Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

19 Septembre 2024

Par : TREIGNER Ael
Titre : Modélisation de séries temporelles et mesure de l'incertitude liée à l'impact du risque climatique sur la mortalité

Confidentialité : Non

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membre présent du jury de l'Institut des Actuaires :
Romain LAILY

Entreprise :
GALEA & Associés

Membres présents du jury de l'EURIA :
Pierre AILLIOT

Directeur de mémoire en entreprise :
LE HO Thomas

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Remerciements

Je tiens à remercier toutes les personnes sans qui la réalisation de mémoire n'aurait pu être possible.

En particulier, je remercie Thomas LE HO pour son accompagnement quotidien, ses nombreuses relectures et remarques sur toute la durée de mon alternance.

Je remercie aussi toutes les personnes de GALEA qui ont pu m'apporter leurs expertises techniques. Je pense spécifiquement à Alexandre EBY, Etienne RAYNAL et Irchad MAMODE VALJEE.

Je remercie l'ensemble du cabinet GALEA pour leur accueil bienveillant.

Merci à l'ensemble du corps enseignant de l'EURIA.

Enfin, merci à ma famille, mes parents et ma grand-mère pour leur soutien et les relectures.

Je n'oublie pas non plus ceux qui ont pu être présent et m'aider à leur manière, notamment un grand merci à l'associé en devenir et à la stagiaire graphisme.

Résumé

L'apparition de nouveaux risques challenge les actuaires dans leurs modélisations, en particulier aux horizons lointains. La quantification de leur impact peut néanmoins être rendue possible grâce à de nouvelles sources d'*open-data*. Par ailleurs, en disposant d'informations suffisantes, des modèles de *machine learning* permettent d'obtenir des prédictions se rapprochant de la réalité.

Des modèles de séries temporelles sont utilisés lorsqu'on souhaite projeter à différents horizons des données ordonnées. Cependant, ces modèles "classiques" ne permettent pas d'intégrer de variables explicatives. Ainsi, des modèles hybrides tel que Prophet, entre modèles de séries temporelles et de *machine learning* peuvent être envisagés. Les résultats en découlant, bien souvent déterministes, donnent notamment aux actuaires la possibilité d'évaluer l'impact de ces risques émergents.

La prise en compte de ces risques, nouveaux en assurance, soulève des interrogations sur la démarche à suivre. Le risque climatique peut rentrer dans cette catégorie et est également suivi par les différents organismes prudentiels. Notamment les stress tests climatiques demandés par les superviseurs visent à quantifier les impacts liés au changement climatique. Il est pour cela nécessaire de recourir à des modèles en capacité d'intégrer des données représentant des facteurs de risques. Des données *open-source* de Météo-France, du DRIAS et de l'INSEE, permettent par exemple d'estimer à long terme l'impact du risque climatique sur la mortalité. Cependant, l'incertitude des résultats augmente avec l'horizon de projection. La quantification de cette incertitude, pourtant considérée comme indispensable dans certains domaines, tel que la médecine, reste secondaire dans le monde de l'actuariat.

L'objectif de ce mémoire est de proposer une réponse méthodologique à l'évaluation du risque climatique à long terme. Plus précisément, ce mémoire cherche à mesurer l'impact du climat sur la mortalité à horizon 2050 et à y apporter une estimation de l'incertitude.

Mots clés : Séries Temporelles - Machine-Learning - Mesure d'incertitude - Mortalité - Risque climatique - Prophet

Abstract

The emergence of new risks is challenging actuaries in their modelling, particularly in the long term. However, quantifying their impact can be achieved thanks to new open-data sources. In addition, machine learning models can be used to obtain predictions that are closer to reality, given the availability of sufficient information.

Time series models are used when you want to project ordered data to different timescales. However, these 'classic' models cannot incorporate explanatory variables. Hybrid models such as Prophet, combining time series and machine learning models, can therefore be considered. The resulting results, which are often deterministic, allow actuaries to assess the impact of emerging risks.

Taking account of these risks, which are new to insurance, raises questions about the approach to be taken. Climate risk may fall into this category, and is also monitored by the various prudential bodies : the climate stress tests requested by supervisors aim to quantify the impact of climate change. This requires the use of models capable of integrating data representing risk factors. Open-source data from Météo-France, DRIAS and INSEE, for example, can be used to estimate the long-term impact of climate risk on mortality. However, the uncertainty of the results increases with the projection horizon. Although quantifying this uncertainty is considered essential in certain fields, such as medicine, it remains of secondary importance in the actuarial world.

The aim of this thesis is to propose a methodological response to the assessment of long-term climate risk. More specifically, this thesis seeks to measure the impact of climate on mortality by 2050 and to provide an estimate of the uncertainty involved.

Keywords : Time series - Data-Science - Uncertainty quantification - Mortality - Climate risk - Prophet

Synthèse

▷ **Introduction :**

L'apparition de nouveaux risques challenge les actuaires dans leurs modélisations, en particulier aux horizons lointains. La quantification de leur impact peut être rendue possible grâce à de nouvelles sources de données, notamment en *open-data*.

En disposant d'informations suffisantes et de modèles adaptés, les actuaires ont la possibilité d'évaluer l'impact de ces risques émergents. En ce sens, les superviseurs émettent de nouvelles demandes, dont certaines pourraient devenir régulières et obligatoires. Ce sont notamment le cas des stress tests climatiques, dont l'objectif est de quantifier les impacts liés au changement climatique.

La prise en compte de ces risques, nouveaux en assurance, soulève tout de même des interrogations sur la démarche à suivre.

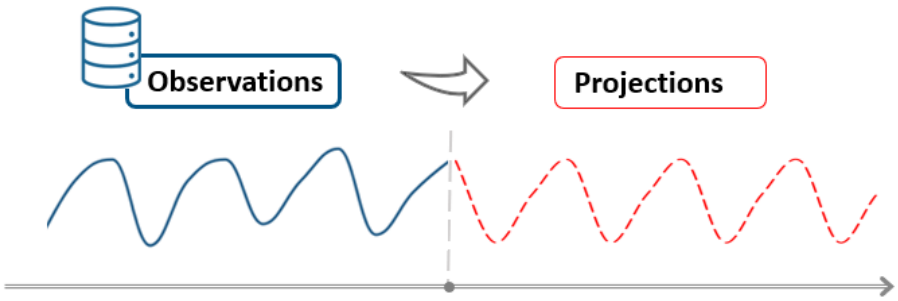
Des modèles de *machine learning* ou de séries temporelles peuvent réaliser des projections à long terme. Cependant, l'incertitude des résultats augmente avec l'horizon de projection. La quantification de cette incertitude semble alors indispensable.

L'objectif est de proposer une réponse méthodologique à l'évaluation du risque climatique à long terme. Plus précisément, ce mémoire cherche à mesurer l'impact du risque climatique sur la mortalité et à y apporter une estimation de l'incertitude.

▷ **Le *machine learning* pour les séries temporelles :**

Afin de projeter les impacts du risque climatique sur les taux de mortalité, il semble nécessaire de recourir à des modèles en capacité d'intégrer des facteurs de risques.

Les projections à long terme sont classiquement envisagées via des modèles de séries temporelles. Toutefois, les modèles traditionnels n'incorporent pas de variables externes, mais analysent les processus $Y = (Y_t)_{t\in\mathbb{N}}$ sous-jacents via différents termes.



Par exemple, dans le cas d'une décomposition additive :

$$Y_t = T_t + S_t + X_t$$

où :
— T_t est le terme de tendance.
— S_t est le terme de saisonnalité.
— X_t est le terme représentant un processus stationnaire.

Le processus $X = (X_t)_{t\in\mathbb{N}}$ est notamment modélisable avec un processus ARMA, afin de capturer à la fois les dépendances linéaires entre les valeurs passées de la série temporelle (via la partie auto-régressive AR) et les dépendances entre les erreurs passées (via la partie moyenne mobile MA).

Des alternatives sont indispensables afin d'intégrer les facteurs de risque climatique. Le modèle Prophet est un hybride entre les modèles de séries temporelles, se basant uniquement sur les observations passées, et les modèles de *machine learning*, qui se basent uniquement sur des variables explicatives extérieures pour leurs prédictions. L'équation du modèle additif est la suivante :

$$y(t) = g(t) + s(t) + h(t) + \beta X(t) + \epsilon_t$$

où :
— $g(t)$ est le terme de tendance.
— $s(t)$ est le terme de saisonnalité.
— $h(t)$ est le terme représentant les jours fériés.
— $\beta X(t)$ sont les apports des régresseurs externes, tels que des variables explicatives climatiques.
— ϵ_t est le terme d'erreur, supposé normalement distribué.

Le modèle Prophet est simple à implémenter, facilement interprétable, permet d'intégrer les variables climatiques désirée et intègre une mesure d'incertitude avec une approche bayésienne.

▷ **État des connaissances :**

Une surmortalité inattendue et persistante est apparue en 2020 avec le Covid, provoquant une rupture de tendance. Il convient alors de se demander si le phénomène est conjoncturel ou structurel. Suite à la recommandation d'experts, une approche conjoncturel sera privilégiée, signifiant que les modélisations chercheront à reprendre la tendance pré-2020.

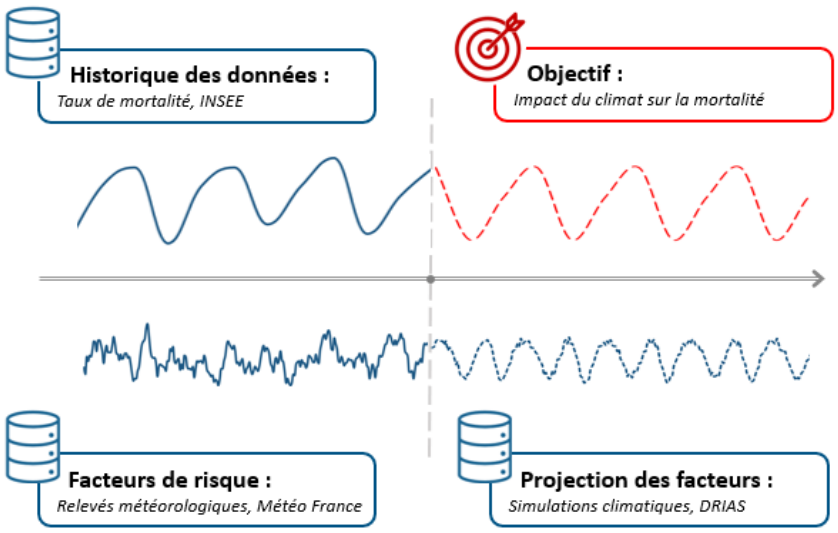
La mortalité humaine est soumise aux facteurs climatiques. Une revue des phénomènes observés permet d'anticiper l'étape de modélisation.

Les décès montrent une saisonnalité annuelle, avec une baisse de la mortalité en été. Cependant, les canicules sont les phénomènes ponctuels les plus impactants. Ce premier constat permet d'anticiper le point suivant : la modélisation des canicules demandera un traitement particulier.

Par ailleurs, la température humide est introduite, à partir de résultats de la littérature médicale, indiquant que certaines conditions croisées de chaleur et d'humidité sont mortelles pour l'homme.

▷ **Données utilisées :**

La projection de l'impact du risque climatique sur la mortalité a nécessité 3 sources de données en *open-data* : l'historique des taux de mortalités, des relevés météorologiques et des projections climatiques.



L'analyse des différentes hypothèses et contraintes sur les regroupements démographiques, temporels et géographiques a mené à privilégier la maille *semaine - département - sexe - tranche d'âge quinquennale*. L'étude se concentrera sur le périmètre 60-79 ans, car les taux de mortalité à ces âges sont les moins volatiles et parmi les plus sensibles au risque climatique.

Historique des données : Taux de mortalité

La base de données permettant d'avoir des taux de mortalité à la maille souhaitée est construite à partir de deux bases de l'INSEE : un fichier de relevés des décès et une estimation de la population.

Les fichiers des décès sont construits à partir des décès remontés mensuellement à l'INSEE par les communes.

Les estimations de la population proviennent des recensements de la population.

En outre, des estimations de l'évolution de la population permettront de passer de taux de mortalité projetés à un nombre de décès attendu lors des interprétations finales.

Différentes estimations seront utilisées, afin d'estimer une part d'incertitude liée à l'évolution démographique.

Facteurs de risque : Relevés météorologiques

Météo France collecte des données climatologique via des réseaux de stations météorologiques. Depuis le 1er janvier 2024, les données concernant un nouveau réseau sont disponibles librement.

Une sélection de ces stations est nécessaire, à partir d'une analyse de leur pertinence et de leur exhaustivité.

Une étape de traitement permet de créer les variables explicatives, telles que l'humidité moyenne sur 30, jours, les températures maximales sur 7 jours, ou une approximation de la température humide maximale sur 7 jours.

Projections des facteurs de risque : Scénarios climatiques

Le DRIAS met à disposition des simulations réalisées à partir de modèles climatologiques se basant sur les trajectoires RCP préconisées par le groupe d'expert du GIEC.

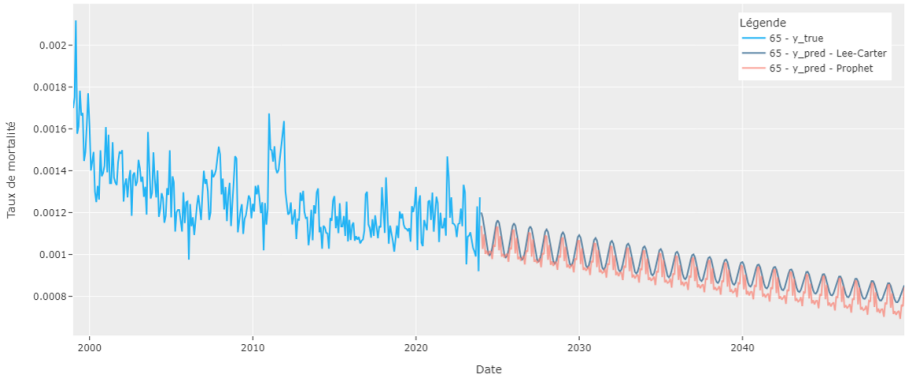
Différentes simulations permettront d'intégrer l'incertitude liée aux scénarios climatiques. Deux simulations seront utilisées : une pour le scénario "optimiste" RCP2.6 et une pour le scénario "pessimiste" RCP8.5.

Un traitement visant à reproduire à l'identique les variables créées à l'étape précédente permet d'avoir une projection des variables explicatives.

▷ **Projections des taux de mortalité :**

Comparaison à un modèle Lee-Carter :

Dans un premier temps, des projections mensuelles par âge réalisées par un modèle Prophet et par un modèle Lee-Carter sont comparées afin de s'assurer de la faisabilité d'une étude de mortalité avec Prophet.



(a) Observations et Projections des taux de mortalité mensuels, 65 ans, Hommes

Âge	Horizon	MAPE
65	2030	4,02%
	2040	4,41%
	2050	5,03%

(b) Erreur relative absolue entre les projections de Prophet et de Lee-Carter

FIGURE 1 – Comparaison des projections des taux de mortalité mensuels des modèles de Lee-Carter et de Prophet

Paramétrage des modèles Prophet :

Par construction, le modèle Prophet impose d'adopter une approche locale, signifiant qu'il sera nécessaire d'entraîner autant de modèles que de segments démographiques considérés, soit 768 modèles (2 sexes, 4 tranches d'âge, 96 départements). Dans ces conditions, le paramétrage devient alors un challenge d'optimisation des modèles sous une contrainte de temps de calcul limité.

Le paramétrage est donc décomposé en 3 étapes.

Dans un premier temps, les hyperparamètres du modèle sont recherchés. Le nombre limité d'hyperparamètres pertinents à cette étape permet de procéder par *grid-search* et *cross-validation*.

Dans un second temps, une méthode visant à quantifier l'apport de l'ajout individuel des régresseurs est mise en place. Des modèles avec régresseur sont évalués et comparés à un modèle de référence, hyperparamétré à partir des résultats de la 1ère étape :

FIGURE 2 – Apport moyen des régresseurs

Deux ensemble de variables sont créés : le *'v1'* avec une ensemble de variables sélectionnés à partir du graphe précèdent et de l'interprétation des variables, et le *'v2'* auquel est ajouté la température humide.

Variable	Add (+)	Mltp (x)	Ensemble 'v1'	Ensemble 'v2'	Signification
t_max_7j	×		×	×	Température maximum sur les 7 derniers jours
t_mean_30j	×		×	×	Température moyenne sur les 30 derniers jours
u_max_7j	×		×	×	Maximum de l'humidité quotidienne moyenne sur les 7 derniers jours
u_max_30j		×	×	×	Maximum de l'humidité quotidienne moyenne sur les 30 derniers jours
u_std_7j		×	×	×	Ecart-type de l'humidité quotidienne moyenne sur les 7 derniers jours
tu_7j	×			×	Température humide, approximée à partir de *t_mean_7j* et *u_mean_7j*

Enfin, la sensibilité des modèles à un ensemble d'hypothèses permettant de prendre en compte la rupture de tendance de 2020 est mesurée. La période d'entraînement, un hyperparamètre lié à la tendance et la présence du Covid vont permettre de trouver un modèle optimisant différents critères qui traduit l'effet conjecturel supposé.

n°	modèle	période observation	*changepoint range*	traitement covid	regresseur	mean_RMSE	mean_MAE
1.2	Prophet	2019	0,99	Non	v2	0,17986	0,14093
2.2	Prophet	2023	0,8	Oui	v2	**0,17848**	**0,14019**
3.0	Prophet	2023	0,8	Non	no	0,17893	0,14038
3.1	Prophet	2023	0,8	Non	v1	0,17879	0,14038
4.0	Prophet	2023	0,99	Non	no	0,17882	0,14033

FIGURE 3 – Extrait d'un tableau de métriques pour des modèles Prophet

▷ **Résultats :**

L'entraînement des modèles, réalisés à partir du paramétrage, permet ensuite de projeter le périmètre étudié, avec le scénario "optimiste" RCP2.6 et avec le scénario "pessimiste" RCP8.5.

Dans la suite, deux points de référence, qui permettront d'évaluer la capacité du modèle à répliquer les canicules, sont à retenir :

— les canicules de 2032 du scénario RCP2.6,
— les canicules de 2027 du scénario RCP8.5.

Apports des scénarios :

Les estimations de population de l'INSEE permettent de passer des taux de mortalité à un nombre de décès. Il est alors possible de sommer les décès en agrégeant les départements, sexes et tranches d'âges, pour avoir une estimation à l'échelle française.

Trois scénarios de populations sont utilisés. Couplées aux deux scénarios RCP, 6 trajectoires sont réalisables par modèle :

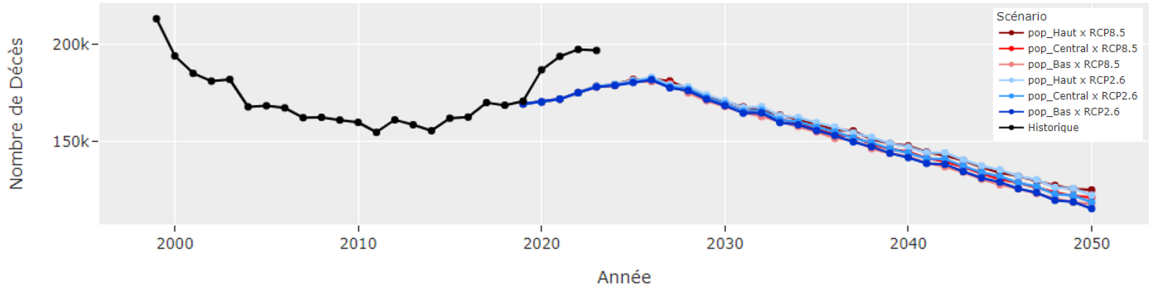


FIGURE 4 – Comparaison des projections du nombre de décès annuel selon les scénarios RCP du DRIAS et les scénarios d'évolution de la population de l'INSEE, 60-79 ans

Il est observé que le modèle considère que la tendance de 2020 est conjecturelle, ce qui provoque un décalage sur 2024 et 2025.

Ces scénarios nous permettent de mesurer une partie de l'incertitude associée. En rapprochant les sorties modèles aux scénarios d'évolution de la population de l'INSEE, l'amplitude du nombre de décès annuel varie de 2% à 8% à horizon 2050.

Ramené à la population, les taux de mortalité observés chuteraient de 12‰ en 2024 à 8‰ en 2050 pour les tranches d'âges de 60 à 79 ans, ce qui concorde avec la tendance historique.

Apports des régresseurs :

Le modèle Prophet ne permet pas de connaître l'apport individuel des régresseurs ajoutés aux modèles. Il est est toutefois possible d'apporter une mesure globale de l'impact du climat en comparant les modèles avec régresseurs à leur modèle de référence, sans régresseurs :

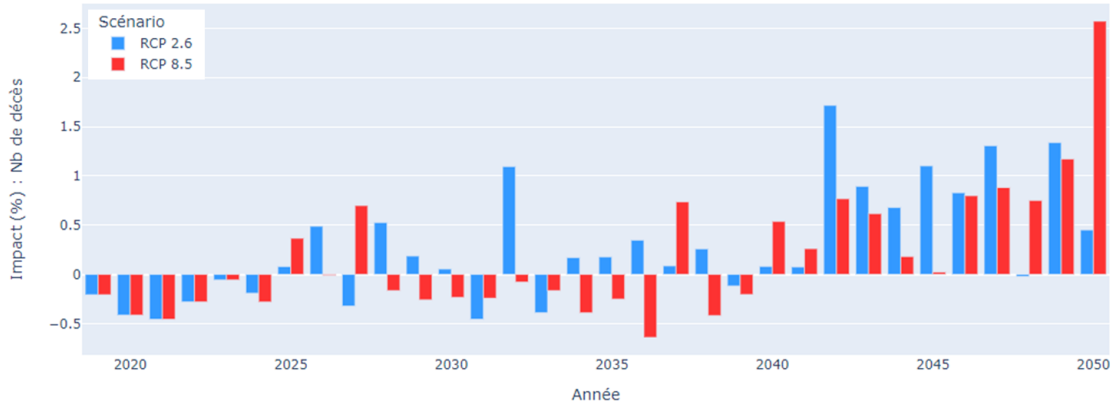


FIGURE 5 – Impact annuel des régresseurs en fonction des scénarios RCP

Les canicules de références en 2027 et 2032 sont répliquées. La fréquence et la sévérité des canicules augmentent avec l'horizon de projection.

Ces effets sont particulièrement évidents en vision hebdomadaire :

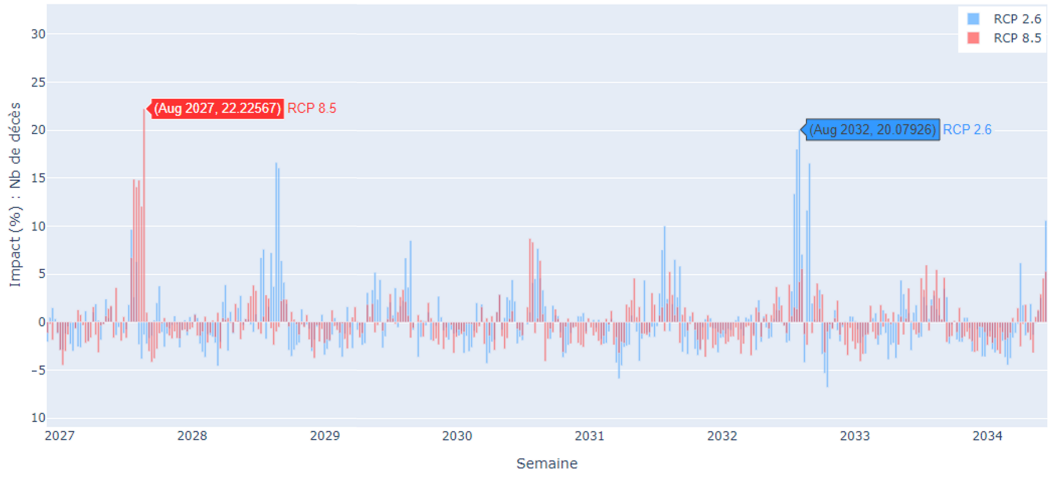


FIGURE 6 – Impact hebdomadaire des régresseurs en fonction des scénarios RCP, 60-79 ans, 2027-2034

Les impacts climatiques sont quantifiables, par exemple pour les canicules suivies :

— en 2027, avec le scénario RCP8.5 : le modèle prévoit 2 796 décès supplémentaires dûs au climat, soit une surmortalité de 12,3%.

— en 2032, avec le scénario RCP2.6 : le modèle prévoit 2 631 décès supplémentaires dûs au climat, soit une surmortalité de 9,8%.

Au global, c'est-à-dire en intégrant à la fois la surmortalité liés aux vagues de chaleurs et la sous-mortalité due aux hivers plus doux, il est attendue 10,15 décès supplémentaires par semaine dus au climat à horizon 2050 d'après le scénario RCP2.6 et 6,02 d'après le scénario RCP8.5. Cela représente des surmortalités respectives de 0,28% et 0,18%.

Mesure d'incertitude :

En plus de l'incertitude apporté par les scénarios RCP et les scénarios de population de l'INSEE, diverses sources d'incertitudes qualitatives sont à retenir, telles que l'incertitude des mesures météorologiques de MétéoFrance.

A l'échelle française, pour la tranche d'âge 60-79 ans, les intervalles de confiance construits avec Prophet induisent des intervalles trop larges pour être interprétés de façon fiable.

Une méthode de *conformal prediction* permettrait d'obtenir un intervalle de prédictions. Son implémentation a été testée avec Neural Prophet, modèle complexifiant l'équation de base du modèle Prophet avec des réseaux neuronaux.

Prophet ne permet de quantifier directement l'impact des régresseurs. Il est possible qu'une part des effets dûs au climat soit intégrés à la saisonnalité.

Un travail approfondi sur la création de variables explicatives serait envisageable pour obtenir de meilleurs résultats.

Synthesis

▷ **Introduction :**

The emergence of new risks challenges actuaries in their modelling, particularly in the long term. Quantifying their impact can be made possible thanks to new sources of data, notably in the form of "open-data".

Armed with sufficient information and suitable models, actuaries can assess the impact of these emerging risks. To this end, supervisors are issuing new requirements, some of which could become regular and mandatory. These include climate stress tests, which aim to quantify the impact of climate change.

Taking account of these risks, which are new to the insurance industry, nevertheless raises questions about the approach to be taken.

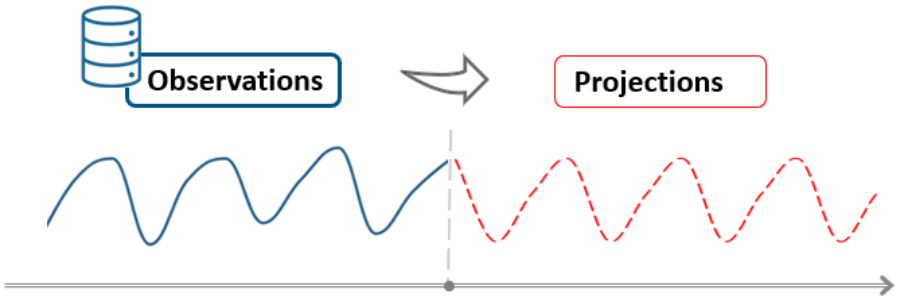
Long-term projections can be made using machine learning or time series models. However, the uncertainty of the results increases with the projection horizon. Quantifying this uncertainty therefore seems essential.

The aim is to propose a methodological response to long-term climate risk assessment. More specifically, this thesis seeks to measure the impact of climate risk on mortality, and to provide an estimate of the uncertainty involved.

▷ **Machine learning for time series :**

In order to project the impact of climate risk on mortality rates, it seems necessary to use models capable of integrating risk factors.

Long-term projections are traditionally made using time-series models. However, traditional models do not incorporate external variables, but analyze the underlying $Y = (Y_t)_{t\in\mathbb{N}}$ processes via different terms.



For example, in the case of additive decomposition :

$$Y_t = T_t + S_t + X_t$$

où :
— T_t is the trend term.
— S_t is the seasonality term.
— X_t is the term representing a stationary process.

The process $X = (X_t)_{t\in\mathbb{N}}$ can be modeled with an ARMA process, in order to capture both linear dependencies between past values of the time series (via the AR auto-regressive part) and dependencies between past errors (via the MA moving average part).

Alternatives are essential to incorporate climate risk factors. The Prophet model is a hybrid between time series models, based solely on past observations, and machine learning models, which rely solely on external explanatory variables for their predictions. The equation for the additive model is as follows :

$$y(t) = g(t) + s(t) + h(t) + \beta X(t) + \epsilon_t$$

where :
— $g(t)$ is the trend term.
— $s(t)$ is the seasonality term.
— $h(t)$ is the term representing public holidays.
— $\beta X(t)$ are the contributions of external regressors, such as climatic explanatory variables.
— ϵ_t is the error term, assumed to be normally distributed.

The Prophet model is simple to implement, easy to interpret, can incorporate the desired climate variables and incorporates a Bayesian uncertainty measure.

▷ **Status of knowledge :**

Unexpected and persistent excess mortality appeared in 2020 with Covid, causing a break in the trend. This raises the question of whether the phenomenon is cyclical or structural. Following the recommendation of experts, a conjunctural approach will be favored, meaning that modeling will seek to resume the pre-2020 trend.

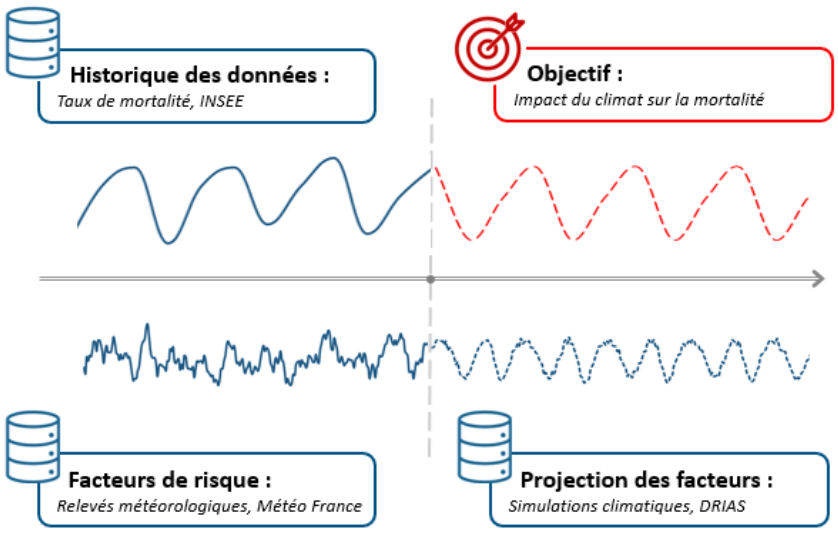
Human mortality is subject to climatic factors. A review of observed phenomena enables us to anticipate the modeling stage.

Deaths show an annual seasonality, with a drop in mortality in summer. However, heatwaves are the most significant one-off phenomena. This first observation allows us to anticipate the next point : modeling heat waves will require special treatment.

To this end, humid temperature is introduced, based on findings in the medical literature indicating that certain cross-conditions of heat and humidity are lethal to humans.

▷ **Data :**

Projecting the impact of climate risk on mortality required 3 sources of data : historical mortality rates, meteorological records and climate projections.



Analysis of the various assumptions and constraints on demographic, temporal and geographic groupings led to the choice of the *week - department - gender - five-year age bracket* grid. The study will focus on the 60-79 age group, as mortality rates at these ages are the least volatile and among the most sensitive to climatic risk.

Data history : Mortality rates

The database used to obtain mortality rates at the desired grid is built from two INSEE databases : a death record file and a population estimate.

Death files are built up from deaths reported monthly to INSEE by local authorities.

Population estimates are derived from population censuses.

In addition, population trend estimates are used to convert projected mortality rates into an expected number of deaths for final interpretation.

Various estimates will be used, in order to estimate the degree of uncertainty associated with demographic change.

Risk factors : Meteorological data

Météo France collects climatological data via networks of weather stations. Since January 1, 2024, data from a new network are freely available.

A selection of these stations is necessary, based on an analysis of their relevance and completeness.

Explanatory variables, such as 30-day average humidity, 7-day maximum temperatures, or an approximation of 7-day maximum wet temperature, are then created.

Risk factor projections : Climate scenarios

DRIAS provides climate model simulations based on the RCP trajectories recommended by the IPCC expert group.

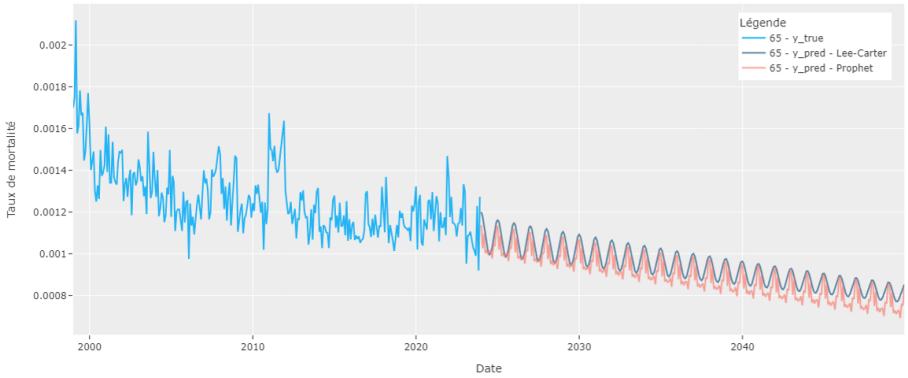
Different simulations are used to incorporate the uncertainty associated with climate scenarios. Two simulations will be used : one for the “optimistic” RCP2.6 scenario and one for the “pessimistic” RCP8.5 scenario.

The variables created in the previous step are reproduced identically to produce a projection of the explanatory variables.

▷ **Projections of mortality rates** :

Comparison with a Lee-Carter model :

As a first step, monthly age-specific projections produced by a Prophet model and by a Lee-Carter model are compared to ensure the feasibility of a mortality study with Prophet.



(a) Observations and Projections of Monthly Mortality Rates, 65 years old, Men

Âge	Horizon	MAPE
65	2030	4,02%
	2040	4,41%
	2050	5,03%

(b) mean absolute relative error between Prophet and Lee-Carter projections

FIGURE 7 – Comparison of Lee-Carter and Prophet projections of monthly mortality rates

Setting up Prophet models :

By design, the Prophet model requires a local approach, meaning that it will be necessary to train as many models as there are demographic segments considered, i.e. 768 models (2 genders, 4 age groups, 96 departments). Under these conditions, parameterization becomes a challenge of model optimization under a constraint of limited computing time.

Parameterization is therefore broken down into 3 stages.

First, the model's hyperparameters are identified. The limited number of relevant hyperparameters at this stage means that we can proceed by grid-search and cross-validation.

In a second step, a method is developed to quantify the contribution of adding individual regressors. Models with regressors are evaluated and compared with a reference model, hyperparameterized from the results of the 1st step :

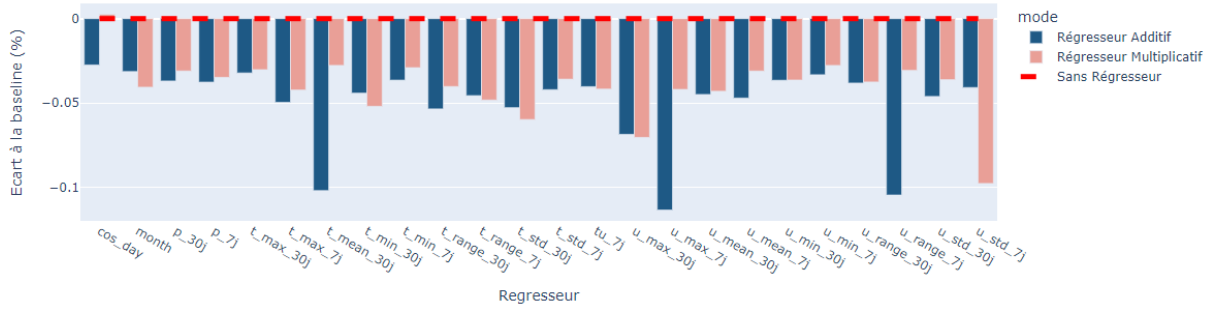


FIGURE 8 – Average contribution of regressors

Two sets of variables are created : the *'v1'* with a set of variables selected from the preceding graph and the interpretation of the variables, and the *'v2'* to which the wet temperature is added.

Variable	Add (+)	Mltp (x)	Ensemble 'v1'	Ensemble 'v2'	Signification
t_max_7j	×		×	×	Température maximum sur les 7 derniers jours
t_mean_30j	×		×	×	Température moyenne sur les 30 derniers jours
u_max_7j	×		×	×	Maximum de l'humidité quotidienne moyenne sur les 7 derniers jours
u_max_30j		×	×	×	Maximum de l'humidité quotidienne moyenne sur les 30 derniers jours
u_std_7j		×	×	×	Ecart-type de l'humidité quotidienne moyenne sur les 7 derniers jours
tu_7j	×			×	Température humide, approximée à partir de *t_mean_7j* et *u_mean_7j*

Finally, the sensitivity of the models to a set of assumptions that allow for the 2020 trend break is measured. The training period, a hyperparameter linked to the trend and the presence of Covid will enable us to find a model optimizing the RMSE that reflects the assumed conjectural effect.

n°	modèle	période observation	*changepoint range*	traitement covid	regresseur	mean_RMSE	mean_MAE
1.2	Prophet	2019	0,99	Non	v2	0,17986	0,14093
2.2	Prophet	2023	0,8	Oui	v2	**0,17848**	**0,14019**
3.0	Prophet	2023	0,8	Non	no	0,17893	0,14038
3.1	Prophet	2023	0,8	Non	v1	0,17879	0,14038
4.0	Prophet	2023	0,99	Non	no	0,17882	0,14033

FIGURE 9 – Extract from a table of metrics for Prophet models

▷ **Results :**

Model training, based on parameterization, then allows us to project the studied perimeter, with the “optimistic” scenario RCP2.6 and with the “pessimistic” scenario RCP8.5.

In the following, two reference points are used to assess the model’s ability to replicate heatwaves :

— the 2032 heatwaves of the RCP2.6 scenario,
— the 2027 heatwaves of the RCP8.5 scenario.

Scenario contributions :

INSEE population estimates can be used to convert mortality rates into numbers of deaths. It is then possible to aggregate deaths by department, gender and age group, to obtain a French-wide estimate.

Three population scenarios are used. Coupled with the two RCP scenarios, 6 trajectories are possible per model :

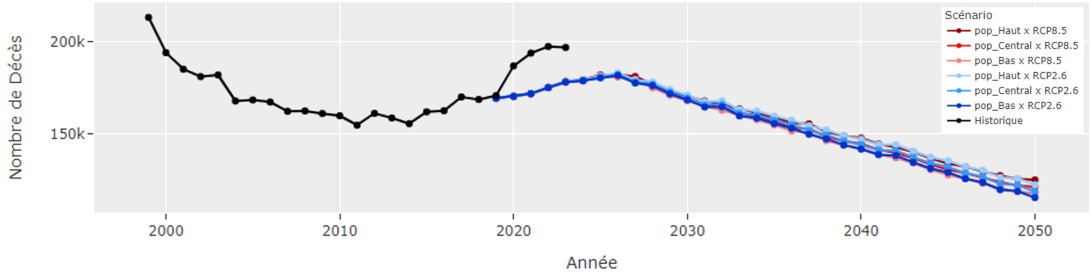


FIGURE 10 – Comparison of annual death projections according to DRIAS RCP scenarios and INSEE population evolution scenarios, 60-79 years

It should be noted that the model considers the 2020 trend to be conjectural, causing a lag in the first 2 years of projection.

These scenarios enable us to measure some of the associated uncertainty. The various INSEE population projection scenarios generate differences ranging from 2% to 8% in the number of annual deaths by 2050.

On a population basis, observed mortality rates would fall from 12‰ in 2024 to 8‰ in 2050 for the 60 to 79 age groups.

Contributions of regressors :

The Prophet model does not allow us to know the individual contribution of the regressors added to the models. However, it is possible to provide an overall measure of the impact of climate by comparing models with regressors to their reference model, without regressors :

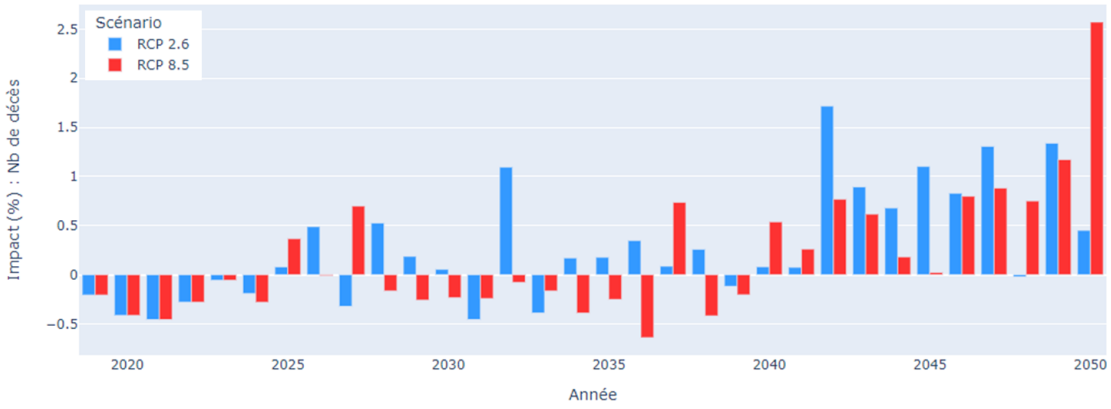


FIGURE 11 – Annual impact of regressors according to RCP scenarios

Reference heatwaves in 2027 and 2032 are replicated. The frequency and severity increase with the projection horizon.

