DÉTECTION STATISTIQUE D'ANOMALIES D'OZONE AU MOYEN D'UN RÉSEAU DE CAPTEUR

Harrou F.[1], *Bobbia M.*[2], *Fillatre L.*[1] *et Nikiforov I.*[1]

[1]Laboratoire de Modélisation et Sûreté des Systèmes (LM2S) STMR/ICD/UTT, UMR CNRS 6279, 12 rue Marie Curie - B.P. 2060 - 10010, Troyes, France

[2]Air Normand, 3 place de la pomme d'or, 76000, Rouen, France

fouzi.harrou@utt.fr, michel.bobbia@airnormand.fr, lionel.fillatre@utt.fr, igor.nikiforov@utt.fr

ABSTRACT

Pollution in our world has become a major problem that is debilitating the entire world slowly, but efficiently. The proposed theoretical solutions are applied to the problem of detection of abnormal ozone (O_3) measurements caused by air pollution or any incoherence between the different network sensors or sensor faults in the framework of regional ozone surveilance network. Dealing with nuisance parameters is an important issue in the framework of statistical anomaly detection in complex systems. The physical nature of nuisance parameters is partially known in many real life situations. This knowledge allows us to define bounds to the values of nuisance. The goal of this paper is to apply the constrained generalized likelihood ratio test to the problem of ozone anomalies detection/isolation in the framework of surveilance network.

RÉSUMÉ

La qualité de l'air est devenue une préoccupation majeure en France et dans le monde. Les outils théoriques proposés ici permettent la détection de mesures d'ozone inhabituelles, qu'elles soient d'origine anthropique (pointes de pollution dues à l'activité humaine) ou qu'elles résultent de disfonctionnement de capteurs (pannes, interférences, …). La méthode proposée permet de détecter des incohérences entre les différentes stations de mesure d'un réseau régional de surveillance de l'ozone.

Dans le cadre de la détection de défaillance à base de modèles dans des systèmes complexes, les anomalies à détecter sont souvent associées à des paramètres de nuisances qui sont indésirables mais physiquement inévitables. Dans de nombreux cas, la nature physique des paramètres de nuisance est connue, ce qui permet de fixer des bornes sur leurs valeurs. L'objectif de cet article est d'appliquer le test du rapport de vraisemblance généralisé avec contraintes aux mesures de stations d'un réseau de surveillance pour de détecter/localiser des anomalies d'ozone.

1 INTRODUCTION

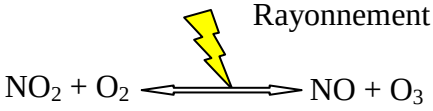
En France comme à l'étranger, l'étude de la qualité de l'air s'est largement diversifiée pour une meilleure connaissance des phénomènes de pollution et de lutte contre la pollution atmosphérique. La qualité de l'air est aujourd'hui un problème multidisciplinaire qui mobilise autant les spécialistes épidémiologiques, les spécialistes en modélisation des transports, en émissions et transformation des polluants, en systèmes géographiques et en détection que les autorités locales et les industriels. Plusieurs recherches épidémiologiques ont montré que la pollution atmosphérique est ainsi devenue en quelques années une question de santé publique préoccupant les hommes politiques et intéressant de nombreux chercheurs.

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. En France, la loi sur l'air et l'utilisation rationnelle de l'énergie du 30 décembre 1996 (loi n° 96-1236) prévoit un ensemble de mesures pour garantir aux citoyens un air de qualité. De nombreuses études épidémiologiques ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme le dioxyde de soufre (SO_2), le dioxyde d'azote (NO_2), l'ozone (O_3) ou des particules sous forme de poussières contenues dans l'air. L'influence de cette pollution est notable sur les personnes sensibles (nouveau-nés, asthmatiques, personnes âgées). La détection des pics de concentration de ces composés est donc un enjeu important pour la santé. Actuellement, parmi les composés surveillés, l'ozone est l'un des plus préoccupants. En France, comme dans la plupart des pays européens, des épisodes de pollution par l'ozone touchant une large partie du territoire pendant la période estivale. On va s'intéresser plus particulièrement à la détection de pics atypiques de pollution par l'ozone, ce qui nous permettra d'intervenir soit en informant la population (ozone avéré), soit en effectuant une maintenance sur le capteur (panne, substance interférente).
Dans cet article, après avoir présenté un bref état de l'art sur la pollution par l'ozone, on discute l'état des lieux de la surveillance de la qualité de l'air en France par les réseaux de stations dans la section 2. On élabore ensuite un modèle explicatif pour décrire la concentration d'ozone pour trois jeux de données réelles qui nous ont été fournies par les réseaux de surveillance : ATMO Champagne-Ardenne et Air Normand en section 3. Puis, on aborde le problème de détection/ localisation d'anomalies en présence de paramètres de nuisances bornés dans la section 4. La section 5 est consacrée aux résultats expérimentaux sur la modélisation des mesures et sur l'application de l'algorithme de détection/localisation d'anomalies d'ozone décrit dans la section 4.

2 POLLUTION PAR L'OZONE

2.1. Généralités

On distingue deux types d'ozone : le « bon » ozone, stratosphérique (90 % de l'ozone), et le « mauvais » ozone, troposphérique (10 % de l'ozone). L'ozone stratosphérique, présent à une altitude comprise entre 13 et 30 kilomètres, constitue la couche d'ozone. Il constitue un filtre naturel qui protège la vie sur la terre de l'action néfaste des ultraviolets. Il absorbe les rayons ultraviolets de longueurs d'onde comprises entre 230 et 300 nm, nocifs pour la matière vivante. Le « trou d'ozone » est une destruction partielle de ce filtre, liée à l'effet « destructeur d'ozone » de certains polluants émis dans la troposphère et qui migrent lentement dans la stratosphère. L'ozone troposphérique est un polluant qui a suscité un intérêt croissant au fil des années. Il s'agit d'un polluant dit « secondaire » formé à la suite de réactions chimiques complexes mettant en jeu les composés organiques volatils et les oxydes d'azote. Il se forme graduellement sous l'action du rayonnement solaire et les plus fortes concentrations d'ozone apparaissent l'été, en périphérie des zones émettrices des polluants primaires, puis peuvent être transportées sur de grandes distances.
L'équation simplifiée de formation de l'ozone est la suivante :



où NO_2 est le dioxyde d'azote, NO est le monoxyde d'azote et O_2 est l'oxygène de l'air. Les oxydes d'azote NO_2 résultent de la combinaison de l'oxygène O_2 avec le monoxyde d'azote NO induit par les activités humaines (combustion des hydrocarbures, pour le transport ou le chauffage...) et des composés organiques volatils (COV) provenant principalement des industries. Les radiations solaires de longueurs d'onde inférieures à 430 nm sont capables de dissocier le NO_2 en une molécule de monoxyde d'azote NO et un atome d'oxygène O. Ce dernier se recombine avec l'oxygène de l'air pour former la molécule d'O_3 ou ozone.

Cette réaction apporte deux informations essentielles :

- En ville, sous l'action du soleil, les oxydes d'azote et les composés organiques volatils produisent de l'ozone mais celui-ci n'est stable que loin du trafic automobile. En effet, à proximité de celui-ci, il se retransforme en dioxyde d'azote. En revanche, l'ozone transporté par les masses d'air loin de la ville reste stable, il s'accumule dans les zones périurbaines ainsi qu'en pleine campagne où la circulation automobile et les activités industrielles sont faibles donc où les oxydes d'azote sont plus rares. Il arrive souvent que les niveaux d'ozone soient plus élevés à la campagne qu'à la ville [1]. En campagne, il n'y a pas de sources de pollution telles que les industries, et il y a moins de voitures, mais on trouve en général des concentrations en ozone très élevées; les polluants précurseurs de l'ozone sont transportés loin de leur source (agglomérations, industries) et réagissent pour former l'ozone en campagne où il s'accumule puisque l'air est moins pollué. L'ozone, faute d'être détruit s'y accumule [2].

- La nuit l'ozone, produit à la lumière du jour (grâce au rayonnement du soleil), disparaît. Le monoxyde d'azote réagit avec l'ozone pour former du dioxyde d'azote ($3NO + O_3 \rightarrow 3NO_2$) de telle façon que tout l'ozone peut être éliminé s'il est en présence de quantités suffisantes de NO, ce qui est souvent le cas dans les régions urbaines. Dans les régions rurales, les concentrations de NO sont généralement trop faibles pour éliminer l'ozone d'une manière appréciable.

D'après ces deux informations essentielles, la conclusion la plus importante est que les oxydes d'azote sont à la fois les polluants précurseurs de l'ozone et aussi les gaz responsables de sa destruction.
Le rythme journalier suivi par l'ozone est de type :

- Le matin, les précurseurs de l'ozone (COV, NO_2, …) sont émis, et s'accumulent dans l'air,
- Vers midi, le soleil entre en action, il provoque les réactions photochimiques et l'ozone commence à être produit,
- L'après midi, la concentration d'ozone est la plus forte,
- La nuit, il n'y a plus de soleil, l'ozone est détruit par le monoxyde d'azote. Le monoxyde d'azote réagit avec l'ozone pour former du dioxyde d'azote.

Les variations diurnes des concentrations en ozone obéissent à un cycle caractéristique, avec un minimum en fin de nuit et un maximum en milieu d'après-midi, comme le montre la figure 1. Cette figure présente les mesures de 10 stations différentes (situées dans un même réseau) pour une même journée. Les 10 courbes ont un comportement journalier type très similaire.

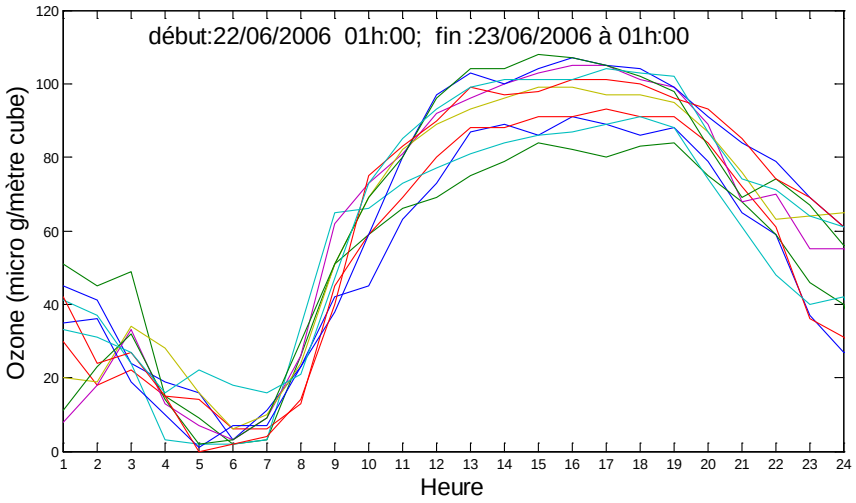


Figure 1 : Exemple de concentration d'ozone journalière.

