Université du Québec
Institut National de la Recherche Scientifique
Centre Eau Terre Environnement

Approches statistiques avancées pour la modélisation des séries chronologiques en régression, appliquées à l'épidémiologie environnementale

Par

Pierre Masselot

Thèse présentée pour l'obtention du grade de
Philosophiae doctor (Ph.D.) en sciences de l'eau

Jury d'évaluation

Président du jury et examinateur interne	Erwan Gloaguen INRS-ETE
Examinatrice externe	Séverine Deguen École des Hautes Études en Santé Publique Rennes (France)
Examinateur externe	David A. Stephens Université McGill
Codirecteur de recherche	Taha B.M.J. Ouarda INRS-ETE
Codirecteur de recherche	André St-Hilaire INRS-ETE
Directeur de recherche	Fateh Chebana INRS-ETE

Résumé

La santé des populations est un des défis majeurs liés à l'adaptation aux changements climatiques. L'effet des vagues de chaleur est notamment déjà visible alors que ces évènements devraient se multiplier dans les années à venir. Les maladies cardiovasculaires représentent une des classes de maladies les plus touchées, tout en étant déjà un problème majeur de santé publique à l'heure actuelle. De plus en plus d'études en épidémiologie environnementale visent à identifier l'effet de la météorologie sur la santé, afin d'anticiper les changements climatiques et mettre en place des alertes appropriées. Les études d'épidémiologie environnementales s'appuient notamment sur des modèles de régression, lesquels sont appliqués avec des données pouvant prendre la forme de séries chronologiques (on peut aussi citer les données de type spatial). Les séries chronologiques violent notamment les hypothèses d'indépendance et de distribution identique des résidus dans la régression, et nécessitent donc des méthodes mieux adaptées. Cette thèse propose donc des méthodologies statistiques visant à répondre aux problèmes créés par l'utilisation de séries chronologiques dans la régression. Les méthodologies consistent toutes en un prétraitement des données puis à l'application de modèles de régression adaptés pour prendre en compte les caractéristiques des données transformées. Elles sont ensuite appliquées à l'étude du lien existant entre la météorologie, en particulier la température et l'humidité, sur la mortalité par maladie cardiovasculaire dans la communauté métropolitaine de Montréal.

Les données sanitaires utilisées dépendent notamment de l'organisation des services médicaux. Or, cette organisation entraîne la présence de bruit dans les données (p. ex. davantage de personnel de jour que de nuit), pouvant rendre plus difficile l'estimation de la relation entre une variable explicative et une réponse. Il est ainsi proposé d'agréger temporellement les séries

de données sanitaires afin de faire ressortir le signal dû à la météorologie, puis d'appliquer un modèle de régression pour série temporelle visant à modéliser la dépendance temporelle dans les résidus. La comparaison de cette méthodologie avec un modèle classique d'épidémiologie environnementale montre qu'elle permet un meilleur ajustement du modèle aux données. La comparaison de diverses stratégies d'agrégation mène cependant à la conclusion que la fenêtre d'agrégation ne doit pas être supérieure à une semaine.

Une problématique plus générale des études concernant des processus naturels est la présence de saisonnalité et tendance entre autres menant à des cas de régression fallacieuse. Il est ainsi proposé dans la thèse de décomposer les différents motifs réguliers présents dans les séries de données par décomposition modale empirique. Les composantes en résultant sont ensuite utilisées dans la régression au lieu des séries de données d'origine, en utilisant la technique du Lasso (opérateur de sélection et réduction par moindres valeurs absolues) pour ne conserver que les composantes les plus importantes pour l'explication de la réponse. L'application de cette méthodologie aux données de mortalité, température et humidité permet de mettre en évidence des aspects de la relation habituellement invisibles dans les modèles statistiques. Cette méthodologie permet ainsi un regard alternatif et détaillé sur la relation entre des séries de données chronologiques.

De nombreux problèmes liés à l'utilisation de séries chronologiques dans la régression tels que l'autocorrélation et la non-stationnarité sont issus du fait qu'elles sont en fait des discrétisations de processus intrinsèquement continus. La thèse propose donc de considérer les séries sanitaires et météorologiques comme des courbes continues en utilisant le cadre de l'analyse de données fonctionnelle. Notamment, les modèles de régression fonctionnelle sont adaptés aux problématiques inhérentes au domaine de l'épidémiologie environnementale. Les

résultats montrent le potentiel de la régression fonctionnelle pour comprendre le lien entre la météorologie et la santé dans sa globalité, en retranscrivant notamment les processus d'adaptation physiologique des individus.

Mots clés : épidémiologie environnementale; régression; séries temporelles; maladies cardiovasculaires; analyse fonctionnelle; décomposition modale empirique.

Abstract

In the context of climate change adaptation, public health management is a major challenge. For instance, the frequency and strength of heatwaves are expected to increase in the future while their effect on mortality is already well-known. Among the affected disease classes are cardiovascular diseases, which are already an important public health issue. Nowadays, many environmental epidemiology studies seek to understand precisely the effect of weather on population health, in order to accurately anticipate the future. Studies of the effect of meteorological factors on health often rely on regression models applied on time series data (although other types of data exist such as spatial data). However, several assumptions of regression models do not hold in presence of time series data, *i.e.* the assumptions of independence and same distribution of the residuals. Therefore, the purpose of the present thesis is to propose a number of regression methodologies addressing several issues caused by the temporal structure of data. The methodologies all rely on data preprocessing, in order to obtain transformed data that could be used in existing and efficient regression methods. They are illustrated on the relationship between weather and cardiovascular mortality in the census metropolitan area of Montréal, Canada.

Health data are often noisy because of organisational factors in hospitals, which complicate the task of estimating the effect of a weather exposure on a health issue. It is herein proposed to temporally aggregate the health response before using it in a regression model. A time series regression model is then used to account for the temporal dependence of data. Comparing this methodology with classical regression models show that it leads to a better fit to health data as well as unveiling the relationship at a the weekly scale than classical regression.

Moreover, several aggregation strategies are tried and it is shown that the best results are obtained using aggregations with small time windows.

Many natural time series contains nonstationary patterns such as seasonality and trend, which could lead to spurious regression. The present thesis proposes to decompose time series data into basic oscillating components through empirical mode decomposition in order to use the components as new variables in a regression model. The use of the Lasso (least absolute shrinkage and selection operator) allows keeping only the most important components in order to explain the health response. The application of this methodology on temperature and humidity related to cardiovascular morbidity unveils little known aspects of the relationship, in addition to providing a good fit of the data. Hence, it is argued that this methodology represents a tool to understand more accurately than classical models any relationship between time-related processes.

Many time series related issues in regression models are due to the fact that time series can be viewed as the discretization of intrinsically continuous processes. Therefore, the present thesis argues for the use functional data analysis which deals with data as continuous curves instead of discrete series. In particular, functional regression models are adapted to the particular issues of environmental epidemiology. The application of such models on the temperature-related cardiovascular mortality shows that they are able to describe an overall relationship. Functional models especially bring a tool that allows representing the physiological adaptation of populations, rarely taken into account in classical models.

Keywords: environmental epidemiology; regression; time series; cardiovascular diseases; functional data analysis (FDA); empirical mode decomposition (EMD).

Avant-Propos

Ce document présente les résultats des travaux de recherche mené pendant ces quatre années de doctorat. Ces travaux ont été menés dans le cadre du plan d'action sur les changements climatiques (PACC) et plus précisément du programme de recherche en santé et changements climatiques 2011-2016 réalisé en partenariat avec l'Institut national de santé publique du Québec (INSPQ). À ce titre, les travaux ont donné lieu aux articles intégrés dans la présente thèse mais également à des rapports techniques appliquant les méthodologies développées à différents cas et différents lieux de la province du Québec. Ces rapports peuvent être trouvés parmi les publications en ligne de l'INRS.

La structure du présent document suit les standards des thèses par article l'INRS-ETE, contenant une première partie synthétisant la problématique générale, les travaux menés et les résultats obtenus, puis une deuxième partie contenant les trois articles produits.

Remerciements

Le document que vous tenez entre vos mains sonne le glas de mon doctorat et, plus généralement, de ma longue vie d'étudiant. Pendant ces quatre années, l'INRS a été le théâtre du début de ma carrière de chercheur, de mon passage à une plus grande maturité (j'espère), mais aussi, et surtout, de nombreuses rencontres m'ayant grandi et façonnées. Au sein de ces couloirs, mes vies professionnelles et personnelles se sont entremêlées, ne me laissant que de bons souvenirs à l'heure de passer une étape cruciale de ma vie de chercheur.

La réalisation de ce doctorat doit beaucoup à mon directeur de thèse, Fateh Chebana, témoin et guide de mes premiers pas dans le monde de la recherche scientifique. Je me considère chanceux d'avoir pu bénéficier de ta présence et de ta disponibilité du début à la fin, afin de me conseiller et m'aider à mener ce projet à bien. Je tiens à remercier mes co-directeurs, Taha Ouarda et André St-Hilaire, pour leurs encouragements constants ainsi que pour l'aide apportée au cours de mes différents travaux. Des remerciements tout particuliers vont à deux grands artisans de mon doctorat, Diane Bélanger et Pierre Gosselin, que j'ai vraiment aimé côtoyer tout au long de ces quatre années. Par votre bonté et votre patience, vous avez facilité et encouragé ma découverte du monde de la santé publique. Merci également à Belkacem Abdous pour ses conseils et son aide précieuse, ainsi qu'à Jean-Xavier Giroux sans qui notre équipe de recherche aurait certainement tourné moins rond. Enfin, j'aimerais remercier Séverine Deguen, David Stephens et Erwan Gloaguen d'avoir accepté d'évaluer cette thèse, en espérant que vous avez apprécié découvrir notre travail.

Si la réalisation d'un doctorat est une chose importante, l'amitié est au moins aussi importante, et Dieu sait que je fus gâté pendant ces quatre belles années. À commencer par les beaux, grands, forts, excellents, brillants Yohann et Lauriane, qui sont tellement géniaux que j'en

viendrais presque à court de superlatifs. On a fait tellement de choses ensemble, à la fois dans et hors de l'INRS, que sans vous ces quatre années auraient clairement été différentes. Oserais-je vraiment me lancer dans la liste de toutes les personnes qui ont compté et façonné cette période ma vie ? Merci donc à Véronique, Marc, Dikra, Silvia, François, William, Étienne, Alexandre, Vivien… **erreur! Trop de données**. Oups. Je tiens également à remercier la machine à café du 3ème étage de l'INRS, pourvoyeuse de carburant sans lequel j'aurais eu beaucoup plus de difficulté à me rendre à la fin de mon doctorat. Un petit merci tout mignon va à Noctalie, mon petit chat, toujours prompt à m'aider à me lever le matin, d'un bon coup de griffe dans le pied. J'envoie également un merci à mes parents, pour avoir soutenu de loin leur fils ayant décidé de traverser l'Atlantique, mais également pour m'avoir appris à être curieux et m'avoir donné le goût d'apprendre, traits m'ayant conduit tout droit vers ce doctorat. Enfin, merci à ma douce et tendre, Isabelle, qui est ce que le Québec a fait de mieux dans ma vie. Ton aide a été précieuse, pour me dire quand je suis trop stressé (parce que oui, tu le sais mieux que moi même) et pour m'aider à me calmer dans ces cas-là.

Contributions

[R1] **Masselot P.**, Chebana F., Bélanger D., St-Hilaire A., Abdous B., Gosselin P., Ouarda T.B.M.J. (2016) Agrégation de la réponse dans la régression – application à la relation entre les maladies cardiovasculaires et la météorologie. INRS-ETE, R1682.

[R2] **Masselot P.**, Chebana F., Bélanger D., St-Hilaire A., Abdous B., Gosselin P., Ouarda T.B.M.J. (2015) Régression EMD avec application à la relation entre les maladies cardiovasculaires et le climat. INRS-ETE, R1594.

[A1] **Masselot P.**, Chebana F., Bélanger D., St-Hilaire A., Abdous B., Gosselin P., Ouarda T.B.M.J. (2017) Aggregating the response in time series regression models with an application to weather-related cardiovascular diseases. Soumis

[A2] **Masselot P.**, Chebana F., Bélanger D., St-Hilaire A., Abdous B., Gosselin P. (2017) EMD-regression with application to weather-related cardiovascular mortality. Soumis.

[A3] **Masselot P.**, Chebana F., Bélanger D., St-Hilaire A., Abdous B., Gosselin P., Ouarda T.B.M.J. (2017) A new look at weather-related health through functional regression. Soumis

Les deux premiers chapitres de la thèse ont généré deux documents : un rapport [R] en français à destination de l'Institut National de Santé Publique du Québec (INSPQ) et un article scientifique [A]. Les rapports contiennent des applications exhaustives des méthodologies aux cas de mortalité et morbidité par maladies cardiovasculaires des communautés métropolitaines de Montréal et Québec en fonctions de cinq variables météorologiques différentes. Ils peuvent être trouvés dans le catalogue en ligne de l'INRS. Les articles mettent plus l'accent sur la contribution statistique des travaux et limitent les résultats aux cas de la mortalité de la communauté métropolitaine de Montréal en fonction d'une ou deux variables météorologiques. Ils constituent les chapitres du présent document.

Le rapport [R1] propose d'agréger temporellement la réponse dans la régression et développe une méthodologie pour ce faire. L'article [A1] développe plus loin la méthodologie en lui ajoutant une version non-linéaire.

Le rapport [R2] développe une méthodologie de régression-EMD pour étudier une relation à différentes échelles temporelles. Ayant bénéficié de multiples révisions, l'article [A2] présente une version raffinée de la méthodologie de base.

L'article [A3] propose l'utilisation de la régression fonctionnelle en épidémiologie environnementale pour étudier l'évolution temporelle de l'effet de la température sur la mortalité par maladies cardiovasculaires.

Les travaux ont été conduits par P. Masselot sous la supervision de F. Chebana en bénéficiant tout au long du processus des conseils réguliers de D. Bélanger et P. Gosselin. A. St-Hilaire, B. Abdous et T.B.M.J. Ouarda ont participé aux réunions en fournissant des idées et des pistes de réflexions et ont aidé à la révision finale des manuscrits.

Table des matières

Liste des tableaux

Liste des figures

PARTIE I :

SYNTHÈSE

Notations mathématiques récurrentes

x_i	Variable explicative, variable quelconque
y_i	Variable réponse
t, s	Indices temporels
$x[t]$, $y[t]$	Séries temporelles
$x_i(t)$, $y_i(t)$	Données fonctionnelles
$f(x_i)$	Fonction de régression
β, $\beta(t)$, $\beta(u,t)$	Coefficient de régression classique, fonctionnel, fonctionnel bidimensionnel
$s(t)$	Fonction lisse du temps pour les confondants non-mesurés
n	Taille des séries
P	Nombre de variables explicatives dans la régression
H	Taille de fenêtre d'agrégation
$\tilde{y}[t]$	Série réponse agrégée
$c_k[t]$	IMF
A_k	Amplitude crête-à-crête de l'IMF $c_k[t]$
S_k	Sensibilité estimée pour l'IMF $c_k[t]$
T	Période considérée pour les variables fonctionnelles
N	Nombre de données fonctionnelles

Sigles récurrents

ARMA	*Autoregressive-moving average*, modèles mixtes autorégressifs et moyenne mobile
CC	Changements climatiques
CIM10	Classification international des maladies, version 10
CMM	Communauté métropolitaine de Montréal
CMQ	Communauté métropolitaine de Québec
DLM	*Distributed lag models*, modèles à effets retardés distribués
DLNM	*Distributed lag nonlinear models*, modèles non linéaires à effets retardé distribués
EMD	*Empirical mode decomposition*, décomposition modale empirique
FDA	*Functional data analysis*, analyse de données fonctionnelles
IMF	*Intrinsic mode function*, fonction modale intrinsèque
INSPQ	Institut national de santé publique du Québec
Lasso	*Least absolute shrinkage and selection operator*, Opérateur de sélection et rétrécissement par moindre valeur absolue
MCV	Maladies cardiovasculaires
MEMD	*Multivariate empirical mode decomposition*, EMD multivarié
R-EMD	Régression EMD
VC	Validation croisée

1. Introduction

La présente section introduit le contexte sanitaire puis explique la problématique d'ordre statistique ayant mené à la réalisation de la thèse.

1.1. Contexte sanitaire

Un des enjeux majeurs du XXIe siècle est l'adaptation aux changements climatiques (CC). D'après le rapport Ouranos (2015), des augmentations des températures moyennes, des températures maximales ainsi que de la durée des vagues de chaleur en été sont attendues au Québec. Une augmentation des précipitations, notamment au printemps et à l'automne, ainsi que des hivers globalement moins froids sont également prévus. Une conséquence déjà observable de ces changements est l'augmentation de la mortalité humaine due aux évènements météorologiques, en particulier lors des vagues de chaleur estivales (e.g. Braga *et al.*, 2001b; Knowlton *et al.*, 2009; Gasparrini et Armstrong, 2011; Morabito *et al.*, 2012; Bustinza *et al.*, 2013; Gasparrini *et al.*, 2015). Les maladies cardiovasculaires (MCV) représentent notamment une classe de maladies particulièrement touchée par la météorologie (e.g. Braga *et al.*, 2002; Doyon *et al.*, 2006; Bayentin *et al.*, 2010; Törő *et al.*, 2010; Phung *et al.*, 2016).

Les MCV regroupent toutes les maladies touchant le cœur et les vaisseaux sanguins, telles que les cardiopathies ischémiques et les accidents vasculaires cérébraux. Il s'agit d'une des principales causes de mortalité dans les pays développés, responsables d'un tiers des décès à l'échelle du Canada (Wielgosz *et al.*, 2009) et d'un quart à l'échelle du Québec (ISQ, 2009), bien que la mortalité par MCV a diminué depuis le milieu des années 90 grâce à de nouveaux traitements (MSSS, 2011). La prévalence des maladies cardiaques ainsi que de certains facteurs de risque importants (p.ex. l'obésité et le diabète) sont cependant en augmentation (Lee *et al.*, 2009), notamment à cause du vieillissement de la population (Tu *et al.*, 2009). Ainsi, les MCV

représentent un problème de santé publique majeur, d'autant plus qu'elles représentent un coût important sur le système de santé publique en étant la deuxième source de dépenses en santé au Canada (Wielgosz *et al.*, 2009).

Les impacts étudiés de la météorologie sur les maladies cardiovasculaires sont principalement liés aux températures extrêmes. En particulier, les vagues de chaleur ont un impact très fort sur la mortalité par MCV avec de forts excès de mortalité et d'admissions hospitalières (p. ex. Knowlton *et al.*, 2009; Nitschke *et al.*, 2011). Les vagues de froids ont également été identifiées comme facteurs de risque sur les MCV, mais avec des expositions plus longues que pour la chaleur (p. ex. Braga *et al.*, 2002; Lan Chang *et al.*, 2004). Il est cependant à noter qu'une augmentation des infarctus du myocarde a été trouvée en lien avec les chutes de pression atmosphérique (Houck *et al.*, 2005). Enfin, de rares études ont également étudié l'impact de l'humidité sur les MCV, sans trouver de relation significative (p. ex. Schwartz *et al.*, 2004).

Avec les changements climatiques, il est donc possible que la tendance à la baisse de la mortalité par MCV se stabilise voire s'inverse. En effet, les canicules, qui devraient survenir plus fréquemment dans le futur, ajoutent un stress important à un système vasculaire déjà affaibli, en jouant sur la pression sanguine (Sawka *et al.*, 2011). De plus, malgré l'atténuation globale de la rudesse hivernale, il y aura encore au Québec des vagues de froid intense (Sillmann *et al.*, 2013) qui représentent également un stress important sur un organisme humain déjà atteint par des MCV (Huynen *et al.*, 2001). La diminution de la fréquence des vagues de froid pourrait également se révéler à double tranchant en rendant celles qui surviennent d'autant plus stressantes pour les organismes (Keatinge, 2002; Kinney *et al.*, 2012).

En résumé, les MCV représentent une classe de maladies à forte prévalence dans la population et sont particulièrement sensibles aux stress météorologiques. Dans un but

d'adaptation aux changements climatiques, il est donc crucial de mieux identifier l'influence des variations météorologique sur les MCV dans la population. C'est dans ce contexte de raffinement de ces connaissances que s'inscrit la présente thèse.

1.2. Problématique

Pour l'étude du lien entre la santé et la météorologie (et, plus largement, l'environnement), la littérature d'épidémiologie environnementale utilise majoritairement des modèles de régression qui permettent d'exprimer statistiquement l'effet d'une variable environnementale x_i sur une issue sanitaire particulière y_i :

$$y_i = f\left(x_i\right) + \varepsilon_i \tag{1}$$

où $f(.)$ est une fonction possiblement non linéaire et ε_i est le résidu de la régression, c.-à-d. la partie de y_i non expliquée par x_i. Les données y_i et x_i $(i = 1,...,n)$ utilisées sont généralement, pour la variable réponse y_i, des totaux de mortalité ou morbidité pour cause de la maladie étudiée et des mesures de variables environnementales pour les variables explicatives x_i (p. ex. Peng et Dominici, 2008). Ces données sont souvent disponibles sur une base quotidienne ou, plus rarement, horaire. Dans tous les cas, les données y_i et x_i utilisées sont sous forme de séries temporelles, donc indexées par le temps $t = 1,...,n$, ce qui sera dénoté ensuite $y[t]$ et $x[t]$ dans la synthèse. L'utilisation de séries temporelles dans un modèle de régression est délicate car si leur structure temporelle n'est pas prise en compte, elle se retrouve dans les résidus $\varepsilon[t]$, violant ainsi une ou plusieurs hypothèses de la régression classique (Hamilton, 1994, chapitre 8). Les problèmes issus de la structure temporelle des données sont présentés ci-dessous et résumés dans le Tableau 1.

Le plus étudié des problèmes liés aux séries temporelles dans la régression (p. ex. Aitken, 1935b; Cochrane et Orcutt, 1949; Pesaran, 1973) est la présence de corrélation entre les observations successives, ou autocorrélation. Les températures sont un bon exemple d'autocorrélation dans la mesure où la température d'un jour est liée à celle du jour d'avant (voire de plusieurs jours). Dans un modèle de régression, la présence d'autocorrélation viole l'hypothèse d'indépendance des résidus. Cela amène à des estimateurs biaisés et non consistants, et dont la variance est sous-estimée, invalidant ainsi les tests d'hypothèse classiques (p. ex. Mizon, 1995). Cet état de fait est également vrai dans le cas de la régression non linéaire (p. ex. Glasbey, 1980).

Dans beaucoup de cas, les séries temporelles peuvent également être non stationnaires. Une série temporelle est dite non stationnaire lorsque la distribution des observations $x[t]$ ou la dépendance entre les observations successives (p. ex. $x[t]$ et $x[t+1]$) dépend du temps t. De nombreuses sources de non stationnarité existent (voir p. ex. Ventosa-Santaulària, 2009), mais les plus communs sont une autocorrélation très forte ainsi que la présence de tendance ou de saisonnalité. La température est encore un excellent exemple car elle présente des saisonnalités quotidiennes et annuelles évidentes ainsi qu'une tendance possible due aux changements climatiques, rendant la distribution des observations dépendantes de la période d'observation. La non stationnarité des données viole l'hypothèse de distribution identique des résidus et mène au phénomène de « régression fallacieuse ». Cette dernière survient lorsqu'une analyse de régression conclut à un lien entre $x[t]$ et $y[t]$ alors les deux séries sont absolument indépendantes (p. ex. Hoover, 2003). Ce phénomène très étudié en économétrie (p. ex. Granger et Newbold, 1974; Phillips, 1986; Phillips, 1998) est bien connu mais les méthodes ou tests développés pour s'en prémunir ciblent généralement une seule catégorie de série temporelle (p. ex. le test de racine

unitaire de Phillips, 1987 ne comprend que les séries qui sont stationnaires par différenciation). Il est à noter que le problème de régression fallacieuse existe également dans le cas de la régression non linéaire (Lee *et al.*, 2005).

Les séries de données sanitaires, telles qu'hospitalisations et décès, incluent souvent des motifs réguliers dus à l'organisation des hôpitaux. Or, de tels motifs sont assimilés à du bruit qui peut être dominant dans les séries temporelles et donc masquer la réponse à une exposition dans les données. Dans le cadre de l'épidémiologie environnementale, ces motifs sont généralement contrôlés en ajoutant des variables au modèle. Cependant, ils peuvent survenir à différentes échelles, et donc résulter en un grand nombre de variables ajoutées pour ceci, ce qui peut résulter en un sur ajustement du modèle.

La mortalité et la morbidité sont généralement associées à divers indicateurs de risque, pouvant être corrélés entre eux (p. ex., température, précipitations, pollution de l'air, comme dans Wong *et al.*, 1999). Or, dans certains cas, cette multicolinéarité peut entraîner des conclusions erronées (Zidek *et al.*, 1996). Il est à noter que ce problème peut être lié à la non stationnarité car des séries ayant une saisonnalité commune, telles que les variables environnementales ayant un cycle été/hiver, sont nécessairement fortement corrélées. Ainsi, ce point doit également être pris en compte dans toute analyse de régression.

Tableau 1 : Résumé des problèmes présents dans les séries de données utilisées en épidémiologie environnementale avec quelques méthodes (utilisées ou proposées) pour y remédier.

Concerne	Problème	Engendre	Remède(s) classique(s)	Proposition de la thèse
$y[t], x[t]$	Autocorrélation	- Violation de l'hypothèse d'indépendance des résidus - Inconsistance des estimateurs - Sous-estimation de la variance des estimateurs	- Moindres carrés généralisés - Préblanchiment	- Régression fonctionnelle [A3]
	Non stationnarité	- Violation de l'hypothèse de distribution identique des résidus - Régression fallacieuse	- Différenciation - Retrait de tendance/saisonnalité	- Régression-EMD [A2] - Régression fonctionnelle [A3]
$x[t]$	Multicolinéarité	- Solution non unique	- Sélection de variables - Régression Ridge - Lasso	- Régression-EMD [A3]
$y[t]$	Bruit dans la série	- Masque le signal important	- Variables dichotomiques	- Agrégation de la réponse [A1]

Considérant les problèmes inhérents aux séries de données environnementales et sanitaires (résumés dans le Tableau 1), il est important pour la recherche en santé publique de disposer de méthodes prenant en compte ces problèmes. En effet, lorsqu'ignorés, ces problèmes méthodologiques peuvent invalider les conclusions tirées d'une étude. Un grand nombre de méthodes existent pour traiter les différents problèmes exposés (p. ex. moindres carrés généralisés, intégration des séries, modèles non linéaires, etc.), mais chaque méthode se concentre sur un aspect précis et ignore les autres. Il est donc important que les chercheurs disposent d'outils statistiques permettant de gérer les problèmes cités ci-dessus afin d'obtenir des résultats valides statistiquement.

