 à Canetto à $h + 24$ après ACP, en conservant les composantes principales jusqu'à expliquer plusieurs pourcentages de la valeur propre totale.

On remarque que la précision des modèles prédictifs est plus grande quand on conserve un grand nombre de composantes principale, et diminue si l'on en supprime trop. La transformation de l'ACP produit des variables orthogonales entre elle, c'est-à-dire linéairement indépendantes. Ceci diminue fortement la redondance qui existe entre les variables, et explique pourquoi la conservation de 100 % des composantes permet d'obtenir des scores parmi les meilleurs. La suppression de composantes est intéressante jusqu'à à peu près 85 % de la valeur propre totale, et c'est d'ailleurs cette configuration qui a produit le meilleur modèle.

5.1.5 Bilan concernant les méthodes de sélection de variables

Nous avons abordé différentes méthodes de sélection de variables, certaines basées sur des critères d'évaluation, certaines faisant appel à des métaheuristiques, et d'autres sur les propriétés de l'ACP. Les métaheuristiques utilisées utilisaient soit une approche « filter » dans le cas des AG, soit une approche « wrapper » dans le cas du recuit simulé.

Ces méthodes ont toutes leurs contraintes et leurs avantages. Nous n'avons présenté ici que certains résultats pour illustrer leur fonctionnement, mais ces techniques ont été utilisées tout au long du reste de nos travaux. Les résultats du modèle prédictif de concentration en PM10 à Canetto à $h + 24$ avec les variables retenues par chaque méthode sont montrés à la figure 5.12.

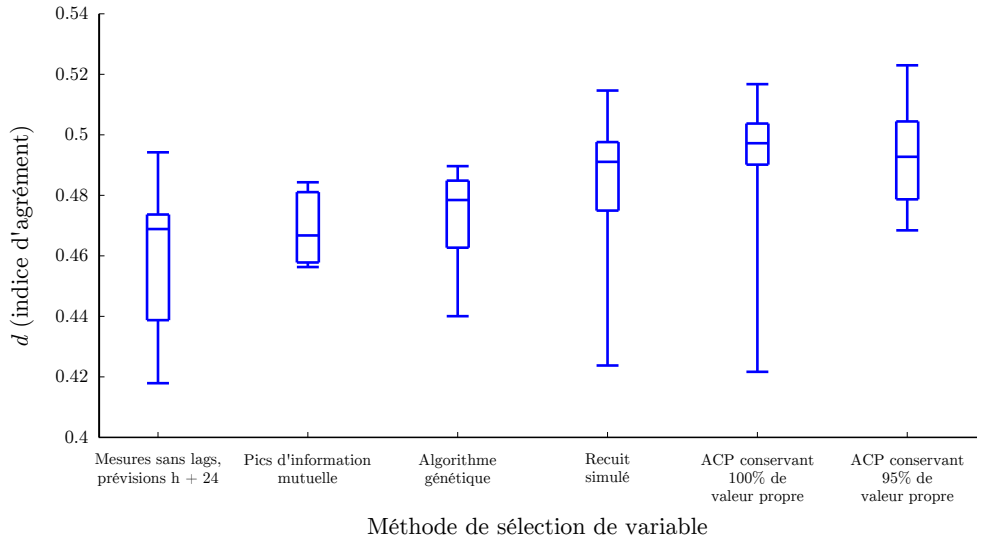


FIGURE 5.12 : Comparaison de la précision obtenue avec diverses méthodes de sélection de variable appliquées à la prévision des concentrations en PM10 à Canetto à h + 24.

En l'absence de métaheuristique utilisée pour la sélection, l'adoption des variables mesurées sans délais et des prévisions du modèle AROME à l'horizon $h + 24$ apparait comme la moins bonne solution. La sélection par l'utilisateur en utilisant l'IM pour sélectionner les meilleurs délais et échéance améliore les résultats, mais demande une étude préalable par l'utilisateur.

Les deux métaheuristiques utilisées ont également amélioré les prévisions. Les algorithmes génétiques, en l'absence de meilleur critère d'évaluation que la fonction C (équation 5.1) de minimisation de redondance et maximisation de la pertinence, demanderaient une régulation de l'algorithme ou d'autres tests empiriques pour paramétrer le critère d'évaluation. Le recuit simulé lui sera préféré. L'approche « wrapper » permet d'adapter les variables directement au modèle prédictif. La paramétrisation de l'algorithme demeure tout de même empirique, problème typique des métaheuristiques.

Finalement, l'usage de l'ACP en sélection de variable apporte plusieurs avantages par rapport aux autres méthodes. Tout d'abord, son utilisation est simple, même s'il faut fixer le nombre de variables que l'on souhaite conserver. Elle présente de bons scores de précision, proches de ceux obtenus par recuit simulé. Nous utiliserons l'ACP pour construire nos modèles pour ces raisons.

Le recuit simulé apparait également comme un outil utile. Une perspective est de l'utiliser dans un cadre plus large que la sélection de variables. Il pourrait permettre de fixer d'autres paramètres de la configuration d'un PMC que ses variables. On pourrait l'appliquer à la sélection d'architecture, ou la sélection d'algorithmes d'apprentissage par exemple.

La sélection de variables permet d'améliorer nos résultats, en ne conservant que les variables les plus utiles aux modèles prévisionnels. Cette étape préliminaire est importante et permet de se conformer au principe de parcimonie. D'autres choix que celui des variables doivent également être pris en compte pour assurer la parcimonie du PMC, notamment le choix du nombre de neurones, et donc du nombre de paramètres du réseau. Nous allons voir à la section suivante une méthode vouée à l'optimisation du nombre de paramètres.

5.2 Elagage

Un PMC comporte un certain nombre de paramètres, ses poids et ses biais, dont le nombre dépend du nombre de neurones et de variables d'entrées. Plus ce nombre de paramètres est grand, plus le PMC va se spécialiser sur ses données d'apprentissage et risquer le sur-apprentissage. S'il est trop réduit en revanche, le modèle sera trop simpliste et ne pourra capter toutes les interactions entre variables lors de l'apprentissage. Toujours pour suivre le principe de parcimonie, il est utile de réguler ce nombre de paramètres. La sélection de variables permet une première régulation, mais il est également possible de réguler les neurones eux-mêmes.

Afin de contrôler le nombre de paramètres d'un RNA, on peut utiliser des méthodes d'élagage (« pruning » en anglais) dont le principe est d'éliminer les paramètres les moins utiles au réseau. Nous avons développé une technique d'élagage basée sur l'utilisation de l'algorithme de LM afin d'éliminer les paramètres inutiles. Cette méthode a été présentée à la « 2nd International Conference on Mathematical Modeling in Physical Sciences » (Voyant *et al.*, 2014). Elle a également fait l'objet d'une publication (Voyant *et al.*, 2015).

La méthodologie est la suivante. L'élagage a lieu avant d'entrainer le PMC. On prélève un sous-échantillon du jeu d'apprentissage, afin de générer un ensemble de N systèmes de m équations, m correspondant au nombre de paramètres du PMC. Chacun des N systèmes représente en effet m exemples de variables d'entrée $\mathbf{x}$ correspondant à une observation de la cible y. Le PMC doit modéliser y tel que :

$$y = f(\mathbf{x}, \omega, \mathbf{b}) = \sum_{j=1}^{m} \omega_{jo} \; . \left(g(\sum_{i=1}^{n} \omega_{i,j} \; . \; \mathbf{x}_i + b_j) \right) + b_o \tag{5.5}$$

avec n le nombre de variables d'entrées, m le nombre de neurones cachés, $\omega_{i,j}$ et $\omega_{j,o}$ respectivement les poids des neurones cachés et du neurone de sortie, et b_i et b_o respectivement les biais des neurones cachés et du neurone de sortie. On peut donc utiliser l'algorithme de LM pour résoudre le système formé par m réalisations de l'équation 5.5 afin d'obtenir les paramètres du PMC qui satisfont le système.

Cette opération est donc réalisée N fois, ce qui permet d'obtenir une distribution de taille N de la valeur des paramètres du réseau. A partir de cet échantillon, on calcule l'intervalle de confiance bootstrap (DiCiccio et Efron, 1996) de la valeur des poids et biais. Tous les paramètres dont les extrémités de la distribution (les 2ème et 98ème centiles) n'ont pas le même signe ont une distribution dont le centre est proche de zéro. Ils sont considérés comme non-significativement différents de zéro et sont supprimés.

Le PMC est ainsi élagué. Son architecture s'en trouve simplifiée, les connexions entre les neurones dont le poids a été élagué étant supprimées. Ce PMC est ensuite entrainé et évalué normalement. Nous l'avons utilisé lors de certaines expériences, notamment celles impliquant un grand nombre de variables d'entrées.

Cependant, les résultats de cet élagage sur nos données de qualité de l'air (présentés à la figure 5.13 pour 15 entrainements dans chaque configuration) ne sont pas aussi bons que ceux que nous avons eu avec d'autres jeux de données météorologiques (voir Voyant *et al.*, 2015).

On voit sur la figure 5.13 que l'élagage dégrade systématiquement les résultats. Avec deux couches cachées, architecture présentant plus de connexions et de possibilités d'élagage, les résultats sont moins bons qu'avec une seule couche cachée. De plus dans cette configuration, sur la quinzaine d'apprentissages menés, tous ont donné des résultats extrêmement proches sauf un, suggérant un piège dans un minimum local lors de l'élagage à l'aide de l'algorithme de LM.

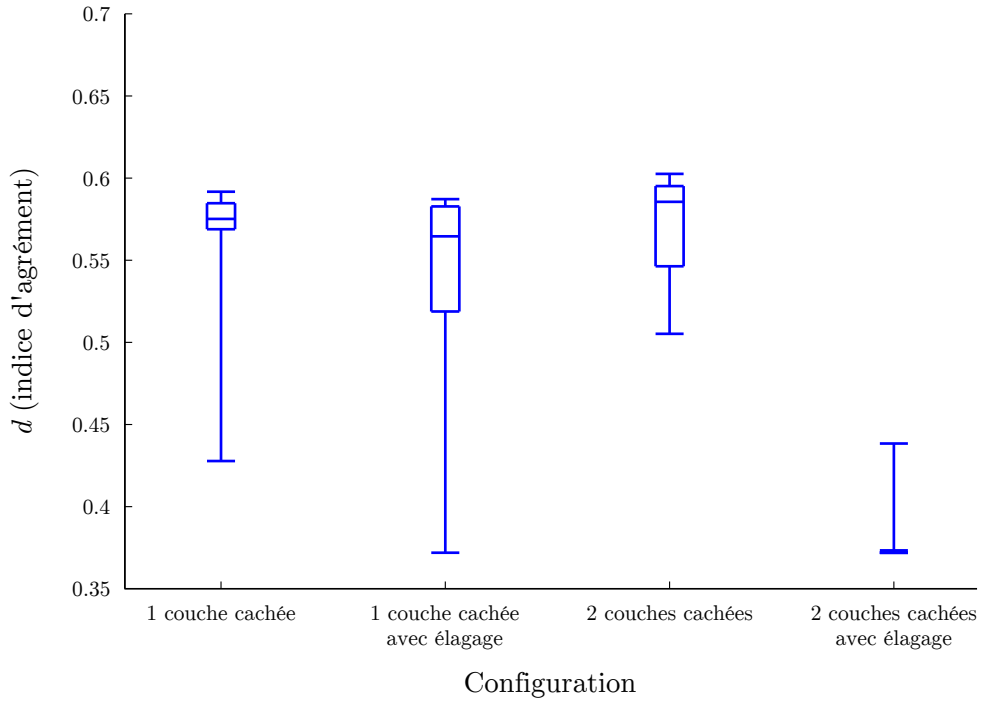


FIGURE 5.13 : Comparaison de la précision obtenue avec une ou deux couches cachées, avec et sans élagage pour la prévision des concentrations en PM10 à Canetto à h + 24.

Bien qu'avec d'autres données, cette méthode ait permis de réduire le nombre de paramètres des PMC sans pénaliser les résultats, nous en avons abandonné l'usage à des fins de prévision de la qualité de l'air.

5.3 Création des modèles prévisionnels pour la Corse

Nous avons abordé deux domaines (sélection de variables et élagage) permettant d'améliorer le fonctionnement des PMC pour la prévision de séries temporelles. Revenons maintenant à la problématique corse de prévision de la qualité de l'air.

Nous allons maintenant nous intéresser au modèle AIRES qui réalise des prévisions en Corse. Ce Chemical Transport Model (CTM) tel qu'il est déployé actuellement n'est pas apparu suffisamment précis pour effectuer seul les prévisions en Corse (voir section 3.3, page 79). Cela a motivé ces travaux afin d'utiliser des modèles statistiques, qui peuvent être meilleurs que le CTM, mais peuvent également fonctionner conjointement avec ce dernier.

Nous présenterons ensuite les modèles construits pour les villes d'Ajaccio (section 5.3.2, page 130) et de Bastia (section 5.3.3, page 134).

5.3.1 Utilisation des sorties du modèle AIRES par le PMC

Nous allons voir comment les PMC peuvent fonctionner avec AIRES, en utilisant ses sorties brutes dans leurs données d'entrée, pour la prévision d'ozone et de PM10. Plusieurs expériences ont été menées afin d'évaluer l'intérêt de l'usage de sorties d'AIRES. Les modèles que nous avons construits utilisent en entrée les données de pollution mesurées par Qualitair Corse ainsi que les sorties de modèles AROME.

Le jeu de données d'entrée « de base » est composé de la mesure du polluant endogène, ainsi que les variables d'AROME suivantes : GEO à 800 hPa, U et V à 10m et à 800 hPa, TK à 2m, ECI, RS, RT, HR à 2m, PA au sol et PA recalculée pour le niveau de la mer. Pour les variables du modèle prédictif, plusieurs échéances ont été sélectionnées ($h + 15$, $h + 20$, $h + 24$), sachant que l'ensemble du jeu sera soumis à une ACP qui réduira sa taille.

Trois cas de figure ont été envisagés. Le premier correspond à l'usage exclusif de ces variables. Le second cas correspond à l'utilisation de ces mêmes variables, en plus de la prévision de concentrations de la variable endogène provenant d'AIRES. Dans le troisième cas, les concentrations de tous les polluants fournis par AIRES sont utilisées. Les échéances des prévisions sont les mêmes que celles utilisées pour AROME. Les polluants dont AIRES fournit les prévisions sont : O_3, PM10, H_2O_2, NO, NO_2, HNO_3, HONO et le PéroxyAcétylNitrate (PAN). Pour chaque cas, dix PMC ont été entrainés afin de comparer les résultats.

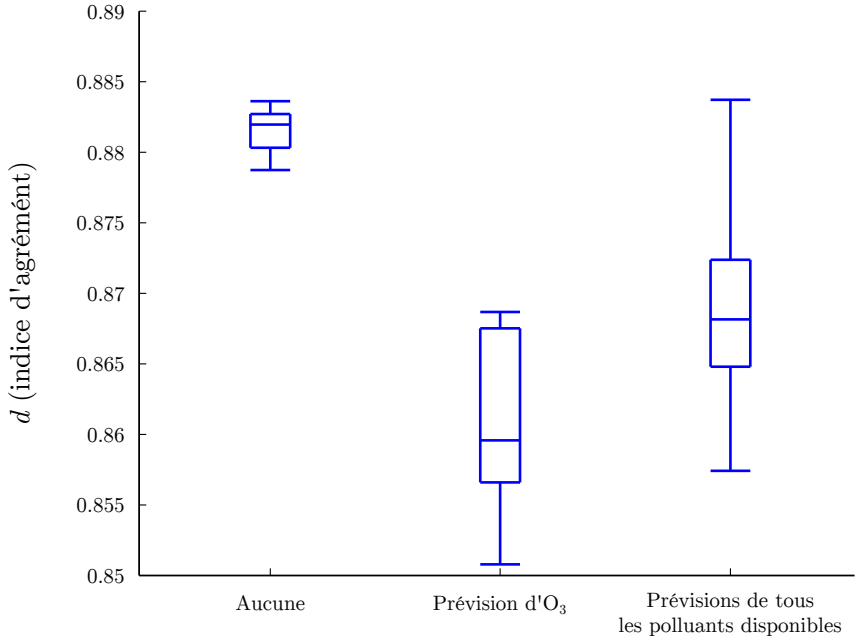


FIGURE 5.14 : Evaluation des prévisions d'ozone à Canetto, en fonction des sorties d'AIRES utilisées en entrée.

La figure 5.14 montre les résultats obtenus avec ces trois jeux de données pour la prévision d'ozone à Canetto. Pour ce polluant, il est assez clair que l'usage de données d'AIRES dégrade la précision des PMC. L'ajout de la seule variable O_3 donne les moins bons résultats, suivis par l'usage de l'ensemble des polluants. Le jeu de données de base l'emporte avec une plus grande stabilité de l'indice d'agrément obtenu par les différents PMC entraînés.

L'état photochimique de l'atmosphère est peut être mieux décrit par l'ensemble des polluants que par l'ozone seul. L'utilisation de variables représentant les NO_x (oxydes d'azote, NO et NO_2) et les espèces réservoir de NO_x comme le PAN peut permettre d'apporter des informations plus fiables que l'ozone seul, dont on a vu la relative imprécision sur la Corse. En tout cas, ce cas montre qu'il faut se poser la question avant d'inclure les sorties d'AIRES à un jeu de données d'entrée.

Les résultats obtenus pour les PM10 à Canetto sont décrits à la figure 5.15. Contrairement à l'ozone, les prévisions de particules utilisant des sorties d'AIRES semblent bénéficier de ces

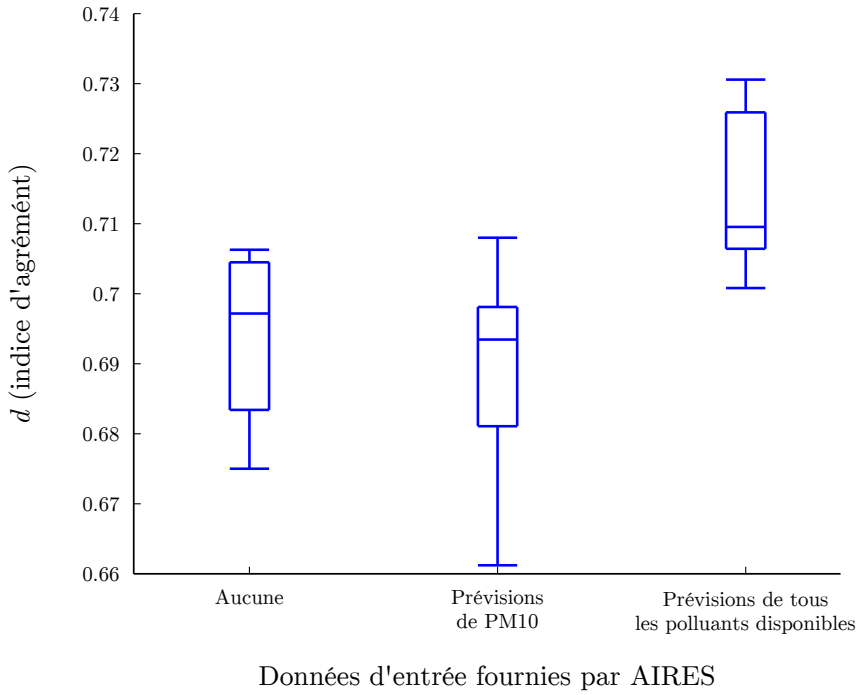


FIGURE 5.15 : Evaluation des prévisions de PM10 à Canetto, en fonction des sorties d'AIRES utilisées en entrée.

données, puisque l'indice d'agrément est globalement meilleur quand tous les polluants sont utilisés en entrée. L'usage de la seule variable de la concentration en PM10 n'apporte par contre pas de gain de précision. L'ensemble des polluants permet d'atteindre au mieux $d = 0.730$ au lieu de 0.706 pour la meilleure configuration sans AIRES. Le gain est faible, mais l'amélioration est robuste sur l'ensemble des dix PMC utilisés (meilleurs médiane et centiles). La précision d'AIRES est peut-être meilleure pour d'autres espèces que les PM10, qui permettent d'apporter au PMC les informations sur l'arrivée de masses d'air transportées. Le réseau de neurones peut alors faire le lien entre ce type d'événement et les niveaux de PM10.

Au final, l'utilisation des prévisions de polluants d'AIRES permet donc d'améliorer les prévisions dans certains cas, mais pas systématiquement. On peut donc envisager de construire le modèle opérationnel avec ces données, mais il faut mener des expérimentations spécifiques pour en valider l'intérêt. Dans le cas où les sorties d'AIRES dégradent la précision, il est possible que les informations délivrées par AIRES en relation avec les niveaux d'ozone soient déjà présentes dans le jeu de données contenant les observations de Qualitair Corse et les prévisions du modèle AROME. On peut tout de même s'attendre à une amélioration des prévisions utilisant des sorties d'AIRES quand le cadastre régional de la Corse sera utilisé.

Dans le cadre de nos expérimentations, nous avons donc utilisé ces données pour la prévision après vérification de la précision qu'elles apportent. De plus, ces données sont disponibles depuis le début de l'année 2011, ce qui peut réduire de manière contraignante la taille du jeu de données d'entrée. En fonction de l'étude réalisée, l'usage de ces sorties n'est donc pas systématique.

5.3.2 Prévision à Ajaccio

Nous avons construit les modèles de prévision de particules et d'ozone à Canetto et à Giraud en suivant la méthodologie présentée à la section 4.5, en utilisant l'ACP pour la sélection de variables. Les configurations sont présentées au tableau 5.1.

Les mesures de polluants correspondent à toutes les séries temporelles de polluants mesurées aux stations Canetto et Sposata. La sélection de variables est assurée par l'ACP, avec différents pourcentages de valeur propre conservée. Dans ces deux expériences, le remplacement des valeurs manquantes s'est avéré plus néfaste qu'utile. Une seconde couche cachée n'a pas permis d'améliorer les résultats dans ce cas. Pour la prévision d'ozone, l'usage des sorties d'AIRES a été écarté après expérimentations, comme on a pu voir à la section précédente. L'usage de plus de délais et d'échéances s'est par contre averé plus efficace.

TABLEAU 5.1 : Configurations obtenues pour la prévision de PM10 et d'ozone à l'horizon $h+24$ à Ajaccio.

Polluant	Variables de base (et délais/échéances)	Sélection de variables	Prétraitements	Nombre de couche cachée	Neurones cachés
PM10	Mesures de polluants (h) Prévisions AROME ($h+23$) Prévisions AIRES ($h+23$)	ACP (97 %)	Stationnarisation Normalisation ACP	1	10
O_3	Mesures de polluants (h, $h-1$, $h-2$, $h-5$) Prévisions AROME ($h+15$, $h+20$, $h+23$)	ACP (80 %)	Stationnarisation Normalisation ACP	1	8

Nous allons maintenant présenter les résultats obtenus avec les meilleures configurations identifiées pour la prévision de PM10 à Ajaccio (station de Canetto) à l'horizon $h+24$. Ils sont présentés à la figure 5.16.

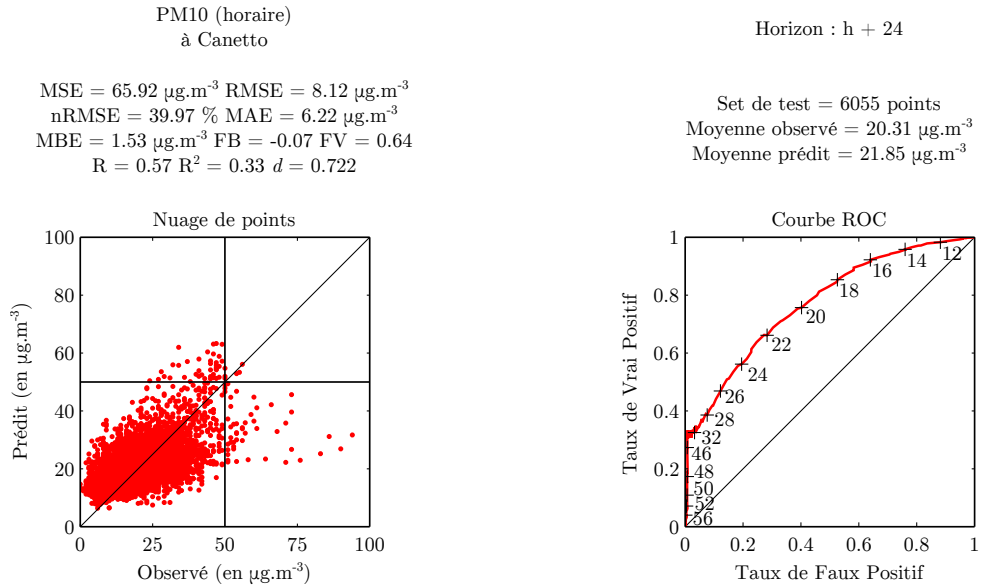


FIGURE 5.16 : Résultats du modèle prévisionnel de concentration horaire de PM10 à Canetto, construit suivant notre méthodologie. Les lignes droites sur le nuage de points indiquent le seuil d'information.

On obtient de bons résultats. L'indice d'agrément d atteint 0.722, un bon score pour la prévision de PM10 24 heures à l'avance. Ce score se traduit par un nuage de points réunis autour de la droite $x = y$. La dispersion autour de cet axe est visible, mais on peut remarquer que certaines des valeurs de concentration approchant le seuil d'information de 50 $\mu g.m^{-3}$ (indiqué sur la figure) sont prédites à des valeurs élevées. Cependant, les plus fortes concentrations, qui

dépassent les 80 µg.m^{-3} (seuil d'alerte) sont ratées par le modèle. Rappelons que les alertes ne sont pas prévues pour des valeurs horaires comme celles utilisées ici mais pour des valeurs journalières. Ce comportement vis-à-vis des valeurs rares et extrêmes reste cependant problématique. On peut voir sur la courbe ROC (pour Receiver Operating Characteristic) comment se comporte le modèle en fonction des seuils de concentration considérés. On y retrouve les capacités du modèle à prévoir les valeurs jusqu'à 46 µg.m^{-3} avec un taux de vrai positif au dessus des 30 %, avant de voir la précision du modèle diminuer.

On peut donc dire que les modèles basés sur des PMC prévoient les concentrations de PM10 avec de bons résultats, mais qui se dégradent de manière problématique pour les plus fortes valeurs de concentrations.

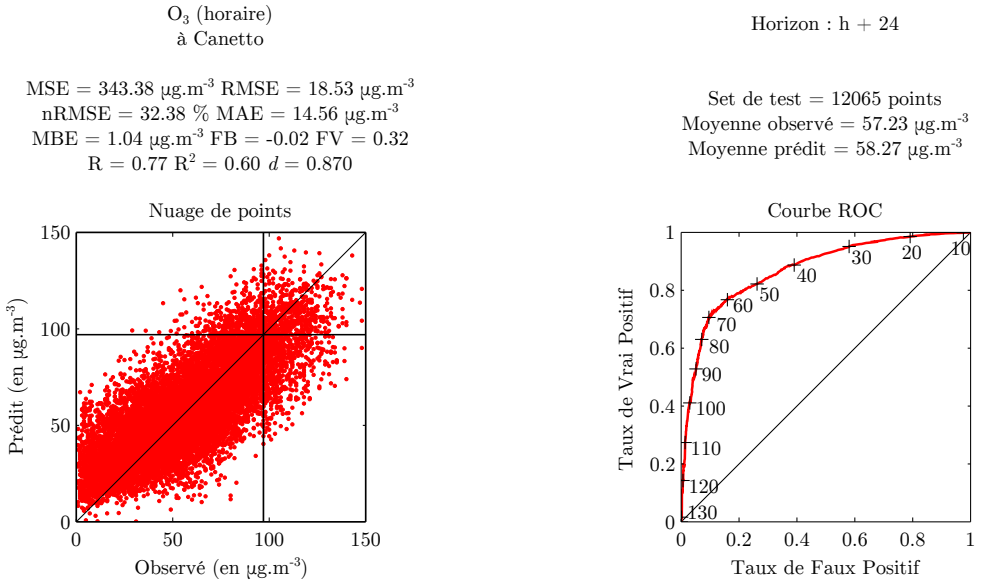


FIGURE 5.17 : Résultats du modèle prévisionnel de concentration horaire d'ozone à Canetto, construit suivant notre méthodologie.

Les résultats des modèles de prévision de concentrations d'ozone sont montrés à la figure 5.17. Les résultats sont bons, tant en termes d'indice d'agrément ($d = 0.87$) que sur le nuage de points, où l'on peut voir que les points suivent bien la droite $x = y$. Le centile 90 (noté C90, 90% des données de l'échantillon sont inférieures à la valeur du C90) des concentrations observées d'ozone est indiqué sur le nuage. On peut voir que même les valeurs les plus élevées sont correctement prédites, contrairement aux PM10. La nRMSE, calculée pour la valeur moyenne observée de 5.27 µg.m^{-3}, est à considérer avec en tête les valeurs d'incertitude de mesure de l'ozone vu à la section 3.2.4 (page 75), estimée à 12.1 % et 10.6 % en 2012 et 2013 respectivement.

La courbe ROC montre de bons résultats de détection jusqu'à la centaine de µg.m^{-3} au-delà de quoi les valeurs mesurées font partie des 10 % les plus élevées. Aucun seuil d'information n'a par contre été détecté, ni observé, sur la période de test mais également depuis le début des mesures en Corse.

Les séries temporelles observées et prédites sont présentées en figure 5.18 pour les PM10 et en figure 5.19 pour l'ozone. On peut y voir que les prévision sont assez bonnes, tout en pouvant sous-estimer les valeurs les plus fortes pour l'ozone, et rater certain pic resserrés pour les PM10. Sur la figure 5.18, on peutvoir un pic prévu pour la matinée du 23 décembre qui n'a pas eu lieu. Ceci illustre un effet de persistance qu'on retrouve parfois sur les prévisions, mais pas systématiquement. Le PMC est alors particulièrement influencé par la variable d'entrée

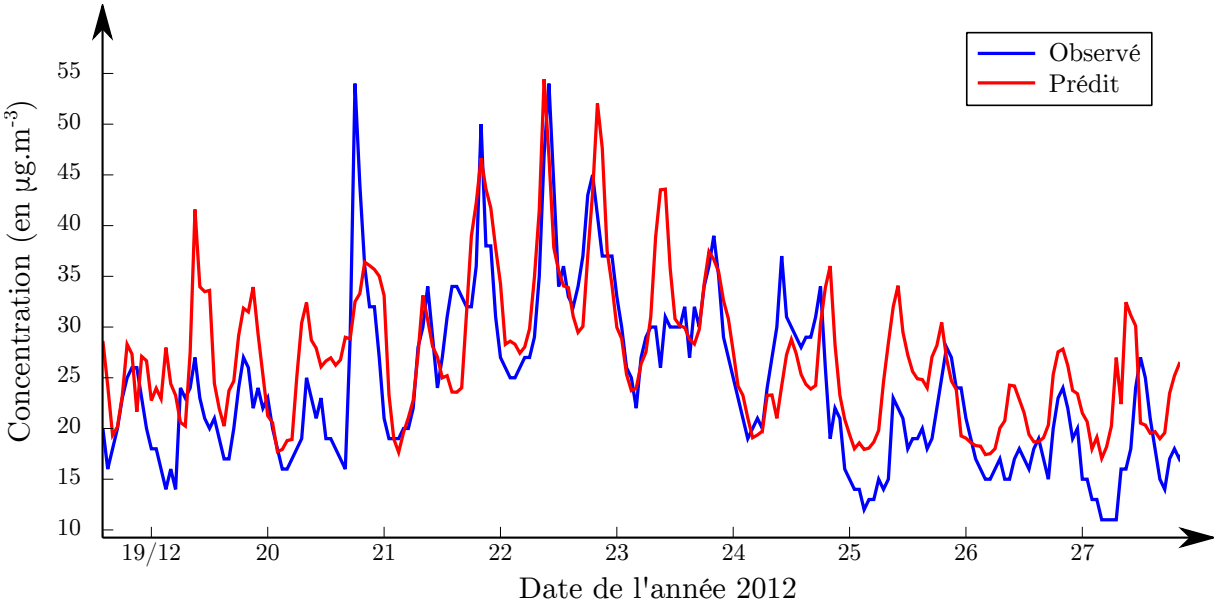


FIGURE 5.18 : Séries temporelles de concentrations horaires de PM10 à Canetto observée et prédite par le modèle prévisionnel.

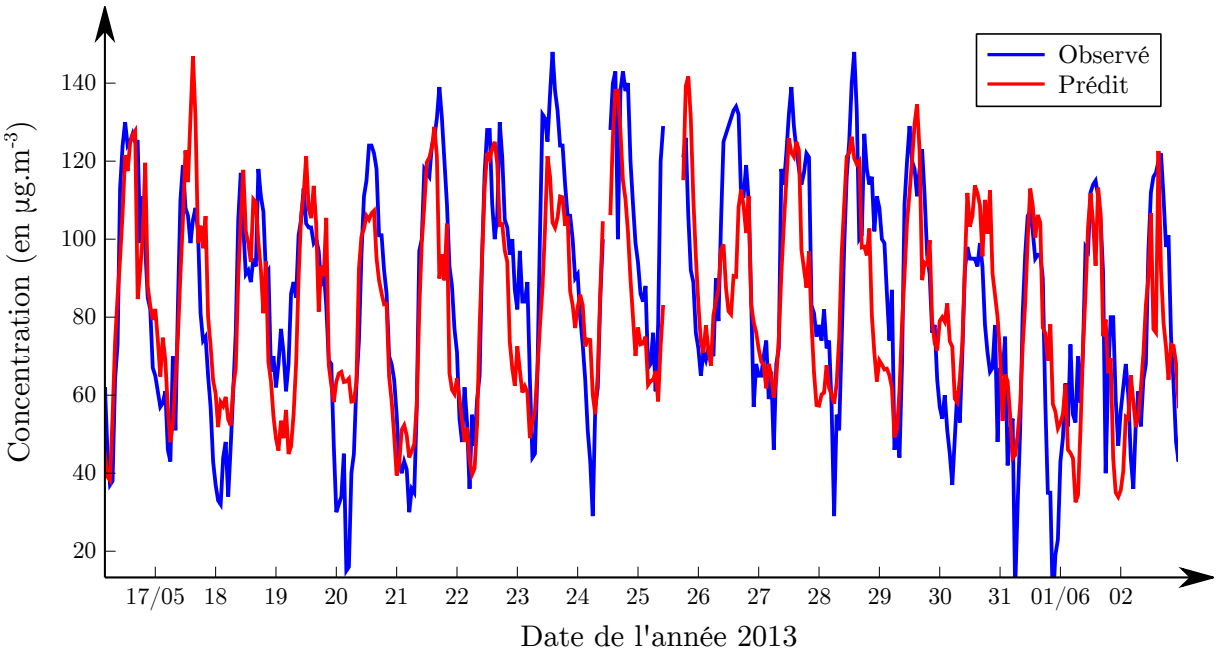


FIGURE 5.19 : Séries temporelles de concentrations horaires d'ozone à Canetto observée et prédite par le modèle prévisionnel.

endogène, qui ici présente la valeur élevée de la veille. Cette variable peut avoir tendance à induire un comportement de modèle de persistance (voir section 2.3.1 page 29), pour lesquels la prévision est égale à la valeurs observée de la veille. L'usage de nombreuses variables exogènes chargées d'information évite ce comportement.

5.3.3 Prévision à Bastia

Nous avons également créé les modèles prévisionnels de PM10 et d'ozone pour la ville de Bastia. Les données utilisées proviennent des stations Giraud et Montesoro, ainsi que de la station Bastia de Météo-France. Il s'agit de prévisions à la station urbaine de Giraud. Les configurations sont présentées au tableau 5.2.

Contrairement au modèle ajaccien, la prévision bénéficie positivement de l'usage des sorties d'AIRES, pour l'ozone comme pour les PM10. Le remplacement des données manquantes n'a pas été retenu, comme a Ajaccio.

TABLEAU 5.2 : Configurations obtenues pour la prévision de PM10 et d'ozone à l'horizon $h+24$ à Bastia.

Polluant	Variables de base (et délais/échéances)	Sélection de variables	Prétraitements	Nombre de couche cachée	Neurones cachés
PM10	Mesures de polluants (h) Prévisions AROME ($h+23$) Prévisions AIRES ($h+23$)	ACP (98 %)	Stationnarisation Normalisation ACP	1	7
O_3	Mesures de polluants (h) Prévisions AROME ($h+23$) Prévisions AIRES ($h+23$)	ACP (96 %)	Stationnarisation Normalisation ACP	1	8

La figure 5.20 montre les résultats de la prévision de PM10, à l'horizon $h+24$. On peut voir sur le nuage de points que la plupart des points observés autour de 50 µg.m^{-3} sont correctement prédits, mais que les valeurs au delà sont sous-estimées par le modèle. On retrouve ce défaut de détection sur la courbe ROC du modèle, où les taux de détections chutent rapidement à partir de 30 µg.m^{-3}.

Si l'on ne s'intéresse pas spécialement aux valeurs élevées, le modèle bastiais a une précision similaire au modèle ajaccien (avec un indice d'agrément d égal à 0.725 pour 0.722 à Ajaccio).

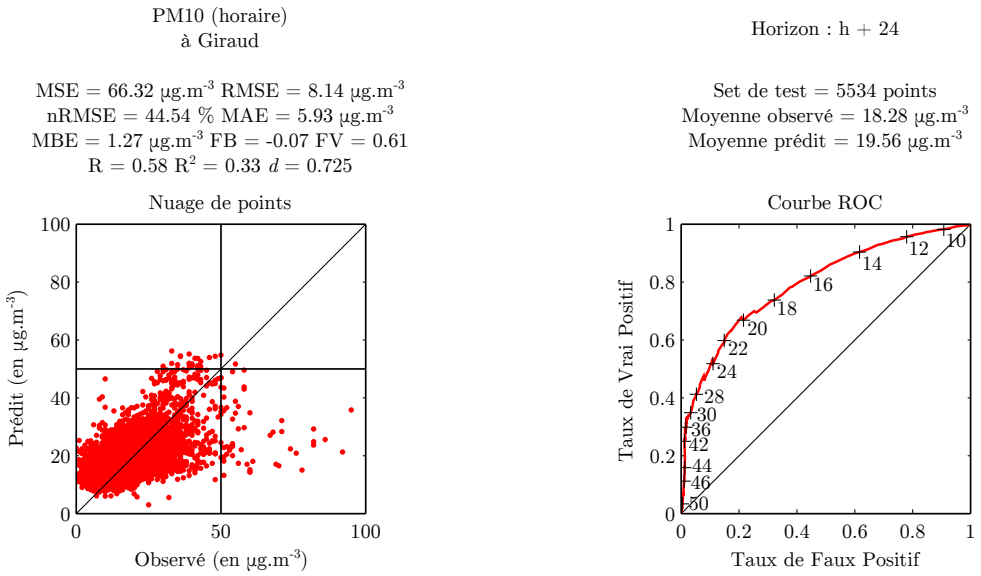


FIGURE 5.20 : Résultats du modèle prévisionnel de concentration horaire de PM10 à Giraud, construit suivant notre méthodologie. Les lignes droites sur le nuage de points indiquent le seuil d'information.

On peut voir les résultats du modèle de prévision d'ozone sur la figure 5.21. La RMSE est plus basse qu'à Ajaccio, à cause de la dynamique de l'ozone à Bastia qui, on l'a vu à la section 3.2.2, voit ses concentrations se maintenir la nuit en raison de l'instabilité de la Couche Limite Atmosphérique (CLA). L'écart des concentrations est donc resserré entre les valeurs maximales et minimales, ce qui induit une baisse de RMSE pour le modèle statistique. L'indice d'agrément d, qui prend en compte les valeurs mesurées, est quant à lui proche de celui d'Ajaccio, 0.86 alors qu'on a 0.87 à Canetto.

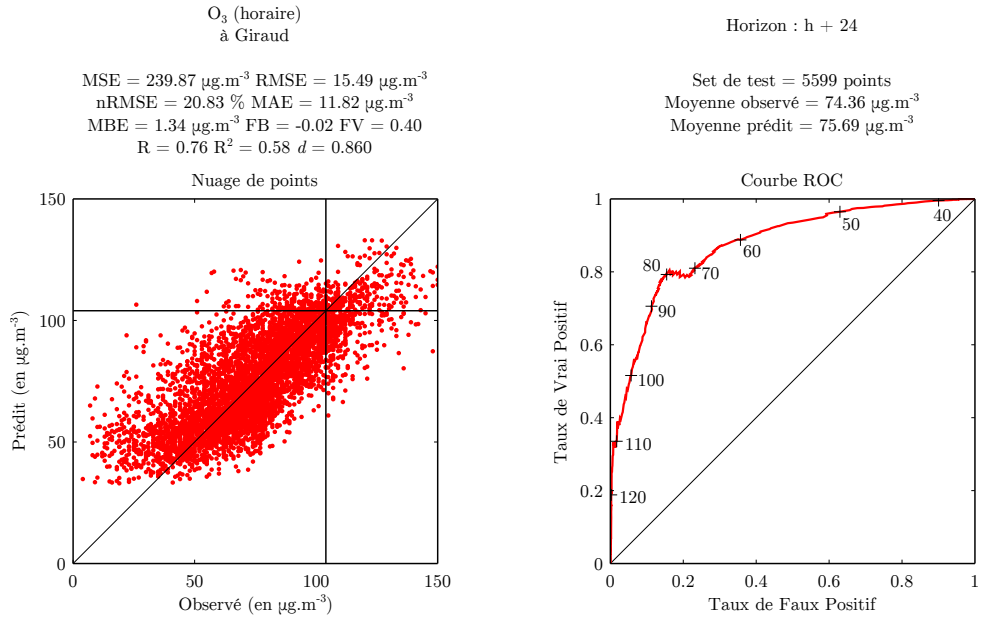


FIGURE 5.21 : Résultats du modèle prévisionnel de concentration horaire d'ozone à Giraud, construit suivant notre méthodologie.

Les séries temporelles observées et prédites des modèles bastiais sont montrées en figure 5.22 pour les PM10 et en figure 5.23 pour l'ozone. On voit pour les PM10 un pic le 30 octobre qui est raté par le modèle. C'est typiquement le genre d'épisode difficile à prévoir pour le PMC. Même s'il s'agit de concentrations horaires et que les alertes se prennent sur une base journalière pour les particules, c'est le type de comportement que l'on souhaite améliorer.

Les séries temporelles d'ozone à la figure 5.23 laissent paraitre une certaine sous-estimation des valeurs élevées. On la distingue également sur le nuage de points de la figure 5.21. Le modèle arrive à prévoir les pics d'ozone (qui restent à des valeurs en deçà du seuil d'information de 180 µg.m−3) mais en minimisant leur valeur.

Les modèles prévisionnels pour les deux polluants et les deux villes ont une précision satisfaisante. Certains épisodes de particules peuvent cependant être ratés. En effet, la rareté des épisodes de fortes concentrations ne favorise pas leur apprentissage par le PMC. De plus, les valeurs les plus élevées ont tendance à être sous-estimées. Le PMC est en effet entrainé pour prévoir les valeurs basses comme les valeurs hautes, ce qui a tendance à moyenner ses sorties.

L'usage des PMC permet tout de même d'améliorer les prévisions fournies par CHIMERE. Même sans les sorties du CTM, ils permettent une prévision du jour au lendemain efficace. On verra à la section 6 une proposition de méthode pour améliorer la détection des plus fortes valeurs.

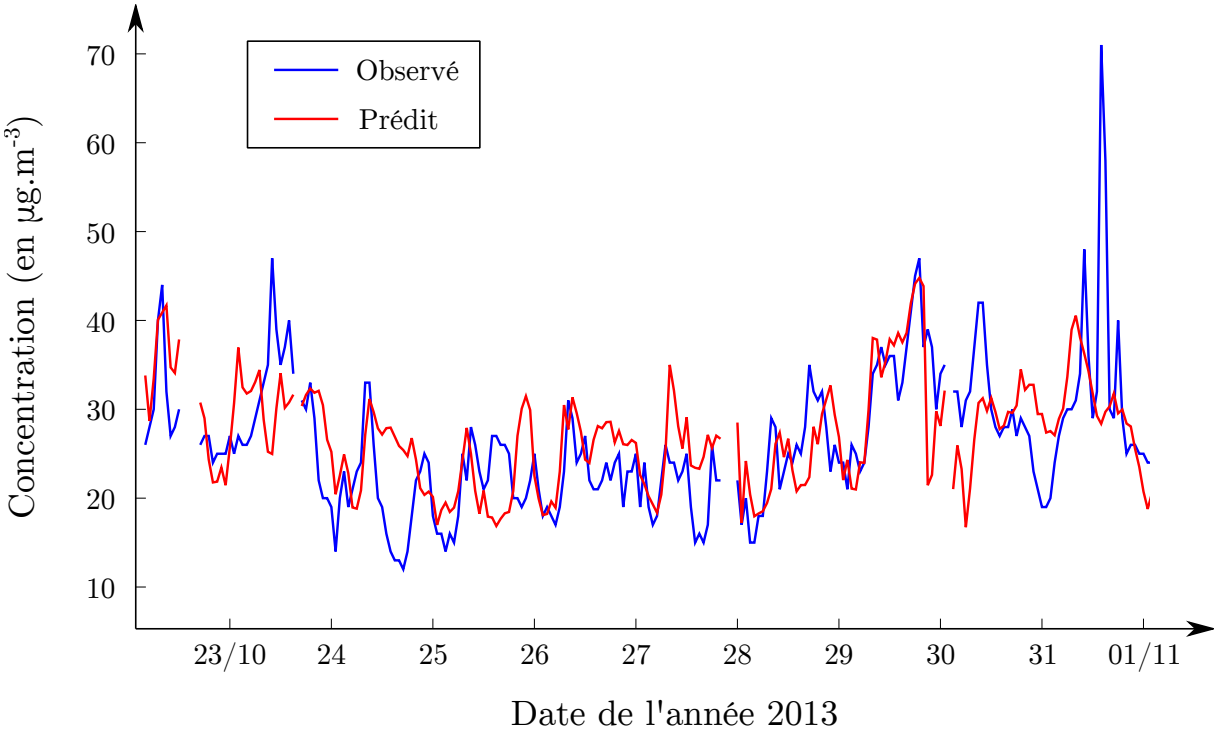


FIGURE 5.22 : Séries temporelles de concentrations horaires de PM10 à Giraud observée et prédite par le modèle prévisionnel.

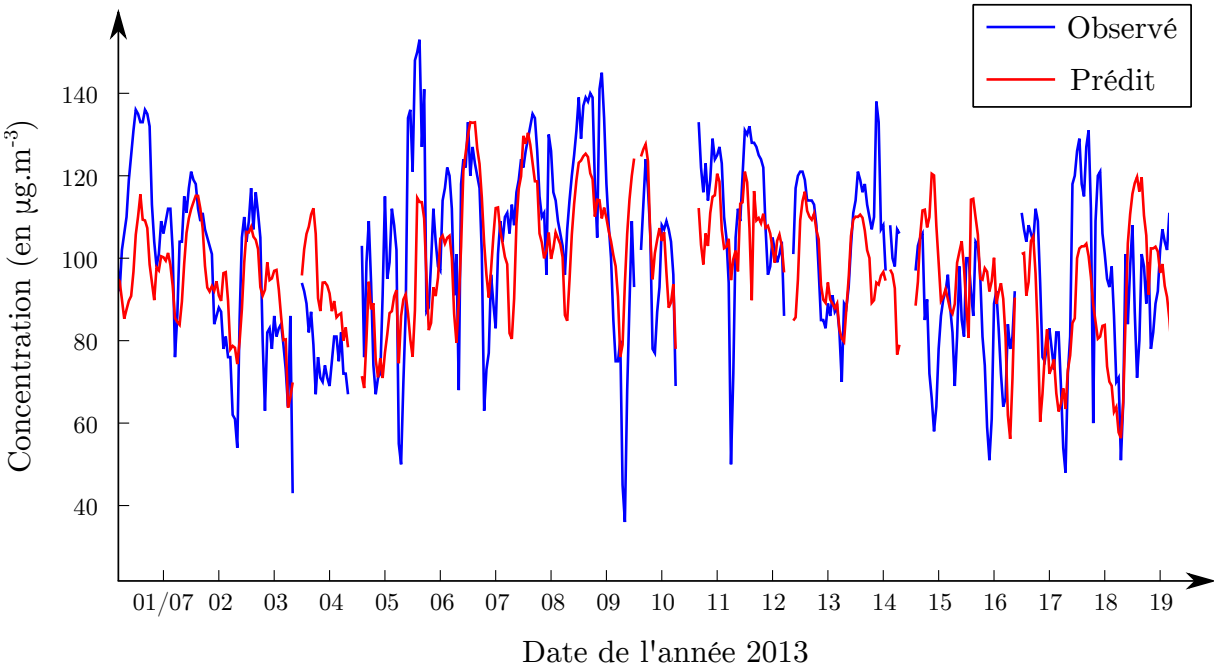


FIGURE 5.23 : Séries temporelles de concentrations horaires d'ozone à Giraud observée et prédite par le modèle prévisionnel.

