Contribution en apprentissage automatique pour la maîtrise des risques

Lassana Coulibaly

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Mathématiques Appliquées

Présentée et soutenue par :

M. LASSANA COULIBALY

le jeudi 17 décembre 2020

Titre :

Contribution en apprentissage automatique pour la maîtrise des risques

---

Ecole doctorale :

Aéronautique-Astronautique (AA)

Unité de recherche :

Laboratoire de Génie de Productions de l'ENIT (E.N.I.T-L.G.P.)

Directeur(s) de Thèse :

M. BERNARD KAMSU-FOGUEM

M. FANA TANGARA

Rapporteurs :

M. MAMADOU MBOUP, UNIVERSITE DE REIMS

M. MOHAMED LEMDANI, UNIVERSITE LILLE 2

Membre(s) du jury :

Mme SYLVIE LE HEGARAT MASCLE, UNIVERSITE PARIS 11, Président

M. BERNARD KAMSU-FOGUEM, ECOLE NATIONALE D'INGENIEUR DE TARBES, Membre

M. FANA TANGARA, UNIV. DES SCIENCES TECH & TECHNO BAMAKO, Membre

M. JULES SADEFO KAMDEM, UNIVERSITE DE MONTPELLIER, Membre

Mme FABIENNE LOHOU, UNIVERSITE PAUL SABATIER, Membre

Table des matières

Table des matières

Table des matières

Dédicace

Je dédie ce travail :
A mes deux parents qui ont toujours veillé sur moi ;
A toute ma famille ;
A tous ceux qui m'ont aidé de près ou de loin à réaliser ce modeste travail.

Remerciements

Tout d'abord je tiens à remercier Allah, le Tout Puissant et miséricordieux, qui m'a donné la force, l'intelligence et la patience d'accomplir ce modeste travail.

Je remercie très chaleureusement mon Directeur de thèse, Monsieur Bernard KAMSU-FOGUEM, d'avoir accepté de diriger cette thèse ; encore merci pour sa disponibilité et ses conseils éclairés. Quel que soit un problème posé durant ma thèse, il m'a toujours soutenu et guidé dans mon travail ; et m'a aidé à trouver des solutions pour avancer.

Je remercie sincèrement Madame Fabienne LOHOU, Encadrante de ce travail ; elle s'est toujours montrée à l'écoute et très disponible tout au long de la réalisation de ce travail, ainsi pour l'inspiration, l'aide et le temps qu'elle a bien voulu me consacrer.

Je remercie sincèrement Monsieur Fana TANGARA, mon Co-directeur de thèse ; il s'est aussi toujours montré disponible pour la réalisation de ce travail et m'a guidé à trouver des solutions pour avancer face à certaines situations difficiles telles que des problématiques scientifiques et autres, je dis merci.

Je remercie aussi Pierre MASSONNAT pour sa disponibilité et son soutien. Il m'a facilité la résolution de tant de problèmes informatiques tels que des problèmes d'algorithme et d'automatisation des astuces qui prenait plus de temps à faire à la main ; merci pour ton efficacité.

Merci mon collègue Boukaye Boubacar TRAORE pour ton intervention inoubliable en début de ma thèse. Tu m'as initié aux outils informatiques dans le domaine d'intelligence artificielle, précisément le Data Mining ; sans quoi le travail me serait très difficile.

Merci mon collègue Tamba Mamadou CAMARA pour tes conseils, tes remarques et suggestions lors de nos travaux en groupe. Je n'ai pas oublié que tu m'as aidé à trouver le courage, le moral et surtout la patience pour surmonter des obstacles durant ce travail.

Merci mon collègue Abdoulaye DIAMOUTENE pour ta disponibilité. Je n'ai pas oublié tes interventions linguistiques dans mes projets d'article. En plus, tes conseils et tes expériences statistiques m'ont beaucoup aidé dans mes approches statistiques.

Merci mes collèges Solemane COULIBALY et Cheick Abdoul Kadir A. KOUNTA, des informaticiens qui ont toujours travaillé avec moi en groupe au laboratoire LGP de l'ENIT. Vos remarques et suggestions, lors de nos travaux en groupe au LGP, m'ont aidé d'améliorer mes différents projets d'article dans ce travail.

Merci à mon collègue Sinaly DISSA pour ta disponibilité, tes conseils et le partage de tes expériences avec moi. Tu as toujours répondu avec calme et patience à mes questions quotidiennes.

Remerciements

Merci à mon collègue Moumouni DIALLO pour ta disponibilité permanente. Je n'ai pas oublié l'aide que tu m'as apporté pendant mes démarches administratives, encore merci.

Je remercie sincèrement l'ambassade de France au Mali de m'avoir accordé cette bourse d'étude doctorale.

Je remercie aussi sincèrement le Gouvernement du Mali d'avoir mis en place ce programme de formation des formateurs de l'enseignement supérieur au Mali.

Je remercie tous ceux et celles qui m'ont aidé et encouragé de près ou de loin dans la réalisation de ce travail, par leur patience, leurs compétences et leurs interventions adéquates aux plans technique, économique et moral. Il me serait difficile de les citer tous. Qu'ils trouvent ici, l'expression de ma reconnaissance.

Résumé

Les changements climatiques entraînent régulièrement des phénomènes menaçant directement l'environnement et l'humanité. Dans ce contexte, la météorologie joue de plus en plus un rôle important dans la compréhension et la prévision de ces phénomènes. Le problème de fiabilisation des observations est essentiel pour le raisonnement numérique et la qualité de la simulation. En plus, l'interopérabilité est importante tant pour les entreprises que pour les services publics traitant des données et des modèles complexes découlant de ces observations. Dans les services météorologiques, la fiabilité des données d'observations est une exigence fondamentale. Les prévisions du temps et du climats sont dépendantes de nombreux phénomènes physiques à différentes échelles de temps et d'espace. Un de ces phénomènes est le transfert d'énergie de la surface vers l'atmosphère qui est considéré un paramètre sensible. Les observations des paramètres sensibles produisent souvent des données qui ne sont pas fiables (données imparfaites). Un meilleur traitement de ces données imparfaites pourra améliorer l'évaluation de la simulation. Nous proposons l'utilisation de méthodes d'apprentissage automatique susceptibles (i) d'améliorer l'évaluation des échanges entre la surface et l'atmosphère dans les modèles numériques de prévision du temps et du climat et (ii) de produire des connaissances pour l'interopérabilité. Cela peut appuyer la communication des services d'observation et les modèles numériques de prévision.

L'objectif de ce travail est de diagnostiquer les modèles numériques de prévision pour chercher les faiblesses de ces modèles dans la simulation des échanges entre la surface et l'atmosphère. Ces échanges sont quantifiés par les flux de chaleur sensible et de chaleur latente. Dans un premier temps, la méthode d'extraction des règles d'association est choisie pour : mettre en évidence les faiblesses du modèle ; effectuer des comparaisons entre les observations effectuées et les simulations réalisées par le modèle numérique pour la détection des variables critiques. Dans un deuxième temps, des processus gaussiens tenant compte des incertitudes sont utilisés pour modéliser les valeurs mesurées afin de rendre la base de données d'observation plus fiable. Cette modélisation est réalisée par un processus d'apprentissage approfondi qui inclut la régression en intégrant les connaissances sur le terrain. Ensuite, un optimiseur a été défini à partir des propriétés sur les transformations géométriques par homothétie. Cet optimiseur permet d'effectuer un ajustement aux données simulées pour mettre à l'échelle le modèle.

Ces méthodes sont déployées sur une base de données contenant des variables mesurées (flux de chaleur sensible et latente, température et humidité de l'air, vitesse et direction du vent, pluie, rayonnement global, etc.) sur le site expérimental du Centre de Recherches Atmosphériques (CRA) qui est l'un des deux sites composant la Plateforme Pyrénéenne d'Observation de l'Atmosphère (P2OA) en France.

Résumé

Les résultats obtenus et exprimés sous forme de règles d'association ont permis de mettre en évidence des faiblesses dans les modèles numériques : d'abord, la mise en évidence des différences (erreurs) entre les observations et les simulations ; ensuite l'analyse des règles générées a montré que les différences importantes sur le rayonnement global sont souvent concomitantes à des différences importantes sur les flux de chaleur sensible et latente. Ceci est souvent dû à des perturbations naturelles (par exemple, emplacement des nuages) qui impactent la qualité des observations/ simulations des flux de chaleur sensible et chaleur latente. Les bénéfices escomptés sont relatifs à la génération de connaissances utiles à l'amélioration de la qualité de la simulation numérique des processus de surface.

En plus, l'optimiseur proposé a donné des résultats satisfaisants. Les valeurs simulées ont été mises à l’échelle à 100% dans le cas des formes similaires et à 98% dans le cas des formes avec présence de pics. Cet optimiseur peut être appliqué à toutes les autres variables pour encore mieux améliorer la fiabilité du modèle numérique de prévision.

Mots clés : Interopérabilité ; Météorologie ; Incertitude ; Fiabilité ; Apprentissage automatique ; Data Mining ; Processus gaussiens ; Transport optimal.

Abstract

Climate change regularly causes phenomena that directly threaten the environment and humanity. In this context, meteorology is playing more and more an important role in the understanding and forecasting of these phenomena. The problems of reliability of the observations is essential for the numerical reasoning and the quality of the simulation. In addition, interoperability is important both for companies and for public services dealing with complex data and models. In meteorological services, the reliability of observational data is a fundamental requirement. Weather and climate predictions are dependent on many physical phenomena on different time and space scales. One of these phenomena is the transfer of energy from the surface to the atmosphere that is a sensitive parameter. Observations of sensitive parameters often produce data that are unreliable (imperfect data). A better treatment of these imperfect data may improve the evaluation of the simulation. We propose the use of machine learning methods that can : (i) improve the evaluation of surface-atmosphere exchanges in numerical weather and climate prediction models and (ii) produce knowledge for interoperability. This can support the communication of observation services and numerical prediction models.

The objective of this work is to diagnose numerical prediction models in order to look for the weaknesses of these models in the simulation of exchanges between the surface and the atmosphere. These exchanges are quantified by sensible and latent heat fluxes. In a first instance, Gaussian processes taking into account uncertainties are used to model the measured values in order to make the observational database more reliable. This modelling is carried out through a thorough learning process that includes regression by integrating field knowledge. Then the extraction method of the association rules is chosen in order to : highlight the weaknesses of the model ; make comparisons between the observations made and the simulations made by the numerical model. Finally, an optimizer has been defined from some properties on geometric transformations in mathematics. This optimizer makes it possible to perform an adjustment to the simulated data in order to minimize simulation errors.

These methods are deployed on a database containing measured variables (sensible and latent heat flux, air temperature and humidity, wind speed and direction, rain, global radiation, etc.) on the experimental site of the Centre de Recherches Atmosphériques (CRA) which is one of the two sites making up the Pyrénéenne Plateforme d'Observation de l'Atmosphère (P2OA) in France.

The results obtained and expressed in the form of association rules have made it possible to highlight certain weaknesses in the numerical models : first, the highlighting of differences (errors) between the observations and the simulations ; then the analysis of the generated rules showed that important differences on global radiation are often concomitant with important differences on sensible and latent heat fluxes. This is often due to natural disturbances (e.g. cloud location) that impact the quality of

observations/simulations of sensible and latent heat fluxes. The expected benefits are related to the generation of useful knowledge to improve the quality of numerical simulation of surface processes.

In addition, the proposed optimizer gave satisfactory results. The simulated values were scaled to 100% in the case of similar shapes and to 98% in the case of shapes with peaks. This optimizer can be applied to all other variables to further improve the reliability of the numerical prediction model.

Key words : Interoperability ; Meteorology ; Uncertainty ; Reliability ; Machine learning ; Data Mining ; Gaussian processes ; Optimal transport.

Liste des figures

Liste des figures et tableaux

Liste des tableaux

Introduction générale

1. Contexte et motivations

Les changements climatiques entraînent régulièrement des phénomènes menaçant directement l'environnement et l'humanité. Il s'agit donc d'une forme de réchauffement climatique dont les causes sont en grande partie anthropiques comme le montrent régulièrement les rapports du GIEC (Groupe International d'Experts sur le Climat). Ce réchauffement climatique est un phénomène global de transformation du climat caractérisé par une augmentation générale des températures moyennes, et qui modifie durablement les équilibres météorologiques et les écosystèmes. De nombreux scientifiques étudient ce phénomène et tentent de comprendre comment les activités des sociétés humaines provoquent ce réchauffement (Perkins et al., 2018). Ces scientifiques sont regroupés au sein du GIEC, et ils publient régulièrement des rapports étudiant l'évolution du réchauffement climatique.

Dans ce contexte, la météorologie joue de plus en plus un rôle important dans la compréhension de ces phénomènes, notamment les phénomènes extrêmes, de voir comment ils vont évoluer avec le changement climatique. La fiabilité des modèles de prévisions météorologiques est une question complexe, car elle dépend de nombreux paramètres et de l'infrastructure technique qui les soutient. De plus une meilleure fiabilité des prévisions météorologiques est importante pour la gestion des recommandations (par exemple des alertes météo éditées par les services météos à destination de la population) en prenant en compte les variables importantes comme les incertitudes, les mesures d'erreur, et les prévisions de charge, etc., (Aguera-Pérez et al., 2018).

Définition 1. La fiabilité d'un modèle numérique de prévision est la probabilité de n'avoir aucune défaillance pendant la durée de simulation. C'est-à-dire, le modèle est dit fiable lorsque toutes les valeurs simulées convergent vers les valeurs mesurées pensant l'évaluation, avec un intervalle de convergence défini au préalable par l'expert du domaine.

Définition 2. Les données observées/mesurées sont dites fiables ou parfaites lorsqu'elles ne sont pas infectées par des fluctuations de mesures et des valeurs manquantes.

Compte tenu de cette importance de la fiabilité des modèles de prévisions météorologiques, des initiatives ont été menées pour son amélioration. Par exemple, la fiabilité des systèmes de prévision d'ensemble météorologique et hydrologique peut être améliorée grâce à l'utilisation des approches statistiques (Abaza et al., 2017). La fiabilité de la prévision de l'humidité du sol peut aussi être augmentée par le couplage du modèle de bilan hydrique du sol et des prévisions météorologiques par satellite (Corbari et al., 2019). En raison des variations interannuelles des niveaux de pollen, un modèle de prévision plus fiable a été élaboré pour fournir des informations précises aux patients et au personnel médical concernant les niveaux de pollen en suspension dans l'air. De ce fait, un algorithme a été développé pour prévoir la quantité totale de pollen dans l'air. Cela

suggérait que la quantité totale de pollen dans l'air au cours d'une saison donnée pouvait être estimée à l'aide des seules données météorologiques des années précédentes (Tseng et al., 2018). Une méthodologie a été présentée par (Doycheva et al., 2017) pour améliorer la fiabilité des prévisions d'inondation en pondérant les éléments de l'ensemble de prévisions en fonction de leurs compétences. En plus, une approche basée sur un apprentissage automatique supervisé a été présentée et testée sur des prévisions de précipitations pour le bassin de la Mulde en Allemagne. En résumé, la maîtrise de la prédiction des phénomènes dangereux (inondations, pollution de l'air, etc) dépend de la fiabilité du modèle numérique de prévision du temps.

En plus, le contexte marqué par un besoin prégnant d'interopérabilité (i.e. capacité des systèmes, unités, matériels à opérer ensemble) a évolué très rapidement au cours de la dernière décennie. Il s'est étendu à partir du domaine des systèmes d'information, qui est essentiellement axé sur la possibilité de communication entre deux ou plusieurs systèmes, appareils ou éléments informatiques (Zutshi et al., 2012), (Borgogno & Colangelo, 2019).

Définition 3. Interopérabilité. En informatique, l'interopérabilité peut se définir comme la capacité, pour deux ou plusieurs systèmes informatiques, à fonctionner ensemble. Elle permet d'assurer l'échange d'informations entre plusieurs systèmes/produits.

Définition 4. Système d'information (SI). C'est un ensemble organisé de ressources qui permet de collecter, stocker, traiter et distribuer de l'information, en général grâce à un réseau d'ordinateurs.

L'interopérabilité est importante tant pour les entreprises que pour les services publics traitant des données et des modèles complexes. Dans les services météorologiques, cette interopérabilité est une exigence fondamentale utile pour la communication dans un cadre collaboratif (Rathore & Patidar, 2019).

La plupart des auteurs ont récemment utilisé les techniques de classification et de régression aux données météorologiques pour la prédiction des phénomènes dangereux, par exemple la prévision de la pollution atmosphérique grave (Liu et al., 2018), (Abaza et al., 2017). Dans le domaine d'apprentissage automatique la gestion des données imparfaites est une préoccupation majeure qui est accentuée dans des situations d'interopérabilité entre les services météorologiques. Ce faisant, notre approche consiste à proposer des méthodes pour réaliser des ajustements aux données météorologiques mesurées/simulées imparfaites dans le but de réduire les fluctuations de mesures et qui pourra appuyer la communication entre les services météorologiques. Le schéma suivant (*Figure 0-1*) illustre un exemple d'aperçu des acteurs interagissant avec les services de Météo-France.

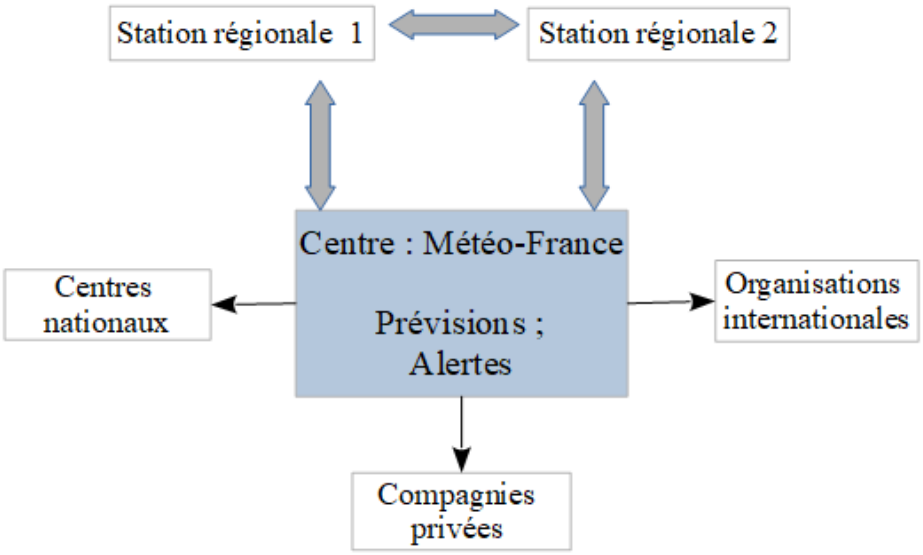


Figure 0-1 : Schéma des acteurs interagissant aux échanges d'informations en météorologie

Les données récoltées sur les différents services des stations météorologiques sont souvent imparfaites (par exemple données manquantes ou bruitées) et peuvent impacter la qualité de la prévision météorologique. Plusieurs entités sont impliquées dans l'échange d'informations relatives aux prévisions météorologiques. On peut notamment citer les stations régionales, des centres nationaux, des organisations internationales et parfois des compagnies privées (*Figure 0-1*). De ce fait, notre approche consiste à améliorer la fiabilité de la modélisation des données météorologiques par des méthodes d'apprentissage automatique. En effet, il est souvent nécessaire de fiabiliser ces données compte tenu de multiples incertitudes inhérentes à la mesure (par exemple données manquantes ou bruitées). Dans notre contexte, cette fiabilisation est réalisée avec un apprentissage automatique basé sur un processus gaussien. Le but est d'améliorer les échanges d'informations sur la base des données mesurées et simulées. Le premier bénéfice obtenu concerne la fiabilité des données mesurées par la modélisation en apprentissage automatique pour la réduction des fluctuations et accidents de mesures. Le second bénéfice est la fiabilité du modèle par l'utilisation d'un optimiseur pour la mise à l'échelle des données simulées (*Figure 0-2*).

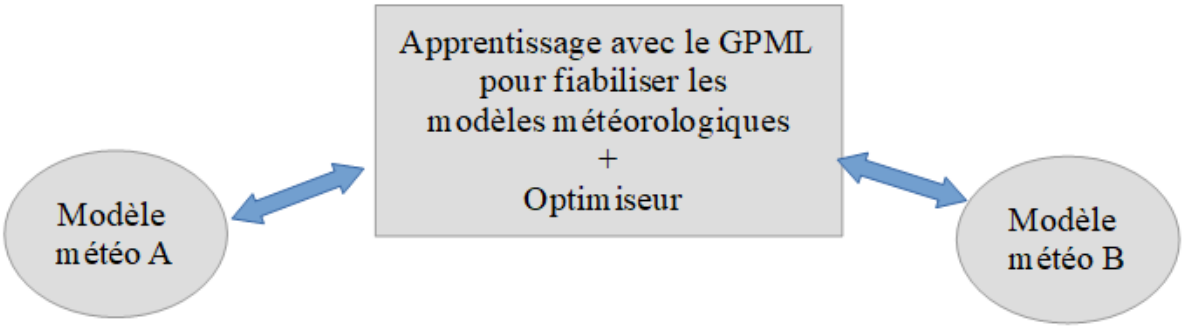


Figure 0-2 : Schéma pour fiabiliser les mesures dans les modèles A et B avec GPML

Modèle météo A peut être une représentation des observations réalisées sur différents sites ;

Modèle météo B peut être une représentation des simulations par le modèle numérique de prévision ;

Optimiseur désigne la fonction définie pour la mise à l'échelle des données simulées ;

Gaussian Processes for Machine Learning (GPML) est une méthode d'apprentissage automatique avec les Processus gaussiens.

2. Objectifs de la thèse

Les incertitudes sont souvent dynamiques dans les mesures des paramètres météorologiques (Safa et al., 2018), (Haymann et al., 2019). En particulier, les valeurs horaires manquantes sont souvent fréquentes dans les mesures des variables météorologiques comme les flux de chaleur sensible et de chaleur latente.

L'incertitude désigne (i) la marge d'imprécision sur la valeur de la mesure d'une grandeur physique et (ii) les valeurs manquantes remplacées.

L'objectif de cette thèse est (i) d'améliorer l'évaluation des échanges entre la surface et l'atmosphère dans les modèles numériques de prévision du temps et du climat ; et (ii) de produire des connaissances pour l'interopérabilité. Il s'agit de diagnostiquer les modèles pour rechercher ses faiblesses dans la simulation des processus de surface (échanges entre la surface et l'atmosphère). Ces échanges sont quantifiés par le rayonnement global, les flux de chaleur sensible et de chaleur latente. Il s'agit de mettre en évidence les faiblesses du modèle ; effectuer des comparaisons entre les observations effectuées et les simulations réalisées par le modèle numérique de prévision ; comprendre et identifier les paramètres qui influencent les biais importants.

3. Approche scientifique

Les paramètres qui quantifient le transfert d'énergie de la surface vers l'atmosphère sont très sensibles. Les données mesurées de ces paramètres sensibles sont souvent imparfaites. Par conséquent, il est nécessaire de chercher des moyens permettant

d'améliorer la qualité des données mesurées. En outre, il existe souvent des biais importants dans la simulation de ces paramètres sensibles comparativement aux mesures sur des périodes correspondantes. Ce faisant, il y a un besoin de méthodes avancées orientées pour une meilleure compréhension des modèles numériques de prévision et l'analyse des principaux paramètres associés. Il est aussi important de rechercher des pistes d'amélioration de la fiabilité de ces modèles.

Dans ce travail, pour atteindre les objectifs définis précédemment, dans un premier temps, nous proposons l'utilisation des méthodes d'extraction de connaissances à partir de données pour contribuer à l'évaluation de la simulation d'un modèle de prévisions numériques du temps. C'est ainsi que la méthode d'extraction des règles d'association est choisie pour effectuer des comparaisons entre les observations effectuées et les simulations numériques réalisées dans le but de mettre en évidence les faiblesses du modèle. Cette méthode comprend trois étapes : (i) un prétraitement est effectué par l'utilisation combinée de l'algorithme des k-plus proches voisins (*k*-NN) et la méthode de la moyenne mobile intégrée et autorégressive (ARIMA) afin d'estimer les valeurs manquantes, suivi d'une technique de réduction de données (p. ex. la discrétisation pour transformer des données quantitatives en des données qualitatives) ; (ii) un traitement est fait par exploration de données avec des règles d'association pour découvrir des relations intéressantes entre les différences (erreurs) des variables mesurées et simulées ; l'apprentissage par extraction des règles permet aussi d'effectuer des comparaisons entre les observations et les modèles de simulation numérique ; (iii) Le post-traitement est effectué par un raisonnement logique et graphique qui facilite la visualisation des liens entre les règles obtenues. Dans un second temps, ce travail de recherche propose l'utilisation de la méthode du processus gaussien pour l'apprentissage automatique (GPML) qui est susceptible d'améliorer les échanges d'informations sur les données. Le GPML, qui prend en compte des incertitudes (Abdelfatah et al., 2018), est utilisé pour modéliser les données dans le but de réaliser un ajustement des valeurs mesurées afin de réduire les fluctuations irrégulières (intégrant éventuellement des accidents de mesures), (Fanoodi et al., 2019), (Chang, 2017). Enfin, un optimiseur est défini à partir des propriétés sur les transformations géométriques par homothétie pour la mise à l'échelle des données simulées par le modèle.

Ces méthodes sont déployées sur une base de données contenant des variables mesurées/simulées (flux de chaleur sensible et latente, température et humidité de l'air, vitesse et direction du vent, pluie, rayonnement global, etc.) sur le site expérimental du Centre de Recherches Atmosphériques (CRA) qui est l'un des deux sites composant la Plateforme Pyrénéenne d'Observation de l'Atmosphère (P2OA) en France.

4. Structuration du document

L'organisation de ce manuscrit de thèse est décrite ci-après. Après l'*introduction générale*, le *chapitre 1* expose, dans la littérature, quelques applications récentes des méthodes du Data Mining, des méthodes du processus gaussien pour l'apprentissage automatique. Ensuite, le *chapitre 2* consiste à une présentation des données

météorologiques disponibles avec une visualisation statistique des variables mesurées et simulées. Une méthodologie est proposée dans le *chapitre 3* pour l'apprentissage automatique basé sur des règles pour la découverte de connaissances dans les données météorologiques avec un cas d'étude. Le *chapitre 4* présente aussi une autre méthodologie proposée pour la mise à l'échelle des données simulées avec un cas d'étude. Enfin, une conclusion générale est donnée dans le chapitre '*Conclusion – Perspectives*'.

Chapitre 1

1.Etat de l'art

1.1 Introduction

Dans le contexte de la maîtrise des risques (par exemple, des alertes météo) et la gestion des incertitudes sur les données météorologiques, le problème de fiabilité des modèles d'observation/simulation est important pour les prévisions météorologiques. De ce fait, l'apprentissage automatique peut être d'une grande utilité dans la fouille des données météorologiques.

Par conséquent, des auteurs ont mis l'accent sur l'application et le rôle d'apprentissage automatique aux données météorologiques dans la compréhension de la prévision des catastrophes naturelles, comme dans (Zhao & Song, 2017), (Azimi et al., 2016), etc. D'autres ont utilisé la notion de classification pour comparer les températures historiques et prévues dans le but de comprendre les variations de température avant, pendant et après des catastrophes naturelles telles que le vortex polaire ou les inondations, voir les détails dans (Chang, 2017).

Ce chapitre va présenter quelques applications récentes des méthodes d'apprentissage automatiques autour de notre problématique définie dans le chapitre *Introduction générale*. Après cette introduction, la notion d'exploration de données est présentée de façon générale dans la *section 1.2*. Ensuite, des études expérimentales de Data Mining avec les données météorologiques sont présentées dans la *section 1.3*. De plus, dans le cadre de la gestion des incertitudes de mesures, la *section 1.4* présente des travaux récents sur la modélisation de données à l'aide des méthodes d'apprentissage profond. En fin, la *section 1.5* donne une conclusion avec les contributions envisagées dans ce document.

1.2 Généralités sur le Data mining

Définition 5. Techniquement, le Data Mining est le procédé permettant de trouver des structures intéressantes ou des patterns à partir des bases de données selon des critères préalablement déterminés. Le Data Mining repose sur des algorithmes complexes et sophistiqués permettant d'extraire une connaissance (par exemple de segmentation ou de corrélation) à partir de grandes masses de données et d'évaluer son usage en aide à la décision (Bandaru et al., 2017).

Définition 6. Le Data Mining peut être aussi défini comme une découverte automatique de modèles intéressants à partir d'ensemble de données de grande taille.

Le Data Mining comprend, en général, trois étapes qui sont : prétraitement, traitement et post-traitement (Ristoski & Paulheim, 2016), (Djenouri & Comuzzi, 2017), (Ajak et al., 2017). Par exemple, dans le contexte de recherche d'association, le schéma suivant illustre le processus d'exploration de données (*Figure 1-1*).

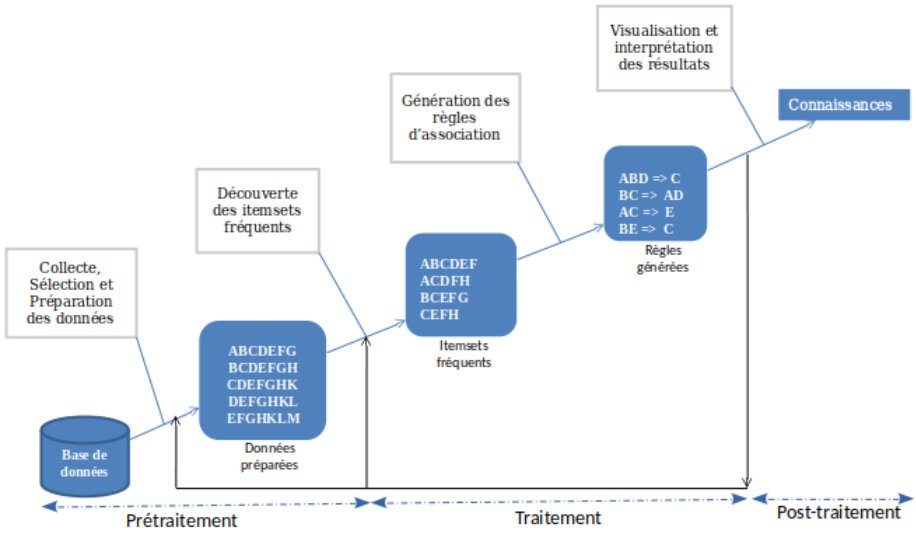


Figure 1-1 : Étapes d'exploration de données en Data Mining

L'étape de prétraitement consiste d'abord à collecter les données à partir des sources, ensuite sélectionner les données pertinentes et les préparer pour qu'elles soient compatibles aux algorithmes de traitement. La phase de traitement permet une identification des motifs contenus dans une base de données. Le post-traitement est la dernière étape qui cherche à faire une meilleure compréhension et interprétation des connaissances générées au profit l'utilisateur.

1.2.1 Phase de prétraitement de données

Le prétraitement est essentiel pour analyser les ensembles de données imparfaites avant l'exploration de données. Les principales taches de prétraitement de données sont les suivantes :

1. **Le nettoyage des données.** Le nettoyage des données supprime les observations contenant du bruit (ou des données incohérentes) et celles avec des données manquantes.
1. **L'intégration de données.** Lorsque les données proviennent de sources différentes.
2. **La sélection de données.** Elle permet de rechercher des données pertinentes dans la base de données.
3. **La transformation des données.** Elle consiste à mettre les données dans une forme appropriée pour les algorithmes de fouille de données.

Dans cette étude, nous avons une base de données constituée des données d'observation provenant d'une seule source (station d'observation du Centre de Recherche Atmosphérique (CRA) de Lannemezan). Ainsi, les taches de prétraitement abordées dans ce travail sont : le nettoyage des données et la transformation des données. Ces taches sont détaillées dans la *section 3.2*.

1.2.2 Phase de traitement (Data Mining)

En Data Mining, il existe plusieurs méthodes : on dénombre cinq variétés du Data Mining que l'on peut illustrer sur le schéma suivant (*Figure 1-3*).

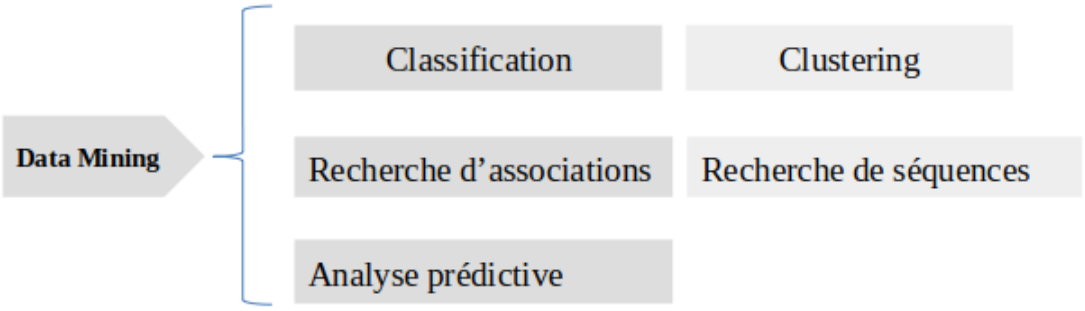


Figure 1-2 : Les différentes tâches de Data Mining

Ces taches sont réparties entre les deux types d'apprentissages automatiques existantes en intelligence artificielle, notamment l'apprentissage automatique **supervisé** et l'apprentissage automatique **non supervisé**.

L'apprentissage supervisé permet d'apprendre une fonction qui se rapproche le mieux de la relation entre entrée et sortie observable dans les données. Dans ce type d'apprentissage, nous avons une connaissance préalable des valeurs de sortie de nos échantillons (par exemple, classification, régression ou analyse prédictive), (De Mauro et al., 2018), (Chemchem & Drias, 2015), (Nisbet et al., 2018).

En revanche, l'apprentissage **non supervisée** permet de déduire la structure naturelle présente dans un ensemble de points de données. Dans ce cas, nous n'avons aucune connaissance préalable sur l'étiquetage des résultats (par exemple, clustering, recherche d'associations/séquences), (Nguyen & Kuo, 2019), (Gan et al., 2020), (Zhongjie Zhang et al., 2018), (Mai et al., 2017), (Huang et al., 2017).

Dans cette partie, quelques tâches de Data Mining seront définies. Nous allons d'abord présenter brièvement la classification et le clustering avant une description détaillée de la notion de recherche d'associations.

1.2.2.1 Classification

La classification d'une collection consiste à répartir les éléments qui composent la collection en catégories ou classes (Chemchem & Drias, 2015). En data mining, la classification s'effectue à l'aide d'un modèle qui repose sur des données historiques. La classification prédictive vise à prédire avec exactitude la classe cible pour chaque enregistrement de nouvelles données, c'est-à-dire les données qui ne sont pas dans les données historiques. Une tâche de classification commence par construire les données (également connu sous le nom entraînement) pour lesquels les valeurs cibles (ou les affectations de classe) sont connues. Les algorithmes de classification utilisent différentes techniques pour trouver des relations entre les valeurs de l'attribut de la variable

explicative et les valeurs de l'attribut cible dans les générations de données. Les *k* plus proches voisins (*k* Nearest Neighbors (*k*-NN)) est l'un de ces algorithmes qui sont très simples à comprendre, en outre, il fonctionne incroyablement bien dans la pratique, notamment dans le domaine de détection d'anomalies comme le soulignent *RONG* et ses collègues (Rong et al., 2020), également pour la catégorisation de texte comme dans le travail de *JIANG* et ses collègues (Jiang et al., 2012). Aussi, il est étonnamment polyvalent et ses applications vont de la vision à l'analyse des protéines en passant par la géométrie algorithmique de graphes. Avec l'algorithme *k*-NN, nous pouvons obtenir des résultats satisfaisants, en outre, son principe de base est très simple et facile à mettre en œuvre. Le *k*-NN est un algorithme d'apprentissage non paramétrique, et il est utilisé lorsque l'ensemble des données n'obéit pas à une fonction déterminée (mélanges gaussiens, etc. linéairement séparable). L'algorithme *k*-NN peut s'expliquer comme suit, dans un premier temps, les données d'apprentissage qui sont déjà classées sont considérées et ensuite pour classer les nouvelles données, il faut calculer la distance de similitudes entre ces nouvelles données et toutes les données d'apprentissage. C'est après que les *k* plus proches voisins sont extraits. En fin de compte, les nouvelles données sont assignées à la classe la plus fréquente de ces voisins.

1.2.2.2 Clustering

Le mécanisme de clustering aux données consiste à mettre les données homogènes dans le même groupe ou une classe afin d'envoyer les données hétérogènes en différents groupes. Dans la littérature, il existe deux principales manières différentes pour regrouper les données : le regroupement hiérarchique et le regroupement par partition. Pour le clustering hiérarchique, les grappes sont emboitées. Cette catégorie de regroupement est utilisée lorsque les données peuvent être séparées en différents niveaux. En outre, la Classification Ascendante Hiérarchique (CAH) est l'algorithme hiérarchique le plus connu, il commence en mettant chaque instance dans un seul cluster après qu'il calcule les différences pour toutes les deux instances pour combiner les clusters qui ont la plus faible distance. Ce processus est répété jusqu'à ce que nous obtenions un seul cluster (Pei et al., 2006), (Ertoz et al., 2004), (Steinbach et al., 2004), (Henriques et al., 2015). Par contre, le regroupement par partitionnement consiste à construire plusieurs partitions séparément puis à les évaluer selon des critères. Le *K*-means est l'un des plus simples algorithmes d'apprentissage de partitionnement qui résout un problème de clustering bien connu, comme le mentionnent *NGUYEN* et ses collègues (Nguyen & Kuo, 2019), (Han et al., 2017). La procédure suit un moyen simple et facile de classer un ensemble de données à travers un certain nombre de grappes fixé au départ. L'idée principale est de définir la gravité k des centres, un pour chaque cluster. Le centre de gravité doit être placé de manière astucieuse parce que le résultat de clustering dépend de leur emplacement dans les grappes. Afin d'optimiser l'efficacité des résultats, il est judicieux de les placer autant que possible loin de l'autre. La prochaine étape est de prendre chaque point appartenant à un ensemble de données et il est associé au centre de gravité le plus proche. Lorsqu'aucun point n'est en cours, la première étape est terminée et un regroupement rapide est effectué. À ce stade, nous devons recalculer les *k* nouveaux centres de gravité de l'amas résultant de l'étape précédente et on répète le processus. L'arrêt de ce dernier

s'effectue lorsqu'on n'observe plus aucun changement des grappes, en d'autres termes lorsque les centres de gravité ne se déplacent plus.

1.2.2.3 Recherche d'association

L'extraction des règles d'association a été introduite pour la première fois par *AGRAWAL* (Agrawal et al., 1993). Une règle d'association peut être définie comme suit.

Définition 5. Soit $I=\{i_1, i_2, i_3, i_4,...\}$ un ensemble d'items. Soit la base de données $D=\{T_1,T_2, T_3, T_4,...\}$ un ensemble de transactions où chaque transaction T est un sous-ensemble de I . Une règle d'association est une inférence de la forme $X \longrightarrow Y$, où X, Y sont des sous-ensembles de I et $X \cap Y = \emptyset$. L'ensemble d'itemsets X est appelé antécédent et Y est appelé conséquent (Narvekar & Syed, 2015).

Les deux propriétés confiance et support sont généralement considérées dans l'extraction des règles d'association.

Définition 6. Soient *freq(X)* le nombre de transactions contenant l'ensemble des éléments X dans la base de données et c*ard(D)* le nombre total de transaction de la base de données. Le support de l'ensemble d'éléments X est défini comme la fraction de toutes les lignes contenant l'ensemble d'éléments, c'est-à-dire *freq(X)/card(D)*. Le support d'une règle d'association est le support de l'union de l'antécédent X et du conséquent Y, c'est-à-dire

$$support\ (X \longrightarrow Y)\ = \frac{freq(X \cup Y)}{card(D)}$$

La confiance d'une règle d'association est définie comme le pourcentage de lignes dans D contenant l'ensemble d'éléments X qui contiennent également l'ensemble d'éléments Y, c'est-à-dire

$$confiance(X \longrightarrow Y) = \frac{support(X \cup Y)}{support(X)}$$

Un ensemble d'items est fréquent si son support est supérieur ou égal à un support minimal spécifié par l'utilisateur. L'extraction de règles d'association consiste à identifier toutes les règles répondant aux contraintes spécifiées par l'utilisateur, telles que le support minimal et la confiance minimale.

Définition 7. Principe d'anti-monotonie : les algorithmes Apriori et FPGrowth sont les plus utilisés pour découvrir les règles d'association souhaitée (Djenouri & Comuzzi, 2017). Ces deux algorithmes utilisent un même principe mathématique, notamment le principe d'anti-monotonie (Henriques et al., 2015), (Apiletti et al., 2017). **Principe** : si un ensemble est non fréquent, alors tous ses sur-ensembles ne sont pas fréquents.

Exemple 1. Si {A} n'est pas fréquent alors {AB} ne peut pas l'être si {AB} est fréquent alors {A} et {B} le sont.

Dans notre base de données observées, nous avons environ 100 000 valeurs réelles mesurées sur deux ans et à chaque heure (soit environ 17 520 transactions). Dans cette section, nous allons nous limiter à 10 transactions, dans le *Tableau 1* ci-dessous, pour expliquer les étapes de traitement par les algorithmes Apriori et FPGrowth. Dans la partie application, nous prendrons en compte la totalité de la base de données de transaction.

