Méthodologie pour estimer la consommation d'énergie dans les bâtiments en utilisant des techniques d'intelligence artificielle

Subodh Paudel

Thèse de Doctorat

Subodh PAUDEL

Mémoire présenté en vue de l'obtention du

grade de Docteur de l'École des Mines de Nantes

sous le label de L'Université Nantes Angers Le Mans

École doctorale : *SCIENCES POUR L'INGENIEUR, GEOSCIENCES, ARCHITECTURE*

Discipline : Thermique, Energétique et Génie des Procédés
Spécialité : *Energétique*
Unité de recherche : *GEPEA UMR CNRS 6144*

Soutenue le 22 Septembre 2016
Thèse N° : *2016 EMNA 0237*

Methodology to Estimate Building Energy Consumption Using Artificial Intelligence

PRESIDENT DU JURY :
BENELMIR Riad, Professeur, Faculté des Sciences et Techniques, Université de Lorraine, France

RAPPORTEURS DE THESE :
MAGOULES Frédéric, Professeur, CentraleSupelec, Université Paris-Saclay, France
VIRGONE Joseph, Professeur, INSA Université Lyon 1, France

MEMBRE DU JURY :
LASSUE Stéphane, Professeur, Faculté des Sciences Appliquées, Université d'Artois, France

INVITE :
ELMTIRI Mohamed, Docteur, Veolia Environnement Recherche et Innovation, Limay, France

DIRECTEUR DE THESE :
LE CORRE Olivier, Docteur (HDR), Ecole des Mines de Nantes, France
CO-DIRECTEUR DE THESE :
KLING Wil, Professeur, Technische Universiteit Eindhoven, Netherlands (Regretté)
KAMPHUIS René, Professeur, Technische Universiteit Eindhoven, Netherlands
RESPONSABLE SCIENTIFIQUE :
LACARRIERE Bruno, Docteur (HDR), Ecole des Mines de Nantes, France

Acknowledgements

I would like to express my gratitude to all those who support during my PhD studies. I would like to thank first SELECT+ program for funding this PhD and helps to pursue my career.

I am indebted to my supervisors Professor Olivier Le Corre and Professor Bruno Lacarrière at Ecole des Mines de Nantes, Professor Wil L. Kling (Late), Professor René Kamphuis and Professor Phuong H. Nguyen at Eindhoven University of Technology and Dr. Mohamed Elmitri at Veolia Recherche et Innovation (VERI). I appreciate their guidance, insights, comments and directions in the research and giving me an opportunity to work in Ecole des Mines de Nantes, Eindhoven University of Technology and Veolia Recherche et Innovation.

I gratefully acknowledge Industrial partner, Veolia, especially Isabelle Verdier, Dr. Mohamed Elmitri, and Dr. Stéphane Couturier. They provide constructive feedback, direction of industrial research and organizing various fruitful meetings.

My personal appreciation to Prof. Frèderic Magoules from Université Paris-Saclay, France and Prof. Joseph Virgone from Université Lyon 1, France for their review and suggestions to improve further corrections. I would like to thank president of jury Prof. Riad Benelmir from Universitè de Lorraine, France and Prof. Stéphane Lassue from Université d'Artois, France for their acceptance to participate in the jury.

I again thank all the people in the Department of Energy System and Environment (DSEE) in Ecole des Mines de Nantes and Department of Electrical Engineering at Energy System Group in Eindhoven University of Technology for their pleasant working environment. I would like to express special thanks to Professor Olivier Le Corre for his guidance, motivation and help in difficult times and during the whole PhD thesis.

I thank my friends- Bishal, Umesh, Sudeep, Rajani and Bhuwan for their help to improve the manuscript. Finally, I would like to express my deepest thanks to my parents, wife Muna, daughter Melinsha, brothers and sisters for their encouragement, endless support and love during my studies.

Subodh PAUDEL

Nantes, June 2016

Dedication

To my family

Extended Abstract in French

Méthodologie pour estimer la consommation d'énergie dans les bâtiments en utilisant des techniques d'intelligence artificielle

S. Paudel

Pour une société de services énergétiques opérant sur les réseaux de chaleur, il est essentiel d'avoir un estimé de la courbe de charge de son réseau en fonction des prévisions météorologiques et du comportement de ces clients. Or, les bâtiments à basse consommation d'énergie rendent de moins en moins précis les anciennes approches, telle que la droite de charge (relation linéaire entre la charge du réseau et la température extérieure), principalement du fait de l'inertie de ces nouvelles enveloppes, voir la **figure 1**. La finalité de cette thèse est de proposer une méthodologie pour estimer le besoin de chaleur de tels bâtiments.

L'estimation du besoin d'un bâtiment peut être abordée de différentes manières :

- Modèle de connaissance (désigné aussi par *boîte blanche*) : en partant des équations de la physique et en connaissant l'ensemble des éléments constituant le système, il est possible d'établir un modèle.
- Modèle de comportement : en partant d'essais spécifiques sur le système, il est possible d'établir un modèle de type entrées/sorties. Il existe plusieurs types de méthodes pour établir un tel modèle : méthode d'identification (moindre carrés), méthode d'auto-régression (type ARX), ou les méthodes d'intelligence artificielle (**AI**) (Réseaux de neurones **ANN**, Machines à vecteur de support **SVM**, Arbre de décision **DT**, ou Forêt aléatoire **RF**), présentées en **annexe B**. Ces modèles sont désignés par l'appellation *boîte noire*. Le tableau 2.5 présente une synthèse de 23 auteurs ayant utilisé l'une de ces méthodes.
- Modèle intermédiaire (désigné par *boîte grise*) : en reprenant les équations de la physique, certains paramètres sont identifiés sur le système étudié.

L'approche *boîte blanche* n'a pas été retenue dans cette thèse car l'opérateur du réseau de chaleur n'a pas nécessairement les données sur la composition des bâtiments alimentés. L'approche *boîte grise* n'a pas été retenue car l'opérateur ne pourrait que très difficilement effectuer des essais chez son client, notamment pour des immeubles résidentiels. L'approche *boîte noire* est donc celle qui semble la plus opportune. Du fait, de complexes interactions entre la température extérieure, le rayonnement solaire (sur les murs et à travers les fenêtres), l'inertie du bâtiment, l'usage et le pilotage de la fourniture de chaleur, les méthodes d'intelligence artificielle ont été retenues.

L'état de l'art des applications des techniques AI appliquées dans les bâtiments est l'objet du **chapitre 2**. Des exemples d'application sur des bâtiments basse consommation n'ont pas été identifiés. Définir une approche méthodologique mettant en œuvre une technique AI pour la prédiction d'un besoin de chaleur d'immeuble basse consommation (fortement inertiel) est l'objet des travaux présentés dans ce manuscrit.

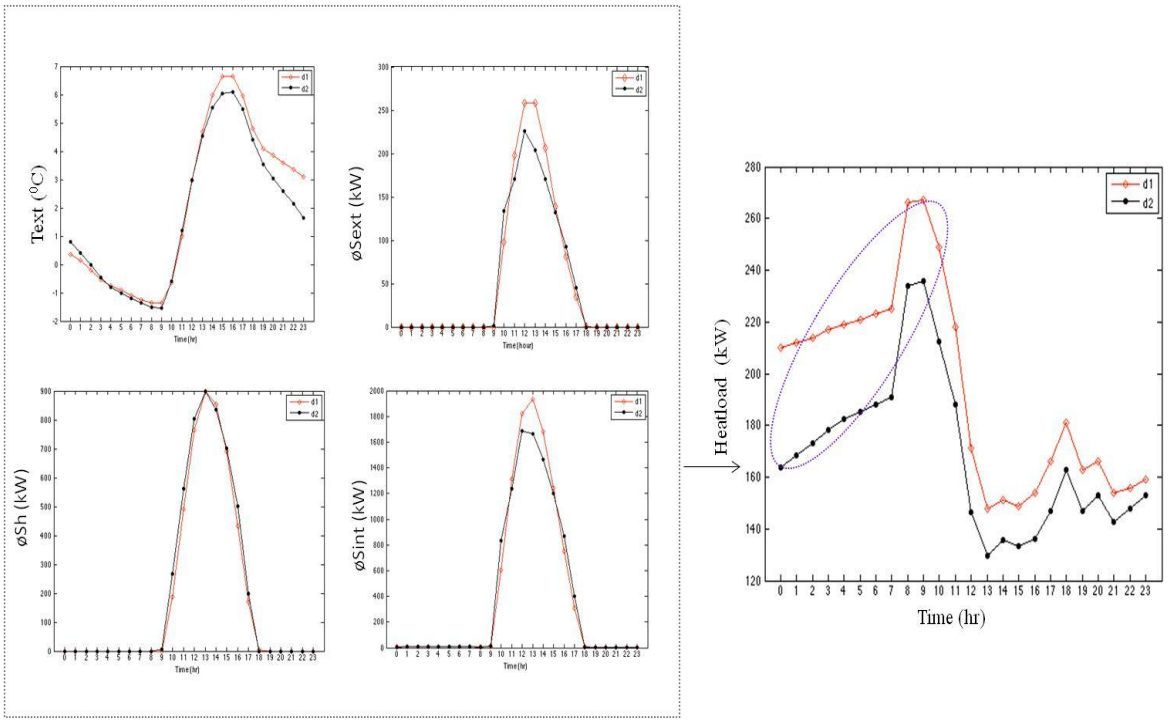

Figure 1 : Besoin de chaleur d'un bâtiment basse consommation (à droite) en fonction des données météorologiques (à gauche température extérieure et rayonnement solaire) pour deux journées différentes D1 et D2.

Le point de départ de cette thèse est :

- L'historique des consommations (horaire moyennée) du client est disponible ;
- L'historique des conditions climatiques (horaire moyennée) est disponible ;
- Une estimation de l'usage (horaire moyenné) du bâtiment est disponible (occupation, température de consigne, etc…) ;
- Une prédiction des conditions météorologiques est disponible.

Le caractère stochastique est *de facto* hors du cadre de ce travail.

Le cadre méthodologique proposé dans cette thèse contient une préparation de la base de données (constituée des historiques de consommations et des conditions météorologiques ainsi que de l'usage du bâtiment), voir la **figure 2**.

Il existe deux possibilités pour mettre en œuvre une technique AI :

1. Utiliser l'ensemble de la base de données et créer un unique modèle AI. Dans le corps du manuscrit, cette approche est désignée par « all data ».
2. Effectuer une présélection dans la base de données en lien avec les conditions climatiques à prédire. Un modèle AI est établi pour chaque nouvelle condition climatique. Dans le corps de ce manuscrit, cette approche est désignée par « relevant data ».

Cette présélection peut être effectuée de différentes manières :

- La méthode « homme de l'art » du degré jour unifié, adaptée pour des bâtiments plutôt anciens, fortement dépendant de la température extérieure (mal isolés et avec des infiltrations)
- Des méthodes mathématiques de traitement du signal permettant de corréler des signaux.

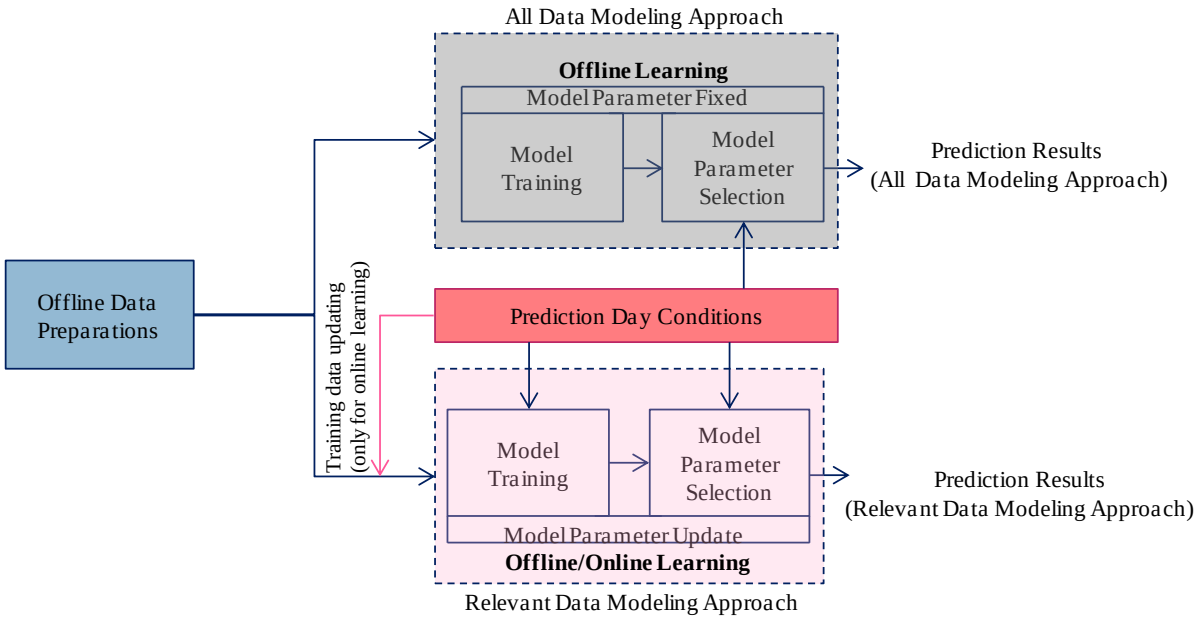

Figure 2 : Cadre méthodologique des approches « all data » et « relevant data »

L'une des difficultés majeures dans la prédiction du besoin de chaleur est de tenir compte de la constante de temps de l'enveloppe du bâtiment.

L'historique de la consommation d'un bâtiment contient l'ajustement, en boucle fermée, de cette consommation au suivi des températures intérieures de consigne, en fonction de son usage. Ainsi, pour des conditions extérieures identiques et pour des valeurs de consignes constantes, le besoin de chaleur n'est pas le même après un changement de température de consigne ou une heure après ce changement, voir la **figure 3**. Pour tenir de ce phénomène, un modèle, désigné par pseudo-dynamique, est l'une des propositions de cette thèse, voir le **chapitre 3** pour plus de détails.

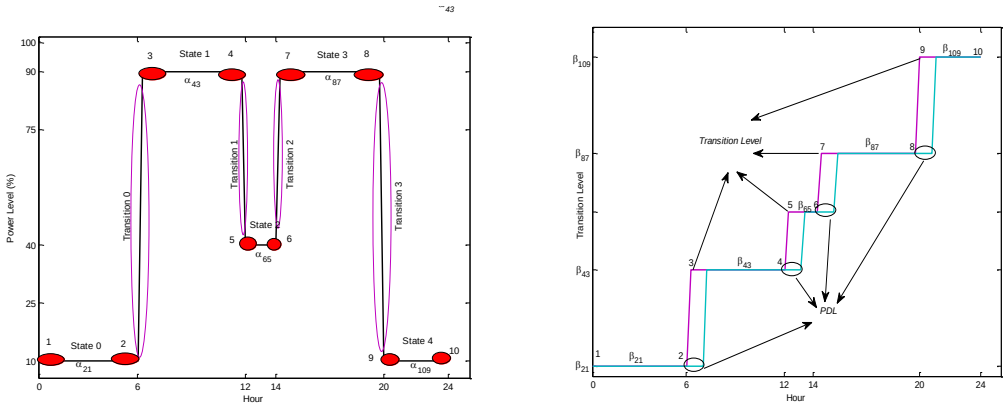

Figure 3 : Changement de consigne et modèle de transition pseudo-dynamique

Ce modèle pseudo-dynamique doit permettre d'indiquer qu'un autre effet (inertie thermique interne) est en cours. La proposition est de créer un **vecteur de transition** indiquant les différents seuils de changement de consigne (ou d'occupation). L'aspect inertiel est lui indiqué en « décalant » ce vecteur de manière à ce qu'en entrée du module « intelligence artificielle » le comportement dynamique puisse être « appris ». Dans le manuscrit, cette connaissance *à priori* du comportement thermique est donc mise en entrée. Ce modèle est décrit par une réponse du premier ordre du bâtiment mais fonctionne aussi pour une réponse oscillante : l'important est

qu'en entrée, les plages horaires soient bien distinguables. Ce modèle de transition est aussi appliqué à l'occupation qui, pour la même raison, pourrait avoir des entrées « constantes » mais dont le besoin de chaleur serait à distinguer (ex : l'occupation des salles de cours dans une école).

L'aspect dynamique plus long dû à l'enveloppe du bâtiment est mis en évidence sur la **figure 4**, correspondant à l'historique de la température extérieure et du flux radiatif des 5 jours précédents les jours D1 et D2 de la **figure 1**.

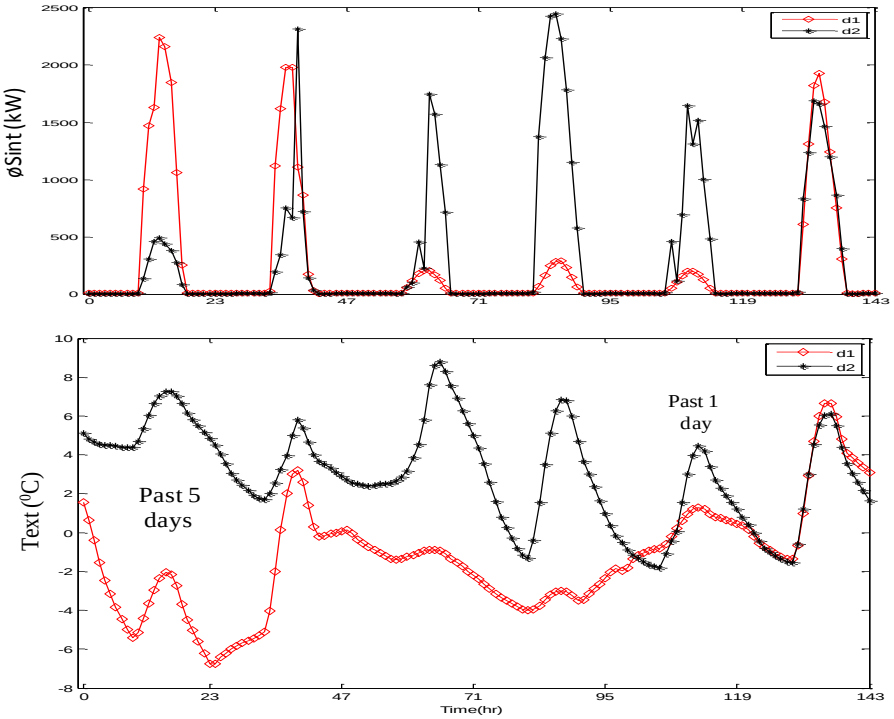

Figure 4 : Historique des 5 jours précédents la prédiction des jours D1 et D2.

Pour intégrer cet aspect deux éléments sont proposés :

- D'une part, calculer les valeurs journalières des variables climatiques sur les jours précédents. Ainsi, il est proposé de mettre en entrée la température journalière extérieure, le flux solaire journalier impactant les murs et celui traversant les fenêtres, voir le **chapitre 3** pour plus de détails.
- D'autre part, faire une décomposition en ondelettes des variables climatiques, voir le **chapitre 3** pour plus de détails.

La **figure 5** schématise l'ensemble de la préparation des données effectuée.

Comme indiqué précédemment deux approches ont été développées : « all data » et « relevant data ». Néanmoins, un travail plus important a été réalisé sur l'approche « relevant data » car les résultats obtenus ont été meilleurs, comme décrit ci-après, et au **chapitre 4**.

L'approche « relevant data » repose sur une sélection de jours dans la base de données ressemblant le plus possible aux conditions météorologiques à prévoir (incluant l'historique). Pour cette sélection il a été envisagé des méthodes mono-variables (comme le degré jours ***HDD*** ou une modification ***mHDD***) et multi-variables (comme le critère de la distance de Fréchet ***FD*** et celui de la déformation temporelle dynamique, ci après ***DTW***) pour les variables climatiques (température extérieure, flux solaire sur les murs et traversant les fenêtres), voir le **chapitre 3**. Quatre techniques d'intelligence artificielle ont été mises en œuvre : réseau de neurones ***ANN***, Machines

à vecteur de support ***SVM***, Arbre de décision ***DT***, ou Forêt aléatoire ***RF*** (l'**annexe B** est une introduction à ces techniques). La méthodologie est décrite dans le **chapitre 3**.

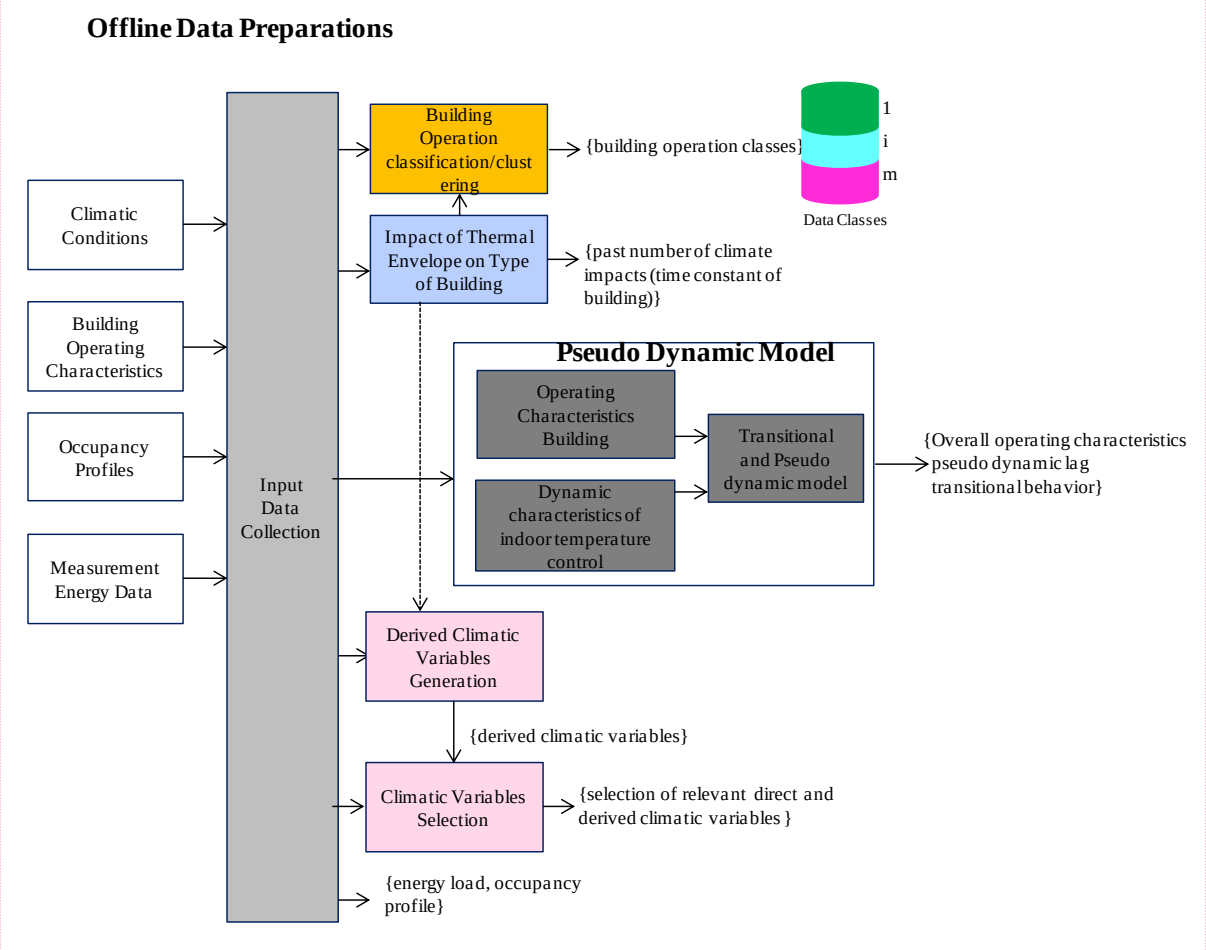

Figure 5 : Préparation des données

L'approche « all data » est comparée avec l'approche « relevant data » sur le **couple** (***DTW*** et ***SVM),*** association donnant les meilleures résultats.

La mise en œuvre de cette méthodologie s'effectue en deux étapes distinctes : la première étape est numérique (la base de données est générée en utilisant TRNSys), voir le **chapitre 4**, la seconde est une application sur le bâtiment de l'Ecole des Mines de Nantes, voir le **chapitre 5**.

Pour la première étape de mise en œuvre, 6 configurations de bâtiments ont été imaginées en concertation avec le centre Veolia Environnement Recherche et Innovation, voir le **tableau 1**. Les quatre premiers cas correspondent au même bâtiment, avec différents degrés d'isolation et différentes constantes de temps d'enveloppes (cas 1 30h, cas 2 53h, cas 3 76h, cas 4 119h, cas 4* 210h), pour un usage de type résidentiel, voir la **figure 6 (gauche)**. Ces quatre premiers cas correspondent à des bâtiments d'une consommation spécifique conventionnelle à une basse consommation. Le cas 5 est un immeuble de bureau, avec un profil d'occupation spécifique, voir la **figure 6 (droite)**, avec une constante de temps de 210 h pour le cas 5 et 219 h pour le cas 5*. Le cas 6 correspond à un immeuble avec un autre type d'occupation de type centre commercial, voir la **figure 6 (bas)**, avec une constante de temps de 219 h pour les cas 6 et 6*.

Descriptions	Case 1	Case 2	Case 3	Case 4[1*]	Case 5[1*]	Case 6[1*]
Floor Surface (m^2)	3333	3333	3333	3333	1372	10521
Number of floor	6	6	6	6	10	1
Total surface (m^2)	20000	20000	20000	20000	13720	10521
External wall South (m^2)	4000	4000	4000	4000	4450	330
External wall North (m^2)	4000	4000	4000	4000	4450	330
External wall West (m^2)	1250	1250	1250	1250	-	330
External Wall East (m^2)	1250	1250	1250	1250	-	330
Floor height (m)	3.2	3.2	3.2	3.2	3.2	3.2
U-value of walls, roofs and floors ($W/m^2.K$)	2	1	0.5	0.25	0.25	0.25
U-value of glazing $W/m^2.K$	2.95	2.95	2.95	1.76	1.43	1.43
Glazing rate on each external wall (%)	25	25	25	25	30	30
Building Type	Residential	Residential	Residential	Residential	Office	Commercial
Single/Multi-zone Type	Single	Single	Single	Single/Multi	Single/Multi	Single/Multi

Tableau 1 : Principales caractéristiques des bâtiments

La modélisation sous TRNSys est une modélisation mono-zone pour les cas sans le symbole (*). Pour les cas 4, 5 et 6, une modélisation multizone a été effectuée, symbolisée avec (*).