1.3. Objectifs et réalisations

Étant donné les problèmes liés aux séries temporelles que l'on peut retrouver en épidémiologie environnementale, l'objectif global de la présente thèse est de proposer des méthodologies de régression applicables malgré la présence des problèmes résumés au Tableau 1. Bien que les méthodologies proposées dans la présente thèse soient appliquées au lien entre la météorologie et les MCV, elles se veulent applicables à n'importe quel autre domaine dans lequel les données sont sous forme de séries temporelles. Le dénominateur commun des méthodologies est qu'elles sont basées sur un traitement préalable des séries de données. L'objectif global est scindé en trois sous-objectifs, donnant lieu à trois méthodologies, toutes trois basées sur des traitements préalables différents illustrés dans la Figure 1. Chacune des méthodologies donne lieu à un chapitre de thèse (et donc un article).

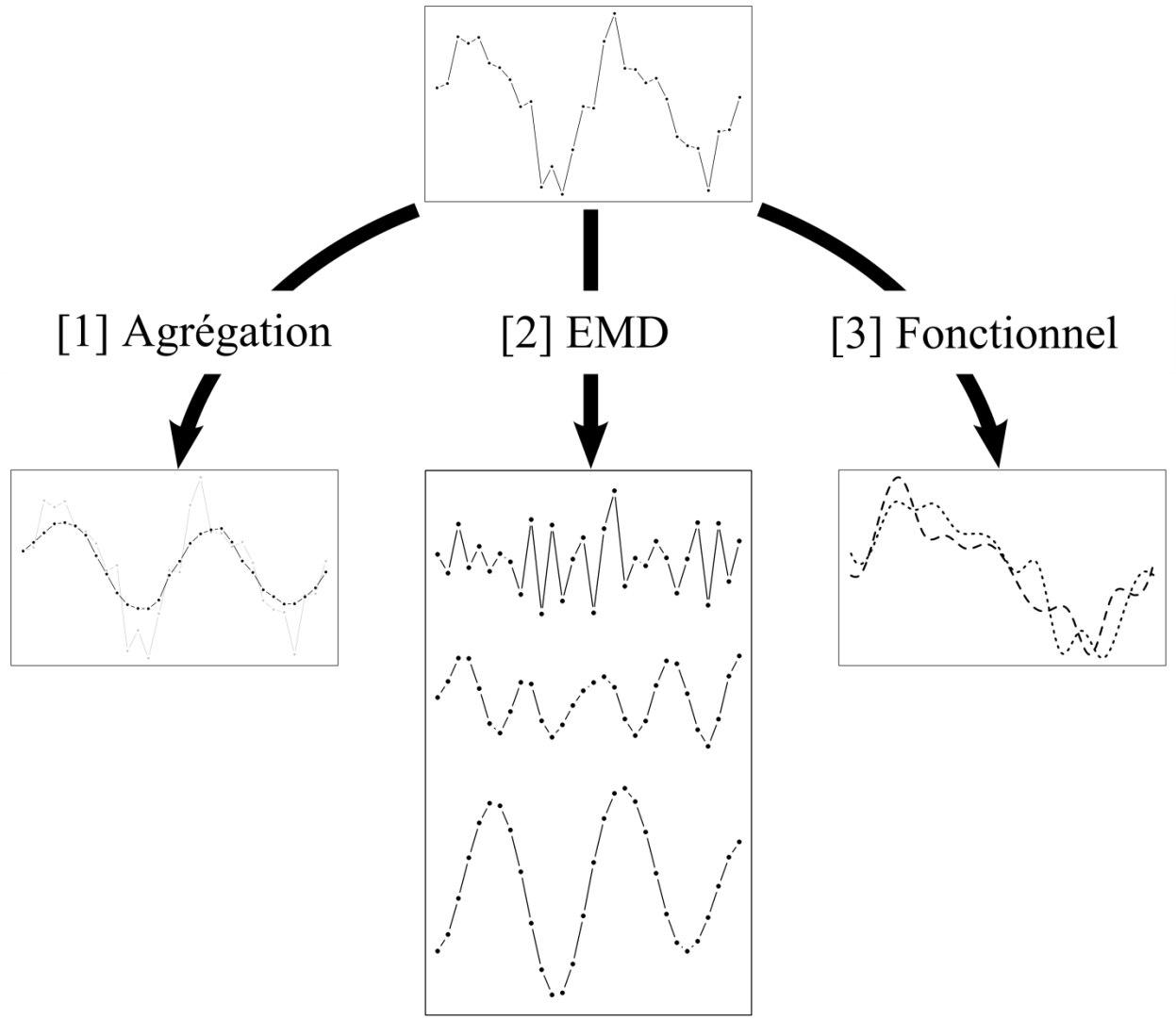


Figure 1 : Illustration des trois traitements appliqués aux données dans chacune des méthodologies développées dans la présente thèse.

Le premier sous-objectif est de s'attaquer au problème des motifs dans les données sanitaires, notamment dus à l'organisation des hôpitaux. Ce problème est traité par agrégation temporelle de la série sanitaire réponse pour supprimer les variations de plus haute fréquence dans la série et donc mettre en évidence le signal d'intérêt (voir Figure 1). À partir de là, une méthodologie de régression pour série réponse agrégée est proposée puis appliquée à

l'explication de la morbidité par MCV par rapport à la température quotidienne moyenne dans la communauté métropolitaine de Montréal (CMM). Cette première méthodologie est développée dans un rapport pour l'Institut national de la santé publique du Québec (INSPQ) [R1] avant de donner lieu à l'article [A1]. À noter que [R1] inclut également l'application de la méthodologie à la morbidité par MCV de la CMM et contient un annexe présentant des résultats pour la mortalité et morbidité de la communauté métropolitaine de Québec (CMQ).

Le deuxième sous-objectif est de répondre plus particulièrement aux problèmes de non stationnarité et de multicolinéarité des séries. Les méthodes classiques pour répondre à ces problèmes (voir Tableau 1) sont principalement basées sur un retrait d'information (p. ex. retrait de tendance). Ainsi, il est choisi ici, non pas de retirer de l'information des séries mais de les décomposer en utilisant la décomposition modale empirique (EMD, Huang *et al.*, 1998). Le but d'EMD est d'extraire d'une série ses modes d'oscillation principaux pour les utiliser ensuite comme nouvelles variables stationnaires et orthogonales dans la régression (voir Figure 1). Ainsi, cette partie de la thèse propose une méthodologie de régression-EMD (R-EMD) puis l'applique à l'estimation de la mortalité par MCV en fonction de la température et de l'humidité quotidiennes moyennes (article [A2]). Comme la première méthodologie, la R-EMD est développée dans le rapport [R2] qui contient également des applications sur la morbidité par MCV de la CMM ainsi que sur la mortalité et morbidité de la CMQ. Les variables explicatives de [R1] incluent également les précipitations quotidiennes totales, la hauteur de neige au sol quotidienne et la pression atmosphérique quotidienne moyenne. Il est à noter que la méthodologie R-EMD a été raffinée pour donner lieu à l'article [A2]. Enfin, un ensemble de fonctions sont en cours de développement pour le logiciel `R` (R Core Team, 2015) afin d'aider à l'application de la méthodologie.

La troisième partie est née de l'observation que les problèmes d'autocorrélation et non stationnarité (Tableau 1) vient du fait qu'une série temporelle est en fait la discrétisation d'un processus intrinsèquement continu. Le troisième sous-objectif est ainsi d'exprimer les séries de données d'épidémiologie environnementale comme des fonctions continues (ou courbes, voir Figure 1). Le but de cette partie est donc d'adapter l'analyse de données fonctionnelles (FDA pour *functional data analysis*), et plus précisément la régression fonctionnelle (Ramsay et Silverman, 2005) au domaine de l'épidémiologie environnementale. L'article [A3] fait donc une présentation de la régression fonctionnelle et en applique deux modèles au lien entre la mortalité par MCV et la température. L'intérêt de cette partie est ainsi dans l'apport des modèles de régression fonctionnelle en épidémiologie environnementale, car les données utilisées s'y adaptent très bien.

1.4. Organisation de la synthèse

La synthèse s'organise comme suit. La section 2 présente une revue de littérature des méthodes de régression utilisées en épidémiologie environnementale puis d'autres méthodes existantes dans la littérature statistique pour gérer les problèmes liés à l'utilisation de séries temporelles dans la régression. Le but est de mettre en évidence les limites de ces différentes approches. Cette même section présente ensuite une revue de littérature des outils statistiques utilisés dans la thèse à savoir, l'agrégation temporelle, la méthode EMD et la régression fonctionnelle. La section 3 présente ensuite un résumé des méthodologies proposées dans cette thèse. Une sélection de résultats de l'application de ces méthodologies au lien entre les MCV et la météorologie est ensuite présentée dans la section 4. La section 5 conclut cette synthèse en exposant les contributions scientifiques de la thèse ainsi que les perspectives ouvertes par celle-ci.

2. Revue de littérature

Cette section décrit la littérature utilisée tout au long de la thèse, en commençant par les méthodes de régression souvent utilisées aussi bien en épidémiologie environnementale qu'en statistiques appliquées. Ensuite, une revue de littérature concernant les outils statistiques utilisés dans la thèse est présentée.

2.1. État de l'art en régression avec séries temporelles

Cette sous-section se concentre sur les modèles de régression autres que ceux qui seront ensuite utilisés dans la thèse. Dans un premier temps, les modèles les plus utilisés en épidémiologie environnementale sont décrits afin de mettre en évidence leurs limites, puis d'autres modèles issus de la littérature statistique sont discutés.

2.1.1. La régression en épidémiologie environnementale

L'utilisation de modèles de régression constitue tout un pan de l'épidémiologie environnementale. À cause des problématiques inhérentes à ce type d'études, certaines pratiques et certains types de modèles se sont imposés. La première problématique importante est le temps de latence entre une exposition environnementale et sa réponse sanitaire qui diffère selon les individus. Ainsi, la mortalité ou morbidité du jour t est généralement modélisée comme dépendante de l'historique sur plusieurs jours de la variable explicative, en utilisant les modèles à effets retardés distribués (DLM pour *distributed lag models*) exprimés (Almon, 1965) :

$$y[t] = \sum_{l=0}^{L} \beta_l x[t-l] + \varepsilon[t] \tag{2}$$

où l correspond au retard entre la variable explicative et la réponse, L est le retard maximum et les β_l sont les coefficients associés à chacun des retards. Les DLMs permettent entre autres de

représenter le cumul d'une exposition (ce qui était réalisé par l'utilisation de moyennes mobiles avant). Dans l'équation (2), les $x[t-l]$ étant fortement colinéaires, une contrainte est ajoutée pour l'estimation des β_l, les forçant à s'ajuster à une fonction lisse (souvent polynomiale, Schwartz, 2000a). Les DLM ont longtemps été étudiés en économétrie (p. ex. Schmidt, 1974; Mitchell et Speaker, 1986; Gelles et Mitchell, 1989) avant d'être introduits en épidémiologie environnementale. Ils ont depuis été grandement utilisés (p.ex. Schwartz, 2000a; Braga *et al.*, 2001a; Schwartz *et al.*, 2004), notamment pour modéliser précisément l'effet de moisson qui désigne un taux de mortalité particulièrement bas après un pic (p.ex. Braga *et al.*, 2001b). L'utilisation de tels modèles a permis de préciser l'ensemble des temps de réponse à un stress environnemental, en mettant notamment en évidence un effet plus persistant du froid que de la chaleur (p.ex. Braga *et al.*, 2002).

La principale limitation des DLM tels qu'exposés dans l'équation (2) est qu'ils ne sont pas capables de retranscrire une relation non linéaire, comme peuvent le faire les modèles additifs généralisés (GAM pour generalized additive models en anglais, Hastie et Tibshirani, 1986). Or, beaucoup de relations étudiées en épidémiologie environnementale ne sont pas linéaires comme en témoigne la grande popularité des GAM (p. ex. Schwartz, 1993; 1994; Dominici *et al.*, 2002; Doyon *et al.*, 2008; Dukić *et al.*, 2012). Par exemple, la relation entre mortalité et température prend souvent une forme de « J » avec une zone de confort aux alentours de 20°C et un fort effet positif des extrêmes (p. ex. Doyon *et al.*, 2006). C'est pourquoi les modèles non linéaires à effets retardés distribués (DLNM pour *distributed lag nonlinear models* en anglais) ont été développés par Gasparrini *et al.* (2010) et s'expriment :

$$y[t] = S\left(x[t]; l\right) + \varepsilon[t] \tag{3}$$

où $S(.;.)$ est une fonction bidimensionnelle possiblement non linéaire, une dimension représentant la variable explicative $x[t]$, et l'autre dimension représentant le retard. Le modèle (3) permet donc à la fois de modéliser une relation non linéaire entre une réponse sanitaire et une exposition et de modéliser l'évolution de cette relation selon le retard. Les DLNM ont déjà été massivement utilisés dans les applications épidémiologiques (p. ex. Vutcovici *et al.*, 2013; Wu *et al.*, 2013; Gasparrini *et al.*, 2015; Yang *et al.*, 2015; Phung *et al.*, 2016). L'utilisation de ces modèles a ainsi permis, par exemple, de mettre en évidence l'impact très important du froid sur une longue période, ce qui ne se voyait pas sur des modèles ne prenant pas en compte le passé complet de l'exposition. De plus, la grande popularité de ces modèles a permis de souligner des impacts différents des températures extrêmes selon les pays (Gasparrini *et al.*, 2015).

Malgré leur pertinence, les DLNM ne traitent pas directement des problèmes du Tableau 1, notamment des problèmes de non stationnarité et multicolinéarité. La non stationnarité est généralement gérée en ajoutant le terme supplémentaire $s(t)$ aux modèles de régression (par exemple (3)), où t représente le temps et $s(.)$ est une fonction lisse, souvent des Splines cubiques (p. ex. Peng et Dominici, 2008). Le terme $s(t)$ permet ainsi de capter la tendance et la saisonnalité de la série réponse $y[t]$ (en épidémiologie, on parle de « prendre en compte les confondants non mesurés »). Si l'ajout de $s(t)$ permet d'obtenir de bons ajustements des modèles de régression, il implique que l'on considère que la variable explicative $x[t]$ n'agit sur $y[t]$ qu'à court terme alors qu'il se peut que ce ne soit pas le cas (p. ex. Burr *et al.*, 2015).

Le problème de multicolinéarité est peu discuté et les études d'épidémiologie environnementale se concentrent en général sur l'étude d'une seule variable explicative. Quand une autre variable environnementale est incluse, elle ne sert en général qu'à titre de variable

confondante et n'est pas modélisée par DLNM (p. ex. l'humidité relative quotidienne dans Phung *et al.*, 2016) et la potentielle multicolinéarité n'est pas discutée. Ceci s'explique par la méthode d'estimation des DLNM dans laquelle la fonction $S(.;.)$ est exprimée en utilisant des Splines bidimensionnelles (Gasparrini *et al.*, 2010). Le nombre de coefficients à estimer pour obtenir $\hat{S}(.;.)$ est donc déjà important dans le cas d'une seule variable explicative et en utiliser plusieurs rend l'estimation très incertaine. Une des réponses à ce problème peut-être l'utilisation de modèles synoptiques (Sheridan, 2002) qui utilisent une classification de types de météo (tropical, modéré, etc...) en utilisant plusieurs variables. Cependant, ces approches très qualitatives ne permettent pas la discrimination des facteurs environnementaux.

2.1.2. Autres méthodes de régression

Les problèmes résumés dans le Tableau 1 apparaissent à divers degrés dans différents domaines et donc de nombreux modèles ont été développés pour les prendre en compte (également évoqués dans le Tableau 1). L'autocorrélation des résidus $\varepsilon[t]$ de la régression a notamment été très étudiée, ayant donné lieu à la « régression pour séries temporelles » (p. ex. Choudhury *et al.*, 1999). Dans celle-ci, un modèle de série temporelle (souvent un mélange autorégressif-moyenne mobile, appelé ARMA pour *autoregressive-moving average*) est ajusté sur les résidus de la régression $\varepsilon[t] = y[t] - f\left(x[t]\right)$ (cf. équation (1)). À noter que la régression pour séries temporelles est légèrement différente des modèles ARMA avec variable exogène (ARMAX) où c'est la variable réponse qui est modélisée comme un ARMA au lieu des résidus (p. ex. Shumway et Stoffer, 2000). Une fois les résidus modélisés, leur autocorrélation peut être estimée et un estimateur de $f(.)$ prenant en compte cette autocorrélation peut être appliqué. L'estimateur de régression pour séries temporelles le plus connu est celui des moindres carrés

généralisés (Aitken, 1935b) qui existe aussi pour $f(.)$ non linéaire (Gallant et Goebel, 1976). D'autres méthodes d'estimation existent cependant comme le préblanchiment (Cochrane et Orcutt, 1949) et l'estimation conjointe des coefficients de régression et du ARMA des résidus par maximum de vraisemblance (Pesaran, 1973; Pagan et Nicholls, 1976).

Les modèles de régression pour séries temporelles concernent le cas de séries autocorrélées mais néanmoins stationnaires. Lorsque les séries ne sont pas stationnaires, les méthodes les plus simples sont l'estimation et le retrait de la tendance et de la saisonnalité par lissage (p. ex. Schwartz *et al.*, 1996) et la différentiation des séries (p. ex. Cryer et Chan, 2008). Or, cela revient à retirer de l'information qui pourrait être pertinente pour comprendre le lien entre santé et météo. D'autres approches existent comme la modélisation par cointégration (Engle et Granger, 1987), ou encore les modèles à hétéroscédasticité conditionnelle (GARCH, pour *generalized autoregressive conditional heteroskdasticity*) lorsque la non stationnarité concerne la variance (Bollerslev, 1986). Dans tous les cas, éviter les problèmes liés à l'utilisation de séries temporelles dans la régression nécessite une bonne connaissance de la forme des séries temporelles (est-ce que les séries ont une saisonnalité, une tendance, sont intégrées, ont une autocorrélation trop importante, etc.). Connaître plus précisément ces propriétés nécessite l'application d'une batterie de tests comme celui de Dickey et Fuller (1979), de racine unitaire (Phillips, 1987) et bien d'autres (voir Ventosa-Santaulària, 2009). Ces étapes alourdissent une analyse, en particulier pour des praticiens non experts en statistiques. Il est à noter que certains de ces modèles ne sont pas inconnus en épidémiologie environnementale, comme les modèles ARMA (Lin et Xiraxagar, 2006) ou GARCH (Modarres *et al.*, 2014) mais ces études restent marginales.

Enfin, la multicolinéarité est sans doute le problème le plus simple à traiter de nos jours dans la mesure où des méthodes applicables directement existent, à savoir la régression Ridge (Hoerl et Kennard, 1970) et le Lasso (Tibshirani, 1996). Ces deux méthodes sont liées étant toutes deux des régressions linéaires biaisées, c.-à-d. que l'ajustement se fait par moindres carrés auxquels est ajoutée une pénalisation de la forme $\lambda\sum_{j=1}^{P}\left|\beta_j\right|^q$, avec $q=2$ pour la régression Ridge et $q=1$ dans le cas du Lasso. Le paramètre λ contrôle la sévérité de la pénalisation et est généralement choisi par validation croisée (VC, Stone, 1974). La régression Ridge est plus adaptée dans le cas où les P variables explicatives peuvent agir sur la réponse, alors que le Lasso est plutôt utilisé dans l'hypothèse où seul un sous-ensemble des variables explicatives agit sur la réponse (Friedman *et al.*, 2010). En effet, contrairement à la régression Ridge, le Lasso force un certain nombre de β_j à être nuls, opérant ainsi une sélection des variables. Le Lasso est très populaire et a vu de nombreuses extensions telles que le Lasso adaptatif qui est plus consistant asymptotiquement (Zou, 2006), le Lasso groupé où les variables sont incluses ou exclues en groupes (Yuan et Lin, 2006a), le Lasso pour modèles linéaires généralisés (Park et Hastie, 2007) ainsi que des homologues non linéaires (Ravikumar *et al.*, 2009; Marra et Wood, 2011). Il est également à noter le développement du « filet élastique » (*elastic net* en anglais) qui autorise $1<q<2$ afin de profiter conjointement des forces de la régression Ridge et du Lasso (Zou et Hastie, 2005). Si la théorie est la pratique sont bien établies pour gérer la multicolinéarité, à notre connaissance aucune méthode similaire existe lorsque les données sont des séries temporelles (on peut citer un Lasso pondéré par des retards, Park et Sakaori, 2013). Le sujet est par exemple absent du livre de Hastie *et al.* (2015) dédié à la sélection de variables.

2.2. Outils statistiques utilisés

Cette section présente et effectue une revue de littérature sommaire des outils statistiques constituant les méthodologies des trois articles [A1;A2;A3].

2.2.1. L'agrégation d'une série temporelle

L'agrégation d'une série temporelle, et de manière générale le lissage, est un sujet très étudié en statistiques (ce sujet tient une place prépondérante dans les livres de, p. ex. Green et Silverman, 1994; Sarda et Vieu, 2000; Schimek, 2000). L'agrégation la plus simple est la moyenne mobile qui consiste à remplacer la valeur d'une série temporelle par la moyenne des H valeurs l'entourant (le paramètre H est souvent appelé « taille de fenêtre » ou « largeur de bande »), c.-à-d. $\tilde{y}[t] = H^{-1} \sum_{i=-(H-1)/2}^{(H-1)/2} y[t+i]$. La moyenne mobile est d'ailleurs régulièrement utilisée en épidémiologie environnementale pour agréger plusieurs retards de la variable explicative (p.ex. Schwartz et Marcus, 1990; Braga *et al.*, 2001a; Sarmento *et al.*, 2011). Le problème de la moyenne mobile est le fait d'attribuer un poids identique ($1/H$) à toutes les valeurs utilisées pour calculer $\tilde{y}(t)$ (Schwartz *et al.*, 1996), rendant un résultat peu lisse. En épidémiologie environnementale, il est plus intuitif de considérer des poids plus faibles pour des jours éloignées. Ce type de pondération peut être obtenu avec une méthode de lissage Nadaraya-Watson qui est similaire à une moyenne mobile mais en attribuant des poids suivant une fonction définie K_H appelée noyau (Nadaraya, 1964; Watson, 1964), c.-à-d. :

$$\tilde{y}[t] = \frac{\sum_{i=-(H-1)/2}^{(H-1)/2} K_H(i)\, y[t+i]}{\sum_{i=-(H-1)/2}^{(H-1)/2} K_H(i)} \qquad (4)$$

Dans le cas où $K_H(i) = 1/H$, on retombe sur la moyenne mobile. De nombreuses fonctions noyaux existent comme le noyau tri-cube (Cleveland *et al.*, 1988) ou le noyau gaussien, mais la plus utilisée est le noyau d'Epanechnikov (Epanechnikov, 1969) qui a la propriété de minimiser l'erreur quadratique d'ajustement de $\tilde{y}[t]$ à $y[t]$ pour un H donné. Wand et Jones (1995, p. 31) font cependant remarquer que le choix du noyau n'a finalement que peu d'impact comparé à l'importance du choix du paramètre H. Une autre méthode d'agrégation connue est la régression locale (ou Loess, Cleveland et Devlin, 1988) qui est très similaire à l'agrégation par noyaux, mais avec l'ajustement d'une fonction polynomiale au lieu de linéaire comme les noyaux. Les agrégations les plus connues sont toutes symétriques, c.-à-d. qu'elles attribuent autant de poids aux $i < 0$ qu'aux $i > 0$ (voir p. ex. Sarda et Vieu, 2000). Cependant, il existe des méthodes par noyaux attribuant plus de poids d'un côté ou de l'autre de la valeur courante, comme le noyau d'Epanechnikov asymétrique (Michels, 1992), ou encore le lissage exponentiel qui utilise uniquement le passé ($i < 0$) pour lisser la série d'intérêt (p. ex. Gardner, 1985).

Un autre type de méthodes pour lisser une série sont les méthodes spectrales qui consistent à exprimer une série par des fonctions oscillantes de base et à retrancher les oscillations de plus hautes fréquences. Les oscillations de base peuvent être de simples sinus et cosinus dans le cas de l'analyse de Fourier (p. ex. Cooley *et al.*, 1969) mais aussi des ondelettes (p. ex. Antoniadis *et al.*, 2001) ou encore des composantes obtenues par EMD (p. ex. Boudraa et Cexus, 2007).

2.2.2. La décomposition modale empirique

Cette section présente la décomposition modale empirique (EMD), introduite par Huang *et al.* (1998) qui représente la transformation appliquée aux données dans la méthodologie de

régression-EMD développée dans la présente thèse. La méthode EMD vise à décomposer une série de données $x[t]$, $t=1,...,n$, en un ensemble de K composantes $c_k[t]$ (voir Figure 1) :

$$x[t]=\sum_{k=1}^{K} c_k[t]+r_K[t] \tag{5}$$

où $c_k[t]$ est appelée « fonction modale intrinsèque » (*intrinsic mode function*, IMF, en anglais) et $r_K[t]$ est le reste de la décomposition que l'on associe à la tendance de la série $x[t]$. Une IMF $c_k[t]$ est définie selon la seule contrainte qu'elle doit osciller symétriquement autour de zéro (tel qu'illustré dans la Figure 1). Cela lui permet de représenter un intervalle de fréquences particulier. L'algorithme EMD est expliqué en détail par Huang et Wu (2008). Il a été développé comme alternative aux décompositions de Fourier et d'ondelettes afin de traiter les séries à la fois non stationnaires (non traitées par Fourier) et non linéaires (qui ne sont traitées ni par Fourier ni par les Ondelettes). En effet, la méthode EMD est entièrement empirique et la définition des IMFs laisse une grande flexibilité dans la décomposition, là où les méthodes de Fourier et d'ondelettes imposent des fonctions spécifiques pour les composantes. La grande force de la méthode EMD est de résulter en un faible nombre de composantes (dans nos applications, K ne dépasse jamais 15) facilement interprétables en pratique (Huang *et al.*, 1999).

Les propriétés d'EMD évoquées ci-dessus ont rendu la méthode très populaire en mathématiques appliquées. Ainsi, de nombreuses études étendant ou améliorant l'algorithme sont parues. En effet, certains détails comme les effets de bord (au début et à la fin de série) et le critère d'arrêt de l'algorithme ont été améliorés (p. ex. Rilling *et al.*, 2003; Rato *et al.*, 2008). L'algorithme a également connu des extensions dont les plus importantes sont l'EMD d'ensemble (EEMD, Wu et Huang, 2009; Mandic *et al.*, 2013) qui permet de s'assurer que les IMFs ont des fréquences bien séparées (ce qui n'est pas toujours le cas avec l'algorithme de base,

Huang *et al.*, 1999) et les versions bivariées puis multivariées de la méthode (MEMD) qui permettent de décomposer conjointement plusieurs variables d'un même processus (Rilling *et al.*, 2007; Rehman et Mandic, 2010). Il est à noter que d'autres modifications de l'algorithme encore peu utilisées dans la littérature ont été imaginées telles que l'EMD statistique (Kim *et al.*, 2012) ou l'EMD d'ensemble avec bruit adaptatif (Torres *et al.*, 2011).