Pendant la saison estivale, les (successions de) jours caractérisés par des températures élevées accompagnées d'un ensoleillement direct et une stagnation des masses d'air, réunissent les conditions
propices à des concentrations élevées d'ozone. En effet l'ozone, principal composant du smog photochimique estival, constitue dans le monde entier le problème le plus préoccupant, car il s'agit du polluant dont les concentrations mesurées excèdent le plus fréquemment le seuil de protection de la santé.
Les effets sur la santé humaine sont directement liés à des concentrations élevées de substances polluantes dans l'atmosphère, parmi lesquels l'ozone. Les affections chroniques du système respiratoire constituent un groupe important de maladies liées, entre autres, aux conditions atmosphériques et à l'exposition prolongée à certaines substances polluantes. Étant donné leur fréquence relativement élevée, leur caractère chronique, leur impact sur les activités normales et le coût de leur traitement, les maladies respiratoires chroniques font peser un lourd fardeau, y compris financier, tant sur les individus que sur la société. Les enfants, les personnes âgées, les asthmatiques et les insuffisants respiratoires sont particulièrement sensibles à la pollution par l'ozone. Les conséquences pour la santé varient selon le niveau d'exposition, le volume d'air inhalé et le taux d'activité. Plusieurs problèmes de santé sont possibles : toux, asthme, gêne douloureuse en cas d'inspiration profonde, mais aussi essoufflement, irritation nasale, oculaire et de la gorge.
Les effets nocifs de l'ozone sur la végétation font l'objet de nombreuses recherches. À forte concentration, l'ozone conduit à la formation de nécroses sur les feuilles, ce qui limite la photosynthèse des végétaux soumis à ces concentrations et provoque à terme des baisses de rendement pour les cultures, voire des dépérissements des écosystèmes. L'étude de ces dégâts foliaires, qui se présentent sous forme de taches à la surface de feuilles, permet d'évaluer le degré d'exposition d'une plante à l'ozone. C'est pourquoi des espèces particulièrement sensible à ce polluant, telles que le tabac, sont utilisées comme *bio-indicateur*. Dans le rapport [2], il est mentioné que la baisse de rendement des cultures aux États-Unis liées à la pollution par l'ozone est estimée entre 1 et 2 milliard de dollars chaque année.

Le seuil d'alerte est le niveau de concentration de substances polluantes dans l'atmosphère au delà duquel une exposition de courte durée présente un risque pour la santé humaine et/ou l'environnement. Dans le cadre de la loi sur l'air, la réglementation française a fixé des seuils de protection pour la santé : un seuil d'information et de recommandations aux personnes sensibles, et un seuil d'alerte à la population.

- Le seuil d'information et de recommandations correspond à une concentration en ozone de 180 $\mu g/m^3$ en moyenne horaire. Ce niveau comprend des actions d'informations de la population, des recommandations sanitaires aux catégories de la population particulièrement sensibles, ainsi que des recommandations visant à réduire certaines des émissions polluantes.
- Le seuil d'alerte correspond à une concentration en ozone de 240 $\mu g/m^3$ en moyenne horaire. De plus, il comprend trois seuils pour la mise en œuvre progressive des mesures d'urgence :

  1er seuil : 240 $\mu g/m^3$ en moyenne horaire dépassé pendant 3 heures consécutives,

  2ème seuil : 300 $\mu g/m^3$ en moyenne horaire dépassé pendant 3 heures consécutives,

  3ème seuil : 360 $\mu g/m^3$ en moyenne horaire.

Ce niveau comprend, outre les actions d'informations et de recommandations, des mesures de restriction ou de suspension des activités concourant à la pollution, y compris, le cas échéant, de la circulation des véhicules.
Le déclenchement de la procédure n'est effectué que si un seuil est dépassé sur deux stations de la région concernée avec un décalage temporel maximal de 3 heures.

2.2. Phénomènes d'ozone atypiques

Généralement, il existe deux types d'anomalies dans les mesures d'ozone : un vrai pic d'ozone et un pic d'ozone anormal lié aux différents phénomènes parasites (pannes ou disfonctionnements de capteurs, autres polluants interférant avec la mesure, conditions atmosphériques anormales, …) [3, 4].
La formation d'un vrai pic d'ozone nécessite certaines conditions : un bon ensoleillement avec une absence quasi totale de nuages – puisque ce sont les rayons ultra violets du soleil qui permettre la formation d'ozone à partir d'autres polluants – et un vent faible qui ne permet pas de disperser la pollution. Ces pics sont larges, d'une durée de plusieurs heures, les temps de réactions relativement longs impliquent une formation progressive du pic. Ces pics ont une forme en cloche.

Les pics d'ozone qualifiés de « faux » pics se démarquent de ces conditions. On peut observer des pics hors de la période estivale, à des concentrations d'ozone très élevées : typiquement de 150 à plus de 600 $\mu g/m^3$, pendant des durées brèves, d'un quart d'heure à une heure. Par exemple, considerons les pics en plein nuit. Ces pics anormaux sont très pointus, ce qui n'est généralement pas le cas pour l'ozone photochimique, puisqu'il faut un certain temps pour que la réaction se déclenche, et pour qu'elle s'arrête lorsque le rayonnement solaire et/ou la concentration des précurseurs diminuent. Ces pics anormaux inexpliqués ont fait l'objet de plusieurs recherches. D'après les études menées à ce sujet par l'INERIS [5], l'apparition de pics d'ozone intenses et brefs peut être due à différents phénomènes :

- intrusion d'ozone stratosphérique dans la troposphère,
- transport d'ozone formé ailleurs (sous forme de bouffée),
- existance d'interférents aux analyseurs d'ozone (mercure, composés organiques, aérosols,…).

2.3. La surveillance de la qualité de l'air en France par les stations de mesures

La loi sur l'air reconnaît un droit fondamental pour le citoyen : respirer un air qui ne nuise pas à sa santé. Elle définit une surveillance de la qualité de l'air à l'aide d'un réseau de surveillance et la rend obligatoire pour les agglomérations de plus de 100 000 habitants. Les réseaux de surveillance de la qualité de l'air sont généralement constitués de plusieurs stations de mesure, réparties sur l'aire géographique concernée. L'air extérieur est prélevé et analysé d'une manière permanente. Chaque station est équipée d'un ou plusieurs analyseurs mesurant chacun un polluant spécifique. Les paramètres météorologiques peuvent également être suivis. Ces réseaux de mesures sont placés sous la responsabilité des Associations Agréées de Surveillance de la Qualité de l'Air (AASQA). Les missions principales de ces associations se définissent dans les directions suivantes :

1. Surveiller la qualité de l'air sur l'ensemble de la région.
2. Analyser et expliquer les phénomènes de pollution atmosphérique afin de connaître avec précision la dispersion des polluants, leur répartition dans le temps et l'espace.
3. Établir des prévisions de qualité de l'air en utilisant des modèles de simulation.
4. Alerter les autorités durant les situations critiques (en cas de pic de pollution atmosphérique) et informer les populations des mesures à suivre pour minimiser les impacts sur la santé.
5. Rendre accessible des données à toute personne (mise en ligne sur internet, données du réseau de mesure, indice de qualité de l'air quotidien et alerte en cas de pic de pollution).

Il existe actuellement 40 réseaux de mesure pour surveiller la qualité de l'air en France. Ils constituent le réseau national de surveillance et d'information sur l'air, ATMO France.

L'épisode de canicule que la France a connu durant l'été 2003 a été doublé d'une pollution par l'ozone exceptionnelle, qui a touché l'ensemble de l'Europe. Les conséquences de cette canicule ont démontré l'importance de disposer de systèmes d'alerte précoces et fiables pour des événements non prévisibles, mais ont également souligné leur insuffisance. Des efforts considérables ont été déployés (et le sont encore) pour doter les AASQA de modèles explicatifs de la dispersion de l'ozone. Mais ces modèles dit déterministes sont parfois loin de la réalité, et dans ce cas il s'avère intéressant de compléter les moyens de détection des pics d'ozone. Notre méthode s'inscrit dans ce cadre, en complément des moyens existants, et devrait permettre de palier à des manques dans les techniques de détection de par leurs différences intrinsèques.

2.4. Sites d'études et données utilisées

Les deux sites étudiés sont la région de Champagne-Ardenne (ATMO Champagne-Ardenne[1]) et la région de Haute-Normandie (Air Normand[2]), pour lesquelles les données nous ont été fournies.

1 http://www.atmo-ca.asso.fr/
2 http://www.airnormand.fr/

Site de la région de Champagne-Ardenne

Le dispositif de surveillance de l'ozone dans la région de Champagne-Ardenne se décline de la façon suivante :

- 4 stations dans l'agglomération rémoise : 2 stations sont de typologie urbaine STA02 (station Murigny) et STA04 (station Mairie) et les deux autres sont des stations périurbaines STA09 (Tinqueux) et STA010 (Bétheny),
- 3 stations dans l'agglomération troyenne : 2 stations sont de typologie urbaine STA031 (station La Tour) et STA033 (station Saint-Parres) et la dernière STA033 est périurbaine à Sainte-Savine,
- 2 stations dans l'agglomération châlonnaise : 1 de typologie urbaine STA022 (station Châlon-en-Champagne) et l'autre est périurbaine STA021 (station Saint Memmie),
- 1 station près de Revin est rurale STA008,
- 1 station à St-Dizier STA049 de typologie urbaine.

La répartition spatiale des stations est illustrée sur la figure 2 ; leurs positions sont déterminées en utilisant leurs coordonnées en Lambert II étendu.

