Traitement des données

4.1 Introduction

Nous avons développé dans les chapitres précédents les différents flux énergétiques et les différents flux d'informations. Ces données sont stockées dans des bases de données et accessibles à tout instant. L'objectif de ce chapitre est de présenter le vecteur d'entrée pour chaque flux énergétique avant l'entraînement des différents modèles mathématiques décrits dans le chapitre 2.

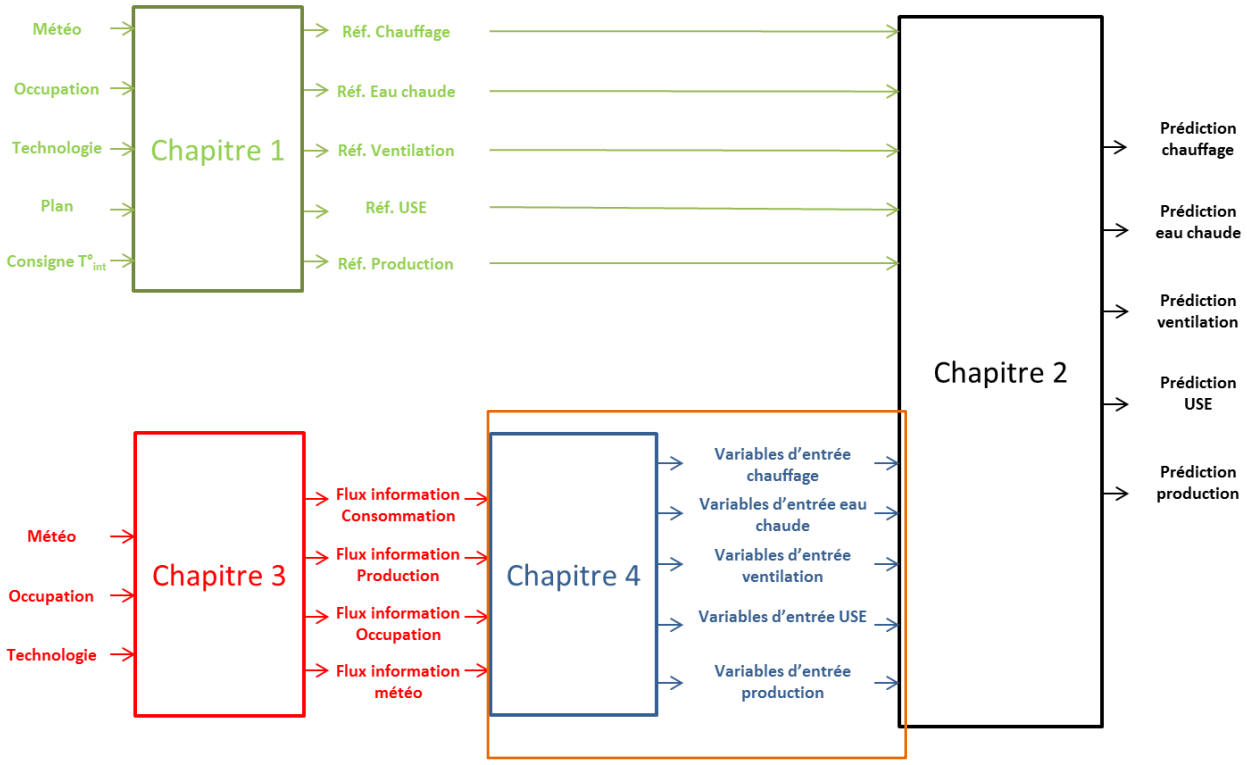


Figure 45 : Aide pour la lecture du document associée aux différents chapitres et au processus générale d'aide à la décision

4.2 Extraction des caractéristiques depuis le compteur global

4.2.1 Préparer et extraire les statistiques de puissance

Le prétraitement des données dépend du choix du capteur sélectionné pour mesurer la consommation d'énergie. Dans cette partie, nous présentons les connaissances tirées depuis le compteur global d'électricité, en particulier au niveau du secteur résidentiel où la puissance active et la

puissance réactive sont mesurées chaque seconde. Les connaissances tirées de cette partie sont généralisables peu importe la fréquence de mesure, seule la fenêtre d'échantillonnage évoluera.

Dans le cas de l'identification de l'ensemble des appareils, une étude à la seconde est nécessaire pour les appareils avec des durées de cycle court comme les convecteurs électriques ou le tambour du lave-linge. Cela permet de soutirer des connaissances liées aux appareils à cycle court : nous parlons de densité d'évènements **[GUE-2011]**.

Ainsi, la fenêtre de temps d'identification des variations de puissance doit être différente selon le problème posé. Dans cette thèse, nous nous intéressons à l'identification du chauffage et de l'eau sanitaire fournis par des pompes à chaleur. Ces appareils ont des cycles lents supérieurs à la minute. Un filtre médian permettra d'enlever du bruit sur le signal électrique et d'enlever les appareils à cycle court qui pourrait se confondre avec des pompes à chaleur. Un filtre médian de taille 10 est appliqué sur chacune des phases nommé p au niveau du compteur global :

$$E_{Consommation\_électrique\_totale}\ t(i) = [Date\ ;$$

$$P_{active}\ t(i) = médiane \sum_{p=1}^{3} \sum_{i=1}^{10} P\ _{active\_norm}\ t(i) + P\ _{active\_norm}\ t(i) + P\ _{active\_norm}\ t(i)\ ;$$

$$P_{réactive}\ t(i) = médiane \sum_{p=1}^{3} \sum_{i=1}^{10} P\ _{réactive\_norm}\ t(i) + P\ _{réactive\_norm}\ t(i) + P\ _{réactive\_norm}\ t(i)]; \textbf{(4.1)}$$

Le lissage permet d'enlever du bruit mais il a comme conséquence de perdre également de l'information (Figures 46 et 47). Comme énoncé ci-dessus, tout dépend de l'identification des appareils électriques recherchés.

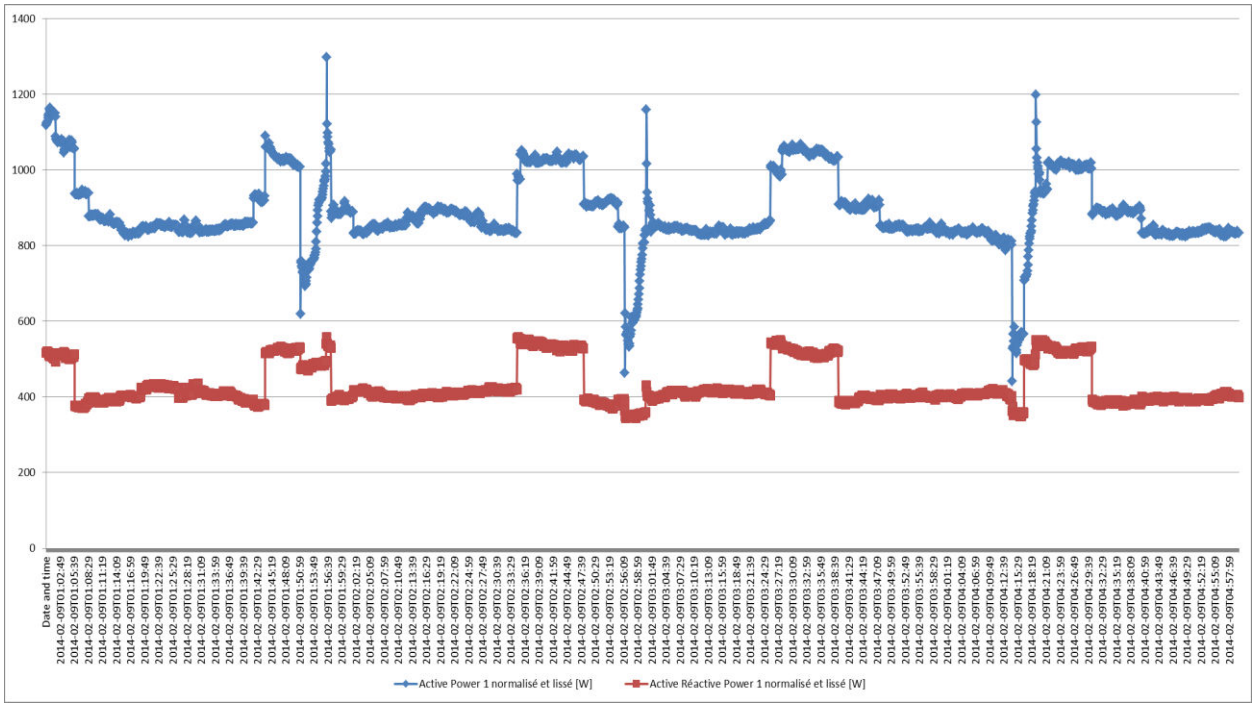


Figure 46 : Puissance active et réactive à la seconde en sortie du compteur d'électricité dans le secteur résidentiel sur une phase