En plus des études visant à améliorer et étendre l'algorithme, de nombreuses publications étudiant ses propriétés ont vu le jour, que ce soit pour l'algorithme de base (p. ex. Flandrin *et al.*, 2004b; Haiyong et Qiang, 2006; Huang *et al.*, 2009; Yang et Yang, 2009; Tsakalozos *et al.*, 2012; Wang et Li, 2012) ou pour l'EEMD (p. ex. Niazy *et al.*, 2009; Zhang *et al.*, 2010; Colominas *et al.*, 2013). Les applications potentielles de l'EMD ont également été grandement discutées, que ce soit pour la détection de tendance (p. ex. Flandrin *et al.*, 2004a; Wu *et al.*, 2007) ou le filtrage d'une série (p. ex. Flandrin *et al.*, 2004a; Boudraa et Cexus, 2007). Pour filtrer une série, il est également à mentionner le test d'hypothèse visant à déterminer les IMFs contenant significativement de l'information (Wu et Huang, 2004). Parmi les applications potentielles, l'utilisation de l'EMD dans la régression n'a fait l'objet que d'une application (Yang *et al.*, 2011b).

Finalement, en plus de passionner nombre de mathématiciens appliqués, la méthode EMD a fait l'objet de nombreuses applications dans des domaines divers tels que l'océanographie (p. ex. Huang *et al.*, 1999), l'hydrologie (p. ex. Lee et Ouarda, 2010; Durocher *et al.*, 2015), la climatologie (p. ex. Lee et Ouarda, 2011; 2012), la géophysique (p. ex. Huang et Wu, 2008), la sismologie (p. ex. Loh *et al.*, 2001), l'astronomie (p. ex. Coughlin et Tung, 2004), l'économétrie (p. ex. Chuanrui, 2010) ou encore la santé (p. ex. Yang *et al.*, 2010; Xie *et al.*, 2014). Ainsi, la méthode EMD crée beaucoup d'émulation et la suite logique est le développement d'une

méthodologie de régression basée sur son utilisation, au même titre que les décompositions de Fourier (p. ex. Treagust *et al.*, 1980) et d'ondelettes (Donoho et Johnstone, 1994).

2.2.3. La régression fonctionnelle

L'analyse de données fonctionnelles (FDA) a été introduite par Ramsay (1982) et popularisée principalement par Ramsay et Silverman (2005). La FDA vise à modéliser statistiquement des données sous forme de courbes $x(t)$, $t \in T$, où T est un domaine continu et non discret comme dans le cas d'une série de données (voir Figure 1). Ainsi, le cadre FDA s'adapte particulièrement bien à l'étude des processus indexés par le temps. Bien que le domaine soit encore relativement jeune, de nombreuses méthodes statistiques classiques ont maintenant une version fonctionnelle, que ce soit les statistiques descriptives et l'analyse en composantes principales (Ramsay, 1982), l'analyse canonique des corrélations (He *et al.*, 2003), la classification (p. ex. Ternynck *et al.*, 2016), les statistiques spatiales (p. ex. Dabo-Niang et Yao, 2007) et la régression (chapitres 12 à 16 de Ramsay et Silverman, 2005).

Plusieurs types de régression fonctionnelle existent : l'ANOVA fonctionnelle (FANOVA), la régression fonctionnelle pour réponse scalaire (RFS), le modèle concurrent et la régression fonctionnelle pour réponse fonctionnelle (RFF). Ces différents modèles sont décrits dans le Tableau 2 et revus en détail par Morris (2015). Dans la présente thèse, seuls les RFS et RFF sont utilisés. En effet, même si le modèle FANOVA a un potentiel d'application en épidémiologie (p. ex. pour classifier des courbes de décès), il ne répond pas particulièrement aux problématiques exposées dans la section 1 car elles ne traitent pas les variables explicatives d'intérêt (météorologie) comme continues. Le modèle concurrent, quand à lui, n'explique un lien entre la réponse fonctionnelle et la variable explicative fonctionnelle qu'au même temps t, ce qui

ne permet pas de considérer le retard entre les deux, alors qu'il s'agit d'un aspect important d'une relation entre la santé et l'environnement.

Tableau 2 : Liste des différents modèles fonctionnels existant.

Nom	Réponse	Variable explicative	Forme du coefficient β	Référence
FANOVA	Fonctionnelle $y_i(t)$	Scalaire x_i	Fonctionnel $\beta(t)$	Brumback et Rice (1998)
RFS	Scalaire y_i	Fonctionnelle $x_i(t)$	Fonctionnel $\beta(t)$	Hastie et Mallows (1993)
Concurrent	Fonctionnelle $y_i(t)$	Fonctionnelle $x_i(t)$	Fonctionnel $\beta(t)$	Hastie et Tibshirani (1993)
RFF	Fonctionnelle $y_i(t)$	Fonctionnelle $x_i(s)$	Surface $\beta(s,t)$	Ramsay et Dalzell (1991)

La régression fonctionnelle pour réponse scalaire (ici désignée par l'acronyme RFS, mais appelée en anglais *functional regression for scalar response* ou encore *scalar-on-function regression*) exprime l'effet de $x_i(t)$ à chaque temps $t \in T$ continu, sur la réponse scalaire y_i. La RFS est le modèle de régression fonctionnelle le plus étudié (Cardot *et al.*, 1999; 2003; Ramsay et Silverman, 2005; Goldsmith *et al.*, 2011; Brockhaus *et al.*, 2015). À la base linéaire, la RFS a vu un certain nombre d'extension, à savoir la version modèle linéaire généralisé (James, 2002), les GAM fonctionnels (McLean *et al.*, 2014) et la régression fonctionnelle non paramétrique (Ferraty et Vieu, 2004). Grâce à ses propriétés attractives, la RFS a été appliquée dans beaucoup de domaines comme l'hydrologie (p. ex. Masselot *et al.*, 2016b), l'écologie (p. ex. Bel *et al.*, 2011; Stewart-Koster *et al.*, 2014), la médecine (p. ex. Ratcliffe *et al.*, 2002a; 2002b), la spectrométrie (p. ex. Ferraty et Vieu, 2002), l'économétrie (p. ex. Sood *et al.*, 2009) et récemment en épidémiologie environnementale (Arisido, 2016, paru pendant la réalisation de

cette étude). À noter que l'étude d'Arisido (2016) se concentre sur la pollution sur des individus exposés, alors que l'article [A3] sur concentre plus sur la météorologie.

Le deuxième type de régression fonctionnelle considérée ici est la régression avec à la fois une réponse et une variable explicative fonctionnelles (RFF, souvent appelée *function-on-function regression* ou *fully functional regression* en anglais). Elle permet d'exprimer l'influence de chaque temps s de la variable explicative $x_i(s)$ sur chaque temps t de la réponse $y_i(t)$, et est donc très générale. La RFF étant un modèle complexe, plusieurs méthodes d'estimation ont vu le jour afin d'obtenir des estimateurs à la fois consistants et efficaces en temps de calcul (Ramsay et Dalzell, 1991; Besse et Cardot, 1996; Yao *et al.*, 2005; Ivanescu *et al.*, 2014; Brockhaus *et al.*, 2015). La RFF étant plus complexe que la RFS, moins d'extensions ont été développées. La régression fonctionnelle historique (RFH) constitue néanmoins un ajout notable car elle vise précisément les données temporelle et exprime la réponse fonctionnelle à tous temps t par rapport au prédicteur aux seuls temps s qui précèdent t. Ce modèle a intéressé différents auteurs avec plusieurs méthodes d'estimation différentes à cause de la forme particulière du coefficient à estimer (Malfait et Ramsay, 2003; Şentürk et Müller, 2010; Kim *et al.*, 2011; Brockhaus *et al.*, 2016). La RFF et la RFH ont moins été appliquées que la RFS, mais on peut tout de même noter des applications en hydrologie (Masselot *et al.*, 2016b), médecine (Hosseini-Nasab et Mirzaei, 2014) et neurologie (Meyer *et al.*, 2015).

Pour finir, il est à noter qu'une méthode d'estimation unifiant tous les modèles du Tableau 2 a récemment été proposée par Brockhaus *et al.* (2015; 2016). Cette méthode est basée sur le boosting (Bühlmann et Hothorn, 2007) et permet d'estimer efficacement n'importe quel modèle fonctionnel. L'autre nouveauté importante de cette méthode d'estimation est la très récente possibilité de considérer plusieurs variables explicatives dans la RFF et la RFH ainsi que d'opérer

une sélection des variables. Il s'agit donc d'une avancée importante pour la régression fonctionnelle, permettant ainsi sa diffusion au domaine de l'épidémiologie environnementale.

3. Résumé des méthodologies proposées

Cette section présente les méthodologies proposées dans la présente thèse.

3.1. Agrégation temporelle de la réponse dans la régression

La méthodologie proposée consiste en deux étapes :

1. Agrégation temporelle de la série réponse $y[t]$ pour obtenir une série lissée $\tilde{y}[t]$;
2. Explication de la série $\tilde{y}[t]$ par les expositions $x_j[t]$ en utilisant un modèle de régression pour séries temporelles.

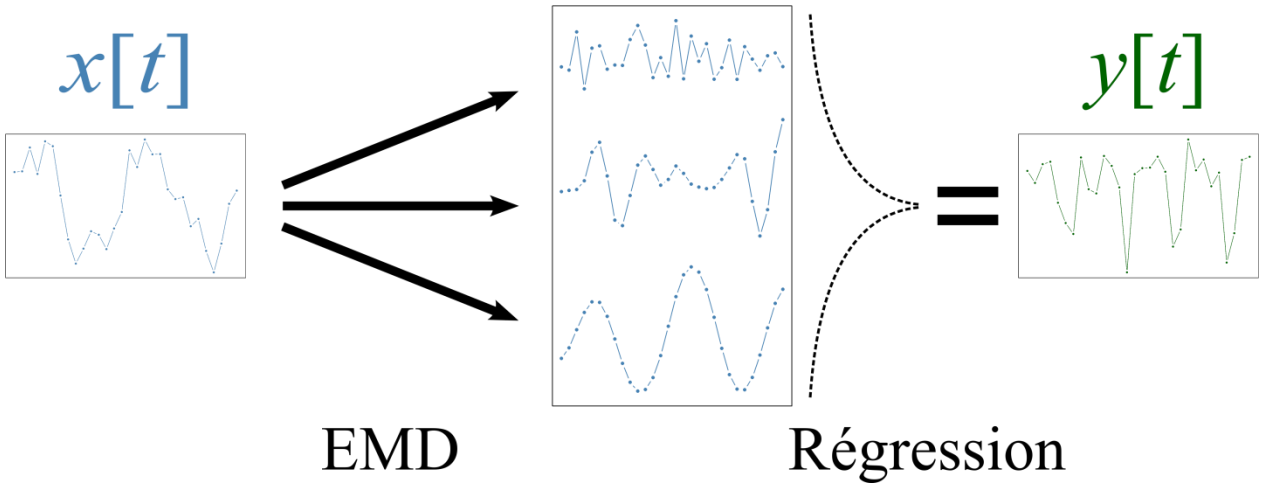
La première étape consiste à l'application d'une des méthodes de lissage revues dans la section 2.2.1. Des agrégations locales (c.-à-d. calculées à partir d'observations proches de t) sont privilégiées afin de représenter l'ensemble de la réponse à une exposition, en plus de réduire le bruit. Un autre avantage des agrégations locales est qu'elles permettent d'utiliser plus facilement des critères comme la VC par blocs pour évaluer la performance de la modélisation. Ainsi, des méthodes utilisant la série entière pour l'agrégation comme le filtrage EMD et par séries de Fourier sont écartées (ces méthodes ont tout de même été testées dans [R1]).

Une fois la série lissée $\tilde{y}[t]$ obtenue, elle est introduite comme réponse dans un modèle de régression. Cependant, deux problèmes émergent : 1) la distribution de $\tilde{y}[t]$ et 2) l'autocorrélation créée. Le premier point possède une réponse finalement assez simple : dans tous les cas, la distribution converge vers une loi de Gauss. En effet, dans le cas des agrégations

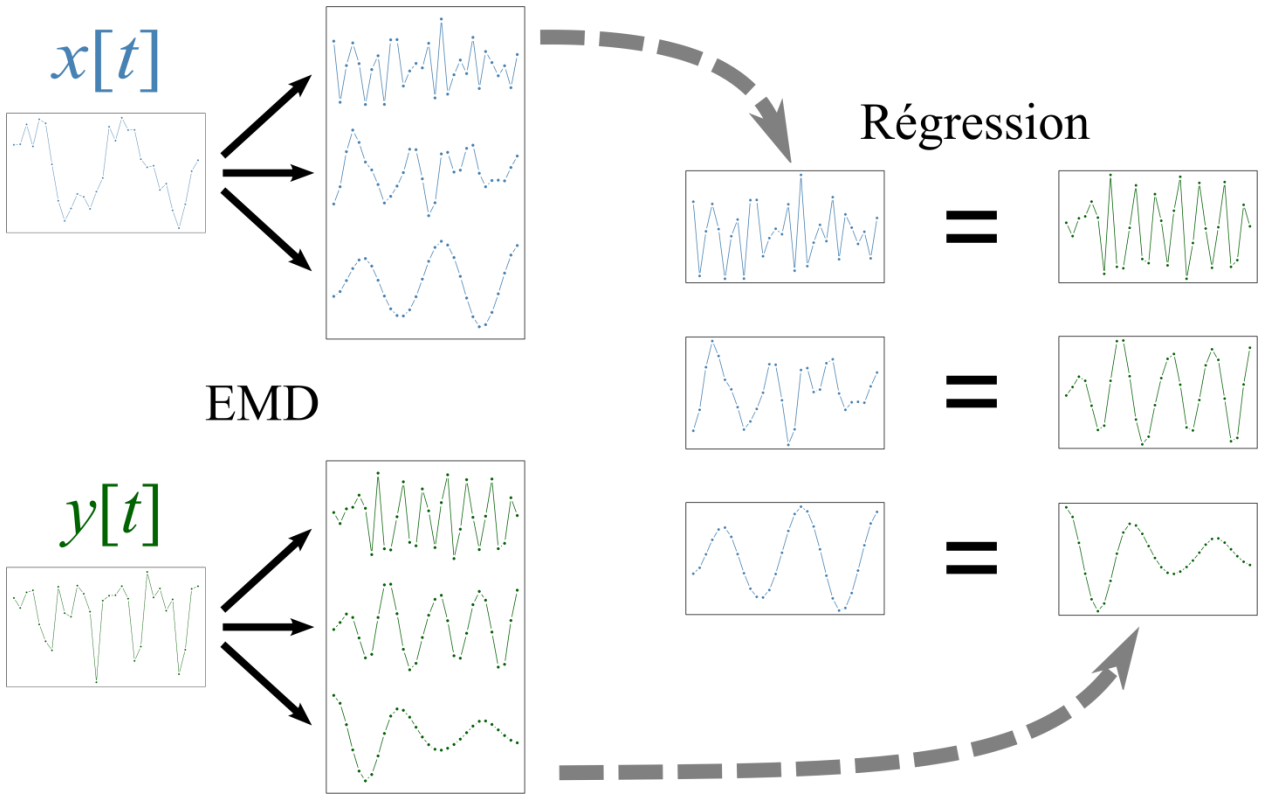
linéaires, cette convergence s'explique par le théorème de la limite centrale généralisé pour des variables dépendantes (p. ex. le chapitre 27 de Billingsley, 1995).

L'agrégation temporelle crée de la dépendance entre les observations successives de $\tilde{y}[t]$ car elles sont calculées avec des observations communes de $y[t]$, sans compter la dépendance déjà présente à la base. Ceci est réalisé en utilisant un modèle de régression pour série temporelle dans laquelle, pour rappel, les résidus sont modélisés par un modèle ARMA (voir section 2.1.2). Le modèle de régression pour série temporelle est estimé par maximum de vraisemblance car les autres méthodes d'estimation (moindres carrés généralisés, préblanchiment) sont moins efficaces (Mizon, 1995; Choudhury *et al.*, 1999). Le principal inconvénient de cette méthode est qu'elle est pour l'instant éprouvée pour le cas linéaire uniquement. Cette méthode permet cependant l'utilisation des DLNM car leur méthode d'estimation par fonctions de base permet de revenir à un cas linéaire (voir section 2.1.1).

À noter qu'une étape préliminaire est nécessaire pour déterminer l'ordre du modèle ARMA pour les résidus (*i.e.* le nombre de coefficients associés à la dépendance temporelle). L'ordre du modèle est obtenu en appliquant une première régression classique afin d'obtenir une série de résidus $\hat{\varepsilon}[t]$, puis en appliquant une procédure de type « stepwise » sur les résidus. Le critère utilisé pour identifier le meilleur modèle est le critère d'information d'Akaike (AIC, Akaike, 1974; Hyndman et Khandakar, 2007).



a) R-EMD1



b) R-EMD2

Figure 2 : Illustration des deux versions de la R-EMD. La R-EMD1 ne décompose que les variables explicatives alors que la R-EMD2 décompose à la fois les variables explicatives et la variable réponse.

3.2. Régression-EMD

La méthodologie de base de la régression-EMD (R-EMD) consiste en deux étapes : 1) décomposition des séries de données par EMD et 2) utilisation des IMFs $c_k[t]$ comme nouvelles variables dans la régression. Cette méthodologie globale se décline en deux versions permettant deux niveaux de détails différents : (1) la R-EMD1 dans laquelle seule les variables explicatives $x_j[t]$ sont décomposées et dont les IMFs $c_{jk}[t]$ servent à expliquer la série réponse $y[t]$, et (2) la R-EMD2 dans laquelle $y[t]$ est également décomposée et chacune de ses IMFs $c_{Yk}[t]$ est expliquée par les IMFs explicatives $c_{jk}[t]$ de même ordre. Les deux versions sont schématisées sur la Figure 2 pour illustrer la différence.

Dans la R-EMD1 (Figure 2a), seules les variables explicatives $x_j[t]$ ($j = 1,...,P$) sont décomposées et leurs IMFs $c_{jk}[t]$ sont utilisées comme variables explicatives dans la régression :

$$y[t] = \sum_{j=1}^{P}\left(\sum_{k=1}^{K}\left(\beta_{jk}^{(1)} c_{jk}[t]\right) + \beta_{j(K+1)}^{(1)} r_{jK}[t]\right) + \varepsilon[t] \qquad (6)$$

où les $\beta_{jk}^{(1)}$ sont les coefficients de régression et $\varepsilon[t]$ est la série des résidus. La R-EMD1 permet de mettre en évidence les principales variations des prédicteurs agissant sur la réponse.

Afin de bonifier l'étude de la relation entre les prédicteurs et la réponse, la R-EMD2 (Figure 2b) décompose à la fois les variables explicatives et la variable réponse. Ainsi, il y a cette fois $K+1$ modèles à ajuster, c.-à-d. que chaque IMF $c_{Yk}[t]$ de la variable à expliquer est la réponse de son propre modèle de régression en fonction des IMFs explicatives de même ordre (et donc de fréquence similaire) :

$$c_{Yk}[t] = \sum_{j=1}^{P} \beta_{jk}^{(2)} c_{jk}[t] + \varepsilon[t], \quad k = 1, ..., K \tag{7}$$

et de même pour la tendance $r_{YK}[t] = \sum_{j=1}^{P} \beta_{j(K+1)}^{(2)} r_{jK}[t] + \varepsilon[t]$. La R-EMD2 peut ainsi permettre de mettre en évidence des relations à des échelles temporelles de faibles amplitudes, qui sont habituellement cachées dans la série de base $y[t]$.

Dans les deux cas (6) et (7), toutes les variables sont décomposées conjointement par la version multivariée de l'algorithme (MEMD). Conceptuellement, ceci permet de prendre en compte l'ensemble des variables météorologiques considérées comme un seul phénomène. Pratiquement, cela permet d'obtenir exactement le même nombre d'IMFs pour chaque variable et d'avoir un alignement des modes, c.-à-d. que les IMFs de même ordre k ont des fréquences similaires (Mandic *et al.*, 2013).

Pour les deux versions de la R-EMD, le Lasso (Tibshirani, 1996) est utilisé au lieu des moindres carrés ordinaires pour estimer les coefficients $\beta_{jk}^{(m)}$, $(m = 1, 2)$ dans les équations (6) et (7). En effet, dans la R-EMD1 le nombre d'IMFs explicatifs est souvent très grand car la décomposition résulte en $K \approx \log_2(n)$ IMFs (p. ex. les séries traitées dans cette thèse donnent jusqu'à 15 IMFs). Ainsi le Lasso permet de sélectionner seulement les IMFs $c_{jk}[t]$ les plus importantes pour l'explication de la réponse $y[t]$. Dans le cas de la R-EMD2, comme toutes les IMFs explicatives sont du même ordre k dans chacune des $K+1$ régressions, elles sont possiblement fortement corrélées car de fréquences similaires. Le Lasso permet donc l'estimation des coefficients $\beta_{jk}^{(2)}$ malgré cette possible corrélation, en procédant à une sélection des variables.

Comme il est expliqué dans l'introduction, il y a souvent un temps de latence entre une variable explicative et la réponse. Ainsi, il est probable que les IMFs $c_{jk}[t]$ doivent être décalées de l_{jk} jours avant de les intégrer dans le modèle de régression, afin de bien représenter leur influence sur la réponse. Le décalage l_{jk} de chaque IMF est celui qui maximise la fonction de corrélation croisée entre l'IMF $c_{jk}[t]$ et la réponse (p.ex. chapitre 11 de Cryer et Chan, 2008). Notez qu'une contrainte est ajoutée sur le décalage l_{jk}, l'obligeant à être inférieur à la période moyenne de $c_{jk}[t]$.

Toutes les IMFs $c_{jk}[t]$ obtenues n'ont pas la même amplitude, celle-ci décroissant généralement quand l'ordre k augmente (Wu et Huang, 2004). Il peut donc être difficile de comparer tous les coefficients $\beta_{jk}^{(1)}$ et $\beta_{jk}^{(2)}$. Pour faciliter la comparaison entre les coefficients, il est d'usage de calculer une mesure de sensibilité en standardisant les coefficients par l'écart-type de leurs variables explicatives. Dans le cas de la R-EMD, comme les modèles ont des séries oscillantes comme variables, cette sensibilité est obtenue en utilisant plutôt l'amplitude moyenne des IMFs, c.-à-d.

$$S_{jk}^{(m)} = \beta_{jk}^{(m)} \times A_{jk}, \quad m = 1,2 \tag{8}$$

où A_{jk} est l'amplitude crête-à-crête de l'IMF $c_{jk}[t]$. La sensibilité $S_{jk}^{(m)}$ représente ainsi l'amplitude des variations de la réponse induites par l'IMF $c_{jk}[t]$, comme illustré sur la Figure 3.

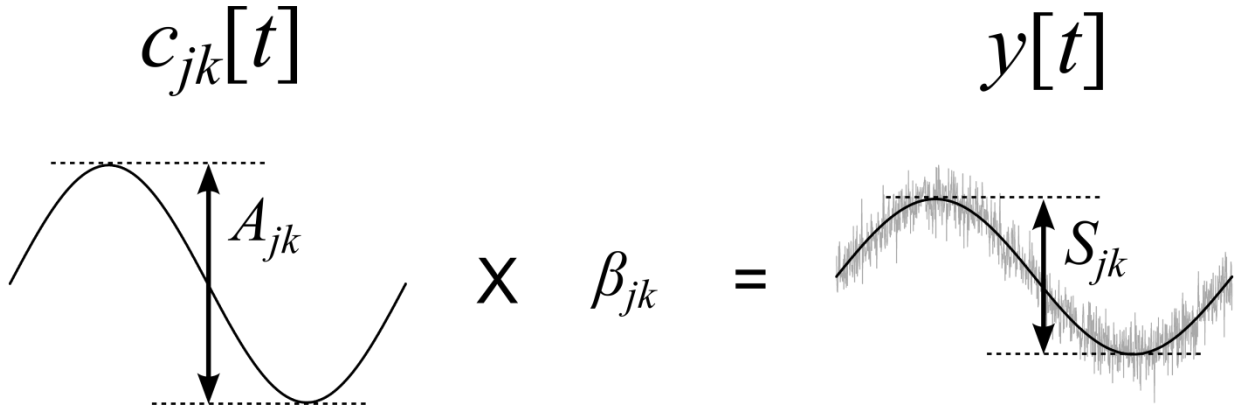


Figure 3 : Interprétation du critère de sensibilité introduit pour la R-EMD. Le calcul est identique pour la R-EMD2, mais dans la figure $y[t]$ est à remplacer par $c_{Yk}[t]$.

3.3. Adaptation de la régression fonctionnelle à l'épidémiologie environnementale

Toute analyse de données fonctionnelle contient au minimum deux étapes :

1) L'obtention des données fonctionnelles par lissage à partir de la série de données observées;
2) L'application de la méthode fonctionnelle d'intérêt (ici la régression) sur les données fonctionnelles obtenues à la première étape.

La première étape consiste en un découpage de la série $x[t]$ ($t = \{1,...,n\}$) en un ensemble de N sous-séries représentant la même période T (comme illustré dans la Figure 1), dont chacune est transformée en donnée fonctionnelle $x_i(t)$ où cette fois $t \in T$ et $i = 1,...,N$. Afin d'obtenir des courbes continues, les données fonctionnelles sont exprimées :

$$x_i(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{9}$$

où les $\phi_k(t)$ ($k = 1,...,K$) sont des fonctions de base connues analytiquement, typiquement des bases B-spline ou des bases de Fourier. Ainsi, l'obtention des données fonctionnelles consiste en l'estimation des coefficients c_k, ce qui se fait par moindres carrés pénalisés (Ramsay et Silverman, 2005, chapitre 5), de manière similaire à la régression Ridge.

Le découpage de la série dépend des données et de l'objectif de l'analyse, mais dans le cas de l'épidémiologie environnementale, deux découpages semblent naturels : le découpage quotidien ($T = [0;24]$ où l'unité correspond à une heure) et le découpage annuel ($T = [0;365]$ où l'unité correspond au jour). Le découpage quotidien représente le cycle jour/nuit de la variable considérée (p. ex. température) qui peut avoir une influence sur la santé humaine (p. ex. Vutcovici *et al.*, 2013). Le découpage annuel est le plus évident à cause de la périodicité annuelle des variables environnementales. Ainsi, une donnée fonctionnelle $x_i(t)$ représente une année de la variable considérée (p. ex. température, humidité), ce qui permet d'avoir un jeu de données stationnaires car la saisonnalité est contenue dans $x_i(t)$.

De nombreuses modélisations de la relation entre la santé et l'environnement sont possibles grâce à la régression fonctionnelle, mais l'article [A3] en propose deux basées respectivement sur les découpages quotidiens et annuels. Dans le premier cas, comme les variables sanitaires sont rarement disponibles avec un pas de temps plus faible qu'une journée, il est proposé d'utiliser la RFS pour exprimer le nombre de cas quotidien (de l'issue sanitaire considérée) en fonction des variations d'une variable explicative au sein d'une journée. Le modèle s'exprime ainsi :

$$\ln(y_i) = s(i) + \beta_0 + \int_0^{24} x_{i-1}(t)\beta_1(t)dt + \varepsilon_i \qquad (10)$$

où y_i est le nombre de cas au jour i, $x_{i-1}(t)$ est la courbe de la variable explicative environnementale du jour précédent, $\beta_1(t)$ est le coefficient fonctionnel de régression indiquant l'effet de $x_{i-1}(t)$ sur y_i à chaque instant $t \in [0;24]$ heures, $s(i)$ est une fonction lisse du temps pour prendre en compte la saisonnalité et la tendance de y_i, et β_0 est l'ordonnée à l'origine du modèle. Le logarithme naturel est utilisé pour la réponse car les issues sanitaires suivent généralement une loi de Poisson (Schwartz *et al.*, 1996), ce qui se vérifie sur les données de mortalité utilisées dans la présente thèse. Le modèle (10) permet ainsi l'explication du lien entre une variable environnementale et la santé à très court terme, ce qui est d'intérêt notamment pour comprendre l'impact sanitaire des vagues de chaleur en été.