Tableau 1 : Exemple de base de données avec 10 transactions

Base de données, minsup=2	
Tid	itemsets
tr1	I_1, I_2, I_5
tr2	I_2, I_4
tr3	I_2, I_3
tr4	I_1, I_2, I_4
tr5	I_1, I_3
tr6	I_2, I_3
tr7	I_1, I_3
tr8	I_1, I_2, I_3, I_5
tr9	I_1, I_2, I_3
tr10	I_6

Tableau 2 : Algorithme Apriori

Définition 8. Apriori est un algorithme itératif qui alterne entre les deux tâches importantes, la première est la génération de candidats à partir d'ensembles d'éléments fréquents de l'itération précédente et la seconde est le balayage de la base de données pour soutenir le comptage des candidats par rapport à chaque transaction.

Dans la k ième itération ($k \geq 2$), les k-itemsets candidats C_k sont générés à partir des (k-1)-itemsets fréquents L_{k-1}, puis les sous-ensembles de k-itemsets de chaque transaction sont comparés aux candidats C_k pour le comptage de support. Les itemsets candidats C_k sont obtenus en joignant conditionnellement L_{k-1} à lui-même, et ensuite l'élagage des itemsets qui ne satisfont pas à la propriété d'anti-monotonie (Singh et al., 2017). Selon cette propriété, tous les ensembles d'items de C_k peuvent être retirés de C_k si l'un de leurs (k-1) sous-ensembles n'est pas présent dans la L_{k-1}.

Exemple 2. En utilisant l'exemple de la base de données transactionnelle ci-dessus (*Tableau 2*), l'algorithme d'Apriori procède comme suite :

- Après avoir scanné la base de données ci-dessus, on obtient :

C_1
Itemset:sup
$\{I_1\}$:6
$\{I_2\}$:7
$\{I_3\}$:6
$\{I_4\}$:2
$\{I_5\}$:2
~~$\{I_6\}$:1~~

Retirer les itemsets non fréquents →

L_1
Itemset:sup
$\{I_1\}$:6
$\{I_2\}$:7
$\{I_3\}$:6
$\{I_4\}$:2
$\{I_5\}$:2

- Générer les candidats C_2 à partir de L_1, scanner la base de données puis appliquer le principe d'anti-monotonie, on obtient :

C_2
Itemset:sup
$\{I_1, I_2\}$:4
$\{I_1, I_3\}$:4
~~$\{I_1, I_4\}$:1~~
$\{I_1, I_5\}$:2
$\{I_2, I_3\}$:4
$\{I_2, I_4\}$:2
$\{I_2, I_5\}$:2
~~$\{I_3, I_4\}$:0~~
~~$\{I_3, I_5\}$:1~~
~~$\{I_4, I_5\}$:0~~

Retirer les candidats non fréquents

L_2
Itemset:sup
$\{I_1, I_2\}$:4
$\{I_1, I_3\}$:4
$\{I_1, I_5\}$:2
$\{I_2, I_3\}$:4
$\{I_2, I_4\}$:2
$\{I_2, I_5\}$:2

- Générer les candidats C_3 à partir de L_2, scanner la base de données puis appliquer le principe d'anti-monotonie, on obtient :

C_3
Itemset:sup
$\{I_1,I_2,I_3\}$:2
$\{I_1,I_2,I_5\}$:2

Retirer les candidats non fréquents →

L_3
Itemset:sup
$\{I_1,I_2,I_3\}$:2
$\{I_1,I_2,I_5\}$:2

- Générer les candidats C_4 à partir de L_3, scanner la base de données puis appliquer le principe d'anti-monotonie, on obtient :

C_4
Itemset:sup
~~$\{I_1,I_2,I_3,I_5\}$:1~~

Retirer les candidats non fréquents →

L_4
Itemset:sup
{ }

- L'algorithme s'arrête ici.

Tableau 3 : Algorithme FPGrowth

Définition 9. A l'opposé d'Apriori qui génère des itemsets candidats et qui les teste pour ne conserver que les itemsets fréquents, FP-Growth construit les itemsets fréquents sans génération de candidats (Khader et al., 2016), (Narvekar & Syed, 2015).

Tout d'abord, il compresse les itemsets fréquents représentés dans la base de données à l'aide des FP-Tree (Frequent-Pattern Tree) dont les branches contiennent les associations possibles des items. Chaque association peut être divisée en fragments qui constituent les itemsets fréquents. La méthode FP-Growth transforme le problème de la recherche de l'itemset fréquent le plus long par la recherche du plus petit et sa concaténation avec le suffixe correspondant (le dernier itemset fréquent de la branche aboutissant à l'item considéré). Ceci permet de réduire le coût de la recherche.

Exemple 3. En utilisant l'exemple de la base de données transactionnelle ci-dessus (*Tableau 3*), l'algorithme FPGrowth procède en 4 étapes :

1. Cette première étape identifie les items dont le support est supérieur ou égal au support minimal choisi (minsup =2). Ceci donne la liste ci-dessous (Tableau 4) :

Tableau 4 : Liste des items identifiés par ordre de support, avec minsup=2

Liste des items identifiés, avec minsup = 2	
Liste items ord	Support
I2	7
I1	6
I3	6
I4	2
I5	2

2. La seconde étape consiste à repositionner les items par ordre décroissant de support, dans chaque itemset fréquent. On utilisera l'ordre des items fréquents dans (Tableau 4) : {I2:7, I1:6, I3:6, I4:2, I5:2}. Ceci donne le tableau ci-dessous (Tableau 5).

Tableau 5 : Positionnement des items par ordre décroissante pour chaque itemset

Ordre des items fréquents : I_2:7, I_1:6, I_3:6, I_4:2, I_5:2		
Tid	Itemsets	Itemsets Ordre
tr1	I_1, I_2, I_5	I_2, I_1, I_5
tr2	I_2, I_4	I_2, I_4
tr3	I_2, I_3	I_2, I_3
tr4	I_1, I_2, I_4	I_2, I_1, I_4
tr5	I_1, I_3	I_1, I_3
tr6	I_2, I_3	I_2, I_3
tr7	I_1, I_3	I_1, I_3
tr8	I_1, I_2, I_3, I_5	I_2, I_1, I_3, I_5

tr9	I_1, I_2, I_3	I_2, I_1, I_3
tr10	I_6	~~I_6~~

3. Ensuite, cette étape est la construction de l'arbre FP-Tree. Le graphique suivant, *Figure 1-4*, illustre le schéma de l'arbre FP-Tree.

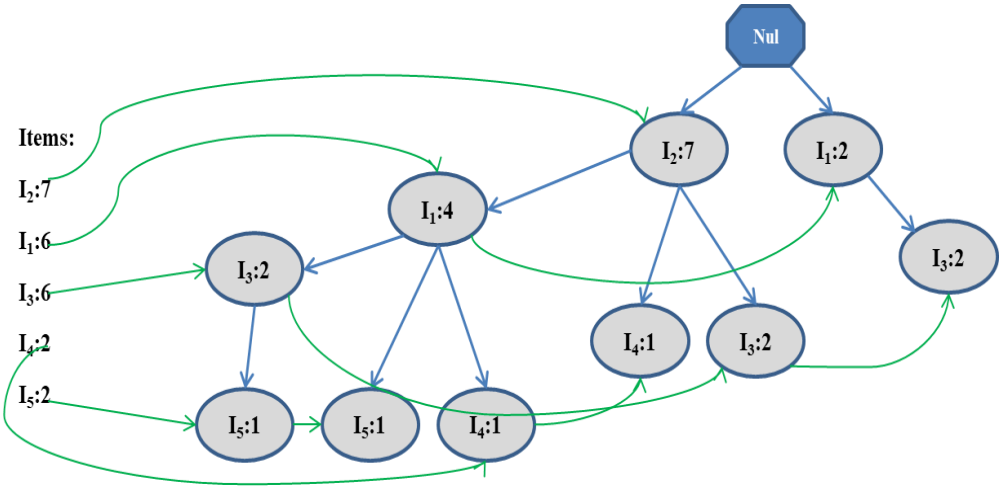


Figure 1-3 : Schéma de l'arbre FP-Tree

4. Cette dernière étape consiste à faire une extraction des patterns fréquents à partir de l'arbre FP-Tree (*Figure 1-4*). Ceci génère les patterns fréquents dans le tableau suivant (*Tableau 6*).

Tableau 6 : Patterns fréquents générés

item	Base de patterns conditionnelle	FP Tree conditionnel	Patterns fréquents générés
I_5	{{I_2,I_1:1},{ I_2,I_1, I_3:1}}	{I_2:2,I_1:2}	{I_2,I_5:2},{I_1,I_5:2},{I_2, I_1,I_5:2}
I_4	{{I_2,I_1:1},{ I_2:1}}	{I_2:2}	{I_2,I_4:2}
I_3	{{I_2,I_1:2},{ I_2:2},{ I_1:2}}	{I_2:4,I_1:4}	{I_2,I_3:4},{I_1,I_3:4},{I_2, I_1,I_3:4}
I_1	{I_2:4}	{I_2:4}	{I_2,I_1:4}

1.2.3 Phase de post-traitement

Les résultats en sortie de l'algorithme de forage de données peuvent être raffinées dans la phase de post-traitement. Celle-ci peut se résumer, dans certains cas, à l'interprétation des connaissances extraites ; ou dans d'autres cas, elle nécessite des traitements supplémentaires sur les connaissances extraites. Le but final de cette phase est de

permettre une meilleure compréhension et interprétation des connaissances générées par les algorithmes de forage de données. Pour faire cela, le moyen le plus simple et efficace est le recours aux techniques de visualisation graphique (Chen et al., 2017), (Blanchard et al., 2007).

1.3 Data Mining avec les données météorologiques

Afin d'enrichir la théorie existante de l'acquisition de règles dans des contextes décisionnels formels, *XIE* et ses collègues proposent trois nouveaux types de règles (Xie et al., 2018) : (i) les règles d'association de décision, (ii) les règles d'association de décision non redondantes et (iii) les règles d'association de décision les plus simples. Ensuite, ils analysent la relation entre ces trois types de règles et développent des méthodes pour les acquérir à partir de contextes de décision formels à une seule échelle. Quelques expériences numériques sont également menées pour comparer les performances de la méthode d'acquisition des règles d'association pour les décisions les plus simples à celle de la méthode existante d'acquisition des règles d'association de décision non redondantes. De plus, les trois nouveaux types de règles sont utilisés pour introduire trois types de cohérence dans des contextes de décision formels multi-échelles. En outre, la notion d'échelle optimale est définie par chaque type de cohérence, et comment sélectionner une échelle optimale est également étudié. Enfin, deux applications de la méthode proposée pour l'acquisition de règles et de sélection de l'échelle optimale sont appliquées à la ville intelligente.

ZHAO and *SONG* ont appliqué une approche de classification dans les données météorologiques (direction du vent, vitesse du vent, température, humidité, précipitations, etc.), entre 2013 et 2015 en Chine et les données de concentration de Particulate Matter (PM) de 2,5 (les particules en suspension dont le diamètre équivalent est inférieur à 25µm) pour la prévision de la pollution atmosphérique grave (Zhao & Song, 2017).

YESILBUDAK et ses collègues ont proposé une approche pour la mise en œuvre d'un classifieur avec l'algorithme des *k* plus proches voisins (*k*-NN) à partir des données météorologiques pour la prédiction de l'énergie éolienne (source d'énergie qui dépend du vent), (Yesilbudak et al., 2017). Une version améliorée de l'algorithme *K*-means est proposée par *AZIMI* et ses collègues pour la classification des données de vent afin de prédire la source d'énergie dépendante du vent (Azimi et al., 2016). Le modèle hybride du Data Mining (DM) a été proposé pour découvrir l'affiliation entre les stations synoptiques de la sécheresse de Tabriz et de Kermanshah (situées en Iran) et les Sea Surface Temperature (SST) de dépersonnalisation des mers Noire, Méditerranéenne et Rouge (Nourani & Molajou, 2017). Ce modèle est une combinaison de deux techniques d'exploration de données (arbre de décision et règles d'association).

Le modèle proposé comportait deux étapes principales : la classification de la SST hors tendance de données, la sélection des groupes les plus efficaces et en extrayant les informations cachées dans les données. Les techniques d'arbre décisionnel, permettant d'identifier les bons traits d'un ensemble de données, ont été utilisées pour la classification et la sélection des groupes les plus efficaces et des règles d'association utilisées pour extraire les informations prédictives cachées à partir des grandes données observées. Les mesures calculées confirment une performance fiable de la méthode hybride de data mining. Le modèle hybride proposé a été appliqué pour la surveillance de la sécheresse dans deux villes iraniennes. Les résultats montrent une corrélation relative entre la Méditerranée, la tendance de Mer Rouge et Noire SSTs, la sécheresse de Tabriz et des stations synoptiques Kermanshah pour que la confiance entre l'Indice de Précipitation Normalisé mensuel (SPI) des valeurs et la tendance SST de mers soit plus haute que 70 et 80% respectivement pour Tabriz et des stations synoptiques Kermanshah.

YU et ses collègues ont développé un algorithme d'apprentissage basé sur des règles faisant appel à Dynamic Time Warping (DTW) (Yu et al., 2018). Cet algorithme étudie la possibilité de détecter des réponses anormales à partir de sondes d'humidité du sol à l'aide de données recueillies du site IG à Milwaukee. Il est possible d'obtenir une meilleure précision en impliquant davantage de caractéristiques liées à l'humidité du sol qui est modifié et en tirant des enseignements d'un plus grand ensemble de données provenant d'observations plus longues ou d'un réseau de surveillance avec plusieurs sondes. Les règles d'association sont développées à la fois sur les caractéristiques environnementales et les caractéristiques d'événement. De telles règles d'association peuvent aider à vérifier efficacement la validité d'un modèle de changement d'humidité du sol. Les résultats suggèrent que cette méthode pourrait être utilisée pour identifier des modèles de changement anormaux par rapport aux observations historiques intra-sites. Bien que développée pour l'humidité du sol, cette méthode pourrait facilement être étendue pour s'appliquer à d'autres ensembles de données environnementales en continu.

Des méthodes d'extraction de données s'efforcent de trouver un équilibre entre l'optimisation des ressources informatiques et l'amélioration de l'efficacité prédictive (p. ex. par le biais de l'exploitation des données de localisation) en raison de son rôle dans différentes applications telles que les systèmes de recommandation à l'aide des services de réseautage social (Valverde-Rebaza et al., 2018).

Dans des expériences de classification multi-label, l'utilisation arbitraire des mesures d'évaluation sans une analyse objective de corrélation ou avec partialité peut conduire à des conclusions trompeuses ; d'où l'utilité d'une analyse de corrélation des mesures d'évaluation pour fournir des suggestions concrètes lors du choix des mesures d'évaluation pour la classification multi-étiquettes (Pereira et al., 2018).

FIGUEIREDO et ses collègues ont proposé une méthode d'extraction de données basée sur les informations rendues et un modèle n-gramme (DERIN) qui vise à améliorer les performances du wrapper en sélectionnant automatiquement la région principale de données à partir d'une page de résultats de recherche et en extrayant ses enregistrements

et attributs en fonction des informations rendues (Figueiredo et al., 2017). La méthode DERIN proposée peut détecter différentes structures d'enregistrement à l'aide de techniques basées sur un modèle à n-grammes. De plus, cette méthode n'a pas besoin d'exemples pour apprendre à extraire les données, elle exécute indépendamment un domaine donné et peut détecter des enregistrements qui ne sont pas des enfants du même élément parent dans l'arborescence Document Object Model (DOM). Les résultats expérimentaux utilisant des pages Web de plusieurs domaines montrent que DERIN est très efficace et donne de bons résultats par rapport à d'autres méthodes. C'est pour améliorer l'importance de la compréhension sur les corrélations entre les données que les méthodes d'extraction de données sont intéressantes. Il s'agit de trouver quelques modèles et règles d'association pour diverses analyses et fournir des aides à la décision tels que les systèmes de recommandations par catégorie et la détermination des changements de comportement possibles (Liao & Chang, 2016).

Dans le cadre des enquêtes sur les événements météorologiques extrêmes, *CHANG* a présenté une analyse de données innovante pour la météo à l'aide du Cloud Computing, intégrant à la fois des services d'application et des services de Sciences de Données ou Data Science (Chang, 2017). L'objectif de son étude est de traiter, d'analyser et de visualiser les données collectées, d'étudier les implications et de rendre compte des résultats significatifs. Dans un premier temps, à partir des données historiques de Sydney, Singapour et Londres, il a présenté une méthode de prévision des températures pour comparer les températures historiques et prévues. Ensuite, il a utilisé MapReduce pour analyser et visualiser les distributions de température aux États-Unis, avant, pendant et après la période de vortex polaire, ainsi comme au Royaume-Uni pendant et après les inondations. Les résultats ont montré, avec l'utilisation des techniques d'optimisation, une amélioration des performances entre 20% et 30% sous 60 nœuds dans le Cloud.

En termes de temps d'exécution et d'espace de stockage, les performances des algorithmes de génération des règles d'association dépendent essentiellement de l'extraction de jeux d'éléments fréquents qui est souvent basée sur la propriété de monotonie pour prédire avec certitude quand un jeu d'élément est fréquent et quand il est peu fréquent (Melucci, 2016).

Les motifs fréquents sélectionnés d'éléments sont utilisés pour générer les règles d'association qui peuvent se classer selon les niveaux de préférence (souvent avec un manque de précédentes informations sur l'utilisateur ou l'histoire de notation utilisateur-point (problème de démarrage à froid)) pour la spécificité de systèmes décisionnels tels que les systèmes de recommandation de services (Gonzalez Camacho & Alves-Souza, 2018).

Dans la littérature sur l'exploitation des règles d'association, différencier les influence de corrélation est connue pour être une tâche difficile (Monteserin & Armentano, 2018), étant donné que différents facteurs (p. ex. similarité et environnement) peuvent induire une corrélation statistique entre les variables et les différences observées. Il est donc utile de prendre des précautions dans l'analyse des influences de la propagation obtenue dans

chaque jeu de données particulier qui peuvent être multifactorielles parfois en raison de caractéristiques intrinsèques (c.-à-d. le nombre de transactions, temps moyen entre les transactions et ainsi de suite).

Les modèles de processus déclaratifs définissent le comportement des processus métier comme un ensemble de contraintes. La découverte de processus déclarative vise à déduire de telles contraintes à partir des journaux d'événements. Les techniques de découverte existantes vérifient la satisfaction des contraintes candidates sur le journal, mais négligent complètement leurs interactions. En conséquence, les contraintes inférées peuvent être contradictoires et leur interaction peut conduire à un modèle de processus incohérent qui n'accepte aucune trace. Dans un tel cas, la sortie s'avère inutilisable à des fins de promulgation, de simulation ou de vérification. De plus, le modèle découvert contient, en général, des redondances qui sont dues à des interactions complexes de plusieurs contraintes et qui ne peuvent pas être corrigées avec les approches d'élagage existantes. Ces problèmes peuvent être abordés en proposant une technique d'exploration de données qui résout automatiquement les conflits au sein des modèles découverts en éliminant les redondances. Le niveau d'intérêt est dicté par des critères de priorisation spécifiés par l'utilisateur sur un ensemble de journaux d'événements du monde réel (Di Ciccio et al., 2017).

1.4 Apprentissage profond pour la modélisation de données

L'intelligence artificielle peut être d'une grande utilité dans la prévision des catastrophes naturelles. Des auteurs ont mis l'accent sur l'application et le rôle de l'apprentissage profond dans la compréhension de la prédiction des paramètres météorologiques.

Jancic et ses collègues ont utilisé les processus gaussiens profonds pour identifier un système dynamique et évaluer la méthode de manière empirique (Jancic et al., 2018). Cette méthode fournit une prévision en avance des valeurs de sortie du système avec son incertitude, qui peuvent être utilisées de manière avantageuse. En particulier, ils ont étudié la prédiction à court terme de la température de l'air dans les basses couches.

Hu et ses collègues ont proposé un modèle hybride constitué de la transformation empirique en ondelettes, de l'algorithme de propagation des attentes et de la régression du processus Gaussien avec le modèle d'observation établi pour la prévision à court terme de la vitesse du vent (Hu et al., 2017). L'approche proposée extrait d'abord des informations utiles d'une série de vitesses de vent à court terme et modélise ensuite l'incertitude inhérente et les caractéristiques dynamiques de la série chronologique de la vitesse du vent. Cette approche a été testée sur les données obtenues d'un parc d'éoliennes situé dans le nord-ouest de la Chine. Les résultats informatiques démontrent que la proposition de l'approche hybride génère une précision prédictive ponctuelle et des performances de prévision par intervalles satisfaisantes.

Le processus gaussien a été utilisé par Abdelfatah et ses collègues avec des entrées incertaines pour obtenir des prévisions approximatives des résultats du modèle (Abdelfatah et al., 2018). Ils ont proposé une performance du modèle empilé non paramétrique dans la composition du modèle et les prédictions en cascade. Ce modèle a été testé dans un problème de feux de forêt et de ressources minérales à l'aide de données réelles, et son application à la prévision de série chronologique est démontrée dans un problème d'advection de bouffée en 2D.

Les processus gaussiens (GP) comprennent un puissant paradigme d'apprentissage automatique basé sur le noyau qui a récemment attiré l'attention de la communauté d'identification de systèmes non linéaires, en particulier en raison de son traitement inhérent de style bayésien de l'incertitude. Cependant, comme les modèles GP standard supposent une distribution gaussienne pour le bruit d'observation, les capacités d'apprentissage et de prédiction de ces modèles peuvent être gravement dégradées lorsque des valeurs aberrantes sont présentes dans les données. Mattos et ses collègues présentent un modèle d'apprentissage robuste basé sur des processus gaussiens (dénommé RGP-t) avec des données contenant des valeurs aberrantes (Mattos et al., 2017). Ce modèle modélise explicitement la couche d'observation et permet d'entraîner la dynamique du système directement à partir de données d'estimation contaminées par des valeurs aberrantes. L'approche proposée a été testée dans plusieurs niveaux de contamination aberrants. Les résultats de simulation obtenus par le modèle RGP-t indiquent une résilience impressionnante aux valeurs aberrantes et une capacité supérieure à apprendre la dynamique non linéaire directement à partir de données fortement contaminées par rapport aux modèles GP existants.

Des modèles de régression de processus gaussiens ont été combinés avec les approches d'imputations multiples par Liu et ses collègues pour effectuer des prévisions d'énergie éolienne à court terme dans le cadre du scénario de données manquantes (Liu et al., 2018). L'algorithme de maximisation des attentes est utilisé pour estimer les composants de mélange de la distribution de données sous-jacente ainsi que pour traiter les données manquantes. Les résultats expérimentaux démontrent que comparativement à plusieurs approches existantes, la méthode proposée est efficace pour la prévision de l'énergie éolienne avec des données manquantes.

Zhu et ses collègues ont mis au point un modèle de diffusion gaussien incertain ou Gaussian Diffusion Model (UGD) pour faciliter la gestion des systèmes de production-émission ou Production-Emission System (PES) dans les villes dépendantes du charbon (Zhu et al., 2019). Le modèle UGD-PES peut non seulement traiter des incertitudes exprimées sous forme de distributions de probabilité, mais aussi quantifier les imprécisions subjectives présentées comme des ensembles flous avec une attitude de préférence du risque. Le modèle UGD-PES développé est ensuite appliqué à une étude de cas de la ville de Yulin dans un contexte multi-industries, multi-émissions et multi-périodes, afin d'identifier les politiques industrie-environnement actuelles dans cette ville dépendant du charbon. Les résultats obtenus peuvent aider les gestionnaires à identifier les conceptions de système souhaitées, ce qui peut être bénéfique pour la conception d'un

mode industriel plus propre et d'une politique optimisée entre l'industrie et l'environnement afin d'alléger la pression sur la gouvernance environnementale atmosphérique due aux émissions industrielles.

Dans le cadre de la planification et l'utilisation de l'énergie éolienne, un nouveau modèle hybride combinant le Shared Weight Long Short-Term Memory Network avec le Gaussian Process Regression (SWLSTM-GPR) a été proposé par Zhang et ses collègues pour obtenir un résultat de prédiction probabiliste fiable de la vitesse du vent (Zhendong Zhang et al., 2019). Ce modèle est appliqué à quatre cas de prévision de la vitesse du vent en Mongolie, Chine et comparé aux méthodes de prédiction de la vitesse du vent à la pointe de la technologie selon quatre aspects : exactitude de la prévision ponctuelle, pertinence de la prédiction des intervalles, prédiction de la probabilité, performances complètes et temps d'entraînement. Le test de fiabilité de SWLSTM-GPR garantit que les résultats de la prévision sont fiables et convaincants. Les résultats expérimentaux montrent que le SWLSTM-GPR peut obtenir une prédiction ponctuelle de haute précision, un intervalle de prédiction approprié et des résultats de prédiction probabilistes fiables avec un temps d'entraînement plus court sur les problèmes de prédiction de la vitesse du vent. Le test de fiabilité de SWLSTM-GPR garantit que les résultats de la prévision sont fiables et convaincants. Les résultats expérimentaux montrent que le SWLSTM-GPR peut obtenir une prédiction ponctuelle de haute précision, un intervalle de prédiction approprié et des résultats de prédiction probabilistes fiables avec un temps d'entraînement plus court sur les problèmes de prédiction de la vitesse du vent.

Dans le contexte d'amélioration de la fiabilité des systèmes de prévision d'ensemble météorologique et hydrologique, l'expérience de Abaza et ses collègues s'applique sur trois grands bassins versants et repose sur la combinaison de deux produits de météorologie pré-diffusés : les rediffusions canadiennes de 4 membres du Centre canadien de prévision météorologique et environnementale ou Canadian Centre for Meteorological and Environmental Prediction (CCMEP) et les rediffusions américaines de 10 membres de la National Oceanic and Atmospheric Administration (NOAA), conduisant à 14 membres à chaque pas de temps (Abaza et al., 2017). Les résultats montrent que les quatre Hydrological Ensemble Prediction Systems (H-EPS) testés conduisent à des valeurs de résolution et de netteté assez similaires, avec un avantage sur Distribution Based Scaling (DBS) + Ensemble Kalman filter (EnKF). L'ensemble Bayesian Model Averaging (BMA) n'est pas en mesure de compenser un quelconque biais laissé dans les prévisions d'ensemble de précipitations. D'autre part, il réussit à calibrer les membres de l'ensemble qui sont par ailleurs sous-dispersés. Si la fiabilité est préférée à la résolution et à la netteté, l'ensemble BMA DBS + EnKF + fonctionne mieux, en utilisant les deux processeurs du système H-EPS. Inversement, pour une résolution et une netteté améliorée, la méthode DBS est la méthode préférée.

1.5 Problématique météorologique abordée

Les phénomènes météorologiques et le climat sont la résultante de nombreux processus dynamiques, thermodynamiques, biologiques à différentes échelles de temps qui

interagissent entre eux. Autant dire que les modèles numériques du temps et du climat doivent représenter une extrême complexité. Nous nous intéresserons dans le présent travail à une petite partie des modèles de prévision mais qui est pourtant essentielle dans les prévisions du temps et du climat. Il s'agit des interactions entre la surface et l'atmosphère, c'est-à-dire comment la surface transmet de la chaleur et de la vapeur d'eau à l'atmosphère, chaleur et vapeur d'eau qui sont les moteurs des phénomènes atmosphériques.

Pour évaluer les échanges simulés par le modèle météorologique entre la surface et l'atmosphère nous utiliserons huit grandeurs que je vais présenter rapidement ici.

- Le rayonnement global qui est l'éclairement énergétique solaire d'une surface horizontale. Le rayonnement global arrive au sol en observant deux types de trajectoire possibles : (i) Celle du rayonnement direct, où le parcours des rayons est celui d'une droite entre le Soleil et notre planète ; (ii) Celle du rayonnement diffus, où le parcours des rayons est modifié par une succession d'obstacles (gouttelettes et cristaux de glace des nuages, aérosols, molécules d'azote et d'oxygène).
- Le deuxième paramètre est la pluviométrie, c'est-à-dire la quantité d'eau tombée à la surface.
- La température de l'air, le module et la direction du vent sont des grandeurs classiques communément utilisées en météorologie et font partie des paramètres étudiés.
- Pour caractériser l'humidité de l'air, nous utiliserons l'humidité relative. Elle s'exprime en pourcentage de vapeur d'eau que la particule d'air peut contenir. Lorsque l'humidité relative est à 50%, la particule contient donc seulement la moitié de la vapeur d'eau qu'elle peut contenir. Lorsque l'humidité relative est à 100%, la particule d'air est saturée en vapeur d'eau et un nuage peut se former.
- Enfin les deux paramètres qui indiquent l'intensité des transferts entre la surface et l'atmosphère sont le flux de chaleur sensible (pour les transferts d'énergie par convection) et le flux de chaleur latente (pour les transferts d'énergie par évaporation). Ces deux flux de chaleur dépendent bien de nombreux paramètres mais en premier lieu du rayonnement global, et des paramètres météorologiques précédemment cités comme la température, l'humidité, le vent et la pluviométrie.

1.6 Conclusion - Contributions

La plupart des auteurs ont utilisé les techniques de classification et de régression aux données météorologiques mesurées pour la prédiction des phénomènes dangereux tels que la prévision de la pollution atmosphérique grave, la prévision de l'énergie éolienne, etc (Liu et al., 2018). La fiabilisation de cette prédiction suscite un intérêt toujours plus vif, étant donné que les modèles numériques ont des problèmes de fiabilité dans la simulation des paramètres météorologiques (Aguera-Pérez et al., 2018), (Abaza et al., 2017). Ce faisant, ce travail consiste à mettre l'accent sur la modélisation statistique des données météorologiques mesurées et simulées tout en proposant deux méthodologies pour contribuer à l'amélioration de la fiabilité dans la simulation des paramètres météorologiques.

- La première méthodologie proposée consiste à utiliser la technique de découverte de corrélation entre les paramètres météorologiques pour (i) comprendre le croisement des différences/erreurs entre modèle et observation ; et (ii) mettre en évidence les faiblesses du modèle numérique de prévision avec précision.
- La seconde approche proposée consiste à élaborer une technique de réduction des erreurs de simulation. Pour cela, dans un premier temps, une méthode d'apprentissage automatique dénommée Gaussian Processes for Machine Learning (GPML) par régression est utilisée pour la modélisation et la prise en compte des incertitudes des variables météorologiques mesurées ; ceci pour faciliter le diagnostic du modèle numérique de prévision. Ensuite, un optimiseur est défini à partir des propriétés sur les transformations géométriques par homothétie. Cet optimiseur permet de transporter des données simulées vers les voisinages des données mesurées pour réduire les erreurs de simulation.

Chapitre 2

2.Présentation et visualisation statistique des données météorologiques mesurées et simulées

2.1 Introduction

En Data mining, le problème de fiabilité dans la découverte de connaissances est important. La fiabilité des informations extraites dépend du choix de la technique de prétraitement de données. Par conséquent, dans le contexte de prétraitement de données météorologiques mesurées/simulées, la visualisation statistique joue un rôle très important.

Dans ce chapitre, l'objectif principal est d'analyser statistiquement les distributions des variables/erreurs, en vue de mieux choisir les bornes des intervalles de discrétisation. Ceci permet de faire une répartition pertinente des échantillons entre les intervalles, et qui pourra donc améliorer la qualité des informations extraites.

Cette méthode de prétraitement est déployée sur une base de données contenant des variables météorologiques mesurées/simulées sur le site expérimental du Centre de Recherches Atmosphériques (CRA) qui est l'un des deux sites composant la Plateforme Pyrénéenne d'Observation de l'Atmosphère (P2OA) en France. Dans ce travail, les variables considérées sont les suivantes : les flux de chaleur sensible et latente, la température de l'air, l'humidité relative de l'air, la vitesse du vent avec sa direction, la pluie, et le rayonnement global.

Après cette introduction, la *section 2.2* désigne la collecte de données mesurées et simulées, suivie des explications sur les incertitudes de la mesure en *section 2.3,* la *section 2.4* présente la distribution de chaque variable avec les distributions des erreurs correspondantes, et en fin une conclusion en *section 2.5*.

2.2 Collecte de données mesurées et simulées

Données mesurées : l'observation est le point de départ de toute prévision météorologique. Les observations sont la matière première utilisée par le météorologiste pour prévoir le temps, et par le climatologue pour étudier le climat (Guillemot, 2010). Cette connaissance permet de comparer le temps d'aujourd'hui à celui d'hier pour prévoir

le temps de demain. Dans ce cas d'étude, les observations disponibles sont des mesures des stations météorologiques de surface du système opérationnel de météo-France. Les variables mesurées sont celles de la station du Centre de Recherche Atmosphérique (CRA) de Lannemezan.

Données simulées : les météorologistes sont aujourd'hui aidés dans leur tâche par les modèles numériques de prévision. Le monde de l'observation météorologique s'est ainsi structuré pour fournir des données de qualité, capables de renseigner ces modèles et par la suite fournir des résultats que constitue la base de données simulées (Koci & Cerny, 2017), (Docine et al., 1999). Les données simulées obtenues sont celles qui correspondent géographiquement à la station de Campistrous, mais le modèle simule la météo sur toute l'Europe. Le modèle concerné couvre donc toute la France et les pays voisins avec une résolution **horizontale** de 1.3 km (cela signifie que les données représentent une moyenne sur un carré de 1.3×1.3 km² et en plus moyennées à l'heure) et 90 niveaux verticaux avec un premier niveau à 5 m au-dessus de la surface. Le domaine géographique couvert est représenté sur la *Figure 2-1*.

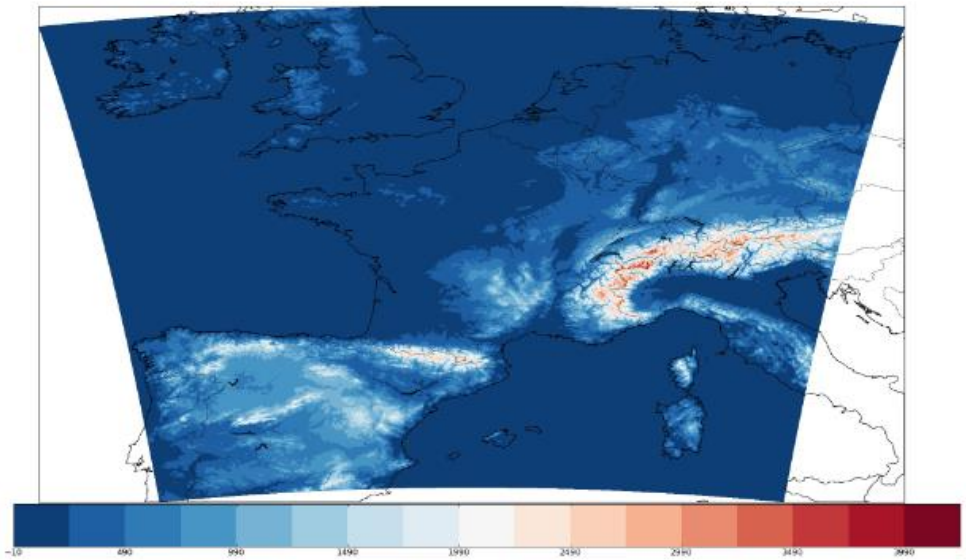


Figure 2-1 : Domaine de calcul du modèle concerné avec une résolution horizontale de 1.3 km et 90 niveaux verticaux au-dessus de la surface, source : http://www.umr-cnrm.fr/

Ce modèle a été conçu pour améliorer la prévision des phénomènes convectifs dangereux, des événements locaux et de la météorologie de basses couches. Il produit des prévisions très détaillées, que les prévisionnistes utilisent pour affiner leurs prévisions à petite échelle, notamment en termes d'anticipation et de localisation des phénomènes météorologiques potentiellement dangereux, tels que les orages, les crues soudaines, les rafales, les précipitations intenses, turbulence, visibilité, etc.

Ces prévisions fournissent des informations locales et précises de température, d'humidité, d'état du sol, d'état du ciel répondant aux besoins des citoyens dans leur vie quotidienne. Le modèle prend en compte les observations de vent produites par la plupart des radars Doppler météorologiques. Il intègre également les observations de

précipitations fournies par ces mêmes radars. Il intègre aussi les données issues des stations automatiques au sol, des radiosondages, des stations GPS, des satellites, etc. Toutes ces données renseignent sur la position des nuages à l'origine des pluies intenses et des orages ainsi que sur la distribution et l'intensité des précipitations sur toute l'Europe.

Mais dans ce cas d'étude, nous nous intéressons à la météorologie de basses couches qui est une petite partie du modèle. Il s'agit des données simulées qui correspondent géographiquement à la station de Campistrous (Centre de Recherche Atmosphérique (CRA) de Lannemezan). Les données de cette station sont assimilées dans le modèle comme tant d'autres en France et en Europe.

Taille des variables : les données mesurées et simulées sont des valeurs horaires sur l'année 2016. L'ensemble des variables considérées constitue une base de données transactionnelles de taille $N = 8784$.

2.3 Incertitudes de la mesure

Les incertitudes ou imprécisions de la mesure sont dues principalement à des phénomènes perturbateurs (C. Li et al., 2016). Ces principaux phénomènes perturbateurs concernant les stations de mesure sont : les imprécisions de l'instrument de mesure, le mauvais fonctionnement des senseurs, les imprécisions liés à la calibration des capteurs, les imprécisions liées à l'installation de la station de mesure (emplacement, ...) et les erreurs d'observation ou de lecture des données. Ces phénomènes perturbateurs sont les quantifiés par les scientifiques en parlant de précision de la mesure. Le tableau suivant résume les mesures de précision de chaque variable considérée.

Tableau 7 : Les mesures de précision des variables

Variables	Précisions
Rayonnement global	20 W/m²
Chaleur sensible	15 W/m²
Chaleur latente	15 W/m²
Humidité relative	5 %
Température	1 °C
Vitesse du vent	1 m/s
Pluie	1 L/m²

2.4 Distribution statistique et visualisation des variables

Cette section consiste à visualiser et analyser statistiquement les données mesurées et les données simulées dans le but de mettre en évidence l'existence des différences (erreurs)

entre les variables mesurées et simulées ; mais aussi mieux comprendre pourquoi un prétraitement rigoureux des données à partir des distributions statistiques des variables considérées. Ainsi, le schéma suivant explique l'importance de la visualisation des distributions statistiques (*Figure 2-2*).

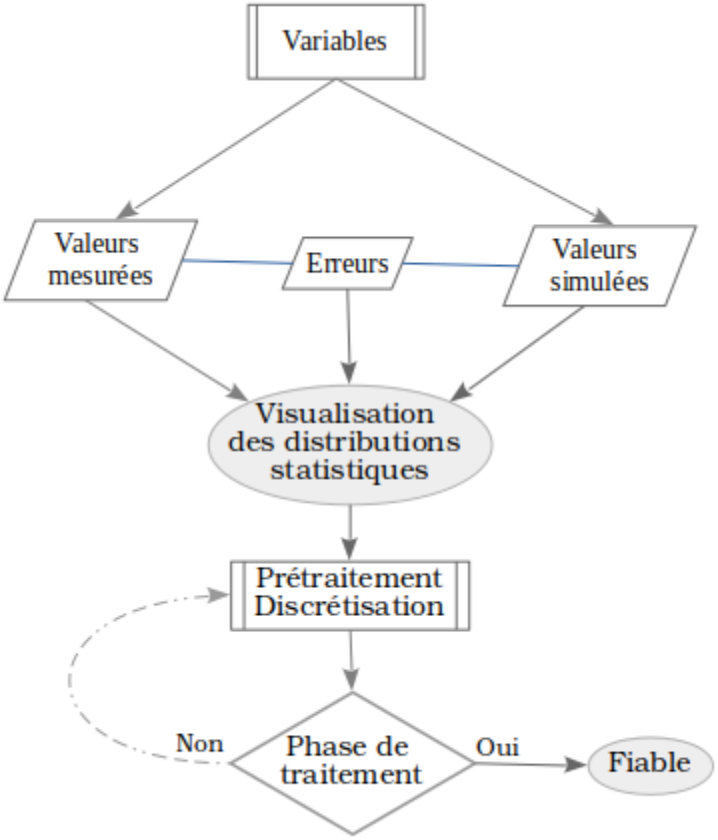


Figure 2-2 : Schéma expliquant pourquoi la visualisation des distributions statistiques

Cette section explique d'abord comment les deux bases de données ont été constituées, notamment la base de données mesurées et la base de données simulées, ensuite elle présente les distributions statistiques des valeurs horaires mesurées et simulées pour chaque variable considérée sur l'année 2016, en plus des distributions des erreurs correspondantes. Concernant la discrétisation en phase de prétraitement, l'approche adoptée est détaillée dans la *section 3.2.2* ; et la phase de traitement pour la découverte de connaissances est décrite dans les *sections 1.2.2.3* et *3.2.3*.

2.4.1 Distribution de la température (tt)

Définition 10. Température. La température, que l'on note ici tt, est considérée comme une grandeur physique liée à la notion imminente de chaud et froid. Elle désigne la manifestation, à l'échelle macroscopique, du mouvement des atomes et molécules. Ainsi une température élevée signifie une grande « agitation » atomique. Elle s'exprime généralement en kelvin (K) ou en degré Celsius (°C).