These effects are particularly evident in the weekly view :

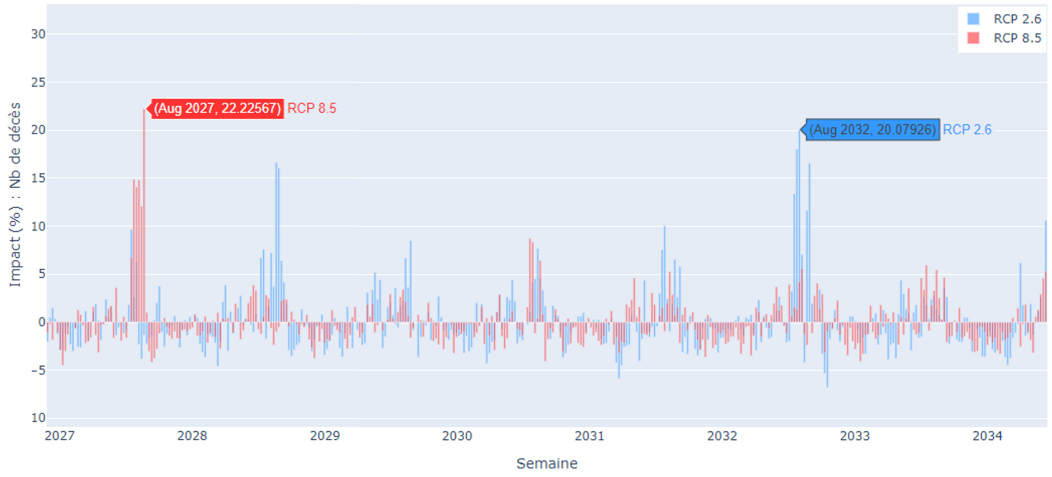


FIGURE 12 – Weekly impact of regressors according to RCP scenarios, 60-79 years, 2027-2034

Climate impacts are quantifiable, e.g. for heatwaves tracked :

— in 2027, with the RCP8.5 scenario : the model predicts 2,796 climate-related deaths, i.e. an excess mortality of 12.3%.

— in 2032, with the RCP2.6 scenario : the model predicts 2,631 climate-related deaths, i.e. an excess mortality of 9.8%.

Overall, i.e. integrating both the excess mortality due to heat waves and the under-mortality due to milder winters, an additional 10.15 climate-related deaths per week are expected by 2050 according to the RCP2.6 scenario, and 6.02 according to the RCP8.5 scenario. This represents excess mortality of 0.28% and 0.18% respectively.

Uncertainty quantification :
In addition to the uncertainty provided by the RCP scenarios and the INSEE population scenarios, various qualitative sources of uncertainty should be taken into account, such as the uncertainty of meteorological measurements by MétéoFrance.

On a French scale, for the 60-79 age group, the confidence intervals calculated predict a weekly death toll of between 400 and 8,000, for predictions of between 3,000 and 4,000.

A *conformal prediction* method could be used to obtain a prediction interval. Its implementation was tested with Neural Prophet, a model that complexifies the basic Prophet equation with neural networks. In the context of this thesis, this complexity requires a computation time that would be unimaginable during parameterization to be applied in full.

▷ **Conclusion :**

The model constructed was used to assess the impact of climatic variables on mortality rates, providing an alternative to duration and time series models.
The expected climatic phenomena were well replicated, as were the known specificities linking mortality rates to climatic variables.
Despite the difficulty caused by the long calculation times required to parameterise the models, it was possible to produce coherent projections that met the requirements of the 2020 trend break.
The various sources of uncertainty were identified throughout the process.

The method seems to underestimate the impact of climate risk on mortality. As all climate hazards are taken into account simultaneously, the models tend to minimise extremes, such as heatwaves. In developed countries, where the climate remains temperate, climate change is still moderate and difficult to model.

Table des matières

Introduction

Les risques afférents aux compagnies du secteur assurantiel peuvent être difficilement quantifiables, particulièrement aux horizons lointains. Bien que le développement des outils à disposition permettent d'obtenir des prédictions de plus en plus proches de la réalité, l'erreur est inhérente à tout modèle mathématique.

La quantification de cette erreur est non seulement liée à la variabilité naturelle des phénomènes observés, mais aussi à l'incapacité des modèles à parfaitement comprendre ces phénomènes et à les répliquer dans leurs prédictions. Les méthodes permettant de mesurer l'incertitude sont parfois sous exploitées en actuariat, et pourraient permettre d'avoir une confiance accrue envers les résultats.

Afin de faciliter le travail des entreprises dans leurs quantifications de ces risques, différentes mesures gouvernementales visent à instaurer "*l'ouverture libre, gratuite et par défaut de toutes les données dont la publication représente un intérêt économique, social, sanitaire ou environnemental*" [1]. Ainsi, plus de données permettant d'estimer certains risques avec fiabilité ont récemment été mises à disposition, notamment, Météo France publie depuis le 1er janvier 2024 différentes données issues des relevés météorologiques de nouvelles stations françaises.

L'impact des risques physiques liés au changement climatique peut par exemple être estimé à partir de ces nouvelles données *open-source*.

Par ailleurs, des projections climatiques ont été réalisées par différents laboratoires d'experts de la modélisation du climat à partir de scénarios émis par le GIEC (Groupe Intergouvernemental d'expert sur l'Evolution du Climat).

Mises à disposition par le DRIAS, ces projections climatiques peuvent être couplées aux données Météo-France afin d'avoir des projections de l'impact du changement climatique sur un périmètre donnée.

L'influence des facteurs climatiques sur la santé humaine est désormais connue et documentée. En actuariat, il peut alors être pertinent de prendre en compte ce risque en assurance Vie.

L'INSEE publie régulièrement différentes données démographiques, permettant notamment de calculer des taux de mortalité. L'impact du climat sur les taux de mortalité est donc quantifiable à partir d'un ensemble de données *open-source*.

Ce mémoire vise à mesurer l'impact et l'incertitude du risque climatique sur la mortalité, principalement via le modèle de séries temporelles Prophet.

Dans un premier temps, les concepts utilisés dans ce mémoire sont présentés. La théorie du *machine learning* pour les séries temporelles est d'abord détaillée. Différentes méthodes de quantification de l'incertitude sont ensuite expliquées. Enfin, les équations des modèles Prophet et Neural Prophet seront développées.

Ensuite, un état des connaissances sur le lien entre la mortalité et le climat sera essentiel afin d'anticiper les résultats finaux.

Les différentes bases de données utilisées seront présentées, ainsi que les traitements nécessaires à leur exploitation.

1. Circulaire n°6264/SG du 27 avril 2021 relative à la politique publique de la donnée, des algorithmes et des codes sources

Dans la dernière partie, les taux de mortalité seront projetés à différentes mailles. Ils seront d'abord comparés à un modèle Lee-carter. Enfin l'impact du climat sera quantifié, en particulier pour les personnes âgées de 60 à 79 ans.

Chapitre 1

Le *machine learning* pour les séries temporelles

Cette partie présente et explique les définitions des concepts de base de ce mémoire. Les séries temporelles et les modèles classiques sont abordés afin de poser un cadre théorique.

Ensuite seront détaillés les concepts et les pratiques usuels d'un projet de machine learning adapté aux séries temporelles.

Enfin, le modèle Prophet qui sera utilisé pour les premières projections sera entièrement explicité.

L'étude des séries temporelles permet de faire des projections d'une série d'observations dans le futur. Les modèles de projections doivent au préalable s'être ajustés aux données, puis être évalués avec des métriques et indicateurs justifiant de leur pertinence.

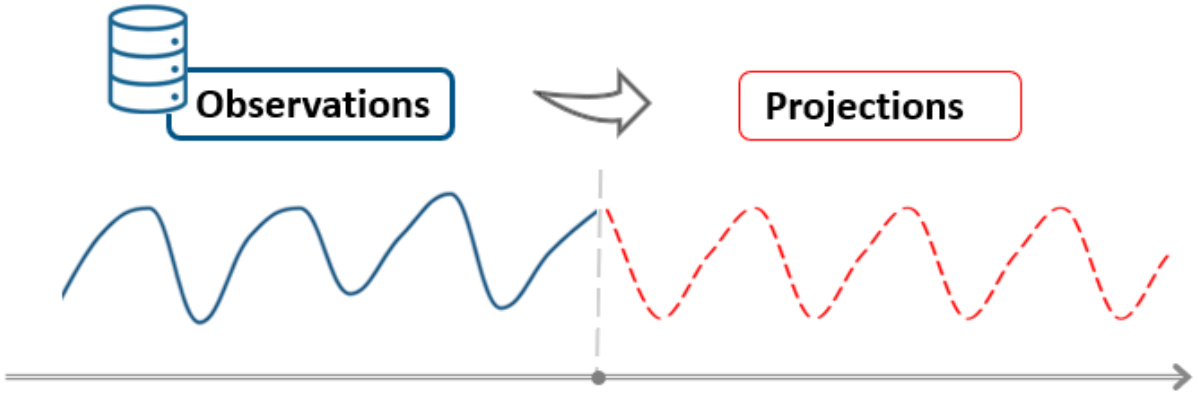


FIGURE 1.1 – Schéma du concept de projection d'une série temporelle

1.1 La théorie des séries temporelles

Une série temporelle est une suite d'observations indexées chronologiquement dans le temps. Ces observations sont des données collectées à intervalles, préférablement réguliers, et peuvent correspondre à n'importe quel format. Par exemple, le prix d'une action, des températures horaires, ou des indicateurs de suivi quotidiens.

L'étude de séries temporelles vise à comprendre les mécanismes sous-jacents pouvant expliquer un phénomène étudié, afin d'obtenir des valeurs futures possibles.
Il sera question de prévisions lorsque les prédictions doivent permettre d'anticiper les prochaines observations à court terme, et de projections lorsque les prédictions doivent représenter une trajectoire envisageable à long terme. Par exemple les prévisions météorologiques sur la semaine à suivre ont pour but de représenter le plus fidèlement possible des variables spécifiques telles que les précipitations, alors que les projections climatiques cherchent à comprendre des tendances globales pour avoir une vue d'ensemble sur plusieurs années.

Dans le cadre de ce mémoire, et comme dans la majorité des applications concrètes, les observations seront à valeurs réelles et disponibles à temps discret. Elles seront notées $(y_t)_{t\in[1,\dots,T]}$, où le nombre d'observations T est la longueur de la série temporelle. Les prédictions issues de modèles seront notées $(\hat{y}_t)_{t>T}$.

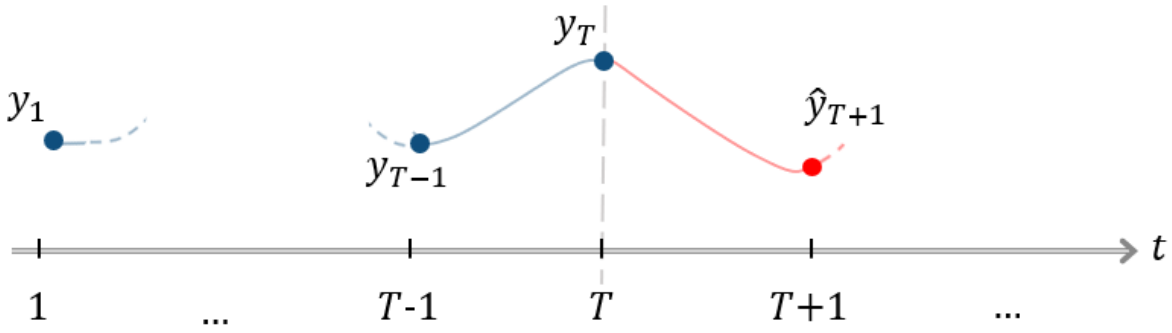


FIGURE 1.2 – Schéma et notation d'une série temporelle

Les séries temporelles peuvent donc être considérées comme des processus stochastiques, c'est-à-dire comme une suite de variables aléatoires $(Y_t)_{t\in\mathbb{N}}$.

Un processus $Y = (Y_t)_{t\in\mathbb{N}}$ est **stationnaire** lorsque la loi qui le définit est constante dans le temps. En absence de corrélation temporelle, un processus est stationnaire si :

$$\begin{cases} \forall t \in \mathbb{N}, & E(Y_t) = m, \quad m \in \mathbb{R} \\ \forall (t,h) \in \mathbb{N}^2, & \text{Cov}(Y_t, Y_{t+h}) = \gamma(h) \end{cases} \tag{1.1}$$

Autrement dit, l'espérance et les covariances des processus stationnaires sont constantes dans le temps. Cette définition correspond à la stationnarité faible, le cas de la stationnarité forte, difficilement vérifiable ne sera pas traité.

En théorie des séries temporelles, il est important de distinguer le cas des séries stationnaires et non-stationnaires. La majorité des séries temporelles "réelles" sont non-stationnaires.

1.1.1 Séries stationnaires : processus ARMA

▷ **Modéliser un processus stationnaire :**

La modélisation d'une série temporelle stationnaire peut comprendre deux termes. La valeur à un instant donné peut dépendre :

— des p dernières valeurs passées. Il conviendra d'introduire un terme auto-régressif d'ordre p : AR(p),

— des q précédentes erreurs aléatoires. Il conviendra d'introduire une moyenne mobile (*moving average*) d'ordre q : MA(q).

Dans tous les cas, il existe des erreurs aléatoires à chaque instant t d'un processus Y, qui sont supposées être des bruits blancs.

Bruits blancs :
Un bruit blanc $\epsilon = (\epsilon_t)_{t\in\mathbb{N}}$ est un cas particulier de processus stationnaire, définie par :

$$\begin{cases} \forall t \in \mathbb{N}, & E(\epsilon_t) = 0 \\ \forall (t,s) \in \mathbb{N}^2, t \neq s, & \text{Cov}(\epsilon_t, \epsilon_s) = 0 \\ \forall t \in \mathbb{N}, & Var(\epsilon_t) = \sigma_\epsilon^2 > 0 \end{cases} \tag{1.2}$$

Ainsi défini, le bruit blanc représente une erreur sans structure temporelle.

Processus auto-régressifs (AR) :
Un processus auto-régressif suppose que la valeur à un instant donné est une combinaison linéaire des valeurs aux temps précédents et d'une erreur aléatoire, représentée par un bruit blanc :

$$AR(p) : \quad Y_t = \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + \cdots + \psi_p Y_{t-p} + \epsilon_t \tag{1.3}$$

où :
— ϵ_t est un bruit blanc,
— $\psi_1, \psi_2, \ldots, \psi_p$ sont les paramètres du modèle qui mesurent l'influence des valeurs passées sur la valeur actuelle,
— p est l'ordre du modèle, qui indique le nombre de valeurs passées à considérer pour prédire la valeur actuelle.

Cette formulation se rapproche d'une régression linéaire, dans laquelle les ψ_i joueraient le rôle de variables explicatives, d'où le nom auto-régressif.

La majorité des séries temporelles rencontrées dans la nature contiennent un terme auto-régressif. En effet, il parait assez intuitif qu'une observation à un moment donné dépende de la dernière observation.

Moyennes mobiles (MA) :
Une moyenne mobile est un processus où la valeur à un instant est une combinaison linéaire des erreurs aléatoires observées aux temps précédents, représentées par des bruits blancs :

$$MA(q) : \quad Y_t = \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q} \tag{1.4}$$

où :
— ϵ_t est un bruit blanc,
— $\phi_1, \phi_2, \ldots, \phi_q$ sont les paramètres du modèle qui mesurent l'influence des erreurs passées sur la valeur actuelle,

— q est l'ordre du modèle, qui indique le nombre de termes d'erreurs passées à considérer pour prédire la valeur actuelle.

Processus ARMA :
Un processus stationnaire peut présenter les deux effets présentés, ce qui nécessite d'introduire le modèle ARMA (auto-régressif *moving-average*). Ce modèle combine aisément les deux équations 1.3 et 1.4 :

$$ARMA(p,q): \quad Y_t = \sum_{i=1}^{p} \psi_i Y_{t-i} + \epsilon_t + \sum_{j=1}^{q} \phi_j \epsilon_{t-j} \tag{1.5}$$

Le modèle ARMA(p, q) est utilisé pour capturer à la fois les dépendances linéaires entre les valeurs passées de la série temporelle (via la partie AR) et les dépendances entre les erreurs passées (via la partie MA), offrant ainsi une approche complète pour modéliser des séries temporelles stationnaires.

▷ **Sélectionner et valider le modèle :**

Les processus stationnaires peuvent être modélisés par un processus ARMA(p,q). Cependant les ordres p et q du modèle ne sont pas connus, il faut donc recourir à une méthode appropriée afin de trouver les paramètres optimaux.

Une analyse graphique des fonctions d'auto-corrélation $\rho(h) = \gamma(h)/\gamma(0)$ (ACF) peut donner des indications sur les ordres à choisir (la fonction γ est introduite en 1.1). Cependant, la recherche d'optimisation de la métrique AIC est généralement privilégiée, car elle apporte une mesure quantitative.

Le Critère d'Akaike (AIC) :
La recherche de paramètres optimaux avec le critère AIC s'approche d'une méthode naïve. Tous les modèles possibles sont entraînés, pour un ensemble de paramètres p et q préalablement choisi, par exemple $(p,q) \in [1,2,3]^2$.

Le critère d'Akaike évalue la qualité relative des modèles, en les comparant entre eux, à partir de la métrique suivante :

$$AIC = 2k - 2ln(L) \tag{1.6}$$

où :
— k est le nombre de paramètres à estimer du modèle, permettant de pénaliser les modèles plus complexes.
— L est le maximum de la fonction de vraisemblance, calculée pour ajuster le modèle aux données. Une vraisemblance plus élevée correspond à un meilleur ajustement du modèle aux données.

L'AIC n'a pas d'interprétation directe, mais permet de sélectionner un modèle à partir d'un critère objectif. L'AIC la plus faible offre le meilleur compromis entre l'ajustement et la complexité du modèle.

Le recours à des techniques utilisées en *data-science* est également envisageable pour sélectionner le meilleur modèle. Ces méthodes seront développées en partie 1.2.4.

Lorsqu'un modèle est sélectionné et ajusté aux données, il est possible de réaliser les premières prédictions. Il vaut toutefois mieux s'assurer que les hypothèses formulées soient bien vérifiées.

Vérification des hypothèses principales :
Les formulations précédentes font l'hypothèse que les termes d'erreurs ϵ sont des bruits blancs. Afin de vérifier cette hypothèse, on peut se rapporter à la fonction d'auto-corrélation **ACF** $\gamma(h)$, qui, pour un bruit blanc doit vérifier :

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} 1 & \text{si } h = 0 \\ 0 & \text{sinon} \end{cases} \tag{1.7}$$

Dans le cas d'un bruit blanc, l'ACF des résidus du modèle doit présenter des valeurs nulles à tous les *lags* h autres que 0 :

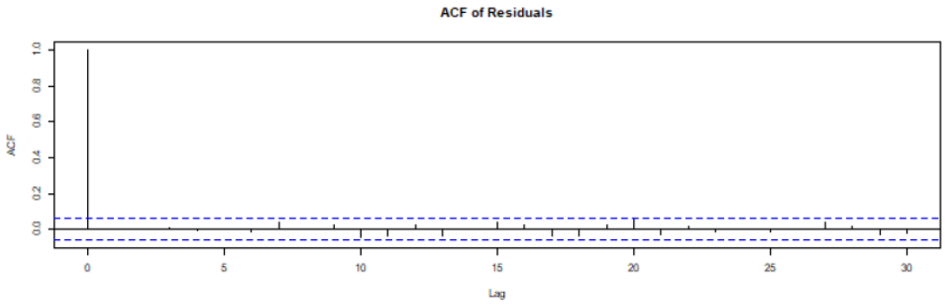


FIGURE 1.3 – Exemple d'ACF d'un bruit blanc

Ce contrôle visuel peut être complété par un test statistique permettant de vérifier formellement que les résidus d'un modèle ne présentent pas une auto-corrélation significative à plusieurs *lags* : le **test de Ljung-Box**.

L'hypothèse nulle, selon laquelle les résidus ne présentent pas d'auto-corrélation jusqu'au *lag* h est basée sur la statistique :

$$Q = T(T+2)\sum_{k=1}^{h} \frac{\hat{\rho}(k)^2}{T-k}$$

où :
— T est le nombre d'observations,
— $\hat{\rho}(k)^2$ est l'estimation de l'auto-corrélation au *lag* k.

Une p_value supérieure à 5% permet d'accepter l'hypothèse que les résidus sont des bruits blanc.

A ce test de bruit blanc, il peut être pertinent de réaliser un autre test pour vérifier que les bruits blancs sont gaussiens. Cette hypothèse est en effet posée lors de la maximisation de la fonction de vraisemblance et pour la construction d'intervalles de confiance.

La méthode graphique la plus courante est de comparer la distribution observée des résidus à celle d'une distribution de référence. Pour cela, les quantiles sont tracés sur les deux axes, pour réaliser un **QQ-plot** (quantile-quantile) :

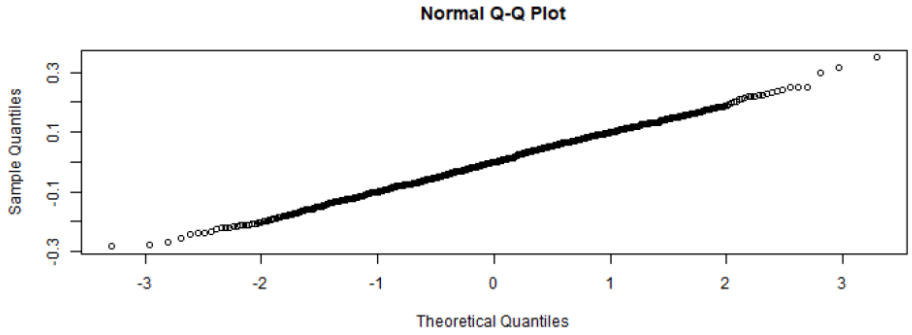


FIGURE 1.4 – Exemple d'un QQ-plot normal

Lorsque que les deux distributions concordent, les points s'alignent sur la diagonale. Et par exemple, des queues plus épaisses ou plus fines que celles de la distribution normale peuvent apparaître comme des courbures dans les extrémités du QQ-plot.

Comme pour les bruits blancs, le contrôle visuel de l'hypothèse gaussienne peut être complété par un test statistique vérifiant que les quantiles observés suivent ceux d'une distribution normale : **le test de Shapiro-Wilk**.

L'hypothèse nulle, selon laquelle les résidus peuvent suivre une loi normale est basée sur la statistique :

$$W = \frac{\left(\sum_{i=1}^{T} a_i Y_{(i)}\right)^2}{\sum_{i=1}^{T}(Y_{(i)} - \bar{Y})^2}$$

où :
— T est le nombre d'observations,
— $\bar{Y}$ est la moyenne des données,
— a_i sont des coefficients spécifiques au test, déterminés en fonction de la distribution normale. [1]

Une p_value supérieure à 5% permet d'accepter l'hypothèse que les résidus sont des bruits blancs gaussiens.

Le modèle ARMA permet de représenter un processus stationnaire. Cependant, bien souvent les séries temporelles ne vérifieront pas la stationnarité.

1. S.S. Shapiro et M.B. Wilk, « *An analysis of variance test for normality (complete samples)* », Biometrika

1.1.2 Séries non-stationnaires : décomposition additive et processus SARIMA

▷ **Décomposer un processus :**

Dans les cas où le processus représentant la série temporelle n'est pas stationnaire, c'est-à-dire qu'il ne vérifie pas les propriétés de la définition 1.1, une décomposition additive est généralement supposée, de la forme suivante :

$$Y_t = T_t + S_t + X_t \tag{1.8}$$

où :
— T_t est le terme de tendance.
— S_t est le terme de saisonnalité.
— X_t est le terme représentant un processus stationnaire.

Cette décomposition représente un modèle paramétrique, dans laquelle $(Y_t)_{t\in\mathbb{N}}$ et $(X_t)_{t\in\mathbb{N}}$ sont des processus stochastiques, alors que $(T_t)_{t\in\mathbb{N}}$ et $(S_t)_{t\in\mathbb{N}}$ sont des composantes non-stationnaires, modélisables par des fonctions déterministes. La composante aléatoire est donc uniquement portée par le processus stationnaire X.

La tendance :
La tendance T_t représente le mouvement général à long terme des observations. Elle permet de capturer des changements structurels globaux, montrant l'évolution moyenne sur une période étendue.

Dans ses formes les plus simples, la tendance peut être modélisée via une régression linéaire, une moyenne mobile ou un polynôme.

La saisonnalité :
La saisonnalité S_t représente les effets récurrents, se répétant à des intervalles réguliers, formant ainsi un cycle de paternes anticipables. Par exemple, avec des données mensuelles, un cycle de 12 périodes représente une saisonnalité annuelle.

Ce terme est généralement modélisé par une série de Fourier, c'est-à-dire un polynôme trigonométrique construit à partir des fonctions cosinus et sinus.

Le terme stationnaire :
En supposant que les termes de tendance et de saisonnalité soient parfaitement représentés, alors le dernier terme d'un processus Y peut être stationnaire.

En effet, si la tendance est entièrement captée par T_t, alors l'espérance de X_t ne doit pas dépendre du temps.
De même, si la saisonnalité S_t représente complètement les cycles récurrents, alors la covariance de X_t ne doit pas être dépendante du temps.
Sous ces deux conditions, en se ramenant à la définiton 1.1, le processus X est bien stationnaire.

Une modélisation suivant ce modèle additif consiste alors à représenter la tendance et la saisonnalité par des fonctions déterministes, où les paramètres doivent être optimisés, puis à modéliser la composante stationnaire en retrouvant le modèle ARMA. Il peut être intéressant

de noter que cette décomposition peut être multiplicative, ce qui ne change pas la méthode.

Une autre approche consiste à appliquer des opérations permettant de se ramener à un processus stationnaire tout en prenant en compte les éventuelles tendance et saisonnalité. Ce modèle correspond alors à un processus SARIMA.

▷ **Intégrer les composantes non-stationnaires au modèle :**

Plutôt que de modéliser les composantes stationnaires séparément, la série temporelle est transformée, à l'aide d'une différentiation, pour retrouver un processus stationnaire.

Différenciation :
La différentiation est une technique qui doit permettre d'éliminer la tendance d'un processus, en calculant des différences entre les observations successives :

$$\Delta Y_t = Y_t - Y_{t-1}$$

$$\Delta Y_t = (1 - B)Y_t$$

où Δ correspond à l'opérateur différentiation et B est l'opérateur retard.

L'équation du modèle ARMA peut être retravaillée, ce qui permettra d'introduire ensuite le modèle ARIMA (*Integrated* ARMA) :

$$ARMA(p, q) : \quad \Phi(B)Y_t = \Psi(B)\epsilon_t$$

où :
— $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ est la fonction d'auto-régression d'ordre p,
— $\Psi(B) = 1 + \psi_1 B - \cdots - \psi_q B^q$ est la fonction de moyenne mobile d'ordre q.

Un processus ARIMA(p,d,q) correspond alors à un processus différencié d fois afin d'être stationnaire, ce qui permet de pouvoir le modéliser par un ARMA(p,q) :

$$ARIMA(p, d, q) : \quad (1 - B)^d \Phi(B)Y_t = \Psi(B)\epsilon_t \tag{1.9}$$

Bien souvent, une composante saisonnière est également à prendre en compte, ce qui demande de passer au modèle SARIMA (*Seasonal* ARIMA).

Composante saisonnière :
Une saisonnalité de période s signifie qu'un cycle de motifs recommence toutes les s observations. Afin de prendre en compte cette saisonnalité, l'opérateur de différentiation d'ordre s est utilisé :

$$\Delta_s Y_t = Y_t - Y_{t-s}$$

$$\Delta_s Y_t = (1 - B^s)Y_t$$

Le modèle SARIMA intègre alors une composante auto-régressive saisonnière d'ordre P et une moyenne mobile saisonnière d'ordre Q, différenciées à l'ordre D à partir de la période s :

$$SARIMA(p,d,q)(P,D,Q)_s : \quad \Phi(B)(1-B)^d \cdot \Phi'(B^s)(1-B^s)^D \cdot Y_t = \Psi(B) \cdot \Psi'(B^s) \cdot \epsilon_t \quad (1.10)$$

où :
— $\Phi'(B^s) = 1 - \phi'_1 B^s - \cdots - \phi'_P B^{P \times s}$ est la fonction d'auto-régression saisonnière d'ordre P,
— $\Psi'(B^s) = 1 + \psi'_1 B^s - \cdots - \psi'_Q B^{Q \times s}$ est la fonction de moyenne mobile saisonnière d'ordre Q.

Avec cet ensemble de termes, le modèle SARIMA permet de prendre en compte une saisonnalité et une tendance comme le ferait une décomposition additive. La sélection des paramètres et la validation du modèle suit le même principe que pour le modèle ARMA.

La modélisation d'une série temporelle via un processus SARIMA est la méthode "classique" la plus courante suivant la théorie mathématique dédiée aux séries temporelles. D'autres approches plus génériques, comme l'utilisation de modèles de *machine learning* est aussi envisageable pour modéliser des séries temporelles, ceci fera l'objet de la prochaine partie. Afin d'anticiper la suite du mémoire, il faut également noter le développement de modèles de *machine learning* spécifiquement dédiés aux séries temporelles, tel que le modèle Prophet, qui sera développé en partie 1.4.

1.2 Les basiques du *machine learning*

Le *machine learning*, ou apprentissage automatique / statistique en français, est une branche de l'intelligence artificielle. Elle est dédiée à la conception, au développement et à la mise en pratique d'algorithmes permettant à des systèmes informatiques d'apprendre à identifier des motifs ou des relations à partir d'un ensemble de données afin de réaliser des prédictions ou de prendre des décisions. L'optimisation et l'amélioration des performances durant l'entraînement des modèles sont basées sur la minimisation d'erreurs statistiques.

La théorie mathématiques du *machine learning* utilise généralement les notations suivantes :
- y **est la variable cible, à expliquer ou à prédire**,
- X **représente l'ensemble des variables explicatives**,
- $\hat{y}$ **est la prédiction du modèle** de la variable cible y, réalisée à partir des variables explicatives X.

Comme il a été vu en partie 1.1, les séries temporelles classiques ne prennent pas en compte de variables explicatives X.
Ce point met en avant l'intérêt de ce mémoire, qui consiste à essayer des modèles de *machine learning* dédiés aux séries temporelles afin de prendre en compte l'influence de variables externes sur des projections.

Il sera question de prédictions dans le cadre de modèles de *machine learning* et de projections dans le cadre spécifique des séries temporelles. Les projections peuvent être considérées comme des prédictions ordonnées dans le temps.

Les modèles de *machine learning* sont classés en deux grandes catégories, en fonction de la façon dont les données sont utilisées pour entraîner les modèles :
- l'apprentissage supervisé, basé sur des données étiquetées, sur lesquelles une réponse correcte est connue.
- l'apprentissage non-supervisé, basé sur des données non-étiquetées, sans réponse prédéfinie.

Apprentissage supervisé :
L'apprentissage supervisé regroupe la majorité des modèles de *machine learning*. L'objectif de ces modèles est de réaliser des prédictions. Pour cela, les données observées sont étiquetées, c'est-à-dire que l'ensemble des variables explicatives représentant une observation sont associées à une valeur cible connue, continue ou discrète.
Dans le cas d'une valeur continue, par exemple un prix, c'est un problème de régression. C'est le type de problème le plus répandu, et également celui rencontré dans ce mémoire avec des taux de mortalité continus. Dans le cas d'une variable discrète, par exemple une indicatrice de fraude (0/1), c'est un problème de classification. Dans les deux cas, les modèles fréquemment utilisés sont des GLM (Modèle Linéaire Généralisé) ou des modèles construits autour d'arbres de décisions, comme RandomForest ou XGBoost. Ces modèles recherchent les interactions entre variables explicatives influant sur la variable cible, et s'évaluent en comparant leurs prédictions à des observations dont l'étiquette est connue.

Apprentissage non-supervisé :
L'apprentissage non-supervisé est plus minoritaire. Les modèles de ce type n'ont pas pour objectif de réaliser des prédictions mais de trouver des structures dans un ensemble de données. Les modèles d'apprentissage non-supervisés les plus communs sont les modèles de *clustering*, qui visent à réaliser des groupes d'observations similaires, souvent en introduisant une notion de distance, tels que les modèles KNN ou DBSCAN. Des méthodes comme l'ACP (Analyse en Composantes Principales), dont le but est réduire la dimensionnalité de données, peuvent aussi correspondre à cette catégorie.

Construction d'un modèle :

Quelle que soit la problématique, le schéma de construction d'un modèle de *machine learning* est identique.

Le point de départ est l'identification du type de modèle à mettre en place et du format de l'éventuelle variable cible.

Ensuite vient la préparation des données en vue de cet objectif, qui se termine par la séparation des données en un jeu de données d'entraînement et un jeu de données de test.

Le modèle peut alors être construit, avec un ensemble de paramètres qui lui sont propres.

Le modèle est ajusté sur les données d'entraînement et est prêt à être évalué. Pour cela, une métrique permettant de mesurer la performance du modèle est définie, puis est calculée en se basant sur les prédictions du modèle appliqué aux données de test. D'autres ensembles de paramètres peuvent ensuite être envisagés pour trouver la combinaison minimisant la métrique.

Ces étapes sont reprises dans le schéma suivant, et vont être détaillées ci-dessous :

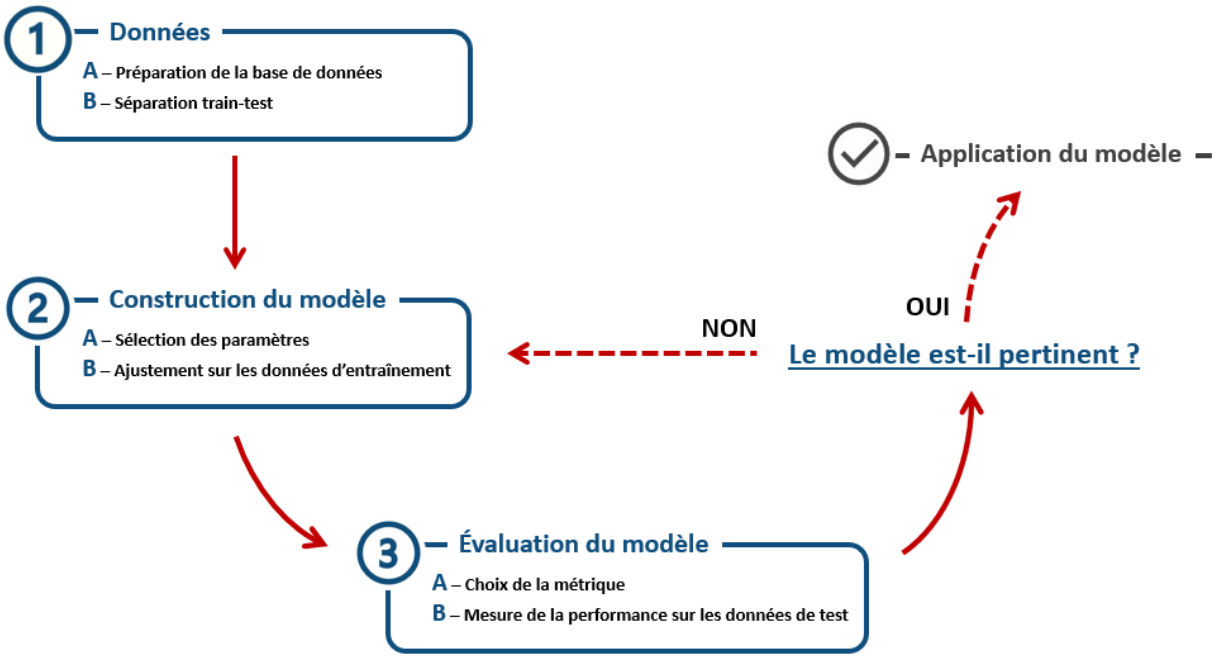


FIGURE 1.5 – Schéma des étapes de la mise en place d'un modèle de *machine learning*

Ce mémoire utilisera uniquement des modèles de régression, l'accent sera donc mis sur ce type de modèle dans la suite.

1.2.1 Préparer les données à la modélisation

▷ **Identifier le format adéquat :**

La construction d'une base de données qui permettra aux modèles d'utiliser les variables explicatives de manière optimale est la première étape pratique d'un processus de modélisation.

Il faut d'abord choisir la maille de l'étude, c'est-à-dire se demander ce que doit représenter une ligne d'observation dans la base de données finale. Dans ce mémoire, la sélection de la maille a une place prépondérante : parfois une ligne de données représentera une semaine ou une année, et pourra également représenter un département ou la France métropolitaine.

Lorsque la maille sera évoquée, les notations suivantes, propres à ce mémoire, seront utilisées : *mailleA - mailleB*.

Les variables explicatives doivent ensuite subir des traitements spécifiques(*features engineering*).

Le premier traitement, qu'il est au moins indispensable de vérifier, est la présence de valeurs manquantes. L'imputation de ces éventuels "trous d'informations" peut aller de méthodes simples, comme des décisions arbitraires, à des modèles complexes de générations de données artificielles.

Ces traitements peuvent ensuite inclure des transformations de données, telles que la normalisation pour modifier l'échelle d'une variable continue, ou l'encodage qui transpose une variable catégorielle à N modalités en $N-1$ colonnes d'indicatrices. Dans l'étude présentée, toutes les variables sont continues, et aucun traitement de ce type ne leur sera appliqué.

La création de nouvelles variables explicatives est également envisageable (par exemple dans le cas de données temporelles, ajouter une colonne représentant la saison).

▷ **Séparer en *train-test* :**

Les bases train-test :
Lorsque que le jeu de données est au format souhaité, avec des variables correspondant aux attentes posées à la fois par la problématique et par le modèle, un dernier travail de données pré-modélisation est impératif : la séparation en base d'apprentissage et base de test (*train-test)*. La séparation train-test est indispensable afin d'évaluer la pertinence d'un modèle d'apprentissage supervisé.

La base d'entraînement, représentant usuellement 70 à 80% des observations, permet au modèle de s'ajuster en apprenant les relations entre variables explicatives X, ce qui permettra de prédire la variable cible y.
La base de test contient le reste de la base de données initiale. Le modèle n'a pas eu connaissance de ces observations. Le modèle ajusté réalise des prédictions à partir des variables explicatives X, et il est alors possible de comparer les prédictions $\hat{y}$ à la réalité y.

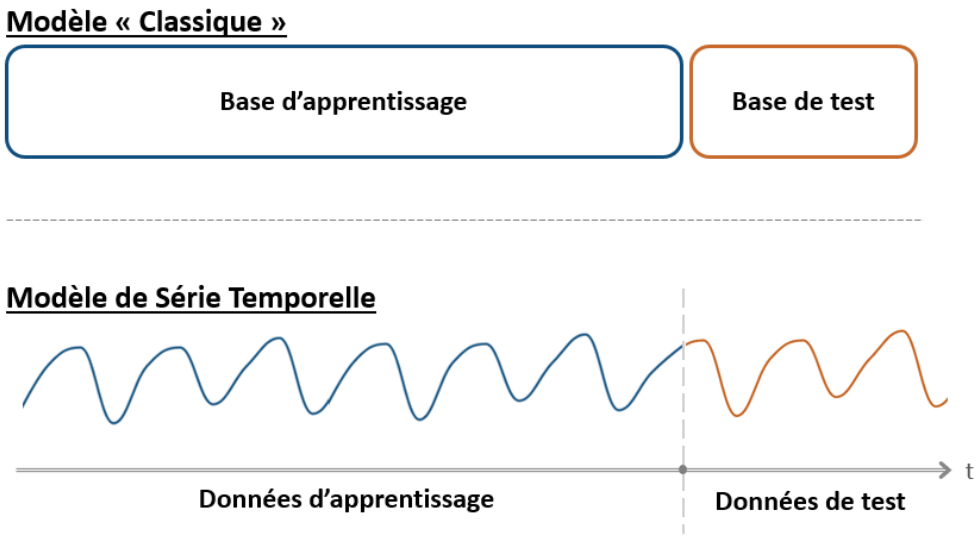


FIGURE 1.6 – Séparation train-test

Dans le cas de séries temporelles, la séparation train-test se fait en prenant en compte le caractère ordonné des données. La base d'entraînement représente les $\sim 75\%$ premières observations, et la base de test représente les $\sim 25\%$ restant. Les projections dans le temps se font dans la continuité de la dernière observation du jeu de test.

Dans les autres cas plus classiques, où les données ne sont pas ordonnées, la séparation train-test se fait aléatoirement. Cependant un biais lié à l'aléatoire existe, car la séparation peut créer des jeux de données non-représentatifs ou distribués non équitablement. Pour atténuer ce biais, un jeu de validation peut être ajouté. Souvent, la validation croisée (*cross-validation)* sera mise en pratique.

La cross-validation :

En réalité, tester la pertinence du modèle sur un unique jeu de données n'est pas optimal. **Multiplier les évaluations permet de réduire la volatilité des résultats et de s'assurer qu'un modèle ne fait pas de sur-apprentissage (*overfitting*)**. Autrement dit, cela permet de s'assurer que les performances du modèle sont stables et ne dépendent pas de la base de test. Pour cela, le concept de *cross-validation* doit être introduit : afin de réduire le biais d'échantillonnage, le modèle est ajusté sur plusieurs jeux d'entraînement et évalué sur plusieurs jeux de test. Ces bases ne se construisent pas de la même façon dans le cas de données ordonnées ou non :

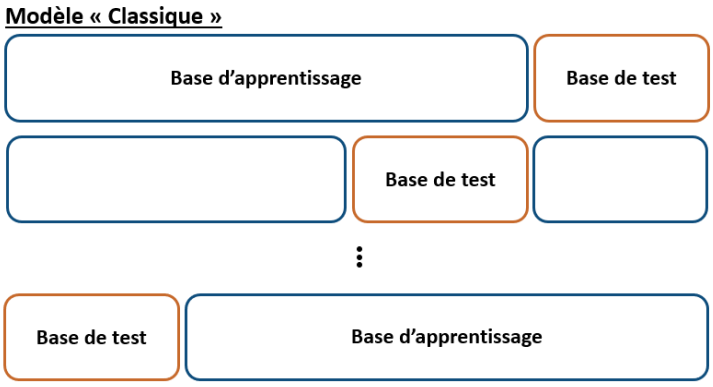


FIGURE 1.7 – Cross-Validation, cas "classique" de données non-ordonnées

Dans le cas de données non-ordonnées, le jeu de données est divisé en k segments aléatoires de tailles égales. Le modèle est testé sur 1 des segments après avoir été entraîné sur les $k-1$ autres. Le modèle peut ainsi être évalué k fois, ce qui permet d'augmenter la robustesse de l'évaluation.

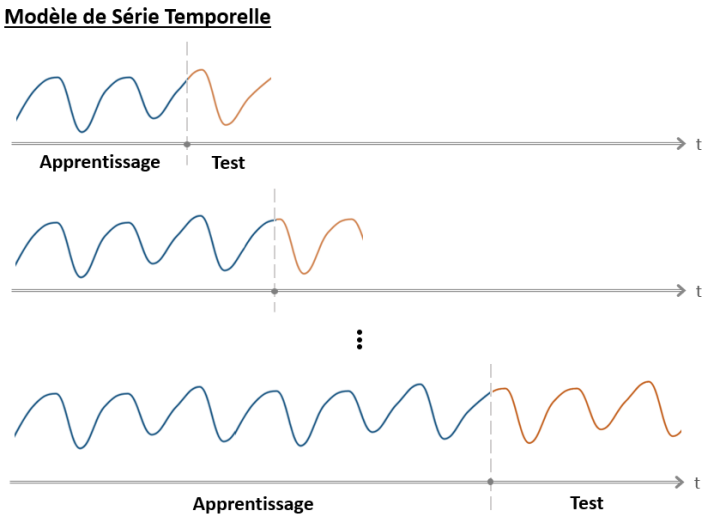


FIGURE 1.8 – Cross-Validation, cas des séries temporelles

Le caractère ordonné des données des séries temporelles empêche de réaliser des segments aléatoires. Afin d'être évalué k fois, la période d'entraînement initiale est réduite, et une période d'horizon est définie. Le modèle est entraîné sur les premières données observées, qui constituent la base d'apprentissage, puis projeté et testé sur la période d'horizon. Une partie de la période d'horizon peut ensuite être ajoutée aux données d'entraînement pour recommencer k fois.