La deuxième modélisation proposée se base sur des données fonctionnelles annuelles, afin de modéliser l'évolution annuelle du lien entre une issue sanitaire et une variable explicative environnementale. Dans cette modélisation, la réponse est également fonctionnelle et la régression utilisée est la RFF :

$$y_i(t) = s(i) + \beta_0(t) + \int_{t-60}^{t} x_i(u)\beta_1(u,t)du + \varepsilon_i(t) \qquad (11)$$

où $y_i(t)$ et $x_i(u)$ $(u,t \in [0;365])$ sont les courbes de la réponse et de la variable explicative de l'année i, $\beta_1(u,t)$ est le coefficient fonctionnel (sous forme de surface car bidimensionnel) donnant l'effet de $x_i(u)$ sur $y_i(t)$, $s(i)$ est une fonction lisse sur les années pour contrôler une éventuelle tendance, $\beta_0(t)$ est la constante du modèle (correspondant à la courbe moyenne des $y_i(t)$), et $\varepsilon_i(t)$ est le résidu fonctionnel. Ce modèle est contraint pour n'expliquer $y_i(t)$ qu'en fonction de $x_i(u)$ où $t-60 < u < t$, c.-à-d. en fonction des 60 jours précédents, lesquels constituent une borne supérieure au temps de latence maximum reporté dans la littérature (qui est

de 40 jours, Peng et Dominici, 2008). Le modèle (11) permet ainsi de modéliser une issue sanitaire en fonction de l'historique d'une variable environnementale, comme dans le cas du DLM, mais également de modéliser l'évolution de cette relation au cours de l'année (car $\beta_1(u,t)$ dépend aussi de t), alors que le DLM considère cette relation comme fixe.

Les deux modèles (10) et (11) proposés se généralisent naturellement à plusieurs variables explicatives fonctionnelles grâce à la méthode d'estimation récente de Brockhaus *et al.* (2015). De plus, il est possible de rajouter des variables explicatives non fonctionnelles pour contrôler les variables potentiellement confondantes (p. ex. une variable dichotomique pour la fin de semaine).

4. Applications réalisées

Toutes les méthodes proposées dans la thèse ont été appliquées au lien entre les maladies cardiovasculaires et la météorologie au Québec. On présente ici une sélection des principaux résultats décrits dans les articles [A1;A2;A3]. Chacun de ces articles présente une application sur la relation entre température et mortalité car il s'agit du cas le plus simple et le plus courant. Seul l'article [A2] inclut également l'humidité comme variable explicative supplémentaire pour bien montrer que la méthodologie de R-EMD a été pensée dans un cadre de régression multiple. Des applications plus complètes et couvrant plusieurs cas (mortalité et morbidité pour les communautés métropolitaines de Québec et Montréal) ont été réalisées avec les deux première méthodologies dans des rapports pour l'Institut national de santé publique du Québec (INSPQ) [R1;R2]. Des applications non présentes dans l'article [A3] ont également été réalisées avec la régression fonctionnelle; elles sont jointes à l'annexe.

4.1. Données

Les données utilisées dans les articles proviennent de la région métropolitaine de Montréal (CMM) qui comprend la ville de Montréal et ses banlieues (voir Figure 4). Il s'agit du plus grand et plus dense bassin de population au Québec ce qui permet d'avoir un grand nombre de cas de MCV dans une aire relativement restreinte pour laquelle la météorologie peut être considérée homogène (Chebana *et al.*, 2012b). Différentes variables mesurées au sein de la CMM sont utilisées selon les applications. Elles sont résumées dans le Tableau 3.

Figure 4 : Emplacement et carte de la communauté métropolitaine de Montréal.

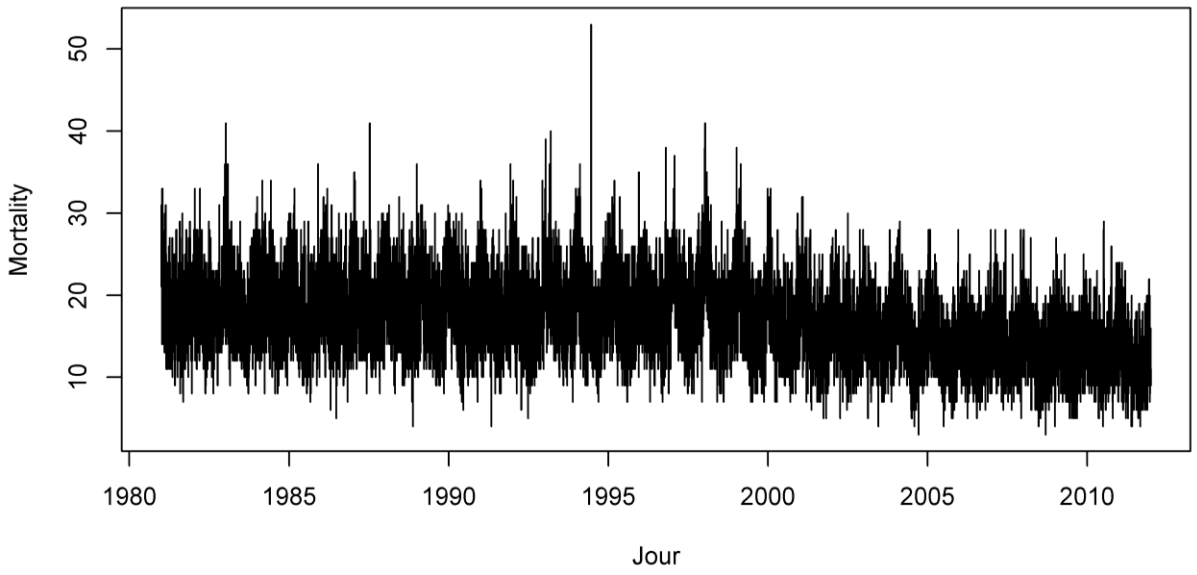
Les données sanitaires sont la mortalité [A2;A3] pour causes de MCV. Les MCV regroupent les maladies coronariennes (codes I20 à I25 dans la dixième version de la Classification internationale des maladies, CIM10, et codes 410-414 en CIM9 avant l'année

2000), l'insuffisance cardiaque (I50 dans la CIM10, 428 en CIM9) et les maladies vasculaires cérébrales (G45, H34.0, H34.1, I60, I61, I63 et I64 dans la CIM10, et 362.3, 430, 431, 434.x, 435.x en CIM9). Les données sont fournies par l'INSPQ; elles étaient disponibles de 1981 à 2011 au moment de la thèse et sont montrées dans la Figure 5a.

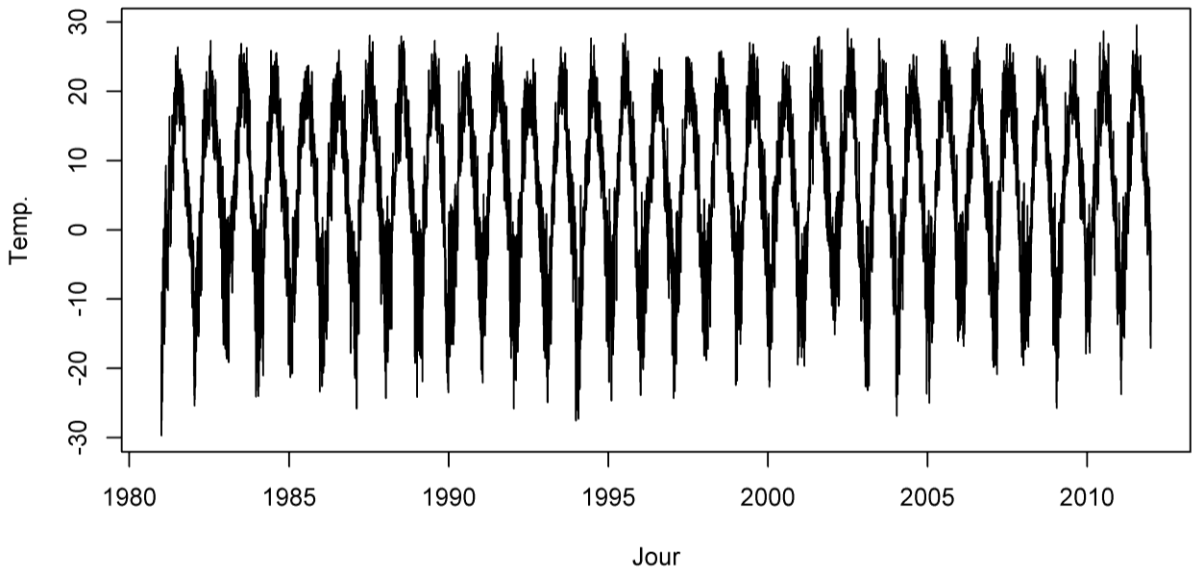
Tableau 3 : Résumé des données utilisées dans chacune des applications présentées dans la thèse.

Article	Application	Réponse	Variable(s) explicative(s)	Pas de temps	Période
[A1]	Agrégation de la réponse	Mortalité	Température	Quotidien	1981-2011
[A2]	R-EMD	Mortalité	Température, Humidité	Quotidien	1981-2011
[A3]	RFS	Mortalité	Température	Quotidien (Mortalité) / Horaire (Température)	2007-2011
[A3]	RFF	Mortalité	Température	Quotidien	1981-2011

Les données météorologiques utilisées proviennent de deux sources différentes. Dans trois applications, elles sont quotidiennes de 1981 à 2011 et issues du portail de téléchargement *Données Accès et Intégration* (DAI) d'Environnement Canada. Ces données sont illustrées sur la Figure 5. Cependant, pour l'application de la RFS (le modèle exprimé par l'équation (10)), des données de températures horaires ont été nécessaires. Elles ont donc été fournies par le *Ministère du Développement durable et de la Lutte contre les changements climatiques* (MDDELCC) et sont disponibles de 2007 à 2011 inclusivement. Dans tous les cas, les données sont celles de multiples stations de mesure dispersées dans le territoire de la CMM, sur lesquelles une moyenne spatiale a été appliquée pour n'obtenir qu'une seule série de données (Giroux *et al.*, 2013).



a) Mortalité pour cause de MCV



b) Température quotidienne moyenne

Figure 5 : Évolution temporelle des données de mortalité pour cause de MCV et de température de la communauté métropolitaine de 1981 à 2011.

4.2. Principaux résultats

Cette section présente les principaux résultats obtenus pour les méthodologies proposées sur les données du Tableau 3. Tous les résultats sommaires ci-dessous sont montrés et discutés avec plus de détails dans les articles correspondant.

4.2.1. Réponse agrégée

Les résultats de l'article [A1] sont présentés en deux étapes. La première étape consiste à effectuer une comparaison d'un DLNM avec réponse agrégée à un DLNM classique. L'agrégation effectuée dans cette étape est la moyenne mobile centrée de 7 jours, comme Sarmento *et al.* (2011). Les DLNM sont ajustés selon le même schéma que Gasparrini *et al.* (2015). Les résultats montrent que la méthodologie de régression pour réponse agrégée s'ajuste mieux aux données que le DLNM classique ($R^2 = 32\%$ contre 22% pour le DLNM classique), mais qu'il n'y pas de gain concernant la prédiction de données. En termes d'interprétation, les surfaces obtenues montrent que les DLNMs avec réponse agrégée se concentrent sur la relation entre température et mortalité à plus long terme, alors que le DLNM classique se concentre davantage sur la relation à court terme. Ainsi, les DLNMs avec réponse agrégée mettent en exergue l'influence de l'hiver avec 5 à 10 jours de retard, alors que le modèle classique met plus en avant l'effet immédiat du froid (ainsi que de la chaleur estivale). Ceci s'explique par le fait que les hautes fréquences sont retirées de la réponse par le lissage. Cependant, il est à noter que les mêmes motifs ressortent lorsque l'on applique la méthodologie avec des tailles de fenêtre plus importantes, ce qui suggère qu'il y a un signal important en hiver avec une semaine de latence. Ce résultat est connu dans la littérature sur le sujet, mais l'utilisation de l'agrégation permet de le mettre en évidence et le modéliser. L'amélioration des performances explicatives indique

également que l'effet court terme pourrait être surestimé habituellement, ce qui est corrigé avec la méthode proposée ici.

La deuxième étape des résultats est d'investiguer des agrégations différentes de la moyenne mobile, à savoir l'agrégation par noyau d'Epanechnikov et de Michels ainsi que le Loess. Les résultats montrent que le choix de l'agrégation influe peu sur les performances, à l'exception du Loess qui nécessite des tailles de fenêtre d'agrégation plus importantes que les autres méthodes pour obtenir des performances similaires. Cependant, la taille de fenêtre H a une influence, car les meilleures performances (d'ajustement aux données notamment) sont réalisées pour des H compris entre 3 et 7 jours. Ces résultats sont cohérents avec la littérature statistique qui met plus l'accent sur la choix de H que du noyau (voir p. ex. Wand et Jones, 1995). Ainsi, ce résultat de peu d'influence du noyau peut être considéré généralisable à d'autres problématiques. Il est cependant recommandé d'utiliser le noyau d'Epanechnikov qui permet, théoriquement, un ajustement optimal pour un H donné et dont la forme représente bien l'évolution de l'effet d'une exposition sur la santé.

4.2.2. Régression-EMD

Dans l'article [A2], la R-EMD1 (6) et la R-EMD2 (7) sont toutes deux appliquées à l'influence de la température et de l'humidité sur la mortalité par MCV. Deux variables explicatives météorologiques sont utilisées conjointement dans cette application pour bien montrer que la R-EMD a été pensée pour un cadre de régression multiple. Afin de présenter les résultats sous une forme compacte, le Tableau 4 montre les sensibilités $S_{jk}^{(m)}$ obtenues dans l'application, mettant en exergue celles qui sont considérées significativement différentes de zéro.

Tableau 4 : Sensibilités obtenues par application de la R-EMD. Les valeurs en rouge sont significativement différentes de zéro à 95%, c.-à-d. que l'intervalle de confiance calculé par bootstrap ne contient pas la valeur zéro.

Ordre k	Période moyenne	Sensibilités $S_{jk}^{(m)}$			
		Température		Humidité	
		R-EMD1	R-EMD2	R-EMD1	R-EMD2
1	3 jours	0.00	0.21	0.13	0.18
2	5 jours	0.00	0.16	0.12	0.11
3	9 jours	0.00	0.00	**-0.14**	0.00
4	16 jours	0.00	0.16	-0.04	**-0.18**
5	1 mois	0.00	0.00	-0.16	**-0.25**
6	2 mois	0.00	0.00	**-0.20**	**-0.22**
7	3 mois	**-0.20**	**-0.43**	-0.01	**-0.12**
8	6 mois	**-0.29**	**-0.63**	0.09	**0.35**
9	1 an	**-4.48**	**-4.22**	0.00	0.00
10	2 ans	0.00	0.00	0.00	**-0.12**
11	4 ans	0.00	**-0.42**	0.00	**-0.22**
12	9 ans	0.00	0.00	0.00	**0.35**
13	25 ans	-	**0.07**	-	**0.66**
r	-	**-4.16**	**-9.05**	0.00	**-4.6**

Les sensibilités $S_{jk}^{(1)}$ de la R-EMD1 et $S_{jk}^{(2)}$ de la R-EMD2 sont relativement cohérentes les unes avec les autres dans la mesure où les IMFs significatifs dans la R-EMD1 le sont aussi dans la R-EMD2, et ce avec des sensibilités d'amplitude similaires, avec la seule exception de l'IMF d'humidité de période 9 jours. La principale différence est que les résultats de la R-EMD2 sont plus détaillés que ceux de la R-EMD1, avec plus d'IMFs significatifs, notamment les IMFs d'humidité d'échelle pluriannuelle.

Les résultats mettent en perspective l'influence de la température et de l'humidité. La première agit principalement à une échelle comprise entre 3 mois et 1 an de périodicité, avec une sensibilité fortement négative. Les sensibilités $S_{Temp;9}^{(2)} = -4.48$ et $S_{Temp;9}^{(2)} = -4.22$ indiquent qu'à cause du froid, l'hiver compte chaque jour un peu plus de 4 décès de plus que l'été. Les sensibilités $S_{Temp;r}^{(m)}$ de la tendance de températures sont également très importantes, suggérant une

forte association entre la hausse des températures moyennes (les changements climatiques) et la baisse de la mortalité observée (une baisse supérieure à 4 décès par jours entre 1981 et 2011). Ce résultat est cependant atténué par le fait que la baisse de mortalité s'explique en partie par l'arrivée de nouveaux traitements (Luepker, 2011), ce qui est difficile à intégrer comme confondant dans un modèle en raison de l'absence de données à ce sujet dans les fichiers médico-administratifs.

L'humidité semble plutôt agir à une échelle comprise entre 2 semaines et 3 mois. Les sensibilités significatives sont toutes négatives, indiquant que des périodes sèches peuvent résulter en une augmentation de la mortalité, en particulier au printemps et à l'automne. En ajoutant les sensibilités correspondantes (cf. Tableau 4), les résultats indique un excès de mortalité pendant ces périodes sèches, compris entre 0.34 décès par jours (ou 1 décès tous les trois jours) selon la R-EMD1 et 0.77 décès par jours selon la R-EMD2. En effet, les IMFs d'humidité avec des sensibilités $S_{Hum;k}^{(m)}$ significatives ont une amplitude plus importante pendant ces saisons. Les périodes sèches semblent donc avoir une plus grande influence sur la mortalité par MCV pendant les périodes de transition entre l'hiver et l'été.

Après l'interprétation des résultats, une comparaison de la R-EMD est réalisée avec les GAM et DLNM qui sont les deux modèles les plus populaires en épidémiologie environnementale. Il est montré que les modèles de R-EMD ont un grand pouvoir explicatif et prédictif. En effet, ils montrent des valeurs de R^2 de 26 et 28 % alors de que les GAM et DLNM montrent des valeurs à 10 et 17 % respectivement. De plus, l'erreur de prédiction estimée par validation croisée est de 19 pour les deux modèles R-EMD alors qu'elle est de 23 et 21 pour les GAM et DLNM respectivement. Ces résultats peuvent être surprenant sachant que la R-EMD est linéaire au contraire des GAM et DLNM. Cependant, il est montré dans [A2] que la relation non

linéaire mise en lumière dans les GAM et DLNM est en fait linéaire par morceaux et représentée par plusieurs IMFs différentes.

4.2.3. Régression fonctionnelle

Deux applications visant à illustrer les avantages de la régression fonctionnelle sont réalisées dans [A3]. La première utilise la RFS pour expliquer le nombre quotidien de décès par MCV en fonction de la courbe de température de la journée précédente, comme exprimé dans l'équation (10). Ce modèle a été réalisé pour différents mois de l'année. Seule son application sur les mois d'été (juin, juillet et août) est toutefois présentée dans l'article (pour les autres périodes, voir l'annexe A). La courbe de température de la journée entière étant utilisée, ça permet d'inclure les températures quotidiennes moyennes, minimales et maximales.

Les résultats montrent une influence positive de la courbe des températures pendant la matinée et pendant la soirée de la journée précédente, alors que l'influence est quasi nulle pendant l'après-midi. Ces résultats suggèrent que la mortalité estivale n'est pas causée par les chaleurs les plus intenses survenant l'après-midi, mais plutôt lorsque la température reste élevée pendant la matinée et la soirée. Cela met en évidence le fait que le corps a besoin de se reposer d'un stress thermique, ce qui n'est pas possible lorsque la chaleur reste élevée jour et nuit. Ce résultat est cohérent d'un point de vue physiologique (p. ex. Sawka *et al.*, 2011) mais n'est habituellement pas visible sur les modèles d'épidémiologie environnementale.

Le modèle de la première application ne peut pas être utilisé sur les données de l'année complète car la relation discutée ci-dessus n'est pas la même durant les autres saisons. Ainsi, la deuxième application utilise la RFF exprimée par l'équation (11) pour expliquer la relation entre la température et la mortalité à plus grande échelle et voir l'évolution de cette relation au cours de l'année.

La surface $\hat{\beta}(u,t)$ estimée montre une relation principalement négative au cours de l'année, montrant un effet du froid principalement. Cet effet est particulièrement fort durant le printemps et l'automne, c.-à-d. des périodes de l'année pendant lesquelles la population est moins préparée aux grands froids. De plus, la relation trouvée avec la température à la fin de l'hiver (février) est nulle voire même positive avec un retard entre 2 semaines et un mois. Ce résultat suggère une adaptation de la population au froid au cours de l'hiver, froid qui a donc moins d'impact sur la santé à la fin de la saison. Ces aspects de la relation entre la température et la mortalité par MCV, qui ne sont pas mis en lumière dans les modèles habituels (p. ex. DLNM), montrent que les modèles fonctionnels sont capables de retranscrire l'adaptation physiologique de la population selon la saison.

Les deux modèles décrits ci-dessus, sont finalement comparés à l'application d'un DLNM sur les mêmes données. La comparaison se fait grâce aux courbes annuelles de R^2 et de l'erreur de prédiction par validation croisée calculées sur chaque jour de l'année. Les résultats montrent que le DLNM a une valeur de R^2 plus élevée que les deux modèles fonctionnels sur quasiment toute l'année. Par contre, l'erreur de prédiction du DLNM est également plus élevée, et le RFS a l'erreur la plus basse sur l'été. Ainsi, la régression fonctionnelle ajuste moins bien les données quotidiennes qu'un DLNM, mais sont plus juste dans la prédiction de données non observées. D'un autre côté, les DLNM pourraient sur-ajuster les données.

5. Conclusion et perspectives

Cette section conclut la synthèse en résumant le travail effectué, puis en mettant l'accent sur les apports scientifiques et les perspectives ouvertes par la thèse.

5.1. Conclusion générale

La compréhension du lien entre la santé humaine et l'environnement est très importante dans un contexte de changements climatiques afin d'anticiper l'évolution de cette relation dans le futur. Une bonne anticipation de cette évolution est cruciale pour favoriser l'adaptation à ces changements, en particulier pour mettre en place des alertes plus précises et donc diminuer les impacts néfastes de l'environnement sur la santé. Au niveau de la recherche, la compréhension du lien santé-environnement passe notamment par des méthodes statistiques, qui doivent rendre compte le plus fidèlement possible de la relation entre une issue sanitaire et une variable explicative environnementale. C'est dans cette perspective que la présente thèse s'inscrit en proposant différentes méthodes statistiques pour répondre à certaines difficultés liées aux méthodes usuellement utilisées en épidémiologie environnementale. Les méthodologies proposées reposent sur un prétraitement des données (agrégation, décomposition et transformation en données fonctionnelles) pour ensuite utiliser les données transformées dans des modèles de régression adaptés.

D'abord, l'agrégation de la variable sanitaire est proposée pour tenir compte du bruit organisationnel dans les données sanitaires. Cette proposition donne suite à une méthodologie de régression applicable lorsque la variable réponse a été agrégée. Cette méthodologie se base notamment sur l'utilisation d'estimateurs de régression pour séries temporelles, afin de prendre en compte l'autocorrélation créée par l'agrégation.

Ensuite, en raison des saisonnalités et tendances importantes contenues dans les données sanitaires et environnementales, une méthode de régression-EMD est développée pour pouvoir étudier la relation d'intérêt à différentes échelles temporelles simultanément. Cette démarche se base sur une décomposition préliminaire des séries de données par EMD pour l'obtention de leurs IMFs. Une façon de prendre en compte les IMFs dans la régression est ensuite proposée, incluant notamment l'utilisation du Lasso pour sélectionner les IMFs importantes et d'un critère de sensibilité pour interpréter les résultats.

Enfin, l'utilisation de la régression fonctionnelle est proposée pour le domaine de l'épidémiologie environnementale. En effet, la nature intrinsèquement continue des processus naturels étudiés rendent pertinente l'utilisation des modèles fonctionnels dans ce domaine. Ainsi, les applications de la régression fonctionnelle présentées, ouvrent la voie à l'utilisation de la statistique fonctionnelle en épidémiologie environnementale afin d'étendre la connaissance des relations entre la santé et l'environnement.

Les méthodes de régression pour réponse agrégée et de R-EMD développées dans la présente thèse se veulent générales. Le but est de pouvoir les appliquer dans n'importe quel domaine lorsque les variables mises en relation ont des données se présentant sous forme de séries, comme les séries temporelles. Ainsi, les articles [A1] et [A2] visent à toucher un public de statisticiens appliqués et d'autres experts intéressés par les méthodes, notamment en santé et en environnement. Un package développé pour le logiciel `R` (R Core Team, 2015) facilitera l'utilisation des méthodes qui y sont décrites et intégreras différents outils d'interprétation.

Les méthodologies proposées sont illustrées plus spécifiquement par leur application au lien entre les MCV et la température (ainsi que l'humidité dans le cas de la R-EMD). Ces applications permettent de montrer comment interpréter les résultats des méthodologies et quels

en sont les avantages. Les applications sont également l'occasion de comparer les méthodologies proposées aux GAMs et aux DLNM qui sont des méthodes plus connues et plus souvent utilisées en épidémiologie environnementale. L'interprétation des résultats, ainsi que les comparaisons permettent de mettre en exergue les avantages des différentes méthodes proposées, et donc les innovations apportées par la thèse.

5.2. Apports à la santé publique et la recherche

Les applications des méthodologies décrites dans cette thèse ont apporté plusieurs éclairages intéressants pour la compréhension du lien entre la météorologie et les MCV et pouvant être utiles tant à la santé publique qu'à la recherche.

L'application de l'agrégation de la réponse illustre très bien que l'influence des températures froides sur la mortalité cardiovasculaire se fait sur un plus long terme que des températures chaudes. À l'inverse, sans agrégation de la réponse, c'est l'influence à court terme des températures chaudes qui ressort. Cette nuance est importante à considérer lors de l'établissement des seuils d'alerte, ce qui n'est pas encore dans la pratique usuelle.

Par une vision différente de la relation santé-environnement, La R-EMD a mis en évidence l'effet certain de l'humidité sur les MCV, notamment pendant les périodes sèches. De façon notable, elle permet aussi de relever l'importance de fortes variations d'humidité sur la mortalité cardiovasculaire, à l'instar des variations de température. En particulier dans les saisons de transition (printemps et automne), lesquelles constituent des périodes de l'année où la population est moins préparée à y faire face. Or, l'humidité est assez peu étudiée dans la littérature en santé environnementale et si elle l'est, c'est généralement en tant que variable potentiellement confondante (p. ex. Phung *et al.*, 2016). L'inclure aux systèmes de surveillance et

de prévention des impacts sanitaires au climat constituerait aussi une avancée majeure qui pourrait améliorer la performance de tels systèmes.