Les graphiques suivants illustrent les distributions de température simulée (*Figure 2-3*) et de température mesurée (*Figure 2-4*) avec un diagramme en bâtons.

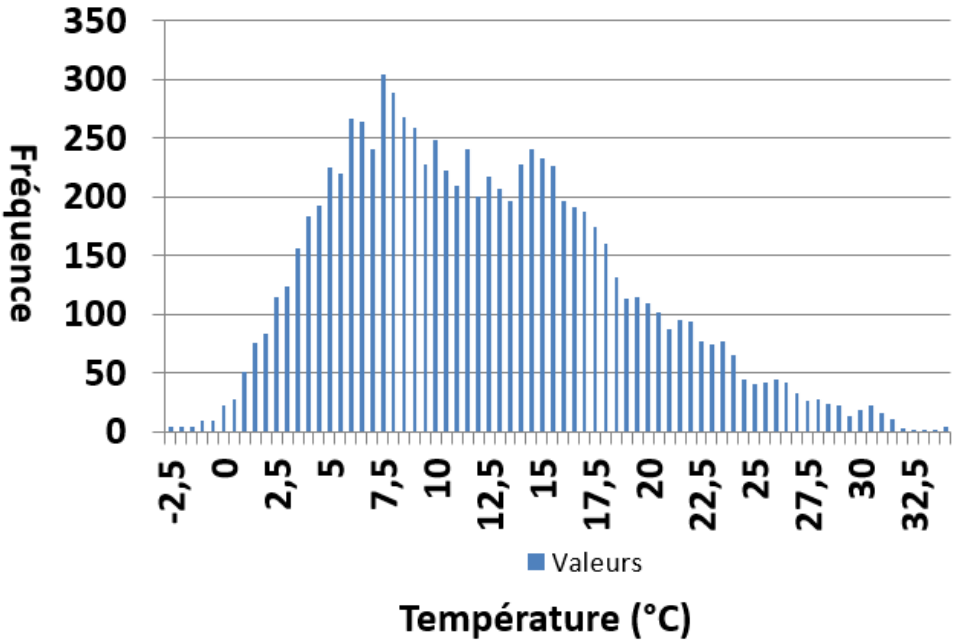


Figure 2-3 : Diagramme de température simulée

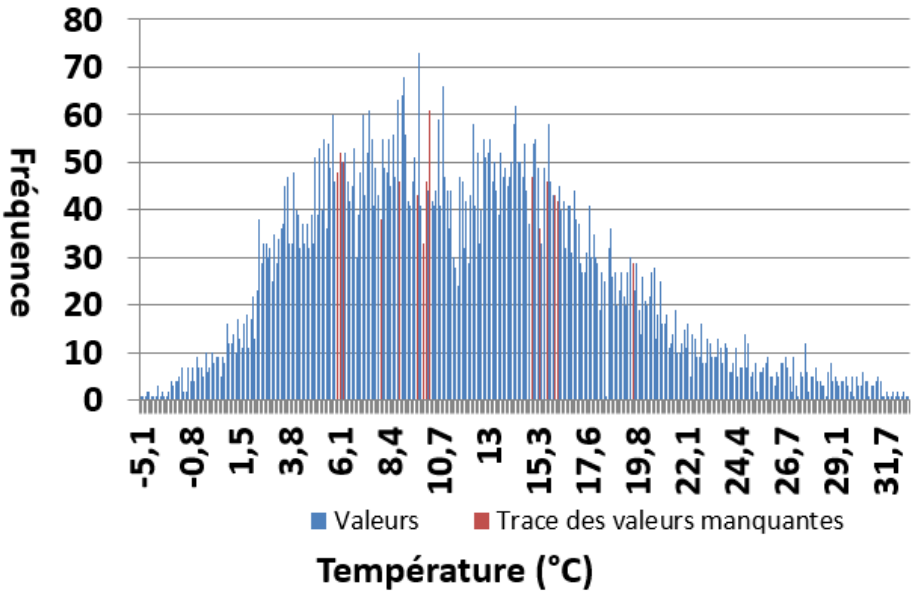


Figure 2-4 : Diagramme de température mesurée

Les deux diagrammes suivent la même forme de distribution. Par observation graphique, la seule différence est que :

$$min\ tt_{simulée} > min\ tt_{mesurée}$$

Les barres en couleur rouge (*Figure 2-4*) représentent la traçabilité des valeurs manquantes imputées. Le graphique suivant (*Figure 2-5*) met en évidence les différences entre température simulée et température observée.

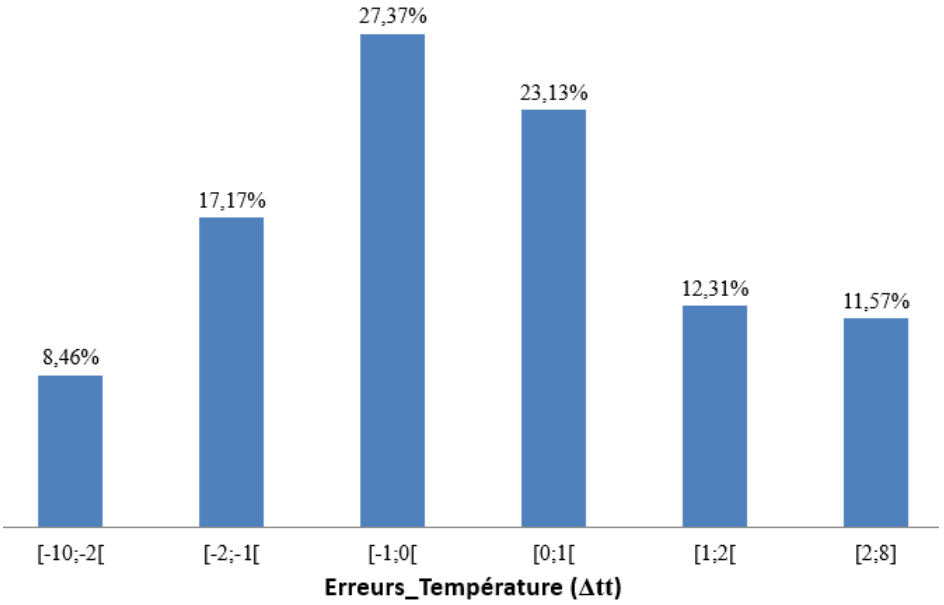


Figure 2-5 : Histogramme des erreurs sur la température

Par observation graphique, environ 50% des biais entre la température simulée et la température mesurée se situe dans l'intervalle $[-1; 0[\cup [0; 1[$. En outre, environ 26% de la différence entre la température simulée et la température observée se trouve dans l'intervalle $[-2; -1[\cup [1; 2[$. Enfin, environ 24% des biais entre la température simulée et la température observée se situe dans l'intervalle $[-10; -2[\cup [2; 8]$.

2.4.2 Distribution de la vitesse du vent (ff)

Définition 11. Vent. En météorologie, le vent désigne le mouvement horizontal de l'air et sa mesure comprend deux paramètres, notamment sa direction et sa vitesse qui sont mesurées par la girouette et l'anémomètre. La vitesse du vent (ff) est exprimée habituellement en km/h ou m/s.

Les graphiques suivants illustrent les distributions de la vitesse du vent simulé (*Figure 2-6*) et mesuré (*Figure 2-7*) avec un diagramme en bâton.

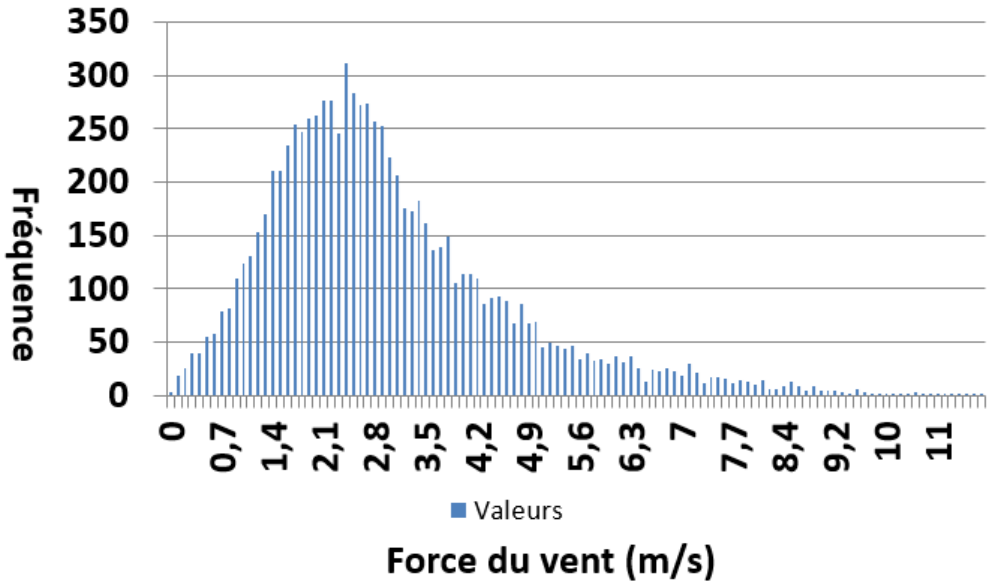


Figure 2-6 : Diagramme de la vitesse du vent simulé

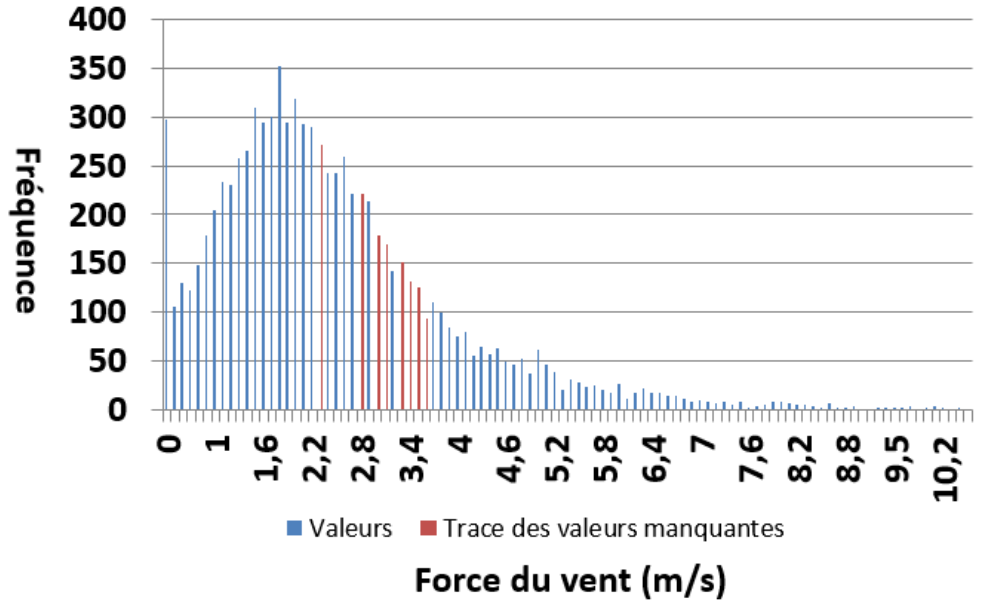


Figure 2-7 : Diagramme de la vitesse du vent mesuré

Les deux diagrammes suivent la même forme de distribution. La différence est que :

$$min\ ff_{simulée} > min\ ff_{mesurée}$$

$$max\ ff_{simulée} > max\ ff_{mesurée}$$

Le graphique suivant (*Figure 2-8*) met en évidence les différences entre la vitesse du vent simulée et la vitesse du vent mesurée.

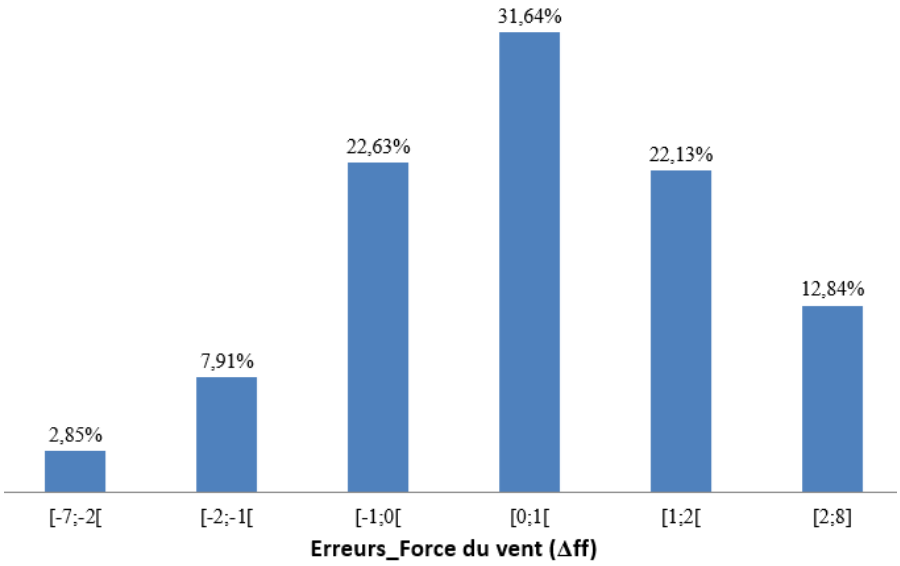


Figure 2-8 : Histogramme des erreurs sur la vitesse du vent

Par observation graphique, environ 54% des biais entre la vitesse du vent simulé et la vitesse du vent mesuré se situe dans l'intervalle $[-1;0[\cup[0;1[$. En outre, environ 11% des biais entre la vitesse du vent simulé et la vitesse du vent mesuré se trouve dans l'intervalle $[-2;-1[\cup[1;2[$. Enfin, environ 35% des biais entre la vitesse du vent simulé et la vitesse du vent mesuré se situe dans l'intervalle $[-7;-2[\cup[2;8]$.

2.4.3 Distribution de l'humidité relative (hu)

Définition 12. Humidité relative. L'humidité relative de l'air, notée hu dans ce travail, est le rapport de la pression partielle de la vapeur d'eau contenue dans l'air sur la pression de vapeur saturante à la même température. Elle est donc une mesure du rapport entre le contenu en vapeur d'eau de l'air et sa capacité maximale à en contenir dans ces conditions. Ce rapport s'exprime en pourcentage (%) et varie en fonction de la température/pression. L'humidité relative de l'air est mesurée à l'aide d'un hygromètre.

Les graphiques suivants illustrent les distributions de l'humidité relative simulée (*Figure 2-9*) et de l'humidité relative mesurée (*Figure 2-10*) avec un diagramme en bâton.

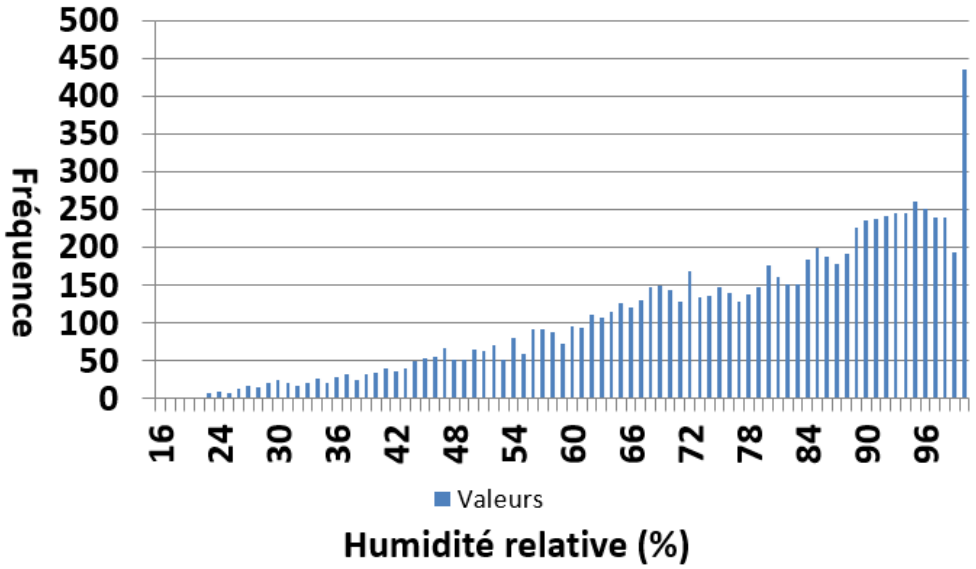


Figure 2-9 : Diagramme de l'humidité relative simulée

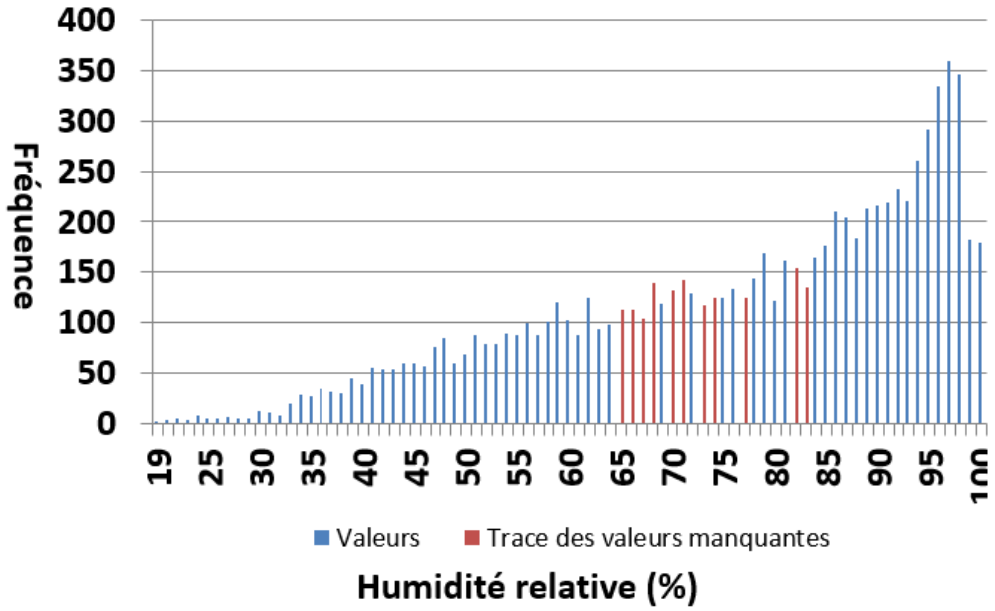


Figure 2-10 : Diagramme de l'humidité relative mesurée

Pour mieux distinguer les différences, le graphique suivant (*Figure 2-11*) illustre la mise en évidence des différences entre l'humidité relative simulée et l'humidité relative mesurée.

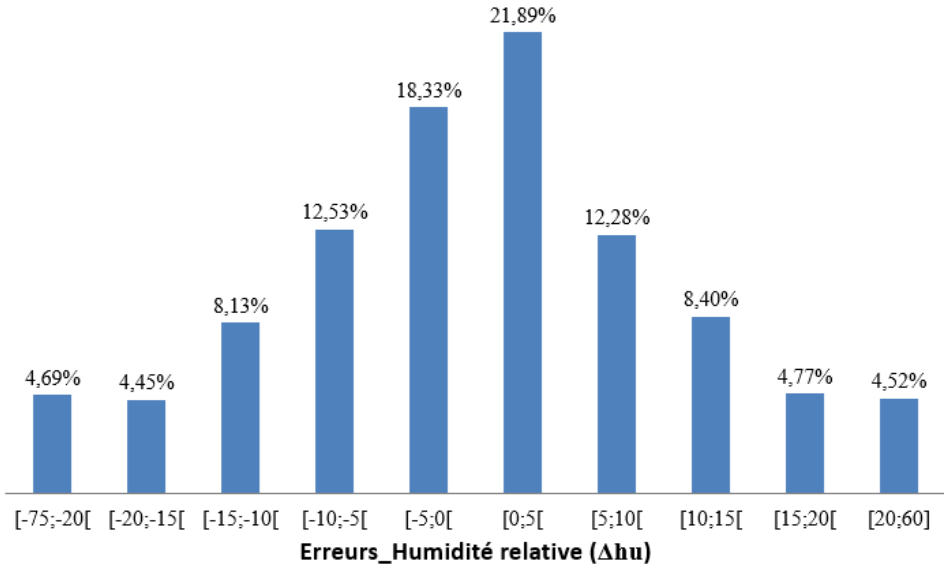


Figure 2-11 : Histogramme des erreurs sur l'humidité relative

Compte tenu de la précision de 5% des mesures de l'humidité relative (*Figure 2-11*), environ 40% de l'humidité relative simulée, sont considérés comme proches des valeurs mesurées, ie $\Delta hu \in [-5; 0[\cup [0; 5[$. Par contre, environ 36% des erreurs se situent dans l'intervalle $[-75; -10[\cup [10; 60]$.

2.4.4 Distribution de la pluie (rr)

Définition 13. La pluie est une des formes de précipitations. Par définition, les précipitations désignent les eaux qui tombent sur la surface de la Terre, tant sous forme liquide (bruine, pluie, averse) que sous forme solide (neige, grésil, grêle) ou déposée (rosée, gelée blanche, givre, …). La quantité d'eau tombée est mesurée durant un certain laps de temps que l'on exprime généralement soit en millimètres (mm), soit en litres par mètre carré (l/m²). 1 mm de précipitations correspond à 1 l d'eau par m². L'intensité de la pluie est la hauteur d'eau précipitée par unité de temps (généralement en mm/h).

Les graphiques suivants illustrent les distributions de la pluie simulée (*Figure 2-12*) et mesurée (*Figure 2-13*) avec un diagramme en bâton.

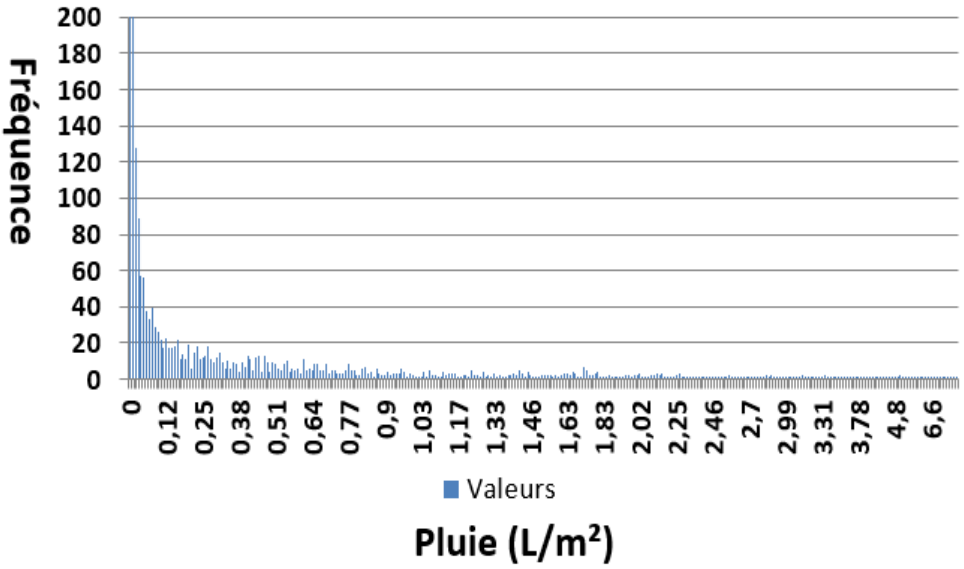


Figure 2-12 : Diagramme de la pluie simulée

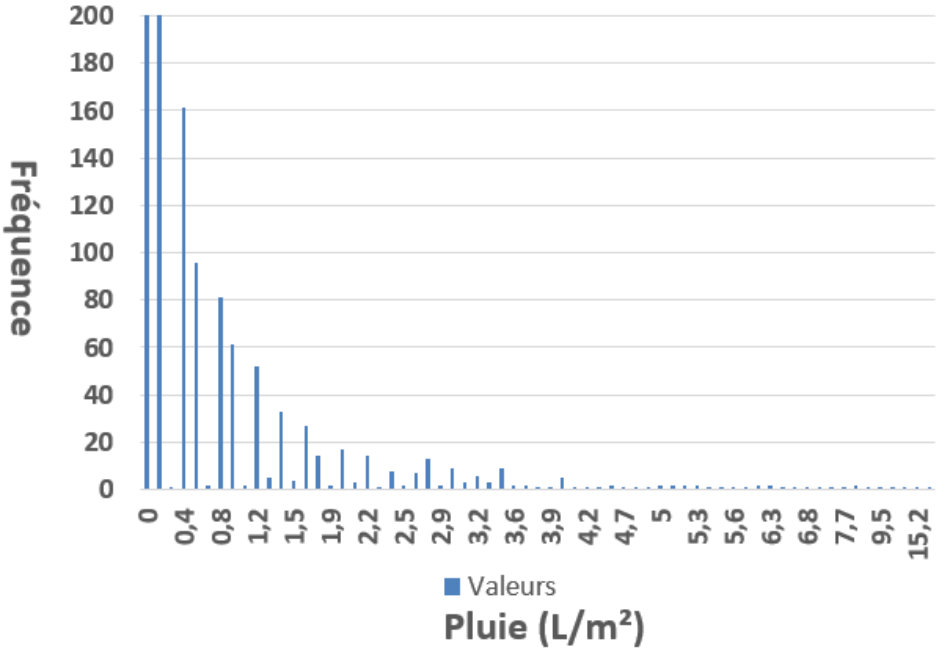


Figure 2-13 : Diagramme de la pluie mesurée

Pour mieux distinguer les différences, le graphique suivant (*Figure 2-14*) illustre la mise en évidence des différences entre pluie simulée et pluie mesurée.

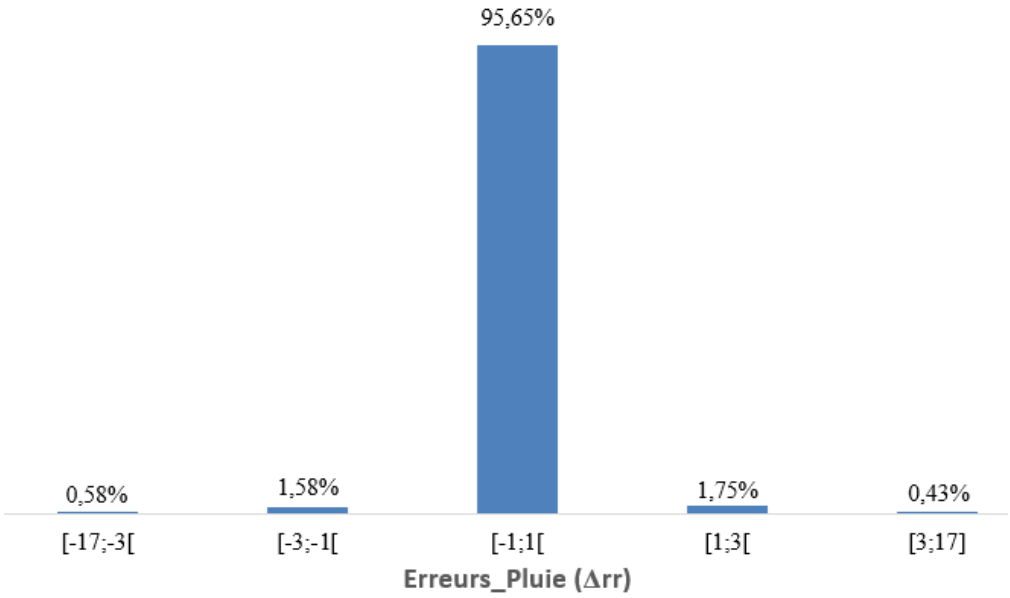


Figure 2-14 : Histogramme des erreurs sur la pluie

Le graphique ci-dessus (*Figure 2-14*) montre que 95% des valeurs simulées convergent vers les valeurs mesurées. En d'autres termes, à 95% $\Delta rr \in [-1; 1[$.

2.4.5 Distribution du rayonnement global (glo)

Définition 14. Rayonnement global. Le rayonnement solaire est l'ensemble des ondes électromagnétiques émises par le soleil. Le rayonnement solaire global peut être défini comme l'énergie rayonnante totale du soleil qui atteint une surface horizontale à la surface de la terre au cours d'une unité de temps précise.

Dans tous les secteurs de la carte d'ensoleillement et de la carte du ciel, le rayonnement global est calculé comme étant la somme des rayonnements direct et diffus.

Les graphiques suivants illustrent les distributions du rayonnement global simulé (*Figure 2-15*) et du rayonnement global mesuré (*Figure 2-16*) avec un diagramme en bâton. Ces distributions montrent comme attendu une très forte probabilité des petites valeurs qui sont celles mesurées la nuit, en début et fin de journée.

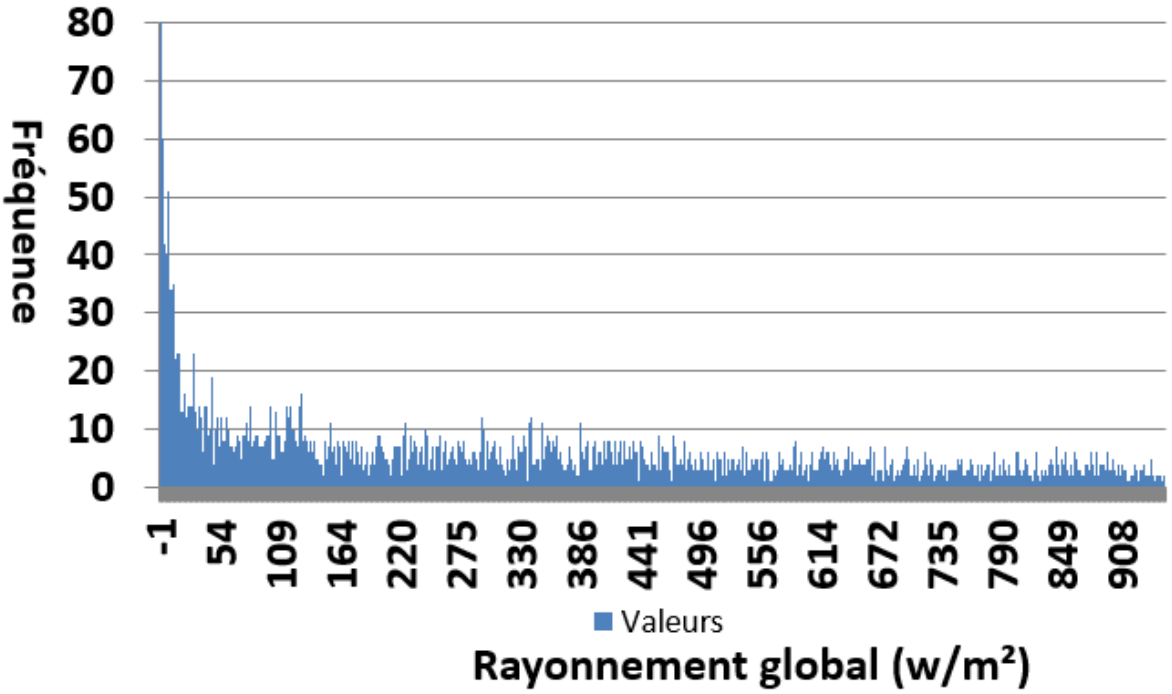


Figure 2-15 : Diagramme du rayonnement global simulé

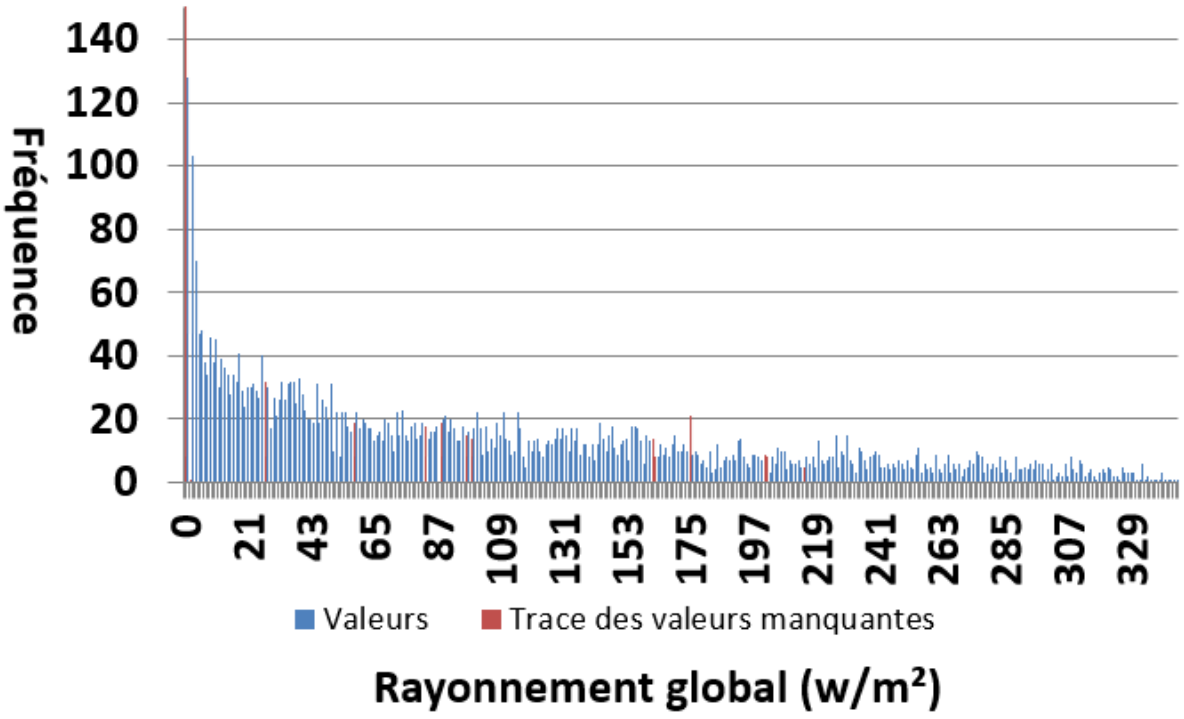


Figure 2-16 : Diagramme du rayonnement global mesuré

Pour mieux distinguer les différences, le graphique suivant (*Figure 2-17*) illustre la mise en évidence des différences entre rayonnement global simulé et rayonnement global mesuré.

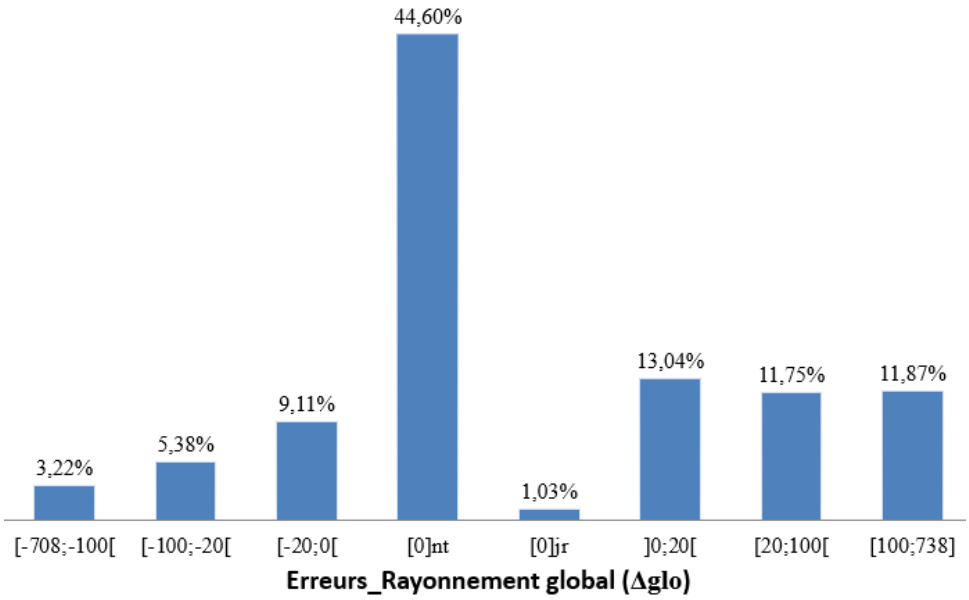


Figure 2-17 : Histogramme des erreurs sur le rayonnement global. Les intervalles affectés des indices nt et jr indiquent respectivement nuit et jour.

Le rayonnement global à la surface est surtout affecté par le lever du soleil, et aussi par plusieurs phénomènes comme les nuages ou la concentration d'aérosols. Les nuages réduisent particulièrement le rayonnement global en fonction de leur profondeur, de leur forme et de leur teneur en eau liquide. Une comparaison temporelle du rayonnement global observé et simulé en un endroit donné dépend fortement de ces paramètres qui varient dans le temps et l'espace. Par conséquent, un biais important entre le modèle et l'observation n'implique pas nécessairement que la simulation n'est pas bonne, mais que la simulation n'a peut-être pas généré le nuage exactement au même endroit ou au même moment que dans la réalité. Comme prévu, la comparaison des valeurs nocturnes (environ 44% des valeurs) conduit à un biais nul puisque le rayonnement global est nul tant dans les modèles que dans les observations. Seul 1% de la comparaison diurne présente un biais nul. Ainsi, compte tenu de la précision de la mesure et de la simulation, un biais de $20\ W/m^2$ semble raisonnable, environ 66% des biais sont dans l'intervalle des faibles erreurs $[-20; 0[\cup [0; 20[_{W/m^2}$; et 32% de la comparaison présente un biais supérieur ou égal à $20\ W/m^2$, dont environ 15% de biais sont dans l'intervalle des erreurs importantes $[-708; -100[\cup [100; 738]_{W/m^2}$, voir (*Figure 2-17*).

2.4.6 Distribution de la chaleur sensible (H)

Définition 15. Chaleur sensible. La chaleur sensible, notée H, est la quantité de chaleur qui est échangée, sans changement d'état physique, entre plusieurs corps. C'est-à-dire, la chaleur sensible est la chaleur émise/absorbée par un corps matériel dont la température varie en conséquence, mais sans changement d'état physique du corps matériel. Elle s'exprime en watt par mètre carré (*w/m²*).

Les graphiques suivants illustrent les distributions de la chaleur sensible simulée (*Figure 2-18*) et de la chaleur sensible mesurée (*Figure 2-19*) avec un diagramme en bâton.

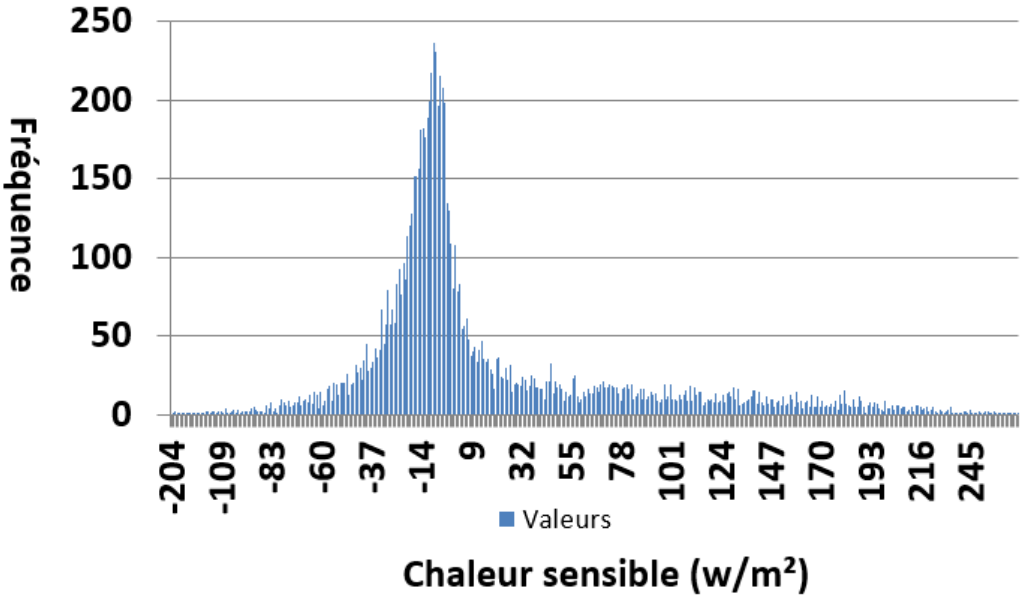


Figure 2-18 : Diagramme de la chaleur sensible simulée

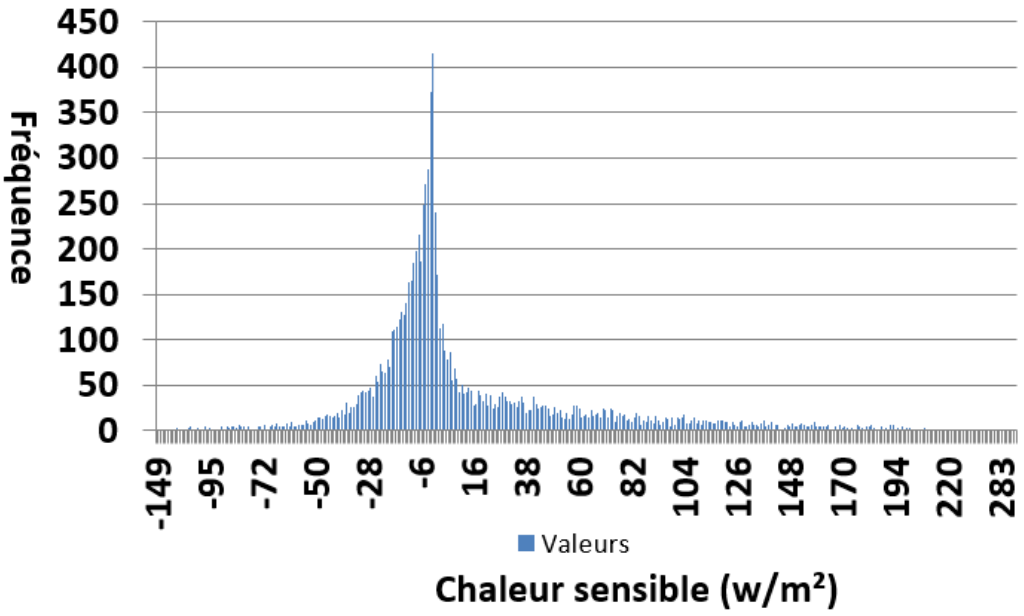


Figure 2-19 : Diagramme de la chaleur sensible mesurée

Pour mieux distinguer les différences, le graphique suivant (*Figure 2-20*) illustre la mise en évidence des différences entre chaleur sensible simulée et chaleur sensible mesurée.