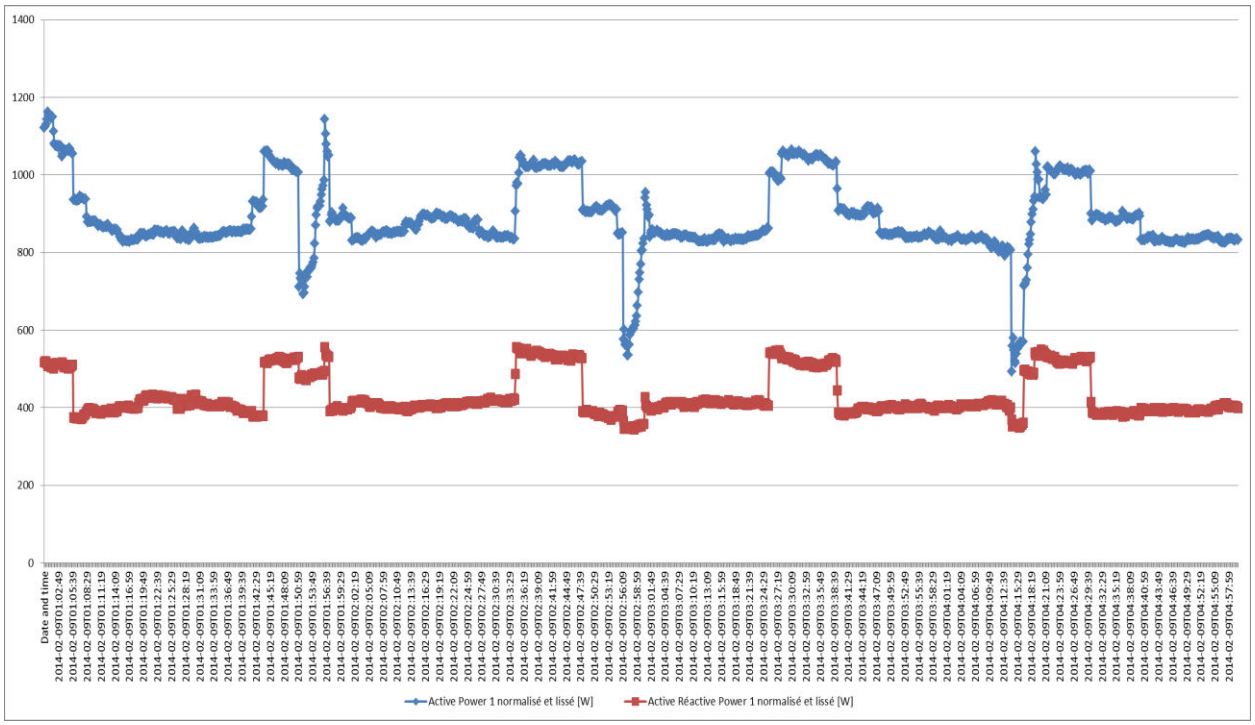


Figure 47 : Puissance active et réactive sans filtre au-dessus et après un filtre médian de taille 10

D'un point de vue électrique, en régime permanent, une variation est caractérisée par la transition entre un état électrique « stable » caractéristique du fonctionnement d'un appareil et un état instable. Dans la littérature, un évènement est une variation d'état au niveau de la seule puissance active et/ou réactive si le capteur récolte cette information **[HART-1992].**

Les seuils de puissance utilisés sont compris entre **40 Watts et 50 Watts** pour une identification complète des charges individuelles et de **15 VAR pour la puissance réactive [HART-1992]**. Cela sera notre point de départ comme seuil pour l'identification d'évènements sur la courbe de charge globale par phase. Cette variation peut se faire à partir d'une moyenne glissante sur une fenêtre de temps à définir selon le problème et les charges à identifier.

$$Evènement\ t(i) = \ [Date, \Delta P\ t(i), \Delta Q\ t(i)] \qquad \textbf{(4.2)}$$

Un évènement est détecté si la puissance active et la puissance réactive dépasse un seuil sur une fenêtre de temps prédéfinie :

$$[\Delta P\ t(i), \Delta Q\ t(i)] = [Date,$$

$$\Delta P\ t(i) = \frac{P_{\ active}\ t(i-1) * (n-1) + \ P_{\ active}\ t(i)}{n} > Seuil\ P_{\ active}\ (40\ Watt),$$

$$\Delta Q\ t(i) = \frac{P_{\ réactive}\ t(i-1) * (n-1) + P_{\ réactive}\ t(i)}{n} > Seuil\ P_{\ réactive}\ (15\ VAR)\ ] \qquad \textbf{(4.3)}$$

Nous avons maintenant en sortie, pour chaque heure, une table avec l'énergie globale normalisée et les différentes variations correspondante à une date donnée :

$$E_{Consommation\_électrique\_totale}\ t(i) = [Date\ ;$$

$$P_{\ active\_normalisé\_phase1}\ t(i) = \left(\frac{230}{V\ t(i)}\right)^2\ P_{\ active}\ t(i),$$
$$P_{\ réactive\_normalisé\_phase1}\ t(i) = \left(\frac{230}{V\ t(i)}\right)^2\ P_{\ réactive}\ t(i)\ ;$$
$$P_{\ active\_normalisé\_phase2}\ t(i) = \left(\frac{230}{V\ t(i)}\right)^2\ P_{\ active}\ t(i),$$
$$P_{\ réactive\_normalisé\_phase2}\ t(i) = \left(\frac{230}{V\ t(i)}\right)^2\ P_{\ réactive}\ t(i)\ ;$$
$$P_{\ active\_normalisé\_phase3}\ t(i) = \left(\frac{230}{V\ t(i)}\right)^2\ P_{\ active}\ t(i),$$
$$P_{\ ré active_{normalisé_{phase3}}\ t(i)} = \left(\frac{230}{V\ t(i)}\right)^2\ P_{\ réactive}\ t(i)\ ;$$
$$\Delta P\ t(i), \Delta Q\ t(i), Date\_Evènement] \qquad \textbf{(4.4)}$$

Statistiques sur les puissances et sur les variations

Les calculs statistiques permettent de caractériser un flux énergétique. Ils sont réalisés sur les puissances et sur les évènements. Ils sont les points d'entrée de la caractérisation passée de chaque flux énergétique.

La moyenne glissante : est une notion statistique, où la moyenne au lieu d'être calculée sur n valeurs fixes, est calculée sur n valeurs consécutives « glissantes ».

$$\bar{x} = \frac{\bar{x}_{n-1} * (n-1) + x_n}{n} \qquad \textbf{(4.5)}$$

La médiane est la valeur à laquelle 50 % des valeurs observées sont inférieures. En supposant que l'on ait, au préalable, rangé les valeurs observées de sorte qu'elles se trouvent indexées suivant l'ordre des valeurs croissantes ($x_1 + x_2 + x_i + x_{i+1}$). Pour un nombre pair *2n* de valeurs, la médiane est la *moyenne* des deux valeurs centrales, soit

$$\frac{x_n + x_{n+1}}{2} \qquad \textbf{(4.6)}$$

ou toute autre valeur strictement comprise entre x_n *et* x_{n+1}

pour un nombre impair *2n+1* de valeurs, la médiane est unique et égale à x_{n+1}

La **variance** est une mesure servant à caractériser la dispersion d'un échantillon ou d'une distribution. Si la série statistique est de moyenne *m* et prend les valeurs x_1, x_2, ..., x_n, sa variance est

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \qquad \textbf{(4.7)}$$

La déviation moyenne, la moyenne géométrique, le minimum et le maximum sont également ajoutés **[BER-2010].**

Identification des habitudes de consommations

Nous allons nous baser sur la variable de temps pour extraire les différentes connaissances qui permettront à nos modèles mathématiques de **définir des habitudes de consommations**. Nous ajoutons, tirés de planning public, des jours particuliers comme les jours fériés ou les vacances scolaires des enfants caractérisés par des variables booléennes :

- ⇨ des **mois** m allant de 1 à 12
- ⇨ des **jours de l'année** allant de 1 à 365
- ⇨ des **jours de la semaine** allant de 1 à 7 (lundi, mardi, mercredi, jeudi, vendredi, samedi et dimanche)
- ⇨ **l'heure h** allant de 0 à 23
- ⇨ des jours correspondant aux **vacances scolaires** cantonales caractérisés par une variable booléenne 0 ou 1.
- ⇨ des jours correspondant aux **jours fériés** cantonaux caractérisés par une variable booléenne 0 ou 1.

Comme décrit dans le chapitre 3, nous ajoutons comme information le **planning de production** pour chacun des flux énergétiques. Dans le cas du secteur résidentiel, cela passe par la connaissance des **consignes données aux pompes à chaleur** où une distinction est réalisée entre les jours de la semaine et les jours du week-end.

Concernant le secteur résidentiel, la consommation électrique globale, le besoin en chauffage et le besoin en eau chaude, nous **distinguons les jours de la semaine et le week-end**. **Concernant le secteur tertiaire et industriel, les jours et les heures en dehors du planning de production sont enlevées** de notre set de données d'entraînement et de test.