Cette influence particulière de la température pendant les saisons de transition se retrouve dans les résultats de l'application de la régression fonctionnelle, notamment la RFF. Les autres résultats importants apportés par la régression fonctionnelle sont le fait que la mortalité estivale survient lorsque la température reste élevée la nuit, et le fait que l'effet du froid s'atténue à la fin de l'hiver comparativement au début de la saison. Tous ces résultats retranscrivent en fait les mécanismes d'adaptation physiologique des populations par rapport à la météorologie ainsi que l'importance des périodes de repos pour l'organisme.

Les avantages des travaux décrits dans cette thèse ne sont pas seulement en termes de résultats épidémiologiques mais aussi en termes de méthodes statistiques. En effet, l'article [A1] pose le problème de la régression lorsque la série de données réponse a été lissée. Une méthodologie est donc proposée pour y répondre, ce qui constitue un ajout à la littérature statistique (bien que le cas d'une agrégation augmentant le pas de temps a été étudié, p. ex. Tiao et Wei, 1976), légitime dans un contexte comme l'épidémiologie environnementale. Ainsi, un avantage annexe de cette étude pourrait être d'amener la problématique dans la littérature statistique théorique.

Au-delà des nouveautés dans les résultats, la R-EMD propose une vision alternative des résultats. Dans la régression classique, les résultats sont interprétés au travers de la fonction de régression (ou des coefficients dans le cas linéaire) indiquant l'impact sur la réponse d'une augmentation d'une unité de la variable explicative. La R-EMD propose d'interpréter les résultats dans le domaine fréquentiel et permet de modéliser le fait qu'une baisse de température de 10°C (p. ex.) n'a pas nécessairement le même effet si elle est progressive ou soudaine. Si l'idée a déjà

été proposée (p. ex. Yang *et al.*, 2011a), la méthodologie de R-EMD permet une application moins naïve avec l'utilisation de méthodes statistiques justifiées et performantes.

L'évolution temporelle de la relation entre deux variables est ce qui justifie l'utilisation de la régression fonctionnelle en épidémiologie environnementale. Elle permet de modéliser le fait que la relation entre une variable environnementale et la santé n'est pas fixe mais dépend du moment de la journée ou de la saison par exemple. Cet aspect permet de prendre en compte l'adaptation physiologique, comme indiqué ci-dessus. Si les aspects d'adaptation des populations sont discutés dans la littérature (p. ex. Kinney *et al.*, 2012; Liu *et al.*, 2015), ils ne sont pas retranscrits dans les modèles classiques. Or, l'adaptation physiologique est un aspect important à prendre en compte pour mettre en place des alertes efficaces. Il s'agit donc d'un avantage important apporté par la régression fonctionnelle.

5.3. Perspectives

Les perspectives sont nombreuses pour chacune des parties de la thèse, tant sur le plan méthodologique que pratique. En effet, les méthodologies développées dans la présente thèse sont des combinaisons de méthodes statistiques existantes, mais il peut être d'intérêt de développer des méthodes théoriques spécifiquement pour les problèmes soulevés au cours de la thèse. Dans la conclusion de [A1], il est précisé que la méthode d'estimation par maximum de vraisemblance ne permet pas l'utilisation de méthodes complexes pour la régression telles que le Lasso, la régression Ridge ou encore une méthode non linéaire. Ainsi, il est important de développer des versions de ces méthodes prenant en compte des structures de dépendance temporelle dans les résidus de la régression. On peut d'ores et déjà noter les travaux d'Alkhamisi (2010), qui propose des estimateurs pour le cas où les résidus suivent au processus autorégressif d'ordre 1, mais plus de recherches sont nécessaires. De plus, la régression pour réponses agrégées

proposée se fait en deux temps, avec le lissage dans un premier temps et la régression dans un deuxième temps. Une perspective pourrait donc être d'estimer conjointement les paramètres de lissage et la régression en fonction des prédicteurs.

La méthodologie R-EMD montre également une limite importante, soit la gestion des retards entre le prédicteur et la réponse. En effet, le retard est en réalité souvent distribué alors que la R-EMD considère un retard simple. Le problème est que ce type de modèles est peu performant dès que plusieurs prédicteurs sont utilisés, ce qui rend impossible l'utilisation d'un grand nombre d'IMFs. Une perspective peut donc être dans le développement d'un hybride entre DLM (ou DLNM) et Lasso.

Les perspectives les plus grandes font suite à l'introduction de la régression fonctionnelle en épidémiologie environnementale, ouvrant la porte à de nombreuses applications potentielles. Notamment, si la RFS et la RFF ont été discutées, la régression avec réponse fonctionnelle et prédicteur scalaire peut-être d'intérêt dans le domaine, pour exprimer la forme de la courbe de mortalité ou de morbidité en fonction de catégories telles que la saison, les jours de semaine la zone climatique, etc… Des travaux plus théoriques sur la statistique fonctionnelle sont nécessaires, notamment au niveau de l'inférence pour développer des tests permettant de déterminer si un prédicteur peut être considéré comme agissant significativement sur l'issue sanitaire d'intérêt.

Enfin, en termes pratiques, il est nécessaire d'appliquer les méthodes développées et discutées dans la présente thèse à différents cas d'études, que ce soit en épidémiologie environnementale ou dans d'autres domaines. Pour l'épidémiologie environnementale, on peut notamment citer les maladies respiratoires et l'influenza pour les issues sanitaires, mais aussi les polluants atmosphériques pour les prédicteurs environnementaux.

PARTIE II :

ARTICLES

Article 1 :

Aggregating the response in time series regression models with an application to weather-related cardiovascular diseases

-

Agrégation de la réponse dans les modèles de régression pour séries chronologiques avec application à l'effet de la météorologie sur les maladies cardiovasculaires

Pierre Masselot[1*], Fateh Chebana[1], Diane Bélanger[1,2], André St-Hilaire[1], Belkacem Abdous[3], Pierre Gosselin[1,2,4], Taha B.M.J. Ouarda[1]

[1]*Institut National de la Recherche Scientifique, Centre Eau-Terre-Environnement, Québec, Canada;*

[2]*Centre Hospitalier Universitaire de Québec, Centre de Recherche, Québec, Canada;*

[3]*Université Laval, Département de médecine sociale et préventive, Québec, Canada;*

[4]*Institut national de santé publique du Québec (INSPQ), Québec, Canada.*

Soumis

Article 2 :

EMD-regression with application to weather-related cardiovascular mortality

-

Régression-EMD avec application à l'effet de la météorologie sur la mortalité par maladie cardiovasculaire

Pierre Masselot[1*], Fateh Chebana[1], Diane Bélanger[1,2], André St-Hilaire[1], Belkacem Abdous[3], Pierre Gosselin[1,2,4]

[1]*Institut National de la Recherche Scientifique, Centre Eau-Terre-Environnement, Québec, Canada;*

[2]*Centre Hospitalier Universitaire de Québec, Centre de Recherche, Québec, Canada;*

[3]*Université Laval, Département de médecine sociale et préventive, Québec, Canada;*

[4]*Institut national de santé publique du Québec (INSPQ), Québec, Canada.*

Soumis

Résumé

De nombreux domaines de recherches utilisent des données sous forme de séries temporelles dans les analyses par régression. De telles données sont souvent multi-échelle et non-stationnaires, ce qui mène à des résultats peu précis et donc moins fiables. Afin de répondre à cette problématique, le présent article présente une méthodologie de régression-EMD qui consiste en l'application de l'algorithme EMD aux séries de données pour utiliser les composantes en résultant dans un modèle de régression. En plus de prendre la problématique de non-stationnarité en compte, la méthodologie proposée agit comme un scan de la relation entre des prédicteurs et la réponse à différentes échelles temporelles, montrant ainsi différents aspects de la relation. À des fins d'illustration, la méthodologie est appliquée à l'étude de l'effet de la température sur la mortalité par maladie cardiovasculaire de la communauté métropolitaine de Montréal, Canada. Cette application met en évidence de nouvelles caractéristiques de la relation et montre de meilleures performances de la régression-EMD par rapport aux modèles additifs généralisés et aux modèles non-linéaires à effet retardé distribué. La régression-EMD proposée est générale et peut-être considérée dans de nombreuses applications. Elle permet de rendre compte des résultats en termes d'échelle temporelle.

Abstract

In many research fields and applications, regression analyses are often performed using time series data. Such data are often multi-scale and non-stationary, leading to a poor accuracy of the resulting regression models and therefore to less reliable results. To manage this issue, the present paper introduces the EMD-regression methodology consisting in applying the *empirical mode decomposition* (EMD) algorithm on data series and then using the resulting components in regression models. The proposed methodology accounts of the issues of non-stationarity of the data series. In addition, this approach acts as a scan of the relationship between a response and the predictors at different time scales, providing new insights about this relationship. To illustrate the methodology, it is applied to the problem of weather-related cardiovascular mortality in Montreal, Canada. This application outlines new features in the relationship and shows that EMD-regression outperforms generalized additive models and distributed lag models in this application. The proposed EMD-regression is general and can be considered in a variety of situations and applications to provide new insights. It allows communicate in terms of time scale effects and results in an increase of the explained proportion of mortality compared to classical methods.

1. Introduction

Climate change adaptation is a major public health challenge of the forthcoming decades since its impact on human health is already observed (Patz *et al.*, 2014). Since climate change will result in a modification of many environmental variables, addressing the climate change challenge necessitates a clear understanding of how these variables affect human health. Therefore, there is an increasing literature assessing the effect of environmental exposures on some health issue through statistical models. Environmental exposures include, for instance, atmospheric pollutants (e.g. Martins *et al.*, 2006; Kan *et al.*, 2010) and weather variables (e.g. Keatinge, 2002; Anderson et Bell, 2009; Bassil *et al.*, 2009; Gasparrini et Armstrong, 2011) while the health issue is either general mortality or morbidity (e.g. Peng *et al.*, 2006; Yu *et al.*, 2012), either specific diseases such as respiratory illnesses (e.g. Qiu *et al.*, 2012) or cardiovascular diseases (e.g. Braga *et al.*, 2002; Houck *et al.*, 2005; Törő *et al.*, 2010; Huang *et al.*, 2012).

The natural way to model the effect of environmental exposure on a health issue is through regression analysis. The data used are often measures of the exposures (*e.g.* measures of ozone level) at regular time intervals (often daily) and a quantity of death or hospitalization cases during the same interval. Hence, such data are under the form of time series, which causes accuracy issues in the regression parameters estimation (Granger et Newbold, 1974). Indeed, natural data, as well as health data, are often non-stationary (*i.e.* the moments vary with time) and some dominant patterns in time series (such as annual cycles) create a large amount of multicollinearity in the exposure time series when several covariates are considered. In a regression analysis, if the model does not take account of this issue, it can lead to an increase in the variability of parameter estimates, making the final result less reliable. Furthermore, this

increases the chances of making wrong conclusions concerning whether or not a predictor influences the response (i.e. the so-called "spurious regression" issue, see Granger et Newbold, 1974; Phillips, 1986). In the time series literature, the usual methods to solve the non-stationarity issue are to remove the trend and the seasonality from the series, to apply a difference operator to the series, or to add a “time variable” to the regression model (especially in the field of environmental epidemiology). However, performing one of these operations means focusing on short time scales while we might also be interested by all the scales (Peng et Dominici, 2008; Burr *et al.*, 2015), especially in times of climate change.

The problematic exposed above comes from the fact that processes as complex as climate or diseases spreading are multiscale, meaning that number of natural time scales are embedded in the data. The weather is an especially good example since a lot of different cycles can be identified such as the daily cycle, the annual cycle, some longer climate cycles etc… Hence, it is relevant to consider several time scales by decomposing the time series in basic oscillating components (Kelsall *et al.*, 1999). In the environmental epidemiology literature, it has been considered by Schwartz (2000b) as well as Dominici *et al.* (2003) in the context of regression. In these two papers, the decomposition is respectively achieved with the seasonal-trend using Loess (STL) algorithm (Cleveland *et al.*, 1990) and the discrete Fourier transform (DFT, Cooley *et al.*, 1969). Note that time series decomposition can also be achieved using the discrete wavelet transform (DWT, Daubechies, 1992) such as performed by some authors in the hydrologic literature (Kucuk et Agiralioglu, 2006; Kişi, 2009). However, all these methods lack practical flexibility and objectivity since it is necessary to fix the time scales of interest in the STL algorithm, as well as for the DFT and the DWT. Another drawback of all these studies is that they did not address the issue of including several predictors in the models. This must be

addressed because many factors can impact the human health and a single predictor seems too simplistic.

In the present paper, we propose to entirely let the data tell us what are the time scales to consider. This goal can be achieved by decomposing time series into data-driven basic oscillating components called intrinsic mode functions (IMFs) using the empirical mode decomposition algorithm (EMD, Huang *et al.*, 1998). Unlike the components resulting from the DFT and DWT, the individual IMFs are flexible enough to catch small irregularities in the series, resulting in a complete decomposition with a limited number of components, while other methods such as the DFT and DWT need a high number of components to represent irregular series (Flandrin *et al.*, 2004b). Since the IMFs represent separate frequency bands, they are orthogonal to each other and few variations of their moments. All these attractive features make EMD the perfect tool for modelling the time-frequency of natural processes (see Huang *et al.*, 2008 for several examples).

The EMD method has received an increasing interest in the literature and has been extensively studied in order to increase its applicability by assessing its properties (e.g. Flandrin *et al.*, 2004b; Haiyong et Qiang, 2006; Rato *et al.*, 2008) and improving the algorithm (e.g. Junsheng *et al.*, 2006; Rilling *et al.*, 2007; Wu et Huang, 2009). Moreover, EMD has been applied in very different fields such as oceanography (Huang *et al.*, 1999), climatology (Lee et Ouarda, 2011; Lee et Ouarda, 2012), hydrology (Lee et Ouarda, 2010; Durocher *et al.*, 2015; Ghouse *et al.*, 2015), seismology (Loh *et al.*, 2001) and econometrics (Zhang *et al.*, 2008). However, the method is little known in epidemiology in spite of recent work (Yang *et al.*, 2010; 2011a; 2013; Xie *et al.*, 2014).

The present paper aims at proposing a general EMD-regression (EMD-R) methodology, which can be used in any regression analysis where data are time series. In this methodology

some or all of the time series data are decomposed through EMD and the resulting IMFs are used in a regression model. Moreover, since the entirety of time scales may not be relevant to understand the effect of interest, we use the Lasso (Tibshirani, 1996) as a variable selection method in order to discard irrelevant IMFs from the model. The model is performed according to two designs: i) only the predictor(s) are decomposed in order to model the most important scales of influence and ii) both the predictor(s) and the response are decomposed in order to look at the effect of the exposure with more details. The second design is expected to improve the explicative power of the model to better understand the effect of interest. The details of the methodology are presented in section 2. To illustrate its properties and strengths, EMD-R is then applied to the public health problem of weather-related cardiovascular diseases (CVD) mortality in the Greater Montreal region (Canada) from 1981 to 2011. The data and results are presented in section 3. The obtained results are discussed in section 4 along with a comparison with common regression models in environmental epidemiology, and finally section 5 concludes.

2. EMD-regression (EMD-R)

The EMD-regression methodology aims at explaining the effect of covariates X_j on a response Y by: 1) decomposing the time series using EMD and 2) using the IMFs as new variables in a sparse regression model, namely the Lasso. Two designs are considered: 1) only the covariate(s) X_j are decomposed and their IMFs are used to explain the original response series (EMD-R1) and 2) both the covariate(s) X_j and the response Y are decomposed and each IMF from Y is the response of its own regression model (EMD-R2), similarly to multivariate regression models **(e.g. Mardia *et al.*, 1979)**. The whole methodology is summarized in Figure 2.1.

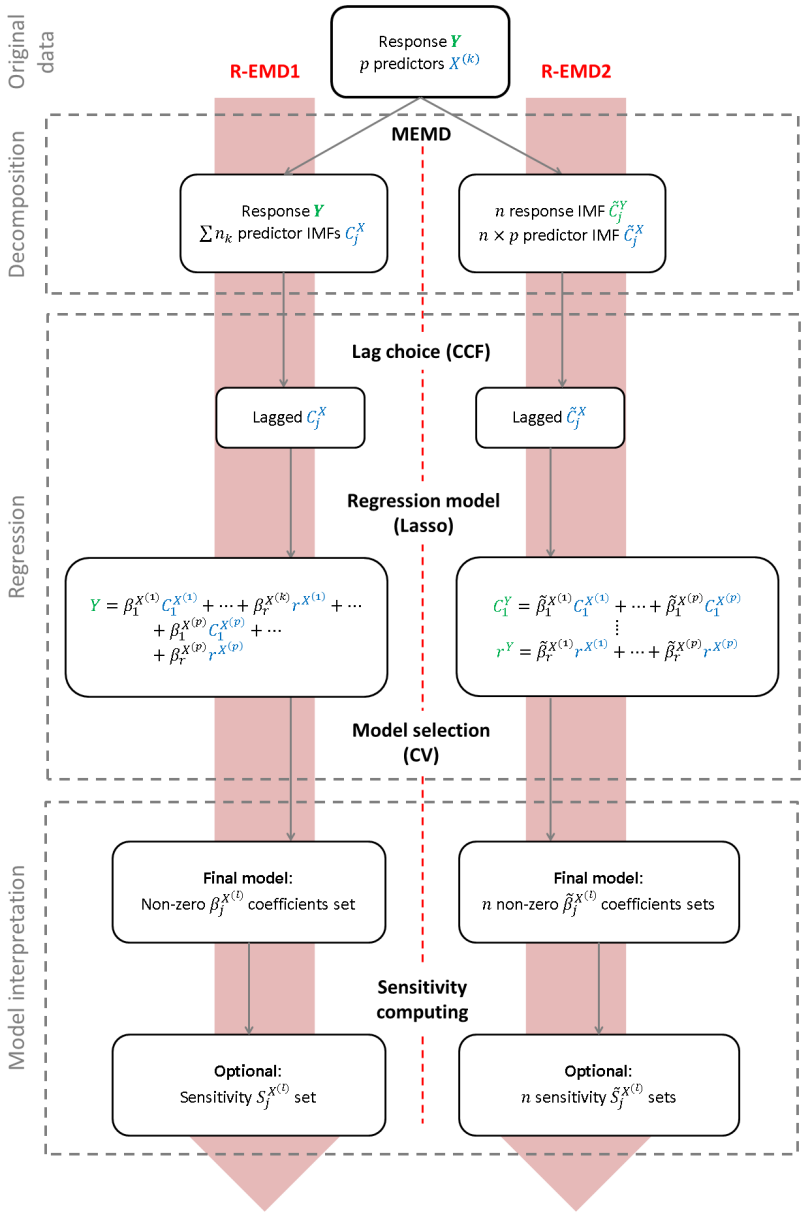


Figure 2.1: EMD-regression (EMD-R) methodology summary.

2.1. Background

2.1.1. Empirical mode decomposition (EMD)

The EMD algorithm (Huang *et al.*, 1998) has been introduced in order to identify a series $X(t)$ oscillation modes $C_k(t)$ without making any *a priori* assumption on the shape of these oscillations. This is achieved by decomposing the series such as

$$X(t) = \sum_{k=1}^{K} C_k(t) + r(t) \tag{16}$$

where the $C_k(t)$ $(k = 1, ..., K)$ are intrinsic mode functions (IMFs), *i.e.* series oscillating around the zero line with symmetric upper and lower envelopes (Huang *et al.*, 1998). This definition of IMFs allows the $C_k(t)$ s to have meaningful physical interpretation while still representing separate frequency bands.

The IMFs $C_k(t)$ are obtained iteratively, by extracting them from the data beginning with the smallest periodicity to the largest, through a sifting process. The sifting process consists in three steps: 1) obtain the upper and lower envelopes $u(t)$ and $l(t)$ of the data series $X(t)$ by respectively connecting the local maxima and minima through a cubic spline in order to compute their mean $m(t) = \big(u(t) - l(t)\big)/2$, 2) subtract this mean $m(t)$ to $X(t)$ to obtain the first IMF prototype $h_1(t) = X(t) - m(t)$ and 3) repeat the steps 1 and 2 on $h_1(t)$ until obtaining a prototype $h_l(t)$ which corresponds to an IMF, according to some chosen stopping criterion (for different stopping criteria, see Rilling *et al.*, 2003; Huang et Wu, 2008). The prototype $h_l(t)$ is then the first IMF $C_1(t)$ and the whole sifting process is performed again on the residue

$r_1(t) = X(t) - C_1(t)$ to obtain $C_2(t)$, then on $r_2(t) = r_1(t) - C_2(t)$ and then again until obtaining a monotonic residue $r_K(t)$. $r_K(t)$ is then considered as the trend of the signal.

When jointly considering $p > 1$ variables, it is useful to obtain the IMFs with the multivariate EMD (MEMD, Rehman et Mandic, 2010). It allows to obtain the exact same number of IMFs for each variable, and to obtain mode alignment, *i.e.* two IMFs of the same order from different variables have similar frequency bands (Rehman *et al.*, 2013). The MEMD algorithm works by making a large number of univariate projections of the multivariate series. The envelopes of univariate projections are computed as in the univariate EMD and form together a multidimensional envelope from which a mean can be computed and subtracted from the series (Rilling *et al.*, 2007). Apart from the envelope computation, the MEMD algorithm is identical to the EMD algorithm.

Finally, to avoid mode mixing (i.e. the mixing of very different frequency bands inside one IMF, Huang *et al.*, 1999), we use the noise-assisted ensemble methods originally developed by Wu et Huang (2009) for the univariate EMD and later extended by Rehman et Mandic (2011) to the multivariate case (NA-MEMD). The NA-MEMD algorithm adds one or several white noise variables to the multivariate signal in order to make MEMD acting as dyadic filter. Once the MEMD has been performed, the dimensions corresponding to the noises in the resulting multivariate IMFs are discarded. The NA-MEMD is used in the EMD-R but, for simplicity, it is still referred as EMD throughout the paper.

2.1.2. The Lasso

The Lasso (for least absolute shrinkage and selection operator, Tibshirani, 1996) is a shrinkage method performing a variable selection while fitting the regression model

$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon$. Instead of the ordinary least squares (OLS), the Lasso fits the model through the minimization of the penalized criterion

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j X_j \right) + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\} \quad (17)$$

where λ is a tuning parameter controlling the severity of the penalization. The greater λ is, the lower the number of covariates X_j remaining in the model at the end is. In practice, λ is automatically chosen through cross-validation (for details, see Friedman *et al.*, 2009). Note that the Lasso can also be fit with when the response is not Gaussian, similarly to generalized linear models (Park et Hastie, 2007; Friedman *et al.*, 2010).

Note that generically, theoretical standard error computation does not exist for Lasso estimates of the β_j s (Lockhart *et al.*, 2014). In practice it is always useful to compute quantities such as p-values or confidence intervals for the estimates, which are based on these standard errors. Thus, to reach this goal, it is necessary to use nonparametric methods. In particular, the bootstrap allows the estimation of a distribution for the β_j s estimates (Chatterjee et Lahiri, 2011). This aspect is discussed with great details in the chapter 6 of Hastie *et al.* (2015). Note that, since the data used in EMD-R are time series, it is necessary to use a block bootstrap method (Lahiri, 1999) to account for the dependence structure of data.

2.2. EMD-regression presentation

As stated in the introduction of this section, the EMD-R methodology contains two designs. In the EMD-R1 design, only the predictors $X_j(t)$ are decomposed and their IMFs $C_{jk}^{(1)}(t)$ are used to explain the original response series $Y(t)$:

$$Y(t) = \sum_{j=1}^{p}\left(\sum_{k=1}^{K} \beta_{jk}^{(1)} C_{jk}^{(1)}(t) + \beta_{jr}^{(1)} r_{jK}^{(1)}(t) \right) + \varepsilon(t) \tag{18}$$

EMD-R1 is similar as the study of Yang *et al.* (2011b) and is meant to give an overall view of the relationship.

The EMD-R2 design is meant to be more accurate and to give additional insights of the relationship at scales with low energy, which are hidden in the original series. In the EMD-R2 design, both the predictors $X_j(t)$ and the response $Y(t)$ are decomposed, leading to the models

$$C_{Yk}^{(2)}(t) = \sum_{j=1}^{p} \beta_{jk}^{(2)} C_{jk}^{(2)}(t) + \varepsilon_k(t) \text{ for } k = 1,...,K \tag{19}$$

and the model $r_{YK}^{(2)}(t) = \sum_{j=1}^{p} \beta_{jr}^{(2)} r_{jK}^{(2)}(t) + \varepsilon_r(t)$. This design is similar to Hu et Si (2013). Since EMD is a complete decomposition (*i.e* there is not any loss of information in the EMD process), a prediction of $Y(t)$ can be obtained with EMD-R2 by summing the predictions of the models in (19), *i.e.* $\hat{Y}(t) = \sum_{k=1}^{K} \hat{C}_{Yk}^{(2)}(t) + \hat{r}_{YK}^{(2)}(t)$. Note that the of the MEMD allows the number of IMFs K to be constant for all the covariates in both designs (18) and (19), but that the predictor's IMFs C_{jk} can be slightly different in EMD-R1 and EMD-R2, depending whether the response $Y(t)$ is one of the decomposed variables or not.

In many applications with time-related data, there may be a lag between an exposure and its response. Thus, in both EMD-R1 and EMD-R2, the predictors IMFs C_{jk} may have to be lagged before inclusion in the regression model. The optimal lag should be chosen by maximizing the (absolute) cross-correlation function (CCF) between C_{jk} and the appropriate

response (Shumway et Stoffer, 2000). Note that we constrain the lag to be lower than the mean period (defined as the mean interval between two maxima, Wu et Huang, 2004) of the corresponding IMF C_{kj}.

It is expected that the variance (or energy) of the IMFs decreases with the frequency (Wu et Huang, 2004). Thus, it may not be convenient to compare directly the β_{jk} estimates. A more meaningful quantity to interpret the results is what we hereafter call the sensitivity S_{jk}. The sensitivity is the estimated coefficient $\hat{\beta}_{jk}$ standardized by the mean peak-to-peak amplitude A_{jk} of C_{jk}, *i.e.*

$$S_{jk} = \hat{\beta}_{jk} \times A_{jk} \tag{20}$$

The mean peak-to-peak amplitude A_{jk} is used instead of the more traditional standard deviation because of the oscillating nature of C_{jk}. Hence, S_{jk} actually indicates the difference C_{jk} makes in the response when going from a minimum to a maximum. This quantity is useful to immediately see which oscillating modes mostly influence the response.

3. Application to weather-related cardiovascular mortality

Among the potentially harmful consequences of climate change, is expected an increase of weather-related mortality. Cardiovascular diseases (CVD) are among the most affected diseases since it has already been observed an impact of extreme weather (e.g. Braga *et al.*, 2002; Bustinza *et al.*, 2013). Indeed, CVD are already the main cause of mortality in Canada and could represent an increasing burden on the Canadian public health system in future years (Wielgosz *et al.*, 2009). Therefore, it is important to well understand the impact of weather on CVD, in order to organize more efficiently adaptation strategies to reduce climate change harmful effects. To

help achieve this purpose, EMD-R is hereby applied to the issue of weather-related cardiovascular mortality in the city of Montreal (Canada). The first sub-section introduces the data and the second presents the results.