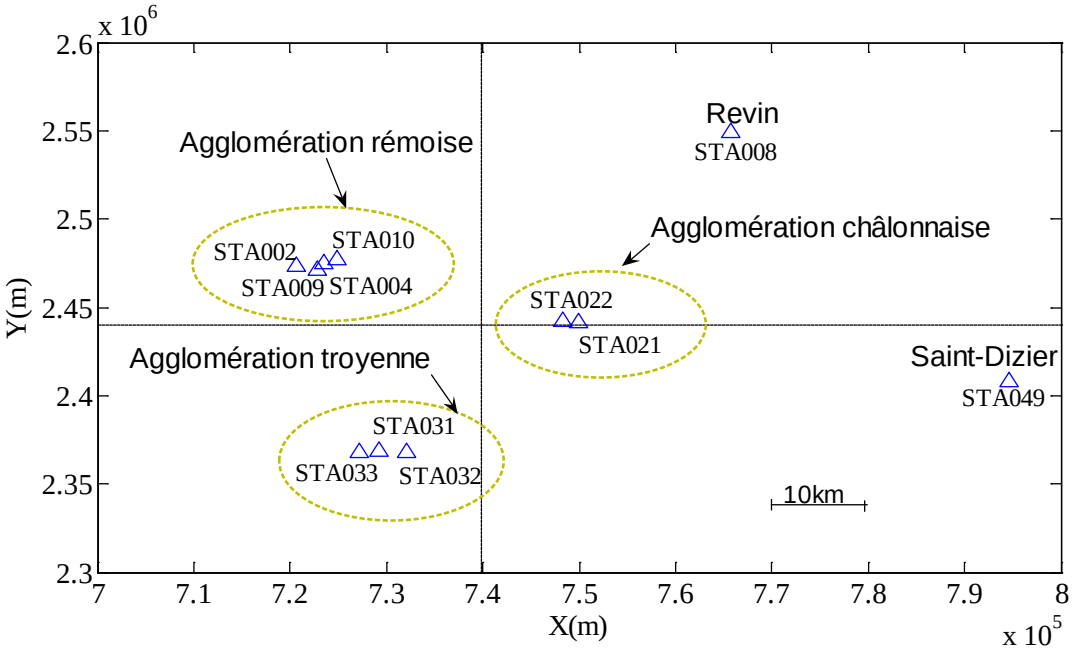


Figure 2 : Réseau de stations de mesures pour la région Champagne-Ardenne.

Site de la Haute-Normandie

Pour assurer la surveillance de l'ozone sur la région Haute-Normandie, on dispose de 10 stations de mesures :

- 3 stations de mesures sur la zone de Rouen : une station est de typologie urbaine CHS et les deux autres sont des stations périurbaines BCO et MEN,
- 4 stations de mesures sur la zone de l'estuaire : une station est de typologie urbaine (MAR), 2 sont périurbaines (MON et SRC) et la dernière est industrielle (ND2). Cette dernière station permet de suivre la formation particulière d'ozone à partir de COV et d'oxydes d'azote d'origine industrielle.
- 3 stations de mesures en continu permettent la surveillance de l'ozone sur le reste du territoire : une station à Evreux de typologie périurbaine (EVI), une dans la Forêt de

Brotonne (BRO) et la dernière est rurale à Elbeuf (ELB).

La répartition spatiale des stations de mesure est illustrée sur la figure 3 ; leur positions sont déterminées en utilisant leurs coordonnées en Lambert II étendu.

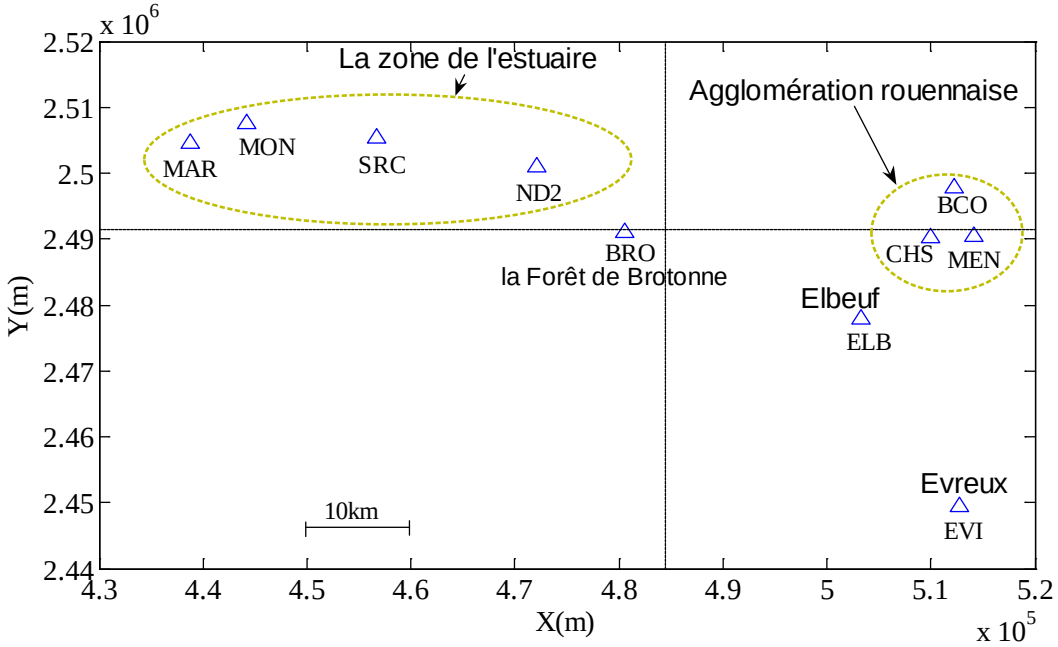


Figure 3 : Réseau de stations de mesures pour la région Haute-Normandie (dans le cas de mesures horaires).

Pour la région de Haute-Normandie, deux jeux de données ont été étudiés : des données de concentrations horaires d'ozone mesurées par le réseau de stations représenté sur la figure 3, et des données de concentrations quart-horaires mesurées par le réseau de stations représenté sur la figure 4.

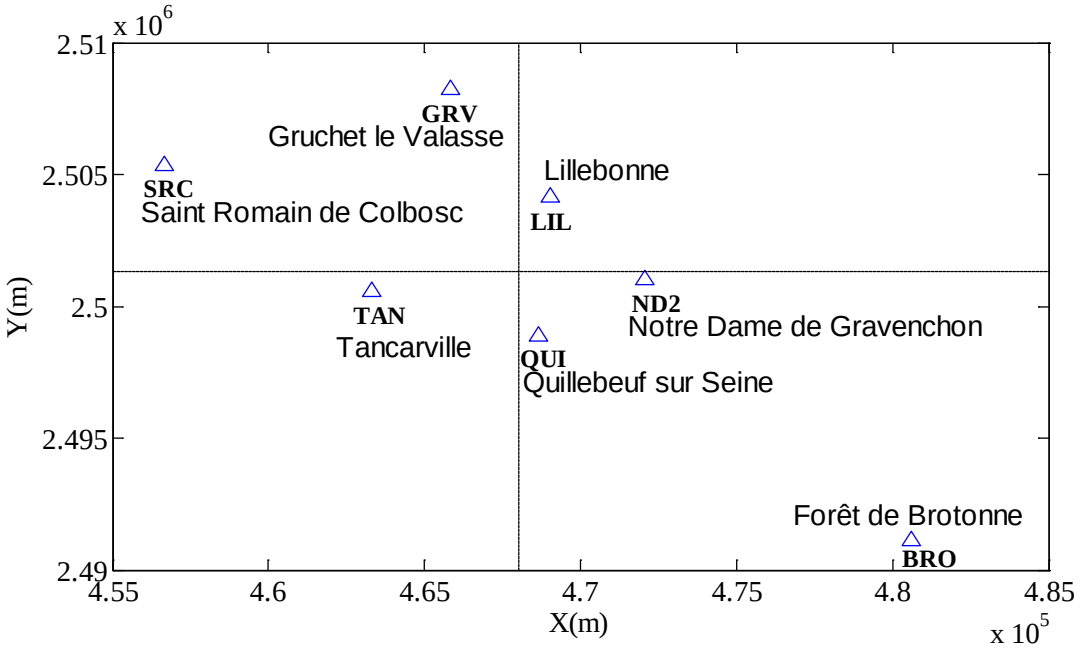


Figure 4 : Réseau de stations de mesures pour la région Haute-Normandie (dans le cas de mesures quart-horaires).

Afin de réaliser la détection d'anomalies d'ozone, on va procéder en deux étapes :

1) Tout d'abord, on introduit un modèle qui explique le mieux possible le phénomène de pollution par l'ozone,
2) puis, en se basant sur le modèle proposé et l'algorithme statistique de détection/localisation développé dans la section 4, on détecte les anomalies d'ozone.

3 MODELISATION STATISTIQUE DE DONNEES REELLES D'OZONE

Un modèle est une image simplifiée de la réalité. On distingue deux types de modèles, modèles descriptifs et modèles prédictifs. On parle de modèles descriptifs lorsque les modèles servent à représenter des données historiques. Dans cette étude, on se limite aux modèles descriptifs ; c'est dans ce cadre que s'inscrit notre travail, dont l'objectif est de construire un modèle statistique pour « expliquer » la concentration d'ozone.

La plupart des modèles rencontrés dans la pratique présentent une dimension temporelle qui peut parfois être dissimulée, mais qui le plus souvent doit être prise en compte. C'est clairement le cas des données de pollution, comme les niveaux de concentrations d'ozone pour lesquels la valeur y_t à l'instant t dépend des valeurs précédentes $y_{t-1}, y_{t-2}, \ldots$. Les modèles scalaires dynamiques AR, MA, ARMA et SARMA [8, 9, 10] sont largement exploités pour modéliser des concentrations d'ozone et d'autres polluants [11, 12,13, 14, 15, 16, 17] à partir d'une station. On trouve également les modèles du type ARMAX (AutoRegressive Moving Average model with eXternal inputs) avec ses variantes (ARX, ARIMAX, SARIMAX, …) qui prennent en compte les variables explicatives (par exemple, les mesures météorologiques) [18, 19] ; dans ce type de modèle, la variable à expliquer est la concentration d'ozone y_t et les variables explicatives sont par exemple la température, la direction du vent,… Finalement, dans le cas d'un réseau des stations d'observations d'ozone, on trouve dans la littérature les modèles vectoriels VAR (Vector AutoRegressive), VMA, VARMA qui assurent le traitement des séries multivariées [20, 21]. Ces derniers modèles permettent un traitement simultané des mesures provenant de plusieurs stations, avec ou sans variables explicatives. Tous ces modèles dynamiques sont descriptifs et prédictifs à la fois. Leur faiblesse principale est la nécessité d'une adaptation permanente en cas d'absence de variables explicatives ou lorsqu'un réseau supplémentaire est nécessaire pour obtenir des mesures météorologiques.

L'originalité du modèle SARMAX proposé dans cet article est la présence des paramètres de nuisance bornés de nature déterministe qui sont inconnus. Le rôle de ces paramètres est de modéliser les facteurs agrégés météorologiques et, de cette façon, de compléter le descriptif statistique du phénomène de pollution par l'ozone. Les conditions météorologiques régionales et le facteur commun d'ensoleillement définissent le comportement régulier commun des concentrations d'ozone mesurées par les différentes stations du réseau mais ils sont inconnus. Évidemment, il est impossible d'extrapoler la concentration d'ozone à partir d'un tel modèle dont une partie des paramètres est inconnue, mais le rôle du modèle proposé dans le test statistique est purement explicatif. En se basant sur la théorie statistique, on peut éliminer les paramètres de nuisance inconnus et construire un test statistique invariant par rapport à ces paramètres pour détecter et localiser des anomalies d'ozone.