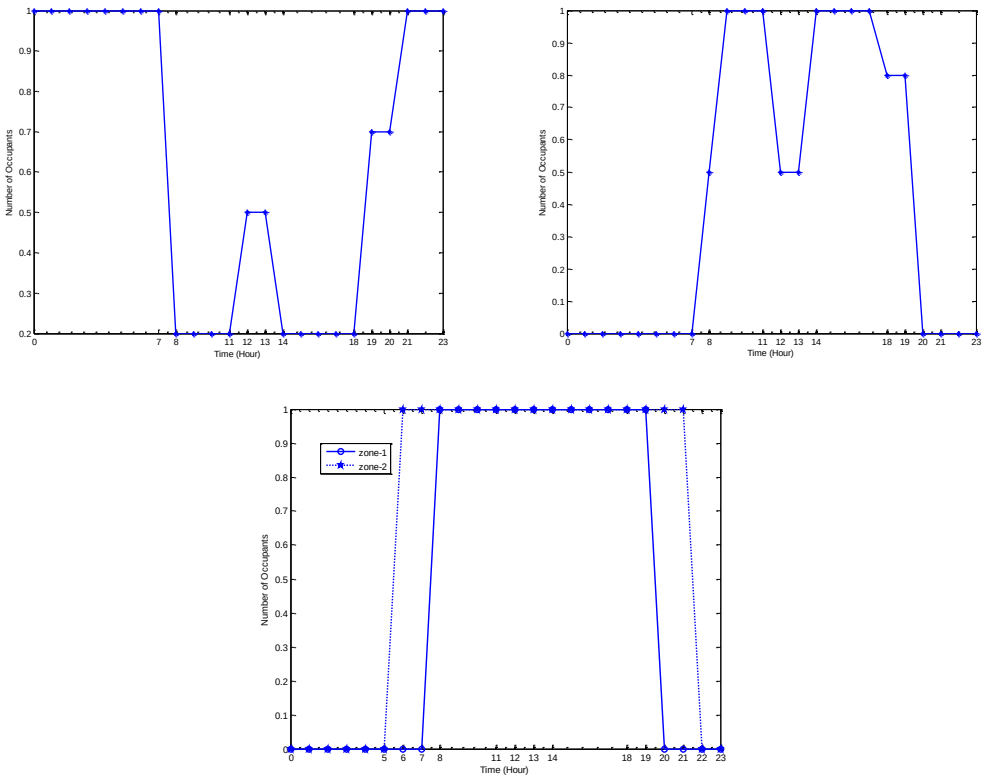

Figure 6 : Profil d'occupation selon les cas

Les profils des flux (éclairage, ventilation) et des conditions opératoires sont tracés sur la **figure 7** (cas 1-4), la **figure 8** (pour le cas 5 et 5*) et la **figure 9** (pour le cas 6 et 6*).

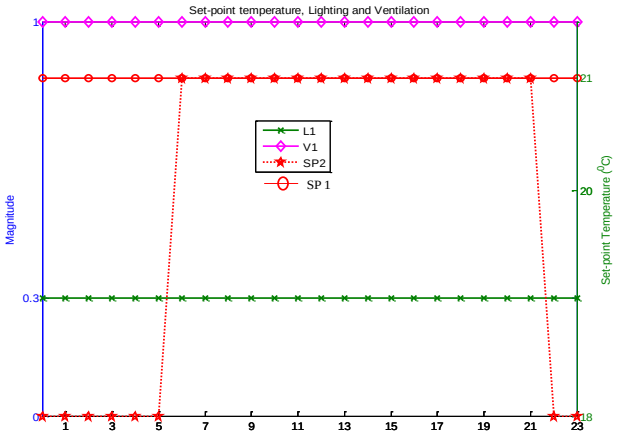

Figure 7 : Profil des flux (éclairage, ventilation) et conditions opératoires pour les cas 1 à 4(4*)

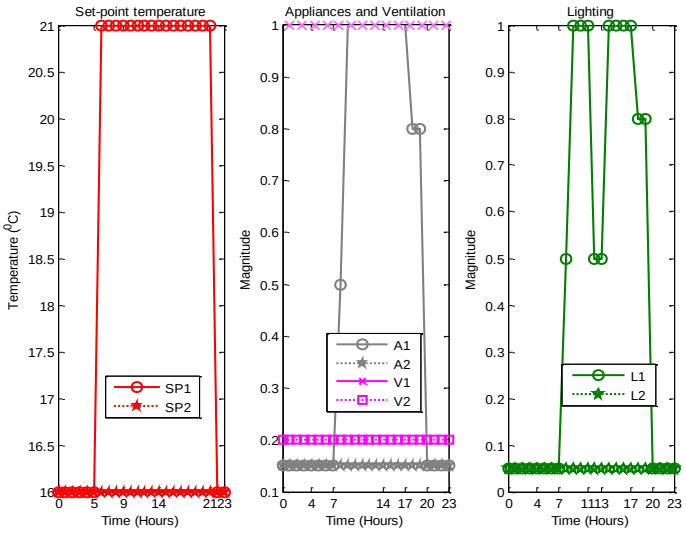

Figure 8 : Profil des flux (éclairage, ventilation) et conditions opératoires pour les cas 5 et 5*

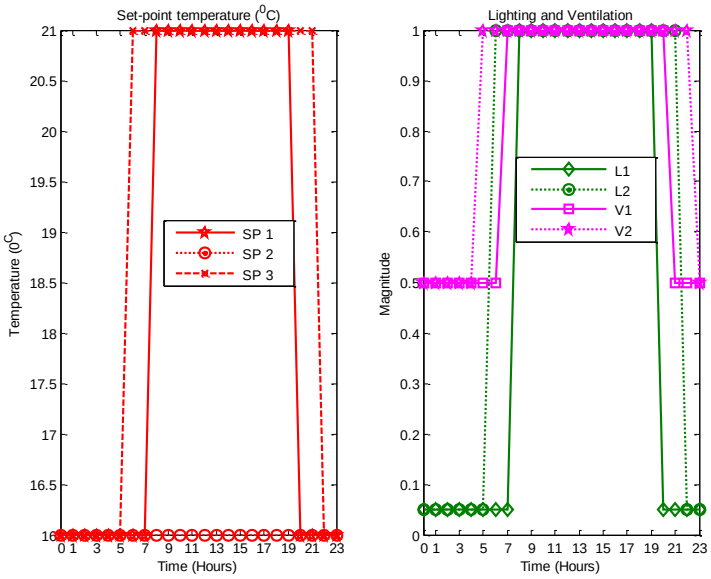

Figure 9 : Profil des flux (éclairage, ventilation) et conditions opératoires pour les cas 6 et 6*

Le besoin de chaleur concerne le bâtiment dans son ensemble et correspond à la puissance thermique délivrée par la sous-station. La demande annuelle spécifique de l'ensemble des cas est indiquée sur la **figure 10**, pour l'ensemble des cas et pour 4 fichiers météorologiques (correspondant à 4 villes : Paris, Lille, Lyon et Clermont-Ferrand). La demande spécifique des cas 1 à 4 passe de 80 kWh/m²/an à 20 kWh/m2/an, permettant de représenter des bâtiments de consommation dite conventionnelle et des bâtiments basse consommation. Les cas 5 et 6 présentent des demande spécifiques de l'ordre de 20-25 kWh/m²/an, correspondant à des bâtiments basse consommation.

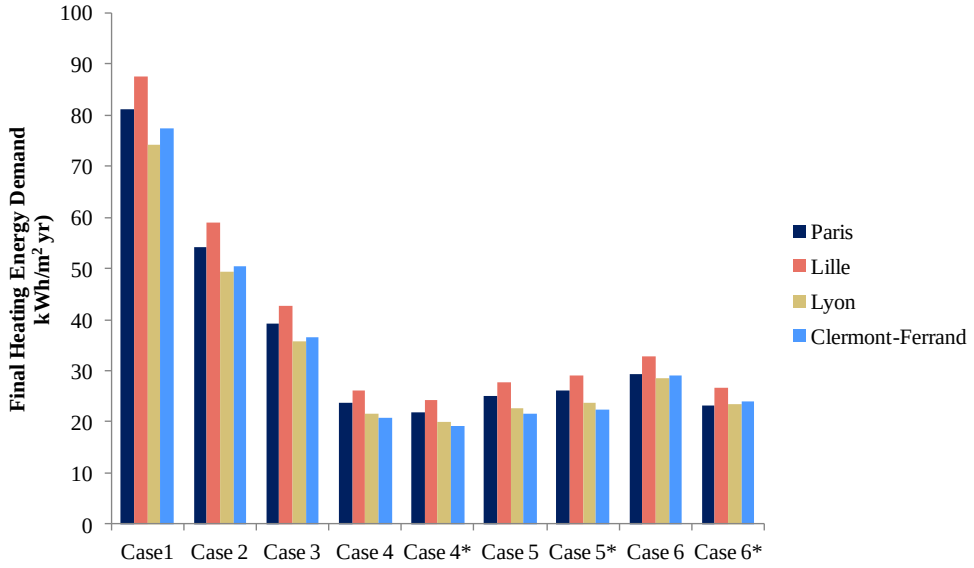

Figure 10 : Demande annuelle spécifique pour l'ensemble des cas : résultats TRNSys.

Préalablement à la mise en ouvre d'une méthode AI, il faut définir la composition des entrées. Huit séries d'entrées ont été considérées, voir le **tableau 2**.

Name		Description	Input Features Scenarios							
			S1	S2	S3	S4	S5	S6	S7	S8
Outputs	P (t)	Heat Load (kW)	×	×	×	×	×	×	×	×
Inputs	Text(t)	External temperature (^{0}C)	×	×	×	×	×	×	×	×
	Text (t-1)	External temperature at 1 hour time delay (^{0}C)	×	×	×	×	×	×	×	×
	øSh(t)	Horizontal solar radiation (kW)	×	×	×	×	×	×	×	×
	øSh(t-1)	Horizontal solar radiation at 1 hours delay (kW)	×	×	×	×	×	×	×	×
	øSh(t-2)	Horizontal solar radiation at 2 hours delay (kW)			×	×	×	×	×	×
	øSext(t)	Solar gain transmitted through window (kW)		×	×	×	×	×	×	×
	øSext(t-1)	Solar gain transmitted through window at 1 hour delay (kW)		×	×	×	×	×	×	×
	øSext(t-2)	Solar gain transmitted through window at 2 hours delay (kW)			×	×	×	×	×	×
	øSint(t)	Solar gain on wall (kW)		×	×	×	×	×	×	×
	øSint(t-1)	Solar gain on wall at 1 hour delay (kW)		×	×	×	×	×	×	×
	øSint(t-2)	Solar gain on wall at 2 hours delay (kW)			×	×	×	×	×	×
	occup	Occupancy profile [0 1]				×	×	×	×	×
	trans	Transitional attributes [0.2 1]					×	×	×	×
	PDL-1	Pseudo dynamic lag 1 [0.2 1]						×	×	×
	PDL-2	Pseudo dynamic lag 2 [0.2 1]							×	×
	Text_TDM	Temporal moving average of external temperature (^{0}C)								×
	øSint_TDM	Temporal moving average of solar gain on wall (kW)								×

Tableau 2 : Composition des entrées pour les cas 1-4

Le besoin de chaleur du cas 3, respectivement 4 (bâtiment faible et basse consommation, profil d'occupation de type résidentiel) est tracé sur la **figure 11**, respectivement sur la **figure 12**. Dans la légende, « actual » correspond aux résultats de TRNSys ; « prediction S7 et S8 » correspondent aux résultats obtenus avec le **couple** (DTW, SVM) avec les séries d'entrées S7 et S8. Trois paires de journées sont reportées et correspondent à trois mois différents.

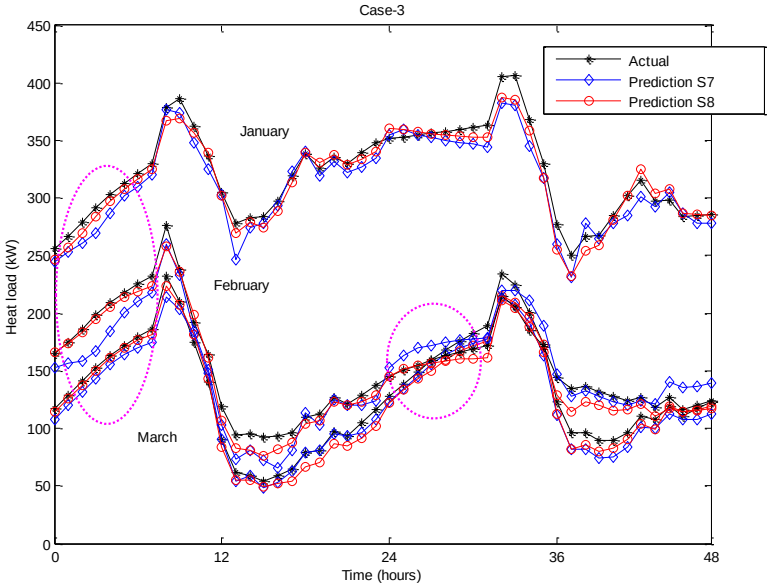

Figure 11: Besoin de chaleur du bâtiment Cas-3 (faible consommation)

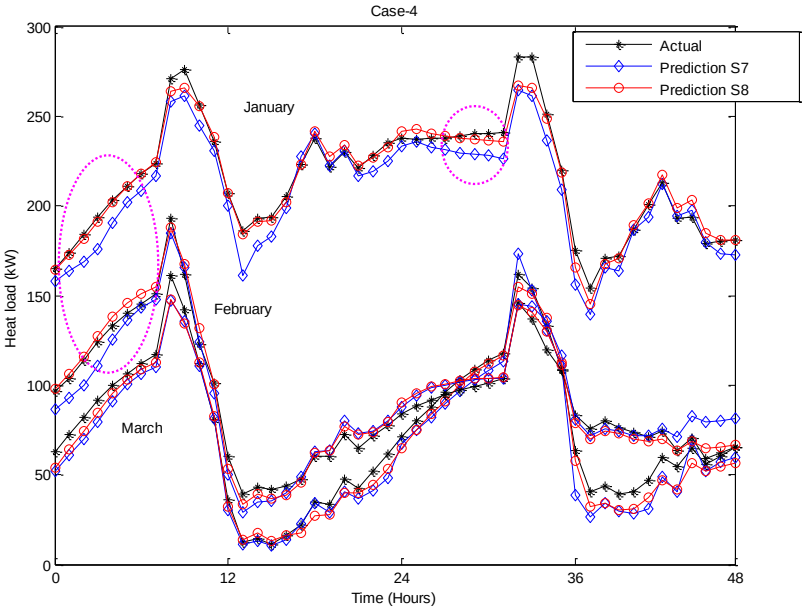

Figure 12: Besoin de chaleur du bâtiment Cas-4 (basse consommation)

Les pics de consommation et les comportements du besoin de chaleur sont bien décrits. Le coefficient de détermination, l'erreur quadratique moyenne sont donnés pour l'ensemble des 8 séries d'entrées pour une année dans le **tableau 3**. La série d'entrées S8 donne des coefficients

presque unitaires indiquant une bonne prédiction. Ceci est particulièrement pertinent pour le cas 4 (bâtiment à basse consommation).

Models	Case 1				Case 2				Case 3				Case 4			
	Median		Overall		Median		Overall		Median		Overall		Median		Overall	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
S1	0.74	20.5	0.98	23.4	0.71	17.7	0.96	23.4	0.67	16.2	0.96	18.3	0.69	13.2	0.93	16.6
S2	0.75	20.1	0.98	23.0	0.69	17.5	0.97	20.5	0.69	16.1	0.96	18.1	0.68	14.1	0.94	15.5
S3	0.77	19.6	0.98	22.8	0.69	17.3	0.97	20.7	0.70	15.9	0.96	18.1	0.70	13.6	0.94	15.5
S4	0.88	14.7	0.99	18.5	0.88	12.3	0.98	16.2	0.90	10.4	0.98	13.6	0.93	7.2	0.97	9.7
S5	0.89	14.2	0.99	17.8	0.88	12.2	0.98	16.6	0.91	10.1	0.98	13	0.94	7.3	0.98	9.3
S6	0.88	14.5	0.99	19.5	0.86	13.1	0.98	17.2	0.90	10.1	0.98	13.4	0.93	7.2	0.97	10.6
S7	0.89	14.1	0.99	17.8	0.88	12.7	0.98	16.2	0.91	10.0	0.98	13.6	0.93	6.9	0.98	8.9
S8	0.93	10.9	0.99	14.1	0.92	9.5	0.99	13.6	0.96	6.0	0.99	8.5	0.97	3.9	0.99	6.0

Tableau 3 : Coefficients de détermination et erreur quadratique moyenne pour le couple (DTW et SVM)

De plus la distribution mensuelle de ces deux coefficients montre que l'erreur la plus importante correspond au mois d'Avril (mois de fin de saison de chauffe), ce qui n'est pas le mois le plus significatif pour un opérateur de réseaux de chaleur, voir la **figure 13**.

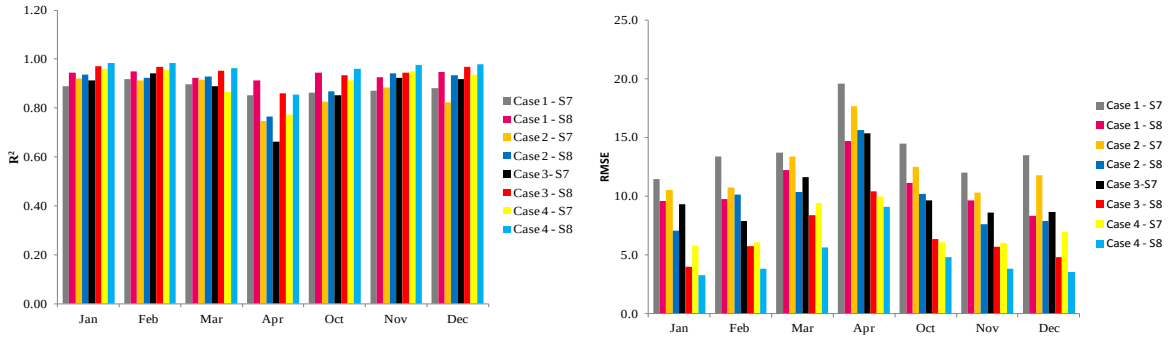

Figure 13 : Distribution mensuelle de l'erreur de prédiction selon les cas 1-4 pour le couple (DTW-SVM)

L'erreur pour l'ensemble des couples (méthode de sélection, méthode d'intelligence artificielle) est reportée dans le **tableau 4** pour la série d'entrées S8. On observe que :

- la méthode de sélection mHDD fonctionne bien pour les bâtiments de consommation conventionnelle (80 kWh/m²/an), représentés par les cas 1-2.
- La méthode DTW donne les meilleurs résultats quelle que soit la méthode d'intelligence artificielle.
- La méthode d'intelligence artificielle SVM donne les meilleurs résultats quelle que soit la méthode de sélection.

La conclusion est donc que le **couple** (DTW associé à SVM) est le choix donnant les meilleurs résultats pour les cas 1-4.

Models	Relevant Data Training Selection Methods	Case 1				Case 2				Case 3				Case 4			
		Median		Overall		Median		Overall		Median		Overall		Median		Overall	
		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
ANN	HDD	0.86	15.8	0.987	18.5	0.81	14.3	0.978	18	0.91	8.4	0.987	10.5	0.81	11.1	0.945	14.3
	Modified HDD	0.93	11.2	0.993	13.9	0.91	10.2	0.986	14.4	0.96	5.9	0.991	9	0.95	6	0.977	9.3
	Frechet Distance	0.93	10.9	0.99	14.1	0.92	9.5	0.99	13.6	0.96	6.1	0.99	8.6	0.93	6.4	0.971	10.3
	DTW	0.95	10.1	0.995	12.2	0.94	8	0.99	11.5	0.96	6	0.99	8.5	0.97	3.9	0.99	6
SVM	HDD	0.86	16.1	0.987	18.6	0.82	14	0.978	17.6	0.94	7.5	0.988	9.8	0.87	9.1	0.963	11.7
	Modified HDD	0.97	7.4	0.996	10.1	0.96	6.4	0.994	9.3	0.96	5.8	0.992	8.2	0.97	4.4	0.989	6.1
	Frechet Distance	0.96	7.2	0.995	10.6	0.95	6.2	0.994	9.6	0.98	5.4	0.994	7	0.98	3.6	0.99	5.2
	DTW	0.97	7.5	0.996	10.4	0.96	6.4	0.994	9.4	0.98	5.1	0.994	7	0.98	3.3	0.993	5.1
BEDT	HDD	0.78	20.1	0.982	22	0.71	17.1	0.972	20.1	0.84	11.7	0.976	14.4	0.81	10.6	0.944	14.4
	Modified HDD	0.89	13.4	0.991	15.4	0.85	12.2	0.984	15.2	0.92	8.9	0.983	12.1	0.94	6.6	0.983	7.9
	Frechet Distance	0.87	13.6	0.99	15.7	0.84	12.4	0.981	16.4	0.91	9.4	0.983	12.2	0.92	6.7	0.974	9.6
	DTW	0.89	13.2	0.992	15.2	0.85	12.2	0.984	15.1	0.91	9.3	0.985	11.2	0.93	6.6	0.98	8.6
RF	HDD	0.86	16.2	0.986	18.4	0.82	13.9	0.978	17.7	0.85	11.8	0.976	14.4	0.87	9.3	0.962	11.9
	Modified HDD	0.94	9.9	0.994	11.9	0.92	8.9	0.991	11.4	0.93	7.8	0.99	9.3	0.96	5.5	0.987	6.8
	Frechet Distance	0.94	10.1	0.992	12.4	0.92	9.3	0.986	14.4	0.94	7.5	0.989	9.7	0.94	5.9	0.978	9.1
	DTW	0.94	9.8	0.995	11.6	0.93	9.2	0.991	11.6	0.94	7.6	0.989	9.3	0.95	5.7	0.986	7.2

Tableau 4 : Erreur pour la combinaison des entrées S8 en fonction du couple (méthode de sélection associé à la méthode AI).

Les résultats obtenus par le couple (DTW, SVM) sont comparés à l'approche « all data », sur trois années de données. Les erreurs obtenues sont reportées dans le **tableau 5**. L'approche « relevant data » (DTW, SVM) donne la meilleure précision. Il faut aussi noter les temps de mise en œuvre pour l'approche « all data » : pour ANN 184h de calculs, pour SVM 75h et pour RF 15h. Par contre, il faut souligner que pour l'approche « relevant data » il faut refaire les calculs pour chaque prévision météorologique. Le temps de calculs est de 3 min. Il parait tout à fait acceptable de relancer ces calculs, même quotidiennement.

Performances	DTW Relevant Data Training		All Data Training							
	SVM		ANN		SVM		BEDT		RF	
	Median	Overall	Median	Overall	Median	Overall	Median	Overall	Median	Overall
R^2	0.98	0.993	0.89	0.971	0.93	0.978	0.86	0.981	0.96	0.993
RMSE	3.3	5.1	8.4	10.3	7.1	9.1	8.9	8.5	4.4	4.9
Model Training Time	3 min 40 sec		184 hours 43 min 6 sec		75 hour 43 min 12 sec		1 hour 37 min 11 sec		15 hour 42 min 18 sec	

Tableau 5 : Comparaison entre l'approche « relevant data » (DTW, SVM) et l'approche « all data »

Les résultats de l'approche « all data » se dégradent significativement si la base de données n'est pas suffisante. Par exemple, le coefficient de détermination progresse de 0.91 à 0.96 (RF).

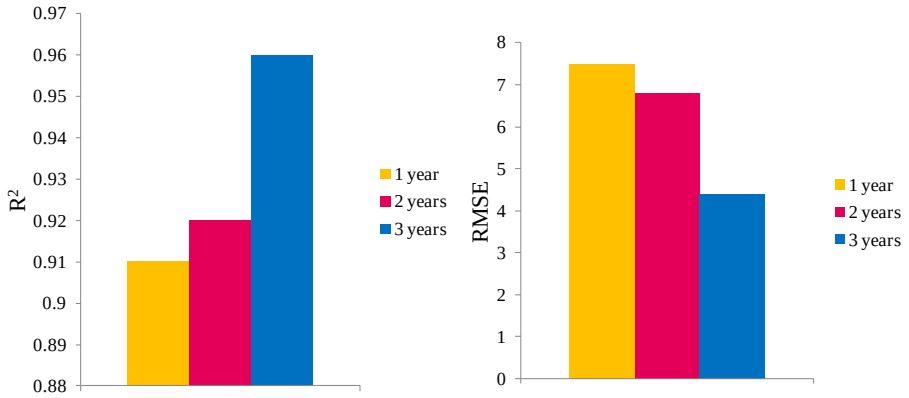

Figure 14 : Coefficient de détermination en fonction de la base de données par l'approche « all data »

Le besoin de chaleur obtenu par l'approche « all data » en utilisant SVM et celui utilisant l'approche « relevant data » couple (DTW, SVM) sont tracés sur la **figure 15**. L'approche « all data » ne transcrit pas bien le profil « actual ».

En considérant la quantité significative de données, le coefficient de détermination et le temps de calculs (nécessitant plusieurs jours), l'approche « all data » est considérée comme inadéquate pour un gestionnaire de réseaux de chaleur.

Pour les cas 1-4, le couple (DTW, SVM) permet d'obtenir les meilleurs coefficients de détermination. Cette observation ne peut pas être une généralisation. Parmi les interrogations, on peut énoncer :

Est-ce toujours valide pour d'autres types d'occupations ?
Est-ce une conséquence de la modélisation très simplifiée dans TRNSys ?

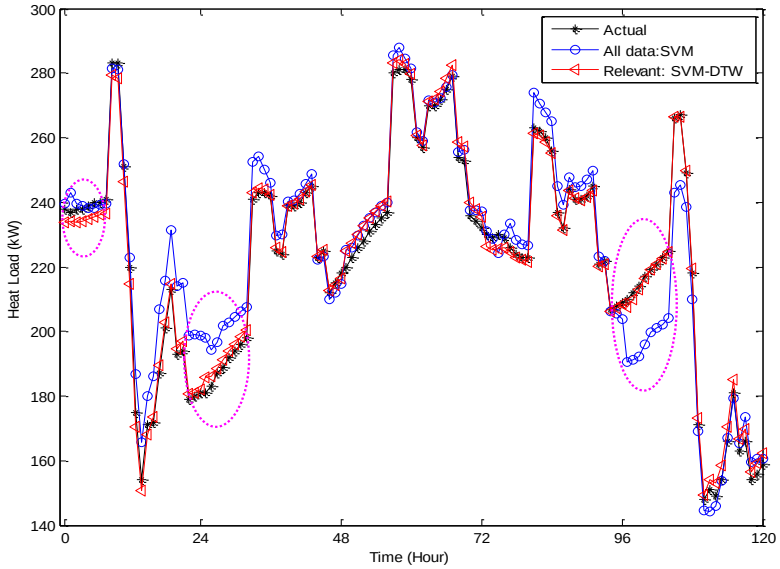

Figure 15 : Comparaison des approches « all data » et « relevant data » avec la méthode SVM

Les cas 5 et 6 ont été conçus comme des premiers éléments de réponse à la première question, respectivement les cas 4*, 5* et 6* à la seconde question.

Comme les conditions opératoires sont différentes, il est nécessaire de définir de nouvelles séries d'entrées. Cinq scénarii d'entrées ont été établis pour les cas 5 et 6, voir le **tableau 6**.

L'approche « relevant data » (DTW, SVM) est naturellement mise en œuvre comme précédemment. Le coefficient de détermination et l'erreur quadratique moyenne sont reportés dans le **tableau 7**. Pour les cas 5 et 6, le coefficient (annuel) est très proche de l'unité : cela montre que l'approche « relevant data » (DTW, SVM) permet de bien prédire le besoin de chaleur.