5.4 Généralisation de notre approche à la région PACA

5.4.1 Prévision en PACA

Nous avons vu jusqu'à présent des exemples de modèles appliqués à la prévision en Corse, c'est-à-dire entraînés avec les données de qualité de l'air de l'île. Ces données, présentées au chapitre 3, présentent peu de dépassements de seuils. Seules les PM10 en on été responsables, et une seule fois pour le seuil d'alerte de 80 $\mu g.m^{-3}$. Cette relativement bonne qualité de l'air gêne l'évaluation du fonctionnement des modèles statistiques pour les fortes valeurs de concentration.

Afin de nous placer dans un contexte plus pollué, nous nous sommes rapprochés d'Air PACA. La qualité de l'air dans la région PACA (Provence-Alpes-Côte d'Azur) pâtit d'une industrialisation dense, notamment dans les Bouches-du-Rhône autour de l'étang de Berre où se concentre un grand nombre de sites industriels (raffinage et stockage d'hydrocarbures, sidérurgie). La région abrite plus de 4.3 millions d'habitants, dont près de trois sur quatre vivent à moins de 20 km de la mer. Marseille et Nice font partie des cinq villes les plus peuplées de France. La région est également fortement touristique, avec plus de 350000 touristes présents en moyenne sur l'année. Le climat de cette région est méditerranéen, jusqu'aux Alpes du sud dans l'arrière-pays où il devient montagnard.

Air PACA gère la plate-forme de prévision AIRES, qui repose sur les modèles WRF (Weather Research and Forecasting model, Michalakes *et al.*, 2004) et CHIMERE et fournit des cartes de prévision jusqu'à J + 2 pour la PACA, ainsi que pour le Languedoc-Roussillon et la Corse. Les conditions aux limites sont forcées à partir du modèle national Prév'air, et les inventaires des émissions des régions PACA et Languedoc-Roussillon sont utilisés. Pour la Corse où un tel inventaire sera disponible fin 2015, c'est pour l'instant l'inventaire européen EMEP (www.ceip.at) qui est utilisé. Les prévisions effectuées par Qualitair Corse utilisent les résultats de cette plate-forme, dont les cartographies sont disponibles en ligne (www.aires-mediterranee.org).

Les prévisions brutes d'AIRES sont corrigées par un modèle statistique afin d'assimiler les observations issues du parc analytique de l'AASQA (Association Agréée de Surveillance de la Qualité de l'Air), à l'image de ce qui a été présenté dans la section 5.3.1. Le modèle Forêt Aléatoire (FA, random forest en Anglais) est celui actuellement en place à Air PACA pour assurer cette assimilation. Ce modèle est utilisé grâce au logiciel R (Liaw et Wiener, 2002). Morgan Jacquinot, ingénieur à Air PACA, gère cette modélisation et a participé à cette étude. Le schéma présenté en figure 5.24 illustre le fonctionnement de la plate-forme AIRES avec ce post-traitement statistique.

5.4.2 Cadre expérimental

Nous avons voulu comparer l'usage des PMC tels que nous les utilisons en Corse avec les FA, pour trois raisons. C'est tout d'abord l'occasion de vérifier si nos modèles ont une précision comparable avec un modèle statistique différent mais déjà en place, afin de le valider. C'est également l'occasion de se confronter à des situations plus polluées qu'en Corse, mais dans un contexte géographique et climatique similaire (littoral, climat méditerranéen, montagnes alpines en arrière-pays). Enfin d'un point de vue pratique, cela permet de préparer un usage opérationnel à Qualitair Corse. En effet, si AIRES n'est pour l'instant pas précis en Corse, les prochaines améliorations du dispositif (notamment grâce à l'usage du cadastre régional des émissions) ont toutes les chances d'améliorer ses résultats. On pourra alors envisager de suivre une méthodologie proche de celle suivie par Air PACA, à savoir un post-traitement statistique des sorties d'AIRES. Les résultats présentés à la section 5.3.1 montrent que pour les PM10, les sorties d'AIRES sont

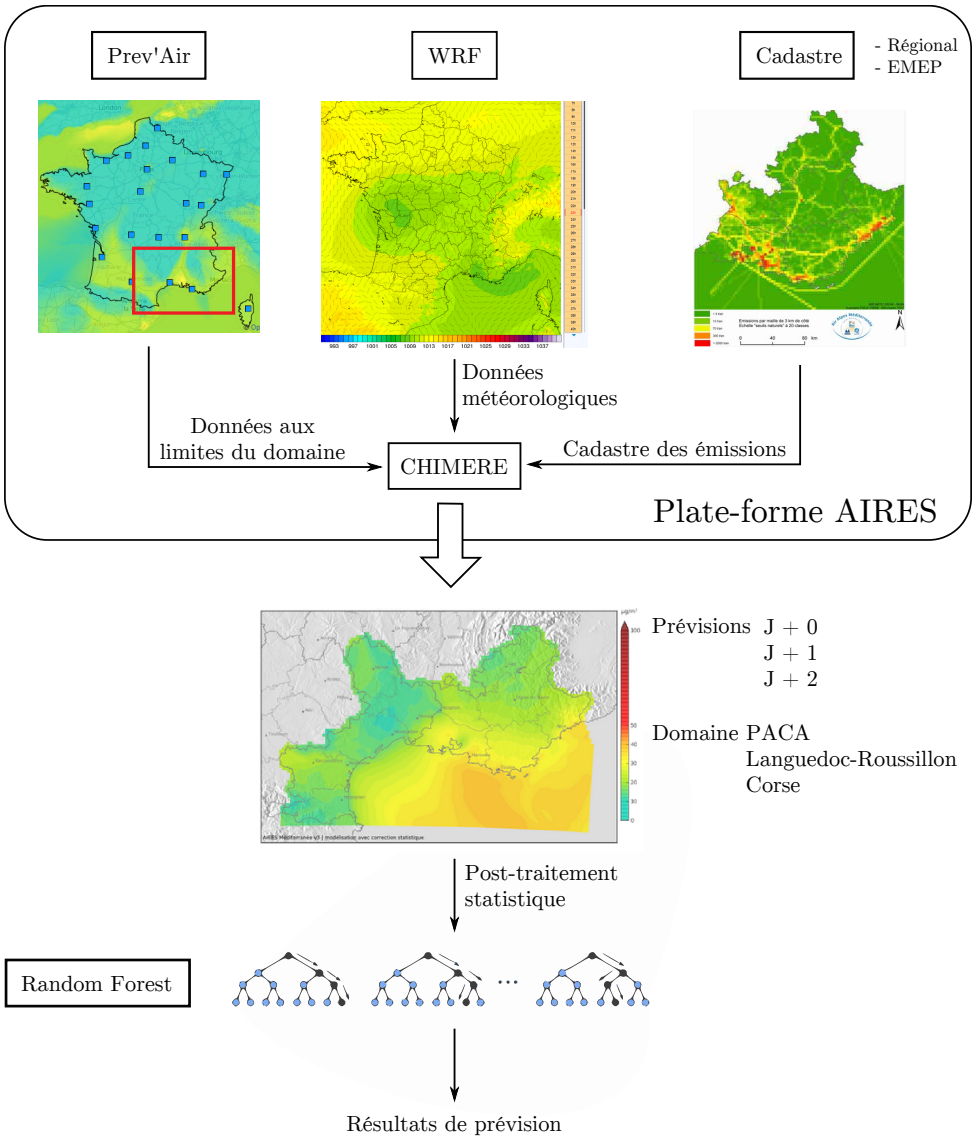


FIGURE 5.24 : Schéma du fonctionnement de la plate-forme AIRES en région PACA.

d'ores et déjà intéressantes, bien que pour l'ozone cela dépende du lieu étudié.

Les FA, introduites à la section 2.3.3, page 34, sont constituées comme leur nom le suggère de plusieurs arbres de décision. Chacun est dédié à modéliser un sous-problème, à partir d'une partie des variables d'entrée. L'ensemble des résultats est réuni pour former la prévision du modèle FA. Ce type de modèle permet d'utiliser un très grand nombre de variables d'entrée, sans les problèmes de parcimonie qui paralyseraient d'autres modèles statistiques multivariés.

De plus lors de cette étude, nous avons eu l'occasion d'utiliser un autre type de modèle neuronal : un PMC dont l'apprentissage suit un schéma à régulation bayésienne, en collaboration avec Philippe Lauret de l'université de la Réunion (laboratoire Physique et Ingénierie Mathématique pour l'Energie, l'environnemeNt et le bâtimenT (PIMENT)). Ce type de schéma d'apprentissage (Lauret *et al.*, 2008) adapte les paramètres du réseau en utilisant le théorème de Bayes. Les paramètres sont initialisés aléatoirement, puis sont adaptés itérativement en prenant en compte les probabilités d'observer les données cibles d'apprentissage, en fonction des paramètres. Cette approche probabiliste de l'apprentissage permet d'assurer les capacités de généralisation des modèles. Elle peut également identifier les variables d'entrée les plus intéressantes pour la prévision

de la cible.

Comparer FA, PMC utilisant l'algorithme LM et PMC utilisant un apprentissage bayésien nous permet d'appréhender plusieurs possibilités d'utilisation des prévisions d'AIRES. Cette étude a porté sur la prévision d'ozone et celle de PM10.

Le jeu de données que nous avons utilisé a été fourni par Air PACA. Pour les PM10, la variable à prévoir est le C90 des moyennes journalières de 23 stations fixes des Bouches-du-Rhône. Pour l'ozone, c'est le C90 du maximum des concentrations horaires que l'on veut prévoir. Pour correspondre au besoin de prévision fixé par l'arrêté ministériel contenant les instructions du gouvernement relatives au déclenchement des procédures préfectorales en cas d'épisodes de pollution (arrêté en annexe A page 196), la prévision est effectuée avec les données disponibles à 9h du matin pour le jour même (prévision à $j + 0$). Le jeu de variables utilisé par les modèles statistiques est constitué des éléments suivants :

- Les prévisions AIRES de la moyenne journalière de concentrations en PM10, ou le maximum des concentrations horaires d'ozone, prévus aux emplacements des stations du département
- Les prévisions AIRES des C90, 80, 70, 60 et 50 des concentrations en PM10 ou en ozone aux stations du département
- Les C90, 80, 70, 60 et 50 des moyennes 00h - 9h des concentrations en PM10 ou ozone mesurées aux 23 stations
- Les C90, 80, 70, 60 et 50 des moyennes journalières des concentrations en PM10 ou ozone mesurées la veille aux 23 stations
- Les prévisions des valeurs journalières moyennes, minimales et maximales de VV, HR, TC, HCL, du flux de chaleur sensible et la vitesse de frottement à chaque station
- Les valeurs journalières moyennes, minimales et maximales des variables météorologiques précédentes sur toutes les stations
- Les prévisions de direction moyenne du vent aux stations, ainsi que la moyenne pour toutes les stations

Les dernières variables de cette liste correspondant aux directions du vent sont exprimées de manière qualitative. La direction est soit exprimée en utilisant les points cardinaux (NE, SSO, etc.), soit en précisant « brise » pour indiquer un régime de brise de mer. La brise de mer prenant une direction différente en fonction du site, nous ne pouvons pas directement transformer ces variables en variables quantitatives. Ceci n'est pas gênant pour un modèle de type FA mais l'est pour un réseau de neurones. Les variables de direction de vent ont donc dû être ignorées pour les modèles neuronaux.

Un jeu de 397 variables a ainsi été créé pour la prévision de PM10. Pour l'ozone, les variables sont plus nombreuses (697), du fait d'un plus grand nombre de stations. Une telle quantité de variables d'entrée n'est pas dérangeante pour un modèle FA, mais l'est pour les réseaux neuronaux, très sensibles à la redondance d'information et à la pertinence des variables (principe de parcimonie). Nous avons donc utilisé l'ACP en prétraitement après normalisation des variables afin d'en réduire le nombre et d'en limiter la colinéarité.

Le jeu de données est constitué de 1038 points, soit 1038 jours c'est-à-dire un peu moins de trois ans de données. Afin de réaliser le jeu de test, nous avons utilisé à peu près 30 % des données, réparties sur l'ensemble de l'échantillon de manière à respecter la distribution du jeu d'apprentissage. Ce jeu est petit par rapport à ce que l'on utilise pour la prévision horaire en Corse, où l'on dispose de données parfois depuis 2006.

Avant de s'intéresser aux modèles statistiques, il convient tout d'abord d'évaluer la précision

du modèle AIRES lui-même, puisqu'il s'agit d'améliorer lesdites prévisions. La figure 5.25 montre l'évaluation des sorties brutes d'AIRES pour les PM10 sur le jeu de test, et la figure 5.26 les résultats obtenus pour la prévision d'ozone.

On peut voir sur le nuage de points que les prévisions brutes des PM10 sous-estiment largement les observations, comme c'est souvent le cas pour les CTM (McKeen *et al.*, 2007; Yu *et al.*, 2008). La forte valeur négative de la MBE traduit cette erreur systématique. La courbe ROC a une allure atypique du fait de cette sous-estimation. Le taux de faux positif est presque tout le temps nul puisque quasiment aucune fausse alerte n'est émise. Le taux de vrais positifs est très bas même pour des seuils peu élevées. Un modèle statistique en post-traitement semble adapté pour limiter l'erreur systématique, puisque une simple correction linéaire semble déjà pouvoir améliorer cette sous-estimation.

Par rapport aux résultats pour les PM10, la précision d'AIRES pour l'ozone est bien meilleure. L'indice d'agrément de 0.948 est élevée, ce qui ce traduit par un nuage de point resserré et une aire sous la courbe ROC grande. Cependant, les rares dépassements du seuil d'information (4 dépassements) ne sont pas prévus.

Ces sorties brutes font donc partie du jeu de données utilisé en entrée par trois modèles à apprentissage différent : le PMC entrainé par LM (PMC_LM), le PMC entrainé par régulation bayésienne (PMC_B) et les FA. Nous allons maintenant présenter les résultats de ces trois modèles pour ces deux polluants.

5.4.3 Principaux résultats

Les résultats des modèles PMC_LM, PMC_B et FA pour la prévision du C90 sont montrés à la figure 5.27, et les scores correspondants au tableau 5.3.

Tableau 5.3 : Scores obtenus par les modèles PMC_LM, PMC_B et FA pour la prévision du centile 90 des concentrations en PM10 des Bouches-du-Rhône.

Modèle	MBE ($\mu g.m^{-3}$)	MAE ($\mu g.m^{-3}$)	RMSE ($\mu g.m^{-3}$)	nRMSE (%)	MAPE (%)	FB	R	d
PMC_LM	0.144	5.64	7.11	19.04	16.14	-0.00	0.848	0.916
PMC_B	-0.085	5.48	7.17	19.21	16.05	0.00	0.845	0.905
FA	0.101	4.46	5.80	15.55	12.74	-0.00	0.902	0.943

Les trois modèles statistiques donnent de bons résultats. L'erreur systématique est quasiment annulée, avec des valeurs de Mean Bias Error (MBE) et de Fractional Bias (FB) petites pour les trois modèles. L'indice d'agrément d est systématiquement supérieur à 0.9. Les scores sont légèrement meilleurs pour le modèle FA, ce qui se traduit par un nuage de points plus resserrés.

Tableau 5.4 : Taux de détections par rapport au seuil d'information pour les PM10 de 50 $\mu g.m^{-3}$ obtenus par les modèles PMC_LM, PMC_B et FA dans les Bouches-du-Rhône.

Modèle	VP	VN	FP	FN
PMC_LM	38	250	9	15
PMC_B	34	256	3	19
FA	39	252	7	14

Le tableau 5.4 donne des éléments concernant la détection des journées pour lesquelles le C90 des concentrations en PM10 dépasse le seuil d'information. On remarque que les plus fortes

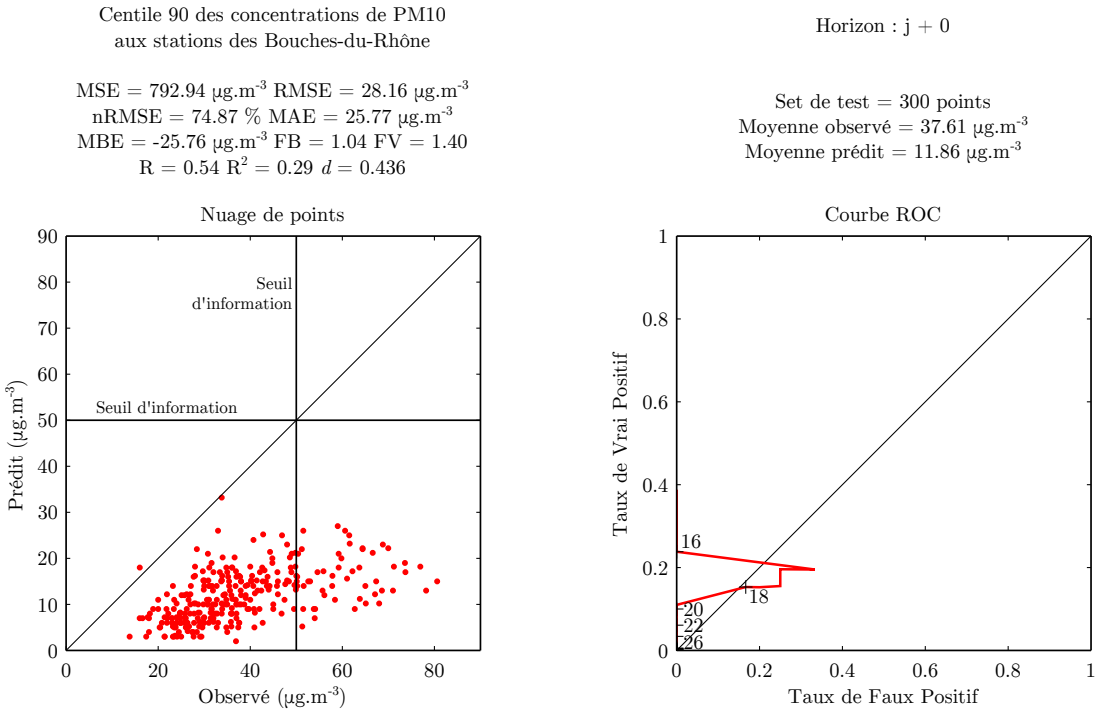


FIGURE 5.25 : Evaluation des prévisions brutes AIRES du centile 90 des PM10 dans les Bouches-du-Rhône.

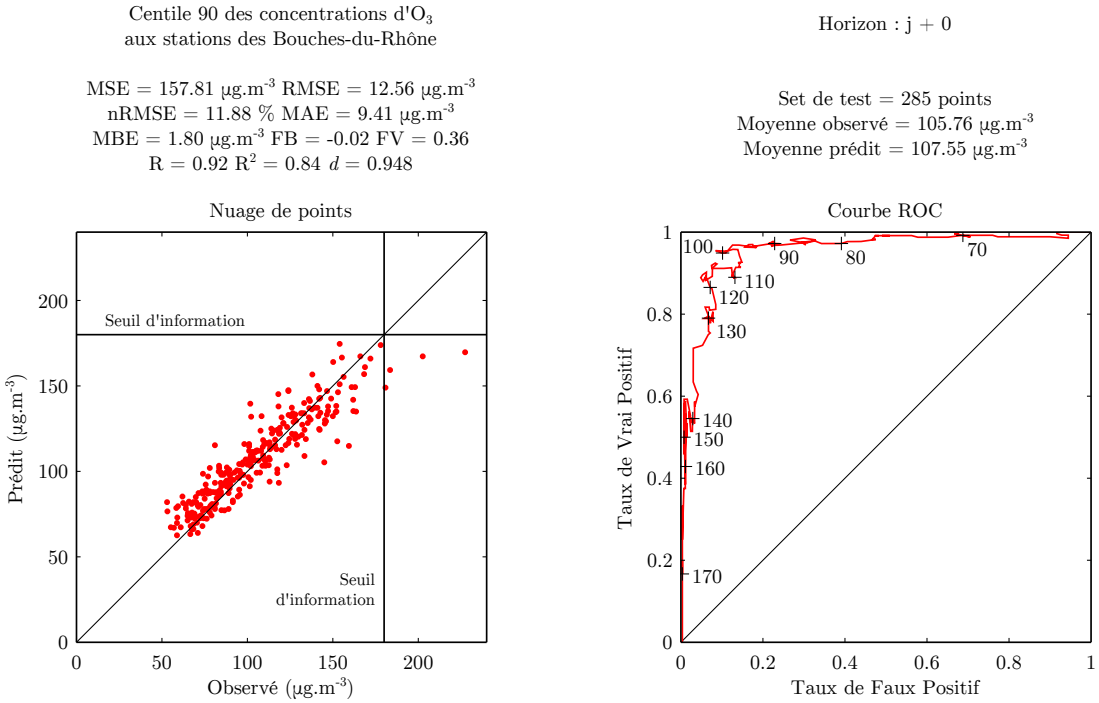
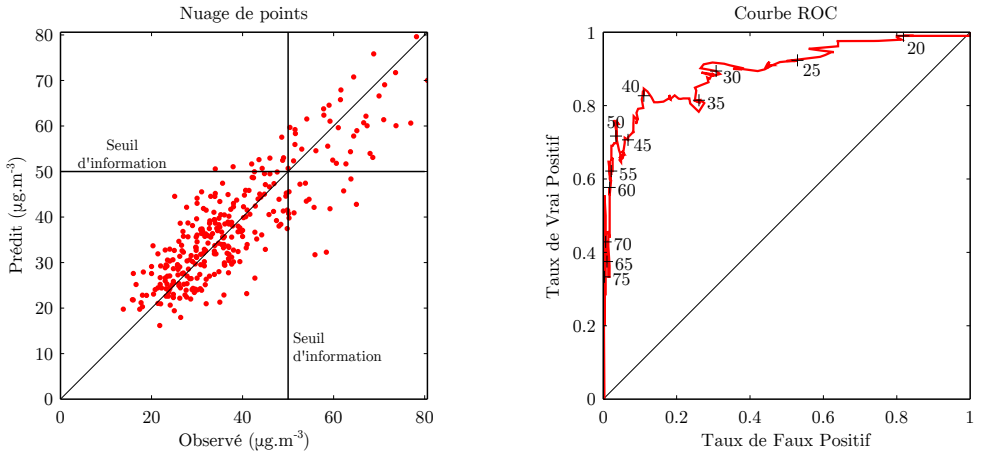
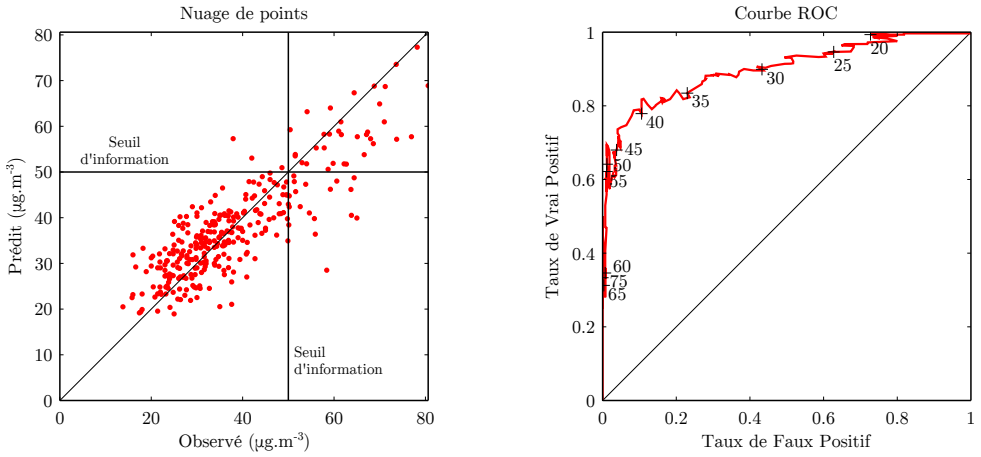


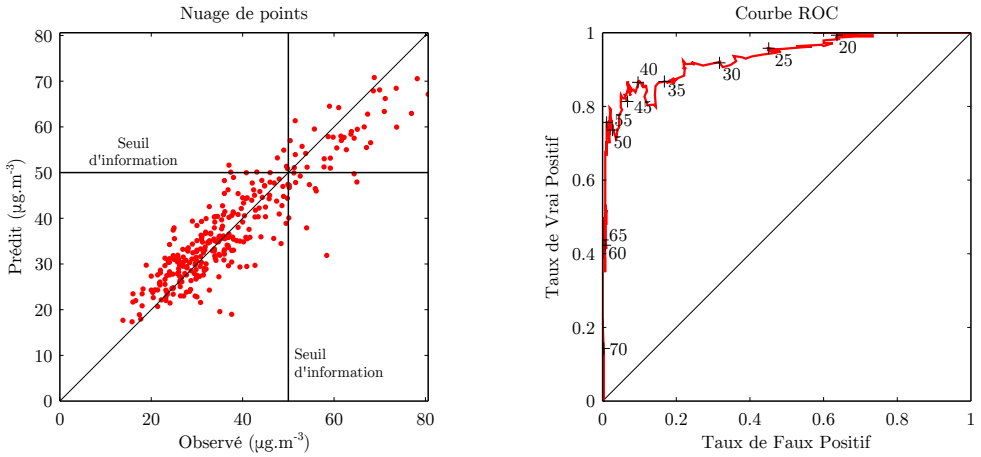
FIGURE 5.26 : Evaluation des prévisions brutes AIRES du centile 90 d'O_3 dans les Bouches-du-Rhône.



PMC_LM



PMC_B



FA

FIGURE 5.27 : Evaluation des prévisions d'AIRES du centile 90 des PM10 dans les Bouches-du-Rhône, avec post-traitement par PMC_LM, PMC_B et FA.

valeurs de PM10 sont correctement prédites. C'est le modèle FA qui bénéficie de la meilleure détection des valeurs dépassant le seuil d'information de 50 $\mu g.m^{-3}$, talonné par PMC_LM, comme le montre le nombre de vrais positifs. Par contre, les deux modèles neuronaux l'emportent sur les FA pour les seuils plus élevés, comme le montrent les Taux de Vrai Positif (TVP) pour 70 et 75 $\mu g.m^{-3}$ sur les courbes ROC. On peut également voir cette différence sur les nuages de points.

Intéressons nous à présent aux résultats obtenus pour la prévision d'ozone dans les Bouches-du-Rhône. Les alertes sont déclenchées à partir des concentrations horaires de ce polluant. Nous avons donc utilisé les prévisions du maximum journalier des valeurs horaires d'ozone.

Le tableau 5.5 montre les scores obtenus par les modèles statistiques. On remarque qu'encore une fois les scores sont assez proches, même si ceux du PMC_LM l'emportent. Les trois modèles améliorent les performances d'AIRES.

Tableau 5.5 : Scores obtenus par les modèles PMC_LM, PMC_B et FA pour la prévision du centile 90 des concentrations en ozone des Bouches-du-Rhône.

Modèle	MBE ($\mu g.m^{-3}$)	MAE ($\mu g.m^{-3}$)	RMSE ($\mu g.m^{-3}$)	nRMSE (%)	MAPE (%)	FB	R	d
PMC_LM	0.36	7.94	11.16	10.56	7.60	-0.00	0.933	0.965
PMC_B	-0.73	8.09	11.40	10.78	7.81	0.01	0.933	0.958
FA	0.77	8.53	12.04	11.38	7.99	-0.01	0.924	0.960

Le seuil d'information de 180 $\mu g.m^{-3}$ (tableau 5.6) n'a été dépassé que quatre fois dans le jeu de test, un nombre insuffisant pour juger les modèles sur la prévision des dépassements. Cependant, le modèle FA a réussi à prévoir l'un de ces dépassements.

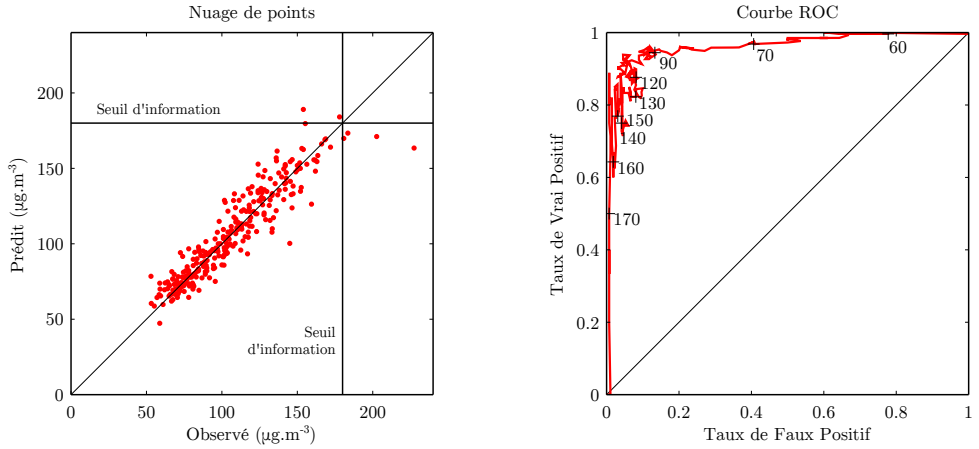
Tableau 5.6 : Taux de détections par rapport au seuil d'information pour l'ozone de 180 $\mu g.m^{-3}$ obtenus par les modèles PMC_LM, PMC_B et FA dans les Bouches-du-Rhône.

Modèle	VP	VN	FP	FN
PMC_LM	0	279	2	4
PMC_B	0	281	0	4
FA	1	278	3	3

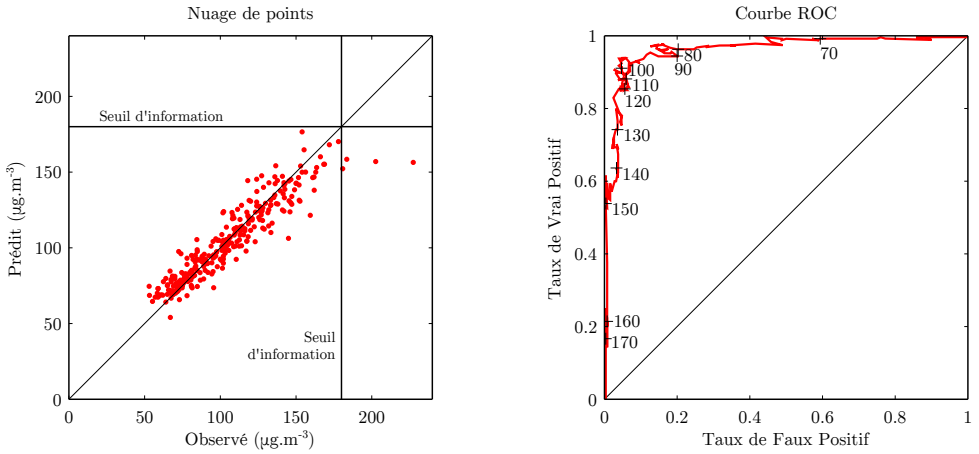
Ces résultats permettent de valider l'approche utilisant PMC_LM. Ce modèle peut s'adapter efficacement à un jeu de données comprenant un grand nombre de variables, plus adapté aux FA, et l'emporte légèrement au niveau des scores sur le PMC_B. Les résultats appliqués à la prévision de particules en utilisant un grand nombre d'observations et de sorties de modèle sont similaires aux autres méthodes, y compris FA actuellement utilisée en PACA. Cela permet d'envisager ce type de modèle opérationnel en Corse, quand AIRES bénéficiera du cadastre régional en cours de finalisation à Qualitair Corse.

5.4.4 Bilan de l'étude

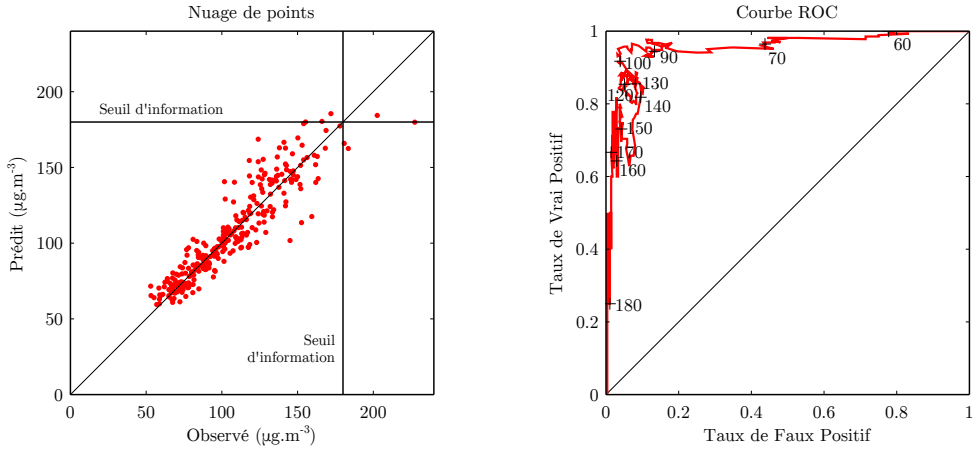
Trois modèles à apprentissage ont été utilisés pour réaliser des prévisions à $j+0$ de PM10 et d'ozone dans les Bouches-du-Rhône. Leurs données d'entrée sont les sorties du modèle AIRES ainsi que du modèle WRF pour la météorologie, en plus des mesures disponibles à 9h, heure à laquelle doit être réalisée la prévision.



PMC_LM



PMC_B



FA

FIGURE 5.28 : Evaluation des prévisions d'AIRES du centile 90 d'ozone dans les Bouches-du-Rhône, avec post-traitement par PMC_LM, PMC_B et FA.

Les modèles utilisés correspondaient au PMC tel que nous l'employons avec les données corses, un PMC à apprentissage bayésien ainsi qu'un modèle FA, actuellement utilisé à Air PACA.

Les résultats montrent que les trois méthodes permettent d'améliorer sensiblement les prévisions d'AIRES, notamment pour la prévision de particules, aux valeurs largement sous-estimées par le modèle déterministe. Des trois modèles, c'est le PMC à apprentissage par l'algorithme LM qui obtient les meilleurs résultats pour tous les scores pour la prévision d'ozone, mais de manière légère. Ce sont les FA qui l'emportent pour les PM10 pour tous les scores, cette fois de manière légèrement plus prononcée. Les FA détectent légèrement plus de dépassement du seuil d'information des deux polluants. Cependant, la détection de seuils de PM10 plus élevés que le seuil d'information à 50 $\mu g.m^{-3}$ (autour de 70 - 75 $\mu g.m^{-3}$) est meilleure avec les PMC.

Le jeu de données est à la base adapté à un usage pour les FA. Certaines variables qualitatives (direction du vent) ont du être exclues du jeu de données pour les PMC, et l'ACP a été nécessaire pour en réduire la taille. Les résultats des trois modèles sont en tout cas satisfaisants, et la possibilité d'utiliser des PMC en post-traitement d'AIRES en Corse est confirmée, après les résultats allant dans ce sens, notamment pour les PM10, vus à la section 5.3.1.

5.5 Conclusion

Nous avons présenté dans ce chapitre des méthodes permettant l'amélioration des résultats obtenus par les PMC pour la prévision de la qualité de l'air. Visant à se conforter au principe de parcimonie, la sélection de variables et l'élagage des RNA ont tout deux été étudiés.

La sélection de variables est un domaine large et plusieurs méthodes existent pour la mener. Nous avons exploré en particulier quatre approches : la première est l'usage par l'utilisateur de l'Information Mutuelle (IM) entre les variables comme critère de sélection. La seconde est l'utilisation d'algorithmes génétiques utilisant une fonction objectif également basée sur l'IM. La troisième est l'utilisation du recuit simulé et la dernière correspond à l'usage de l'ACP pour se défausser des composantes principales de moindre valeur propre.

L'IM, notion issue de la théorie de l'information, permet de quantifier la quantité d'information partagée par deux séries temporelles. Elle est donc particulièrement utile quand on cherche à conserver les variables les plus pertinentes par rapport à une variable cible. Son usage permet effectivement d'optimiser les jeux de données d'entrée, qu'elle soit calculée par l'utilisateur pour chaque variable ou qu'elle soit à la base de la fonction objectif d'une métaheuristique utilisant l'approche « filter ». Son principal défaut cependant est son temps de calcul, qui nous empêche de calculer en un temps raisonnable l'IM d'un jeu de variable d'entrée complet avec la cible, bien que le calcul soit théoriquement possible, et ce quelle que soit la méthode d'estimation de l'IM envisagée. Les fonctions objectifs de substitution utilisant l'IM n'ont pas permis un usage optimal des AG pour la sélection de variables.

Les recuits simulés ont donnés des résultats plus intéressants. Nous les avons utilisés *via* l'approche « wrapper », basée sur l'évaluation des modèles prévisionnels construits. Cette approche a permis d'obtenir des résultats satisfaisants, et nous laisse des perspectives intéressantes. On peu envisager d'utiliser des recuits simulés en « wrapper » dans le but d'assurer, en plus du choix des variables, l'ensemble des choix de configuration évoqués en section 4.5. Une telle automatisation rendrait l'usage des PMC complètement indépendant des choix de l'utilisateur. Cependant, une étude poussée visant à diminuer les temps de calculs des recuits ou un matériel informatique avec plus de ressources nous serait alors nécessaire pour faire face à la dimension

du problème.

L'ACP a finalement donné des résultats très intéressants, comparée aux autres méthodes. Cette dernière étant déjà utilisée en tant que prétraitement, son usage est plus simple que celui du recuit simulé et sera conservé comme méthode de sélection de variables pour nos travaux futurs. En tant que perspective, nous prévoyons une étude de sélection de variables mais aussi de configuration automatique à l'aide des recuits simulés. Cela permettrait de fixer l'ensemble des points de configuration à l'aide de la métaheuristique.

Une méthode d'élagage permettant de se conformer au principe de parcimonie a également été présentée. Elle est basée sur l'usage de l'algorithme de LM pour identifier les paramètres du modèle qui tendent à être non-significativement différents de zéro, afin de les supprimer et d'alléger le PMC. Bien qu'efficace pour la prévision d'autres variables, cette méthodologie a cependant été abandonnée dans le cadre de la prévision de la qualité de l'air, pour laquelle ses résultats ne sont pas probants.

Ce chapitre nous a permis d'ajouter une étape de sélection de variable à notre méthodologie de construction de modèle. Le modèle classique sera construit en exécutant les opérations suivantes, dans l'ordre :

- Création du jeu de données initial
- Projection des variables circulaires
- Désaisonnalisation des variables à composante périodique
- Suppression ou remplacement des points comportant des données manquantes
- Calcul des composantes principales et réduction du nombre de variables d'entrée
- Initialisation des paramètres par algorithme de Nguyen-Widrow
- Apprentissage par algorithme de Levenberg-Marquardt
- Prévision utilisant le modèle entrainé avec les données du jeu de test
- Post-traitement pour rétablir la sortie de modèle après normalisation et stationnarisation
- Evaluation

Des modèles ainsi construits ont été utilisés dans deux contextes différents, en Corse et en PACA. En premier lieu nous avons appliqué notre méthodologie à la construction de modèles opérationnels en Corse, permettant de prédire les niveaux de pollution en Ozone et PM10 à horizon $h + 24$ à Ajaccio et à Bastia. L'usage de variables issues de la plate-forme AIRES s'est avéré bénéfique, mais pas dans toutes les situations.

Dans un deuxième temps, nous avons comparé nos modèles avec ceux actuellement en fonction chez Air PACA, les FA qui utilisent en entrée un grand nombre de variables issues de l'ensemble du parc analytique de l'AASQA, et des sorties brutes de la plate-forme prédictive AIRES. Les résultats on montré la portabilité des PMC dans ce contexte. Cette adaptabilité permet d'envisager l'ajout de nombreuses séries temporelles exogènes aux entrées des PMC, quand elles deviendront suffisamment longues (Venaco, Cap Corse, etc.) pour adopter en Corse une prévision à une échelle plus large que la ville.

Nous avons construit jusqu'ici des modèles optimisés pour la prévision dans sa globalité. Aucune distinction n'est faite entre les différentes conditions météorologiques qu'on peut rencontrer dans les données, les résultats sont donc meilleurs pour les situations les plus fréquentes dans le jeu d'apprentissage. Les pics de pollution, qui sont rares, ne sont pas spécialement favorisés. Leurs valeurs étant les plus extrêmes dans les jeux de données de concentrations, les PMC ont du mal à les prévoir.

La prévision de séries temporelles n'a pas forcément pour but de prévoir correctement certaines situations typiques plus que d'autres. C'est cependant le cas de la prévision de la qualité

de l'air, où les situations de fortes concentrations sont les plus préoccupantes. C'est dans ces conditions que l'impact des polluants atmosphériques sur la santé des populations est maximal, et que des procédures parfois lourdes sont déclenchées pour limiter les émissions.

Pour cette raison, nous allons au prochain chapitre nous focaliser sur la partie de nos travaux qui touche plus particulièrement la prévision des valeurs extrêmes, correspondant aux pics de pollution. A partir des méthodes présentées dans ce chapitre, nous proposerons une méthode d'hybridation de modèles dont le but sera de favoriser la détection des pics de pollution par rapport aux situations classiques.

Chapitre 6

Proposition de modèles hybrides pour la détection de pics de pollution

La prévision de la qualité de l'air peut être menée grâce à des modèles statistiques, et nous avons développé une méthodologie au chapitre 4 permettant de construire les PMC et d'obtenir de bons résultats de prévision. Nous avons cependant pu constater deux problèmes propres à la modélisation statistique. Tout d'abord la sous-estimation des valeurs extrêmes, liée au fait que les paramètres des modèles sont fixés pendant l'apprentissage pour améliorer les prévisions dans toutes les situations. Ceci a tendance à moyenner les résultats, surestimant les valeurs les plus faibles, mais surtout sous-estimant les valeurs les plus fortes. Mais à cela s'ajoute la mauvaise représentation des phénomènes rares dans les jeux de données d'apprentissage, comme le sont les pics de pollution en Corse. Les relations entre les variables propres à ces phénomènes pèsent moins lourd dans l'apprentissage que les relations liant ces mêmes variables dans des conditions rencontrées plus fréquemment.

Pour nous adapter à ces difficultés, nous avons développé une méthodologie en lien avec ces variables. Il s'agit de coupler un modèle de classification avec différents PMC prédictifs. La tâche du classifieur est de diviser les données pour isoler les situations qui apparaissent comme spécifiques. Pour chacune de ces situations, un PMC dédié est alors entrainé, qui sera spécialisé dans la prévision pour le régime atmosphérique correspondant. On qualifie les modèles ainsi formés (plusieurs PMC sélectionnés après une étape de classification) de « modèles hybrides ». Un tel modèle est illustré à la figure 6.1.

On a vu à la section 1.2 (page 10) certains des régimes atmosphériques particuliers qui influencent la qualité de l'air. Ces régimes ont un impact majeur sur les concentrations des polluants, vu que leur dynamique dépend des conditions atmosphériques et pas uniquement des émissions. Quand on veut connaître les concentrations en polluants, savoir dans quel régime atmosphérique on se trouve simplifie le problème. On sait par exemple qu'en situation d'inversion thermique on a des chances d'observer des concentrations élevées en particules, ou que de nuit en plaine les concentrations d'O_3 vont être basses. Un PMC qui serait entrainé sur un régime particulier pourrait en tirer un gain de précision.

Plusieurs possibilités existent pour classifier les variables. Tout d'abord, on peut le faire soi-même. Sur la base de connaissances propres à la dynamique des polluants, des règles de séparation peuvent être définies pour isoler les situations connues pour être propices aux pics. Cette approche est cependant biaisée par les aprioris de l'utilisateur, et nous la comparerons avec des méthodes moins dépendantes de nos préconceptions.

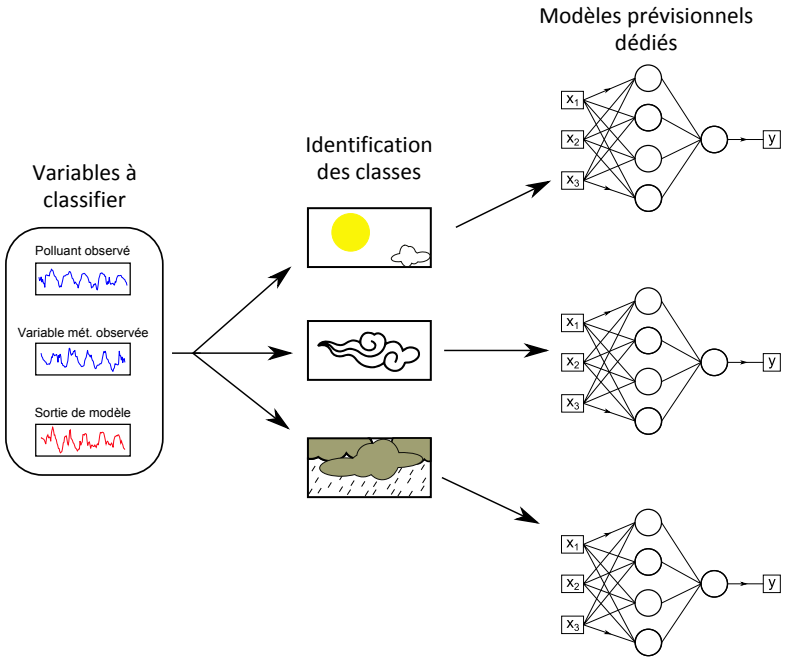


FIGURE 6.1 : Illustration de l'usage d'un modèle hybride.

Les modèles de classification non-supervisée (modèles de clustering) regroupent les données en fonctions de critères de dissimilarité. Cette hybridation entre clustering et régression a été utilisée en prévision de séries temporelles (Fan *et al.*, 2009; Goia *et al.*, 2010; Cherif *et al.*, 2013), mais rarement pour la qualité de l'air. Pour la pollution atmosphérique certains travaux ont envisagé ce type d'hybridation (Kolehmainen *et al.*, 2000; Lu *et al.*, 2006; Poggi et Portier, 2011) mais avec des modèles prédictifs différents des nôtres ou appliqués à d'autres types de prévisions.

Le modèle hybride que nous voulons construire est composé du modèle de classification et de l'ensemble des PMC entrainés, qu'on appelle parfois des modèles « locaux » (Cherif *et al.*, 2013). Pour réaliser une prévision grâce à ce modèle hybride ou pour l'évaluer, il faut procéder en deux temps. D'abord prévoir le régime, puis utiliser le PMC local correspondant pour prévoir la concentration. Pour pouvoir classifier le régime atmosphérique de manière prospective, nous utiliserons des sorties du modèle AROME (Seity *et al.*, 2011), fournissant des prévisions météorologiques jusqu'à l'horizon $h + 30$.

Pour mener ces études, l'application Aria Base que nous avons développée (voir section 7.1, page 7.1) a été d'une grande aide. Elle a permis de diriger la division des jeux de données, mais surtout d'automatiser les apprentissages des modèles prédictifs sur les sous-échantillons. L'évaluation des modèles hybrides a également été fortement facilitée.

Nous utiliserons principalement les courbes ROC présentées à la section 2.4.3 page 44 pour évaluer les capacités de détection de pics de pollution des modèles. L'approche consistant à utiliser des règles définies par l'expérimentateur sera abordée à la section 6.1. On présentera ensuite les résultats obtenus avec plusieurs méthodes de classification non-supervisées, à la section 6.2, avant de conclure sur l'intérêt des modèles hybrides pour la prévision de pics de pollution.

6.1 Division par l'utilisateur

L'occurrence de pics de pollution est due à plusieurs facteurs ; des facteurs d'émission, de transport, de dispersion. Certaines situations météorologiques sont connues pour favoriser l'apparition de pics. Nous allons donc expérimenter la division de données basée sur ce type de connaissances pour former les modèles hybrides.