L'objectif de cette section est de proposer un modèle expliquant le phénomène de pollution par l'ozone. Il est établit dans la littérature [1-4, 11-14, 19-21] que ce phénomène a un caractère périodique, avec la période 24 h, (ou « saisonnier », en utilisant le langage de la statistique des séries temporelles), que les mesures d'ozone sont fortement auto et spatialement corrélées. Notre propre expérience avec les données réelles enregistrées sur des réseaux de stations régionales (de la région Champagne-Ardenne et Haute-Normandie) confirme complètement cette opinion. D'après l'analyse des données réelles, le choix d'une structure simple du modèle SARMAX (1,1,1) a été fait : des polynômes AR $(1-B\alpha)$ et AR saisonnier $(1-B^{24}\alpha_s)$ d'ordre 1, un polynôme MA (moyenne glissante) $(1-B\beta)$ d'ordre 1 et un terme explicatif linéaire Hx_t. Ce choix a été confirmé par les tests à base des données réelles (voir la section 5).

Supposons que le réseau des stations d'observations d'ozone est composé de n stations installées aux positions géographiques (φ_i, λ_i), $i = 1,...,n$. Nous considérons que la Terre est localement plate et que le champ scalaire d'ozone, mesuré chaque heure aux positions géographiques (φ_i, λ_i), $i = 1,...,n$, est modélisé par l'équation dynamique stochastique suivante :

$$(1-B^{24}\alpha_s)(1-B\alpha)Y_t = Hx_t + (1-B\beta)\xi_t \ [+\theta] \ , \qquad (1)$$

où $Y = (y_1,...y_n)^T \in R^n$ est le vecteur composé de mesures d'ozone de n stations à l'instant du temps discret t, $x_t = (x_1, x_2, x_3)_t^T$ est le vecteur des entrées inconnues (paramètres de nuisance), H est la matrice $(n \times 3)$, ξ_t est le vecteur aléatoire issu d'une distribution normale $N(0,\Sigma)$, B est l'opérateur du retard : $BY_t = Y_{t-1}$, α_s est le coefficient autorégressif « saisonnier » (avec la période $24\,h$), α est le coefficient autorégressif « régulier », β est le coefficient de moyenne glissante, θ est le vecteur d'anomalies.

Dans l'équation (1), le terme Hx_t décrit l'impact des paramètres de nuisance inconnus (impact géographique et ensoleillement) sur la variation des concentrations d'ozone $y_1,...y_n$. Les lignes $(h_{i,1}\ h_{i,2}\ h_{i,3})$ de la matrice H sont calculées de la façon suivante :

$$h_{i,1} = \frac{X_i - \bar{X}}{\sigma_X}, \quad h_{i,2} = \frac{Y_i - \bar{Y}}{\sigma_Y}, \quad h_{i,3} = 1 \ ,$$

où $X_i = X_i(\varphi_i, \lambda_i)$ et $Y_i = Y_i(\varphi_i, \lambda_i)$ sont les coordonnées de la i-ème station en Lambert II étendu, $\bar{X}$ et $\bar{Y}$ sont les moyennes des X_i et Y_i, $i = 1, \dots, n$, respectivement et σ_X et σ_Y sont les ecart-types des X_i et Y_i, $i = 1, \dots, n$, respectivement.

Les entrées x_1, x_2 représentent le facteur géographique agrégé (corrélation surfacique due à la condition météorologique régionale) et l'entrée x_3 représente le facteur commun d'ensoleillement des stations du réseau. Ces entrées définissent le comportement régulier des concentrations d'ozone mesurées par les différentes stations du réseau mais elles sont inconnues et, donc, doivent être éliminées du problème de détection de pics d'ozone atypiques. L'altitude n'est pas prise en compte dans l'impact géographique car elle a un impact négligeable sur la variation de concentration d'ozone.

On peut réécrire le modèle (1) de mesure sous la forme suivante :

$$\overbrace{(1-B\beta)^{-1}(1-B^{24}\alpha_s)(1-B\alpha)Y_t}^{Z_t} = H\overbrace{(1-B\beta)^{-1}x_t}^{X_t} + \xi_t \ [+\theta] \qquad (2)$$

ou sous la forme compacte :

$$Z_t = HX_t + \xi_t[+\theta].$$

Comme nous l'avons déjà remarqué, il existe deux types d'anomalies dans les mesures d'ozone : un vrai pic d'ozone et un pic d'ozone anormal lié aux différents phénomènes parasites (pannes ou disfonctionnements de capteurs, autres polluants interférant avec la mesure, conditions atmosphériques anormales, …) [3, 4]. Un vrai pic d'ozone se caractérise par une hausse simultanée du niveau d'ozone sur toutes les stations du réseau, par contre, un pic d'ozone anormal se caractérise par la perte de « synchronisation » entre les mesures faites sur les différentes stations.

Pour détecter de façon fiable les deux types d'anomalies, il faut imposer des contraintes sur les composantes du vecteur HX sous la forme suivante :

$$h_i = \varphi_i x_1 + \lambda_i x_2 + x_3 \le b_i, \quad i = 1,...,n, \qquad (3)$$

où les paramètres b_i sont définis en fonction du niveau acceptable d'ozone, des paramètres α_s, α et β. En l'absence d'information sur le comportement du vecteur x, le rôle des contraintes est de limiter l'augmentation maximale « autorisée » pour la variable h_i afin de favoriser la détection d'une hausse simultanée du niveau d'ozone sur toutes les stations du réseau.

4 DETECTION STATISTIQUE D'ANOMALIES

Cette section est dédiée au problème de détection (localisation) d'anomalies d'ozone à base des mesures transformées $(Z_t)_{t\ge 1}$:

$$Z_t = (1-B\beta)^{-1}(1-B^{24}\alpha_s)(1-B\alpha)Y_t.$$

On peut imaginer deux approches différentes : la détection (pure) d'anomalie, c'est-à-dire le signalement de la présence d'une anomalie dans les mesures des stations du réseau, et la détection/localisation d'une anomalie, c'est-à-dire la détection de la présence d'anomalie et l'indentification de la station responsable. L'approche proposée traite les données temporelles de manière successive et indépendante dans le temps. De ce fait, pour simplifier les notations, l'indice t du temps sera omis dans les sous-sections suivantes.

4.1 Méthodes statistiques de détection et localisation d'anomalies

Dans la pratique statistique [23,24] on se trouve souvent dans la situation où la distribution P des observations Z_t dépend de deux paramètres θ et X_t, c'est-à-dire que : $Z_t \sim P_{\theta,X_t}$. Ici θ est un paramètre «informatif » et X_t un paramètre de « nuisance ». On souhaite tester l'hypothèse de base $H_0 = \{P_{\theta,X} \mid \theta \in \Theta_0, X \in R^m\}$ contre l'hypothèse alternative $H_1 = \{P_{\theta,X} \mid \theta \in \Theta_1, X \in R^m\}$, $\Theta_0 \cap \Theta_1 = \emptyset$, en considérant X comme inconnu. La conception d'algorithmes de détection (et localisation) nécessite la maîtrise de deux exigences essentielles contradictoires qui sont, d'une part, une «sensibilité» suffisante aux anomalies θ que l'on veut détecter (et localiser) et, d'autre

part, une «insensibilité» suffisante aux paramètres de nuisance X (perturbations, erreurs et incertitudes). Dans cette situation, la difficulté du problème de détection (et localisation) ne se caractérise plus par le rapport traditionnel signal à bruit de mesure (aléatoire) mais plutôt comme le rapport signal informatif à paramètre de nuisance déterministe.

La qualité d'un test statistique δ pour choisir entre H_0 et H_1 est définie par la probabilité de fausse alarme $\sup_{\theta\in\Theta_0,X} P_{\theta,X}(\delta = H_1) = \alpha$ et par la fonction de puissance $\beta = P_{\theta,X}(\delta = H_1)$, $\theta \in \Theta_1$, c'est-à-dire par la probabilité de bonne détection [23,24]. La solution idéale est un test invariant Uniformément le Plus Puissants (UPP), c'est-à-dire un test qui maximise la puissance β simultanément pour toutes les valeurs θ et $\forall X \in R^m$. Malheureusement, l'usage des tests UPP dans le traitement du signal est très limité car leur existence est conditionnée soit par une famille de distribution dépendant d'un paramètre scalaire θ, soit par le Rapport de Vraisemblance (RV) monotone et une contre-hypothèse H_1 unilatérale, soit par l'existence d'une famille exponentielle scalaire et une hypothèse de base H_0 bilatérale [23,24]. Pour cette raison la communauté statistique actuellement cherche d'autres tests avec les critères d'optimalité comportant des contraintes supplémentaires ou ayant un caractère minimax.

Le problème de détection pure d'anomalies dans un système stochastique linéaire avec des paramètres de nuisance a été étudié dans la littérature et des tests optimaux invariants UPP avec la puissance constante sur une surface dans l'espace paramétrique ont été développés [24-27]. Le problème de détection et localisation d'anomalies est plus compliqué, mais certains résultats (sous-) optimaux pour des signaux forts ont été proposés [28,29]. Enfin, l'impact de nuisance bornée sur la qualité du test du RV Généralisé (RVG) a été également étudié en [6,7].

4.2 Algorithme de détection (pure) : RVG avec contraintes

Considérons le modèle de régression linéaire suivant :

$$Z = HX + \theta + \xi \tag{4}$$

où $Z \in R^n$, est le vecteur de mesures, H est la matrice de rang plein colonne de taille $(n \times m)$ caractérisant le système inspecté avec $m < n$, $X \in D$ est le paramètre d'état inconnu, non aléatoire, à valeurs bornées (paramètre de nuisance), $\theta \in R^n$ est le paramètre informatif (anomalie ou cible) et $\xi \sim \mathrm{N}(0,\Sigma)$ est un bruit de mesure Gaussien centré de matrice de covariance $\Sigma = diag(\sigma_1^2,...,\sigma_n^2)$ connue. Le domaine D est un parallélotope défini par l'ensemble des contraintes (3). L'objectif est de proposer un test statistique pour choisir entre l'hypothèse de base:

$$H_0 = \{Z \sim \mathrm{N}(HX,\Sigma); X \in D\} \text{ (absence d'anomalies)}$$

et l'hypothèse alternative

$$H_1 = \{Z \sim \mathrm{N}(HX + \theta,\Sigma); X \in D, \theta \in \Theta\} \text{ (présence d'anomalies)}$$

où $\Theta = \{\theta : \forall (X, X') \in D^2, HX + \theta \neq HX'\}$. Dans ce problème de décision, un vecteur θ est donc considéré comme une anomalie si son apparition, conjointement avec la présence d'un paramètre de nuisance X, ne peut pas être assimilée à un paramètre de nuisance "fictif " X' acceptable sous l'hypothèse H_0. Le rapport de vraisemblance (RV) est une statistique de décision

dans une large classe des problèmes de test d'hypothèses. Lorsque le paramètre de nuisance est présent alors ce n'est pas possible d'utiliser directement le RV. Dans ce cas, il est nécessaire d'utiliser une méthode basée sur l'estimation des paramètres de nuisance au même titre que les paramètres d'intérêt. Les lecteurs intéressés peuvent se reporter à la littérature sur le sujet [6-7,25-29]. Dans le cas du problème de test d'hypothèses avec des paramètres de nuisance bornés, on propose d'utiliser le test du RVG avec contraintes (ou RVGC) [6,7] qui est basé sur la maximisation du RV par rapport aux paramètres d'intérêt θ et de paramètres de nuisance $X \in D$. Les étapes de l'algorithme de détection du RVGC pour choisir entre H_0 et H_1 sont données par le tableau suivant :

Tableau 1 : Algorithme de détection d'anomalies.