1.2.2 Construire le modèle

▷ **Sélectionner les paramètres :**

Un modèle de *machine learning* est un algorithme qui définit un ensemble de décisions et de méthodes permettant d'apprendre à partir des données. Les paramètres du modèle définissent la manière dont il apprend, et peuvent changer significativement son comportement. Par exemple dans le cas de modèles d'arbres de décisions, la profondeur des arbres et le nombre de feuilles sont paramétrables.

Une configuration de ces paramètres non adaptée aux données d'entraînement peut mener à de faibles performances, ou au contraire à des performances trop élevées dans le cas de sur-apprentissage. D'un autre côté, il existe des combinaisons de paramètres qui conviennent mieux aux données, et qui permettront d'avoir de meilleures prédictions.

Il est donc important de bien comprendre chacun des paramètres lors de l'initialisation d'un modèle. Pour cela, rechercher la documentation d'un modèle est préférable.

▷ **Ajuster aux données d'entraînement :**

Avec une base de données d'entraînement et un ensemble de paramètres, le modèle peut commencer à apprendre à partir des données. Cette phase d'apprentissage dépend des modèles. Souvent déterministe, l'algorithme sous-jacent est suivi, dans le but de minimiser une fonction de coût, généralement basée sur l'écart entre les valeurs observées y et les valeurs prédites $\hat{y}$.

A l'issue de la phase d'apprentissage, le modèle peut être évalué sur les données de test.

1.2.3 Évaluer le modèle

▷ **Comprendre les métriques :**

Les projections du modèle sur les données de test sont nécessaires à son évaluation. Cette évaluation est réalisée à l'aide de métriques : des mesures objectives de l'erreur, quantifiant l'écart entre les prédictions $\hat{y}$ et les valeurs réelles y.

Pour les régressions, les métriques les plus courantes sont le coefficient de détermination R^2, l'erreur quadratique moyenne (*Mean Squared Error*, MSE) ou l'erreur absolue moyenne (*Mean Absolute Error*, MAE). L'erreur relative absolue (*Mean Absolute Percentage Error*, MAPE) est particulièrement intéressante pour son interprétation dans le cadre de séries temporelles. **Les quelques métriques utilisées pour ce mémoire sont décrites ci-dessous, à partir des notations introduites en début de partie.**

MSE & RMSE :
L'erreur quadratique moyenne (MSE) est la moyenne des carrés des erreurs :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Les erreurs sont élevées au carré : la MSE pénalise les erreurs importantes. C'est donc une métrique intéressante dans les situations où les valeurs éloignées font l'objet d'une attention particulière.

L'erreur quadratique moyenne de la racine (RMSE) est la racine carrée de la MSE :

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Tout comme la MSE, les erreurs importantes ont un poids conséquent. Cependant, les erreurs sont ensuite ramenées à la même échelle que les données, ce qui facilite l'interprétation par rapport à la MSE. La RMSE est donc une métrique interprétable, pénalisant les erreurs de prédictions importantes, utilisable dans n'importe quel cadre.

MAPE :
L'erreur relative absolue (MAPE) est la moyenne des valeurs absolues des erreurs relatives, exprimée en pourcentage :

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$$

La MAPE n'utilise pas de terme quadratique : les erreurs importantes ne sont pas pénalisées. C'est une métrique très facilement interprétable. Par exemple une MAPE de 10% signifie que les prédictions sont en moyenne à 10% des valeurs réelles. Dans un contexte de séries temporelles, suivre la MAPE dans le temps permet de visualiser l'erreur relative en prenant en compte la tendance et la saisonnalité.

MASE :
L'erreur absolue moyenne à l'échelle (*Mean Absolute Scaled Error*, MASE) est une métrique utilisée spécifiquement dans le contexte de séries temporelles, basée sur la MAE, et prenant en compte la saisonnalité dans son calcul :

$$\text{MASE} = \frac{\text{MAE}_{\text{model}}}{\text{MAE}_{\text{naïve}}} = \frac{\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\frac{1}{n-m}\sum_{i=m+1}^{n}|y_i - y_{i-m}|}$$

Le terme au numérateur représente l'erreur absolue moyenne (MAE) du modèle. Le terme au dénominateur représente la MAE d'un modèle naïf de saisonnalité m, c'est-à-dire un modèle qui reprend à chaque date t la valeur observée en $t - m$ pour réaliser sa prédiction.

Cette métrique est robuste aux valeurs aberrantes et ne dépend pas de l'échelle des données. De plus, elle n'est pas non plus sensible aux observations nulles ($y_i = 0$) comme le serait la MAPE. Elle permet de comparer facilement des modèles entre eux : une MASE de 1 signifie que le modèle performe aussi bien que le modèle naïf, et donc plus la MASE est faible, meilleur est le modèle. Cependant son calcul dépend de la période de saisonnalité, ce qui complique l'évaluation dans le cas de modèle avancé détectant plusieurs saisonnalités.

1.2.4 Sélectionner le modèle

▷ **L'hyperparamétrage :**

Un modèle avec un ensemble de paramètre est testé, plusieurs fois dans le cas de *cross-validation*, puis évalué à partir d'une métrique adaptée. Afin de déterminer si un modèle est "meilleur" qu'un autre, il faut pouvoir le comparer.

Pour cela, un nouveau modèle est ajusté avec un nouvel ensemble de paramètres, et toutes autres choses égales par ailleurs. Ce processus de recherche des paramètres optimum, couplé à la *cross-validation* et à l'évaluation quantitative des performances via une métrique est appelé hyperparamétrage. En pratique, cela est facilement automatisable dans des boucles en spécifiant la liste des paramètres à tester et toutes les combinaisons envisageables. Alors que la *cross-validation* permet d'éviter le sur-apprentissage, l'hyperparamétrage permet d'éviter le sous-apprentissage.

▷ **Le compromis biais-variance :**

Il existe deux sources d'erreurs dans la modélisation en *machine learning* : le biais et la variance.

Supposons qu'il existe une relation expliquant chaque observations y_i :

$$y_i = f(X_i) + \epsilon_i$$

où ϵ_i est le bruit, tel que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, et X_i représente les variables explicatives.

L'erreur de prédiction quadratique du modèle peut alors s'écrire de la façon suivante (démonstration en Annexe 1) :

$$\mathbb{E}[(y - \hat{y})^2] = \text{Biais}^2 + \text{Variance} + \sigma^2$$

Les termes de variance et de biais dépendent de la complexité du modèle :

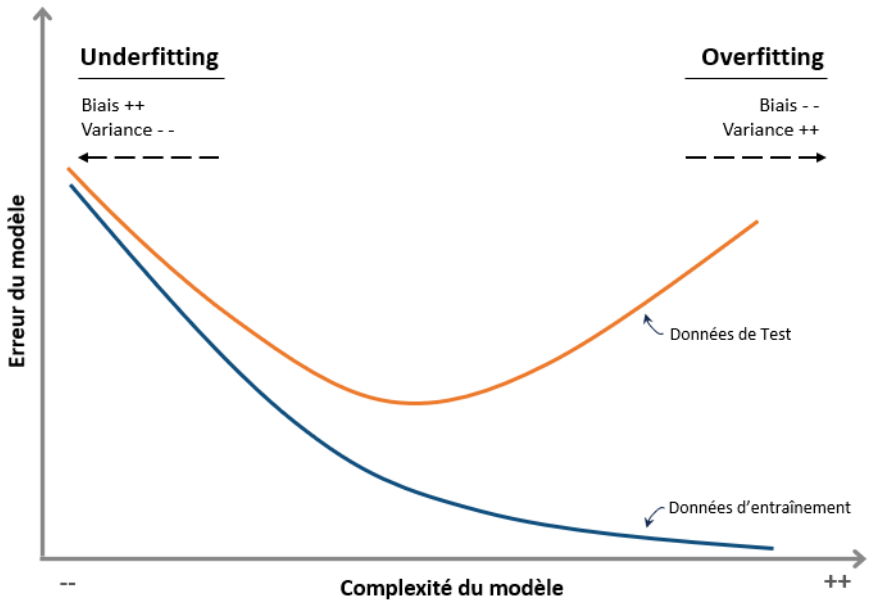
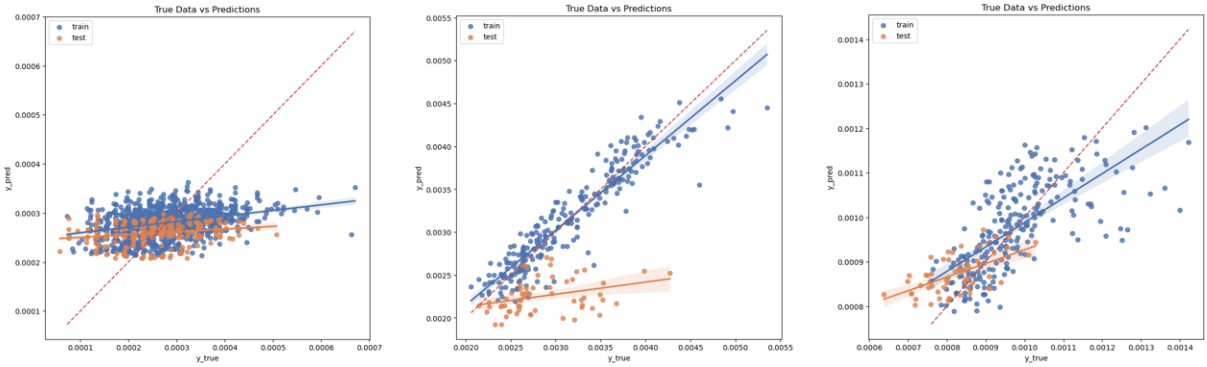


FIGURE 1.9 – Compromis biais-variance

La variance peut être due à un modèle trop complexe, en sur-apprentissage (*overfitting*). Le sur-apprentissage se produit lorsque le modèle s'adapte trop étroitement aux données d'entraînement. Le modèle ne capte pas seulement les interactions entre variables mais aussi le bruit ϵ_i imprévisible. Ce cas de figure peut être évité avec la *cross-validation*, en simplifiant le modèle ou en augmentant le volume de données d'entraînement.

Le biais peut être, à l'opposé, dû à un modèle trop simple, en sous-apprentissage (*underfitting*). Le sous-apprentissage se produit lorsque le modèle ne capture pas assez la complexité des variables explicatives. Autrement dit, les prédictions du modèle sont trop éloignées des observations. Ce cas de figure peut être évité avec l'hyperparamétrage, ou en complexifiant le modèle.

Dans ce mémoire, afin d'analyser la pertinence d'un modèle, des projections réalisées sur la période d'entraînement et sur la période de test seront comparées aux observations. Par exemple :



(a) Manque d'amplitude (b) *Overfitting* (c) Modèle acceptable

FIGURE 1.10 – Comparaison y_true contre y_pred sur les données de train-test

Ce graphe permet de valider visuellement si les prédictions sont proches des observations. Dans le cas d'un modèle oracle, c'est-à-dire qui n'a aucune erreur de prédiction, alors $y = \hat{y}$ et tous les points s'alignent sur la diagonale. Plus les points s'éloignent de la diagonale, moins les prédictions sont justes. Lorsque les points s'alignent horizontalement, alors les prédictions manquent d'amplitude, ce qui peut signifier une saisonnalité trop "lisse".

Tracer à la fois les projections sur l'entraînement (en bleu), et sur la période de test (orange) permet d'identifier le sur-apprentissage. Cela peut être le cas lorsque les projections sur l'entraînement s'alignent sur la diagonale mais que les projections sur la période de test s'en éloignent. Toutefois, coupler ce graphique à un tracé de la série temporelle est indispensable pour s'assurer de son interprétation. Par exemple, un changement de comportement sur la période de test, comme l'effet Covid, peut renvoyer un graphe similaire sans pour autant correspondre à un cas de sur-apprentissage.

Avec tous ces éléments, un modèle de *machine learning* pour les séries temporelles peut être mis en place, de la phase de préparation des données à la phase de sélection du modèle après hyperparamétrage.

Dans le cadre de ce mémoire, on s'intéresse non seulement aux projections des modèles mais également à la mesure de l'incertitude et à la création d'intervalles de prédictions. Certaines méthodes intrinsèques aux modèles permettent de quantifier cette incertitude, mais il faudra aussi parfois se ramener aux propriétés statistiques nécessaires à la construction d'intervalles.

1.3 L'évaluation de l'incertitude

L'incertitude est une composante inhérente de toute prédiction, en particulier dans le domaine du *machine learning*. Cette incertitude peut provenir de plusieurs sources et peut avoir des conséquences significatives, notamment dans des applications critiques comme l'imagerie médicale ou la sécurité des véhicules autonomes. Sans point d'attention sur ces incertitudes, les modèles prédictifs tendent à produire des intervalles de prédiction soit trop larges, soit trop étroits, ce qui peut entraîner des décisions erronées. Par exemple, des prédictions imprécises concernant la fréquence ou la gravité des sinistres peuvent entraîner une sous-estimation ou une surestimation des provisions nécessaires.

La mesure et la gestion de l'incertitude sont essentielles pour garantir la qualité des décisions basées sur les prédictions. Les prédictions sont souvent réalisées à partir de données incomplètes ou de modèles qui ne peuvent pas capturer entièrement la complexité du monde réel. **Cela peut inclure des erreurs dues à des mesures imprécises, des biais d'échantillonnage, des hypothèses fortes dans les modèles, les paramètres et l'évaluation des modèles, la variabilité naturelle des données ou encore des changements de comportements aléatoires.**

Pour améliorer la robustesse des prévisions et la crédibilité des conclusions, il semble approprié de fournir des intervalles associés aux prédictions. Ces intervalles permettent d'intégrer une partie de la variabilité des données et des erreurs de prédictions, ce qui doit refléter la prise en compte des incertitudes. **La construction d'intervalles apporte une quantification de l'incertitude en fournissant une plage de valeurs probables pour les paramètres ou projections.**

Différentes méthodes permettent de mesurer l'incertitude :
- la méthode "classique" fréquentiste, basé sur l'échantillon observé,
- l'approche bayésienne, qui consiste à mettre à jour les hypothèses initiales avec la distribution observée des données,
- la régression de quantile, qui se concentre directement sur la prédiction de quantiles.,
- la prédiction conforme : une méthode agnostique introduisant des scores de conformité,
- la méthode CQR (*Conformalized Quantile Regression*), couplant les 2 précédentes,
- des méthodes spécifiques aux modèles déployés. Ce sera le cas pour l'incertitude de la tendance du modèle Prophet, développée en partie 1.4.3.

La régression de quantile et la méthode CQR ne seront pas détaillées.

1.3.1 Mesurer l'incertitude

Dans le cadre de modèles en *machine learning*, il est crucial de distinguer deux concepts clés : l'intervalle de confiance et l'intervalle de prédiction. Ces deux types d'intervalles mesurent l'incertitude, mais servent des objectifs différents et fournissent des informations distinctes.

L'intervalle de confiance :
Un intervalle de confiance donne une plage de valeurs qui, avec un certain niveau de probabilité, doit contenir la véritable valeur d'une valeur estimée, telle qu'une moyenne ou un coefficient de régression. Dans le cadre étudié, il permet notamment d'évaluer l'incertitude associée à l'estimation d'un paramètre d'un modèle ou d'une statistique mesurée. Par exemple, un intervalle de confiance pour la moyenne d'une population indique l'étendue dans laquelle la moyenne réelle de cette population est estimée se situer avec un certain niveau de confiance (par exemple, 95%). L'intervalle de confiance prend en compte la variance due à l'échantillon dans l'estimation de la variable d'intérêt.

L'intervalle de prédiction :

Un intervalle de prédiction concerne une observation individuelle inconnue. Il s'agit d'une plage de valeurs dans laquelle une observation inconnue est estimée se situer avec une certaine probabilité. Cet intervalle évalue l'incertitude concernant une observation spécifique, en tenant compte non seulement de la variabilité de l'échantillon, inhérente aux données utilisées pour l'estimation, mais aussi de la variance introduite par le modèle prédictif lui-même. Il est donc par définition plus large que l'intervalle de confiance.

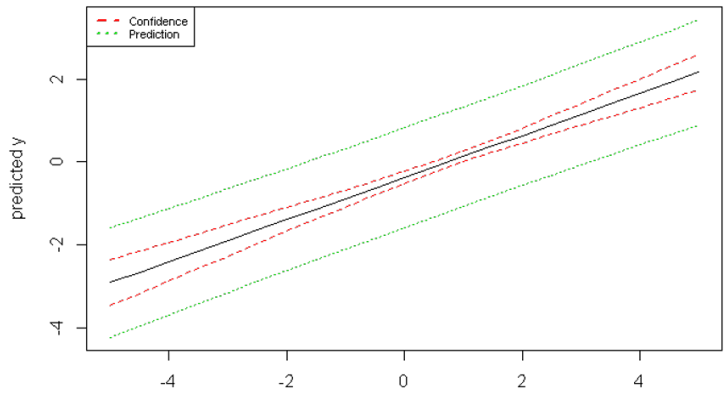


FIGURE 1.11 – Exemple d'intervalles de confiance et de prédiction (Source : The Uncertainty of Predictions, C. Stevenson)

Pour résumer :

— l'intervalle de confiance est utilisé pour estimer la précision des paramètres du modèle. Il fourni une mesure de l'incertitude liées aux données.
— l'intervalle de prédiction est utilisé pour estimer la variabilité des observations individuelles inconnues. Il fournit une mesure de l'incertitude liées aux données et au modèle.

L'incertitude à mesurer et à suivre lors de projections de séries temporelles est donc l'intervalle de prédiction, qui donnera une estimation des trajectoires envisageables pour un niveau de confiance défini, généralement 95%.

Les métriques pour l'incertitude :

Tout comme pour les prédictions, la pertinence des intervalles peut être évaluée. Un intervalle à un niveau de confiance de 95% peut possiblement être respecté pour seulement 70% des prédictions. Un autre intervalle à 95% comprendra à coup sûr la valeur souhaitée s'il couvre l'étendue des valeurs possibles. Ces deux exemples introduisent les notions de couverture et de largeur.

La couverture (*coverage*) mesure la proportion de y_{true} bien compris dans l'intervalle de prédiction. Pour un échantillon de test de taille n, avec un intervalle $[y_{lower}\,;\,y_{upper}]$, la couverture vaut :

$$\textit{Coverage} = \frac{\text{Nombre de y\_true dans l'intervalle}}{\text{Nombre total de y\_true}} = \frac{\sum_{i=1}^{n} \mathbb{1}(y_{\text{true},i} \in [y_{\text{lower},i}; y_{\text{upper},i}])}{n}$$

Pour évaluer la pertinence d'un intervalle, il est important de vérifier si cette couverture est proche du niveau de confiance spécifié. **Contrairement aux différents types de modèles, où il existe plusieurs métriques plus ou moins adaptées aux différents cas de figure rencontrés, la couverture est toujours la métrique la plus importante pour l'évaluation de la pertinence d'intervalles.**

Cette métrique sera toutefois souvent complétée par la largeur de l'intervalle (*width*). La largeur de l'intervalle mesure l'étendue de l'intervalle autour de l'estimation du paramètre :

$$\textit{Width} = \text{Limite Supérieure} - \text{Limite Inférieure}$$

Une largeur trop grande peut indiquer un manque de précision dans l'estimation, tandis qu'une largeur trop petite peut suggérer une sous-estimation de l'incertitude. La largeur de l'intervalle est donc une indication directe de la précision et de la fiabilité de l'estimation.

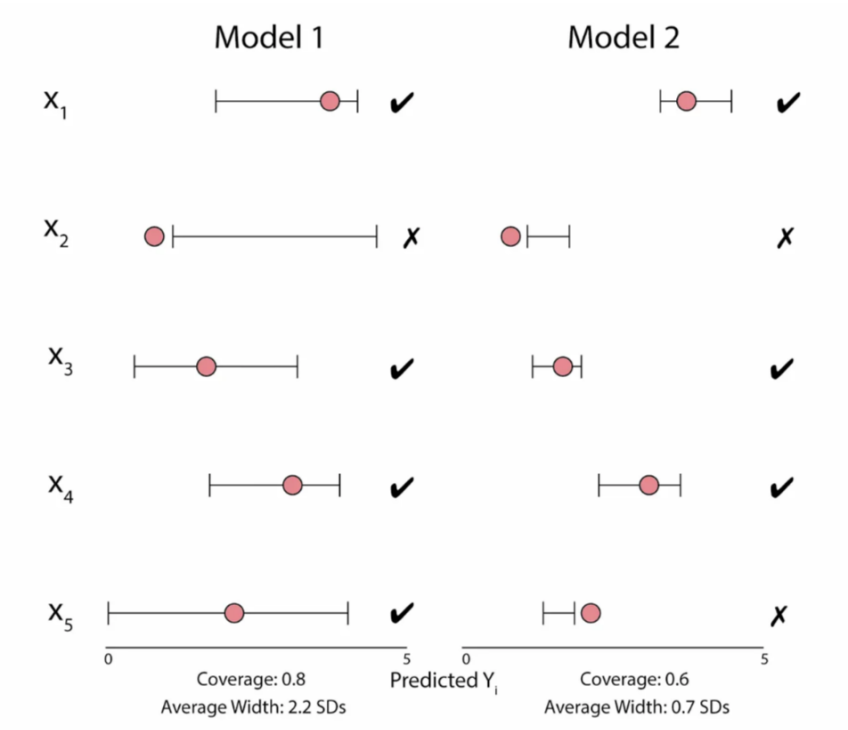


FIGURE 1.12 – Exemple de calcul de *Coverage* et *Width* (Source : How to evaluate Probabilistic Forecasts, V. Manokhin)

Ces deux mesures fournissent des informations importantes sur l'efficacité et la fiabilité des intervalles de prédiction dans la gestion de l'incertitude. Dans le cas de séries temporelles, conserver une couverture stable sur le long terme entraînera généralement une augmentation de la largeur d'intervalle.

1.3.2 Méthode fréquentiste

En statistique fréquentiste, une probabilité est interprétée comme la fréquence relative d'un événement si une expérience pouvait être répétée un nombre infini de fois. Contrairement à d'autres approches, les paramètres ne possèdent pas de distribution propre et leur valeur est constante, bien que non observée directement. L'incertitude est donc mesurée en considérant les paramètres du modèle comme des quantités fixes mais inconnues. **La statistique fréquentiste correspond à une vision déterministe des paramètres**.

Dans la méthode fréquentiste, les intervalles de confiance et de prédiction sont construits pour évaluer l'incertitude associée aux estimations basées sur les observations. Les analyses sont centrées sur les échantillons de données observés. Les paramètres du modèle sont estimés en utilisant des statistiques descriptives telles que la moyenne et la variance, et les erreurs sont minimisées par des techniques comme les moindres carrés ou la méthode du maximum de vraisemblance. Les intervalles au niveau de confiance $1-\alpha$ prennent donc la forme suivante :

$$ITV_{1-\alpha} = \hat{\theta} \pm z_{\alpha/2} \cdot SE(\hat{\theta}) \tag{1.11}$$

où :
— $\hat{\theta}$ est l'estimation du paramètre ou la prédiction réalisée,
— $z_{\alpha/2}$ est le quantile pour un niveau de confiance $1-\alpha$ d'une distribution de référence, généralement normale ou de Student,
— $SE(\hat{\theta})$ est l'erreur standard de $\hat{\theta}$. Ce terme permet de prendre en compte la variabilité due à l'échantillon dans le cas d'intervalle de confiance, mais également la variance due au modèle pour les intervalles de prédiction.

Ces calculs déterministes permettent d'avoir une estimation simple des intervalles de prédiction, mais généralement ne vérifient pas la couverture au niveau $1-\alpha$.

1.3.3 Approche bayésienne

La vision fréquentiste et la vision bayésienne représentent deux approches fondamentales en statistique pour l'inférence et l'analyse des données. En statistique bayésienne, la probabilité est interprétée comme un degré de croyance ou une mesure de la confiance en un événement. Cette interprétation est subjective et peut être mise à jour avec de nouvelles informations. **La statistique bayésienne correspond à une vision stochastique des paramètres**.

Les méthodes bayésiennes impliquent l'utilisation de connaissances sur les données pour supposer une distribution a priori d'un paramètre, représentant les croyances. Des données observées, ou données antérieures, sont également indispensables pour mettre à jour les croyances initiales. Une distribution postérieure est calculée à partir de la formule de base de l'inférence bayésienne :

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} \tag{1.12}$$

où :
— $P(\theta)$ est la probabilité *a priori* : la probabilité initialement supposée des paramètres θ avant d'observer les données. Les paramètres sont vus comme des variables aléatoires.

— $P(D|\theta)$ est la vraisemblance, correspondant à la probabilité d'observer les données D étant donné les paramètres θ.

— $P(D)$ est la vraisemblance marginale, servant de facteur de normalisation.

— $P(\theta|D)$ est la probabilité postérieure, c'est-à-dire la probabilité des paramètres θ sachant les données D. La distribution postérieure représente la distribution supposée des paramètres après mise à jour grâce aux données observées. Elle n'est généralement pas calculable directement et demande d'utiliser des méthodes d'échantillonage, comme la méthode MCMC (Markov Chain Monte-Carlo).

L'incertitude sur ce paramètre est ensuite capturée à partir des échantillons, permettant de construire les distributions postérieures. L'ensemble des échantillons est projeté, ce qui permet de calculer des quantiles $q_{1-\frac{\alpha}{2}}$. Un intervalle de crédibilité est enfin construit à partir de ces quantiles.

L'intervalle de crédibilité quantifie la probabilité que le paramètre étudié se trouve dans une plage de données, à partir des croyances a priori.

L'approche fréquentiste, correspondant à une vision déterministe des paramètres, permet d'obtenir l'estimation MAP (Maximum A Priori). Cette estimation MAP est égale au mode (observation la plus fréquente) de la distribution postérieure obtenue avec l'approche bayésienne, correspondant à une vision stochastique.

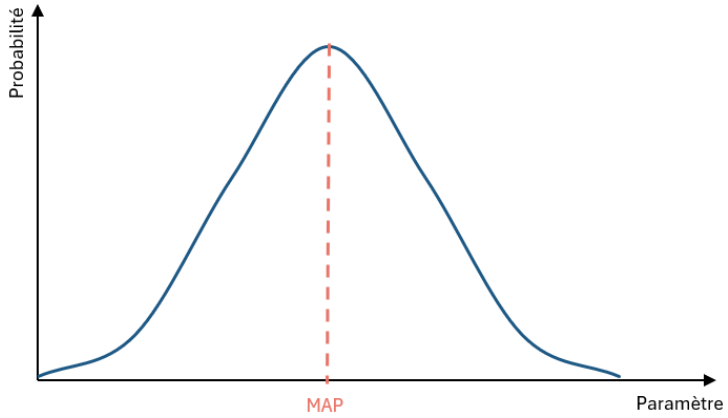


FIGURE 1.13 – Exemple de distribution postérieur et de l'estimation MAP

L'intervalle de crédibilité se rapproche donc d'un intervalle de confiance, et non d'un intervalle de prédiction. Il permet d'obtenir des meilleurs intervalles de confiance que la méthode fréquentiste, et est notamment à privilégier pour les modèles paramétriques.

Dans la pratique, les hypothèses spécifiées peuvent ne pas être valables, ce qui limite la pertinence des méthodes bayésiennes. Les intervalles obtenus peuvent donc ne pas respecter dans tous les cas le niveau de confiance ou la couverture revendiquée.

1.3.4 *Conformal Prediction*

La prédiction conforme permet d'obtenir des mesures d'incertitude pour des prédictions individuelles. Toutefois, contrairement aux méthodes précédentes, les intervalles renvoyés sont valides, c'est-à-dire qu'ils couvrent forcément le niveau de confiance renseigné au moment de la construction.

De plus, le concept de cette méthode est très simple et son implémentation, peu coûteuse en calculs, est agnostique, signifiant qu'elle peut être appliqué à n'importe quel modèle déjà entraîné.

Lorsque la méthode est intégrée au modèle, la base de données initiale est d'abord séparée en trois, suivant le même principe que la séparation train-test présenté en figure 1.6. Des jeux de données train-calibration-test, représentant respectivement 70%-20%-10% approximativement de la base de départ sont créés.

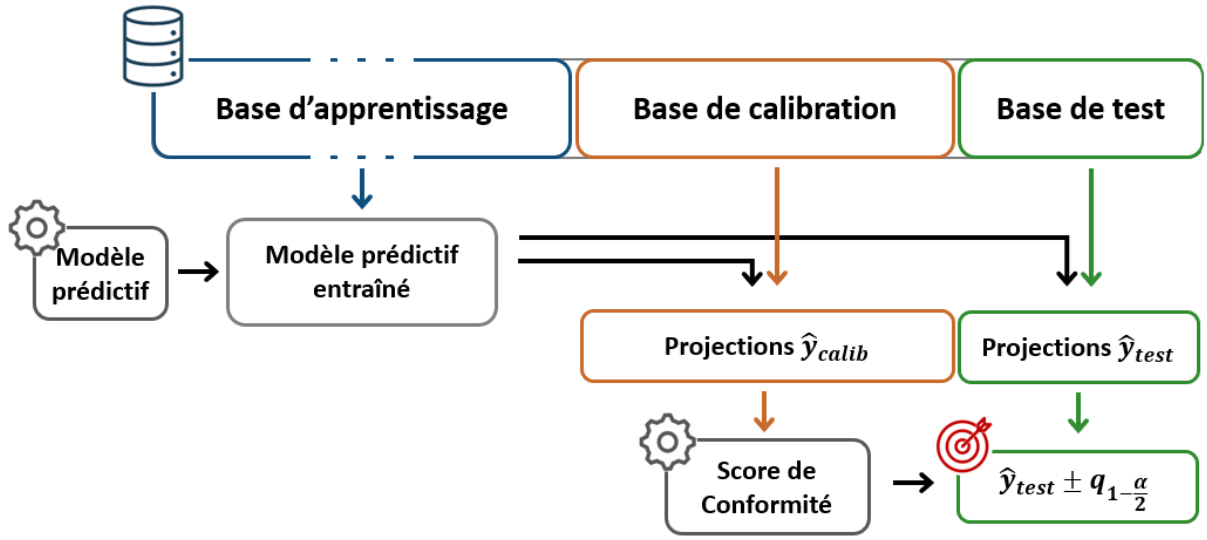


FIGURE 1.14 – Construction d'un intervalle de prédiction avec une méthode de *Conformal Prediction*

Le modèle de prédiction est d'abord ajusté sur la base d'entraînement. Des prédictions peuvent alors être réalisées sur la base de calibration pour obtenir $\hat{y}_{calib}$. De même, des prédictions sur la base de test donnent $\hat{y}_{test}$.

La méthode de prédiction conforme demande de choisir un score de conformité C_i, qui permettra de mesurer l'erreur de prédiction individuelle sur la base de calibration. Dans le cas de la méthode naïve, les scores de conformités sont :

$$C_i = |y_i - \hat{y}_{calib,i}|$$

Les calculs de score de conformité permettent d'obtenir une distribution des erreurs de prédiction, à partir de laquelle il est possible d'estimer les quantiles $q_{1-\frac{\alpha}{2},i}$.

Il est enfin possible d'assembler les prédictions $\hat{y}_{test}$ avec les quantiles $q_{1-\frac{\alpha}{2},i}$ pour obtenir un intervalle de prédiction.

Les métriques pour les intervalles peuvent être vérifiées. Par construction, la couverture sera forcément validée sur la base de calibration.

Lorsque la méthode est appliqué à un modèle déjà entraîné, la calibration peut se faire sur un segment de la base d'apprentissage.

Cette méthode agnostique ne demande pas d'hypothèse préalable sur la distribution des données, ne prend en compte aucun paramètre et garantit la validité théorique des intervalles à la construction, quel que soit le nombre d'observations.

1.4 Le modèle Prophet

Prophet est un modèle de séries temporelles développé par une équipe de *data-scientists* de Facebook. Présenté dans leur article "*Forecasting at scale*" [2], le modèle a pour objectif d'apporter une alternative aux modèles plus classiques de séries temporelles, tels que SARIMA, tout en étant accessible à des analystes, experts de leur domaine, et non seulement à des data-scientists, experts des modèles de projections. La librairie associée est disponible sur R et Python depuis mars 2017, et est régulièrement mise à jour depuis. Son implémentation a été réalisée avec Python dans cette étude, comme pour la majorité des travaux présentés.

Les créateurs du modèle le présentent ainsi : "*Prophet est une procédure de projection des séries temporelles basée sur un modèle additif dans lequel les tendances non linéaires sont ajustées avec une saisonnalité annuelle, hebdomadaire et quotidienne, avec les effets de jours fériés et de régresseurs externes. Il fonctionne mieux avec des séries temporelles ayant de forts effets saisonniers et plusieurs saisons de données historiques. Prophet est robuste aux données manquantes et aux changements de tendance, et gère généralement bien les valeurs aberrantes.*" [3]

L'équation du modèle additif derrière Prophet est la suivante :

$$y(t) = g(t) + s(t) + h(t) + \sum_{i=1}^{k} \beta_i x_i(t) + \epsilon_t \tag{1.13}$$

où :
— $g(t)$ est le terme de tendance.
— $s(t)$ est le terme de saisonnalité.
— $h(t)$ est le terme représentant les jours fériés.
— β_i sont les coefficients des régresseurs externes.
— $x_i(t)$ sont les régresseurs externes.
— ϵ_t est le terme d'erreur, supposé normalement distribué.

Le modèle a été construit avec une vision orientée *business*, de façon à pouvoir être interprétable par des experts métiers et enrichi par leurs connaissances, dans un processus d'*analyst-in-the-loop*. Un modèle performant est facilement réalisable, grâce à la combinaison de l'analyse statistique traditionnelle aux visualisations des effets individuels des composantes du modèle.

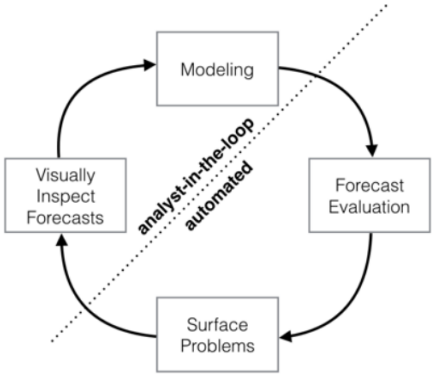


FIGURE 1.15 – Processus *analyst-in-the-loop* (Source : Forecasting at scale, S. J. Taylor & B. Letham, 2017

2. Forecasting at scale, S. J. Taylor & B. Letham, 2017
3. https://facebook.github.io/prophet/

1.4.1 Pourquoi Prophet ?

▷ **Les spécificités de Prophet :**

L'intégration de régresseurs externes aux prédictions est une caractéristique qui peut rendre le modèle Prophet particulièrement intéressant. Un régresseur externe $x_i(t)$ peut être n'importe quelle variable additionnelle ayant la forme d'une série temporelle, qui est censée influencer la série temporelle d'étude $y(t)$. Ces régresseurs pourront être utilisés pour ajouter aux prédictions l'influence de séries temporelles parallèles telles que des projections de facteurs économiques ou des indicateurs de marché. **Dans le cadre de ce mémoire, l'ajout de mesures et de projections de température, d'humidité et de précipitation permettra de quantifier l'impact du climat sur la mortalité.**

Sa forme matricielle permet d'introduire le terme $\mathbf{X}(t)$ représentant le vecteur de variable explicative, l'équation du modèle équivaut alors à :

$$y(t) = g(t) + s(t) + h(t) + \boldsymbol{\beta}\mathbf{X}(t) + \epsilon_t \tag{1.14}$$

où :
— $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_k]$ est le vecteur des coefficients des régresseurs externes.
— $\mathbf{X}(t) = [x_1(t), x_2(t), \ldots, x_k(t)]^\intercal$ est le vecteur des régresseurs externes à l'instant t. Dans le cadre classique des modèles de machine learning, ce vecteur représente les variables explicatives.

Étant un modèle additif, dans lequel chaque terme est traité simultanément, Prophet ne requiert pas que la série temporelle d'étude soit stationnaire avant de l'analyser. L'éventuelle non-stationnarité est captée de manière flexible à la fois par le terme de saisonnalité, qui permet d'avoir plusieurs saisonnalités pour une même série, et à la fois par le terme de tendance, qui peut être ajusté dynamiquement via des tendances linéaires par morceaux ou via une tendance évolutive saturée. Les détails de ces termes seront présentés en partie 1.4.2.

Prophet se présente alors comme un modèle intermédiaire entre les modèles de séries temporelles traditionnels, qui se basent uniquement sur les observations passées pour prédire le futur, et des modèles de machine learning, qui se basent uniquement sur des variables explicatives extérieures pour leurs prédictions.

▷ **Comparaison à d'autres modèles :**

D'autres modèles permettant de projeter des séries temporelles tout en prenant en compte des informations extérieures existent mais n'ont pas été conservés pour cette étude.

Le modèle SARIMAX :
SARIMAX (*Seasonal AutoRegressive Integrated Moving Average with eXogenous factors*) est une extension du modèle ARIMA, qui inclut des effets saisonniers et des variables explicatives externes.

Comme son nom l'indique, SARIMA est un modèle basé directement sur la théorie des séries temporelles, avec l'intégration de termes auto-régressifs, de *moving average*, et de différenciation pour la tendance et la saisonnalité, comme vu en partie 1.1.2.
L'unique paramètre de saisonnalité s doit être spécifié avec des connaissances sur la série étudiée ou après analyses des auto-corrélations. Le modèle SARIMA a donc des difficultés à modéliser correctement des saisonnalités plus complexes.
Par exemple, une série de données quotidiennes pour laquelle un paramètre de saisonnalité égal à 30 aurait été choisi, indique un cycle mensuel. Cependant, si la même série de données

présente également un cycle hebdomadaire, un paramètre de saisonnalité égale à 7 serait aussi pertinent. La prise en compte de plusieurs saisonnalités est impossible avec ce modèle.

Les variables exogènes sont ensuite ajoutées au modèle SARIMA comme une composante additionnelle linéaire, similairement au modèle Prophet :

$$y_t = SARIMA(y_t) + \boldsymbol{\beta}\mathbf{X}(t)$$

L'intervalle de confiance calculé par SARIMAX pour une prévision $\hat{y}_t$ à un niveau de confiance $(1-\alpha)$ est donné par des propriétés statistiques de l'approche fréquentiste (cf partie 1.3.2) :

$$[\hat{y}_t - z_{\alpha/2} \cdot \hat{\sigma}_t; \hat{y}_t + z_{\alpha/2} \cdot \hat{\sigma}_t]$$

où :
— $\hat{y}_t$ est la prévision moyenne pour la période t.
— $z_{\alpha/2}$ est la valeur critique de la distribution normale pour un niveau de confiance $(1-\alpha)$.
— $\hat{\sigma}_t$ est l'erreur standard de la prévision pour la période t.

L'intervalle en découlant est conséquent, étant donné la prise en compte des incertitudes associées aux nombreux paramètres estimés du modèle et à la variance des résidus dans le calcul de l'erreur standard.

SARIMAX est un des modèles classiques de série temporelle les plus complets, mais il est peu flexible dans l'estimation de ses paramètres, et demande beaucoup de connaissances mathématiques pour une implémentation juste.

Le modèle DeepAR :
DeepAR est un modèle de séries temporelles développé par une équipe de *data-scientists* d'Amazon. Présenté pour la première fois en avril 2017, juste après le lancement de Prophet, il est mis à disposition en open-source en 2019 pour Python [4]. Ce modèle de Deep Learning repose sur une architecture de réseau neuronal récurrent (RNN) basé sur des cellules Long Short-Term Memory (LSTM). Ces cellules apprennent des distributions de probabilités de plusieurs séries à chaque pas de temps. Le modèle renvoie donc des projections de distributions de probabilités, ce qui permet d'avoir des intervalles de confiance pertinents immédiatement.

DeepAR est capable d'entraîner un unique modèle sur un ensemble de séries temporelles liées entre elles, permettant ainsi d'apprendre des paternes communs à travers ces séries. Cette capacité à gérer plusieurs séries temporelles simultanément est particulièrement intéressante dans le cadre de ce mémoire, car cela pourrait permettre d'avoir un seul modèle entraîné sur les 96 départements de France métropolitaine, et non 96 modèles à entraîner.

Comme Prophet, DeepAR permet l'ajout de régresseurs externes et de variables spécifiques comme les jours fériés pour affiner les prédictions. En revanche, n'étant pas un modèle traitant les données de manière ordonnée, DeepAR crée automatiquement des séries temporelles additionnelles en fonction de la granularité des données cibles, comme les jours du mois ou les jours de l'année, ce qui aide le modèle à apprendre des paternes temporels spécifiques. Ces séries temporelles additionnelles peuvent être traitées comme des variables explicatives temporelles, au même titre que les variables climatiques.

DeepAR est un modèle de Deep Learning adapté aux séries temporelles, parmi les plus performants. Spécialement conçu pour des scénarios complexes impliquant de nombreuses séries temporelles, ce modèle de prévisions probabilistes peut présenter une certaine variabilité dans les résultats. De plus, n'étant pas un modèle additif, ses résultats sont difficilement interprétables.

4. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks, D. Salinas, V. Flunkert & J. Gasthaus, 2019

1.4.2 L'équation du modèle

Pour rappel, l'équation du modèle est décomposée en différentes composantes dépendantes du temps :

$$y(t) = g(t) + s(t) + h(t) + \sum_{i=1}^{k} \beta_i x_i(t) + \epsilon_t$$

Les formules présentées reprennent les notations introduites dans l'article de présentation du modèle "Forecasting at scale" [5].

Le modèle ajuste un ensemble de fonctions linéaires et non-linéaires pour ses différents termes, à la façon d'un modèle additif généralisé (GAM), dans lequel le temps serait un régresseur. Prophet approche la problématique des projections de séries temporelles comme un exercice d'ajustement à des distributions (*curve-fitting*).

▷ **Le terme de tendance :**

L'objectif de la modélisation de la tendance est de comprendre la manière dont a évolué une population d'étude dans le temps de manière macroscopique, pour ensuite pouvoir poursuivre cette évolution dans le futur.