Nous avons vu aux chapitres 1 et 3 comment agissent les phénomènes tels que l'inversion thermique ou les brises sur les niveaux de polluants en Corse. Nous avons également vu que la date et l'heure, qu'on peut fournir aux PMC sous forme d'indices temporels, ont un lien indirect avec les concentrations. Enfin, nous avons vu que certains vents dominants sont propices à des épisodes de transport. Nous allons tenter d'identifier la meilleure manière de diviser les données dont nous disposons afin d'isoler les situations propices aux pics de pollution à partir des variables décrivant ces phénomènes. Cette division permet de créer des sous-échantillons, chacun servant de données d'apprentissage pour un PMC spécialisé.

Cette division des données se fera en suivant des règles définies par l'utilisateur, permettant de classifier un point de donnée dans le groupe lui correspondant. Par exemple, imaginons la règle suivante : si nous sommes l'hiver, classer le point dans la classe 1. Si nous somme l'été, vérifier s'il pleut. S'il oui classer le point dans la classe 2, sinon dans la classe 3.

Cette démarche peut se représenter par un graphique sous forme d'arbre (voir figure 6.2). C'est la représentation utilisée par les arbres de décision, mais il ne s'agit pas ici de ce type de modèle puisque c'est l'utilisateur qui définit les règles de division. Chaque embranchement de l'arbre correspond à une division du jeu de données en fonction d'un critère choisi. Chaque feuille finale de l'arbre correspond à une classe, et à un sous-groupe de l'échantillon.

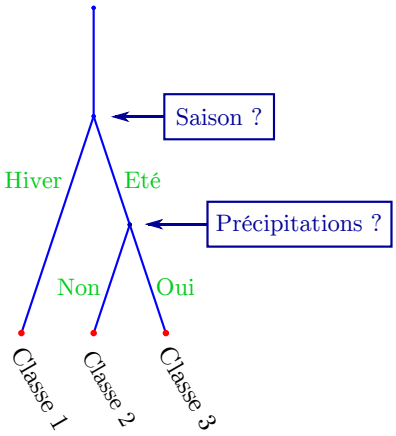


FIGURE 6.2 : Illustration de règles de séparation des données établies par l'utilisateur.

Nous avons utilisé notre application Aria Base afin de générer ces sous-groupes facilement et de gérer les modèles qui leur sont liés (voir section 7.1.2 page 170). Nous avons mené un certain nombre d'essais, notamment en séparant les données en fonction de la saison ou de l'heure de la journée, de la direction du vent au sol ou à 800hPa, et de la stabilité de la couche limite, représentée par HCL ou ECI. On a ainsi voulu isoler des situations propices aux pics de pollution comme les heures de pointes, les situations de stabilité de la couche limite, les vents susceptibles de favoriser le transport de masses d'air depuis le sud de la France ou le Sahara, etc. Nous nous sommes particulièrement focalisés sur l'O_3 et les PM10 à la station de Canetto, qui bénéficie des plus longues séries temporelles.

Les résultats que nous avons obtenus avec cette méthodologie sont mitigés. La précision en

termes de RMSE ou de d est souvent dégradée, pour une capacité de détection des pics équivalente ou moindre. La division en sous-échantillons semble plus pénaliser les modèles prédictifs que les aider à se spécialiser sur les phénomènes en liens avec les fortes concentrations. Ces résultats sont présentés à l'annexe B (page 203).

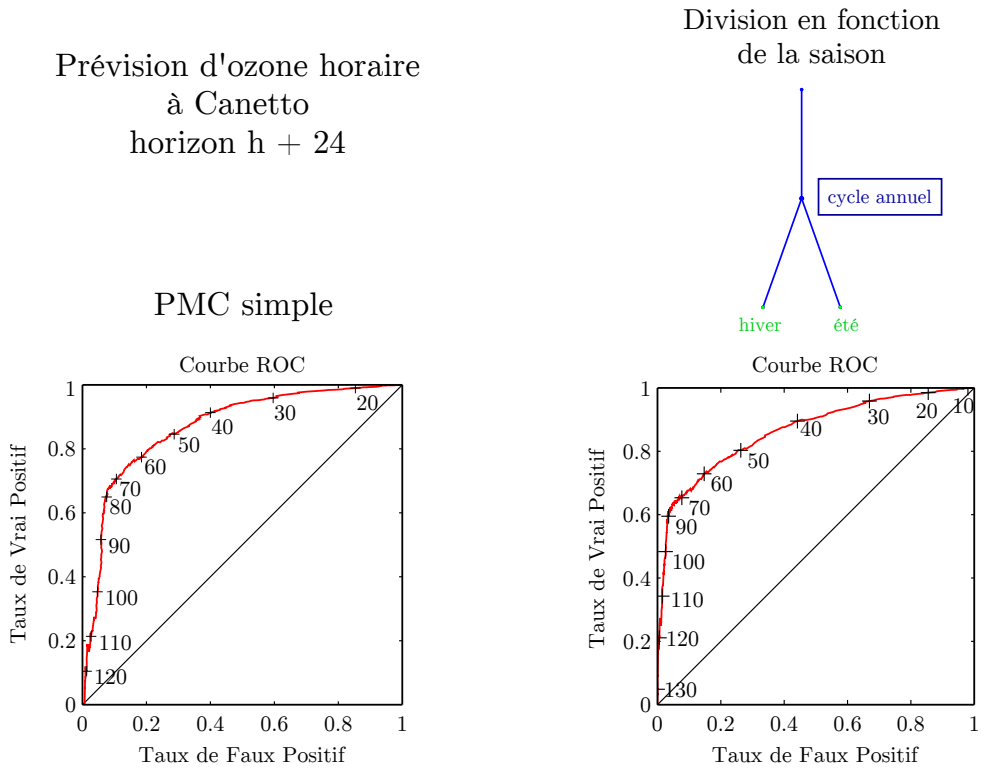


FIGURE 6.3 : Prévision d'ozone avec un PMC simple et un modèle hybride avec division des données suivant la saison, été ou hiver.

On peut tout de même citer un exemple qui fonctionne bien, dont les résultats sont montrés en figure 6.3. Il s'agit de la séparation des données en fonction de la saison (été ou hiver) pour la prévision d'ozone. Ce résultat n'est pas surprenant, tant les différences de concentrations entre l'été et l'hiver pour ce polluant sont importantes, la production photochimique d'ozone étant fortement liée à l'ensoleillement. Certaines études sur la prévision statistique de ce polluant n'utilisent d'ailleurs que les données estivales pour cette raison.

Les résultats obtenus par division « utilisateur » laissent penser qu'il est possible d'améliorer la détection des pics avec des modèles hybrides. Après les résultats obtenus pour lesquels seul l'ozone a pu bénéficier d'une augmentation des taux de détection, nous avons expérimenté l'hybridation des PMC avec des modèles de clustering.

6.2 Clustering

Nous avons appliqué des méthodes de classification non-supervisées à la division de nos jeux de données. L'aspect non-supervisé de l'apprentissage permet de s'assurer de l'absence de biais dus à nos aprioris sur le fonctionnement atmosphérique. Un tel apprentissage est basé non pas sur des règles des utilisateurs portant sur les données elles-mêmes, mais sur des règles de division ou de regroupement, utilisant des notions de distances entre les points de données. Nous avons appliqué ces méthodes sur des sorties de modèle AROME, afin de tenter de capter les régimes

météorologiques favorisant les pics de pollution.

Afin de déterminer les différentes classes représentant les différents régimes météorologiques, nous avons utilisé des modèles de partitionnement, ou clustering. Ce sont des modèles qui séparent les données en différents groupes par apprentissage non-supervisé. Le modèle prévisionnel résultant est donc un modèle hybride, constitué d'un modèle de partitionnement et de différents modèles prédictifs. Il existe plusieurs types de modèles de partitionnement.

L'usage de clustering à l'aide de Self-Organizing Map (SOM, cartes auto-organisatrices en français, également appelées cartes de Kohonen) préalable à l'usage d'un RNA prédictif est une pratique qu'on retrouve dans la prévision de série temporelle, au delà de la qualité de l'air (Cherif *et al.*, 2011; Beccali *et al.*, 2004; Sánchez-Maroño *et al.*, 2003; Walter *et al.*, 1990).

Pour la prévision de la qualité de l'air, les SOM ont été utilisées pour assurer une classification non-supervisée préalable à l'utilisation de PMC dès les années 2000 (Kolehmainen *et al.*, 2000). Après le partitionnement, un PMC par classe est entrainé et la combinaison de leurs résultats est utilisée comme prévision. Les auteurs ont constaté la difficulté de prévoir les particules, plus importante que pour la prévision des gaz, mais n'ont pas comparé leurs modèles hybrides avec un modèle de référence. L'avantage de l'usage préalable des SOM par rapport à un simple PMC n'est donc pas évalué.

Le seul autre exemple d'hybridation entre clustering et RNA pour la prévision de la qualité de l'air est à notre connaissance celle de Lu *et al.* (2006). Au delà de l'usage de RNA, on retrouve également d'autres modèles régressifs de prévision de la qualité de l'air hybridés avec des modèles de clustering. L'étude de Poggi et Portier (2011) utilise des modèles de régression linéaire entrainés sur des données séparées par clustering pour prévoir les concentrations journalières de PM10 un jour à l'avance. L'appartenance des données à des classes est obtenue grâce à un modèle de mélanges gaussiens (Grün et Leisch, 2007).

Les auteurs ont par ailleurs expérimenté deux méthodes d'assignation des points à une classe. La méthode « dure » consistant à assigner les points à la classe leur correspondant le mieux, et une méthode « floue » où pour chaque point, une appartenance à chaque classe est calculée, fonction de sa distance au centre de cette classe. Les prévisions sont réalisées avec les modèles prédictifs correspondant à toutes les classes, et le résultat est obtenu en pondérant les sorties de chaque modèle par l'appartenance à la classe correspondante. Plus lourde, cette méthodologie qu'on peut imaginer plus appropriée puisqu'elle évite une assignation à un unique cluster s'avère cependant similaire au niveau des scores.

L'approche de clustering de Lu *et al.* (2006) pour entrainer ensuite les PMC est basée sur l'utilisation de SOM en combinaison à l'algorithme de k-means qui permet la diminution de dimension du jeu de données. Les résultats de ces approches sont prometteurs puisque des améliorations de précision sont rapportées. Cependant cette étude est limitée à la prévision d'un seul polluant, la concentration horaire maximale d'ozone pour le jour même. Nous avons repris ce type d'expérimentation, en nous focalisant sur trois polluants : l'ozone, les PM10 et le NO_2.

Nous avons utilisé deux approches différentes pour notre problématique. Nous avons tout d'abord expérimenté la méthode de partitionnement mixte SOM/k-means de Lu *et al.* (2006), que nous développerons dans la section suivante. Nous avons également expérimenté une autre approche de partitionnement, la Classification Ascendante Hiérarchique (CAH), décrite à la section 6.2.2, page 155. Une fois ces deux méthodes présentées, nous verrons quels sont les résultats obtenus en matière de prévision d'évènements de dépassement.

6.2.1 Self-Organizing Map et k-means

Les SOM ont été introduits par Kohonen (1982), d'où leur désignation alternative de « cartes de Kohonen ». Ce sont des modèles à apprentissage non-supervisé, souvent considérés comme faisant partie de la famille des réseaux de neurones. Les cartes de Kohonen sont constituées d'une couche de neurones, qu'on peut en général représenter disposés sur une grille en deux dimensions et interconnectés à leurs voisins. Ces neurones n'ont pas de « poids » ni de « biais », comme on nomme les paramètres des neurones de PMC. Chaque neurone d'une SOM possède un paramètre, qu'on appelle sa « position ». La grille formée par la carte est paramétrée par la position de ses neurones, dans l'espace des variables d'entrées.

Lors de l'apprentissage, la position des neurones va changer. Ils vont se « déplacer » pour recouvrir au mieux l'espace formé par les points de données d'apprentissage (voir l'illustration à la figure 6.4). L'apprentissage se fait généralement en utilisant l'algorithme de Kohonen. Il s'agit d'apprentissage compétitif. Chaque point du jeu de données d'apprentissage est fourni en entrée de la SOM. Dans l'espace des variables, ce point correspond à une position définie par la valeur de chacune des variables. Le neurone dont la position est la plus proche du point gagne la compétition. Il est « rapproché » de la position du point, et ses neurones voisins sont dans une moindre mesure entrainés également. Quand tous les points de données ont été utilisés pour l'apprentissage, les positions des neurones sont optimisées. Ceci correspond à un apprentissage point par point, ce qu'on appelle un apprentissage « on line ». On peut également utiliser un apprentissage prenant en compte l'ensemble des points du jeu d'apprentissage, et corrigeant itérativement la position des poids, un apprentissage « batch ». Cette approche se base sur une formulation de l'état d'équilibre atteint à la fin de l'approche « on-line » (Fort *et al.*, 2001). Ce type d'approche a l'avantage de converger plus rapidement. C'est celle que nous avons utilisée.

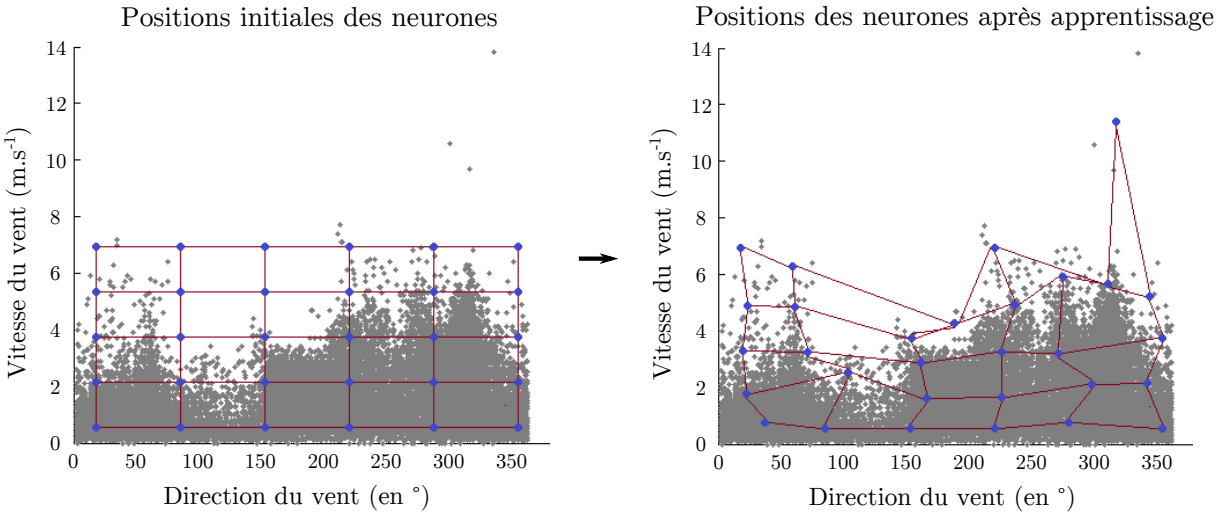


FIGURE 6.4 : Illustration de l'apprentissage d'une SOM sur deux variables, les direction et vitesse du vent dont les points sont représentés en gris (données de la station Sposata).

Chaque neurone de la SOM entrainée représente une classe, et chaque point de données fait partie de la classe du neurone dont il est le plus proche. Outre leur utilisation pour partitionner des données, les SOM peuvent par exemple être utilisées pour réduire la taille d'un jeu de données, en ne conservant que les points correspondant à la position des neurones après apprentissage.

Une fois qu'on dispose de la carte entrainée, on a classifié nos données en autant de classes que la carte a de neurones. Si l'on désire avoir un nombre de classes plus restreint, (le nombre

de neurones pouvant facilement dépasser la centaine) on peut par la suite utiliser un autre algorithme de clustering dont le nombre de classes final fait partie des paramètres. Par exemple, on peut utiliser l'algorithme k-means, (Lloyd, 1982; Forgy, 1965), ou la projection de Sammon (Sammon, 1969). Nous avons adopté l'algorithme k-means afin de reprendre la configuration utilisée par Lu *et al.* (2006) qui a donné de bons résultats à Taïwan pour l'ozone.

L'algorithme des k-means permet d'identifier k classes, définies par leur centre de gravité. Après une initialisation de ces centres de gravité, les points de données sont assignés au centre dont elles sont le plus proches en termes de distance euclidienne. Ce procédé minimise la distance quadratique intra-groupe D :

$$D = \sum_{i=1}^{k} \sum_{\mathbf{x} \in Q_i} ||\mathbf{x} - \mathbf{g}_i||^2 \tag{6.1}$$

avec $\mathbf{x}$ les points de données, Q_i le $i^{\text{ème}}$ groupe et $\mathbf{g}_i$ son centre de gravité. Quand les points ont été assignés au groupe, cela modifie son centre de gravité qui est recalculé, et le processus continue itérativement jusqu'à ce que les centres de gravité restent stables.

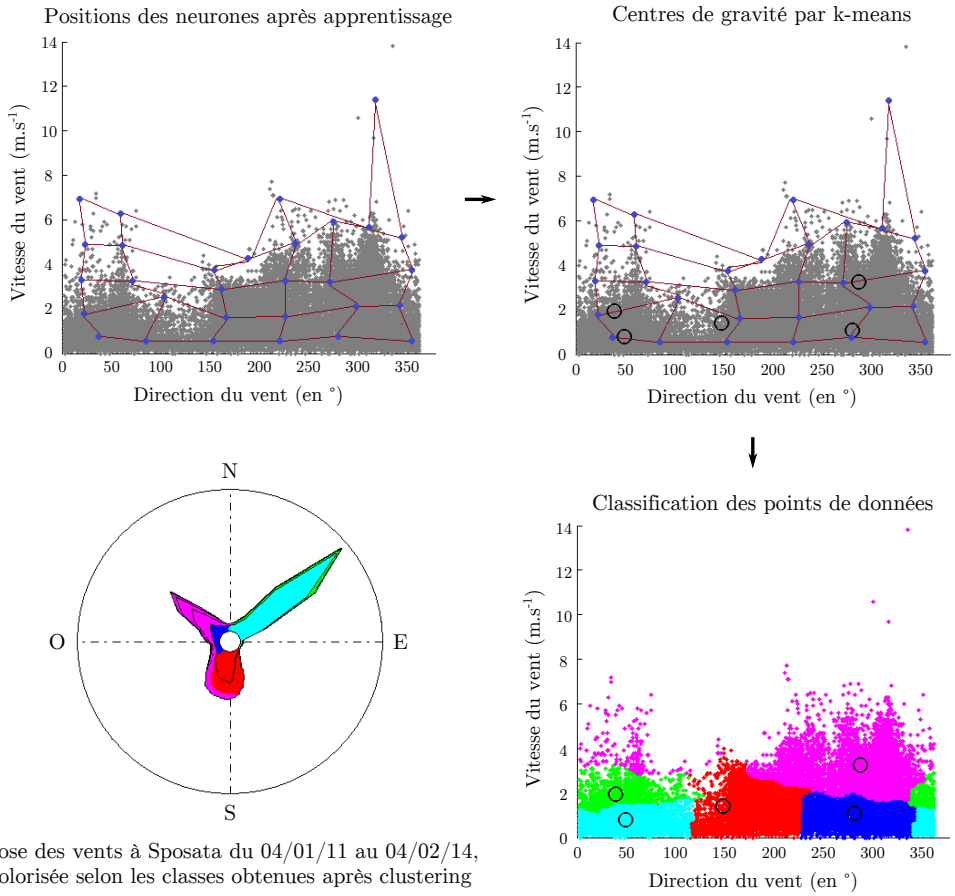


FIGURE 6.5 : Utilisation de l'algorithme du k-means à 5 classes sur les positions de neurones de SOM entrainée, suivi de l'assignation des points à leur classe respective, et rose des vents correspondant aux données utilisées colorisées selon les classes obtenues (données de la station Sposata).

La figure 6.5 illustre le processus de l'utilisation de SOM suivie de l'algorithme du k-means pour obtenir cinq classes à partir de mesures de vitesse et force du vent à Sposata. L'utilisation

des cosinus et sinus de la direction du vent permet de garder la continuité entre les valeurs proches de 0° et celle proche de 360°. La rose des vents équivalente aux données est également visible, colorisée selon la classe des points obtenue par le clustering. On peut y voir que les classes correspondent à des cas typiques sur cette rose des vents, avec en bleu clair les vents nord-est de force faible ou moyenne, en vert de force plus grande, en rouge les vents du sud, en bleu foncé les vents nord-ouest de faible vitesse et en rose les vents les plus forts, principalement nord-ouest.

L'avantage de l'utilisation des SOM plutôt que l'utilisation directe de ce type d'algorithme sur le jeu de données initial est le temps de calcul. Le principe de voisinage des neurones, et le fait que les neurones attirent leurs voisins avec eux quand ils se déplacent confèrent une vitesse de convergence aux cartes de Kohonen que n'ont pas les algorithmes tels que le k-means, ce qui met en question leur usage direct sur de larges jeux de données.

Avant l'étape de clustering, les données circulaires sont donc projetées et toutes les données sont normalisées, pour des raisons similaires à l'utilisation de ces prétraitements lors de la construction de modèles prédictifs (voir section 4.4 page 98). L'ACP a également des liens avec les processus de clustering de la famille des k-means. L'utilisation de composantes principales facilite la recherche des centres de gravité par l'algorithme des k-means (Ding et He, 2004). Nous utilisons donc l'ACP en prétraitement avant le clustering des données par SOM/k-means.

6.2.2 Classification Ascendante Hiérarchique

La Classification Ascendante Hiérarchique (CAH) est une autre méthode de clustering, qu'on peut appliquer à la distinction de régimes à partir de données météorologiques (Yu *et al.*, 2015). Elle n'a cependant pas encore été utilisée en prévision de la qualité de l'air. C'est une méthode itérative qui rassemble petit à petit les points de données en groupes. Chaque point de donnée est initialement assimilé à un groupe. Une métrique doit être définie pour représenter la distance entre les groupes. Il peut par exemple s'agir d'une distance euclidienne, métrique que nous avons utilisée. Un critère doit également être défini, qui utilise la distance entre les groupes pour choisir à chaque itération les deux groupes les plus proches. Nous avons utilisé le critère de Ward (Ward, 1963). Ce critère identifie les groupes de manière à minimiser leur variance intra-groupe, ce qui revient à maximiser la variance extra-groupe σ^2_{eg} d'après le théorème de Huygens (Saporta, 2006). On a :

$$\sigma^2_{eg} = \frac{1}{n} \sum_{i=1}^{k} n_i \, ||\mathbf{g} - \mathbf{g}_i||^2 \tag{6.2}$$

avec n la taille de l'échantillon, $\mathbf{g}$ son centre de gravité, k le nombre de groupes, n_i le nombre de points dans le groupe i et $\mathbf{g}_i$ son centre de gravité.

Ces deux groupes choisis sont fusionnés avant de passer à l'itération suivante. L'algorithme se poursuit pendant $n - 1$ itérations, n étant la taille de l'échantillon. A la fin du procédé, il ne reste plus qu'un unique groupe. On s'arrête à l'itération adéquate afin d'avoir le nombre de groupes désiré. On peut illustrer cette classification à l'aide d'un dendrogramme. Ce procédé est illustré en figure 6.6. L'approche ascendante, ou bottom-up, a son pendant descendant, ou top-down, qui consiste à partir d'un unique cluster et à le diviser itérativement jusqu'à ce que tous les points soient séparés. Nous avons préféré utiliser l'approche montante qui diffère plus de l'approche des k-means. Une approche descendante, lors de sa première itération, suit en effet le même objectif que l'algorithme des k-means avec $k = 2$, c'est-à-dire séparer en deux l'échantillon tout en suivant le même objectif de réduction de la variance intra-groupe. A l'inverse, la version

ascendante n'obtient les deux derniers groupes qu'à la fin du processus.

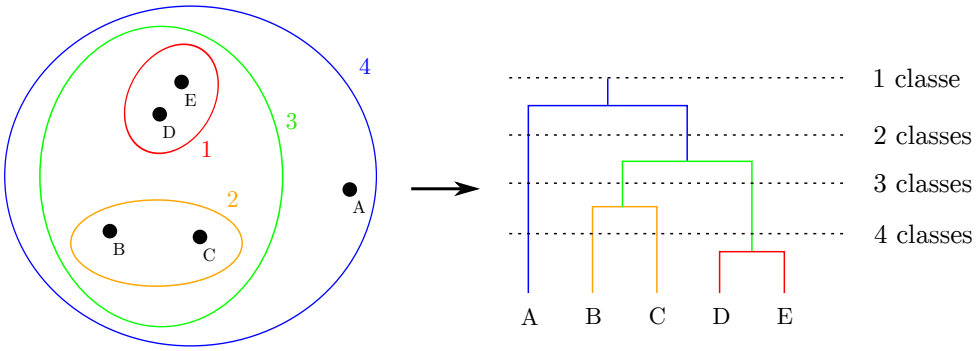


FIGURE 6.6 : Utilisation d'une classification ascendante hiérarchique de cinq points (A, B, C, D et E), avec à droite le dendrogramme correspondant.

6.2.3 Résultats obtenus par les modèles hybrides

Nous avons construit deux modèles hybrides, qui après une première phase de classification non-supervisée fournissent k sous-échantillons de données. Chaque sous-échantillon permet de créer un PMC prédictif. Ces modèles ont été entrainés pour réaliser des prévisions à l'horizon $h + 24$. Nous verrons que pour les PM10 et l'ozone, certains modèles hybrides arrivent a mieux prévoir les pics de pollution que le simple PMC.

Nous avons expérimenté l'hybridation de PMC et de clustering avec les données provenant des deux stations urbaines de l'île : Canetto et Sposata. Quelques statistiques sur les mesures effectuées à ces stations sont présentées au tableau 6.1. On y indique notamment le C90. Les concentrations des polluants de l'indice ATMO dépassent rarement les seuils d'information et d'alerte en Corse, quand on compare avec d'autres régions françaises. Utiliser le C90 des mesures permet donc d'avoir un seuil pour situer les mesures les plus élevées effectuées sur un site.

TABLEAU 6.1 : Statistiques sur les concentrations horaires d'O_3, de NO_2 et de PM10 et sur les moyennes sur 24h glissantes des concentrations de PM10 à Canetto et Sposata.

Station	Polluant	Moyenne ($\mu g.m^{-3}$)	Ecart-type ($\mu g.m^{-3}$)	Min ($\mu g.m^{-3}$)	Max ($\mu g.m^{-3}$)	C90 ($\mu g.m^{-3}$)
Canetto (Ajaccio)	O_3	58.25	29.20	0	166	98
	PM10	24.23	11.28	0	165	31
	PM10 *	*24.23*	*8.38*	*6*	*82*	*28*
	NO_2	21.48	16.29	0	128	45
Giraud (Bastia)	O_3	75.89	23.44	1	164	108
	PM10	22.31	10.73	0	149	31
	PM10 *	*22.31*	*8.04*	*5*	*79*	*28*
	NO_2	15.09	12.69	0	130	32

* *moyennes sur 24 heures glissantes*

Les études citées au début de la section 6.2 (Kolehmainen *et al.*, 2000; Lu *et al.*, 2006; Poggi et Portier, 2011) ont utilisé des mesures météorologiques pour réaliser le partitionnement. Nous utiliserons des sorties du modèle AROME, à l'échéance de notre modèle prédictif. La classification correspondra donc au régime météorologique attendu au moment de la prévision

$(h + 24)$, et non pas au régime actuel (à h). Les sorties de modèles seront celles au point de la grille du modèle le plus proche de la station dont proviennent les mesures de pollution (latitude de 41.925°N et longitude de 8.725°E pour la station Canetto, latitude de 42.7°N et longitude de 9.45°E pour la station Giraud). Les variables utilisées sont : TK, PA, U, V, HR, P, ECI, NEB, GEO, RS et RT. L'ensemble de ces variables doit permettre d'identifier les différents régimes météorologiques grâce à nos deux techniques de clustering. Les composantes du vent permettent d'isoler les régimes de vents particuliers, comme par exemple le Sirocco susceptible d'être chargé en particules, ou les vents en provenance du nord pouvant amener des masses d'air polluées du continent. L'Epaisseur de la Couche d'Inversion (ECI) permet d'identifier les situations de stabilité dues à une inversion thermique importante.

La HCL a été dans un premier temps incluse dans cet ensemble. Cependant, cette variable n'a été disponible qu'un an après les autres, son utilisation réduisant d'une année le jeu de données. Après une étude préliminaire, il est apparu que cette réduction du jeu était néfaste à la précision des modèles résultants et elle a due être supprimée du jeu de données. Notons également que, pour les même raisons, les sorties du modèle AIRES n'ont pas été intégrées au jeu de données. Il n'est donc pas possible de comparer les résultats obtenus avec ces modèles hybrides et ceux que donnent des modèles simples bénéficiant des sorties du CTM.

Pour utiliser les modèles de clustering présentés aux sections précédentes, il est nécessaire de fixer le nombre de classes que l'on veut obtenir. Les études citées au début de la section 6.2 se sont basées pour cela sur des indices quantifiant la pertinence de ce nombre de classes. Lu *et al.* (2006) ont utilisé l'indice de Davies–Bouldin (Davies et Bouldin, 1979) qui se calcule une fois le clustering effectué, mais ne prend pas en compte les performances des modèles prédictifs. Poggi et Portier (2011) ont utilisé le BIC (Bayesian Information Criterion), critère utilisé en sélection de modèle favorisant les modèles ayant la plus grande vraisemblance mais pénalisant un trop grand nombre de paramètres et de trop larges échantillons. Le BIC est donc calculé après entrainement des modèles prédictifs, pour rétrospectivement choisir le nombre de classes approprié. Nous avons préféré évaluer les modèles selon les méthodes présentées à la section 2.4 page 38. Des tests préliminaires montrent qu'au dessus de cinq classes, les performances des modèles chutent. Nous avons donc mené notre étude en utilisant les modèles entrainés sur une classe (modèle sans clustering) et de deux à cinq classes (modèles hybrides).

Chaque PMC prédictif a été entrainé dix fois, et c'est le meilleur d'entre eux (avec le plus fort indice d'agrément d) qui a été choisi pour représenter sa classe dans le modèle hybride. On dénomme le modèle constitué d'un simple PMC sans clustering sPMC. Le modèle hybride après un partitionnement par CAH sera noté hPMC et celui après partitionnement par SOM/k-means kPMC. Ces modèles sont illustrés à la figure 6.7.

Les données complètes utilisées couvrent les années de 2009 à 2014. Une année a été dédiée au jeu de validation (2011), une autre au jeu de test (2012) et le reste au jeu d'apprentissage. On note que données d'entrainement, de validation et de test des différents PMC formant un modèle hybride sont nécessairement différentes, puisqu'on les a justement partitionnées, mais les modèles hybrides sont eux évalués sur la même année de test complète. Nous avons entrainé les modèles à l'aide de l'algorithme de LM. La gestion des sous-groupes de données, l'entrainement des modèles correspondants, l'évaluation des modèles hybrides résultant ainsi que l'affichage des résultats ont été gérés grâce à notre application Aria Base.

Les scores des modèles sPMC pour les trois polluants sont montrés au tableau 6.2. Les tableaux 6.3 et 6.4 montrent ces scores pour les modèles kPMC et hPMC respectivement. Les scores sont également fournis pour les concentrations moyennes sur 24 heures glissantes en PM10, afin de correspondre avec les moyennes servant de base pour les alertes en France. Quatre modèles

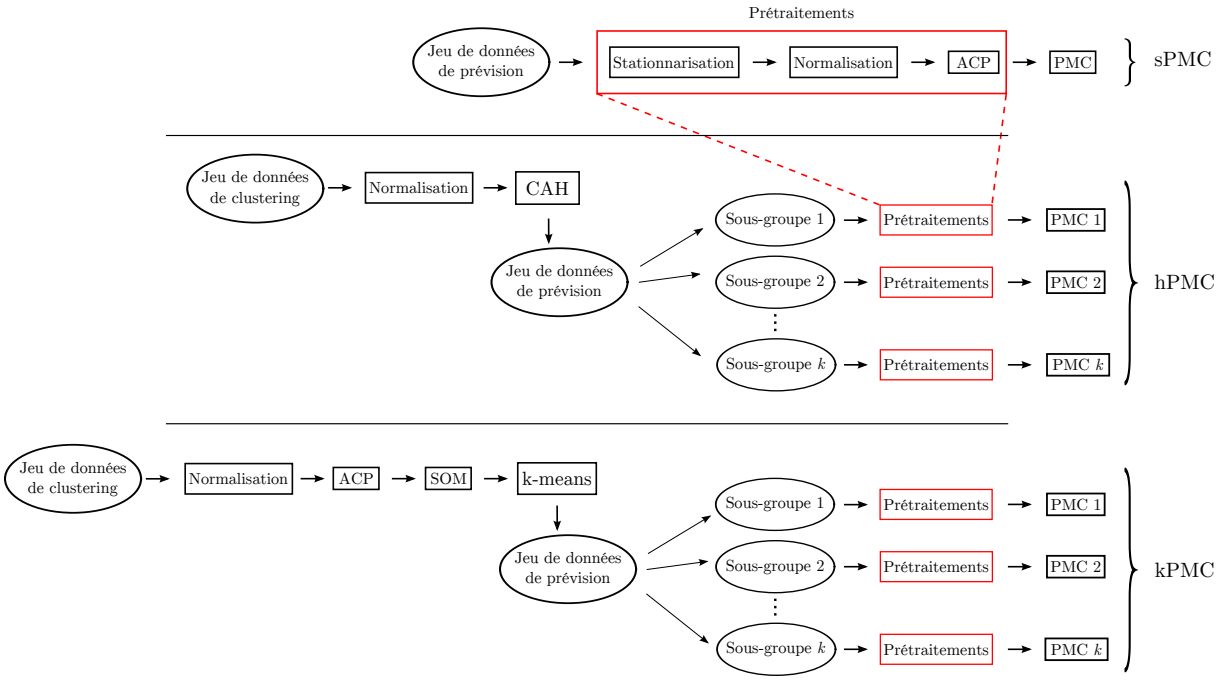


FIGURE 6.7 : Illustration des différentes étapes suivies par le modèle classique (sPMC) et des modèles hybrides CAH (hPMC) et SOM/k-means (kPMC).

hybrides ont été construits pour chaque polluant, puisqu'on a expérimenté les nombres de classes entre deux et cinq. Ceux qui sont montrés dans les tableaux sont ceux présentant le meilleur d.

TABLEAU 6.2 : Evaluation du modèle sPMC pour les prévisions à $h + 24$ des concentrations horaires d'O_3, de PM10 et de NO_2, ainsi que des concentrations moyennes sur 24h glissantes pour les PM10. d et R sont sans dimension.

Station	Polluant	RMSE ($\mu g.m^{-3}$)	nRMSE (%)	MAE ($\mu g.m^{-3}$)	d	R
Canetto	PM10	7.40	37.02	5.77	0.736	0.599
	*PM10**	*4.42*	*22.04*	*3.56*	*0.827*	*0.752*
	O_3	18.65	31.57	14.69	0.870	0.766
	NO_2	12.10	55.84	8.58	0.805	0.669
Giraud	PM10	7.49	38.89	5.70	0.728	0.606
	*PM10**	*4.73*	*24.52*	*3.72*	*0.834*	*0.759*
	O_3	15.84	20.53	12.23	0.837	0.743
	NO_2	10.90	71.97	7.30	0.735	0.592

* *moyennes sur 24 heures glissantes*

Les résultats présentés dans ces tableaux montrent que le modèle de référence sPMC obtient de meilleurs scores que les modèles hybrides (à l'exception de hPMC qui obtient le même indice d'agrément d que sPMC pour la moyenne sur 24 heures glissantes de PM10 à Canetto). Ceci peut s'expliquer par le fait que la division des sets crée des jeux de données plus petits, qui favorisent le sur-apprentissage des modèles prédictifs. Or, bien que la spécialisation de ces modèles soit recherchée, on sait qu'elle conduit facilement à une dégradation des capacités de généralisation. Cependant, nous verrons que malgré ces scores, calculés sur l'ensemble du set de test, les modèles hybrides peuvent surpasser sPMC pour la détection des pics.

TABLEAU 6.3 : Evaluation du modèle kPMC pour les prévisions à $h + 24$ des concentrations horaires d'O_3, de PM10 et de NO_2, ainsi que des concentrations moyennes sur 24h glissantes pour les PM10. d et R sont sans dimension.

Station	Pollutant	Nombre de classes	RMSE ($\mu g.m^{-3}$)	nRMSE (%)	MAE ($\mu g.m^{-3}$)	d	R
Canetto	PM10	4	8.28	41.32	6.48	0.681	0.499
	*PM10**		*4.66*	*23.26*	*3.83*	*0.807*	*0.726*
	O_3	2	22.56	36.04	17.53	0.808	0.662
	NO_2	3	15.42	74.81	10.95	0.672	0.452
Giraud	PM10	3	8.10	42.31	6.10	0.662	0.509
	*PM10**		*5.00*	*26.09*	*3.90*	*0.797*	*0.720*
	O_3	2	16.99	21.81	13.17	0.818	0.699
	NO_2	2	11.96	79.79	8.21	0.642	0.474

* *moyennes sur 24 heures glissantes*

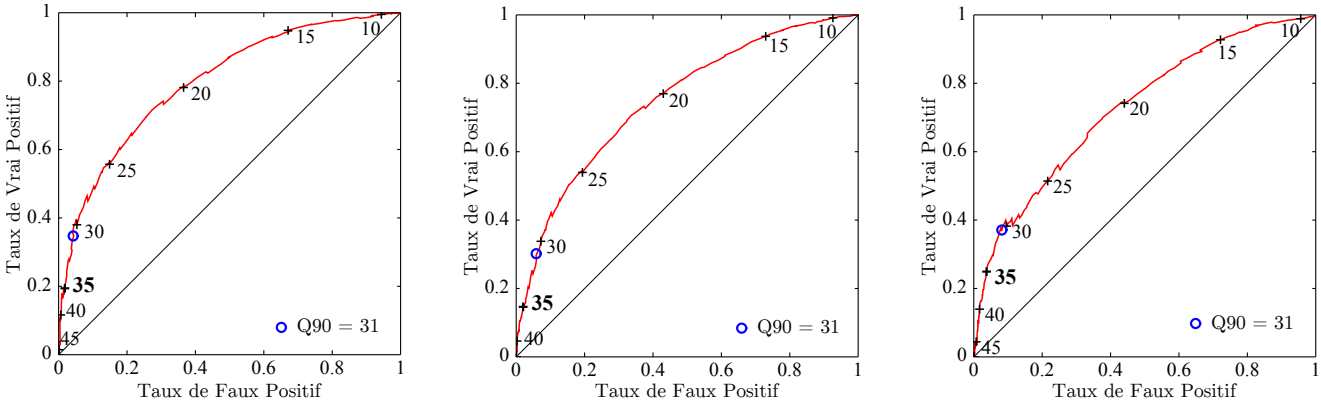
TABLEAU 6.4 : Evaluation du modèle hPMC pour les prévisions à $h + 24$ des concentrations horaires d'O_3, de PM10 et de NO_2, ainsi que des concentrations moyennes sur 24h glissantes pour les PM10. d et R sont sans dimension.

Station	Polluant	Nombre de classes	RMSE ($\mu g.m^{-3}$)	nRMSE (%)	MAE ($\mu g.m^{-3}$)	d	R
Canetto	PM10	3	8.72	43.51	6.79	0.649	0.437
	*PM10**		*4.53*	*22.59*	*3.65*	*0.826*	*0.728*
	O_3	4	24.22	38.69	19.03	0.779	0.613
	NO_2	4	17.57	85.26	12.82	0.591	0.326
Giraud	PM10	2	7.91	41.36	6.02	0.696	0.545
	*PM10**		*4.74*	*24.77*	*3.71*	*0.831*	*0.754*
	O_3	2	16.74	21.48	12.93	0.826	0.711
	NO_2	2	12.34	82.23	8.52	0.655	0.464

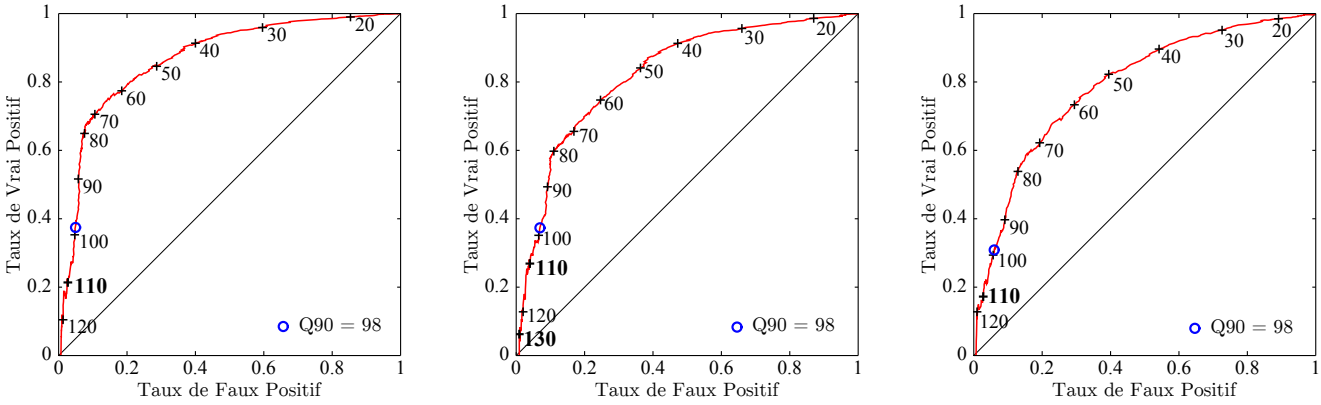
* *moyennes sur 24 heures glissantes*

Intéressons-nous à cette capacité de détection des évènements à forte concentration. Les figures 6.8 et 6.9 montrent les courbes ROC de ces modèles, pour la station Canetto d'Ajaccio et la station Giraud de Bastia respectivement. Les modèles hybrides ont chacun été entrainés avec de 2 à 5 classes, seul le meilleur modèle est montré ici avec son nombre de classes indiqué. Cette fois pour les modèles hybrides, nous avons montré les résultats non pas des modèles au meilleur d, mais ceux dont les courbes ROC montrent les meilleurs taux de détection pour les fortes concentrations.

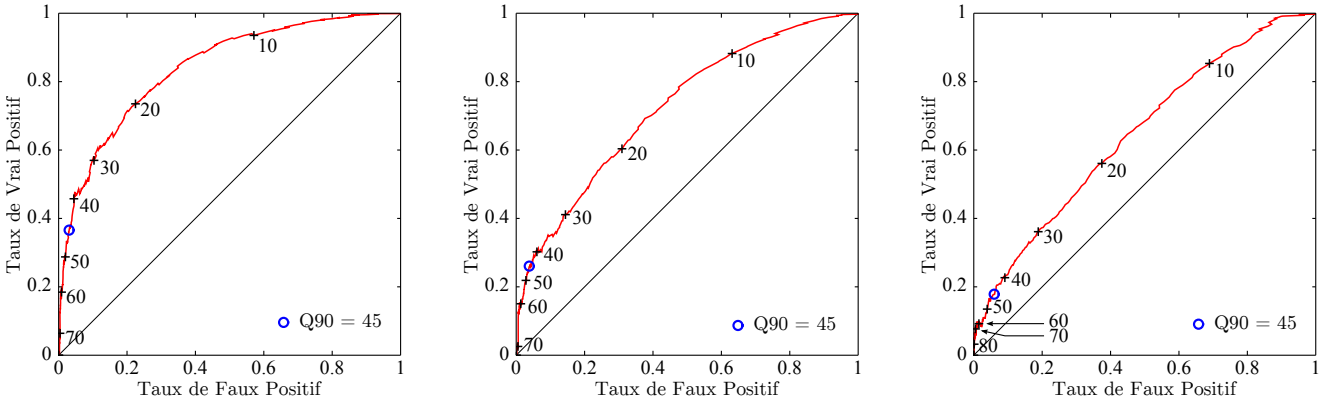
Sur les courbes ROC, on s'intéresse aux valeurs de seuil les plus élevées, en bas à gauche du graphique. C'est l'augmentation du TVP qui est recherchée, pour les seuils les plus élevés. L'examen de ces courbes mène à deux constats principaux. Le premier est que les meilleurs modèles en termes de scores généraux ne sont pas nécessairement ceux ayant les meilleurs taux de détection pour les fortes concentrations. Ce constat n'est pas surprenant, et souligne l'intérêt de l'utilisation de représentations comme les courbes ROC à la place de simples indices. Les meilleurs modèles hybrides ne sont donc pas les mêmes que ceux des tableaux 6.3 et 6.4.



PM10, sPMC | PM10, kPMC (2 clusters) | PM10, hPMC (2 clusters)

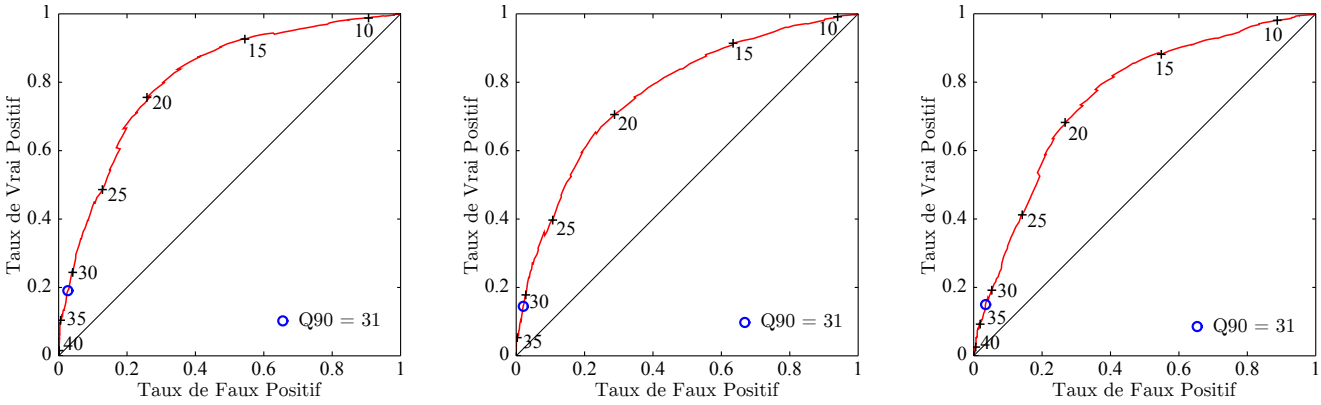


O_3, sPMC | O_3, kPMC (2 clusters) | O_3, hPMC (3 clusters)

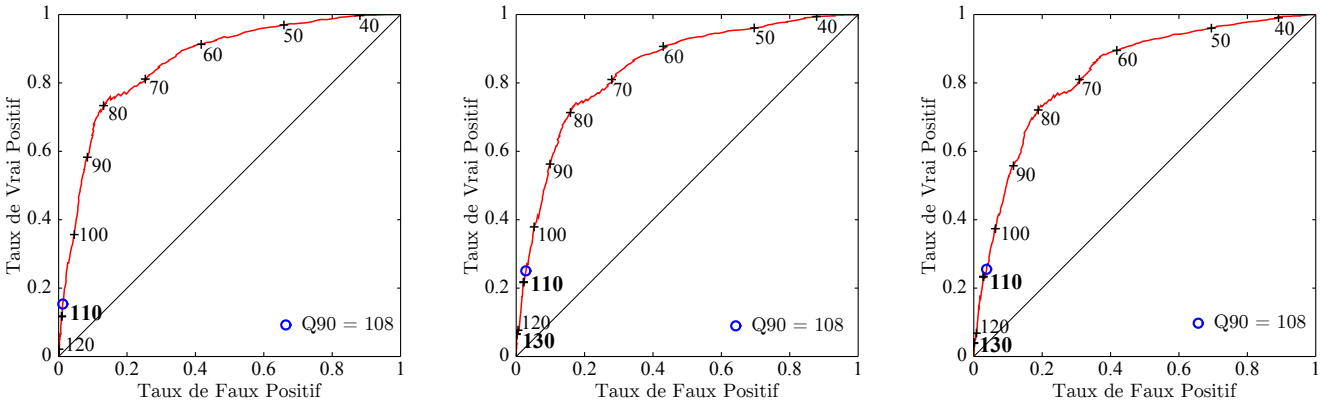


NO_2, sPMC | NO_2, kPMC (4 clusters) | NO_2, hPMC (5 clusters)

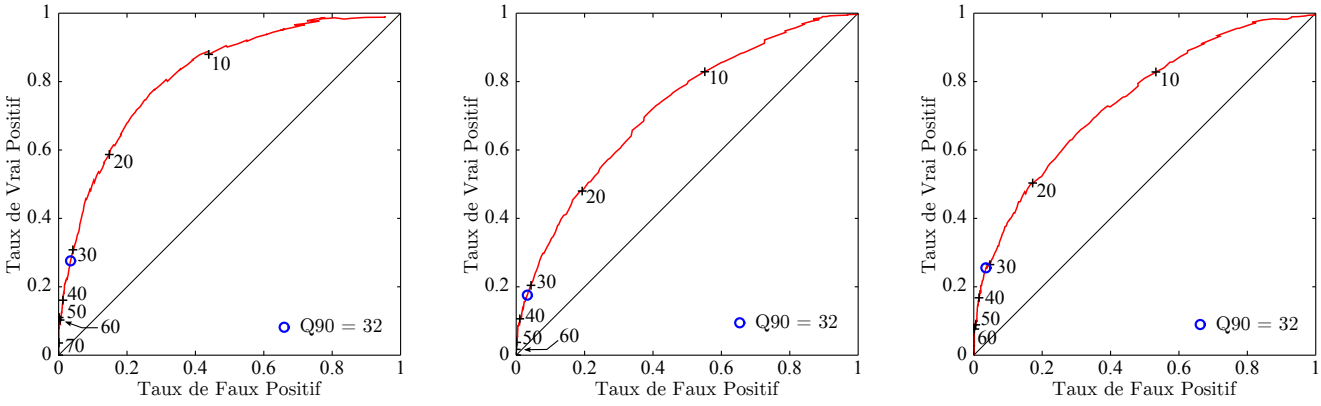
FIGURE 6.8 : Courbes ROC des modèles de prévision à $h + 24$ de concentrations en PM10, O_3 et NO_2 à la station de Canetto (Ajaccio). Les modèles hybrides comportant de 2 à 5 classes ont été créés et les résultats montrés ici sont ceux présentant les meilleurs taux de détection pour les fortes concentrations. Certains points sont mis en valeur en gras. Les seuils indiqués sont en $\mu g.m^{-3}$.