Etape	Action
1.	Initialisation : - H (matrice $(n \times m)$), Σ (matrice diagonale, n éléments), α_s (réel), α (réel), β (réel), $B = (b_1,...,b_n)^T$ (vecteur, n éléments).
2.	Calculer le vecteur $Z_t = (1 - B\beta)^{-1}(1 - B^{24}\alpha_s)(1 - B\alpha)Y_t$.
3.	Calculer la fonction de décision $\Lambda(Z) = 2\log \dfrac{\sup_{\theta \in R^n, X \in D} f_{HX+\theta}(Z)}{\sup_{X \in D} f_{HX}(Z)}$ où $f_{HX+\theta}(Z) = \dfrac{1}{(2\pi)^{\frac{n}{2}}(\det \Sigma)^{\frac{1}{2}}} \exp\left\{-\dfrac{1}{2}(Z - HX - \theta)^T \Sigma^{-1}(Z - HX - \theta)\right\}$ est la densité de probabilité des mesures Z.
4.	Calculer le seuil de décision $h(\alpha)$ de telle façon que la probabilité maximale de fausse alarme soit bornée par une valeur α fixée a priori [6, 7, 26, 27]: $\sup_{X \in D} P_{0,X}(\delta(Y) \neq H_0) = \sup_{X \in D} P_{0,X}(\Lambda(Z) \geq h(\alpha)) = \alpha$.
5.	Décision : si $\Lambda(Z) \geq h(\alpha)$ alors décider qu'il y a une anomalie.

4.3 Algorithme de détection/localisation

Souvent, il est également intéressant de savoir où l'anomalie se trouve ou quelle station est responsable d'un pic d'ozone atypique. S'il s'agit d'une panne de capteur, il est nécessaire d'organiser une opération de maintenance et dans le cas d'un pic d'ozone atypique, l'information géographique est également très utile pour une analyse approfondie de la pollution. Dans ce paragraphe on présente l'algorithme de détection/localisation d'anomalies dans les mesures d'ozone sous l'hypothèse qu'une seule station de mesure est contaminée à la fois. Les lecteurs intéressés peuvent se reporter à la littérature sur le sujet [28-29].
L'objectif du détecteur/localisateur est de détecter la présence du vecteur d'anomalie θ et d'identifier la station de mesure responsable de cette anomalie. Le problème de

détection/localisation se résume à tester l'hypothèse :

$$H_0 = \{ Z \sim \mathrm{N}(HX, \Sigma); X \in D \}$$ (absence d'anomalies)

contre les hypothèses alternatives suivantes

$$H_i = \{ Z \sim N(HX + \theta_i, \Sigma); X \in D_i, \theta_i \in \Theta_i \}$$ (anomalie à la $i^{\text{ème}}$ station) $i = 1,...,n$,

où $\Theta_i = \{ \theta_i : \forall (X, X') \in D_i^2, HX + \theta_i \neq HX' \}$ et $\theta_i = (0,\ldots,0,\nu_i,0,\ldots,0)^T$.

Pour résoudre le problème de détection/localisation, le test RVGC a été proposé [28]. Les étapes de l'algorithme de détection/localisation du RVGC sont données par le tableau suivant :

Tableau 2 : Algorithme de détection/localisation d'anomalies.

Etape	Action
1.	Initialisation : - H (matrice $(n \times m)$), Σ (matrice diagonale (n éléments)), α_s (réel), α (réel), β (réel), c_i (réel), d_i (réel).
2.	Calculer le vecteur $Z_t = (1 - B\beta)^{-1}(1 - B^{24}\alpha_s)(1 - B\alpha)Y_t$.
3.	Calculer le vecteur de parité V_t $V_t = WZ_t$ avec W satisfaisant les conditions suivantes : $W^T H = 0$, $WW^T = I_n - H(H^T H)^{-1} H^T$ et $W^T W = I_{n-m}$.
4.	Calculer le seuil de décision $h(\alpha)$ de telle façon que la probabilité maximale de fausse alarme soit bornée par une valeur α fixée priori [28] : $\sup_{X \in D} P_{0,X}(\delta(V_t) \neq H_0) = \sup_{X \in D} P_{0,X}(\Lambda(V_t) \geq h(\alpha)) = \alpha$.
5.	Calculer le test RVGC $\delta(V_t) = \begin{cases} H_0 \; si \; \Lambda(V_t) = \max\limits_{1 \leq l \leq n} \dfrac{f_{WT_l}(V_t)}{f_{WT_0}(V_t)} < h(\alpha) \\ H_v \quad si \quad v = \arg\max\limits_{1 \leq l \leq n} \left\{ \dfrac{f_{WT_l}(V_t)}{f_{WT_0}(V_t)} \geq h(\alpha) \right\} \end{cases}$ où $\mathrm{T}_l = (0,\ldots,0,\tilde{\nu}_l / \sigma_l,0,\ldots,0)^T$ et $\mathrm{T}_0 = (0,\ldots,0,\tilde{\nu}_j / \sigma_j,0,\ldots,0)^T$ sont exprimés par les formules suivantes : $\tilde{\nu}_l = \arg\max\limits_{|\nu_l| \geq d_l} \{ f_{WT_l}(V_t) \} = \arg\min\limits_{|\nu_l| \geq d_l} \left\{ \left\| V_t - W_l \dfrac{\nu_l}{\sigma_l} \right\|_2^2 \right\}$, $(\tilde{\nu}_j, j) = \arg\min\limits_{1 \leq i \leq n} \min\limits_{|\nu_i| \leq c_i} \left\{ \left\| V_t - W_i \dfrac{\nu_i}{\sigma_i} \right\|_2^2 \right\}$ où les coefficients $0 \leq c_l \leq d_l$, , $l = 1,\ldots,n$ déterminent la sélectivité et la robustesse du test RVG par rapport à l'anomalie provenant de la $l^{\text{ème}}$ station et W_j est la colonne j de la matrice W.

Le vecteur V_t s'appelle le vecteur de parité. Il est obtenu à partir des observations intiales en éliminant la partie linéaire HX du modèle (1). Cette élimination se fait au moyen d'une

projection linéaire (via la matrice *W*) sur l'espace orthogonale au sous-espace engendré par la matrice H. En absence d'anomalies, l'espérance mathématique de V_t est nulle. Dans le cas contraire, si la $j^{ème}$ station présente une anomalie alors l'espérance mathématique du vecteur de parité V_t est $W_j \boldsymbol{v_j}$ où W_j désigne la colonne *j* de la matrice *W*. Le détecteur/localisateur cherche alors à identifier quel est le « profil anormal » W_j présent dans V_t. Le paramètre c_j définit l'intensité maximale d'une anomalie W_j considérée comme acceptable, c'est-à-dire d'une anomalie qui n'est pas suffisamment significative pour justifier le déclenchement d'une alarme. C'est un paramètre de robustesse. Le paramètre d_j définit l'intensité minimale d'une anomalie W_j considérée comme significative. C'est un paramètre de sélectivité. Contraindre l'intensité minimale de chaque anomalie a pour effet de garantir un écart minimum entre toutes les anomalies possibles, ce qui permet au détecteur/localisateur de mieux localiser la véritable anomalie. Les explications détaillées sont présentées en [28].

5 RESULTATS EXPERIMENTAUX

5.1 Données réelles et l'apprentissage des modèles

Dans cette étude, on va traiter trois jeux de données de concentrations en ozone. Le premier jeu de données correspond aux mesures de concentrations horaires d'ozone observées en région Champagne-Ardenne. Les deux derniers correspondent aux concentrations horaires et quarts-horaires d'ozone observées en région Haute-Normandie. Les données sont traitées avec le logiciel Matlab. Au préalable, les portions de données utilisées pour identifier les modèles ont été analysées par des experts, pour éliminer les pics d'ozone atypiques et les pannes de capteurs.

Avant toute analyse satistique, il est intéréssant de représenter les données. Les figures 5 et 6 présentent les concentrations horaires d'ozone entre le 08/06/2007 à 09:00 et le 19/06/2007 à 09:00 observées en région Champagne-Ardenne (C-A), ainsi que leur fonction d'autocorrélation (ACF). Toutes les stations possèdent un profil très similaire, noté STA X sur les figures, à l'exception des stations STA008 et STA049 qui ont des profil légèrement différents.

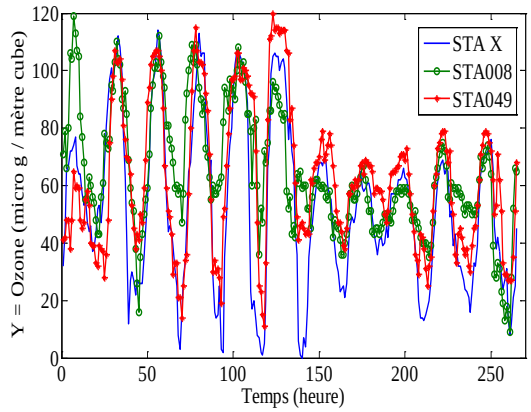


Figure 5 : Concentrations horaires d'ozone Y (C-A).

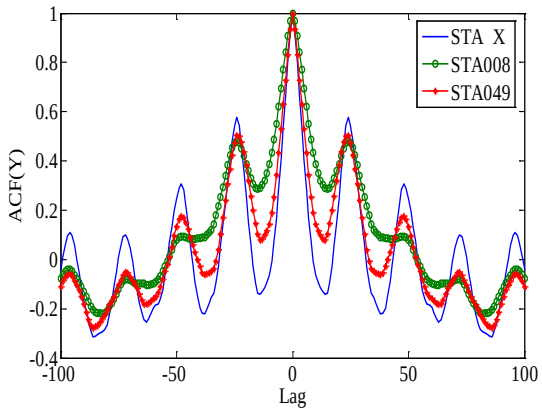


Figure 6 : ACF de Y.