Si nous reprenons notre vecteur d'entrée, nous avons maintenant :

$$\begin{aligned} E_{\text{Consommation\_électrique\_totale}}\, t(i) = [&\text{Date},\\ &\text{Statistiques Puissance}_{\text{Phase 1}}\, t(i),\\ &\text{Statistiques Variations}_{\text{Phase 1}}\, t(i)\\ &\text{Statistiques Puissance}_{\text{Phase 2}}\, t(i),\\ &\text{Statistiques Variations}_{\text{Phase 2}}\, t(i)\\ &\text{Statistiques Puissance}_{\text{Phase 3}}\, t(i),\\ &\text{Statistiques Variations}_{\text{Phase 3}}\, t(i)\\ &\text{Année}_j\,, \text{Mois}_m\,, \text{Jour}_j, \text{Semaine}_{0.1}, \text{Heure}_h, \text{Jour fériés}_{0.1}, \text{Vacances scolaires}_{0.1},\\ &\Delta P\, t(i), \Delta Q\, t(i), \text{Date\_Variations}] \end{aligned} \quad \textbf{(4.8)}$$

Un modèle d'entraînement sera créé pour les jours k allant de lundi au vendredi (de 1 à 5) et un autre modèle pour le samedi et le dimanche (6 et 7) :

$$E_{\text{Consommation\_électrique\_totale}}\, t(i) = \sum_{j=1,k=1,w=0,h=1}^{j=365,k=5,w=1,h=24} E_{\text{Consommation\_électrique\_totale}}\, t(i) \quad \textbf{(4.9)}$$

$$E_{Consommation\_électrique\_totale}\ t(i) = \sum_{j=1,k=6,w=0,h=1}^{j=365,k=7,w=1,h=24} E_{Consommation\_électrique\_totale}\ t(i) \quad \textbf{(4.10)}$$

4.2.2 Identification du chauffage et de l'eau chaude

La plus grande difficulté dans l'identification des appareils électriques de manière non intrusive est sa généralisation. Dans les études qui ont porté sur des centaines de logements **[ENE-2010] [ADE-2010]**, les données collectées ont souvent portées que sur la puissance active et sur la consommation d'énergie mais les analyses présentent des variations extrêmes entre les logements. Il devient ainsi difficile avec une fréquence à la seconde de séparer les évènements, en particulier ceux liés à des faibles puissances (généralement en dessous de la puissance du réfrigérateur-congélateur de 250 Watt en pic) mais qui représentent la plus grande part dans la répartition parmi l'ensemble des évènements détectés (Figure 48) **[DUF-14]**. Leurs nombres sont conséquents et correspondent à des appareils à un état (lampes, congélateurs) ou un sous-état spécifique (compresseur d'une pompe à chaleur) (Figure 43).

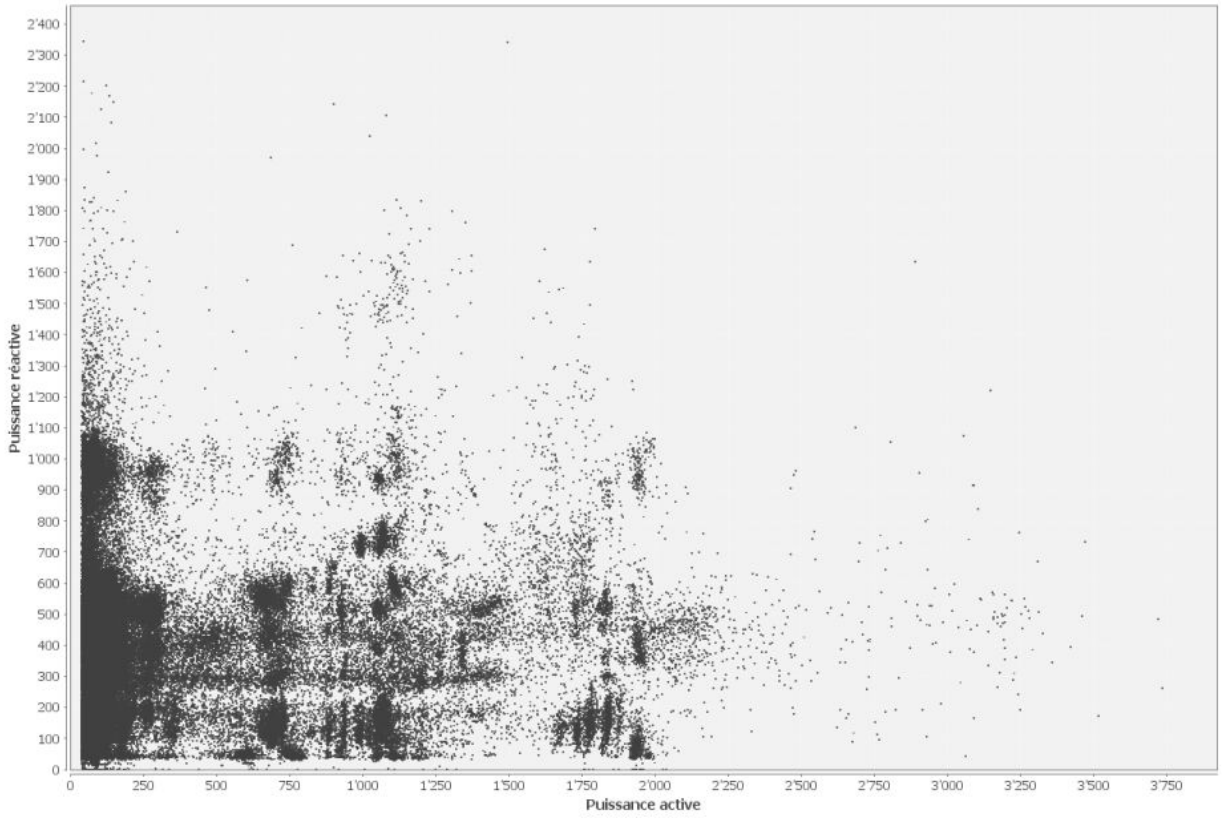


Figure 48 : Evènements détectés sur chacune des phases en sortie du compteur global entre le 1 Janvier et le 18 Juin 2014

Pour l'électroménager, une étude à la seconde est nécessaire pour tirer parti du nombre d'évènements liés aux appareils électriques à forte densité. Cette notion est définie dans **[GUE-2009]**. Elle permet par exemple d'extraire les variations dites rapides d'un convecteur électrique ou d'un lave-linge. Ainsi, le type de traitement avant l'identification des appareils de manière individuelle dépend de l'appareil visé. Nous nous sommes basés sur le secteur résidentiel car nous avions accès aux bases de contenant la puissance active et réactive d'appareils électriques.

Nous nous sommes concentrés sur les plus gros consommateurs et avec le plus gros potentiel de pilotage : le chauffage et l'eau chaude. Ces usages énergétiques sont fournis par une pompe à chaleur, appareil électrique qui a un cycle de fonctionnement lent. Les variations entre une variation positive et négative sont supérieures à la minute dans nos cas d'étude. Notre méthode peut se résumer en six étapes (Figure 49) :

- ⇨ Création d'une base de données d'appareils électriques avec la puissance active et réactive
- ⇨ Décomposition des appareils multi-états en sous états
- ⇨ Définition des différents évènements à partir de la puissance active et réactive sur chacune des phases (3 points de mesure) à partir d'un filtre médian à la minute
- ⇨ Recherche spécifique par classe : Pompe à chaleur/Autre
- ⇨ Ajouts de seuils spécifiques liés à la puissance active et réactive
- ⇨ Identification des évènements dans le plan actif/réactif à partir d'une base de données d'apprentissage et de méthodes d'identification non supervisé