3.1. Data

The data used are from the Greater Montreal area (geographical location shown on Figure 2.2). This area is the densest population basin of the province of Quebec, allowing enough CVD death cases in a relatively small area. This allows the model to be relevant since the weather can be considered homogeneous inside this small and flat area.

In the present application, the response series (Y) is the mortality (M) series, *i.e.* the daily number of CVD deaths from 1981 to 2011 included (a total of $n = 11322$ days). CVD includes ischaemic heart diseases (I20-I25 in the tenth version of the international classification of diseases, ICD-10), heart failure (I50 in the ICD-10), cerebrovascular disease and Transient cerebral ischaemic attacks (G45, H34.0, H34.1, I60, I61, I63 and I64 in the ICD-10).

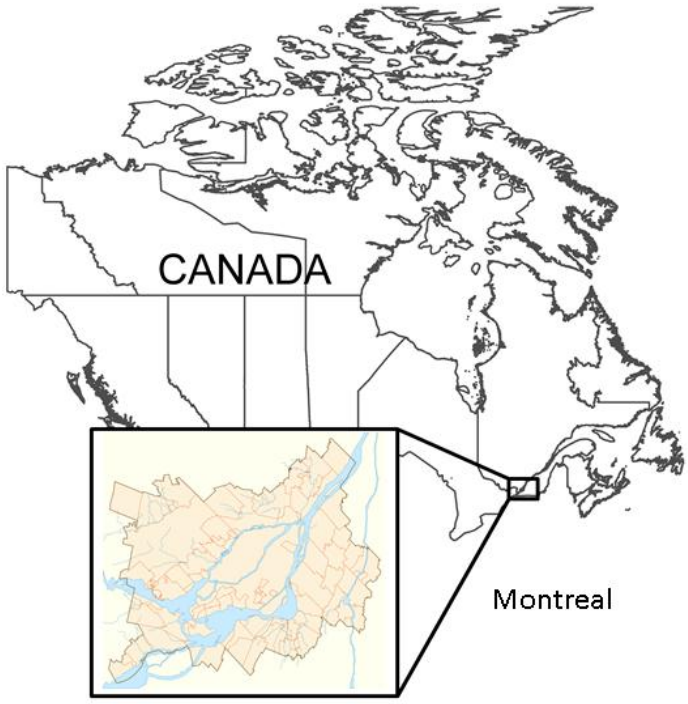


Figure 2.2: Map and location of the study region, *i.e.* the greater Montreal in the province of Quebec, Canada.

The predictor series (X_j s) are daily weather variables in the same days as the CVD deaths data. To illustrate the fact that the EMD-R methodology can be applied when several predictors are considered, temperatures (*T*) and relative humidity (*H*) are chosen in the present application. Temperatures represent the most studied predictor in environmental epidemiology (Gasparrini *et al.*, 2015) and humidity is also sometimes considered as a predictor (Schwartz *et al.*, 2004) or as a confounder (Yang *et al.*, 2015; Phung *et al.*, 2016). Weather data series are measured from many stations spread over the Greater Montreal area. However, in order to have only one time series for each predictor, the spatial mean is computed on the values of all the stations. Note that, in a previous study, weighted kriging was considered instead of spatial mean, but it has been concluded that it did not improve the models (Giroux *et al.*, 2013).

3.2. Results

The present section shows the results obtained by applying the EMD-R models described in section 2.2 on the data introduced in section 3.1. For interpretation purposes, in the following an IMF is denoted by $C_{X\,f}^{(m)}$ for a given variable X (*i.e.* *M*, *T* or *H* respectively for mortality, temperature and humidity), *f* is its mean periodicity in days and m represents the model ($m=1$ means that we are in the EMD-R1 context and $m=2$ represents the EMD-R2). The mean periodicity *f*, computed as the mean difference between two successive peaks in the IMF (Wu et Huang, 2004), is more useful than the order *j* for interpretation (although the two are related since the order *j* increases with the periodicity *f*). For instance, $C_{T365}^{(1)}$ is the temperature IMF representing the annual cycle (*i.e.* a periodicity of 365 days) used in the EMD-R1 model. This change in notation also impacts the quantities associated to the IMFs such as the amplitudes, regression coefficients and sensitivities now respectively denoted $A_{X\,f}^{(m)}$, $\beta_{X\,f}^{(m)}$ and $S_{X\,f}^{(m)}$.

For the present application, the parameters of the NA-MEMD are set according to the advices of Rehman *et al.* (2013). The stopping criterion for the sifting process is the one of Rilling *et al.* (2003) and two white noise variables with a variance equal to 10% the variance of the data are added to perform NA-MEMD. The number of projections for computing the multidimensional envelopes is set to 128.

3.2.1. Interpreting the results

To give an example of EMD decomposition, Figure 2.3 shows the temperature IMFs $C_{T\,f}^{(1)}$ obtained to use as predictors in the EMD-R1 model. It can be seen that the frequency of each IMF is indeed regular but that the amplitude may vary inside one IMF. Figure 2.3 confirms that EMD results in components that can be interpreted, since the extremely regular annual cycle as

well as the increasing trend induced by climate change, are well represented. The IMFs shown in Figure 2.3, along with the humidity IMFs $C_{H\,f}^{(1)}$ are then the predictors of the EMD-R1 model.

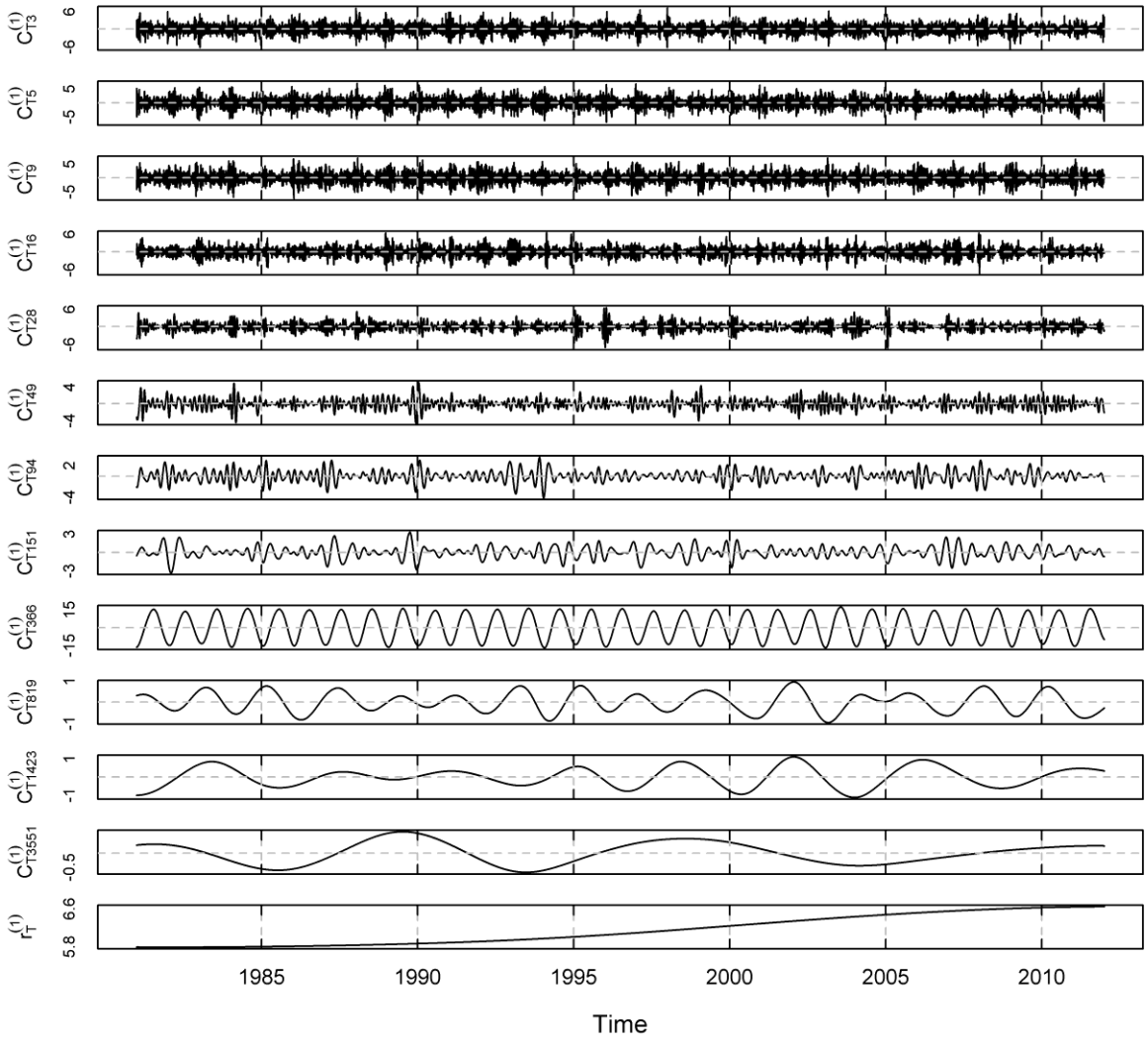
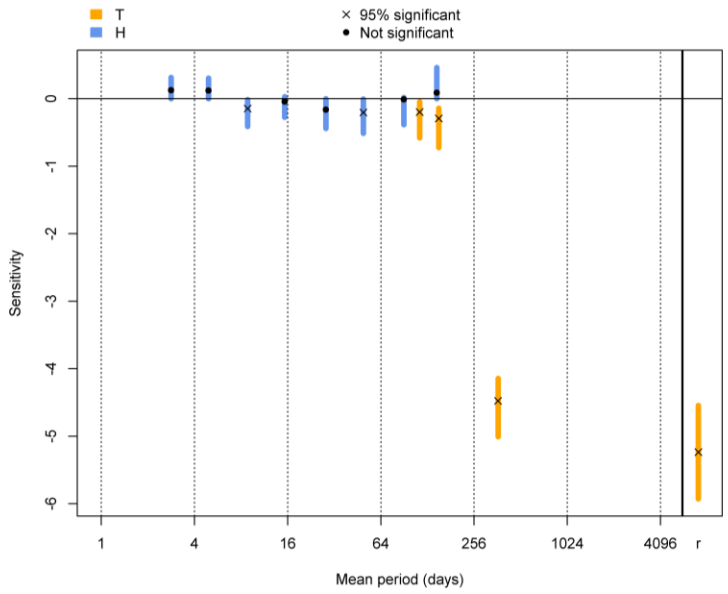


Figure 2.3: Temperatures IMFs $C_{Tf}^{(2)}$ by the MEMD on the trivariate signal where the variables are cardiovascular mortality, temperatures and humidity, for the EMD-R2 model.

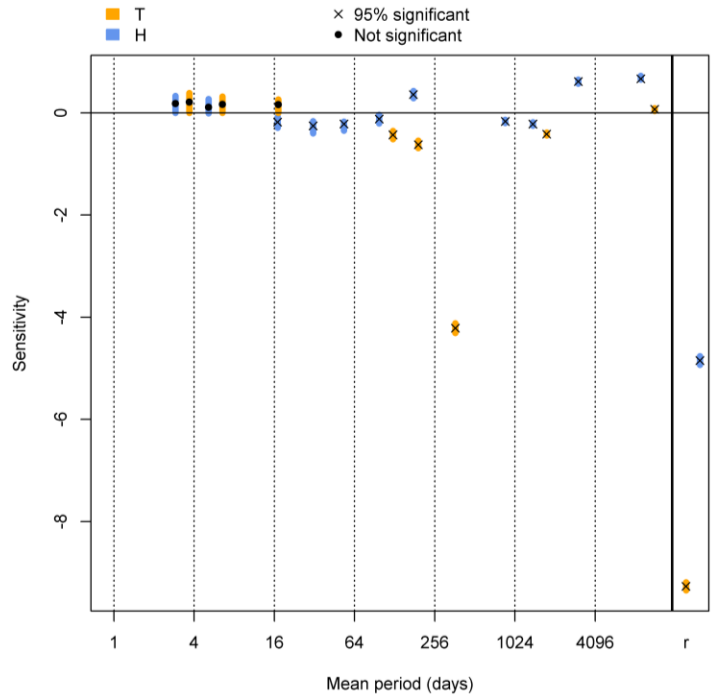
The sensitivities $S_{X\,f}^{(m)}$ estimated by the EMD-R1 and EMD-R2 models (18) and (19) are shown in Figure 2.4 along with 95% confidence intervals (CI) computed through 500 moving-

block bootstrap replications. Note that, the "one standard error rule" in cross-validation for choosing the parameter λ in (17) has been used in the case of EMD-R1 (e.g. Krstajic *et al.*, 2014). This rule takes account of the uncertainty of the cross-validation and allows obtaining the sparsest model possible since the goal of the EMD-R1 is to depict an overall image of the relationship.

A comparison of the results of EMD-R1 (Figure 2.4a) and EMD-R2 (Figure 2.4b) shows that both results are similar, only with greater details for EMD-R2. Focusing on Figure 2.4a indicates that there is a clear separation between humidity (in blue), which affects the mortality at short time scales, and temperatures (orange) which affects the mortality at large time scales. At the shortest time scales, humidity presents positive sensitivities, but note that in this case the CIs reach the zero line, meaning that we are not confident about this effect. However, at slightly larger time scales (i.e. $C_{H9}^{(1)}$ and $C_{H49}^{(1)}$), there are sensitivities between -0.15 and -0.20 with CIs not containing the zero value. Cumulating them, this means that there is an extra death every 3 days during the dryer periods compared to the humid ones. Figure 2.5 shows the mean amplitude of $C_{H9}^{(1)}$ and $C_{H49}^{(1)}$ on one year (the top two panels) and indicates that their amplitude during the months of March to May is twice the amplitude of the remaining of the year. Therefore, the dry effect on mortality is more important during spring season.



a) *EMD-R1*



b) *EMD-R2*

Figure 2.4: Non-null sensitivities obtained by EMD-R1 (panel a) and EMD-R2 (panel b) according to the mean period of the associated IMF. The points are the Lasso estimates on the whole dataset while the segments indicate 95% confidence intervals computed using 500 block bootstrap replications. "x"s represent the estimations for which the confidence interval does not reach the zero line, which can be seen as the "significant" coefficients. The x axis is in binary logarithmic scale because of the dyadic nature of EMD.

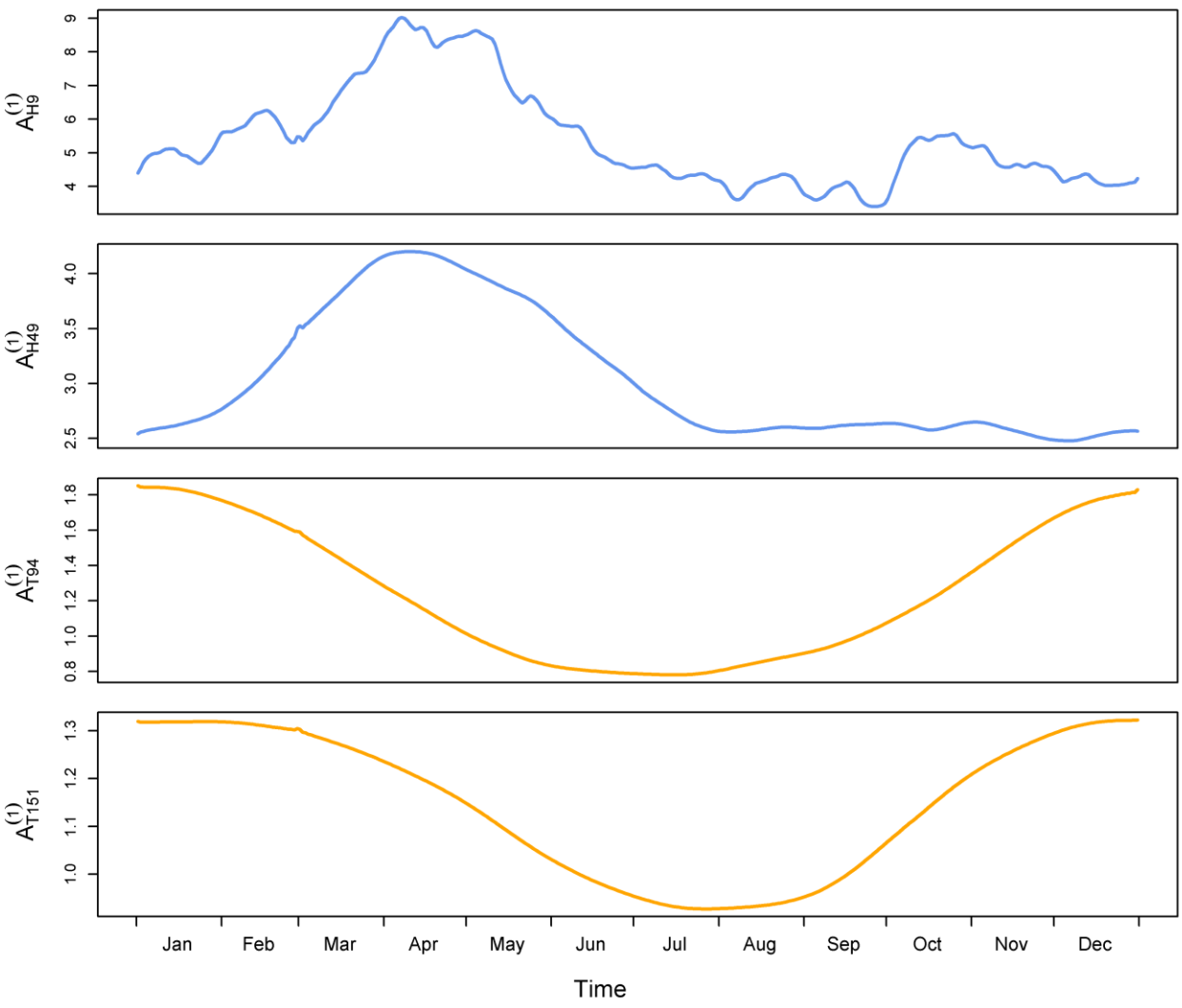


Figure 2.5: Mean amplitude on one year for several significant IMFs. The amplitudes are computed by Hilbert-Huang transform (Huang *et al.*, 1998).

According to Figure 2.4, temperatures have an effect on mortality only at periodicities greater than 3 months. There is especially the obvious annual effect with $S^{(1)}_{T365} = -4.47$ meaning that there is an increase of more than 4 deaths a day when $C^{(1)}_{T365}$ at its minimum (*i.e* winter) relatively than when it is at its maximum (*i.e.* summer). Beside the annual effect, there are also slight negative effects of $C^{(1)}_{T94}$ and $C^{(1)}_{T151}$. Figure 2.5 (two bottom panels) indicates that these

IMFs have their highest amplitudes during winter which means that they aggravate the already strong winter effect. Finally, the strongest effect reported by Figure 2.4 is the effect of the increasing temperature trend (*i.e.* the global warming), which is associated to the decreasing trend cardiovascular mortality. Therefore, it would seem that the global warming tends to mitigate the winter effect.

Besides the results of EMD-R1, EMD-R2 gives more insights on the relationship. Figure 2.4b shows positive sensitivities for the shortest periodicities temperatures IMFs ($C_{T3}^{(2)}, C_{T5}^{(2)}$ and $C_{T17}^{(2)}$), but with CI reaching the zero line meaning that we are not too confident about this effect. The biggest difference between EMD-R1 and EMD-R2 results, is that the latter indicates effects at very large time scales. There are especially high sensitivities for humidity at 8-years and 25-years scales ($S_{H3059}^{(2)} = 0.35$ and $S_{H9017}^{(2)} = 0.66$). Finally, Figure 2.4b also indicates that the humidity trend $r_H^{(2)}$ joins itself to the temperatures trend $r_T^{(2)}$ to explain the trend of mortality.

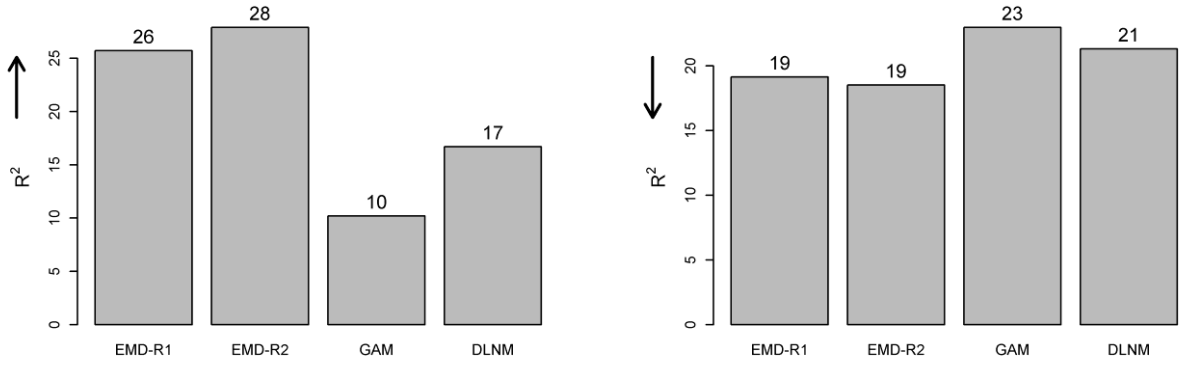
This application has been purposely kept simple to focus on the methodology itself. However, note that in more complete epidemiologic studies, some variables such as age and gender must be controlled. Models for different age classes and for both genders have been performed in the present case study. Nevertheless, no real differences have been noticed from the general results presented above in this section. In addition, models dealing with separating winter and summer have also been performed since the effect of weather can change according to the season. The obtained results do not provide any added value to the present paper. Hence, they were not presented here to limit the complexity of the application. Further models and results on this topic are presented in the extensive application performed in the technical report of Masselot *et al.* (2015).

3.2.2. Performance assessment and comparison

The main difference between EMD-R1 and EMD-R2 is that the latter retains more weather IMFs as predictors and hence, more information than the EMD-R1. However, EMD-R1 is easier to interpret because it is more parsimonious. The two models can be more objectively compared between them and with two commonly used models in environmental epidemiology: generalized additive models (GAM, Hastie et Tibshirani, 1986) and distributed lag nonlinear models (DLNM, Gasparrini *et al.*, 2010). GAM are commonly used in environmental epidemiology for having brought to attention the nonlinear J, U or V-shapes relationships between mortality and temperatures (e.g. Braga *et al.*, 2001b; Bayentin *et al.*, 2010). DLNMs are also nonlinear models allowing in addition the use of several lags for each predictor, in order to model more precisely the induction time between an exposure and the response (e.g. Li *et al.*, 2013; Vanos *et al.*, 2014). The comparison between the models is performed through the R^2 and generalized cross-validation (GCV, Craven et Wahba, 1978) criteria. They respectively represent the explicative and the predictive power of the models.

The values of the R^2 and GCV criteria for each model are shown in Figure 2.6. One can see that the EMD-R1 and EMD-R2 models display the best R^2 values with 26% and 28% respectively, while GAM and DLNM have R^2 values of 10% and 16% respectively (Figure 2.6a). The GAM and DLNM R^2 values are consistent with the values usually found in the literature. The higher score for EMD-R2 shows that this model is more accurate than EMD-R1, since its details allow explaining a larger proportion of the response's variance. Note that these R^2 scores are particularly high knowing that the weather is actually one of many factors (and not the main) affecting the CVD mortality. Other factors include, for instance, physical exercise, obesity, dietary habits and smoking (Institut national de santé publique du Québec, 2006).

The GCV scores (Figure 2.6b) lead to the conclusion that EMD-R models have lower prediction error than GAM and DLNM (although not large differences since GAM and DLNM have scores of 23 and 21 versus scores of 19 for both EMD-R1 and EMD-R2).



a) R^2 *b) GCV criterion*

Figure 2.6: Comparison of the performance criteria R^2 and GCV of the two EMD-R models as well as GAM and DLNM applied on the same data.

4. Discussion

The results on the weather-related cardiovascular mortality issue, described in section 3.2, already shows one advantage of the EMD-R: it is able to display some hidden aspects of the relationship. In this case, the effect of humidity found during spring season and at very large time scales (*i.e.* periodicities of several years) is quite new in the field of environmental epidemiology. Indeed, similar studies often concentrate on temperatures (e.g. Patz *et al.*, 2014). In section 3.2.1, the effect of temperature is mainly an effect of winter cold. The readers used to the field of environmental epidemiology could be surprised about the non-significance of short periodicity temperature IMFs, since the effect of heat waves is the most documented one in the literature (e.g. Chebana *et al.*, 2012b; Bustinza *et al.*, 2013). However, heat waves are extreme events, not

necessarily well represented by single IMFs, and need particular statistical methods to be studied. The present finding that the main effect is the constant cold is actually consistent with the global study of (Gasparrini *et al.*, 2015).

The performance comparison of section 3.2.2 shows that EMD-R offers an improvement over reference methods in the environmental epidemiology, both for explanation and prediction. This might seem odd since GAM and DLNM are both nonlinear while the EMD-R is only linear, and that the relationship between temperatures and mortality is well known for being nonlinear. This is illustrated by Figure 2.7 which shows the J-shaped GAM function obtained for temperatures on the cardiovascular mortality for the data described in section 3.1. However, note that this function is piecewise linear and that all the pieces can be found in Figure 2.4b. Indeed, the coefficient $\hat{\beta}^{(2)}_{T365} = -0.14$ corresponds to the main piece in box 1 which has a slope of approximately -0.15. Moreover, the positive coefficients $\hat{\beta}^{(2)}_{T3}$, $\hat{\beta}^{(2)}_{T5}$ and $\hat{\beta}^{(2)}_{T17}$ (although not significant) correspond to the increasing piece of box 2 and the negative coefficients $\hat{\beta}^{(2)}_{T94}$ and $\hat{\beta}^{(2)}_{T151}$ correspond to the slightly decreasing piece of box 3. The usefulness of nonlinear models is that they summarize several effects in one curve but EMD-R is able to provide details about the different parts of a nonlinear relationship.

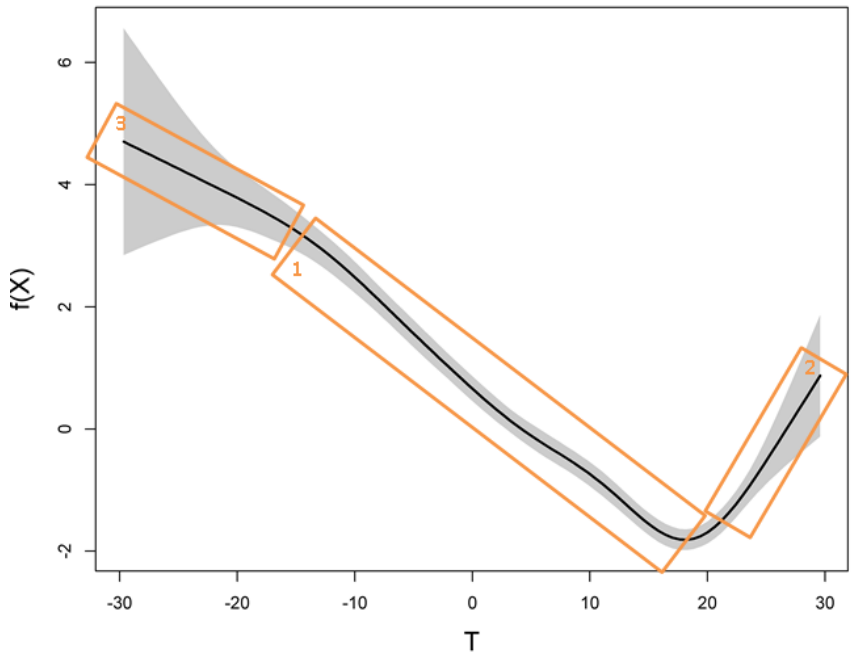


Figure 2.7: GAM function depicting the whole relationship between temperatures and cardiovascular mortality. Boxes indicate the linear pieces of the function.