Les concentrations quart-horaires en ozone observées du 26/06/2006 à 13:00 au 29/06/2006 à 21:00 en région Haute-Normandie (H-N) ainsi que leur fonction d'autocorrélation sont représentées par les figures 7 et 8. Pour une meilleure lisibilité des figures, seules les courbes des trois stations ND2, SRC et QUI (voir Figure 4) sont tracées. Les autres stations présentent des courbes presque identiques.

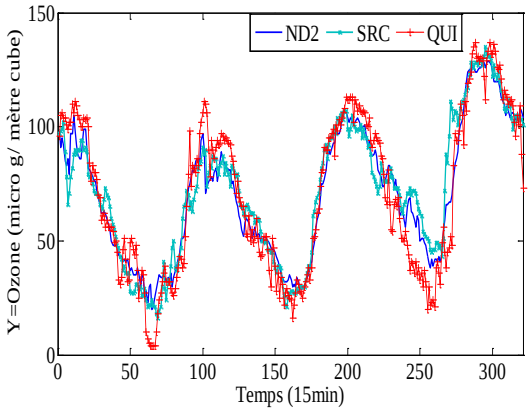


Figure 7 : Concentrations quart-horaires d'ozone Y (H-N).

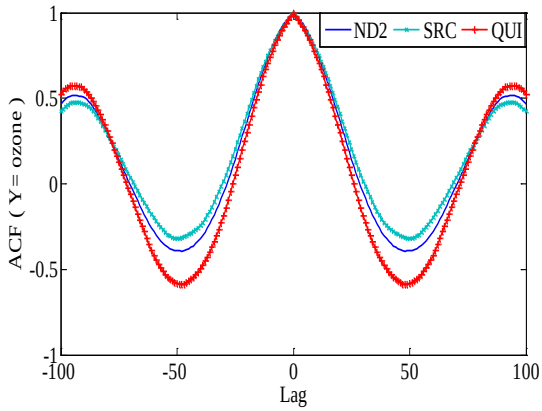


Figure 8 : ACF de Y .

Les figures 9 et 10 montrent les concentrations horaires d'ozone mesurées entre le 21/09/2005 à 11:00 et le 04/10/2005 à 10:00 observées en région Haute-Normandie ainsi que leur fonction d'autocorrélation. Pour une meilleure lisibilité des figures, seules les courbes des trois stations CHS, MAR et ND2 (voir Figure 3) sont tracées. Les autres stations présentent des courbes presque identiques.

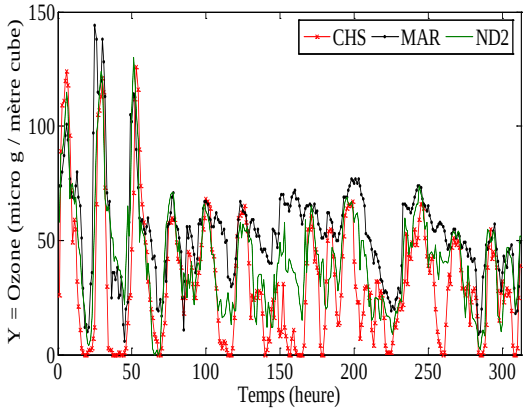


Figure 9: Concentrations horaires d'ozone Y (H-N).

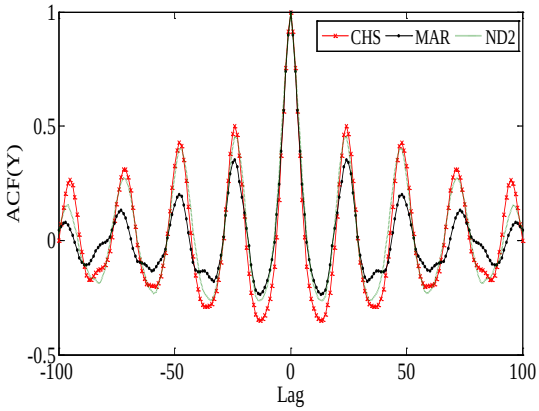


Figure 10: ACF de Y .

Ces graphiques sont intéréssants à plusieurs titres. Pour les trois jeux de données, on peut remarquer l'existence d'un cycle journalier de la concentration d'ozone. Cette variation périodique est due au cycle du rayonnement solaire (jour/nuit) qui est intimement lié au mécanisme de formation de ce polluant. La journée, le rayonnement solaire et les températures plus élevées se traduisent par un taux de production d'ozone supérieur au taux de destruction. La nuit, avec la chute de l'intensité solaire, la destruction de l'ozone devient supérieure à sa production.

On remarque aussi qu'il existe une forte similitude des fonctions d'autocorrélation entre la

majorité des stations de chaque réseau, sauf pour les stations STA008 et STA049 du réseau de Champagne-Ardenne . La courbe ACF de la station STA008 est présentée sur la figure 6 en trait avec des cercles et celle de la station STA049 en trait avec des étoiles. En effet, ces deux stations ne se trouvent pas dans une grosse agglomération. La concentration d'ozone devient plus élevée hors grosse agglomération à cause de la faible présence du destructeur d'ozone. Ce phénomène est expliqué plus en détails en commentaire du Tableau 5.

En utilisant la méthode de Box & Jenkins [8], les paramètres du modèle SARMAX (1,1,1)

$$(1-B^{24}\alpha_s)(1-B\alpha)Y_t = HX_t + (1-B\beta)\xi_t$$

pour chaque jeu de données sont décrits dans le tableau 3.

Tableau 3 : Paramètres du modèle.

Paramètres du modèle	Données horaire C-A	Données quart-horaire H-N	Données horaire H-N
α_s	0.08	0.0211	0.0727
α	0.77	0.8374	0.7544
β	-0 .065	-0.0457	-0.0664

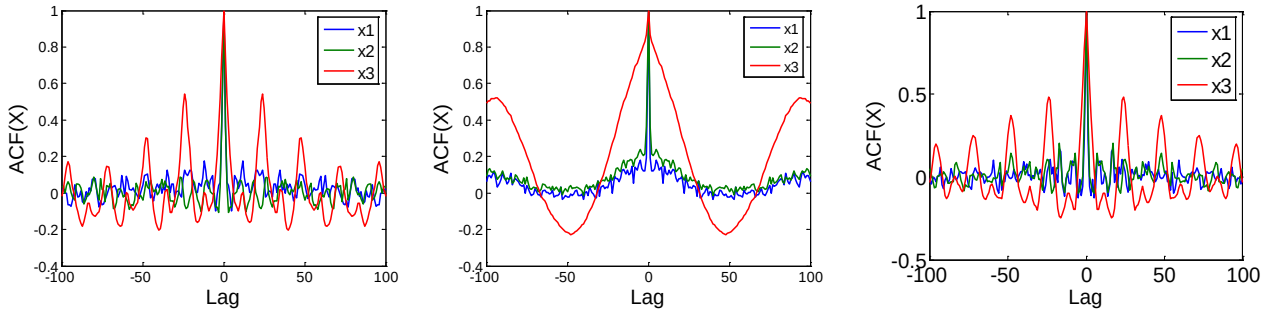
Il nous a semblé intéressant de connaître l'impact des entrées x_1 et x_2 (impact géographique) et d'entrée x_3 (le facteur commun météorologique). Le tableau suivant décrit les variances pour chaque variable et ceci pour les trois jeux de données traitées.

Tableau 4 : Variance de $\tilde{X}$.

Variance	Données horaire C-A	Données quart-horaire H-N	Données horaire H-N
var($\tilde{x}_1$)	11.84	20.70	14.1
var($\tilde{x}_2$)	12.02	21.33	5.1
var($\tilde{x}_3$)	63.79	43.88	601.4

où $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)^T = (H^T H)^{-1} H^T Z$ est la solution par la méthode du maximum de vraisemblance.

La figure 11 représente les fonctions d'autocorrélation du $\tilde{X}$ pour chaque jeu de données.



a) b) c)

Figure 11 : ACF de $\tilde{X}$, a) O_3 (C-A horaire), b) O_3 (H-N quart-horaire), c) O_3 (H-N horaire).

L'examen du tableau 3 et des graphiques d'autocorrélation de $\tilde{X}$ representés sur la figure 11 révèle immédiatement l'importante influence de la troisième composante du $\tilde{X}$ sur la concentration d'ozone. Cette dernière représente le facteur commun entre les différentes stations du réseau, c'est-à-dire l'impact des facteurs météorologiques (ensoleillement, températures, …). Aussi, on constate que l'impact du facteur géographique est moins important. Pour autant, l'absence de colonnes correspondant aux facteurs géographiques (x_1, x_2) dans la matrice H conduit à une dégradation de la qualité du modèle et celle du détecteur.

5.2 Evaluation des modèles

En modélisation, l'analyse des résidus $\hat{\xi}_t = Z_t - H\tilde{X}_t$ constitue une étape primordiale pour la validation du modèle. Cette étape est essentiellement fondée sur des méthodes graphiques (Histogramme de fréquence, Q-Q Plot, P-P Plot,…) et des tests statistiques (Test de Jarque-Bera, Test de Lilliefors,…) [30, 31]. Les résidus d'un modèle sont qualifiés de bons s'ils possèdent diverses propriétés : normalité, homoscédasticité, et indépendance.
Dans cette section, une analyse des résidus a été faite pour les trois modèles proposés. Tout d'abord, il est nécessaire de vérifier si la distribution des résidus est réellement approximable par une distribution normale. Dans cette étude, la vérification de la normalité des résidus se fait par l'examen de la droite de Henry et de l'histogramme. Ce dernier est l'outil graphique le plus simple qui permet de vérifier visuellement la normalité des résidus. On rappelle que pour la droite de Henry, le caractère normal de la distribution des résidus se reconnaît à la qualité de l'alignement des points. Si la distribution de la variable dans l'échantillon était de Gauss, les points seraient parfaitement alignés. La vérification de la normalité des résidus par la droite de Henry et par l'histogramme pour les données horaires de la région Champagne-Ardenne est illustrée par les figures 12 et 13.

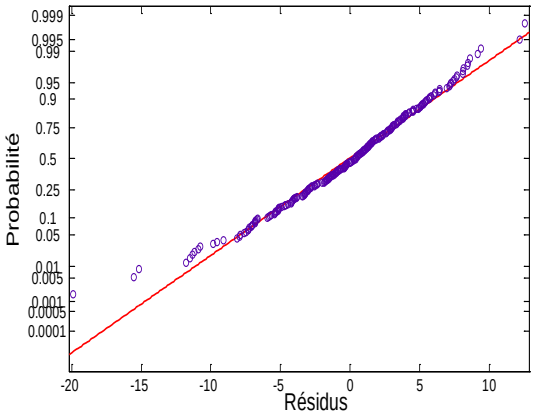


Figure 12: Droite de Henry (C-A horaire).