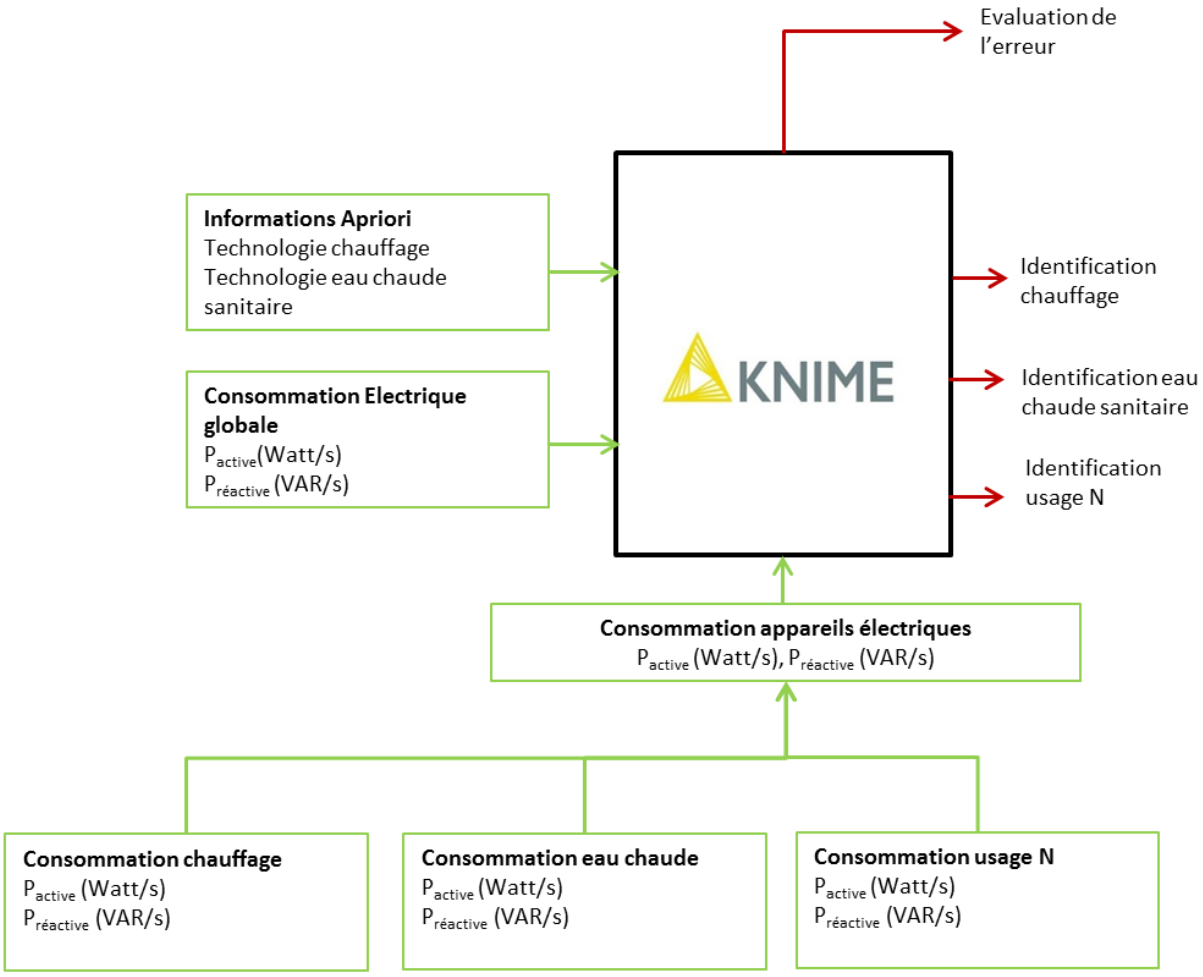


Figure 49 : Identification des appareils électriques par usages à partir de la consommation électrique globale en sortie du compteur et d'une base de données d'apprentissage

L'étude commence hors ligne par une meilleure décomposition des appareils électriques dans le but d'améliorer l'identification. Nous allons décomposer les appareils multi-états par une classification non supervisée. L'identification doit porter sur les appareils qui représentent soit le plus grand potentiel de flexibilité, soit la plus grande part dans la consommation d'énergie du système, soit les deux.

Notre objectif est de collecter la puissance active et réactive d'un maximum d'appareils utilisés dans le secteur résidentiel et tertiaire. Cette base de données représentera notre set d'entraînement de notre modèle d'identification. Concernant la campagne de mesure réalisée par la HES-SO Fribourg **[RID-2015]**, deux sessions d'acquisition d'une heure avec une fréquence de collecte de 10 secondes sur 100 appareils domestiques divisés dans 10 catégories ont été réalisées (Tableau 13).

Appareils électriques	Paramètres	Unité	Fréquence de collecte	Historique
10 congélateurs	[Puissance active, Puissance réactive]	[Watt, VAR]	Seconde	1 mois
10 réfrigérateurs				
systèmes Hi-fi (lecteurs de CD)				
10 lampes				
10 machines à café				
10 Micro-ondes				
10 Imprimantes				
10 Ordinateurs fixes				

Tableau 13 : Base de données d'appareils électriques [RID-2015]

Appareils électriques	Paramètres	Unité	Fréquence de collecte	Historique
Pompe à chaleur Eau chaude	[Puissance active, Puissance réactive]	[Watt, VAR]	Seconde	1 mois
Pompe à chaleur Chauffage				
Lave-Vaisselle				
Lave-linge				

Tableau 14 : Base de données d'appareils électriques dans la maison test [DUF-2015]

Chaque appareil à plusieurs états présent dans nos bases de données est décomposé en sous états par un apprentissage supervisé.

Chaque pompe à chaleur de notre base de données (une pompe à chaleur liée à l'eau chaude et une autre liée au chauffage) est décomposée en plusieurs états. Nous avons choisi une décomposition en deux états pour se focaliser sur les variations à haute puissance en relation avec le condenseur.

Ces variations sont les données d'entrée d'un apprentissage supervisé développé dans le chapitre 4.4. Cela nous permet de créer des classes de manière supervisée ou non supervisée. Dans le cas de la pompe à chaleur, nous utilisons un algorithme, le 2-means **[BER-2010]** pour réaliser la classification qui va créer deux classes de manière itérative centrées chacune d'elles autour d'une moyenne.

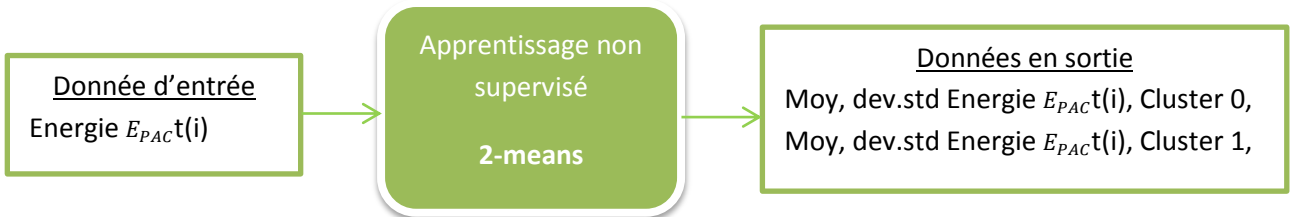


Figure 50 : Donnée d'entrée et de sortie de l'algorithme de décomposition de la signature électrique d'une pompe à chaleur en 2 états

Nous utilisons les évènements détectés en sortie des compteurs de la PAC liée au chauffage et celle liée à l'eau chaude comme données d'entrée pour identifier les classes de fonctionnement. La valeur absolue est calculée pour chaque évènement identifié. La connaissance liée à l'enclenchement ON ou déclenchement OFF de l'évènement est gardée.

Nous avons en sortie différents capteurs liés à nos pompes à chaleur les différentes variations liées à la puissance active et réactive :

$$Variations\ t(i) = [|\Delta P\ t(i), \Delta Q\ t(i)|, ON/OFF]. \quad \textbf{(4.11)}$$

Une identification à partir de l'algorithme SVM **[DUF-2015]** est réalisée sur la fenêtre de temps choisie par le centre de pilotage. Le détail de la méthodologie de prédiction du chauffage et de l'eau chaude par pompe à chaleur est décrit dans l'annexe 3.

En sortie nous avons l'énergie de chauffage et d'eau chaude consommée pendant l'heure considérée :

$$\begin{aligned} E_{Consommation\_électrique\_totale}\ t(i) = [&Date,\\ &Statistiques\ Puissance_{\ Phase\ 1}\ t(i),\\ &Statistiques\ Variations_{\ Phase\ 1}\ t(i)\\ &Statistiques\ Puissance_{\ Phase\ 2}\ t(i),\\ &Statistiques\ Variations_{\ Phase\ 2}\ t(i),\\ &Statistiques\ Puissance_{\ Phase\ 3}\ t(i),\\ &Statistiques\ Variations_{\ Phase\ 3}\ t(i),\\ &E_{Chauffage}\ t(i),\\ &E_{ECS}\ t(i),\\ Année_j\ , Mois_m\ , Jour_j, Semaine_{0.1}, Heure_h, &Jour\ fériés_{0.1}, Vacances\ scolaires_{0.1},] \end{aligned} \quad \textbf{(4.12)}$$

4.3 Créer son propre modèle de prédictions météorologiques

Les données météorologiques en particulier pour la production solaire décentralisée et pour le chauffage sont essentielles particulièrement les données prédites pour l'heure suivante.

Nous avons remarqué que les valeurs de prédiction fournies par MétéoSuisse comprenaient des erreurs en luminosité et en température extrêmes correspondant à des phénomènes locaux difficilement prévisibles. Nous définissons les erreurs liées aux prédictions fournies de température et de luminosité qui correspondent chaque heure à la différence entre la réalité et la prédiction (Equations 4.16 et 4.17). Notre objectif sera d'apprendre l'erreur sur des années d'historique pour l'appliquer sur nos années utilisées en test.

$$Erreur_{Luminosité} = Luminosité_{réelle} - Luminosité_{prédite} \quad \textbf{(4.13)}$$

$$Erreur_{Température} = Température_{réelle} - Température_{prédite} \quad \textbf{(4.14)}$$

Par exemple pour la luminosité, l'erreur est égale à la luminosité réelle moins la luminosité prédite. La dispersion de chaque erreur est gaussienne avec des écarts de plus de 1000 W/m2 pour la luminosité et de 24 degrés pour la température (Figures 51 et 52).