Input Features	Case-5				Case-6			
	Median		Overall		Median		Overall	
Scenarios	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
S1	0.220	101.6	0.410	108.4	0.947	22.7	0.954	27.8
S2	0.982	17.4	0.973	23.3	0.987	11.1	0.986	15.7
S3	0.976	18.1	0.971	24.1	0.983	12.1	0.984	16.5
S4	0.978	16.8	0.976	21.8	0.992	9.0	0.990	13.2
S5	0.978	16.6	0.975	22.6	0.991	9.0	0.988	14.2

Tableau 7 : Coefficients de détermination et erreur quadratique moyenne selon les compositions des entrées

Name		Description	Scenarios				
			S1	S2	S3	S4	S5
Outputs	P (t)	Heat Load (kW)	×	×	×	×	×
Inputs	Text(t)	External temperature (^{0}C)	×	×	×	×	×
	Text (t-1)	External temperature at 1 hour time delay (^{0}C)	×	×	×	×	×
	øSh(t)	Horizontal solar radiation (kW)	×	×	×	×	×
	øSh(t-1)	Horizontal solar radiation at 1 hour time delay (kW)	×	×	×	×	×
	øSh(t-2)	Horizontal solar radiation at 2 hours time delay (kW)	×	×	×	×	×
	øSext(t)	Solar gain transmitted through window (kW)				×	×
	øSext(t-1)	Solar gain transmitted through window at 1 hour delay (kW)				×	×
	øSext(t-2)	Solar gain transmitted through windows at 2 hours delay (kW)				×	×
	øSint(t)	Solar gain on wall (kW)				×	×
	øSint(t-1)	Solar gain on wall at 1 hour delay (kW)				×	×
	øSint(t-2)	Solar gain on wall at 2 hours delay (kW)				×	×
	occup	Occupancy profile [0 1]	×	×	×	×	×
	oper	Operational characteristics [0 1]		×	×	×	×
	trans	Transitional attributes [0.2 1]		×	×	×	×
	PDL-1	Pseudo dynamic lag 1 [0.2 1]		×	×	×	×
	PDL-2	Pseudo dynamic lag 2 [0.2 1]		×	×	×	×
	Text_TDM	Temporal moving average of external temperature (^{0}C)			×		×
	øSh_TDM	Temporal moving average of horizontal solar radiation (kW)			×		
	øSint_TDM	Temporal moving average of solar gain on wall (kW)					×

Tableau 6 : Composition des entrées pour les cas 5-6

La **figure 16**, respectivement **17**, illustre que l'approche « relevant data » (DTW, SVM) reproduit fidèlement le comportement du besoin de chaleur obtenu des simulations de TRNSys dans le cas d'un immeuble de bureaux (cas 5), respectivement dans le cas d'un immeuble de type centre commercial (cas 6).

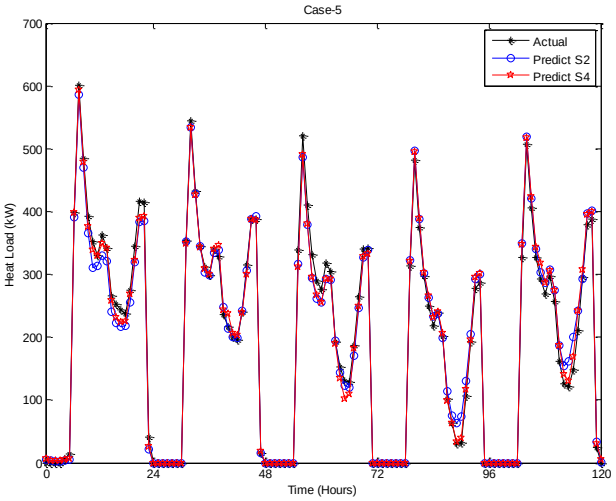

Figure 16 : Prédiction du besoin de chaleur dans le cas 5 (immeuble de type bureaux)

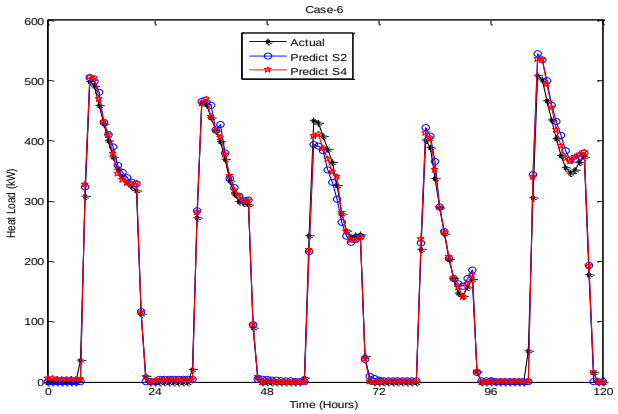

Figure 17 : Prédiction du besoin de chaleur dans le cas 6 (immeuble de type centre commercial)

Comme indiqué précédemment, les cas 4* (occupation « résidentielle »), 5* (occupation « immeuble de bureaux ») et 6* (occupation « centre commercial ») correspondent à une modélisation plus raffinée dans TRNSys. Les cinq séries d'entrées (de la méthode AI) ont été testées sur le couple (DTW-SVM). Le coefficient de détermination et l'erreur quadratique permettent d'affirmer que les résultats reproduisent assez fidèlement le besoin de chaleur (généré par TRNsys) quelle que soit la typologie d'occupation indépendamment de la modélisation dans TRNSys, voir le **tableau 8**.

Input Features	Case-4*				Case-5*				Case-6*			
	Median		Overall		Median		Overall		Median		Overall	
Scenarios	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
S1	0.730	25.1	0.842	28.7	0.245	72.8	0.452	84.1	0.900	19.0	0.922	26.7
S2	0.979	7.2	0.975	11.5	0.982	11.6	0.975	17.8	0.971	9.9	0.980	13.4
S3	0.980	7.5	0.977	11	0.976	13.4	0.974	18.3	0.970	10.3	0.978	14.2
S4	0.981	7.2	0.975	11.3	0.985	11.0	0.979	16.5	0.988	6.0	0.991	9.0
S5	0.985	7.1	0.978	10.7	0.980	10.6	0.980	16.2	0.979	8.3	0.986	11.6

Tableau 8 : Coefficients de détermination et erreur quadratique moyenne selon les compositions des entrées

Le besoin de chaleur « actual » (généré par TRNSys) et celui prédit par l'approche « relevant data » (DTW, SVM) pour les séries d'entrées S2 et S4 sont tracés sur une semaine, **figure 18** pour le cas 4*, **figure 19** pour le cas 5* et **figure 20** pour le cas 6*.

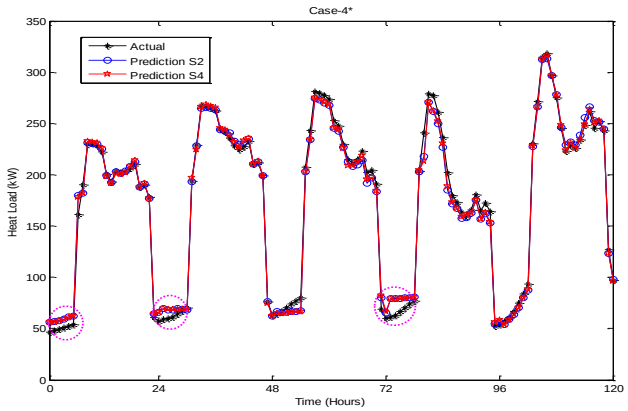

Figure 18: Besoin de chaleur estimé (TRNSys) et prédit par l'approche « relevant data » (DTW, SVM) pour une occupation de type résidentielle, cas 4*

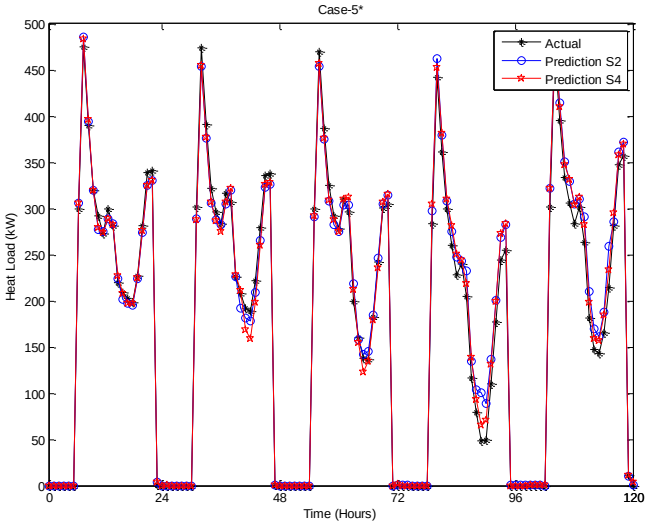

Figure 19: Besoin de chaleur estimé (TRNSys) et prédit par l'approche « relevant data » (DTW, SVM) pour une occupation de type « immeuble de bureaux », cas 5*

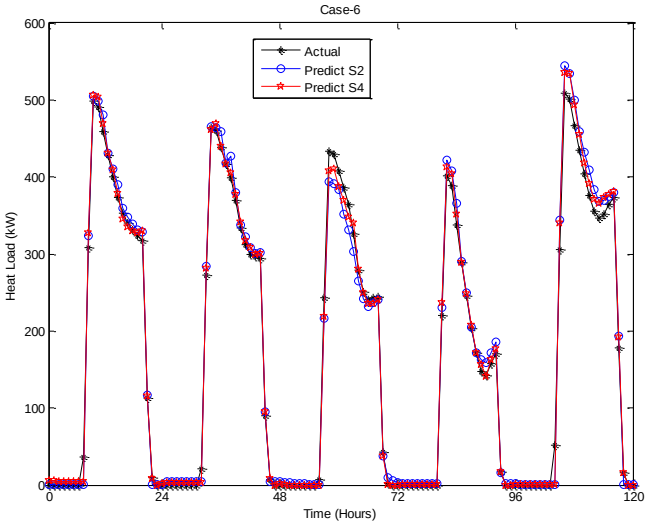

Figure 20: Besoin de chaleur estimé (TRNSys) et prédit par l'approche « relevant data » (DTW, SVM) pour une occupation de type « centre commercial », cas 6*

L'approche « relevant data » utilisant la méthode de sélection DTW associée à la technique d'intelligence artificielle SVM donne des prédictions de besoin de chaleur avec des coefficients de détermination proche de l'unité.

Est-ce que cette approche reste aussi consistante sur des données de consommation réelles ?

A l'Ecole des Mines de Nantes, les besoins thermiques sont enregistrés et peuvent servir de support de cas test, appelé Cas EMN.

Le profil d'occupation est simplifié et correspond à celui de bureaux, voir la **figure 21 (haut)**. La température de consigne globale est aussi simplifiée, voir la **figure 21(bas)**.

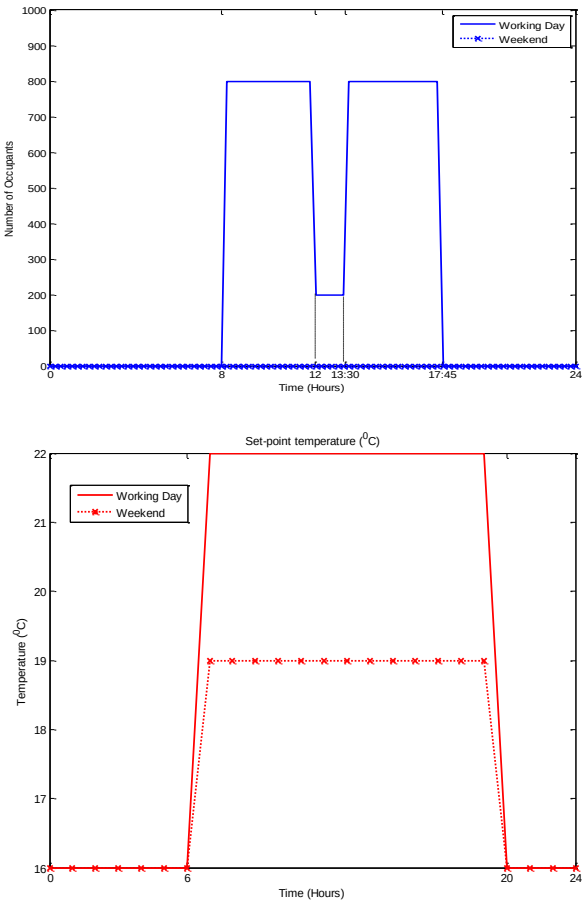

Figure 21 : Profil simplifié d'occupation et de la température de consigne « globalisée » du cas EMN

Le **vecteur de transition** est représenté sur la **figure 23.**

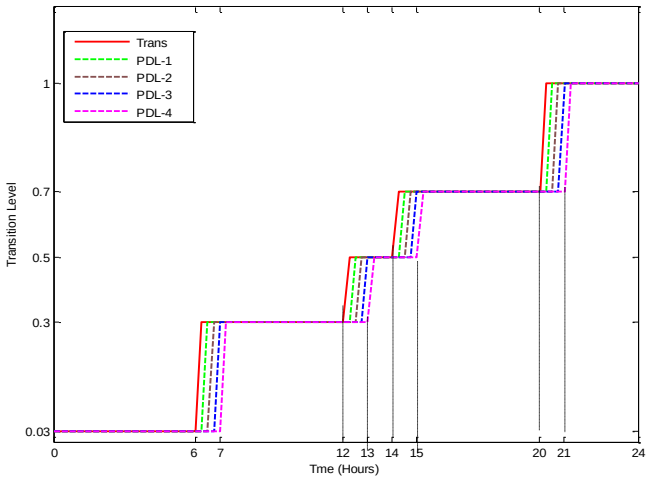

Figure 23: Vecteur de transition du cas EMN

Neuf séries d'entrées ont été considérées dans le cas EMN, voir **tableau 9**.

Name		Description	Scenarios								
			S1	S2	S3	S4	S5	S6	S7	S8	S9
Outputs	P(t)	Heat Load (kW)	×	×	×	×	×	×	×	×	×
Inputs	Text(t)	External temperature (^{0}C)	×	×	×	×	×	×	×	×	×
	Text (t-1)	External temperature at past 15 min delay (^{0}C)	×	×	×	×	×	×	×	×	×
	Text (t-2)	External temperature at past 30 min delay (^{0}C)	×	×	×	×	×	×	×	×	×
	occup	Occupancy profile [0 to 1]		×	×	×	×	×	×	×	×
	oper	Operational characteristics [0 1]			×	×	×	×	×	×	×
	trans	Transitional characteristics [0.2 1]				×	×	×	×	×	×
	PDL-1	Pseudo dynamic lag 1 [0.2 1]					×	×	×	×	×
	PDL-2	Pseudo dynamic lag 2 [0.2 1]						×	×	×	×
	PDL-3	Pseudo dynamic lag 1 [0.2 1]							×	×	×
	PDL-4	Pseudo dynamic lag 2 [0.2 1]								×	×
	Text_TDM	Temporal moving average of external temperature (^{0}C)									×

Tableau 9 : Configuration des entrées considérées pour le cas EMN

L'approche "relevant data" (DTW, SVM) est appliquée à ces 9 séries d'entrées. Les coefficients de détermination et l'erreur quadratique moyenne sont moins proches de l'unité (qu'à partir des simulations TRNSys), voir le **tableau 10**.

Models	Median		Overall	
	R^2	RMSE	R^2	RMSE
S1	0.14	101.6	0.53	97.0
S2	0.44	77.3	0.67	82.2
S3	0.71	55	0.67	81.5
S4	0.72	57.9	0.73	73.6
S5	0.72	54.3	0.74	71.4
S6	0.73	52	0.83	63.7
S7	0.76	50.8	0.80	58.8
S8	0.77	48.9	0.85	54.5
S9	0.77	48.9	0.85	54.0

Tableau 10 : Coefficient de détermination et erreur quadratique moyenne pour l'approche « relevant data » (DTW, SVM) pour le cas EMN

La précision est certes moins correcte mais d'une part le modèle de transition, celui de la température de consigne globale sont assez grossiers et d'autre part il existe aussi de nombreux phénomènes négligés (dont le rayonnement solaire non disponible dans la base de données des années antérieures). En outre, des problèmes de cohérence de mesures ont aussi été observés.

Le besoin de chaleur mesuré « actual » et celui obtenu en utilisant l'approche « relevant data » SVM associé à chaque technique de sélection (HDD, mHDD, FD, DTW) est tracé pour 120 heures consécutives, voir la **figure 24**. L'allure générale est bien prédite : le dernier jour est plutôt bien estimé (un jour de WE). Les jours mardi, mercredi et vendredi sont aussi plutôt bien décrits. Pour le jeudi (dans ce cas), un écart significatif est observé : l'occupation de l'école est différente du reste de la semaine (pas de cours l'après midi et une activité pédagogique le matin différente du reste de la semaine). Ce point pourrait donc être affiné.

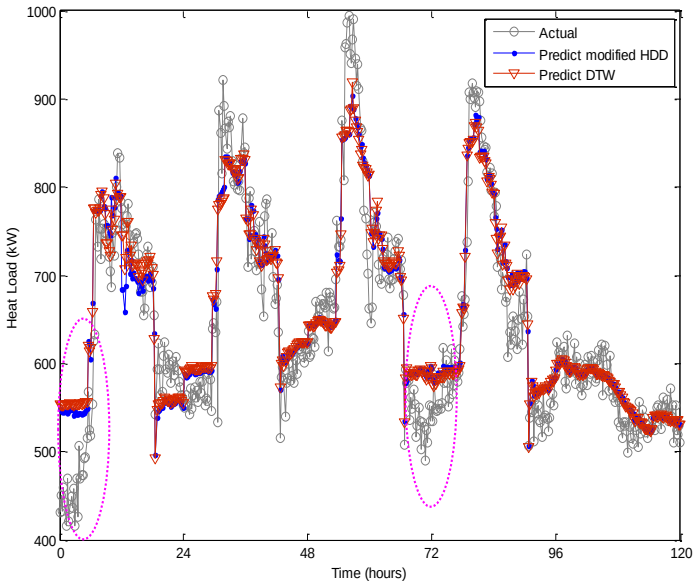

Figure 24: Besoin de chaleur mesuré et prédit pour le cas EMN, SVM associé à chaque méthode de sélection.

Les coefficients de détermination et l'erreur quadratique moyenne pour l'approche « relevant data » SVM en fonction de la méthode de sélection sont indiqués dans le **tableau 11**.

Models	HDD				Modified HDD				Frechet Distance				DTW			
	Median		Overall		Median		Overall		Median		Overall		Median		Overall	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
ANN	0.753	47.4	0.850	55	0.751	49.4	0.853	54.4	0.751	50.8	0.843	56.2	0.77	48.9	0.850	54.5
SVM	0.772	48.8	0.862	52.7	0.781	46.8	0.862	52.9	0.749	51.3	0.84	56.5	0.751	50.1	0.856	53.7
BEDT	0.653	48.7	0.833	58	0.707	53.8	0.822	59.9	0.729	53.8	0.833	58	0.652	51.5	0.819	60.5
RF	0.704	49.5	0.834	57.9	0.735	53.6	0.839	56.9	0.675	49.5	0.837	57.4	0.704	54.5	0.836	57.6

Tableau 11 : Coefficient de détermination et erreur quadratique moyenne pour l'approche « relevant data » en fonction de la méthode de sélection

Les temps de calculs pour l'élaboration d'un modèle prédictif par l'approche « relevant data » sont présentés sur la **figure 25**. A gauche, le choix de la technique AI est associé à la sélection DTW, à droite la technique AI est SVM. Les 15 minutes de CPU requises pour la mise au point d'un modèle lors d'une prévision météorologiques semblent acceptables.

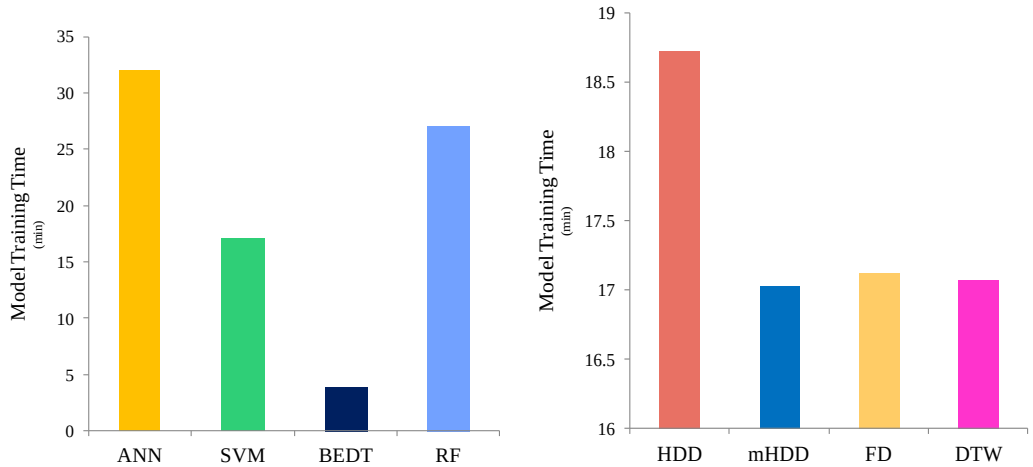

Figure 25 :Temps de calcul CPU, à gauche en fonction du choix de la méthode AI, à droite en fonction de la méthode de sélection.

L'approche « relevant data » est comparée à l'approche «all data », voir le **tableau 12**.

Building Functioning Type	Performances	DTW Relevant Data Training		All Data Training					
		SVM		ANN		SVM		BEDT	
		Median	Overall	Median	Overall	Median	Overall	Median	Overall
Working Day	R^2	0.80	0.86	0.55	0.84	0.56	0.85	0.54	0.83
	RMSE	51.5	57.6	91.2	60	90	58.9	92.7	63
	Model Training Time	15 min 3 sec		14 hour 18 min 5 sec		3 hour 15 min 45 sec		42 min 27 sec	
Weekend	R^2	0.63	0.82	0.41	0.75	0.48	0.81	0.33	0.66
	RMSE	36.3	42.7	53	50.1	44.1	43.9	50.1	59.2
	Model Training Time	12 min 48 sec		2 hour 57 min 12 sec		22 min 27 sec		26 min 14 sec	

Tableau 12: Coefficient de détermination et erreur quadratique moyenne pour l'approche « relevant data » et « all data »

Le besoin de chaleur mesuré et prédit (« revelant data » SVM-DTW et « all data » SVM) sont tracés sur la **figure 26**.

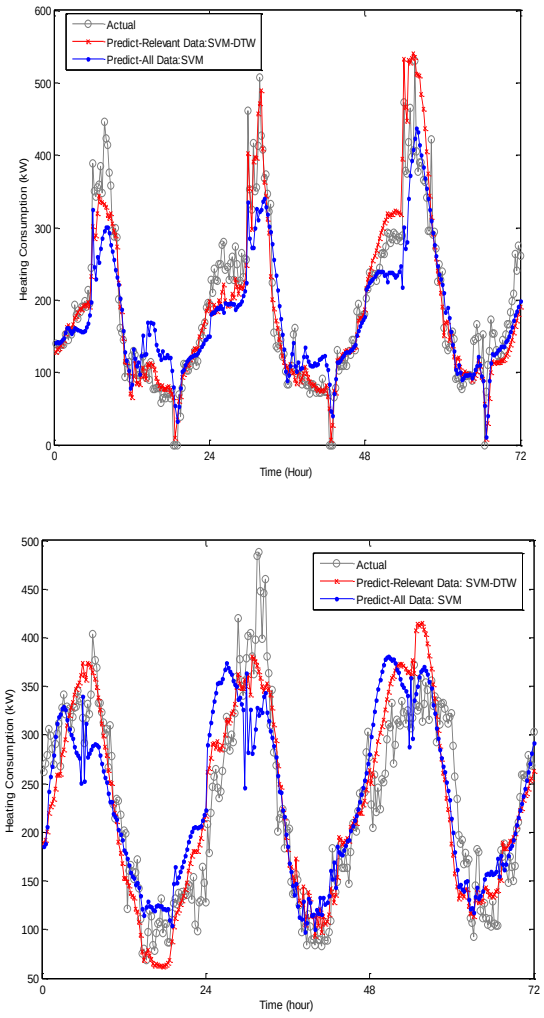

Figure 26: Besoin de chaleur mesuré et prédit par l'approche « relevant data » SVM-DTW et « all data » SVM

L'approche « relevant data » (DTW-SVM) reste assez performante pour le cas EMN (et est la plus performante parmi les couples testés).

Les résultats du cas EMN ne sont pas pleinement satisfaisants :

- La base de données n'incluait pas le flux solaire.
- De nombreuses mesures aberrantes ont été détectées.
- Un vecteur de transition et une température de consigne globalisée pourraient aussi être repris.
- L'occupation pourrait aussi être affinée, en fonction des activités pédagogiques.

Pour conclure la méthodologie « relevant data » est schématisée dans la **figure 27**. Elle se met en œuvre suivant les 7 étapes suivantes :

Etape-1: Classification des conditions d'exploitation et d'usage de l'immeuble

Etape-2: Mise au point du modèle pseudo-dynamique de transition

Etape-3: Choix des variables climatiques les plus explicatives (incluant des effets inertiels par le biais de grandeurs moyennées)

Etape-4: Configurations des entrées (pour la phase d'apprentissage puis de prédiction)

Etape-5: Pré-analyse de la base de données (application du traitement d'ondelettes pour pondérer les variables climatiques entre-elles)

Etape-6: Sélection des jours les plus ressemblants aux conditions à prédire

Etape-7: Prédiction du besoin de chaleur.

Il est recommandé d'utiliser la technique la technique « Machines à vecteur de support » (SVM) associée à une extraction de la base de données des jours les plus ressemblants par la sélection de la déformation temporelle dynamique (DTW).