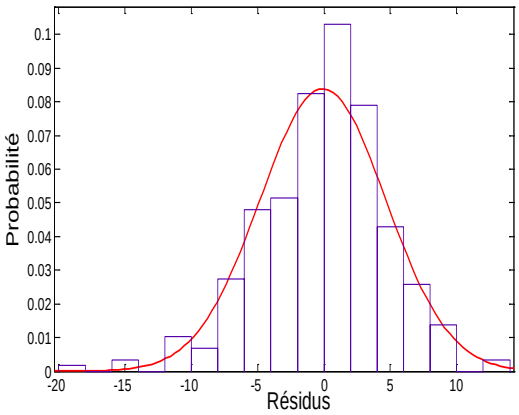


Figure 13 : Histogramme des résidus.

Les figures 14 et 15 illustrent la vérification de la normalité des résidus pour les données de concentrations quarts-horaires d'ozone en région Haute-Normandie.

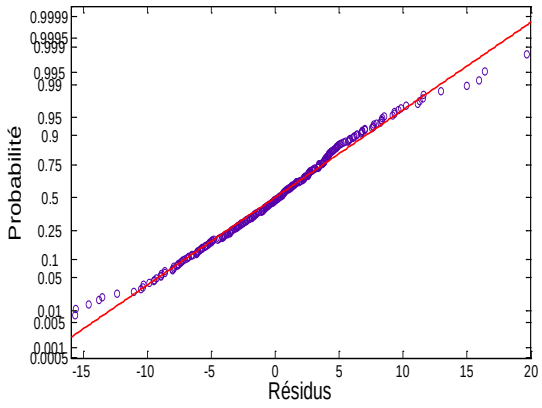


Figure 14: Droite de Henry (H-N quart-horaire).

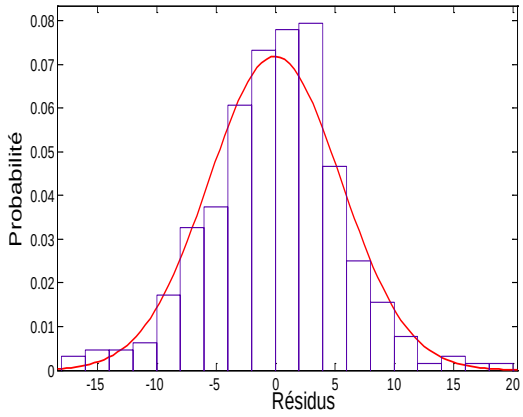


Figure 15 : Histogramme des résidus.

Enfin, la vérification de la normalité des résidus par la droite de Henry et par l'histogramme pour les données horaires de la région Haute-Normandie est illustrée par les figures 16 et 17.

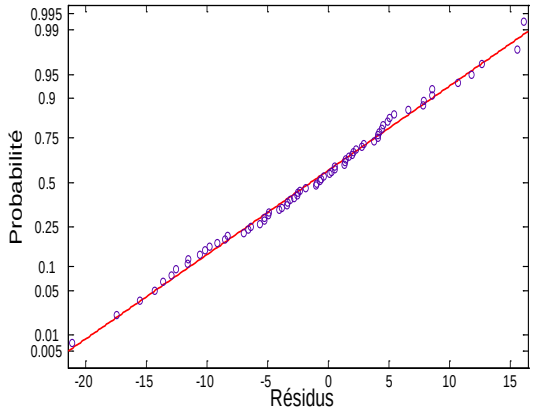


Figure 16: Droite de Henry (H-N horaire).

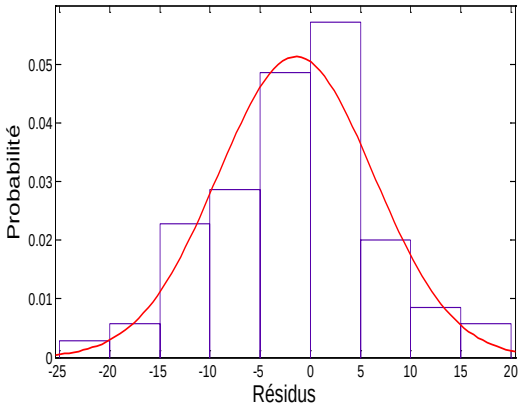
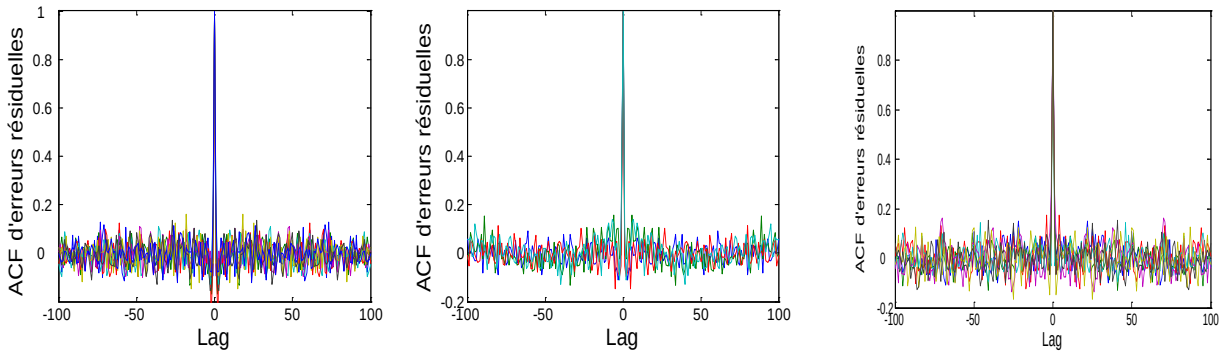


Figure 17 : Histogramme des résidus.

Les figures précédentes montrent que l'approximation de la distribution empirique des résidus par une loi normale est satisfaisante.

Ensuite, on va vérifier l'indépendance des résidus (plus précisément, l'absence d'autocorrélation). Si les résidus respectent cette hypothèse, alors l'ACF empirique des résidus ne doit pas recéler de valeurs systématiques significativement différentes de 0 pour des $\text{lag} \neq 0$.



a) b) c)

Figure 18 : ACF d'erreurs résiduelles : a) O_3 (C-A horaire), b) O_3 (H-N quart-horaire), c) O_3 (H-N horaire).

D'après la figure 18, les résidus sont approximativement non corrélés. Comme les résidus sont normalement distribués et non corrélés, on en déduit que le modèle correspond à la réalité.

Finalement, pour étudier le cas spécial des stations STA008 et STA049 du réseau de Champagne-Ardenne, l'hypothèse d'une matrice de covariance Σ scalaire a été remplacée par l'hypothèse plus réaliste que les variances résiduelles des stations STA008 et STA049 sont élevées par rapport aux variances d'erreurs résiduelles des autres stations. Les résultats du calcul sont présentés dans le tableau 5.

Tableau 5 : Variance empirique de $\hat{\xi}$.

Station	STA002	STA004	STA009	STA010	STA021	STA022	STA008	STA031	STA032	STA033	STA049
var(ξ)	22.3	22.3	22.3	22.3	22.3	22.3	191	22.3	22.3	22.3	241.5

D'après ce tableau, on remarque que les variances d'erreurs résiduelles correspondant aux stations STA008 et STA049 sont effectivement largement élevées par rapport aux variances d'erreurs résiduelles des autres stations. Contrairement aux autres stations, STA008 et STA049 se situent hors d'une grosse agglomération. Dans un tel environnement, il n'y a pas (ou peu) d'émission d'oxydes d'azote, aussi appelées NOx. Les NOx urbains produisent l'ozone mais ils participent également à sa destruction. Dans une grosse agglomération, cela crée un équilibre plus ou moins en faveur de l'ozone en fonction des conditions climatiques. En particulier, durant la nuit, l'ozone n'est plus produit (par manque de lumière et de chaleur) mais est détruit par les NOx. Par le jeu des déplacement des masses d'air, l'ozone est transporté hors de l'agglomération où il va se stocker parce qu'il ne rencontre pas de NOx pour le détruire, notamment la nuit. Les niveaux en ozone ne vont donc pas diminuer comme dans une grosse agglomération. Seuls les stations STA008 et STA049 ont ce comportement. Les autres stations ont des niveaux d'ozone qui baissent largement la nuit. Elles sont donc naturellement mieux adaptées au modèle proposé qui se base sur un facteur commun météorologique. Ceci explique le niveau plus élevé des variances résiduelles des deux stations STA008 et STA049.
Par contre, l'hypothèse d'une matrice de covariance Σ scalaire semble être acceptable pour le réseau de Haute-Normandie. Toutes les stations de ce réseau ont un comportement similaire.

5.3 Résultats de détection et localisation

Dans cette dernière partie, on présente les résultats obtenus par l'application du détecteur/localisateur RVGC aux différents jeux de données (ces échantillons sont bien sûr différents des données utilisées pour l'apprentissage du modèle). On va pouvoir évaluer les résultats de détection/localisation obtenus par le test RVGC à l'aide d'une comparaison avec les résultats de détection d'anomalies manuels déclarées par des experts d'Air Normand. On teste l'algorithme de détection/localisation RVGC sur trois échantillons. Pour définir le seuil de décision $h = h(\alpha)$, la probabilité de fausse alarme a été choisie de valeur $\alpha = 10^{-4}$, cela donne le seuil $h = 14.89$. Ce choix signifie que le taux de fausse alarme est une fausse alarme par 10^4 mesures d'ozone (soit 104 jours dans le cas de concentrations quart-horaires et 416 jours dans le cas de concentrations horaires). Le premier échantillon couvre la période allant du 11 juin 2006 au 09 juillet 2006, soit une période de 27 jours. Le deuxième échantillon couvre la période allant du 19 Août 2006 au 8 septembre 2006, soit une période de 21 jours. Le dernier couvre la période allant du 9 septembre 2006 au 10 octobre soit une période de 29 jours.

Les résultats de détection/localisation sont donnés dans le tableau 6. Les sept premières colonnes présentent les résultats d'analyse des experts d'Air Normand. La première colonne présente la date d'une anomalie constatée par les experts d'Air Normand. La deuxième et troisième colonne présentent le temps et l'intensité maximale d'un pic. La colonne 4 présente le nom de station où l'anomalie a eu lieu et les colonnes 5, 6 et 7 présentent le début, la fin et la durée de cette anomalie. Les colonnes 8 et 9 donnent les résultats de détection/localisation. Si le résultat est « oui », alors il s'agit une bonne détection et/ou localisation. Si le résultat est « non », alors c'est une non détection et/ou une fausse localisation. Dans la dernière colonne quelques explications et commentaires sont fournis par rapports aux résultats de détection/localisation automatique.