Pour correspondre aux différents cas observables dans la réalité, la tendance $g(t)$ du modèle peut être implémentée de deux manières différentes :

— une tendance non-linéaire, de croissance logistique saturée, évolutive dans le temps,
— une tendance linéaire par morceaux, avec sélection automatique des points de rupture.

Le modèle de croissance est à choisir lors de l'initialisation du modèle, en fonction des connaissances métiers de la série étudiée, ou après une phase d'hyperparamétrage ayant permis d'identifier celle s'ajustant au mieux aux observations de test.

La tendance logistique saturée :

Dans la nature, la croissance d'une population au sein d'un écosystème est généralement assimilable à un modèle non-linéaire de croissance qui sature à un seuil de capacité C, représentant la population maximale.

Ce type de croissance est représenté dans sa forme la plus basique par un modèle de croissance logistique :

$$g(t) = \frac{C}{1 + \exp(-k(t - m))}$$

où :

— C est la capacité maximale. Elle correspond au plateau supérieur de la série temporelle, atteint asymptotiquement à long terme. Par exemple, dans le cas des études réalisées par les équipes de Facebook, ce seuil peut être égal au nombre de personnes ayant un compte.

— k est le taux de croissance, contrôlant la rapidité avec laquelle la série temporelle approche de la capacité C. Un k élevé signifie que la série atteindra rapidement sa capacité maximale, tandis qu'une valeur de k plus faible indique une croissance plus lente.

— m est une valeur de décalage temporel, représentant le moment où la moitié de la capacité C a été atteinte.

5. Forecasting at scale, S. J. Taylor & B. Letham, 2017

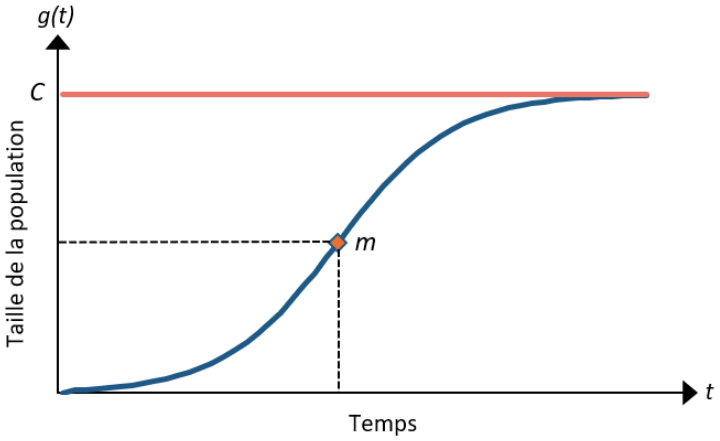


FIGURE 1.16 – Croissance logistique basique

Cette forme présente toutefois deux problèmes majeurs, qui peuvent être corrigés en augmentant la dépendance temporelle dans la modélisation : ni le seuil de capacité C, ni le taux de croissance k ne sont dépendants du temps.

La capacité maximale de la population peut évoluer au cours du temps, le paramètre C est donc remplacé par une capacité évolutive $C(t)$. En effet, pour reprendre l'exemple des équipes de Facebook, le nombre de personnes ayant un compte Facebook est croissant avec le temps, par conséquent faire l'hypothèse d'une capacité constante C peut être problématique sur un horizon lointain. En pratique, la capacité est alors une série temporelle à définir en parallèle de la série d'étude, pour associer une valeur de $C(t)$ à chaque $y(t)$ sur la période d'entraînement, et nécessite de définir des valeurs estimées de $C(t)$ sur la période de projection.

Le taux de croissance peut également être amené à changer au cours du temps en fonction de l'évolution de facteurs externes, le paramètre k est donc remplacé par un taux de croissance $k(t)$ ajusté à différents points de rupture. Dans l'exemple suivi, le nombre de nouveaux comptes a probablement augmenté plus vite avec le développement des smartphones, puis a ralenti avec l'arrivée d'autres réseaux sociaux. Un certain nombre de points de rupture peuvent alors être automatiquement placés à intervalles réguliers, représentant les dates auxquelles la tendance peut changer.

Supposons qu'il y ait S points de rupture aux temps s_j, $j \in [1, S]$. On appelle δ_j les ajustements du taux de croissance initial k. Ainsi, le taux de croissance en t vaut $k + \sum_{j|t>s_j} \delta_j$:

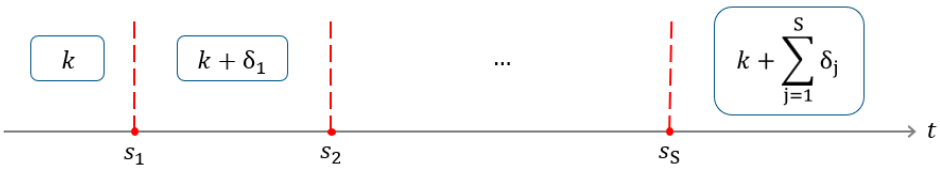


FIGURE 1.17 – Valeur du taux de croissance dans le temps

Un vecteur des ajustements du taux $\boldsymbol{\delta} \in \mathbb{R}^S$ est défini, ainsi qu'un vecteur d'indicatrices $\boldsymbol{a}(t) \in \{0;1\}^S$ tel que :

$$a_j(t) = \begin{cases} 1, & \text{if } t \geq s_j, \\ 0, & \text{sinon} \end{cases} \tag{1.15}$$

Ainsi, le taux de croissance linéaire par morceaux du modèle de tendance logistique saturée est défini par :

$$k(t) = k + \boldsymbol{a}(t)^\intercal \boldsymbol{\delta}$$

La valeur de décalage m doit par conséquent être également corrigée pour être en cohérence avec les différents changements de taux de croissance. Un vecteur des ajustements du paramètre de décalage temporel $\boldsymbol{\gamma} \in \mathbb{R}^S$ doit être défini de la façon suivante pour conserver la continuité de la tendance :

$$\gamma_j = \left(s_j - m - \sum_{l<j} \gamma_l \right) \left(1 - \frac{k + \sum_{l<j} \delta_l}{k + \sum_{l \leq j} \delta_l} \right)$$

Ainsi, le décalage temporel ajusté pour correspondre aux changements de taux de croissance vaut :

$$m(t) = m + \boldsymbol{a}(t)^\intercal \boldsymbol{\gamma}$$

Enfin, le modèle amélioré de croissance logistique saturée utilisé par Prophet est défini par l'équation :

$$g(t) = \frac{C(t)}{1 + \exp(-(k + \boldsymbol{a}(t)^\intercal \boldsymbol{\delta})(t - (m + \boldsymbol{a}(t)^\intercal \boldsymbol{\gamma})))} \tag{1.16}$$

Ce type de tendance est typiquement adapté aux séries temporelles ne présentant pas une tendance linéaire.

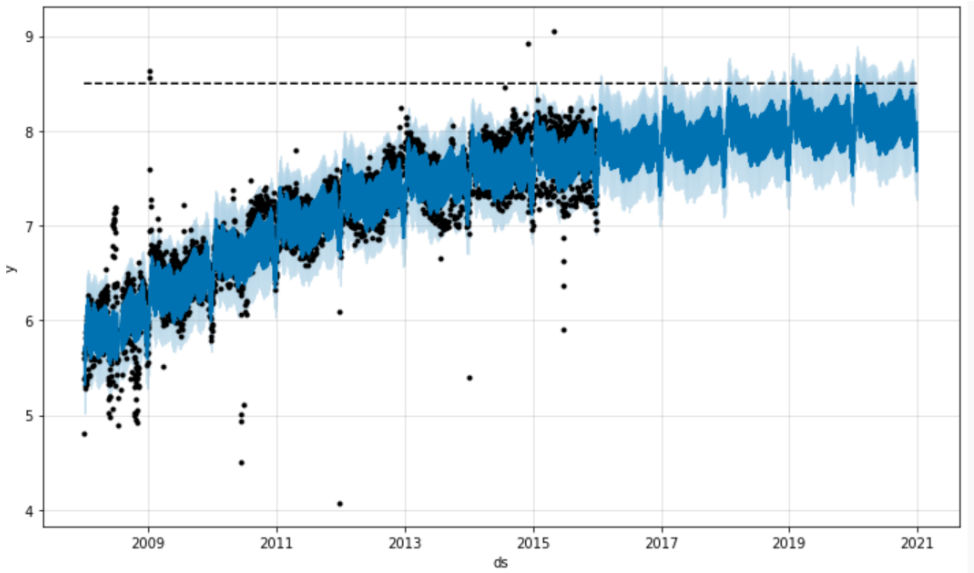


FIGURE 1.18 – Exemple de série temporelle avec une tendance logistique (Source : https://facebook.github.io/prophet/docs/saturating_forecasts.html)

La tendance linéaire par morceaux :

Pour les séries temporelles présentant une croissance linéaire, Prophet propose une implémentation similaire. Un modèle de croissance linéaire basique, tel que :

$$g(t) = kt + m$$

est adapté afin de proposer un taux de croissance dépendant du temps, de la façon suivante :

$$g(t) = (k + \boldsymbol{a}(t)^\intercal \boldsymbol{\delta})t + (m + \boldsymbol{a}(t)^\intercal \boldsymbol{\gamma}) \tag{1.17}$$

où :
— k, $\boldsymbol{a}(t)$, $\boldsymbol{\delta}$ et m sont identiques à ceux du modèle de croissance logistique,
— $\boldsymbol{\gamma}$, le vecteur des ajustements de m est défini par $\gamma_j = -s_j \delta_j$, afin de conserver la continuité.

Il en résulte un modèle de tendance linéaire par morceaux, modifié à certains points de ruptures.

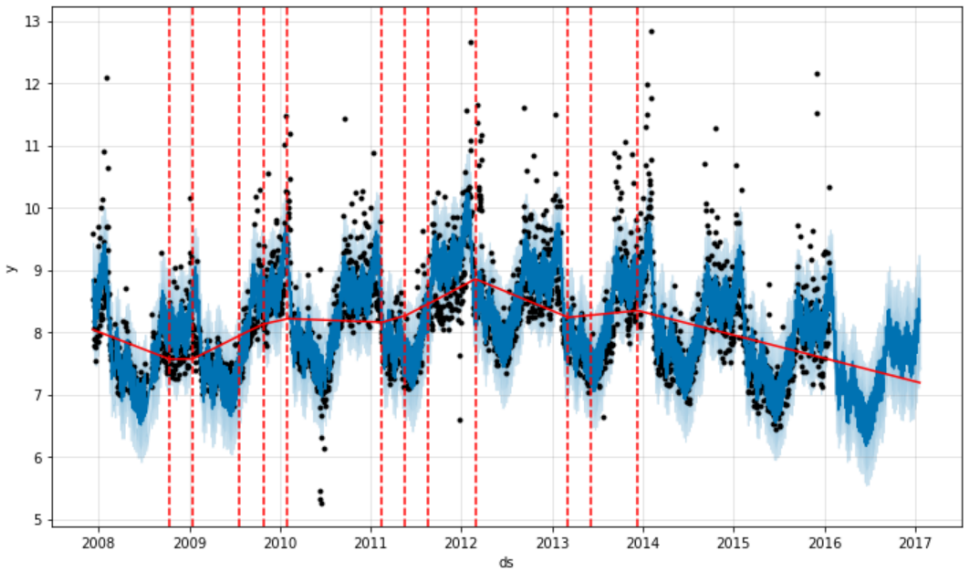


FIGURE 1.19 – Exemple de série temporelle avec une tendance linéaire par morceaux (Source : https://facebook.github.io/prophet/docs/trend_changepoints.html)

▷ **Le terme de saisonnalité :**

La saisonnalité représente les variations périodiques détectables dans les données observées. Modéliser la saisonnalité est essentiel afin de capter les schémas récurrents et reproductibles, pour ensuite pouvoir les poursuivre dans le futur.

Le modèle de saisonnalité de Prophet est conçu de façon à permettre l'intégration de plusieurs saisonnalités de différentes périodicités, de manière additive ou multiplicative.

Chaque terme est associé à une série de Fourier, d'ordre et de périodicité paramétrables, défini de la manière suivante :

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right) \tag{1.18}$$

où :
— N est l'ordre de la série de Fourier, qui détermine le nombre de termes dans la somme. Plus l'ordre est important et plus la série sera capable de s'ajuster à des paternes complexes, mais plus le risque de sur-apprentissage augmentera.

— P est la périodicité de la fonction, c'est-à-dire la période du signal. Par exemple, dans le cas de données quotidiennes, une valeur de P égale à $365,25$ représente une périodicité annuelle.

— a_n et b_n sont les coefficients de Fourier, qui déterminent l'amplitude des composantes cosinus et sinus respectivement.

L'introduction des $2N$ coefficients de Fourier nécessite d'utiliser des notations matricielles, tels que le vecteur des coefficients, supposé suivre une loi normale centrée :

$$\mathbf{\Pi} = [a_1, b_1, \ldots, a_N, b_N]$$

et le vecteur des saisonnalités :

$$\boldsymbol{Y}(t) = [\cos(\frac{2\pi(1)t}{P}), \sin(\frac{2\pi(1)t}{P}), \ldots, \cos(\frac{2\pi(N)t}{P}), \sin(\frac{2\pi(N)t}{P})]^\intercal$$

Ainsi, chaque saisonnalité est définie par :

$$s(t) = \mathbf{\Pi}\boldsymbol{Y}(t)$$

Par défaut, le modèle propose d'ajouter deux termes de saisonnalité :
— saisonnalité annuelle, correspondant aux paramètres $P = 365,25$ et $N = 10$,
— saisonnalité mensuelle, correspondant aux paramètres $P = 30,5$ et $N = 3$.
D'autres termes peuvent être ajoutés, en spécifiant l'ordre de la série de Fourier N et la périodicité P.

Les saisonnalités sont supposées additives, c'est-à-dire indépendantes de la tendance, mais spécifier qu'une tendance est multiplicative est possible, ce qui signifie que l'amplitude des paternes varie en suivant la tendance. En considérant uniquement les termes de tendance et saisonnalité de la série temporelle y, cela revient à considérer que le modèle suit l'équation suivante :
— $y(t) = g(t) + s(t)$, pour une saisonnalité additive,
— $y(t) = g(t)(1 + s(t))$, pour une saisonnalité multiplicative.

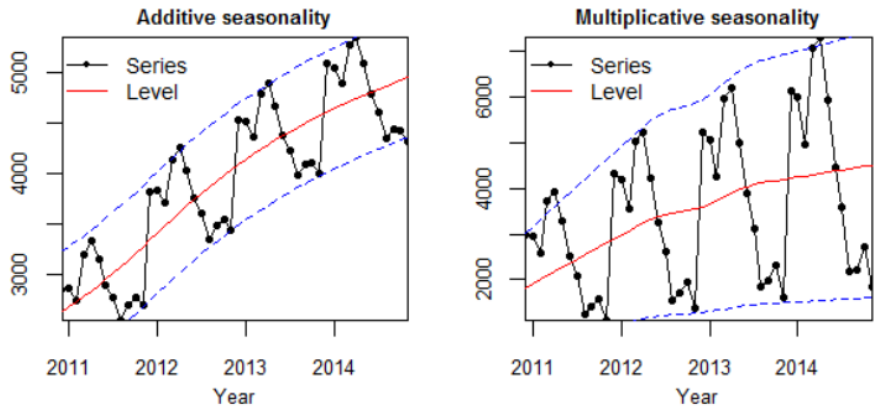


FIGURE 1.20 – Différence entre une saisonnalité additive et multiplicative (Source : *Additive and multiplicative seasonality*, N. Kourentzes, 2014)

▷ **Le terme des jours fériés et évènements exceptionnels :**

Certains évènements sont prévisibles mais ne suivent pas un paterne régulier. D'autres sont prévisibles mais trop ponctuels pour être modélisables par une série de Fourier sans sur-apprentissage. Le modèle Prophet intègre un terme pour correspondre spécifiquement à ces cas particuliers, ce qui demande de pouvoir définir les dates D_i de ces évènements, dans le passé et dans le futur. Par exemple, tous les 1er janvier ou tous les lundis de Pâques.

Les dates D_i correspondent à une ou plusieurs dates, auxquelles sont associées des valeurs κ_i, qui traduisent le décalage de prédiction. Les κ_i sont supposés suivre une loi centrée. Avec les notations matricielles, il faut définir le vecteur d'indicatrices $\boldsymbol{Z(t)}$ suivant :

$$\boldsymbol{Z(t)} = [\mathbb{1}_{t \in D_i} | i \in \mathbb{N}]^\intercal$$

Ainsi, le terme des évènements exceptionnels est défini par :

$$h(t) = \boldsymbol{\kappa Z(t)}$$

Le modèle propose d'intégrer automatiquement les jours fériés d'un pays. Cependant, dans le cas de données non-quotidiennes, ces informations sont peu pertinentes.

Cette fonctionnalité est toutefois particulièrement intéressante pour isoler l'effet d'un évènement en particulier. En définissant une fenêtre de jours exceptionnels sur les données observées, et en choisissant de ne pas les répéter dans le futur, l'impact perturbateur d'un évènement non-reproduisible est limité. Cette possibilité sera typiquement envisagée pour isoler l'impact du Covid.

▷ **Le terme des régresseurs externes :**

Comme présenté précédemment, notamment via l'équation (1.14), un terme dans l'équation du modèle Prophet est dédié à la prise en compte de régresseurs externes, représentant des séries temporelles parallèles à la série d'étude, ayant un pouvoir explicatif sur celle-ci. Les régresseurs x_i nécessitent d'avoir une projection existante, réalisée, si possible, par des experts du sujet (par exemple, cf. partie 3.1.4).

Les coefficients β_i, supposés suivre des lois normales centrées, quantifient la force et la direction de l'impact des régresseurs $x_i(t)$ sur la variable cible $y(t)$. Comme pour la tendance, les régresseurs sont supposés additifs, c'est-à-dire indépendants de la tendance, mais ils peuvent être ajoutés de manière multiplicative individuellement.

En considérant uniquement les termes de tendance et des régresseurs dans l'équation du modèle, cela revient aux deux cas suivants :

— $y(t) = g(t) + x_i(t)\beta_i$, pour un régresseur additif :
Cela signifie qu'une augmentation d'une unité de x_i en t produit une augmentation de β_i sur y en t.

— $y(t) = g(t)(1 + x_i(t)\beta_i)$ pour un régresseur multiplicatif :
Cela signifie qu'une augmentation d'une unité de x_i en t produit une augmentation de $g(t) * \beta_i$ sur y en t.

1.4.3 L'incertitude du modèle

Le modèle propose des mesures de l'incertitude, calculées de manière spécifique pour la tendance, et avec une méthode agnostique pour la saisonnalité. Par défaut, seule l'incertitude de la tendance est implémentée, mais les paramètres du modèle permettent d'avoir une incertitude globale.

▷ **L'incertitude dans la tendance :**

Le terme de tendance de Prophet est basé sur le concept de points de rupture, permettant de mettre à jour les paramètres de taux croissance du modèle dans le temps, dans le cas d'une croissance logistique comme dans le cas d'une croissance linéaire. Le placement de ces points de rupture permet de faire des simulations sur les possibles évolutions des changements de tendance futurs.

Sélection des points de rupture :

Les points de rupture peuvent être disposés manuellement à l'initialisation du modèle à des dates s_j connues ayant une signification particulière, par exemple le début d'un confinement.

Cependant, par défaut les points sont placés à des intervalles réguliers sur la période d'entraînement. Les changements de taux δ_i associés à ces dates sont distribués suivant une loi de Laplace :

$$\delta_i \sim \text{ Laplace}(0, \tau)$$

où τ est le paramètre d'échelle et contrôle la flexibilité du modèle à modifier le taux de croissance k :

$$\tau = 0 \Rightarrow \forall j : \delta_j = 0 \Leftrightarrow \forall t : k(t) = k$$

Ainsi, un paramètre d'échelle nul équivaut à n'avoir aucun point de rupture.

Après placement des points, seuls les plus significatifs sont conservés. Au final, il y aura S points de rupture répartis sur le nombre T d'observations de la période d'entraînement.

Projection des points de rupture :

Pour les projections de la série temporelle, le taux de croissance est gardé constant, égal à $k(T)$. Le calcul de l'incertitude suppose qu'il y aura de nouveaux points de rupture dans le futur, de telle sorte que le taux de croissance continuera d'évoluer. L'hypothèse que la fréquence des points de rupture et leur amplitude sera constante dans le futur est faite.

Sur la période d'entraînement on avait S points sur T observations, on peut donc placer à chaque date de la période de projection un point de rupture fictif $s_{j,j>T}$ avec la probabilité $\frac{S}{T}$, pour conserver la même fréquence.

Afin que les ajustements de taux futurs $\delta_{j,j>T}$ aient la même amplitude que ceux passés, le paramètre d'échelle de la loi de Laplace est à estimer. Pour cela, la méthode du maximum de vraisemblance est utilisée pour obtenir le paramètre $\lambda = \frac{1}{S}\sum_j^S |\delta_j|$.

Les points de rupture futurs sont calculés comme suit :

$$\forall j > T, \quad \begin{cases} \delta_j = 0 & \text{avec la probabilité } \frac{T-S}{T}, \\ \delta_j \sim \text{Laplace}(0, \lambda) & \text{avec la probabilité } \frac{S}{T}. \end{cases}$$

Il est possible de faire un ensemble de simulations qui permettront d'obtenir une distribution de la tendance future, et donc d'avoir un intervalle de prédiction.

▷ **L'incertitude dans la saisonnalité :**

L'incertitude liée à la saisonnalité est calculable mais n'est pas disponible par défaut dans les paramètres du modèle, car cela demande de passer d'un cadre déterministe à un cadre stochastique en utilisant une approche bayésienne, ce qui provoque donc une augmentation importante de la complexité du code.

Afin d'obtenir des intervalles de prédictions, il faut estimer des distributions des paramètres a_i et b_i du terme de saisonnalité (vu en 1.18). A cette fin, une régression bayésienne est mise en place, qui a pour but d'estimer la distribution postérieure des paramètres a_i et b_i à partir des distributions a priori (cf partie 1.3.3).

Des intervalles de prédictions à différents niveaux de confiance sont calculés avec des saisonnalités simulées à partir de tirage des distributions postérieures.

1.5 Le modèle Neural Prophet

Neural Prophet est un modèle de séries temporelles développé par une autre équipe de *data-scientists* de Facebook. Présenté en novembre 2021 dans leur article "*NeuralProphet : Explainable Forecasting at Scale*" [6], le modèle vise a perfectionner le modèle Prophet en introduisant des composantes de *deep learning*.

Equation du modèle :
Le principal défaut du modèle Prophet se trouve dans son incapacité à modéliser les dépendances locales. En effet, l'équation du modèle n'introduit aucun terme auto-régressif. Neural Prophet cherche à améliorer le modèle afin d'identifier les effets locaux immédiats. Pour cela, **un terme auto-régressif et un terme de régresseurs retardés (aussi appelés covariables) sont introduits** à l'équation du modèle Prophet :

$$y(t) = \text{Prophet} + AR(t,p) + L(t) + \epsilon_t$$

où :

— $AR(t,p)$ est l'effet à l'instant t du terme auto-régressif d'ordre p. Son implémentation dans le modèle est disponible de deux manières, soit linéairement, soit en introduisant des réseaux neuronaux.
Dans le cas linéaire, ce terme correspond exactement à celui présenté à l'équation 1.3 lors de la présentation des bases des séries temporelles.

— $L(t)$ correspond aux covariables, c'est-à-dire aux régresseurs avec effets retardés, n'ayant pas de projections préalables. Son implémentation est similaire au terme $AR(t,p)$.
Les k covariables sont projetées avec un module auto-régressif, puis l'effet est ajouté additivement à la réponse du modèle, ainsi $L(t) = \sum_{j=1}^{k} L_j(x'_j(t-1), \cdots, x'_j(t-p_j))$.

Alors que le modèle Prophet est une méthode hybride, introduisant des composantes propres aux modèles de séries temporelles classiques à un modèle de *machine learning*, **Neural Prophet est une méthode hybride, entre séries temporelles, *machine learning* et *deep learning***.

Utilisation :
Sans les termes auto-régressifs, les deux modèles ont exactement la même structure théorique. Toutefois, l'architecture des codes ayant permis leur implémentation en Python diffère. En effet, Neural Prophet est écrit avec Pyro [7] afin d'utiliser des réseaux neuronaux, et Prophet est écrit avec Stan [8] pour la construction d'intervalles utilisant l'approche bayésienne. **Avec des paramètres équivalents, les deux modèles permettent donc d'avoir des résultats quasiment identiques.**

Avec l'introduction de termes auto-régressifs, et en particulier en utilisant l'approche *deep learning*, Neural Prophet performe bien mieux que son prédécesseur. Cependant, aucune modélisation se servant de réseaux neuronaux ne sera réalisée dans ce mémoire, non seulement car cela demanderait de présenter et d'introduire des méthodes de *deep learning*, mais également pour des raisons opérationnelles, car les modèles Neural Prophet demandent environ 5 fois plus de temps d'entraînement.

Estimation d'intervalles :
Etant donné qu'il n'est pas conçu pour implémenter des méthodes bayésiennes, Neural Prophet ne permettait pas d'estimer des intervalles à ses débuts. Pour combler ce manque, des méthodes de *conformal predictions* ont ensuite été intégrées. **Des intervalles de prédictions construits à partir de *conformal predictions* sont ainsi facilement estimés.**

6. NeuralProphet: Explainable Forecasting at Scale, O. Taylor et al. , 2021
7. Pyro : *Deep Universal Probabilistic Programming.* https://pyro.ai
8. Stan : *Probabilistic Programming with Bayesian inference.* https://mc-stan.org/

Chapitre 2

Lien Mortalité - Climat

L'application des méthodes et modèles présentés dans ce mémoire vise à estimer l'impact des projections climatiques sur la mortalité. Bien que ces méthodes soient pensées pour des applications diverses, une connaissance du sujet sous-jacent est indispensable afin de préparer la modélisation et d'analyser les résultats.

Cette partie présente les impacts connus des variations climatiques sur la mortalité. Les modèles classiquement utilisés pour traiter ce lien mortalité-climat seront ensuite évoqués.

L'impact des variations climatiques sur la mortalité n'est plus à identifier. Parmi les premiers travaux, M. Gover présente en 1938 son étude sur la surmortalité durant les vagues de chaleurs [1]. L'objectif des études récentes est d'identifier précisément les différentes sources de surmortalité, de quantifier leurs impacts et de les modéliser afin d'anticiper les périodes à venir.

1. M. Gover, *Mortality during periods of excessive temperature*, 1938

2.1 État des connaissances

Les modèles de mortalité classiques prennent seulement en compte l'exposition et les décès pour calculer des taux de mortalité, en segmentant généralement les populations étudiées par sexe et par âge. Dans un environnement stable, les modèles prospectifs qui considèrent uniquement l'historique des taux de mortalité passés peuvent effectivement être efficients pour prédire le futur, cependant cette hypothèse de stabilité n'est pas vérifiée avec des données réelles.

La mortalité humaine est soumise à des facteurs externes qui évoluent dans le temps, telles que les conditions climatiques. Cet effet se vérifie avec les observations : une surmortalité est remarquée sur les dernières années observées, c'est-à-dire un excédent de décès par rapport au nombre attendu, qui est obtenu via un modèle supposant des conditions extérieures constantes.

En 2022, l'INSEE a remonté une surmortalité de 53 800 décès, soit +8,7% par rapport à leurs estimations réalisées en 2019. Cette surmortalité est en augmentation pour la 3ème année consécutive, d'après leurs données.

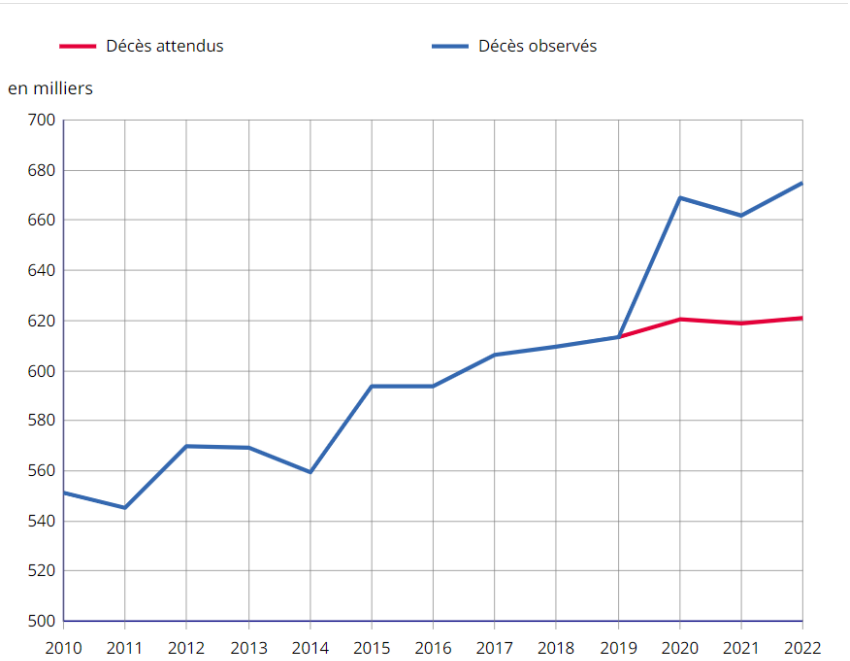


FIGURE 2.1 – Décès observés de 2010 à 2022 et attendus de 2020 à 2022, (Source : https://www.insee.fr/fr/statistiques/7628176)

En introduction de l'article, il est expliqué que "*l'année a compté de manière inhabituelle deux épisodes de grippe, en raison d'une épidémie tardive en mars-avril et précoce en décembre. Des épisodes de fortes chaleurs en été ont occasionné davantage de décès en 2022 qu'en 2021.*" [2].

Les principales sources de mortalité dues au climat sont connues. D'un côté les vagues de froid, causes de maladies, provoquent une grande partie de la surmortalité, d'un autre côté des pics de chaleur et canicules peuvent ponctuellement ajouter une surmortalité additionnelle. **Il convient alors d'étudier l'effet saisonnier du lien mortalité-climat.**

2. INSEE, "53 800 décès de plus qu'attendus en 2022"

2.1.1 Saisonnalité des décès

Les décès ne sont pas répartis uniformément durant l'année en France. Sur les journées les plus froides de l'année, c'est-à-dire aux alentours du 20 janvier, le nombre de décès journalier est près de 50% supérieur à celui des journées comptant le moins de décès. Les journées les plus chaudes, 6 mois plus tard vers le 20 juillet, peuvent présenter des pics dûs aux canicules.

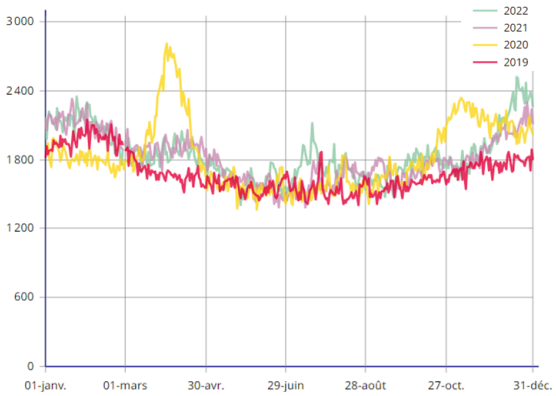


FIGURE 2.2 – Nombre de décès quotidien depuis 2019, (Source : https://www.insee.fr/fr/statistiques/6206305)

Cette forme en 'U' démontre une corrélation négative évidente entre la mortalité et les températures. Toutefois le problème de la modélisation des canicules semble se poser, étant donné que hors vagues d'extrême chaleur, les périodes estivales présentent les plus faibles taux de mortalité.

La **hausse de mortalité en hiver** est principalement due aux maladies, telle que les épidémies grippales, mais également à d'autres pathologies pour lesquelles le climat hivernal représente un facteur aggravant, certaines maladies cardio-vasculaires ou respiratoires par exemple.

La **baisse de mortalité en été** est due à l'effet inverse, c'est-à-dire à un climat globalement moins favorable au développement de maladies. Les vagues de chaleur et canicules peuvent cependant créer des **pics de mortalité**. Le taux de mortalité journalier le plus important a ainsi été relevé pendant la canicule de 2003. Bien que ces pics soient assez peu présents sur l'historique en France, ils pourraient devenir de plus en plus récurrents.
Contrairement à la mortalité hivernale, ces épisodes de surmortalité sont généralement suivis d'épisodes de sous-mortalité : l'**effet moisson**. En effet, la surmortalité due aux canicules correspondrait en partie aux décès des personnes fragiles, qui seraient décédées dans un futur proche. L'effet direct des canicules est donc réduit si la maille temporelle n'est pas assez fine. A la maille *mensuelle*, l'impact d'une canicule de 5 jours suivie d'un effet moisson serait moindre.

Cette saisonnalité, provoquant des taux de mortalité plus élevés en hiver et plus faibles en été, sauf lors de canicules, est globalement **plus marquée chez les femmes**, plus sensibles aux variations climatiques. Les trajectoires sur une période donnée des taux de mortalité chez les femmes présenteront plus d'amplitude que les mêmes trajectoires pour les taux chez les hommes. Les taux de mortalité chez les hommes restent toutefois légèrement plus importants au global, signe de leur espérance de vie plus faible.

En décomposant la population par âge ou tranches d'âge, il apparaît que **la saisonnalité est d'autant plus forte avec l'âge**, étant pratiquement nulle aux âges les plus faibles. Alors que les personnes de plus de 60 ans sont touchées par les maladies hivernales, les hommes dans la vingtaine peuvent présenter une surmortalité légèrement plus importante en été, due aux accidents en vacances.

Tous ces effets semblent montrer que, peu importe l'influence du climat, l'utilisation de modèles de *machine learning* nécessitera de modéliser séparément les différentes catégories démographiques.

2.1.2 Phénomènes climatiques impactant la mortalité

Différents phénomènes climatiques intenses ont également un impact conséquent sur la mortalité. N'étant pas récurrents, ces effets ne peuvent pas être anticipés via un modèle considérant uniquement l'historique de la mortalité.

Vagues de froid :

Des vagues de froid peuvent provoquer un pic de surmortalité en hiver, en particulier chez les populations vulnérables. Le plan grand froid est une mesure mise en place afin d'alerter et de protéger ces populations, de novembre à mars. Le plan défini trois niveaux de risque, qui peuvent être adaptés en fonction des années et des départements, par exemple :

— temps froid : les températures sont positives en journée, mais comprises entre 0°C et -5°C de nuit sur au moins 2 jours,
— grand froid : les températures de jours sont négatives, et comprises entre -5°C et -10°C de nuit,
— froid extrême : les températures de jours sont négatives, et inférieures à -10°C de nuit.

Ces niveaux de risque permettent d'apporter une information supplémentaire sur les périodes particulièrement dangereuses en hiver.

Canicules :

A l'instar du plan grand froid, le plan national canicule vise à protéger et alerter les populations des fortes chaleurs. Pour cela, le plan s'organise autour de 4 axes, actifs du 1er juin à mi-septembre :

— Axe 1 : Prévenir les effets d'une canicule,
— Axe 2 : Protéger les populations par la mise en place de mesures de gestion adaptées aux niveaux de vigilance météorologique,
— Axe 3 : Informer et communiquer,
— Axe 4 : Capitaliser les expériences.

Les différents niveaux de vigilance de l'axe 2 sont basés sur les définitions des canicules de Météo-France.

— Niveau 1 - Vert : veille saisonnière.
— Niveau 2 - Jaune : avertissement, un des seuils est ponctuellement dépassé (fig. 2.3).
— Niveau 3 - Orange : alerte canicule, Météo France déclare une canicule.
— Niveau 4 - Rouge : mobilisation maximale, canicule intense et durable.

Météo France utilise des indicateurs biométéorologiques (IBM), qui correspondent aux moyennes sur trois jours des températures minimales (IBM_{min}) et maximales (IBM_{max}). Une canicule est déclenchée lorsque ces indicateurs dépassent les seuils départementaux. Par exemple, à Toulouse en 2015, le seuil supérieur IBM_{max} était à 36°C et le seuil inférieur IBM_{min} à 21°C :

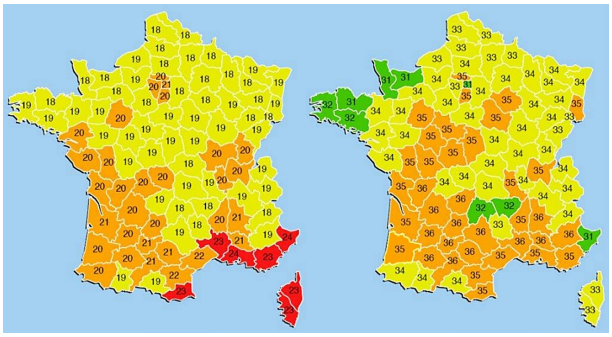


FIGURE 2.3 – Seuils d'alerte IBM_{min} et IBM_{max} par département, été 2015, (Source : Santé publique France, Évaluation de la surmortalité pendant les canicules de 2015)

A partir des écarts entre les indicateurs et les seuils départementaux correspondant, des calculs d'intensité et de sévérité des canicules pourraient être envisagés, permettant de comparer les canicules entre elles.

La surmortalité liée aux canicules est basée sur le nombre de décès observés sur toute la durée d'une canicule, à laquelle sont ajoutés les 3 jours suivants, afin de prendre en compte d'éventuels effets non-immédiats. Le calcul se fait donc au minimum sur 6 jours : les 3 jours minimum nécessaires à la déclaration d'une canicule et les 3 suivants pour l'effet retardé. **Le calcul de surmortalité sur 7 jours, à une maille *hebdomadaire*, paraît donc être adapté à l'observation des canicules.**

Température Humide :

La température n'est pas la seule variable climatique ayant un impact sur la santé, il est admis dans la littérature médicale que **certaines conditions croisées de chaleur et d'humidité sont mortelles pour l'homme**.

L'humidité relative (exprimée en %) mesure la quantité de vapeur d'eau présente dans l'air par rapport à sa capacité maximale à en contenir. Le maximum est de 100%, après quoi l'évaporation devient impossible et le phénomène de condensation apparaît.

Ainsi, la transpiration ne s'évapore plus à 100% d'humidité relative, ne permettant plus au corps de se refroidir. Dans ces conditions, et au delà de 31°C, la température centrale du système humain augmente d'un degrés toutes les 45 minutes [3].

Cet effet avait fait l'objet d'une publication du cabinet Galea & Associés sur la surmortalité journalière, à partir des mêmes données que celles utilisées pour ce mémoire :

3. S.C. Sherwood, *An adaptability limit to climate change due to heat stress*, 2010

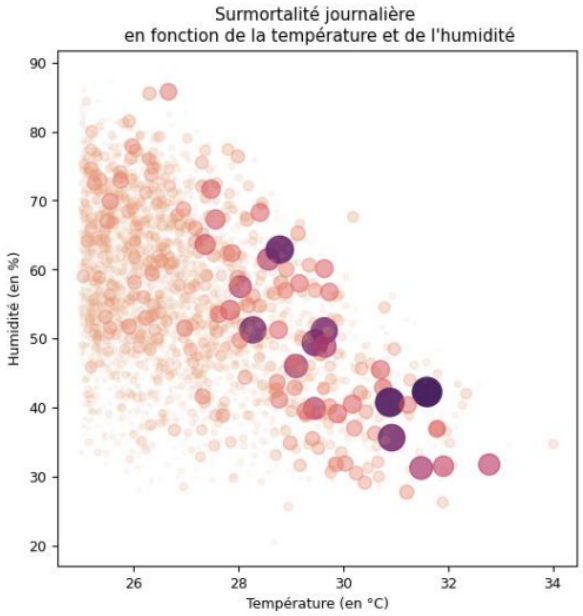


FIGURE 2.4 – Surmortalité journalière selon les températures et l'humidité journalières moyennes (Source : Galea & Associés, Impact de la météo et plus particulièrement des pics de chaleur humide sur la surmortalité, 2023)

Sur ce graphique, la taille et la couleur des points sont fonction de la surmortalité, et "*au-delà de la tendance à l'augmentation de la surmortalité avec la température, il apparaît que les fortes surmortalités se concentrent sur les hauts taux d'humidité, en haut du nuage de points*".

Afin d'étudier cet effet croisé, il peut être intéressant d'introduire la notion de **température humide** (*wet bulb temperature)*, correspondant à un état où l'air est saturé en humidité. Cette température humide T_w peut être approximée à partir de seulement la température de l'air T et de l'humidité relative $RH\%$ via l'équation de Stull [4] :

$$\begin{aligned} T_w = & T \arctan(0.151977 \times (RH\% + 8.313659)^{0.5}) \\ & + \arctan(T + RH\%) \\ & - \arctan(RH\% - 1.676331) \\ & + 0.00391838 \times RH\%^{1.5} \times \arctan(0.023101 \times RH\%) \\ & - 4.686035 \end{aligned} \tag{2.1}$$

L'utilisation de cette variable pourrait permettre aux modèles de comprendre l'interaction température-humidité correspondant directement au phénomène spécifique décrit.

4. R. Stull, *Wet-Bulb Temperature from Relative Humidity and Air Temperature*, 2011

2.1.3 Autres effets sur la mortalité

Les éléments précédemment évoqués semblent montrer l'existence d'une température "optimale".

Cette température, appelée **MMT (*Minimum Mortality Temperature*)** serait d'environ 18°C en France, et en augmentation constante [5]. Cette température serait autour des 16°C en Finlande et de 24°C en Grèce [6], suggérant une **acclimatation des populations** à leur environnement géographique, mais également dans le temps. Cette acclimatation semble évidente, et serait due à l'adaptation des populations, de leur mode de vie et de leurs infrastructures. Elle apparaît toutefois compliquée à prendre en compte dans des modèles prospectifs.

Par ailleurs, les infrastructures et l'environnement proche auraient aussi un impact sur la mortalité.

Les **îlots de chaleur urbains (ICU)** sont des zones, où les températures sont significativement plus importantes que dans les zones rurales avoisinantes, à cause de l'activité humaine et de l'architecture urbaine. Les conditions de vie y sont différentes, conséquence de ce microclimat, ce qui a un impact indirect sur la mortalité.

Une étude de Santé publique France [7]a quantifié le risque relatif de décès au sein de ces ICU en fonction, notamment, du taux de végétalisation, du taux d'imperméabilisation des sols et d'un indice de défaveur sociale. Si les résultats montrent une augmentation de la mortalité en fonction de ces caractéristiques, elles ne semblent pas applicables à une étude géographique à échelle macroscopique.

2.2 Modèles de référence

Les études se penchant sur l'intégration et la quantification d'un risque sur une série temporelle sont de plus en plus fréquentes, notamment concernant le lien entre la mortalité et le climat. Voici une liste non-exhaustive de modèles utilisés dans ce contexte.