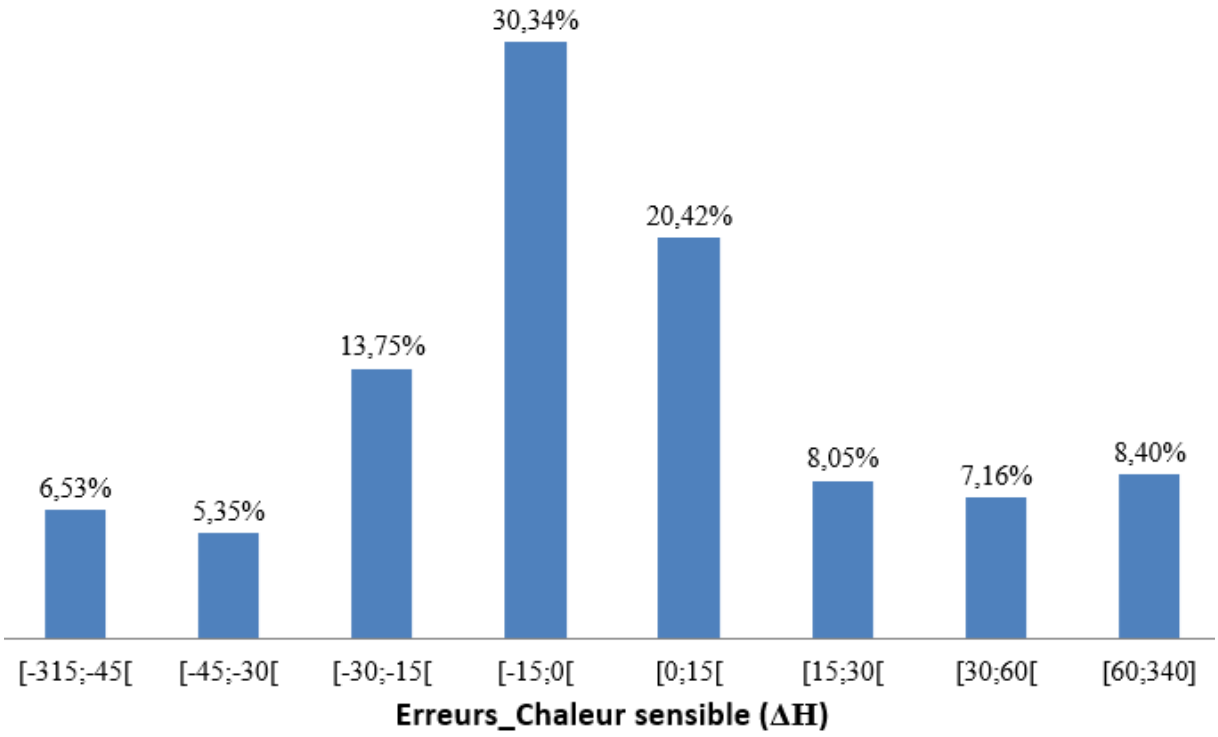


Figure 2-20 : Histogramme des erreurs sur la chaleur sensible

En considérant la mesure de précision de 15 W/m^2, le graphique ci-dessus (*Figure 2-20*) montre que 51% des valeurs simulées convergent vers les valeurs mesurées, ie à 51% $\Delta H \in [-15; 0[\cup [0; 15[_{W/m^2}$. En d'autres termes, à 51% nous avons une simulation à faible erreur du modèle ; et à 49% le modèle diverge, dont environ 15% de biais sont dans l'intervalle des erreurs importantes, ie à 15% $\Delta H \in [-315; -45[\cup [60; 340]_{W/m^2}$.

2.4.7 Distribution de la chaleur latente (LE)

Définition 16. Chaleur latente. La chaleur latente (ou enthalpie de changement d'état), notée LE, est la quantité d'énergie fournie pour changer l'état physique d'une masse (solide, liquide ou gazeux). Elle s'exprime aussi en watt par mètre carré (w/m²).

Les graphiques suivants illustrent les distributions de la chaleur latente simulée (*Figure 2-21*) et de la chaleur latente mesurée (*Figure 2-22*) avec un diagramme en bâton.

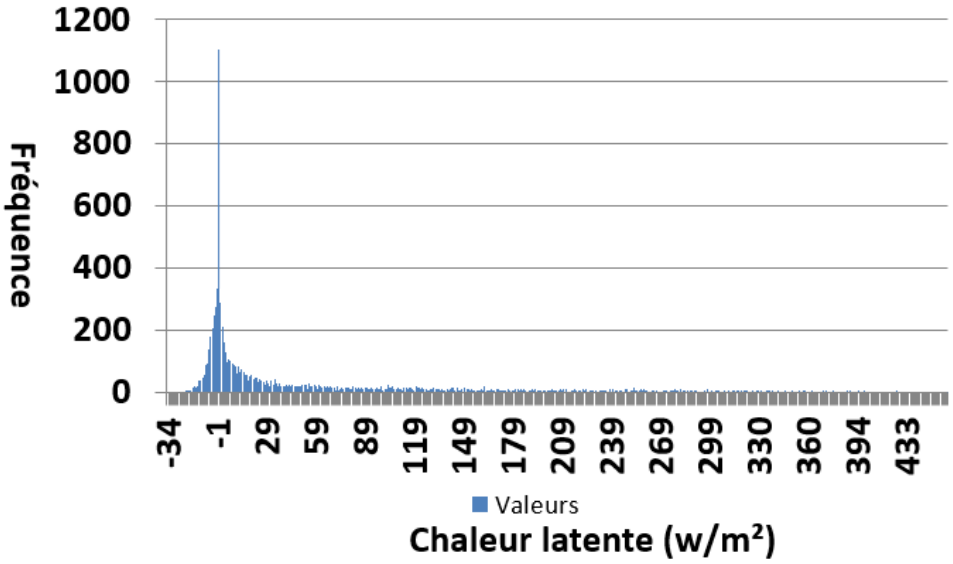


Figure 2-21 : Diagramme de la chaleur latente simulée

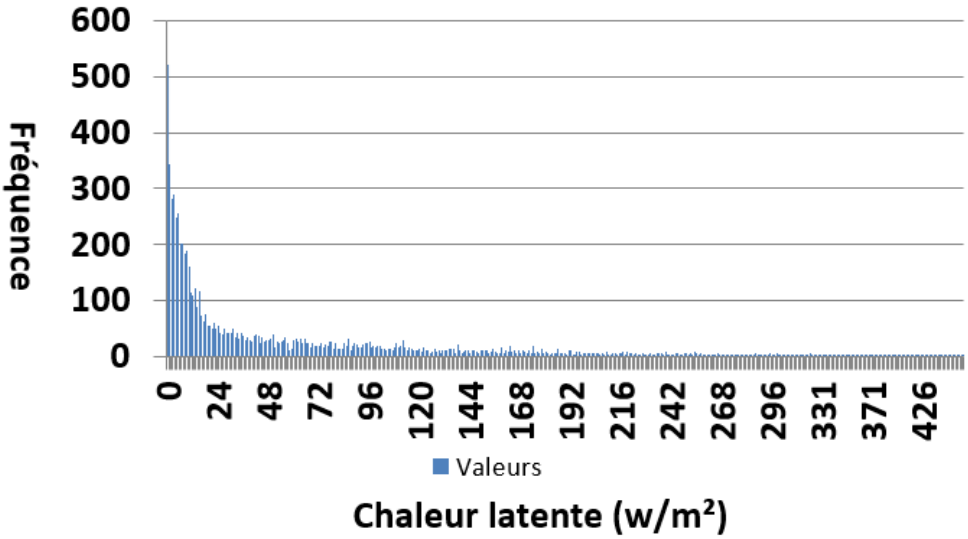


Figure 2-22 : Diagramme de la chaleur latente mesurée

Contrairement aux observations des flux de chaleur latente (*Figure 2-22*), on observe des valeurs négatives dans les flux de chaleur latente simulée (*Figure 2-21*). Ceci peut être interprété comme un des facteurs de dysfonctionnement du modèle. Le graphique suivant (*Figure 2-23*) illustre la mise en évidence des différences entre chaleur latente simulée et chaleur latente mesurée.

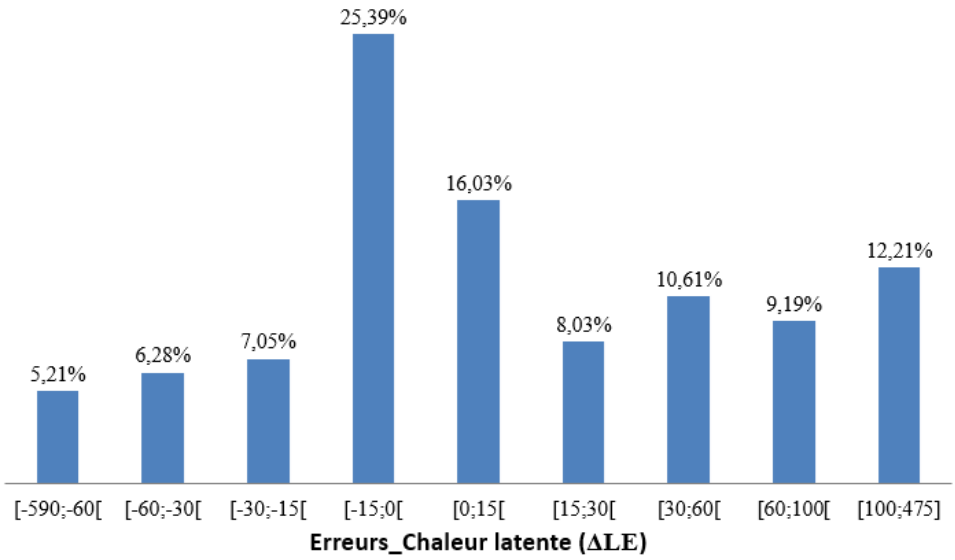


Figure 2-23 : Histogramme des erreurs sur la chaleur latente

De façon analogue, compte tenu de la mesure de précision de $15\ w/m^2$, le graphique ci-dessus (*Figure 2-23*) montre qu'environ 41% des valeurs simulées convergent vers les valeurs mesurées, ie à 41% $\Delta LE \in [-15; 0[\cup [0; 15[_{W/m^2}$. En d'autres termes, à 41% nous avons une simulation à faible erreur du modèle ; et à 59% le modèle diverge, dont environ 26% de biais sont dans l'intervalle des erreurs importantes, ie à 26% $\Delta LE \in [-590; -60[\cup [60; 475]_{W/m^2}$, voir (*Figure 2-23*).

2.5 Conclusion

Ce chapitre a présenté des données mesurées et simulées pour analyser des distributions de données météorologiques collectées dans le but de rendre ces données compréhensibles. Cette analyse de distribution, pour chaque variable (ou erreur) considérée, a permis d'adopter une technique de réduction/compression de données. Ceci donne le bon choix des bornes des intervalles de discrétisation en vue de faire une répartition pertinente des échantillons entre les intervalles, et qui pourra donc améliorer la qualité des informations extraites dans la fouille de données.

Chapitre 3

3.Apprentissage automatique basé sur des règles pour la découverte de connaissances dans les données météorologiques

3.1 Introduction

Dans le contexte de la prise des décisions complexes (par exemple des alertes météo), une meilleure fiabilité des prévisions météorologiques est importante (Aguera-Pérez et al., 2018). Cette fiabilité peut être améliorée (i) par des outils de traitement statistique (Abaza et al., 2017), et (ii) par des méthodes d'apprentissage automatique. Par exemple, une méthodologie, basée sur l'apprentissage automatique, a été présentée dans (Doycheva et al., 2017) pour améliorer la fiabilité des prévisions d'inondations. En plus, cette approche a été testée sur des prévisions de précipitations pour le bassin de la Mulde en Allemagne.

En résumé, la maîtrise de la prédiction des phénomènes dangereux (inondations, pollution de l'air, etc.) dépend de la fiabilité du modèle numérique de prévision du temps. Il y a donc un besoin de méthodes avancées orientées pour une meilleure compréhension de ce modèle et l'analyse des principaux paramètres associés. L'objectif de ce travail de recherche est d'utiliser l'extraction des connaissances à partir de données pour contribuer à l'évaluation de la simulation des processus de surface par le modèle numérique de prévision. Eu égard à son expressivité intuitive, la méthode d'extraction des règles d'association est choisie pour effectuer des comparaisons entre les observations effectuées et les simulations numériques réalisées. Cette méthode est déployée sur deux bases de données contenant des variables mesurées et simulées qui sont présentées dans le chapitre 2.

La méthodologie proposée dans ce chapitre comprend trois étapes : (i) prétraitement avec des techniques de gestion des données manquantes et de discrétisation (transformation des données quantitatives en des données qualitatives) ; (ii) traitement avec le Data Mining via des règles d'association pour découvrir des relations intéressantes entre les variables météorologiques mesurées/simulées ; (iii) post-traitement utilisant un raisonnement logique pour analyser ces règles et visualiser leurs liens pour une interprétation ultérieure.

Après l'introduction dans ce chapitre, la *section 3.2* présente la méthodologie proposée pour la découverte de connaissances. Ensuite, une étude de cas est abordée dans la *section 3.3*. La *section 3.4* consiste à discuter sur les résultats obtenus. Enfin, la *section 3.5* donne une conclusion.

3.2 Méthodologie adoptée pour la découverte de connaissances

Dans cette section, une méthodologie est proposée pour l'évaluation de la simulation des processus de surface par les modèles de prévision du temps. Elle comprend trois étapes : prétraitement, traitement et post-traitement. Le schéma suivant (*Figure 3-1*) illustre la méthodologie proposée :

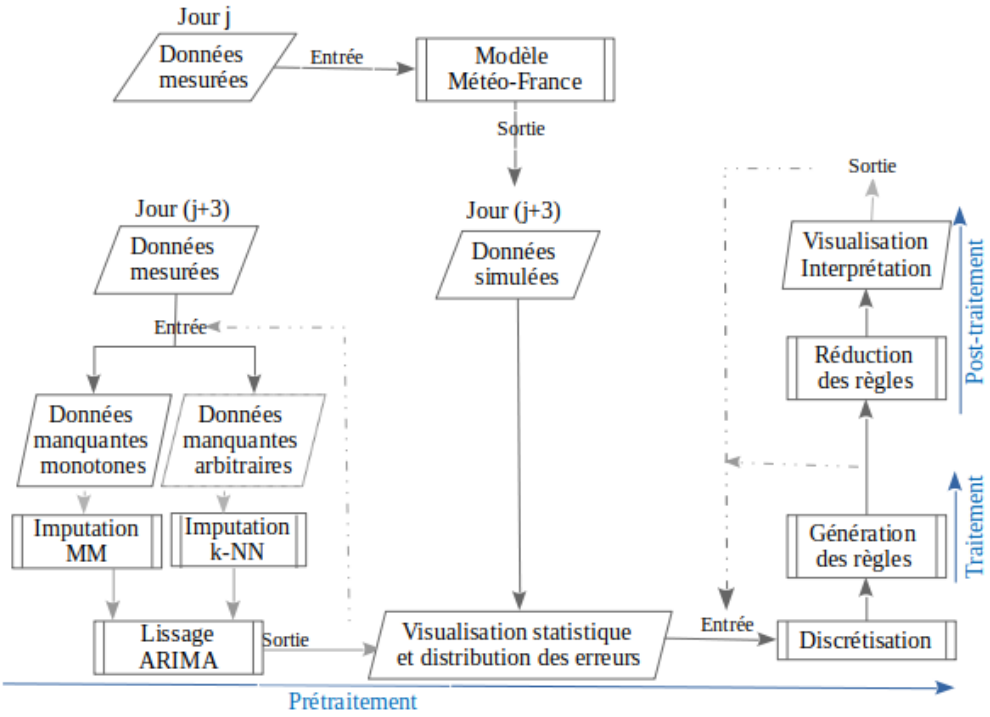


Figure 3-1 : Méthodologie proposée pour la découverte de connaissances

Cette méthodologie permet d'aborder la problématique suivante par une méthode du Data Mining (découverte de règles d'association) : diagnostiquer et évaluer la simulation du modèle numérique de prévision pour comprendre et identifier les paramètres qui influent sur les erreurs importantes dans la simulation des flux de chaleur sensible et chaleur latente. Les phases de prétraitement et post-traitement relèvent des contributions de cette thèse avec :

- la première étape (prétraitement) explique la technique de gestion des données imparfaites et la technique de discrétisation adoptée ; le processus de collecte de données, la visualisation statistique des données et les distributions des erreurs sont décrits dans le chapitre précédent (voir *section 2.2*).

- la seconde étape est la phase de traitement qui consiste à l'extraction des règles d'association dans l'exploration des données mesurées/simulées et des erreurs.
- la troisième étape est la phase de post-traitement qui consiste à réduire le nombre de règles générées pour faciliter la visualisation et l'interprétation des règles. Les informations extraites sont relatives à la mise en évidence des faiblesses du modèle numérique de prévision ; compréhension des paramétrisations liées aux erreurs importantes et l'identification des paramètres critiques.

3.2.1 Imputation des valeurs manquantes

La Base de données simulées qui est constituée des variables météorologiques simulées fournie par le modèle est complète. Par contre, la Base de données mesurées contient des valeurs horaires manquantes. Le graphique suivant (*Figure 3-2*) illustre le taux de pourcentage des valeurs manquantes dans chaque variable mesurée :

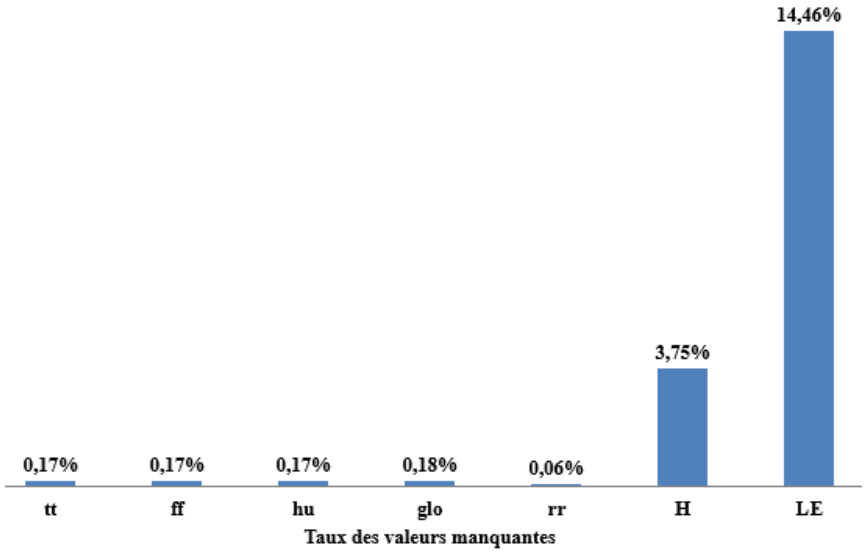


Figure 3-2 : Taux de pourcentage des valeurs manquantes sur la base de données mesurées durant l'année 2016

Un nombre important de valeurs horaires manquantes ont été repérées dans la base de données mesurées. Les valeurs manquantes sont plus fréquentes dans les mesures horaires de la chaleur sensible et la chaleur latente. Les flux convectifs sont difficiles à mesurer et il est normal que ce soient les paramètres pour lesquels il manque le plus de données. Nous avons deux types de valeurs manquantes (monotone et arbitraire) dans notre base de données mesurées. Le schéma suivant (*Figure 3-3*) illustre ces deux types de valeurs manquantes.

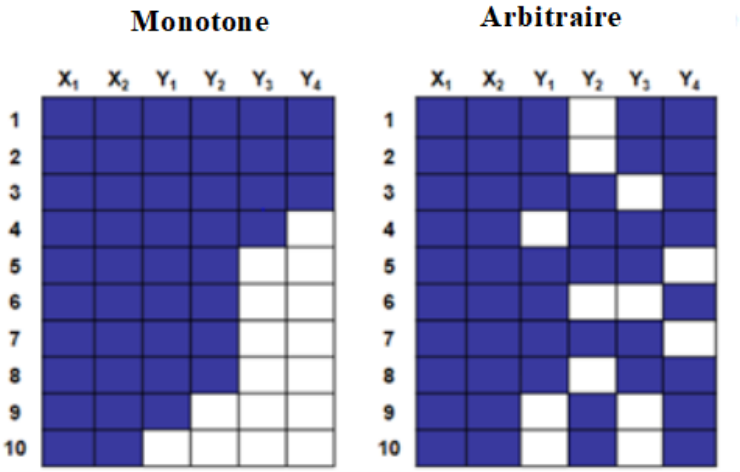


Figure 3-3 : Exemple d'aperçu des valeurs manquantes monotones et arbitraires existantes dans la base de données mesurées durant l'année 2016

Où *X* un vecteur de variables complètes et *Y* un vecteur de variables incomplètes ; les cases blanches représentent les valeurs manquantes. Nous perdons plus d'informations en supprimant des séries d'observations infectées par des valeurs manquantes. Par conséquent, il est nécessaire de remplacer ces valeurs manquantes en utilisant des méthodes d'imputation. Dans ce cas d'étude, deux méthodes d'imputation ont été adoptées : la méthode des *k* plus proches voisins (*k*-NN), et la méthode de la moyenne mobile.

3.2.1.1 Méthode des *k*-plus proches voisins (*k*-NN)

La méthode d'imputation des plus proches voisins (*k*-NN) modélise et prévoit l'approximation des données manquantes, elle est beaucoup plus utilisée dans la littérature comme (Y. Zhang et al., 2019), (Bugata & Drotar, 2019), (Colak et al., 2013). Cette méthode consiste à exécuter l'algorithme suivant :

Tableau 8 : Algorithme des k plus proches voisins (k-NN)

1. Choix d'un entier k : $1 \leq k \leq n$;
2. Calculer les distances $d(Y_{i*}, Y_i)$, $\quad i = 1, \dots, n$; Y_{i*} est la i^{ème} observation contenant la valeur manquante y_{i*} ;
3. Retenir les k observations $y_{i1}, \dots, y_{ik}$ pour lesquelles ces distances sont les plus petites ;
4. Affecter aux valeurs manquantes la moyenne des valeurs des k voisins :

$$y_{i*} = \frac{1}{k}(y_{i1} + \cdots + y_{ik})$$

Le *k*-NN donne un résultat satisfaisant dans le cas des valeurs manquantes arbitraires. Mais pour le cas des valeurs manquantes monotones, le *k*-NN ne donne pas un résultat

satisfaisant, voir (*Figure 3-5*). De ce fait, la méthode moyenne mobile est utilisée pour remplacer les valeurs manquantes monotones.

3.2.1.2 Méthode de la moyenne mobile (MM)

Par définition, une moyenne mobile permet de « lisser » une série de valeurs exprimées en fonction du temps (série chronologique). Elle permet d'éliminer les fluctuations les moins significatives (Chang, 2017). On calcule des moyennes mobiles d'ordre 1, d'ordre 2, d'ordre 3, etc. L'ordre est le nombre de périodes (années, trimestres, mois, ...) sur lesquelles la moyenne mobile est calculée. Elle est aussi beaucoup plus utilisée dans la littérature pour le traitement des données imparfaites comme (Chang, 2017), (N. Zhang et al., 2020).

Les variations de la moyenne mobile comprennent : des formes simples, centrées et pondérées. Nous utilisons la méthode moyenne mobile centrée pour remplacer les valeurs manquantes en intégrant les valeurs du voisinage sans changement brusque, car il s'agit des observations aux mêmes heures. Pour une série de valeurs horaires d'une variable météorologique $X = (x_1,\ x_2, \dots, x_n)$, la moyenne mobile centrée est calculée par la formule suivante :

$$MM_{p,t} = \frac{1}{p}\left[\frac{x_{t-k}}{2} + \sum_{i=-k+1}^{i=k-1} x_{t+i} + \frac{x_{t+k}}{2}\right] \qquad (3.2.1)$$

Où la période $p = 2k$ $(p = 2k + 1)$; t est l'indice horaire avec $2 \leq t \leq n - 1$;

Dans ce cas d'étude, on considère une variable météorologique X contenant n observations horaires avec des valeurs manquantes monotones. L'imputation nécessite une transformation au préalable pour regrouper les observations par similarité temporelle. Ainsi, l'imputation par la moyenne mobile procède en deux étapes :

Etape 1 : on regroupe les *n* observations par similarité temporelle (*k* groupes journaliers similaires et *p* groupes horaires similaires). Ce regroupement consiste à faire une transformation des *n* valeurs horaires sous la forme d'une matrice dont les lignes représentent *k* groupes journaliers et les colonnes représentent *p* groupes horaires.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i-1,1} & x_{i-1,2} & \cdots & x_{i-1,j} & \cdots & x_{i-1,p} \\ x_{i,1} & x_{i,2} & \cdots & x_{i^*,j^*} & \cdots & x_{i,p} \\ x_{i+1,1} & x_{i+1,2} & \cdots & x_{i+1,j} & \cdots & x_{i+1,p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{k,1} & x_{k,2} & \cdots & x_{k,j} & \cdots & x_{k,p} \end{bmatrix} \qquad (3.2.2)$$

Où le couple (i, j) représente (jour, heure) ; $X_{i^*j^*}$ est une valeur manquante dans la $i^{ème}$ ligne et $j^{ème}$ colonne ; $p = 24$ par défaut ;

$k = 31$, pour les mois : janvier, mars, mai, juillet, aout, octobre et décembre ;

$k = 30$, pour les mois : avril, juin, septembre et novembre ;

$k = 28/29$, pour le mois de février.

Etape 2 : on applique la moyenne mobile centrée aux groupes horaires contenant des valeurs manquantes pour l'imputation. La matrice *X* est choisie de façon à permettre que X_{i*j*} ne figure pas sur la première et dernière ligne (ie : $2 \leq i \leq k-1$). Ainsi, la valeur manquante (X_{i*j*}) à l'heure j sera donc calculée de la manière suivante :

$$\begin{cases} \mathrm{X_{i*j*}} \leftarrow 0\ (initialisation) \\ \mathrm{X_{i*j*}} \leftarrow \frac{1}{p}\left[\frac{x_{t-k,j}}{2} + \sum_{i=-k+1}^{i=k-1} x_{t+i,j} + \frac{x_{t+k,j}}{2}\right] \end{cases} \qquad (3.2.3)$$

Dans le cas d'imputation des valeurs manquantes monotones, la méthode moyenne mobile (MM) est meilleure par rapport à la méthode *k*-NN qui sera justifiée dans la partie expérimentale. Compte tenu de la complexité des phénomènes spatio-temporels, la robustesse du processus de remplacement des valeurs manquantes peut être augmentée en utilisant des modèles d'ajustement. Pour cela, un ajustement a été réalisé avec les modèles autorégressifs et à moyenne mobile pour réduire les fluctuations irrégulières (intégrant éventuellement des accidents de mesures), (Fanoodi et al., 2019), (Chang, 2017). Une étude de cas est présentée dans la partie expérimentale (section 4.4.1)

3.2.1.3 Méthode autorégressive (ARIMA)

Les données météorologiques sont des séries chronologiques. Un modèle autorégressif à moyenne mobile intégrée (ARIMA) est une généralisation d'un modèle autorégressif à moyenne mobile (ARMA). Le modèle autorégressif et moyenne-mobile (ARMA) est un outil pour mieux comprendre les données, soit pour prédire les points futurs de la série (prévisions). Dans certains cas où les données montrent des preuves de non-stationnarité (fluctuations irrégulières dans les données), il peut être appliqué une ou plusieurs fois pour éliminer la non-stationnarité (Chang, 2017).

Étant donné une série temporelle X_t, un modèle autorégressif et moyenne-mobile d'ordres (p,q), noté ARMA(p,q), est un processus temporel discret ($X_t, t \in \mathbb{N}$) vérifiant :

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \qquad (3.2.4)$$

où les φ_i et θ_i sont les paramètres du modèle et les ε_i sont les termes d'erreur. Un modèle autorégressif AR(p) est un ARMA(p,0) et un modèle moyenne mobile MA(q) est un ARMA(0,q).

Dans le cadre de la gestion des valeurs manquantes, la méthode d'imputation *k*-NN a des inconvénients face aux valeurs manquantes monotones. Cela impacte la qualité de la discrétisation tout en affectant les résultats globaux (règles d'association générées). Pour

minimiser les inconvénients de *k*-NN face aux valeurs manquantes monotones, la méthode moyenne mobile est utilisée pour remplacer les valeurs manquantes monotones. En plus, la méthode ARMA a été utilisée pour ajuster et éliminer la non-stationnarité des données de la série temporelle.

3.2.2 Transformation de données par discrétisation

Dans ce cas d'étude, la discrétisation est l'étape la plus importante dans l'extraction des règles d'association. En effet, la pertinence des informations extraites dépend du choix des bornes des intervalles. Dans notre base de données, chaque variable est constituée d'une gamme considérable de valeurs numériques. La discrétisation est une technique de réduction de données (Garcia, Ramirez-Gallego, et al., 2016). Elle permet de faire une transformation des valeurs numériques (attributs numériques) en des intervalles de données (attributs nominaux) pour que les données soient compatibles aux formats de données des algorithmes. Cette transformation permet aussi de réduire l'espace de recherche et d'augmenter l'efficacité des algorithmes (Khader et al., 2016).

Ainsi, il faut définir le nombre d'intervalles attendus et l'amplitude associée à chaque intervalle. Pour que la distribution en fréquence ait un sens, il faut que chaque intervalle comprenne un nombre suffisant de valeurs (nombre d'unités statistiques). Pour cela, nous procédons en deux étapes :

Etape 1. Nombre d'intervalles attendus.

Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille n, notamment les règles de STURGE et de YULE :

$$\text{STURGE:} \quad Nombre\ de\ classes = 1 + (3{,}3 \log n)$$

$$\text{YULE:} \quad Nombre\ de\ classes = 2{,}5\sqrt[4]{n}$$

Lorsque n est petit, ces deux règles donnent approximativement le même nombre d'intervalles. Mais quand n est très grand, le nombre d'intervalles de YULE est strictement supérieur au nombre d'intervalles de STURGE.

En plus, compte tenu des contraintes du domaine, nous utiliserons dans la suite la règle de STURGE pour déterminer le nombre d'intervalles attendu pour chaque variable.

Etape 2. Bornes des intervalles.

Les bornes des intervalles sont déterminées à partir des propriétés sur les suites arithmétiques pour chaque type de distribution, en vue de faire une répartition pertinente des échantillons entre les intervalles. Les variables considérées suivent différents types de distributions statistiques, notamment une distribution uniforme, une distribution normale, une distribution asymétrique à gauche (ou à droite).

3.2.2.1 Distribution normale

On parle d'une distribution normale, lorsqu'une variable météorologique X suit la loi normale (forme en cloche), par exemple la distribution de la température (*Figure 2-4*). En règle générale, on choisit des intervalles de même longueur (amplitude).

$$Amplitude = \frac{\max(X) - \min(X)}{Nombre\ de\ classes}$$

3.2.2.2 Distribution asymétrique à gauche

Lorsqu'une distribution s'étire vers les fortes valeurs d'une variable météorologique X, c'est le cas d'une asymétrie à gauche, par exemple la distribution du rayonnement global (*Figure 2-16*). Dans ce cas, pour avoir une répartition pertinente, les bornes des intervalles sont choisies de sorte que les longueurs des intervalles suivent une progression arithmétique croissante de raison A. C'est-à-dire : $\max(X) - \min(X) = A + 2A + 3A + \cdots + NA$, où N = nombre d'intervalles.

3.2.2.3 Distribution asymétrique à droite

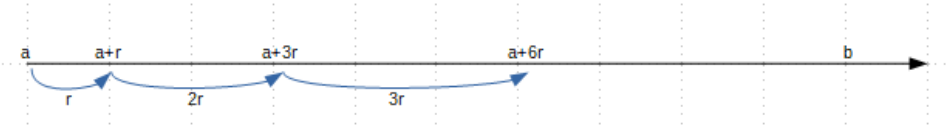
Lorsqu'une distribution est étirée vers les faibles valeurs d'une variable météorologique X, c'est le cas d'une asymétrie à droite, par exemple la distribution de l'humidité (*Figure 2-10*). Dans ce cas, les longueurs des intervalles suivent une progression arithmétique décroissante de raison $-A$. C'est-à-dire : $\max(X) - \min(X) = NA + (N-1)A + \cdots + 3A + 2A + A$, où N = nombre d'intervalles.

Proposition de quelques algorithmes de discrétisations et de codification :

On note X une variable météorologique contenant n observations : $X = (x(1), x(2), \ldots, x(n))$.

Procédé : Il faut choisir d'abord un intervalle $[a, b]$ de telle sorte que les $x(i), i = 1 \text{à} n$, appartiennent à $[a, b]$; ensuite définir le nombre d'intervalles attendu N ; ainsi, les algorithmes ci-dessous sont exécutés selon les cas suivants :

- Série dissymétrique sur un intervalle $[a, b]$: cas où les amplitudes des intervalles sont inégales et suivent une progression arithmétique :



$$b = a + \sum_{j=1}^{N} jr$$

Où N = nombre d'intervalles attendus et r = raison de la progression arithmétique.

Tableau 9 : Algorithme de discrétisation d'une série dissymétrique

Paramètres d'entrée : a ; r ; N ; p ; $b = a + r$
Pour $j = 2$ à N
Pour $i = 1$ à n
Si $x(i) \in [a, b]$, faire
$x(i) = p$;

Sinon, faire

$a = b$;

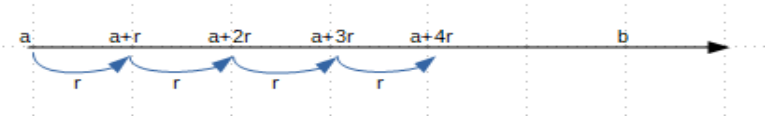
$b = b + jr$;

$x(i) = x(i) + 1$;

Fin

Fin

- Série normale/uniforme sur un intervalle $[a, b]$: cas où les amplitudes des intervalles sont égales :



$$b = a + \sum_{j=1}^{N} r$$

Où N = nombre d'intervalles attendus et r = amplitude des intervalles.

Tableau 10 : Algorithme de discrétisation d'une série normale

Paramètres d'entrée : a ; r ; N ; p ; $b = a + r$

Pour $j = 2$ à N

Pour $i = 1$ à n

Si $x(i) \in [a, b[$ alors

$x(i) = p$

Fin si

Fin pour

$a = b$

$b = b + r$

$P = p + 1$

Fin

Fin

Ainsi, nous obtenons une base de données préparées qui est adaptée aux algorithmes de la fouille de données.

3.2.3 Règles d'association

La notion de règle d'association a été présentée sous le nom de Generalized Unary Hypotheses Automaton (GUHA) en 1966 par Petr Hájek et ses collègues (Hajek et al., 1966). En 1993, le concept de règle d'association a été popularisé par un article de Rakesh Agrawal (Agrawal et al., 1993). Le lecteur trouvera une description détaillée sur la notion de règle d'association dans (Djenouri & Comuzzi, 2017).

Parmi les techniques d'extraction des informations et des connaissances de données, la recherche de règles d'association semble être celle qui permet une représentation factuelle des corrélations. Elle détecte le motif caché, en fouillant le contenu d'une base de données de taille énorme (Narvekar & Syed, 2015).

La génération des règles d'association peut être décomposée en deux étapes : détermination des itemsets fréquents, génération des règles d'association (Djenouri & Comuzzi, 2017). Ce processus d'extraction des règles d'association est détaillé dans la *section 1.2.2*.

Règles rares : une règle d'association rare est une règle d'association générée à partir d'un l'ensemble de données et qui satisfait aux critères exigeant une valeur de support inférieure au seuil de support et une valeur de confiance supérieure au seuil de confiance (Wulandari et al., 2019).

LIFT : le Lift est une mesure d'évaluation de la qualité des règles générées. Cette mesure d'évaluation est utilisée pour mesurer le degré de dépendance entre l'antécédent et le conséquent d'une règle d'association (Wulandari et al., 2019), (McNicholas et al., 2008). En d'autres termes, cette mesure calcule une valeur déterminant la nature de la relation d'influence entre l'antécédent et le conséquent de la règle. Le Lift est calculé par la formule suivante :

$$Lift = \frac{\sup(X \rightarrow Y)}{\sup(X)\sup(Y)} = \frac{\text{conf}(X \rightarrow Y)}{\sup(Y)}$$

Le *Lift* est > 1 indique une dépendance positive, ce qui implique que l'occurrence de l'antécédent a un effet positif sur l'occurrence du conséquent. En d'autres termes, l'antécédent et le conséquent dépendent l'un de l'autre. Cela signifie que l'antécédent et le conséquent apparaissent plus souvent ensemble que prévu.
Le *Lift* est < 1 indique une dépendance négative et signifie que l'occurrence de l'antécédent a un effet négatif sur l'occurrence du conséquent.
Une règle n'a aucun intérêt si son *Lift* est égale à 1, ce qui indique également que l'antécédent et le conséquent sont indépendants l'un de l'autre.
A noter qu'une dépendance positive entre l'antécédent et le conséquent rend la règle potentiellement utile pour prédire le conséquent.

3.2.4 Technique de réduction du nombre de règles générées

Avec un minsup et un minconf choisis, ce cas d'étude consiste à extraire les Règles d'Associations Rares ou Rare Association Rules (RAR) comme indiqué par Wulandari et ses collègues (Wulandari et al., 2019). Ceci génère un nombre important de règles d'association. Pour faciliter la visualisation, l'analyse et l'interprétation des règles en

post-traitement, il est nécessaire de trouver des moyens de réduction du nombre de règles générées. Cependant, pour réduire le nombre de règles générées, nous proposons quelques propriétés sur la théorie des ensembles en mathématiques.

Propriétés : Considérons les règles R1, R2 et R3 définies par :

$$R1 : \ A \rightarrow Y$$

$$R2 : \ B \rightarrow Y$$

$$R3 : \ X \rightarrow Y$$

Où A, B, X sont des antécédents et Y est la conséquence des règles.

P1 : si $A \cap B = A \iff A \subset B$, alors R2 est considérée comme une extension de R1, on conserve la règle R1 ;

P2 : si $A \cap B = \phi \ et \ A \cup B \subset X \implies$ A et B sont disjoints, alors R3 est une jonction de R1 et R2, on conserve les règles R1 et R2 ;

P3 : si $A \cap B = \phi$, alors A et B sont disjoints, on conserve les règles R1 et R2.

Le schéma ci-après (*Figure 3-4*) illustre une visualisation graphique pour mieux comprendre l'application de ces propriétés.

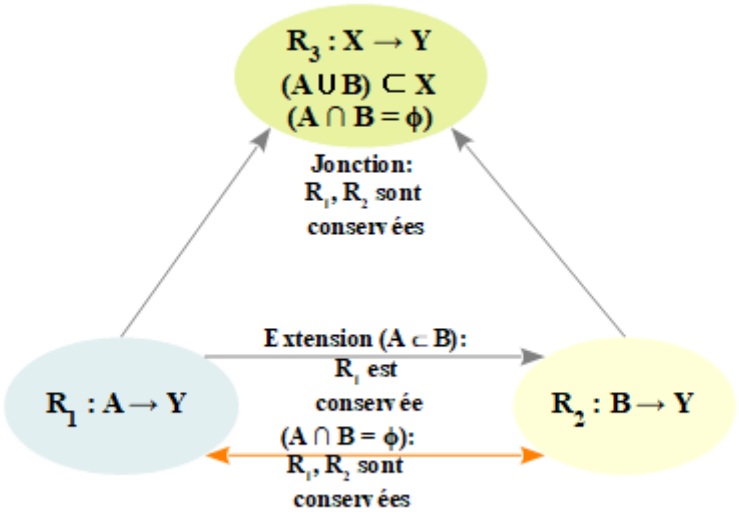


Figure 3-4 : Schéma de base pour la réduction des règles

La figure illustre le schéma de base pour la réduction du nombre de règles générées, mais une variété de configurations est possible. Un exemple de cas d'étude est présenté dans la partie expérimentale (*section 3.3*).

3.3 Application sur quelques variables mesurées / simulées

Dans cette section, nous présentons une étude de cas avec l'application de la technique de fouille de données pour diagnostiquer le modèle numérique de prévision. Ceci pour comprendre et identifier les paramétrisations qui influent sur les erreurs extrêmes dans la simulation des variables météorologiques. L'ensemble des variables utilisées dans ce diagnostic sont celles présentée dans la *section 1.5*. Ce cas d'étude est un processus de trois étapes (*Figure 3-1*) : un prétraitement des valeurs aberrantes ; suivi d'un traitement de données avec la génération des règles d'association ; un post-traitement pour la visualisation, la réduction et l'interprétation des règles générées.