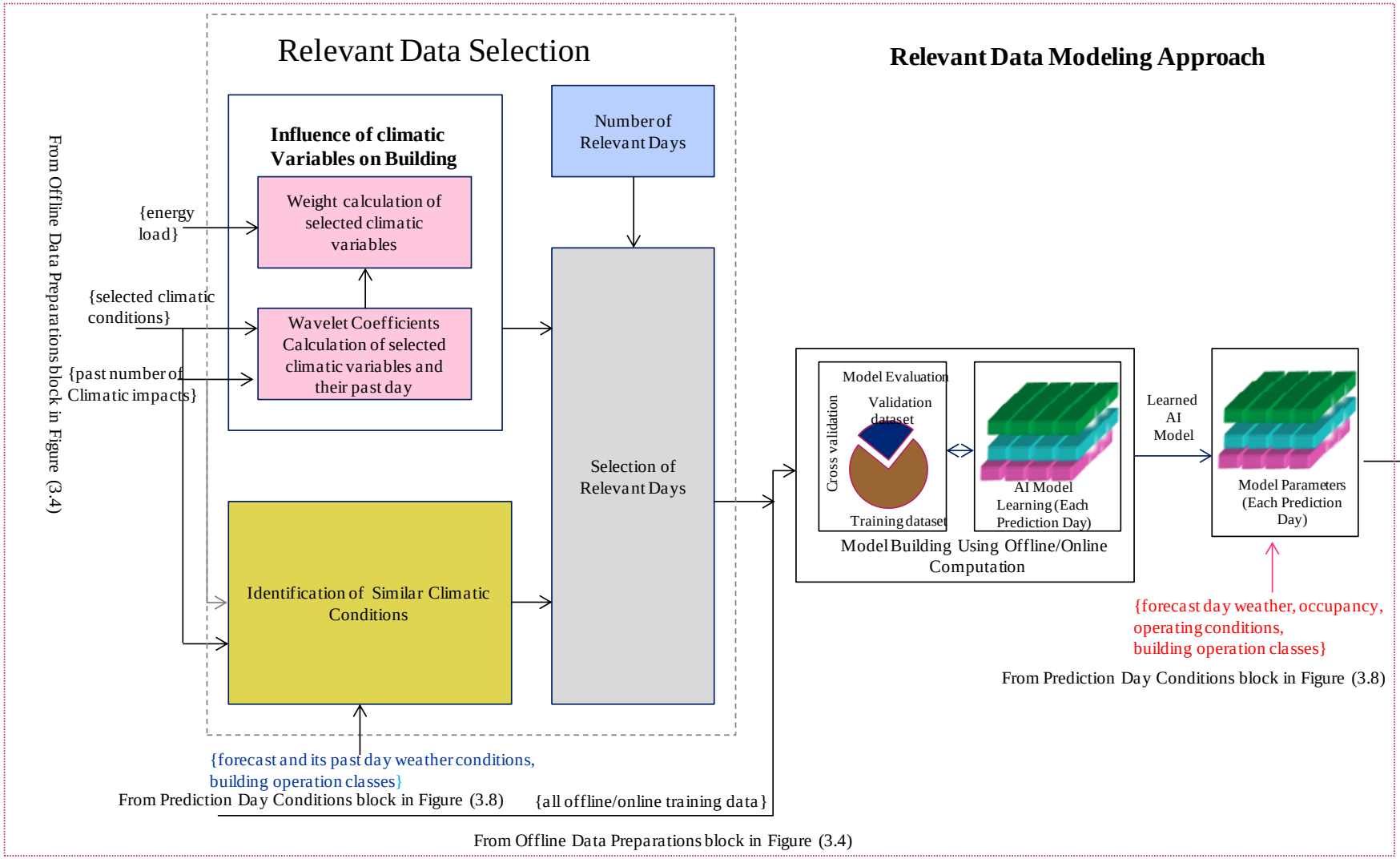

Figure 27: Méthodologie pour la mise en œuvre d'une technique AI par sélection de jours similaires

Contents

Acknowledgements i
Extended Abstract in French v
Nomenclature 5
List of tables 11
List of figures 13
Personal references 17
Summary of Contribution 19
Chapter 1: Introduction 21
1.1 General Background 21
1.2 Research Problems 22
1.3 Research Objectives 23
1.4 Research Framework 23
1.5 Manuscript Outlines 24
Chapter 2: Low Energy Building Modelling 27
2.1 Low Energy Building (LEB) 27
2.1.1 LEB concepts 27
2.1.2 Evolution of LEBs 29
2.1.3 Factors affecting LEB 32
2.2 Building Energy Model 35
2.2.1 Introduction 35
2.2.2 White-box Model 38
2.2.3 Gray-box Model 44
2.2.4 Black-box Model 46
2.3 Applications to Building Energy Modeling by Black Box Model 50
2.3.1 All data modeling approach 52
2.3.2 Relevant data modeling approach 58
2.4 Conclusion 64
Chapter 3: Artificial Intelligence for LEB Modelling 69
3.1 Modeling Approaches 69
3.1.1 Introduction 69
3.1.2 Assumptions 71

3.1.3 Proposed Approaches 72
3.2 Offline Data Preparations 73
3.2.1 Building Operation Classification/Clustering 75
3.2.2 Pseudo Dynamic Model 75
3.2.3 Derived Climatic Variables Generation 78
3.2.4 Climatic Variables Selection 79
3.3 Prediction Day Conditions 81
3.4 All Data Modeling Approach 82
3.5 Relevant Data Modeling Approach 83
3.5.1 Identification of Similar Climatic Conditions 86
3.5.2 Influence of Climatic Variables on Building 91
3.5.3 Selection of Relevant Days 93
3.6 Artificial Intelligence Model 94
3.6.1 Artificial Neural Network 94
3.6.2 Support Vector Machine 97
3.6.3 Boosted Ensemble Decision Tree 98
3.6.4 Random Forest 98
3.6.5 Practical Aspects in AI 99
3.7 Conclusion 100
Chapter 4: Application to Building Simulation 103
4.1 Buildings Characteristics 103
4.1.1 Buildings Description 103
4.1.2 Climatic Conditions 104
4.1.3 Occupancy Profile 105
4.1.4 Building Operating Conditions 107
4.2 Simulation Data Generation 110
4.3 Application of AI Modeling Methodology for CB to LEB 112
4.3.1 Introduction 112
4.3.2 Recommendation for Applying the Methodology “Step by Step” 113
4.4 Conclusion 148
Chapter 5: Application- Real Building 149
5.1 Buildings Characteristics 149
5.1.1 Building Description 149
5.1.2 Data Collection 149
5.1.3 Occupancy Profile 149

5.1.4 Building Operating Conditions 150
5.2 Application of AI Modeling Methodology 151
5.2.1 Introduction 151
5.2.2 Recommendation for Applying the Methodology “Step by Step” 151
5.3 Conclusion 165
Chapter 6: Summary and Future Works 167
6.1 Summary 167
6.2 Future Works 170
Bibliography 171
Appendix A- Steady State Model 183
Appendix B – Machine Learning based Artificial Intelligence 187
B.1 Artificial Neural Network 187
B.2 Support Vector Machine 193
B.3 Decision Tree 196
B.4 Random Forest 198
B.5 Ensemble 199
B.6 Practical Aspects in Artificial Intelligence 202
Appendix C- Building Operation Classification/Clustering 207
Abstract 208
Résumé 208

Nomenclature

Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMA	Autoregressive with Moving Average
ARx	Autoregressive with Exogeneous, i.e., External, Inputs
BBC	Batiment Basse Consummation
BEMS	Building Energy Management System
BT	Boosting Tree
CB	Convectional Building
CDD	Cooling Degree Day
CPU	Central Processing Unit
CVA	Cannonical Variate Analysis
DOF	Degree of Freedom
DT	Decision Tree
DTW	Dynamic Time Warping
EMN	Ecole des Mines de Nantes
EPBD	Energy Performance Building Directive
ESCOs	Energy Services Companys
FD	Frechet Distance
GA-ANFIS	Genetic Algorithm Adaptive Network Fuzzy Interfaces System
HDD	Heating Degree Day
HPE	Haute Performance Enérgetique
HPE EnR	Haute Performance Enérgetique Energie Renouvelable
HVAC	Heating, Ventilation and Air-Conditioning
k-NN	K-Nearest Neighbor
LEB	Low Energy Building
LSM	Least Square Method
MAPE	Mean Absolute Percentage Error
MHDD	Modified Heating Degree Day

MLP	Multi Layer Perceptron
MLR	Multi Linear Regression
NZEB	Nearly Zero Energy Building
OOB	Out of Bag
PAHU	Primary Air Handling Unit
PCA	Principal Component Analysis
PDM	Pseudo Dynamic Model
PDL	Pseudo Dynamic Lag
PEB	Passive Energy Building
RBF	Radial Basis Function
RES	Renewable Energy Sources
RF	Random Forest
RMSE	Root Mean Square Error
SVM	Support Vector Machine
SVR	Support Vector Regression
THPE	Trés Haute Performance Enérgetique
THPE EnR	Trés Haute Performance Enérgetique Energie Renouvelable
VLEB	Very Low Energy Building

Variables

a	Wavelet low frequency coefficients	[-]
A	Area	[m^2]
B	Number of trees in random forest	[-]
bn	Number of bins	[-]
C	Thermal capacity	[J/K]
Coeff_xx	Coefficient of climatic variables	[-]
COP_c	Coefficient of performance of cooling system	[%]
d	Wavelet high frequency coefficients	[-]
D^{-1}	Time delay	[hour]
DD_c	Cooling degree-hour	[^{0}C/h]
DD_h	Heating degree-hour	[^{0}C/h]
E_c	Cooling energy consumption	[kWh]
E_h	Heating energy consumption	[kWh]
f_o	Shape factor	[m^{-1}]
k-value	Thermal conductivity	[W/m.K]

l	Number of relevant days	[-]
$\dot{m}_{air}$	Mass flow rate of air	[kg/s]
n	Number of training data	[-]
N	Number of training days	[-]
$N_{bin,j}$	Number of hours of occurrence of the jth bin	[-]
$N_{Eqn,tr}$	Number of training equations	[-]
N_h	Number of hidden neurons	[-]
$N_{h,max}$	Maximum hidden neurons	[-]
N_i	Number of input neurons	[-]
N_o	Number of output neurons	[-]
N_{ocup}	Number of occupants	[-]
$N_{r,inf}$	Air infiltration rate changes per hour	[h^{-1}]
N_θ	Number of model parameters	[-]
occup	Occupancy profile	[-]
P	Power	[W]
pg	Performance goal	[-]
Q_A	Annual energy consumption	[kWh]
$q_{A,peak}$	Heat peak load	[W/m^2]
$\dot{Q}_{env}$	Heat gain or loss through envelope components	[W]
$\dot{Q}_{h/c}$	Heating or cooling demand	[W]
$\dot{Q}_{int}$	Internal heat gain due to occupants, lighting and appliances	[W]
$\dot{Q}_{ocup,b}$	Heat generation rate from occupants	[W/m^2]
$\dot{Q}_{out}$	Heat loss from the air-zone	[W]
$\dot{Q}_{sol}$	Solar heat gain through transparent building components	[W]
Q_{st}	Heat energy storage	[J]
$\dot{Q}_{source}$	Heat gain inside the air-zone	[W]
$\dot{Q}_{ven}$	Ventilation heat gain or loss due to air exchange	[W]
r	Correlation indexes	[-]
R	Thermal resistance	[m^2.K/W]
R^2	Coefficient of determination	[-]
r_{xy}	Cross-correlation indexes for time series x and y	[-]
S_x	Sample standard deviations of time series x	[-]
t	Time	[hour]

T	Temperature	[^{0}C]
T_b	Base temperature	[^{0}C]
T_{fg}	Exterior floor temperature	[K]
$T_{sol-air}$	Sol-air temperature	[K]
$T_{set-point}$	Set-point temperature	[^{0}C]
trans	Transitional attributes	[-]
u	Number of past day climate impacts	[-]
U	Overall thermal heat loss coefficient	[W/m^2.K]
UA	Overall heat loss coefficient	[W/K]
v	Significant climatic variables	[-]
wc	Desired weight	[-]
z	Decomposition length	[-]
n_m	Number of trees for decision tree	[-]
β_m	Number of leaf in each decision tree	[-]
υ	Learning parameter of decision tree	[-]
φ_{th}	Threshold values	[-]
Φ	Number of lags	[-]
β_0	Initial energy load level	[-]
$\Delta\beta$	Step size of transition of energy load	[-]
$\emptyset_{Sh}$	Horizontal solar radiation	[W]
$\emptyset_D$	Direct solar radiation	[W]
$\emptyset_{Sext}$	Solar gain transmitted through windows	[W]
$\emptyset_{Sint}$	Solar gain on walls	[W]
T_{ext_TDM}	Temporal moving average of external temperature	[^{0}C]
$\emptyset_{Sint\_TDM}$	Temporal moving average of solar gain on walls	[W]
τ	Time constant	[hour]
ρ	Density	[kg/m^3]
ΔS	Thickness	[m]
C_p	Specific heat capacity	[J/kg.K]
$\emptyset_{in}$	Heat flux entering the controlled volume	[W]
$\emptyset_{out}$	Heat flux leaving the controlled volume	[W]
$\emptyset_{source}$	Dissipated amount of heat flux from the control surface	[W]
τ_g	Transmittance on glass plane	[-]

α_s	Solar absorptivity	[-]
G_t	Solar radiation incident on surface	[W]
ε_s	Emissivity	[-]
h_o	Heat transfer coefficient on exterior envelope	[W/m^2.K]
ζ_v	Factor of ventilation system	[-]
$\dot{q}_v$	Volumetric flow of ventilation air	[m^3/hour]
$\mathcal{M}$	Sampling length of data in a day	[-]
η_h	Seasonal average efficiency of heating equipment	[%]

Subscripts

air	Air
app	Appliances
buil	Building
env	Envelope components
e, ext	External
f	Floor
g	Glazing
in	Internal or Interior
lit	Lighting
md	Modified
ocup	Occupancy
rf	Roof
sky	Sky
steady	Steady state
surr	Surrounding
w	Wall
win	Window
z	Zone

List of tables

Table 2.1: LEBs in different parts of the Europe [6] 28
Table 2.2: Migration pathways from CBs to LEBs in Europe (in terms of annual energy consumption) [7] 30
Table 2.3: Typical U-values and R-values of CBs and LEBs in Europe [13] 33
Table 2.4: Summary of input variables and time step of prediction using gray-box model 45
Table 2.5: Summary of input features and time step of prediction using all data modeling approach 57
Table 2.6: Comparison of relevant data with fixed/updated training with all data modeling approach 59
Table 2.7: Summary of input features used to select relevant data 63
Table 2.8 Description of input features used in literatures for building energy consumption prediction 65
Table 2.9: Comparison of white-box, gray-box and black-box prediction models 66
Table 4.1: Description of buildings 104
Table 4.2: Description of materials use for buildings 104
Table 4.3: Summary statistics of climatic conditions at different locations 105
Table 4.4: Summary of time constant for different building types 112
Table 4.5: Summary of input and output variables of different scenarios 118
Table 4.6: Parameters of SVM used for weight calculation 119
Table 4.7: Summary of ANN parameters 126
Table 4.8: Comparison of different input features scenarios for different cases using DTW relevant data modeling approach based on ANN 126
Table 4.9: Summary of SVM, BEDT and RF parameters 131
Table 4.10: Performance of AI models using different relevant data modeling approaches for different cases 132
Table 4.11: Comparison of model performance of DTW based relevant data modeling approach using SVM with all data modeling approach using ANN, SVM, BEDT and RF 134
Table 4.12: Building operation classes for Case-5 and Case-6 building 137
Table 4.13: Summary of input and output variables of different scenarios 139

Table 4.14: Prediction performance of different scenarios for Case-5 and Case-6 building based on DTW relevant data modeling approach using SVM 141
Table 4.15: Prediction performance of heating load for different scenarios and cases based on DTW relevant data modeling approach using SVM 145
Table 5.1: Summary of input and output variables of different scenarios 155
Table 5.2: Parameters of SVM used for weight calculation 156
Table 5.3: Comparison of different scenarios based on DTW relevant data modeling approach using ANN 158
Table 5.4: Performance of different AI models using HDD, modified HDD, FD and DTW relevant data selection method 160
Table 5.5: Comparison of model performance of DTW relevant data modeling approach using SVM with all data modeling approach using ANN, SVM, BEDT and RF 163

List of figures

Figure 1.1: Global building energy consumption [2] 21
Figure 1.2: Annual energy consumption in each sector in France [4] 22
Figure 1.3: Summary of our research framework 24
Figure 1.4: Summary of manuscript outlines 25
Figure 2.1: Summary of evolution from CB to LEBs ([3], [6]) 31
Figure 2.2 : Comparison of different LEBs with CB (general context in Central Europe) 31
Figure 2.3: Factors affecting LEB 32
Figure 2.4: White-box, gray-box and black-box models 37
Figure 2.5: Scheme of energy flows in a building 39
Figure 2.6: Simple illustration of building energy model using lumped resistance and capacitance 41
Figure 2.7: Under-fitting, reasonable-fitting (just right) and over-fitting of data 46
Figure 2.8: Concept of all data and relevant data with fixed training approach to build a model . 51
Figure 2.9: Concept of relevant training with updated training approach to build a model 51
Figure 3.1: Illustration of thermal dynamic behavior in building 70
Figure 3.2: Past 5 day behavior of Text and ØSint from d_1 and d_2 days 71
Figure 3.3: Whole framework of all data and relevant data modeling approach based on offline and online learning for energy load prediction 73
Figure 3.4: Preparations of offline data 74
Figure 3.5: Overall operating characteristics of building (for a day) 76
Figure 3.6: Transitional and pseudo dynamic characteristics (for a day) 77
Figure 3.7: Dynamic characteristics of indoor temperature control in a building 78
Figure 3.8: Prediction day conditions 81
Figure 3.9: Framework of all data modeling approach based on offline learning for energy load prediction 82
Figure 3.10: Framework of proposed relevant data modeling approach based on online/offline learning 85
Figure 3.11: Illustration of dynamic time warping to select similar climatic variables (e.g., external temperature) 89

Figure 3.12: Illustration of Frechet distance to select similar climatic variables (e.g., external temperature) 90
Figure 4.1: Occupancy profile of single-zone Case1 - Case4 building 105
Figure 4.2: Occupancy profile of single-zone Case-5 building 106
Figure 4.3: Occupancy profile of multi-zone Case-6 building 107
Figure 4.4: Operating conditions of Case 1-Case 4 building 108
Figure 4.5: Operating conditions of Case-5 building 109
Figure 4.6: Operating conditions of Case-6 building 109
Figure 4.7: Summary statistics of the solar gain transmitted through the windows for four climatic locations (Paris, Lille, Lyon and Clermont-Ferrand) 110
Figure 4.8: Summary statistics of the solar gain on the walls for four climatic locations (Paris, Lille, Lyon and Clermont-Ferrand) 111
Figure 4.9: Final energy demand for CBs and LEBs 111
Figure 4.10: Classification of building operation classes (Case-4) 113
Figure 4.11: Functioning profile of building (Case-4) 113
Figure 4.12: Transitional and pseudo dynamic characteristics during two consecutive day 114
Figure 4.13: Pseudo dynamic transitional effects on heating load during two consecutive days 115
Figure 4.14: Correlation indexes on climatic conditions for CBs to LEB 116
Figure 4.15: Cross-correlation indexes to select external temperature dynamics for CBs to LEB 117
Figure 4.16: Influence of the number of the past climatic conditions days selection on different building cases 120
Figure 4.17: Influence of past day external temperature Text and solar gain on walls ØSint from prediction day on daily average heating load for Case-4 building 121
Figure 4.18: Performance while fitting wavelet coefficients using LSM based on regression and SVM based on linear kernel 122
Figure 4.19: Influence of climatic conditions on different buildings using SVM based on linear kernel and LSM based on regression 122
Figure 4.20: Individual weight distribution of prediction and past day climatic conditions for different building types (using SVM based on linear kernel) 123
Figure 4.21: Performance of scenarios S7 and S8 for different heating months 128
Figure 4.22: Prediction of heating load from input scenarios S7 and S8 for some random days in different months for Case-3 building 128
Figure 4.23: Prediction of heating load from input scenarios S7 and S8 for some random days in different months of Case-4 building 129

Figure 4.24: Influence of number of days data in accuracy of prediction model 130
Figure 4.25: Model training CPU-time from different AI models 132
Figure 4.26: Model training CPU-time from different relevant data modeling approaches 133
Figure 4.27: Influence of training size data using all data modeling approach using RF 134
Figure 4.28: Prediction of heating load based on DTW relevant data modeling approach using SVM with all data modeling approach using SVM for some random days 135
Figure 4.29: Classification of building operation classes (Case-5 and Case-6 buildings) 137
Figure 4.30: Correlation indexes of direct and derived climatic variables of Case-5 and Case-6 buildings 138
Figure 4.31: Prediction of heating load using scenarios S2 and S4 of Case-5 building (some random days in January) 141
Figure 4.32: Prediction of heating load using scenarios S2 and S4 of Case-6 building (some random days in January) 142
Figure 4.33: Classification of building operation classes (Case-4*, Case-5* and Case-6*) 143
Figure 4.34: Correlation indexes of direct and derived climatic variables of Case-4*, Case-5* and Case-6* building 144
Figure 4.35: Prediction of heating load using scenarios S2 and S4 of Case-4* building (some random days in January) 146
Figure 4.36: Prediction of heating load using scenario S2 and S4 of Case-5* building (some random days in January) 146
Figure 4.37: Prediction of heating load using scenarios S2 and S4 of Case-6* building (some random days in January) 147
Figure 5.1: Occupancy profile for working days and weekend 150
Figure 5.2: Operating conditions of building for working day 150
Figure 5.3: Heating power demand and occupancy profile during working days 151
Figure 5.4: Classification of building operation classes 152
Figure 5.5: Functioning profile of building 152
Figure 5.6: Transitional and pseudo dynamic characteristics during a day 153
Figure 5.7: Pseudo dynamic transitional effects on heating load 154
Figure 5.8: Cross-correlation indexes to select external temperature dynamics 155
Figure 5.9: Influence of past 1 day of external temperature Text on daily average heating load using SVM based on linear kernel and LSM based on regression 157
Figure 5.10: Influence of relevant days data on the accuracy of prediction model 159
Figure 5.11: Prediction of heating load using different relevant data selections based on SVM (for some random days) 161

Figure 5.12: Model training CPU-time using different AI models using DTW..... 162
Figure 5.13: Model training CPU-time using different relevant data modeling approach using SVM..... 162
Figure 5.14: Prediction of heating load based on all data and DTW relevant data modeling approach using SVM for working days..... 164
Figure 5.15: Prediction of heating load based on all data and DTW relevant data modeling approach using SVM for weekend..... 164

Personal references

Journal Publication

- Subodh Paudel, Mohamed Elmtiri, Stéphane Couturier, Phuong H. Nguyen, René Kamphuis, Bruno Lacarrière, Olivier Le Corre, A relevant data selection method for energy consumption prediction of low energy building based on support vector machine, Energy and Buildings, March 2016, *Under Revision*.
- Subodh Paudel, Mohamed Elmtiri, Wil L. Kling, Olivier Le Corre, Bruno Lacarrierè, Pseudo dynamic transitional modeling of building heating energy demand using artificial neural network, Energy and Buildings, 70, 81-93, 2014.

Conference Publication

- Subodh Paudel, Phuong H. Nguyen, Wil L. Kling, Mohamed Elmtiri, Bruno Lacarrierè, Olivier Le Corre, Support Vector Machine in Prediction of Building Energy Demand Using Pseudo Dynamic Approach, Proceedings of ECOS 2015- The 28th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems, June 30- July 3, 2015, Pau, France.

Summary of Contribution

This manuscript makes several contributions:

- It provides detail understanding of physics behind thermal energy transfer process in the buildings and differentiates thermal energy performance criteria for conventional to low energy buildings.
- It reviews and compares the building energy demand estimation and the prediction model.
- It introduces different machine learning artificial intelligence model namely neural network, support vector machine, decision tree and random forest to predict thermal load of building.
- It proposes novel pseudo dynamic model to include transitional behavior of occupancy and building operating conditions[1].
- It modifies the traditional degree-day method to new degree-day method (propose in this manuscript) to include variation of energy load weight effect at different time intervals during a day.
- It proposes novel relevant data selection method to select small representative day data from the given database. Because of this fewer day data representation, it provides flexibility in adjusting the prediction model to rely in a dynamic environment due to lower computational complexities CPU[2]-time.

[1] Refers to the set of values, for example, set-point temperature and ventilation schedule during a day
[2] Central processing unit

Chapter 1: Introduction

1.1 General Background

The total energy consumption globally accounts around 7200 Mtoe (Mega Tonnes Oil Equivalents) [1]. Out of these, the only building sector represents one-third of energy consumption and space heating, space cooling and water heating that accounts for 60% of final energy consumption [2] (Figure 1.1). This total energy consumption building considerably increased from 2002 to 2012 and contributes to large greenhouse gases (GHG) emissions. Similarly, GHG emission from building sector can be observed worldwide, for example, Europe contributes to 40% and United States to 48% [3].

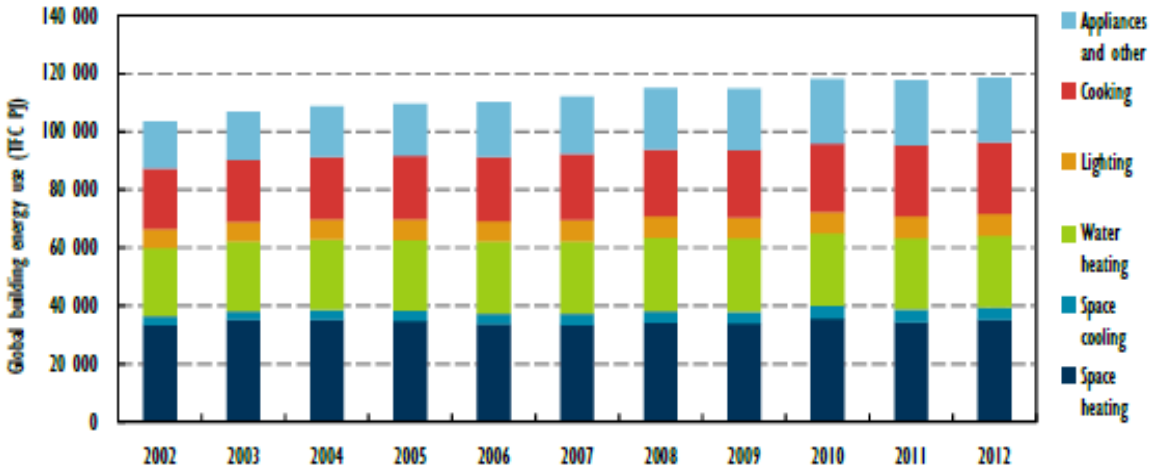

Figure 1.1: Global building energy consumption [2]

In France, the annual energy consumption in different sectors is increasing for last 40 years as shown in Figure (1.2). Apart from industry and transport, the building sector (residential/commercial) is responsible for the largest portion of the energy consumption. The energy that is spent to heat the residential buildings accounts for 40% of total energy demand including electricity, hot-water and air-conditioning. This total energy consumption from building further contributes to 25% of GHG emission. Energy efficiency standards in building thus have drawn significant attention and awareness to focus on reducing annual energy consumption.

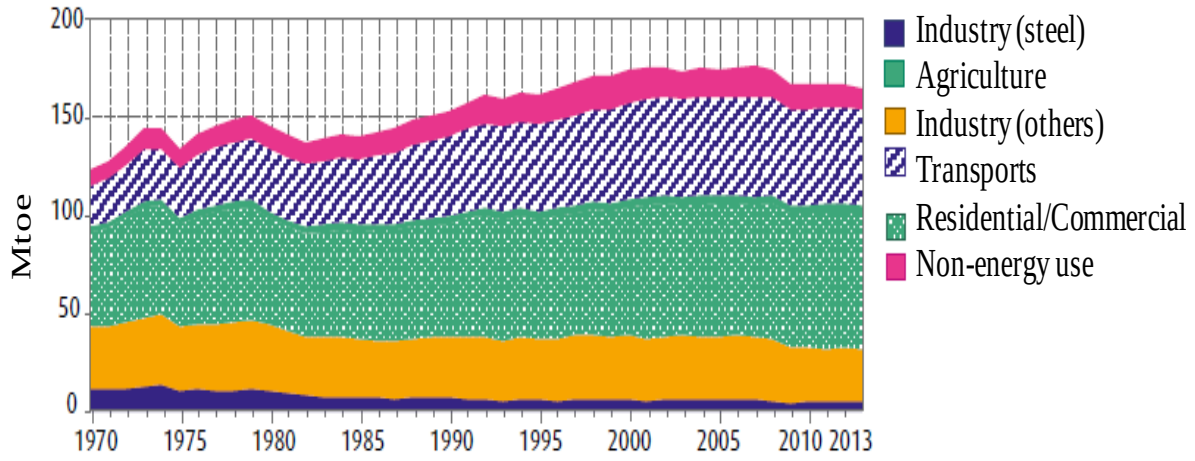

Figure 1.2: Annual energy consumption in each sector in France [4]

In order to address these issues, many developed and developing countries are focusing their attention on energy performances and are migrating from these Conventional buildings (CBs) towards an energy-efficient buildings particularly low energy building (LEB). In order to standardize the building energy performance, the European Commission has formulated an Energy Performance Building Directive (EPBD) and this directive requires all the buildings to be nearly zero energy buildings by 2020. Similarly, Japan plans to implement nearly zero energy building for newly constructed public buildings by 2020 and the US by 2030. However, successful implementation of energy-efficient building requires a radical step in enhancing energy efficiency by improving building envelope (e.g., insulating wall cavities, increasing the quantity of insulation for roof, using high efficiency windows/glazing, compacting building shape), using higher efficient heating and cooling equipments, using renewable sources (solar thermal and electrical renewable energy system, e.g., solar photovoltaic and wind energy) integration in the building, use of intelligent energy management system, improvement in indoor thermal comfort etc. From the improvement of energy conservation point of view, estimation and prediction of energy consumption of building is therefore more noteworthy.