2.2.1 CSDL (*Constrained Segmented Distributed Lag Model*)

Le modèle CSDL est pensé pour prendre en compte des effets retardés de la température sur la mortalité [8]. Des effets non-linéaires sont introduits en définissant des seuils de températures minimum ψ_1 et maximum ψ_2 au-delà desquels l'interaction température-mortalité comprend un terme supplémentaire.

Le modèle comprend alors trois termes :

$$\log(\mathbb{E}(y_t)) = x_t^\intercal \delta + \sum_{l_1=0}^{L_1} \beta_{1l_1}(z_{t-l_1} - \psi_1)_- + \sum_{l_2=0}^{L_2} \beta_{2l_2}(z_{t-l_2} - \psi_2)_+$$

5. J.R. Barrett, *Increased Minimum Mortality Temperature in France: Data Suggest Humans Are Adapting to Climate Change*, 2015

6. W.R. Keating et al., *Heat related mortality in warm and cold regions of Europe: observational study*, 2000

7. Santé Publique France, Influence de caractéristiques urbaines sur la relation entre température et mortalité en Île-de-France, 2020

8. V.M.R Muggeo, *The Constrained Segmented Distributed Lag Parameterization*, *Journal of Statistical Software*, 2010

où,

— y_t : Le nombre de décès attendu au moment t.
— $x_t^\intercal \delta$: Les variables explicatives temporelles x_t, comme le jour de la semaine ou le mois, multipliées par un vecteur de coefficients δ.
— z_{t-l_i} : La température relevée au moment $t - l_i$.
— ψ_i : Les seuils de température à définir, au-delà desquels une modification de l'interaction avec la mortalité est attendue.
— $\beta_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{il_i}, \ldots, \beta_{iL_i})^\intercal$: Les courbes modélisant les relations température-mortalité pour le froid ($i = 1$) et le chaud ($i = 2$), proportionnellement à l'écart de température $(z_{t-l_i} - \psi_i)$.
— $\sum_{l_i=0}^{L_i}, i \in \{1, 2\}$: La somme des effets sur tous les *lags* l_i, correspondant au nombre de périodes temporelles à prendre en compte pour les températures dépassant les seuils de température ψ_i.

Le modèle CSDL permet à la fois de bien modéliser les canicules et les vagues de froid. Cependant, l'introduction d'autres variables, comme l'humidité, n'est pas possible, ce qui peut en faire un modèle trop rigide pour prendre en compte toutes les variables explicatives disponibles.

2.2.2 DLNM (*Distributed Lag Non-Linear Model*)

Le modèle DLNM est une extension des GLM (*Generalized Linear Models*), conçu pour évaluer l'effet retardé (*lagged effect*) et potentiellement non-linéaire d'une variable sur une série temporelle. Présenté pour la 1ère fois en 2010, l'application associée modélise la relation température - mortalité à New-York sur la période 1987-2000 [9].

Le modèle est construit en partant des constats précédemment évoqués :
— l'effet d'un phénomène climatique, comme une canicule, peut se manifester sur plusieurs jours après l'exposition, nécessitant l'introduction d'un effet retardé (*lagged effect*).
— les interactions entre la cible (la mortalité) et le risque auquel elle est exposée (la température) ne sont pas linéaires. Par exemple, une augmentation d'1°C n'a pas le même effet à 20°C et à 35°C. La réponse doit donc pouvoir comprendre une fonction non-linéaire du risque.

A partir ce ces éléments, le modèle prend la forme suivante :

$$log(\mathbb{E}(y_t)) = \alpha + g(t) + \sum_{l=0}^{L} f(x_{t-l}) \cdot h(l) + \sum_{p=0}^{P} z_t$$

où,
— y_t : La réponse observée à un moment donné t, par exemple le nombre de décès.
— α : L'ordonnée à l'origine, représentant le niveau moyen de la réponse.
— $\sum_{l=0}^{L}$: La somme des effets sur tous les *lags* l, 0 correspondant à un effet immédiat et L étant le *lag* maximum pris en compte.
— $f(x_{t-l})$: Une fonction non linéaire de l'exposition x au temps $t - l$.
— $h(l)$: La fonction de lag qui décrit comment l'effet de l'exposition x change en fonction du temps de retard l. Cela peut représenter, par exemple, l'affaiblissement ou l'amplification de l'effet dans le temps.
— $g(t)$: Le terme contrôlant la saisonnalité et la tendance à long terme.
— $\sum_{p=0}^{P} z_t$: Les potentielles p autres variables, avec un effet linéaire non-retardé.

Le modèle DLNM est un modèle paramétrique complexe et puissant, mais difficile à mettre en place, étant donné le nombre conséquent de paramètres à estimer et les connaissances préalables nécessaires dans le choix des *lags* à définir.

9. A. Gasparrini et al., *Distributed lag non-linear models*, *Statistics in medicine*, 2010

Chapitre 3

Présentation des données et de leurs traitements

Cette partie présente le type de données et les traitements adaptés à la mise en place des modèles à venir. Cette trame est suivie et illustrée avec les données démographiques et climatiques utilisées pour ce mémoire. Les détails des traitements, en particulier sur la qualité des données et les contrôles de cohérences ne seront pas tous détaillés.

La première partie pratique de tout projet de Data-Science est l'analyse et le traitement des données. Cette étape peut être la plus chronophage, en particulier lorsque plusieurs sources de données sont nécessaires. Pour autant elle n'est pas à négliger, car la performance des modèles dépend directement de la qualité des données traitées.

La projection de l'impact d'un facteur de risque demande au minimum 3 sources de données :

— **L'historique des données** : données dont la sensibilité aux aléas est étudiée, tels que des portefeuilles d'assurés ou des investissements. Pour ce mémoire, ce sont différentes données démographiques de l'INSEE, permettant de calculer un taux de mortalité.

— **Les facteurs de risque** : représentant les aléas auxquels sont exposés les données. Ils sont nécessaires à l'entraînement des modèles. Des relevés météorologiques fournis par Météo France serviront ici d'historique pour les facteurs de risque climatique.

— **Une projection de ces facteurs** : permettant de fournir les variables explicatives nécessaires à la projection des données exposées au risque. Il est préférable qu'elle ait été créée par des experts du sujet. Dans le cas du risque climatique, les différents scénarios de projections climatiques publiés par le DRIAS sont adaptés.

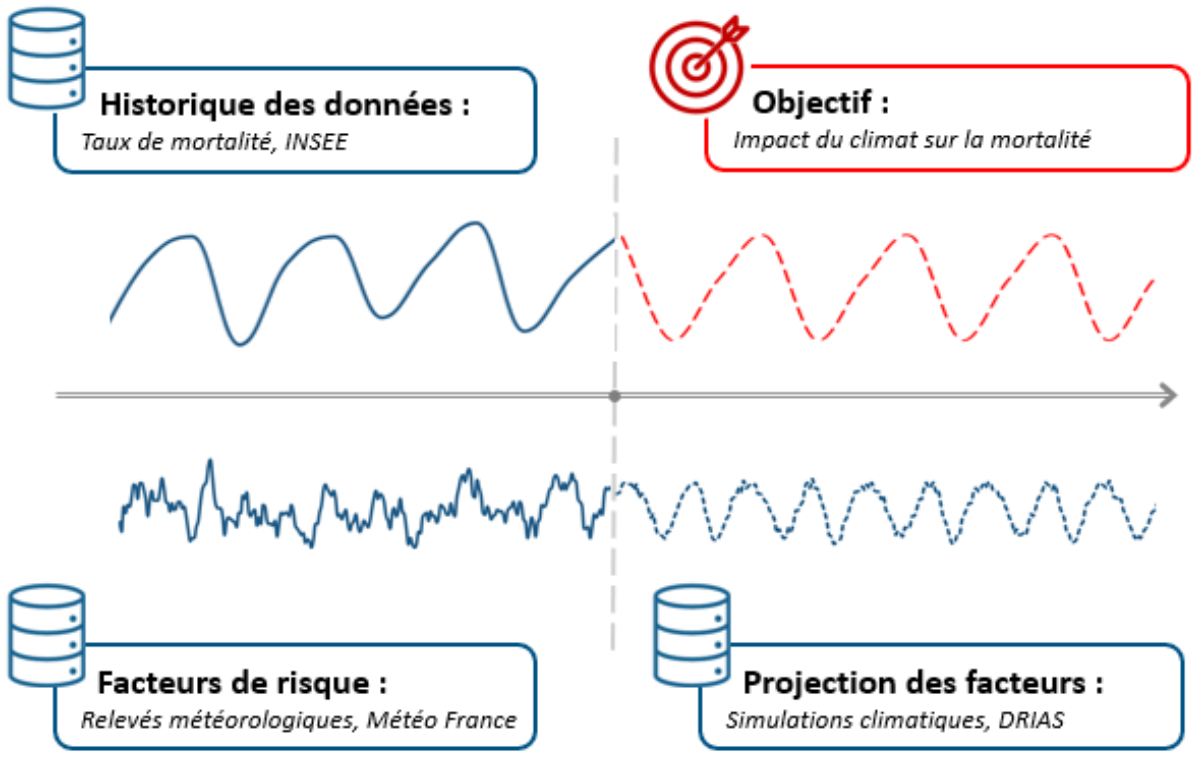


FIGURE 3.1 – Schéma des données nécessaires

En amont, un travail pour s'assurer que les différentes sources de données seront compatibles entre elles est indispensable. En effet, les bases de données devront pouvoir être agrégées et croisées selon un format et une maille identique.

Dans le cadre de ce mémoire, ce qui sera appelé "maille" correspond à un ensemble de choix ou d'hypothèses réalisés sur les regroupements démographiques (sexe et âge), temporels (journalier, hebdomadaire, mensuel ou annuel) et géographiques (départements, régions ou France métropolitaine). Les différentes combinaisons présentent leurs avantages et inconvénients :

		Avantages	Inconvénients
Temporelle	· Journalier	· Utilisation maximale des données climatiques	· Volumétrie conséquente
	· Hebdomadaire	· Données climatiques exploitables · Format plus adapté aux études de mortalité	· Perte d'informations sur les évènements ponctuels
	· Mensuel / Annuel	· Faible volumétrie · Format classique	· Données climatiques non exploitables
Spatiale	· Département	· Utilisation cohérentes des données climatiques	· Volumétrie conséquente (96 départements = 96 modèles)
	· Région	· Réduction du nombre de modèles (96 → 13 modèles)	· Hypothèses climatiques fortes · Perte d'informations localisées
	· France métropolitaine	· Faible volumétrie (1 modèle) · Format classique	· Données climatiques non exploitables
Démographique	· Sexe	· Impact physiologique du climat	· Double le nombre de modèles
	· Âge	· Format classique : comparaison avec des modèles de durée possible	· Volumétrie conséquente (90 âges = 90 modèles)
	· Tranche d'âge	· Réduction du nombre de modèles	· Hors du cadre classique des modèles de durée

Ainsi, la maille optimale pour cette étude semble être la combinaison *jour - département - sexe - âge*. Toutefois, cette possibilité est à écarter non seulement pour des raisons techniques, à cause du volume de données et du nombre de modèles à entraîner, mais également pour des raisons pratiques, car en affinant le périmètre au maximum le nombre d'observations peut devenir non-représentatif et mener à un sur-apprentissage des modèles.

Au regard des différentes possibilités, et rétroactivement après différentes tentatives, la maille privilégiée serait la combinaison *semaine - département - sexe - tranche d'âge*. Par la suite, les tranches d'âges 60-79 feront l'objet d'une attention particulière.

Un modèle à la maille *mois - France - sexe - âge* est aussi envisagé, car il permettrait de valider la méthode en comparant les résultats avec ceux d'un modèle de durée classique tel que Lee-Carter.

3.1 Présentation des données

3.1.1 INSEE : Fichier des décès

L'INSEE publie tous les mois l'ensemble des décès remontés par les communes de France depuis 1970 [1]. Ces fichiers mensuels comprennent les décès dont l'INSEE a pris connaissance, avec les informations suivantes :

— **Nom**, **Prénom** et **Sexe** de la personne décédée,
— **Dates** de naissance et de décès,
— **Code Postaux** et **Libellés** des lieux de naissance et de décès.

Ces publications obligatoires sont réalisées par l'INSEE dans le cadre de leur mission de service public. Ainsi, bien qu'elles contiennent des informations nominatives, ces données ne sont pas considérées comme étant à caractère personnel relevant de la vie privée. Des versions annuelles reprenant les douze fichiers mensuels sont également disponibles pour les années complétées.

Les informations fournies mensuellement par les communes comprennent les informations concernant les décès dont elles ont eu connaissances depuis leur dernier envoi. Par conséquent, les fichiers peuvent contenir des décès ayant eu lieu plusieurs mois auparavant. Ainsi, tous travaux souhaitant étudier ces données jusqu'à l'année N doivent utiliser les premiers fichiers mensuels de l'année N+1 pour réduire le biais des décès survenus mais non déclarés à l'INSEE. Dans le cas de ce mémoire, les données sont étudiées de 1990 à 2023 avec les fichiers mensuels de 1990 à mai 2024, ce qui représente 20 millions de lignes.

Ces données permettront d'avoir un nombre de décès à la maille *jour - sexe - commune - âge*.

3.1.2 INSEE : Estimation de la population

L'INSEE estime annuellement la population française à la maille *sexe - tranche d'âge de 5 ans - département* depuis 1975 [2]. L'estimation de la population est également disponible à la maille *sexe - âge - France* [3]. Ces deux sources utilisent les mêmes données, sont parfaitement cohérentes entre elles, mais sont présentées à des mailles différentes.

En fonction de l'année, ces estimations peuvent provenir à la fois des recensements, de statistiques réalisées à partir des états civils, et d'une estimation du solde migratoire.

— De 1975 à 1999, et comme depuis 1946 (année de création de l'INSEE), l'ensemble des communes de France était recensé de manière exhaustive à intervalles réguliers.

— De 2000 à 2005, aucun recensement de la population n'a eu lieu. L'estimation de la population est alors réalisée à partir des recensements précédents, en prenant en compte l'évolution de la population.
Cette évolution vient de deux effets : l'excédent naturel et le solde migratoire. L'excédent naturel correspond à la différence entre le nombre de naissance et le nombre de décès, et est calculable à partir des états civils. Le solde migratoire ne peut pas être directement obtenu : il correspond à la différence entre la variation de population recensée entre deux années et l'excédent naturel correspondant. Le solde migratoire est calculé sur l'historique des recensements et estimé de 2000 à 2008.

1. INSEE, Fichiers des personnes décédées depuis 1970
2. INSEE, Estimation de la population au 1 janvier 2024
3. INSEE, Pyramides des âges au 1 janvier 2024

La population annuelle est estimée, et sera ensuite corrigée rétroactivement avec les recensements à venir.

— De 2006 à 2020, la population est estimée à partir des résultats des nouveaux recensements. Débutés en 2004, ils concernent les petites communes et 40% des autres foyers, sur une période centrée de 5 ans.

— De 2021 à 2024, les résultats des recensements ne sont pas encore disponibles ou définitifs. La population est estimée avec les résultats provisoires des recensements. Ces résultats sont révisés annuellement et deviennent définitifs en N-3.

Ces données permettent d'avoir annuellement un nombre de personnes par sexe en fonction de différents périmètres. Couplées aux données précédentes, et en faisant l'hypothèse que la population reste constante au cours de l'année, **un taux de mortalité est calculable à la maille *jour/semaine/année - sexe - département - tranche d'âge* pour les modèles optimaux, mais aussi à la maille *jour/semaine/année - sexe - France - âge* pour des modèles comparables à des modèles de durée.**

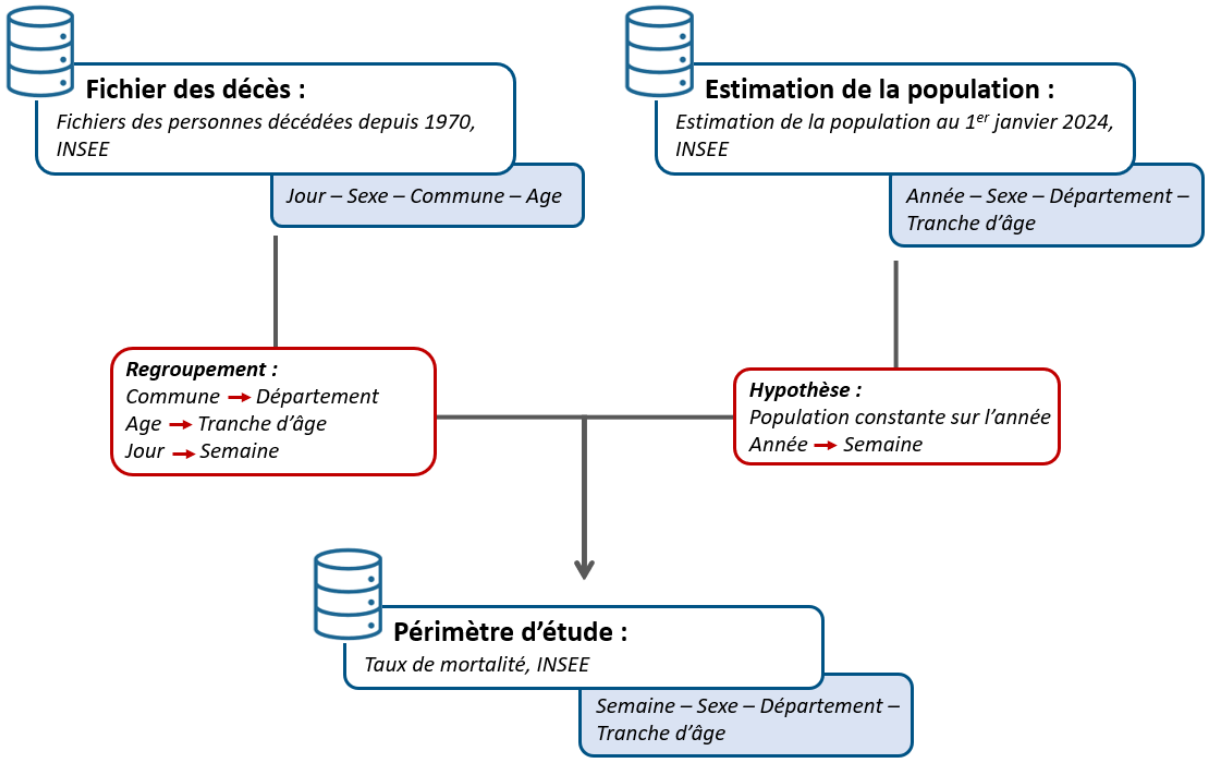


FIGURE 3.2 – Exemple de construction d'une base de données démographique, à partir de données INSEE

3.1.3 Météo-France : Données climatiques

Météo-France est le service météorologique et climatique de l'Etat, et a notamment pour mission la collecte de données climatologiques via leur réseaux d'infrastructures d'observation, la transformation et la mise à disposition de ces données.

Jusqu'au 1er janvier 2024, Météo-France publiait gratuitement uniquement les données climatologiques correspondant à son réseau d'infrastructures *SYNOP*. Ce réseau comprend une quarantaine de stations synoptiques professionnelles destinées à la collecte de données à grande échelle. L'utilisation de ces données nécessitait de réaliser des regroupements de départements afin de pouvoir associer au moins une série d'observations par département.

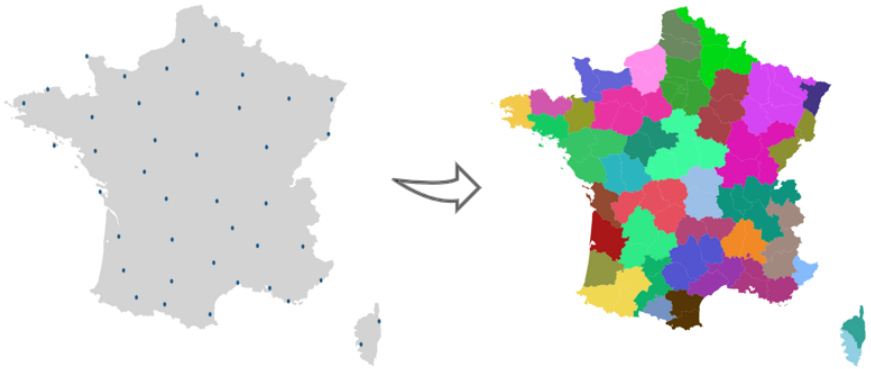


FIGURE 3.3 – Regroupement géographique à partir des stations du réseau *SYNOP*

La circulaire du Premier ministre du 27 avril 2021 relative à la politique publique de la donnée a pour objectif d'aider les transitions écologiques et énergétiques en facilitant l'accès à certaines données publiques. Ainsi, depuis le 1er janvier 2024, toutes les données publiques de Météo-France sont utilisables gratuitement [4]. Les données climatologiques sont désormais disponibles avec des réseaux d'infrastructures plus précis, avec les stations du réseau *RADOME* et du réseau *ETENDU*. Les plus récentes sont mises à jour quotidiennement.

Les stations du réseau *ETENDU* ne faisaient pas toutes partie des stations officielles de Météo-France lors de leur installation, et certaines sont destinées à la mesure d'un paramètre en particulier. Comptant près de 1000 stations, ce réseau permet d'obtenir des informations spécifiques à l'échelle locale. De par leur grand nombre de valeurs manquantes et de leur incertitude, ces relevés ne sont pas utilisables dans l'étude de ce mémoire.

Les stations du réseau *RADOME* (Réseau Automatique de DOnnées MEtéorologiques) sont environ 550, et permettent de mesurer des paramètres simples, tels que le vent, la température, l'humidité, la pression et les précipitations, à l'échelle sub-départementale. Les stations du réseau *SYNOP* sont incluses dans celles du réseau *RADOME*. Dans la majorité des cas, l'utilisation de ces données nécessite de réaliser des regroupements de stations afin de pouvoir associer plusieurs séries d'observations pour chaque département individuellement. Ce réseau nous permet ainsi d'avoir des données fiables à l'échelle départementale, voir sub-départementale.

4. Météo-France, Données climatologiques de base - quotidiennes

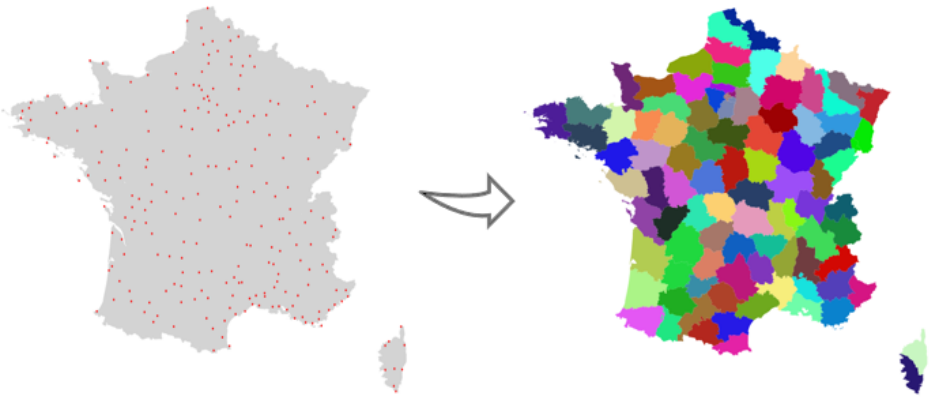


FIGURE 3.4 – Regroupement géographique à partir des stations du réseau *RADOME*

Les fichiers sont disponibles par département pour la période 1950-2024. Ils contiennent par jour et par station près de 70 variables, dont notamment :

- La **latitude** et la **longitude**, afin de localiser la station,
- Les **maximum**, **minimum** et **moyenne** sur la journée de la **température**, de l'**humidité** et de la **force du vent**,
- La quantité de **précipitation** tombée en 24h,
- Le **rayonnement** et **cumul des UV**,
- Des indicatrices d'occurrence de **brouillard**, de **neige** ou d'**orage**.

Certaines d'entre-elles ne sont pas utilisables car elles ne sont pas mesurées par toutes les stations, ou n'ont pas de projections réalisées par le DRIAS.

Après traitements, ces données permettrons d'avoir un historique du risque à la maille *semaine - département*.

3.1.4 DRIAS : Projections climatiques

Le Groupe Intergouvernemental d'experts sur l'Evolution du Climat (GIEC) évalue et synthétise les travaux scientifiques liés au changement climatique. Dans son 5ème rapport, publié en 2014, le GIEC sélectionne quatre scénarios de forçage radiatif (dit scénarios RCP : Representative Concentration Pathways) qui servent de base à des modèles de projections climatiques. Cette sélection a été effectuée parmi un panel de plus de 300 scénarios publiés dans la littérature scientifique.

Dans son 6ème rapport, publié en 2023, le GIEC présente cinq scénarios socio-économiques (dit scénarios SSP : Shared Socioeconomic Pathways) plus représentatifs de trajectoires envisageables. Ces nouveaux scénarios prennent en compte de nouvelles hypothèses, tels que des indices de développement humain, d'éducation, ou de croissance économique. Ces scénarios ne seront pas développés dans la suite de ce mémoire, car **les modèles climatiques de référence sont basés sur les scénarios RCP pour le moment**.

Le DRIAS est une initiative dédiée à la mise à disposition de projections climatiques élaborées par des laboratoires de modélisation du climat. Il doit aider et accompagner les acteurs nationaux concernés par le changement climatique afin d'avoir accès aisément aux informations nécessaires à leurs modélisations et études d'impact. Les projections réalisées par différents modèles grâce aux scénarios préconisés par le GIEC sont disponibles sur le site du DRIAS [5].

5. DRIAS, les futurs du climats

Ces simulations sont réalisées à partir d'un scénario de référence qui sert à un modèle climatique global, puis à un modèle climatique régional, qui nécessite une méthode de descente d'échelle et de correction de biais. Afin d'évaluer des incertitudes, les études doivent se reposer sur plusieurs simulations scénarios-modèles-méthodes de correction de biais.

▷ **Les scénarios RCP :**

Les scénarios de forçage radiatif RCP (Representative Concentration Pathways) représentent différentes trajectoires d'évolution associées aux changements de la composition de l'atmosphère, prenant en compte différentes hypothèses socio-économiques envisageables.
Le forçage radiatif est la différence entre l'énergie radiative reçue (venant à plus de 99,5% du soleil), et celle émise par le système atmosphère/Terre, due à des facteurs d'évolution du climat tels que la concentration des gaz à effet de serre. Un forçage radiatif positif tend à réchauffer la surface. Il se mesure en W/m².

Le scénario RCP8.5, le plus pessimiste, trace un futur sans politique de régulation des émissions. Environ 10% des scénarios envisagés le dépasse. Il correspond à un forçage radiatif de l'ordre de 8,5 W/m², ce qui se traduit par une élévation de la température moyenne de +4°C à horizon 2100.

A l'opposé, le scénario RCP2.6, le plus favorable, est le seul conforme aux objectifs de l'accord de Paris. Il dépasse environ 10% des scénarios proposés. Le forçage radiatif de 2,6 W/m² se traduit par un réchauffement global qui restera inférieur à 2°C par rapport aux températures pré-industrielles à horizon 2100.

Les scénarios RCP4.5 et RCP6.0 sont des voies intermédiaires, dans lesquelles les émissions continuent de croître pendant quelques décennies avant de se stabiliser.
Ces scénarios de forçage radiatif sont utilisés en entrée de modèles climatiques globaux (GCM) dont l'objectif est de simuler l'évolution du climat à l'échelle mondiale.

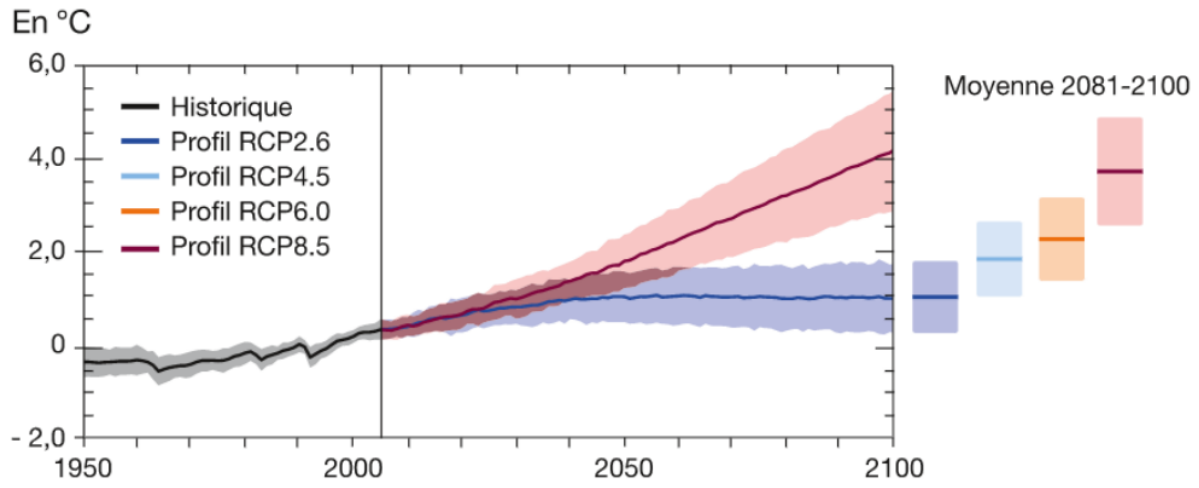


FIGURE 3.5 – Trajectoires des scénarios RCP sélectionnés par le GIEC (Source : GIEC, 5ème rapport, 2014)

▷ **Les modèles climatiques :**

Les modèles de prévisions météorologiques et les modèles climatiques se basent sur des principes physiques de circulation et d'échange d'énergie pour calculer des paramètres atmosphériques, telles que les températures ou précipitations.

Les modèles de prévisions météorologiques ont pour objectif la projection sur un horizon proche en s'assurant de la cohérence avec les observations passées. Ils sont utilisés pour anticiper la météo à 15 jours. Les modèles climatiques simulent des évolutions plausibles sur un horizon allant jusqu'à plusieurs siècles. Ces derniers ont permis de créer les simulations mises

à dispositions par le DRIAS. Ils se décomposent en un modèle de climat global (GCM), un modèle de climat régional (RCM) avec une méthode de descente d'échelle, et une méthode de correction de biais.

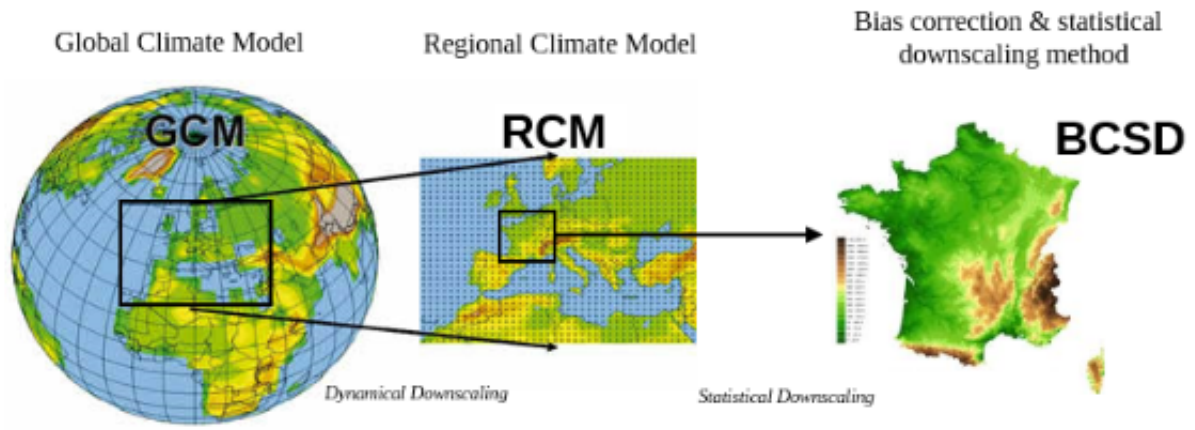


FIGURE 3.6 – Étapes de construction d'un modèle climatique, (Source : DRIAS)

Les modèles climatiques globaux (GCM) :

Les GCM prennent en entrée un scénario de forçage radiatif RCP pour modéliser au mieux les flux d'énergie entrants et sortants. Le climat simulé est la résultante de l'ajustement entre énergie reçue et énergie perdue, calculée en prenant en compte les évolutions et interactions entre les différents milieux du système atmosphère/Terre (l'océan, les glaciers, les couches atmosphériques, ...). La distribution statistique des variables météorologiques simulées doit être similaire à celle observée. L'objectif est de simuler en tout point du globe un ensemble de paramètres climatiques, mais actuellement la résolution spatiale des modèles les plus précis est d'environ 100km.

Les modèles climatiques régionaux (RCM) :

Pour des études d'impact, il est nécessaire d'avoir des données climatiques à des emplacements précis correspondant aux stations de mesure, ce qui n'est pas possible avec des GCM seuls. Pour résoudre ce problème, il est indispensable de procéder à des descentes d'échelle. Les méthodes de descente d'échelle permettent d'augmenter artificiellement la résolution spatiale des modèles climatiques. Un modèle climatique régional (RCM) fonctionne avec une résolution plus fine sur une zone spécifique de l'atmosphère et utilise les données d'un GCM pour ses conditions aux frontières. Concrètement, cela consiste à utiliser en entrée une simulation d'un GCM avec les données atmosphériques d'une région spécifique supprimées, pour fournir des contraintes de bordures au RCM. Puis, cette "zone vide" est remplacée par la simulation d'un RCM, calculée à une maille plus fine avec plus de détails.

Les méthodes de correction de biais :

Les projections climatiques régionalisées ne peuvent pas être utilisées directement pour des études d'impact à l'échelle locale en raison de deux problèmes principaux : elles sont biaisées par rapport aux observations et leur échelle spatiale est trop grossière. Des méthodes de correction de biais sont mises en œuvre point par point pour corriger les variables du modèle, afin d'aligner la distribution statistique des données du modèle sur celle des données observées.

La façon la plus complète d'analyser le comportement en une station est de comparer les fonctions de densité probabiliste pour chaque saison et variable. Le diagramme quantile-quantile, aussi appelé Q-Q plot, est à la fois un outil d'analyse et une méthode de correction : une fonction de transfert associe chaque centile du modèle au centile observé, indépendamment pour chaque variable. Au-delà du dernier centile estimé et en deçà du 1er, une correction constante est appliquée.

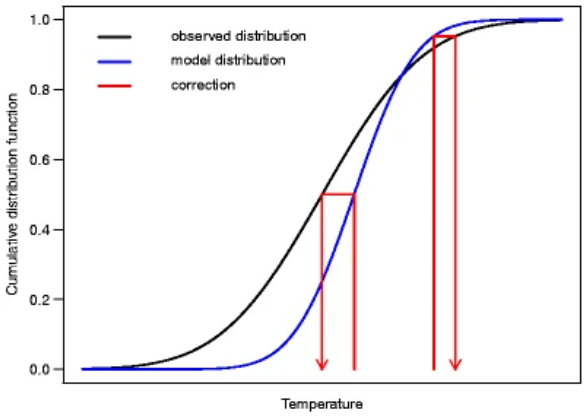


FIGURE 3.7 – Correction de biais avec un QQ-plot, (Source : Maraun, D. Bias Correcting Climate Change Simulations - a Critical Review (2016))

Cette méthode présente plusieurs limites :
— elle ne préserve pas l'interdépendance entre les variables,
— elle ne tient pas compte des spécificités climatiques des variables,
— et l'hypothèse de stationnarité entre la fonction de répartition cumulée (CDF) du modèle régional et la CDF du climat de référence est difficilement vérifiable.

D'autres techniques de correction de biais existent pour les projections climatiques, comme la méthode ADAMONT. C'est une consolidation de la méthode d'ajustement statistique quantile-quantile qui ajuste les biais selon quatre types de temps par saison. Le changement climatique se traduit alors par une modification de la fréquence des types de temps. L'hypothèse de stationnarité a donc plus de chance d'être vérifiée au sein d'un même type de temps. De plus, un traitement particulier est effectué pour conserver au mieux la cohérence entre les variables, par exemple entre la transition de la neige à la pluie.

Les sorties des modèles :
Les simulations climatiques corrigées sont disponibles avec une résolution de 8km x 8km sur le portail DRIAS et seront utilisées pour représenter les projections de risque dans ce mémoire. Il est important de considérer les différentes sources d'incertitude associées à ces projections, à savoir celles liées aux scénarios, aux données observées, aux modèles climatiques globaux et régionaux, ainsi qu'aux méthodes de correction de biais.

▷ **Les simulations :**
Le portail DRIAS met à dispositions 30 simulations différentes pour les données DRIAS-2020, créées à partir des scénarios RCP. Ces simulations correspondent à des combinaisons scénarios-GCM-RCM-méthodes de correction de biais. Seuls les trois scénarios RCP2.6, RCP4.5 et RCP8.5 ont été utilisés pour produire des simulations régionales. Les utilisateurs ont ainsi le choix parmi 8 simulations pour le RCP2.6, 10 pour le RCP4.5 et 12 pour le RCP8.5.
Des guides et des indicateurs, basés notamment sur les différences de températures par saison entre une période donnée et une période de référence permettent de sélectionner un nombre limité de simulation. Par exemple, pour les automnes du scénario RCP2.6, le graphe suivant permet d'identifier les simulations les plus clivantes.

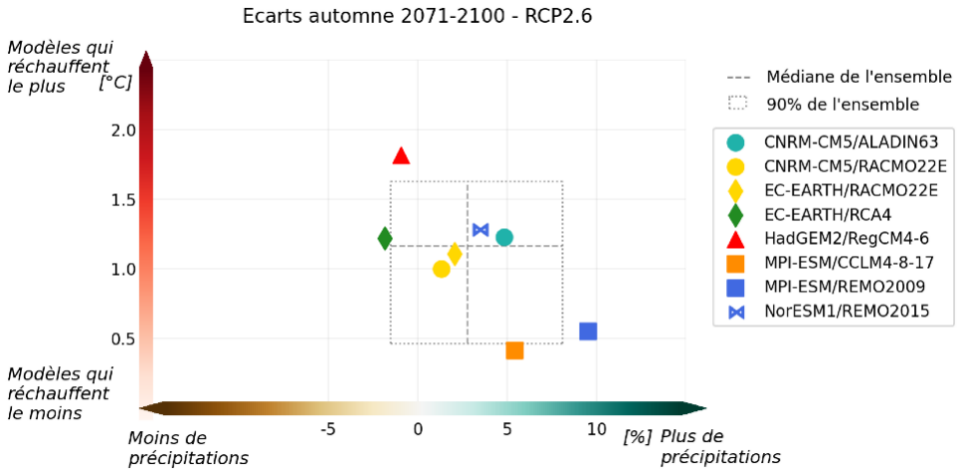


FIGURE 3.8 – Diagramme $\Delta P/\Delta T$, saison automnale, RCP2.6, (Source : DRIAS, Aide à la sélection)

Dans les travaux à suivre, l'idéal serait de conserver au moins 3 simulations par scénario, soit un total de 9 jeux de données. Chaque scénario aurait une simulation pessimiste, une simulation médiane et une positive, permettant d'avoir un intervalle d'incertitude lié aux données climatiques par scénario.

Etant donné le temps de calcul conséquent, seules les simulations "CNRM-CM5/ALADIN63" des scénarios RCP2.6 et 8.5 seront utilisées. Ces simulations correspondent à des simulations médianes des scénarios extrêmes.

Les extractions :

Les extractions sont personnalisables pour correspondre à la maille spatiale souhaitée, permettant de sélectionner des stations parmi 19 160 points de la carte. Les données sont disponibles sur la période temporelle allant de 2006 à 2099. Un choix parmi un ensemble de variables climatiques est également proposé.

Cette étude demande d'extraire les données sur la période 2023-2050. **Pour correspondre au groupement par département des données MétéoFrance, 96 points sont choisis, correspondant aux centres de chaque département.**

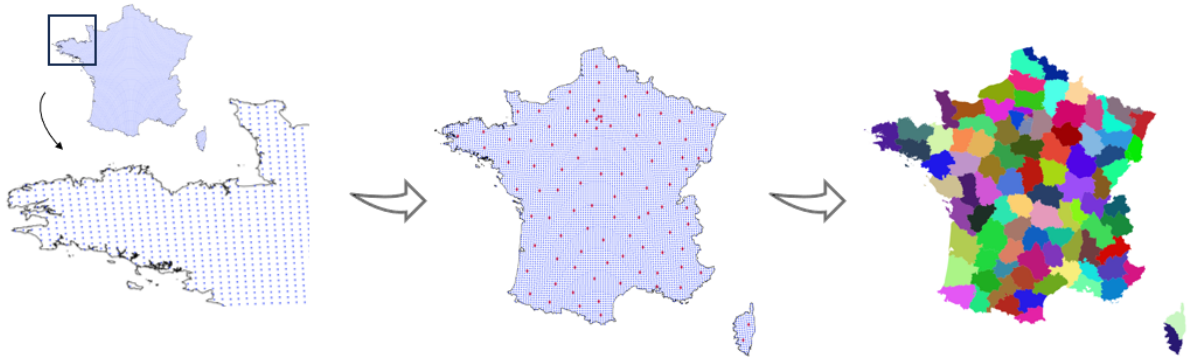


FIGURE 3.9 – Regroupement géographique à partir des points de la grille du DRIAS

Les fichiers sont fournis pour la période renseignée. Ils contiennent par jour et par station une dizaine de variables, dont notamment :

— La **latitude** et la **longitude**, afin de localiser la station,
— Les **maximum**, **minimum** et **moyenne** sur la journée de la **température**
— Une valeur d'**humidité** et de **vitesse du vent** pour la journée,
— La quantité de **précipitation** tombée en 24h,
— Et, selon les modèles, des valeurs de **rayonnement** et d'**évapo-transpiration**.

Après traitements, ces projections permettront d'avoir une projection du risque à la maille *semaine - département*.

3.1.5 INSEE : Projections de la population

Les données précédentes permettent de réaliser l'étude souhaitée qui consiste à prendre en compte l'impact du climat dans la projection de taux de mortalité.

A des fins d'interprétations, quantifier cet impact en nombre de décès peut être pertinent. Cela demande d'avoir des projections de la population à la maille *sexe - tranche d'âge - département*.

L'INSEE réalise chaque année des projections de la population en France avec ses modèles "Omphale" (Outil Méthodologique de Projection d'Habitants, d'Actifs, de Logements et d'Élèves). Ces projections sont disponibles à plusieurs échelles, et notamment à la maille évoquée. Leur dernière publication, du 8 janvier 2024, propose des scénarios de projections à horizon 2070 [6].

Le modèle de projection Omphale de l'INSEE est régulièrement mis à jour depuis sa création. Des détails sur la construction et les hypothèses du modèle avaient été publiés lors de la sortie du modèle de 2010 sur le site de l'INSEE [7]. Les projections utilisées ont été réalisées à partir de la version 2022, qui comprend quelques adaptations.