3.3.1 Étape de prétraitement des données

Cette partie de prétraitement présente d'abord un cas d'étude sur la comparaison des méthodes d'imputation, suivi d'un exemple de lissage, ensuite un tableau de discrétisation des variables et des erreurs considérées.

3.3.1.1 Comparaison des méthodes d'imputation - Lissage de données

Les données météorologiques mesurées sont imparfaites en raison de la présence des valeurs manquantes et des valeurs hors portées (bruits). Alors que ce cas d'étude exige une base de données sans valeurs manquantes. De plus, les méthodes d'ajustement de données ne prennent pas en compte les valeurs manquantes. Par conséquent, il est nécessaire dans un premier temps, d'utiliser des méthodes d'imputation pour remplacer ces valeurs manquantes. Ensuite, une méthode de lissage est utilisée pour ajuster les pics de mesures. Le graphique suivant (*Figure 3-5*) illustre la comparaison de deux méthodes d'imputation, notamment le *k* plus proches voisins (*k*-NN) et la moyenne mobile (MM). Cette comparaison portera sur les données manquantes de type monotone de la température du mois de janvier.

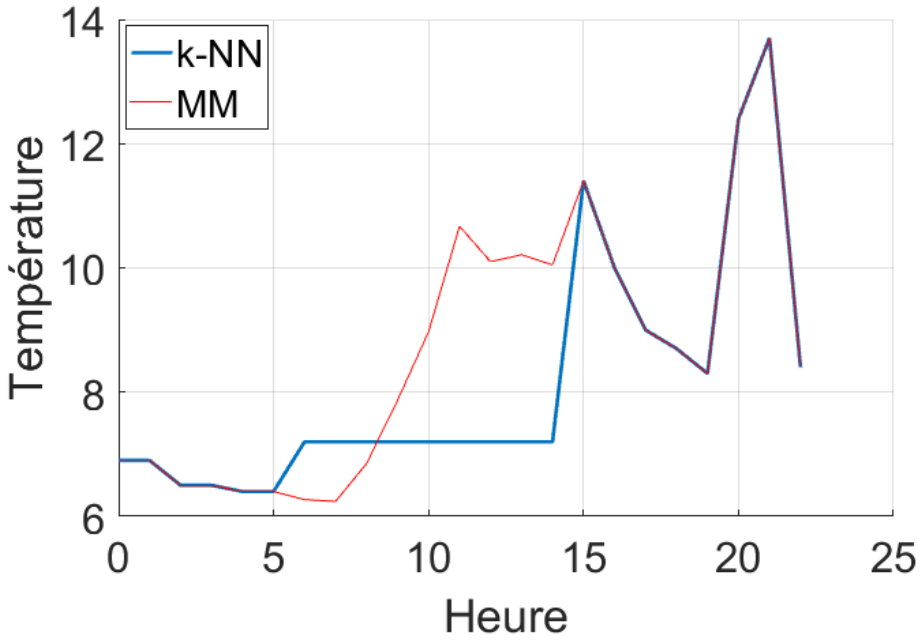


Figure 3-5 : Comparaison des méthodes k-NN et MM

Les valeurs manquantes remplacées sont les valeurs horaires de 6h à 14h. Le trait bleu représente les valeurs remplacées par la méthode des *k* plus proches voisins (*k*-NN). Le trait rouge représente les valeurs remplacées par la méthode moyenne mobile (MM). Dans ce cas d'étude sur [6h, 14h], nous remarquons que le *k*-NN fournit des valeurs de température constante. Par contre, le MM donne des valeurs de température variable. Compte tenu de la complexité des phénomènes spatio-temporelles, la robustesse du processus de remplacement des valeurs manquantes peut être augmentée en utilisant des modèles d'ajustement. Pour cela, un ajustement a été réalisé avec les modèles à moyenne mobile intégrée autorégressive ou Autoregressive Integrated Moving Average (ARIMA) pour réduire les fluctuations irrégulières (intégrant éventuellement des accidents de mesures), (Fanoodi et al., 2019), (Chang, 2017). Ainsi, la figure suivante illustre un exemple de cas d'ajustement par ARIMA(p,0,q) sur la même période des valeurs de température (*Figure 3-6*).

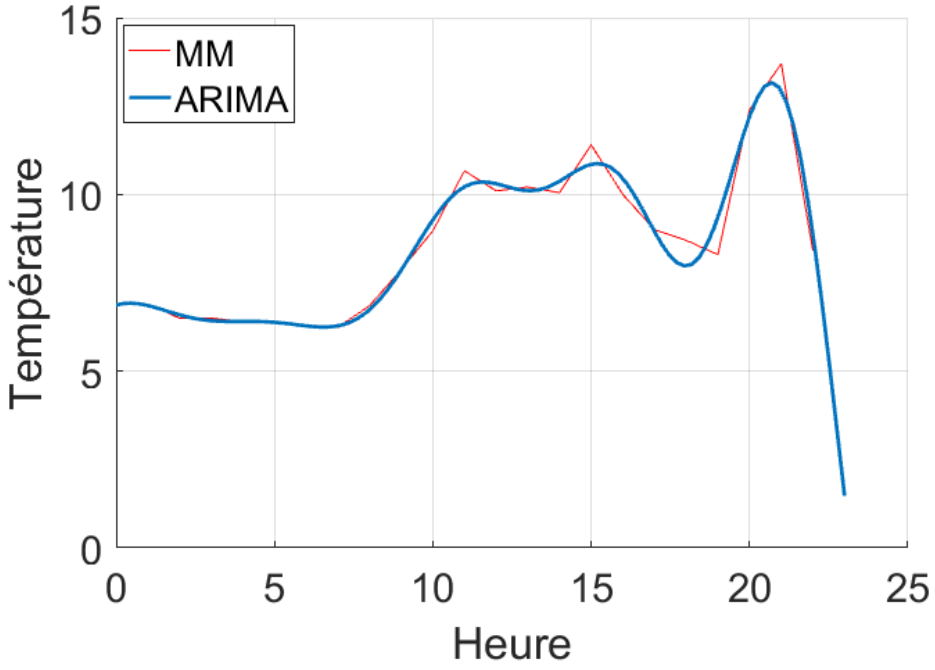


Figure 3-6 : Lissage des valeurs par ARIMA(1,0,1)

La courbe rouge (*Figure 3-6*) est la même que celle figurant sur (*Figure 3-5*). Cette courbe rouge représente les valeurs mesurées et les valeurs manquantes remplacées de la température. Ces valeurs contiennent quelques pics de mesures. Ainsi, ARIMA est utilisé pour modéliser ces valeurs prédictives dans le but d'ajuster les pics de mesures. Ces valeurs ajustées et lissées sont représentées sur la (*Figure 3-6*) sous forme de courbe continue en bleue. Cette courbe est une estimation des fluctuations de valeurs mesurées.

3.3.1.2 Discrétisation des variables et des erreurs de simulation

En utilisant la technique de discrétisation proposée dans la *section 3.2.2* et les distributions statistiques des variables dans la *section 2.2*, le tableau suivant présente la discrétisation de chaque variable considérée et aussi des erreurs de simulation (*Tableau 11*).

Tableau 11 : Discrétisation des variables et des erreurs de simulation

Discrétisation des variables						
Catégorisation de la température						
Température tt	T_1	T_2	…	T_{n-1}	T_n	$n > 0$
(amplitudes égales)	[-6;-3[	[-3;0[		[30;33[	[33;36]	
Catégorisation de la vitesse du vent						
Vitesse du vent ff	F_1	F_2	…	F_{p-1}	F_p	$p > 0$
(amplitudes inégales)						

	[0;0.5[	[0.5;1[		[9.6;11.2[	[11.2;13]	
Catégorisation de l'humidité						
Humidité hu	U_1	U_2	…	U_{q-1}	U_q	$q > 0$
(amplitudes inégales)						
	[9; 22[	[22;34[		[97;99[	[99;100]	
Catégorisation de la pluie						
Pluie rrl	R_1	R_2	…	R_{j-1}	R_j	$j > 0$
(amplitudes inégales)						
	[0;0]	]0;0.2[		[13.2;15.6[	[15.6;18.2]	
Catégorisation de la direction du vent						
Direction du vent dd	D_1	D_2	…	D_{k-1}	D_k	k>0
(amplitudes inégales)	$]330; 360] \cup [0; 30]$	]30;60]		$] 240; 300]$	]300;330]	
Discrétisation des erreurs de simulation						
Catégorisation des erreurs de température						
Δtt	ΔT_1	ΔT_2	…	$\Delta T_{n'-1}$	$\Delta T_{n'}$	$n' > 0$
(amplitudes inégales)	[-10;-2[	[-2;-1[		[1;2[	[2;8]	
Catégorisation des erreurs de vitesse du vent						
Δff	ΔF_1	ΔF_2	…	$\Delta F_{p'-1}$	$\Delta F_{p'}$	$p' > 0$
(amplitudes inégales)						
	[-7;-2[	[-2;-1[		[1;2[	[2;8]	
Catégorisation des erreurs d'humidité						
Δhu	ΔU_1	ΔU_2	…	$\Delta U_{q'-1}$	$\Delta U_{q'}$	$q' > 0$
(amplitudes inégales)						
	[-75;-20[	[-20;-15[		[15;20[	[20;60]	
Catégorisation des erreurs du rayonnement global						
Δglo	ΔR_1	ΔR_2	…	$\Delta R_{j'-1}$	$\Delta R_{j'}$	$j' > 0$
(amplitudes inégales)						

	[-708;-100[	[-100;-20[		[20;100[	[100;738]	
Catégorisation des erreurs de chaleur sensible						
ΔH	ΔH_1	ΔH_2	…	$\Delta H_{k'-1}$	$\Delta H_{k'}$	$k' > 0$
(amplitudes inégales)	[-315;-45[	[-45;-30[		[30;60[	[60;340]	
Catégorisation des erreurs de chaleur latente						
ΔLE	ΔLE_1	ΔLE_2	…	$\Delta LE_{k'-1}$	$\Delta LE_{k'}$	$l' > 0$
(amplitudes inégales)	[-590;-60[	[-60;-30[		[60;100[	[100;475]	
Où les entiers n, p, q, j, k, n', p', q', j', k' sont des nombres d'intervalles.						

Remarque : En météorologie, on peut diviser la direction du vent en huit secteurs de direction d'où vient le vent (Naughton et al., 2018) : le vent venant du Nord $(]330°, 360°] \cup [0°, 30°])$, le vent venant du Nord-Est $(]30°, 60°])$, le vent venant de l'Est $(]60°, 120°])$, le vent venant du Sud-Est $(]120°, 150°])$, le vent venant du Sud $(]150°, 210°])$, le vent venant du Sud-Ouest $(]210°, 240°])$, le vent venant de l'Ouest $(]240°, 300°])$, le vent venant du Nord-Ouest $(]300°, 330°])$.

Par la suite, nous utiliserons différentes abréviations et notations dans le tableau suivant :

Tableau 12 : Abréviations et notations

Abréviations
Δtt : Erreur sur la température
Δff : Erreur sur la vitesse du vent
Δhu : Erreur sur l'humidité relative
Δglo : Erreur sur le rayonnement global
ΔH : Erreur sur la chaleur sensible
ΔLE : Erreur sur la chaleur latente
dd : Direction du vent
N : Vent venant du Nord (]330°;360°]∪[0°;30°])
NE : Vent venant du Nord-Est (]30°;60°])
E : Vent venant de l'Est (]60°;120°])
SE : Vent venant du Sud-Est (]120°;150°])
S : Vent venant du Sud (]150°;210°])
SO : Vent venant du Sud-Ouest (]210°;240°])
O : Vent venant de l'Ouest (]240°;300°])
NO : Vent venant du Nord-Ouest (]300°;330°])

3.3.2 Étape de traitement (génération des règles d'association)

Dans cette section, nous allons générer d'abord les règles d'association à partir des données simulées et mesurées, dans le but de comparer le modèle numérique de prévision et les observations pour mettre en évidence les faiblesses du modèle. Ensuite, nous générons les règles d'association en explorant les différences (ΔV= erreurs entre données simulées et données mesurées), dans le but d'établir le croisement entre les différences (ΔV) des variables considérées pour comprendre et détecter les paramétrisations liées à des erreurs importantes dans la simulation des variables telles que, par exemple, la chaleur sensible et la chaleur latente.

3.3.2.1 Génération des règles d'association à partir des données observées et simulées pour la mise en évidence des différences entre modèle et observations

L'objectif ici est de comparer les données simulées par le modèle et les données mesurées. Pour un min*sup* et un min*conf* choisis, l'exploration des deux bases de données mesurées et simulées avec l'algorithme FP-Growth donne des résultats différents figurant respectivement dans les tableaux (*Tableau 13*) et (*Tableau 14*). Les règles générées sont des cas où la pluie apparaît dans le conséquent des règles générées. Avec ces critères, 12 règles ont été générées à partir des données mesurées (*Tableau 13*) et 62 règles ont été générées à partir des données simulées (*Tableau 14*).

Tableau 13 : Règles générées dans la base de données mesurées

	Antécédent : tt, dd, ff, hu, glo						**Conséquent**	**SUP**	**CONF**
R1	tt[3à6[	O	ff[1à2[	hu[97à99[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R2	tt[3à6[	O	ff[1à2[	hu[97à99[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R3	tt[3à6[	O	ff[4à4,2[	hu[90à94[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R4	tt[3à6[	O	ff[4,2à4,6[	hu[94à97[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R5	tt[6à9[	O	ff[1à2[	hu[94à97[	glo[0à0]	==>	rr[0,6à1,2[	2	100 %
R6	tt[6à9[	O	ff[2à3[	hu[99à100[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R7	tt[6à9[	O	ff[3à4[	hu[97à99[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R8	tt[6à9[	O	ff[7à8,2[	hu[97à99[	glo[0à0]	==>	rr[0,6à1,2[	2	100 %
R9	tt[9à12[	O	ff[2à3[	hu[97à99[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R10	tt[9à12[	O	ff[6à7[	hu[97à99[	glo[24à40[	==>	rr[0,2à0,6[	2	100 %
R11	tt[12à15[	N	ff[1à2[	hu[97à99[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R12	tt[15à18[	NO	ff[2à3[	hu[97à99[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %

Tableau 14 : Règles générées dans la base de données simulées

	Antécédent : tt, dd, ff, hu, glo						rr	SUP	CONF
R1	tt[0à3[	O	ff[2à3[	hu[94à97[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R2	tt[0à3[	O	ff[2à3[	hu[99à100[	glo[0à0]	==>	rr[1,2à2[	2	100 %
R3	tt[0à3[	O	ff[3à4[	hu[85à90[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R4	tt[3à6[	E	ff[5,2à6[	hu[85à90[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R5	tt[3à6[	O	ff[1à2[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	3	100 %
R6	tt[3à6[	O	ff[2à3[	hu[90à94[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R7	tt[3à6[	O	ff[2à3[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R8	tt[3à6[	O	ff[3à4[	hu[85à90[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R9	tt[3à6[	O	ff[3à4[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	3	100 %
R10	tt[3à6[	O	ff[3à4[	hu[99à100[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R11	tt[3à6[	O	ff[5,2à6[	hu[94à97[	glo[0à0]	==>	rr[1,2à2[	2	100 %
R12	tt[3à6[	O	ff[7à8,2[	hu[85à90[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R13	tt[3à6[	O	ff[9,6à11,2[	hu[79à85[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R14	tt[6à9[	E	ff[3à4[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R15	tt[6à9[	S	ff[1à2[	hu[72à79[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R16	tt[6à9[	S	ff[1à2[	hu[94à97[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R17	tt[6à9[	S	ff[1à2[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	4	100 %
R18	tt[6à9[	S	ff[2à3[	hu[90à94[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R19	tt[6à9[	S	ff[2à3[	hu[94à97[	glo]0à4[	==>	rr]0à0,2[	2	100 %
R20	tt[6à9[	S	ff[2à3[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R21	tt[6à9[	S	ff[3à4[	hu[55à64[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R22	tt[6à9[	O	ff[1à2[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	5	100 %
R23	tt[6à9[	O	ff[2à3[	hu[90à94[	glo[60à84[	==>	rr]0à0,2[	2	100 %
R24	tt[6à9[	O	ff[2à3[	hu[94à97[	glo[0à0]	==>	rr[0,2à0,6[	2	100 %
R25	tt[6à9[	O	ff[2à3[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R26	tt[6à9[	O	ff[3à4[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R27	tt[6à9[	O	ff[3à4[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	3	100 %
R28	tt[6à9[	O	ff[3à4[	hu[99à100[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R29	tt[6à9[	O	ff[4,6à5,2[	hu[90à94[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R30	tt[6à9[	O	ff[5,2à6[	hu[85à90[	glo[180à220[	==>	rr]0à0,2[	2	100 %
R31	tt[6à9[	O	ff[5,2à6[	hu[94à97[	glo[0à0]	==>	rr[1,2à2[	2	100 %
R32	tt[6à9[	O	ff[6à7[	hu[90à94[	glo[0à0]	==>	rr[2à3[	2	100 %
R33	tt[9à12[	S	ff[2à3[	hu[64à72[	glo[0à0]	==>	rr]0à0,2[	3	100 %
R34	tt[9à12[	S	ff[2à3[	hu[79à85[	glo[0à0]	==>	rr]0à0,2[	2	100 %

	Antécédent : tt, dd, ff, hu, glo						rr	SUP	CONF
R35	tt[9à12[	O	ff[3à4[	hu[90à94[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R36	tt[9à12[	O	ff[5,2à6[	hu[72à79[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R37	tt[9à12[	O	ff[5,2à6[	hu[97à99[	glo[24à40[	==>	rr[0,2à0,6[	2	100 %
R38	tt[9à12[	O	ff[7à8,2[	hu[85à90[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R39	tt[9à12[	O	ff[7à8,2[	hu[94à97[	glo]0à4[	==>	rr[0,2à0,6[	2	100 %
R40	tt[9à12[	O	ff[7à8,2[	hu[94à97[	glo[4à12[	==>	rr[0,2à0,6[	2	100 %
R41	tt[12à15[	N	ff[2à3[	hu[64à72[	glo>364	==>	rr]0à0,2[	2	100 %
R42	tt[12à15[	N	ff[2à3[	hu[94à97[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R43	tt[12à15[	N	ff[3à4[	hu[64à72[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R44	tt[12à15[	S	ff[1à2[	hu[90à94[	glo[0à0]	==>	rr]0à0,2[	3	100 %
R45	tt[12à15[	SO	ff[1à2[	hu[90à94[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R46	tt[12à15[	SO	ff[1à2[	hu[94à97[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R47	tt[12à15[	O	ff[2à3[	hu[85à90[	glo[180à220[	==>	rr]0à0,2[	2	100 %
R48	tt[12à15[	O	ff[3à4[	hu[85à90[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R49	tt[12à15[	O	ff[3à4[	hu[94à97[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R50	tt[12à15[	O	ff[4à4,2[	hu[72à79[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R51	tt[12à15[	O	ff[4à4,2[	hu[94à97[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R52	tt[12à15[	NO	ff[2à3[	hu[97à99[	glo[40à60[	==>	rr]0à0,2[	2	100 %
R53	tt[12à15[	NO	ff[7à8,2[	hu[64à72[	glo>364	==>	rr]0à0,2[	2	100 %
R54	tt[15à18[	S	ff[3à4[	hu[64à72[	glo>364	==>	rr]0à0,2[	2	100 %
R55	tt[15à18[	O	ff[2à3[	hu[79à85[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R56	tt[15à18[	O	ff[2à3[	hu[85à90[	glo[60à84[	==>	rr]0à0,2[	2	100 %
R57	tt[15à18[	O	ff[2à3[	hu[90à94[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R58	tt[15à18[	O	ff[3à4[	hu[55à64[	glo>364	==>	rr]0à0,2[	2	100 %
R59	tt[15à18[	O	ff[3à4[	hu[64à72[	glo>364	==>	rr]0à0,2[	2	100 %
R60	tt[15à18[	O	ff[3à4[	hu[79à85[	glo[0à0]	==>	rr]0à0,2[	2	100 %
R61	tt[15à18[	O	ff[3à4[	hu[94à97[	glo[0à0]	==>	rr]0à0,2[	3	100 %
R62	tt[24à27[	E	ff[3à4[	hu[45à55[	glo>364	==>	rr]0à0,2[	2	100 %

Sur les 12 et 62 règles obtenues après exploration des deux bases de données mesurées et simulées, toutes les variables considérées (température, direction vent, module vent, humidité, pluie, rayonnement global) apparaissent localement en certaines valeurs. Les règles générées ont pour conséquent la pluie sur cinq intervalles (]0à0.2[; [0.2à0.6[; [0.6à1.2[; [1.2à2[et [2à3[) et toutes les autres variables apparaissent en antécédent des règles. La figure ci-dessous illustre la comparaison des fréquences de pluie mesurée et simulée, dans le conséquent des règles générées (*Figure 3-7*).

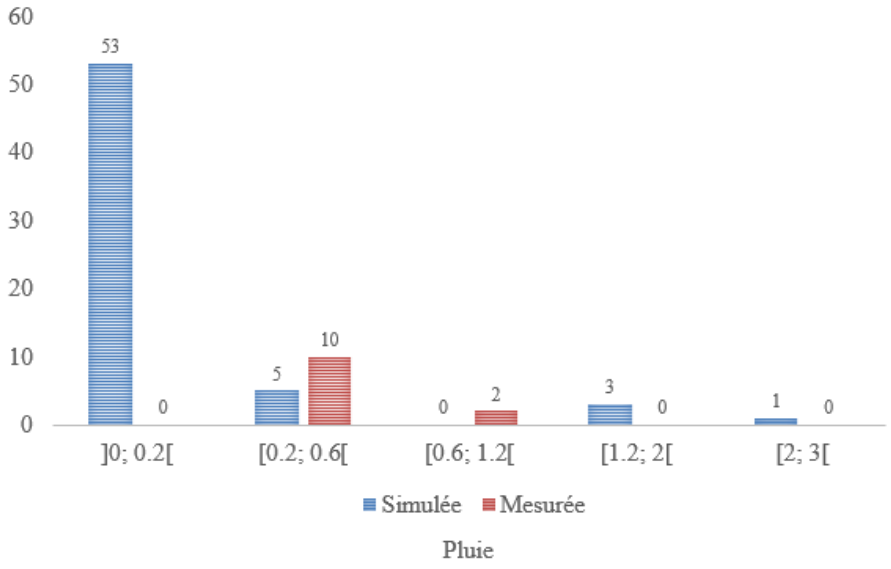


Figure 3-7 : Mise en évidence des différences à partir des fréquences d'intervalles particuliers de pluie sur le conséquent des règles générées à partir de données simulées et observées

Ce graphique (*Figure 3-7*) représente la comparaison des fréquences d'intervalles particuliers de pluie sur le conséquent des règles générées à partir de données simulées et mesurées. Il apparaît une certaine différence entre le nombre de règles générées (12 règles) avec les données mesurées relativement au nombre de règles engendrées (plus de 60 règles) avec les données simulées. Cette différence semble s'expliquer par deux facteurs : (i) premièrement l'intervalle]0;0.2[est beaucoup plus fréquent dans les données simulées que dans les données mesurées ; (ii) secondairement l'on a choisi un même couple de minimum de support et confiance (min*sup* = 2 , min*conf* = 98%) pour l'extraction des règles dans les deux bases de données simulées et mesurées. On observe donc sur (*Figure 3-7*) que l'item]0; 0.2[est fréquent (53 fois) dans le conséquent des règles générées à partir seulement des données simulées. Alors que l'item [0.2; 0.6[est fréquent à la fois dans le conséquent des règles générées à partir des données simulées et observées. Mais la fréquence de l'item [0.2; 0.6[domine plus fortement dans les observations que dans le modèle. Ces constats peuvent être interprétés comme un des facteurs de faiblesse du modèle dans la simulation de la variable pluie. Les informations extraites sont relatives à la compréhension de la simulation des processus de surface, à la mise en évidence des différences entre le modèle numérique de prévision et les observations par apprentissage automatique. Elles sont aussi cruciales pour l'aide à la décision notamment durant les alertes météo en période de fortes pluies pour faciliter, par exemple, une meilleure gestion des crues (Pereira et al., 2018), (Arnaud et al., 2017).

3.3.2.2 Génération des règles d'association à partir des erreurs sur les différentes variations saisonnières en 2016

L'objectif ici est de trouver la relation entre les erreurs de simulation pour différentes variables, éventuellement pour comprendre et identifier les paramètres liés à des erreurs importantes dans la simulation des variables. Pour un minimum de support (min sup) et

un minimum de confiance (min conf) choisis, l'exploration de la base des erreurs avec l'algorithme FP-Growth a généré des règles pour toutes les variations saisonnières. Ces informations extraites sont des règles qui ont pour conséquents : erreurs sur la chaleur sensible (ΔH) et erreurs sur la chaleur latente (ΔLE). Les règles générées pour les saisons printemps, Eté, Automne et Hiver figurent respectivement dans (*Tableau 15*), (*Tableau 16*), (*Tableau 17*) et (*Tableau 18*).

Tableau 15 : Règles générées pour le Printemps (à partir du 20 Mars)

	Antécédent					**Conséquent**	**SUP**	**CONF**	**LIFT**
R1	Δglo[-708à-100[				==>	ΔH[-315à-45[	25	56,82 %	5,92
R2	Δtt[-2à-1[	Δglo[-708à-100[			==>	ΔH[-315à-45[	22	52,38 %	5,46
R3	Δff[0à1[	Δglo[-708à-100[			==>	ΔH[-315à-45[	20	55,56 %	5,79
R4	Δhu[15à20[	Δglo[-708à-100[			==>	ΔH[-315à-45[	18	66,67 %	6,95
R5	Δtt[-10à-2[	Δff[-7à-2[	ΔLE[100à475]		==>	ΔH[-45à-30[	3	100,00 %	19,31
R6	Δtt[-1à0[	Δhu[-10à-5[	ΔLE[-15à0[		==>	ΔH[-30à-15[	6	50,00 %	4,43
R7	Δtt[0à1[	Δglo]0à20[			==>	ΔH[-15à0[	30	50,85 %	2,43
R8	Δglo]0à20[	ΔLE[0à15[			==>	ΔH[-15à0[	26	50,00 %	2,39
R9	Δtt[0à1[	Δff[1à2[	ΔLE[-15à0[		==>	ΔH[-15à0[	13	54,17 %	2,59
R10	Δhu[-5à0[	Δglo]0à20[	ΔLE[0à15[		==>	ΔH[-15à0[	13	68,42 %	3,28
R11	Δtt[0à1[	Δhu[-5à0[	ΔLE[-15à0[		==>	ΔH[-15à0[	12	50,00 %	2,39
R12	Δhu[0à5[	Δgnt[0à0]	ΔLE[0à15[		==>	ΔH[0à15[	27	51,92 %	2,43
R13	Δtt[-1à0[	Δhu[0à5[	ΔLE[0à15[		==>	ΔH[0à15[	22	56,41 %	2,64

	Antécédent					Conséquent	SUP	CONF	LIFT
R14	Δtt[-1à0[	Δff[1à2[	Δgnt[0à0]		==>	ΔH[0à15[	21	50,00 %	2,34
R15	Δtt[0à1[	Δff[-1à0[	Δgnt[0à0]		==>	ΔH[0à15[	20	60,61 %	2,84
R16	Δtt[0à1[	Δff[-1à0[	ΔLE[-15à0[		==>	ΔH[0à15[	18	58,06 %	2,72
R17	Δglo[100à738]	ΔLE[0à15[			==>	ΔH[15à30[	5	50,00 %	6,18
R18	Δtt[-2à-1[	Δff[0à1[	ΔLE[60à100[		==>	ΔH[30à60[	8	53,33 %	5,30
R19	Δglo[100à738]				==>	ΔH[60à340]	125	65,79 %	4,85
R20	Δff[0à1[	Δglo[100à738]			==>	ΔH[60à340]	72	54,14 %	3,99
R21	Δff[0à1[	Δglo[100à738]	ΔLE[100à475]		==>	ΔH[60à340]	54	71,05 %	5,24
R22	Δhu[-5à0[	Δglo[100à738]			==>	ΔH[60à340]	44	52,38 %	3,86
R23	Δtt[1à2[	Δglo[100à738]			==>	ΔH[60à340]	38	52,05 %	3,84
R24	Δtt[-10à-2[	Δglo[-708à-100[	ΔH[-315à-45[		==>	ΔLE[-590à-60[	14	51,85 %	9,28
R25	Δtt[-1à0[	Δff[1à2[	Δgnt[0à0]	ΔH[-30à-15[	==>	ΔLE[-30à-15[	4	57,14 %	10,13
R26	Δgnt[0à0]				==>	ΔLE[-15à0[	225	51,14 %	2,46
R27	Δtt[0à1[	Δgnt[0à0]			==>	ΔLE[-15à0[	74	52,48 %	2,52
R28	Δgnt[0à0]	ΔH[-15à0[			==>	ΔLE[-15à0[	74	58,73 %	2,82
R29	Δff[2à8]	Δgnt[0à0]			==>	ΔLE[-15à0[	59	54,63 %	2,62
R30	Δff[-1à0[	Δgnt[0à0]			==>	ΔLE[-15à0[	52	57,14 %	2,74

	Antécédent					**Conséquent**	**SUP**	**CONF**	**LIFT**
R31	Δff[0à1[	Δgnt[0à0]	ΔH[0à15[		==>	ΔLE[0à15[	17	50,00 %	3,21
R32	Δtt[-1à0[	Δhu[0à5[	Δgnt[0à0]	ΔH[0à15[	==>	ΔLE[0à15[	16	55,17 %	3,54
R33	Δhu[-5à0[	Δglo]0à20[	ΔH[-15à0[		==>	ΔLE[0à15[	13	68,42 %	4,39
R34	Δtt[-1à0[	Δff[0à1[	Δgnt[0à0]	ΔH[0à15[	==>	ΔLE[0à15[	11	61,11 %	3,92
R35	Δff[0à1[	Δhu[0à5[	Δgnt[0à0]	ΔH[0à15[	==>	ΔLE[0à15[	10	71,43 %	4,58
R36	Δff[2à8]	Δhu[-75à-20[	ΔH[-315à-45[		==>	ΔLE[15à30[	4	50,00 %	6,77
R37	Δff[1à2[	Δglo[20à100[	ΔH[-15à0[		==>	ΔLE[30à60[	8	53,33 %	5,39
R38	Δtt[2à8]	ΔH[30à60[			==>	ΔLE[60à100[	7	58,33 %	5,39
R39	Δtt[-1à0[	Δglo[20à100[	ΔH[30à60[		==>	ΔLE[60à100[	7	50,00 %	4,62
R40	Δtt[2à8]	Δglo[100à738]	ΔH[30à60[		==>	ΔLE[60à100[	7	63,64 %	5,88
R41	Δglo[100à738]				==>	ΔLE[100à475]	190	51,08 %	2,61
R42	ΔH[60à340]				==>	ΔLE[100à475]	150	64,38 %	3,28
R43	Δglo[100à738]	ΔH[60à340]			==>	ΔLE[100à475]	125	68,68 %	3,50
R44	Δff[0à1[	Δglo[100à738]			==>	ΔLE[100à475]	76	57,14 %	2,91
R45	Δff[0à1[	ΔH[60à340]			==>	ΔLE[100à475]	64	73,56 %	3,75

Tableau 16 : Règles générées pour l'Eté (à partir du 20 Juin)

	Antécédent					Conséquent	SUP	CONF	LIFT
R1	Δff[-1à0[	Δhu[-5à0[	ΔLE[-590à-60[		==>	ΔH[-315à-45[	8	61,54 %	12,66
R2	Δtt[2à8]	Δff[2à8]	ΔLE[-15à0[		==>	ΔH[-45à-30[	9	56,25 %	12,73
R3	Δff[2à8]	Δglo[-20à0[			==>	ΔH[-30à-15[	9	56,25 %	4,35
R4	Δtt[2à8]	Δglo]0à20[			==>	ΔH[-15à0[	23	54,76 %	2,39
R5	Δtt[2à8]	Δff[-1à0[	Δgnt[0à0]		==>	ΔH[-15à0[	19	65,52 %	2,86
R6	Δtt[2à8]	Δff[-1à0[	ΔLE[-15à0[		==>	ΔH[-15à0[	18	66,67 %	2,91
R7	Δff[0à1[	Δhu[-5à0[	Δgnt[0à0]		==>	ΔH[-15à0[	17	60,71 %	2,65
R8	Δtt[2à8]	Δff[-1à0[	Δgnt[0à0]	ΔLE[-15à0[	==>	ΔH[-15à0[	15	71,43 %	3,12
R9	Δtt[0à1[	Δff[1à2[	ΔLE[0à15[		==>	ΔH[0à15[	10	52,63 %	3,15
R10	Δtt[-1à0[	Δff[1à2[	ΔLE[0à15[		==>	ΔH[0à15[	8	50,00 %	2,99
R11	Δtt[-1à0[	Δglo]0à20[	ΔLE[-15à0[		==>	ΔH[0à15[	8	53,33 %	3,19
R12	Δff[-1à0[	Δglo]0à20[	ΔLE[-15à0[		==>	ΔH[0à15[	7	53,85 %	3,22
R13	Δhu[-5à0[	Δglo]0à20[	ΔLE[-15à0[		==>	ΔH[0à15[	7	63,64 %	3,80
R14	Δtt[-1à0[	Δff[0à1[	Δhu[-10à-5[	Δglo[20à100[	==>	ΔH[15à30[	5	62,50 %	6,62
R15	Δglo[100à738]	ΔLE[0à15[			==>	ΔH[30à60[	7	50,00 %	4,67
R16	Δtt[0à1[	Δff[-2à-1[	Δhu[-5à0[		==>	ΔH[30à60[	6	60,00 %	5,60

	Antécédent					Conséquent	SUP	CONF	LIFT
R17	Δglo[100à738]	ΔLE[-15à0[			==>	ΔH[30à60[	5	71,43 %	6,67
R18	Δglo[100à738]				==>	ΔH[60à340]	202	64,74 %	3,60
R19	ΔLE[100à475]				==>	ΔH[60à340]	184	61,95 %	3,44
R20	Δglo[100à738]	ΔLE[100à475]			==>	ΔH[60à340]	119	78,29 %	4,35
R21	Δtt[2à8]	Δglo[100à738]			==>	ΔH[60à340]	94	80,34 %	4,46
R22	Δff[0à1[	Δglo[100à738]			==>	ΔH[60à340]	80	64,00 %	3,56
R23	Δff[-1à0[	ΔH[-315à-45[			==>	ΔLE[-590à-60[	14	53,85 %	7,12
R24	Δglo[-708à-100[	ΔH[-315à-45[			==>	ΔLE[-590à-60[	11	73,33 %	9,69
R25	Δff[-1à0[	Δhu[-5à0[	ΔH[-315à-45[		==>	ΔLE[-590à-60[	8	100,00 %	13,22
R26	Δhu[-5à0[	Δgnt[0à0]	ΔH[30à60[		==>	ΔLE[-60à-30[	5	83,33 %	11,61
R27	Δtt[0à1[	Δhu[-10à-5[	Δgnt[0à0]	ΔH[15à30[	==>	ΔLE[-30à-15[	5	50,00 %	6,81
R28	Δgnt[0à0]				==>	ΔLE[-15à0[	278	54,62 %	2,28
R29	Δgnt[0à0]	ΔH[-15à0[			==>	ΔLE[-15à0[	110	58,20 %	2,43
R30	Δff[0à1[	Δgnt[0à0]			==>	ΔLE[-15à0[	95	57,93 %	2,42
R31	Δtt[2à8]	Δgnt[0à0]			==>	ΔLE[-15à0[	90	73,17 %	3,05
R32	Δff[1à2[	Δgnt[0à0]			==>	ΔLE[-15à0[	83	53,90 %	2,25
R33	Δtt[1à2[	Δgnt[0à0]			==>	ΔLE[-15à0[	70	67,96 %	2,84

Antécédent						Conséquent	SUP	CONF	LIFT
R34	Δtt[-1à0[	Δff[0à1[	Δhu[0à5[	Δgnt[0à0]	==>	ΔLE[0à15[	7	53,85 %	4,54
R35	Δtt[-1à0[	Δhu[0à5[	Δgnt[0à0]	ΔH[0à15[	==>	ΔLE[0à15[	7	53,85 %	4,54
R36	Δff[2à8]	Δhu[-10à-5[	Δglo]0à20[		==>	ΔLE[0à15[	5	62,50 %	5,26
R37	Δff[1à2[	Δglo[20à100[	ΔH[-15à0[		==>	ΔLE[30à60[	8	66,67 %	7,10
R38	ΔH[60à340]				==>	ΔLE[100à475]	184	56,44 %	3,44
R39	Δglo[100à738]	ΔH[60à340]			==>	ΔLE[100à475]	119	58,91 %	3,59
R40	Δff[0à1[	ΔH[60à340]			==>	ΔLE[100à475]	79	60,31 %	3,68
R41	Δtt[2à8]	Δglo[100à738]			==>	ΔLE[100à475]	72	61,54 %	3,75
R42	Δtt[2à8]	ΔH[60à340]			==>	ΔLE[100à475]	59	59,00 %	3,60
R43	Δtt[2à8]	Δglo[100à738]	ΔH[60à340]		==>	ΔLE[100à475]	59	62,77 %	3,83

Tableau 17 : Règles générées pour l'Automne (à partir du 22 Septembre)

Antécédent						Conséquent	SUP	CONF	LIFT
R1	Δtt[-10à-2[	Δff[0à1[	ΔLE[-60à-30[		==>	ΔH[-315à-45[	5	100,00 %	11,41
R2	Δtt[-10à-2[	Δhu[5à10[	Δglo[-20à0[		==>	ΔH[-315à-45[	5	62,50 %	7,13
R3	Δtt[2à8]	Δff[2à8]	Δhu[-75à-20[		==>	ΔH[-315à-45[	4	66,67 %	7,60
R4	Δtt[0à1[	Δff[0à1[	ΔLE[-60à-30[		==>	ΔH[-45à-30[	4	57,14 %	7,95
R5	Δhu[-10à-5[	Δglo[-20à0[			==>	ΔH[-30à-15[	8	50,00 %	3,16
R6	Δtt[2à8]	Δff[0à1[	Δgnt[0à0]		==>	ΔH[-30à-15[	8	57,14 %	3,62
R7	Δtt[2à8]	Δff[1à2[	Δgnt[0à0]		==>	ΔH[-30à-15[	8	80,00 %	5,06

	Antécédent					Conséquent	SUP	CONF	LIFT
R8	Δtt[1à2[	ΔLE[-15à0[			==>	ΔH[-15à0[	23	53,49 %	2,14
R9	Δhu[-15à-10[	ΔLE[-15à0[			==>	ΔH[-15à0[	17	51,52 %	2,06
R10	Δtt[0à1[	Δff[-1à0[	Δgnt[0à0]		==>	ΔH[-15à0[	17	51,52 %	2,06
R11	Δtt[0à1[	Δff[-1à0[	ΔLE[-15à0[		==>	ΔH[-15à0[	17	51,52 %	2,06
R12	Δhu[-20à-15[	Δgnt[0à0]			==>	ΔH[-15à0[	16	59,26 %	2,37
R13	Δglo]0à20[	ΔLE[-15à0[			==>	ΔH[0à15[	25	52,08 %	2,41
R14	Δhu[5à10[	ΔLE[-15à0[			==>	ΔH[0à15[	19	52,78 %	2,44
R15	Δhu[5à10[	Δgnt[0à0]	ΔLE[-15à0[		==>	ΔH[0à15[	14	60,87 %	2,81
R16	Δtt[-1à0[	Δff[-2à-1[	Δgnt[0à0]		==>	ΔH[0à15[	10	50,00 %	2,31
R17	Δtt[-1à0[	Δff[1à2[	ΔLE[-15à0[		==>	ΔH[0à15[	10	52,63 %	2,43
R18	Δtt[-2à-1[	Δff[1à2[	Δhu[5à10[		==>	ΔH[15à30[	5	50,00 %	5,32
R19	Δhu[15à20[	Δgnt[0à0]			==>	ΔH[15à30[	4	66,67 %	7,09
R20	Δff[0à1[	Δglo[100à738]	ΔLE[60à100[		==>	ΔH[30à60[	7	63,64 %	9,05
R21	Δhu[5à10[	Δglo[100à738]			==>	ΔH[30à60[	6	54,55 %	7,76
R22	Δtt[0à1[	Δglo[100à738]			==>	ΔH[60à340]	27	52,94 %	10,16
R23	Δff[1à2[	Δglo[100à738]			==>	ΔH[60à340]	23	53,49 %	10,26
R24	Δhu[-10à-5[	Δglo[100à738]			==>	ΔH[60à340]	22	56,41 %	10,82
R25	Δglo[100à738]	ΔLE[30à60[			==>	ΔH[60à340]	21	50,00 %	9,59
R26	Δtt[0à1[	Δhu[-10à-5[	Δglo[100à738]		==>	ΔH[60à340]	17	77,27 %	14,82
R27	Δtt[0à1[	Δglo[100à738]	ΔLE[30à60[		==>	ΔH[60à340]	13	72,22 %	13,85
R28	Δglo[-708à-100[	ΔH[-315à-45[			==>	ΔLE[-60à-30[	5	55,56 %	8,08
R29	Δglo[-708à-100[	ΔH[-30à-15[			==>	ΔLE[-60à-30[	5	62,50 %	9,09
R30	Δff[-1à0[	Δgnt[0à0]	ΔH[15à30[		==>	ΔLE[-30à-15[	4	57,14 %	6,96