5. Conclusion

The present paper introduces a methodology for EMD-regression when dealing with time series data (and more generally all data with autocorrelation). The purpose of the EMD-R is to understand a relationship between variables from a different point a view, *i.e.* from a time scale point of view. This point of view acknowledges the complexity of many real-world time series which contains important information quantity through their variations. With EMD-R, it is possible to assess the influence of all of these time scales on the response of interest, while the classical regression will only depict the relationship at the dominant time scale. The main interest of EMD-R is to interpret and communicate the results in terms of time scales. Moreover, using the Lasso to perform the regression analysis in both EMD-R1 and EMD-R2 allows using several

predictors in the analysis. Besides the core of the EMD-R methodology (*i.e.* EMD and the Lasso), the present paper also proposes a number of tools helping in the interpretation, illustrated in the application results. These tools are the sensitivity (20) with an associated plot (illustrated in Figure 2.4) and the amplitude plot (illustrated in Figure 2.5).

As a perspective, the motivating application of weather-related cardiovascular mortality appeals for a more complete analysis controlling for some variables, *e.g.* atmospheric pollutants. Note that all the tools used in the field of environmental epidemiology, such as the computation of relative risks and the use of a time variable to control for unmeasured confounders, can be used in the context of EMD-R. The present paper intended to introduce the EMD-R in its basic form but for a general context to show its applicability and benefits in fields where regression is used and data are time dependent.

Despite its usefulness, EMD-R is not intended to be an alternative to the most used regression methods but to complete the existing methods. Indeed, EMD-R results are difficult to interpret "at a glance" like, *e.g.* GAM functions. Moreover, the methodology is intended to study the relationship at different time scales through regular pattern. The application of section 3 suggests that EMD-R is not able to detect the effect of extreme events, which are not necessarily regular. Hence, depending on the goal, it may be necessary to complete this analysis with extreme statistics. Another methodological limitation lies in the choice of lags. Indeed, the present methodology choose only one lag per IMF, but in many cases like in environmental epidemiology, the lags are distributed, meaning that the effect of an exposure is not only on a single day (Schwartz, 2000a). Therefore, a perspective would be to develop a statistical method able to estimate distributed lags as well as perform a variable selection, for instance a mix between the group Lasso (Yuan et Lin, 2006b) and the lag weighted Lasso (Park et Sakaori,

2013). Finally, although this aspect has not been brought in the paper, note that existing tools allow forecasting future oscillations of the IMFs (e.g. Kurbatskii *et al.*, 2011; Lee et Ouarda, 2011). Therefore another perspective is to use or adapt these tools to forecast the predictors IMF in order to provide a forecast of the response.

Acknowledgements

The authors are thankful to the Fonds Vert du Québec for funding this study and to the Institut national de santé publique du Québec for data access. The authors also wish to thank Jean-Xavier Giroux (INRS-ETE) for his help on the database establishing as well as Yohann Chiu (INRS-ETE) for all his relevant comments during the project. All the analyses of the present paper have been performed with the open source R software (R Core Team, 2015). The R functions developed by the authors for the study are freely available upon request.

Article 3 :

A new look at weather-related health through functional regression

-

Un nouveau regard sur l'effet de la météorologie sur la santé par régression fonctionnelle

Pierre Masselot[1*], Fateh Chebana[1], Diane Bélanger[1,2], André St-Hilaire[1], Belkacem Abdous[3], Pierre Gosselin[1,2,4], Taha B.M.J. Ouarda[1]

[1]*Institut National de la Recherche Scientifique, Centre Eau-Terre-Environnement, Québec, Canada;*

[2]*Centre Hospitalier Universitaire de Québec, Centre de Recherche, Québec, Canada;*

[3]*Université Laval, Département de médecine sociale et préventive, Québec, Canada;*

[4]*Institut national de santé publique du Québec (INSPQ), Québec, Canada.*

Soumis

Cet article a dû être retiré de la version électronique en raison de restrictions liées au droit d'auteur.

BIBLIOGRAPHIE

Aitken, A. C. (1935a). On least squares and linear combination of observations. Proc. Roy. Soc. Edin. A.

Aitken, A. C. (1935b). "On least squares and linear combination of observations." Proceedings of the Royal Society of Edinburgh **55**: 42-48.

Akaike, H. (1974). "A new look at the statistical model identification." Automatic Control, IEEE Transactions on **19**(6): 716-723.

Alkhamisi, M. A. (2010). "Ridge Estimation in Linear Models with Autocorrelated Errors." Communications in Statistics - Theory and Methods **39**(14): 2630-2644.

Almon, S. (1965). "The Distributed Lag Between Capital Appropriations and Expenditures." Econometrica **33**(1): 178-196.

Anderson, B. G. et M. L. Bell (2009). "Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States." Epidemiology (Cambridge, Mass.) **20**(2): 205.

Antman, E., J.-P. Bassand, W. Klein, M. Ohman, J. L. Lopez Sendon, L. Rydén, M. Simoons et M. Tendera (2000). "Myocardial infarction redefined—a consensus document of The Joint European Society of Cardiology/American College of Cardiology committee for the redefinition of myocardial infarctionThe Joint European Society of Cardiology/ American College of Cardiology Committee∗." Journal of the American College of Cardiology **36**(3): 959-969.

Antoniadis, A., J. Bigot et T. Sapatinas (2001). "Wavelet estimators in nonparametric regression: a comparative simulation study." Journal of Statistical Software **6**(6): 1-83.

Arisido, M. W. (2016). "Functional measure of ozone exposure to model short-term health effects." Environmetrics **27**(5): 306-317.

Armstrong, B. (2006). "Models for the relationship between ambient temperature and daily mortality." Epidemiology **17**(6): 624-631 610.1097/1001.ede.0000239732.0000250999.0000239738f.

Barreca, A. I. et J. P. Shimshack (2012). "Absolute Humidity, Temperature, and Influenza Mortality: 30 Years of County-Level Evidence from the United States." American Journal of Epidemiology **176**(suppl 7): S114-S122.

Bassil, K. L., D. C. Cole, R. Moineddin, A. M. Craig, W. Y. Wendy Lou, B. Schwartz et E. Rea (2009). "Temporal and spatial variation of heat-related illness using 911 medical dispatch data." Environmental Research **109**(5): 600-606.

Bayentin, L., S. El Adlouni, T. Ouarda, P. Gosselin, B. Doyon et F. Chebana (2010). "Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada." International Journal of Health Geographics.

Bel, L., A. Bar-Hen, R. Petit et R. Cheddadi (2011). "Spatio-temporal functional regression on paleoecological data." Journal of Applied Statistics **38**(4): 695-704.

Bergmeir, C. et J. M. Benítez (2012). "On the use of cross-validation for time series predictor evaluation." Information Sciences **191**(0): 192-213.

Besse, P. C. et H. Cardot (1996). "Approximation spline de la prevision d'un processus fonctionnel autorégressif d'ordre 1." Canadian Journal of Statistics **24**(4): 467-487.

Billingsley, P. (1995). Probability and Measure.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity." Journal of Econometrics **31**(3): 307-327.

Boudraa, A. O. et J. C. Cexus (2007). "EMD-Based Signal Filtering." Instrumentation and Measurement, IEEE Transactions on **56**(6): 2196-2202.

Box, G. E. P. et G. M. Jenkins (1976). Time series analysis: forecasting and control. San Francisco, Calif.

Braga, A. L. F., A. Zanobetti et J. Schwartz (2001a). "The Lag Structure Between Particulate Air Pollution and Respiratory and Cardiovascular Deaths in 10 US Cities." Journal of Occupational and Environmental Medicine **43**(11): 927-933.

Braga, A. L. F., A. Zanobetti et J. Schwartz (2001b). "The Time Course of Weather-Related Deaths." Epidemiology **12**(6): 662-667.

Braga, A. L. F., A. Zanobetti et J. Schwartz (2002). "The effect of weather on respiratory and cardiovascular deaths in 12 U.S. cities." Environmental health perspectives.

Brewer, M. J., A. Butler et S. L. Cooksley (2016). "The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity." Methods in Ecology and Evolution **7**(6): 679-692.

Brockhaus, S., M. Melcher, F. Leisch et S. Greven (2016). "Boosting flexible functional regression models with a high number of functional historical effects." Statistics and Computing: 1-14.

Brockhaus, S. et D. Ruegamer (2016). FDboost: Boosting Functional Regression Models.

Brockhaus, S., F. Scheipl, T. Hothorn et S. Greven (2015). "The functional linear array model." Statistical Modelling **15**(3): 279-300.

Brumback, B. A. et J. A. Rice (1998). "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves." Journal of the American Statistical Association **93**(443): 961-976.

Bühlmann, P. et T. Hothorn (2007). "Boosting Algorithms: Regularization, Prediction and Model Fitting." Statistical Science **22**(4): 477-505.

Burnham, K. P. et D. R. Anderson (2004). "Multimodel Inference: Understanding AIC and BIC in Model Selection." Sociological Methods & Research **33**(2): 261-304.

Burr, W. S., G. Takahara et H. H. Shin (2015). "Bias correction in estimation of public health risk attributable to short-term air pollution exposure." Environmetrics **26**(4): 298-311.

Bustinza, R., G. Lebel, P. Gosselin, D. Belanger et F. Chebana (2013). "Health impacts of the July 2010 heat wave in Quebec, Canada." BMC Public Health **13**(1): 56.

Cardot, H., F. Ferraty et P. Sarda (1999). "Functional linear model." Statistics & Probability Letters **45**(1): 11-22.

Cardot, H., F. Ferraty et P. Sarda (2003). "Spline estimators for the functional linear model." Statistica Sinica **13**(3): 571-592.

Chatterjee, A. et S. N. Lahiri (2011). "Bootstrapping Lasso Estimators." Journal of the American Statistical Association **106**(494): 608-625.

Chebana, F., S. Dabo-Niang et T. B. M. J. Ouarda (2012a). "Exploratory functional flood frequency analysis and outlier detection." Water Resources Research **48**(4): W04514.

Chebana, F., B. Martel, P. Gosselin, J.-X. Giroux et T. B. Ouarda (2012b). "A general and flexible methodology to define thresholds for heat health watch and warning systems, applied to the province of Québec (Canada)." International journal of biometeorology **57**(4): 631-644.

Chiu, Y., F. Chebana, B. Abdous, D. Bélanger et P. Gosselin (2016). "Mortality and morbidity peaks modeling: An extreme value theory approach." Statistical Methods in Medical Research: 0962280216662494.

Choudhury, A. H., R. Hubata et R. D. St. Louis (1999). "Understanding time-series regression estimators." American Statistician **53**(4): 342-348.

Chuanrui, F. (2010). Forecasting Exchange Rate with EMD-Based Support Vector Regression. Management and Service Science (MASS), 2010 International Conference on.

Cleveland, R. B., W. S. Cleveland, J. E. McRae et I. Terpenning (1990). "STL: A seasonal-trend decomposition procedure based on loess." Journal of Official Statistics **6**(1): 3-73.

Cleveland, W. S. et S. J. Devlin (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." Journal of the American Statistical Association **83**(403): 596-610.

Cleveland, W. S., S. J. Devlin et E. Grosse (1988). "Regression by local fitting." Journal of Econometrics **37**(1): 87-114.

Cochrane, D. et G. H. Orcutt (1949). "Application of Least Squares Regression to Relationships Containing Auto- Correlated Error Terms." Journal of the American Statistical Association **44**(245): 32-61.

Colominas, M. A., G. Schlotthauer, M. E. Torres et P. Flandrin (2013). "Noise-assisted EMD methods in action." Advances in Adaptive Data Analysis.

Cooley, J. W., P. A. W. Lewis et P. D. Welch (1969). "The Fast Fourier Transform and Its Applications." IEEE Transactions on Education **12**(1): 27-34.

Coughlin, K. T. et K. K. Tung (2004). "11-Year solar cycle in the stratosphere extracted by the empirical mode decomposition method." Advances in Space Research **34**(2): 323-329.

Craven, P. et G. Wahba (1978). "Smoothing noisy data with spline functions." Numerische Mathematik **31**(4): 377-403.

Cristobal, J. A. C., P. F. Roca et W. G. Manteiga (1987). "A Class of Linear Regression Parameter Estimators Constructed by Nonparametric Estimation." (2): 603-609.

Cryer, J. D. et K.-S. Chan (2008). Time series analysis: with applications in R, Springerverlag New York.

Dabo-Niang, S. et F. Ferraty (2008). Functional and operatorial statistics, Springer.

Dabo-Niang, S. et A.-F. Yao (2007). "Kernel regression estimation for continuous spatial processes." Mathematical Methods of Statistics **16**(4): 298-317.

Daubechies, I. (1992). Ten lectures on wavelets, SIAM.

Davis, R. E., E. Dougherty, C. McArthur, Q. S. Huang et M. G. Baker (2016). "Cold, dry air is associated with influenza and pneumonia mortality in Auckland, New Zealand." Influenza and other respiratory viruses **10**(4): 310-313.

Dickey, D. A. et W. A. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series With a Unit Root." Journal of the American Statistical Association **74**(366): 427-431.

Dominici, F., A. McDermott, S. L. Zeger et J. M. Samet (2002). "On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health." American Journal of Epidemiology **156**(3): 193-203.

Dominici, F., A. McDermott, S. L. Zeger et J. M. Samet (2003). "Airborne Particulate Matter and Mortality: Timescale Effects in Four US Cities." American Journal of Epidemiology **157**(12): 1055-1065.

Donoho, D. L. et J. M. Johnstone (1994). "Ideal spatial adaptation by wavelet shrinkage." Biometrika **81**(3): 425-455.

Doyon, B., D. Bélanger et P. Gosselin (2006). "Effets du climat sur la mortalité au Québec méridional de 1981 à 1999 et simulations pour des scénarios climatiques futurs." Institut national de santé publique du Québec.

Doyon, B., D. Bélanger et P. Gosselin (2008). "The potential impact of climate change on annual and seasonal mortality for three cities in Québec, Canada." International Journal of Health Geographics **7**: 23.

Dukić, V., M. Hayden, A. Forgor, T. Hopson, P. Akweongo, A. Hodgson, A. Monaghan, C. Wiedinmyer, T. Yoksas, M. Thomson, S. Trzaska et R. Pandya (2012). "The Role of Weather in Meningitis Outbreaks in Navrongo, Ghana: A Generalized Additive Modeling Approach." Journal of Agricultural, Biological, and Environmental Statistics **17**(3): 442-460.

Durocher, M., T. S. Lee, T. B. M. J. Ouarda et F. Chebana (2015). "Hybrid signal detection approach for hydro-meteorological variables combining EMD and cross-wavelet analysis." International Journal of Climatology: n/a-n/a.

Engle, R. F. et C. W. J. Granger (1987). "Co-Integration and Error Correction: Representation, Estimation, and Testing." Econometrica **55**(2): 251-276.

Epanechnikov, V. A. (1969). "Non-Parametric Estimation of a Multivariate Probability Density." Theory of Probability & Its Applications **14**(1): 153-158.

Ferraty, F. et P. Vieu (2002). "The functional nonparametric model and application to spectrometric data." Computational Statistics **17**(4): 545-564.

Ferraty, F. et P. Vieu (2004). "Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination." Nonparametric Statistics **16**(1-2): 111-125.

Flandrin, P., P. Goncalves et G. G. Rilling (2004a). Detrending and denoising with empirical mode decompositions**:** XXXV-2310 p.

Flandrin, P., G. Rilling et P. Goncalves (2004b). "Empirical mode decomposition as a filter bank." Signal Processing Letters, IEEE **11**(2): 112-114.

Friedman, J., T. Hastie et R. Tibshirani (2009). The elements of statistical learning.

Friedman, J., T. Hastie et R. Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software **33**(1): 1-22.

Gallant, A. R. et J. J. Goebel (1976). "Nonlinear regression with autocorrelated errors." Journal of the American Statistical Association **71**(356): 961-967.

Gardner, E. S. (1985). "Exponential smoothing: The state of the art." Journal of Forecasting **4**(1): 1-28.

Gasparrini, A. et B. Armstrong (2011). "The impact of heat waves on mortality." Epidemiology (Cambridge, Mass.) **22**(1): 68.

Gasparrini, A. et B. Armstrong (2013). "Reducing and meta-analysing estimates from distributed lag non-linear models." BMC Medical Research Methodology **13**(1): 1.

Gasparrini, A., B. Armstrong et M. G. Kenward (2010). "Distributed lag non-linear models." Statistics in Medicine **29**(21): 2224-2234.

Gasparrini, A., Y. Guo, M. Hashizume, E. Lavigne, A. Zanobetti, J. Schwartz, A. Tobias, S. Tong, J. Rocklöv, B. Forsberg, M. Leone, M. De Sario, M. L. Bell, Y.-L. L. Guo, C.-f. Wu, H. Kan, S.-M. Yi, M. de Sousa Zanotti Stagliorio Coelho, P. H. N. Saldiva, Y. Honda, H. Kim et B. Armstrong (2015). "Mortality risk attributable to high and low ambient temperature: a multicountry observational study." The Lancet **386**(9991): 369-375.

Gelles, G. M. et D. W. Mitchell (1989). "An approximation theorem for the polynomial inverse lag." Economics Letters **30**(2): 129-132.

Ghouse, B., T. B. M. J. Ouarda et P. R. Marpu (2015). "Long-term projections of temperature, precipitation and soil moisture using non-stationary oscillation processes over the UAE region." International Journal of Climatology: n/a-n/a.

Giroux, J.-X., F. Chebana, D. Bélanger, E. Gloaguen, T. B. M. J. Ouarda et S.-H. A. (2013). Projet M1 : Comparaison de l'utilisation des moyennes spatiales à celle du krigeage, appliquée à la relation mortalité par MCV - météorologie, au Québec, de 1996 à 2007., INRS-ETE.

Glasbey, C. A. (1980). "Nonlinear Regression with Autoregressive Time Series Errors." Biometrics **36**(1): 135-139.

Goldsmith, J., J. Bobb, C. M. Crainiceanu, B. Caffo et D. Reich (2011). "Penalized Functional Regression." Journal of Computational and Graphical Statistics **20**(4): 830-851.

Gonzalez-Manteiga, W. et A. Martinez-Calvo (2011). "Bootstrap in functional linear regression." Journal of Statistical Planning and Inference **141**(1): 453-461.

Granger, C. W. J. et P. Newbold (1974). "Spurious regressions in econometrics." Journal of Econometrics **2**(2): 111-120.

Green, P. J. et B. W. Silverman (1994). Nonparametric regression and generalized linear models: a roughness penalty approach, Chapman & Hall London.

Grineski, S. E., J. M. Herrera, P. Bulathsinhala et J. G. Staniswalis (2015). "Is there a Hispanic Health Paradox in sensitivity to air pollution? Hospital admissions for asthma, chronic obstructive pulmonary disease and congestive heart failure associated with NO2 and PM2.5 in El Paso, TX, 2005–2010." Atmospheric Environment **119**: 314-321.

Haiyong, Z. et G. Qiang (2006). Research on Properties of Empirical Mode Decomposition Method. Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on.

Hamilton, J. (1994). Time-series analysis, Princeton Univerity Press.

Hastie, T. et C. Mallows (1993). "[A Statistical View of Some Chemometrics Regression Tools]: Discussion." Technometrics **35**(2): 140-143.

Hastie, T. et R. Tibshirani (1986). "Generalized Additive Models." Statistical Science **1**(3): 297-310.

Hastie, T. et R. Tibshirani (1993). "Varying-Coefficient Models." Journal of the Royal Statistical Society. Series B (Methodological) **55**(4): 757-796.

Hastie, T., R. Tibshirani et M. Wainwright (2015). Statistical learning with sparsity: the lasso and generalizations, CRC Press.

He, G., H.-G. Müller et J.-L. Wang (2003). "Functional canonical analysis for square integrable stochastic processes." Journal of Multivariate Analysis **85**(1): 54-77.

Hoerl, A. E. et R. W. Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics **12**(1): 55-67.

Hoover, K. D. (2003). "Nonstationary Time Series, Cointegration, and the Principle of the Common Cause." The British Journal for the Philosophy of Science **54**(4): 527-551.

Hosseini-Nasab, M. et Z. Mirzaei (2014). "Functional analysis of glaucoma data." Statistics in Medicine **33**(12): 2077-2102.

Houck, P. D., J. E. Lethen, M. W. Riggs, D. S. Gantt et G. J. Dehmer (2005). "Relation of Atmospheric Pressure Changes and the Occurrences of Acute Myocardial Infarction and Stroke." The American Journal of Cardiology **96**(1): 45-51.

Hu, W. et B. C. Si (2013). "Soil water prediction based on its scale-specific control using multivariate empirical mode decomposition." Geoderma **193–194**(0): 180-188.

Huang, C., A. G. Barnett, X. Wang et S. Tong (2012). "Effects of Extreme Temperatures on Years of Life Lost for Cardiovascular Deaths: A Time Series Study in Brisbane, Australia." Circulation: Cardiovascular Quality and Outcomes **5**(5): 609-614.

Huang, N. E., Z. Shen et S. R. Long (1999). "A New View of Nonlinear Water Waves : The Hilbert Spectrum1." Annual Review of Fluid Mechanics **31**(1): 417-457.

Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung et H. H. Liu (1998). "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis." Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **454**(1971): 903-995.

Huang, N. E. et Z. Wu (2008). "A review on Hilbert-Huang transform: Method and its applications to geophysical studies." Reviews of Geophysics **46**(2): n/a-n/a.

Huang, N. E., Z. Wu, S. R. Long, K. C. Arnold, X. Chen et K. Blank (2009). "On instantaneous frequency." Advances in Adaptive Data Analysis **1**(02): 177-229.

Huang, T.-L., W.-X. Ren et M.-l. Lou (2008). The orthogonal Hilbert-Huang transform and its application in earthquake motion recordings analysis. The 14th World Conference on Earthquake Engineering. Beijing, China.

Huynen, M. M., P. Martens, D. Schram, M. P. Weijenberg et A. E. Kunst (2001). "The impact of heat waves and cold spells on mortality rates in the Dutch population." Environmental health perspectives **109**(5): 463-470.

Hyndman, R. (2015). "forecast: Forecasting functions for time series and linear models." R package version 6.2.

Hyndman, R. J. et Y. Khandakar (2007). Automatic time series for forecasting: the forecast package for R, Monash University, Department of Econometrics and Business Statistics.

Institut national de santé publique du Québec (2006). "Les maladies du coeur et les maladies vasculaires cérébrales : prévalence, morbidité et mortalité au Québec." INSPQ.

IPCC (2013). Climate Change: The Physical Science Basis. C. U. Press.

IPCC (2014). Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White (eds.)]. Cambridge, United Kingdom and New York, NY, USA, Cambridge University Press.

ISQ, I. d. l. s. d. Q. (2009). Décès selon les principaux groupes de causes, sexes réunis, Québec, 2000-2009.

Ivanescu, A., A.-M. Staicu, F. Scheipl et S. Greven (2014). "Penalized function-on-function regression." Computational Statistics: 1-30.

James, G. M. (2002). "Generalized linear models with functional predictors." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(3): 411-432.

Junsheng, C., Y. Dejie et Y. Yu (2006). "Research on the intrinsic mode function (IMF) criterion in EMD method." Mechanical Systems and Signal Processing **20**(4): 817-824.

Kan, H., C.-M. Wong, N. Vichit-Vadakan et Z. Qian (2010). "Short-term association between sulfur dioxide and daily mortality: The Public Health and Air Pollution in Asia (PAPA) study." Environmental Research **110**(3): 258-264.

Keatinge, W. R. (2002). "Winter mortality and its causes." International Journal of Circumpolar Health **61**(4).

Kelsall, J. E., S. L. Zeger et J. M. Samet (1999). "Frequency Domain Log-linear Models; Air Pollution and Mortality." Journal of the Royal Statistical Society: Series C (Applied Statistics) **48**(3): 331-344.

Khaliq, M. N., T. Ouarda, P. Gachon et L. Sushama (2011). "Stochastic modeling of hot weather spells and their characteristics." Climate Research **47**(3): 187-199.

Kim, D., K. Kim et H.-S. Oh (2012). "Extending the scope of empirical mode decomposition by smoothing." EURASIP Journal on Advances in Signal Processing **2012**(1): 1-17.

Kim, K., D. Şentürk et R. Li (2011). "Recent history functional linear models for sparse longitudinal data." Journal of Statistical Planning and Inference **141**(4): 1554-1566.

Kingman, J. F. C. (2005). Poisson Processes. Encyclopedia of Biostatistics, John Wiley & Sons, Ltd.

Kinney, P., M. Pascal, R. Vautard et K. Laaidi (2012). "La mortalité hivernale va-t-elle diminuer avec le changement climatique?" Bulletin Épidémiologique Hebdomadaire **12-13**.

Kişi, Ö. (2009). "Wavelet regression model as an alternative to neural networks for monthly streamflow forecasting." Hydrological processes **23**(25): 3583-3597.

Knowlton, K., M. Rotkin-Ellman, G. King, H. G. Margolis, D. Smith, G. Solomon, R. Trent et P. English (2009). "The 2006 California heat wave: impacts on hospitalizations and emergency department visits." Environ Health Perspect **117**(1): 61-67.

Krstajic, D., L. J. Buturovic, D. E. Leahy et S. Thomas (2014). "Cross-validation pitfalls when selecting and assessing regression and classification models." Journal of Cheminformatics **6**(1): 10.

Kucuk, M. et N. Agiralioglu (2006). "Wavelet Regression Technique for Streamflow Prediction." Journal of Applied Statistics **33**(9): 943-960.

Kurbatskii, V. G., D. N. Sidorov, V. A. Spiryaev et N. V. Tomin (2011). "On the Neural Network Approach for Forecasting of Nonstationary Time Series on the Basis of the Hilbert-Huang Transform." Automation and Remote Control **72**(7): 1405-1414.

Lahiri, S. N. (1999). "Theoretical Comparisons of Block Bootstrap Methods." The Annals of Statistics **27**(1): 386-404.

Lan Chang, C., M. Shipley, M. Marmot et N. Poulter (2004). "Lower ambient temperature was associated with an increased risk of hospitalization for stroke and acute myocardial infarction in young women." Journal of Clinical Epidemiology **57**(7): 749-757.

Lee, D. S., M. Chiu, D. G. Manuel, K. Tu, X. Wang, P. C. Austin, M. Y. Mattern, T. F. Mitiku, L. W. Svenson, W. Putnam, W. M. Flanagan, J. V. Tu et f. t. C. C. O. R. Team (2009). "Trends in risk factors for cardiovascular disease in Canada: temporal, socio-demographic and geographic factors." Canadian Medical Association Journal **181**(3-4): E55-E66.