Le modèle est décrit comme étant basé sur une "méthode des composantes", utilisant des hypothèses sur trois termes pour décrire les variations de population :

— la fécondité, à partir de l'indice conjoncturel de fécondité (ICF), mesurant le nombre moyen d'enfants qu'une femme pourrait avoir au cours de sa vie.
— les migrations, avec des données d'émigration vers l'étranger et d'immigration depuis l'étranger.
— la mortalité, en se basant sur l'évolution de l'espérance de vie (EDV) à la naissance.

Chacune des composantes présente trois scénarios, permettant d'avoir 27 scénarios possibles. Dans les faits, seuls les 11 suivants sont disponibles librement :

6. INSEE, Projections de population 2018-2070
7. INSEE, Le modèle de projection démographique Omphale 2010

Scénarios standards	Hypothèses retenues		
	Fécondité	Espérance de vie	Migrations avec l'étranger
Central	Centrale	Centrale	Centrale
Population haute	Haute	Haute	Haute
Population basse	Basse	Basse	Basse
Fécondité haute	Haute	Centrale	Centrale
Fécondité basse	Basse	Centrale	Centrale
Espérance de vie haute	Centrale	Haute	Centrale
Espérance de vie basse	Centrale	Basse	Centrale
Migrations hautes	Centrale	Centrale	Haute
Migrations basses	Centrale	Centrale	Basse
Population jeune	Haute	Basse	Haute
Population âgée	Basse	Haute	Basse

FIGURE 3.10 – Scénarios de projections de la population (Source : hhttps ://www.insee.fr/fr/information/2571308)

Les différentes composantes peuvent être directement choquées ou progressivement intégrées à la démographie du scénario de développement standard (DSDS), par exemple pour la version utilisée :

Composantes	Hypothèse centrale	Hypothèse basse	Hypothèse haute
Fécondité	Baisse de l'ICF parallèle à la tendance centrale de la DSDS : 1,87 à 1,8 de 2018 à 2023 puis constance jusqu'en 2070	Baisse de l'ICF parallèle à la tendance basse de la DSDS : 1,87 à 1,6 de 2018 à 2030 puis constance jusqu'en 2070	Hausse de l'ICF parallèle à la tendance haute de la DSDS : 1,87 à 2 de 2018 à 2030 puis constance jusqu'en 2070
Espérance de vie	Gains d'espérance de vie parallèle à la tendance centrale France entière de la DSDS : EDV Femmes 85,4 ans en 2018 et 90 ans en 2070. EDV Hommes : 79,5 ans en 2018 et 87,5 ans en 2070	Gains d'espérance de vie parallèle à la tendance basse France entière de la DSDS : EDV Femmes 85,4 ans en 2018 et 86,5 ans en 2070. EDV Hommes : 79,5 ans en 2018 et 84,0 ans en 2070	Gains d'espérance de vie parallèle à la tendance haute France entière de la DSDS : EDV Femmes 85,4 ans en 2018 et 93,5 ans en 2070. EDV Hommes : 79,5 ans en 2018 et 91,0 ans en 2070
Migrations avec l'étranger	France entière : + 87 000 par an jusqu'en 2020 Puis +70 000 par an jusqu'en 2070	France entière : + 87 000 par an jusqu'en 2020 Puis +20 000 par an jusqu'en 2070	France entière : + 87 000 par an jusqu'en 2020 Puis +120 000 par an jusqu'en 2070

FIGURE 3.11 – Hypothèse des composantes (Source : https ://www.insee.fr/fr/information/2571308)

Le modèle utilise cet ensemble d'hypothèses, les états civils et le recensement de 2018 pour calculer des indices de fécondité, des soldes migratoires et des taux de mortalité. Divers aménagements ont été effectués pour adapter les données au Covid.

La projection est ensuite réalisée à un pas quinquennal, correspondant à la période de recensement (voir partie 3.1.2), puis annualisée.

Enfin, les données projetées sont retraitées avec une méthode de "calage" qui a pour but de garantir l'additivité du modèle. Sans cette étape, la somme de toutes les populations départementales n'est pas égale à la population totale.

A partir de ces données, et en conservant l'hypothèse de constance des populations sur l'année, il sera possible de passer à des taux de mortalité à la maille *sexe - tranches d'âge - département - semaine* à un nombre de décès. Les décès seront alors sommables, pour quantifier l'impact du climat sur la population à l'échelle nationale. Similairement aux scénarios DRIAS, considérer au minimum le scénario central et ceux de population haute et basse permettra d'estimer une incertitude.

3.2 Traitements

Cette partie a pour objectif la préparation des données afin qu'elles puissent être exploitées efficacement par les modèles, en transformant les données brutes en un format propre et structuré. Une base cohérente et nettoyée de toute anomalie qui pourrait pénaliser l'apprentissage est impérative.

Obtenir des données propres nécessite de réaliser plusieurs opérations, telles que :

— le nettoyage des données, pour éliminer les erreurs,
— l'analyse de valeurs manquantes, pour avoir une base complète,
— la recherche des éventuels doublons, pour garantir l'unicité des identifants.

Clairement définir le format souhaité est une priorité pour avoir une base structurée. La base finale doit contenir au minimum les informations suivantes :

— **Les identifiants** : caractéristiques de l'observation permettant de l'identifier de manière unique. Ces caractéristiques dépendent de la maille d'étude, par exemple pour ce mémoire elles comprendront **la date, le sexe, le département et la tranche d'âge**.

— **La variable cible** : information à projeter avec nos modèles de séries temporelles. Elle est communément représentée par la lettre y_t dans les équations de modèles. Dans cette étude, la variable cible est **un taux de mortalité**.

— **Les variables explicatives** : informations liées à l'observation. Pour les modèles de séries temporelles, ces variables explicatives apportent une information supplémentaire à la variable cible projetée. Elles sont communément représentées par la lettre $\mathbf{X}_t$ dans les équations de modèles. Dans le cas présenté, les variables explicatives comprennent notamment des informations climatiques sur **la température, l'humidité ou les précipitations**.

Identifiant *(maille)*				**y** *(cible)*	**X** *(var. explicatives)*		
semaine	département	sexe	tr_âge	taux_mortalité	température	humidité	précipitation

FIGURE 3.12 – Format cible de la base de données finale

3.2.1 Création de la base démographique

La première étape vise à créer la variable cible : les taux de mortalité. Plusieurs bases INSEE sont alors agrégées, comme présenté à la figure 3.2.

Base de données INSEE - Fichier des décès

La base présentée en partie 3.1.1, comprend des fichiers annuels des personnes décédées de 1990 à 2023, à concaténer ensemble. A ces fichier annuels sont ajoutés des fichiers mensuelles, des périodes de janvier à mai 2024, afin de prendre en compte des décès survenus en 2023 mais non déclarés.

Après imports et concaténations, la base comprend 19.9 millions de lignes. Le nettoyage des données a permis de remonter les erreurs suivantes :

— 10k lignes incomplètes,
— 480k lignes avec des anomalies, telles que des dates de décès antérieurs aux dates de naissance.

Étant données le volume conséquent de données et la difficulté d'imputation des valeurs, ces lignes sont supprimées de la base.

Les âges au décès sont calculés en millésime, à partir des dates de naissance et de décès. Le calcul en millésime suppose une répartition uniforme des âges de décès dans l'année, ce qui est acceptable pour une population de cette taille :

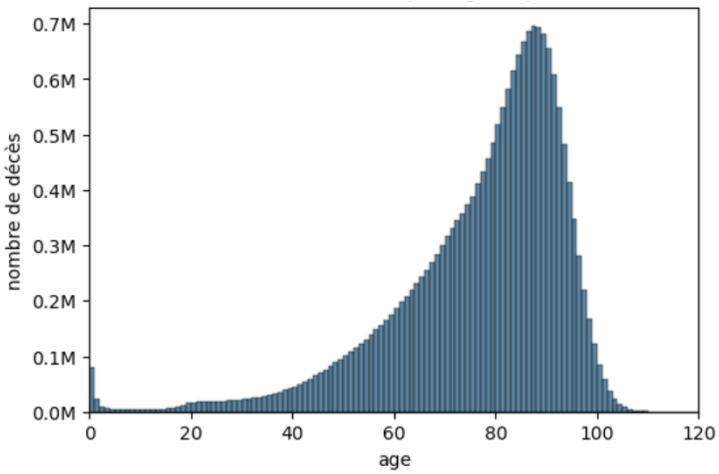


FIGURE 3.13 – Distribution du nombre de décès par âge depuis 1990

Des regroupements sont ensuite à réaliser pour créer plusieurs bases aux différentes mailles souhaitées.

La maille démographique distinguera toujours le sexe.

Lorsque les taux de mortalités sont à calculer par âge, ce qui sera le cas en partie 4.1, les âges supérieurs à 99 ans sont réduits à 99, conformément au formalisme de l'INSEE dans ses estimations de la population par âge. La maille *âge* comprend alors 100 modalités, allant de 0 à 99 ans.

Lorsque les taux de mortalité sont à calculer par tranches d'âges, ce qui sera le cas pour le cœur de l'étude, des groupes de 5 ans sont réalisés, conformément aux projections par tranche d'âges de l'INSEE (présentées en partie 3.1.5). La maille *tranche d'âges* comprend alors 20 modalités, allant de 0-4 à 95+ ans. Les tranches d'âges de 60 à 79 ans feront l'objet d'un suivi spécifique.

Pour la maille géographique, le département est obtenu à partir du code postal de la commune de décès. Seuls les décès concernant la France métropolitaine sont conservés. Un regroupement sur les départements est ensuite réalisable si la maille étudiée est la France métropolitaine, comme en partie 4.1.

Chaque décès correspond à une ligne dans les différentes bases obtenues (*sexe - âge - France* ou *sexe - tranche d'âges - département* par exemple). Un nombre de décès est alors calculable à la maille temporelle souhaitée (*jour / semaine / mois / année*). A la maille la plus fine, c'est-à-dire *sexe - âge - département - jour*, la base comprend plus de 47.5 millions de lignes.

A cette étape du traitement, les bases ont le format suivant :

Identifiant *(maille)*				**y** *(cible)*			**X** *(var. explicatives)*		
semaine	département	sexe	tr_âge	nombre_décès	population	taux_mortalité	température	humidité	précipitation
X	X	X	X	X					

FIGURE 3.14 – Format de la base de données après traitement de la base de décès INSEE

Base de données INSEE - Estimation de la population

Les secondes bases de l'INSEE nécessaires, présentées en partie 3.1.2, permettent d'obtenir des estimations de la population aux mailles souhaitées. Aucun traitement n'est nécessaire, la base étant déjà au format adéquat.

Identifiant *(maille)*				**y** *(cible)*			**X** *(var. explicatives)*		
semaine	département	sexe	tr_âge	nombre_décès	population	taux_mortalité	température	humidité	précipitation
X	X	X	X		X				

FIGURE 3.15 – Format de la base de données après traitement de la base de population INSEE

Agrégation

Les deux bases peuvent être agrégées afin d'obtenir la base démographique à la maille souhaitée, avec les taux de mortalité qui serviront d'historique d'entraînement aux modèles :

Identifiant *(maille)*				**y** *(cible)*			**X** *(var. explicatives)*		
semaine	département	sexe	tr_âge	nombre_décès	population	taux_mortalité	température	humidité	précipitation
X	X	X	X	X	X	X			

FIGURE 3.16 – Format de la base de données après traitement de la base de population INSEE

A cette étape, des projections de taux de mortalité sans intégrer de variables explicatives sont réalisables. La base représentant l'historique du risque doit être créée suivant le même format.

3.2.2 Création de la base d'historique météorologique

L'ajout de variables explicatives demande une source d'informations additionnelle qui doit être compatible. Les données météorologiques, présentées en partie 3.1.3, permettent de créer la base de variables explicatives de cette étude.

Base de données Météo-France - Données climatiques

Entre le réseau *ETENDU* et le réseau *SYNOP*, 2105 stations sont disponibles, identifiables avec les caractéristiques suivantes :

— Un numéro d'identification,
Les premiers chiffres correspondent au département, comme pour les codes postaux, permettant une localisation macroscopique rapide.
— Le nom de la station,
— Le réseau auquel elle appartient, '*ETENDU*' ou '*RADOME*',
— La date d'ouverture de la station,
— Sa géolocalisation, avec la latitude, longitude et altitude. Ces données permettront une localisation précise.

Une sélection de ces stations météorologiques est nécessaire :

— seules les stations du réseau *RADOME* sont conservées, leurs relevés étant plus fiables.
— les stations des départements d'outre-mer ne sont pas conservées.

— un filtre sur la date d'ouverture des stations est réalisé. Après les filtres précédents, les stations ouvertes au 1er janvier 1990 étaient au nombre de 289 (387 en 1999 et 561 en 2010). Rétroactivement, avec les traitements à suivre, il est décidé de conserver uniquement les 387 stations ouvertes au 1er janvier 1999, afin d'avoir au minimum une station pertinente par département.

La période d'entraînement commencera donc au minimum en 1999, bien que les données démographiques soient disponibles depuis 1990.

Les données sont disponibles par départements. Deux fichiers sont à concaténer par départements, correspondant aux variables météorologiques "classiques", telles que les températures, forces du vent ou précipitations, et aux variables "secondaires", telles que la pression, l'humidité, l'évapotranspiration ou la grêle.

Après agrégation des fichiers départementaux, **un premier tri parmi les variables est effectué afin de ne conserver que celles ayant leur correspondance dans les variables projetées par le DRIAS**. En effet, les variables qui n'ont pas de projection ne pourront pas être utilisées en tant que variables explicatives par les différents modèles.

Il est ensuite indispensable d'étudier l'exhaustivité des variables conservées. Pour cela, les taux de valeurs manquantes sont calculés par station. Par exemple, pour une sélection de variables, pour les stations des département 30 à 39 :

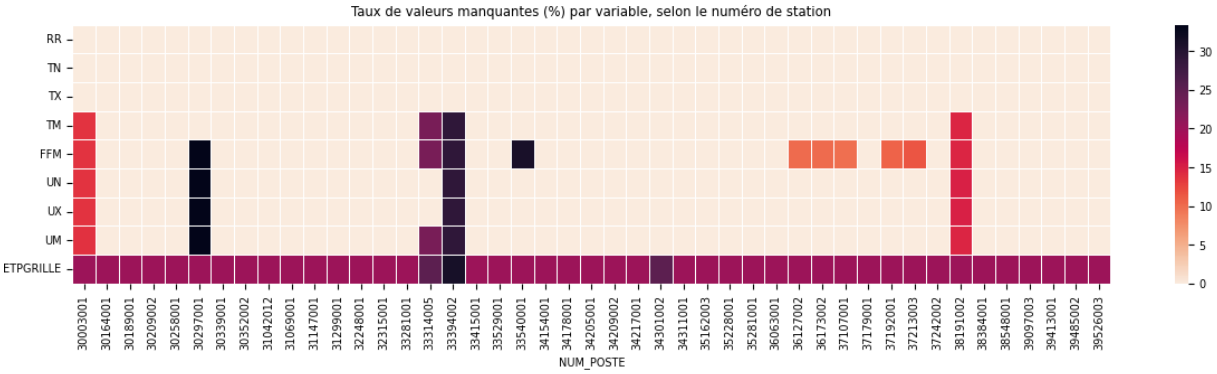


FIGURE 3.17 – Exemple d'étude de valeurs manquantes par stations et variables

Cet exemple permet notamment de certifier que la variable ETPGRILLE (évapotranspiration calculée au point de grille le plus proche) n'est pas à conserver.
De même, la première station, numéro 30003001, ne sera pas pertinente.

Cette étude permet de supprimer 70 stations. Les quelques valeurs manquantes restantes sont imputées par propagation avant, c'est-à-dire en supposant que chaque valeurs manquantes sont identiques à la précédente valeur observée.

Par rapport à l'objectif de création de base, la base est au format suivant :

Identifiant *(maille)*				**y** *(cible)*			**X** *(var. explicatives)*		
semaine	département	sexe	tr_âge	nombre_décès	population	taux_mortalité	température	humidité	précipitation
							X	X	X

FIGURE 3.18 – Format de la base de données Météo France après nettoyage

Traitement des variables explicatives

A l'issue de cette étape, les variables explicatives respectant à la fois les critères d'exhaustivité évoqués et ayant leur équivalent dans les projections du DRIAS sont conservées. Les variables suivantes pourront ensuite être retravaillées, notamment pour correspondre à la maille désirée :

— RR : quantité de précipitation tombée en 24h (de 6h du jour J et 6h du jour J+1), en [mm].
— TN, TM, TX : température minimale (N), moyenne (M) et maximale (X) de la journée sous abri, en °C.
— UN, UM, UX : humidité relative horaire minimale (N), moyenne (M) et maximale (X) de la journée, en %.

De premières visualisations sont alors réalisables, par exemple pour les variables TM et UM :

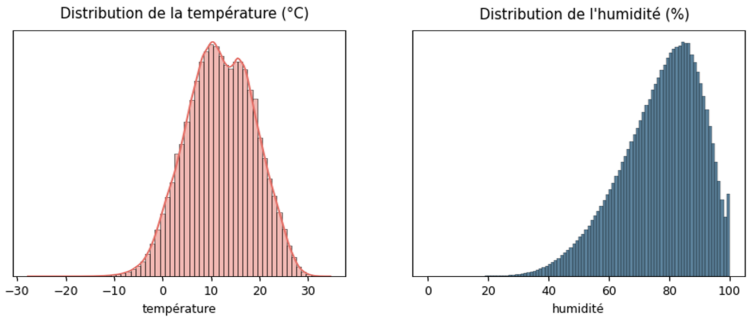


FIGURE 3.19 – Distribution des températures et humidités journalières moyennes observées sur l'ensemble des stations conservées, 1999-2023

Ce premier graphe permet d'observer la distribution des températures moyennes, avec deux pics identifiables à 10°C et 16°C .

Le second graphe présente l'humidité relative (cf point 2.1.2). Le pic a 100% peut être dû aux erreurs de mesure ou de conversion : l'humidité spécifique étant mesurée en $[kg_{eau}]/[kg_{air}]$ et est convertie en humidité relative à partir de la pression atmosphérique et de l'altitude. Les valeurs excédant 100% ont probablement été écrêtées.

Afin d'être utilisées à une maille temporelle moins fine, notamment à la maille *hebdomadaire*, des variables cohérentes doivent être créées. Ce travail permettra également, doit une moindre mesure, de prendre en compte des effets retardés ou non-immédiats.

Par exemple, pour la température, les variables suivantes sont ajoutées aux variables explicatives :

— 't_mean_7j' et 't_mean_30j' : moyenne de la variable TM sur les 7 / 30 derniers jours.
— 't_min_7j' et 't_min_30j' : minimum de la variable TN sur les 7 / 30 derniers jours.
— 't_max_7j' et 't_max_30j' : maximum de la variable TX sur les 7 / 30 derniers jours.
— 't_std_7j' et 't_std_30j' : écart-type de la variable TM sur les 7 / 30 derniers jours.
— 't_range_7j' et 't_range_30j' : différence entre les variables 't_max_7j' et 't_min_7j' (de même pour 30 jours).

Des variables similaires peuvent être créées pour l'humidité et les précipitations. Les variables sur les 30 derniers jours demandent d'avoir un mois supplémentaire d'historique de données.

La base est alors utilisable à la maille *hebdomadaire*, voire *mensuelle* :

Identifiant *(maille)*				**y** *(cible)*			**X** *(var. explicatives)*		
semaine	département	sexe	tr_âge	nombre_décès	population	taux_mortalité	température	humidité	précipitation
X							X	X	X

FIGURE 3.20 – Format de la base de données Météo France après travail des variables explicatives

Ces variables explicatives sont disponibles par stations, le passage à la maille *département* demande donc de travailler les variables géographiques.

Agrégation à la maille *département*

Les stations météorologiques peuvent être associées à leur département à partir de leur numéro d'identification. A partir de là, trois cas se présentent par département :

— le département contient une seule station. Aucun traitement n'est alors nécessaire, il est supposé que les relevés de la station en question représentent entièrement le département.

— le département comprend plusieurs stations. Ce sera le cas de la majorité des départements. Plus d'informations fiables sont disponibles : l'hypothèse que le climat est uniforme au sein des département est moins forte. Les variables explicatives sont alors moyennées pour ne conserver qu'une seule valeur par département.

— le département ne comprend aucune station. Ce sera le cas pour 5 petits départements, comme les Hauts-de-Seine et le Territoire de Belfort. La localisation exacte des stations est alors utilisée pour calculer la distance du centre du département à chaque station. Il est ensuite supposé que la station la plus proche du centre du département représentera le climat de celui-ci, en plus du sien. Le département est alors assimilable au premier cas, où une seule station est disponible.

Ce regroupement de station par département permet d'avoir une base à la maille souhaitée :

Identifiant *(maille)*				**y** *(cible)*			**X** *(var. explicatives)*		
semaine	département	sexe	tr_âge	nombre_décès	population	taux_mortalité	température	humidité	précipitation
X	X						X	X	X

FIGURE 3.21 – Format de la base de données Météo France traitée

La base d'historique météorologique est alors complète, et prête à être agrégée à la base démographique à partir du couple d'identifiant semaine - département :

Identifiant *(maille)*				**y** *(cible)*			**X** *(var. explicatives)*		
semaine	département	sexe	tr_âge	nombre_décès	population	taux_mortalité	température	humidité	précipitation
X	X	X	X	X	X	X	X	X	X

FIGURE 3.22 – Format finale de la base de données de l'historique

3.2.3 Traitement des bases de projections climatiques

Pour rappel, les extractions des projections du DRIAS des variables explicatives climatiques ont été présentées en partie 3.1.4.

Les données sont déjà exemptes de valeurs manquantes. L'unique traitement préalable nécessaire est d'associer chaque point à un département à partir de ses coordonnées. .

Base de données DRIAS - Projections climatiques

Les extractions permettent bien d'obtenir les précipitations en [mm] et les températures maximales, minimales et moyennes de la journée en °C. Cependant la seule variable d'humidité disponible correspond à une mesure d'humidité spécifique, exprimé en $[kg_{eau}]/[kg_{air}]$.

La conversion de l'humidité spécifique vers l'humidité relative demande de connaître la pression de l'air et la pression partielle de vapeur d'eau, qui dépendent de l'altitude. Ces données n'étant pas disponibles au sein des extractions, il est nécessaire de recourir à une approximation de la pression dépendant uniquement de la température.

Pour cela, l'approximation de Nadeau est utilisée [8], en supposant que la pression de l'air équivaut à la pression standard de 1013,25 hPa :

$$\mathrm{HS} = \frac{0,622\ p_{\mathrm{sat}}(\theta)\ \mathrm{HR}}{101\,325 - p_{\mathrm{sat}}(\theta)\ \mathrm{HR}} \text{ et } p_{\mathrm{sat}}(\theta) = \exp\left[23,3265 - \frac{3\,802,7}{\theta + 273,18} - \left(\frac{472,68}{\theta + 273,18}\right)^2\right]$$

où,

— θ est la température,

— $p_{sat}(\theta)$ est la pression de vapeur saturée, en Pa.

Les mêmes graphes que ceux présentés avec les données historiques Météo France sont réalisables :

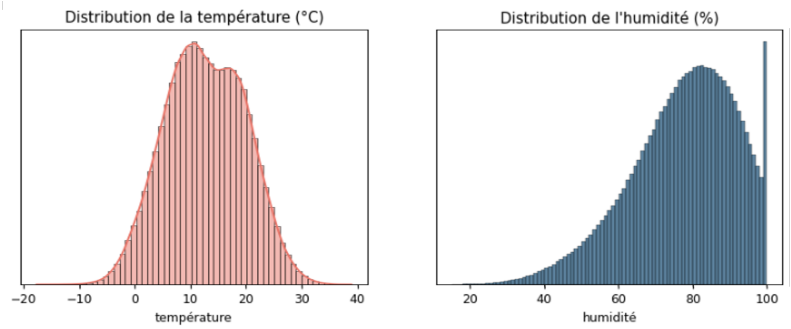


FIGURE 3.23 – Distribution des températures et humidités journalières moyennes projetées sur l'ensemble des départements, 2023-2050, simulation ALADIN63 RCP2.6

Sur le premier graphique, les deux pics observés à 10°C et 16°C sont bien identifiables. Comme attendu, la distribution des températures journalières moyennes projetées subie une translation vers les températures élevées par rapport à celles observées.

La distribution de l'humidité journalière moyenne projetée semble identique à celle observée. Le pic à 100% d'humidité relative, déjà relevé sur la distribution observée, est amplifié, possiblement à cause de l'approximation nécessaire à la conversion.

Enfin, le traitement réalisée afin d'obtenir des variables sur 7 et 30 jours sur les données Météo France est reproduit à l'identique. Ces travaux sont réalisés indépendamment pour chaque simulation DRIAS utilisée.

8. J.P. Nadeau, Séchage : des processus physiques aux procédés industriels, 1995

Les manipulations décrites ont permis d'avoir une base de données historique, aux différentes mailles souhaitées, comprenant des informations démographiques et des variables explicatives climatiques.

Couplée aux jeux de données représentant les variables climatiques projetées par le DRIAS, des taux de mortalités hebdomadaires peuvent également être projetés.

3.3 Statistiques descriptives : Taux de mortalité historiques

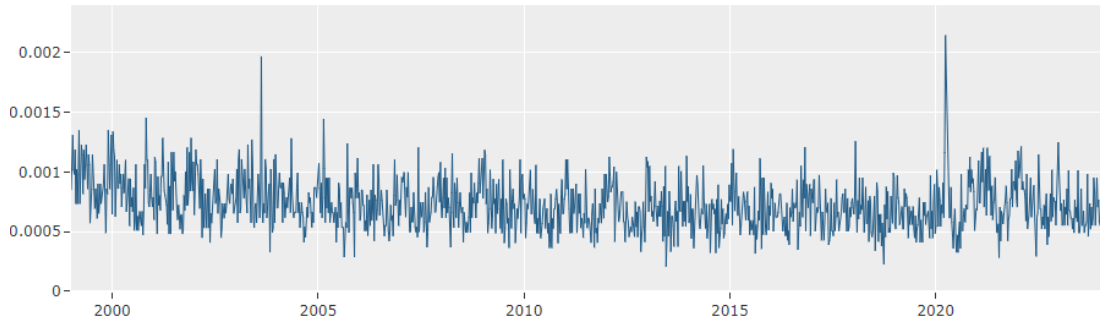
La base de données historique permet d'observer les taux de mortalités aux différentes mailles. Ces observations peuvent donner de premières informations sur les différents comportements attendus.

Comparaison des sexes :

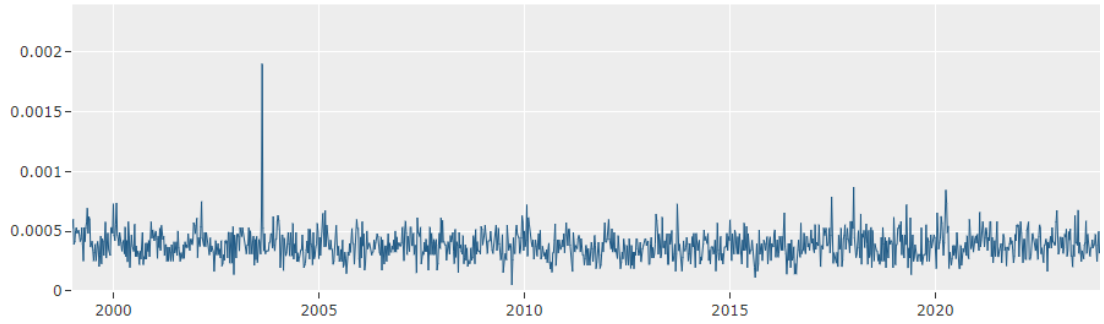
Il est observé que l'espérance de vie des hommes est plus faible.

D'après l'état des connaissances, présenté en partie 2.1, les taux de mortalités des femmes seraient plus influencés par le climat.

En comparant des taux de mortalités d'hommes et de femmes, pour une même tranche d'âge et un même département, il est donc attendu que les taux de mortalités des hommes soient plus importants que ceux des femmes, qui présenteront une saisonnalité plus marquée :



(a) Taux de mortalité, Hommes



(b) Taux de mortalité, Femmes

FIGURE 3.24 – Comparaison de taux de mortalité, tranche 75-79, département 75.

Pour cet exemple, les taux de mortalités des hommes sont plus importants que ceux des femmes, comme attendu.

Toutefois, la saisonnalité ne semble pas plus ou moins marquée pour un des deux sexes.

Le pic dû à la canicule de 2003 est ici plus prononcé chez les femmes, alors que celui lié au Covid l'est plus chez les hommes.

Comparaison des tranches d'âges :

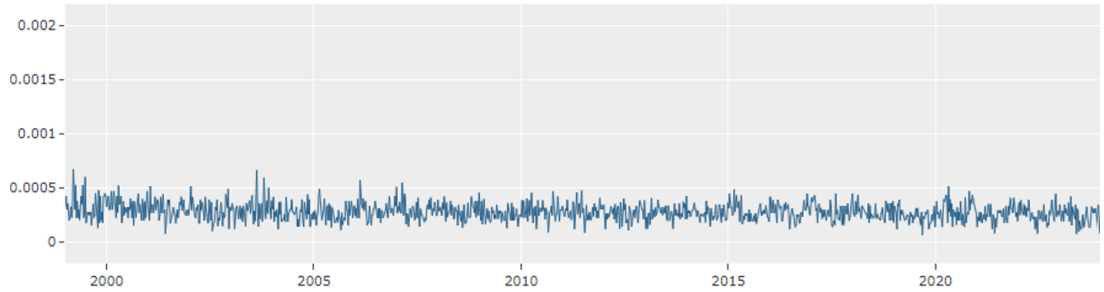
Pour un même sexe et un même département, la comparaison des taux de mortalité par tranches d'âges doit permettre de visualiser les différents effets connus :

— les taux de mortalités augmente avec l'âge, en particulier après 60 ans.

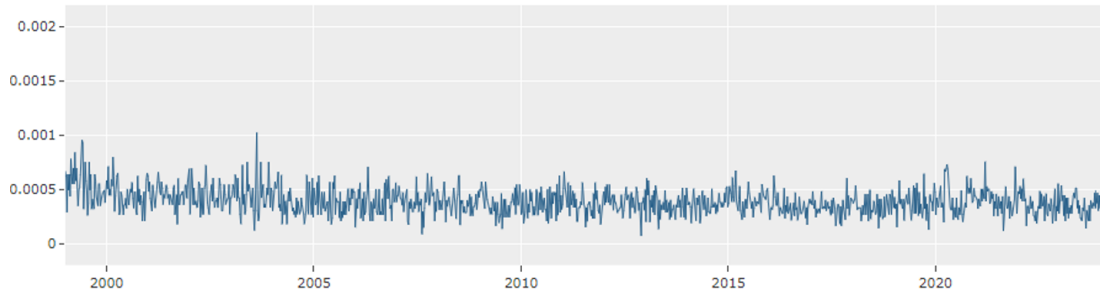
— la saisonnalité se renforce avec l'âge, signifiant que la période de l'année et en particulier le climat ont un impact plus important aux âges élevés.

— les évènements exceptionnels, tels que la canicule de 2003 ou le Covid, provoquent des pics de mortalité plus importants aux âges élevés.

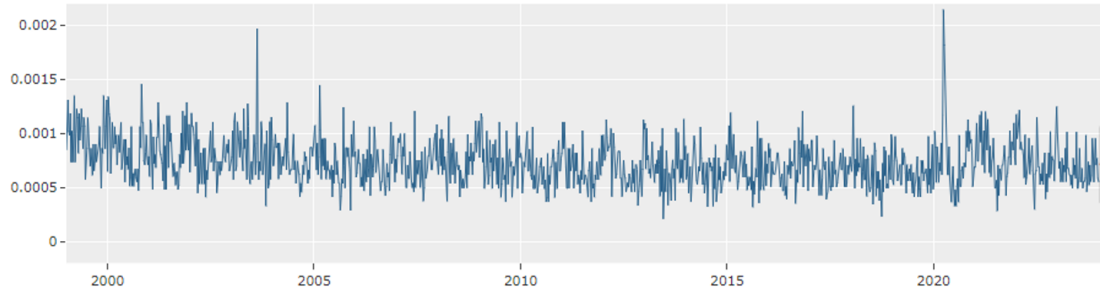
— la volatilité des taux de mortalité augmente aux âges extrêmes. En effet, plus de décès sont déclarés pour une population moins importante.



(a) Taux de mortalité, 60-64 ans



(b) Taux de mortalité, 65-69 ans



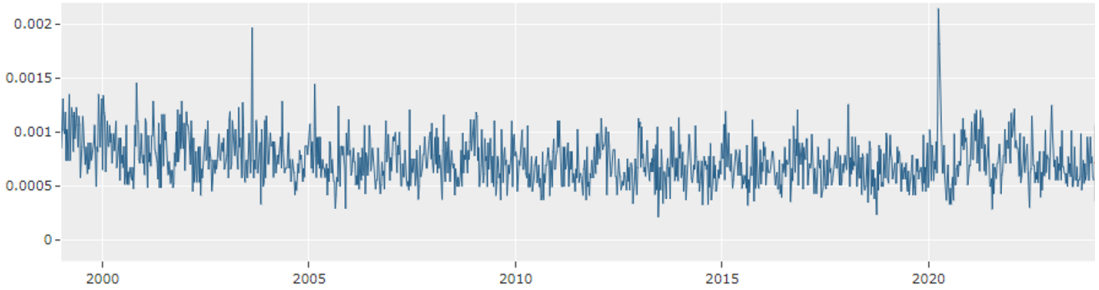
(c) Taux de mortalité, 75-79 ans

FIGURE 3.25 – Comparaison de taux de mortalité, hommes, département 75.

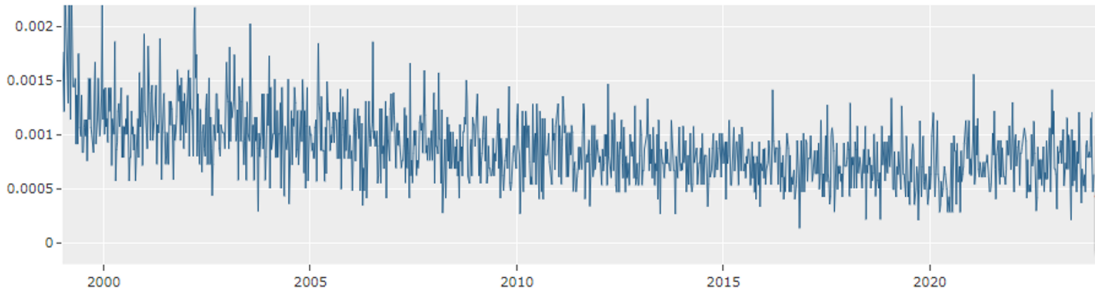
Les effets décrits sont bien observés :
— les évènements rares et la saisonnalité s'intensifient,
— les taux de mortalité et la volatilité augmente.
Il semble donc plus pertinent de se concentrer sur la modélisation d'âges élevés.

Comparaison des départements :

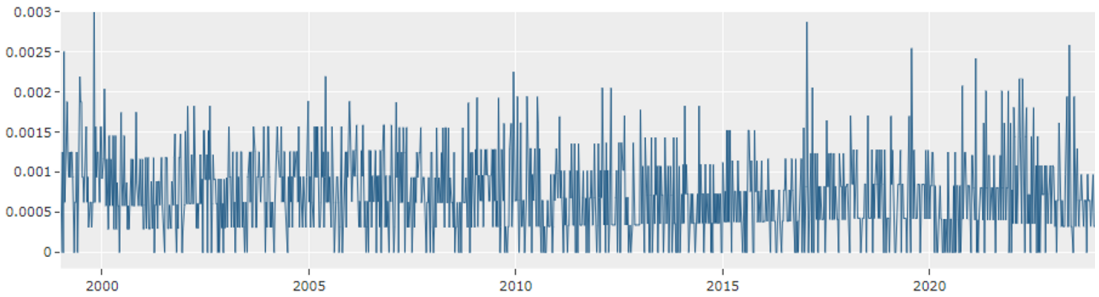
Pour un même sexe et une même tranche d'âge, il est possible de comparer les taux de mortalités par département. La pyramide des âges par département peut avoir une influence, notamment entre les départements ruraux et urbains. L'impact du département sur la mortalité pourrait être plus marqué par la suite, lorsque les variables climatiques seront ajoutées à l'étude.



(a) Taux de mortalité, département 75 - Paris



(b) Taux de mortalité, département 29 - Finistère



(c) Taux de mortalité, département 23 - Creuse

FIGURE 3.26 – Comparaison de taux de mortalité, hommes, tranche 75-79 ans.

Les taux de mortalité semblent augmenter pour les départements plus ruraux. La volatilité augmente pour les départements moins peuplés. Les pics de mortalités sont moins visibles.

Les modélisations seront probablement plus fiables pour les départements urbains.

Les bases de données ont été sélectionnées, importées, nettoyées puis retravaillées pour correspondre au format souhaité. Les taux de mortalité calculés sur l'historique sont cohérents aux attentes. L'entraînement des modèles de projections peut débuter.

Chapitre 4

Application : projection des taux de mortalité

Cette partie présente les différentes projections réalisées, et des détails sur :
- *la raison de leur mise en place,*
- *leur mise en pratique,*
- *l'interprétation des résultats.*

Tout d'abord, un modèle de durée prospectif, Lee-Carter, sera appliqué et comparé à des modèles de Prophet afin de s'assurer que la méthode à suivre est cohérente avec celles spécifiques aux études de mortalité.

Puis la maille sera affinée, ce qui permettra d'intégrer les variables climatiques à un modèle Prophet et de quantifier l'impact du climat sur la mortalité.

Enfin, un modèle Neural Prophet sera testé afin d'essayer d'améliorer les résultats et d'apporter une mesure de l'incertitude, calculée avec une méthode de prédictions conformes.

Comme pour tout modèle prédictif, le processus de projection des séries temporelles est circulaire, impliquant des itérations entre le modèle et l'analyse des résultats. Après l'application d'un modèle, des ajustements peuvent être nécessaires pour améliorer la précision. Cette approche permet d'optimiser et d'affiner les projections jusqu'à obtenir des résultats pertinents. Toutes les projections présentées dans cette partie ont suivi ce processus.

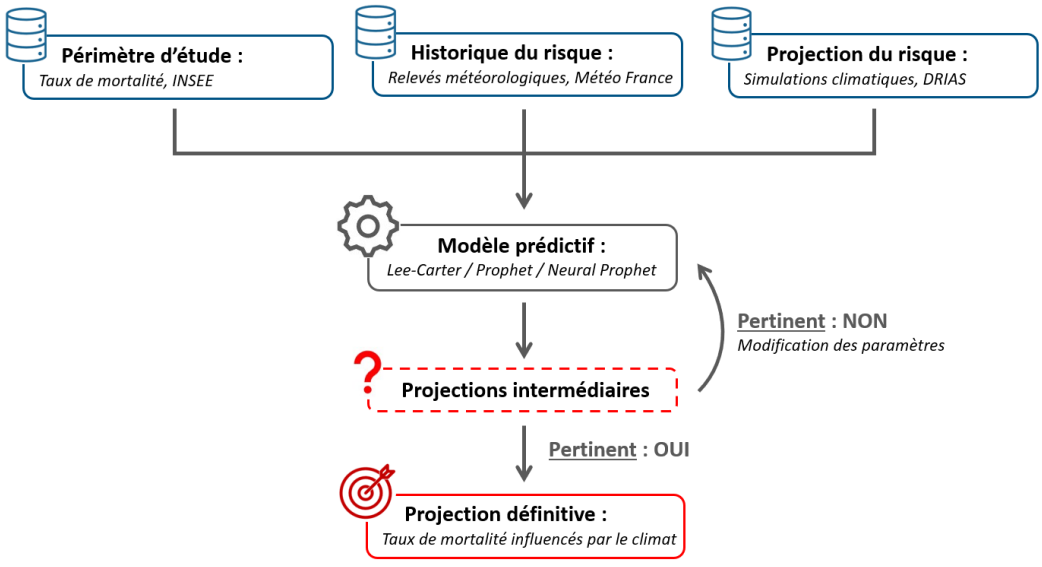


FIGURE 4.1 – Schéma du processus circulaire de projections

4.1 Comparaison à un modèle de durée prospectif : Lee-Carter

Ce mémoire n'a pas pour objectif de présenter ni d'éprouver des modèles de durée. Cette section vise à comparer un modèle de durée "classique" à la méthode basée sur Prophet, qui est le centre de ce mémoire. À partir d'une maille adaptée aux deux modèles, des projections similaires peuvent être confrontées. Ce rapprochement permettra de s'assurer que les projections du modèle de Prophet, appliquées à des taux de mortalité, sont cohérentes et conciliables avec les résultats obtenus via un modèle de durée.

Les modèles de durée n'étant pas le sujet principal de ce mémoire, les notations seront reprises, mais les détails mathématiques ne seront pas développés.

4.1.1 Les modèles de durée

Les modèles de durée sont des outils statistiques utilisés pour prédire le temps écoulé jusqu'à un événement, en fonction de diverses variables tels que l'âge ou le sexe. En actuariat, ces modèles sont principalement appliqués sur la mortalité pour estimer des durées de vie, mais peuvent aussi permettre d'estimer des durées en incapacité ou en arrêt de travail.

Les modèles de durée prospectifs, tel que le modèle de Lee-Carter, prennent en compte l'évolution de la mortalité dans le temps, afin de projeter des estimations dans le futur. Cette dépendance temporelle nécessite généralement le recours à des modèles de séries temporelles.

Dans le cadre de ce mémoire, les taux de mortalités calculés correspondent aux taux de décès bruts $m(t,x)$ suivant :

$$m(t,x) = m_{x,t} = \frac{\text{Nombre de personnes décédées entre les âges x et x+1 durant l'année t}}{\text{Population estimée de personnes d'âge x durant l'année t}}$$

Les périodes temporelles généralement considérées dans le contexte des modèles de durée sont exprimées en années. Les formules et résultats restent néanmoins équivalents lorsque les données sont à une échelle temporelle plus réduite, comme des mois.

Le périmètre étudié étant conséquent, l'hypothèse que la population est stationnaire est envisageable. Autrement dit, la pyramide des âges, représentant la répartition de la population par âge et par sexe à un moment donné, est supposée rester constante dans le temps. Dans cette situation, $m_{x,t} = \mu_{x,t}$, où $\mu_{x,t}$ est le taux instantané de mortalité d'un individu d'âge x en t, défini par :

$$\mu_t = \lim_{u \to 0} \frac{\mathbb{P}(T \leq t+u \mid T > t)}{u}$$

où T est une variable aléatoire dans $[0; +\infty[$ représentant la durée en vie d'un individu d'âge quelconque.