	Antécédent					Conséquent	SUP	CONF	LIFT
R31	Δgnt[0à0]				==>	ΔLE[-15à0[	247	60,54 %	2,02
R32	Δgnt[0à0]	ΔH[-15à0[			==>	ΔLE[-15à0[	92	70,77 %	2,36
R33	Δtt[-1à0[	Δgnt[0à0]			==>	ΔLE[-15à0[	80	56,34 %	1,88
R34	Δtt[0à1[	Δgnt[0à0]			==>	ΔLE[-15à0[	78	60,47 %	2,02
R35	Δff[-1à0[	Δgnt[0à0]			==>	ΔLE[-15à0[	70	72,16 %	2,41
R36	Δff[2à8]	Δgnt[0à0]	ΔH[0à15[		==>	ΔLE[0à15[	7	53,85 %	3,00
R37	Δtt[-1à0[	Δhu[0à5[	Δgnt[0à0]	ΔH[0à15[	==>	ΔLE[0à15[	7	50,00 %	2,79
R38	Δtt[0à1[	Δff[2à8]	Δhu[-5à0[	Δgnt[0à0]	==>	ΔLE[0à15[	7	53,85 %	3,00
R39	Δtt[0à1[	Δff[-1à0[	ΔH[-30à-15[		==>	ΔLE[0à15[	6	66,67 %	3,72
R40	Δtt[2à8]	Δhu[-75à-20[	ΔH[-30à-15[		==>	ΔLE[0à15[	6	60,00 %	3,35
R41	Δff[1à2[	ΔH[60à340]			==>	ΔLE[30à60[	14	51,85 %	4,53
R42	Δtt[-1à0[	Δglo[100à738]	ΔH[30à60[		==>	ΔLE[60à100[	5	62,50 %	7,19
R43	Δff[-2à-1[	Δglo[100à738]			==>	ΔLE[100à475]	4	57,14 %	9,16
R44	Δhu[-20à-15[	Δglo[100à738]			==>	ΔLE[100à475]	5	55,56 %	8,90
R45	Δglo[100à738]	ΔH[-30à-15[			==>	ΔLE[100à475]	4	50,00 %	8,01

Tableau 18 : Règles générées pour l'Hiver (à partir du 21 Décembre)

	Antécédent					Conséquent	SUP	CONF	LIFT
R1	Δtt[-10à-2[	ΔLE[-590à-60[			==>	ΔH[-315à-45[	9	60,00 %	6,32
R2	Δhu[-10à-5[	Δglo[-20à0[			==>	ΔH[-30à-15[	7	63,64 %	4,91
R3	Δff[1à2[	Δgnt[0à0]	ΔLE[-15à0[		==>	ΔH[-15à0[	28	50,91 %	2,06
R4	Δtt[2à8]	Δgnt[0à0]			==>	ΔH[-15à0[	20	57,14 %	2,31
R5	Δtt[2à8]	ΔLE[-15à0[			==>	ΔH[-15à0[	16	57,14 %	2,31
R6	Δtt[-1à0[	Δff[1à2[	ΔLE[-15à0[		==>	ΔH[-15à0[	16	50,00 %	2,02
R7	Δtt[0à1[	Δhu[-5à0[	Δgnt[0à0]		==>	ΔH[-15à0[	16	53,33 %	2,16

	Antécédent					Conséquent	SUP	CONF	LIFT
R8	Δff[0à1[	Δhu[0à5[	Δgnt[0à0]		==>	ΔH[0à15[	18	56,25 %	2,72
R9	Δtt[-1à0[	Δff[0à1[	Δhu[0à5[		==>	ΔH[0à15[	13	50,00 %	2,42
R10	Δff[-1à0[	Δhu[-5à0[	ΔLE[-15à0[		==>	ΔH[0à15[	11	55,00 %	2,66
R11	Δgnt[0à0]	ΔLE[-60à-30[			==>	ΔH[0à15[	10	50,00 %	2,42
R12	Δtt[-2à-1[	Δff[0à1[	Δgnt[0à0]		==>	ΔH[0à15[	9	52,94 %	2,56
R13	Δtt[0à1[	Δglo[100à738]	ΔLE[60à100[		==>	ΔH[30à60[	9	56,25 %	6,87
R14	Δglo[100à738]				==>	ΔH[60à340]	15	55,56 %	9,35
R15	Δtt[-1à0[	Δglo[100à738]	ΔLE[60à100[		==>	ΔH[60à340]	12	52,17 %	8,78
R16	Δtt[0à1[	Δff[0à1[	Δglo[100à738]		==>	ΔH[60à340]	11	55,00 %	9,26
R17	Δff[1à2[	ΔLE[100à475]			==>	ΔH[60à340]	9	60,00 %	10,10
R18	Δtt[0à1[	ΔLE[100à475]			==>	ΔH[60à340]	8	72,73 %	12,24
R19	Δglo[-708à-100[	ΔH[-315à-45[			==>	ΔLE[-590à-60[	9	52,94 %	14,61
R20	Δtt[-10à-2[	Δglo[-708à-100[	ΔH[-315à-45[		==>	ΔLE[-590à-60[	6	54,55 %	15,05
R21	Δff[-2à-1[	Δglo[-708à-100[			==>	ΔLE[-590à-60[	4	50,00 %	13,80
R22	Δtt[-10à-2[	Δff[-7à-2[			==>	ΔLE[-60à-30[	6	50,00 %	7,84
R23	Δff[-2à-1[	Δhu[15à20[			==>	ΔLE[-30à-15[	5	50,00 %	6,70
R24	Δgnt[0à0]				==>	ΔLE[-15à0[	248	50,00 %	1,72
R25	Δgnt[0à0]	ΔH[-15à0[			==>	ΔLE[-15à0[	92	59,74 %	2,06
R26	Δhu[0à5[	Δgnt[0à0]			==>	ΔLE[-15à0[	73	57,03 %	1,97
R27	Δgnt[0à0]	ΔH[0à15[			==>	ΔLE[-15à0[	73	55,73 %	1,92
R28	Δff[1à2[	Δgnt[0à0]			==>	ΔLE[-15à0[	55	52,88 %	1,82
R29	Δff[2à8]	Δhu[-5à0[			==>	ΔLE[0à15[	19	52,78 %	2,60
R30	Δtt[0à1[	Δff[2à8]	Δgnt[0à0]		==>	ΔLE[0à15[	19	57,58 %	2,84
R31	Δtt[-1à0[	Δff[2à8]	ΔH[-15à0[		==>	ΔLE[0à15[	11	52,38 %	2,58
R32	Δtt[1à2[	Δhu[-5à0[			==>	ΔLE[0à15[	9	52,94 %	2,61

	Antécédent					Conséquent	SUP	CONF	LIFT
R33	Δtt[0à1[	Δff[2à8]	ΔH[-45à-30[		==>	ΔLE[0à15[	9	52,94 %	2,61
R34	Δff[1à2[	Δglo[100à738]	ΔH[60à340]		==>	ΔLE[100à475]	8	50,00 %	15,00
R35	Δtt[0à1[	Δhu[10à15[	ΔH[60à340]		==>	ΔLE[100à475]	5	71,43 %	21,43
R36	Δhu[10à15[	Δglo[100à738]	ΔH[60à340]		==>	ΔLE[100à475]	4	50,00 %	15,00
R37	Δtt[0à1[	Δhu[10à15[	Δglo[100à738]	ΔH[60à340]	==>	ΔLE[100à475]	4	66,67 %	20,00
R38	Δglo[100à738]	ΔH[-315à-45[			==>	ΔLE[100à475]	3	60,00 %	18,00

Les règles générées sont des règles d'association les plus fiables relativement au seuil de support déterminé ($\min sup = 80$) ; En plus des règles d'association rares (ie, les règles dont les supports sont inférieurs à $\min sup$. Un nombre important de règles ont été générées. En plus, toutes ces règles ont pour conséquents des erreurs sur la chaleur sensible (ΔH) et sur la chaleur latente (ΔLE). Le Lift de chaque règle générée est supérieur à 1, cela indique une dépendance positive. Ce qui implique que l'occurrence de l'antécédent a un effet positif sur l'occurrence du conséquent. Cependant, pour faciliter l'analyse et l'interprétation, il est important de réduire le nombre de règles générées. Ainsi, en utilisant le schéma de base pour la réduction du nombre de règles (*Figure 3-4*), le schéma ci-dessous (*Figure 3-8*) illustre une visualisation de la réduction de l'ensemble des règles générées au printemps dans le tableau (*Tableau 15*).

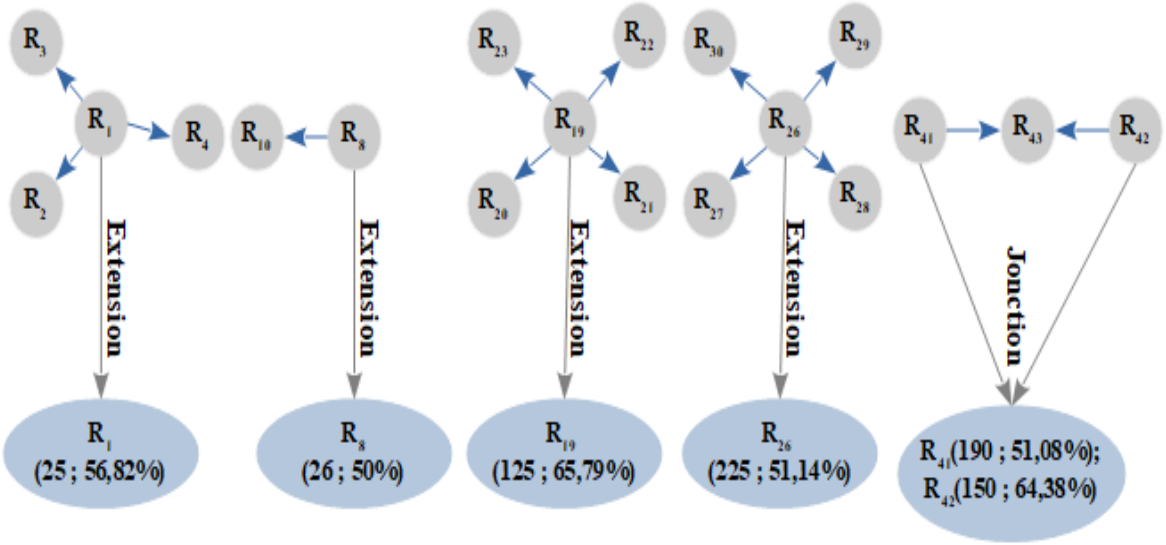


Figure 3-8 : Visualisation des règles réduites

Les règles R1, R2, R3 et R4 ont toutes une conséquence commune. En outre, l'intersection de leurs antécédents vaut l'antécédent de R1. Alors, les règles R2, R3, R4

peuvent être considérées comme des extensions de R1. Ainsi, parmi les règles (R1, R2, R3 et R4), seule la règle R1 sera conservée et considérée pour l'analyse et l'interprétation. De façon analogue, pour l'analyse et l'interprétation : la règle R8 représente la règle R10 ; la règle R19 représente les règles (R20, R21 et R23) ; la règle R26 représente les règles (R27, R28, R29 et R30) ; et pour le cas de la jonction, les règles R41 et R42 représentent la règle R43.

3.3.3 Étape de post-traitement (Visualisation et interprétation des résultats)

3.3.3.1 Interprétation statique des règles générées

Afin d'illustrer l'interprétation statistique des règles d'association obtenues, nous choisissons une règle (R19) dans (*Tableau 12*), ce qui suit :

R19 (Δglo[100à738] ==> ΔH[60à340]), avec le couple (support : 125 , confiance : 66%) :

Cette règle signifie que les instances de différences (erreur sur le rayonnement global variant entre 100 w/m² et 738 w/m²) et l'erreur sur la chaleur sensible variant entre 60 w/m² et 340 w/m² apparaissent simultanément cent vingt-cinq fois (i.e. 125/ 24*365 valeurs).

De plus, 66% des différences (erreur sur le rayonnement global variant entre 100 w/m² et 738 w/m²) ont pour conséquent l'erreur sur la chaleur sensible variant entre 60 w/m² et 340 w/m².

3.3.3.2 Interprétation sémantique des règles générées

Afin d'illustrer l'interprétation sémantique des règles, les tableaux synthétiques suivants présentent la dépendance statistique avec les précisions entre les erreurs sur les flux de chaleur sensible (ou de chaleur latente) et les erreurs sur d'autres variables (par exemple, le rayonnement global, ...), pour chaque variation saisonnière dans l'année.

Pendant le printemps, 45 règles ont été générées (*Tableau 15*). Parmi ces règles, 5 règles ont pour conséquent un biais grand sur la chaleur sensible ΔH([60à340]) et 5 règles ont pour conséquence un biais grand sur la chaleur latente ΔLE([100à475]).

Parmi les 5 règles qui ont pour conséquent un biais grand sur la chaleur sensible ΔH, il y a une concomitance avec un biais grand sur le rayonnement global Δglo. On note que la simulation de l'écoulement du vent a une faible concomitance avec la simulation de la chaleur sensible. Ceci est résumé dans (*Tableau 19*) qui suit :

Tableau 19 : Tableau synthétique des 5 règles avec conséquent : ΔH ([60à340])

Antécédents : Δglo[100à738[, Δtt[1à2[, Δff[0à1[	Conséquent : sup=125 ΔH[60à340]

	Fréquence :	*Simulation* :	*Simulation* :
Δglo[100à738[	5/5	Biais grand	Biais grand
Δtt[1à2[	1/5	Biais assez-faible	Biais grand
Δff[0à1[	2/5	Biais faible	Biais grand

Parmi les 5 règles qui ont pour conséquent un biais grand sur chaleur latente ΔLE, il y a une concomitance avec un biais grand sur rayonnement global Δglo, car Δglo apparaît 3 fois en antécédent avec un biais grand et l'erreur sur la vitesse du vent Δff apparaît 2 fois en antécédent avec un biais faible. Ceci est résumé dans (*Tableau 20*) qui suit :

Tableau 20 : Tableau synthétique des 5 règles avec conséquent : ΔLE ([100à475[)

Antécédents : Δglo[100à738], Δff[0à1[			Conséquent : sup=125 ΔH[60à340]
	Fréquence :	*Simulation* :	*Simulation* :
Δglo[100à738]	3/5	Biais grand	Biais grand
Δff[0à1[	2/5	Biais faible	Biais grand

En résumé, dans ce cas d'étude pour toutes les variations saisonnières dans l'année 2016, ces interprétations ont permis de comprendre et d'identifier les paramètres les plus influents sur les biais des flux de chaleur sensible et de flux de chaleur latente. En conclusion, le graphique ci-dessous résume l'interprétation sémantique de l'ensemble des règles générées (*Figure 3-9*).

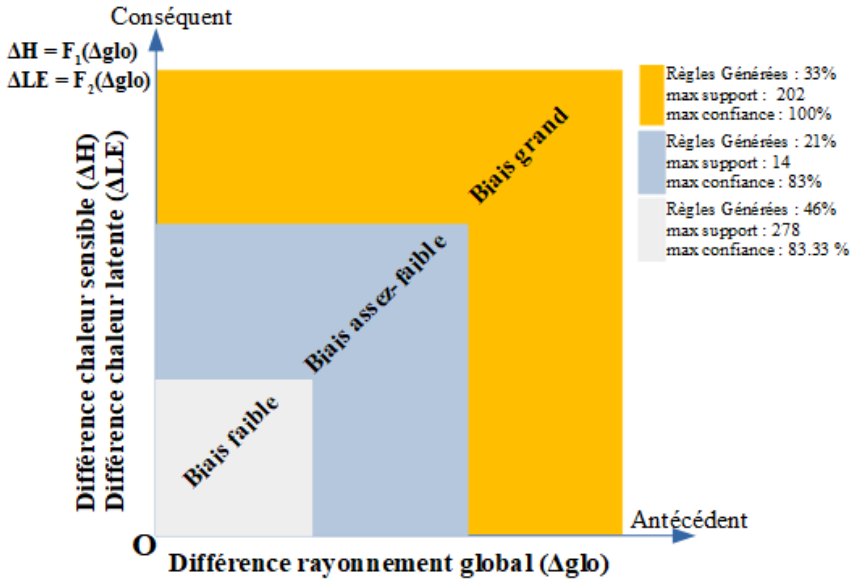


Figure 3-9 : Interprétation sémantique des règles

3.4 Discussion sur les résultats obtenus

Dans ce chapitre, le Data Mining est appliqué pour l'extraction des connaissances dans des bases de données météorologiques (mesurées/simulées). Dans le prétraitement des données météorologiques, une transformation des données a été faite à priori et ensuite le modèle autorégressif (ARMA) a été utilisé pour la gestion des données manquantes monotones.

Dans l'étape de prétraitement, une méthode de discrétisation est adoptée pour déterminer les items pertinents et d'obtenir un format de données qui peut être traité par la méthode d'extraction considérée ; les différences (erreurs) existantes entre les deux bases de données mesurées et simulées ont été mises en évidence. A l'étape de traitement, ces nouveaux paramètres d'erreurs ont été explorés pour identifier les paramètres qui influent sur les erreurs importantes dans la simulation des flux de chaleur sensible et chaleur latente. Les résultats obtenus conduisent à un nombre important de règles d'association pertinentes pour les interprétations. Dans l'étape de post-traitement, une méthode de réduction du nombre de règles a été adoptée pour faciliter l'interprétation des résultats. Les résultats montrent que la qualité de la simulation du rayonnement global a une dépendance statistique avec la qualité de simulation des flux de chaleur sensible et de chaleur latente. Les autres paramètres ont un impact statistiquement négligeable sur la simulation de la chaleur sensible et la chaleur latente. Ces informations sont relatives à la compréhension des échanges d'énergies entre la surface et l'atmosphère. Ceci pourra aider à prendre des mesures nécessaires pour améliorer la qualité de la simulation, mais aussi pour faciliter les instructions efficaces des décideurs (Pereira et al., 2018).

La réduction des règles en utilisant des propriétés ensemblistes fournie des idées pour la génération automatique des moyens de factorisations des règles avec la mise en évidence

des formes de méta règles. Ceci peut s'intégrer dans un processus algorithmique d'obtention de règles sur le critère d'une forme de relation d'ordre souhaitée. C'est dans cette optique que s'inscrit la sélection et le classement des règles en se basant sur une relation d'ordre décrivant la dominance ou la non dominance des règles d'associations (Bouker, 2012). La sélection des règles obéit à deux conditions : (i) toute règle dominée ne peut être mieux classée qu'une règle non dominée, (ii) deux règles non dominées doivent être rangées selon un degré de similarité relativement à une règle de référence fournie par l'utilisateur. Le classement des règles s'effectue en partant de l'ensemble des règles les plus pertinentes (c'est-à-dire les règles non dominées) et l'utilise pour identifier l'ensemble classé suivant (c'est-à-dire le *successeur*). En parallèle, une autre perspective duale reste possible. Il repose sur le fait de partir de l'ensemble des règles les moins pertinentes (c'est-à-dire les règles qui ne dominent pas les autres règles) et de les utiliser pour identifier l'ensemble précédent de règles classées dénommé ensemble *prédécesseur*.

Ce principe algorithmique pourrait être amélioré par prise en compte de nos propriétés ensemblistes de jonction, extension et éventuellement permutation. En réalité, le déploiement de ces propriétés a nécessité la vérification de l'égalité ou de l'inclusion ensembliste des antécédents et des conclusions des règles. Cette vérification peut se traduire par une mesure relative à des conditions d'inclusions ensemblistes. En outre, l'avantage de notre forme de domination des règles est la mise en œuvre des moyens d'évaluation aisément compréhensibles par l'utilisateur car faisant apparaître clairement la comparaison des ensembles d'antécédents ou de conclusion des règles. Des investigations plus poussées sur l'intégration d'autres techniques mathématiques permettraient certainement des avancées théoriques dans une forme de description hiérarchique des règles avec des liens sémantiques formalisées.

3.5 Conclusion

Ce chapitre a proposé une méthodologie progressive d'évaluation de la simulation d'un modèle de prévisions numériques du temps. Cette méthodologie comprend trois étapes : (i) prétraitement avec le nettoyage de données, et des techniques de réduction (par exemple discrétisation pour déterminer les items pertinents avec une transformation des données quantitatives en des données qualitatives), (ii) exploration des erreurs pour découvrir des relations intéressantes entre les erreurs sur les variables considérées, et (iii) post-traitement avec l'analyse et l'interprétation des règles générées pour comprendre et identifier les paramètres qui influent sur les erreurs importantes dans la simulation des flux de chaleur sensible et chaleur latente.

Particulièrement, dans l'étude de cas cible, les données météorologiques ne sont pas parfaites. Dans le cadre de la gestion des valeurs manquantes, la méthode d'imputation *k*-NN a des inconvénients face aux valeurs manquantes monotones. Cela impacte la qualité de la discrétisation en affectant les résultats globaux (règles d'association générées). Pour minimiser les inconvénients de *k*-NN face aux valeurs manquantes monotones, la méthode moyenne mobile a été utilisée pour remplacer les valeurs manquantes monotones. Ensuite, un ajustement a été réalisé avec les modèles à moyenne mobile

intégrée autorégressive pour remplacer des valeurs manquantes et réduire les fluctuations irrégulières (intégrant éventuellement des accidents de mesures). En plus, la structure de données météorologiques n'est pas compatible avec l'algorithme utilisé (notamment FP-Growth). De ce fait, une approche de discrétisation a été adoptée en prétraitement pour transformer les données originales sous forme d'intervalles (items) codés et pour obtenir ainsi un format qui pourrait être importé dans le traitement de données. Cette transformation permet de réduire l'espace de recherche et le temps d'exécution de l'algorithme. En plus, quelques notions des suites numériques en mathématiques ont été utilisées pour déterminer les bornes des intervalles. Cela nous a permis de faire une répartition pertinente des échantillons entre les intervalles.

Par exemple, avec un couple de mesures (min sup=225 ; min conf=50%) choisi, l'exploration de la base des erreurs a généré un nombre important de règles d'associations rares pour toutes les variations saisonnières. Vu ce nombre important de règles générées, des propriétés mathématiques sur la théorie des ensembles ont été utilisées pour réduire le nombre de règles générées et pour faciliter la visualisation et l'interprétation des résultats, voir (*Figure 3-4*) et (*Figure 3-8*). Le graphique (*Figure 3-9*) résume la visualisation et l'interprétation sémantique des règles générées sur toutes les variations saisonnières (printemps, été, automne et hiver) en 2016. Dans ce cas d'étude, l'approche utilisée nous a permis de comprendre le comportement des autres paramètres associés et d'identifier le plus influent avec précision.

En résumé, le résultat obtenu montre qu'un dysfonctionnement du rayonnement global, avec une fiabilité (max support= 225) et une précision (max confiance= 100%), est souvent dû à des perturbations naturelles (par exemple, le passage des nuages au-dessus des capteurs). Ceci pourra produire une importante erreur de simulation sur le modèle numérique de prévision. Ces genres de perturbations impactent, souvent, la qualité des observations/simulations des flux de chaleur sensible et chaleur latente (*Figure 3-9*).

L'incertitude est très dynamique sur les mesures de quelques grandeurs météorologiques (Jancic et al., 2018) ; ceci impacte la qualité de la simulation du modèle numérique de prévision. Une perspective de ce chapitre pourrait être de définir un optimiseur/correcteur qui pourra minimiser les erreurs de simulation. Par conséquent, le chapitre 4 suivant propose une méthodologie pour minimiser les erreurs de simulation.

Chapitre 4

4.Transport optimal avec les processus gaussiens pour minimiser les erreurs de simulation

4.1 Introduction

Les composantes des échanges d'énergies à la surface jouent un rôle très important dans le réchauffement climatique (Safa et al., 2018). La prévision de ces composantes par la modélisation numérique est nécessaire pour la compréhension de l'évolution des phénomènes dangereux liés au réchauffement climatique. Mais sur les données que nous étudions, la précision du modèle numérique est très difficile dans la simulation des échanges d'énergies à la surface. Ceci est souvent dû à des incertitudes sur des observations lors de l'évaluation de la simulation.

De nombreux scientifiques étudient le problème d'incertitude sur les observations et tentent de maximiser la précision (Zhendong Zhang et al., 2019). Par exemple, dans le cadre de modélisation de l'incertitude géo-spatiale, un réseau de processus gaussiens a été utilisé pour prédire en intégrant les incertitudes sur les observations à plusieurs étapes (Abdelfatah et al., 2018).

Par conséquent, il est important de rechercher des moyens permettant de fournir des pistes d'amélioration des modèles numériques de prévision du temps. L'objectif de ce travail de recherche est de proposer une approche pour contribuer à l'évaluation de la simulation d'un modèle numérique de prévision du temps. Cette approche est déployée sur deux bases de données contenant des variables mesurées et simulées. Ces variables sont présentées dans le *chapitre 2*.

Plus particulièrement, les caractéristiques dynamiques de la nature peuvent engendrer des difficultés dans la bonne modélisation des coordonnées des phénomènes observés et ceci peut être la source de l'incertitude dans les valeurs modélisées des flux de chaleur sensible et de chaleur latente. De ce fait, dans un premier temps, le Processus Gaussien pour l'apprentissage automatique dénommée Gaussian Processes for Machine Learning (GPML) est utilisé pour définir un prédicteur en modélisant par régression les données mesurées. Cette modélisation prend en compte les incertitudes et facilite le diagnostic du modèle numérique de prévision. Dans un second temps, le concept de transport optimal est utilisé en définissant un optimiseur à partir de quelques transformations géométriques ayant des fondements mathématiques. Cet optimiseur sert à effectuer un transport optimal

simultané par homothétie des données simulées (*source*) vers les voisinages des données mesurées (*cible*), ceci pour réduire les erreurs de simulation.

Après cette introduction, la *section 4.2* présente l'approche proposée pour la minimisation des erreurs de simulation. Ensuite, la *section 4.3* désigne une étude de cas. Les résultats sont présentés dans la *section 4.3.1*, qui est suivie d'une discussion sur les résultats obtenus dans la *section 4.3.2*. Enfin, la *section 4.4* donne une conclusion.

4.2 Méthodologie adoptée pour minimiser les erreurs de simulation

Dans cette section, une méthodologie est proposée pour la modélisation des variables météorologiques mesurées et la minimisation des erreurs produites, lors de la simulation, par le modèle numérique de prévision. Le schéma suivant illustre l'approche proposée (*Figure 4-1*).

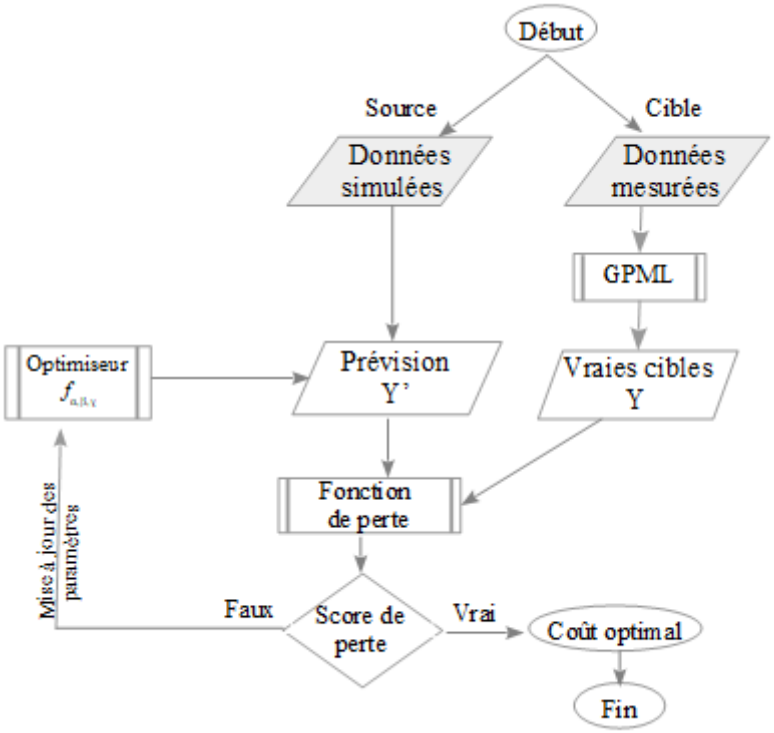


Figure 4-1 : Méthodologie proposée pour minimiser les erreurs

Cette méthodologie permet d'aborder en deux étapes la problématique par une méthode d'apprentissage automatique (GPML) (Jancic et al., 2018), en combinaison avec le concept de Transport Optimal (Monge, 1781), (Kantorovich, 1982), (Courty et al., 2015), (Bernard et al., 2018) : dans un premier temps, il s'agit d'établir un optimiseur f jouant sur les données simulées (Y') pour les transporter vers le voisinage des données mesurées. Ceci permettra de minimiser (corriger) les erreurs de simulation (il s'agit d'une mise à l'échelle les données simulées). Ensuite, il s'agit de trouver une fonction prédictive Y

(notée g) en modélisant les données mesurées. Cette fonction g facilitera le diagnostic du modèle numérique de prévision pour comprendre et identifier les anomalies lors de l'évaluation de la simulation des variables météorologiques. Dans cette approche, les variables mesurées/simulées considérées sont présentées dans le *chapitre 2*. Dans cette section, nous abordons d'abord le concept de transport optimal. Ensuite, nous présentons la conception de l'optimiseur, la spécification des contraintes de séries temporelles, et enfin une description du GPML.

4.2.1 Notion de Transport Optimal

Le problème du transport optimal initial a été introduit et étudié par *MONGE* en 1781 (Monge, 1781).

Le transport optimal est un ensemble de procédés qui permet à la fois de définir une notion géométrique et naturelle de distance entre deux distributions de probabilités ; et de transformer une distribution en une autre à moindre coût.

La déclaration du problème initiale de Monge (*PM*) est la suivante :

$$(PM): \begin{cases} \inf\limits_{T:X\to X} \int\limits_{X} c(x,T(x))u(x)dx \\ \text{sous contrainte:} \\ \forall B \subset X, \int\limits_{T^{-1}(B)} u(x)dx = \int\limits_{B} v(x)dx \end{cases} \quad (4.2.1.1)$$

où X est un sous-ensemble de $\mathbb{R}^2$, u et v sont deux fonctions positives définies sur X et $c(\cdot\,,\,\cdot)$ est une distance euclidienne dans l'énoncé du problème initial de Monge ; les fonctions u et v représentent respectivement la hauteur du domaine source et celle du domaine cible.

Les insuffisances de contraintes ont rendu le problème de Monge très difficile telles que : (i) la contrainte de conservation de la masse locale ; (ii) la fonctionnalité optimisée par le problème de Monge est non symétrique, ce qui pose des difficultés lorsqu'on étudie l'existence de solutions à son problème.

Pour surmonter ces difficultés, Kantorovich a énoncé un problème avec un plus grand espace de solutions, c'est-à-dire un assouplissement du problème de Monge, où la masse peut être à la fois divisée et fusionnée (Kantorovich, 1982). L'idée consiste à supposer que le graphique de T est une fonction g définie sur $X \times Y$ qui indique pour chaque couple de points $x \in X, y \in Y$ combien de matière passe de x à y. Cependant, il n'est pas possible d'utiliser les fonctions standards pour représenter le graphique de T. Mais lorsqu'on pense au graphe d'une fonction univariée $x \mapsto f(x)$, il est défini sur $\mathbb{R}^2$ mais *concentré* sur une courbe. Pour cette raison, il faut utiliser des mesures. Ainsi, il faut

chercher maintenant une mesure γ pris en charge par l'espace produit $X \times Y$. Le problème de Kantorovich (PK) est énoncé comme suit :

$$\inf_{\gamma}\left\{\int_{X\times Y} c(x,y)d\gamma \mid \gamma \geq 0 \text{ et } \gamma \in \Pi(\mu,\nu)\right\} \qquad (4.2.1.2)$$

où $\Pi(\mu,\nu) = \{\gamma \in P(X \times Y) | (PX)\#\gamma = \mu; (PY)\#\gamma = \nu\}$; (PX) et (PY) désignent les deux projections $(x,y) \in X \times Y \longmapsto x$ et $(x,y) \in X \times Y \longmapsto y$, respectivement.

Les deux mesures $(PX)\#\gamma$ et $(PY)\#\gamma$ obtenues en faisant avancer γ par les deux projections sont appelées les marges de γ. Les mesures γ de l'ensemble admissible $\Pi(\mu, \nu)$, c'est-à-dire les mesures ayant μ et ν comme marginaux, sont appelées plans de transport optimaux.

Dans le contexte de l'adaptation de domaine non supervisée, *Courty* et ses collègues ont proposé la méthode du transport optimal de distribution conjointe dénommée Joint Distribution Optimal Transportation (JDOT) (Courty et al., 2015). L'idée est de considérer le problème de transport optimal entre les distributions sur le produit cartésien de l'espace des caractéristiques avec l'espace des étiquettes, au lieu de considérer uniquement les distributions de l'espace des caractéristiques. Dans ce contexte, la mesure source μ_s et les mesures cibles ν_t sont des mesures sur l'espace du produit. Nous notons les couples (x^s, y^s) et (x^t, y^t) des échantillons de μ_s et ν_t respectivement, et ceci est illustré par le graphique suivant (*Figure 4-2*) :

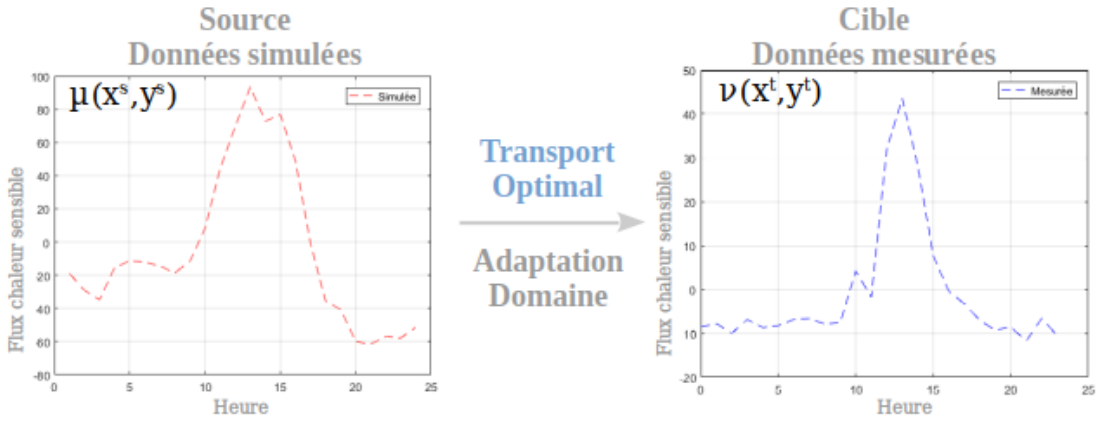


Figure 4-2 : Transport Optimal - Adaptation de domaine

Le coût (distance) généralisé associé à cet espace peut être naturellement exprimé comme une combinaison pondérée des coûts dans les espaces des entrées x (abscisses) et des étiquettes y (ordonnées). Ceci s'exprime par la formule suivante

$$d(x_i^s, y_i^s ; x_j^t, y_j^t) = \alpha c(x_i^s, x_j^t) + \beta L(y_i^s, y_j^t) \qquad (4.2.1.3)$$

pour le i^{ème} échantillon du support de μ_s et le j^{ème} échantillon du support de ν_t, la fonction de coût $c(x_i^s, x_j^t)$ choisie dépend de l'activité d'entrainement. Les fonctions coûts $c(\cdot\,,\,\cdot)$

et $L(\cdot\ ,\ \cdot)$ sont respectivement des distances entre les caractéristiques source-cible et entre les étiquettes source-cible.

On peut donc avoir des fonctions de coût relatives à la classification (par exemple, la norme 1 ou à la régression (par exemple, la norme 2). Les paramètres α et β sont deux scalaires pondérant les contributions relatives des différences entre les caractéristiques et entre les étiquettes.

Dans le cadre de l'adaptation de domaine non supervisé, les étiquettes y_j^t sont inconnues et il s'agit d'entraîner un classifieur (estimateur) $h: X \to Y$ pour estimer le label $h(x_j^t)$ pour chaque échantillon cible. C'est ainsi qu'avec le couple $(x^t, h(x_j^t))$ des échantillons issus de la distribution cible, nous pouvons définir la perte de la manière suivante :

$$d(x_i^s, y_i^s; x_j^t) = \alpha c(x_i^s, x_j^t) + \beta L(y_i^s, h(x_j^t)) \qquad (4.2.1.4)$$

Dans notre contexte, les étiquettes y_j^t sont connues. Au lieu d'utiliser un classifieur h pour l'estimation des échantillons cible comme Courty fait dans (Courty et al., 2015), on définit plutôt un optimiseur f jouant sur les échantillons source pour effectuer le transport des étiquettes de la source (*valeurs simulées*) vers les voisinages des étiquettes de la cible (*valeurs mesurées*). Ceci donne la formulation suivante (4.2.1.5) :

$$d(x_i^s, y_i^s\ ; x_j^t, y_j^t) = \alpha c(x_i^s, x_j^t) + \beta L(f(y_i^s), y_j^t) \qquad (4.2.1.5)$$

Compte tenu de la similarité des indices horaires entre, les différences entre les caractéristiques sont nulles avec la fonction de coût $c(x_i^s, x_j^t) = 0$; f est l'optimiseur qui sera défini dans la *section 4.2.2*.

4.2.2 Conception de l'optimiseur

Les données mesurées et simulées ont des ordres de grandeurs significativement différents. Ce chapitre propose alors une méthode de mise à l'échelle pour les données simulées. Soit X une variable météorologique mesurées, les fonctions considérées en entrée-sortie de l'approche sont les suivantes :

- En entrée de l'approche proposée (*Figure 4-1*), nous avons l'ensemble des valeurs horaires mesurées que l'on note $X = Obs$:

$$X : \begin{cases} A^p \longrightarrow B^p \\ \mathrm{t} \longmapsto Obs(\mathrm{t}) \\ o\grave{u}\ \mathrm{t}=(t_1, t_2,\ \ldots, t_p), \in \mathrm{A} \subset \mathrm{N}\ \text{et}\ \mathrm{B} \subset \mathbb{R} \end{cases} \qquad (4.2.2.1)$$

et sa simulation par le modèle numérique est notée $Y' = Mod$:

$$Y' : \begin{cases} A^p \longrightarrow B^p \\ \mathrm{t} \longmapsto Mod(\mathrm{t}) \\ o\grave{u}\ \mathrm{t}=(t_1, t_2,\ \ldots, t_p), \mathrm{A} \subset \mathrm{N}\ \text{et}\ \mathrm{B} \subset \mathbb{R} \end{cases} \qquad (4.2.2.2)$$

- En sortie de l'approche (*Figure 4-1*), nous obtenons la fonction prédictive $Y = \text{GPML(X)} = g$ en modélisant les valeurs mesurées, elle est définie par :

$$Y : \begin{cases} B^p \longrightarrow C^p \\ \text{X} \longmapsto g(\text{X}) \approx X \\ o\grave{u}\ \text{X}=(x_1, x_2, \ \ldots, x_p),\ \text{B} \subset \mathbb{R} \text{ et C} \subset \mathbb{R} \end{cases} \qquad (4.2.2.3)$$

et l'optimiseur f défini par :

$$\begin{cases} \text{B}^p \longrightarrow \text{C}^p \\ Z \longmapsto f_n(Z) = \text{sign} \times \left(\frac{\beta_n}{\gamma_n}\right)^{\alpha_n} \times Z \end{cases} \qquad (4.2.2.4)$$

où $\alpha_n,\ \beta_n, \gamma_n$ sont des réels strictement positifs ; $Z = (z_1, z_2, \ \ldots, z_p) \in \text{B}^p \subset \mathbb{R}^p$ et $C^p \subset \mathbb{R}^p$; les composants z_i désignent les valeurs simulées par le modèle numérique de prévision de la variable mesurée X , ie Z= Modèle (X) ; sign $= \pm 1$.

Cet optimiseur est une fonction paramétrique généralisée à partir d'une série de transformations effectuées sur différents types de formes géométriques rencontrées. Le terme $k = \text{sign} \times \left(\frac{\beta_n}{\gamma_n}\right)^{\alpha_n}$ est appelé rapport d'homothétie de centre mobile O situé sur l'axe des abscisses. L'objectif principal de l'approche proposée est de minimiser la fonction d'erreur en entrée $d(t) = Mod(t) - Obs(t)$. Le problème revient à trouver un optimiseur f telle que la fonction d'erreur à la sortie $d = f(Mod) - g(Obs)$ soit dans l'intervalle d'optimalité qui est connue pour chaque variable météorologique.

Les courbes représentatives des fonctions Obs et Mod présentent souvent des formes géométriques complexes, par exemple les courbes des données mesurées et simulées qui sont représentées dans la *Figure 4-2*. La composée des fonctions Obs et g, notée $(g \circ Obs)$, représente la modélisation des valeurs mesurées et produit une forme géométrique connue, notée C_{Obs}. Pour minimiser la fonction coût (erreur) d, il est donc nécessaire d'utiliser des propriétés sur les transformations géométriques par homothétie pour effectuer un déplacement des valeurs simulées vers les voisinages des valeurs mesurées. Ce déplacement s'effectuera en appliquant un optimiseur f (expression (4.2.2.4)). La composée des fonctions f et *Mod, notée* $(f \circ Mod)$, transporte les valeurs simulées vers le voisinage des valeurs mesurées.