PM10, sPMC | PM10, kPMC (5 clusters) | PM10, hPMC (5 clusters)



O_3, sPMC | O_3, kPMC (2 clusters) | O_3, hPMC (3 clusters)



NO_2, sPMC | NO_2, kPMC (5 clusters) | NO_2, hPMC (2 clusters)

FIGURE 6.9 : Courbes ROC des modèles de prévision à $h + 24$ de concentrations en PM10, O_3 et NO_2 à la station de Giraud (Bastia). Les modèles hybrides comportant de 2 à 5 classes ont été créés et les résultats montrés ici sont ceux présentant les meilleurs taux de détection pour les fortes concentrations. Certains points sont mis en valeur en gras. Les seuils indiqués sont en $\mu g.m^{-3}$.

Le second constat est que les modèles hybrides surpassent en général les capacités du sPMC pour les PM10 et l'ozone, mais jamais pour le dioxyde d'azote. Pour les PM10, on a les TVP les plus élevés pour les concentrations au dessus du C90 à Canetto avec hPMC (2 classes), bien qu'à Bastia le sPMC garde les meilleurs résultats. Pour l'ozone, kPMC obtient les meilleures performances à Ajaccio. Les deux modèles hybrides ont des résultats similaires au dessus du C90 à Bastia et y surpassent sPMC. On observe que lorsqu'un modèle hybride surpasse sPMC pour les fortes concentrations, c'est toujours lorsqu'il est composé de deux ou trois classes, quels que soient le site et le polluant.

Malheureusement, la division des jeux de données telle que nous l'avons menée ne semble pas profiter aux modèles de prévision de NO_2, dont les meilleurs modèles sont systématiquement de type sPMC. On peut d'ailleurs se poser la question dans ce cas du choix des variables utilisées pour la classification non-supervisée. Cette classification était pensée de manière à refléter le régime météorologique du lendemain, en utilisant des sorties d'AROME, proches du sol ou à plus haute altitude. Cela semble convenir pour les particules et l'ozone, touts deux sujets à des épisodes de transport à méso-échelle (entre la dizaine et le millier de kilomètres), et très sensibles aux conditions météorologiques (formation, dispersion, dépôt, etc.). A Giraud et Canetto qui sont des stations urbaines, le NO_2, bien qu'également soumis à ces conditions météorologiques, a des concentrations qui dépendent plus des émissions anthropiques locales. Un jeu de données plus représentatif des émissions, avec par exemple des indices temporels voire des mesures de trafic, pourrait être plus adapté à la classification pour les modèles hybrides de NO_2.

Pour la prévision des PM10, on utilise les sorties de nos modèles pour calculer les moyennes sur 24 heures glissantes des concentrations, afin de coïncider avec les moyennes utilisées de manière opérationnelle pour émettre les alertes. Intéressons-nous au site de Canetto, qui est à la fois celui où un modèle hybride surpasse sPMC, et à la fois le site le plus concerné par les dépassements de seuil de ce polluant. Les résultats de sPMC et du meilleur modèle hybride hPMC sont montrés en utilisant cette moyenne, pour sPMC en figure 6.10 et pour hPMC en figure 6.11.

L'amélioration des taux de détection pour les valeurs fortes est visible sur ces figures. hPMC peut ainsi prévoir plus de 35 % des dépassements de 39 $\mu.g.m^{-3}$, alors que sPMC ne prévoit aucun des dépassements au dessus de 36 $\mu.g.m^{-3}$. Cette tendance à moins sous-estimer les concentrations se voit par une hausse de la MBE qui passe de 1.64 $\mu.g.m^{-3}$ pour sPMC à 2.20 $\mu.g.m^{-3}$ pour hPMC, ainsi que par le FB qui passe de -0.08 à -0.10. La précision de hPMC en termes de RMSE ou de d est par contre moindre que celle de sPMC, mieux optimisé pour l'ensemble des situations. On remarque également une diminution de la Fractional Variance (FV) entre sPMC et hPMC (0.56 et 0.39 respectivement), indiquant que la variance des sorties de ce dernier est moins diminuée par rapport à celle des observations que pour sPMC. Cela illustre le fait que hPMC a moins tendance à sous-estimer les valeurs extrêmes (hautes ou basses), ce qui va dans le sens de la prévision des pics de pollution.

La figure 6.12 montre les séries temporelles mesurée et prédite par les trois modèles. On peut voir que hPMC détecte mieux le pic autour du 18 mars 2012. Les prévisions y apparaissent relativement bonnes.

Pour conclure, on peut dire que l'utilisation de classification non-supervisée pour séparer les données peut être bénéfique pour la prévision de pics de pollution aux PM10 ou à l'ozone. L'utilisation de l'algorithme des k-means précédée d'une réduction du jeu de données par SOM, ainsi que la classification ascendante hiérarchique donnent tous deux des modèles capables de mieux détecter les fortes concentrations. C'est tout de même la CAH qui apporte les meilleures performances. La méthodologie de clustering étant également plus simple à mettre en place que

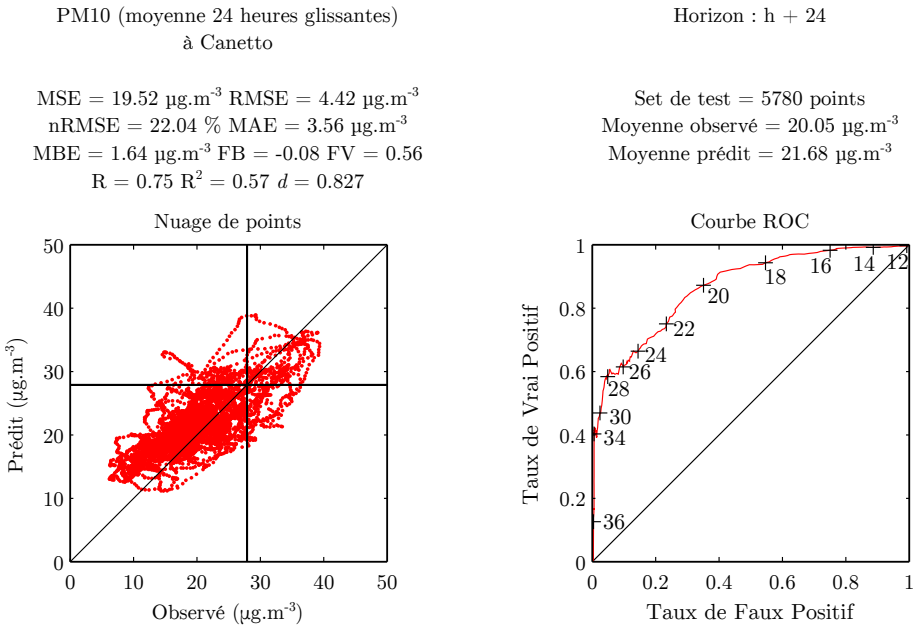


FIGURE 6.10 : Résultat du modèle sans classification préalable (sPMC) pour des moyennes sur 24h glissantes de concentration de PM10 à Canetto.

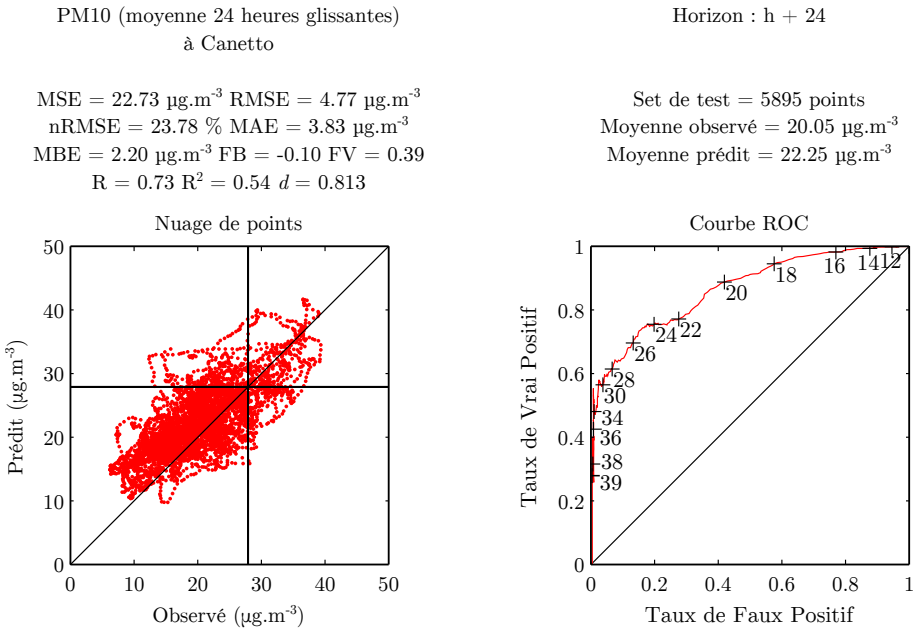


FIGURE 6.11 : Résultat du modèle hybride avec classification ascendante hiérarchique préalable (hPMC à 2 classes) pour des moyennes de 24h glissantes de concentration de PM10 à Canetto.

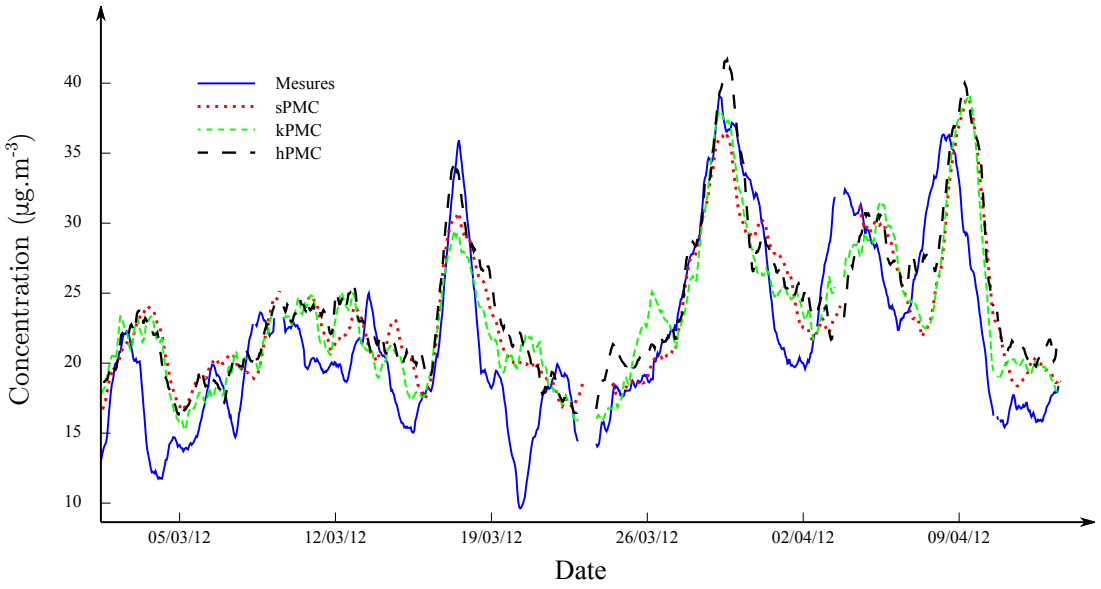


FIGURE 6.12 : Séries temporelles de moyennes sur 24h glissantes de PM10 à Canetto mesurée et prévues par sPMC, kPMC et hPMC.

celle de kPMC, qui utilise deux modèles successifs, nous préfèrerons utiliser hPMC.

Le nombre de classes apparait devoir se limiter à 2, voire 3, ce qui coïncide avec l'étude menée par Poggi et Portier (2011) utilisant des modèles linéaires. Nous préconisons l'utilisation exploratoire de ces deux méthodes de clustering dans les études de prévision de la qualité de l'air basées sur des modèles statistiques.

Un des intérêts majeurs de cette méthodologie est qu'elle permet d'améliorer la détection des événements extrêmes et rares des modèles alors que ces pics sont très peu présents dans les données corses. Cette rareté des fortes concentrations est difficile à surmonter pour des modèles à apprentissage automatique. Les modèles hybrides pourraient se comporter différemment sur des jeux de données contenant plus d'épisodes de forte pollution. L'amélioration de la détection des pics est en tout cas un atout pour la prévision en Corse, et occulte la légère baisse de scores généraux comme la RMSE ou d.

Le fait qu'on obtienne de meilleures capacités de détection avec de moins bons scores moyens amène indirectement des perspectives vis-à-vis de l'apprentissage. Les algorithmes à direction de descente que nous avons utilisés sont programmés pour optimiser une certaine fonction objectif, qui correspond dans notre cas à la MSE. La Neural Network Toolbox que nous avons utilisée ne propose que cette fonction ainsi que la Mean Absolute Error (MAE). Il serait intéressant de développer un algorithme de LM, basé sur une fonction mettant d'avantage en avant la précision pour les fortes concentrations, comme par exemple une MSE calculée uniquement pour les observations dépassant le C90.

6.3 Conclusion

Nous avons dans ce chapitre présenté nos résultats de prévision de la qualité de l'air avec des modèles hybrides, utilisant plusieurs PMC sur des sous-échantillons de données. Une augmentation des taux de détection des hautes concentrations de nos modèles est observée dans la plupart des cas. Les résultats que nous avons obtenus, en cour de publication (Tamas *et al.*, sous presse), nous permettront d'améliorer les modèles utilisés par Qualitair Corse. Cette méthode

permet une amélioration dans un contexte compliqué, où peu de pics de pollution sont présents dans les données.

Plusieurs choix doivent être considérés pour utiliser ces modèles hybrides. Le nombre de classes (entre deux et trois) ainsi que la méthode (basée sur les k-means ou la classification ascendante hiérarchique) ne peuvent être fixés *a priori*. Cependant, pour les PM10, la CAH avec deux classes donne les meilleurs résultats. Cette alternative à la méthode proposée par Lu *et al.* (2006) (pour la prévision d'ozone) est plus à-même d'améliorer la prévision de pics de particules.

La division des données suivant des règles ayant pour but d'identifier les situations connues pour être propices aux pics fonctionne en général moins bien que la classification non-supervisée pour les PM10. Mais les résultats obtenus avec le clustering motivent une reprise des expériences sur la division par l'utilisateur, en examinant d'autres types de division et en se limitant à deux ou trois classes. Pour l'ozone, des cas simples comme la division des données en fonction de la saison apportent cependant déjà une amélioration des taux de détection.

Une perspective intéressante est l'application de métaheuristiques comme le recuit simulé pour rechercher les meilleurs configurations de modèles hybrides (meilleur algorithme de clustering, nombre de classes), en optimisant la fonction objectif pour qu'elle prenne en compte les taux de détection des pics plutôt que les scores usuels.

L'étude de sensibilité des modèles à l'aide d'outils comme les courbes ROC s'est avérée nécessaire pour étudier en détail le comportement des modèles, en s'affranchissant du choix d'un seuil en particulier. Les modèles hybrides sont implémentés au sein de l'application Aria Web qui est déployée à Qualitair Corse. Les courbes ROC des modèles sont systématiquement fournies aux prévisionnistes. Connaissant ainsi le comportement du modèle, il est plus facile d'interpréter les résultats des modèles prédictifs.

L'usage de notre application Aria Base aura grandement aidé la réalisation de ces expériences. La gestion de la division des jeux, de l'entrainement des modèles et de l'évaluation des modèles hybrides aura été des plus utiles. Une de nos perspectives est donc de continuer l'étude des modèles hybrides avec cette application. Le clustering provoquant une diminution du volume de données nuisant aux résultats, on s'attend également à une amélioration quand plus de données seront disponibles.

Chapitre 7

Développements logiciels

Après avoir introduit le contexte qui a mené vers l'élaboration de ce travail de doctorat, nous avons eu l'occasion de présenter les résultats que nous avons obtenus pour la prévision de la qualité de l'air en Corse à l'aide de modèles à apprentissage. Mener ces expériences a demandé un travail de programmation que nous allons maintenant présenter.

La prévision à l'aide de réseaux de neurones a de nombreux avantages, qu'on a vus à la section 2.3.4 (page 35), mais possède le défaut de nécessiter un grand nombre d'opérations de paramétrage. Plusieurs algorithmes d'apprentissage existent et doivent être envisagés (section 4.3.4 page 97), des prétraitements sont nécessaires (section 4.4 page 98), le choix des variables est une problématique en soi (section 5.1 page 113). Identifier une configuration optimale demande donc de nombreuses expérimentations avec les RNA, et d'autant plus si on utilise d'autre méthode pour améliorer les résultats, comme l'hybridation des modèles présentée au chapitre 6 (page 148). Nous avons adapté notre méthodologie de travail à cette contrainte.

Nous avons principalement utilisé le langage de programmation Matlab, avec sa Neural Network Toolbox. Afin d'utiliser efficacement cette boîte à outil et de prendre en compte nos contraintes, nous avons développé, sous Matlab, une application dont l'objectif est de nous permettre de mener nos expériences. Nous avons appelé le prototype d'application « Aria Base ».

Aria Base a été conçu pour réaliser les tâches suivantes :

- Récupération et gestion de l'ensemble des données
- Configuration complète des modèles prédictifs
- Gestion des modèles hybrides
- Conduite automatique des expériences préconfigurées
- Evaluation détaillée des modèles
- Archivage des modèles et de leur évaluation
- Export des modèles

L'outil Aria Base est un outil de recherche, mais nous verrons qu'il permet également de fournir les modèles opérationnels, utilisés par l'AASQA. Quand avec le temps, l'AASQA a enregistré suffisamment de nouvelles données, il permet de réentrainer les modèles statistiques afin de les maintenir à jour. Mais cela reste un outil technique, fait pour être utilisé par un modélisateur statisticien. Une autre application a été développée, dédiée à la réalisation des prévisions quotidiennes.

Cette autre application est appelée « Aria Web ». Il s'agit d'une application web, hébergée sur le serveur de Qualitair Corse. Elle a pour fonction de faire fonctionner les modèles neuronaux entrainés importés depuis « Aria Base ». En plus de cela, elle est chargée de regrouper et

d'archiver différentes cartes de prévisions qui sont disponibles sur internet, afin de réunir toutes les informations utiles. C'est une application d'aide à la décision, prévue pour être utilisée par les prévisionnistes de l'AASQA. Aria Web sera décrite à la section 7.2, page 173.

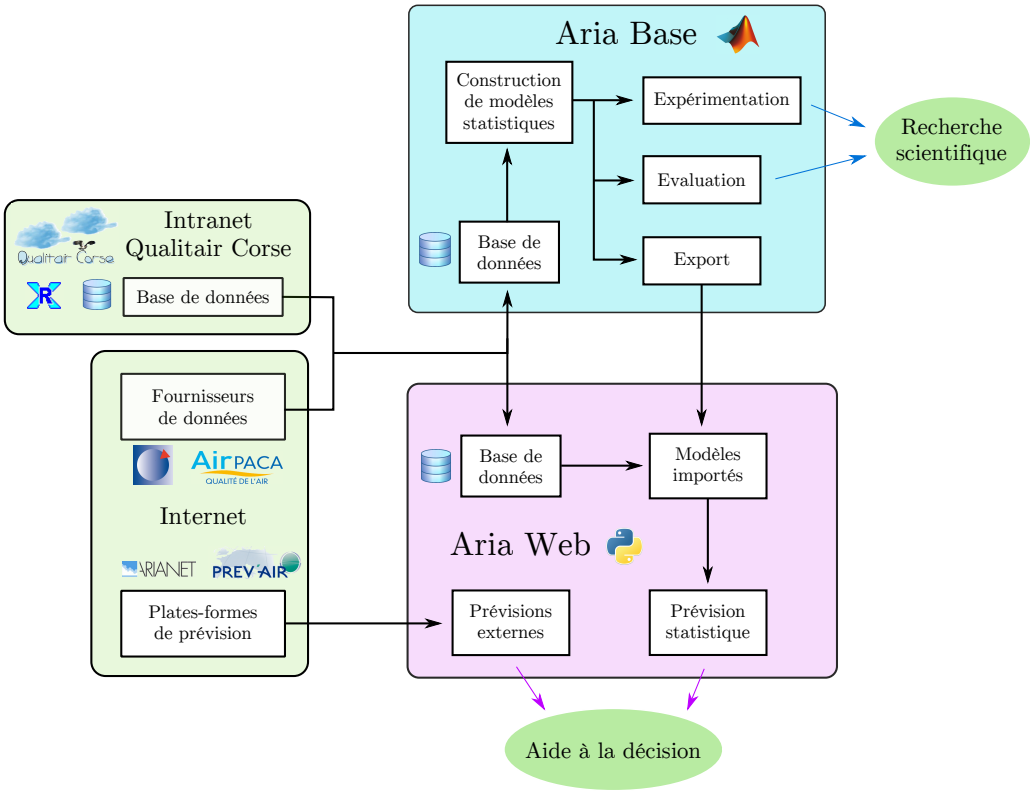


FIGURE 7.1 : Coordination entre les applications Aria Base et Aria Web.

Les rôles respectifs de ces deux applications, illustrées à la figure 7.1, couvrent l'ensemble de la problématique de ce travail de thèse. Leurs tâches respectives s'inscrivent dans un contexte différent, ce qui explique la programmation de deux applications séparées. La première est dédiée à un modélisateur statisticien, et lui permet la recherche du meilleur modèle statistique de prévision de la qualité de l'air adapté à une problématique donnée et de le maintenir à jour. Elle est axée vers le calcul numérique, alors que la seconde est un outil d'aide à la décision, voué à de l'affichage synthétisé de résultats, axé vers le web et s'adressant aux prévisionnistes eux-mêmes.

7.1 Application de recherche « Aria Base »

Aria Base a été développée sous Matlab (version 2012a) et utilise les fonctions de la Neural Network Toolbox (fr.mathworks.com/products/neural-network/), version 7.0.3.

Matlab est un environnement de développement ainsi qu'un langage de programmation dit de quatrième génération, développé par la société MathWorks. Ce langage est adapté à un usage par des scientifiques de différents horizons, qui ne sont pas informaticiens.

Au delà de la facilité de prise en main de Matlab, c'est également pour ses capacités de calcul matriciel et de manipulation des matrices que Matlab connait un certain succès (Matlab est la contraction de « Matrix Laboratory »). Ce sont ces dernières raisons qui nous ont incités à l'utiliser pour nos travaux. Pour une description détaillée des méthodes de calcul numérique avec Matlab, on pourra consulter le livre de Lindfield et Penny (2012). Matlab est en développement

continu depuis sa création en 1970. Il permet (comme de nombreux langages) la programmation orientée objet, que nous utiliserons pour nos travaux. De nombreuses fonctionnalités et boîtes à outils (« toolboxes ») ont été ajoutées. Le tracé de graphique est particulièrement souple, et il existe des possibilités de développement d'interface graphique.

La NNToolbox est une boîte à outil de Matlab, qui fournit des outils nécessaires à l'expérimentation à l'aide de réseaux de neurones. Elle permet la création de réseaux neuronaux, leur entrainement ainsi que leur évaluation. Cette boîte à outil propose une interface graphique qui facilite sa prise en main. Il reste heureusement possible de gérer les modèles directement depuis l'invite de commande ou depuis différents scripts ou fonctions développés par l'utilisateur. Mais l'interface graphique proposée ne permet pas d'accéder à l'intégralité des paramètres des modèles. On peut voir sur la figure 7.2 un exemple d'interface graphique de cette boîte à outil proposant de configurer le nombre de neurones cachés d'un PMC.

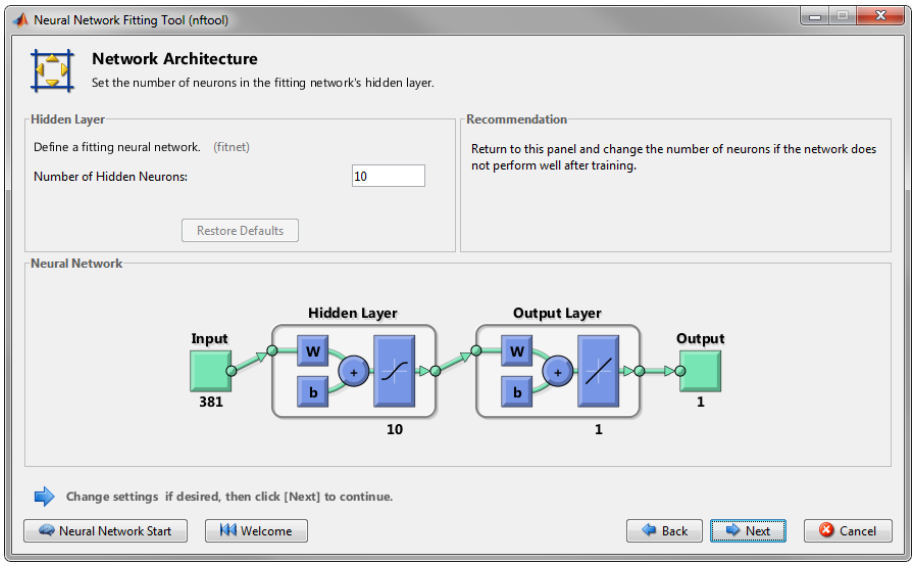


FIGURE 7.2 : Capture d'écran d'une interface graphique de la Neural Network Toolbox proposant une sélection du nombre de neurones cachés.

La NNToolbox permet de créer des modèles neuronaux sous forme d'objet, dont l'ensemble des paramètres est indiqué dans ses attributs. Sont ainsi renseignés le nombre de données d'entrées, la structure de chaque couche, les paramètres et fonctions de transfert des neurones de chaque couche ou les algorithmes d'apprentissage à utiliser. Certains prétraitements peuvent également être gérés, ainsi que la division des jeux de données ou les décalages temporels de certaines variables.

Les paramètres des réseaux de neurones (poids et biais) sont représentés par des matrices organisées par couches. Les connexions entres les différentes couches sont représentées par des matrices booléennes. Les méthodes de l'objet permettent plusieurs actions, comme déclencher l'apprentissage du réseau, simuler la sortie du réseau après apprentissage ou en dessiner la structure.

Se reposer entièrement sur les fonctions d'une boîte à outils comporte certains désavantages. Certaines fonctions ne sont pas entièrement décrites, et leur script est inaccessible ou difficile à relire et interpréter. De plus, certains paramètres et certaines données ne sont pas accessibles à l'utilisateur. Ceci peut poser problèmes quand on veut créer des configurations qui ne sont pas prévues par la boîte à outil. Cela peut aussi empêcher l'exportation de données ou de paramètres en dehors de Matlab.

A ces difficultés s'ajoutent la relativement grande dimensionnalité des paramètres de configuration à optimiser. Tous les détails de configuration devant être fixés (architecture, initialisation des paramètres, apprentissage, variables d'entrée, prétraitements, stratégies d'optimisation, etc.) influent sur les résultats des prévisions, de manière interdépendante. Optimiser par exemple le nombre de neurones cachés n'a plus de sens si l'on supprime ensuite des variables par soucis de parcimonie, ou encore la meilleure configuration pour prévoir l'ozone sur un site ne sera pas la même que celle obtenue pour un autre site. Cette contrainte implique la réalisation d'un grand nombre d'expériences pour converger vers les meilleurs compromis de configuration, méthodologie qui est évoquée à la section 5.3 (page 128). Un archivage de ces expériences est également nécessaire pour comparer les modèles et les réutiliser.

Pour toutes ces raisons, nous avons rapidement envisagé de créer notre propre application, adaptée à nos besoins expérimentaux. De la NNToolbox de Matlab, nous avons uniquement utilisé les fonctionnalités permettant de créer et entrainer les RNA. La gestion des données, leur prétraitement, la gestion des jeux de variables, l'évaluation des modèles, l'archivage des expériences (configuration complète et résultats d'évaluation) ont été réalisés par nos fonctions, via l'interface graphique développée pour Aria Base.

Chaque fonctionnalité de l'application Aria Base est présentée en détail à l'annexe C, page 211, qui peut être considérée comme le manuel utilisateur de ce prototype d'application de recherche.

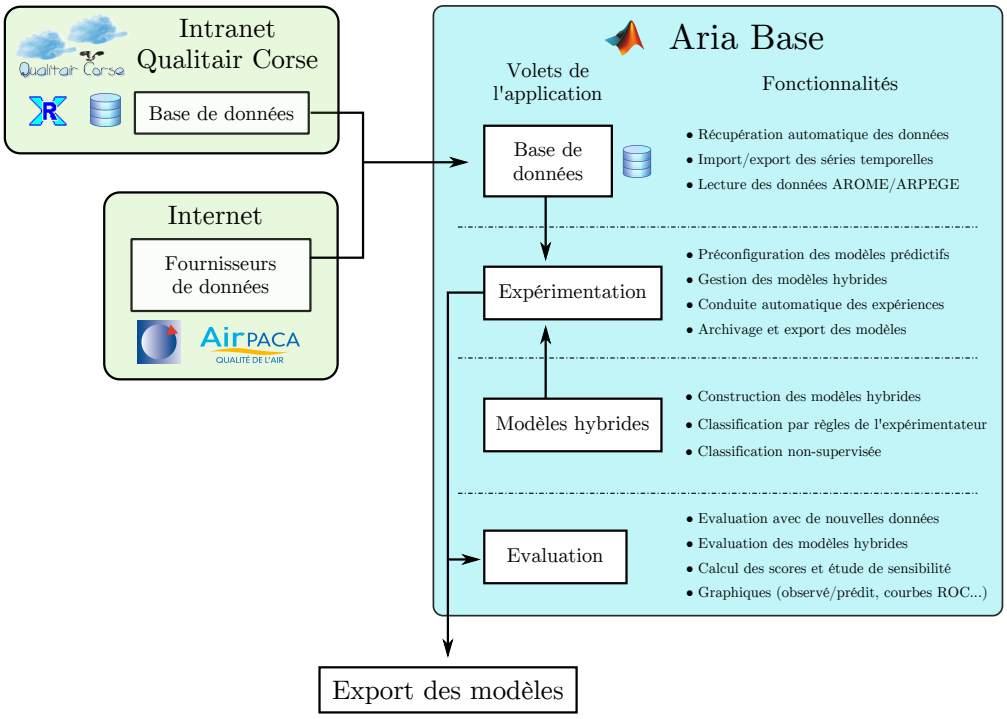


FIGURE 7.3 : Structure et fonctionnalités de l'application Aria Base.

Les fonctionnalités d'Aria Base sont présentées à la figure 7.3. Elles assurent deux grandes tâches, que sont la gestion des données et celles des modèles, que nous allons maintenant présenter.

7.1.1 Gestion des données

Les modèles statistiques sont basés sur les interactions existant entre plusieurs variables. Les données décrivant ces variables doivent être récupérées et gérées. C'est par la gestion des données qu'a commencé le développement de Aria Base. Cette gestion permet la récupération et le formatage des données, leur stockage, leur visualisation et leur export.

Les données proviennent de fournisseurs différents (Qualitair Corse, Météo-France, Air PACA) et leur formatage n'est pas le même. De plus, les données doivent être récupérées régulièrement, au fur et à mesure qu'elles sont produites. Chaque format est donc geré par Aria Base, et une récupération automatique via des serveurs FTP (File Transfer Protocol) est en place. Cela permet à l'expérimentateur d'accéder aux variables dont il a besoin, pour les utiliser directement avec Aria Base ou les exporter vers l'espace de travail de Matlab, ou sous un autre format (xls par exemple). La figure 7.4 montre le volet « Base de données » de l'application qui permet ces opérations.

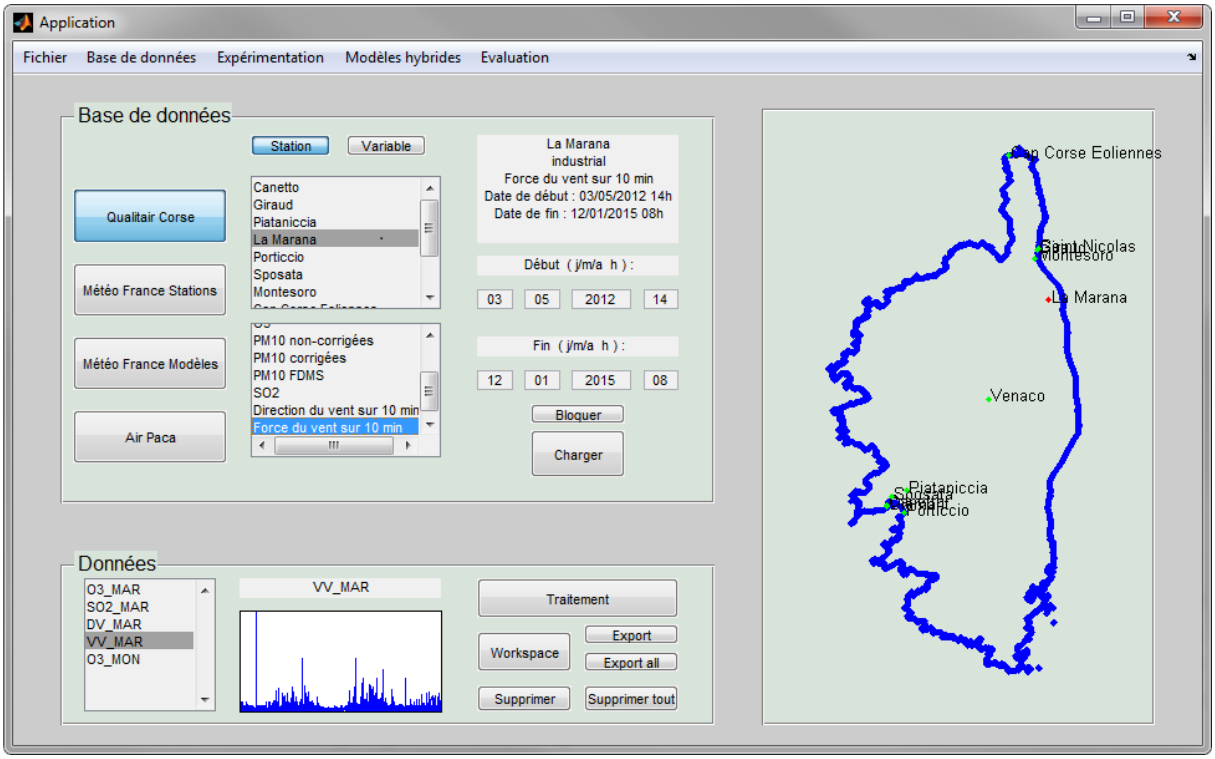


FIGURE 7.4 : Capture d'écran de la page de gestion des données.

Au delà de la gestion des séries temporelles, propre au travail de prévision statistique, ce volet permet également de visualiser directement les données telles qu'elles sont présentées dans les fichiers issus des fournisseurs. En effet, il n'y a aucun autre moyen à Qualitair Corse de travailler directement avec par exemple les fichiers bruts d'AIRES, qui peuvent être utiles pour d'autres études.

7.1.2 Gestion des modèles

La gestion des modèles prévisionnels est fortement améliorée par l'utilisation d'Aria Base. On a vu notamment aux chapitres 4 et 5 l'étendue des détails de configuration qui devaient être fixés avant l'usage d'un RNA. Le plus souvent, modifier l'un des éléments de configuration peut

impliquer de devoir refixer certains autres. Il est nécessaire, quand on travaille avec ce type de modèle, d'effectuer un grand nombre d'expériences et de les archiver avant d'obtenir la meilleure configuration. Ces expériences consistent en :

- la configuration complète des modèles
- leur apprentissage automatique
- l'évaluation des modèles

Le volet « Expérimentation » de l'application permet de mener ces expériences. Il est possible de configurer tous les détails de l'expérience, concernant les données (choix des variables, délais, prétraitements, post-traitements, jeux d'apprentissage, de validation, de test, etc.), concernant les modèles (architecture, algorithmes d'initialisation et d'apprentissage, opération particulière comme l'élagage, etc.). Une fois l'ensemble de ces points fixé, la configuration est validée et rejoint la liste des configurations prévues. Elles peuvent être sauvegardées sur le disque dur. L'interface de ce volet est présentée à la figure 7.5.

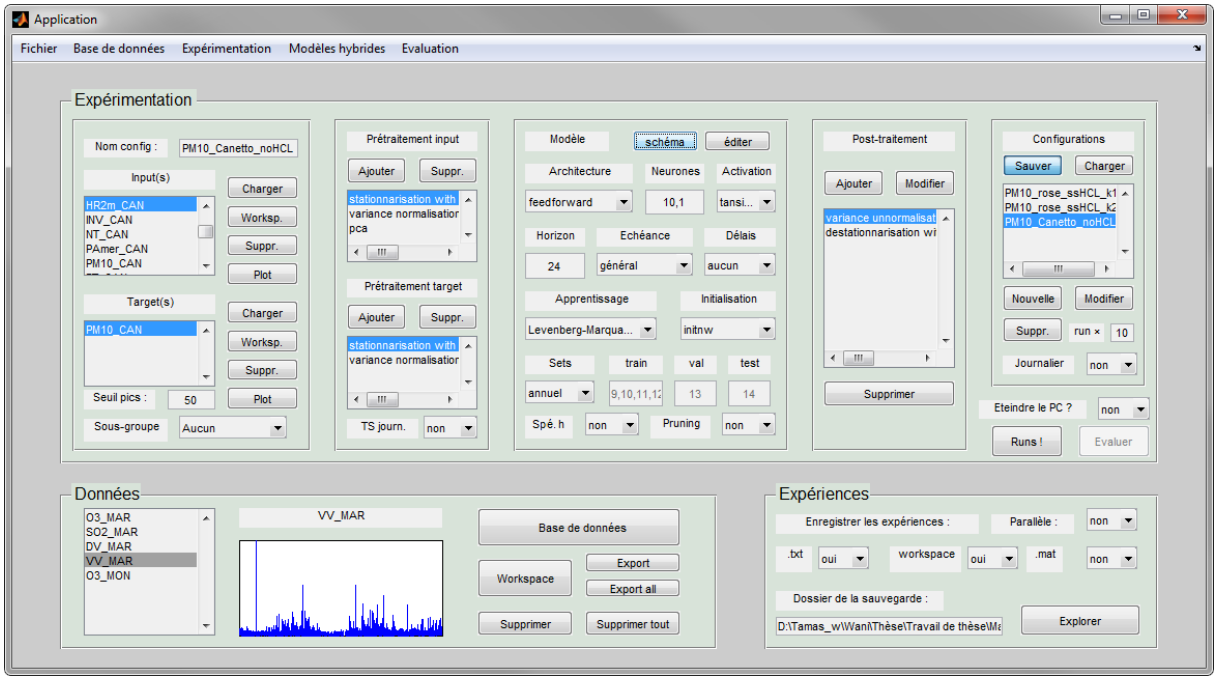


FIGURE 7.5 : Capture d'écran du volet « Expérimentation ».

On peut alors lancer l'ensemble des expériences ainsi pré-configurées, en répétant chacune autant de fois que nécessaire. Il est possible d'utiliser les périodes d'inactivité, comme les nuits, pour mener les calculs. L'entrainement d'un réseau de neurones étant relativement court sur un ordinateur portable classique (entre 30 secondes et une demi-heure, en fonction de la complexité du modèle), on peut ainsi mener le nombre requis d'expériences. Cela permet de palier le principal défaut des RNA, le besoin de fixer empiriquement un grand nombre de points de configuration en conservant une démarche rigoureuse. Près d'une dizaine de milliers de RNA ont ainsi été entrainés et évalués par Aria Base durant ces travaux de doctorat.

Les modèles hybrides présentés au chapitre 6 (page 148) demandent plus d'opérations, afin de faire fonctionner plusieurs PMC conjointement avec le modèle de classification. Un volet de l'application est dédié à leur construction. On peut créer une classification en fixant soi-même des règles de division des données, ou utiliser un modèle de classification non-supervisée (clustering) comme l'algorithme des k-means ou la classification ascendante hiérarchique.

L'évaluation de base (calcul des indices d'erreur) est effectuée automatiquement après chaque entrainement de RNA dans le volet « Expérimentation ». Nous avons également ajouté un volet « Evaluation », qui permet d'obtenir plus d'informations sur les performances des modèles en général (matrices de contingence, courbes ROC, etc.) mais surtout d'évaluer correctement les modèles hybrides (qui sont constitués de plusieurs PMC et utilisent un classifieur). Une capture d'image de ce volet est présenté en figure 7.6.

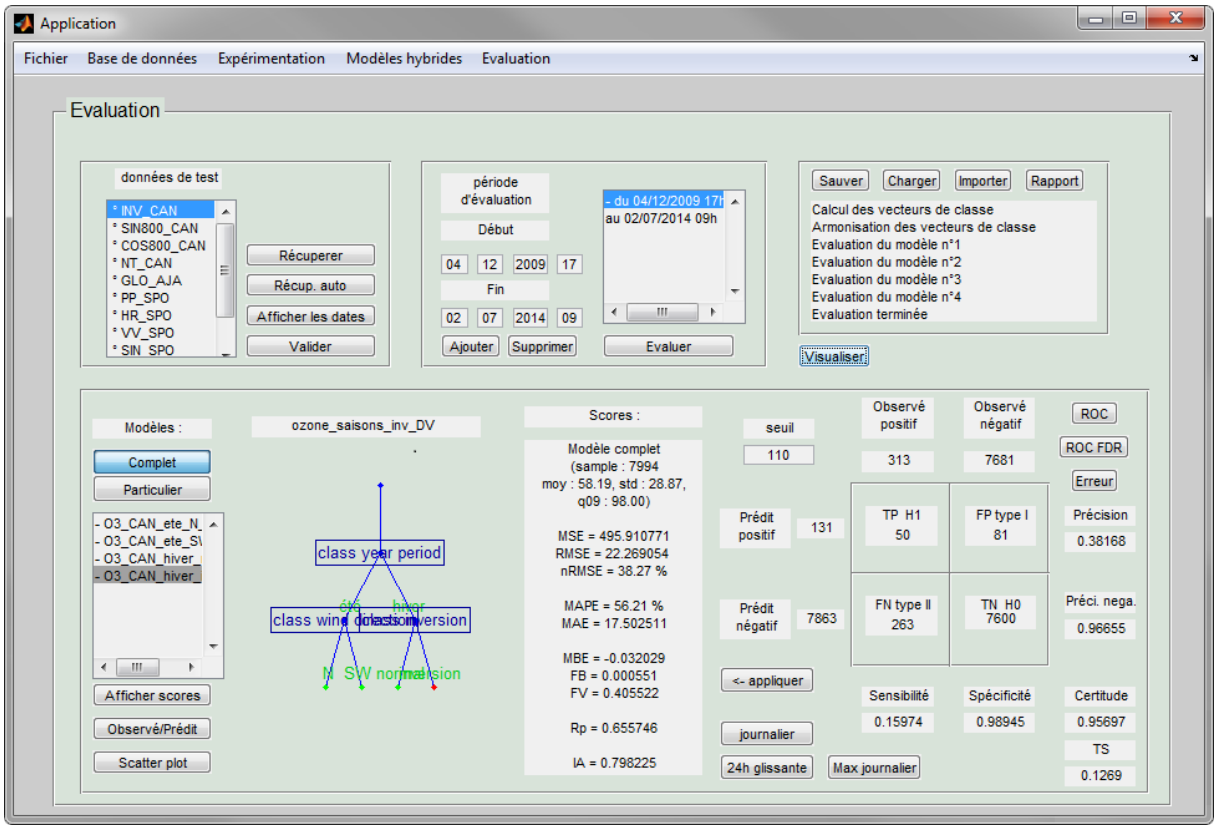


FIGURE 7.6 : Capture d'écran du volet « Evaluation ».

Les modèles hybrides bénéficient pleinement de l'usage des fonctions d'Aria Base. Des opérations comme le découpage en différents jeux de données (apprentissage, validation, test) parfois discontinus, avec cohérence entre les différents RNA constituant un modèle hybride seraient difficiles à mener sans gestion automatique. De plus, l'archivage de tous les points de configuration et des résultats obtenus à chaque expérience apporte une traçabilité et une rigueur au travail, importantes quand on souhaite inter-comparer des modèles prévisionnels.

L'ensemble des capacités de l'application Aria Base nous a permis de travailler avec des jeux de données renouvelés, et avec le nombre d'expériences élevé et la rigueur nécessaire à l'usage des PMC. L'export des modèles obtenus est possible, notamment vers Aria Web, application dédiée au fonctionnement opérationnel de ces modèles.

Nous allons à présent nous intéresser à cette application, dédiée à la prévision opérationnelle et destinée au personnel de l'AASQA.

7.2 Application d'aide à la décision « Aria Web »

Ce travail de doctorat a pour but d'améliorer la prévision de la qualité de l'air réalisée quotidiennement à Qualitair Corse. Jusqu'à présent, la prévision se fait à partir des prévisions fournies par des plates-formes gérées par d'autres organismes. Les prévisions de modèles comme Prév'air, AIRES ou Skyron (voir chapitre 2.2 page 22) librement accessibles sur internet sont consultées, ainsi que les prévisions météorologiques et les données de qualité de l'air du jour. Le prévisionniste réalise alors sa propre prévision empiriquement à partir de ces éléments.

Afin de simplifier ce travail, nous avons développé l'application Aria Web. Elle a pour but de faciliter la prévision, en réunissant de manière claire, au même endroit, toutes les informations pertinentes (mesures de Qualitair Corse et du réseaux des AASQA, prévisions, etc.). Elle doit également permettre l'analyse *a posteriori* d'épisodes de pollution en archivant ces informations. En effet, les résultats de prévisions fournis par des organismes externes ne sont que brièvement disponibles. Leur conservation est donc utile pour toute étude rétrospective, en vue par exemple d'analyser les raisons et configurations d'un pic de pollution.

Les fonctionnalités de cette application ont été définies en collaboration avec le personnel de Qualitair Corse chargé de la prévision, afin de correspondre au mieux aux besoins opérationnels. Nous avons voulu une application capable de :

- Faire fonctionner les modèles statistiques générés par Aria Base
- Regrouper les informations à consulter avant d'effectuer les prévisions
- Conserver les informations nécessaires aux études rétrospectives

L'application fait fonctionner de manière opérationnelle les RNA importés depuis l'application Aria Base. Elle doit également réunir de manière claire les informations disponibles importantes à prendre en compte : prévisions d'autres modèles disponibles sur internet, dernières données mesurées sur le parc de Qualitair Corse. Le prévisionniste a ainsi toutes les informations pour réaliser la prévision transmise au public et aux autorités. Ces fonctionnalités sont présentées à la figure 7.7.