1.2 Research Problems

LEBs are new concept being considered as a solution for the built environment to satisfy high-energy efficiency standards and to improve an energy performance. These are still progress in research and the technology is basically focused on improving thermal performance of envelope by adding layers of materials with very low thermal conductivity (W/m.K), thereby obtaining building envelope with low U-value (thermal transmittance, $W/m^2.K$) or high R-value (thermal

resistance m^2.K/W). This lower U-value decreases the annual heating requirements and introduces large time constant in building. Because of large time constant as well as large heat capacity, it slows the rate of heat transfer between interior of building and outdoor environment and alters the indoor climate in building regardless of sudden changes in climatic conditions. In addition, the estimation of energy demand complexity increases due to the non-linear relationship between the energy demand and other factors such as solar gains, internal gains (occupancy and lighting) and changes in climatic and operating conditions of building. Therefore, estimation and prediction of thermal energy demand of LEB is quite complex.

1.3 Research Objectives

For an energy services company, it is essential to know the estimate of energy demand by knowing forecasted weather and behavior of customer. So, the objective of this research is to estimate the thermal energy demand of LEBs based on forecast weather and behavior of customer. The range of prediction is from hours to couple of days or even for longer periods depending upon the forecast range of climatic conditions. The specific objectives of this research are:

- To develop a prediction model using few available data for control and system management.
- To analyze the behavioral change of a prediction model for different kinds of buildings: CB and LEB, office, commercial and residential building, and single-zone and complex-zone building.

1.4 Research Framework

Our research work is mainly involved in development of a prediction model for LEB during the operation phase of buildings[3]. In order to build the prediction model for CB to LEBs, the work will focuses to understand the heat and energy transfer in the building and the principle behind CB and LEBs. Then, the work will emphasize review on existing prediction models for building thermal load. These studies lead us to understand the advantages and drawbacks of each prediction model and suggest the criteria to select the model. Finally, the research will focus on to integrate building non-linear dynamics due to large time constant and other second order factors (internal gains, solar gains, changes in climatic/operating conditions of building etc.) in the selected model. The summary of the research framework is shown in Figure (1.3).

[3] Refers to the phase when the activities of building operations started to be used

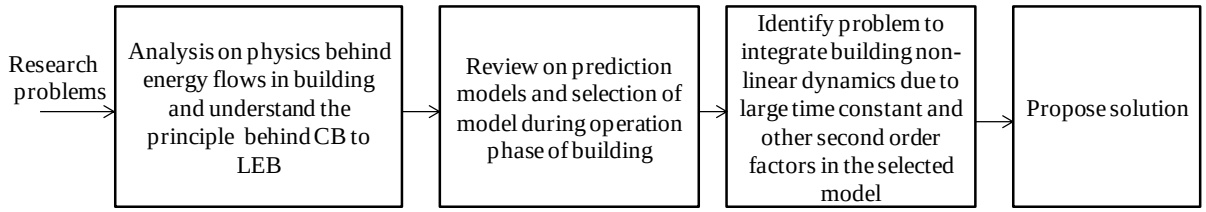

Figure 1.3: Summary of our research framework

1.5 Manuscript Outlines

This manuscript is organized as shown in Figure (1.4). **Chapter 2** will provide about LEB concepts and its evolution trends while migrating from CB to LEBs in Europe. It also explains the factors that govern energy-efficient measures for LEB. At last, it presents review and state of art on building energy models. It also compares different building energy models based on several factors and suggests criteria to select model during the operation phase of building.

Chapter 3 proposes an artificial intelligence (AI) model for modeling the LEB. It introduces two kinds of modeling approaches: "**all data**" and "**relevant data**". It then discusses different steps to prepare data for both kinds of modeling approaches. For instance, firstly, it describes classification/clustering methods to classify building operation according to week type. Secondly, it detail on novel pseudo dynamic model to generate additional input to model and to encompass building indoor characteristics. Finally, it provides derived climatic conditions generation steps and describes climatic variables selection method to identify most important climatic conditions that governs the building load.

Then it describes on different "**relevant data**" modeling approaches to select small representative datasets to build an **AI model**. These relevant data modeling methods are based on simplified physical methods: heating-degree-day (HDD) and modified HDD, and pattern recognition methods: Frechet distance (FD) and Dynamic time warping (DTW). Finally, different machine learning AI models: Artificial Neural Network (ANN), Support Vector Machine (SVM), Boosted Ensemble Decision Tree (BEDT) and Random Forest (RF) and their practical aspects are highlighted.

Chapter 4 discusses the application of methodology to predict thermal load for simulation building. It further describes the step-by-step process to apply the methodology for single-zone building model. It also provides comparison of different **AI models** and "**relevant data**" modeling methods. In addition, it also compares "**all data**" and "**relevant data**" modeling

approaches. Finally, the methodology is also evaluated with different occupancies profiles and multi-zone building models.

Chapter 5 presents the application of the methodology on real building to predict thermal load using "**relevant data**" modeling approach and compare with "**all data**" approaches.

Chapter 6 draws a summary and recommends future steps.

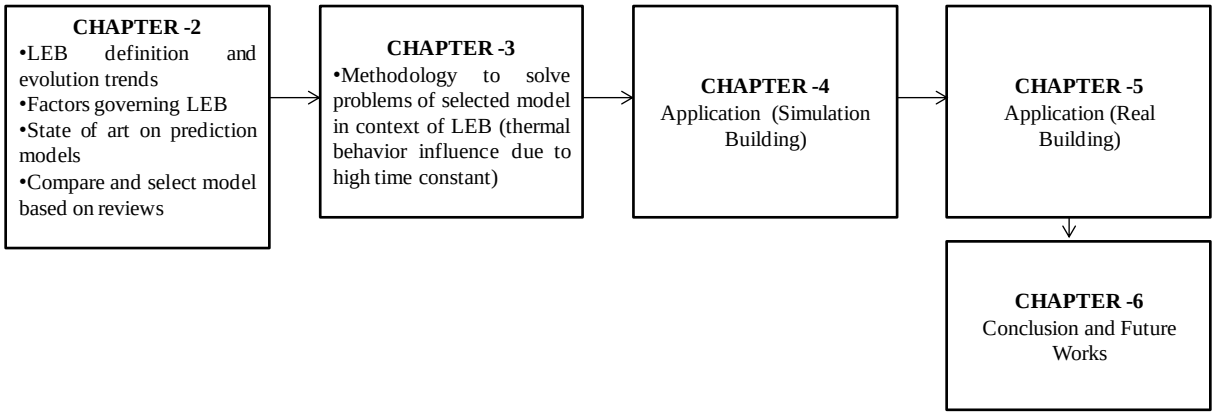

Figure 1.4: Summary of manuscript outlines

Chapter 2: Low Energy Building Modelling

2.1 Low Energy Building (LEB)

2.1.1 LEB concepts

The conventional building (CB) in this manuscript refers to a building with low insulation and high air leakage requiring high energy consumption. Besides high energy consumption, this type of building has several drawbacks such as a large peak power driven by weather conditions, environmental consequences (due to high energy requirement) and imbalance thermal comfort. There is no clear definition of LEB, but LEB refers to the building with high thermal performances or a building with a significant reduction in the annual energy consumption for the same level of thermal comfort compared to CB. Those high thermal performances are achieved by a good design, for example, the building compactness, fulfilling the building codes and standards, providing quality of thermal comfort, reliable building operations and using active and passive technologies. The active technologies are based on the use of mechanical, electrical and electronic equipments like the use of renewable sources, the use of heat pumps coupled with air or ground in the building. The passive technologies are based on building thermal envelope improvements like the thermal insulation, the quality of windows and the use of natural light to pass to interior spaces during a day. Due to the integrated design solutions, LEBs reduces 80% of the operational cost [5].

Generally, thermal performances are measured based on a significant reduction in annual energy consumption. This energy consumption can be a primary energy or directly measured which in this case is called end-use. The primary energy quantifies the energy resources at the generation or production sites and end-use energy consumption is measured at the final level of buildings.

The LEBs are defined by a low annual energy consumption and a low heat peak at the end-use and the annual energy consumption (Q_A) and the heat peak load ($q_{A,peak}$) link is given by the Equation (2.1) [6]:

$$\frac{Q_A}{A_{buil,env}} = a_0 + a_1 f_{o,buil} \quad (2.1)$$

$$q_{A,peak} \leq b_0 + b_1 f_{o,buil} \quad (2.2)$$

$$f_{o,buil} = \frac{A_{buil,env}}{V_{o,buil}} \quad (2.3)$$

Where, $f_{o,buil}$ is the building shape factor (m^{-1}), $A_{buil,env}$ is the external wall area of building construction (m^2) and $V_{o,buil}$ is the external volume of the heated space in the building (m^3). The values for a_0, a_1, b_0 and b_1 are reported in Table (2.1).

LEBs characterization varies with the location of the countries due to the weather conditions. Mumovic and Santamouris [6] linked the building shape factor $f_{o,buil}$ in terms of annual heating energy consumption and the heat peak load for different countries in Europe (see Table 2.1). It clearly illustrates that LEBs standardization varies in Europe according to the building shape factor.

Country	$Q_A/A_{buil,env}$ ($kWh/m^2.year$)	$q_{A,peak}$ (W/m^2)
Austria	24.55 + 81.82 $f_{o,buil}$	3.11 + 10.36 $f_{o,buil}$
Germany	26+ 13 $f_{o,buil}$	2.25 + 1.6 $f_{o,buil}$
Slovenia	45+40 $f_{o,buil}$	6+5.33 $f_{o,buil}$
Rest of Europe	13.64 + 45.45 $f_{o,buil}$	1.73 +5.76 $f_{o,buil}$

Table 2.1: LEBs in different parts of the Europe [6]

In Central Europe, LEBs are standardized for an improvement in energy consumption of 30% to 50% to CB. Such LEBs have an annual heating energy consumption of 40-60 kWh/m^2.year. In Czech Republic, LEBs are characterized by the U-value of building envelope and their U-value should be improved by 66% to CB. Similar trend of increasing performances of building envelope is seen in Germany and they defined 30% to 45% improvement in the quality of building envelope [7].

In France, five labels of energy performances are defined: Haute Performance Energétique (HPE), HPE EnR (Energie Renouvelable), THPE (Très Haute Performance Energétique), THPE EnR and BBC (Bâtiment basse consummation). The effinergie[4] standardized the LEBs (BBC) criteria by an average annual requirement for heating, cooling, ventilation, hot water and lighting of ≤40-65

[4] French association for environment which promote low energy consumption buildings

kWh/m^2.year (in primary energy and depending on location) for new dwellings [7] compare to CB (190 kWh/m^2.year in primary energy)[5].

2.1.2 Evolution of LEBs

Due to thermal energy standards, LEBs are also emerging to very low energy building (VLEB) or passive energy building (PEB) and nearly zero energy building (NZEB) in the Europe.

VLEBs or PEBs focus on passive technologies and provide an equilibrium indoor climate in summer and winter without the need of conventional heating system. These PEBs provide more effective "free heat gains" from solar radiations. The other requirements of PEB are reduction in unwanted air leakage through building fabric and limiting thermal transfer (U-value) of building envelopes. According to Feist[6], "*A passive house is a building, for which thermal comfort can be achieved solely by post-heating or post-cooling of the fresh air mass, which is required to achieve sufficient indoor air quality conditions-without the need for additional recirculation of air*". PEB in terms of annual energy consumption (Q_A) and peak heating load ($q_{A,peak}$) is given by [6]:

$$\frac{Q_A}{A_{buil,env}} \leq 4.1 + 13.64\, f_{o,buil} \tag{2.4}$$

$$q_{A,peak} \leq 0.5 + 1.73\, f_{o,buil} \tag{2.5}$$

PEBs characterization also varies with the countries. In central Europe, the maximum specific supply air heating load should be ≤10 W/m^2 and maximum annual heating energy consumption should be ≤15 kWh/m^2.year (in end-use) to achieve thermal comfort without using a conventional heating system. In addition, to fulfill PEBs requirement, the air-tightness should be ≤0.6 h^{-1} and percentage of time operative temperature (above 20 ^{0}C) should be around 10% [3]. In Czech Republic, PEBs are characterized by the U-values of building envelope and the criteria are: U-value of wall less than or equals to 0.3 W/m^2.K, U-value of roof equals to 0.12 W/m^2.K and U-value of window equals to 0.8 W/m^2.K [7].

NZEBs are LEBs which are integrated to on-site renewable energy sources (RES) to meet the remaining energy of building tending to be zero energy consumption requirements. There are several definitions of NZEB and are called by different names like zero energy building, nearly zero energy building, net zero energy building, energy positive building, zero carbon building etc. According to European Commission EPBD, NZEB is defined as a building which has very high

[5] http://www.concept-bio.eu/the-thermal-regulation-2005-rt2005.php
[6] http://www.passipedia.org/basics/the_passive_house_-_definition

energy performances in which nearly zero or very low amount of energy required is covered by a significant amount of energy from renewable sources including energy from renewable sources produced on-site or nearby [8]. Torcellini et al. [9] define a zero energy building as a building that could meet an energy requirement at relatively low cost from on-site generation of renewable sources. This further implies that this building could produce a significant renewable energy to meet or surplus annual energy requirement to achieve net zero energy consumption and/or zero carbon emissions. These NZEBs have the similar characteristics between LEBs and VLEBs/PEBs. The only major differences are NZEBs produces on-site RES or integrate RES on buildings to make an energy consumption requirement zero.

The summary of evolution from CB to different LEBs is shown in Figure (2.1)-Figure (2.2). It can be noticed that while the building is migrating from CB to LEBs, building characteristics: insulation, thermal performances of windows/glazing and air-tightness go on increasing. Figure (2.2) further illustrates that building envelope loss goes on decreasing while the building is transforming from CB to LEBs.

The migration pathways to LEBs for different countries in Europe are shown in Table (2.2). The energy performance improvements are based on annual heating energy consumption. It can be observed that most of the European countries will migrate to NZEB by 2020. For instance, in Denmark, the national regulation impelled to reduce energy consumption by 25% in 2010, 50% in 2015 and 75% in 2020 compared to CB. In France, national regulation aimed to reduce the energy consumption by 50% in 2012 and planned to migrate to NZEB by 2020. Similarly, the national regulation targeted to reach PEB by 2013 in United Kingdom and reveals that LEBs had been implemented already.

Country/Year	2008	2010	2012	2013	2015	2016	2020
Austria					VLEB		
Denmark		25%			50%		*75%
Hungary			LEB				NZEB
France			50%				NZEB
Germany	30%		60%				NZEB
Netherlands		25%			50%		NZEB
United Kingdom		25%		*44%		NZEB	

*: VLEB/PEB

Table 2.2: Migration pathways from CBs to LEBs in Europe (in terms of annual energy consumption) [7]

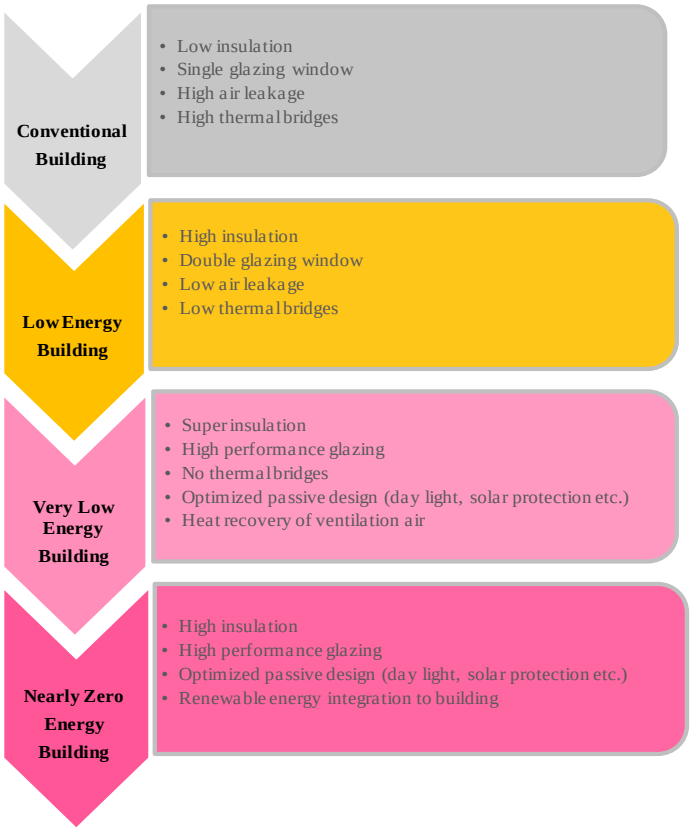

Figure 2.1: Summary of evolution from CB to LEBs ([3], [6])

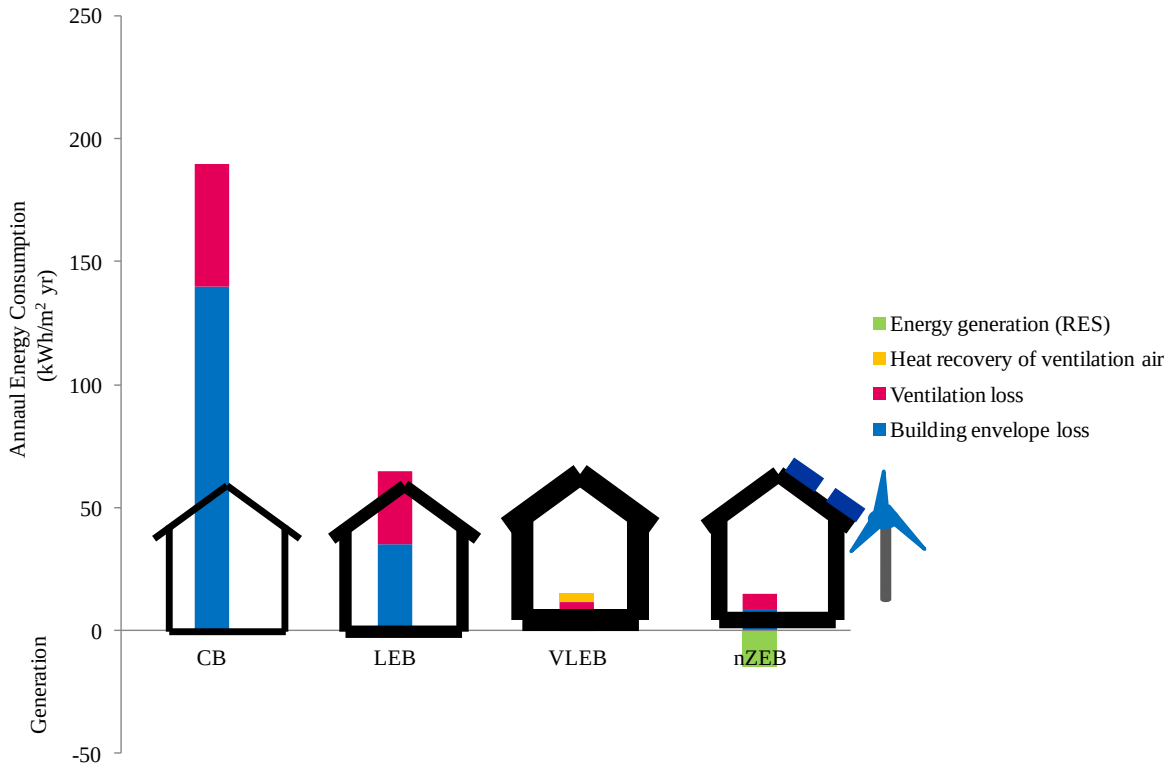

Figure 2.2 : Comparison of different LEBs with CB (general context in Central Europe)

2.1.3 Factors affecting LEB

The energy performance of a building depends on the design factors, thermo-physical properties of building construction, climatic conditions and building operating conditions. Apart from climatic conditions, occupancy and building operating conditions, the major factors affecting both CB and LEB are shown in Figure (2.3).

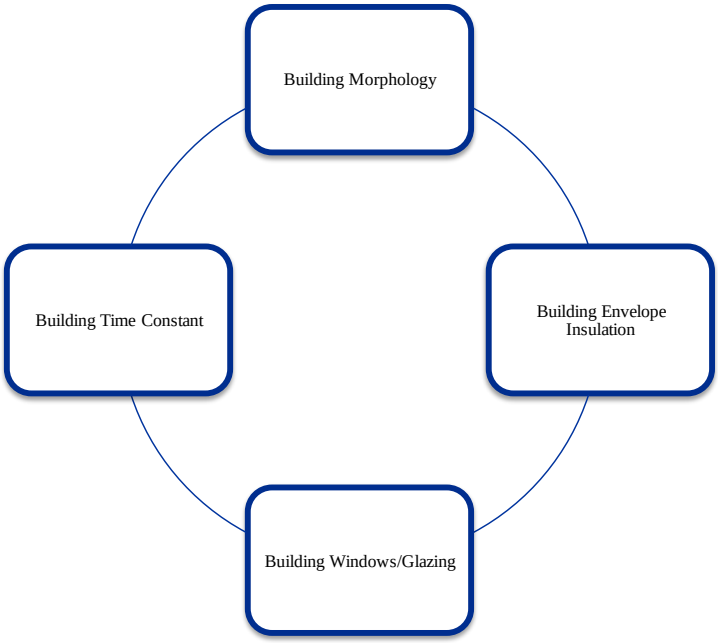

Figure 2.3: Factors affecting LEB

Building Morphology

Building morphology is an important indicator to determine energy demand of building. The building's shape also defines the morphology and is given by Equation (2.3). A compact building has less thermal envelope hence has less heat losses and this decreases its final energy demand. The building shape factor thus can be reduced by a compact building design. LEBs generally have a small shape factor. More detail about building morphology is given in Pessenlehner and Mahdavi [10].

Building Envelope Insulation

The building envelope is the boundary between outdoor and indoor. It consist external walls, floors, roofs, ceiling, windows and doors. These envelope components play a major role in improving an energy efficiency of building. A study on thermal building envelope components and passive energy savings can be found in Sadineni et al. [11].

The insulation is one of the main material components of building envelope. It helps to reduce the heat transfers between indoor and external weather conditions mainly driven by the temperature gradients. This is usually obtained by adding layers with very low thermal conductivity (W/m.K) in the envelope contributing low U-values (thermal transmittance, $W/m^2.K$) or high R-value (thermal resistance $m^2.K/W$). The high insulation materials help to maintain an equilibrium indoor condition due to slow heat transfer between envelopes and indoor conditions. "Super insulation" and "Over insulation" is a suggested approach to reduce the heat transfer loss in building envelope (U-values ranging from 0.15 to 0.10 $W/m^2.K$) [12].

The typical U-values and R-values of building envelope components for CB and LEBs are shown in Table (2.3). One can see that the most significant differences are in the U-values of walls and roofs, and shows that the insulation in LEB has been increased by at least 10% to CB.

Elements of Building	Type of Building	U-value ($W/m^2.K$)	R-value ($m^2.K/W$)
Wall	CB	2.5	0.4
	LEB	0.25	4
	VLEB/PEB	0.15	6.67
Roof	CB	1.9	0.53
	LEB/PEB	0.15	6.67
Glazing	CB	4.8	0.21
	LEB	2	0.37
	VLEB/PEB	0.8	1.25

Table 2.3: Typical U-values and R-values of CBs and LEBs in Europe [13]

Building Window/Glazing

The area of windows/glazing provides a significant role as a means to provide "free heat gains" in building with solar gains. It also balances the thermal comfort and illumination inside the building. The window-to-wall ratio is used as an indicator to evaluate thermal performance and Li et al. [14] mentioned that it is possible to reduce the amount of heat gain/loss by simply lowering the window-to-wall ratio. Persson et al. [15] highlight that efficient windows (higher glazing and lower U-value materials) would have a significant contribution on energy demand reduction compared to highly insulated wall without windows.

Table (2.3) shows the U-value and R-value of glazing for different types of building in Europe and these classifications are based on glazing components. For instance, CB normally have single and/or double glazing, LEB have double glazing and PEB have triple glazing. It can be observed

that CB has a glazing insulation 6 times worst than well insulated glazing (U-value of glazing in CB 4.8 W/m^2.K compared to 0.8 W/m^2.K in PEB).

Building Time Constant

The time constant of building is another important design factor to evaluate the performance of a building. It is a measure of the thermal response of building and is defined as a function of total heat capacity of building and insulation level. A high building time constant is achieved by combining high heat capacity and low U-value. It thus determines the effect of dampening of indoor temperature fluctuations corresponding to the external temperature. It is independent on the size of buildings, for instance, large and small buildings can have the same rates of response to temperature changes. Generally, LEBs envelope has a large time constant, more than 100 hours [16]. A first approximation of the global time constant of building (τ_{buil} expressed in h) is given by[7]:

$$\tau_{buil} = \frac{C_{buil}}{UA_{buil}} \approx \frac{\sum_j C_{env,j}}{\sum_j \frac{A_{env,j}}{R_{env,j}}} = \frac{\sum_i C_{env,j}}{\sum_j A_{env,j} U_{env,j}} \tag{2.6}$$

where, $C_{env,j}$ is the thermal capacity for each envelope construction component j (J/K). $U_{env,j}$ (W/m^2.K), $A_{env,j}$ (m^2), $R_{env,j}$ (m^2.K/W) are the U-values, i.e., heat loss factors, area and thermal resistance of each envelope construction component j including ventilation respectively. UA_{buil} is the overall building heat loss coefficient (W/K) and C_{buil} is the total heat capacity of a building. The total heat capacity of a building depends on thickness and surface area of envelope components and can be estimated by summing heat capacities of building envelope layers in contact and is given by:

$$C_{buil} = \sum_j \rho_{env,j} \, C_{p,env,j} \, A_{env,j} \, \Delta S_{env,j} \tag{2.7}$$

Where, $\rho_{env,j}$ is the density of building envelope j (kg/m^3), $C_{p,env,j}$ is the specific heat capacity of building envelope j (J/kg.K) and $\Delta S_{env,j}$ is the thickness of building envelope j (m). Higher heat capacity of a building means higher thermal mass and slower response to heat up a building. The heavy weight construction material in the building such as clay, concrete, stone have high thermal capacity and are main attributes of the thermal mass.

[7] See the next section about the main assumptions for Equation (2.6) and the symbol ≈ means all the heat resistances are obviously not in series.

2.2 Building Energy Model

2.2.1 Introduction

Building energy demand can be estimated and predicted using two modeling approaches: top-down and bottom-up approaches [17]. The top-down modeling approach estimates the long-term total energy demand and is based on macroeconomic indicators (gross domestic product, unemployment and inflation), energy price and general climate. It also roughly reduced to the scale of a district or a building. On the contrary, bottom-up approach estimates the individual energy demand of a building and aggregates it for the whole energy demand at the scale of a district. Top-down modeling approach thus fails to consider discontinuous advances in technology, and bottom-up modeling approach fails to take into account some effects such as the profusion of human behaviors. Since the aim of this research is the prediction of energy demand or consumption for different types of buildings, the prediction model based on bottom-up modeling approach is only reviewed.

There are several approaches to model the building energy based on partial and ordinary differential equations, steady and unsteady equations, design and control models. In this study, building models are classified into three categories: **white-box**, **black-box** and **gray-box** as in [18][19][20][21][22][23] and the general overview of these three models is shown in Figure (2.4).

Definition 2.1: Building parameters – Model parameters

> *The factors, for example, window to wall ratio, U-value of building envelope etc. that influences the energy demand of a building are defined as building parameters. The sets of input values given to a model e.g., hidden neurons in neural network, kernel function in support vector machine, number of trees in decision tree etc. (see Appendix B.1, B.2 and B.3) for a black-box model or thermal resistance R and thermal capacitance C for a gray-box model (see Section 2.2.3) are called the model parameters.*

- First, the **white-box** models estimate the energy demand from detailed physical understanding of building and imply numerous degrees of non-linearity. These are also called fundamental models since they derive from fundamental principle of energy balance in buildings.