Pour la résolution du problème posé, il serait nécessaire de reformuler le problème à un problème d'optimisation.

Définition 17. L'optimisation est une branche des mathématiques cherchant à modéliser, à analyser et à résoudre analytiquement ou numériquement les problèmes qui consistent à minimiser ou maximiser une **fonction coût** sur un ensemble (B. Li et al., 2019).

Ce cas d'étude est un problème de minimisation de la fonction d'erreur (fonction coût / distance) entre valeurs simulées et valeurs mesurées. Le problème revient donc à :

$$\text{minimiser} \left(\sqrt{\sum_{0}^{p} \left[f_{\alpha,\beta,\gamma}\left(Mod(t_i)\right) - g\left(Obs(t_i)\right) \right]^2} \right), \qquad (4.2.2.5)$$

$$\text{sous contraintes :} \begin{cases} f_{\alpha,\beta,\gamma}(X) = \text{sign} \times \left(\frac{\beta}{\gamma}\right)^{\alpha} \times X \\ g(\text{X}) \cong X \\ \quad \alpha > 0,\ \beta > 0,\ \gamma > 0 \\ \quad \text{a} \leq \max\left\{ |d(t_i)|_{i=0,\ldots,n} \right\} \leq b \\ a,\ b \text{ sont connus} \end{cases} \qquad (4.2.2.6)$$

$d(t_i) = f_{\alpha,\beta,\gamma}\left(Mod(t_i)\right) - g\left(Obs(t_i)\right)$ désigne l'erreur à l'instant i ;

Le problème revient à trouver les réels α, β et γ tels que la fonction coût d_{Ouput} soit minimale, avec

$$\text{d}_{\text{ouput}} = \sqrt{\sum_{0}^{p} [d(t_i)]^2}. \qquad (4.2.2.7)$$

Les paramètres α, β et γ sont choisis selon les cas suivants :

Cas 1 : le modèle sous-estime (OM' > OM)

C'est le cas d'un agrandissement du rapport (*Figure 4-3*). Il faut donc choisir les paramètres α, β et γ tels que : $\beta > \gamma$, γ fixé, faire varier α et β.

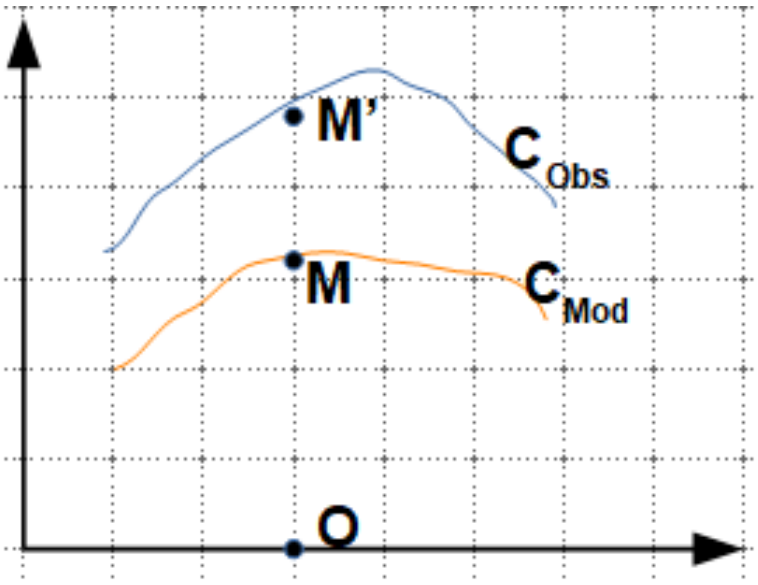


Figure 4-3 : Sous-estimation du modèle

Cas 2 : le modèle surestime (OM' < OM)

C'est le cas d'une réduction du rapport (*Figure 4-4*). Il faut donc choisir les paramètres α, β et γ tels que : $\beta < \gamma$, β fixé, faire varier α et γ.

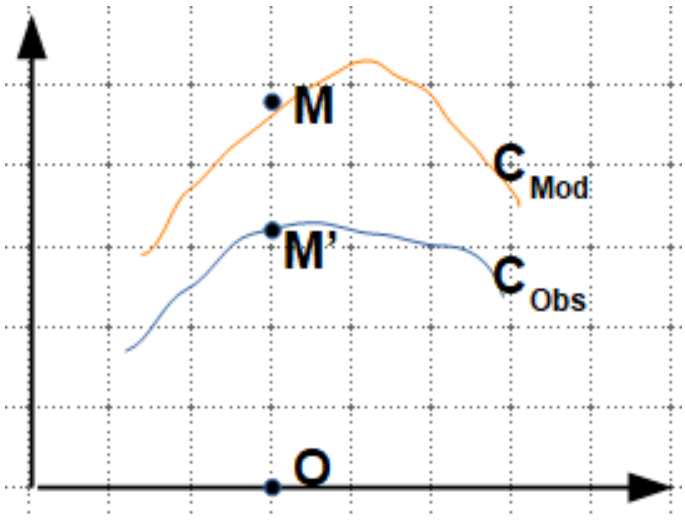


Figure 4-4 : Surestimation du modèle.

Cas 3 : Le modèle prend des valeurs nulles (O = M)

C'est le cas des points invariants où le centre O est confondu au point M (*Figure 4-5*). Il n'y a donc pas de transformation possible en ce point, car la distance OM est nulle. Dans de tel cas, on effectue d'abord un changement de variable par translation, par rapport à un vecteur u de norme supérieure à |min(Mod)|. Ce changement de variable permet à l'optimiseur d'effectuer le déplacement des points invariants :

$$\begin{cases} \|\vec{u}\| > |\min(Mod)| \\ \mathrm{OP} = \mathrm{OM} + \|\vec{u}\| \\ \mathrm{OP'} = \mathrm{OM'} + \|\vec{u}\| \end{cases}$$

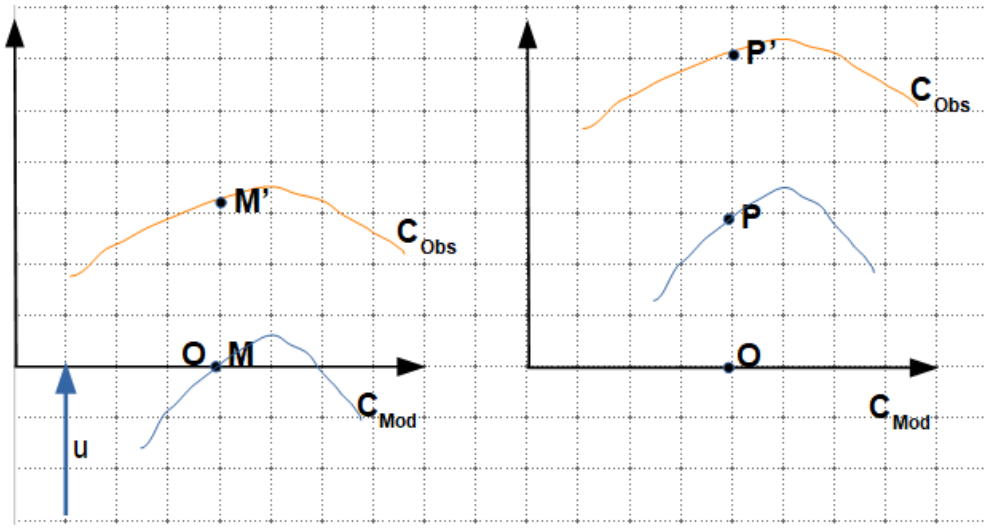


Figure 4-5 : Changement de variable par translation

Remarque : Dans le cas où les points M et M' sont du même côté par rapport au centre O, alors sign $= +1$. Dans le cas où le centre O se situe entre M et M', alors sign $= -1$.

Tableau 21 : Algorithme d'optimisation

ti est l'indice horaire ; j est le nombre d'itérations ; $\alpha,\ \beta\ et\ \gamma$ sont les paramètres de l'optimiseur ; $[a, b]$ est l'intervalle de convergence/optimalité

$$d_j = d(t_j) = f_{\alpha,\beta,\gamma}\left(Mod(t_j)\right) - g\left(Obs(t_j)\right)$$

1. **Input** : $\alpha > 0,\ \beta > 0,\ \gamma > 0$
2. Initialisation : $\alpha = 1\ ; \beta = 1\ ;\ \gamma = 1$
3. Begin
4. Pour chaque t_i variant de 1 à 24
5. j=1
6. Tant que $d_j \notin [a, b]$
7. j=j+1
8. Calculer paramètres $\alpha_j\ ; \beta_j\ et\ \gamma_j$
9. Mettre à jour l'optimiseur f
10. Mettre à jour d_j
11. Fin tant que
12. Sauvegarder $\alpha_j\ ; \beta_j\ et\ \gamma_j$
13. Fin pour

4.2.3 Contraintes des séries temporelles

Cette section donne un bref aperçu des classes de contraintes existantes. Chaque classe de contraintes est décrite de la manière suivante : Il faut d'abord donner la convention de dénomination utilisée pour construire systématiquement le nom des contraintes. Ensuite, il est nécessaire de fournir le pattern d'argument commun à toutes les contraintes appartenant à une classe. De même, il est utile d'énoncer le but des contraintes d'une classe. En fin, il est essentiel d'illustrer l'utilisation de ces contraintes par un exemple où nous calculons quelques informations provenant d'une série chronologique fixe, et un autre exemple où nous générons des séries temporelles satisfaisant à une propriété donnée. Quelle que soit la classe à laquelle elles appartiennent, toutes les contraintes décrites dans la littérature ont une série temporelle finie X=(x_1, x_2, …, x_n) avec plusieurs arguments, où chaque x_i est un nombre entier fixe ou une variable de domaine (Arafailova, 2018), (Arafailova et al., 2018). On déduit de X la séquence des valeurs de signature S=(s_1, s_2, …, s_n) via les contraintes de signature ($x_i < x_{i+1} \Leftrightarrow s_i = '<'$) ; ($x_i = x_{i+1} \Leftrightarrow s_i = '='$) et ($x_i > x_{i+1} \Leftrightarrow s_i = '>'$) pour tout i ∈ {1,...,n-1}. Les noms des contraintes utilisées sont systématiquement construits à l'aide de la grammaire suivante, où les littéraux utilisent des minuscules qui sont indiquées entre guillemets, tandis que tous les noms des contraintes seront en majuscules.

Définition 18. Un **PEAK** est une occurrence maximale de l'expression régulière

$$'<(=|<)*(>|=)*>' \qquad (4.2.3.1)$$

dans la séquence de valeurs de signature associée à une série chronologique. Le quantificateur ' * ' correspond à ce qui le précède, répété zéro ou plusieurs fois.

Étant donné une séquence S={'<' ;'=' ;'>' } , une occurrence du pattern PEAK est la sous-séquence maximale de S qui correspond à l'expression régulière (4.2.3.1). La partie (A) et les parties (B) du graphique (*Figure 4-6*) représentent respectivement l'Automate d'états finis associé au pattern PEAK ainsi qu'un exemple de son exécution sur une série chronologique.

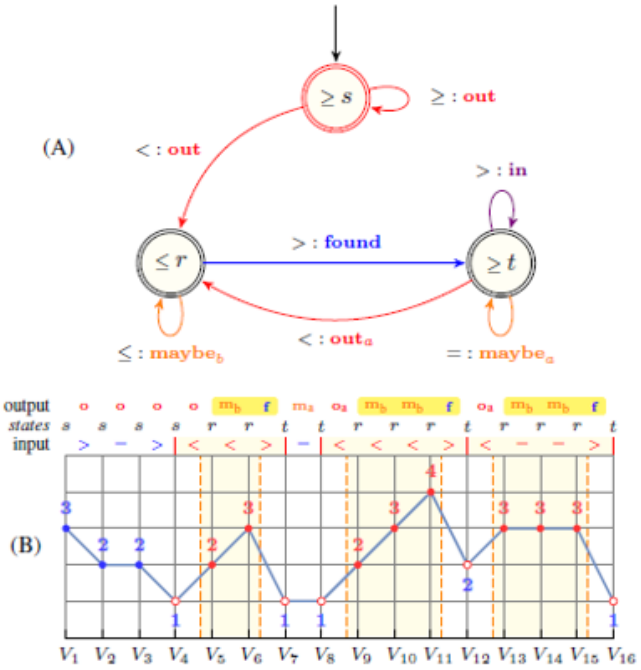


Figure 4-6 : Traducteur associé au pattern PEAK

Dans le graphique (*Figure 4-6*), le schéma (A) représente l'Automate d'états finis pour le pattern PEAK : '<(=|<)*(>|=)*>' ; le schéma (B) illustre l'exécution du pattern PEAK sur une série chronologique (les sorties o, mb, f , ma, oa sont des raccourcis pour out, maybeb, found, maybea outa).

Définition 19. Une **GORGE** est une occurrence minimale de l'expression régulière :

$$'(>|(>(=|>)*>))(<|<(=|<)*<)' \qquad (4.2.3.2)$$

Le quantificateur ' * ' correspond à ce qui le précède, répété zéro ou plusieurs fois.

Étant donné une séquence S={ '<' ;'=' ;'>' } , une occurrence du pattern GORGE est la sous-séquence maximale de S qui correspond à l'expression régulière (4.2.3.2). Les graphiques (A) et (B) représentent respectivement l'Automate d'états finis associé au pattern GORGE ainsi qu'un exemple de son exécution sur une série chronologique.

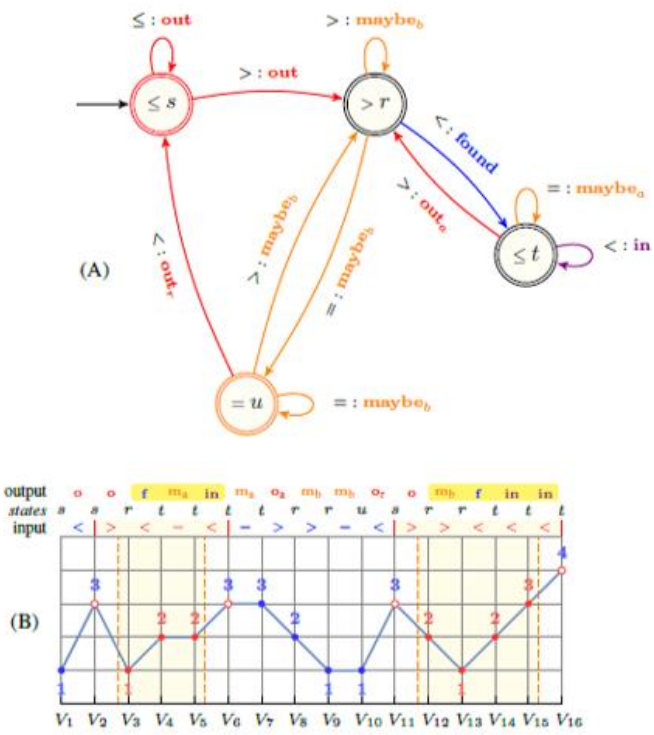


Figure 4-7 : Traducteur associé au pattern GORGE

Dans le graphique (*Figure 4-6*), le schéma (A) représente l'Automate d'états finis pour le pattern GORGE : '(>| (>(= | >)* >)) (< | < (= | <)* <)' ; le schéma (B) illustre l'exécution du l'Automate d'états finis sur une série chronologique (dans le cadre de la sortie o, f, ma, oa, mb, ou sont des raccourcis pour out, found, maybea, outa, maybeb, outr).

Exemple 1. L'expression régulière générique d'un PEAK est la suivante (4.2.3.1) : '< (= | <)* (> | =)*>'. On a sur le graphique (*Figure 4-8*) une occurrence du pattern PEAK, car cette séquence d'expression spécifique < (>)* respecte l'expression générique (4.2.3.1).

Exemple 2. L'expression régulière générique d'une GORGE est la suivante (4.2.3.2) : (>| (> (= | >)* >)) (< | < (= | <)* <). On a sur le graphique (*Figure 4-9*) une occurrence d'une GORGE, car cette séquence d'expression spécifique (>) (<) respecte l'expression générique (4.2.3.2).

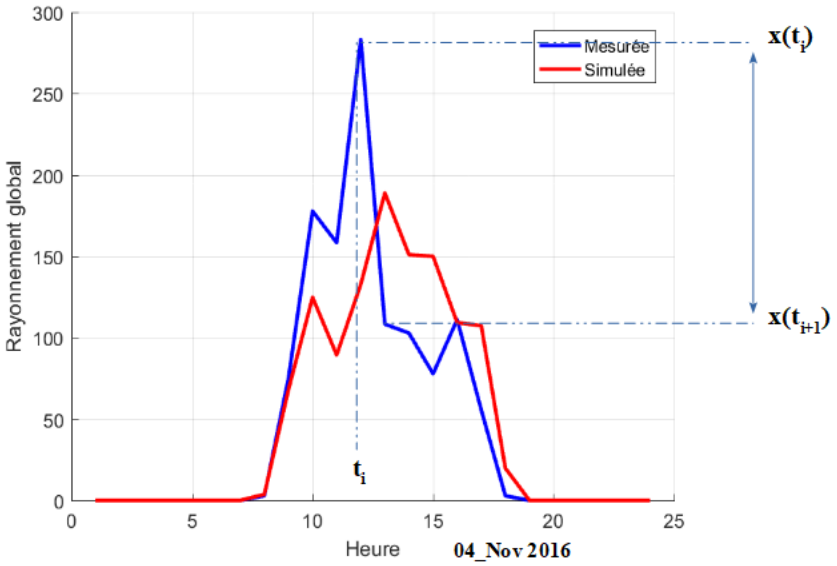


Figure 4-8 : Exemple de pattern PEAK avec le rayonnement global

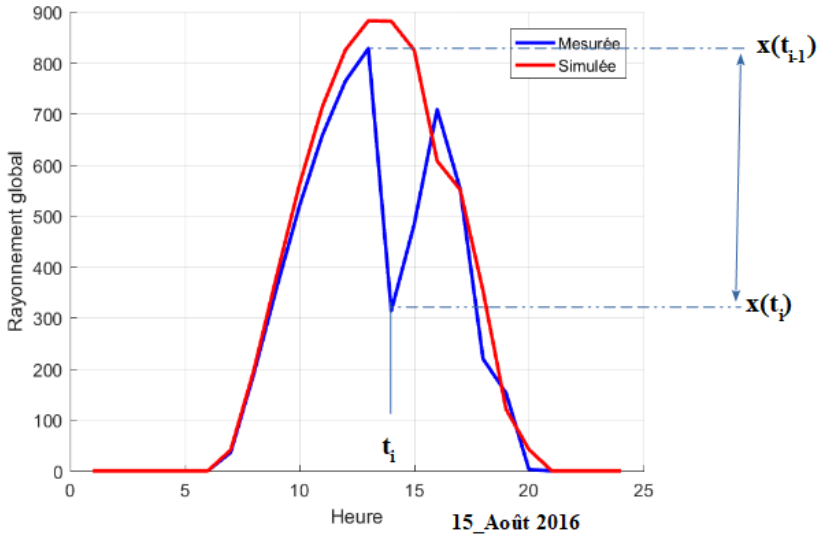


Figure 4-9 : Exemple de pattern GORGE avec le rayonnement global

Un classifieur automatique a été créé en utilisant les Automates d'états finis pour les patterns PEAK et GORGE des contraintes de séries temporelles. Ce classifieur permet de détecter le jour et l'heure de la survenue d'un PEAK ou d'une GORGE. Un algorithme de classification est proposé ci-dessous (*Tableau 22*).

Tableau 22 : Algorithme de classification des types d'erreurs

1. Début
2. Xj=(x_1, x_2, …, x_n) est la série mesurée le jour j et C_j son graphe.
3. Erreur1 = types d'erreur de formes géométriques similaires
4. Erreur2 = types d'erreur de formes géométriques similaires (ou non similaires) avec un pic (PEAK)

5. Erreur3 = types d'erreur de formes géométriques avec présence de bruits (cas à exclure)
6. Seuil1 et seuil2 sont définis en fonction de l'erreur maximale avec l'expert du domaine.
7. Pour j = 1 à N (chaque jour)
8. Pour i = 1 à n (chaque heure)
9. Si max ($|x_{i+1} - x_i|$,$|x_i - x_{i-1}|$) > seuil2 Faire
10. C_j inclus dans Error3
11. Sinon si max ($|x_{i+1} - x_i|$,$|x_i - x_{i-1}|$) ∈ [seuil1, seuil2] Faire
12. C_j inclus dans Error2
13. Sinon C_j inclus dans Error1
14. Fin Si
15. Fin Pour
16. Fin Pour
17. Retour Error1, Error2, Error3
18. Fin

Un exemple de cas d'étude de ce classifieur est présenté dans la partie expérimentale sur les grandeurs suivantes : la Chaleur sensible (H), la Chaleur latente (LE) et le rayonnement global (glo).

4.2.4 Processus gaussiens pour l'apprentissage automatique (GPML)

Dans cette section, nous allons présenter quelques propriétés sur les processus gaussiens dans la modélisation des données.

4.2.4.1. Modélisation avec le GPML

Les modèles GP sont des modèles probabilistes, non paramétriques basées sur les principes de probabilité bayésienne. GPs peuvent être considérées comme des méthodes du noyau avec une interprétation bayésienne (Rasmussen & Williams, 2005). Un modèle GP ne se rapproche pas du système modélisé en ajustant les paramètres des fonctions de base choisies, mais implique une relation entre les données mesurées. L'utilisation des propriétés et des modèles GP pour la modélisation sont minutieusement décrits dans (Kocijan, 2016), (J. Q. Shi & Choi, 2011), (Rasmussen & Williams, 2008), (Rasmussen & Williams, 2005).

Les modèles GP peuvent être utilisés pour la régression (Sharifzadeh et al., 2019), où la tâche est de déterminer un mappage à partir d'un ensemble de vecteurs de régression N D-dimensions représenté par la matrice de régression $X = [x_1, x_2, \dots, x_N]^T$ vers un vecteur de données de sortie $y = [y_1, y_2, \dots, y_N]^T$. Les sorties sont généralement considérées comme des réalisations bruyantes d'une fonction $f(x_i)$. Un modèle GP suppose que la sortie est une réalisation d'un GP avec une fonction de densité de probabilité conjointe :

$$p(y) = \mathcal{N}(m, K), \qquad (4.2.4.1)$$

avec la moyenne m et la matrice de covariance $K = [K_{ij}]$ en fonctions des entrées x. Habituellement, la fonction moyenne est définie comme étant égale à 0, tandis que la fonction de covariance ou le noyau $K_{ij} = \mathcal{C}(x_i, x_j)$ définit les caractéristiques du processus à modéliser, c.-à-d. la stationnarité statistique, la douceur, etc.

La valeur de la fonction de covariance $\mathcal{C}(x_i, x_j)$ exprime la corrélation entre les différentes sorties f(x_i) et f(x_j) par rapport aux entrées x_i et x_j. En supposant que les données statistiquement stationnaires sont contaminées par le bruit blanc, la fonction de covariance la plus couramment utilisée est la composition de la fonction de covariance exponentielle carrée (SE) avec « détermination automatique de la pertinence » (ARD) hyperparamètres et fonction de covariance constante en supposant un bruit blanc (Kocijan, 2016) :

$$\mathcal{C}(x_i, x_j) = \sigma_f^2 exp\left[-\frac{1}{2}(x_i - x_j)^T \Lambda^{-1}(x_i - x_j)\right] + \delta_{ij}\sigma_n^2, \qquad (4.2.4.2)$$

où Λ^{-1} est une matrice diagonale $\Lambda^{-1} = \text{diag}(l_1^{-2}, \dots, l_D^{-2})$ des hyperparamètres ARD, σ_f^2 et σ_n^2 sont des hyperparamètres de la fonction de covariance, et $\delta_{ij} = 1$ si $i = j$ et 0 sinon. Les hyperparamètres peuvent être écrits comme un vecteur $= \left[l_1^{-2}, \dots, l_D^{-2}, \sigma_f^2, \sigma_n^2\right]^T$. La propriété ARD signifie que l_i^{-2} (i = 1, … , D) indique l'importance des entrées individuelles. Si l_i^{-2} est égal ou proche de zéro, cela signifie que les entrées dans la dimension i ne contiennent que peu d'informations et pourraient éventuellement être éliminées. D'autres fonctions de covariance adaptées à diverses applications peuvent être trouvées dans, par exemple, (Kocijan, 2016).

L'objectif commun de la régression est de prédire la sortie y dans un lieu d'essai non observé x compte tenu des données d'entraînement, d'une fonction moyenne connue et d'une fonction de covariance connue C. La distribution prédictive de la sortie peut être obtenue en utilisant la règle de Bayes. L'effet des hyperparamètres inconnus θ doit être pris en compte. Ceci conduit à une tâche de calcul très exigeante, parfois insoluble. Une solution approximative fréquemment utilisée pour le problème du calcul consiste à estimer les hyperparamètres en maximisant la probabilité marginale de la règle de Bayes. Les détails de l'inférence des hyperparamètres se trouvent dans (Rasmussen & Williams, 2005), (Kocijan, 2016).

Une fois que les valeurs des hyperparamètres sont obtenues, la distribution normale prédictive de la sortie pour une nouvelle entrée d'essai peut être calculée en utilisant

$$\mu(y^*) = k(x^*)^T K^{-1} y, \qquad (4.2.4.3)$$

$$\sigma^2(y^*) = \kappa(x^*) - k(x^*)^T K^{-1} k(x^*), \qquad (4.2.4.4)$$

où $k(x^*) = [C(x_1, x^*), \dots, C(x_N, x^*)]^T$ est le $N \times 1$ vecteurs de covariance entre l'essai et les cas d'entraînement, et $\kappa(x^*) = C(x^*, x^*)$ est la covariance entre l'entrée d'essai elle-même.

Une prévision du modèle GP, en plus de la valeur moyenne (4.2.4.3), fournit également de l'information sur la fiabilité de la prévision à l'aide de la variance de prévision (4.2.4.4). Habituellement, la confiance dans la prédiction est interprétée avec un intervalle de 2σ. Cet intervalle de confiance met en évidence les zones de l'espace d'entrée où la

qualité des prévisions est mauvaise, en raison du manque de données ou de données bruyantes, en indiquant un intervalle de confiance plus large autour de la moyenne prévue.

4.2.4.2 Modèle de variable latente du processus gaussien

Un type de modèle GP est le Modèle de Variable Latente du Processus Gaussien dénommé Gaussian Process Latent Variable Model (GP-LVM). Le GP-LVM est connu à l'origine comme une méthode de réduction de la dimensionnalité qui s'est avérée très robuste dans les applications de données volumineuses (A. C. Damianou et al., 2016), (Lawrence, 2006). Historiquement, le GP-LVM a été introduit pour les besoins de l'apprentissage non supervisé (A. Damianou, 2015) ; cependant, la transition vers l'apprentissage supervisé exige un effort minimum d'application de la croyance préalable à l'entrée.

L'objectif principal d'un modèle GP-LVM dans les deux cas d'apprentissage, outre la réduction de la dimensionnalité, est de trouver une relation mathématique entre le vecteur d'entrée de haute dimension $X \in \mathbb{R}^{N\times D}$ et l'espace dimensionnel inférieur des variables latentes $Y \in \mathbb{R}^{N\times P}$. Le principal défi ici est le vecteur X non observé des valeurs d'entrée dans le cas de l'apprentissage non supervisé. Le modèle GP-LVM offre une solution élégante au défi en traitant le vecteur d'entrée comme des variables latentes et en déployant simultanément des GPs indépendants de P comme croyance préalable (Lawrence, 2006) : $f(X) = (f_1(X), \dots, f_P(X))$ telle que

$$f_j(X) \sim GP\big(0, C(X,X)\big), \qquad j = 1, \dots, P. \tag{4.2.4.5}$$

La difficulté de la méthodologie bayésienne est la propagation de la probabilité de la croyance antérieure p(X) par la fonction de correspondance non linéaire f.

L'algorithme d'optimisation nécessite le calcul de la probabilité conjointe

$$p(y) = \int p(y|f)(p(f|X)p(X)dX)df, \tag{4.2.4.6}$$

où y est un vecteur de cibles, qui correspond, par exemple, à un modèle de régression avec une seule sortie. Comme il s'avère que, les entrées X de la matrice K du noyau sont contenues dans la probabilité conjointe (4.2.4.6) d'une manière non linéaire très complexe, laissant l'intégration sur le domaine X dans la plupart des cas insoluble. Pour éviter cette difficulté de traitement, on utilise une méthodologie standard de variation bayésienne pour estimer la probabilité marginale de p(y) avec une limite inférieure de variation (A. Damianou, 2015). Cette probabilité marginale p(y) sert à propager toutes les incertitudes liées aux données d'entrée.

4.2.4.3 Processus gaussien profond (Deep-GP)

Le Deep-GP a été introduit comme une approche non paramétrique souple de l'apprentissage profond (A. Damianou & Lawrence, 2013), (A. Damianou, 2015), (A. C. Damianou et al., 2016). Le Deep-GP se compose de L couches cachées de variables latentes h_L. Le Deep-GP possède trois types de nœuds comme illustré sur le schéma suivant (*Figure 4-10*) :

- Nœuds Feuilles : $Y \in \mathbb{R}^{N\times D}$ (observée) ;
- Espace Latent Intermédiaire : $X_h \in \mathbb{R}^{N\times Q_h}$ *h=1, … ,L-1, où L* est le nombre de couches cachées ;

- Nœud Parent latent : $Z = X_L \in \mathbb{R}^{N \times Q_z}$ (peut ne pas être observée et potentiellement contraint avec apriori).

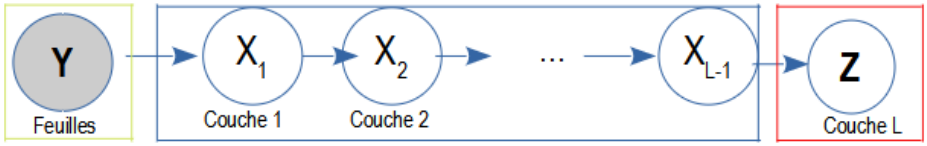


Figure 4-10 : Schéma des trois nœuds du Deep-GP

Les processus gaussiens régissent les mappages entre les couches. En termes simples, un modèle de GP profond est constitué de GP imbriqués, dans notre cas de GP-LVM, où les sorties d'un GP sont traités comme des entrées d'un autre GP. Le schéma suivant (*Figure 4-11*) représente un exemple de Deep-GP avec deux couches cachées.

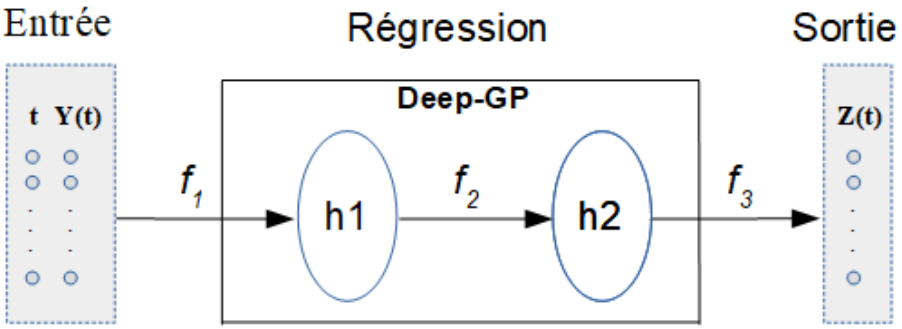


Figure 4-11 : Modélisation de données par le DeepGP avec deux couches cachées

La relation suivante (4.2.4.7) est la formule générale du Deep-GP

$$Z = f_{1:L} + \epsilon = f_L\left(f_{L-1}(\dots f_1(Y))\right) + \epsilon, \qquad (4.2.4.7)$$

où chaque f_i est un modèle de GP indépendant et le nombre L de couches cachées, aussi appelé la profondeur d'un modèle d'apprentissage profond ; les variables *t* et *Y(t)* désignent les données d'entraînement du modèle Deep GP ; *Z* désigne les valeurs estimées en sortie (*Figure 4-11*). Le terme ε désigne le bruit blanc avec une distribution normale qui est ajouté aux sorties de chaque couche (A. Damianou, 2015). La distribution conjointe d'un modèle GP profond avec des couches cachées en L peut être écrit comme (4.2.4.8), pour chaque ensemble de variables latentes $h_l,\ l = 1, \dots, L$.

$$P(y, \{h_l\}_{l=1}^{L}) = P(y/h_l)P(h_L/h_{L-1}) \dots P(h_2/h_1)P(h_1), \qquad (4.2.4.8)$$

où les $(h_l)_{l=1,\dots,L}$, peuvent être interprétées comme les représentations *latente*s des données. Dans cette étude, la modélisation des processus gaussiens est poursuivie avec des logiciels.

4.3 Cas d'étude sur les données du rayonnement global

Cette partie explique d'abord les types d'erreurs (différences) qui existent entre les deux bases de données, notamment la base de données observées (mesurées) et la base de données simulées ; ensuite, les résultats d'un cas d'étude suivi d'une discussion sont présentées. Les mesures/simulations des variables disponibles qui peuvent être utilisées pour la modélisation sont : la chaleur sensible (H), la chaleur latente (LE), la température (tt), l'humidité relative (hu), le rayonnement solaire global (glo), la vitesse du vent (ff), etc. Ces variables sont présentées dans le chapitre 2. Parmi ces variables mesurées sur l'année 2016, le biais est important dans les simulations des flux de chaleur sensible (H) et de chaleur latente (LE). Ceci est souvent dû à une perturbation naturelle (emplacement des nuages) du rayonnement global.

Dans les calendriers journaliers des variables sélectionnées ci-dessus, les formes géométriques rencontrées sont classifiées en trois catégories : formes géométriques similaires ; formes géométriques similaires (ou non similaires) avec pics ; formes géométriques avec présence de bruits (valeurs aberrantes).

Cas 1 (Erreur1**)**. Formes géométriques similaires : la forme graphique des valeurs mesurées est semblable à la forme graphique des valeurs simulées. Dans ce cas, la minimisation de la fonction coût (fonction d'erreur) atteint l'optimal à 100%.

Cas 2 (Erreur2**)**. Formes géométriques similaires/non similaires avec pics : la forme graphique des valeurs mesurées n'est pas forcement semblable à la forme graphique des valeurs simulées. En plus, il y a souvent la présence de pics sur la forme graphique des valeurs mesurées. La correction du modèle numérique aux niveaux des pics peut engendrer la maximisation de la fonction coût. Dans ce cas, la minimisation de la fonction coût (fonction d'erreur) atteint l'optimal à 98%.

Cas 3 (Erreur3**)**. Formes géométriques avec présence de bruits : les formes graphiques des valeurs mesurées avec présence de bruits (valeurs écartées) sont des cas à exclure.

L'utilisation de l'algorithme (*Tableau 22*) donne une classification des erreurs en trois catégories sur les grandeurs suivantes (*Tableau 23*) : la Chaleur sensible (H), la Chaleur latente (LE) et le rayonnement global (glo).

.

Tableau 23 : Classification des types d'erreurs sur le rayonnement global (glo), les flux de Chaleur sensible (H) et de Chaleur latente (LE)

Mois/2016	Erreur1			Erreur2			Erreur3		
	H	LE	glo	H	LE	glo	H	LE	glo
Janvier	16	14	30	12	10	2	3	7	0
Février	17	13	23	9	11	6	3	5	0
Mars	15	23	23	14	5	8	2	3	0
Avril	11	20	25	17	9	5	2	1	0
Mai	18	17	23	12	13	8	1	1	0
Juin	11	15	25	18	13	5	1	2	0
Juillet	8	17	24	18	10	7	5	4	0
Août	12	19	27	12	8	4	7	4	0
Septembre	15	15	26	13	14	4	2	1	0
Octobre	11	17	26	17	13	5	3	1	0
Novembre	12	13	22	12	16	8	6	1	0
Décembre	19	20	33	11	10	1	1	1	0

Remarques : La différence entre le rayonnement global mesuré et simulé a des origines plurielles, et peut par exemple être due à une position des nuages dans les modèles météo qui n'est pas exactement celle de la réalité. C'est-à-dire que le passage des nuages au-dessus des capteurs peut engendrer des fluctuations des mesures. Ces mesures qui fluctuent ou qui ont de grandes variations ne sont pas nécessairement fausses puisque la nature n'est pas toute lisse avec des mécanismes intrinsèquement complexes. Il peut éventuellement y avoir des différences entre certains emplacements des nuages dans la simulation par rapport à la réalité. De telles différences d'emplacements peuvent être causes de la présence des biais importants sur le rayonnement global, mais aussi sur les flux de chaleur sensible et de chaleur latente. Ainsi, par la suite, le rayonnement global sera utilisé pour tester l'approche proposée.

4.3.1 Résultats

Cette section présente les résultats de l'approche proposée pour l'amélioration de la fiabilité du modèle numérique de prévision dans la simulation du rayonnement global (glo) :

4.3.1.1 Cas 1. Formes géométriques similaires

Le graphique suivant (*Figure 4-12*) illustre la modélisation (régression avec le GPML) du rayonnement global mesuré et le transport des valeurs simulées vers le voisinage des valeurs mesurées en vue de minimiser la fonction coût (erreur sur 24 heures).

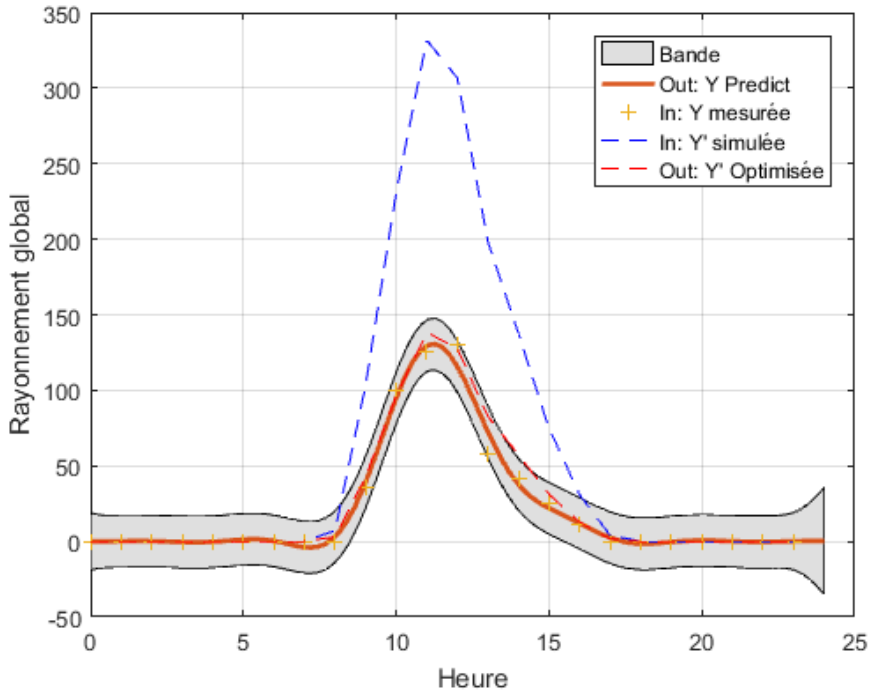


Figure 4-12 : Modélisation par régression et minimisation des erreurs de simulation du rayonnement global (cas 1)

L'implémentation a été faite avec le logiciel MATLAB. La bande grise est la surface qui canalise les données prises en compte pour la modélisation. Le prédicteur g est la courbe continue en rouge qui représente la modélisation des valeurs mesurées du rayonnement global (en croix orange). La courbe discontinue en bleue représente les valeurs simulées à l'entrée de l'approche proposée. La courbe discontinue en rouge est la sortie de l'approche proposée. Cette courbe est obtenue en effectuant un transport simultané par homothétie des valeurs simulées vers les voisinages des valeurs mesurées. Ce transport a été fait en jouant avec les paramètres de l'optimiseur (correcteur) f. Le graphique suivant (*Figure 4-13*) illustre la minimisation de la fonction coût (erreur) sur le rayonnement global.

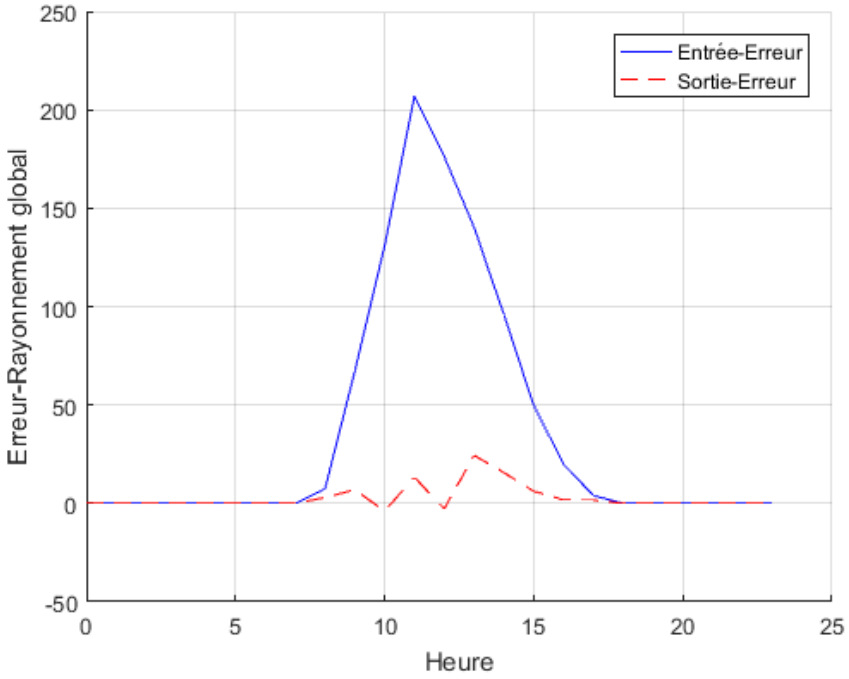


Figure 4-13 : Comparaison des fonctions coût sur le rayonnement global en entrée et sortie de l'approche proposée (cas 1)

La fonction coût (distance sur 24 heures) à l'entrée vaut d = 357 w/m² et celle de la sortie vaut 34 w/m², avec $(\alpha,\ \beta, \gamma) = (1, 1, 2.4)$.