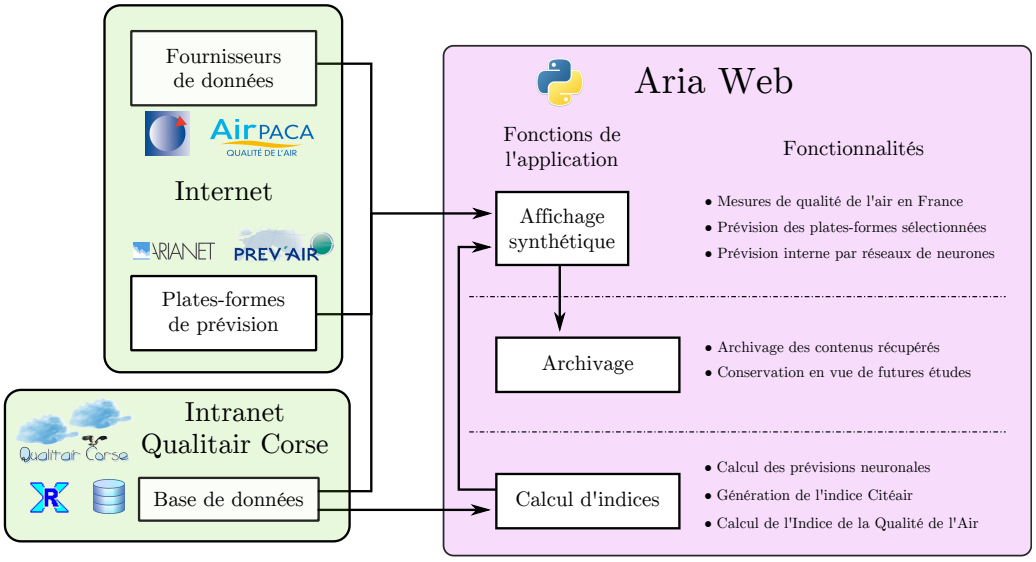


FIGURE 7.7 : Structure et fonctionnalités de l'application Aria Web.

Le développement de cette application web a été réalisé en collaboration avec Aurélia Balu, alors alternante dans le cadre du master Systèmes d'Information et Internet de l'université de

Corse. Elle a été développée avec des outils du monde des logiciels libres, dont la gratuité était une contrainte forte pour que l'application soit utilisable par l'AASQA.

L'application est codée dans le langage Python et utilise le framework web Django qui est bien documenté et OpenSource. Elle utilise par ailleurs la librairie pyair, développée par Lionel Roubeyrie à Limair, l'AASQA du Limousin. Cette librairie propose des outils permettant l'échange de données avec le logiciel XR qui gère les données mesurées par Qualitair Corse (Roubeyrie, 2013). L'application utilise en interne le système de gestion de base de données libre PostgreSQL.

Hébergé sur le serveur de Qualitair Corse, le contenu de l'application web est accessible depuis n'importe quel navigateur web, en interne de l'AASQA.

Nous allons présenter cette application web et ses capacités. Nous verrons d'abord ses fonctionnalités permettant de réaliser les prévisions à l'aide de RNA et des données nécessaires. Nous présenterons ensuite les possibilités de récupération et d'affichage des prévisions issues d'autres organismes et disponibles sur internet, ainsi que l'affichage des indices de qualité de l'air proposé avant de conclure sur l'intérêt et les perspectives de cette application.

7.2.1 Prévision à l'aide de RNA

L'application web permet l'exécution journalière automatique des RNA paramétrés et sélectionnés par le modélisateur via l'application Aria Base. Ces modèles fournissent les prévisions de concentration de PM10 et d'ozone. On peut utiliser des modèles simples autant que des modèles hybrides.

Les modèles choisis, importés depuis l'application Aria Base, comportent les indications permettant de les faire fonctionner, indiquant les informations suivantes :

- Variables d'entrée
- Prétraitements et post-traitements à appliquer et leurs paramètres
- Descriptions des couches de neurones (nombre de couches, connexions, nombre de neurones, fonctions d'activation)
- Poids et biais des neurones

Dans le cas de modèle hybride, c'est l'ensemble des RNA prédictifs qui sont importés, accompagnés du modèle de classification permettant de classer les données récupérées. Avant chaque prévision, l'application sélectionne grâce à ce classifieur le modèle prédictif à utiliser.

Aria Web récupère régulièrement et automatiquement les données d'entrées du modèle, qui sont mises à disposition par nos fournisseurs. Ces données sont récupérées depuis les serveurs FTP des organismes fournisseurs. Elles sont conservées le temps nécessaire pour effectuer les prévisions (durée paramétrable). Elles sont ensuite supprimées pour libérer l'espace disque.

Quand le modèle prédictif a été défini, l'application lance l'exécution des modèles prédictifs automatiquement et régulièrement avec ces données. Il est possible de lancer manuellement une prévision si l'on veut s'assurer de bénéficier de la prévision utilisant les dernières données récupérées. La figure 7.8 montre la page qui présente les dernières données mesurées et affiche les prévisions de modèle neuronal.

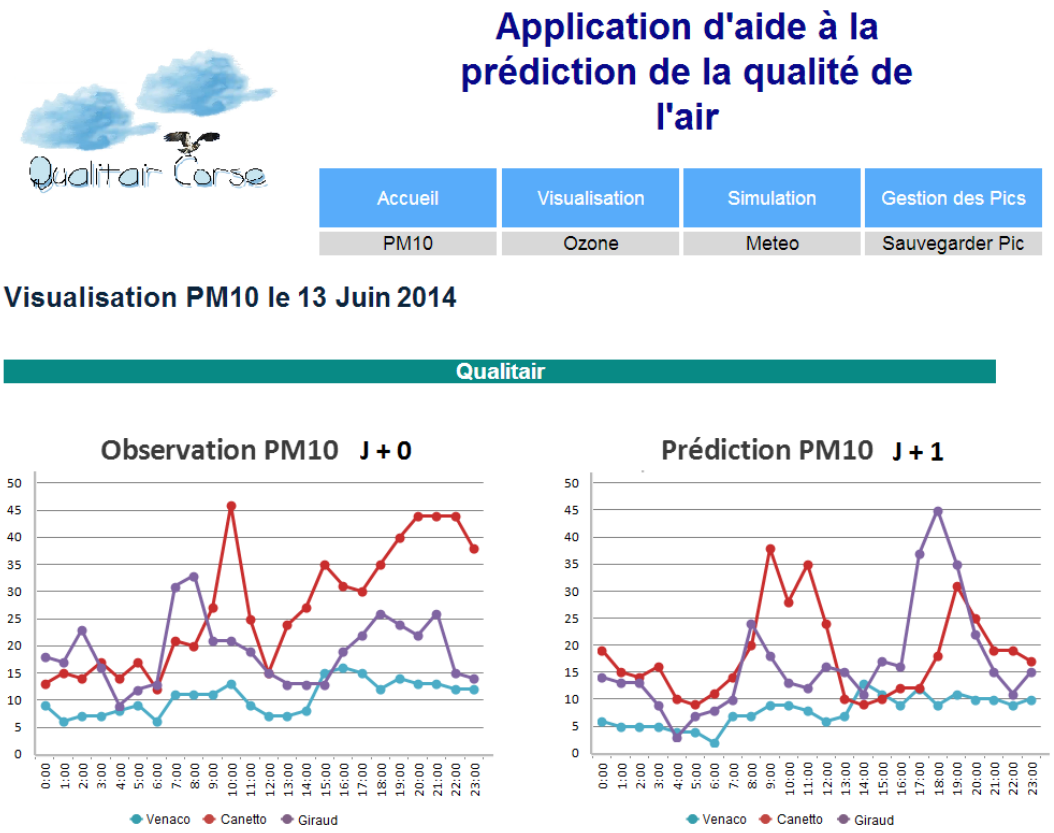


FIGURE 7.8 : Capture d'écran de la page de visualisation des prévisions statistiques.

7.2.2 Agrégation et visualisation de données

Les prévisionnistes de Qualitair Corse consultent également les cartographies de prévisions mises à disposition sur internet afin de bénéficier de toutes les informations disponibles. L'application regroupe au sein d'une même page les cartographies et indices proposés par d'autres structures.

L'application agrège les contenus et récupère les mises à jour dès qu'elles sont effectuées sur les sites émetteurs. On peut ainsi récupérer des images présentant des cartographies de pollution, ou des animations qui en montrent l'évolution (voir figure 7.9. Les images sont conservées un certain temps, puis supprimées pour libérer l'espace disque.

la liste des contenus récupérés est évolutive. Les utilisateurs peuvent entrer l'adresse URL (Uniform Resource Locator) de toute page présentant des cartes de prévision, ou plus généralement n'importe quel type d'image, pour que l'application les récupère.

Au delà de la prévision qui doit être réalisée quotidiennement, les AASQA réalisent des études rétrospectives, qui concernent ou non des épisodes de pollution en particulier. Les cartographies de prévisions d'autres organismes pouvant ne plus être disponibles sur internet, l'application permet de conserver certains contenus. Lors d'un épisode ou quelque temps après, l'utilisateur peut indiquer que toutes les données doivent être archivées. Les contenus téléchargés, les données des fournisseurs mais surtout les cartes agrégées montrant les prévisions sont alors toutes conservées sans limite de temps, jusqu'à ce que l'utilisateur indique que l'épisode est terminé.

L'application permet également de visualiser différents indices de pollutions (voir figure 7.10). Tout d'abord, elle calcule l'Indice de la Qualité de l'Air (IQA) à partir des données de Quali-

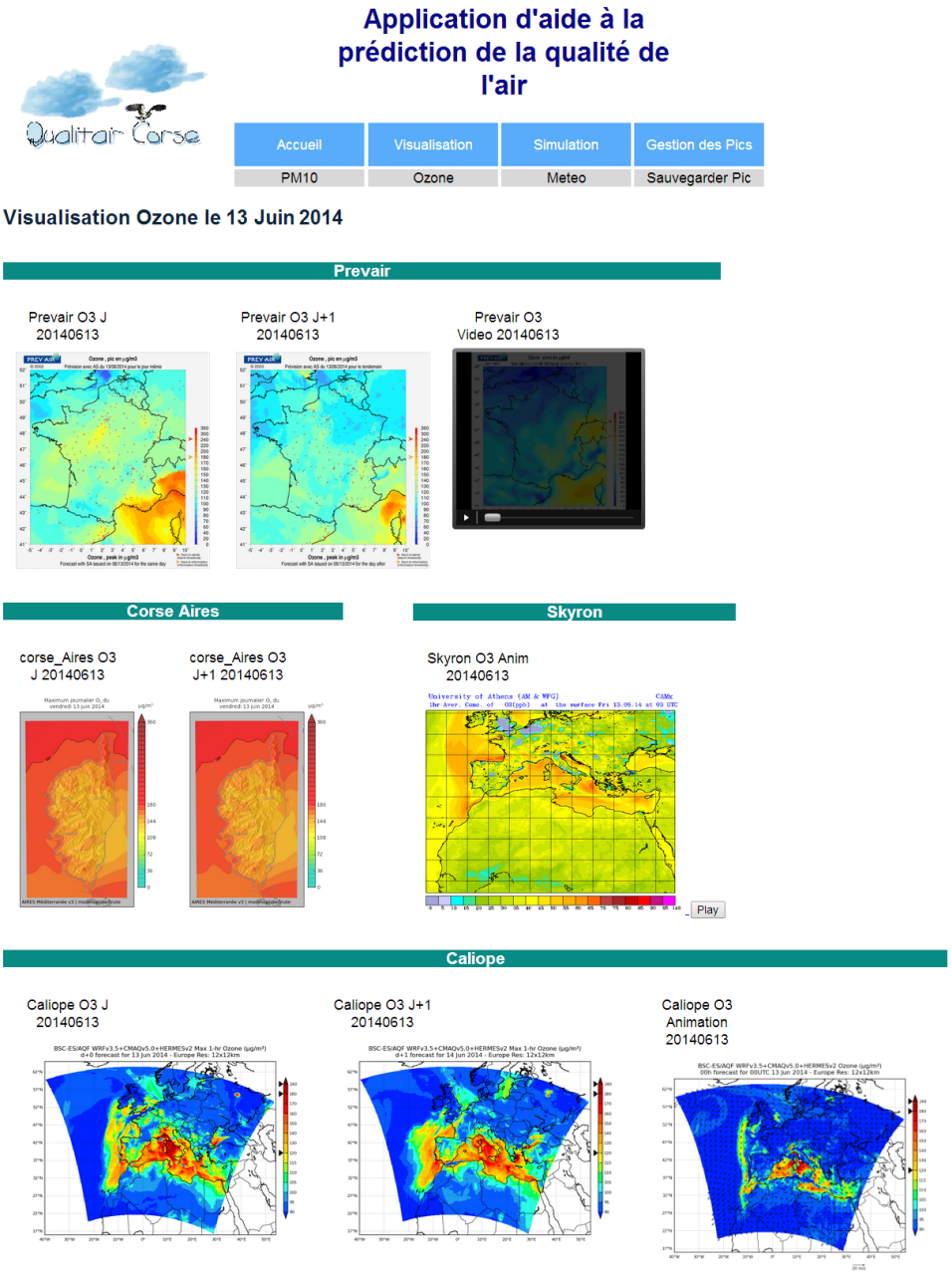


FIGURE 7.9 : Capture d'écran de la page d'agrégation de cartographies de prévisions.

tair Corse. L'indice européen Citeair (Common Information To European AIR) est également calculé. Cet indice a été développé afin de fournir une information simple, prenant en compte la pollution à proximité du trafic et qui soit comparable à travers l'Europe, donc compatible avec les méthodes de mesure de chaque réseau de surveillance de la qualité de l'air.

C'est un indice compris entre 0 pour une très bonne qualité de l'air et 100 pour une pollution très élevée. Il peut dépasser cette borne supérieure, auquel cas il est indiqué comme supérieur à 100. Son calcul se base sur le calcul de plusieurs sous-indices propres à plusieurs polluants. Ces sous-indices dépendent des concentrations de manière non-linéaire. Le calcul de l'indice Citeair est spécifié à l'annexe E. Le site internet www.airqualitynow.eu permet de consulter cet indice en temps réel sur toute l'Europe.

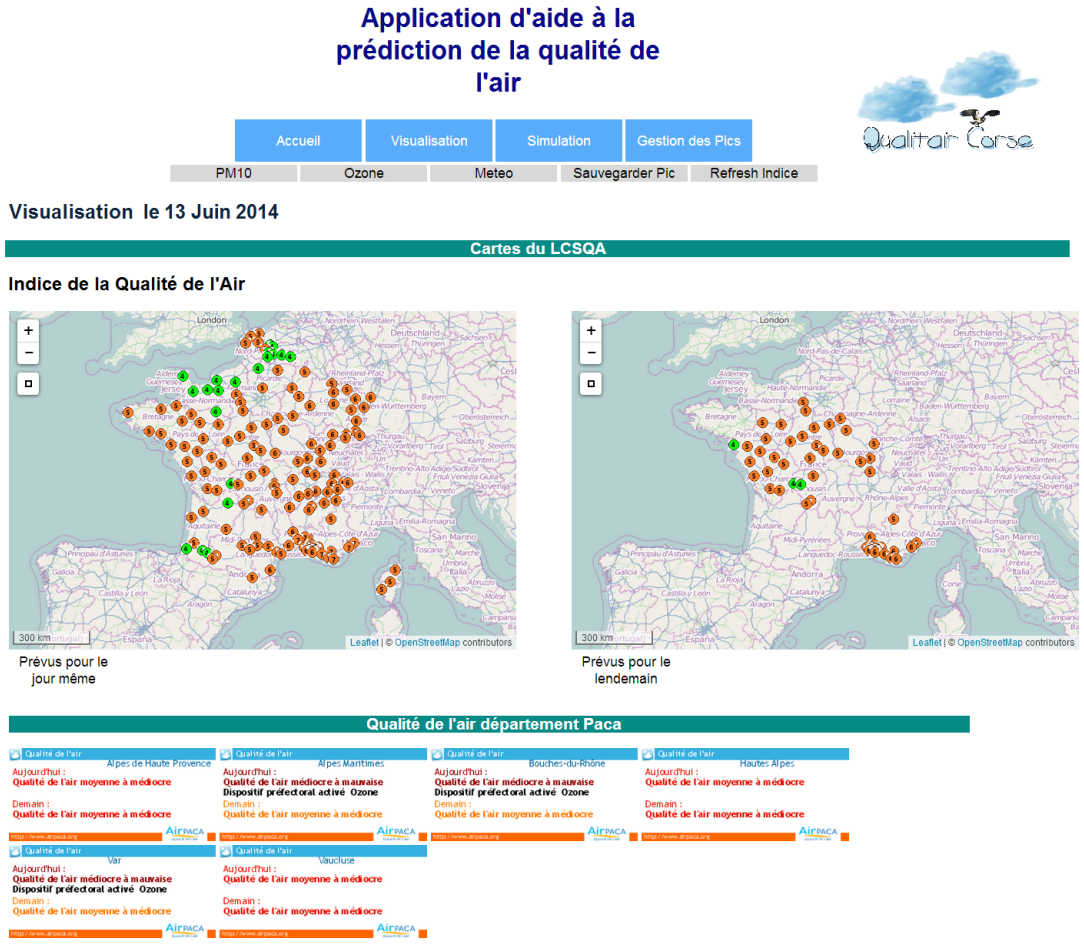


FIGURE 7.10 : Capture d'écran de la page de visualisation d'indices de qualité de l'air.

7.3 Conclusion

Afin de répondre à la problématique de la prévision de la qualité de l'air en Corse, nous avons travaillé sur les modèles neuronaux et en particulier le PMC, et proposé une méthodologie de modèle hybride utilisant également des classifieurs. Pour mener nos travaux, nous avons développé l'application Aria Base sous Matlab. Elle permet également d'utiliser nos modèles de manière opérationnelle, conjointement avec l'application d'aide à la décision Aria Web.

Aria Base utilise la NNToolbox afin de faciliter l'expérimentation à l'aide de RNA pour la prévision de la qualité de l'air, mais aussi pour d'autres domaines pour lesquels on pourrait utiliser des RNA. Elle permet tout d'abord de gérer et faire évoluer les jeux de données que nous utilisons. Elle permet également de mener les expérimentations à l'aide de RNA de manière simple en laissant tout choix de configuration possible. La planification d'un ensemble d'expériences utilisant plusieurs configurations prédéfinies est particulièrement utile pour entrainer de larges ensembles de réseaux de neurones, permettant d'identifier les meilleures configurations possibles. Enfin, la gestion des modèles hybrides et de leur évaluation nous a permis d'élaborer plusieurs types de modèles hybrides et de proposer une méthodologie pour améliorer les taux de détection des pics de pollution, particulièrement ardus à prévoir.

La poursuite d'un « meilleur modèle » possédant la « meilleure configuration » pour prévoir l'évolution des polluants ne nous est pas apparue être la meilleure stratégie, étant donné la nature de notre problème. Un modèle statistique est un outil puissant mais ne peux prétendre

suffire seul à une telle problématique. Les sites de mesure sont différents les uns des autres, les polluants ont des dynamiques différentes, nos jeux de données évoluent et peuvent mettre en valeurs de nouvelles interactions entre les variables. Les meilleurs modèles en termes d'indice d'erreur type Root Mean Squared Error (RMSE) sur l'ensemble du jeu de test ne seront pas forcément les meilleurs modèles pour la détection des pics. De plus, l'étude de l'état de l'art de la prévision statistique de la qualité de l'air a révélé que plutôt que de converger vers un modèle optimal, une multitude de solutions existait pour optimiser les résultats de prévisions statistiques.

Cette application fait partie de notre réponse au problème de la prévision des pics de pollution en Corse. Un outil évolutif permettant d'identifier expérimentalement la meilleure configuration dans un contexte donné. Grâce à lui, nous avons eu l'occasion d'entrainer près de dix-mille RNA lors de ces travaux de doctorat. Mais il ne permet pas seulement de travailler de manière empirique. Il permet également les investigations nécessaires à la découverte de solutions génériques, permettant l'amélioration des prévisions quel que soit le contexte, en faisant un véritable outil de recherche.

Beaucoup de perspectives de développement se présentent à ce prototype. Tout d'abord, sa mise à disposition sur internet serait intéressante, pour permettre à d'autres organismes de travailler sur la prévision avec ces outils. Une seconde perspective serait l'ajout de plusieurs autres types de modèles à apprentissage à la page de gestion des modèles, ainsi que les prétraitements leur étant nécessaires. Cela offrirait une plus grande diversité d'expérimentations. D'autres opérations que nous avons menées, comme la sélection de variables à l'aide de métaheuristique, pourraient également y être implémentées. En temps qu'utilisateurs de cette application, nous avons également eu l'occasion d'identifier quelques améliorations mineures à apporter. Développée sous Matlab 2012a, cette application pourrait être remaniée pour bénéficier des fonctionnalités des dernière versions du logiciel (l'actuelle étant la 2015a). La possibilité d'utiliser le GPU (processeur graphique) de l'ordinateur pour accélérer le calcul pourrait être envisagée. Au niveau des performances, une optimisation de la mémoire vive utilisée lors des expériences nous semble également possible et bénéfique.

L'application « Aria Web » répond elle aux besoins opérationnels de prévision de Qualitair Corse. Elle permet de faire fonctionner automatiquement les modèles entrainés par Aria Base. Toutes les données nécessaires sont automatiquement récupérées de la base de données de Qualitair Corse, ou sur les serveurs d'autres fournisseurs. Les prévisions peuvent être réalisées de manière à être disponibles selon les spécificités de l'arrêté ministériel du 26 mars 2014 relatif au déclenchement de procédures préfectorales en cas d'épisodes de pollution de l'air ambiant (présenté en annexe A). En cas d'évolution de la législature, elle est facilement adaptable.

Elle propose également un accès facilité aux résultats de prévisions d'autres plates-formes, disponibles sur internet. Elle agrège et calcule également d'autres informations, comme les indices de qualité de l'air émis ou prévus par d'autres AASQA. La consultation facilitée de l'ensemble de ces informations permet aux prévisionnistes d'une AASQA de disposer du plus d'informations possible au moment d'émettre les prévisions.

Au-delà de la prévision, l'agrégation de ces contenus permet de fournir des éléments utiles lors d'études rétrospectives, fréquemment effectuées au sein des AASQA. Ainsi l'application Web archive les contenus jugés représentatifs par le prévisionniste qui peut ainsi se constituer au fil du temps une banque de données riche et pertinente.

Plusieurs possibilités de développements ultérieurs existent pour cette application web. Premièrement, il serait intéressant de permettre le support d'autres types de modèles statistiques que les RNA. Ce type d'évolution pourrait être pensé de pair avec l'application Matlab qui four-

nit les modèles prêts à l'utilisation. Deuxièmement, un accès contrôlé au contenu de l'application depuis internet (voire depuis une application mobile) pourrait venir remplacer le seul accès possible depuis le réseau interne de l'AASQA. Cette problématique requiert les compétences de sécurité informatique appropriées, ce qui nous a poussés dans ces travaux de doctorat à rester sur un simple accès interne, en intranet. Mais un accès sécurisé à distance pourrait apporter de la facilité d'utilisation, notamment lors des périodes d'astreinte hors du lieu de travail. Enfin, une troisième perspective pourrait consister en une mise à disposition de l'outil, ou de certaines de ses capacités, au public.

Conclusion générale

La prévision de la qualité de l'air est une problématique importante, car l'anticipation des pics de pollution permet d'en limiter l'impact en prenant à temps des mesures de restriction d'émission de polluants. La limitation de l'ampleur de ces pics préserve la santé de la population, qui n'a d'autre choix que celui de respirer l'air auquel elle est exposée.

De nombreux modèles existent, qui suivent des paradigmes différents. Ce sont les modèles statistiques à apprentissage automatique qui sont ressortis comme étant les plus adaptés pour améliorer la prévision réalisée à Qualitair Corse, association chargée de la surveillance de la qualité de l'air sur l'île. Les PMC, utilisés correctement, donnent des résultats qui surpassent les modèles actuellement disponibles sur la Corse.

Une fois configurés, ces modèles se paramétrisent automatiquement lors d'un apprentissage supervisé. Les données d'apprentissage leurs servent d'exemples, à partir desquels ils tentent de représenter les relations sous-jacentes existantes entre les variables par une modification des poids et des biais que comportent leurs neurones. Si on a pris soin d'éviter le sur-apprentissage, qui a lieu si l'entrainement est trop poussé et que le modèle apprend « par cœur » les données d'apprentissage, on obtient un modèle capable de généraliser les relations observées avec de nouvelles données. On peut ainsi s'en servir de modèle prédictif.

Mais l'optimisation de tels modèles est délicate et fastidieuse. Il existe plusieurs stratégies d'apprentissage de ces réseaux de neurones, plusieurs architectures possibles. La sélection de variables doit être soignée, et les prétraitements possibles pour les données d'entrées sont nombreux.

C'est un travail largement interdisciplinaire qui nous a permis d'apporter une réponse à cette problématique. Nous avons exploré un grand nombre de pistes permettant d'améliorer les performances de base des prévisions effectuées par les PMC. Nous avons appliqué des métaheuristiques à la sélection de variables pour optimiser les jeux de données utilisés. Nous avons investigué des méthodes d'élagage afin de limiter la complexité des réseaux de neurones et respecter le principe de parcimonie. Et surtout, nous nous sommes donné les moyens de mener le grand nombre d'expériences nécessaires à l'isolation de la meilleure configuration possible, en fonction des données disponibles et suivant les performances des modèles. Les méthodes que nous avons utilisées sont issues du domaine de l'apprentissage automatique (machine learning), de l'optimisation, de la statistique, mais aussi des sciences atmosphériques.

Parmi les plus intéressantes pistes que nous avons identifiées, l'utilisation de modèles hybrides semble prometteuse. Cette méthodologie consiste en l'utilisation d'un classifieur pour identifier un régime météorologique à venir et d'utiliser un PMC, spécialiste de ce régime, pour effectuer la prévision. L'avantage de cette méthode est que l'on travaille avec des modèles prédictifs dont les paramètres (poids et biais) sont fixés pour représenter les interactions entre les variables typiques de chaque régime identifié. L'inconvénient est que les jeux de données sont subdivisés, limitant le nombre de données disponibles pour chaque modèle et risquant de favoriser le surapprentissage. Il

est apparu que la méthode développée a donné des résultats intéressants, réussissant à améliorer la détection des rares fortes concentrations présentes dans les jeux de données mesurés en Corse.

Ces modèles statistiques n'ont pas vocation à n'être utilisés qu'avec des données mesurées. Nous avons le plus souvent utilisé des sorties du modèle AROME de Météo-France en entrée, et l'avons également couplé avec AIRES. Cette plate-forme de prévision provenant d'Air PACA est basée sur le CTM CHIMERE et produit des prévisions sur le Languedoc-Roussillon, la PACA et la Corse. Des études conjointes avec Air PACA ont montré que nos modèles pouvaient être utilisés efficacement pour assurer l'assimilation statistique des données mesurées sur la PACA. Le PMC a su s'adapter à un contexte de prévision différent (prévision sur toute une région avec de très nombreuses variables disponibles). Pour la prévision des particules, les données issues d'AIRES améliorent également les prévisions en Corse, et on s'attend à d'autres progrès liés à la prochaine livraison d'un cadastre corse des émissions qui manque encore pour l'instant à AIRES.

Plutôt que d'avancer un modèle comme étant le meilleur, une configuration comme étant la plus optimisée, nous avons pris acte de la relativité de ce type de solution. La configuration optimale n'est pas la même pour chaque polluant, ni pour chaque station de mesure. Même à un endroit donné, la meilleure configuration change quand plus de données deviennent disponibles, ce qui améliore l'apprentissage et rend certaines options plus intéressantes qu'elles ne l'étaient avec moins de données. Toute nouvelle série temporelle décrivant un phénomène en lien avec la qualité de l'air peut être envisagée pour rejoindre les données d'entrées. La liste de ce qui peut influencer la configuration des PMC est encore longue.

Plus qu'un modèle en particulier, c'est une démarche que nous proposons comme réponse à la problématique de la prévision des pics de pollution en Corse. Une démarche qui a l'avantage d'être adaptable à l'évolution de la situation (nouvelles stations de mesure, évolution de la réglementation, agrandissement des jeux de données, etc.) et dont nous fournissons les outils nécessaires à la mise en place.

L'application Aria Base permet de mener aisément les expériences d'optimisation de la configuration des PMC pour la prévision d'un polluant. Elle permet de maintenir à jour ses jeux de données, de mener les expériences et de les archiver. Les méthodes qui ont été présentées dans ces travaux sont implémentées au sein de cette application, ce qui facilite leur réutilisation future. Les modèles qui sont entrainés et évalués peuvent être exportés et utilisés de manière opérationnelle.

Pour ce faire, l'application d'aide à la décision Aria Web a été développée. Elle réalise la prévision statistique grâce à la récupération automatique des données nécessaires, mais permet aussi d'agréger les résultats des modèles de prévision d'autres organismes. Elle fournit ainsi les éléments nécessaires à la réalisation quotidienne des prévisions.

Ces deux applications fournissent à Qualitair Corse les outils nécessaires à la réalisation des prévisions qui font partie de ses missions. La première est à disposition d'un modélisateur statisticien qui pourra entretenir les modèles et en créer des nouveaux en fonction des besoins de l'AASQA. Il lui sera également possible de continuer les investigations pour améliorer les performances obtenues.

La seconde est à destination des prévisionnistes de l'AASQA, qui ne sont pas forcément experts en modèles statistiques. Les prévisions statistiques ainsi que les prévisions récupérées automatiquement sur internet leur apportent tous les éléments dont ils ont besoin pour réaliser la prévision qui sera ensuite diffusée au public et transmise aux autorités.

Ces applications permettent en particulier de s'adapter à l'arrêté du 26 mars 2014 relatif

au déclenchement de procédures préfectorales en cas d'épisode de pollution de l'air ambiant. La circulaire d'application de cet arrêté demande que les prévisions soient réalisées à midi heure locale pour le reste de la journée et pour le lendemain. Ceci est possible en créant les PMC appropriés avec Aria Base et en les utilisant avec Aria Web.

Les travaux que nous avons menés ouvrent plusieurs perspectives, qui concernent trois aspects distincts de ces travaux. Pour continuer à améliorer la qualité de la prévision, on pourra tout d'abord s'intéresser aux données utilisées par les modèles. D'autres perspectives concernent l'amélioration des modèles eux-mêmes. Enfin, l'amélioration des applications développées est également envisagée.

Voyons tout d'abord les perspectives concernant les données. On a vu l'intérêt des PMC (mais aussi d'autres modèles comme les FA) pour assurer l'assimilation statistique qui améliore les prévisions des CTM. On pourrait élargir l'origine des données utilisées à plusieurs CTM et modèles météorologiques (AROME, Skyron, Prév'air, Polyphemus, etc.) afin d'adopter une méthodologie de prévision d'ensemble. Ce type d'approche permettait d'allier les qualités des différents modèles déterministes et d'améliorer la robustesse des prévisions.

Toujours concernant les données utilisées, on pourra envisager l'usage des séries temporelles qui sont encore trop courtes, ce qui pénalise l'apprentissage, mais qui pourront atteindre une taille suffisamment importante, améliorant ainsi les performances de prévision. La hauteur de la couche limite issue du modèle AROME en est un exemple, les mesures de la station rurale de Venaco ou du Centre d'Observation Régional pour la Surveillance du Climat et de l'environnement Atmosphérique et océanographique en Méditerranée occidentale (CORSiCA) dans le Cap Corse en sont un autre. Les mesures de la station rurale et du site du Cap Corse permettraient de fournir aux PMC les informations permettant d'identifier les épisodes de transport, n'étant que très peu influencées par les sources locales de polluants.

L'usage de données autres que des mesures de polluants ou des mesures météorologiques, comme des mesures de trafic routier ou des données de consommation électrique, peut également apporter de l'information concernant cette fois les émissions locales pour parfaire la prise en compte de ces sources dans la prévision des concentrations. Enfin, l'usage d'observations satellites, après traitement pour en constituer des séries temporelles représentatives des niveaux de polluants près du sol, peut également être envisagé.

Intéressons-nous aux perspectives concernant les modèles prévisionnels. Concernant les PMC, on a vu que des métaheuristiques, notamment le recuit simulé grâce à son approche « wrapper », étaient utiles pour la problématique de la sélection de variables. Nous avons ensuite préféré l'ACP qui permet de facilement diminuer le volume de variables, tout en les transformant de manière à réduire la redondance entre elles. Mais le recuit simulé reste une puissante méthode d'optimisation qui pourrait être appliquée à d'autres aspects de la configuration des PMC.

En plus de la sélection de variables, les recuits simulés pourraient par exemple gérer l'architecture du réseau (choix du nombre de couches cachées et de neurones). Dans le cadre de la création de modèles hybrides, la métaheuristique pourrait servir à isoler la meilleure technique de clustering, et le nombre de classes à utiliser, pour améliorer la détection des pics de pollution. Des ressources de calcul plus conséquentes que celles dont nous avons bénéficié pour ces travaux (ordinateur portable classique) seraient alors nécessaires.

Toujours concernant les modèles, l'utilisation d'alternatives aux PMC serait intéressante. Les FA ont montré de bons résultats, notamment par leurs capacités à utiliser en entrée un jeu de données constitué d'un grand nombre de séries temporelles, éventuellement qualitatives. Leur usage peut devenir intéressant dans le cadre de la récupération de nombreuses nouvelles variables, notamment issues de CTM.

D'autres modèles comme les Support Vector Machine (SVM) semblent également prometteurs à appliquer à la prévision de la qualité de l'air. Bien que ces modèles soient souvent utilisés pour des problèmes de classification/régression, on note peu d'applications à la prévision de la qualité de l'air. Ils demandent également un travail de configuration mais pourraient apporter des résultats intéressants.

Enfin, des perspectives d'évolution des deux applications développées existent également. L'application Aria Base tout d'abord, n'est pour l'instant qu'un prototype, outil de travail dont nous avons été les seuls utilisateurs. Après ces travaux, il serait intéressant de passer à une version plus aboutie, à l'installation aisée sur d'autres ordinateurs et à l'ergonomie plus travaillée. Ses performances de calcul peuvent également être améliorées, par exemple en optimisant la gestion de la mémoire vive ou en utilisant le GPU de l'ordinateur. Au niveau des expérimentations, de nouveaux modèles comme ceux que l'on vient de citer (FA, SVM) peuvent facilement être ajoutés aux modèles utilisables par Aria Base. Aria Base pourrait alors être diffusée et utilisée dans de nombreux domaines.

L'application Aria Web possède elle aussi des pistes d'améliorations. Elle devra tout d'abord évoluer avec Aria Base si des fonctionnalités sont ajoutées à cette dernière. La mise à disposition d'une partie de ses fonctionnalités au public, depuis internet, serait un projet intéressant. Le public pourrait ainsi suivre au mieux l'évolution de la qualité de l'air mesurée sur le parc de Qualitair Corse, ainsi que les prévisions effectuées. Ce pourrait être un outils de communication efficace pour l'association. Enfin, sa mise à disposition à d'autres AASQA permettrait de faciliter le travail de prévision, notamment pour les petites structures n'ayant pas les ressources pour faire fonctionner une plate-forme de prévision utilisant un CTM.

Bibliographie

Abdul-Wahab S.A. et Al-Alawi S.M. 2002. "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks." *Environmental Modelling & Software* 17 (3), pages 219 – 228.

Académie des sciences. 1993. *Ozone et propriétés oxydantes de la troposphère : essai d'évaluation scientifique*. Paris : Technique et Documentation.

Agirre-Basurko E., Ibarra-Berastegi G. et Madariaga I. 2006. "Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area." *Environmental Modelling & Software* 21, pages 430 – 446.

Beccali M., Cellura M., Lo Brano V. et Marvuglia A. 2004. "Forecasting daily urban electric load profiles using artificial neural networks." *Energy Conversion and Management* 45 (18 - 19), pages 2879 – 2900.

Bell M.L., Davis D.L. et Fletcher T. 2004. "A Retrospective Assessment of Mortality from the London Smog Episode of 1952 : The Role of Influenza and Pollution." *Environmental Health Perspectives* 112 (1), pages 6 – 8.

Ben-Ameur W. 2004. "Computing the Initial Temperature of Simulated Annealing." *Computational Optimization and Applications* 29 (3), pages 369 – 385.

Bertsimas D. et Tsitsiklis J. 1993. "Simulated Annealing." *Statistical Science* 8 (1), pages 10 – 15.

Bessagnet B., Hodzic A., Vautard R., Beekmann M., Cheinet S., Honoré C., Liousse C. et Rouil L. 2004. "Aerosol modeling with CHIMERE—preliminary evaluation at the continental scale." *Atmospheric Environment* 38 (18), pages 2803 – 2817.

Bessagnet B., Menut L., Aymoz G., Chepfer H. et Vautard R. 2008. "Modeling dust emissions and transport within Europe : The Ukraine March 2007 event." *Journal of Geophysical Research* 113, pages D15202.

Bouttier F. 2007. "Arome, avenir de la prévision régionale." *La Météorologie* 8 (58), pages 12 – 20.

Box G.E.P., Jenkins G.M. et Reinsel G.C. 1994. *Time series analysis, forecasting and control*. Englewood Cliffs, N.J. : Prentice Hall.

Breiman L. 1996. "Bagging predictors." *Machine Learning* 24 (2), pages 123 – 140.

Breiman L. 2001*a*. "Random Forests." *Machine Learning* 45 (1), pages 5 – 32.

Breiman L. 2001*b*. “Statistical Modeling : The Two Cultures.” *Statistical Science* 16 (3), pages 199 – 231.

Broyden C.G. 1970. “The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations.” *IMA Journal of Applied Mathematics* 6 (1), pages 76 – 90.

Brunelli U., Piazza V., Pignato L., Sorbello F. et Vitabile S. 2007. “Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy.” *Atmospheric Environment* 41 (14), pages 2967 – 2995.

Cai M., Yin Y. et Xie M. 2009. “Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach.” *Transportation Research Part D: Transport and Environment* 14 (1), pages 32 – 41.

Carnevale C., Finzi G., Pisoni E., Singh V. et Volta M. 2011. “An integrated air quality forecast system for a metropolitan area.” *Journal of Environmental Monitoring* 13 (12), pages 3437 – 3447.

CBA CAFE. 2005. Baseline Analysis 2000 to 2020. Rapport technique, .

Cerny V. 1985. “Thermodynamical Approach to the Traveling Salesman Problem : An Efficient Simulation Algorithm.” *Journal of Optimization Theory and Applications* 45 (1), pages 41 – 51.

Chelani A. et Devotta S. 2006. “Air quality forecasting using a hybrid autoregressive and nonlinear model.” *Atmospheric Environment* 40 (10), pages 1774 – 1780.

Cherif A., Cardot H. et Boné R. 2011. “SOM time series clustering and prediction with recurrent neural networks.” *Neurocomputing* 74 (11), pages 1936 – 1944.

Cherif A., Cardot H. et Boné R. 2013. Hierarchical Clustering for Local Time Series Forecasting. Dans *Neural Information Processing*, ed. M. Lee, A. Hirose, Z.-G. Hou et R. M. Kil. Vol. 8227 Berlin, Heidelberg : Springer Berlin Heidelberg pages 59 – 66.

Coman A., Ionescu A. et Candau Y. 2008. “Hourly ozone prediction for a 24-h horizon using neural networks.” *Environmental Modelling & Software* 23 (12), pages 1407 – 1421.

Comrie A.C. 1997. “Comparing Neural Networks and Regression Models for Ozone Forecasting.” *Journal of the Air & Waste Management Association* 47 (6), pages 653 – 663.

Corani G. 2005. “Air quality prediction in Milan : feed-forward neural networks, pruned neural networks and lazy learning.” *Ecological Modelling* 185 (2 - 4), pages 513 – 529.

Cornuéjols A., Kodratoff Y. et Miclet L. 2002. *Apprentissage artificiel : concepts et algorithmes*. Paris : Eyrolles.

Cortes C. et Vapnik V. 1995. “Support-vector networks.” *Machine Learning* 20 (3), pages 273 – 297.

Davies D.L. et Bouldin D.W. 1979. “A cluster separation measure.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (2), pages 224 – 227.

Díaz-Robles L. A., Ortega J. C., Fu J. S., Reed G. D., Chow J. C., Watson J. G. et Moncada-Herrera J. A. 2008. “A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas : The case of Temuco, Chile.” *Atmospheric Environment* 42 (35), pages 8331 – 8340.

de Mattos Neto P.S.G., Madeiro F., Ferreira T.A.E. et Cavalcanti G.D.C. 2014. "Hybrid intelligent system for air quality forecasting using phase adjustment." *Engineering Applications of Artificial Intelligence* 32, pages 185 – 191.

Debry E. et Mallet V. 2014. "Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM10 on the Prev'Air platform." *Atmospheric Environment* 91, pages 71 – 84.

Di Biagio C., Doppler L., Gaimoz C., Grand N., Ancellet G., Raut J.-C., Beekmann M., Borbon A., Sartelet K., Attié J.-L., Ravetta F. et Formenti P. 2015. "Continental pollution in the Western Mediterranean Basin : vertical profiles of aerosol and trace gases measured over the sea during TRAQA 2012 and SAFMED 2013." *Atmospheric Chemistry and Physics Discussions* 15 (6), pages 8283 – 8328.

DiCiccio T.J. et Efron B. 1996. "Bootstrap confidence intervals." *Statistical Science* 11 (3), pages 189 – 212.

Ding C. et He X. 2004. "K-means Clustering via Principal Component Analysis." Dans *21st International Conference on Machine Learning*. Banff, Canada : Russ Greiner et Dale Schuurmans.

Dreyfus G. 2004. *Réseaux de neurones : méthodologie et applications*. Paris : Eyrolles.

Dreyfus G. 2008. *Apprentissage statistique : réseaux de neurones, cartes topologiques, machines à vecteurs support*. Paris : Eyrolles.

Dréo J., Pétrowsky A. et Siarry P. 2003. *Métaheuristiques pour l'optimisation difficile*. Paris : Eyrolles.

Dulac F., Sciare J., Nicolas J.B., Petit E., Hamonou E., Ramonet M., Roberts G., Bourrianne T., Mallet M., Pont V., Lambert D. et Léon J.-F. 2013. "The new Mediterranean background monitoring station of Ersa, Cape Corsica : A long term Observatory component of the Chemistry-Aerosol Mediterranean Experiment (ChArMEx)." *Geophysical Research Abstracts* 15 (EGU2013-12108).

Dutot A., Rynkiewicz J., Steiner F. et Rude J. 2007. "A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions." *Environmental Modelling & Software* 22 (9), pages 1261 – 1269.

Englert N. 2004. "Fine particles and human health—a review of epidemiological studies." *Toxicology Letters* 149 (1 - 3), pages 235 – 242.

Fahlman S.E. et Lebiere C. 1990. "The Cascade-Correlation Learning Architecture." Dans *Advances in Neural Information Processing Systems II*. San Mateo : D.S. Touretzky pages 524 – 532.

Fan S., Liao J.R., Yokoyama R., Chen L. et Lee W.-J. 2009. "Forecasting the Wind Generation Using a Two-Stage Network Based on Meteorological Information." *IEEE Transactions on Energy Conversion* 24 (2), pages 474 – 482.

Fawcett T. 2006. "An introduction to ROC analysis." *Pattern Recognition Letters* 27 (8), pages 861 – 874.

Fernando H.J.S., Mammarella M.C., Grandoni G., Fedele P., Di Marco R., Dimitrova R. et Hyde P. 2012. "Forecasting PM10 in metropolitan areas : Efficacy of neural networks." *Environmental Pollution* 163, pages 62 – 67.

Fletcher R. 1970. "A new approach to variable metric algorithms." *The Computer Journal* 13 (3), pages 317 – 322.

Forgy E.W. 1965. "Cluster analysis of multivariate data : efficiency versus interpretability of classifications." *Biometrics* 21, pages 768 – 769.

Fort J.-C., Cottrell M. et Letremy P. 2001. "Stochastic on-line algorithm versus batch algorithm for quantization and self organizing maps." Dans *Neural networks for signal processing XI*. North Falmouth, MA : IEEE pages 43 – 52.

Frénay B., Doquire G. et Verleysen M. 2013. "Is mutual information adequate for feature selection in regression ?" *Neural Networks* 48, pages 1 – 7.

Fromage A. et Gilibert E. 1997. "Prévision des épisodes d'ozone : état de l'art dans le monde." *Pollution Atmosphérique* 39 (154), pages 52 – 59.

Gardner M.W. et Dorling S.R. 1998. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." *Atmospheric Environment* 32 (14 - 15), pages 2627 – 2636.

Gardner M.W. et Dorling S.R. 1999. "Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London." *Atmospheric Environment* 33 (5), pages 709 – 719.

Gardner M.W. et Dorling S.R. 2000. "Statistical surface ozone models : an improved methodology to account for non-linear behaviour." *Atmospheric Environment* 34, pages 21 – 34.

Gautam A.K., Chelani A.B., Jain V.K. et Devotta S. 2008. "A new scheme to predict chaotic time series of air pollutant concentrations using artificial neural network and nearest neighbor searching." *Atmospheric Environment* 42 (18), pages 4409 – 4417.

Goia A., May C. et Fusai G. 2010. "Functional clustering and linear regression for peak load forecasting." *International Journal of Forecasting* 26 (4), pages 700 – 711.

Goldfarb D. 1970. "A family of variable-metric methods derived by variational means." *Mathematics of Computation* 24 (109), pages 23 – 26.

Goyal P., Chan A.T. et Jaiswal N. 2006. "Statistical models for the prediction of respirable suspended particulate matter in urban cities." *Atmospheric Environment* 40 (11), pages 2068 – 2077.

Grün B. et Leisch F. 2007. "Fitting finite Mixtures of Generalized Linear Regressions in R." *Computational Statistics & Data Analysis* 51, pages 5247 – 5252.

Hao J. 2005. Input selection using Mutual Information – Applications to time series prediction. Rapport technique, Helsinki University of Technology.

He H.-D., Lu W.-Z. et Xue Y. 2014. "Prediction of particulate matters at urban intersection by using multilayer perceptron model based on principal components." *Stochastic Environmental Research and Risk Assessment* .

Hebb D.O. 1949. *The Organization of Behavior : A Neuropsychological Theory.* New-York : Wiley.

Hervé-Bazin B. 2004. *Le risque cancérogène du plomb : évaluation du risque cancérogène lié à l'exposition professionnelle au plomb et à ses composés inorganiques.* Paris ; Les Ulis, France : Institut national de Recherche et de Sécurité ; EDP Sciences.

Hestenes M.R. 1980. *Conjugate Direction Methods in Optimization.* New York, NY : Springer.

Hestenes M.R. et Stiefel E. 1952. "Methods of Conjugate Gradients for Solving Linear Systems." *Journal of Research of the National Bureau of Standards* 49 (6), pages 409 – 436.

Holland J.H. 1975. *Adaptation in Natural and Artificial Systems.* Ann Arbor : University of Michigan Press.

Hooyberghs J., Mensink C., Dumont G., Fierens F. et Brasseur O. 2005. "A neural network forecast for daily average PM concentrations in Belgium." *Atmospheric Environment* 39 (18), pages 3279 – 3289.

Hornik K., Stinchcombe M. et White H. 1989. "Multilayer feedforward networks are universal approximators." *Neural Networks* 2 (5), pages 359 – 366.

Hrust L., Klaić Z.B., Križan J., Antonić O. et Hercog P. 2009. "Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations." *Atmospheric Environment* 43 (35), pages 5588 – 5596.

Ibarra-Berastegi G., Elias A., Barona A., Saenz J., Ezcurra A. et Diaz de Argandoña J. 2008. "From diagnosis to prognosis for forecasting air pollution using neural networks : Air pollution monitoring in Bilbao." *Environmental Modelling & Software* 23, pages 622 – 637.

ICPP. 2014. Climate Change 2014 : Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Rapport technique, ICPP.

IEEE Computer Society, Microprocessor Standards Committee, Institute of Electrical and Electronics Engineers et IEEE-SA Standards Board. 2008. *IEEE standard for floating-point arithmetic.* New York, NY : Institute of Electrical and Electronics Engineers.

Ionescu A. 2013. "Prévision statistique des concentrations de particules dans l'air : tour d'horizon des principaux outils mathématiques." *Pollution atmosphérique, climat, santé, société* 217 (URL : http ://lodel.irevues.inist.fr/pollution-atmospherique/index.php ?id=873).