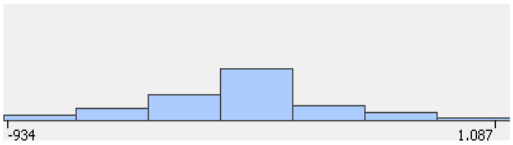
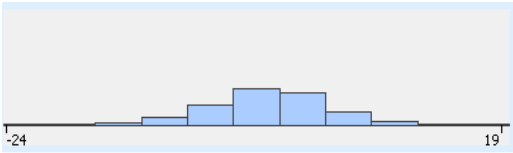
Column	Min	Mean	Median	Max	Std. Dev.	Histogram
Erreur luminosité	-933.9867	0.0173	3.4681	1'087.4983	344.0568	
Erreur température	-23.6817	-0.9533	-0.7967	19.258	5.1153	

Figure 51 : Statistique et distribution de l'erreur de prédiction en luminosité et en température, station Aigle

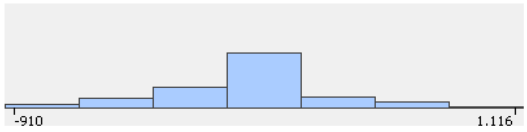
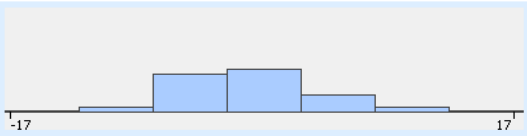
Column	Min	Mean	Median	Max	Std. Dev.	Histogram
Erreur luminosité	-910	5.4369	0.0	1'115.8493	304.1716	
Erreur température	-16.599	-0.466	-0.9812	17.344	4.322	

Figure 52 : Statistique et distribution de l'erreur de prédiction en luminosité et en température, station Sion

Puisque nous avons plusieurs années de données réelles et prédites, nous allons créer notre propre modèle de prédiction de la température et de la radiation lumineuse. Ainsi, nous avons créé un modèle mathématique de l'erreur par station météo et par donnée d'entrée (température, luminosité) (Figure 53).

Cette erreur est apprise par une méthode non paramétrique détaillée dans le chapitre 2. Pour cette étape, nous utilisons un ensemble d'arbres de décision. Les nouvelles prédictions de luminosité et de température corrigées seront nos nouvelles données d'entrée.

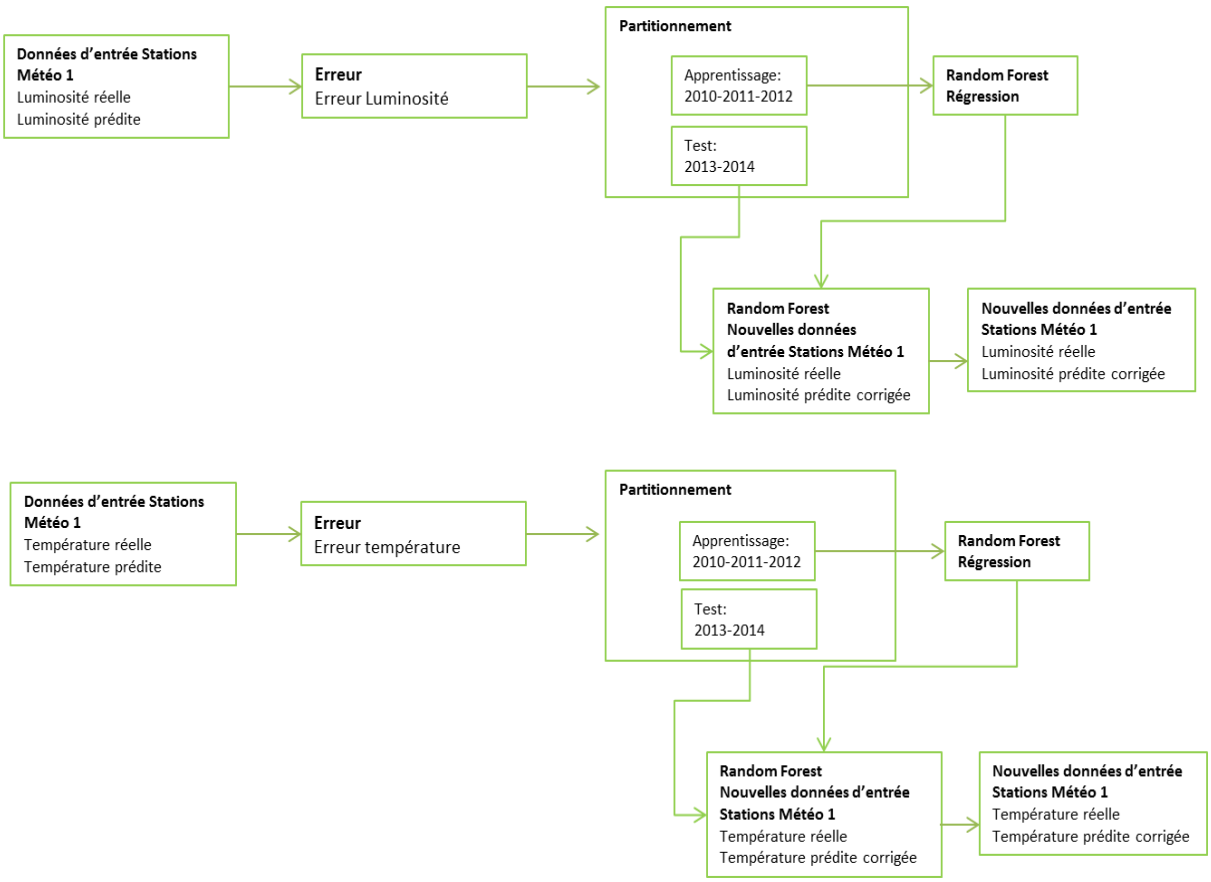


Figure 53: Méthodologie de réduction de l'erreur prédite météorologique provenant de MétéoSuisse

Pour caractériser l'erreur, un historique basé sur les données réelles et prédites est construit. Cet historique représente les données d'entrée de notre modèle mathématique. Les données d'apprentissage correspondent aux mesures du 1 Janvier 2010 au 31 Août 2012. Les données de test vont du 1 septembre 2012 au 31 Mars 2014 (Figure 54). Cette répartition a été réalisée pour avoir les données corrigées qui correspondent à nos périodes de test pour la prédiction des flux énergétiques.

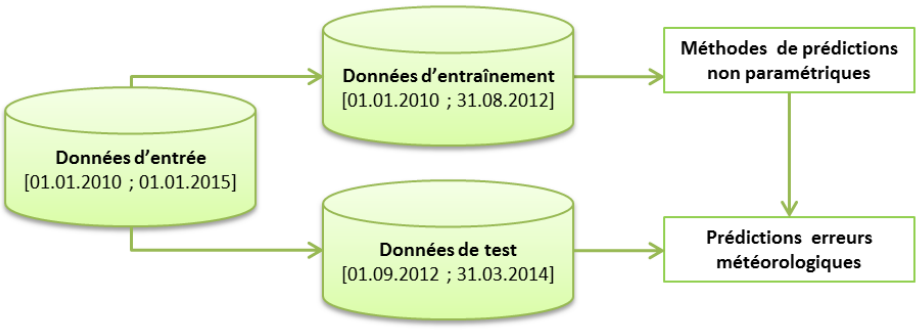


Figure 54 : Répartition des données d'entrée pour la correction des erreurs météorologiques

En sortie du modèle mathématique, une erreur est prédite pour chaque heure correspondant à notre set de test représenté par les figures 49 et 50.

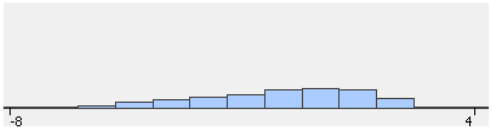
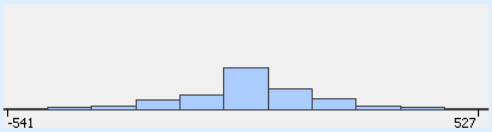
Column	Min	Mean	Median	Max	Std. Dev.	Histogram
Nouvelle erreur température	-8.3284	-1.0864	-0.7661	4.1215	2.1042	
Nouvelle erreur luminosité	-540.8744	10.9844	9.5803	527.4162	147.8339	

Figure 55: Statistiques et distribution de la nouvelle erreur de prédiction en luminosité et en température, station Aigle

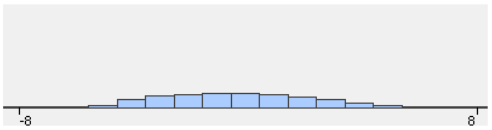
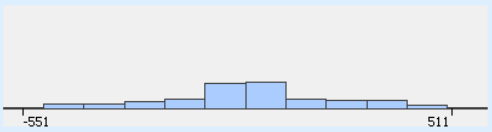
Column	Min	Mean	Median	Max	Std. Dev.	Histogram
Nouvelle erreur température	-8.4099	-0.7491	-0.823	7.6446	2.5826	
Nouvelle erreur luminosité	-550.5096	7.1758	0.0	510.8784	203.8268	

Figure 56: Statistiques et distribution de la nouvelle erreur de prédiction en luminosité et en température, station Sion