4.3.1.2 Cas 2. Formes géométriques similaires/non similaires avec pics

Le graphique suivant (*Figure 4-14*) illustre la modélisation du rayonnement global mesuré et le transport des valeurs simulées vers le voisinage des valeurs mesurées.

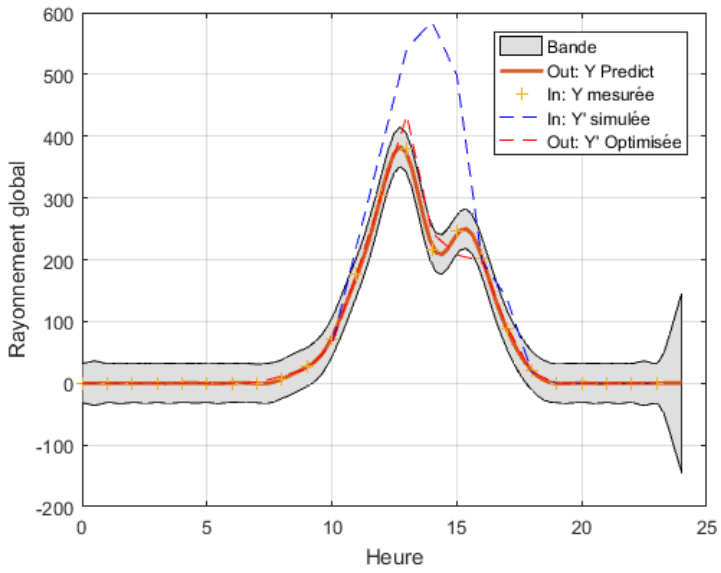


Figure 4-14 : Modélisation par régression et minimisation des erreurs de simulation du rayonnement global (cas 2)

La courbe discontinue en bleue représente les valeurs simulées à l'entrée de l'approche proposée, les points en croix orange représentent les valeurs mesurées à l'entrée. La courbe continue en rouge représente la modélisation des valeurs mesurées du rayonnement global. La courbe discontinue en rouge est la sortie de l'approche proposée. Cette courbe est obtenue en effectuant un transport simultané par homothétie des valeurs simulées vers les valeurs mesurées. Ce transport a été fait en jouant avec les paramètres de l'optimiseur (correcteur) f . Le graphique suivant (*Figure 4-15*) illustre la minimisation de la fonction coût (erreur) sur le rayonnement global.

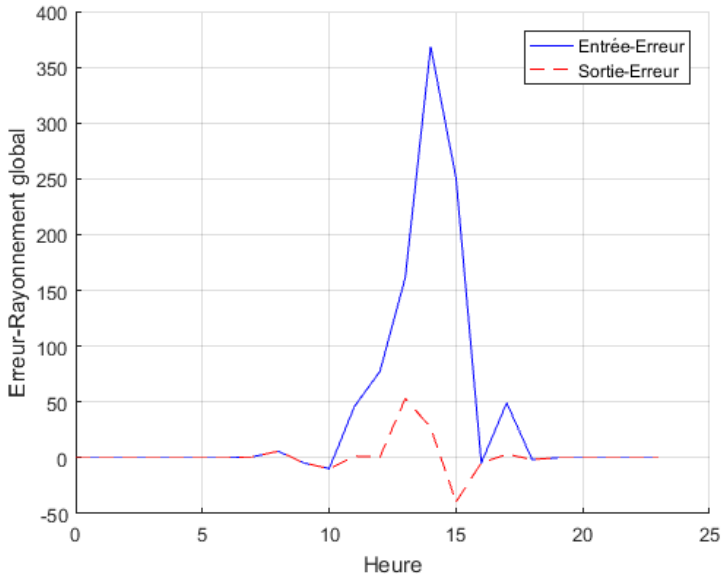


Figure 4-15 : Comparaison des fonctions coût sur le rayonnement global en entrée et sortie de l'approche proposée (cas 2)

La fonction coût (distance sur 24 heures) à l'entrée vaut d = 485 w/m² et celle de la sortie vaut 73 w/m², avec

$$(\alpha,\ \beta, \gamma) = \begin{cases} (1, 2, 2.5) \text{ sur } [0, 14] \\ (1, 1, 2.4) \text{ sur } [15, 23] \end{cases}$$

En sortie de l'approche, la fonction coût est minimale. En effet, dans le cas des formes géométriques similaires, 100% des erreurs horaires $((d_i)_{i=0,\dots,23})$ sont dans l'intervalle d'optimalité du rayonnement global $[-20$ w/m², $+20$ w/m²$]$ défini par l'expert du domaine (*Figure 4-13*) ; et dans le cas des formes géométriques similaires/non similaires avec pics, 98% des erreurs horaires $((d_i)_{i=0,\dots,23})$ sont quasiment dans l'intervalle d'optimalité du rayonnement global $[-20$ w/m², $+20$ w/m²$]$ (*Figure 4-15*).

La comparaison de la fonction coût à l'entrée et sortie de l'approche est effectuée sur 5 jours successifs de simulation :

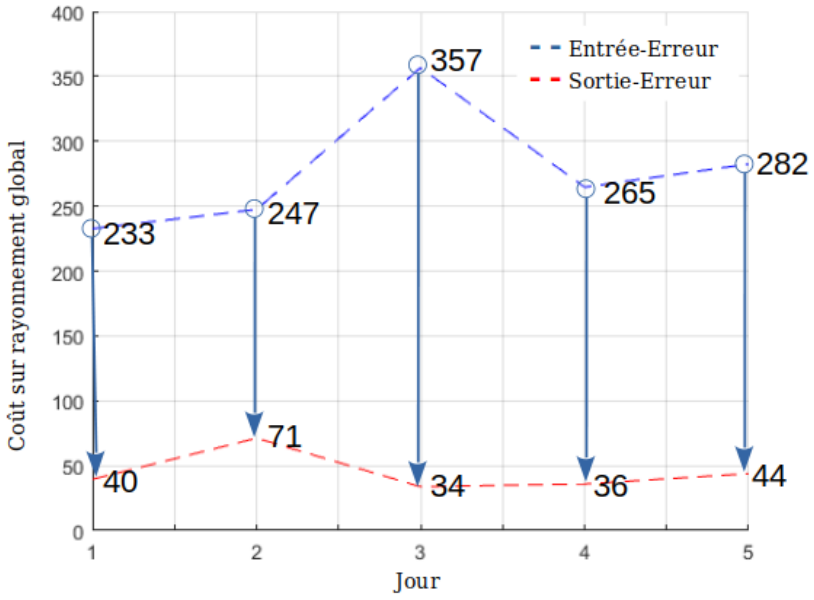


Figure 4-16 : Comparaison de la fonction coût sur le rayonnement global en entrée et sortie de l'approche proposée pour 5 jours successifs en hiver

En 5 jours successifs de simulation, la fonction coût sur le rayonnement global a été réduite d'environ 84% (*Figure 4-16*).

4.3.2 Discussion sur les résultats obtenus

Le transport optimal classique détecte l'optimal d'une fonction coût dans le transport d'une distribution A (*source*) vers une distribution B (*cible*) (Courty et al., 2015), (Lévy & Schwindt, 2018), (Dieci & Walsh III, 2019). Il existe plusieurs possibilités de déplacement des éléments de A vers les éléments de B. Autrement dit, le Transport Optimal original (version de Gaspard Monge) transforme une application surjective en une application bijective avec la contrainte de non fractionnement des masses. En effet, l'étape finale du transport optimal consiste à conserver la distance minimale dans chaque groupe de distances calculées. Cette finalité est le point de départ de notre approche avec le rajout d'un optimiseur pour réduire au mieux les distances optimales.

Notre cas d'étude consiste à transporter, par correspondance, des valeurs réels d'un domaine A (*valeurs simulées*) vers les voisinages des valeurs réels d'un domaine B (*valeurs mesurées*). Bien que le transport optimal soit adapté à cette problématique, il existe des situations dans lesquelles l'intuition peut se révéler avantageuse. Par exemple, dans notre cas l'impression de l'ensemble de données sur un poster a permis d'identifier la linéarité des types d'erreur d'un seul coup d'œil. L'approche proposée a bien minimisé la fonction coût (erreurs) avec environ 84% de réduction d'erreurs sur le rayonnement global.

4.4 Conclusion

Ce chapitre a proposé une méthodologie pour le diagnostic et de minimisation d'erreur d'un modèle de prévisions numériques du temps. Cette méthodologie comprend deux étapes : (i) modélisation (par régression) des données mesurées pour définir un prédicteur de diagnostic, (ii) un optimiseur est défini et associé au modèle pour la réduction des erreurs de simulation.

Particulièrement, dans l'étude de cas cible, l'incertitude est souvent dynamique dans les données météorologiques mesurées. De ce fait, le processus gaussien pour l'apprentissage automatique (GPML), qui prend en compte l'incertitude, est utilisé pour modéliser les données mesurées. Ceci offre l'opportunité de définir un prédicteur qui facilite le diagnostic du modèle numérique de prévision. Ensuite, un optimiseur (correcteur) a été défini à partir de quelques propriétés sur les transformations géométriques en mathématiques. Cet optimiseur est associé au modèle pour effectuer un transport simultané par homothétie des valeurs simulées vers les valeurs mesurées. Ceci fournit donc un moyen astucieux pour minimiser les erreurs de simulation.

En résumé, après la simulation du rayonnement global par le modèle numérique de prévision, l'approche proposée dans cette étude nous a permis de comprendre, d'identifier les anomalies du modèle et apporter des corrections sur les erreurs de simulation. Dans tous les cas qui ont été traités, l'approche proposée a bien marché et a donné des résultats satisfaisants. En 5 jours successifs de simulation du rayonnement global, il y a eu une réduction d'environ 84% d'erreurs sur le rayonnement global (*Figure 4-16*). Pour toutes les variations saisonnières dans l'année, cette approche peut être appliquée à toutes les autres variables météorologiques pour encore mieux améliorer la fiabilité du modèle numérique de prévision.

L'exigence de précision du modèle est très élevée dans la simulation des variables météorologiques. La présente approche a pu minimiser avec satisfaction les différences. Une perspective de ce travail pourrait être d'approfondir l'utilisation de l'optimiseur (correcteur) pour tester l'influence du type d'erreur sur l'efficacité du simulateur.

Conclusion et perspectives

Contributions

Cette thèse a proposé deux méthodologies progressives d'évaluation de la simulation des processus de surface par les modèles numériques de prévisions du temps.

La première méthodologie comprend trois étapes : (i) prétraitement avec le nettoyage de données, et des techniques de réduction (par exemple discrétisation pour déterminer les items pertinents avec une transformation des données quantitatives en des données qualitatives), (ii) exploration des erreurs pour découvrir des relations intéressantes entre les erreurs sur les variables considérées, et (iii) post-traitement avec l'analyse et l'interprétation des règles générées pour comprendre et identifier les paramètres qui influent sur les grosses erreurs dans la simulation de flux de chaleur sensible et chaleur latente.

La seconde méthodologie consiste à diagnostiquer et minimiser les erreurs importantes dans la simulation des paramètres météorologiques. Cette méthodologie comprend deux étapes : (i) modélisation (par régression) des données mesurées pour définir un prédicteur de diagnostic, (ii) utilisation du transport optimal pour minimiser les erreurs de simulation ; pour cela, un optimiseur a été défini, à partir de quelques transformations géométriques élémentaires en mathématiques, pour transporter les données simulées (source) vers les voisinages des données mesurées (cible).

Résultats

Dans l'étude de cas cible, les données météorologiques ne sont pas parfaites. Dans le cadre de la gestion des valeurs manquantes, la méthode d'imputation *k*-NN a des inconvénients face aux valeurs manquantes monotones. Cela impacte la qualité de la discrétisation en affectant les résultats globaux (règles d'association générées).

Pour minimiser les inconvénients de *k*-NN face aux valeurs manquantes monotones, la méthode moyenne mobile a été utilisée pour remplacer les valeurs manquantes monotones. Ensuite, un ajustement a été réalisé avec les modèles autorégressifs, intégrés et à moyenne mobile pour remplacer des valeurs manquantes et réduire les fluctuations irrégulières (intégrant éventuellement des accidents de mesures).

En plus, la structure de données météorologiques n'est pas compatible avec l'algorithme utilisé (notamment FPGROWTH). De ce faite, une approche de discrétisation a été adoptée en prétraitement pour transformer les données originales sous forme d'intervalles (items) codés et pour obtenir en fin un format qui pourrait être importé dans le traitement de données. Cette transformation permet de réduire l'espace de recherche et le temps d'exécution de l'algorithme. En plus, quelques notions des suites numériques en

mathématiques ont été utilisées pour déterminer les bornes des intervalles. Cela nous a permis de faire une répartition pertinente des échantillons entre les intervalles.

Par exemple, avec un couple de mesures (min sup=225, min conf=50%) choisi, l'exploration de la base des erreurs a généré 45 règles d'associations rares pour la saison printemps (*Tableau 15*). Vu le nombre important de règles générées, des propriétés mathématiques (*Figure 3-4*) sur la théorie des ensembles ont été utilisées pour réduire le nombre de règles générées et pour faciliter la visualisation et l'interprétation des résultats. Le graphique (*Figure 3-9*) résume la visualisation et l'interprétation sémantique des règles générées sur toutes les variations saisonnières (printemps, été, automne et hiver) en 2016. Dans ce cas d'étude, l'approche utilisée nous a permis de comprendre le comportement des autres paramètres associés et d'identifier le plus influent avec précision.

En résumé, le résultat montre qu'un dysfonctionnement du rayonnement global, avec une fiabilité (max sup=225) et une précision (max conf=100%), est souvent dû à des perturbations naturelles (par exemple, le passage des nuages au-dessus des capteurs). Cela impacte la qualité des observations/simulations des flux de chaleur sensible et chaleur latente (*Figure 3-9*).

La seconde approche proposée nous a permis de mieux comprendre, de mettre en évidence les faiblesses du modèle et d'apporter des corrections sur les erreurs de simulation. Dans tous les cas qui ont été traités, l'approche proposée a été opportune et a donné des résultats satisfaisants. En quelques jours successifs de simulation du rayonnement global, il y a eu une réduction d'environ 84% d'erreurs sur le rayonnement global. Cette approche peut être appliquée à toutes les autres variables météorologiques considérées.

Perspectives

L'incertitude est très dynamique sur les mesures des grandeurs météorologiques (Jancic et al., 2018) ; ceci impacte la qualité de la simulation des processus de surface par le modèle numérique de prévision. C'est ce qui explique la difficulté de faire des prévisions météorologiques fiables au-delà de quelques jours. Il s'agit en effet de résoudre des systèmes d'équations différentielles ordinaires/partielles, dont la condition initiale n'est connue qu'avec une marge d'erreur : les incertitudes de mesure sur température, pression, vitesse du vent, etc.

La présente approche a pu minimiser avec satisfaction les erreurs de simulation. Alors, des perspectives de ce travail pourront être les suivantes :

1. L'investigation sur plusieurs années pour valider des hypothèses sur la date de survenues des erreurs ; en outre, il faudrait davantage éprouver l'utilisation de l'optimiseur (correcteur) pour tester l'influence de la nature des erreurs traitées sur l'efficacité des corrections du simulateur ; on devrait aussi favoriser l'amélioration de la compréhension des principes sous-jacents aux équations

différentielles gouvernant les phénomènes météorologiques à l'aide des optimiseurs.

2. L’examen plus détaillé des équations différentielles mérite également un intérêt dans la simulation numérique des paramètres météorologiques. Des équations différentielles gouvernant les phénomènes météorologiques, sont résolues par des méthodes d’analyse numérique comme par exemple la méthode des différences finies. Nous pouvons aussi utiliser la théorie des perturbations pour la résolution de ces équations différentielles avec l’ajout des paramètres de perturbations aux coefficients des équations considérées. Ces paramètres pourront prendre en compte des perturbations naturelles (par exemple, le passage des nuages au-dessus des capteurs).

Bibliographie

Abaza, M., Anctil, F., Fortin, V., & Perreault, L. (2017). On the incidence of meteorological and hydrological processors : Effect of resolution, sharpness and reliability of hydrological ensemble forecasts. *Journal of Hydrology*, *555*, 371–384.

Abdelfatah, K., Bao, J., & Terejanu, G. (2018). Geospatial uncertainty modeling using Stacked Gaussian Processes. *Environmental Modelling and Software*, *109*, 293–305.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large database. *Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data, ACM Press*, 207–216.

Aguera-Pérez, A., Palomares-Salas, J. C., González de la Rosa, J. J., & Florencias-Oliveros, O. (2018). Weather forecasts for microgrid energy management : Review, discussion and recommendations. *Applied Energy*, *228*, 265–278.

Ajak, A. D., Lilford, E., & Topal, E. (2017). Application of predictive data mining to create mine plan flexibility in the face of geological uncertainty. *Resources Policy*.

Apiletti, D., Baralis, E., Cerquitelli, T., Garza, P., Pulvirenti, F., & Venturini, L. (2017). Frequent Itemsets Mining for Big Data : A Comparative Analysis. *Big Data Research*, *9*, 67–83.

Arafailova, E. (2018). Functional description of sequence constraints and synthesis of combinatorial objects. *Discrete Mathematics [cs.DM]*. *Ecole nationale supérieure Mines-Télécom Atlantique, 2018IMTA0089*, HAL Id : tel-01962957, version 1.

Arafailova, E., Beldiceanu, N., Douence, R., Carlsson, M., Flener, P., Pearson, J., Rodriguez, M. A. F., & Simonis, H. (2018). Global Constraint Catalog : Time-Series Constraints. *IMT Atlantique (LS2N-CNRS), Nantes, France ; RISE SICS, Kista, Sweden ; Uppsala University, Uppsala, Sweden ; School of Computing, National University of Singapore ; Insight Centre for Data Analytics, University College Cork, Ireland*, *Volume II*.

Arnaud, P., Cantet, P., & Odry, J. (2017). Uncertainties of flood frequency estimation approaches based on continuous simulation using data resampling. *Journal of Hydrology*, *554*, 360–369.

Azimi, R., Ghofrani, M., & Ghayekhloo, M. (2016). A hybrid wind power forecasting model based on data mining and wavelets analysis. *Energy Conversion and Management*, *127*, 208–225.

Bandaru, S., Ng, A. H. C., & Deb, K. (2017). Data mining methods for knowledge discovery in multi-objective optimization : Part A - Survey. *Expert Systems with Applications*, *70*, 139–159.

Bernard, F., Iollo, A., & Riffaud, S. (2018). Reduced-order model for the BGK equation based on POD and optimal transport. *Journal of Computational Physics*, *373*, 545–570.

Blanchard, J., Pinaud, B., Kuntz, P., & Guillet, F. (2007). A 2D–3D visualization support for human-centered rule mining. *Computers & Graphics*, *31*, 350–360.

Bibliographie

Borgogno, O., & Colangelo, G. (2019). Data sharing and interoperability : Fostering innovation and competition through APIs. *Computer Law and Security Review*, *35*, 105314.

Bouker, S. (2012). Ranking and selecting association rules based on dominance relationship. *LIMOS - Laboratoire d'Informatique, de Modélisation et d'optimisation des Systèmes*.

Bugata, P., & Drotar, P. (2019). Weighted nearest neighbors feature selection. *Knowledge-Based Systems*, *163*, 749–761.

Chang, V. (2017). Towards data analysis for weather cloud computing. *Knowledge-Based Systems*, *127*, 29–45.

Chemchem, A., & Drias, H. (2015). From data mining to knowledge mining : Application to intelligent agents. *Expert Systems with Applications*, *42*, 1436–1445.

Chen, W., Xie, C., Shang, P., & Peng, Q. (2017). Visual analysis of user-driven association rule mining. *Journal of Visual Languages Computing*, *42*, 76–85.

Chizat, L., Peyré, G., Schmitzer, B., & Vialard, F.-X. (2018). Unbalanced optimal transport : Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, *274*(11), 3090–3123.

Colak, I., Sagiroglu, S., Demirtas, M., & Yesilbudak, M. (2013). A data mining approach : Analyzing wind speed and insolation period data in Turkey for installations of wind and solar power plants. *Energy Conversion and Management*, *65*, 185–197.

Corbari, C., Salerno, R., Ceppi, A., Telesca, V., & Mancini, M. (2019). Smart irrigation forecast using satellite LANDSAT data and meteo-hydrological modeling. *Agricultural Water Management*, *212*, 283–294.

Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2015). Optimal Transport for Domain Adaptation. *arXiv :1507.00504 [cs]*, *v2*.

Damianou, A. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty. *University of Sheffield*, *phd-thesis*.

Damianou, A. C., Titsias, M. K., & Lawrence, N. D. (2016). Variational inference for latent variables and uncertain inputs in Gaussian processes. *Machine Learning Research*, *17*, 1–62.

Damianou, A., & Lawrence, N. (2013). Deep Gaussian Processes. *Artificial Intelligence and Statistics*, 207–215.

Damodaran, B. B., Flamary, R., Seguy, V., & Courty, N. (2020). An Entropic Optimal Transport loss for learning deep neural networks under label noise in remote sensing images. *Computer Vision and Image Understanding*, *191*, 102863.

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for Big Data professions : A systematic classification of job roles and required skill sets. *Information Processing and Management*, *54*, 807–817.

Di Ciccio, C., Maggi, F. M., Montali, M., & Mendling, J. (2017). Resolving inconsistencies and redundancies in declarative process models. *Information Systems*, *64*, 425–446.

Dieci, L., & Walsh III, J. D. (2019). The boundary method for semi-discrete optimal transport partitions and Wasserstein distance computation. *Journal of Computational and Applied Mathematics*, *353*, 318–344.

Djenouri, Y., & Comuzzi, M. (2017). Combining Apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem. *Information Sciences*, *420*, 1–15.

Bibliographie

Docine, L., Andrieu, H., & Creutin, J. D. (1999). Evaluation of a simplified dynamical rainfall forecasting model from rain events simulated using a meteorological model. *Physics and Chemistry of the Earth, Part B : Hydrology, Oceans and Atmosphere*, *24*, 883–887.

Doycheva, K., Horn, G., Koch, C., Schumann, A., & Konig, M. (2017). Assessment and weighting of meteorological ensemble forecast members based on supervised machine learning with application to runoff simulations and flood warning. *Advanced Engineering Informatics*, *33*, 427–439.

Ertoz, L., Steinbach, M., & Kumar, V. (2004). Finding topics in collections of documents : A shared nearest neighbor approach. *Springer*, *Clustering and Information Retrieval*, 83–103.

Fanoodi, B., Malmir, B., & Jahantigh, F. F. (2019). Reducing demand uncertainty in the platelet supply chain through artificial neural networks and ARIMA models. *Computers in Biology and Medicine*, *113*, 103415.

Figueiredo, L. N. L., Assis, G. T. de, & Ferreira, A. A. (2017). A data extraction method based on rendering information and n-gram. *Information Processing and Management*, *53*, 1120–1138.

Fournier-Viger, P., Faghihi, U., Nkambou, R., & Nguifo, E. M. (2012). CMRules : Mining sequential rules common to several sequences. *Knowledge-Based Systems*, *25*, 63–76.

Gan, W., Lin, J. C.-W., Zhang, J., Chao, H.-C., Fujita, H., & Yue, P. S. (2020). ProUM : Projection-based utility mining on sequence data. *Information Sciences*, *513*, 222–240.

Garcia, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, *98*, 1–29.

Garcia, S., Ramirez-Gallego, S., Luengo, J., Benitez, J. M., & Herrera, F. (2016). Big data preprocessing : Methods and prospects. *Big Data Analytics*, *1*, 9.

Gonzalez Camacho, L. A., & Alves-Souza, S. N. (2018). Social network data to alleviate cold-start in recommender system : A systematic review. *Information Processing and Management*, *54*, 529–544.

Guillemot, H. (2010). Connections between simulations and observation in climate computer modeling. Scientists practices and bottom-up epistemology lessons. *Studies in History and Philosophy of Science Part B : Studies in History and Philosophy of Modern Physics*, *41*, 242–252.

Hajek, P., Havel, & Chytil. (1966). The GUHA method of automatic hypotheses determination. *Computing*, *1*, 293–308.

Han, X., Quan, L., Xiong, X., Almeter, M., Xiang, J., & Lan, Y. (2017). A novel data clustering algorithm based on modified gravitational search algorithm. *Engineering Applications of Artificial Intelligence*, *61*, 1–7.

Haymann, N., Lukyanov, V., & Tanny, J. (2019). Effects of variable fetch and footprint on surface renewal measurements of sensible and latent heat fluxes in cotton. *Agricultural and Forest Meteorology*, *268*, 63–73.

Henriques, R., Antunes, C., & Madeira, S. C. (2015). A structured view on pattern mining-based biclustering. *Pattern Recognition*, *48*(12), 3941–3958.

Hu, J., Wang, J., & Xiao, L. (2017). A hybrid approach based on the Gaussian process with t-observation model for short-term wind speed forecasts. *Renewable Energy*, *114*, 670–685.

Bibliographie

Huang, C., Lu, R., & Choo, K.-K. R. (2017). Secure and flexible cloud-assisted association rule mining over horizontally partitioned databases. *Journal of Computer and System Sciences*, *89*, 51–63.

Jancic, M., Kocijan, J., & Grasic, B. (2018). Identification of Atmospheric Variable Using Deep Gaussian Processes. *IFAC-PapersOnLine*, *51*, 43–48.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, *39*, 1503–1509.

Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). *Mining association rules for the quality improvement of the production process*. Expert Systems with Applications.

Kantorovich, L. V. (1982). Functional Analysis. *Pergamon Press*, *Second Edition*.

Khader, N., Lashier, A., & Yoon, S. W. (2016). Pharmacy robotic dispensing and planogram analysis using association rule mining with prescription data. *Expert Systems with Applications*, *57*, 296–310.

Koci, J., & Cerny, R. (2017). Effect of Weather Data Selection on Simulated Moisture and Temperature Fields in Building Envelopes in Central Europe. *Energy Procedia*, *132*, 514–519.

Kocijan, J. (2016). Modelling and control of dynamic systems using Gaussian process models. *Springer*.

Lawrence, N. D. (2006). Learning for Larger Datasets with the Gaussian Process Latent Variable Model. University of Manchester. *School of Computer Science*.

Lee, A. J. T., Lin, W., & Wang, C. (2006). Mining association rules with multi-dimensional constraints. *Journal of Systems and Software*, *79*(1), 79–92.

Lévy, B., & Schwindt, E. L. (2018). Notions of optimal transport theory and how to implement them on a computer. *Computers and Graphics*, *72*, 135–148.

Li, B., Sun, Y., Aw, G., & Teo, K. L. (2019). Uncertain portfolio optimization problem under a minimax risk measure. *Applied Mathematical Modelling*, *76*, 274–281.

Li, C., Qi, X., & Song, D. (2016). Real-time schedule recovery in liner shipping service with regular uncertainties and disruption events. *Transportation Research Part B : Methodological*, *93*, 762-788.

Liao, S., & Chang, H. (2016). A rough set-based association rule approach for a recommendation system for online consumers. *Information Processing and Management*, *52*, 1142–1160.

Liu, T., Wei, H., & Zhang, K. (2018). Wind power prediction with missing data using Gaussian process regression and multiple imputation. *Applied Soft Computing*, *71*, 905–916.

Mai, T., Vo, B., & Nguyen, L. T. T. (2017). A lattice-based approach for mining high utility association rules. *Information Sciences*, *399*, 81–97.

Mattos, C. L. C., Dai, Z., Damianou, A., Barreto, G. A., & Lawrence, N. D. (2017). Deep recurrent Gaussian processes for outlier-robust system identification. *Journal of Process Control*, *60*, 82–94.

McNicholas, P. D., Murphy, T. B., & ORegan, M. (2008). Standardising the lift of an association rule. *Computational Statistics and Data Analysis*, *52*, 4712–4721.

Melucci, M. (2016). Utilising a statistical inequality for efficiently finding term sets. *Information Processing and Management*, *52*, 1086–1121.

Bibliographie

Monge, G. (1781). Memoire sur la theorie des deblais et des remblais. *Histoire de l'Academie Royale des Sciences de Paris*.

Monteserin, A., & Armentano, M. G. (2018). Influence-based approach to market basket analysis. *Information Systems*, *78*, 214–224.

Narvekar, M., & Syed, S. F. (2015). An optimized algorithm for association rule mining using FP tree. *Procedia Computer Science : International Conference on Advanced Computing Technologies and Applications (ICACTA- 2015)*, *45*, 101–110.

Naughton, O., Donnelly, A., Nolan, P., Pilla, F., Misstear, B. D., & Broderick, B. (2018). A land use regression model for explaining spatial variation in air pollution levels using a wind sector based approach. *Science of The Total Environment*, *630*, 1324–1334.

Nguyen, T. P. Q., & Kuo, R. J. (2019). Partition-and-merge based fuzzy genetic clustering algorithm for categorical data. *Applied Soft Computing*, *75*, 254–264.

Nisbet, R. Ph. D., Miner, G. Ph. D., Yale, K. D. D. S., & J.D. (2018). Chapter 3—The Data Mining and Predictive Analytic Process. *Handbook of Statistical Analysis and Data Mining Applications, Second Edition*, 39–54.

Nourani, V., & Molajou, A. (2017). Application of a hybrid association rules/decision tree model for drought monitoring. *Global and Planetary Change*, *159*, 37–45.

Pei, T., Zhu, Ax., Zhou, C., Li, B., & Qin, C. (2006). A new approach to the nearest-neighbour method to discover cluster features in overlaid spatial point processes. *International Journal Geographical Information Science*, *20*, 153–168.

Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. C. (2018). Correlation analysis of performance measures for multi-label classification. *Information Processing and Management*, *54*, 359–369.

Perkins, K. M., Munguia, N., Moure-Eraso, R., Delakowitz, B., Giannetti, B. F., Liu, G., Nurunnabi, M., Will, M., & Velazquez, L. (2018). International perspectives on the pedagogy of climate change. *Journal of Cleaner Production*, *200*, 1043–1052.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rasmussen, C. E., & Williams, C. K. I. (2008). *Gaussian processes for machine learning* (3. print). MIT Press.

Rathore, A., & Patidar, N. P. (2019). Reliability assessment using probabilistic modelling of pumped storage hydro plant with PV-Wind based standalone microgrid. *International Journal of Electrical Power and Energy Systems*, *106*, 17–32.

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery : A comprehensive survey. *Web Semantics : Science, Services and Agents on the World Wide Web*, *36*, 1–22.

Rong, H., Teixeira, A. P., & Guedes Soares, C. (2020). Data mining approach to shipping route characterization and anomaly detection based on AIS data. *Ocean Engineering*, *198*, 106936.

Safa, B., Arkebauer, T. J., Zhu, Q., Suyker, A., & Irmak, S. (2018). Latent heat and sensible heat flux simulation in maize using artificial neural networks. *Computers and Electronics in Agriculture*, *154*, 155–164.

Bibliographie

Sharifzadeh, M., Sikinioti-Lock, A., & Shah, N. (2019). Machine-learning methods for integrated renewable power generation : A comparative study of artificial neural networks, support vector regression, and Gaussian Process Regression. *Renewable and Sustainable Energy Reviews*, *108*, 513–538.

Shi, J. Q., & Choi, T. (2011). Gaussian process regression analysis for functional data. *CRC Press*.

Shi, L., Luo, Z., Matthews, W., Wang, Z., Li, Y., & Liu, J. (2019). Impacts of urban microclimate on summertime sensible and latent energy demand for cooling in residential buildings of Hong Kong. *Energy*, *189*, 116208.

Singh, S., Garg, R., & Mishra, P. K. (2017). Performance optimization of MapReduce-based Apriori algorithm on Hadoop cluster. *Computers Electrical Engineering*, *67*, 348–364.

Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian Processes using Pseudo-inputs ; Gatsby Computational Neuroscience Unit ; University College London. *Neural Information Processing Systems (NIPS)*.

Steinbach, M., Ertoz, L., & Kumar, V. (2004). The Challenges of Clustering High Dimensional Data. *Springer, New Directions in Statistical Physics*, 273–309.

Tseng, Y.-T., Kawashima, S., Kobayashi, S., Takeuchi, S., & Nakamura, K. (2018). Algorithm for forecasting the total amount of airborne birch pollen from meteorological conditions of previous years. *Agricultural and Forest Meteorology*, *249*, 35–43.

Valverde-Rebaza, J. C., Roche, M., Poncelet, P., & Lopes, A. de A. (2018). The role of location and social strength for friendship prediction in location-based social networks. *Information Processing and Management*, *54*, 475–489.

Villani, C. (2009). Optimal Transport : Old and new. *Berlin - Springer*, *338*.

Wulandari, C. P., Ou-Yang, C., & Wang, H.-C. (2019). Applying mutual information for discretization to support the discovery of rare-unusual association rule in cerebrovascular examination dataset. *Expert Systems with Applications*, *118*, 52–64.

Xie, J., Yang, M., Li, J., & Zheng, Z. (2018). Rule acquisition and optimal scale selection in multi-scale formal decision contexts and their applications to smart city. *Future Generation Computer Systems*, *83*, 564–581.

Yesilbudak, M., Sagiroglu, S., & Colak, I. (2017). A novel implementation of *k*-NN classifier based on multi-tupled meteorological input data for wind power prediction. *Energy Conversion and Management*, *135*, 434–444.

Yu, Z., Bedig, A., Montalto, F., & Quigley, M. (2018). Automated detection of unusual soil moisture probe response patterns with association rule learning. *Environmental Modelling and Software*, *105*, 257–269.

Zhang, N., Lin, A., & Yang, P. (2020). Detrended moving average partial cross-correlation analysis on financial time series. *Physica A : Statistical Mechanics and its Applications*, *542*, 122960.

Zhang, Y., Cao, G., Wang, B., & Li, X. (2019). A novel ensemble method for k-nearest neighbor. *Pattern Recognition*, *85*, 13–25.

Zhang, Zhendong, Ye, L., Qin, H., Liu, Y., Wang, C., Yu, X., Yin, X., & Li, J. (2019). Wind speed prediction method using Shared Weight Long Short-Term Memory Network and Gaussian Process Regression. *Applied Energy*, *247*, 270–284.

Bibliographie

Zhang, Zhongjie, Pedrycz, W., & Huang, J. (2018). Efficient mining product-based fuzzy association rules through central limit theorem. *Applied Soft Computing*, *63*, 235–248.

Zhao, C., & Song, G. (2017). Application of data mining to the analysis of meteorological data for air quality prediction : A case study in Shenyang. *IOP Conference Series : Earth and Environmental Science*, *81*, 012097.

Zhu, Y., Tong, Q. L., Yan, X. X., & Li, Y. X. (2019). Development of an uncertain Gaussian diffusion model with its application to production-emission system management in coal-dependent city- a case study of Yulin, China. *Energy Procedia*, *158*, 3253–3258.

Zutshi, A., Grilo, A., & Jardim-Goncalves, R. (2012). The Business Interoperability Quotient Measurement Model. *Computers in Industry*, *63*, 389–404.

Annexe

Article accepté sur 'Springer', conférence I-ESA 2020 :

COULIBALY L., KAMSU-FOGUEM B. et TANGARA F., (2020). Learning with gaussian processes for interoperable weather data modeling. *Springer-Nature*, International conference, I-ESA 2020.

Abstract.

Interoperability is important for both businesses and public services dealing with complex data and models. In meteorological services, reliability is a fundamental requirement. In particular, the uncertainties are dynamic in the observations of sensible heat and latent heat. It is therefore necessary to model the measured data, using the machine learning methods, to provide a reliable database. This can support communication between weather services and observation models. We propose the use of artificial intelligence methods that can improve the production of knowledge for interoperability. Gaussian processes considering uncertainties are used to model the measured values to make the involved databases more reliable. This modeling is done through a deep learning process that includes regression by integrating knowledge in the field.

Keywords : Collaborative services, Model reliability, Deep Learning, Gaussian Processes, Meteorological Models.

Article publié sur 'Future Generation Computer Systems' :

COULIBALY L., KAMSU-FOGUEM B. et TANGARA F., (2020). Rule-based machine learning for knowledge discovering in weather data. *Future Generation Computer Systems*, volume 108, Pages 861-878. https://doi.org/10.1016/j.future.2020.03.012

Abstract.

The Climate change trains regularly some phenomena threatening directly the environment and the humanity. In this context, meteorology plays a more important role in the control of these phenomena. It is thus important to search resources allowing to contribute to the improvement of the numerical model for the predictions of weather and climate.

The objective of this work is to look for the weaknesses of the models in the simulation of exchanges between the surface and the atmosphere. These exchanges are quantified by sensible and latent heat fluxes. The preprocessing is done through the combined use of k-nearest neighbors algorithm (*k*-NN) and Autoregressive integrated moving average (ARIMA) model in order to estimate missing values. The processing is performed with the learning of the association rules and the knowledge extracted enables us to make some comparisons between observations and simulations by the numerical model. The

postprocessing is made by logical and graphical reasoning that facilitates the visualization of links between the obtained rules.

This method is deployed on a database containing measured variables (sensible and latent heat flux, temperature and humidity of the air, wind speed and direction, rain, global radiation, etc.) at the experimental site of the Centre de Recherches Atmosphériques (CRA) which is one of the two sites composing the Pyrenean Plateforme for the Observation of the Atmosphere (P2OA) in France. The obtained and expressed results in the form of association rules have made it possible to highlight that the differences between model and observations from a surface flux point of view are often concomitant with an important difference on global radiation. The expected profits are relative to the generation of knowledge useful for the improvement in the quality of the prediction with a better analysis of the important concomitant factors during errors on a weather model.

Keywords : Data mining ; Association rule learning ; Prediction ; Time series ; Meteorology ; Land-atmosphere exchanges.

Article proposé sur 'SN Computer Science' :

COULIBALY L., KAMSU-FOGUEM B. et TANGARA F., (2020). Explainability with association rule learning for weather forecast. *SN Computer Science.*

Abstract.

The reliability of the weather forecast models is a complex issue, since it depends on numerous parameters and the technical infrastructure which supports them. In doing so, there is a need of advanced works oriented towards the better understanding of these models and the analysis of main associated parameters. Our approach is to study the applicability of the extracted association rules to provide a clearer understanding of atmospheric exchanges.

In this work, the proposed methodology is based on the discovery of the interesting interpretable relationships between measured meteorological parameters at the Atmospheric Research Center of Lannemezan (South-West of France).

Classical evaluation methods use Principal Component Analysis (PCA) to reduce the number of variables and make information less redundant. On the other hand, the advantage of our approach is that the extracted rules are a compression of interpretable useful knowledge with precision without repetition of information. The generated association rules with their statistical and semantic interpretations have globally highlighted the possibilities of explicit analysis of meteorological parameters. This study showed that among the generated relevant rules, three parameters (temperature, humidity, wind speed) have a high frequency in the antecedents of the rules and that the only consequence is rain. This is useful for the identification of potential improvements and gaps in the existing models of atmospheric observations, in particular to understand the related parameterizations to the productivity of the rain phenomenon.

Keywords : Data mining; Imperfect data; Preprocessing; Frequent item sets; Forecasting models; Meteorology.

Article proposé sur 'Big Data Research' :

COULIBALY L., KAMSU-FOGUEM B. et TANGARA F., (2020). Machine Learning based Gaussian processes for Improved Weather Simulation. *Big Data Research.*

Abstract.

Global radiation, latent heat and sensible heat fluxes are major components of the Earth's energy balance, which play an important role in the water cycle and global warming. There are different methods for measuring or estimating these two components, including observations and the numerical forecasting model. In addition, the accuracy of the model is very difficult with high complexity in the simulation of these two meteorological variables (latent heat and sensible heat). Gaussian processes for machine learning (GPML), which use a hierarchical structure to enable adequate identification of very complex systems, can be used to identify the mapping between the input and output values of the simulation systems. With the given mapping function, a predictor is defined by modeling (regression) to make reliable measured data.

In this case study, this predictor is used to facilitate the diagnosis of the numerical forecasting model. Then, an optimizer is associated to reduce weather simulation errors. This optimizer has been defined by some properties on geometric transformations in mathematics. The proposed approach has generated satisfactory results. The cost function is 100% optimal in the case of similar forms and is 98% optimal in the case of forms with the presence of spikes. In addition, with 5 successive days of global radiation simulation, there has been a reduction of approximately 84% of global radiation errors. This approach can be applied to all other variables to further improve the reliability of the numerical forecasting model in the weather simulation systems.

Keywords : Gaussian processes ; Minimization ; Meteorology ; Global radiation ; Uncertainty ; Reliability.