Jacob Daniel J. 1999. *Introduction to atmospheric chemistry.* Princeton, N.J : Princeton University Press.

Jiang D., Zhang Y., Hu X., Zeng Y., Tan J. et Shao D. 2004. "Progress in developing an ANN model for air pollution index forecast." *Atmospheric Environment* 38 (40), pages 7055 – 7064.

Journal Officiel du 29 mars 2014. 2014. "Arrêté du 26 mars 2014 relatif au déclenchement des procédures préfectorales en cas d'épisodes de pollution de l'air ambiant.".

Juhos I., Béczi R. et Makra L. 2003. "Comparison of atificial intelligence prediction techniques in NO and NO2 concentrations' forecast." *Acta Climatologica et Chorologica* 36 - 37 (45 - 56), pages 45 – 55.

Kallos G., Nickovic S., Papadopoulos A., Jovis D., Kakaliagou O., Misirlis N., Boukas L., Mimikou N., Sakellaridis G., Papageorgiou J., Anadranistakis E. et Manousakis M. 1997. "The regional weather forecasting system SKIRON : An overview.". University of Athens : Proceedings of the symposium on regional weather prediction on parallel computer environments pages 109 – 122.

Kao J.-J. et Huang S.-S. 2000. "Forecasts Using Neural Network versus Box-Jenkins Methodology for Ambient Air Quality Monitoring Data." *Journal of the Air & Waste Management Association* 50 (2), pages 219 – 226.

Kappos A.D., Bruckmann P., Eikmann T., Englert N., Heinrich U., Höppe P., Koch E., Krause G.H.M., Kreyling W.G., Rauchfuss K., Rombout P., Schulz-Klemp V., Thiel W.R. et Wichmann H.-E. 2004. "Health effects of particles in ambient air." *International Journal of Hygiene and Environmental Health* 207 (4), pages 3998 – 407.

Kelley C. T. 2003. *Solving nonlinear equations with Newton's method.* Fundamentals of algorithms Philadelphia : Society for Industrial and Applied Mathematics.

Khazaee P., Mozayani N. et Jahed Motlagh M. 2008. "A Genetic-Based Input Variable Selection Algorithm Using Mutual Information and Wavelet Network for Time Series Prediction." Dans *IEEE International Conference on Systems, Man and Cybernetics.* IEEE.

Kirkpatrick S., Gelatt C.D. et Vecchi M.P. 1983. "Optimization by Simulated Annealing." *Science* 220 (4598), pages 671 – 680.

Knaapen A.M., Borm P.J.A., Albrecht C. et Schins R.P.F. 2004. "Inhaled particles and lung cancer. Part A : Mechanisms." *International Journal of Cancer* 109 (6), pages 799 – 809.

Kohavi R. et John G.H. 1997. "Wrappers for feature subset selection." *Artificial Intelligence* 97 (1 - 2), pages 273 – 324.

Kohonen T. 1982. "Self-organized formation of topologically correct feature maps." *Biological Cybernetics* 43 (1), pages 59 – 69.

Kolehmainen M., Martikainen H. et Ruuskanen J. 2001. "Neural networks and periodic components used in air quality forecasting." *Atmospheric Environment* 35, pages 815 – 825.

Kolehmainen M., Martikainen H., Hiltunen T. et Ruuskanen J. 2000. "Forecasting air quality parameters using hybrid neural network modelling." *Environmental Monitoring and Assessment* 65, pages 277 – 286.

Kraskov A., Stögbauer H. et Grassberger P. 2004. "Estimating mutual information." *Physical Review E* 69 (6), pages 066138.

Kukkonen J., Olsson T., Schultz D.M., Baklanov A., Klein T., Miranda A.I., Monteiro A., Hirtl M., Tarvainen V., Boy M., Peuch V-H., Poupkou A., Kioutsioukis I., Finardi S., Sofiev M., Sokhi R., Lehtinen K.E.J., Karatzas K., San José R., Astitha M., Kallos G., Schaap M., Reimer E., Jakobs H. et Eben K. 2012. "A review of operational, regional-scale, chemical weather forecasting models in Europe." *Atmospheric Chemistry and Physics* 12 (1), pages 1 – 87.

Kukkonen J., Partanen L., Karppinen A., Ruuskanen J., Junninen H., Kolehmainen M., Niska H., Dorlind S., Chatterton T., Foxall R. et Cawley G. 2003. "Extensive evaluation of neural

network models for the prediction of NO2 and PM 10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki." *Atmospheric Environment* 37, pages 4539 – 4550.

Kurt A., Gulbagci B., Karaca F. et Alagha O. 2008. "An online air pollution forecasting system using neural networks." *Environment International* 34 (5), pages 592 – 598.

Lauret P., Fock E., Randrianarivony R.N. et Manicom-Ramsamy J-F. 2008. "Bayesian neural network approach to short time load forecasting." *Energy Conversion and Management* 49 (5), pages 1156 – 1166.

Legg P.A., Rosin P.L., Marshall D. et Morgan J.E. 2007. "Improving accuracy and efficiency of registration by mutual information using Sturges' histogram rule." Dans *Proceedings of Medical Image Understanding and Analysis*. University of Aberystwyth : pages 26 – 30.

Levenberg K. 1944. "A method for the solution of certain non-linear problems in least squares." *Quarterly Journal of Applied Mathmatics* 2 (2), pages 164 – 168.

Li W. 1990. "Mutual Information Functions versus Correlation Functions." *Journal of Statistical Physics* 60 (5 - 6), pages 823 – 837.

Liaw A. et Wiener M. 2002. "Classification and Regression by randomForest." *R News* 2 (3), pages 18 – 22.

Lindfield G. R. et Penny J. E. T. 2012. *Numerical methods : using MATLAB*. 3rd ed. Waltham, MA : Academic Press.

Lloyd S. 1982. "Least squares quantization in PCM." *IEEE Transactions on Information Theory* 28 (2), pages 129 – 137.

Léon J.-F., Augustin P., Mallet M., Bourrianne T., Pont V., Dulac F., Fourmentin M., Lambert D. et Sauvage B. 2015. "Aerosol vertical distribution, optical properties and transport over Corsica (western Mediterranean)." *Atmospheric Chemistry and Physics Discussions* 15 (6), pages 9507 – 9540.

Lu H-C., Hsieh J-C. et Chang T-S. 2006. "Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network." *Atmospheric Research* 81 (2), pages 124 – 139.

Mallet V., Quélo D., Sportisse B., Ahmed de Biasi M., Debry é., Korsakissok I., Wu L., Roustan Y., Sartelet K., Tombette M. et Foudhil H. 2007. "Technical Note : The air quality modeling system Polyphemus." *Atmospheric Chemistry and Physics* 7 (20), pages 5479 – 5487.

Marquardt D.W. 1963. "An algorithm for least-squares estimation of nonlinear parameters." *Journal of the Society for Industrial & Applied Mathematics* 11 (2), pages 431 – 441.

McCollister G.M. et Wilson K.R. 1975. "Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants." *Atmospheric Environment* 9 (4), pages 417 – 423.

McCulloch W.S. et Pitts W. 1943. "A logical calculus of the ideas immanent in nervous activity." *The Bulletin of Mathematical Biophysics* 5 (4), pages 115 – 133.

McKeen S., Chung S. H., Wilczak J., Grell G., Djalalova I., Peckham S., Gong W., Bouchet V., Moffet R., Tang Y., Carmichael G. R., Mathur R. et Yu S. 2007. "Evaluation of several $PM_{2.5}$ forecast models using data collected during the ICARTT/NEAQS 2004 field study." *Journal of Geophysical Research* 112 (D10).

Menut L., Bessagnet B., Khvorostyanov D., Beekmann M., Blond N., Colette A., Coll I., Curci G., Foret G., Hodzic A., Mailler S., Meleux F., Monge J.-L., Pison I., Siour G., Turquety S., Valari M., Vautard R. et Vivanco M. G. 2013. "CHIMERE 2013 : a model for regional atmospheric composition modelling." *Geoscientific Model Development* 6 (4), pages 981 – 1028.

Metropolis N., Rosenbluth A.W., Rosenbluth M.M. et Teller A.H. 1953. "Equation of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21 (6), pages 1088 – 1092.

Michalakes J., Dudhia J., Gill D., Henderson T., Klemp J., Skamarock W. et Wang W. 2004. "The weather research and forecast model : software architecture and performance." Dans *Proceedings of the 11th ECMWF Workshop on the Use of High Performance Computing In Meteorology*. Vol. 25 Reading, UK : G. Mozdzynski.

Milionis A.E. et Davies T.D. 1994. "Regression and stochastic models for air pollution—I. Review, comments and suggestions." *Atmospheric Environment* 28 (17), pages 2801 – 2810.

Minsky L.M. et Papert S.A. 1969. *Perceptrons : an introduction to computational geometry*. Cambridge : The MIT Press.

Mishra D. et Goyal P. 2015. "Development of artificial intelligence based NO2 forecasting models at Taj Mahal, Agra." *Atmospheric Pollution Research* 6 (1), pages 99 – 106.

Møller M.F. 1993. "A scaled conjugate gradient algorithm for fast supervised learning." *Neural Networks* 6 (4), pages 525 – 533.

Moon Y., Rajagopalan B. et Lall U. 1995. "Estimation of mutual information using kernel density estimators." *Physical Review E* 52 (3), pages 2318 – 2321.

Nguyen D. et Widrow B. 1990. "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights." Dans *IJCNN International Joint Conference on Neural Networks*. Vol. 3 San Diego, CA : IEEE pages 21 – 26.

Nicolas J.B., Sciare J., Petit J.-E., Bonnaire N., Féron A., Dulac F., Hamonou E., Gros V., Mallet M., Lambert D., Sauvage S., Léonardis T., Tison E., Colomb A., Fresney E., Pichon J.-M., Bouvier L., Bourrianne T. et Roberts G. 2013. "New insights on aerosol sources and properties of Organics in the west Mediterranean basin." *Geophysical Research Abstracts* 15 (EGU2013-12852).

Niemeyer L.E. 1960. "Forecasting air pollution potentia." *Monthly Weather Review* 88 (3), pages 88 – 96.

Niska H., Hiltunen T., Karppinen A., Ruuskanen J. et Kolehmainen M. 2004. "Evolving the neural network model for forecasting air pollution time series." *Engineering Applications of Artificial Intelligence* 17 (2), pages 159 – 167.

Niska H., Rantamäki M., Hiltunen T., Karppinen A., Kukkonen J., Ruuskanen J. et Kolehmainen M. 2005. "Evaluation of an integrated modelling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations." *Atmospheric Environment* 39 (35), pages 6524 – 6536.

Notton G. 2015. "Importance of islands in renewable energy production and storage : The situation of the French islands." *Renewable and Sustainable Energy Reviews* 47, pages 260 – 269.

Ordieres J.B., Vergara E.P., Capuz R.S. et Salazar R.E. 2005. "Neural network prediction model for fine particulate matter (PM2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua)." *Environmental Modelling & Software* 20, pages 547 – 559.

Paschalidou A.K., Karakitsios S., Kleanthous S. et Kassomenos P.A. 2011. "Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models : implications to local environmental management." *Environmental Science and Pollution Research* 18 (2), pages 316 – 327.

Pechony O. et Shindell D. T. 2010. "Driving forces of global wildfires over the past millennium and the forthcoming century." *Proceedings of the National Academy of Sciences* 107 (45), pages 19167 – 19170.

Pederzoli A., Mircea M., Finardi S., di Sarra A. et Zanini G. 2010. "Quantification of Saharan dust contribution to PM10 concentrations over Italy during 2003–2005." *Atmospheric Environment* 44 (34), pages 4181 – 4190.

Peng H., Long F. et Ding C. 2005. "Feature Selection Based on Mutual Information : Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." Dans *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 27 pages 1226 – 1238.

Perez P. 2012. "Combined model for PM10 forecasting in a large city." *Atmospheric Environment* 60, pages 271 – 276.

Perez P. et Reyes J. 2001. "Prediction of Particlulate Air Pollution using Neural Techniques." *Neural Computing & Applications* 10 (2), pages 165 – 171.

Perez P. et Reyes J. 2002. "Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile." *Atmospheric Environment* 36, pages 4555 – 4561.

Perez P. et Reyes J. 2006. "An integrated neural network model for PM10 forecasting." *Atmospheric Environment* 40 (16), pages 2845 – 2851.

Perez P. et Salini G. 2008. "PM2.5 forecasting in a large city : Comparison of three methods." *Atmospheric Environment* 42 (35), pages 8219 – 8224.

Petit R. H. 2005. "Transport of Saharan dust over the Caribbean Islands : Study of an event." *Journal of Geophysical Research* 110 (D18).

Peuch V-H., Amodei M., Barthet T., Cathala M-L., Josse B., Michou M. et Simon P. 2009. "MOCAGE : MOdèle de Chimie A Grande Echelle.". Vol. 48 Toulouse : Météo-France pages 662 – 689.

Pfeiffer H., Baumbach G., Sarachaga-Ruiz L., Kleanthous S., Poulida O. et Beyaz E. 2009. "Neural modelling of the spatial distribution of air pollutants." *Atmospheric Environment* 43 (20), pages 3289 – 3297.

Pichon J.-M., Colomb A., Gheusi F., Sauvage S., Pont V., Tison E., Bordier F., Grignon G., Savelli J-L., Dulac F., Sciare J., Nicolas J.B., Bourrianne T. et Bouvier L. 2013. "Real-time measurement of reactive gases (NO, NO2, O3, CO) at ERSA, Cape Corsica, a long term Observatory." *Geophysical Research Abstracts* 15 (EGU2013-12622).

Poggi J-M. et Portier B. 2011. "PM10 forecasting using clusterwise regression." *Atmospheric Environment* 45 (38), pages 7005 – 7014.

Polydoras G.N., Anagnostopoulos J.S. et Bergeles G.C. 1998. “Air quality predictions : dispersion model vs Box-Jenkins stochastic models. An implementation and comparison for Athens, Greece.” *Applied Thermal Engineering* 18, pages 1037 – 1048.

Raes F., Van Dingenen R., Vignati E., Wilson J., Putaud J.-P., Seinfeld J.H. et Adams P. 2000. “Formation and cycling of aerosols in the global troposphere.” *Atmospheric Environment* 34 (25), pages 4215 – 4240.

Robeson S.M. et Steyn D.G. 1990. “Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations.” *Atmospheric Environment. Part B. Urban Atmosphere* 24 (2), pages 303 – 312.

Rodriguez S., Querol X., Alastuey A., Kallos G. et Kakaliagou O. 2001. “Saharan dust contributions to PM10 and TSP levels in Southern and Eastern Spain.” *Atmospheric Environment* 35 (14), pages 2433 – 2447.

Rosenblatt F. 1958. “The perceptron : A probabilistic model for information storage and organization in the brain.” *Psychological Review* 65 (6), pages 386 – 408.

Roubeyrie L. 2013. “Package pyair.”. http ://pythonhosted.org/PyAir/.

Rouil L., Honoré C., Bessagnet B., Malherbe L., Meleux F., Vautard R., Beekmann M., Flaud J-M., Dufour A., Martin D., Peuch A., Peuch V-H., Elichegaray C., Poisson N. et Menut L. 2009. “Prev’air : An Operational Forecasting and Mapping System for Air Quality in Europe.” *Bulletin of the American Meteorological Society* 90 (1), pages 73 – 83.

Rumelhart D.E., Hinton G.E. et Williams R.J. 1986. “Learning representations by back-propagating errors.” *Nature* 323 (6088), pages 533 – 536.

Ryan W.F. 1995. “Forecasting severe ozone episodes in the Baltimore metropolitan area.” *Atmospheric Environment* 29, pages 2352 – 2310.

Sammon J.W. 1969. “A Nonlinear Mapping for Data Structure Analysis.” *IEEE Transactions on Computers* C-18 (5), pages 401 – 409.

Saporta G. 2006. *Probabilités, analyse des données et statistique.* Paris : Editions Technip.

Schlink U., Dorling S., Pelikan E., Nunnari G., Cawley G., Junninen H., Greig A., Foxall R., Eben K., Chatterton T., Vondracek J., Richter M., Dostal M., Bertucco L., Kolehmainen M. et Doyle M. 2003. “A rigorous inter-comparison of ground-level ozone predictions.” *Atmospheric Environment* 37 (23), pages 3237 – 3253.

Seity Y., Brousseau P., Malardel S., Hello G., Bénard P., Bouttier F., Lac C. et Masson V. 2011. “The AROME-France Convective-Scale Operational Model.” *Monthly Weather Review* 139 (3), pages 976–991.

Sfetsos A. et Vlachogiannis D. 2010. “A new methodology development for the regulatory forecasting of PM10. Application in the Greater Athens Area, Greece.” *Atmospheric Environment* 44 (26), pages 3159 – 3172.

Shanno D. F. 1970. “Conditioning of quasi-Newton methods for function minimization.” *Mathematics of Computation* 24 (111), pages 647–647.

Shannon C.E. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27 (3 - 4), pages 379 – 423, 623 – 656.

Shekarrizfard M., Karimi-Jashni A. et Hadad K. 2012. "Wavelet transform-based artificial neural networks (WT-ANN) in PM10 pollution level estimation, based on circular variables." *Environmental Science and Pollution Research* 19 (1), pages 256 – 268.

Shi J.P. et Harrison R.M. 1997. "Regression modelling of hourly NOx and NO2 concentrations in urban air in London." *Atmospheric Environment* 31 (24), pages 4081 – 4094.

Slini T., Kaprara A., Karatzas K. et Moussiopoulos N. 2006. "PM10 forecasting for Thessaloniki, Greece." *Environmental Modelling & Software* 21 (4), pages 559 – 565.

Slini T., Karatzas K. et Moussiopoulos N. 2002. "Statistical analysis of environmental data as the basis of forecasting : an air quality application." *Science of The Total Environment* 288 (3), pages 227 – 237.

Sánchez-Maroño N., Fontenla-Romero O., Alonso-Betanzos A. et Guijarro-Berdiñas B. 2003. Self-Organizing Maps and Functional Networks for Local Dynamic Modeling. Dans *ESANN*. Evere, Belgique : pages 39 – 44.

Sousa S.I.V., Martins F.G., Alvim-Ferraz M.C.M. et Pereira M.C. 2007. "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations." *Environmental Modelling & Software* 22, pages 97 – 103.

Sportisse B. 2008. *Pollution atmosphérique : des processus à la modélisation*. Paris : Springer-Verlag.

Stadlober E., Hörmann S. et Pfeiler B. 2008. "Quality and performance of a PM10 daily forecasting model." *Atmospheric Environment* 42 (6), pages 1098 – 1109.

Stull R.B. 1988. *An Introduction to Boundary Layer Meteorology*. Dordrecht : Springer Netherlands.

Sturges H.A. 1926. "The Choice of a Class Interval." *Journal of the American Statistical Association* 21 (153), pages 65 – 66.

Suárez Sánchez A., García Nieto P.J., Riesgo Fernández P., del Coz Díaz J.J. et Iglesias-Rodríguez F.J. 2011. "Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain)." *Mathematical and Computer Modelling* 54 (5-6), pages 1453 – 1466.

Tamas W., Notton G., Paoli C., Nivet M.-L. et Voyant C. sous presse. "Hybridization of air quality forecasting models using machine learning and clustering : an original approach to detect pollutant peaks." *Aerosol and Air Quality Research* .

Tamas W., Notton G., Paoli C., Voyant C., Nivet M-L. et Balu A. 2014. "Urban ozone concentration forecasting with artificail neural network in Corsica." *Mathematical Modelling in Civil Engineering* (1), pages 33 – 41.

Viotti P., Liuti G. et Di Genova P. 2002. "Atmospheric urban pollution : applications of an artificial neural network (ANN) to the city of Perugia." *Ecological Modelling* 148 (1), pages 27 – 46.

Voyant C., Tamas W., Nivet M.-L., Notton G., Paoli C., Balu A. et Muselli M. 2015. "Meteorological time series forecasting with pruned multi-layer perceptron and two-stage Levenberg-Marquardt method." *International Journal of Modelling, Identification and Control* 23 (3), pages 287 – 294.

Voyant C., Tamas W., Paoli C., Balu A., Muselli M., Nivet M-L. et Notton G. 2014. "Time series modeling with pruned multi-layer perceptron and 2-stage damped least-squares method." Dans *Journal of Physics : Conference Series*. Vol. 490 p. 012040.

Walter J., Riter H. et Schulten K. 1990. "Nonlinear prediction with self-organizing maps." Dans *IJCNN International Joint Conference on Neural Networks*. Vol. 1 IEEE pages 589 – 594.

Wang W., Xu Z. et Weizhen Lu J. 2003. "Three improved neural network models for air quality forecasting." *Engineering Computations* 20 (2), pages 192 – 210.

Wang X.-K. et Lu W.-Z. 2006. "Seasonal variation of air pollution index : Hong Kong case study." *Chemosphere* 63 (8), pages 1261 – 1272.

Ward J.H. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58 (301), pages 236 – 244.

Widrow B. et Hoff M.E. 1960. "Adaptive Switching Circuits." Dans *IRE Wescon convention record, part 4*. New-York : IRE pages 96 – 104.

Willmott C.J. 1982. "Some comments on the evaluation of model performance." *Bulletin American Meteorological Society* 63 (11), pages 1309 – 1313.

Wolff G.T. et Lioy P.J. 1978. "An Empirical Model for Forecasting Maximum Daily Ozone Levels in the Northeastern U.S." *Journal of the Air Pollution Control Association* 28 (10), pages 1034 – 1038.

Yao X. 1999. "Evolving artificial neural networks." *Proceedings of the IEEE* 87 (9), pages 1423 – 1447.

Yi J. et Prybutok V.R. 1996. "A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area." *Environmental Pollution* 92 (3), pages 349 – 357.

Yu H.-L., Chen B.-L., Chiu C.-H., Lu M.-M. et Tung C.-P. 2015. "Analysis of space–time patterns of rainfall events during 1996–2008 in Yilan County (Taiwan)." *Stochastic Environmental Research and Risk Assessment* 29 (3), pages 929 – 945.

Yu S., Mathur R., Schere K., Kang D., Pleim J., Young J., Tong D., Pouliot G., McKeen S. A. et Rao S. T. 2008. "Evaluation of real-time $PM_{2.5}$ forecasts and process analysis for $PM_{2.5}$ formation over the eastern United States using the Eta-CMAQ forecast model during the 2004 ICARTT study." *Journal of Geophysical Research* 113 (D6).

Zainuddin Z. et Pauline O. 2011. "Modified wavelet neural network in function approximation and its application in prediction of time-series pollution data." *Applied Soft Computing* 11 (8), pages 4866 – 4874.

Zhang G., Eddy Patuwo B. et Hu M.Y. 1998. "Forecasting with artificial neural networks : The state of the art." *International Journal of Forecasting* 14 (1), pages 35 – 62.

Zhang Y., Bocquet M., Mallet V., Seigneur C. et Baklanov A. 2012*a*. "Real-time air quality forecasting, part I : History, techniques, and current status." *Atmospheric Environment* 60, pages 632 – 655.

Zhang Y., Bocquet M., Mallet V., Seigneur C. et Baklanov A. 2012*b*. "Real-time air quality forecasting, part II : State of the science, current research needs, and future prospects." *Atmospheric Environment* 60, pages 656 – 676.

Annexe A

Arrêté ministériel relatif au déclenchement des procédures préfectorales en cas d'épisodes de pollution de l'air ambiant

Décrets, arrêtés, circulaires

TEXTES GÉNÉRAUX

MINISTÈRE DE L'ÉCOLOGIE, DU DÉVELOPPEMENT DURABLE ET DE L'ÉNERGIE

Arrêté du 26 mars 2014 relatif au déclenchement des procédures préfectorales en cas d'épisodes de pollution de l'air ambiant

NOR : *DEVR1400449A*

La ministre des affaires sociales et de la santé, le ministre de l'intérieur, le ministre du redressement productif, le ministre de l'écologie, du développement durable et de l'énergie, le ministre de l'agriculture, de l'agroalimentaire et de la forêt et le ministre délégué auprès du ministre de l'écologie, du développement durable et de l'énergie, chargé des transports, de la mer et de la pêche,

Vu la directive 2008/50/CE du Parlement européen et du Conseil du 21 mai 2008 concernant la qualité de l'air ambiant et un air pur pour l'Europe ;

Vu code de l'environnement, notamment ses articles L. 221-6, L. 222-4 à L. 222-7, L. 223-1, L. 223-2, R. 221-1, R. 221-4 à R. 221-8, R. 222-13 à R. 222-36 et R. 223-1 à R. 223-4 ;

Vu le code de la route, notamment ses articles R. 311-1 et R. 411-19 ;

Vu le code de la sécurité intérieure, notamment ses articles R.* 122-4, R.* 122-5 et R.* 122-8 ;

Vu le décret n° 2004-374 du 29 avril 2004 relatif aux pouvoirs des préfets, à l'organisation et à l'action des services de l'Etat dans les régions et départements ;

Vu l'arrêté du 11 juin 2003 relatif aux informations à fournir au public en cas de dépassement ou de risque de dépassement des seuils de recommandation ou des seuils d'alerte ;

Vu l'arrêté du 21 octobre 2010 relatif aux modalités de surveillance de la qualité de l'air et à l'information du public ;

Vu l'avis de la commission consultative d'évaluation des normes en date du 6 février 2014 ;

Vu l'avis du commissaire à la simplification en date du 28 février 2014,

Arrêtent :

Art. 1er. – Au sens du présent arrêté, on entend par :

« Episode de pollution de l'air ambiant » : période au cours de laquelle le niveau d'un ou de plusieurs polluants atmosphériques est supérieur au seuil d'information et de recommandation (épisode de pollution d'information et de recommandation) ou au seuil d'alerte (épisode de pollution d'alerte).

« Persistance d'un épisode de pollution aux particules PM10 » : épisode de pollution aux particules PM10 caractérisé par constat de dépassement du seuil d'information et de recommandation (modélisation intégrant les données des stations de fond) durant deux jours consécutifs et prévision de dépassement du seuil d'information et de recommandation pour le jour même et le lendemain. En l'absence de modélisation des pollutions, un épisode de pollution aux particules PM10 est persistant lorsqu'il est caractérisé par constat d'une mesure de dépassement du seuil d'information et de recommandation sur station de fond durant trois jours consécutifs. Dans ce cas, les constats peuvent être observés sur des stations de fond différentes au sein d'une même superficie retenue pour la caractérisation de l'épisode de pollution.

« Procédure préfectorale d'information et de recommandation » : ensemble de pratiques et d'actes administratifs pris par l'autorité préfectorale lors d'un épisode de pollution d'information et de recommandation, comprenant des actions d'information et de communication et des recommandations qu'elle peut mettre en œuvre elle-même ou déléguer aux organismes agréés de surveillance de la qualité de l'air.

« Procédure préfectorale d'alerte » : ensemble de pratiques et d'actes administratifs pris par l'autorité préfectorale lors d'un épisode de pollution d'alerte, comprenant aussi bien des actions d'information et de communication et des recommandations qu'elle peut mettre en œuvre elle-même ou déléguer aux organismes agréés de surveillance de la qualité de l'air que des mesures réglementaires de réduction des émissions de polluants qu'elle met en œuvre elle-même.

« Station de fond » : station de mesure de la qualité de l'air de type urbaine, périurbaine ou rurale permettant le suivi de l'exposition moyenne de la population aux phénomènes de pollution atmosphérique. Son

emplacement, hors de l'influence directe d'une source de pollution, permet de mesurer, pour un secteur géographique donné, les caractéristiques chimiques représentatives d'une masse d'air moyenne dans laquelle les polluants émis par les différents émetteurs ont été dispersés.

Art. 2. – Un épisode de pollution est caractérisé :

– soit à partir d'un critère de superficie, dès lors qu'une surface d'au moins 100 km^2 au total dans une région est concernée par un dépassement de seuils d'ozone, de dioxyde d'azote et/ou de particules PM10 estimé par modélisation en situation de fond ;

– soit à partir d'un critère de population :

  – pour les départements de plus de 500 000 habitants, lorsqu'au moins 10 % de la population du département sont concernés par un dépassement de seuils d'ozone, de dioxyde d'azote et/ou de particules PM10 estimé par modélisation en situation de fond ;

  – pour les départements de moins de 500 000 habitants, lorsqu'au moins une population de 50 000 habitants au total dans le département est concernée par un dépassement de seuils d'ozone, de dioxyde d'azote et/ou de particules PM10 estimé par modélisation en situation de fond ;

– soit en considérant les situations locales particulières portant sur un territoire plus limité, notamment les vallées encaissées ou mal ventilées, les zones de résidence à proximité de voiries à fort trafic, les bassins industriels.

En l'absence de modélisation de la qualité de l'air, un épisode de pollution peut être caractérisé par constat d'une mesure de dépassement d'un seuil sur au moins une station de fond.

Art. 3. – En cas d'épisode de pollution caractérisé conformément à l'article 2 du présent arrêté, les procédures préfectorales visées par le présent arrêté sont déclenchées de manière à prendre effet le jour même ou le lendemain.

Lorsque le dépassement de seuil qui permet de caractériser l'épisode de pollution est issu d'une modélisation, le déclenchement des procédures préfectorales se fait sans attendre la confirmation par mesure dudit dépassement de seuil.

Art. 4. – Les modalités de déclenchement des procédures préfectorales d'information et de recommandation et d'alerte en cas d'épisode de pollution, relatives au polluant dioxyde de soufre, sont définies par arrêté préfectoral ou interpréfectoral.

Art. 5. – La mise en œuvre des actions d'information, de communication et de recommandation et des mesures réglementaires de réduction des émissions de polluants circonscrites à un département relève du préfet de département, sous réserve des compétences du préfet de zone de défense et de sécurité mentionnées à l'article R.* 1311-7 du code de la défense.

Le préfet de zone de défense et de sécurité, conformément aux dispositions du code de la défense précitées, prend les mesures de coordination nécessaires lorsque intervient une situation de crise ou que se développent des événements d'une particulière gravité, quelle qu'en soit l'origine, de nature à porter atteinte à l'environnement et que cette situation ou ces événements peuvent avoir des effets dépassant ou susceptibles de dépasser le cadre d'un département. Il prend pour cela les mesures de police administrative nécessaires à l'exercice de ce pouvoir. A ce titre, il assure la coordination zonale en continu des épisodes de pollution et établit un document-cadre relatif aux procédures préfectorales et aux actions particulières de dimension interdépartementale dans sa zone.

Le préfet de département prend un arrêté déclinant le document-cadre à l'échelle de son département. Afin de tenir compte de la nécessité de déclencher des actions de réduction des émissions dans les territoires plus grands que les seuls départements concernés par des dépassements, cet arrêté peut être interpréfectoral. Le document-cadre relatif aux procédures préfectorales et aux actions particulières de dimension interdépartementale établi par le préfet de zone de défense et de sécurité peut prévoir les cas dans lesquels l'arrêté interpréfectoral est pris.

Cet arrêté préfectoral ou interpréfectoral organise le dispositif à respecter en cas d'épisode de pollution. Il décrit les modalités de déclenchement des procédures prévues dans le présent arrêté et précise le rôle des acteurs, le contenu de l'information à diffuser conformément à l'article R. 221-8 du code de l'environnement, les modalités de diffusion, les recommandations et les mesures réglementaires de réduction des émissions des polluants.

L'arrêté préfectoral ou interpréfectoral établit la liste des actions d'information, de communication et de recommandation et des mesures réglementaires de réduction des émissions de polluants, qui inclut *a minima* celles listées en annexe du présent arrêté. Il adapte ces actions et ces mesures aux particularités locales et précise pour chacune d'elles les circonstances et les caractéristiques des épisodes de pollution causant leur déclenchement.

Art. 6. – Lorsqu'il est informé d'un épisode de pollution par l'organisme agréé de surveillance de la qualité de l'air, conformément à l'arrêté préfectoral ou interpréfectoral cité ci-dessus et dans les formes notamment prévues à l'article R. 223-2 du code de l'environnement, le préfet ou, à Paris, le préfet de police déclenche, pour le département concerné par la nécessité de mettre en œuvre des actions d'information, de communication et de recommandation et/ou de mesures réglementaires de réduction des émissions, une procédure adaptée au(x) polluant(s) et au(x) seuil(s) réglementaire(s) concerné(s), telle que précisée ci-après.

Dans la procédure d'information et de recommandation, le préfet déclenche des actions d'information du public, des maires, des établissements de santé et établissements médico-sociaux, des professionnels concernés et des relais adaptés à la diffusion de cette information ainsi que des diffusions de recommandations sanitaires et de recommandations visant à limiter les émissions des sources fixes ou mobiles de pollution atmosphérique concourant à l'élévation de la concentration du polluant considéré.

Dans la procédure d'alerte, le préfet déclenche, d'une part, des actions d'information du public, des maires, des établissements de santé et établissements médico-sociaux, des professionnels concernés et des relais adaptés à la diffusion de cette information, ainsi que des diffusions de recommandations sanitaires et de recommandations visant à limiter les émissions des sources fixes ou mobiles de pollution atmosphérique concourant à l'élévation de la concentration du polluant considéré et, d'autre part, des mesures réglementaires de restriction ou de suspension de certaines activités concourant à l'élévation de la concentration du polluant considéré, y compris, le cas échéant, de la circulation des véhicules, en application du chapitre III du titre II du livre II du code de l'environnement.

Pour les épisodes de pollution aux particules PM10, la procédure d'information et de recommandation évolue en procédure d'alerte en cas de persistance de l'épisode.

Art. 7. – En cas d'épisode de pollution à l'ozone ou aux particules PM10, les actions d'information, de communication et de recommandation et les mesures réglementaires de réduction des émissions de polluants qui ne sont pas relatives aux transports s'appliquent soit à l'ensemble du département, soit à un bassin d'air proportionné à la zone de pollution, défini, le cas échéant, dans le document-cadre relatif aux procédures préfectorales et aux actions particulières de dimension interdépartementale établi par le préfet de zone et justifié en prenant en considération les caractéristiques topographiques et les circulations d'air sur le territoire concerné.

En cas d'épisode de pollution au dioxyde d'azote, les actions d'information, de communication et de recommandation et les mesures réglementaires de réduction des émissions de polluants qui ne sont pas relatives aux transports peuvent être limitées à une zone habitée concernée par la pollution.

Les actions d'information, de communication et de recommandation et les mesures réglementaires de réduction des émissions de polluants relatives aux transports peuvent être limitées à l'échelle du réseau de transport concerné par la pollution.

Art. 8. – Les informations données par le préfet à la population en cas de procédures préfectorale d'information et de recommandation ou de procédures préfectorales d'alerte comprennent :

– le ou les polluants concernés ;
– la valeur du seuil dépassé ou risquant d'être dépassé et la définition de ce seuil ou, le cas échéant, pour les particules PM10, l'information du déclenchement de la procédure par persistance ;
– le type de procédure préfectorale déclenchée (d'information et de recommandation ou d'alerte) ;
– l'aire géographique concernée et la durée prévue du dépassement, en fonction des données disponibles ;
– l'explication du dépassement (causes, facteurs aggravants, etc.) lorsqu'elle est connue ;
– des prévisions concernant l'évolution des concentrations (amélioration, stabilisation ou aggravation) ;
– les recommandations de réduction des émissions et, le cas échéant, les mesures réglementaires mises en œuvre ;
– les recommandations sanitaires prévues à l'article R. 221-4 du code de l'environnement et un court rappel des effets sur la santé de la pollution atmosphérique ;
– l'aire géographique de mise en place des actions d'information, de communication et de recommandation et des mesures réglementaires de réduction des émissions de polluants.

Le préfet peut confier à l'organisme agréé de surveillance de la qualité de l'air la diffusion de ces informations. Les modalités de cette diffusion sont définies par arrêté préfectoral ou interpréfectoral.

Lors d'un épisode de pollution, le préfet met en œuvre, parmi les recommandations et mesures réglementaires de réduction des émissions listées dans l'arrêté préfectoral ou interpréfectoral cité à l'article 5 du présent arrêté, celles qui sont les mieux adaptées et proportionnées aux caractéristiques de la pollution constatée ou prévue. La population exposée, l'aire géographique et la durée de l'épisode de pollution peuvent être considérées pour la gradation des actions d'information, de communication et de recommandation et des mesures réglementaires de réduction des émissions de polluants.

Art. 9. – En cas d'épisode de pollution, l'organisme agréé de surveillance de la qualité de l'air informe le préfet compétent au moins une fois par jour sur la pollution atmosphérique constatée et prévue.

L'organisme agréé de surveillance de la qualité de l'air tient informé le préfet et l'agence régionale de santé de l'évolution de l'épisode de pollution.

En cas d'épisode de pollution, les informations relatives à l'état du dispositif préfectoral et aux mesures réglementaires de réduction de polluants sont saisies en temps réel dans un outil national de suivi établi par le ministère en charge du développement durable.

Art. 10. – L'arrêté du 17 août 1998 relatif aux seuils de recommandation et aux conditions de déclenchement de la procédure d'alerte et l'arrêté du 11 juin 2003 relatif aux informations à fournir au public en cas de dépassement ou de risque de dépassement des seuils de recommandation ou des seuils d'alerte sont abrogés.

Art. 11. – Le présent arrêté entre en vigueur le 1er juillet 2014.

Art. 12. – La ministre des affaires sociales et de la santé, le ministre de l'intérieur, le ministre du redressement productif, le ministre de l'écologie, du développement durable et de l'énergie, le ministre de l'agriculture, de l'agroalimentaire et de la forêt et le ministre délégué auprès du ministre de l'écologie, du développement durable et de l'énergie, chargé des transports, de la mer et de la pêche, sont chargés, chacun en ce qui le concerne, de l'exécution du présent arrêté, qui sera publié au *Journal officiel* de la République française.

Fait le 26 mars 2014.

Le ministre de l'écologie, du développement durable et de l'énergie,
PHILIPPE MARTIN

La ministre des affaires sociales et de la santé,
MARISOL TOURAINE

Le ministre de l'intérieur,
MANUEL VALLS

Le ministre du redressement productif,
ARNAUD MONTEBOURG

Le ministre de l'agriculture, de l'agroalimentaire et de la forêt,
STÉPHANE LE FOLL

Le ministre délégué auprès du ministre de l'écologie, du développement durable et de l'énergie, chargé des transports, de la mer et de la pêche,
FRÉDÉRIC CUVILLIER

ANNEXE

RECOMMANDATIONS ET MESURES RÉGLEMENTAIRES DE RÉDUCTION DES ÉMISSIONS PAR GRAND SECTEUR D'ACTIVITÉ POUVANT ÊTRE PRISES PAR LE PRÉFET EN CAS D'ÉPISODE DE POLLUTION DE L'AIR AMBIANT

Les actions et mesures sont adaptées aux circonstances locales et aux caractéristiques de chaque épisode de pollution.

Cette annexe ne contient pas de recommandations d'ordre sanitaire.

I. – Recommandations en cas d'activation du niveau d'information et de recommandation ou du niveau d'alerte

1. *Secteur agricole*

Recommander de décaler dans le temps les épandages de fertilisants minéraux et organiques ainsi que les travaux du sol, en tenant compte des contraintes déjà prévues par les programmes d'actions pris au titre de la directive 91/676/CEE.

Recommander de recourir à des procédés d'épandage faiblement émetteurs d'ammoniac.

Recommander de reporter la pratique de l'écobuage ou pratiquer le broyage.

Recommander de suspendre les opérations de brûlage à l'air libre des sous-produits agricoles.

Recommander de reporter les activités de nettoyage de silo ou tout événement concernant ce type de stockage susceptible de générer des particules, sous réserve que ce report ne menace pas les conditions de sécurité.

Recommander de recourir à des enfouissements rapides des effluents.

2. *Secteur résidentiel et tertiaire*

Recommander d'arrêter l'utilisation de certains foyers ouverts, appareils de combustion de biomasse non performants ou groupes électrogènes.

Recommander de reporter l'utilisation de barbecue à combustible solide (bois, charbon, charbon de bois) à la fin de l'épisode de pollution.

Recommander de maîtriser la température dans les bâtiments (chauffage en hiver et climatisation en été).

Déconseiller, lors de travaux d'entretien ou de nettoyage effectués par la population ou les collectivités locales, d'utiliser des outils non électriques (tondeuses, taille-haie...) ainsi que d'utiliser des produits à base de solvants organiques (white-spirit, peinture, vernis décoratifs, produits de retouche automobile...).

Rappeler l'interdiction du brûlage à l'air libre des déchets verts.

3. *Secteur industriel*

Sur la base de plans d'actions en cas d'épisode de pollution de l'air définis par le préfet en concertation avec les acteurs concernés et contenant une étude préalable d'impact économique et social, recommander aux installations industrielles la mise en œuvre de dispositions de nature à réduire les rejets atmosphériques, y compris la baisse de leur activité, sous réserve que les conditions de sécurité soient préservées et que les coûts induits ne soient pas disproportionnés pour les acteurs publics et privés au regard des bénéfices sanitaires attendus.

Recommander de reporter certaines opérations émettrices de COV (travaux de maintenance, dégazage d'une installation, chargement ou déchargement de produits émettant des composants organiques volatils en l'absence de dispositif de récupération des vapeurs) à la fin de l'épisode de pollution.

Recommander de reporter certaines opérations émettrices de particules ou d'oxydes d'azote à la fin de l'épisode de pollution.

Recommander de reporter le démarrage d'unités à l'arrêt à la fin de l'épisode de pollution.

Recommander la mise en fonctionnement de systèmes de dépollution renforcés, lorsqu'ils sont prévus, pendant la durée de l'épisode de pollution.

Recommander la réduction de l'activité sur les chantiers générateurs de poussières et la mise en place de mesures compensatoires (arrosage, etc.) durant l'épisode de pollution.

Recommander de réduire l'utilisation de groupes électrogènes pendant la durée de l'épisode de pollution.

4. *Secteur des transports*

Recommander de développer des pratiques de mobilité relatives à l'acheminement le moins polluant possible des personnes durant l'épisode de pollution : covoiturage, utilisation de transports en commun, réduction des déplacements automobiles non indispensables des entreprises et des administrations, adaptation des horaires de travail et, lorsque cela est possible, télétravail.

Recommander aux autorités organisatrices des transports de faciliter ou de faire faciliter l'utilisation des parkings relais de manière à favoriser l'utilisation des systèmes de transports en commun aux entrées d'agglomération.

Recommander de s'abstenir de circuler avec certaines catégories de véhicules en fonction de leur numéro d'immatriculation ou certaines classes de véhicules polluants définis selon la classification prévue à l'article R. 318-2 du code de la route, hormis les véhicules d'intérêt général visés à l'article R. 311-1 du code de la route.

Promouvoir auprès des acteurs concernés l'humidification, l'arrosage ou toute autre technique rendant les poussières moins volatiles et limitant leur remise en suspension. Cette opération est recommandée aux abords des axes routiers et dans tous autres lieux pertinents, soit avec récupération simultanée des poussières par aspiration ou par tout autre moyen, soit avec évacuation dans les eaux usées après avoir vérifié l'horaire le plus pertinent pour cet arrosage et hors période de gel ou de restriction des ressources en eau.

Sensibiliser le public aux effets négatifs sur la consommation et les émissions de polluants de la conduite « agressive » des véhicules et de l'usage de la climatisation ainsi qu'à l'intérêt d'une maintenance régulière du véhicule.

Recommander d'abaisser temporairement de 20 km/h les vitesses maximales autorisées sur les voiries localisées dans la zone concernée par l'épisode de pollution, sans toutefois descendre en-dessous de 70 km/h.

Recommander aux collectivités territoriales compétentes de rendre temporairement gratuit le stationnement résidentiel.

Recommander aux autorités organisatrices des transports de pratiquer ou de faire pratiquer des tarifs plus attractifs pour l'usage des transports les moins polluants (vélo, véhicules électriques, transports en commun...).

II. – **Mesures réglementaires de réduction des émissions de polluants en cas d'activation du niveau d'alerte**

1. *Secteur agricole*

Interdire les épandages de fertilisants minéraux et organiques ainsi que les travaux du sol, en tenant compte des contraintes déjà prévues par les programmes d'actions pris au titre de la directive 91/676/CEE. En cas de permanence de plus de trois jours de l'épisode de pollution et lorsque l'absence d'intervention sur les parcelles ou les cultures pénaliserait significativement la campagne culturale en cours ou entraînerait un non-respect d'autres dispositions réglementaires définies au titre du présent code, ces interdictions sont levées par le préfet. Le préfet peut alors, si la gravité de l'épisode de pollution l'exige, encadrer ces pratiques (limitation horaire dans la journée, recours à certaines techniques telles que l'injection, la rampe à pendillard ou l'enfouissement immédiat,...).

Interdire la pratique de l'écobuage.

Interdire, en cas d'un tel épisode de pollution de l'air ambiant, toute opération de brûlage à l'air libre des sous-produits de culture agricoles.

Rendre obligatoire le report des activités de nettoyage de silo ou tout événement concernant ce type de stockage susceptible de générer des particules, sous réserve que ce report ne menace pas les conditions de sécurité.

Rendre obligatoire le recours à des enfouissements rapides des effluents.

2. *Secteur résidentiel et tertiaire*

Interdire l'utilisation de certains foyers ouverts, appareils de combustion de biomasse non performants ou groupes électrogènes.

Interdire l'utilisation de barbecue à combustible solide.

Interdire totalement le brûlage des déchets verts à l'air libre : suspension des éventuelles dérogations.

3. *Secteur industriel*

Sur la base de plans d'actions en cas d'épisode de pollution de l'air définis par le préfet en concertation avec les acteurs concernés et contenant une étude préalable d'impact économique et social, rendre obligatoire pour les installations industrielles et les chantiers générateurs de poussières la mise en œuvre de dispositions de nature à réduire les rejets atmosphériques, y compris la baisse de leur activité, sous réserve que les conditions de sécurité soient préservées et que les coûts induits ne soient pas disproportionnés pour les acteurs publics et privés au regard des bénéfices sanitaires attendus.

Rendre obligatoire le report de certaines opérations émettrices de COV (travaux de maintenance, dégazage d'une installation, chargement ou déchargement de produits émettant des composants organiques volatils en l'absence de dispositif de récupération des vapeurs) à la fin de l'épisode de pollution.

Rendre obligatoire le report de certaines opérations émettrices de particules ou d'oxydes d'azote à la fin de l'épisode de pollution.

Rendre obligatoire le report du démarrage d'unités à l'arrêt à la fin de l'épisode de pollution sous réserve que les coûts induits ne soient pas disproportionnés.

Rendre obligatoire la mise en fonctionnement de systèmes de dépollution renforcés, lorsqu'ils sont prévus, pendant la durée de l'épisode de pollution.

4. *Secteur des transports*

Intensifier les contrôles de pollution des véhicules (y compris les deux-roues).

Limiter, voire interdire, la circulation dans certains secteurs géographiques, comme les zones urbaines denses, à certaines catégories de véhicules en fonction de leur numéro d'immatriculation ou certaines classes de véhicules polluants définis selon la classification prévue à l'article R. 318-2 du code de la route, hormis les véhicules d'intérêt général visés à l'article R. 311-1 du code de la route.

Limiter le trafic routier des poids lourds en transit dans certains secteurs géographiques, voire les en détourner en les réorientant vers des itinéraires de substitution lorsqu'ils existent, en évitant toutefois un allongement significatif du temps de parcours.

Abaisser temporairement de 20 km/h les vitesses maximales autorisées sur les voiries localisées dans la zone concernée par l'épisode de pollution, sans toutefois descendre au-dessous de 70 km/h.

Modifier le format des épreuves de sports mécaniques (terre, mer, air) en réduisant les temps d'entraînement et d'essais.

Raccorder électriquement à quai les navires de mer et les bateaux fluviaux en substitution à la production électrique de bord par les groupes embarqués, dans la limite des installations disponibles.

Limiter l'utilisation des moteurs auxiliaires de puissance des avions (APU) au strict nécessaire.

Utiliser les systèmes fixes ou mobiles d'approvisionnement électrique et de climatisation/chauffage des aéroports pour les aéronefs, dans la mesure des installations disponibles.

Réduire les émissions des aéronefs durant la phase de roulage par une attention particulière aux actions limitant le temps de roulage.

En cas de pic de pollution prolongé, le ministre chargé de l'aviation civile prend les mesures nécessaires pour tenir compte de la pollution due aux mouvements d'aéronefs et, le cas échéant, au transport terrestre associé.