Les données météorologiques corrigées sont intégrées à notre vecteur d'entrée :

$$
\begin{aligned}
E_{\text{Consommation\_électrique\_totale}}\, t(i) = [&\text{Date},\\
&\text{Statistiques Puissance}_{\text{Phase 1}}\, t(i),\\
&\text{Statistiques Variations}_{\text{Phase 1}}\, t(i)\\
&\text{Statistiques Puissance}_{\text{Phase 2}}\, t(i),\\
&\text{Statistiques Variations}_{\text{Phase 2}}\, t(i),\\
&\text{Statistiques Puissance}_{\text{Phase 3}}\, t(i),\\
&\text{Statistiques Variations}_{\text{Phase 3}}\, t(i),\\
&E_{\text{Chauffage}}\, t(i),\\
&E_{\text{ECS}}\, t(i),\\
&\text{Données}_{\text{Réelles}}\_\text{Météo}_{\text{Station}_1}\, t(i),\\
&\text{Données}_{\text{Prédites}}\_\text{Météo}_{\text{Station}_1}\, t(i),\\
&\text{Données}_{\text{Réelles}}\_\text{Météo}_{\text{Station}_2}\, t(i)\\
&\text{Données}_{\text{Prédites}}\_\text{Météo}_{\text{Station}_2}\, t(i),\\
&\text{Données}_{\text{Réelles}}\_\text{Météo}_{\text{Station}_3}\, t(i),\\
&\text{Données}_{\text{Prédites}}\_\text{Météo}_{\text{Station}_3}\, t(i),\\
&\text{Données}_{\text{Réelles}}\_\text{Météo}_{\text{Station}_4}\, t(i),\\
&\text{Données}_{\text{Prédites}}\_\text{Météo}_{\text{Station}_4}\, t(i),\\
&\text{Données}_{\text{Réelles}}\_\text{Météo}_{\text{Station}_5}\, t(i),\\
&\text{Données}_{\text{Prédites}}\_\text{Météo}_{\text{Station}_5}\, t(i),\\
&\text{Année}_j, \text{Mois}_m, \text{Jour}_j, \text{Semaine}_{0.1}, \text{Heure}_h, \text{Jour fériés}_{0.1}, \text{Vacances scolaires}_{0.1}]
\end{aligned}
\qquad \textbf{(4.15)}
$$

4.4 Classification

4.4.1 Apprentissage supervisé

Le processus se passe en deux phases. Lors de la première phase (hors ligne, dite d'apprentissage), il s'agit de déterminer un modèle des données étiquetées. La seconde phase (en ligne, dite de test) consiste à prédire l'étiquette d'une nouvelle donnée, connaissant le modèle préalablement appris. Cette méthodologie est utilisée dans le cas de la labellisation de l'eau chaude et du chauffage. L'objectif est de définir deux classes ON et OFF qui correspondent respectivement à l'allumage et à l'arrêt du système de production de chauffage et d'eau chaude.

A partir du besoin en chauffage, un apprentissage supervisé est réalisé sur les données d'entraînement et appliqué sur les données de test. Si l'énergie de chauffage est en dessus d'un seuil, elle est labellisée ON, sinon OFF. Une classification binaire est réalisée à partir de l'énergie de chauffage pour définir les quarts d'heure et les heures où le chauffage est allumé ou éteint (Figure 51). Si le besoin en chauffage ou en eau chaude à l'instant t(i) est en dessous d'un seuil prédéfini, nous considérons le besoin comme OFF. Le même apprentissage est réalisé pour l'eau chaude sanitaire à partir de l'énergie par heure.

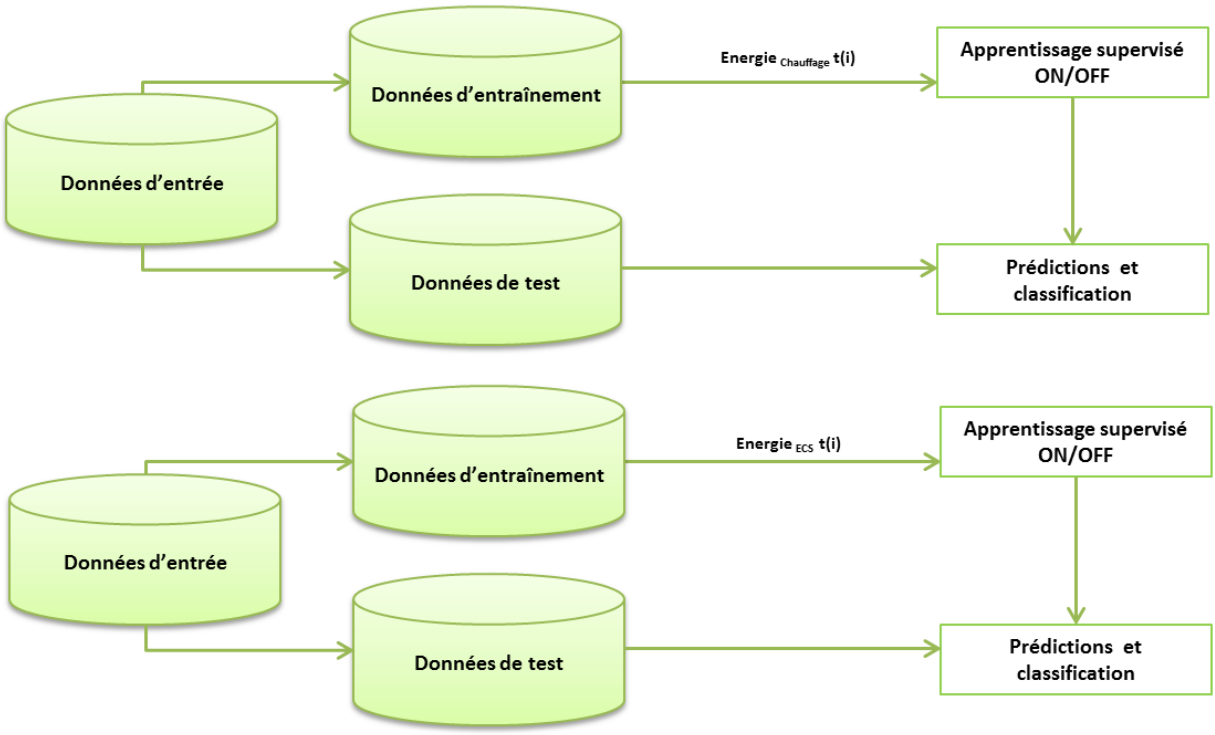


Figure 57 : Méthodes de classification des données pour le besoin en eau chaude sanitaire et en chauffage

Un historique des consommations et des données météorologiques est créé par heure pour chacune des variables d'entrée caractérisées. Cet historique correspond au besoin en chauffage heure

par heure sur les deux jours précédents la prédiction à calculer. Par exemple, pour une prédiction à l'heure, nous avons :

$$E_{Chauffage}\ t(i+1) = [Date\ ;$$

$$\begin{gathered}
Statistiques\ Puissance_{\ Phase\ 1}\ t(i),\dots\ , Statistiques\ Puissance_{\ Phase\ 1}\ t(i-48),\\
Statistiques\ Variations_{\ Phase\ 1}\ t(i),\dots, Statistiques\ Variations_{\ Phase\ 1}\ t(i-48)\\
Statistiques\ Puissance_{\ Phase\ 2}\ t(i),\dots. Statistiques\ Puissance_{\ Phase\ 2}\ t(i-48),\\
Statistiques\ Variations_{\ Phase\ 2}\ t(i),\dots, Statistiques\ Variations_{\ Phase\ 2}\ t(i-48),\\
Statistiques\ Puissance_{\ Phase\ 3}\ t(i),\dots, Statistiques\ Puissance_{\ Phase\ 3}\ t(i-48),\\
Statistiques\ Variations_{\ Phase\ 3}\ t(i),\dots,Statistiques\ Variations_{\ Phase\ 3}\ t(i-48),\\
Données_{Réelles\ _Météo_{Station}}\ \ t(i),\dots., Données_{Réelles\ _Météo_{Station}}\ \ t(i-48),\\
Données_{Prédites\ _Météo_{Station}}\ \ t(i+1),\dots, Données_{Prédites\ _Météo_{Station}}\ \ t(i-48)
\end{gathered}$$

$$Année_j\ , Mois_m\ , Jour_j, Semaine_{0.1}, Heure_h, Jour\ fériés_{0.1}, Vacances\ scolaires_{0.1},$$

$$Classes_{ON/OFF}\ ] \qquad \textbf{(4.16)}$$

Un historique des consommations et des données météorologiques est également créé par heure pour chacune des variables d'entrée caractérisées pour l'eau chaude sanitaire:

$$E_{ECS}\ t(i+1) = [Date\ ;$$

$$\begin{gathered}
Statistiques\ Puissance_{\ Phase\ 1}\ t(i),\dots\ , Statistiques\ Puissance_{\ Phase\ 1}\ t(i-48),\\
Statistiques\ Variations_{\ Phase\ 1}\ t(i),\dots, Statistiques\ Variations_{\ Phase\ 1}\ t(i-48)\\
Données_{Réelles\ _Météo_{Station}}\ \ t(i),\dots., Données_{Réelles\ _Météo_{Station}}\ \ t(i-48),\\
Données_{Prédites\ _Météo_{Station}}\ \ t(i+1),\dots, Données_{Prédites\ _Météo_{Station}}\ \ t(i-48)
\end{gathered}$$

$$Année_j\ , Mois_m\ , Jour_j, Semaine_{0.1}, Heure_h, Jour\ fériés_{0.1}, Vacances\ scolaires_{0.1},$$

$$Classes_{ON/OFF}\ ] \qquad \textbf{(4.17)}$$

4.4.2 Apprentissage non supervisé

Quand le système ne dispose que d'exemples non labellisés et que le nombre de classes n'a pas été prédéterminé, on parle d'apprentissage non supervisé. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données.