These kinds of model are based on prior knowledge and the model structure is completely dependent on physical principles. In addition, the models are built using detailed physical principles and have advantages for understanding the building energy system and energy flows. The model does not depend on measurement or in-situ experiments and their parameters (see **definition 2.1**) have direct physical meaning thus measurement data is not required to make new prediction for such models.

- Second, the **black-box** models only depend on empirical data or data acquired from dynamic thermal energy simulations. So, these are also called data-driven model or inverse model since they predict the behavior from known measurement system (or from numerical simulations). They draw functional relationship of variables (model structure) and building parameters (see **definition 2.1**) are learned from measurement or empirical data. For developing such models, no prior knowledge is required and model parameters have no physical sense. Such models are more suitable for adapting the future environment hence are useful during the operation phase of building.

- Third, the **gray-box** models estimate the building energy demand combining physical understanding using model order reduction and data fitting techniques obtained from empirical data. For such model, the model structure strongly depends on prior knowledge (e.g. models are represented in the form of differential equation represented by lumped resistance and capacitance networks). In addition, model parameters, for example, thermal resistance R and thermal capacitance C assigned to the elements in the zone are determined from the empirical or measurement data.

Definition 2.2: Input Features

Input features are defined by the sets of inputs (e.g., external temperature, occupancy) that are used to build a model. These are also called by the name "input variables" without any distinction.

One can concludes from Figure (2.4) that the internal structure of building energy model and physical interpretation of parameters of model go on decreasing while the model is transforming from **white-box** to **black-box**. In this section, reviews of each models based on input features (see **definition 2.2**) use and their application in building load prediction is presented.

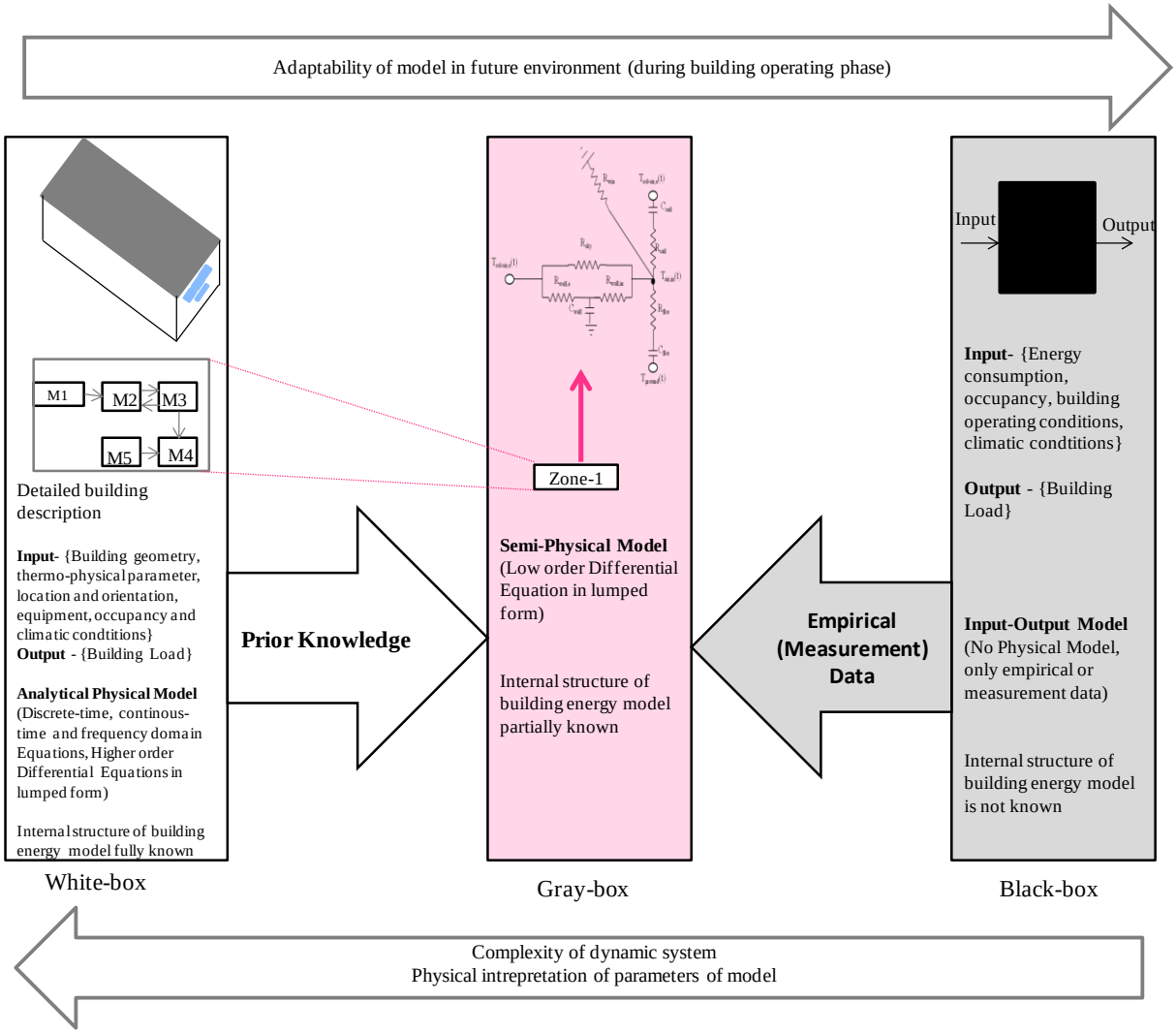

Figure 2.4: White-box, gray-box and black-box models

Definition 2.3: White-box - Black-box – Gray-box

> *A model is classified as a **white-box** model (also called knowledge model) when equations are based on physics and all input features are already exist, a model is a **black-box** model (behavior model) when it is an input-ouput model based on in-situ experiments. A model is considered as a **gray-box** model, when some experiments are required in-situ to identify some parameters of the physical equations.*

Nevertheless, Berthou, page 15 [24], indicates that the borders between gray and black box models are fuzzy. Berthou, page 47 [24] also indicates that for a R4C2 (or R3C2 without a variable mechanical ventilation) it is difficult to attribute the heat flux (between air and pieces of furniture in one way and wall, ceiling, and floor in other way) and so such a model is not a **white-box** model but it is classified as a **gray-box** model.

Whatever those fuzzy borders, artificial intelligence methods (artificial neural network, support vector machine etc.) belong to the **black-box** model and requires in-situ measurements.

2.2.2 White-box Model

Based on complexity of these heat transfer equations, **white-box** models are broadly classified into two categories: steady state models and dynamic models. A steady-state model neglects the important aspects of time constant of a building. A study by Al-Homoud [25] summarizes the simple physical methods to estimate energy demand of a building. The steady state models are outlined in Appendix A.

The zone modeling applied to building is only described in the following in agreement with the subject of this research.

Definition 2.4: Thermal Zone

> *A "thermal zone" is defined by a confined volume in which the inside temperature is assumed homogeneous, so all the thermal properties are constant. Consequently, the inside mass can be viewed as one point with a mass limited by the volume.*
> *A thermal zone can be defined for a room or sets of rooms and this is mainly a modeling assumption.*

The heat flux stored within the controlled volume in definite interval of time is equal to the amount of heat flux entering in the studied volume, the heat flux exiting from that volume and the heat flux dissipated in that volume and is expressed in Equation (2.8):

$$\frac{d}{dt}Q_{st} = \emptyset_{in}(t) - \emptyset_{out}(t) + \emptyset_{source}(t) \tag{2.8}$$

Where, Q_{st} is the stored amount of heat energy in the controlled volume (in J), $\emptyset_{in}(t)$ is the heat flux entering the controlled volume (in W), $\emptyset_{out}(t)$ is the heat flux leaving from the controlled volume (in W) and $\emptyset_{source}(t)$ is the dissipated amount of heat flux from the surface of controlled volume (in W). In Equation (2.8), ($\emptyset_{out}(t)$ -$\emptyset_{in}(t)$) signifies the amount of heat gain/loss from the controlled volume and this gain/loss may be in the representation of conduction transfer, solar radiation, ventilation and internal sources.

Zonal Modeling

The **white-box** building energy models are built with multiple thermal zones (see **definition 2.4**) and in this section single zone building is considered. Figure (2.5) shows the heat transfer between buildings and external climatic environment. The indoor volume is bordered by its envelope (external walls and windows) which separates it to the external climatic conditions. This building is equipped with HVAC system to provide heating and/or cooling by fresh air circulating between the indoor zone and the air handling unit (AHU) through air ducts. Heat flows out from the zone when the indoor temperature of building is above the outside temperature. Heat is also transferred through the zone envelope such as walls, layers of materials and windows. Inside the zone envelope, three types of heat transfer occurs: conduction, convection and radiation, for example, heat is flowed by conduction in envelope. The solar radiation is transmitted and reflected back through transparent glazing and is also absorbed by the indoor surfaces. Due to the presence of occupants, the use of electrical lighting and other appliances, heat is added in the zone. It is also noticed that radiation heat transfer occurs through external and internal envelope on its surroundings.

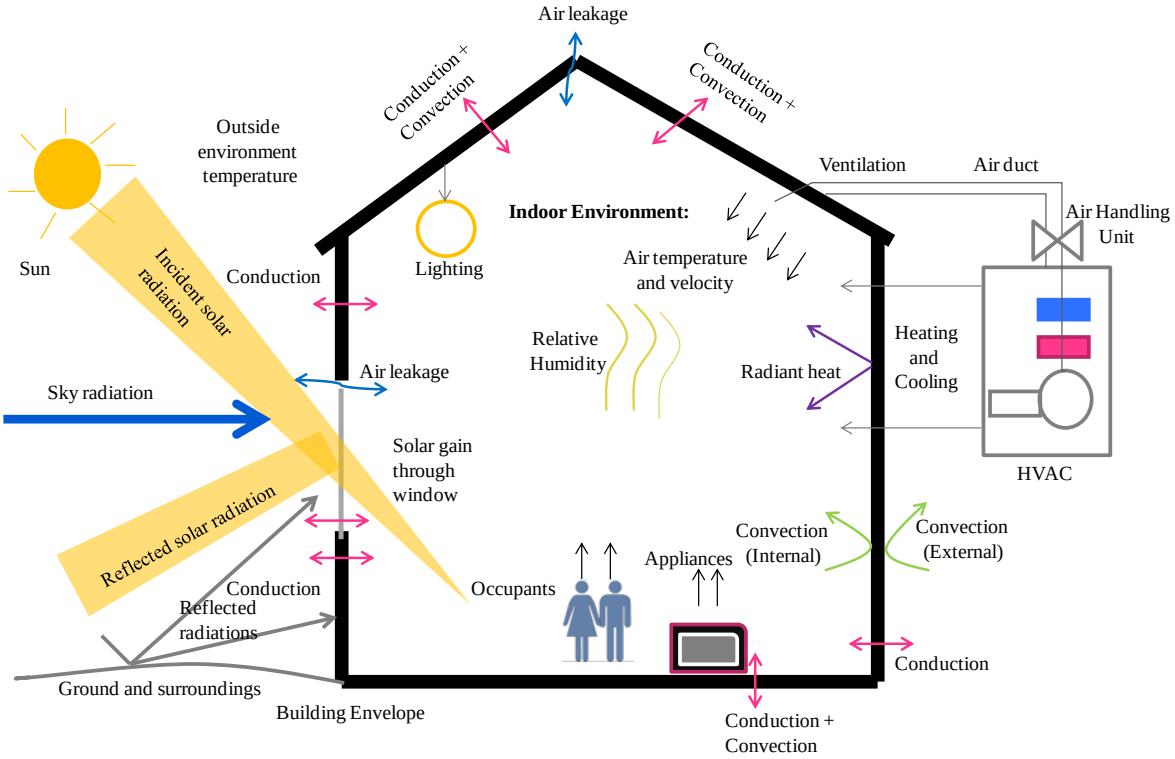

Figure 2.5: Scheme of energy flows in a building

The energy balance of the in-air zone is influenced by heat transfer through building envelope, air flows through ventilation, air flows through leakage and internal gains. Due to this heat and air

flows inside the building, thermal behavior of a building are changing with time. These thermal behaviors eventually result in changing temperature in the indoor environment. The energy balance of in-air zone can be modeled as a single zone with several assumptions:

- The air inside the zone is perfectly mixed with the state of indoor air resulting in uniform air distribution in the zone due to which the zone has the same properties such as temperature and humidity.
- The surface envelopes in the zone are supposed to have uniform surface temperature, uniform solar irradiance and uniform radiant gain.
- Thermal bridges are neglected.
- The furniture inside the zone (e.g. chairs, tables etc.) and internal partitions are not considered and do not have any influence in indoor climate.

General representation of heat balance under steady state conditions of an in-air zone is given by:

$$\sum \dot{Q}_{h/c}(t) = \sum \dot{Q}_{int}(t) + \sum \dot{Q}_{source}(t) - \sum \dot{Q}_{out}(t) \qquad (2.9)$$

Where, $\dot{Q}_{h/c}(t)$ is the heating or cooling required to balance heat in the in-air zone (in W), $\dot{Q}_{int}(t)$ is the internal heat gain due to occupants, lighting and appliances (in W), $\dot{Q}_{source}(t)$ is the sum of heat gain inside the in-air zone (in W), $\dot{Q}_{out}(t)$ is the sum of heat loss from the in-air zone (in W). The in-air zone is bordered by firstly opaque envelope components such as wall, roof and basement floor and by secondly transparent envelope components like windows and glazing surfaces. Considering ventilation in the zone, Equation (2.9) can be further modified as:

$$\sum \dot{Q}_{h/c}(t) = \sum \dot{Q}_{int}(t) + \sum_{i} \dot{Q}_{sol,i}(t) \pm \sum_{j} \dot{Q}_{env,j}(t) \pm \dot{Q}_{ven}(t) \qquad (2.10)$$

Where, $\dot{Q}_{sol,i}(t)$ is the solar heat gain through transparent envelope components *i* (in W), $\dot{Q}_{env,j}(t)$ is the heat gain or loss through zone envelope components *j* (e.g. walls, roof, window etc.) (in W) and $\dot{Q}_{ven}(t)$ is the ventilation heat gain or loss due to air exchange (in W). Heat transfer through envelope (wall, roof and window) thus can be modeled by considering interactions with the external and internal environment and is dominated by conduction and convection heat transfer processes.

Figure (2.6) shows a simple representation of building energy model using lumped resistance for illustration where 2R1C network represents building envelope walls, 1R1C network represents roof and 3R2C represents floors [26]. The window is simply represented by thermal resistance without storage due to its fast response of heat transfer. To be noted that **white-box** models are

usually built with complex lumped resistances model resulting in large number of building parameters (see **definition 2.1**). This lower order RC thermal network shown in Figure (2.6) is just for an illustration. The effect of solar radiation on an opaque building envelope components like walls and roof are considered by replacing environmental external temperature by sol-air temperature. This sol-air temperature takes into account incident solar radiation, radiation exchange with sky and the surrounding surfaces and is given by [27]:

$$T_{sol-air}(t) = T_{air,e}(t) + \frac{\alpha_{s,env}\, G_t(t)}{h_o} - \frac{\varepsilon_{s,env}\, \sigma \left(T_{air,e}^4(t) - T_{surr}^4(t)\right)}{h_o} \qquad (2.11)$$

Where, $T_{sol-air}(t)$ is the sol-air temperature (in K), $T_{air,e}(t)$ is the external air temperature (in K), $\alpha_{s,env}$ is the solar absorptivity on the envelope surface (-), $G_t(t)$ is the solar radiation incident on building envelope surfaces (in W), h_o is the heat transfer coefficient on the exterior envelope surfaces (in W/m^2.K), $\varepsilon_{s,env}$ is the emissivity on the envelope surface (-) and $T_{surr}(t)$ is the surrounding temperature (in K). The heat transfer through envelope *j* can be represented as:

$$\sum_j \dot{Q}_{env,j}(t) = A_{win}U_{win}\left(T_{air,in}(t) - T_{air,e}(t)\right) + A_wU_{w,in}\left(T_{air,in}(t) - T_w(t)\right)$$
$$+ A_fU_f\left(T_{air,in}(t) - T_{fg}(t)\right) + A_{rf}U_{rf}\left(T_{air,in}(t) - T_{sol-air}(t)\right) \qquad (2.12)$$

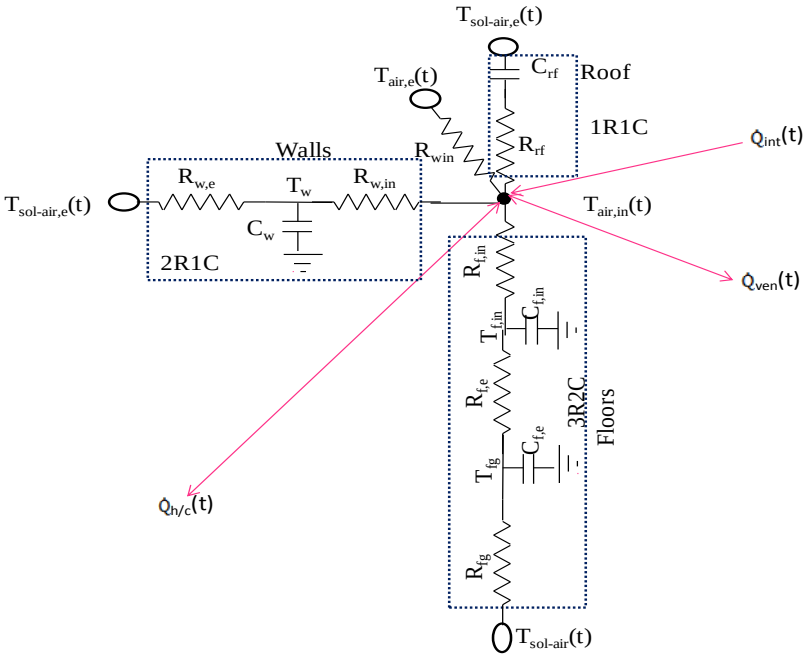

Figure 2.6: Simple illustration of building energy model using lumped resistance and capacitance

Heat capacity of each envelope components represents the thermal storage in zone and energy balance equation of each component is given by Equation (2.13) – (2.16):

$$\frac{dT_w(t)}{dt} = \frac{A_w}{C_w}\left[\frac{T_{air,in}(t) - T_w(t)}{R_{w,in}} + \frac{T_w(t) - T_{sol-air,e}(t)}{R_{w,e}}\right] \quad (2.13)$$

$$\frac{dT_f(t)}{dt} = \frac{A_f}{C_{f,in}}\left[\frac{T_{air,in}(t) - T_{f,in}(t)}{R_{f,in}} + \frac{T_{f,in}(t) - T_{fg}(t)}{R_{f,e}}\right] \quad (2.14)$$

$$\frac{dT_g(t)}{dt} = \frac{A_f}{C_{f,e}}\left[\frac{T_{f,in}(t) - T_{fg}(t)}{R_{f,e}} + \frac{T_{fg}(t) - T_{sol-air,e}(t)}{R_{fg}}\right] \quad (2.15)$$

$$\frac{dT_{rf}}{dt} = \frac{A_{rf}}{C_{rf}}\left[\frac{T_{air,in}(t) - T_{sol-air,e}(t)}{R_{rf}}\right] \quad (2.16)$$

Where, $T_w(t)$ is the wall temperature (in K), $T_{air,in}(t)$ is the in-air temperature (in K), $T_{f,in}(t)$ is the internal floor temperature (in K) and $T_{fg}(t)$ is the exterior floor temperature (in K). A_w, A_{win}, A_f, A_{rf} are the areas of walls, windows, floor and roof respectively (in m^2). $U_{w,in}$, U_{win}, U_f and U_{rf} are the U-values of indoor walls, windows, floor and roof respectively (in $W/m^2.K$). $R_{w,in}$, $R_{w,e}$, $R_{f,e}$, $R_{f,in}$, R_{fg} and R_{rf} are the thermal resistances of indoor walls, outside walls, exterior floor resistance, interior floor, ground conduction coefficient and roof respectively (in $m^2.K/W$). $C_{air,in}$, C_w, $C_{f,in}$, $C_{f,e}$ and C_{rf} are the heat capacities of in-air, walls, interior floor, exterior floor and roof respectively (in J/K). The heat gain in the zone due to solar radiation assuming *i* number of window/glazing in the zone is given by:

$$\sum_i \dot{Q}_{sol,i}(t) = \sum_i \tau_{g,i}\alpha_{g,i}A_{g,i}G_t(t) \quad (2.17)$$

Where, $A_{g,i}$ is the area of glazing *i* (in m^2), $\tau_{g,i}$ is the transmittance on glass plane *i* (-) and $\alpha_{g,i}$ is the solar absorptance on a glass plane *i* (-).

The heat demand due to ventilation system $\dot{Q}_{vent}(t)$ can be simplified as:

$$\dot{Q}_{vent}(t) = \zeta_v\, \dot{q}_v\, \rho_{air,in}\, C_{p,air,in}\, \left(T_{air,in}(t) - T_{air,e}(t)\right) \quad (2.18)$$

Where, ζ_v is the factor of ventilation system (e.g. HVAC system), $\dot{q}_v$ is the volumetric flow of ventilation air (in m^3/h), $\rho_{air,in}$ is the indoor air density (in kg/m^3) and $C_{p,air,in}$ is the specific heat capacitiy of the indoor air (in J/kg.K).

The internal gain of the zone represents the heat gain from occupants and their activity, lighting and appliances uses. The internal gain ($\dot{Q}_{int}(t)$ expressed in W) is given as:

$$\dot{Q}_{int}(t) = \dot{Q}_{ocup}(t) + \dot{Q}_{lit}(t) + \dot{Q}_{app}(t) \quad (2.19)$$

Where, $\dot{Q}_{ocup}(t)$ is the heat gain due to occupancy (in W), $\dot{Q}_{lit}(t)$ is the heat gain due to lighting (in W) and $\dot{Q}_{app}(t)$ is the heat gain from appliances (in W). The heat gain due to occupancy in the zone is given as:

$$\dot{Q}_{ocup}(t) = N_{ocup}\, \dot{Q}_{ocup,b}(t) \tag{2.20}$$

Where, N_{ocup} is the number of occupants (-) and $\dot{Q}_{ocup,b}$ is the heat generation rate from occupants in the zone. The heat gain due to lighting is given as:

$$\dot{Q}_{lit}(t) = \sum_{j} A_{fz}\, P_{lit,j}\,(t) \tag{2.21}$$

Where, A_{fz} is the floor area of the zone (in m^2) and $P_{lighting,j}$ is the specific electric power demand of light *j* in the zone (in W/m^2). The heat gain due to appliances is given as:

$$\dot{Q}_{app}(t) = \sum_{j} A_{fz}\, P_{app,j}(t) \tag{2.22}$$

Where, $P_{app,j}$ is the specific heat gain due to appliances *j* in the zone (W/m^2).

In order to model these transient behaviors of building energy model, nowadays there are several detailed simulation tools available. These simulation tools model the energy and fluid flows including HVAC and plant control system inside the building system in a dynamic way. Many tools have focused on individual components and whole building components, however, there are still limited tools developed to integrate the building systems like EnergyPlus [28], ESP-r [29], IBPT [30], SIMBAD [31], TRNsys[8] etc. These simulation tools developed are modular and transparent. They provide benefits to the individual developers to extend their own model and modify the existing models. Crawley et al. [32] made a comparison of twenty building energy simulation tools (DOE-2.1, EnergyPlus, ESP-r, TRNSYS, etc.). They found that comparison is difficult not only because of the programming language, but also what the given tools able to take into account can be different from one to another. They suggested that the modular capabilities of tools and requirement of future system help to proper choice the individual simulation tools.

Detailed simulation methods take into account the transient or change in the surroundings of a given system with inclusion of building thermal characteristics. Different numerical methods are implemented such as continuous-time (laplacian domain), discrete-time (z-domain) and frequency

[8] TRNSYS 17, a Transient system simulation program. http://sel.me.wisc.edu/trnsys/feature

domain or even higher order lumped to model the physical variables of building components. For example, these methods solve the heat transfer equations governing Fourier equation using finite difference methods and complex transfer functions equations to undertake building dynamics [33]. These detail simulation methods therefore consists several hundreds to thousands of equations in order to model detail air flows and heat transfer of building. It models the whole building components and its integrated sub-components and system considering physical properties of building. Since these detailed simulation tools are build with detailed mathematical equations governing the physical phenomena, they capture the thermal dynamic behavior of buildings efficiently.

It can be concluded that detailed methods uses complex physics based analytical model and are quite good to estimate and predict thermal energy demand for different building types including LEBs. However, such kind of model requires large number of building parameters and seems feasible only during an early phase design of new building rather than operation phase of the building.

2.2.3 Gray-box Model

Gray-box model is a combination between **white-box** and statistics (see **section 2.2.4**). It combines prior physical knowledge of building to model heat dynamics and determine the model structure, and then data fitting techniques to estimate model parameters (see **definition 2.1**) from empirical or measurement data. It includes heat gains or loss through thermal envelope based on temperature difference and overall heat transfer coefficient of wall, roof and glazing. The simplification of gray-box model is low-order thermal network model in the form of electrical resistance–capacitance (RC) circuit. The parameters R and C are modeled in differential equations are transformed into state space model to determine the transfer function. For estimating coefficient matrices of state space model, the boundary conditions are formulated based on prior knowledge of building geometry and materials and then optimization algorithm is used to find the parameter that reflects the physical information.

Bacher et al. [34] proposed short-term heat load forecasting for single family houses located in Denmark. The model was built with the data from sixteen houses. RC low pass transfer function model was developed to represent the heat dynamics. Ogunsola and Song [35] proposed 2R2C model (i.e., two resistances and two capacitances) for the office building located in University of Oklahoma for thermal load prediction. Their results were compared with real measurement and EnergyPlus, and found that their model and EnergyPlus provided similar cooling load prediction

with some under predictions. It was also observed that their model captured many fluctuations which were not captured by EnergyPlus.

Similarly, Wang and Xu [36] proposed 3R2C to model building envelopes including external walls and ceiling/roof and 2R2C to model building internal mass including floors, partitions and internal walls. The parameters of 3R2C model were determined based on frequency response characteristics and parameters of 2R2C model were identified from building operation data using genetic algorithm. Their results provided considerable accuracy of 90% for cooling load.

The higher order R6C2 model was also proposed by Berthou et al. [37] to estimate thermal energy demand. In their model, occupancy profile, ventilation set-point, temperature set-point and solar gains (solar gain on walls and solar gain transmitted through windows)[9] were used to identify parameters of R and C using interior point algorithm. Their model had an accuracy of 84% with an energy error below 2% for heating and cooling load estimation. They used the same model parameters during a year and concluded that R6C2 model was efficient for estimating thermal energy demand for a whole year where thermal power needs are high.

Rather than thermal analogous RC network, Lu et al. [38] proposed a model combining physical and statistical approach to predict heating energy consumption of heterogeneous buildings. The physical model was based on physics of energy flows in a building where they modeled thermal envelope, solar, ventilation, occupancy, lighting and appliances. Then the stochastic time series models were formulated based on lagged value of heating load, indoor and external temperature. The parameters of heterogeneity for different building types were obtained using convex hull technique and their results showed considerable accuracy during the prediction.

The summary of input variables and time step of prediction used in the literatures are detailed in Table (2.4).