Dans le cas d'une étude sur une population spécifique, la notion de troncature et de censure serait à prendre en compte pour calculer un estimateur de $\mu_{x,t}$:

$$\widehat{\mu}_{x,t} = \frac{D_{x,t}}{L_{x,t}}$$

où :
— $D_{x,t}$ est le nombre de personnes décédées parmi les individus ayant l'âge x en t,
— $L_{x,t}$ est l'exposition associée.

La deuxième hypothèse nécessaire est que les taux instantanés $\mu_{x,t}$ soient supposés constants par morceaux, c'est-à-dire :

$$\mu(t+s, x+u) = \mu(t,x), \forall s, u \in [0;1[$$

Ainsi le taux de mortalité $q_{x,t}$ est approché par la relation suivante (démontrée en annexe 2) :

$$q_{x,t} = 1 - exp(-\mu_{x,t}) = 1 - exp(-m_{x,t}) \tag{4.1}$$

En pratique les taux de mortalités évoqués sont proches de zéro, donc l'approximation $m_{x,t} \approx q_{x,t}$ est vérifiée, ce qui permettra de comparer des projections de taux de mortalité $q_{x,t}$, obtenus avec un modèle de Lee-Carter, à des projections de taux de décès bruts $m_{x,t}$, obtenus via des modèles de séries temporelles. Cette approximation n'est toutefois plus vérifiée pour les âges élevés.

4.1.2 Le modèle de Lee-Carter

Le modèle de Lee-Carter est un modèle prospectif développé en 1992 [1], basé sur une décomposition multiplicative permettant de projeter des taux de mortalité par âge dans le futur à partir des données historiques observées.

Le modèle permet d'obtenir des taux de mortalité projetés $q_{x,t}$ à partir des taux observés $\widehat{\mu}_{x,t}$, de la relation 4.1 et de l'équation du modèle :

$$\ln \widehat{\mu}_{x,t} = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t} \tag{4.2}$$

où :
— $\widehat{\mu}_{x,t}$ est le taux de mortalité instantané observé à l'âge x en t.

— α_x est l'effet d'âge, correspondant à la valeur moyenne des $\ln \widehat{\mu}_{x,t}$. Ce terme constant dans le temps capture la tendance de mortalité à chaque âge, et peut être interprété comme l'effet moyen de l'âge sur le taux de mortalité.

— β_x est le coefficient d'ajustement par âge. Ce terme désigne la sensibilité du taux instantané à l'évolution temporelle globale capturée par les κ_t. Autrement dit, il quantifie l'impact de l'évolution temporelle sur le taux de mortalité pour un seul âge x. Ce paramètre représente un des atouts du modèle de Lee-Carter, car d'autres modèles de projection, tel que Prophet, ne peuvent pas prendre en compte cet effet.

— κ_t est l'effet temporel. Ce terme représente l'évolution temporelle générale. C'est le terme permettant de projeter dans le futur des $\mu_{x,t}$ grâce à un modèle de série temporelle,

— $\epsilon_{x,t}$ est le terme d'erreur aléatoire. Ce sont les fluctuations non expliquées par le modèle, supposées suivre une loi normale centrée de variance constante.

1. R.D. Lee, L.R. Carter, "Modeling and forecasting US mortality", Journal of the American statistical association, 1992

En pratique, le modèle est particulièrement utilisé pour son efficacité et sa facilité d'implémentation. Toutefois, il nécessite un nombre important de données pour être pertinent et l'hypothèse d'homoscédasticité nécessaire pour l'estimation des paramètres avec la méthode des moindres carrés n'est pas toujours vérifiée. Enfin, les changements soudains de tendance ou autres effets temporels inhabituels nécessitent d'adapter le modèle pour le coupler à un modèle de série temporelle plus adapté qu'un ARIMA, utilisé dans la majorité des cas pour projeter les κ_t.

4.1.3 Projections des taux par âge

Afin de réaliser un modèle de Lee-Carter pertinent, les données doivent de préférence être à la maille démographique *sexe - âge*, et non *sexe - tranche d'âge*.

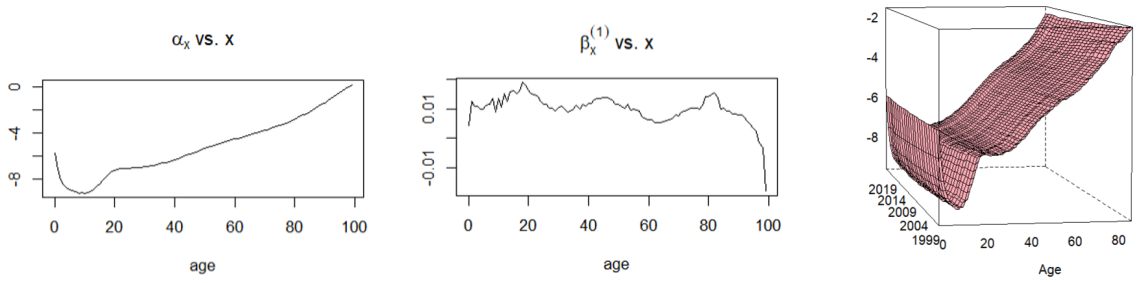
Le modèle de Lee-Carter est pensé pour être utilisé à la maille temporelle *annuelle*, mais peut néanmoins être adapté pour des données *mensuelles* par exemple. Le modèle Prophet est au contraire pensé pour la maille *quotidienne*, mais est parfaitement utilisable avec des données espacées dans le temps, pouvant représenter une maille *hebdomadaire, mensuelle ou annuelle*.

Affiner la maille géographique n'a pas d'intérêt pour ces projections, agréger toutes les données départementales ensemble à l'échelle nationale permet d'augmenter le nombre d'observations.

Avec ces informations, les projections sont réalisées à la maille *sexe - âge - France métropolitaine*, d'abord *annuellement*, puis *mensuellement*.

▷ **Projections annuelles :**

L'utilisation de données *annuelles* convient parfaitement au cadre défini pour le modèle de Lee-Carter. Les paramètres α_x, β_x et κ_t s'adaptent facilement aux données. Cependant, afin que le modèle soit rigoureux, il faudrait lisser les paramètres α_x et β_x afin de s'assurer que les surfaces de mortalité soient bien lisses et ne présentent pas d'aspérités.



(a) Paramètre α_x estimé (b) *Paramètre β_x estimé* (c) Surface de mortalité non-lisse

FIGURE 4.2 – Modèle annuel, paramètres estimés non-lissés et surface de mortalité, Hommes

La rupture de tendance en 2020 est clairement visible à la figure suivante 4.3, introduisant un nouveau questionnement : est-ce un phénomène conjoncturel ou structurel ?

Si cette hausse est conjoncturelle, c'est-à-dire seulement attendue sur quelques années, alors le modèle ne doit pas accorder trop de poids à ces données pendant l'entraînement. Si le phénomène est structurel, il est envisageable que cette hausse perdure, auquel cas le modèle doit répliquer cette hausse. Le paramètre temporel κ_t permet de la visualiser aisément, un simple modèle ARMA semble plutôt renvoyer une projection considérant que le phénomène est conjoncturel :

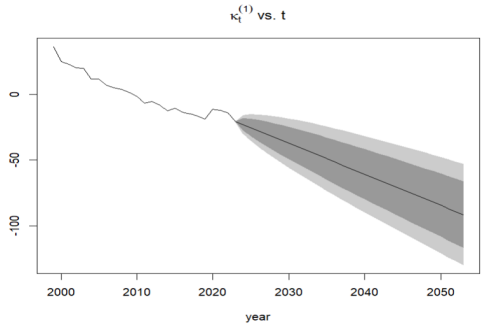


FIGURE 4.3 – Estimation et projection de κ_t, Hommes

Le paramètre κ_t est projeté à l'horizon 2050. A la maille annuelle, la projection se rapproche d'une droite.

En poussant la modélisation avec Lee-Carter, notamment en lissant les paramètres et en cherchant un modèle de projection des κ_t permettant de mieux prendre en compte la rupture de tendance, de meilleures projections pourraient être obtenues.

Bien que cela soit possible, la modélisation de données annuelles s'éloignent du cadre optimal de Prophet. En effet, le modèle a été conçu pour des données *quotidiennes*, ce qui lui permet de prendre en compte de multiples saisonnalités complexes. De plus, étant un modèle de *machine learning*, l'ajustement aux données suit le principe de train-test, ce qui demande un volume de données conséquent. Or, dans le cas de cette étude, utilisant des données de 1999 à 2023, passer à la maille *annuelle* revient à considérer seulement 25 observations, ce qui ne permet probablement pas à un modèle de *machine learning* d'apprendre correctement.

Il est techniquement possible de réaliser un modèle de Prophet avec des données *annuelles*, mais cela ne permet pas d'utiliser pleinement le potentiel du modèle. Pour comparer un modèle Prophet avec un modèle de Lee-Carter, réduire la maille temporelle sera privilégié.

▷ **Projections mensuelles :**

Lee-Carter :

L'utilisation de données *mensuelles* sort du cadre classique des modèles de durée. Les paramètres peuvent tout de même s'adapter aux données.

Le passage à une maille sub-annuelle permet d'observer une saisonnalité annuelle sur les κ_t.

Un modèle plus complexe que ceux implémentés par défaut est donc nécessaire. Il sera choisi de projeter les κ_t avec un modèle SARIMA, modèle usuellement privilégié pour des séries temporelles avec saisonnalité. Un SARIMA optimal est sélectionné en utilisant l'AIC (présenté en 1.1.1) :

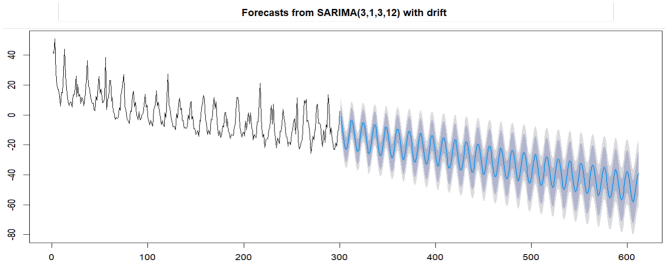


FIGURE 4.4 – Paramètres κ_t estimés avec Lee-Carter et projetés avec SARIMA, données mensuelles, Hommes

Les taux de mortalités $q_{x,t}$ sont ensuite reconstruits. Par exemple, pour les hommes, à 75 ans :

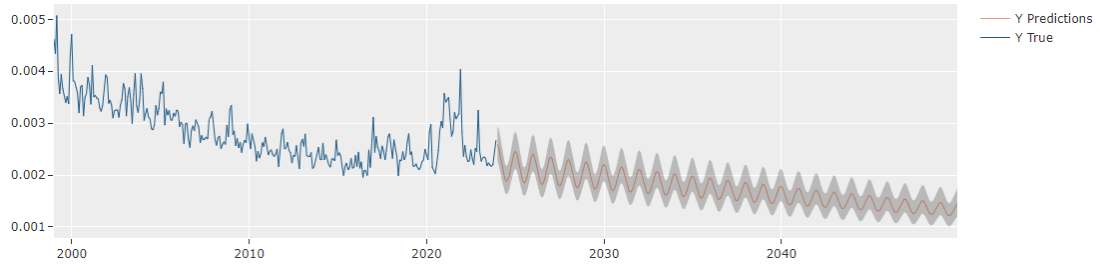


FIGURE 4.5 – Projection des taux de mortalité mensuels par Lee-Carter, Hommes, 75 ans

La période 1999-2023 représente les données mensuelles observées, ayant servi à l'ajustement des paramètres du modèle. Les projections sont réalisées pour chaque sexe par âge jusqu'en 2050.

Il est important de remarquer que l'intervalle de confiance associé à la projection prend uniquement en compte la tendance du modèle SARIMA utilisé pour les κ_t, ce qui ne représente qu'une faible partie de l'incertitude réelle. Il faudrait ajouter à cela un intervalle de confiance sur l'estimation des paramètres α_x et β_x et utiliser un intervalle de prédiction plutôt qu'un intervalle de confiance pour les κ_t.

Ces projections peuvent être comparées par âge et par sexe. L'exemple suivant présente ces projections pour les hommes, aux âges 60, 65 et 75 ans :

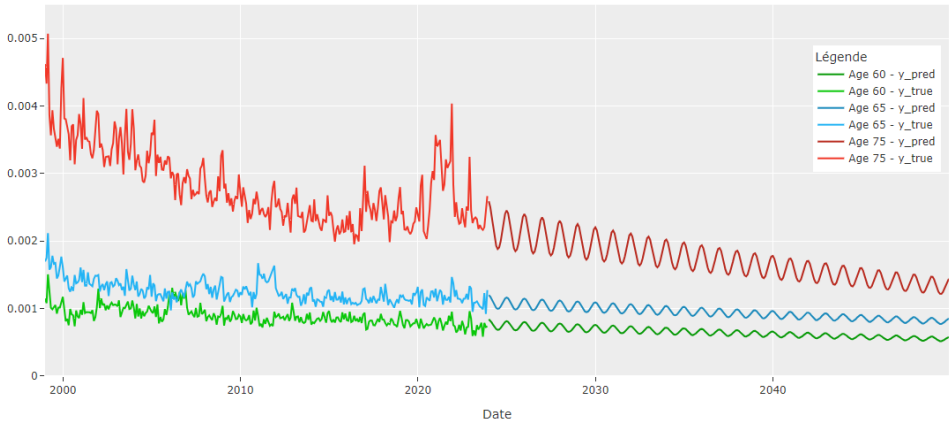


FIGURE 4.6 – Observations et projections avec un modèle de Lee-Carter à différents âges, Hommes

Les taux de mortalité projetés convergent à long terme, signifiant une poursuite de l'amélioration de l'espérance de vie. Les projections affichent une saisonnalité annuelle multiplicative, apportée par la projection des κ_t par le modèle SARIMA.

Prophet :

Les mêmes données sont retravaillées pour correspondre au formalisme de Prophet à la maille *mensuelle*. Afin de comparer les deux méthodes, les projections auraient pu être réalisées sur les κ_t du modèle précédent de Lee-Carter pour reconstruire les $q_{x,t}$. Cependant, par la suite le modèle Prophet sera directement utilisé pour projeter des taux bruts observés. Il a donc été décidé de projeter également les taux observés, comme évoqué avec l'équation 4.1 vu précédemment.

Avec 300 observations, le modèle peut s'entraîner suffisamment et ajouter une saisonnalité, pour produire un résultat satisfaisant :

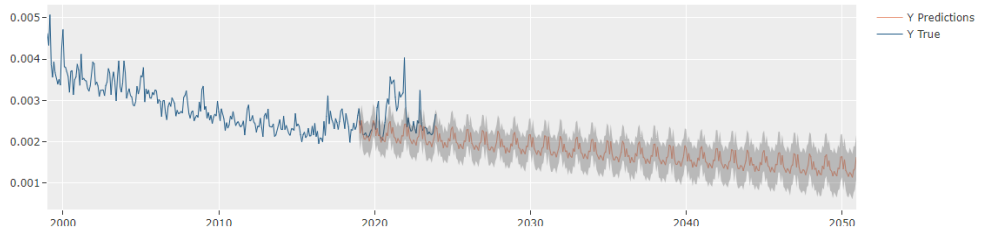


FIGURE 4.7 – Projection des taux de mortalité mensuels par Prophet, Hommes, 75 ans

L'intervalle de confiance associé reprend également uniquement le terme de tendance. Le paramètre permettant de calculer l'incertitude liée aux autres termes n'ayant pas d'utilité ici.

Avec plusieurs modèles, les projections peuvent être comparées entre elle par âges. Ainsi, avec 3 modèles, coïncidant avec les résultats présentés pour le modèle de Lee-Carter, Prophet réalise les projections suivantes :

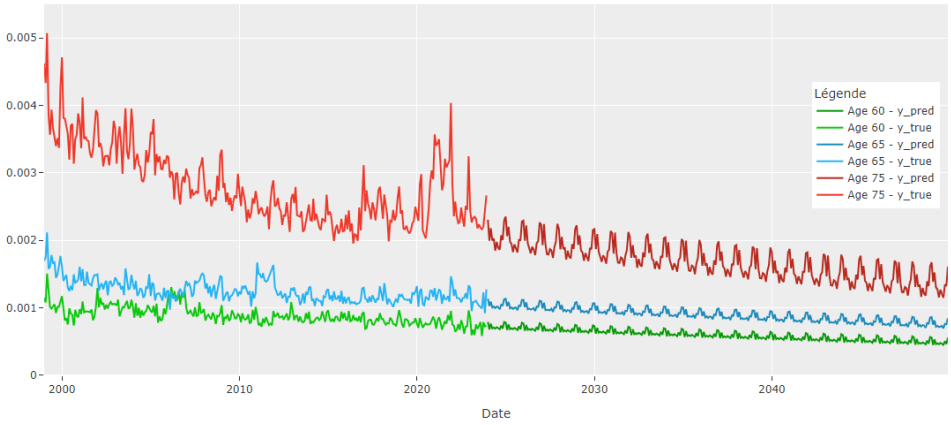


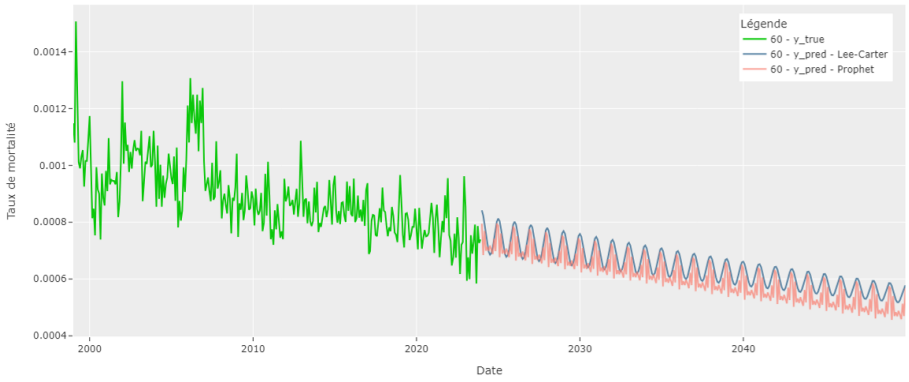
FIGURE 4.8 – Observations et projections avec des modèles Prophet à différents âges, Hommes

Chaque modèle est paramétré individuellement, mais tous détectent une saisonnalité annuelle complexe.

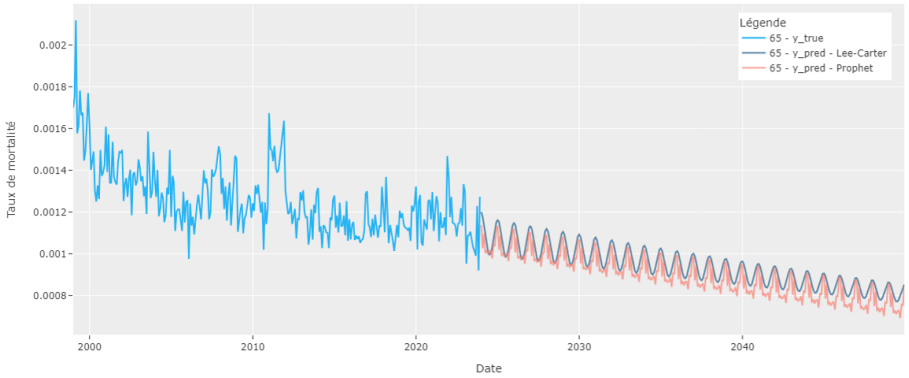
Dans la suite du mémoire, les âges seront regroupés en tranches de 5 ans, ce qui réduira la volatilité due à l'effet cohorte.

4.1.4 Comparaison et validation de la méthode

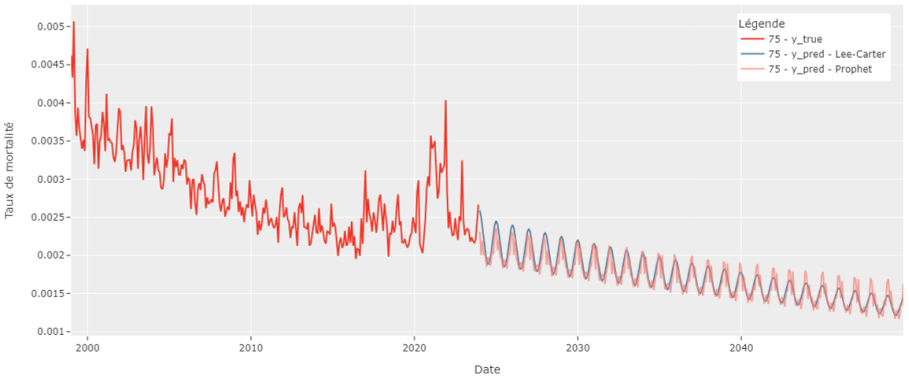
Les deux modèles ont permis de faire des projections à des mailles identiques. Pour conserver l'exemple suivi des hommes de 60, 65 et 75 ans, les résultats obtenus peuvent être comparés :



(a) Projections 60 ans, Hommes



(b) Projections 65 ans, Hommes



(c) Projections 75 ans, Hommes

FIGURE 4.9 – Comparaison des projections des taux de mortalité mensuels des modèles de Lee-Carter et de Prophet à différents âges, hommes

Une simple analyse visuelle permet de remarquer que les deux modèles s'accordent globalement sur la tendance et la saisonnalité. Les projections de Prophet captent cependant une saisonnalité plus complexe.

Les écarts de projections sont mesurés en calculant des MAPE (présenté en partie 1.2.3), représentant la différence absolue des prédictions de Prophet relativement à celles de Lee-Carter. Cela permet d'obtenir une estimation interprétable des écarts entre les deux modèles à plusieurs horizons :

Âge	Horizon	MAPE
60	2030	4,95%
	2040	6,04%
	2050	7,41%
65	2030	4,02%
	2040	4,41%
	2050	5,03%
75	2030	5,63%
	2040	5,14%
	2050	5,38%

FIGURE 4.10 – MAPE des modèles de Lee-Carter et Prophet à différents horizons

Ainsi, à horizon 2050, les prédictions des taux de mortalités mensuels des hommes de 65 ans par le modèle Prophet sont en moyennes à 5,03% des prédictions du modèle Lee-Carter.

Ce rapprochement nous permet de confirmer que le modèle de Prophet peut être adapté pour la projection de taux de mortalité. Il peut être noté que les modélisations par âge peuvent être imprécises, à cause de la sensibilité du modèle aux valeurs aberrantes, en particulier à la fin de sa période d'entraînement.

Le regroupement en tranche d'âge permettra de contourner ce problème, en plus de réduire le nombre de modèles à entraîner et d'augmenter le volume d'observations par segment étudié.

Au contraire, affiner la maille temporelle et géographique réduira le volume d'observations par segment. Toutefois, passer à une maille *semaine - département* permet d'intégrer les variables extérieures climatiques. De plus, passer à la maille hebdomadaire augmente le nombre de segments sur lesquels les modèles s'entraîneront.

La partie suivante, dédiée à l'intégration des variables climatiques aux modèles Prophet se fera à la maille *sexe - tranche d'âge - semaine - département*. Avec l'augmentation du nombre de données d'entraînement et l'intégration de régresseurs externes, il est attendu que les résultats à suivre soient de meilleure qualité.

4.2 Intégration du risque climatique aux projections avec Prophet

Cette partie fait appel aux éléments précédemment présentés, à savoir :
— la théorie du *machine learning*, en 1.2,
— la théorie du modèle Prophet et de son équation, en 1.4,
— les connaissances préalables sur le lien mortalité-climat, en 2.1,
— et la construction des données et des variables explicatives,en 3.2.

Ces différents éléments permettent d'aboutir au cœur de ce mémoire, à savoir la quantification de l'impact du climat sur les projections de taux de mortalité hebdomadaires avec le modèle Prophet.

Comme la majorité des modèles de séries temporelles classiques, le modèle Prophet ne permet pas d'avoir une approche multivariée, c'est-à-dire qu'il n'est pas possible de modéliser un lien de corrélation entre plusieurs séries temporelles parallèles. **Un modèle est incapable d'entraîner simultanément plusieurs séries temporelles similaires, distinguées par une variable, tels que le sexe, le département ou la tranche d'âge.**

Les détails de la maille ne peuvent pas non plus être considérées comme des variables explicatives comme pourrait le faire un modèle de *machine learning*. En effet, le modèle Prophet, comme les modèles de séries temporelles classiques, considère que les dates servent d'indexation, et ne sont pas des variables. Les dates doivent être uniques dans le jeu de données d'entraînement.

Cela contraint à approcher l'étude avec une vision locale plutôt que globale. **Il faudra utiliser un modèle par combinaison *sexe - tranche d'âge - département*, soit** $2 \times 20 \times 96 = 3840$ **modèles pour modéliser l'ensemble de la population de France métropolitaine avec une seule simulation climatique du DRIAS.**

Etant donné la puissance et le temps de calcul conséquent que cela représente, certaines hypothèses ou projections se contenteront d'étudier une population plus réduite par la suite. **En particulier, les tranches d'âges de 60 à 79 ans seront l'objet principal des interprétations finales.**

En effet, d'une part ces âges représentent des populations pour lesquelles le climat peut commencer à avoir un impact significatif.

D'autre part, hors de cette tranche d'âge, les données observées peuvent être trop volatiles.

Les modélisations nécessitent inévitablement de passer par une étape de paramétrage, bien souvent coûteuse en temps de calcul. Dans le cas présenté, le paramétrage prend en compte trois composantes :

— Les hyperparamètres. L'optimisation des résultats demande de réaliser un hyperparamétrage.

— La période d'entraînement. La rupture de tendance de 2020 demande une attention particulière.

— Le choix des régresseurs : les variables explicatives du modèle Prophet. Toutes celles créées ne sont pas à conserver, une sélection est nécessaire.

Dans l'idéal, il faudrait que ces composantes soient prises en charge simultanément lors du paramétrage (le choix des régresseurs pourrait être différent en fonction des hyperparamètres). Cela ne sera toutefois pas envisagé : l'entraînement des 3840 modèles par simulations prendrait plusieurs jours. **Les choix de ses paramètres se feront donc progressivement, en faisant l'hypothèse que les trois composantes décrites sont indépendantes.**

4.2.1 Hyperparamétrage

L'hyperparamétrage doit permettre de trouver une combinaison d'hyperparamètres optimisant les métriques d'un modèle. Néanmoins, il apparaît que le nombre important de modèles à réaliser ne permette pas un hyperparamétrage "complet". En effet, entraîner des modèles non-hyperparamétrés et les projeter suivant les deux simulations du DRIAS prend déjà près de 6 heures.

Le modèle Prophet comprend une liste assez restreinte d'hyperparamètres, dont les concepts sont présentés en partie 1.4. A partir de connaissances préalables ou de différents tests, certains hyperparamètres sont arbitrairement définis :

— '*growth*' : la tendance est à croissance logistique, et non linéaire, ce qui permet notamment de définir des seuils maximum et minimum.

— '*saisonnality_mode*' : la saisonnalité est multiplicative et non additive. En effet, une tendance à la baisse des taux de mortalités par âge est synonyme d'une augmentation de l'espérance de vie résiduelle. Les améliorations de l'espérance de vie sont en grande partie dues aux progrès de la médecine, notamment pour le traitement de pathologies découlant de maladies hivernales.

Une approche rigoureuse est adoptée pour d'autres hyperparamètres pouvant avoir une influence non négligeable. C'est en particulier le cas pour les hyperparamètres impactant la tendance. **En effet, la rupture de tendance de 2020 peut être conjecturelle, il sera alors attendu que la hausse des taux de mortalité après 2020 ne soit pas poursuivi, ou structurelle, signifiant au contraire que la hausse des taux se poursuivra.** Pour cela, un hyperparamétrage est effectué :

— '*changepoint_range*' défini la période sur laquelle les points de rupture de tendance peuvent être placés, compris entre 0 et 1.

La valeur par défaut de 0.8 signifie que la tendance n'évoluera pas sur les derniers 20% de la période d'entraînement.

Rapprocher ce paramètre de 1 peut permettre d'augmenter la plage de données utilisées, au risque de sur-apprentissage. De plus le modèle considérera que la rupture de tendance est structurelle et non conjecturelle.

— '*changepoint_prior_scale*' défini l'intensité des changements de tendance aux points de rupture.

Se contenter de deux hyperparamètres permet d'évaluer avec précision plusieurs centaines de modèles. Chacun est évalué sur les 12 combinaisons sélectionnées d'hyperparamètres. Une *heatmap* présentant la répartition des combinaisons permet de voir les plus fréquentes :

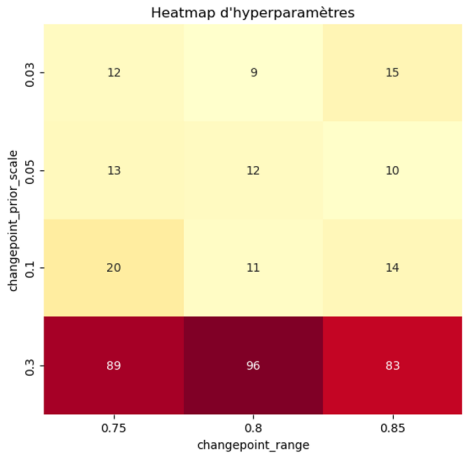


FIGURE 4.11 – Heatmap d'hyperparamètres, femmes, 60-79 ans, 30 départements

La répartition de *changepoint_range* semble uniforme, alors que la répartition de *changepoint_prior_scale* semble converger vers des valeurs élevées. Bien que les métriques semblent privilégier *changepoint_prior_scale* = 0.3, **il sera décidé de conserver les paramètres par défaut, à savoir *changepoint_prior_scale* = 0.1 pour éviter le sur-apprentissage que cela pourrait amener**.

4.2.2 Sélection des régresseurs

En amont de la modélisation, les variables explicatives créées, décrites en partie 3.2, représentent seulement des variables potentiellement intéressantes. Par construction, beaucoup d'entre elles sont très fortement corrélées. Une sélection est donc nécessaire afin de ne pas surcharger les modèles d'un surplus d'informations redondantes. Pour cela, une méthode devant apporter un critère de sélection individuel des régresseurs est d'abord envisagée.

L'objectif est d'apporter une mesure quantitative permettant d'évaluer la pertinence de l'ajout d'un régresseur à un modèle Prophet.

Pour une seule combinaison *sexe - tranche d'âge - département* :

— Un modèle Prophet permet d'obtenir des métriques de référence.
— Les n différents régresseurs possibles sont ajoutés individuellement au modèle de référence deux fois : additivement et multiplicativement (rappel partie 1.4.2).
— Les $2n$ modèles, avec un unique régresseur chacun, sont évalués puis comparés au modèle de référence.

Ces modèles sont entraînés et testés sur la période 1999-2019 afin de ne pas prendre en compte la rupture de tendance de 2020.

Cette méthode permet de voir l'apport individuel des régresseurs pour un des modèles possibles.

Afin de déterminer si certains régresseurs pourraient être ajoutés, quelque soit la maille considérée, la méthode est répétée avec plusieurs centaines de modèles différents. Il est alors possible de calculer les écarts de performance des modèles avec régresseurs avec les performances de leur modèle de référence. L'apport moyen, en pourcentage, apporté par les régresseurs est alors estimé, par exemple :

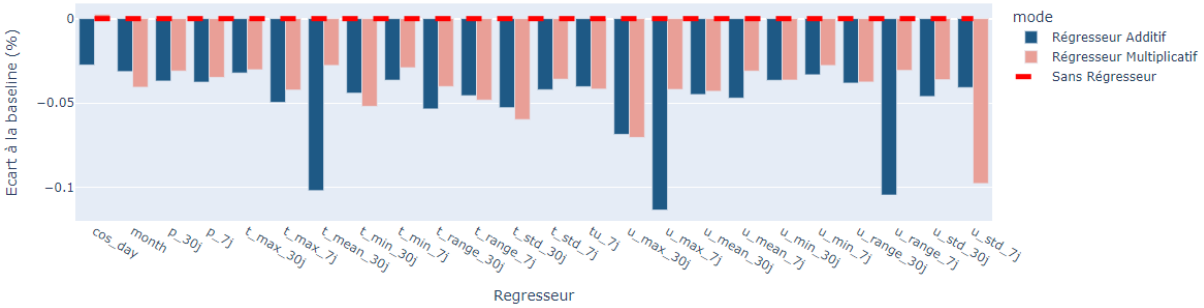


FIGURE 4.12 – Apport moyen des régresseurs

Ce graphe est toutefois à interpréter de manière informative, et n'apporte pas un critère de sélection fixe pour plusieurs raisons :

— La période de test peut contenir des valeurs anormales en fonction de la maille étudiée, par exemple le Covid sur les tranches d'âges élevées. L'effet de variables explicatives peut alors être "gommé", notamment parce que la période de test est restreinte.

— Les différentes variables climatiques ont un impact différent sur les populations. Les régresseurs ressortant comme améliorant le modèle ne seront peut-être pas les mêmes en Ile-de-France ou dans le Rhones-Alpes.

A partir ce cette étude il est décidé de conserver 5 variables explicatives, à la fois pour leur performance et pour leur interprétabilité. Elles formeront le premier ensemble de régresseurs *'v1'*.

Un second ensemble de régresseurs *'v2'* est envisagé, en y ajoutant la température humide, présentée en partie 2.1.2. Cette variable représente un phénomène spécifique connu ayant un impact sur la mortalité.

Les différents régresseurs qui seront utilisés sont récapitulés dans le tableau suivant :

Variable	Add (+)	Mltp (x)	Ensemble 'v1'	Ensemble 'v2'	Signification
t_max_7j	×		×	×	Température maximum sur les 7 derniers jours
t_mean_30j	×		×	×	Température moyenne sur les 30 derniers jours
u_max_7j	×		×	×	Maximum de l'humidité quotidienne moyenne sur les 7 derniers jours
u_max_30j		×	×	×	Maximum de l'humidité quotidienne moyenne sur les 30 derniers jours
u_std_7j		×	×	×	Ecart-type de l'humidité quotidienne moyenne sur les 7 derniers jours
tu_7j	×			×	Température humide, approximée à partir de *t_mean_7j* et *u_mean_7j*

4.2.3 Période d'entraînement

Un équilibre est à trouver entre sur-apprentissage et performance, sans prendre de décision arrêtée sur la rupture de tendance.

Les dernières observations sont les plus importantes en séries temporelles, toutefois, dans ce cas étudié, leur comportement est différent du reste de la série à cause de cette rupture de tendance. Il faut alors se demander quelles sont les observations à utiliser pour l'entraînement des modèles.

Pour cela, des hypothèses doivent être testées. Différents modèles sont alors entraînés :
— Sur des périodes d'observation différentes : allant de 1999 jusqu'en 2019 ou en 2023 ;
— Intégrant différentes valeurs du paramètre *changepoint_range* (0.80 par défaut ou 0.99) ;
— Le modèle Prophet acceptant les valeurs manquantes, en supprimant ou non la période de février 2020 à juin 2021 afin de prendre en compte la surmortalité due au Covid.

Les modèles sont ensuite testés a différents horizons sur toute la période d'observation. Cette nouvelle méthode permet également de comparer les ensembles de régresseurs sélectionnés.

Étant donné le nombre conséquent de modèles a entraîner (192 par tranche d'âge et ensemble d'hypothèses), il est encore décidé de se concentrer uniquement sur les tranches d'âges entre 60 et 79 ans.

Pour chaque combinaison d'hypothèses (période d'observation - *changepoint_range* - traitement Covid - ensemble de régresseur), $4 \times 192 = 768$ modèles sont entraînés puis évalués. Les métriques calculées sont moyennées pour déterminer les hypothèses qui seraient les plus pertinentes :

Le modèle conservé est le n°2.2, entraîné sur la période 1999-2023, avec traitement de la période Covid, et avec l'hyperparamètre *changepoint_range* fixé à 0.8, signifiant que les

n°	modèle	période observation	*changepoint range*	traitement covid	regresseur	mean_RMSE	mean_MAE
1.0	Prophet	2019	0,99	Non	no	0,18046	0,14138
1.1	Prophet	2019	0,99	Non	v1	0,17989	0,14097
1.2	Prophet	2019	0,99	Non	v2	0,17986	0,14093
2.0	Prophet	2023	0,8	Oui	no	0,17890	0,14046
2.1	Prophet	2023	0,8	Oui	v1	0,17857	0,14026
2.2	Prophet	2023	0,8	Oui	v2	**0,17848**	**0,14019**
3.0	Prophet	2023	0,8	Non	no	0,17893	0,14038
3.1	Prophet	2023	0,8	Non	v1	0,17879	0,14038
3.2	Prophet	2023	0,8	Non	v2	0,17876	0,14036
4.0	Prophet	2023	0,99	Non	no	0,17882	0,14033
4.1	Prophet	2023	0,99	Non	v1	0,17860	0,14029
4.2	Prophet	2023	0,99	Non	v2	0,17857	0,14028

FIGURE 4.13 – Tableau de métriques pour des modèles Prophet

derniers points de rupture ont été placés avant le Covid, en fin 2019. Le second ensemble de régresseurs *'v2'*, comprenant la température humide en plus des 5 autres régresseurs sélectionnés, est celui apportant les meilleurs performances.

4.2.4 Projections individuelles

Le modèle optimal est calibré et permettra de réaliser les projections souhaitées. Afin de s'assurer de la fiabilité des hypothèses retenues, plusieurs modèles sont appliqués. Seuls les résultats du modèle n°2.2 sont présentés par la suite.

Afin de quantifier l'impact du climat sur la mortalité, chaque modèle doit être entraîné au minimum deux fois : avec et sans les régresseurs choisis. Dans un second temps, afin de quantifier l'impact des scénarios climatiques, chaque modèle entraîné doit projeter deux fois les taux de mortalité : une première fois à partir de la simulation ALADIN63 du scénario "optimiste" RCP2.6 et une seconde fois avec la simulation ALADIN63 du scénario "pessimiste" RCP8.5.

Au sein de la littérature scientifique actuelle (ou du moins actuarielle), il semble difficile d'estimer un impact du climat sur la mortalité en hiver, au contraire des épisodes caniculaires, mieux documentés.

Dans la suite, deux points de référence sont à retenir :
— les canicules de 2032 du scénario RCP2.6,
— les canicules de 2027 du scénario RCP8.5.
Il est à noter que bien que les deux scénarios seront comparées, les simulations des scénarios RCP ne cherchent pas à être cohérentes entre-elles : aucune canicule particulière n'est attendue en 2032 pour le RCP8.5 par exemple.

Les différents phénomènes décrits à plusieurs reprises dans ce mémoire, à savoir l'influence de l'âge, du sexe ou du département, se retrouvent bien dans les projections individuelles.

Comme attendu, les effets les plus marqués et les plus fiables se retrouvent sur la tranche d'âge 60-79, par exemple :

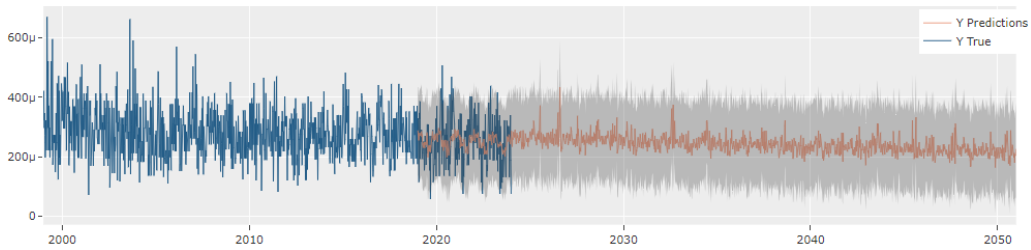


FIGURE 4.14 – Projection des taux de mortalité, hommes, 60-64 ans, département 75, RCP2.6

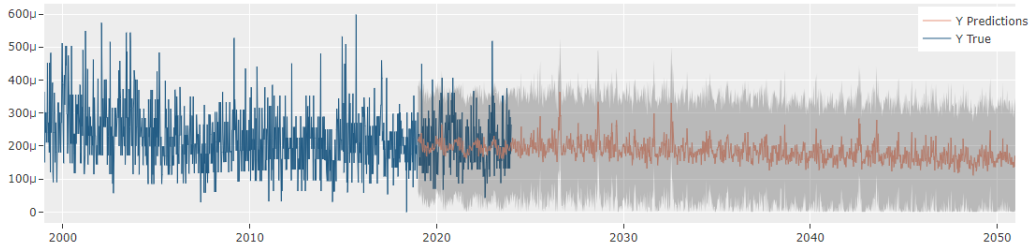


FIGURE 4.15 – Projection des taux de mortalité, femmes, 70-74 ans, département 69, RCP2.6

Une tendance à la baisse est observée, ainsi qu'un pic en août 2032, correspondant à la canicule attendue par ce scénario climatique. Il est intéressant de remarquer que l'intervalle de

confiance, obtenu avec l'approche bayésienne de Prophet, est de largeur relativement constante.

Cet effet sera plus visible en passant des taux à un nombre de décès, grâce aux estimations de croissance de la population de l'INSEE, présentées en partie 3.1.5.

4.2.5 Agrégation et comparaison par scénario

En supposant les populations constantes par années, les estimations de la population de l'INSEE sont disponibles à la maille souahitée (*sexe - tranches d'âges - département*), permettant de passer facilement d'un taux de mortalité à un nombre de décès. **Il est alors possible de sommer les décès en agrégeant les départements, sexes et tranches d'âges, pour avoir une estimation à l'échelle française :**

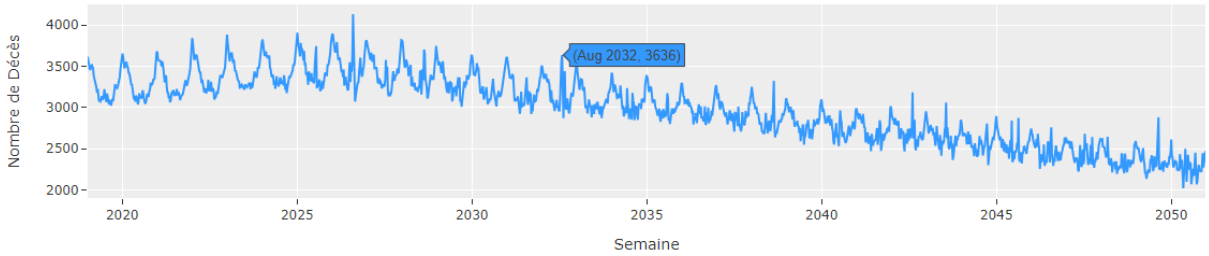


FIGURE 4.16 – Projection du nombre de décès hebdomadaire, RCP2.6, 60-79 ans

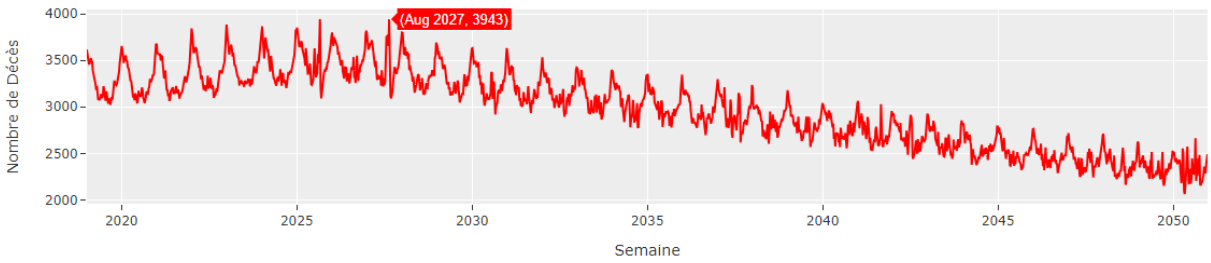


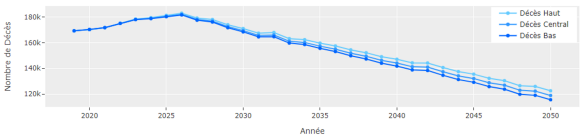
FIGURE 4.17 – Projection du nombre de décès hebdomadaire, RCP8.5, 60-79 ans

Cette représentation permet de confirmer plusieurs effets :

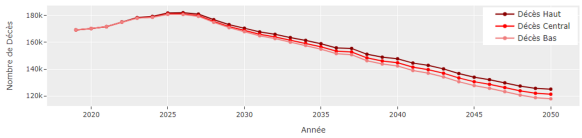
— La tendance à la baisse observée se poursuit à partir de 2025.
— La saisonnalité est bien marquée, avec des périodes hautes tous les hivers.
— Les canicules évoquées sont bien visibles sur l'exemple présenté : en 2027 pour le RCP8.5 et en 2032 pour le RCP2.6. Les nombres de décès hebdomadaires dépassent alors sur une courte période les décès hivernaux.
— La baisse de l'amplitude à la fin de la période, couplée à la multiplication des épisodes de fortes chaleurs, conduit à une augmentation de la volatilité sur les dernières années.