Dans cette thèse, nous utilisons l'algorithme EM (espérance-maximisation) proposé par **[VIC-2014]**. C'est un algorithme itératif qui permet de trouver les paramètres de vraisemblance maximum d'un modèle probabiliste lorsque ce dernier dépend de variables latentes non observables. L'algorithme d'espérance-maximisation comporte :

- ⇨ une étape d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées,
- ⇨ une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi. On dispose de n observations, considérées comme les réalisations de n variables aléatoires indépendantes et identiquement distribuées (X_1,…., X_n).

On note $L_\theta(x_i)$, $L(x_i ; \theta)$, ou plus souvent $L(x_i \mid \theta)$, dans la littérature anglophone, la densité de X_i . On appelle vraisemblance de l'échantillon, la densité jointe de (X_1,…., X_n):

$$L_n(x_1, \ldots\ldots, x_n) = \prod_{i=1}^{n} L(x_i ; \theta) \qquad \textbf{(4.18)}$$

Dans le cas particulier d'une loi discrète, cela se ramène à :

$$L_n(x_1, \ldots\ldots, x_n) = P_\theta \{(X_1 = x_1, \ldots\ldots, X_n = x_n)\} = \prod_{i=1}^{n} P_\theta \{ X_i = x_i \} \qquad \textbf{(4.19)}$$

L'estimateur de θ est :

$$\theta = \arg max \; L_n(x_1, \ldots\ldots, x_n; \theta) \qquad \textbf{(4.20)}$$

Cette méthodologie est utilisée dans le cas de la labellisation de la consommation électrique globale dans les secteurs résidentiel, tertiaire et industriel et de la production décentralisée solaire. En sortie, chaque classe est représentée par une moyenne, une déviation standard et sa répartition dans le set de données d'entraînement et de test.

La consommation électrique globale

A partir de la consommation électrique globale, un apprentissage non supervisé est réalisé. Cet apprentissage est réalisé pour chacun des systèmes étudiés, peu importe le secteur d'activité. Il est réalisé à partir de l'énergie électrique en sortie du compteur. Pour cet apprentissage, la donnée d'entrée utilisée est la consommation électrique globale et le modèle de classification est ensuite appliqué sur les données de test (Figure 52).

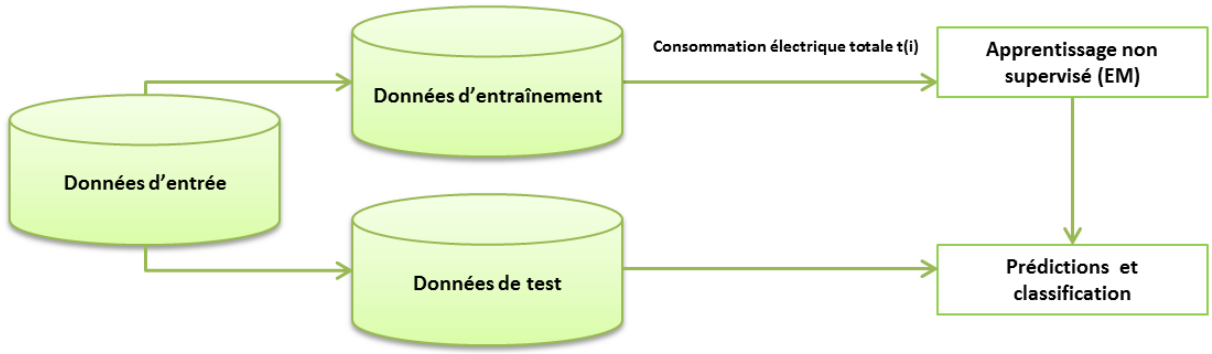


Figure 58 : Méthodes de classification des données pour la consommation électrique totale

En sortie, chaque ligne de notre vecteur d'entrée est labellisé par une classe représentée par une gaussienne avec une moyenne et une déviation standard comme décrit en exemple dans le tableau 15.

Par exemple, nous avons pour le secteur résidentiel 6 classes en sortie de l'apprentissage décrites dans le tableau 15. L'objectif de notre outil sera également de prédire dans quelle classe se

trouvera la consommation électrique l'heure suivante. Un historique des consommations et des données météorologiques est créé par heure pour chacune des variables d'entrée caractérisées :

$$E_{Consommation\_électrique\_totale}\ t(i) = [Date\ ;$$

$$Statistiques\ Puissance_{\ Phase\ 1}\ t(i),\dots, Statistiques\ Puissance_{\ Phase\ 1}\ t(i-48),$$
$$Statistiques\ Variations_{\ Phase\ 1}\ t(i),\dots, Statistiques\ Variations_{\ Phase\ 1}\ t(i-48)$$
$$Données_{Réelles}\ \_Météo_{Station}\ \ t(i),\dots., Données_{Réelles}\ \_Météo_{Station}\ \ t(i-48),$$
$$Données_{Prédites}\ \_Météo_{Station}\ \ t(i+1),\dots, Données_{Prédites}\ \_Météo_{Station}\ \ t(i-48)$$

$$Année_j\ , Mois_m\ , Jour_j, Jour\_Semaine_{js}\ , Heure_h, Jour\ fériés_{0.1}, Vacances\ scolaires_{0.1},$$

$$Classes_c\ ] \qquad \textbf{(4.21)}$$

Energie à prédire	Sortie apprentissage	Moyenne	Déviation standard	Répartition %
Classification de l'énergie électrique, Secteur résidentiel	0	0.5817	0.0504	30
	1	0.821	0.1371	11
	2	1.2571	0.2616	9
	3	2.0145	0.3985	11
	4	3.0272	0.1062	5
	5	3.3657	0.5458	35

Tableau 15 : Répartition des différentes familles caractérisant la consommation électrique dans nos différents cas d'étude

La production-décentralisée-solaire

A partir de l'énergie produite, de la luminosité locale à Sierre et de la luminosité régionale à Sion distante de 15 km du site de test, nous réalisons un apprentissage non supervisé (Figure 53). L'algorithme converge vers 11 classes décrites dans le tableau 16.

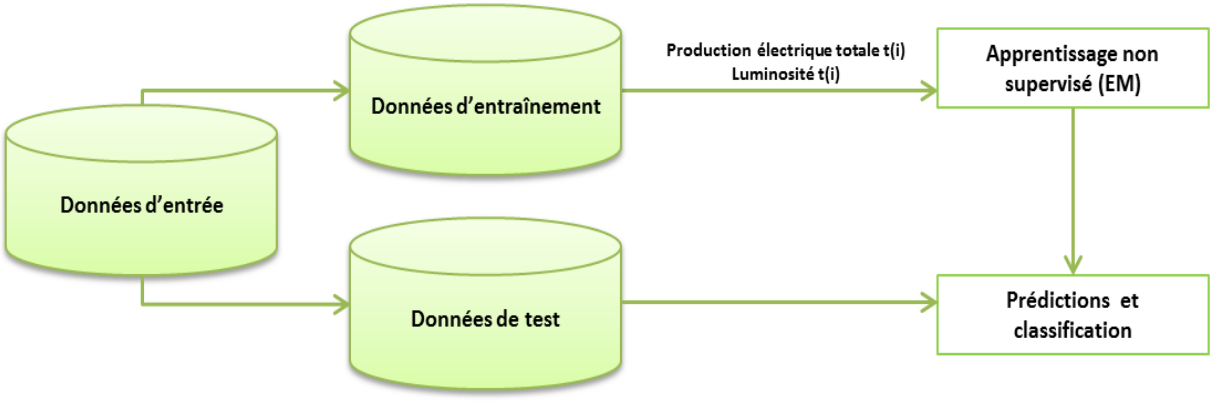


Figure 59 : Méthodes de classification des données pour la production décentralisée solaire

Energie à prédire	Sortie apprentissage	Moyenne	Déviation standard	Répartition %
Production décentralisée solaire	0	1.0798	0.0794	3
	1	2.9921	1.3607	6
	2	7.8636	3.8702	9
	3	15.5819	6.9769	5
	4	21.4828	9.7723	5
	5	46.0692	14.7446	35
	6	46.7811	15.2881	16
	7	84.1565	12.5396	7
	8	101.8406	24.4283	17
	9	140.2383	13.679	9
	10	151.414	9.0872	4

Tableau 16 : Classes décrites par une moyenne, une déviation standard et leurs répartitions dans le set d'entraînement caractérisant la production solaire en kWh

Nous regroupons les familles les plus proches pour faciliter la classification. Nous regroupons les 11 familles en 5 familles que nous avons décrites dans le tableau 15 et schématisé par la figure 54.