Annexe B

Résultats des modèles hybrides après division par l'utilisateur

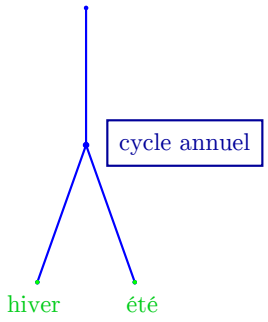
Cette annexe présente les résultats des principales expériences évoquées à la section 6.1 page 6.1. Il s'agit de modèles hybrides de prévision, constitués de plusieurs PMC, chacun entrainé sur un sous-ensemble du jeu de données d'apprentissage. La division du jeu de données est effectuée suivant une règle définie par l'utilisateur, illustré par un arbre. Ces schémas de résultats sont générés par l'application présenté à la section 7.1, page 7.1.

O_3 (horaire)
à Canetto
Horizon : h + 24

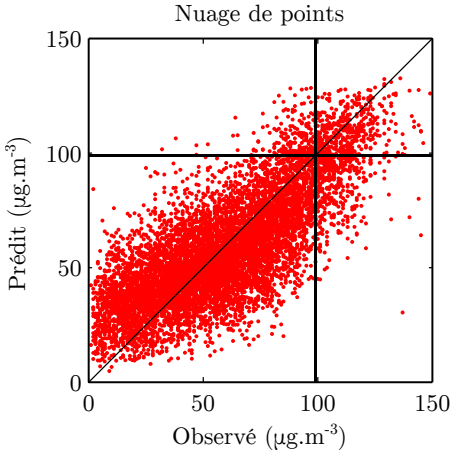
Set de test = 8425 points
Moyenne observé = 59.38 µg.m^{-3}
Moyenne prédit = 59.02 µg.m^{-3}

MSE = 342.06 µg.m^{-3} RMSE = 18.49 µg.m^{-3}
nRMSE = 31.14 % MAE = 14.64 µg.m^{-3}
MBE = -0.37 µg.m^{-3} FB = 0.01 FV = 0.38
R = 0.78 R^2 = 0.60 d = 0.871

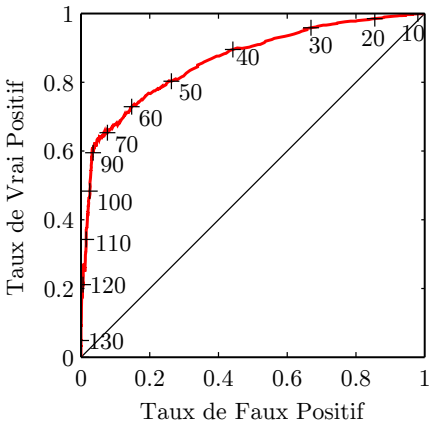
Division en fonction
de la saison



Nuage de points



Courbe ROC

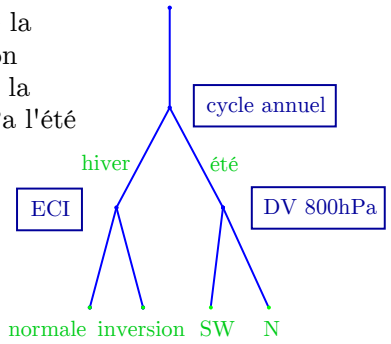


O_3 (horaire)
à Canetto
Horizon : h + 24

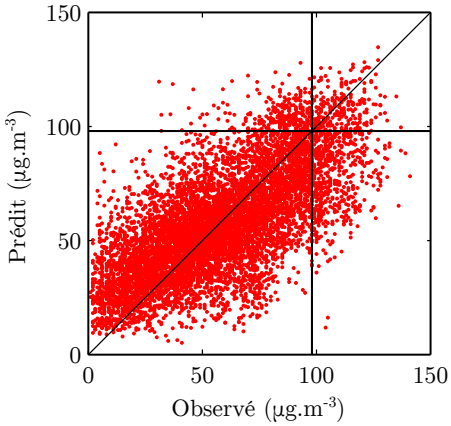
Set de test = 7913 points
Moyenne observé = 58.27 µg.m^{-3}
Moyenne prédit = 57.96 µg.m^{-3}

MSE = 439.91 µg.m^{-3} RMSE = 20.97 µg.m^{-3}
nRMSE = 35.99 % MAE = 16.40 µg.m^{-3}
MBE = -0.31 µg.m^{-3} FB = 0.01 FV = 0.40
R = 0.70 R^2 = 0.49 d = 0.823

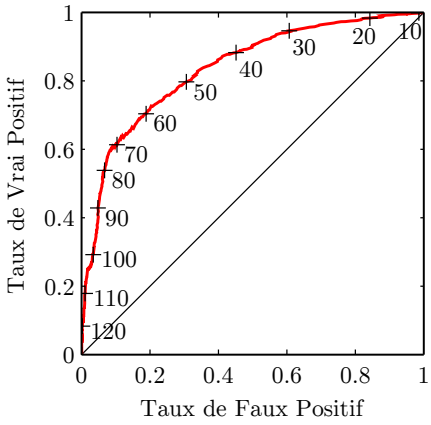
Division en fonction de la saison, et de l'inversion thermique l'hiver et de la direction du vent à 800 hPa l'été



Nuage de points



Courbe ROC

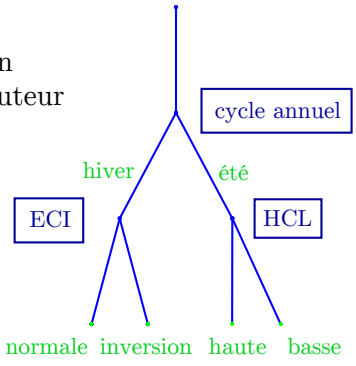


O_3 (horaire)
à Canetto
Horizon : h + 24

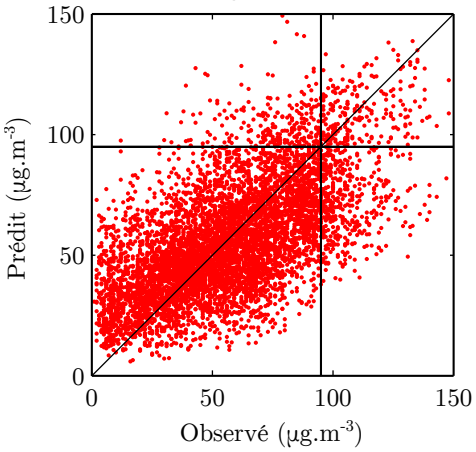
Set de test = 6651 points
Moyenne observé = 57.42 µg.m^{-3}
Moyenne prédit = 56.82 µg.m^{-3}

MSE = 507.88 µg.m^{-3} RMSE = 22.54 µg.m^{-3}
nRMSE = 39.25 % MAE = 17.64 µg.m^{-3}
MBE = -0.60 µg.m^{-3} FB = 0.01 FV = 0.39
R = 0.63 R^2 = 0.40 d = 0.784

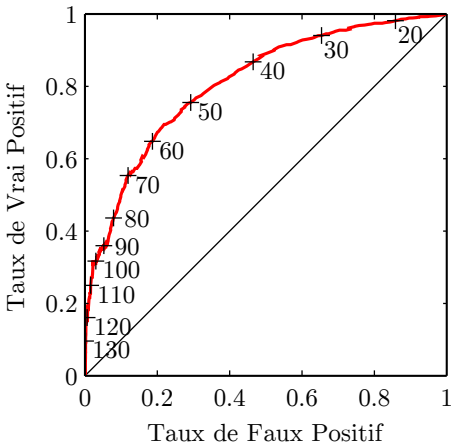
Division en fonction de la saison, et de l'inversion thermique l'hiver et de la hauteur de la couche limite l'été



Nuage de points



Courbe ROC

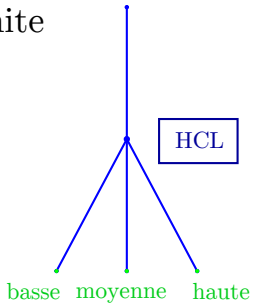


O3 (horaire)
à Canetto
Horizon : h + 24

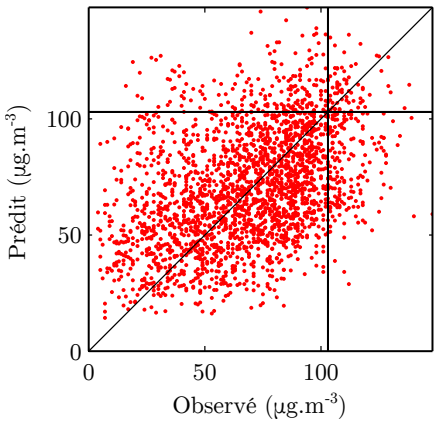
Set de test = 2367 points
Moyenne observé = 68.69 µg.m^{-3}
Moyenne prédit = 70.67 µg.m^{-3}

MSE = 792.86 µg.m^{-3} RMSE = 28.16 µg.m^{-3}
nRMSE = 40.99 % MAE = 22.08 µg.m^{-3}
MBE = 1.98 µg.m^{-3} FB = -0.03 FV = 0.23
R = 0.42 R^2 = 0.18 d = 0.663

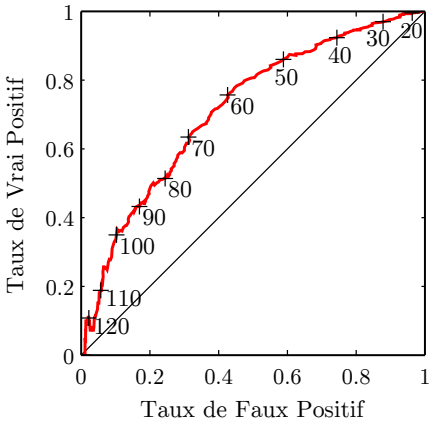
Division en fonction de la hauteur de la couche limite



Nuage de points



Courbe ROC

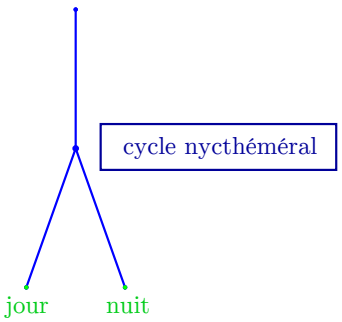


O$_3$ (horaire)
à Canetto
Horizon : h + 24

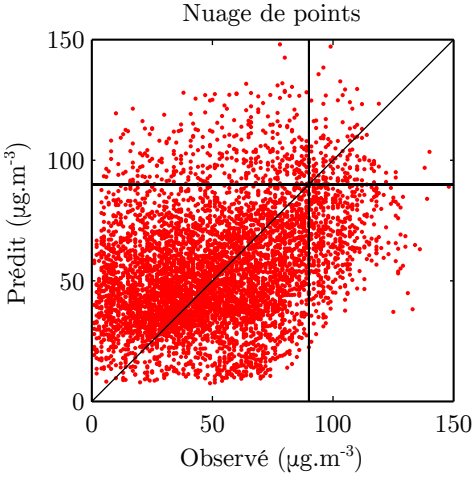
Set de test = 5791 points
Moyenne observé = 52.00 µg.m^{-3}
Moyenne prédit = 54.26 µg.m^{-3}

MSE = 873.40 µg.m^{-3} RMSE = 29.55 µg.m^{-3}
nRMSE = 56.83 % MAE = 23.35 µg.m^{-3}
MBE = 2.25 µg.m^{-3} FB = -0.04 FV = 0.35
R = 0.35 R^2 = 0.12 d = 0.612

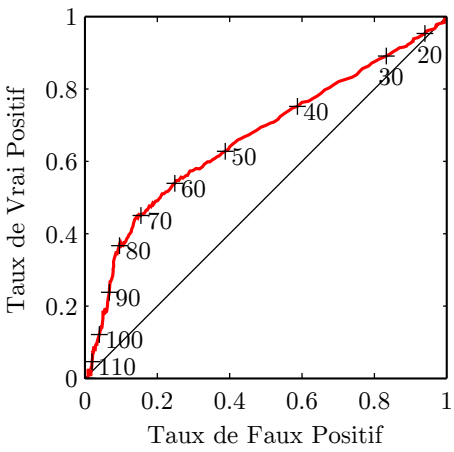
Division en fonction du cycle nycthéméral



Nuage de points



Courbe ROC

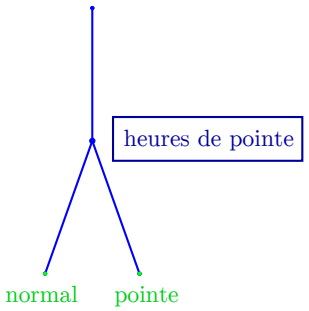


O_3 (horaire)
à Canetto
Horizon : h + 24

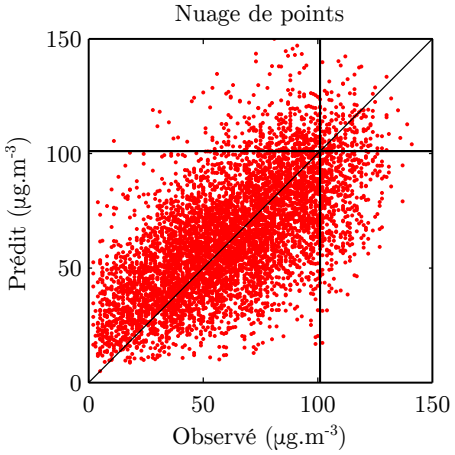
Set de test = 5974 points
Moyenne observé = 62.68 µg.m^{-3}
Moyenne prédit = 64.00 µg.m^{-3}

MSE = 536.70 µg.m^{-3} RMSE = 23.17 µg.m^{-3}
nRMSE = 36.96 % MAE = 18.11 µg.m^{-3}
MBE = 1.32 µg.m^{-3} FB = -0.02 FV = 0.25
R = 0.64 R^2 = 0.41 d = 0.793

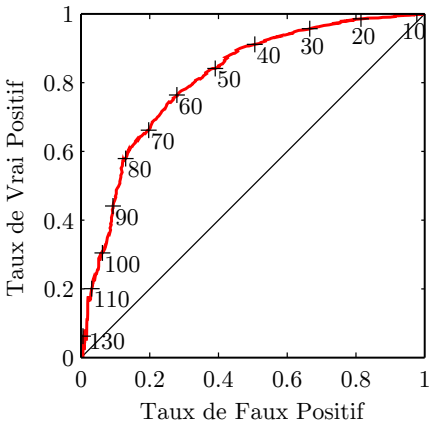
Division en fonction du trafic routier



Nuage de points



Courbe ROC

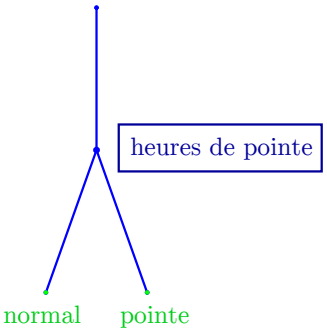


PM10 (moyenne 24 heures glissantes)
à Canetto
Horizon : h + 24

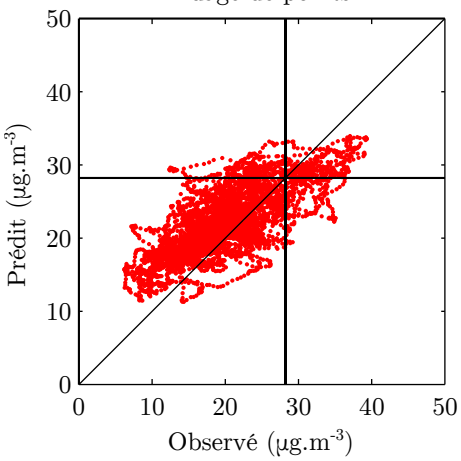
Set de test = 5902 points
Moyenne observé = 20.03 µg.m^{-3}
Moyenne prédit = 22.18 µg.m^{-3}

MSE = 22.01 µg.m^{-3} RMSE = 4.69 µg.m^{-3}
nRMSE = 23.42 % MAE = 3.86 µg.m^{-3}
MBE = 2.14 µg.m^{-3} FB = -0.10 FV = 0.70
R = 0.74 R^2 = 0.54 d = 0.795

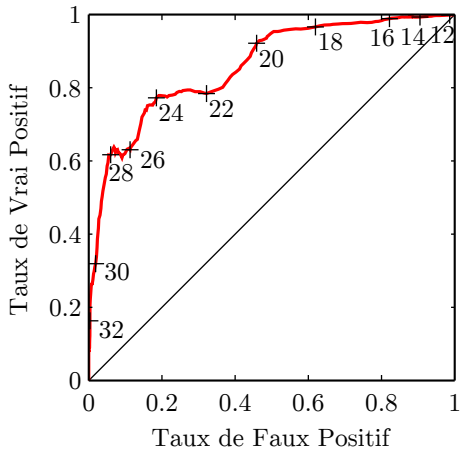
Division en fonction du trafic routier



Nuage de points



Courbe ROC



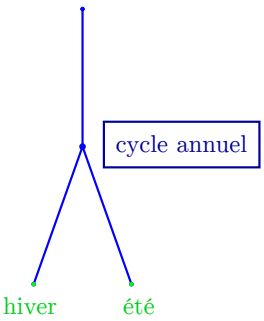
PM10 (moyenne 24 heures glissantes)
à Canetto

Horizon : h + 24

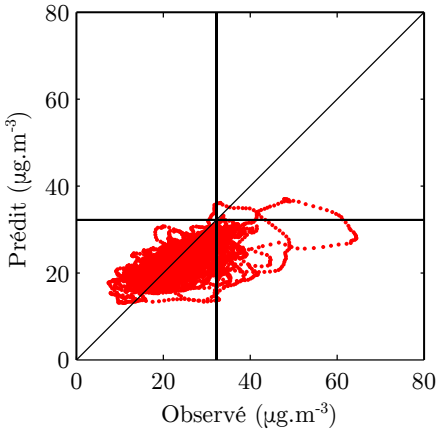
Set de test = 8277 points
Moyenne observé = 23.30 µg.m-3
Moyenne prédit = 22.13 µg.m-3

MSE = 31.20 $\mu g.m^{-3}$ RMSE = 5.59 $\mu g.m^{-3}$
nRMSE = 23.97 % MAE = 4.07 $\mu g.m^{-3}$
MBE = -1.17 $\mu g.m^{-3}$ FB = 0.05 FV = 1.14
R = 0.63 R^2 = 0.40 d = 0.702

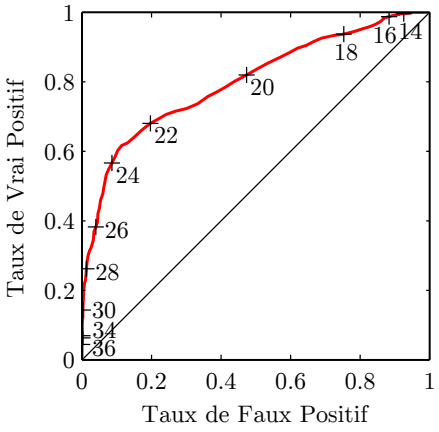
Division en fonction de la saison



Nuage de points



Courbe ROC



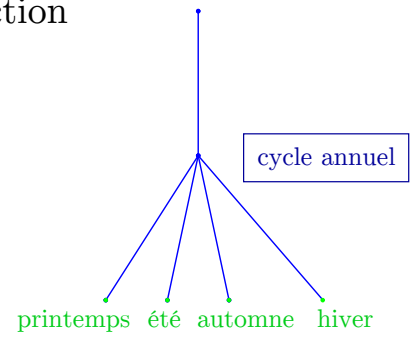
PM10 (moyenne 24 heures glissantes)
à Canetto

Horizon : h + 24

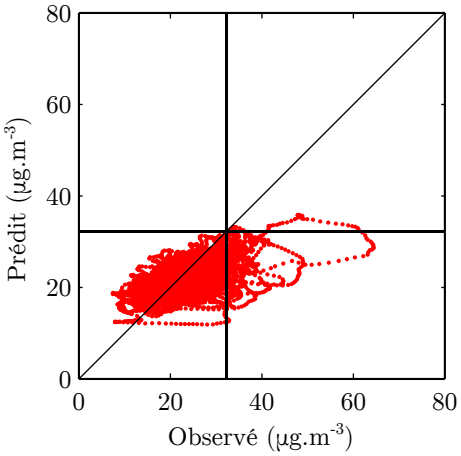
Set de test = 8286 points
Moyenne observé = 23.30 $\mu g.m^{-3}$
Moyenne prédit = 21.81 $\mu g.m^{-3}$

MSE = 33.23 $\mu g.m^{-3}$ RMSE = 5.76 $\mu g.m^{-3}$
nRMSE = 24.74 % MAE = 4.19 $\mu g.m^{-3}$
MBE = -1.48 $\mu g.m^{-3}$ FB = 0.07 FV = 1.11
R = 0.61 R^2 = 0.37 d = 0.689

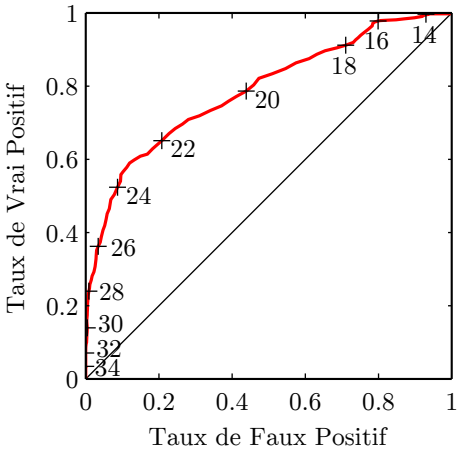
Division en fonction de la saison



Nuage de points



Courbe ROC

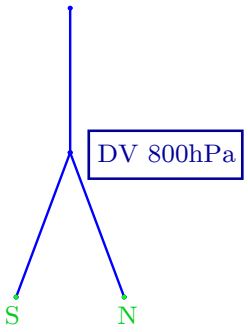


PM10 (moyenne 24 heures glissantes)
à Canetto
Horizon : h + 24

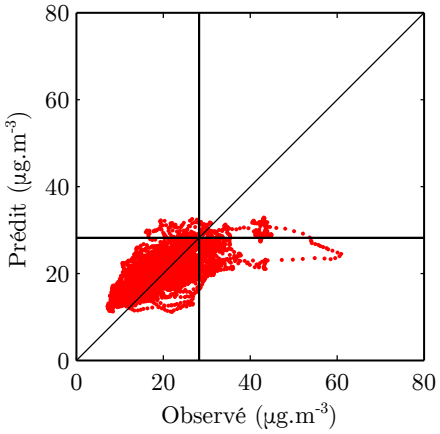
Set de test = 6815 points
Moyenne observé = 20.04 µg.m^{-3}
Moyenne prédit = 20.66 µg.m^{-3}

MSE = 26.66 µg.m^{-3} RMSE = 5.16 µg.m^{-3}
nRMSE = 25.77 % MAE = 3.88 µg.m^{-3}
MBE = 0.62 µg.m^{-3} FB = -0.03 FV = 0.93
R = 0.66 R^2 = 0.43 d = 0.750

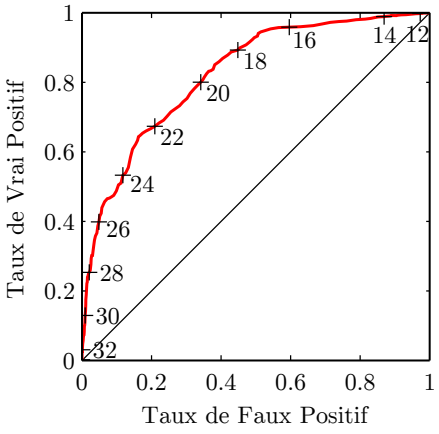
Division en fonction de la direction du vent à 800 hPa



Nuage de points



Courbe ROC

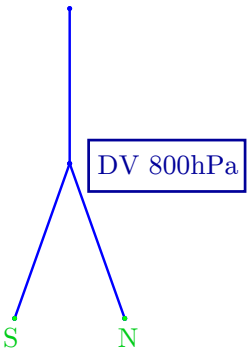


PM10 (moyenne 24 heures glissantes)
à Caneto
Horizon : h + 24

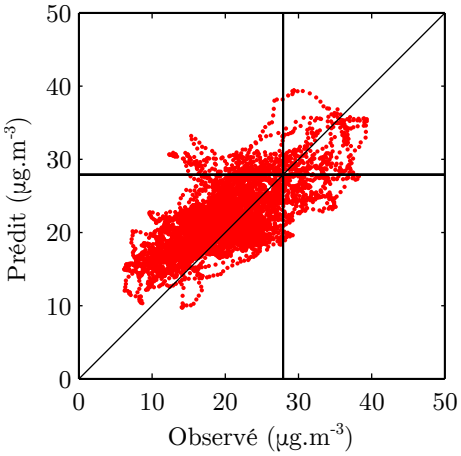
Set de test = 6627 points
Moyenne observé = 20.07 µg.m^{-3}
Moyenne prédit = 22.08 µg.m^{-3}

MSE = 22.18 µg.m^{-3} RMSE = 4.71 µg.m^{-3}
nRMSE = 23.46 % MAE = 3.85 µg.m^{-3}
MBE = 2.01 µg.m^{-3}FB = -0.10 FV = 0.61
R = 0.70 R^2 = 0.49 d = 0.785

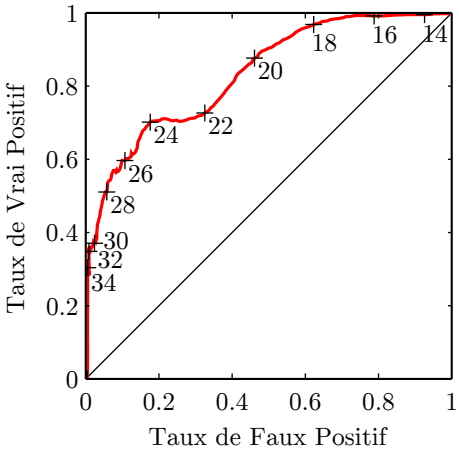
Division en fonction de la direction du vent à 800 hPa (sans HCL en entrée)



Nuage de points



Courbe ROC

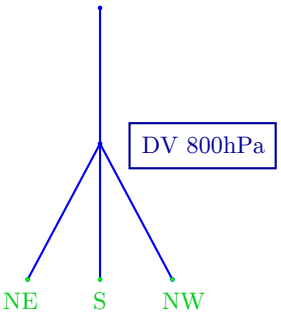


PM10 (moyenne 24 heures glissantes)
à Canetto
Horizon : h + 24

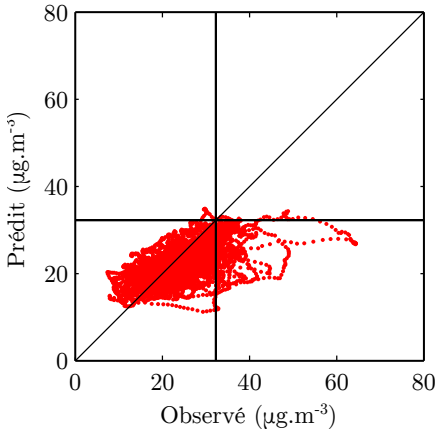
Set de test = 8268 points
Moyenne observé = 23.30 µg.m^{-3}
Moyenne prédit = 21.83 µg.m^{-3}

MSE = 33.24 µg.m^{-3} RMSE = 5.77 µg.m^{-3}
nRMSE = 24.74 % MAE = 4.14 µg.m^{-3}
MBE = -1.47 µg.m^{-3} FB = 0.07 FV = 1.05
R = 0.60 R^2 = 0.36 d = 0.696

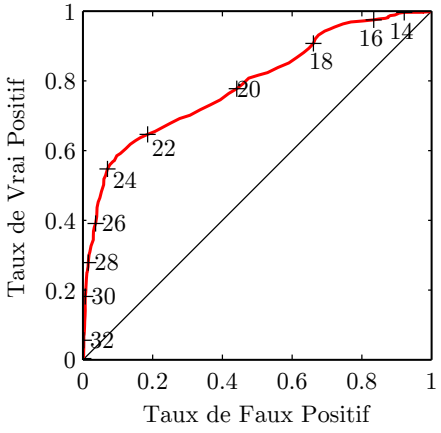
Division en fonction de la direction du vent à 800 hPa



Nuage de points



Courbe ROC

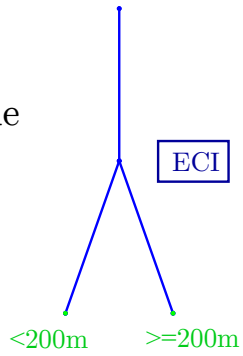


PM10 (moyenne 24 heures glissantes)
à Canetto
Horizon : h + 24

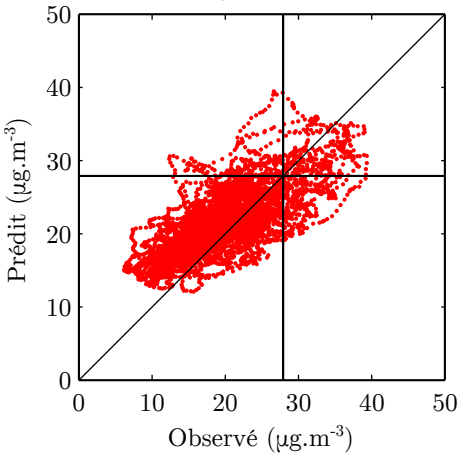
Set de test = 6627 points
Moyenne observé = 20.07 µg.m^{-3}
Moyenne prédit = 21.78 µg.m^{-3}

MSE = 20.98 µg.m^{-3} RMSE = 4.58 µg.m^{-3}
nRMSE = 22.82 % MAE = 3.72 µg.m^{-3}
MBE = 1.71 µg.m^{-3} FB = -0.08 FV = 0.64
R = 0.70 R^2 = 0.50 d = 0.790

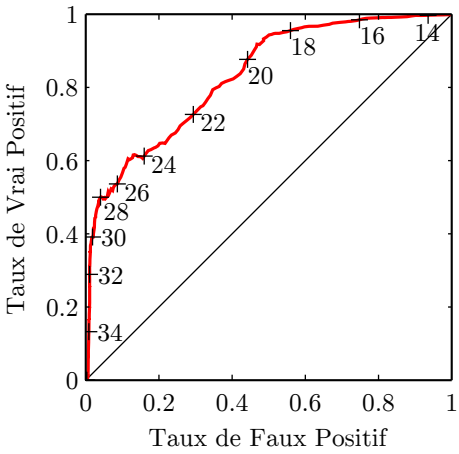
Division en fonction de l'épaisseur de la couche d'invertion thermique



Nuage de points



Courbe ROC

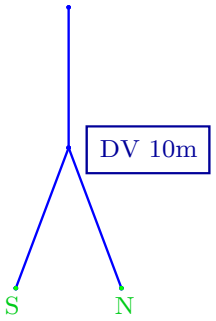


PM10 (moyenne 24 heures glissantes)
à Canetto
Horizon : h + 24

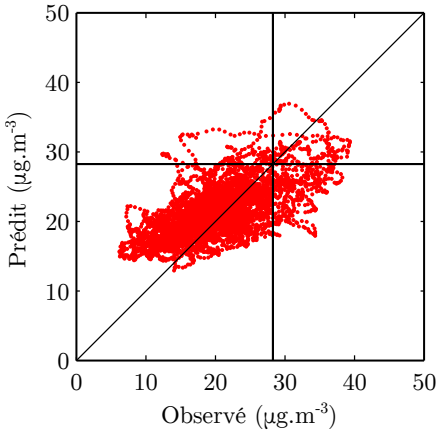
Set de test = 6196 points
Moyenne observé = 20.07 µg.m^{-3}
Moyenne prédit = 21.43 µg.m^{-3}

MSE = 23.42 µg.m^{-3} RMSE = 4.84 µg.m^{-3}
nRMSE = 24.12 % MAE = 3.90 µg.m^{-3}
MBE = 1.36 µg.m^{-3} FB = -0.07 FV = 0.94
R = 0.66 R^2 = 0.43 d = 0.738

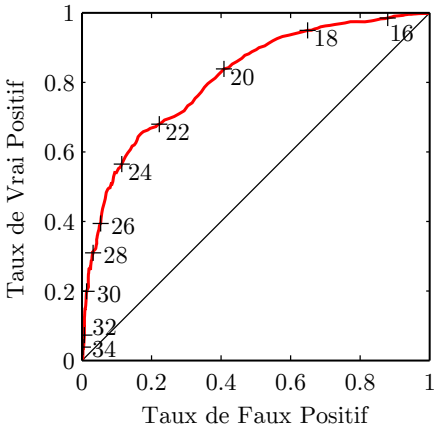
Division en fonction de la direction du vent à 10m



Nuage de points



Courbe ROC

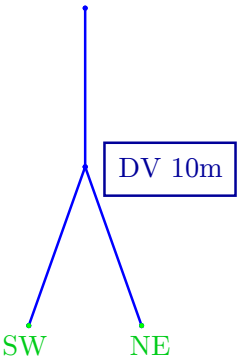


PM10 (moyenne 24 heures glissantes)
à Giraud
Horizon : h + 24

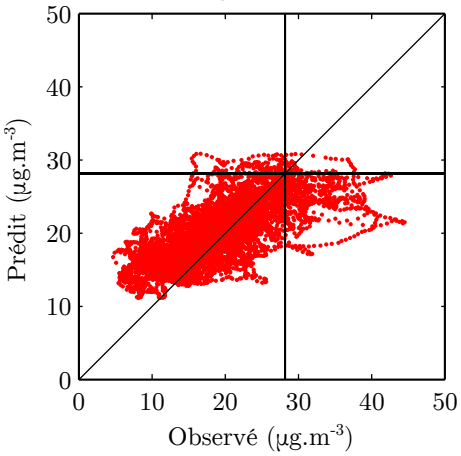
Set de test = 7462 points
Moyenne observé = 19.24 µg.m^{-3}
Moyenne prédit = 20.21 µg.m^{-3}

MSE = 25.05 µg.m^{-3} RMSE = 5.01 µg.m^{-3}
nRMSE = 26.02 % MAE = 3.91 µg.m^{-3}
MBE = 0.98 µg.m^{-3} FB = -0.05 FV = 0.93
R = 0.72 R^2 = 0.52 d = 0.782

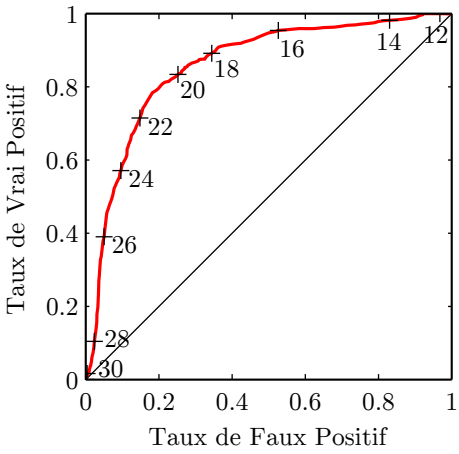
Division en fonction de la direction du vent à 10m



Nuage de points



Courbe ROC



Annexe C

Fonctionnement de l'application Aria Base

Nous présenterons ici en détails les possibilités offertes par l'application Aria Base que nous avons développée sous Matlab.

Nous nous pencherons d'abord sur la première des tâches de notre application, qui consiste en la gestion de nos jeux de données. Tous au long du doctorat, nous avons reçus des données quotidiennement de la part de nos fournisseur. Cette application a permis de les récupérer automatiquement et de les ajouter aux données existantes. Nos données ont été regroupées en un système de fichier faisant office de base de données interne.

Ensuite, nous aborderons la modélisation à l'aide de réseaux de neurones avec cette application. On verra quels choix de configuration sont proposés à l'utilisateur, et quelles possibilités de gestion des expériences sont offertes. Les possibilités d'archivage des résultats et d'export des modèles seront développées.

Nous présenterons le volet dédié à la gestion des modèles hybrides, c'est-à-dire d'un ensemble de RNA spécialisés chacun sur une classe de données, définie par un modèle de classification. Constitués de plusieurs modèles prédictifs, la construction de ces modèles hybrides gagne en effet à être encadrée. Ce volet permet de superviser les opérations de divisions des données, supervisées ou non.

Enfin le volet « Evaluation » qui permet une étude poussée des performances des modèles et qui gère l'évaluation des modèles hybrides sera présenté.

C.1 Volet « Base de données »

Nous utilisons des données de plusieurs fournisseurs pour la prévision de la qualité de l'air. Tout d'abord, les données de concentration en polluant et certaines mesures météorologiques proviennent des stations de Qualitair Corse. Ensuite, la plupart des mesures météorologiques et les sorties des modèles AROME et ARPEGE sont fournies par Météo-France, dans le cadre d'une convention passée avec Qualitair Corse. Air PACA enfin nous fournit les sorties du modèle AIRES, également par le biais d'une convention entre les AASQA.

Les fichiers de données des fournisseurs ont chacun un format différent. Les données des stations Météo-France se présentent sous forme de fichiers textes regroupant l'ensemble des variables sur l'ensemble des stations. Afin d'obtenir les données à temps pour réaliser les prévisions

dans le cadre de l'arrêté du 26 mars 2014 (voir annexe A), il est nécessaire d'en disposer 11h UTC, car la prévision doit être disponible pour midi (précisé par la circulaire d'application de l'arrêté). Pour cela, nous avons fait en sorte de recevoir les données à 9h UTC soit à 10h heure d'hiver et 11h heure d'été, afin de réaliser les prévisions pour midi, comme le demande la circulaire. La figure C.1 montre comment l'application permet de visualiser le contenu d'un fichier de données de stations, via le menu « Fichier ».

Les données des modèles ARPEGE et AROME (voir section 3.2.6, page 78) sont elles fournies au format GRIB (pour General Regularly-distributed Information in Binary form) et correspondent à des fichiers textes regroupant l'ensemble des prévisions du modèle entre $h+0$ et $h+48$ (ARPEGE) ou $h+30$ (AROME), avec un pas de trois heures. Les données sont constituées en blocs qui regroupent les prévisions d'une variable sur toute la grille de point horizontale, à une altitude donnée pour une échéance donnée, ces informations étant décrites dans un en-tête qui précède le bloc. La grille de points d'ARPEGE est constituée d'un point tout les 0.25° de latitude et de longitude (environ 25 km), AROME a une grille contenant un point tous les 0.025° de latitude et de longitude (environ 2.5 km). La figure C.2 montre le volet de l'application qui assure la lecture des fichiers GRIB.

Cette lecture de fichier permet de visualiser les données que l'on récupère, mais ce qui nous intéresse le plus c'est la récupération des données et la mise à jour de séries temporelles pour chaque variable. Les données que nous récupérons sous forme de séries temporelles sont réunies dans un dossier. Chaque série est sauvegardée au format de matlab .mat et est accessible via la page « Base de données » de l'application, à laquelle on accède par le menu éponyme. La figure C.3 présente une capture d'écran de la page permettant cette gestion.

Cette page permet d'avoir accès à l'ensemble des données qui composent la base. Dans le cadre « Base de données », on accède aux données après sélection de l'organisme fournisseur, puis soit par station, soit par type de variable. Des informations concernant la station sont alors données, et sa position sur la carte s'affiche. Sont indiqués le type de station, les dates de début et fin de la série temporelle. On peut choisir les dates de début et de fin de la série que l'on veut charger. Ces dates peuvent être fixées de manière à extraire d'autres séries temporelles avec ces mêmes dates.

Les séries temporelles sont alors chargées dans la mémoire de l'application. Le cadre « Données », qui est présent sur tous les autres volets de l'application, indique les séries temporelles chargées, et trace un aperçu des séries temporelles. On peut les exporter au format .xls ou les charger dans l'espace de travail Matlab.

Météo-France et Air PACA mettent leurs données à disposition sur un serveur FTP. Matlab permet d'effectuer des transferts par FTP, ce qui nous a conduits à automatiser la récupération des fichiers quotidiens, via cette page. Les fichiers des stations de Météo-France sont alors lus et leurs données sont fusionnées aux séries temporelles existantes dans la base. L'import peut également se faire via un fichiers xls.

Dans le cas des sorties de modèle, l'ensemble des points auxquels sont disponibles les variables est trop important pour se permettre d'extraire automatiquement les données et de les stocker sous forme de fichier .mat sans utiliser une trop grosse part d'un disque dur d'ordinateur portable (dans un fichier d'AROME, une variable à une altitude donnée possède 67771 points de données). La récupération automatique de données télécharge donc uniquement les fichiers bruts de sortie de modèle. L'utilisateur de l'application peut extraire de ces fichiers la série temporelle de son choix correspondant à une variable pour une coordonnée géographique (par exemple les coordonnées d'une station de Qualitair Corse).

Cette gestion des données permet de garder à jour ses jeux de variables. C'est important car

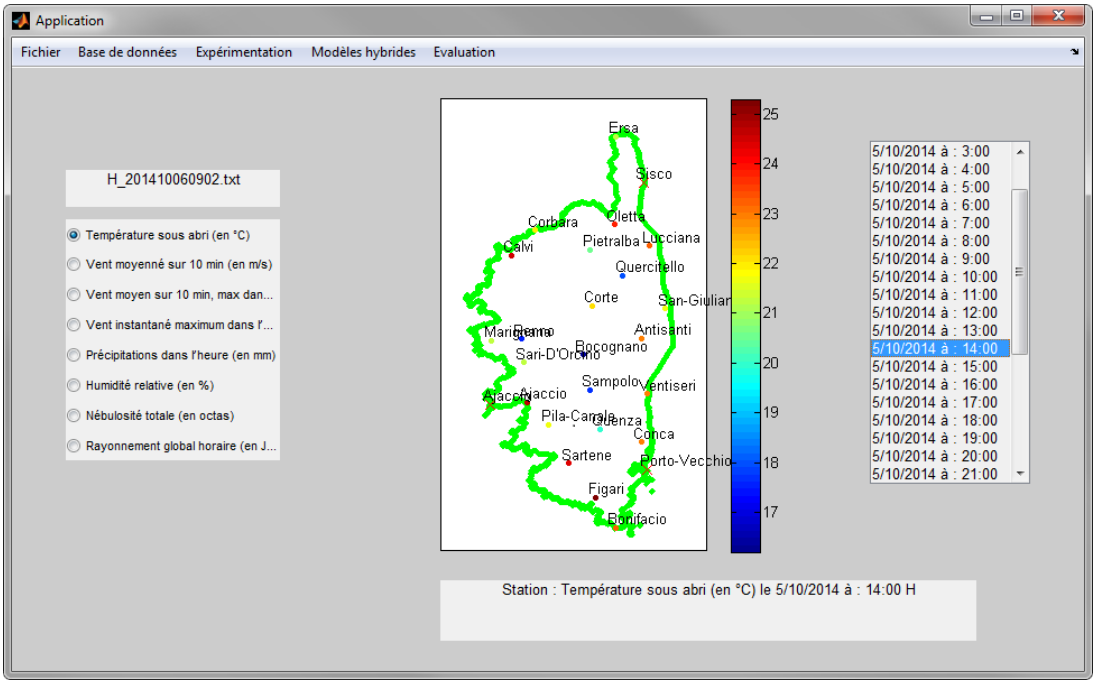


FIGURE C.1 : Capture d'écran de la page de visualisation des données de stations Météo-France. Ici, visualisation de la température le 5 octobre à 14h sur toutes les stations disponibles en corse.

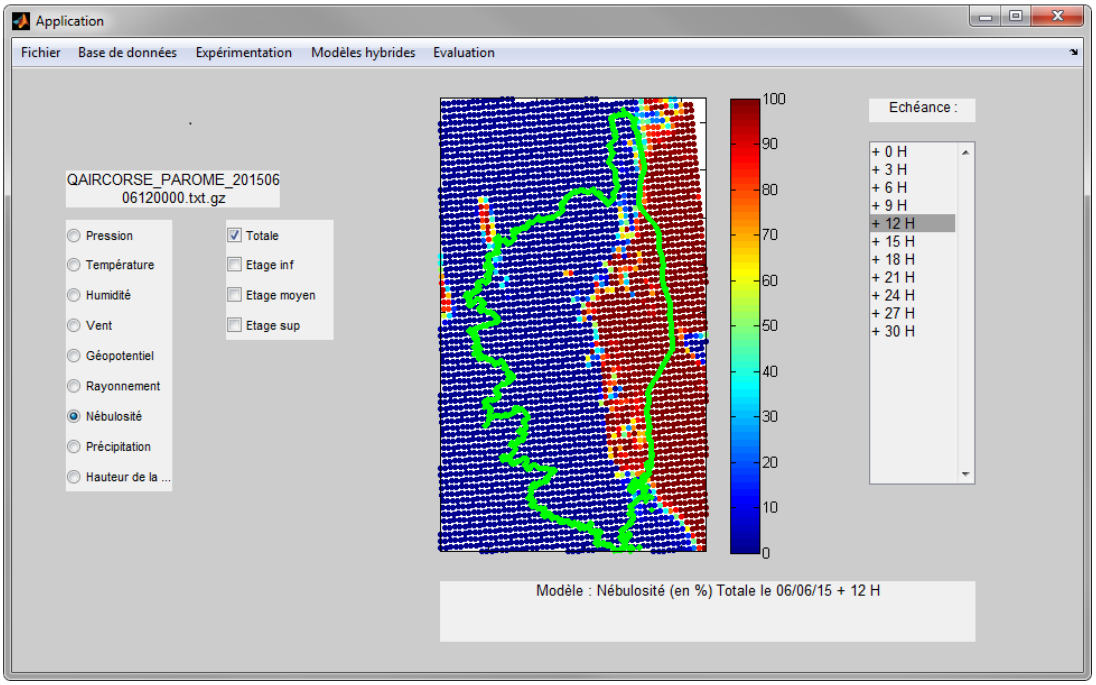


FIGURE C.2 : Capture d'écran de la page de visualisation des données de modèles au format GRIB.

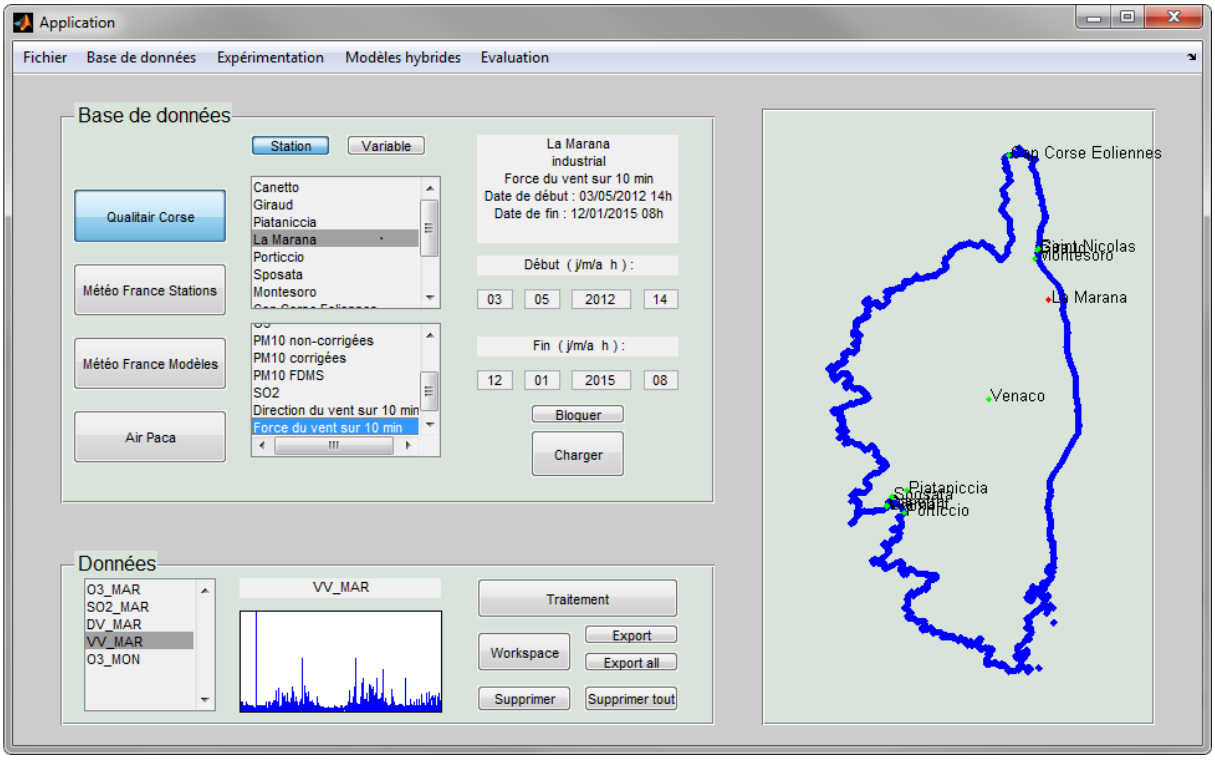


FIGURE C.3 : Capture d'écran du volet « Base des données ».

la taille de l'échantillon est cruciale pour l'apprentissage des réseaux neuronaux. Cela permet, quand un modèle opérationnel a été créer, de le réentrainer facilement quand suffisamment de nouvelles données sont disponibles, afin d'éviter qu'il ne devienne obsolète. Voyons maintenant la partie de l'application qui permet de créer les modèles prédictifs.