Prenons l'exemple des deux premières lignes pour décrire la façon de lire ce tableau. La première ligne signifie que la station LIL (voir la sous-section 2.4 et la figure 4 pour la description des stations) a mesuré un niveau anormale d'ozone 12/06/2006 entre 11:30 et 12:45 pour une durée totale de 0:45 minutes. Le pic de cette anomalie a eu lieu à 11:45 avec un niveau d'intensité maximale de 141 $\mu g / m^3$. C'est une vérité terrain établie par des experts du domaine. Le détecteur RVGC a détecté et localisé correctement cette anomalie (voir les colonnes « Détection » et « Localisation »). Dans la seconde ligne, les stations LIL et ND2 ont présentés toutes les deux des anomalies le 13/06/2006. Le RVGC a détecté ces anomalies mais il n'est pas parvenu à les localiser correctement (la colonne « Explications » fournit quelques commentaires sur l'échec de détection et/ou localisation en cas de non détection et/ou fausse localisation). En comparant les résultats obtenus par le détecteur RVGC et les résultats déclarés par Air Normand, on remarque que le détecteur RVGC a détecté presque la totalité des anomalies.

Tableau 6 : Résultats de détection/localisation

Date	Détection/localisation Air Normand						RVGC		Explications et commentaires
	Heure	Intensité	Lieux	Début	Fin	Durée	Détection	Localisation	
12/06/2006	11:45	141	LIL	11:30	12:15	0:45	oui	oui	
13/06/2006	13:15	168	LIL	12:30	13:45	1:15	oui	non	Fausse localisation
		181	ND2	12:15	14:15	2:00	oui	non	
17/06/2006	08:00	132	SRC	7:15	10:15	3:00	non	non	Plusieurs anomalies simultanées
	08:30	141	TAN	8:00	9:00	1:00	oui	oui	
23 /06/2006	14:15	137	LIL	13:00	15:00	2:00	non	non	Plusieurs anomalies simultanées
	14:30	126	ND2	13:00	15:15	2:15	non	non	
	14:45	127	QUI	13:15	15:15	2:00	non	non	
30/06/2006	08:00	144	TAN	7:15	8:15	1:00	oui	oui	
03/07/2006	08:15	244	TAN	8:15	9:15	1:00	oui	oui	
	10:15	242		9:15	11:15	2:00	oui	oui	
	9:30	179	LIL	9:00	10:15	1:15	oui	oui	
	10:00	166	QUI	9:15	10:15	1:00	oui	oui	
04/07/2006	07:45	201	ND2	6:30	10:00	3:00	oui	oui	
05/09/2006	09:45	180	LIL	7:45	10:45	3:00	oui	oui	
	09:45	115	TAN	8:15	11:00	2:45	oui	non	Fausse localisation
06/09/2006	09:45	182	LIL	8:15	10:30	2:15	oui	oui	
	11:15	168		10:30	13:15	2:45	oui	oui	
	14:00	168	ND2	13:15	15:00	1:45	oui	non	Fausse localisation
	14:30	168	GRV	13:45	15:00	1:15	oui	non	
10/09/2006	09:30	167	QUI	7:30	10:00	2:30	oui	oui	
	09:45	146	LIL	8:45	10:30	1:45	oui	non	Fausse localisation
	11:00	180	TAN	10:15	11:30	1:15	non	non	Plusieurs anomalies simultanées
	12:00	166	GRV	11:30	12:45	1:15	non	non	

L'atout important de ce détecteur se situe au niveau de la rapidité : il faut peu de temps à l'algorithme pour rendre son verdict. C'est la raison pour laquelle les réseaux de surveillance de la qualité de l'air montrent un intérêt pour la détection/localisation automatique des pics d'ozone atypiques (ou des pannes de capteurs).

6 CONCLUSION

Dans cet article, on a dressé un bref état des lieux sur les connaissances actuelles en matière de pollution par l'ozone troposphérique. Ensuite, une présentation succincte de la surveillance de la qualité de l'air en France par les stations de mesures a été exposée. Après ces rappels, deux réseaux de surveillance des régions Champagne-Ardenne et Haute-Normandie ont été présentés. Cette étude a donnée lieu à l'élaboration d'un modèle de concentrations d'ozone multidimensionnel dynamique (du type SARMAX) de la pollution par l'ozone pour un réseau de surveillance d'ozone régional. Ce modèle a été testé sur la base de données des régions Champagne-Ardenne et Haute-Normandie. Ensuite, les méthodes de détection/localisation d'anomalies à base d'un réseau régional de surveillance ont été appliquées aux mesures d'ozone des régions Champagne-Ardenne et Haute-Normandie. La comparaison des résultats de détection/localisation automatique avec le traitement manuel d'experts en surveillance montre que l'approche proposée est assez efficace. La détection/localisation automatique d'anomalies

d'ozone permettrait d'une part d'alerter les spécialistes de maintenance des réseaux de surveillance sur les pannes de capteurs de stations et, d'autre part, de mettre en place des mesures préventives contre les pics d'ozone atypiques.

7 REMERCIEMENTS

Les auteurs tiennent à remercier le Conseil Régional de Champagne-Ardenne et le Fonds Social Européen pour leur aide financière. Aussi, les auteurs tiennent à remercier l'association Air Normand ainsi que l'association ATMO Champagne-Ardenne pour avoir fourni l'ensemble des données nécessaires à cette étude.

8 RÉFÉRENCES BIBLIOGRAPHIQUES

1. Brönnimann S. and Neu U.,(1997), Weekend-weekday differences of near-surface ozone concentrations for different meteorological conditions. Atmospheric Environment, 31(8) : 1127-1135.
2. Ifen., 1997, L'Ozone, un polluant voyageur, les données de l'environnement N° 29, avril/mai, Rapport de l'Institut Français de l'Environnement.
3. Detournay A., 2006, Etude des pointes d'ozone sur le secteur de Port Jérôme, Rapport de stage, Association Air Normand.
4. Detournay A., Le Meur S. and Delmas V., 2007, Compréhension du phénomène de pointes « atypiques » d'ozone observé autour de la zone industrielle de Port-Jérôme en Haute-Normandie, France, Pollution Atmosphérique, Octobre-Décembre - N° 196, pp. 405-422.
5. Zdanevitch I., 2001, Etude d'épisodes inexpliqués d'ozone. Rapport LCSQA, convension 41/2000. INERIS, Paris.
6. Harrou F., Fillatre L. and Nikiforov I., 2009, Bounded nuisance rejection and redundant sensor network. The Eighth International Conference on System Identification and Control Problems, Moscow, IPU: 786- 95.
7. Harrou F., Fillatre L. and Nikiforov I., 2009, Detection statistique de pics de pollution d'ozone au moyen d'un reseau de capteur. CIRI 2009, Reims.
8. Box G. and Jenkins G., 1976, Time Series Analysis : Forecasting and Control. Holden-Bay, San Francisco.
9. Brockwell P. J. and Davis R. A., 1996, Introduction to Time Series and Forecasting. New York : Springer.
10. Brilinger D.R., 1981, Time Series, Data Analysis and Theory. Rinehart& Winston, New-York.
11. Wang, X.K., and Lu, W.Z., 2006, Seasonal variation of air pollution index: Hong Kong case study. Chemosphere, 63(8) : 1261-1272.
12. Vázquez M.G., Sánchez J.A. and Ayala F.J.G., 2005, Tropospheric ozone prediction in mexico city. Journal of the Mexican Chemical Society, 49(1): 2-9.
13. Wang S., 2007, Time series analysis of air pollution in the city of Bakersfield, California. Master of science in statistics, University of California.
14. Chaloulakou A., Assimacopoulos D. and Lekkas T., 1999, Forecasting daily maximum ozone concentrations in the athens basin. Environmental Monitoring and Assessment, Springer, 56(1): 97–112.

15. Chien K.y., 2008, Model application to air pollution data of SCCX. Master of science in statistics, University of California.
16. Prybutok V.R., Yi J. and Mitchell D., 2000, Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. European Journal of Operational Research, 122(1): 31-40.
17. Abdollahian M., and Foroughi R., 2005, Optimal statistical model for forecasting ozone. International Conference on Information Technology: Coding and Computing, 2005. ITCC : 169-173.
18. Schlink U., John S. and Herbarth O., 2002, Transfer-function models predicting ozone in urban air. Technical report, UFZ-Centre for Environmental Research, Germany.
19. Oh S.C , Sohn S.H., Yeo Y.K. and K.S Chang., 1999, A study on the prediction of ozone formation in air pollution, .Korean Journal of Chemical Engineering, 16(1) : 144-149.
20. De Luna X. and Genton M.G., 2005, Predictive spatio-temporal models for spatially sparse environmental data. Statistica Sinica, 15: 547-568.
21. Gilbert P.D., 1995, Combining var estimation and state space model reduction for simple good predictions. Journal of Forecasting, 14: 229–250.
22. Little R.J.A., and Rubin, D.B., 1987, Statistical analysis with missing data., J. Wiley. New-York.
23. T. Ferguson, Mathematical Statistics: A Decision Theoretic Approach. Academic Press, 1967.
24. E. Lehman, 1986,Testing Statistical Hypotheses, Second Edition. Chapman & Hall.
25. L. Scharf and B. Friedlander, 1994, Matched subspace detectors, IEEE Trans. Signal Processing, vol.42, no.8, pp. 2146-2157.
26. M. Fouladirad and I. Nikiforov, 2005, Optimal statistical fault detection with nuisance parameters, Automatica, vol.41, no.7, pp. 1157-1171.
27. L. Fillatre and I. Nikiforov, 2007, Non-bayesian detection and detectability of anomalies from a few noisy tomographic projections, IEEE Trans. Signal Processing, vol.55, no.2, pp. 401-413.
28. Nikiforov, I., 2009, Fault Detection and Isolation Based on the Constrained GLR Test. 13th IFAC Symposium on Information Control Problems in Manufacturing, INCOM, Moscow, Russia, June 3-5: 713-718.
29. L. Fillatre and I. Nikiforov, 2011, A New Criterion for Optimal Constrained Minimax Detection and Classification, ICASSP, Prague, May 22-27.
30. Sneyers, R., 1974, Sur les tests de normalité. Revue de Statistique Appliquée, 22 (2) : 29-36.
31. Morice, E. Tests de normalité d'une distribution observée. Revue de Statistique Appliquée, 20 no. 2 (1972) : 5-35.