Lee, M., F. Nordio, A. Zanobetti, P. Kinney, R. Vautard et J. Schwartz (2014). "Acclimatization across space and time in the effect of temperature on mortality: a time-series analysis." Environmental Health.

Lee, T. et T. B. M. J. Ouarda (2010). "Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes." Journal of Geophysical Research: Atmospheres **115**(D13): n/a-n/a.

Lee, T. et T. B. M. J. Ouarda (2011). "Prediction of climate nonstationary oscillation processes with empirical mode decomposition." Journal of Geophysical Research: Atmospheres **116**(D6): n/a-n/a.

Lee, T. et T. B. M. J. Ouarda (2012). "An EMD and PCA hybrid approach for separating noise from signal, and signal in climate change detection." International Journal of Climatology **32**(4): 624-634.

Lee, Y.-S., T.-H. Kim et P. Newbold (2005). "Spurious nonlinear regressions in econometrics." Economics Letters **87**(3): 301-306.

Li, T., R. M. Horton et P. L. Kinney (2013). "Projections of seasonal patterns in temperature-related deaths for Manhattan, New York." Nature Clim. Change **3**(8): 717-721.

Lin, H. C. et S. Xiraxagar (2006). "Seasonality of hip fractures and estimates of season-attributable effects: a multivariate ARIMA analysis of population-based data." Osteoporosis International **17**(6): 795-806.

Lipfert, F. W. (1993). "A critical review of studies of the association between demands for hospital services and air pollution." Environ Health Perspect.

Lipfert, F. W. et C. J. Murray (2012). "Air pollution and daily mortality: A new approach to an old problem." Atmospheric Environment **55**(0): 467-474.

Liu, C., Z. Yavar et Q. Sun (2015). "Cardiovascular response to thermoregulatory challenges." American Journal of Physiology - Heart and Circulatory Physiology **309**(11): H1793-H1812.

Lockhart, R., J. Taylor, R. J. Tibshirani et R. Tibshirani (2014). "A significance test for the Lasso." Annals of Statistics **42**(2): 413-468.

Loh, C.-H., T.-C. Wu et N. E. Huang (2001). "Application of the Empirical Mode Decomposition-Hilbert Spectrum Method to Identify Near-Fault Ground-Motion Characteristics and Structural Responses." Bulletin of the Seismological Society of America **91**(5): 1339-1357.

Luepker, R. V. (2011). "Cardiovascular disease: rise, fall, and future prospects." Annual review of public health **32**: 1-3.

Malfait, N. et J. O. Ramsay (2003). "The historical functional linear model." Canadian Journal of Statistics **31**(2): 115-128.

Mandic, D. P., N. U. Rehman, W. Zhaohua et N. E. Huang (2013). "Empirical Mode Decomposition-Based Time-Frequency Analysis of Multivariate Signals: The Power of Adaptive Data Analysis." Signal Processing Magazine, IEEE **30**(6): 74-86.

Mardia, K. V., J. T. Kent et J. M. Bibby (1979). Multivariate analysis, Academic press.

Marra, G. et S. N. Wood (2011). "Practical variable selection for generalized additive models." Computational Statistics & Data Analysis **55**(7): 2372-2387.

Martins, L. C., L. A. A. Pereira, C. A. Lin, U. P. Santos, G. Prioli, O. d. C. Luiz, P. H. N. Saldiva et A. L. F. Braga (2006). "The effects of air pollution on cardiovascular diseases: lag structures." Revista de Saúde Pública **40**: 677-683.

Masselot, P., F. Chebana, D. Bélanger, A. St-Hilaire, B. Abdous, P. Gosselin et T. B. M. J. Ouarda (2015). Régression EMD avec application à la relation entre les maladies cardiovasculaires et le climat, INRS-ETE.

Masselot, P., F. Chebana, D. Bélanger, A. St-Hilaire, B. Abdous, P. Gosselin et T. B. M. J. Ouarda (2016a). Agrégation de la réponse dans la régression – application à la relation entre les maladies cardiovasculaires et la météorologie. Québec, Canada, Institut National de la Recherche Scientifique.

Masselot, P., S. Dabo-Niang, F. Chebana et T. B. M. J. Ouarda (2016b). "Streamflow forecasting using functional regression." Journal of Hydrology **538**: 754-766.

McLean, M. W., G. Hooker, A.-M. Staicu, F. Scheipl et D. Ruppert (2014). "Functional Generalized Additive Models." Journal of Computational and Graphical Statistics **23**(1): 249-269.

Meyer, M. J., B. A. Coull, F. Versace, P. Cinciripini et J. S. Morris (2015). "Bayesian Function-on-Function Regression for Multilevel Functional Data." Biometrics **71**(3): 563-574.

Michels, P. (1992). "Asymmetric Kernel Functions in Non-Parametric Regression Analysis and Prediction." Journal of the Royal Statistical Society. Series D (The Statistician) **41**(4): 439-454.

Mitchell, D. W. et P. J. Speaker (1986). "A simple, flexible distributed lag technique." Journal of Econometrics **31**(3): 329-340.

Mizon, G. E. (1995). "A simple message for autocorrelation correctors: Don't." Journal of Econometrics **69**(1): 267-288.

Modarres, R., T. Ouarda, A. Vanasse, M. G. Orzanco et P. Gosselin (2014). "Modeling climate effects on hip fracture rate by the multivariate GARCH model in Montreal region, Canada." International journal of biometeorology **58**(5): 921-930.

Morabito, M., F. Profili, A. Crisci, P. Francesconi, G. Gensini et S. Orlandini (2012). "Heat-related mortality in the Florentine area (Italy) before and after the exceptional 2003 heat wave in Europe: an improved public health response?" International journal of biometeorology **56**(5): 801-810.

Morris, J. S. (2015). "Functional Regression." Annual Review of Statistics and its Applications **2**.

MSSS, M. d. l. S. e. d. S. s. (2011). Pour guider l'action: Portrait de santé du Québec et de ses régions : les statistiques. L. g. d. Québec.

Nadaraya, E. A. (1964). "On Estimating Regression." Theory of Probability & Its Applications **9**(1): 141-142.

Niazy, R., C. F. Beckmann, J. M. Brady et S. M. Smith (2009). "Performance evaluation of ensemble empirical mode decomposition." Advances in Adaptive Data Analysis **1**(02): 231-242.

Nitschke, M., G. R. Tucker, A. L. Hansen, S. Williams, Y. Zhang et P. Bi (2011). "Impact of two recent extreme heat episodes on morbidity and mortality in Adelaide, South Australia: a case-series analysis." Environ Health **10**(1): 42.

Ouranos (2015). Vers l'adaptation : synthèse des connaissances sur les changements climatiques au Québec. Édition 2015. Montréal, Québec : Ouranos.

Pagan, A. R. et D. F. Nicholls (1976). "Exact Maximum Likelihood Estimation of Regression Models with Finite Order Moving Average Errors." The Review of Economic Studies **43**(3): 383-387.

Park, H. et F. Sakaori (2013). "Lag weighted lasso for time series model." Computational Statistics **28**(2): 493-504.

Park, M. Y. et T. Hastie (2007). "L1-regularization path algorithm for generalized linear models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69**(4): 659-677.

Patz, J. A., H. Frumkin, T. Holloway, D. J. Vimont et A. Haines (2014). "Climate change: challenges and opportunities for global health." JAMA.

Peng, R. D. et F. Dominici (2008). Statistical methods for environmental epidemiology with R: A case study in air pollution and health, Springer.

Peng, R. D., F. Dominici et T. A. Louis (2006). "Model choice in time series studies of air pollution and mortality." Journal of the Royal Statistical Society: Series A (Statistics in Society) **169**(2): 179-203.

Pesaran, M. H. (1973). "Exact Maximum Likelihood Estimation of a Regression Equation with a First- Order Moving-Average Error." The Review of Economic Studies **40**(4): 529-535.

Phillips, P. C. B. (1986). "Understanding spurious regressions in econometrics." Journal of Econometrics **33**(3): 311-340.

Phillips, P. C. B. (1987). "Time Series Regression with a Unit Root." Econometrica **55**(2): 277-301.

Phillips, P. C. B. (1998). "New Tools for Understanding Spurious Regressions." Econometrica **66**(6): 1299-1325.

Phung, D., Y. Guo, P. Thai, S. Rutherford, X. Wang, M. Nguyen, C. M. Do, N. H. Nguyen, N. Alam et C. Chu (2016). "The effects of high temperature on cardiovascular admissions in the most populous tropical city in Vietnam." Environmental Pollution **208, Part A**: 33-39.

Qiu, H., I. T.-s. Yu, L. Tian, X. Wang, L. A. Tse, W. Tam et T. W. Wong (2012). "Effects of coarse particulate matter on emergency hospital admissions for respiratory diseases: a time-series analysis in Hong Kong." Environmental health perspectives **120**(4): 572-576.

R Core Team (2015). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Racine, J. (2000). "Consistent cross-validatory model-selection for dependent data: hv-block cross-validation." Journal of Econometrics **99**(1): 39-61.

Ramsay, J. (1982). "When the data are functions." Psychometrika **47**(4): 379-396.

Ramsay, J., H. Wickham, S. Graves et G. Hooker (2011). "fda: Functional data analysis." R package version **2**(6).

Ramsay, J. O. et C. Dalzell (1991). "Some tools for functional data analysis." Journal of the Royal Statistical Society. Series B (Methodological): 539-572.

Ramsay, J. O. et B. W. Silverman (2005). Functional data analysis, Wiley Online Library.

Ratcliffe, S. J., G. Z. Heller et L. R. Leader (2002a). "Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression." Statistics in Medicine **21**(8): 1115-1127.

Ratcliffe, S. J., L. R. Leader et G. Z. Heller (2002b). "Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression." Statistics in Medicine **21**(8): 1103-1114.

Rato, R. T., M. D. Ortigueira et A. G. Batista (2008). "On the HHT, its problems, and some solutions." Mechanical Systems and Signal Processing **22**(6): 1374-1394.

Ravikumar, P., J. Lafferty, H. Liu et L. Wasserman (2009). "Sparse additive models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **71**(5): 1009-1030.

Rehman, N. et D. P. Mandic (2010). "Multivariate empirical mode decomposition." Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science **466**(2117): 1291-1302.

Rehman, N. et D. P. Mandic (2011). "Filter Bank Property of Multivariate Empirical Mode Decomposition." Signal Processing, IEEE Transactions on **59**(5): 2421-2426.

Rehman, N. U., C. Park, N. E. Huang et D. P. Mandic (2013). "EMD Via MEMD: Multivariate Noise-Aided Computation of Standard EMD." Advances in Adaptive Data Analysis **05**(02): 1350007.

Rilling, G., P. Flandrin, P. Gonalves et J. M. Lilly (2007). "Bivariate Empirical Mode Decomposition." Signal Processing Letters, IEEE **14**(12): 936-939.

Rilling, G., P. Flandrin et P. Gonçalvés (2003). On empirical mode decomposition and its algorithms. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP.

Roberts, S. (2005). "Using moving total mortality counts to obtain improved estimates for the effect of air pollution on mortality." Environmental health perspectives.

Sarda, P. et P. Vieu (2000). Kernel Regression. Smoothing and Regression, John Wiley & Sons, Inc.**:** 43-70.

Sarmento, S. M., T. G. Verburg, S. M. Almeida, M. C. Freitas et H. T. Wolterbeek (2011). "Robustness of different regression modelling strategies in epidemiology: a time-series

analysis of hospital admissions and air pollutants in Lisbon (1999–2004)." Environmetrics **22**(1): 86-97.

Sawka, M. N., L. R. Leon, S. J. Montain et L. A. Sonna (2011). Integrated Physiological Mechanisms of Exercise Performance, Adaptation, and Maladaptation to Heat Stress. Comprehensive Physiology, John Wiley & Sons, Inc.

Schimek, M. G. (2000). Smoothing and regression: approaches, computation, and application, Wiley.

Schmidt, P. (1974). "A Modification of the Almon Distributed Lag." Journal of the American Statistical Association **69**(347): 679-681.

Schwartz, J. (1993). "Air Pollution and Daily Mortality in Birmingham, Alabama." American Journal of Epidemiology **137**(10): 1136-1147.

Schwartz, J. (1994). "Nonparametric Smoothing in the Analysis of Air Pollution and Respiratory Illness." The Canadian Journal of Statistics / La Revue Canadienne de Statistique **22**(4): 471-487.

Schwartz, J. (2000a). "The distributed lag between air pollution and daily deaths." Epidemiology **11**(3): 320-326.

Schwartz, J. (2000b). "Harvesting and Long Term Exposure Effects in the Relation between Air Pollution and Mortality." American Journal of Epidemiology **151**(5): 440-448.

Schwartz, J. et A. Marcus (1990). "Mortality and pollution in London: a time series analysis." American Journal of Epidemiology **131**(1): 185-194.

Schwartz, J., J. M. Samet et J. A. Patz (2004). "Hospital Admissions for Heart Disease: The Effects of Temperature and Humidity." Epidemiology **15**(6): 755-761.

Schwartz, J., C. Spix, G. Touloumi, L. Bachárová, T. Barumamdzadeh, A. le Tertre, T. Piekarksi, A. Ponce de Leon, A. Pönkä, G. Rossi, M. Saez et J. P. Schouten (1996). "Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions." Journal of Epidemiology and Community Health **50**(Suppl 1): S3-11.

Schwarz, G. (1978). "Estimating the dimension of a model." The Annals of Statistics **6**(2): 461-464.

Şentürk, D. et H.-G. Müller (2010). "Functional Varying Coefficient Models for Longitudinal Data." Journal of the American Statistical Association **105**(491): 1256-1264.

Sheridan, S. C. (2002). "The redevelopment of a weather-type classification scheme for North America." International Journal of Climatology **22**(1): 51-68.

Shumway, R. H. et D. S. Stoffer (2000). Time series analysis and its applications, Springer New York.

Sillmann, J., V. V. Kharin, F. W. Zwiers, X. Zhang et D. Bronaugh (2013). "Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections." Journal of Geophysical Research: Atmospheres **118**(6): 2473-2493.

Slonosky, V. C. (2015). "Daily minimum and maximum temperature in the St-Lawrence Valley, Quebec: two centuries of climatic observations from Canada." International Journal of Climatology **35**(7): 1662-1681.

Soneja, S., C. S. Jiang, J. Fisher, C. R. Upperman, C. Mitchell et A. Sapkota (2016). "Exposure to extreme heat and precipitation events associated with increased risk of hospitalization for asthma in Maryland, USA." Environmental Health **15**.

Sood, A., G. M. James et G. J. Tellis (2009). "Functional regression: A new model for predicting market penetration of new products." Marketing Science **28**(1): 36-51.

Stewart-Koster, B., J. D. Olden et K. B. Gido (2014). "Quantifying flow-ecology relationships with functional linear models." Hydrological Sciences Journal **59**(Hydrological Science fo Environmental Flows): 629-644.

Stone, M. (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions." Journal of the Royal Statistical Society. Series B (Methodological) **36**(2): 111-147.

Sugg, M. M., C. E. Konrad et C. M. Fuhrmann (2016). "Relationships between maximum temperature and heat-related illness across North Carolina, USA." International journal of biometeorology **60**(5): 663-675.

Suissa, S., S. Dell'Aniello, D. Suissa et P. Ernst (2014). "Friday and weekend hospital stays: effects on mortality." European Respiratory Journal **44**(3): 627-633.

Szpiro, A. A., L. Sheppard, S. D. Adar et J. D. Kaufman (2014). "Estimating acute air pollution health effects from cohort study data." Biometrics **70**(1): 164-174.

Ternynck, C., M. A. Ben Alaya, F. Chebana, S. Dabo-Niang et T. B. M. J. Ouarda (2016). "Streamflow hydrograph classification using functional data analysis." Journal of Hydrometeorology.

Thomas, D. C. (2009). Statistical methods in environmental epidemiology, Oxford University Press.

Tiao, G. C. et W. S. Wei (1976). "Effect of Temporal Aggregation on the Dynamic Relationship of Two Time Series Variables." Biometrika **63**(3): 513-523.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological) **58**(1): 267-288.

Todeschini, R., V. Consonni, A. Mauri et M. Pavan (2004). "Detecting "bad" regression models: multicriteria fitness functions in regression analysis." Analytica Chimica Acta **515**(1): 199-208.

Törő, K., J. Bartholy, R. Pongrácz, Z. Kis, É. Keller et G. Dunay (2010). "Evaluation of meteorological factors on sudden cardiovascular death." Journal of Forensic and Legal Medicine **17**(5): 236-242.

Torres, M. E., M. A. Colominas, G. Schlotthauer et P. Flandrin (2011). A complete ensemble empirical mode decomposition with adaptive noise. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE.

Treagust, D. F., W. Randall et G. E. Folk (1980). "A fourier regression analysis of body temperature of the American opossum, Didelphis virginiana." Journal of Interdisiplinary Cycle Research **11**(2): 135-143.

Tsakalozos, N., K. Drakakis et S. Rickard (2012). "A formal study of the nonlinearity and consistency of the Empirical Mode Decomposition." Signal Processing **92**(9): 1961-1969.

Tsay, R. S. et G. C. Tiao (1984). "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models." Journal of the American Statistical Association **79**(385): 84-96.

Tu, J. V., L. Nardi, J. Fang, J. Liu, L. Khalid, H. Johansen et f. t. C. C. O. R. Team (2009). "National trends in rates of death and hospital admissions related to acute myocardial

infarction, heart failure and stroke, 1994–2004." Canadian Medical Association Journal **180**(13): E118-E125.

Tukey, J. W. (1977). "Exploratory data analysis."

Vanos, J. K., S. Cakmak, L. S. Kalkstein et A. Yagouti (2014). "Association of weather and air pollution interactions on daily mortality in 12 Canadian cities." Air Quality, Atmosphere & Health: 1-14.

Ventosa-Santaulària, D. (2009). "Spurious Regression." Journal of Probability and Statistics **2009**.

Vutcovici, M., M. Goldberg et M.-F. Valois (2013). "Effects of diurnal variations in temperature on non-accidental mortality among the elderly population of Montreal, Québec, 1984–2007." International journal of biometeorology: 1-10.

Wand, M. M. P. et M. C. Jones (1995). Kernel smoothing, Crc Press.

Wang, J.-L. et Z.-J. Li (2012). "What about the asymptotic Behavior of the Intrinsic Mode Functions as the Sifting times Tend to infinity?" Advances in Adaptive Data Analysis **4**(01n02).

Watson, G. S. (1964). "Smooth regression analysis." Sankhyā: The Indian Journal of Statistics, Series A: 359-372.

Wielgosz, A., M. Arango, C. Bancej, A. Bienek, H. Johansen, P. Lindsay, W. Luo, A. Luteyn, C. Nair, P. Quan, P. Stewart, P. Walsh et G. Webster (2009). Suivi des maladies du coeur et des accidents vasculaires cérébraux au Canada. A. d. l. s. p. d. Canada.

Wong, H., R. C. Wu, G. Tomlinson, M. Caesar, H. Abrams, M. W. Carter et D. Morra (2009). "How much do operational processes affect hospital inpatient discharge rates?" Journal of Public Health **31**(4): 546-553.

Wong, T. W., T. S. Lau, T. S. Yu, A. Neller, S. L. Wong, W. Tam et S. W. Pang (1999). "Air pollution and hospital admissions for respiratory and cardiovascular diseases in Hong Kong." Occupational and Environmental Medicine **56**(10): 679-683.

Wu, W., Y. Xiao, G. Li, W. Zeng, H. Lin, S. Rutherford, Y. Xu, Y. Luo, X. Xu, C. Chu et W. Ma (2013). "Temperature–mortality relationship in four subtropical Chinese cities: A time-series study using a distributed lag non-linear model." Science of The Total Environment **449**(0): 355-362.

Wu, Z. et N. E. Huang (2004). "A study of the characteristics of white noise using the empirical mode decomposition method." Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **460**(2046): 1597-1611.

Wu, Z. et N. E. Huang (2009). "Ensemble empirical mode decomposition : A noise-assisted data analysis method." Advances in Adaptive Data Analysis **01**(01): 1-41.

Wu, Z., N. E. Huang, S. R. Long et C.-K. Peng (2007). "On the trend, detrending, and variability of nonlinear and nonstationary time series." Proceedings of the National Academy of Sciences **104**(38): 14889-14894.

Xie, G., Y. Guo, S. Tong et L. Ma (2014). "Calculate excess mortality during heatwaves using Hilbert-Huang transform algorithm." BMC Medical Research Methodology.

Yang, A. C., J.-L. Fuh, N. E. Huang, B.-C. Shia, C.-K. Peng et S.-J. Wang (2011a). "Temporal Associations between Weather and Headache: Analysis by Empirical Mode Decomposition." PLoS ONE **6**(1): e14612.

Yang, A. C., N. E. Huang, C.-K. Peng et S.-J. Tsai (2010). "Do Seasons Have an Influence on the Incidence of Depression? The Use of an Internet Search Engine Query Data as a Proxy of Human Affect." PLoS ONE **5**(10): e13728.

Yang, A. C., S.-J. Tsai et N. E. Huang (2011b). "Decomposing the association of completed suicide with air pollution, weather, and unemployment data at different time scales." Journal of Affective Disorders **129**(1–3): 275-281.

Yang, A. C., C.-H. Yang, C.-J. Hong, Y.-J. Liou, B.-C. Shia, C.-K. Peng, N. E. Huang et S.-J. Tsai (2013). "Effects of Age, Sex, Index Admission, and Predominant Polarity on the Seasonality of Acute Admissions For Bipolar Disorder: A Population-Based Study." Chronobiology international **30**(4): 478-485.

Yang, C., X. Meng, R. Chen, J. Cai, Z. Zhao, Y. Wan et H. Kan (2015). "Long-term variations in the association between ambient temperature and daily cardiovascular mortality in Shanghai, China." Science of The Total Environment **538**: 524-530.

Yang, Z. et L. Yang (2009). "A new definition of the intrinsic mode function." World academy of science.

Yao, F., H.-G. Muller et J.-L. Wang (2005). "Functional linear regression analysis for longitudinal data." The Annals of Statistics(6): 2873-2903.

Yu, W., K. Mengersen, X. Wang, X. Ye, Y. Guo, X. Pan et S. Tong (2012). "Daily average temperature and mortality among the elderly: a meta-analysis and systematic review of epidemiological evidence." International journal of biometeorology **56**(4): 569-581.

Yuan, M. et Y. Lin (2006a). "Model Selection and Estimation in Regression with Grouped Variables." Journal of the Royal Statistical Society. Series B (Statistical Methodology) **68**(1): 49-67.

Yuan, M. et Y. Lin (2006b). "Model selection and estimation in regression with grouped variables." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(1): 49-67.

Zanobetti, A., M. P. Wand, J. Schwartz et L. M. Ryan (2000). "Generalized additive distributed lag models: quantifying mortality displacement." Biostatistics **1**(3): 279-292.

Zhang, J., R. Yan, R. X. Gao et Z. Feng (2010). "Performance enhancement of ensemble empirical mode decomposition." Mechanical Systems and Signal Processing **24**(7): 2104-2123.

Zhang, X., K. K. Lai et S.-Y. Wang (2008). "A new approach for crude oil price analysis based on Empirical Mode Decomposition." Energy Economics **30**(3): 905-918.

Zidek, J. V., H. Wong, N. Le et R. Burnett (1996). "Causality, measurement error and multicollinearity in epidemiology." Environmetrics **7**(4): 441-451.

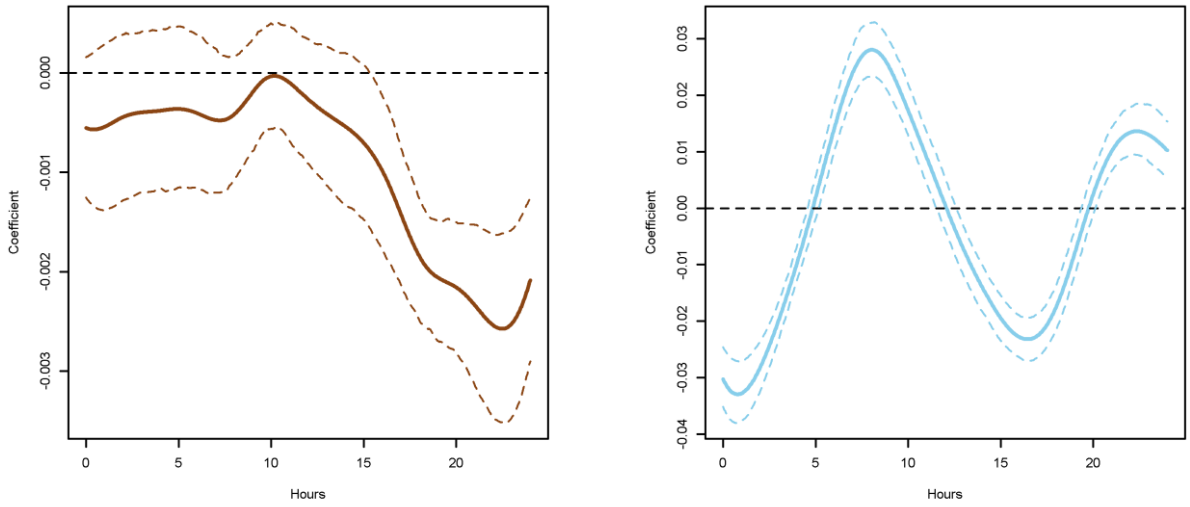
Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties." Journal of the American Statistical Association **101**(476): 1418-1429.

Zou, H. et T. Hastie (2005). "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2): 301-320.

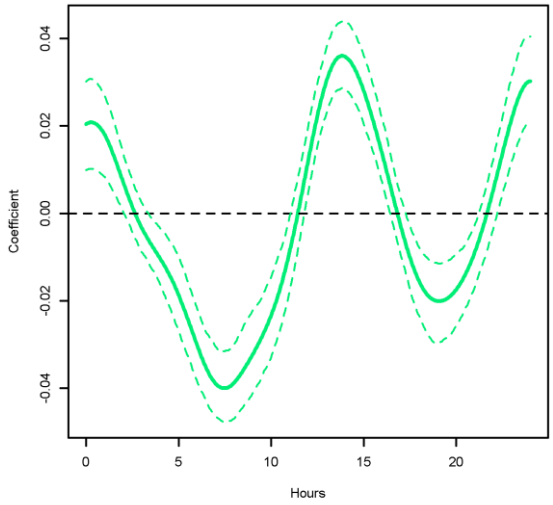
ANNEXES

Annexe A : modèles saisonnier avec RFS

Les résultats montrés ci-dessous ont été produits pendant les travaux de l'article [A3] mais ne furent finalement pas inclus pour des raisons de place et de clarté. Il s'agit du modèle RFS tel que décrit dans la section 3.2.1. de l'article [A3] appliqué sur les saisons d'automne (septembre à novembre), d'hiver (décembre à mars) et de printemps (avril et mai).



a) Automne *b) Hiver*



c) Printemps

Figure A.1: Coefficients fonctionnels estimé pour l'application 1 de l'article [A3]. Les lignes pointillées indiquent l'intervalle de confiance à 95% estimé par wild bootstrap.