Il est également possible de sommer les décès hebdomadaires pour passer à une vision annuelle.

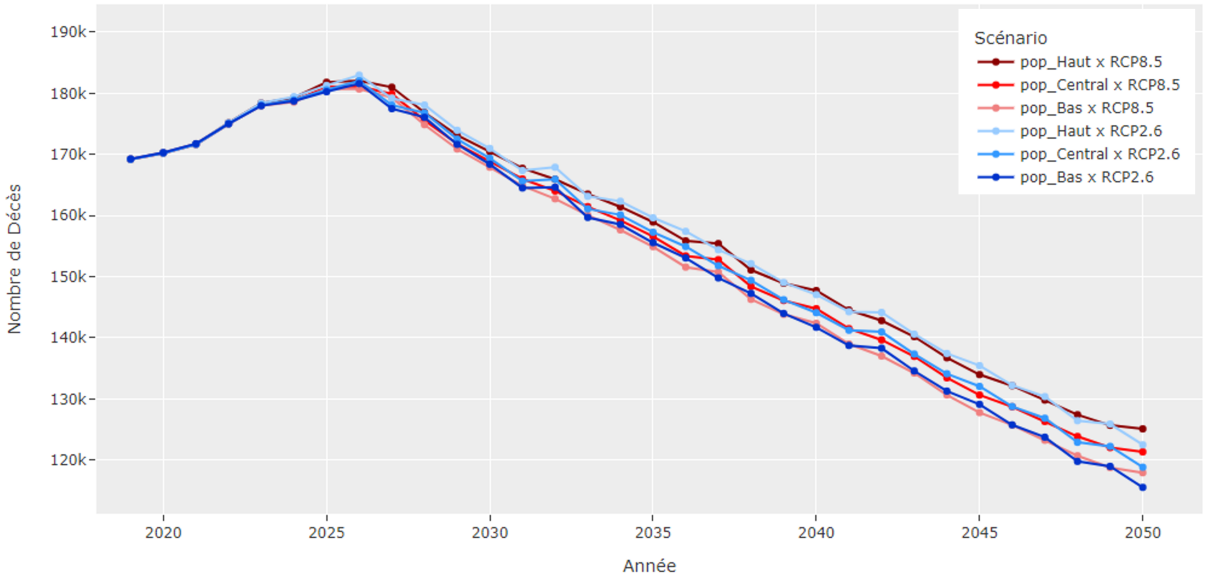
Trois scénarios de projections des populations de l'INSEE sont utilisés, permettant d'avoir une incertitude liée à l'évolution de la population. Couplés aux deux scénarios RCP, **6 trajectoires sont réalisables pour un même modèle**, par exemple :



(a) Projections annuelles, RCP2.6



(b) Projections annuelles, RCP8.5



(c) Comparaison des projections annuelles

FIGURE 4.18 – Comparaison des projections du nombre de décès en vision annuelle selon les scénarios RCP du DRIAS et les scénarios d'évolution de la population de l'INSEE, 60-79 ans

Il est intéressant de remarquer que les projections des scénarios RCP2.6 (le scénaro optimiste) et le 8.5 (le scénario pessimiste) ne sont pas clivantes. En effet, les deux scénarios sont conçus à horizon 2100 (voir figure 3.5).

Les différents scénarios de projections d'évolution de la population de l'INSEE provoquent des écarts allant de 2% à 8% du nombre de décès annuel à horizon 2050.

Les décès historiques sont ajoutés au graphe, afin de se rendre compte de la problématique de la rupture de tendance :

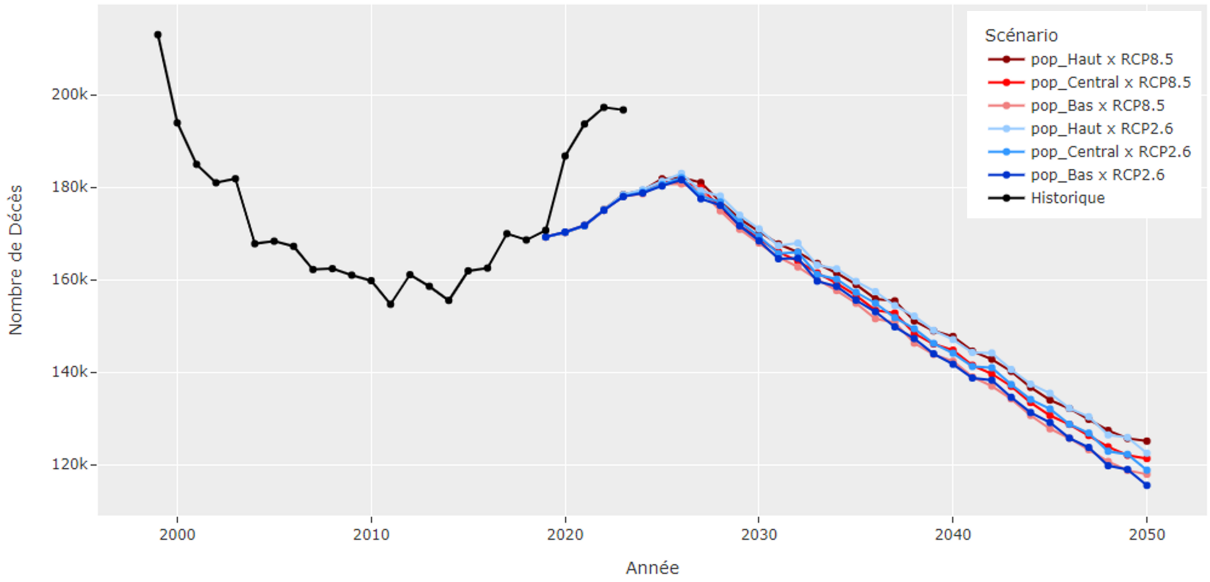


FIGURE 4.19 – Observations et projections du nombre de décès en vision annuelle selon les scénarios DRIAS et INSEE, 60-79 ans.

Ce même graphique, représenté en taux de mortalité pour 1000 habitants, permet d'interpréter les projections sans les hypothèses de projections de population. **Ainsi, de 2024 à 2050, les taux de mortalité observés chuteraient de 12‰ à 8‰ pour les tranches d'âges de 60 à 79 ans** :

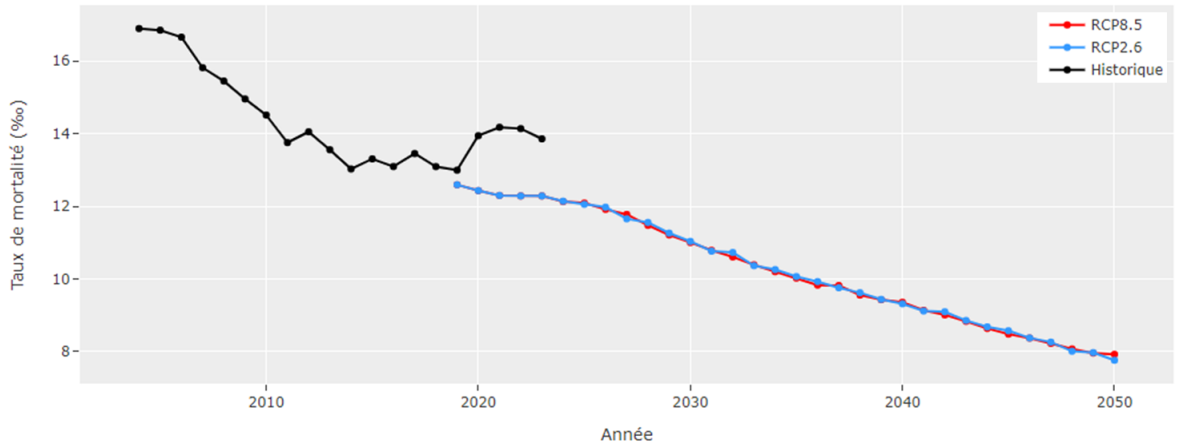


FIGURE 4.20 – Observations et projections de taux de mortalité , 60-79 ans.

Le modèle n°2.2 s'est entraîné avec toute la période d'observation (Covid traité) mais n'a pas modifié sa tendance après 2019. Il en résulte un modèle prenant en compte la rupture de tendance de 2020, mais considérant qu'elle est conjecturelle. La baisse des taux devrait reprendre après quelques années, comme sur la période 2000-2010. Bien que le décalage entre les observations et les prédictions du modèles jusqu'en 2025 soit conséquent, la tendance semble cohérente sur un horizon plus lointain.

Cet écart sur les premières années s'explique en vision hebdomadaire par les pics consécutifs occurrents en hiver sur les dernières années d'observation :

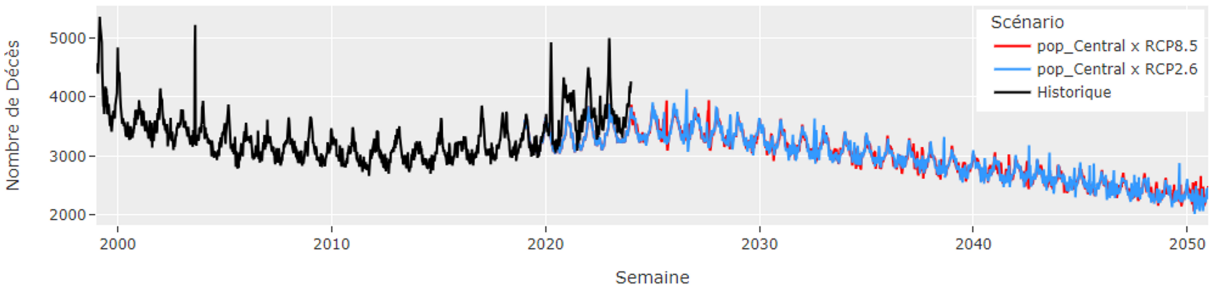


FIGURE 4.21 – Observations et projections du nombre de décès en vision hebdomadaire selon les scénarios DRIAS et INSEE, 60-79 ans.

Ce modèle capte bien les effets souhaités, sans avoir fait de sur-apprentissage en captant les pics de la fin de l'historique. L'évaluation de l'impact des régresseurs permettra de le valider.

▷ **Quantification de l'impact du climat :**

L'apport des régresseurs permet de quantifier l'impact du climat. Cependant Prophet ne permet pas de connaître l'apport individuel des régresseurs ajoutés aux modèles. Il est donc nécessaire de calculer la différence de mortalité entre un modèle sans régresseurs et un des modèles avec régresseurs présentés. Ainsi une mesure globale de l'impact du climat peut être apportée.

En vision annuelle, pour le modèle n°2.2 retenu, l'impact estimé peut être visualisé :

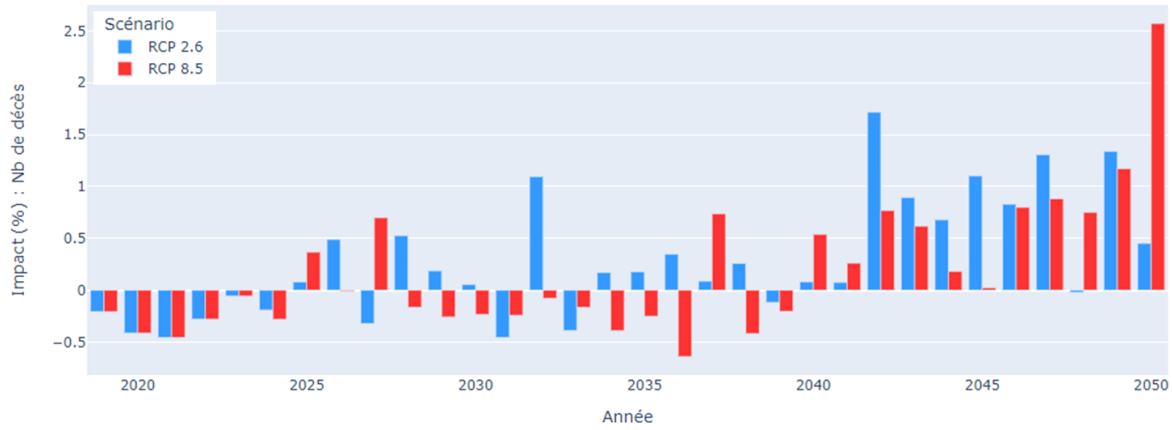


FIGURE 4.22 – Impact annuel des régresseurs en fonction des scénarios RCP

Les canicules de références en 2027 et 2032 sont bien visibles. La fréquence et la sévérité des canicules augmentant avec l'horizon de projection sont également facilement identifiables.

Si les années de surmortalité semblent être dus aux canicules, les années de sous-mortalité peuvent être expliquées par des hivers plus doux qu'habituellement, diminuant le nombre de décès hivernaux. Pour cela, la valeur moyenne annuelle du régresseur de température humide 'tu_7j' est ajoutée au graphe :

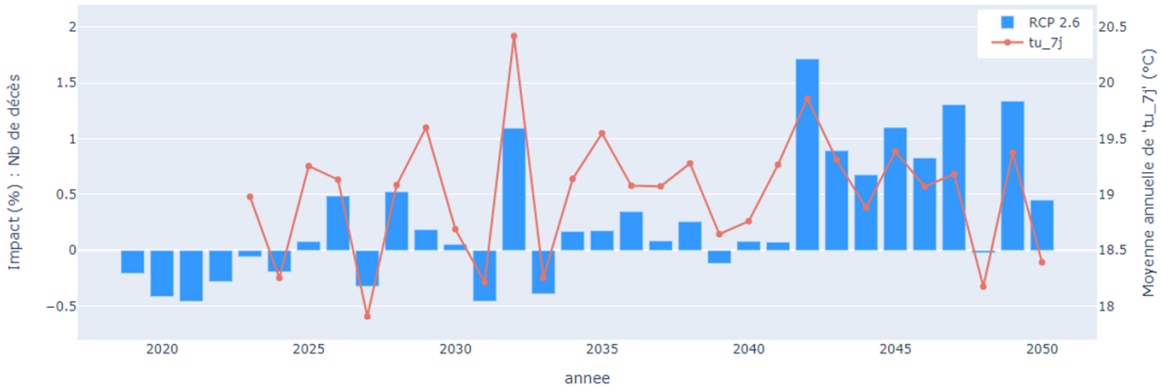


FIGURE 4.23 – Impact hebdomadaire des régresseurs et moyenne annuelle des températures humides, RCP2.6

Les effets des canicules sont particulièrement évident en vision hebdomadaire :

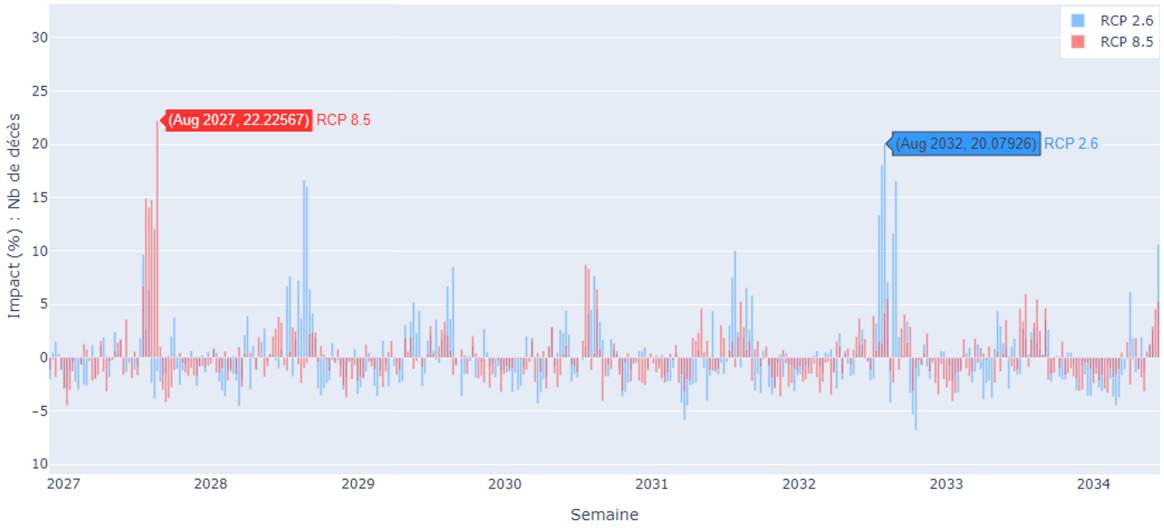


FIGURE 4.24 – Impact hebdomadaire des régresseurs en fonction des scénarios DRIAS, 60-79 ans, 2027-2034

Les impacts climatiques sont quantifiables, par exemple pour les canicules suivies :
- **en 2027, avec le scénario RCP8.5 : le modèle prévoit 2 796 décès dûs au climat sur la tranche d'âge 60-79 ans.**
- **en 2032, avec le scénario RCP2.6 : le modèle prévoit 2 631 décès dûs au climat sur la tranche d'âge 60-79 ans.**

A horizon 2050, les effets des régresseurs gagnent en amplitude, et les phénomènes extrêmes semblent de plus en plus fréquents :

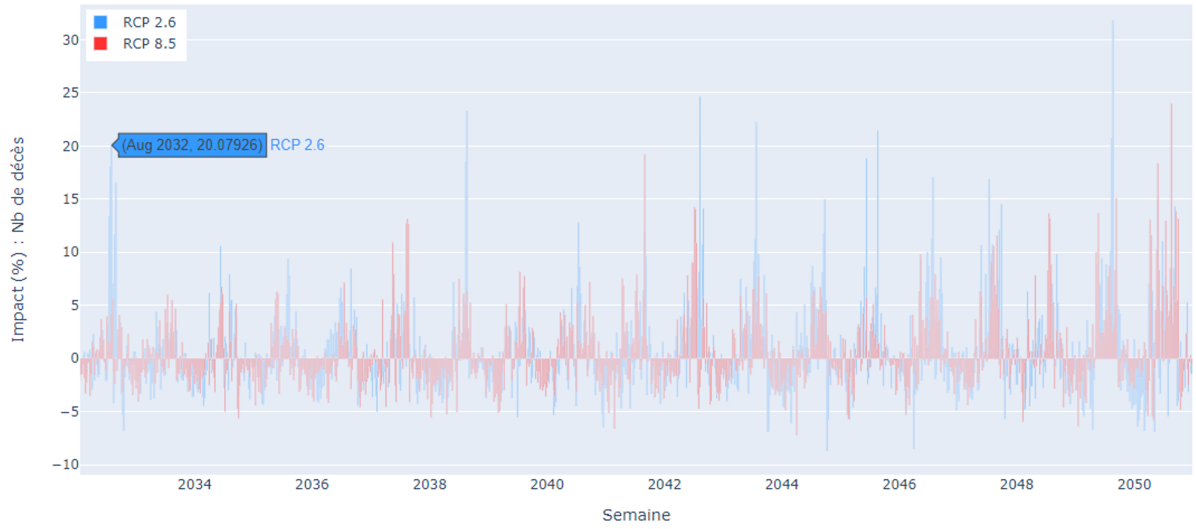


FIGURE 4.25 – Impact hebdomadaire des régresseurs en fonction des scénarios DRIAS, 60-79 ans, 2034-2050

Ceci se confirme par le nombre moyen de décès hebdomadaires dus au climat sur différents horizons :

Période	Scénario	Nb décès hebdomadaire dus au climat
2024-2050	RCP2,6	10,15
	RCP8,5	6,02
2030-2050	RCP2,6	11,8
	RCP8,5	7,14
2035-2050	RCP2,6	14,57
	RCP8,5	11,57
2040-2050	RCP2,6	19,24
	RCP8,5	19,06

FIGURE 4.26 – Nombre moyen de décès hebdomadaires dus au climat

Les moyennes sont croissantes avec l'horizon, ce qui montre encore une fois l'augmentation des phénomènes extrêmes après 2040.

Sur le périmètre considéré, le scénario RCP2.6 provoque une surmortalité hebdomadaire plus importante que sur le scénario RCP8.5.

Toutefois, le nombre de décès hebdomadaire sur la période 2040-2050 du scénario RCP8.5 rattrape celui du scénario RCP2.6.

Cela est cohérent avec les trajectoires des scénarios RCP : alors que le RCP2.6 voit une progression rapide puis constante à partir de 2050, le RCP8.5 suit une trajectoire linéaire jusqu'en 2100. Ainsi, il est attendu à ce que le nombre de canicules du RCP2.6 à horizon proche soit plus important que celui de RCP8.5, bien que ce soit le scénario "optimiste".

La volatilité des données par département ainsi que les différences de populations réduisent la confiance qu'il est possible d'avoir envers certaines projections des 768 modèles nécessaires à la modélisation départementale des tranches d'âges 60-79. La création d'intervalles de confiance ou de prédiction peut justement aider à la prise de décision dans ces situations.

▷ **Mesure de l'incertitude :**

Les mesures d'incertitude implémentées à Prophet ont été présentées en 1.4.3 :

— l'incertitude dans la tendance est captée avec une méthode personnalisée, consistant à réaliser des simulations avec des points de rupture futurs probables,

— l'incertitude de la saisonnalité recours à un calcul d'intervalle de confiance bayésien.

L'approche bayésienne permet d'avoir des intervalles de confiance de largeur constante mais flexible, comme vu à la figure 4.15.

En supposant que toutes les mailles de la population aient des comportements identiques, alors les bornes des intervalles peuvent être considérées comme étant additives.

Dans les faits, cette hypothèse est vérifiée aux grands âges : un évènement rare, provocant un pic de mortalité dépassant l'intervalle de confiance à 95% pour une maille de la population, provoquera probablement le même effet ailleurs. Par exemple la canicule de 2003 ou le covid sont en dehors de l'intervalle à 95% pour toutes les personnes âgées de plus de 60 ans, quelque soit leur sexe ou département.

En ajoutant les scénarios extrêmes de projection des populations de l'INSEE, et les simulations des différents scénarios du DRIAS, un intervalle de confiance est obtenu pour le nombre de décès hebdomadaire :

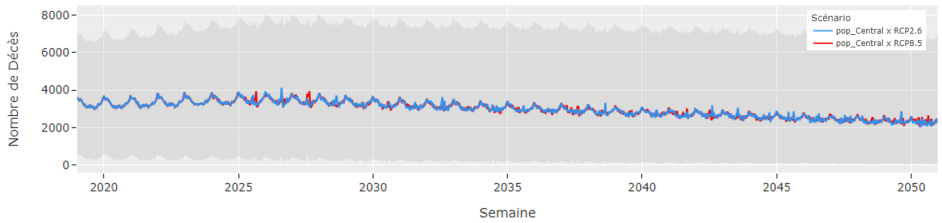


FIGURE 4.27 – Intervalle de confiance sur le nombre de décès hebdomadaire, 60-79 ans

Prenant en compte toutes les sources d'incertitudes quantitatives disponibles, le niveau de confiance de cet intervalle dépasse sûrement le niveau spécifié de 95%. **Alors que les nombres de décès hebdomadaires des âges 60-79 ans sont compris entre 3000 et 4000 avant 2030, l'intervalle construit en annonce un nombre compris entre 400 et 8000.**

Il n'est pas envisageable de fournir un intervalle de confiance annuel. Si agréger les projections en nombre de décès est possible, agréger les bornes de la maille *hebdomadaire* à la maille *annuelle* revient à considérer que les quantiles à 97,5% peuvent être atteints tous les jours sur une année, ce qui va à l'encontre même du principe de construction d'un intervalle de confiance.

D'autres sources d'incertitudes qualitatives sont également à retenir, telles que l'incertitude des mesures météorologiques de MétéoFrance, l'incertitude réflexive liée à l'élaboration des scénarios RCP par le GIEC, ou encore l'incertitude des modèles de projections climatiques.

Introduire une autre approche avec Neural Prophet permettrait de calculer des intervalles de prédictions plus fiables que celui construit.

4.3 Ouverture avec Neural Prophet

Le modèle Neural Prophet, présenté brièvement en partie 1.5, doit améliorer le modèle Prophet en y ajoutant de nouveaux termes. Le modèle se complexifie afin d'intégrer des termes auto-régressifs, faisant appel à des réseaux neuronaux. **La complexité temporelle était déjà la principale limite de l'étude. L'application à quelques modèles permet de confirmer que la méthode n'est pas transposable et permet d'utiliser les méthodes de *conformal predictions* intégrées au modèle.**

▷ **Paramétrage du modèle :**
L'ajout de termes auto-régressifs pourrait améliorer les résultats. Cependant, dans le contexte de l'étude, les temps de calculs nécessaires ne permettent pas de les utiliser.

En théorie, sans ajouter de termes auto-régressifs, les deux modèles seraient identiques. Ainsi, le paramétrage des meilleurs modèles Prophet est reproduit sur un ensemble réduit de modèles Neural Prophet. Sur la tranche d'âge 60-79 ans, les performances suivantes sont obtenues :

n°	modèle	période observation	*changepoint range*	traitement covid	regresseur	mean_RMSE	mean_MAE
1.0	Prophet	2019	0,99	Non	no	0,18046	0,14138
1.1	Prophet	2019	0,99	Non	v1	0,17989	0,14097
1.2	Prophet	2019	0,99	Non	v2	0,17986	0,14093
2.0	Prophet	2023	0,8	Oui	no	0,17890	0,14046
2.1	Prophet	2023	0,8	Oui	v1	0,17857	0,14026
2.2	Prophet	2023	0,8	Oui	v2	0,17848	0,14019
3.0	Neural Prophet	2023	0,8	Oui	v1	0,17623	0,13622
3.1	Neural Prophet	2023	0,8	Oui	v2	0,17622	0,13622
3.2	Neural Prophet	2023	0,8	Oui	no	**0,17520**	**0,13548**
4.0	Neural Prophet	2019	0,99	Non	no	0,17619	0,13635
4.1	Neural Prophet	2019	0,99	Non	v1	0,17749	0,13718
4.2	Neural Prophet	2019	0,99	Non	v2	0,17753	0,13723

FIGURE 4.28 – Tableau de métriques pour des modèles Neural Prophet

Suivant la même méthodologie appliquée à Prophet, un nombre de décès annuels à horizon 2050 est calculé :

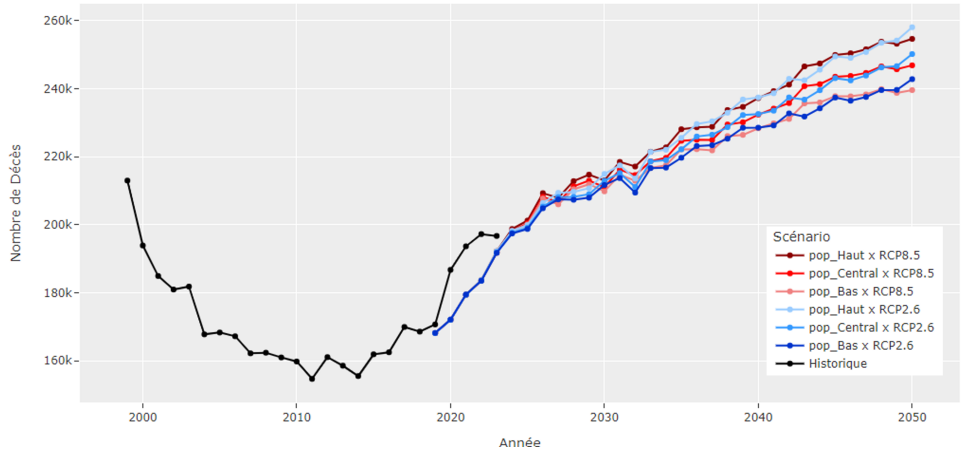


FIGURE 4.29 – Observations et projections du nombre de décès en vision annuelle, 60-79 ans, modèle Neural Prophet

Malgré l'attention portée aux hyperparamètres de tendance et à la période d'entraînement, **les modèles Neural Prophet envisagés ne répliquent pas la rupture de tendance de 2020 conjecturellement comme souhaité, mais structurellement.**

▷ **Quantification de l'incertitude :**

Des hyperparamétrages réalisés sur un nombre très restreint de modèles a permis d'avoir des intervalles de prédictions pertinents, construit avec la méthode de *conformal prediction* intégrée au modèle :

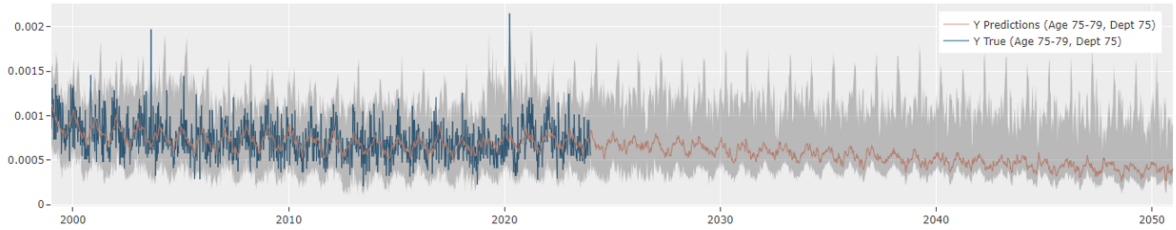


FIGURE 4.30 – Projections et intervalles de prédictions des taux de mortalité, hommes, 75-79 ans, département 75, Neural Prophet

L'intervalle obtenu apporte plusieurs améliorations à ceux obtenus avec les méthodes bayésiennes de Prophet :

— l'intervalle n'est pas centré. En effet, la méthode de *conformal prediction* utilise les quantiles des résidus, ce qui permet d'identifier que la borne inférieur est stable : il ne peut pas y avoir d'évènement soudain provoquant une sous-mortalité conséquente.

— l'intervalle de prédiction suit la saisonnalité. Hors canicules et Covid, les évènements pouvant dépasser le quantile supérieur à 97,5% ont lieu lors du pic hivernale annuel. L'intervalle de prédictions réplique cet effet naturellement.

— les deux effets cumulées conduisent à un intervalle à la fois moins large et plus précis que celui de Prophet.

La méthode de *conformal prediction* est particulièrement bien adaptée à l'étude, et permettrait d'avoir de réels intervalles de prédictions fiables et pertinents. Cependant, son implémentation n'est pas encore intégrée à Prophet, et Neural Prophet est un modèle trop complexe pour un tel volume de données.

Neural Prophet semble pouvoir être plus performant que Prophet, notamment grâce aux termes auto-régressifs utilisant des réseaux neuronaux. L'ajout de régresseurs retardés pourrait permettre de modéliser plus précisément d'autres phénomènes connus, tel que l'effet moisson.

Les méthodes de *conformal predictions* implémentées semblent être plus efficaces que les autres évoquées au long de ce mémoire. Seulement l'approche naïve a été testé ici mais la *Conformalized Quantile Regression* (CQR) est également implémentée au modèle, et serait a priori plus performante.

Dans cette étude, la complexité temporelle a été la contrainte principale. L'application de toutes les méthodes intégrées au modèle Neural Prophet en ont été fortement entravés. Malgré cela, des modèles Prophet ont pu apporter des résultats satisfaisants et cohérents.

Conclusion

Dans une démarche de gestion des risques, les organismes d'assurance se doivent d'identifier et de mesurer les risques auxquels ils font face. Le risque que fait peser la crise climatique en cours sur les risques de longévité et de mortalité est maintenant bien identifié, mais les quantifications de ce risque à long terme sont encore au stade de développement.

Ce mémoire s'inscrit dans la démarche d'amélioration de l'appréhension actuarielle de ce risque de mortalité/longévité. Nous avons proposé une méthode numérique et statistique innovante se basant sur Prophet, modèle récent de de machine learning pour les séries temporelles, et son extension utilisant des réseaux de neurones Neural Prophet, pour projeter, à partir des données *open-source* démographiques de l'INSEE et climatiques de Météo France, les taux de mortalité à horizon 2050 selon deux trajectoires de réchauffement.

Ces modèles permettent, comme démontré dans le mémoire, de capter des ruptures de tendances et des saisonnalités complexes.

Dans une démarche actuarielle, nous nous sommes également intéressés aussi à la mesure des incertitudes autour des projections en mobilisant l'outil mathématiques qu'est la "prédiction conforme".

Sur le périmètre retenu, les personnes âgées de 60 à 79 ans en France métropolitaine, les modèles utilisés ont évalué une hausse des taux de mortalité annuels de l'ordre de 0,25%, effets combinés d'une hausse lors des pics de chaleur et d'une baisse lors d'hivers plus doux. Bien que loin de la valeur du choc S2 sur la mortalité, ce résultat moyen ne doit pas masquer les pics de surmortalité qui auront lieu lors des épisodes de fortes canicules. On retrouve en effet pour les années les plus marquées par les canicules, comme l'année 2042 pour le scénario RCP2.6 et 2050 pour le scénario RCP8.5, des variations de 1.6% et 2.5% des taux de mortalité, en relative cohérence avec le taux de 1,83% proposé par l'ACPR comme choc pour les 65-74 ans pour le scénario court terme qui fait l'hypothèse de 2 canicules très fortes en 2023 et 2024.

De tels résultats nous montrent que dans un pays développé comme la France, l'impact direct des températures sur la mortalité reste contenu. Toutefois, la modélisation des extrêmes climatiques évolue, il convient donc à la communauté actuarielle de continuer à suivre ce risque, à perfectionner la modélisation des impacts, en particulier sur les jours les plus chauds et les plus humides qui peuvent avoir des impacts très forts, comme l'a montré la canicule de 2003 : augmentation de la mortalité totale de 55% entre le 1er et le 15 août selon Santé Publique France [2].

Les travaux effectués ont également permis de mettre en évidence les forces des outils utilisés, à la fois Prophet et Neural Prophet, ainsi que des méthodes de mesures d'incertitude, telle la prédiction conforme. Ces outils pourraient s'appliquer à de nombreuses autres problématiques actuarielles : étude des rachats et résiliations, écoulement des provisions, impacts de variables socio-économiques sur la mortalité, ...

La suite de cette étude pourrait être double. D'un point de vue pratique, une telle étude pourrait être reproduite dans le cadre de scénarios ORSA pour l'évaluation interne des risques. Appliqués à un portefeuille d'assurés, l'organisme d'assurance pourra ainsi évaluer la mortalité

2. Santé Publique France, Impact sanitaire de la vague de chaleur en France survenue en août 2003, 2019

spécifique à son portefeuille se voir évoluer du fait de la crise climatique en intégrant les impacts sur l'ensemble de l'année, hiver et été. Une intégration de tels scénarios est à la fois une demande du régulateur et semble être une nécessité en termes de gestion du risque assurantiel.

D'un point de vue académique, des réflexions pourraient être menées pour une projection jointe de l'ensemble de la population, non permises avec l'outil Prophet, en intégrant les forces des deux modèles utilisés. Des améliorations peuvent aussi être faite pour améliorer la modélisation des queues de distribution et des impacts extrêmes.

Bibliographie

ACPR, Présentation des hypothèses de l'exercice climatique assurances 2023

A. N. Angelopoulos & S. Bates, A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification

Y. Romano, E. Patterson & E. J. Candès, Conformalized Quantile Regression

A. Ringeade & J-P. Vidal, Les GCM, des grosses machines pour simuler le climat futur, mais pas que . . ., Février 2024

Canouï-Poitrine F, Cadot E, Spira A., Surmortalité liée à la canicule d'août 2003 à Paris, France

DRIAS, Les méthodes de correction

DRIAS, Les données disponibles

DRIAS, Aide à la sélection, Scénario d'émission RCP2.6

Galea & Associés, Impact de la météo et plus particulièrement des pics de chaleur humide sur la surmortalité, 2023

INSEE, Fichiers des personnes décédées depuis 1970

INSEE, Le modèle de projection démographique Omphale 2010

INSEE, "53 800 décès de plus qu'attendus en 2022"

INSEE, Projections de population 2018-2070

INSEE, Estimation de la population au 1 janvier 2024

INSEE, Pyramides des âges au 1 janvier 2024

M. Gover, *Mortality during periods of excessive temperature*, 1938

W. Robson, The Math of Prophet, 2019

IDEA, Les algorithmes de Machine Learning pour la prévision des séries temporelles -Partie I-, 2024

N. Kourentzes, *Additive and multiplicative seasonality*, 2014

Météo-France, Données climatologiques de base - quotidiennes

Météo France, Canicule, pic ou vague de chaleur ?, 2023

Maraun, D. Bias Correcting Climate Change Simulations - a Critical Review (2016)

M. Berk, Prophet vs. NeuralProphet, 2021

M. Dei, (Almost) everything you should know to use Facebook Prophet, 2020

J.C. Orduz, Forecasting Weekly Data with Prophet, 2021

V. Flunkert & J. Gasthaus, DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks, D. Salinas, 2019

V. Manokhin, Conformal Prediction forecasting with MAPIE, 2022

V. Manokhin, How to evaluate Probabilistic Forecasts)

V. Manokhin, How to predict quantiles in a more intelligent way, 2021

S. J. Taylor & B. Letham, Forecasting at scale, 2017

S.C. Sherwood, *An adaptability limit to climate change due to heat stress*, 2010

S.S. Shapiro & M.B. Wilk, « *An analysis of variance test for normality (complete samples)*», Biometrika

R. Stull, *Wet-Bulb Temperature from Relative Humidity and Air Temperature*, 2011

Santé Publique France, Influence de caractéristiques urbaines sur la relation entre température et mortalité en Île-de-France, 2020

Santé publique France, Évaluation de la surmortalité pendant les canicules de 2015

Santé Publique France, Système d'alerte canicule et santé : principes, fondements et évaluation, 2023

Stan : *Probabilistic Programming with Bayesian inference*. https://mc-stan.org/

Pyro : *Deep Universal Probabilistic Programming*. https://pyro.ai
The Uncertainty of Predictions, C. Stevenson)

Table des figures

Annexes

1 Démonstration : décomposition biais-variance de l'erreur quadratique

L'égalité présentée en 1.2.4 décompose l'erreur de prédiction quadratique d'un modèle. Supposons qu'il existe une relation expliquant chaque observations y_i :

$$y_i = f(X_i) + \epsilon_i$$

où ϵ_i est le bruit, tel que $\epsilon_i \sim \mathcal{N}(0,\, \sigma^2)$.

L'objectif de la modélisation est de trouver une fonction $\hat{f}$ qui généralise cette hypothèse. Ainsi, $\hat{f}(X) = \hat{y}$.

Pour simplifier la notation de la décomposition biais-variance, les notations suivantes sont abrégées : $f = f(X)$ et $\hat{f} = \hat{f}(X)$.

Pour toute variable aléatoire X, la variance est définie par :

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

f étant déterministe, son espérance est évidente :

$$\mathrm{E}[f] = f$$

Ainsi, étant donné que le bruit est supposé suivre une loi centrée :

$$\mathrm{E}[y] = \mathrm{E}[f + \epsilon] = \mathrm{E}[f] = f$$

En utilisant ce terme pour le calcul de la variance, et connaissant les paramètres de ϵ :

$$\mathrm{Var}[y] = \mathrm{E}[(y - \mathrm{E}[y])^2] = \mathrm{E}[(y - f)^2] = \mathrm{E}[(f + \epsilon - f)^2] = \mathrm{E}[\epsilon^2] = \sigma^2$$

Avec ces termes, l'erreur de prédiction quadratique peut s'écrire :

$$\begin{aligned}
\mathrm{E}[(y - \hat{f})^2] = \mathrm{E}[y^2 + \hat{f}^2 - 2y\hat{f}] & \\
&= \mathrm{E}[y^2] + \mathrm{E}[\hat{f}^2] - 2\mathrm{E}[y\hat{f}] \quad \text{par linéarité} \\
&= (\mathrm{Var}[y] + f^2) + (\mathrm{Var}[\hat{f}] + \mathrm{E}[\hat{f}]^2) - 2f \cdot \mathrm{E}[\hat{f}] \\
&= \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] + (f - \mathrm{E}[\hat{f}])^2 \\
&= \sigma^2 + \mathrm{Var}[\hat{f}] + \mathrm{Biais}[\hat{f}]^2 \quad \text{par définition du biais}
\end{aligned}$$

2 Démonstration : relation entre $q_{x,t}$ et $\mu_{x,t}$

La relation correspondant à l'équation 4.1 est utilisée pour comparer les projections de $q_{x,t}$ des modèles de Lee-Carter aux projections de Prophet.

L'objectif est de montrer :

$$q_{x,t} = 1 - \exp\left(-\int_0^t \mu_{x+s}\, ds\right)$$

Hypothèse :
La fonction de hasard $\mu(t+x)$ est constante par morceaux : $\mu(t+s, x+u) = \mu(t,x)$, $\forall s, u \in [0;1[$.

$$\begin{aligned}
1 - {}_tq_x &= {}_tp_x \\
&= \frac{S(t+x)}{S(x)} \\
&= \exp\left(-\int_0^t \mu(x+s)\, ds\right) \qquad (4.3) \\
&= \exp\left(-\int_0^t \mu(s)\, ds\right) \quad \text{(par hypothèse)} \\
&= \exp(-t \cdot \mu(x))
\end{aligned}$$

Ainsi, on obtient :

$${}_tq_x = 1 - \exp(\mu(x))t$$

Et, pour $t = 1$:

$$q_x = 1 - \exp(\mu_x)$$