C.2 Volet « Expérimentation »

La page consacrée à la création de modèle permet de gérer les expérimentations nécessaires à la configuration du réseau. Cette page est présentée à la figure C.4. Cette page est composée de plusieurs cadres. Tout d'abord, le cadre « Données » en bas à gauche est le même que dans la page « Base de données » et montre les données qui on été chargées de la base et peuvent être utilisées ici.

Le cadre 1 de la figure C.4 propose tout d'abord de choisir le nom de la configuration, qui sera utilisé pour l'archivage. Il permet de choisir les séries temporelles en entrée du modèle et ses cibles, c'est à dire les variables qui vont être prédites. Ces données peuvent avoir été chargées précédemment depuis la base de donnée de l'application et venir du cadre « Données », ou être importées de l'espace de travail Matlab, soit séparément, soit groupées. Les séries temporelles peuvent être horaires ou journalière.

Il est possible d'indiquer une valeur faisant office de seuil de concentration, à partir duquel les valeurs sont considérées comme un dépassement de seuil. Si c'est le cas, lors de l'évaluation des scores de détection seront calculés pour ce seuil (voir section 2.4.3 page 2.4.3). On peut également choisir un « sous-groupe », c'est-à-dire un sous-échantillon de l'ensemble du jeu de données formé par les séries temporelles sélectionnées. Cette fonctionnalité est utilisée pour entrainer les modèles spécifiques à une classe lors de la confection de modèles hybrides tels que

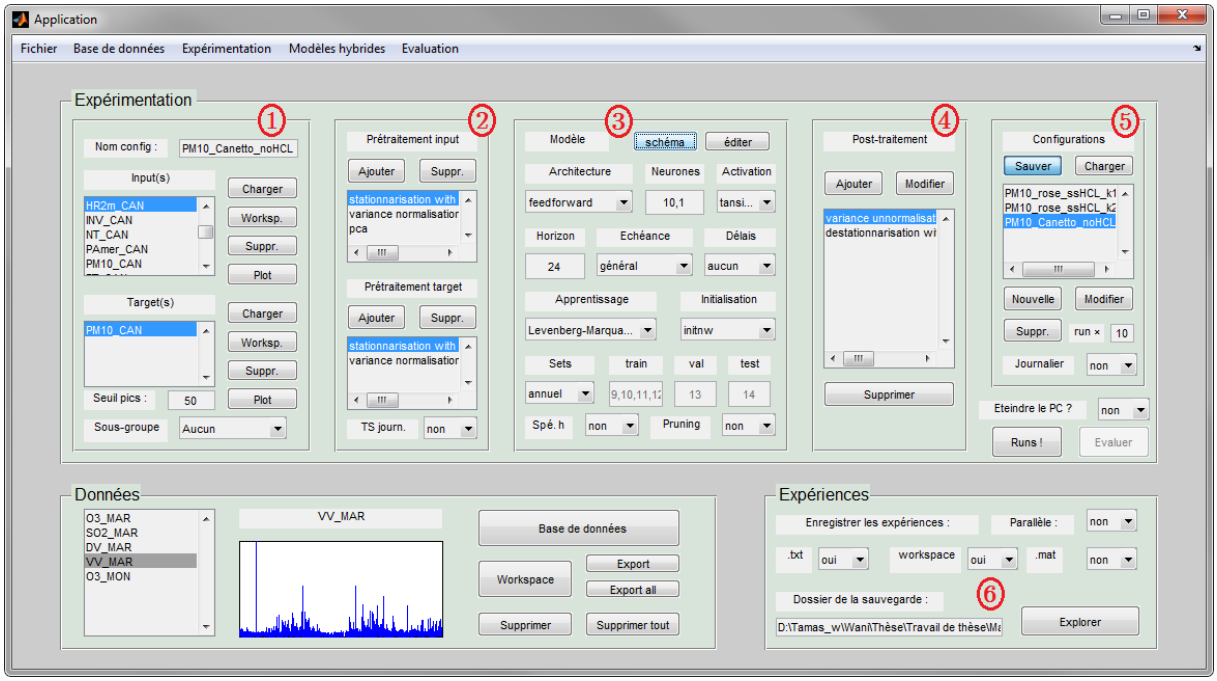


FIGURE C.4 : Capture d'écran du volet « Expérimentation ».

ceux présentés au chapitre 6 (page 148).

Une fois les entrées et les cibles choisies, le cadre 2 permet d'indiquer tous les prétraitements qui devront être appliqués, dans l'ordre, respectivement aux entrées et aux cibles. Quand un prétraitement est sélectionné, l'utilisateur précise à quelles séries temporelles il est destiné ainsi que ses éventuels paramètres. L'option « TS journ. » en bas du cadre permet de transformer des séries temporelles horaires en séries journalières.

Le cadre 3 permet la paramétrisation de modèle lui-même. Il est prévu pour l'usage de RNA issus de la NNToolbox, mais d'autres modèles peuvent facilement être substitués de manière compatible avec le reste de l'application. On peut choisir le type d'architecture du réseau (« feed-forward », bouclé, Radial Basis Function (RBF), Time-Delay Neural Network (TDNN),...), le nombre de couches et le nombre de neurones qui les composent ainsi que les fonctions de transfert des neurones. Le réseau de neurone peut être visualisé grâce au bouton « schéma ». Si l'on veut modifier plus profondément la structure du réseau, le bouton « éditer » permet d'exporter ce dernier vers l'espace de travail de Matlab, ou l'on peut accéder directement à l'objet créé par la NNToolbox, avant que l'application ne le récupère.

On peut ensuite choisir l'horizon de la prévision, afin de décaler dans le temps la série temporelle cible du nombre de pas de temps nécessaire. Sans horizon défini, la cible n'est pas décalée dans le temps, ce qui est adapté à d'autres problèmes que la prévision de série temporelle, par exemple à des problèmes d'estimation. Les données d'entrée peuvent également être décalées, si besoin. Dans le cas d'entrées qui correspondent à des sorties de modèle prédictif, on peut avoir intérêt à utiliser une échéance de prévision en particulier. On peut également vouloir utiliser une série temporelle avec un certain délai (voir section 4.2 sur le fonctionnement du PMC). Ces deux opérations sont proposées. Les séries temporelles sélectionnées seront ensuite décalées d'autant de pas de temps que nécessaire.

On choisi également ici l'algorithme d'apprentissage. Sont disponibles LM, DG, BFGS et SCG qui sont tous proposés par la NNToolbox. On peut également choisir la méthode d'initialisation des neurones (aléatoire, algorithme de Nguyen-Widrow).

Plusieurs options sont ensuite proposées pour définir la séparation entre les données d'apprentissage, les données de validation et les données de test. On peut tout d'abord les séparer par bloc, en indiquant le pourcentage de données qui doit être alloué à chaque ensemble. Dans l'ordre chronologique, un premier bloc de données est alloué à l'apprentissage, ensuite un autre l'est pour la validation et les données restantes sont consacrées au test. Les jeux de données peuvent également être constitués aléatoirement, toujours en respectant le pourcentage relatif pour chaque ensemble et en respectant les éventuels délais entres les variables. On peut constituer les ensembles de manière entrelacée. Du premier au dernier point des séries temporelles (dans l'ordre), chaque point de donnée est successivement assigné à un ensemble, de manière à respecter les proportions données (par exemple, trois points pour le jeu d'apprentissage, puis le suivant pour le jeu de validation, puis le suivant pour le jeu de test, puis les trois suivant pour le jeu d'apprentissage, etc.). Il est également proposé de séparer les données par années. On choisit dans ce cas pour chaque année présente dans le jeu de données à quel ensemble elle sera allouée. Enfin, il est possible d'importer depuis l'espace de travail une série temporelle définissant les dates qui doivent correspondre à chaque jeu.

L'option « Spé h » permet de spécialiser la prévision à une certaine heure du jour. Au lieu d'entrainer un modèle à la prévision à un certain horizon, on le spécialise dans la prévision pour un certain horaire. On peut ainsi obtenir un modèle entrainé pour prévoir, avec les données de minuit, les concentrations à 14h. Cette option est un reliquat d'une expérience (non fructueuse) visant à spécialiser les réseaux de neurones sur des phénomènes propres à certaines heures du jour. L'option « Pruning » permet d'utiliser la méthodologie d'élagage présentée à la section 5.2 (page 127).

Le cadre 4 propose les options de post-traitements. Elles sont nécessaires si la cible a subit un prétraitement, afin que les prévisions correspondent à la variables cible non prétraitée. Ces post-traitement utilisent les éléments de configuration ou certains résultats des prétraitements qui sont enregistrés (par exemple, la variance et la moyenne d'une série temporelle qui a été normalisée).

Tous ces éléments permettent de préparer l'apprentissage d'un RNA. Quand la saisie est terminée, la configuration est fixée mais aucun calcul numérique n'a encore été réalisé, l'expérience est prête à être lancée. Le bouton « Nouvelle » du cadre 5 crée un objet, construit pour recueillir l'ensemble de ces éléments de configuration. Ce type d'objet, nommé « Archive », possède des attributs propres à accueillir toutes les informations nécessaires au lancement de l'expérience. Ses méthodes lui permettent de réaliser toutes les opérations qui ont été prévues.

Cet objet peut être utilisé sous Matlab indépendamment de l'application. Lorsque l'on désire lancer l'expérience, les opérations prévues sont réalisées dans l'ordre adéquat :

- Décalage temporel des entrées
- Décalage temporel des cibles
- Prétraitement des entrées
- Prétraitement des cibles
- Construction du RNA
- Option de sélection d'un sous-groupe (modèle hybride)
- Option de spécialisation horaire
- Option de transformation en séries temporelles journalières
- En cas de non-gestion des valeurs manquantes en prétraitement, suppression des NaN
- Division des données en jeu d'apprentissage, de validation et de test
- Option d'élagage du RNA
- Apprentissage avec « early stopping »

– Evaluation du modèle avec le jeu de test, subissant les pré et post-traitements

On remarque que le modèle est construit après les prétraitements. Ceci est prévu pour le cas où une opération comme une ACP réduirait le nombre de variables d'entrée, qui doit être défini avant de créer un modèle avec la NNToolbox.

A la suite de ces opération, le modèle est prêt a être utilisé de manière opérationnelle, en lui soumettant de nouvelles données d'entrée. Toutes les informations sont conservées (durée de l'expérience, données utilisées, données des pré et post-traitements subis, information sur l'apprentissage, résultats de l'évaluation, etc.). Il est alors possible d'exporter ce réseau vers l'application opérationnelle que l'on a présentée à la section 7.2, qui elle se charge de réaliser les prévisions quotidiennement avec le modèle entrainé et les nouvelles données.

Quand l'objet « Archive » est créé par le bouton « Nouvelle », la configuration correspondante rejoint la liste des configurations prêtes à être lancées, dans la liste du cadre 5. On peut alors créer autant de nouvelles configurations et les ajouter à cette liste, les modifier ou en supprimer. Quand toutes les configurations à expérimenter son prêtes, on peut sauvegarder cet ensemble grâce au bouton « Sauver » en haut du cadre 5. Il est à l'inverse possible de charger un ensemble de configuration préalablement sauvegardé via le bouton « Charger ».

Avant de lancer les expériences, on peut indiquer combien de fois on veut lancer chaque configuration. On peut ainsi entrainer et évaluer chaque modèle plusieurs fois et conserver l'ensemble des résultats, ce qui permet d'étudier la robustesse de l'apprentissage. En effet, l'initialisation des paramètres comprenant une part d'aléatoire, la même configuration donnera des résultats différents. Le bouton « Run ! » lance l'ensemble des expériences, répétées le nombre de fois indiqué. Une fois l'expérience terminée, le bouton « Evaluer » permet de passer à la page d'évaluation présentée plus bas. L'option « Eteindre le PC ? » propose d'arrêter l'ordinateur quand l'ensemble des calculs et des sauvegardes aura été réalisé.

L'option « Journalier » est prévu pour un cas particulier, celui où l'on veut entraîner vingt-quatre modèles de prévision horaire afin d'avoir une prévision sur une journée entière. Activée, elle facilite l'affichage des résultats correspondants.

Le temps d'entrainement d'un RNA est relativement cours, de l'ordre de la minute à la dizaine de minutes. Mener les expériences avec cette application permet d'entrainer un grand nombre de RNA, par exemple la nuit. Travailler ainsi est intéressant, car de très nombreux paramètres sont importants pour l'utilisation de RNA et il est souvent nécessaire de tester plusieurs configurations.

D'autres options accompagnent ces expériences, qui sont présentées dans le cadre « Expériences » (cadre 6). Tout d'abord, on peut paralléliser ces calculs. Ceci est géré de manière simple grâce à la boucle « parfor » de Matlab. Cette boucle est une alternative à la classique boucle « for », utilisable dans le cas où chaque itération de la boucle est indépendante, c'est-à-dire n'a pas besoin des résultats d'itérations précédentes. Avec « parfor », les itérations sont menées en parallèle. Chaque calcul en parallèle est effectué par un « worker » correspondant à un cœur.

Les autres options concernent la sauvegarde des résultats. L'application peut tout d'abord écrire le résumé des opérations effectuées par chaque modèle dans un log. Toutes les informations permettant de connaitre la configuration et ses résultats sont renseignées dans le log, dont on précise l'adresse dans le champ prévu. Les différents objets « Archive » qui contiennent les modèles entrainés peuvent également être exportés vers l'espace de travail de Matlab, ou sauvegardés sur le disque dur de l'ordinateur.

L'ensemble de ces fonctionnalités nous a facilité l'exploration d'un grand nombre de possibilités de configuration durant nos travaux. Elle permet, quand on fait face à un nouveau problème

de prévision, de mener les expériences nécessaires à l'isolement d'un modèle prédictif optimal en définissant une liste d'expériences à mener. Elle permet également de maintenir à jour des modèles, avec la mise à dispositions des nouvelles données des jeux de variables utilisés, grâce à la page de gestion des données.

C.3 Volet « Modèles hybrides »

Lors de nos travaux, nous nous sommes tournés vers l'usage de modèles hybrides (chapitre 6 page 148). Du fait du couplage entre les modèles classifieur et prédictifs qui composent un modèle hybride, les opération d'apprentissage et d'évaluation sont plus délicates que dans le cas d'un RNA seul. L'entrainement de ces modèles se produit en deux phases. La première consiste en une séparation des données en plusieurs classes. La seconde consiste en l'entrainement d'un PMC pour chacune des classes de données. Pour l'évaluation, la classe correspondant à chaque point du jeu de test est identifiée préalablement et le PMC prédictif adéquat est utilisé pour chaque point.

Nous avons prévu les fonctionnalités permettant de gérer efficacement les étapes nécessaires à l'obtention et l'évaluation de tels modèles. Ce volet permet de définir la manière dont on veut assurer la séparation des données ; il est présentée à la figure C.5.

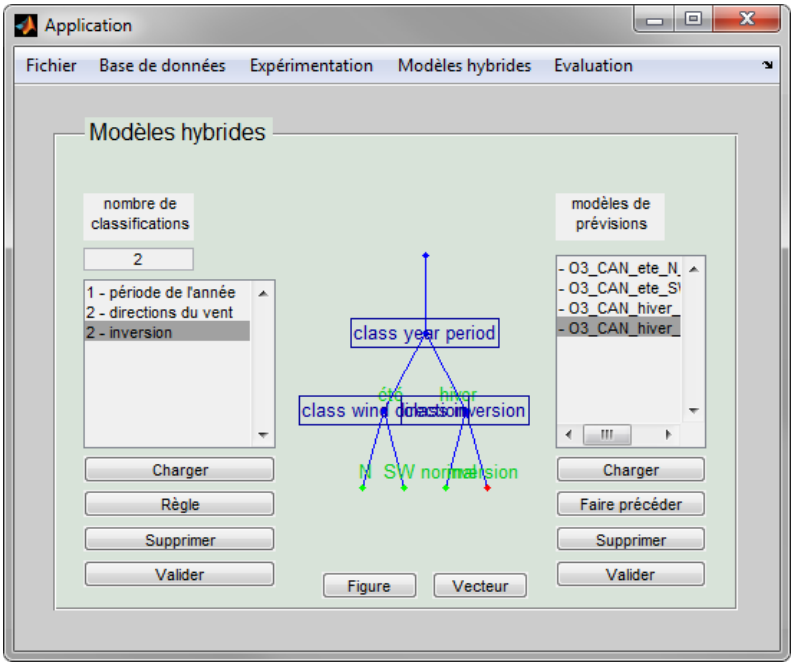


FIGURE C.5 : Capture d'écran du volet « Modèles hybrides ».

La première liste accompagnée de son graphique permettent de réaliser une division des données définie par l'utilisateur. On peut ajouter le nombre de règles de division souhaité, elles seront représentées sous forme d'arbre de décision sur le graphique attenant. La profondeur de l'arbre doit être indiquée (nombre de classification). Le bouton « Règle » permet de définir une règle de classification (voir section 6.1 page 150) correspondant à une branche de l'arbre. Cette opération peut se faire de plusieurs manières, la première étant générique, les suivantes correspondant à des cas fréquemment utilisés :

– Sélection de valeurs limites d'une série temporelle séparant les classes

– Directions de vent
– Epaisseur d'inversion thermique
– périodes de l'année
– heures du jour

Une fois l'ensemble des règles défini, le bouton « Valider » lance le calcul d'une série temporelle dont les valeurs sont des entiers indiquant la classe correspondante. Etablir ce type de règle correspond à faire le travail réalisé par un arbre de décision type CART (Classification And Regression Tree). Les règles définies par un utilisateur peuvent être moins performantes que celles déterminées par un modèle statistique. Nous avons également utilisé des modèles de classification et de partitionnement, qui ne sont pas encore inclus dans ce prototype d'application. Les séries temporelles de classes issues d'autres modèles peuvent être importées via le bouton « Charger ».

Sous la figure, le bouton « Figure » permet d'afficher l'arbre sur une figure indépendante et plus grande. Le bouton « Vecteur » permet d'exporter vers l'espace de travail de Matlab la série temporelle décrivant les classes.

C.4 Volet « Evaluation »

Ce volet permet d'évaluer les modèles, simples ou hybrides, c'est à dire avec ou sans classifieur. Une évaluation des modèles est déjà réalisée après l'apprentissage, mais ce volet permet d'avoir plus de détails, des graphiques, et peut gérer le cas des modèles hybrides avec lesquels plusieurs RNA sont impliqués dans la prévision. Une image de ce volet est présentée à la figure C.6.

On peut charger les modèles à évaluer qui apparaissent dans la liste en bas à gauche (cadre 4), entrainés via la page adéquate de l'application, depuis celle-ci ou depuis l'espace de travail de Matlab.

Les variables qu'il est nécessaire de récupérer pour faire fonctionner les modèles s'affichent dans la liste en haut à gauche (cadre 1). Elles peuvent être récupérées une par une par l'utilisateur depuis l'espace de travail, ou depuis les « Archives » contenant les modèles et leurs données d'apprentissage de validation et de test, via le bouton « Récupérer ». Le bouton « Récup. auto » permet lui de parcourir la base de donnée puis l'espace de travail pour chaque variable, afin de récupérer celle ayant le même nom que ceux figurant dans les métadonnées de l'objet « Archive ».

A partir des « Archives », l'application identifie les dates qui ne peuvent pas être utilisées pour l'évaluation, c'est-à-dire celles faisant partie du jeu d'apprentissage ou du jeu de validation. Les autres dates sont celles du set de test ou de nouvelles données, et sont consultables via le bouton « Afficher les dates ».

L'activation du bouton « Valider » finalise le jeu de données et identifie les dates utilisables pour l'évaluation. La première et la dernière date de la période d'évaluation s'affichent dans le cadre 2. L'utilisateur peut alors choisir un nombre d'intervalle de dates qu'il souhaite conserver pour l'évaluation. Cela permet par exemple d'évaluer un modèle sur une période d'intérêt particulier. Il ajoute par défaut l'ensemble des dates, et le bouton « Evaluer » lance l'évaluation de des modèles pour le jeu de test ainsi constitué. Les modèles sont évalués séparément, ainsi que le modèle hybride formé par l'ensemble.

Le cadre 3 consiste en une console permettant d'indiquer l'état du déroulement de l'évaluation. Il indique les étapes de l'évaluation (qui peut durer quelques dizaines de secondes),

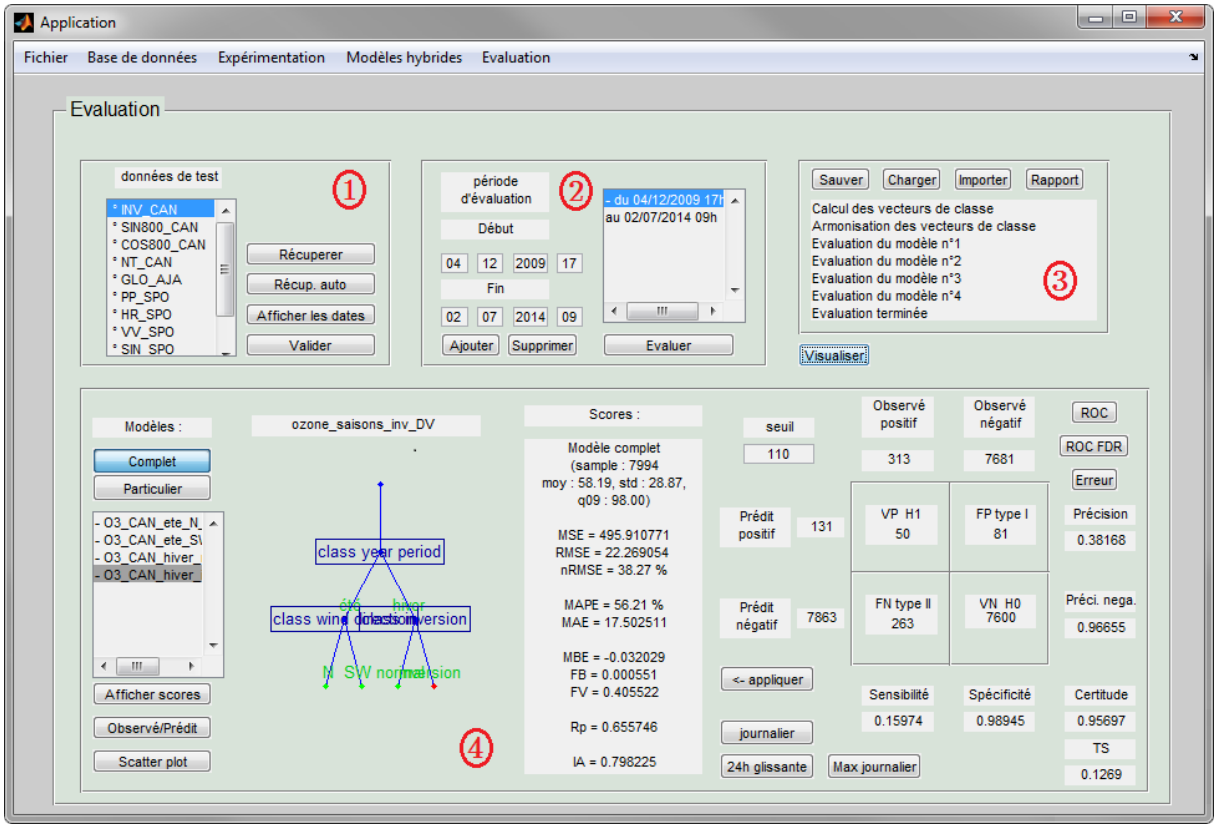


FIGURE C.6 : Capture d'écran du volet « Evaluation ».

les problèmes qui peuvent survenir (données incompatibles, etc.). Ce cadre abrite les boutons « Sauver » et « Charger », qui permettent la sauvegarde de cette évaluation et sa consultation ultérieure. Dans le cas où l'on dispose déjà de données permettant l'évaluation, c'est-à-dire d'une série temporelle représentant la variable cible, et une autre représentant les sorties de modèles correspondantes, le bouton « Importer » permet de charger ces données. Il est ainsi possible d'utiliser les capacités d'affichage de résultats d'évaluation pour d'autres modèles. Une fois l'évaluation terminer, le bouton « Visualiser », rend les résultats visibles, dans le cadre 4.

Dans ce cadre, on peut tout d'abord choisir quels sont les résultats que l'on souhaite afficher, ceux d'un modèle simple, ou ceux du modèle hybride. La structure interne du modèle hybride est rappelée sur la figure sous forme d'arbre.

Le bouton « Observé/Prédit » permet de tracer dans une nouvelle fenêtre les deux séries temporelles, la cible et la prévision pour un aperçu visuel des performances des modèles. Le bouton « Scatter plot » trace le nuage de points correspondant.

Les scores sont présentés au milieu du cadre 4. La taille de l'échantillon de test est rappelée ainsi que ses moyennes, écart-type et C90. L'ensemble des indices de performance que nous utilisons (présenté à la section 2.4) est indiqué.

La partie de droite permet une évaluation de la prévision des dépassements de seuil. Son centre est occupé par une matrice de contingence, indiquant le nombre de Vrai Positif (VP), Vrai Négatif (VN), Faux Positif (FP) et Faux Négatif (FN). Il faut fixer un seuil de concentration en haut à gauche de la matrice, à partir duquel la situation est considérée comme un dépassement de seuil. A l'extérieur de la matrice, le nombre de valeurs prédites positives et prédites négatives est indiqué, ainsi que le nombre de valeurs observées positives et négatives.

Plusieurs statistiques sont calculées à partir de ces données et sont indiquées. Il s'agit de la précision (ou « Positive Predictive Value », PPV) qui représente la part des prévisions de dépassement s'étant avérée juste et de formule :

$$PPV = \frac{VP}{VP + FP} \tag{C.1}$$

La précision négative (ou « Negative Predictive Value », NPV) correspond elle au taux de prévisions de non-dépassement qui a été vérifié. Elle a pour formule :

$$NPV = \frac{VN}{VN + FN} \tag{C.2}$$

La sensibilité (ou Taux de Vrai Positif (TVP), « True Positive Rate » en anglais) qui est représentée en ordonnée sur les courbes ROC) correspond à la part des dépassements réels qui a été prédite et a pour formule :

$$TVP = \frac{VP}{VP + FN} \tag{C.3}$$

La spécificité (ou « True Negative Rate », TNR) s'exprime par :

$$TNR = \frac{VN}{VN + FP} \tag{C.4}$$

Le Taux de Faux Positif (FPR), qui est représenté en abscisse sur les courbes ROC, est égale à $1 - TNR$.

La certitude (ou « Accuracy », ACC) est également indiquée. Il s'agit de la part de prévisions justes, telle que :

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{C.5}$$

Enfin, le « Threat Score » (TS), ou « Critical Succes Index », représente la confiance que l'on peut avoir dans la prévision d'un dépassement. On a :

$$TS = \frac{VP}{VP + FP + FN} \tag{C.6}$$

En plus de ces informations, qui sont toutes uniquement valables pour un seuil donné, il est possible de tracer trois sortes de graphiques avec les boutons en haut à droite qui donnent des informations pour tous les seuils possibles. Premièrement les courbes ROC, que l'on a utilisées dans ce manuscrit. On peut aussi tracer ce que l'on a appelé les « ROC FDR ». Il s'agit de courbe présentant le TVP en ordonnée, tout comme les courbes ROC, mais présentant le « False Discovery Rate » (FDR) en abscisse ($FDR = 1 - PPV$). L'intérêt de ces courbes est d'avoir en abscisse des informations concernant uniquement les observations en deçà du seuil, et en ordonnée les observations atteignant ou dépassant le seuil. Le bouton « Erreur », enfin, permet de tracer l'un des indices d'erreur à choisir parmi tous ceux présentés à la section 2.4, en ne conservant que les observations au dessus d'un certain seuil, qui varie entre 0 et le maximum observé.

Il est également possible d'utiliser le bouton « appliquer » pour utiliser le seuil indiqué en haut à gauche de la matrice avec les scores affichés au milieu de la page. Ces scores seront alors calculés uniquement pour les valeurs du jeu de test supérieures ou égales à ce seuil, au lieu de l'ensemble des données de test. Les boutons en dessous (« Journalier », « 24h glissante » et « Max

journalier ») peuvent être mis en surbrillance. Tant qu'ils sont activés, toutes les informations fournies par l'application (indices d'erreur, étude de sensibilité, courbes ROC, nuages de points, courbes observé VS prédit) seront calculées non pas pour des données horaires, mais pour les valeurs journalières qui correspondent. Ces valeurs journalières sont respectivement la moyenne (entre 00h et 23h), les moyennes sur 24h glissantes, et la valeur maximale journalière.

Une dernière possibilité est laissée à l'utilisateur. L'usage du bouton « Rapport » dans le cadre 3 génère une fenêtre regroupant les indices d'erreurs du modèle, un nuage de point et une courbe ROC.

L'ensemble de ces fonctionnalités permet d'évaluer précisément un modèle. Elles permettent d'illustrer la précision via des figures, de donner différents indices de précision et de mener une étude de sensibilité. Elles sont utiles pour sélectionner un modèle (simple, hybride ou tout modèle extérieur via la fonction importer) en fonction de ses attentes et de son besoin de précision.

Annexe D

Algorithmes d'apprentissage BFGS et SCG

Nous allons voir à cette annexe le fonctionnement de ces deux algorithmes d'apprentissage, qui ont été utilisés dans nos travaux mais auxquels nous avons préféré l'usage de l'algorithme de Levenberg – Marquardt (LM). Nous utiliserons les notations présentées à la section 4.3.1 (page 91).

D.1 Algorithme de BFGS

La méthode de Newton est basée sur l'utilisation d'une approximation de $\nabla S(\mathbf{p}+\mathbf{a})$ présentée à l'équation 4.14 (page 93).

A chaque itération k, il est possible de trouver le vecteur d'incrément $\mathbf{a}_k$ qui annule ∇S, indiquant un extremum de la surface d'erreur :

$$\begin{aligned} \nabla S_{k+1} &= \nabla S_k + \mathbf{H}_{S_k}\,\mathbf{a}_k = 0 \\ \mathbf{a}_k &= -\mathbf{H}_{S_k}^{-1}\,\nabla S_k \end{aligned} \tag{D.1}$$

Si $\mathbf{H}_{S_k}$ est définie positive, alors l'extremum de $S(\mathbf{p})$ est un minimum et l'adoption du nouveau vecteur de paramètre $\mathbf{p}_k + \mathbf{a}_k$ minimise l'erreur du modèle.

En réalité, le calcul numérique des $\mathbf{H}_{r_i}$ est trop coûteux pour pouvoir être réalisé à chaque itération d'un algorithme d'optimisation. C'est pourquoi certaines méthodes utilisent des estimations de cette matrice hessienne pour fonctionner, on les appelle en conséquence les méthodes quasi-Newton.

Avec les méthodes quasi-Newton, on tente de rejoindre à chaque itération le minimum global de la fonction d'erreur à partir d'informations du second ordre, ce qui converge plus vite vers le minimum global que la méthode de la descente de gradient.

L'algorithme de BFGS, des noms de ses auteurs qui l'ont indépendamment découvert à la fin des années 60 (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) est une méthode de quasi-Newton également largement utilisée. Cette méthode calcule à chaque itération la nouvelle valeur de $\mathbf{H}_S$ à partir de l'ancienne.

Pour s'assurer que l'estimation de $\mathbf{H}_S$ soit acceptable, il est calculé de manière à vérifier la condition sécante. Cette condition correspond à s'assurer qu'à chaque itération, $\mathbf{H}_S$ soit estimé

de manière à ce que le gradient ∇S soit vrai au point actuel $\mathbf{p}$ ainsi qu'au point suivant $\mathbf{p}+\mathbf{a}$. On a à partir de (D.1) :

$$\begin{aligned} \mathbf{H}_{S_k} &= \frac{\nabla S_{k+1} - \nabla S_k}{\mathbf{a}_k} = \frac{\mathbf{q}_k}{\mathbf{a}_k} \\ \mathbf{a}_k &= \mathbf{H}_{S_k}^{-1}\,\mathbf{q}_k \end{aligned} \tag{D.2}$$

en notant $\mathbf{q}_k$ la différence entre les gradients de deux itérations consécutives.

En plus de la condition sécante, la méthode BFGS vérifie qu'à chaque itération, $\nabla^2 S$ soit symétrique, ainsi que définie positive, ce qui garantit qu'on adopte une direction de descente et non de montée. Cette condition et la condition sécante laissent encore plusieurs possibilités pour $\mathbf{H}_{S_k}$. Ce dernier est donc choisi de manière à être le plus proche de $\mathbf{H}_{S_{k-1}}$, d'après une certaine métrique. On a donc trois conditions :

- $\mathbf{a}_k = \mathbf{H}_{S_k}^{-1}\,\mathbf{q}_k$ la condition sécante
- $\mathbf{H}_{S_k}^{-1}$ symétrique
- $min\|\mathbf{H}_{S_k}^{-1} - \mathbf{H}_{S_{k-1}}^{-1}\|$

Le respect de toutes ces conditions conduit fastidieusement à l'expression suivante de l'approximation de l'inverse de $\mathbf{H}_{S_{k+1}}$, nécessaire dans (D.2) :

$$\mathbf{H}_{S_{k+1}}^{-1} = \left(\mathbf{I} - \frac{\mathbf{a}_k\mathbf{q}_k^{\mathbf{T}}}{\mathbf{q}_k^{\mathbf{T}}\mathbf{a}_k}\right)\mathbf{H}_{S_k}^{-1}\left(\mathbf{I} - \frac{\mathbf{a}_k^{\mathbf{T}}\mathbf{q}_k}{\mathbf{q}_k^{\mathbf{T}}\mathbf{a}_k}\right) + \frac{\mathbf{a}_k\mathbf{a}_k^{\mathbf{T}}}{\mathbf{q}_k^{\mathbf{T}}\mathbf{a}_k} \tag{D.3}$$

avec $\mathbf{I}$ la matrice identité, $\mathbf{a}_k$ l'incrément obtenu lors de la précédente itération, et $\mathbf{q}$ la différence entre les deux gradients précédemment calculés.

A chaque itération, $\mathbf{H}_{S_{k+1}}^{-1}$ est donc mise à jour à partir de son estimation précédente, de l'incrément précédent et du gradient précédent, sans nécessiter le calcul d'informations du second ordre. Il est nécessaire d'initialiser $\mathbf{H}_{S_k}^{-1}$, qui peut prendre la valeur de $\mathbf{I}$ qui est bien symétrique. Dans ce cas, la première itération de l'algorithme correspond à celle d'une méthode de descente de gradient, mais la convergence se raffine au cours des itérations avec l'estimation de $\mathbf{H}_S$.

D.2 Algorithmes de gradients conjugués

Nous allons maintenant décrire les méthodes basées sur les gradients conjugués, dont l'algorithme SCG (Scaled Conjugate Gradient) que nous avons utilisé fait partie. Partons de l'équation de la méthode de Newton, décrite plus haut et qu'on rappelle ici :

$$\nabla S + \mathbf{H}_S\,\mathbf{a} = 0 \tag{4.14}$$

Pour résoudre un tel système, on peut utiliser la méthode du gradient conjugué, qui va rechercher la valeur optimale de $\mathbf{a}$. Si $\mathbf{A}$ est une matrice symétrique et définie positive, deux vecteurs $\mathbf{u}$ et $\mathbf{v}$ sont conjugués par rapport à $\mathbf{A}$ si :

$$\mathbf{u}^{\mathbf{T}}\,\mathbf{A}\,\mathbf{v} = 0 \tag{D.4}$$

Pour appliquer la méthode du gradient conjugué, proposée par Hestenes et Stiefel (1952), on calcule itérativement les n vecteurs $\mathbf{p}_i$, chacun étant conjugué avec le précédent. Cela évite de rencontrer un problème qu'on a avec la méthode de descente de gradient simple, quand une

direction de descente est prise plusieurs fois lors de différentes itérations. Là, le but est d'utiliser des directions conjuguées, orthogonales entre elles. On obtient donc au bout de n itérations un système conjugué de n vecteurs orthogonaux, qui forment une base de $\mathbb{R}^n$. Le vecteur $\mathbf{a}$ menant à l'optimisation des paramètres peut donc s'exprimer dans cette base sous la forme

$$\mathbf{a} = \sum_{k=1}^{n} \mathbf{p}_k \, \alpha_k \tag{D.5}$$

à partir de quoi en utilisant (4.14) on peut obtenir l'expression des valeurs de α_k, le pas de l'itération tel que $\mathbf{p}_{k+1} = \mathbf{p}_k + \alpha_k \mathbf{a}_k$:

$$\alpha_k = -\frac{\mathbf{a}_k^{\mathbf{T}} \nabla S_k}{\mathbf{a}_k^{\mathbf{T}} \mathbf{H}_{S_k} \mathbf{a}_k} \tag{D.6}$$

La méthode du gradient conjugué (Hestenes, 1980) consiste à initialiser le premier incrément $\mathbf{a}_1$ en le prenant égal à l'opposé du gradient ∇S. Ce premier pas est donc le même que celui d'une descente de gradient. Par la suite, on s'assure à chaque itération que la nouvelle direction de descente soit conjuguée avec la précédente. Chaque nouveau vecteur $\mathbf{a}_k$ correspond à la direction du nouveau gradient comme pour une descente de gradient, mais projetée de manière à être conjuguée à la précédente pour que l'ensemble des vecteur $\mathbf{a}_k$ puissent former une base de $\mathbb{R}^n$. Pour s'assurer que l'on adopte bien une direction de descente $\mathbf{a}_k$ conjuguée avec la précédente, on utilise :

$$\mathbf{a}_{k+1} = -\nabla S_{k+1} + \beta_k \mathbf{a}_k \tag{D.7}$$

avec

$$\beta_k = -\frac{|\nabla S_{k+1}|^2 - \nabla S_{k+1}^{\mathbf{T}} \nabla S_k}{\mathbf{a}_k^{\mathbf{T}} \nabla S_k} \tag{D.8}$$

A chaque itération, on adopte donc les nouveaux paramètres $\mathbf{p}_{k+1}$ tel que

$$\mathbf{p}_{k+1} = \alpha_k \, \mathbf{a}_{k+1} \tag{D.9}$$

On remarque que la méthode du gradient conjugué nécessite le couteux calcul de $\mathbf{H}_S$ intervenant en (D.6) pour fixer le pas α_k à chaque itération. Afin d'éviter d'estimer la hessienne, plusieurs méthodes de recherche linéaire pour fixer α_k existent. La méthode basée sur le gradient conjugué la plus utilisée est la méthode du SCG. Le terme $\mathbf{H}_S \, \mathbf{a}_k$ au dénominateur de α_k est estimé sans recherche linéaire, via un procédé de mise à l'échelle peu coûteux (Møller, 1993).

Annexe E

Calcul de l'IQA et de l'indice Citeair

L'Indice de la Qualité de l'Air (IQA) est calculé chaque jour à partir de mesures de stations fixes. Cet indice est calculé à partir de différents sous-indices, chacun propre à un polluant en particulier (PM10, ozone, dioxyde d'azote et dioxyde de soufre). L'indices Citeair est égal au sous-indice le plus élevé, afin de toujours représenter le ou les polluants les plus néfastes. Les moyennes journalières des PM10 sont utilisées, et le maximum des moyennes horaires du jour sont utilisés pour les autres polluants.

Tableau E.1 : Calcul des sous-indices de l'Indice de la Qualité de l'Air (IQA).

Sous-indice	Qualificatif	Concentrations en polluant			
		PM10 ($\mu g.m^{-3}$)	O_3 ($\mu g.m^{-3}$)	NO_2 ($\mu g.m^{-3}$)	SO_2 ($\mu g.m^{-3}$)
1	Très bon	0 - 6	0 - 29	0 - 29	0 - 39
2	Très bon	7 - 13	30 - 54	30 - 54	40 - 79
3	Bon	14 - 20	55 - 79	55 - 84	80 - 119
4	Bon	21 - 27	80 - 104	85 - 109	120 - 159
5	Moyen	28 - 34	105 - 129	110 - 134	160 - 199
6	Médiocre	35 - 41	130 - 149	135 - 164	200 - 249
7	Médiocre	42 - 49	150 - 179	165 - 199	250 - 299
8	Mauvais	50 - 64	180 - 209	200 - 274	300 - 399
9	Mauvais	65 - 79	210 - 239	275 - 399	400 - 499
10	Très Mauvais	≥ 80	≥ 240	≥ 400	≥ 500

L'indice Citeair est également calculé à partir de différents sous-indices, chacun propre à un polluant. L'indices Citeair est égal au sous-indice le plus élevé. Il existe deux types d'indices Citeair, en fonction de l'emplacement.

L'indice « roadside » (proximité du trafic routier) et l'indice « background » (niveau de fond), qui diffèrent au niveau de calcul uniquement dans le choix des variables utilisées (O_3 et SO_2 en plus pour le « background »). Ils ont chacun des polluants mandataires (obligatoires pour le calcul, indiqués en gras), et des polluants auxiliaires, à ajouter si disponibles.

Indice « roadside » : Applicable aux stations trafic. Egal au maximum des sous-indices des polluants suivants : **NO_2 (maximum journalier)**, **PM10 (maximum journalier)**, **PM10 (moyenne journalière)**, PM2.5 (maximum journalier), PM2.5 (moyenne journalière) et CO (moyenne 8h glissantes).

Indice « background » : Applicable à toute station non-trafic. Egal au maximum des sous-

indices de : **NO_2 (maximum journalier)**, **PM10 (maximum journalier)**, **PM10 (moyenne journalière)**, **O_3 (maximum journalier)**, PM2.5 (maximum journalier), PM2.5 (moyenne journalière), CO (moyenne 8h glissantes) et SO_2 (maximum journalier).

Le tableau E.2 indique la méthode de calcul de chaque sous-indice de polluant.

Tableau E.2 : Calcul de l'indice européen de qualité de l'air Citeair.

Polluant	Intervalle de concentration [1]	Calcul du sous-indice [2]
NO_2 (maximum journalier)	[0 100]	$sup(C/2)$
	]100 200]	$50 + sup((C-100)/4)$
	> 200	$75 + (sup([NO2] - 200)/8)$
PM10 (maximum journalier)	[0 50]	C
	]50 90]	$50 + sup(((C-50)/40) \times 25)$
	> 90	$75 + sup(((C-90)/90) \times 25)$
PM10 (moyenne journalière)	[0 30]	$sup((C/30) \times 50)$
	]30 50]	$50 + sup(((C-30)/20) \times 25)$
	> 50	$75 + sup(((75 + sup((C-200)/8) - 50)/50) \times 25)$
PM2.5 (maximum journalier)	[0 15]	$sup((C/15) \times 25)$
	]15 30]	$25 + sup(((C-15)/15) \times 25)$
	]30 55]	$50 + sup(((C-30)/25) \times 25)$
	> 55	$75 + sup(((C-55)/55) \times 25)$
PM2.5 (moyenne journalière)	[0 10]	$sup((C/10) \times 25)$
	]10 20]	$25 + sup(((C-10)/10) \times 25)$
	]20 30]	$50 + sup(((C-20)/10) \times 25)$
	> 30	$75 + sup(((C-30)/30) \times 25)$
O_3 (max. journalier)	$\forall$	$inf(C/60 \times 25)$
SO_2 (maximum journalier)	[0 50]	$inf(C/2)$
	> 50	$inf((C/8) + 37.5)$
CO (moyenne 8h glissante)	[0 5000]	$inf(C/200)$
	]5000 10000]	$inf(C/100) - 25$
	> 10000	$inf(75 + sup((C-200)/8)/400) + 50$

[1] Concentrations en $\mu g.m^{-3}$

[2] C la concentration, inf l'arrondit à l'inférieur et sup l'arrondit au supérieur

Abstract

The objective of this doctoral work is to develop a forecasting model able to correctly predict next day pollutant concentrations in Corsica. We focused on PM10 and ozone, the two most problematic pollutants in the island. The model had to correspond to the constraints of an operational use in a small structure like Qualitair Corse, the local air quality monitoring network.

The prediction was performed using artificial neural networks. These statistical models offer a great precision while requiring few computing resources. We chose the MultiLayer Perceptron (MLP), with input data coming from pollutants measurements, meteorological measurements, chemical transport model (CHIMERE via AIRES platform) and numerical weather prediction model (AROME).

The configuration of the MLP was optimized prior to machine learning, in accordance with the principle of parsimony. To improve forecasting performances, we led a feature selection study. We compared the use of genetic algorithms, simulated annealing and principal component analysis to optimize the choice of input variables. The pruning of the MLP was also implemented.

Then we proposed a new type of hybrid model, combination of a classification model and various MLPs, each specialized on a specific weather pattern. These models, which need large learning datasets, allow an improvement of the forecasting for extreme and rare values, corresponding to pollution peaks. We led unsupervised classification with self organizing maps coupled with k-means algorithm, and with hierarchical ascendant classification. Sensitivity analysis was led with ROC curves.

We developed the application "Aria Base" running with Matlab and its Neural Network Toolbox, able to manage our datasets, to lead rigorously the experiments and to create operational models.

We also developed the application "Aria Web" to be used daily by Qualitair Corse. It is able to lead automatically the prevision with MLP, and to synthesize forecasting information provided by other organizations and available on the Internet.

Keywords : Forecasting ; Multilayer Perceptron ; Artificial Neural Network ; PM10 ; Ozone

Résumé

L'objectif de ces travaux de doctorat est de développer un modèle prédictif capable de prévoir correctement les concentrations en polluants du jour pour le lendemain en Corse. Nous nous sommes intéressés aux PM10 et à l'ozone, les deux polluants les plus problématiques sur l'île. Le modèle devait correspondre aux contraintes d'un usage opérationnel au sein d'une petite structure, comme Qualitair Corse, l'association locale de surveillance de la qualité de l'air.

La prévision a été réalisée à l'aide de réseaux de neurones artificiels. Ces modèles statistiques offrent une grande précision tout en nécessitant peu de ressources informatiques. Nous avons choisi le Perceptron MultiCouche (PMC), avec en entrée à la fois des mesures de polluants, des mesures météorologiques, et des sorties de modèles de chimie-transport (CHIMERE via la plate-forme AIRES) et de modèles météorologiques (AROME).

La configuration des PMC a été optimisée avant leur apprentissage automatique, en conformité avec le principe de parcimonie. Pour en améliorer les performances, une étude de sélection de variables a été au préalable menée. Nous avons comparé l'usage d'algorithmes génétiques, de recuits simulés et d'analyse en composantes principales afin d'optimiser le choix des variables d'entrées. L'élagage du PMC a été également mis en œuvre.

Nous avons ensuite proposé un nouveau type de modèle hybride, combinaison d'un classifieur et de plusieurs PMC, chacun spécialisé sur un régime météorologique particulier. Ces modèles, qui demandent un large historique de données d'apprentissage, permettent d'améliorer la prévision des valeurs extrêmes et rares, correspondant aux pics de pollution. La classification non-supervisée a été menée avec des cartes auto-organisatrices couplées à l'algorithme des k-means, ainsi que par classification hiérarchique ascendante. L'analyse de sensibilité à été menée grâce à l'usage de courbes ROC.

Afin de gérer les jeux de données utilisés, de mener les expérimentations de manière rigoureuse et de créer les modèles destinés à l'usage opérationnel, nous avons développé l'application « Aria Base », fonctionnant sous Matlab à l'aide de la Neural Network Toolbox.

Nous avons également développé l'application « Aria Web » destinée à l'usage quotidien à Qualitair Corse. Elle est capable de mener automatiquement les prévisions par PMC et de synthétiser les différentes informations qui aident la prévision rendues disponibles sur internet par d'autres organismes.

Mots-clés : Prévision ; Perceptron Multicouche ; Réseaux de Neurones Artificiels ; PM10 ; Ozone