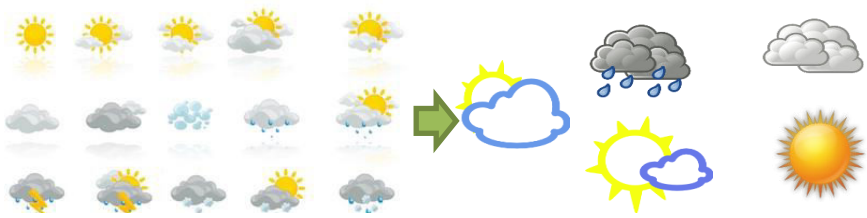


Figure 60 : Représentation du regroupement des classes dans la classification de la production décentralisée solaire

Energie à prédire	Sortie apprentissage	Moyenne	Déviation standard	Répartition %
Production décentralisée solaire	0	1.0798	0.0794	18
		2.9921	1.3607	
		7.8636	3.8702	
	1	15.5819	6.9769	10
		21.4828	9.7723	
	2	46.0692	14.7446	24
		46.7811	15.2881	
	3	84.1565	12.5396	24
		101.8406	24.4283	
	4	140.2383	13.679	13
		151.414	9.0872	

Tableau 17 : Nouvelles classes décrites par une moyenne, une déviation standard et leurs répartitions dans le set d'entraînement

Pour la classification, nous devons nous assurer d'une répartition équitable entre les différentes classes en particulier celle qui requiert un intérêt particulier comme par exemple le pic de puissance de

consommation ou de production. Dans le cas de la production solaire, nous remarquons par exemple que les heures de la classe 10 ne représentent que 4% des données d'entraînement mais 10 % après regroupement. Sans ce regroupement, une étape nommé « boostsrap » est nécessaire. Elle permet de recopier une partie des données sous représentés pour permettre à nos algorithmes itératifs d'avoir une part plus équitable dans les données d'entraînement. Nous comprenons que ces étapes sont particulièrement adaptées à notre problématique de prédictions de variations brutales de la consommation ou de la production et que des algorithmes comme le AdaBoost ou le Gradient Boosting Tree décrit dans le chapitre 2 qui intègre ces mécanismes automatiquement seraient particulièrement adaptés.

En sortie de la classification, chaque heure de notre vecteur d'entrée est caractérisée par une classe et un historique des données de productions et météorologiques sont également créés :

$$
\begin{aligned}
&E_{ENRE}\ t(i+1) = [\text{Date} ; \\
&\text{Statistiques Puissance}_{\text{Phase 1}}\ t(i), \ldots, \text{Statistiques Puissance}_{\text{Phase 1}}\ t(i-48), \\
&\text{Statistiques Variations}_{\text{Phase 1}}\ t(i), \ldots, \text{Statistiques Variations}_{\text{Phase 1}}\ t(i-48), \\
&\text{Données}_{\text{Réelles}}\text{\_Météo}_{\text{Station-Local}}\ t(i), \ldots, \text{Données}_{\text{Réelles}}\text{\_Météo}_{\text{Station-Local}}\ t(i-48), \\
&\text{Données}_{\text{Réelles}}\text{\_Météo}_{\text{Station}}\ t(i), \ldots, \text{Données}_{\text{Réelles}}\text{\_Météo}_{\text{Station}}\ t(i-48), \\
&\text{Données}_{\text{Prédites}}\text{\_Météo}_{\text{Station}}\ t(i+1), \ldots, \text{Données}_{\text{Prédites}}\text{\_Météo}_{\text{Station}}\ t(i-48), \\
&\text{Année}_{j}, \text{Mois}_{m}, \text{Jour}_{j}, \text{Semaine}_{0.1}, \text{Heure}_{h}, \\
&\text{Classes}_{c}\ ]
\end{aligned}
\qquad \textbf{(4.22)}
$$

4.5 Réduction du nombre de variables

Comme énoncé dans le préambule, l'une des questions liée à la prédiction énergétique de flux est le nombre de données et ainsi le nombre de capteurs dont nous avons besoin pour maintenir un niveau acceptable de prédictions.

En sortie des différentes implémentations des arbres de décision, il est possible de sortir les variables qui sont utilisées pour la construction des arbres. Ainsi, nous calculons pour chaque flux énergétique le nombre de fois qu'une variable a été utilisée pour la création des ensembles d'arbres de décision au niveau 1, au niveau 2 et au niveau 3 :

$$Evaluation = Nb\_Utilisation_{Niveau1} + Nb\_Utilisation_{Niveau2} + Nb\_Utilisation_{Niveau3} \qquad \textbf{(4.23)}$$

Un classement par ordre décroissant est ensuite réalisé et le calcul prédictif par flux sera réalisé en faisant diminuer le nombre de variables basées selon leurs importances dans la construction des arbres.

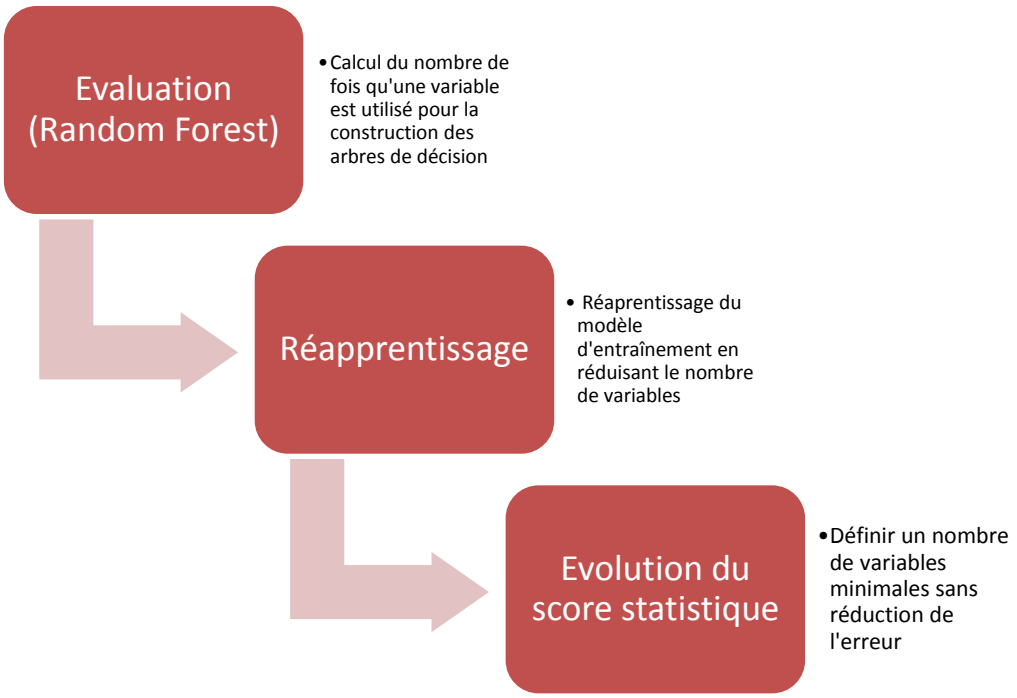


Figure 61 : Méthode d'analse de l'impact du nombre de variables d'entrée sur la précision de la prédiction

4.6 Conclusion

Nous avons défini le traitement en sortie des différents capteurs et l'aggrégation réalisée des caractéristiques tirées des capteurs pour prédire nos différents flux énergétiques.

Nous pouvons repartir de **la problématique globale pour la prédiction de n'importe quel flux énergétique :**

- ⇨ **Quelles données ?** Nous avons défini les données d'entrée par flux énergétique avec l'étude des variations de puissance à la seconde et les statistiques liés directement à la puissance. Pour le cas de la consommation électrique dans le secteur résidentiel, une pré-étape d'identification non intrusive des appareils est réalisée. Cette identification permet la prédiction de l'eau chaude et du chauffage de manière non intrusive également. Enfin, les données météorologiques par le biais de 4 stations cantonales sont corrigées et utilisés également comme données d'entrée.

- ⇨ **À quelles fréquences de collecte ?** L'impact de la collecte des données à la minute ou à l'heure pour une prédiction pour l'heure suivante est mesuré.

- ⇨ **Avec quel système d'information (appareil, stockage…) ?** Nous utilisons différents connecteurs avec du modbus ou du Zigbee comme protocole de communication pour collecter les données à la seconde. Elles sont ensuite stockées dans des bases de données MongoDB. Les données météorologiques sont accessibles par API grâce à l'application de MétéoSuisse. Elles sont stockées dans une autre base de données, influx DB.

- ⇨ **Avec quelles méthodes de prédictions ?** Des méthodes linéaires et non linéaires par apprentissage sont testées.

- ⇨ **Avec quels modèles mathématiques ?** Des algorithmes linéaires (Régression linéaire, ARIMA) et non linéaires (MLP, PNN, Random Forest et Gradient Boosteed Tree) sont testés.

- ⇨ **Pour quels niveaux de prédictions ?** Les résultats sont présentés dans le chapitre 5.

- ⇨ **Sur quel horizon ?** Nous avons choisi comme horizon de prédiction l'heure suivante pour réaliser notre bilan énergétique prédictif.