S.N.	Author and Year	Type of Model	Features Used for Modeling										Time Step	Type of Applications
			External temperature	Solar radiation	Relative Humidity	Wind speed	Occupancy Profile	Function representing H,D,M,A,S	Operational behavior	Other Parameters	Indoor /Set-point temperature	Previous Hours/Day Energy Load		
1	Bacher et al. [34]	Gray-box×	×	×		×		×					H	Heating
2	Ogunsola & Song [35]	Gray-box×								1*:			H	Cooling
3	Wang and ×u [36]	Gray-box×	×	×	×		×		×	2*:	×		H	Cooling
4	Berthou et al. [37]	Gray-box×	×				×			3*:	×		H	Heating and cooling
5	Lu et al. [38]	Gray-box×	×	×			×				×	×	H	Heating

Table 2.4: Summary of input variables and time step of prediction using gray-box model

[9] Solar gain on wall and solar gain transmitted through window are based on geometrical information of building considering the shading effect, time of day and cloud cover data

1*: building envelopes and internal load components
2*: water flow rate, return and supply water temperature difference, indoor humidity, air flow rate, internal gains
3*: ventilation set-point, solar gains on walls, solar gain transmitted through windows
H: Hourly, D: Daily, M: Monthly, A: Annually, S: Seasonally

It can be thus concluded that **gray-box** models provide greater feasibility compared to **white-box** models due to the requirement of fewer features during the operation phase. Nevertheless, the complexity of model increases to fit the parameters of differential equations for large multi-zone building.

2.2.4 Black-box Model

Black-box models rely on a set of input and output data. For such model, the model parameters (see **definition 2.1**) are identified by statistical analysis between inputs and outputs measurements. It is also called input-output model since it maps their dependencies.

Remark 2.1:

Two main drawbacks of black-box model implementation can preliminary highlighted:

- *The objective function can have a lot of minima: there is no evidence for a global minimization.*
- *The results depend on the initial point.*

Remark 2.2:

It can be very difficult to obtain the model parameters of the black-box model. These models are prone to either under-fitting or over-fitting of data to obtain parameters shown in Figure (2.7) where θ represents parameters while fitting x input and y output. The under-fitting of model is due to improper design of model to fit the data. The over-fitting of model is due to complex behavior of data and tries to fit as much as possible. Reasonable fitting is in-between under- and over-fitting.

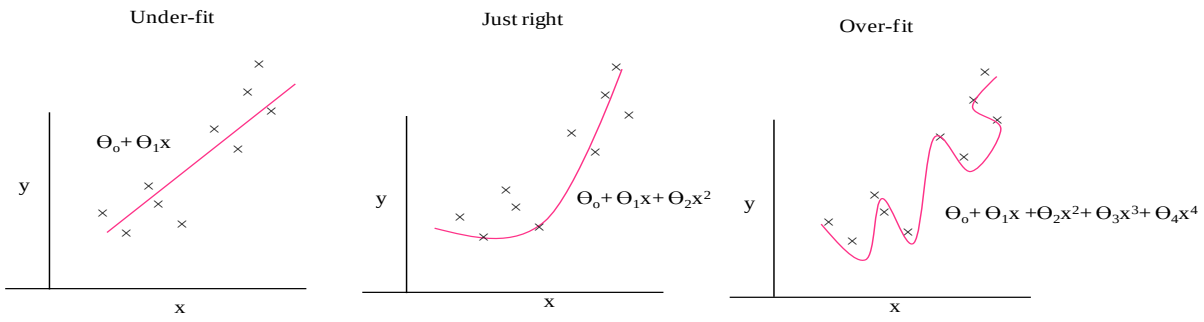

Figure 2.7: Under-fitting, reasonable-fitting (just right) and over-fitting of data

Three methods are reviewed hereafter to fit the **black-box** model:

- Regression methods
- Autoregressive model
- Machine learning methods (Artificial intelligence, see **Appendix B**): ANN, SVM, RF etc.

The applications of these models in the prediction of building energy demand or consumption are highlighted below:

Definition 2.5: Learning phase – Validation phase - Testing phase

For building energy prediction from black-box model, measurements or numerical behavior of input-output data used to build a model are called a learning phase and a part of training data reserved to select parameters of model are called validation phase. Testing data are the prediction day conditions data which are unknown in future to predict building energy demand or consumption are called testing phase. The training phase estimate the parameters of model whereas validation phase refers to the selection of best parameters of model by verifying if any increase in accuracy over training data actually yields a validation accuracy or not.

Definition 2.6: Batch Learning - Sequential Learning

In building energy model, if all the inputs-outputs training data are presented and model parameters are updated thereafter then such type of learning mechanism is defined as batch learning. Whereas, if model parameters are updated with each input-output training data presented, then such type of learning mechanism is defined as sequential or incremental learning.

Remark 2.3:

Even if a black-box model is an input-output model, the explanatory input features are the choice of the modeler and some useful statistical tools exist such as the principal component analysis and semi-physical understanding. So, a pre-treatment of the data can drive to the choice of those input features.

As an example, Lam et al. [39], Olofsson et al. [40], Olofsson and Andersson [41], Chaowen and Dong [42], Wan et al. [43] and Li et al.[44] used principal component analysis (PCA) in order to transform input data (climatic conditions like dry bulb temperature, wet bulb temperature, global solar radiation, clearness index, wind speed, humidity etc.) into principal components before developing prediction model. Lam et al.

[39] deduced several climatic conditions using PCA before predicting a long-term energy consumption of a building located in Hong Kong, China. They identified that clearness index and wind speeds were less significant than dry and wet bulb temperature and global solar radiation for cooling load. They also found that principal components had high correlation (R^2=0.87-0.96) with cooling load.

Yokoyama et al. [45] assumed building dynamics as a first order model and then applied first order differential operation on a training dataset to remove the trend and periodic changes of energy consumption and climatic conditions (external temperature and humidity). Later they used this converted dataset to estimate cooling load. They found that model performance has been increased (relative error ≈8%) while comparing without pre-treatment (relative error ≈11.3%). In order to reduce the degree of variations of energy consumption from seasonal behavior, Deb et al. [46] divided the training data into classes (very low, low, medium, high and very high) according to energy consumption as a pre-treatment step. Later, they used this training data to estimate cooling energy consumption for institution buildings. Their model after pre-treatment exhibits R^2 of 0.94 during prediction conditions, however, they did not compare their results without pre-treatment of input-output data.

Remark 2.4:

Besides pre-treatment of data to simplify the input features choice from statistical method, classification of data (in order to represent building operations according to week type) are the choice of modeler.

For instance, Lam et al. [47] and Gaitani et al. [48] used PCA ;Li et al. [49] used canonical variate analysis (CVA) ; Gao et al. [50], Santamouris et al. [51] and Gaitani et al. [48] used clustering analysis to classify the energy consumption data into different classes or group. As an example, Li et al. (2010) applied CVA to analyze the building operation classes of office building. They used six input variables: mean and peak daily energy consumption and dynamical change of energy consumption coefficients that are obtained from auto-regression model for classification. Their CVA results clearly distinguished two kinds of building operation classes: working day and weekend.

Remark 2.5:

Thermal performance of building depends on time dependent and independent variables. Time independent variables are the design variable that depends on building geometry such as building shape, zone height, envelope area, floor area, window to external area etc.; and thermo-physical factors such as building materials, thermal insulation etc. Time dependent variables are those variables that are varied according to time, for example, climatic conditions, occupant dynamics and operating building characteristics (set- point temperature, lighting and natural ventilation rate).However, the effects of both time dependent and independent variables greatly effects the performance of prediction model.

If all the input features, i.e., time dependent and independent variables (see **definition 2.2**) are used, then it increases the number of training data. This will further results in model complexity and increases the model training CPU-time. Generally, three types of feature selection methods: filter, wrapper and embedded are widely used ([52] and [53]). Filter method selects the input feature based on highest statistical correlation features only. Wrapper method selects the feature based on the accuracy of each feature in the prediction model (e.g, ANN, SVM etc.). Embedded method selects the best combination features evaluating the accuracy of features in the prediction model. It discards the lowest weight feature from the input feature. It is similar to wrapper method but it avoids multiple training of same feature. Therefore, feature selection is also the choice of modeler.

For instance, the feature selection were performed in literatures [54] [55] [56] and [57] to select significant input variables. Zhao and Magoulés [54] performed correlation coefficients and regression gradient guided based on k-nearest neighbor (k-NN) feature selection method on 23 features of building (climatic conditions, water mains temperature, zone total internal heat gain, number of occupants, window heat gain/loss on each wall, zone mean air temperature, zone infiltration volume, district heating outlet temperature and total heat gain from people, light and electricity etc.). They found that similar accuracy can be achieved with 12 features only compared to 23 initial features for predicting heating load of building. They also found that regression gradient guided based on k-NN selects the best feature than correlation coefficient methods.

Similarly, Kuisak et al. [55] used correlation index and boosting decision tree algorithm (see Appendix B.3 for decision tree and B.5.2 for boosting) to determine significant input features for cooling load. They identified that different features are significant for both methods and there was a slight increment in performance $\approx 1\%$ while using significant

features compared to all features in both methods. Jovanovic et al. [56] performed forward selection method based on linear regression to select the combination of best input features for daily heating load of buildings. Initially, they ranked the input features according to higher correlation indexes and evaluated the accuracy from first highest correlation indexes. Similarly, they add one by one other remaining highest correlation inputs and select the best combination. Out of several input combination, they identified that mean daily wind speed and minimum daily temperature are insignificant features for heating load. However, they found that model after forward selection had more error $\approx 0.2\%$ compared to all the input features. Autocorrelation to identify building load dynamics is also used. Zhang et al. [57] determined important input features for heating load using autocorrelation. They investigated heating load of previous hour of the same day and same hour of the previous day and they found that previous 2 hours and last 3 days have highest correlation indexes thus considered as an input features of the model. Unfortunately, they did not compare their results with and without feature selection.

2.3 Applications to Building Energy Modeling by Black Box Model

The building energy modeling based on black box model can be built considering all available data and few data.

Definition 2.7: All data– Relevant data

The approach is defined all data if all the available data (measurement or empirical behavior of building data) are used for model training to determine the parameters of model. For such model, the parameters are fixed for the considered building independently of the prediction day and future environment conditions. The approach is defined relevant data if the pre-selection of data is done initially for model training based on prediction day and future environment conditions. For such model, the data used for model training are reduced based on the relevance and parameters of model are changed for the considered building by each prediction day conditions. This type of approach is also named by few representative data since it selects small data to build a model.

Definition 2.8: Featuring database

A daily database is a collection of days for the input features. Those days have a sampling time of one hour. Consequently, the features are averaged on the sampling time.

The concept of "**all data**" and "**relevant data**" modeling approach for the featuring database (see **definition 2.8**) is shown in Figure (2.8-2.9).

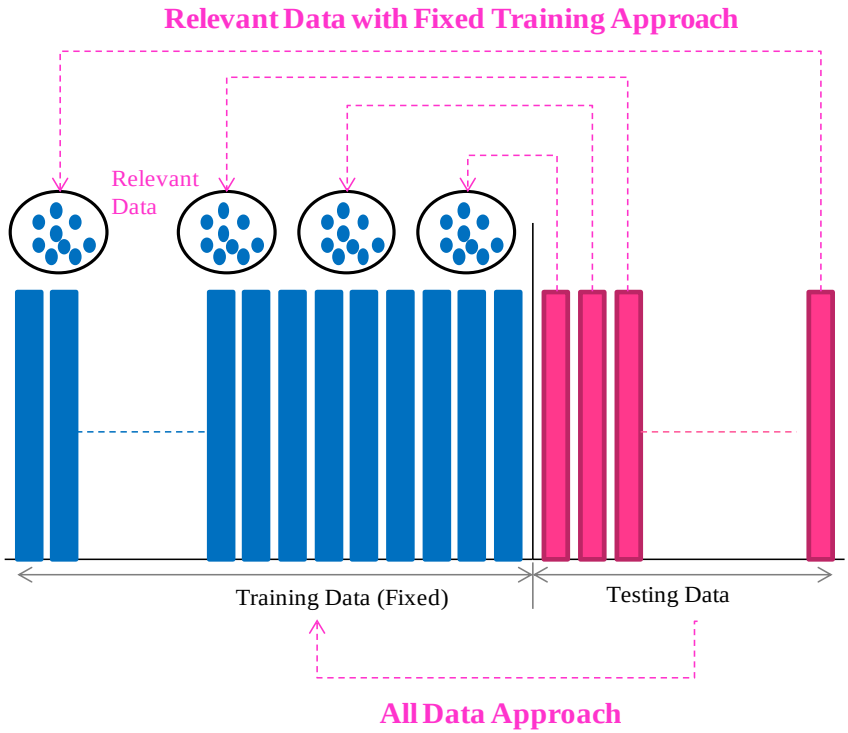

Figure 2.8: Concept of all data and relevant data with fixed training approach to build a model

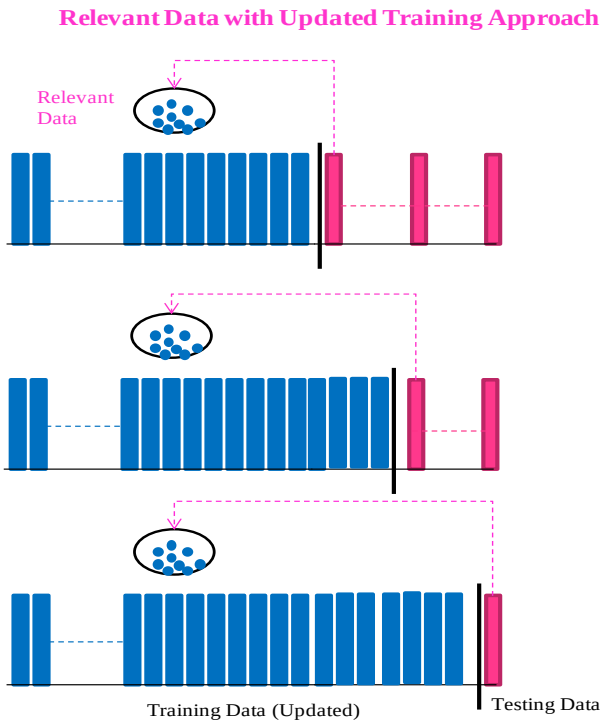

Figure 2.9: Concept of relevant training with updated training approach to build a model

It can be seen that "**all data**" modeling approach uses entire data (e.g., 365 days featuring database) to build a model whatever the prediction day conditions (e.g., climatic conditions, occupancy etc.). The "**relevant data**" modeling approaches uses few representative day data (e.g., 10 days featuring database) from the two kinds of training data: fixed and updated. In the "**relevant data with fixed training**" approach, few representative day data are selected from fixed all available data for model training for each prediction day conditions. On the other hand, in the "**relevant data with updated training**" approach, training data are updated after each prediction day conditions so that few representative training selected can be updated from all given training data if it is relevant to build a model for each prediction day conditions.

2.3.1 All data modeling approach

In this section, only applications related to building load prediction using "**all data**" modeling approach are reviewed.

Regression Methods

Cho et al. [58] used a simple regression model to predict heating load for building located in Daejon, S. Korea. They used single external temperature input features and found that the model was highly sensitive with the length of measurement data. They concluded that the regression model requires more training data to increase the performance of the model.

A second order polynomial regression was used by Catalina et al. [59] to estimate monthly and annual heating energy demand of a residential building. Different scenarios were evaluated by varying the shape factor, U-values, the time constant and the ratio of window to floor areas. The prediction results show a high accuracy with an error less than 3.2%.

Autoregressive Method (ARX): Dynamic Model

Yun et al. [60] used ARX (autoregressive with exogeneous, i.e., external, inputs) time and temperature indexed model with occupancy profile to predict hourly thermal load of building. They used three periods to represent the building energy dynamics: day period (8 AM to 9 PM), transition period (6 AM to 8 AM) and night period (9 PM to 6 AM). They identified that the external temperature is the dominant variable for thermal load. They also found that past hour energy load and occupancy are significant variable during the transition period.

However, a statistical **black-box** model [58] does not precisely represent hourly or in fraction of minute of building load. In order to find the best optimum fitness of data, these statistical models

require more effort and time. Polynomial approximation [59] is rationalized than statistical models due to its non-linear mapping between input and output function, nonetheless it is computationally heavy in terms of curse of dimensionality[10]. The dynamic autoregressive models [60] are not suitable if the prediction range is long horizon since the prediction of building load values depends on predicted values (e.g., building load of previous hours) and errors might be accumulated.

Neural Network

Ben-Nakhi and Mahmoud [61] investigated different buildings based on occupancy density (low-high) and building geometry to predict cooling load using external temperature of previous day using general regression neural network. They showed that neural network is able to predict heating load with good fit (R^2=0.986) while considering single external temperature variable for different building configurations.

Furthermore, thermo-physical parameters of building were investigated using neural network [62] [63] [64] to determine if neural network could be beneficial for different types of building. For instance, Kalogirou et al. [62] estimated minimum and maximum daily heating and cooling load for 9 buildings and their results showed that neural network can be used for thermal load prediction of buildings with different construction. Similarly, Yan and Yao [63] used different climatic zones using thermal envelope building parameter (see **definition 2.1**) including heat transfer coefficient and two other input features heating degree day (HDD) and cooling degree day (CDD) to predict heating and cooling energy consumption. Their results showed an average deviation of 1.7% and 2.9% while compared with actual values of heating and cooling energy consumption. They concluded that neural network can be used for adapting from one known building to another unknown that have different climates and heat transfer coefficients. The effect of insulation thickness and composition of insulation materials, i.e., insulation thermal conductivity (K-value) were investigated by Naji et al. [64] to predict total heating energy demand using extreme machine learning. They identified that energy demand were significantly affected by the properties of insulation materials rather than thickness of wall materials of building. They noticed that with the increasing of thickness of insulating materials affects the other materials in less amount leading to slight decrease in energy demand.

[10] Define complexity of model i.e., with the linear increase of input variables, the complexity of the model goes on increasing

Recurrent neural network, which uses internally generated predicted output to make further output, was used by Kalogirou and Bojic [65] to predict energy consumption for passive solar building in Cyprus. Their results showed higher accuracy (R^2=0.999) for unknown conditions. In addition, dynamics of occupant behavior was included by Kowk and Lee [66] to predict cooling load for office building in Hong Kong, China. Apart from climatic conditions, they used percentage of total building occupancy area to distinguish for working/non-working period and electrical power consumption of the primary air handing unit (PAU) of the ventilation system for indicating occupancy dynamics. They found that with the inclusion of dynamics of occupancy and occupancy area leads to increase in performance (R^2=0.43 to R^2=0.95) compared to only climatic conditions input features and justified that the influence of occupancy is significant for thermal load.

The comparison between different models: neural network with other statistical and physical models were also performed. Tso and Yau [67] made comparison of different **black-box** models: stepwise linear regression, neural network and decision tree for electricity energy consumption in Hong Kong, China. They found that during summer seasons, decision tree performs slightly better than other two methods whereas in winter seasons, neural network performed better than other two models. Comparison between physics based method (finite difference method) with neural network was performed by Ekici and Aksoy [68] to predict heating energy consumption in buildings. They used physical and geometrical input features and their results showed that neural network perform average 94-98% accuracy in comparison to physical method. Similarly, comparison between simulation tools Energy Plus and neural network was proposed by Neto and Fiorelli [69] to estimate energy consumption of buildings in Sao Paulo, Brazil. Their result showed that neural network was slightly more accurate than Energy Plus when comparing with real data.

Support Vector Machine

Zhang et al. [57] used SVM to predict heating load in Daqing city, China. In their work, 120 days from 2007 to 2008 of heating load data were used for training data and last one day was used to evaluate test condition. They performed autocorrelation of heating load of previous hour of the same day and same hour in the previous days and found that previous 2 hours and last 3 days have non-linear thermal dynamics. Later, trained SVM model errors were corrected using Markov chains probability. Their results further illustrated that such models were suitable for pure dynamic model.

Comparison of SVM with ANN was performed by Li et al. [70] to predict cooling load of office building in Guangzhou, China. They found that both SVM and ANN performed higher accuracy. However, the results also revealed that SVM has better accuracy of 0.02% than the ANN.

Decision Tree

Yu et al. [71] used decision tree to classify and estimate Japanese residential building energy use intensity levels into either high or low values. Their results demonstrated that decision tree method correctly classify and predict energy demand with 93% and 92% accuracy on training and test data.

Random Forest

Tsanas and Xifara [72] used random forest to predict heating and cooling load of residential building. Their results showed that RF has higher accuracy with mean absolute errors deviations of 0.51 and 1.42 for heating and cooling load respectively. They also compared their results with linear regression model and identified that RF have higher accuracy due to their capacity of relevance of input variable determination (association strength of variable and their response) and redundancy (association strength between variables, i.e., multi-collinearity effects) unlike regression model.

Hybrid/Ensemble Method

Hybrid methods use fusion of several models whereas ensemble method used outputs of several models to make a final prediction. Ensemble prediction methods combine output of different models by simply averaging, weighted based averaging and median based averaging. The advantages of ensemble model are that it compensates the errors by combining their outputs thus performed better results than individual one ([56], [73]).

Kusiak et al. [55] made a comparison of ensemble neural network model with 9 other machine learning techniques (Decision Tree: CART, CHAID, exhaustive CHAID, and boosting tree; multivariate adaptive regression splines (MARS), RF, SVM, neural network and k-NN) for steam load prediction in Iowa City, USA. They found that ensemble neural network had better performance than other machine learning models. Fan et al. [74] used ensemble of machine learning models to predict peak and total energy consumption of buildings in Hong Kong, China. They compared eight models: statistical methods (MLR and ARIMA), and machine learning based on SVM, RF, multi-layer perceptron neural network, boosting tree (BT), MARS, and k-

Nearest neighbors (k-NN). Their results revealed that ensemble model have MAPE error of 2.3% and 2.9% for daily peak and total energy consumption respectively. Out of eight models, RF and SVM have best performance, so largest weights of them were integrated in the ensemble model. The traditional statistical models, i.e., MLR and ARIMA had poor performance and they had small weights in the ensemble model. Jovanovic et al. [56] proposed ensemble of various neural networks (feed-forward neural networks, radial basis function neural network and adaptive neuro-fuzzy interface system) for the prediction of daily heating energy consumption in Norweign University of Science and Technology campus buildings located in Norway. They found that ensemble method performed better than the individual model.

Xuemei et al. [75] proposed hybrid ARMA and multi-layer perceptron neural network to predict hourly cooling load. The residual errors obtained from neural network model were further used to predict from ARMA model for correcting the cooling load. Li et al. [76] proposed hybrid genetic algorithm-adaptive network-based fuzzy interface system (GA-ANFIS) to predict energy consumption of buildings and compared with neural network. Their results showed that the performance of hybrid GA-ANFIS model was better than neural network. Wang and Meng [77] proposed ARMA-neural network to predict hourly energy consumption of Hebei, China. The residuals of ARMA were further input to neural network. Their results revealed that hybrid model has good accuracy (MAPE=0.3%) compared to individual model (neural network: MAPE=4.0%, ARMA: MAPE=3.5%).

In the application of building with different construction using geometrical input features, Chou and Bui [73] applied ensemble model and compared with different data-driven models: SVM, neural network, classification and regression trees, chi-squared automatic interaction detector, general linear regression to predict heating and cooling load . Their results showed that SVM and ensemble model (combination of averaging the results from neural network and SVM) had better results for heating and cooling load respectively.

The input variables and time step of prediction that are used in the literatures using "**all data**" modeling approaches are summarized in Table (2.5). It can be concluded that machine learning based **AI model** using "**all data**" approach had been widely applied using limited physical features of building during the operation phase. The literatures ([62], [63] and [64]) also applied for building with different construction using thermo-physical and geometrical inputs features and have higher performance compared to statistical models ([73] and, [74]). The main advantage of such **AI model** using "**all data**" approach is once the model has been built, energy operator or building operator does not require knowledge on physical systems.

S.N.	Authors	Type of Model	Input Features Used for Modeling									Time Step
			External temperature	Solar radiation	Relative Humidity	Wind speed	Occupancy Profile	Function representing H,D,M,A,S	Other Parameters	Indoor /Set-point temperature	Previous Hours/Day Energy Load	
1	Cho et al. [58]	Regression	×					×				A
2	Catalina et al. [59]	Polynominal regression	×	×					1*:	×		M, A
3	Yun et al. [60]	ARX	×	×	×	×	×	×			×	H
4	Ben-Nakhi & Mahmoud [61]	Neural network	×									H, D
5	Kalogirou et al. [62]	Neural network	×	×		×		×	2*:			D
6	Yan and Yao [63]	Neural network							3*:			A
7	Naji et al. [64]	Neural network							4*:			A
8	Kalogirou and Bojic [65]	Neural network						×	5*:			H
9	Kwok and Lee [66]	Neural network	×	×	×	×			6*:			H
10	Tso and Yau [67]	Neural network							7*:			A
11	Ekici and Aksoy [68]	Neural network							8*:			A
12	Neto and Fiorelli [69]	Neural network	×	×	×			×				D
13	Zhang et al. [57]	Support vector machine									×	H
14	Li et al. [70]	Support vector machine	×	×	×				9*:			H
15	Yu et al. [71]	Decision tree	×						10*:		×	A
16	Tsanas and Xifara [72]	Random Forest							11*:			-
17	Kusiak et al. [55]	Ensemble	×	×								D
18	Fan et al. [74]	Ensemble	×		×			×	12*:		×	D
19	Jovanovic et al. [56]	Ensemble	×	×	×	×		×			×	D
20	Chou and Bui [73]	Ensemble							11*:			A
21	Xuemei et al. [75]	Hybrid	×	×	×							H
22	Li et al. [76]	Hybrid	×	×			×	×			×	H
23	Wang and Meng [77]	Hybrid									×	H

Table 2.5: Summary of input features and time step of prediction using all data modeling approach

1*: building envelope U-value, window to floor area ratio, building time constant
2*: structural characteristics of heat transfer by indicating wall and roof type
3*: heat transfer coefficients in building envelopes; window to wall ratio; building shade coefficient, orientation, solar absorption, air change rate, shading coefficient of window in each orientation, HDD and CDD
4*: insulation thickness, Insulation k-value
5*: insulation, thickness, heat transfer coefficient
6*: rainfall, bright sunshine duration, occupancy area and air unit power consumption
7*: housing type, household characteristics and appliances ownership
8*: building transparency ratio, insulation thickness and orientation
9*: last 2 hours dealy of outside air temperature and last 1 hour delay of solar radiation
10*: house type, construction type, floor area, heat loss coefficient, space heating, hot water supply
11*: relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution
12*: pressure, cloud, rainfall, evaporation, number of hours of reduced visibility
H: Hourly, D: Daily, M: Monthly, A: Annually, S: Seasonally

2.3.2 Relevant data modeling approach

The small representative data selected from the sets of all data is sufficient to build a predictive model. There are three major reasons to consider "**relevant data**" compared to "**all data**" modeling approach.

- Firstly, all data used for model training contain similarities and dissimilarities of input patterns behaviors and some of the information might be redundant.
- Secondly, a predictive model takes a lot of time for model training when all the data are used.
- Finally, with the adaptability of growing this model in the future, the newest environment and climatic conditions have probable more useful information, which is not considered in "**all data**" modeling approach due to its computational complexities. The effect of this new information is neglected to update the model parameter. In order to update the model parameters in "**all data**" modeling approach, the initial learning algorithm should be modified to complex learning algorithm.

Definition 2.9: Online Learning- Offline Learning

Learning mechanism in all data modeling approach is called offline learning since model parameters are not updated with new datasets. The approach relevant data uses both offline learning and online learning: The offline learning selects few representative data from fixed all available data whereas online learning selects few representative data from updated all available data so that it updates the model parameter with new dataset and adapt to changing environment.

The comparison of "**relevant data**" with fixed/ updated training with "**all data**" modeling approach is shown in Table (2.6). It can be seen that "**relevant data**" modeling approach has many advantages in terms of probability to have redundant information, computing CPU-time and updating of model parameters. It is also seen that "**relevant data with fixed training**" uses offline input featuring database leads to learning mechanism offline (see **definition 2.9**). Whereas, "**relevant data with updated training**" uses online input featuring database result in learning mechanism online.

Characteristics	All data Approach	Relevant data with Fixed Training Approach	Relevant data with Updated Training Approach
Input featuring databse	Whole database Fixed Offline	Selected sub-database Fixed Offline	Selected sub-database Updated Online
Proabability to have redudant information	High	Low	Low
Computing CPU-time	High	Low	Low
Model parameters	Fixed	Updated	Updated

Table 2.6: Comparison of relevant data with fixed/updated training with all data modeling approach

Remark 2.6:

> *The number of training data significantly influences the accuracy of prediction model. For instance, Withdrow and Kamenetsky* [78] *recommend that the ratio of training data should be at least ten times greater than the input features.*

The review works on "**relevant data**" modeling approach to select representative days data for model training applied to energy consumption prediction regardless of type of model used are discussed below:

Similar Climatic Conditions

Several studies have been carried out to select relevant day data based on similar trends of climatic conditions for model training applied to electrical energy consumption: ([79], [80], [81] , [82] and [83]). For example, Chen et al. [79] used weekday and climatic index of wind chill temperature and humidity; other literatures ([80]and [81]) used day indicator, maximum and minimum external temperature; and Jain et al.([82]) used day-type, maximum external temperature and humidity of prediction day to select relevant day data for model training. All these methods determined the similarity of selected individual variable based on the Euclidean

distance between prediction day with training day and they are further multiplied by weight factors of individual selected variable. The weight factors of selected variables were determined using least **square method (LSM) based on regression model**. Mu et al. [83] used day type, weather type, week type, maximum and minimum temperature change and date difference of prediction day to select relevant day for model training. In their work, the weights of selected variables were estimated from correlation coefficients of training and prediction day data. Their results showed that relevant day data based on similar climatic conditions for model training improved the performance, e.g., root mean square error (RMSE) of 0.84 with "**relevant data**" modeling approach compared to RMSE of 4.5 with "**all data**" modeling approach.

Similar Energy Load and Climatic Conditions

Several studies have been carried out to select relevant day based on electrical energy load and climatic conditions data of prediction day for model training. For example, Mandal et al. [84] determined the similarity between prediction day and relevant day based on electrical load, load deviation and deviation of external temperature of previous day from prediction day. They determined the weight factors of selected variables using **LSM based on regression model**. Their results revealed greater accuracy with 2.5% mean absolute error. He et al. [85] used similar trend and day similarity degree to select relevant day for model training. In their work, similar degree-day was calculated from cosine similarity angle between electrical load with the day to be predicted and training day data, and trend similarity with daily average energy load. Their prediction results showed that "**relevant data**" approach is better (MSE[11] ≈ 2.5%) than "**all data**" modeling approach (MSE ≈ 4%).

Heating Degree-Day and Cooling Degree-Day

HDD and CDD was used by Roldàn-Blay et al. [86] to select relevant day for model training and later used trained model for predicting electrical load. In their work, estimated HDD and CDD of prediction day were used to select similar HDD and CDD of training day as a relevant day. Their results showed mean absolute percentage error (MAPE) of 2% while using "**relevant data**" modeling approach. Unfortunately, they did not compare their methods with "**all data**" modeling approach.

[11] Mean Square Error

Clustering

Several studies have been conducted ([87], [88], [89], [90], [91], [92], [93], [94] and [95]) based on clustering /classification methods to select relevant day from particular cluster/classes for model training. The different clustering methods were used by many authors (Jain and Satish [87]: SVM classifier; Ghanbari et al. [88]: k-means clustering algorithm; Pasila [89]: Fuzzy c-means clustering; Yadav and Srinivasan [90]: Kohonen self-organizing map (SOM) clustering) to cluster daily average electrical load of training day and further selected particular cluster based on estimated daily average load of prediction day to define relevant day. As an example, Ghanbari et al. [88] achieved slight increment in prediction performance while applying "relevant data" modeling approach based on clustering compared to "all data" modeling approach, i.e., without clustering (All data: MAE- 1.4%; Relevant data: MAE- 0.6%).

Several authors used different methods (Sun [91]: deterministic annealing clustering algorithm; Duan [92]: ant colony clustering method; Marin et al. [93]: self-organizing map classifier) to cluster electrical load data into different groups. Later on, the load from the previous day's prediction was used to select particular clusters as relevant data for model training. As an example, Marin et al. [93] obtained 15 clusters and their prediction results showed absolute percentage error below 2.3% for all clusters. Grenda and Macukow [94] used SOM to identify different classes of district heating load as relevant day data based on daily average heating energy demand with the assumption that similar daily average customers will have similar thermal properties of buildings. Their results showed acceptable standard deviation error rates of 0.0019.

Pattern Recognition

The SOM clustering based on external temperature and electrical load of the previous day was used by Tafreshi et al. [95] to find similar patterns as relevant day for given predicted external temperature and estimated daily average energy load. The average error rates were ≈ 1.5% while testing with 1 year data.

Reference Day

Reference day based on similarities of occupancy profile was used by Sun et al. [96] to select representative day for model training. Their selection of reference day varied during a week. As an example, the working days (e.g., Wednesday, Thursday and Friday) have similar occupancies with the previous days thus previous days were selected as a reference day. In case of weekend

and Monday, last weekend and Monday was selected as a reference day because of similar occupancy profile. Their predicted R^2 value was 0.89 and observed that their predicted results had significant deviations with the actual measurement values due to their large deviations of weather difference while selecting a reference day.

Sliding Window and Accumulated Training

The selection of representative day based on fixed amount of data also called sliding window and retraining "**all data**" modeling approach with each new update measurement data called accumulated training or incremental learning (see **definition 2.6**) were purposed. For instance, Gonzalez and Zamarreno [97] predicted electricity consumption of building using sliding window of 21 days data for model training using neural network. Their results revealed good fit for working day period whereas their prediction results were below the actual measurement values for weekend. It might be due to 21 day window did not cover the peak energy consumption of data for particular prediction day conditions. Similarly, Yang et al. [98] used fixed sliding window and accumulating training. Compared to accumulative training, their results based on sliding window had better performance for real measurements.

The summary of input features used to select relevant day data and the model used for training are detailed in Table (2.7). It can be concluded that most of the "**relevant data**" modeling approaches used to select small representative days data were based on daily average energy load of the previous day for a given predicted day, daily average energy load of predicted day and initial energy load of predicted day ([84], [85], [87], [88], [89], [90], [99], [91], [92], [93], [94] and [95]). In addition, if the learning mechanism of prediction model of energy demand of building is not only for a day ahead, but also for a longer period in advance, then prediction methods will rely on previously predicted daily average energy load values and errors will be accumulated thus it is not pragmatic during operation phase of building. Furthermore, many review works of "**relevant data**" modeling approaches were based on electricity load and methodology applied to electricity load would not have similar behavior to thermal energy consumption because of thermal inertia and set-point temperature behavior in building. It is also observed the possibility to consider smaller representative data using recent training data or using fixed sliding window ([97] and [98]), but the prediction conditions might not reflect the seasonal variations of energy demand using few or large window sizes. Also, adaptive model is seen based on retraining the model with new update of training data [98]. If the new training data increases then the adjustment of model parameters is difficult due to this recent change. In addition, sometimes this recent training data changes might be more informative but their effect have less impact in updating model training.

SN	Selection Method Type	Authors	Model Type	Significant Input Features							Weight Determination Method	Problems
				External Temperature	Wind Velocity	Humidity	Energy Load	Occupancy	Day Type Categorization	Other Parameters		
1	Similar Climatic Conditions	Chen et al. [79]	Wavelet and neural network	×	×	×			×		LSM	Electricity
		Sun et al. [80]	FSVR and linear extrapolation	×					×		LSM	Electricity
		Senjyu et al. [81]	Neural network	×					×		LSM	Electricity
		Jain et al.[82]	FLC and ant colony	×		×			×		LSM	Electricity
		Mu et al. [83]	Neural network	×						1*	Correlation Coefficients	Electricity
2	Similar Energy Load and Climatic Conditions	Mandal et al. [84]	Neural network	×			×		×	2*	LSM	Electricity
		He et al. [85]	Neural network				×				Correlation Coefficients	Electricity
3	HDD & CDD	Roldàn-Blay et al. [86]	Neural network	×							-	Electricity
4	Clustering	Jain and Satish [87]	Support vector machine				×				-	Electricity
		Ghanbari et al. [88]	Genetic fuzzy and ANF				×				-	Electricity
		Pasila [89]	Neuro-fuzzy				×				-	Electricity
		Yadav and Srinivasan [90]	Auto regression (AR)				×				-	Electricity
		Sun [91]; Duan [92]; Marin et al. [93];	Neural network				×				-	Electricity
		Grzenda and Macukow [94]	SOM and evolutionary NN				×					Heating
5	Pattern Recognition	Tafreshi et al. [95]	Neural network	×			×				Patterns search by SOM	Electricity
6	Reference Day	Sun et al. [96]	3*					×			-	Cooling
7	Sliding window and Accumulative Training	Gonzalez and Zamarreno [97]; Yang et al. [98]	Neural network							4*	-	Electricity

Table 2.7: Summary of input features used to select relevant data

1*: Weather type and date difference
2*: Energy load deviations
3*: Calibration based on reference day
4*: Fixed window size and model focus to update parameters with new measurement data
FLC: fuzzy logic interface
FSVR: fuzzy support vector regression
ANF: adaptive neuro fuzzy
SOM: self organizing map
LSM: Least square method

Remark 2.7:

The individual selected features of building has different weight on building load and influence of building load is estimated by least square method (LSM) method based on regression model in literatures ([84], [91], [79], [81] and [82]*).*

2.4 Conclusion

This chapter provides general concept on LEB and draw a benchmark to compare with CB. It also provides evolutions of LEB trends in Europe. Then it reviews on input features for building energy consumption prediction and found that energy consumption of building depends on several factors: climatic conditions, geometrical parameters, thermo-physical parameters and building operating conditions. The short-description of input features used in the literatures is summarized in Table (2.8).

It then reviews three widely used prediction model namely **white-box**, **gray-box** and **black-box** model to estimate and predict thermal energy consumption of building. The **white-box** models are based on fundamental principle of building physics. **Black-box** model are solely based on measurement or empirical data. **Gray-box** model are just combination of **white-box** and **black-box**. The proper choice of these three models depends on the purpose, prior knowledge and available data.

The summary and comparison of these models based on input data, modular experience, calibration effort and training data requirement, etc. are shown in Table (2.9). The four kinds of artificial intelligence **black-box** models: neural network, support vector machine, decision tree and random forest are only considered for further discussion since they are more suitable for non-linear problems. It can be noticed that when the model goes from **white-box** to **black-box**, the input features goes on decreasing, calibration goes on decreasing and training data sets goes on increasing. It is clear that **white-box** model (detailed energy simulation) should be used when there is a requirement of extensive information of building characteristics. These kinds of models are suitable for an early stage in new buildings for estimating thermal energy consumption. **Gray-box** model, moreover, requires detail understanding of building thermal dynamics but overcomes the limitation of **white-box** model due to structural complexity.

Input Features		Descriptions (with respect to heating energy consumption of building)
Climatic Conditions	dry bulb temperature	determines thermal response of building and amount of heat gain/loss through building envelope; increases heating energy consumption when it is lower
	wet bulb temperature	determines humidification
	solar radiation	means free heat gains which lowers the energy use for heating and increases the heat gains due to increase of it
	sol-air temperature	the equivalent of outside air temperature that provides similar heat transfer due to outside air, solar radiation and radiative heat exchange with sky and surroundings
	humidity	affects the latent load
	clearness index	sky or cloud conditions and blocks the solar radiation, thus affects the shading through windows/glazing
	wind speed and directions	affects natural ventilation and outside surface building envelope; increases the heating energy consumption thus impacts the hygrothermal response of building envelope
Geometrical Parameters	building location and orientation	latitutde, longitude and affects the solar gain
	window to floor area ratio	affects the lighting pass through the building
	shape factor /relative compactness	shape of building type and affects the energy consumption and standards due to heat loss through the surface thus decreases heating energy consumption if it is higher
	transparency ratio	percent of wall covered by the window and determine solar gain effect in building
	area (surface, wall, roof, glazing)	affects total energy consumption
Thermo-physical Parameters	heat transfer coefficient of walls, roof and glazing	determines the energy consumption and indoor environment; it decreases the requirement of heating energy consumption if thermal heat transfer coefficient of envelopes is lower
	solar gain on wall	affects the wall capacitance through insulation
	solar gain transmitted through window	affects the thermal mass and indoor air temperature of building
	time constant	ratio of thermal capacity of the building to the overall heat loss coefficient; and affected by the building envelope and thermal mass of the building
	thermal inertia	quantify in terms of thermal mass, i.e., heat capacity and density, and higher the heat capacity higher will be thermal inertia and balances the indoor environment
	base temperature	the temperature which determines whether heating or cooling requires
	shading coefficients	determines the solar heat gain
	materials of walls, roofs, floors and windows	signify the time constant and thermal inertia of building
Building Operating Conditions	internal gains	gains from occupants, lighting and appliances; it decreases heating energy consumption if it is higher
	indoor temperature, humdity	indoor temperature and moisture variation for thermal comfort
	ventilation/infiltration rate (mass flow rate)	air flow rate from outside to the building and signify heat loss from the building thus it increases heating energy consumption if its losses from ventilation is higher
	return and supply temperature	temperature of water supply/return through pipe to/from the building
	function representating H, D, M & S	signify the energy use time
	AHU power consumption	electrical power consumption required for Air Handling Unit

Table 2.8 Description of input features used in literatures for building energy consumption prediction

H: Hour, D: Day, M: Month and S: Season

Features Specification	White-box	Gray-box	Black-Box				
			Linear Regression	Machine Learning based AI Techniques			
				Neural Network	Support Vector Machine	Decision Tree	Random Forest
Input data	●●●●	●●●	●	●	●	●	●
Modeller experience	●●●● (1*,3*)	●●● (1*,2*)	●	●● (2*)	●● (2*)	●● (2*)	●● (2*)
Simplicity of calibration (in terms of parameters)	●	●●●	●●●●	●●●	●●●●	●●●●	●●●●
Training data	-	●●●	●●●●	●●●●	●●●●	●●●●	●●●●
Model Training Time	●●●	●●●	●●●●	●●●●	●●●	●●	●●
Requirement of building physical information	●●●●	●●●	●	●	●	●	●
Physical intrepretation of parameters	●●●●	●●●	●	●	●	●	●
Model complexity	●●●●	●●●	●	●●●	●●●	●●	●●
Accuracy	●●●●	●●●●	●●	●●●●	●●●●	●●●●	●●●●
Adaptability of model parameters	●	●	●	●●●●	●●●●	●●●●	●●●●
Application during operation phase	●●	●●●	●●	●●●●	●●●●	●●●●	●●●●
Multicollinearity effects from input data	●	●	●●●●	●●●	●●	●●	●●
Uncertainity	●●●●	●●●	●●●	●	●	●	●

1*: Familar with building thermal dynamics
2*: Familar with statisical concepts
3*: Familar with building thermal simulation tools

Notations:
●●●●: Very high
●: Very low

Table 2.9: Comparison of white-box, gray-box and black-box prediction models

Thus, it can be concluded that both **white-box** and **grey-box** model are highly parameterized due to their interactions between systems on various mode of heat transfer requiring more input information. All the physical thermal properties of building are not always known and cost effective, and hence impracticable for Energy Services Company (ESCOs) and/or building energy management system (BEMS) for planning and control use during the operation phase. On the other hand, **black-box** models can be used when few input features are available and is extensively used for adaptive model in the future. Linear statistical regressions based **black-box** models are easier method and do not require expertise knowledge. Machine learning based **AI model** have greater accuracy but suffer from physical interpretation. Neural network requires large number of training data. In contrast, SVM has a huge advantage in representative training data. Random forest and decision tree requires small CPU-time for model training.

However, machine learning based artificial intelligence techniques have several advantages compare to **white-box** and **gray-box** model during the operation phase.

- Firstly, they require fewer parameters of building which might be practicable during operation phase.
- Secondly, they are good in learning the response of building energy system.
- Finally, they have a strong capability of being adaptive to update the model parameter to take into account dynamic environment of future conditions.

Nevertheless, these artificial intelligence models based on "**all data**" modeling approaches has several drawbacks due to redundancy of input information, complexities in model training and adaptability to updating the model parameters in future environment. The review works that are based on small representative data selection known as "**relevant data**" modeling approaches are suitable for adaptive model to update the parameters of model but still they have some limitations. First, the methods that focus on selection of few representative data do not consider past day climatic conditions due to large time constant of building (for example, more than 100 hours in LEBs [16] which is an essential factor for LEBs. Second, these methods do not consider the solar gain impact. Finally, the methods that are based on daily average energy load of prediction day or previous day to select representative data are not suitable for LEBs. If the learning mechanism of prediction model of energy demand of building is not only for a day ahead, but also for a longer period in advance, then the prediction methods will rely on previously predicted daily average energy load values and errors will be accumulated thus it is not pragmatic for real application.

The next chapter will bridge the aforementioned research gap and discuss the methodological framework to predict building energy consumption from hours to couple of days (or even longer periods depending upon the forecast range of climatic conditions). It describes the "**relevant data**" modeling approach that uses few representative data for model training using machine learning model: artificial neural network, support vector machine, boosted ensemble decision tree and random forest. It also explain methodological framework of "**all data**" modeling approach using machine learning model.

Chapter 3: Artificial Intelligence for LEB Modelling

3.1 Modeling Approaches

3.1.1 Introduction

The estimation of the heating load of a LEB is more challenging due to its large time constant, for detail see **section 1.2**, **chapter 1**. As an example, the estimation of heating load is quite different for two similar climatic conditions days d_1 and d_2 (under the same occupancy profile and the same building operating conditions schedule), see Figure (3.1). As shown in Figure (3.1), the days d_1 and d_2 have a similar climatic conditions (External temperature T_{ext}, horizontal solar radiation $\emptyset_{Sh}$, solar gain on walls $\emptyset_{Sint}$ and solar gain transmitted through windows $\emptyset_{Sext}$) but their heat demand during morning (0:9 hour) are quite different generally. It can be observed that external temperature T_{ext} during 0:9 hour interval is quite similar in those two days and illustrates that this time interval does not fully explain the behavior of the heating load. Hence, this reveals that past day climatic dynamics are also significant.

However, the climatic variables do not have the same dynamic effect. For instance, the previous days of the solar radiation transmitted through windows ($\emptyset_{Sext}$) have no impact in the prediction day since it has a fast response to the indoor temperature changes. Since the energy inputs from external heat transfer (by conduction/convection with the external temperature T_{ext} and by solar radiation on walls $\emptyset_{Sint}$) are stored by walls heat capacity for a long period, these variables are useful to find the suitable number of past time delay dynamics. The horizontal solar radiation ($\emptyset_{Sh}$) is not useful to explain this behavior because its effect is already considered in both solar gain transmitted through windows $\emptyset_{Sext}$ and solar gain on walls $\emptyset_{Sint}$. This further concludes that the past day dynamics of external temperature T_{ext} and solar gain on walls $\emptyset_{Sint}$ might be more important to understand the heat demand behavior.

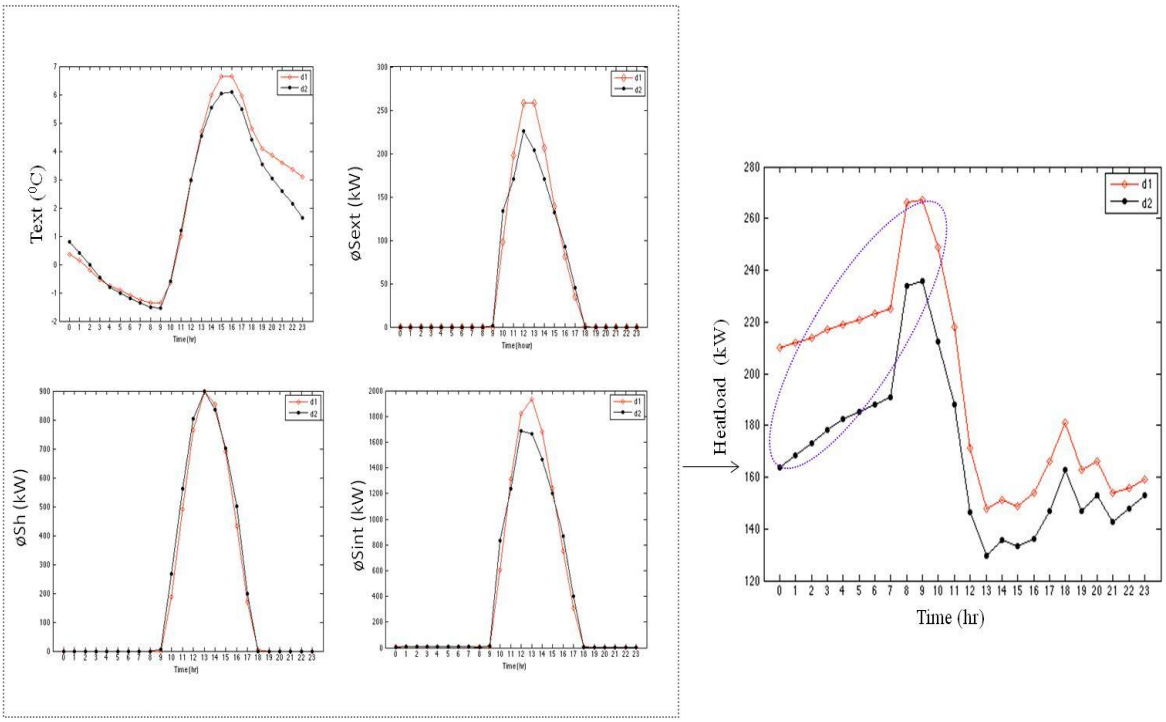

Figure 3.1: Illustration of thermal dynamic behavior in building

Coming back to the example, the past 5 days period is plotted for an analysis since the non-linear dynamic response of LEB is more than 100 hours. The time patterns of external temperature T_{ext} and solar gain on walls $\emptyset_{Sint}$ for the previous 5 days of the days d_1 and d_2 are shown in Figure (3.2). It can be noticed that T_{ext} and $\emptyset_{Sint}$ behaviors of the past days of d_1 and d_2 are quite different. For instance, the time patterns of T_{ext} and solar gain on walls $\emptyset_{Sint}$ in past d_1 day is quite lower in magnitude than the ones of d_2 days for last 3 days. In contrast, on the past 4-5 days period, the solar gain on walls $\emptyset_{Sint}$ for the day d_1 is quite larger in magnitude than the one for the day d_2. Moreover, the time patterns of T_{ext} for the d_1 is lower than the one for the day d_2, thus the effect of the conduction/convection heat transfer (reflected by T_{ext}) is compensated by solar radiation ($\emptyset_{Sint}$). Because of this difference of T_{ext} and $\emptyset_{Sint}$ in past days, the heating loads in d_1 and d_2 days are quite different. Figure (3.2) hence reveals that the past 3 days of T_{ext} and $\emptyset_{Sint}$ is more informative to represent non-linear behavior of building for the days d_1 and d_2 than the period of 4-5 days.

The research questions are:

1. *How to "introduce" such kinds of dynamic behavior in AI model to predict the energy consumption couple of days for a LEB?*
2. *What are the most significant features for different types of building?*
3. *How does the number of past days climatic variables influences the prediction accuracy of energy consumption of buildings?*

4. *How does number of data used for model training influences the performance of machine learning AI model?*

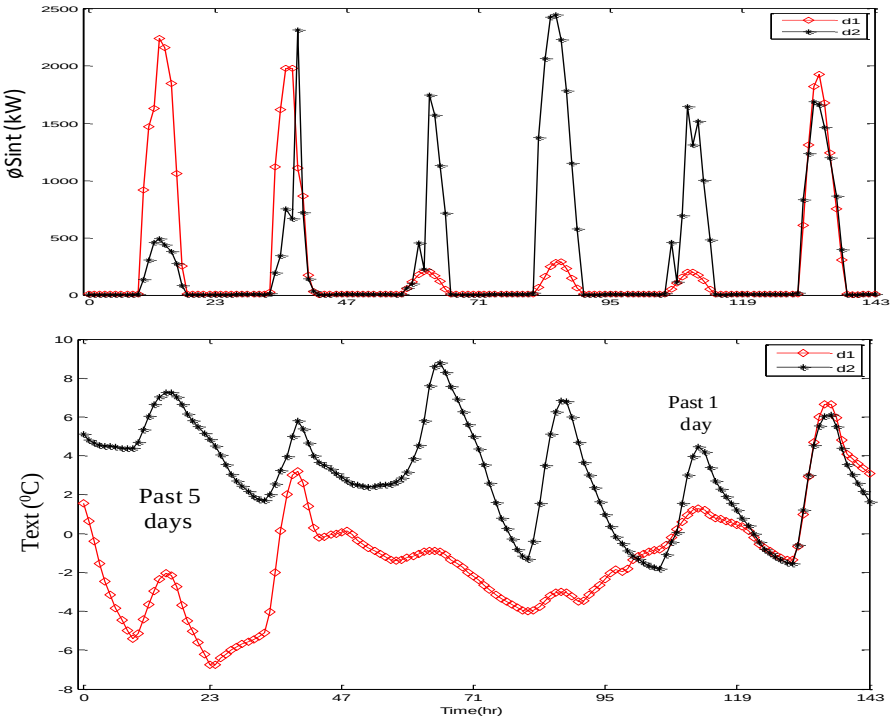

Figure 3.2: Past 5 day behavior of T_{ext} and $\emptyset_{Sint}$ from d_1 and d_2 days

3.1.2 Assumptions

1. The sampling time is fixed. All the data are then averaged on this sampling time. Any stochastic behaviors are not considered.
2. The following data are available:
    - Climatic conditions (external temperature, solar radiation etc.)
    - Derived climatic conditions (e.g., solar gain on walls and solar gain transmitted through windows) obtained using main characteristics (window area, orientation etc.) and location (latitude and longitude)
    - Occupancy profiles (represented by a fixed pattern)
    - Building operating conditions (e.g., set-point temperature, lighting, ventilation imposed by Air Handling Unit)
    - Thermal energy consumption (heating or cooling): either from measurement for an existing building or from numerical simulation for a project

3. Any additional energy production (e.g., solar and wind energy integration to the building) is also out of the scope of this research.

Remark 3.1:

The dynamics of occupants have a large impact in the energy consumption but an estimation of an occupant's behavior is complex due to stochastic nature of occupant's therefore we have used fixed and repeated schedule for all days.

For the *prediction day* (or the couple of days), the **forecasted** weather conditions, the occupancy and the building operation conditions are also assumed available.

Those features are the key inputs for the development of a black box model and so for an application of a machine learning based **AI model**.

3.1.3 Proposed Approaches

In this work, we consider two main approaches during the model training:

- All available data are used to build a model and this approach is named "**all data"** in the following (see **definition 2.7**, **Chapter 2**). The **AI model** consists to fixed parameters of a building and this later is independent to the prediction day conditions.

- A pre-selection in the database is first done. A set of days (e.g., 10 days) data is brought out with the most similar weather conditions than the forecasting ones of the day for model training. Consequently, an **AI model** is defined for each predicted day. This approach is named "**relevant data"** in the following (see **definition 2.7**, **Chapter 2**). Here the training database can be updated day by day by including the training data before the prediction day (or the couple of days).

Definition 3.1: Model parameter selection

The task of choosing the best parameters of AI model is called model parameter selection.

The whole framework of the model training methodologies "**all data**" and "**relevant data**" for an energy load prediction is shown in Figure (3.3).

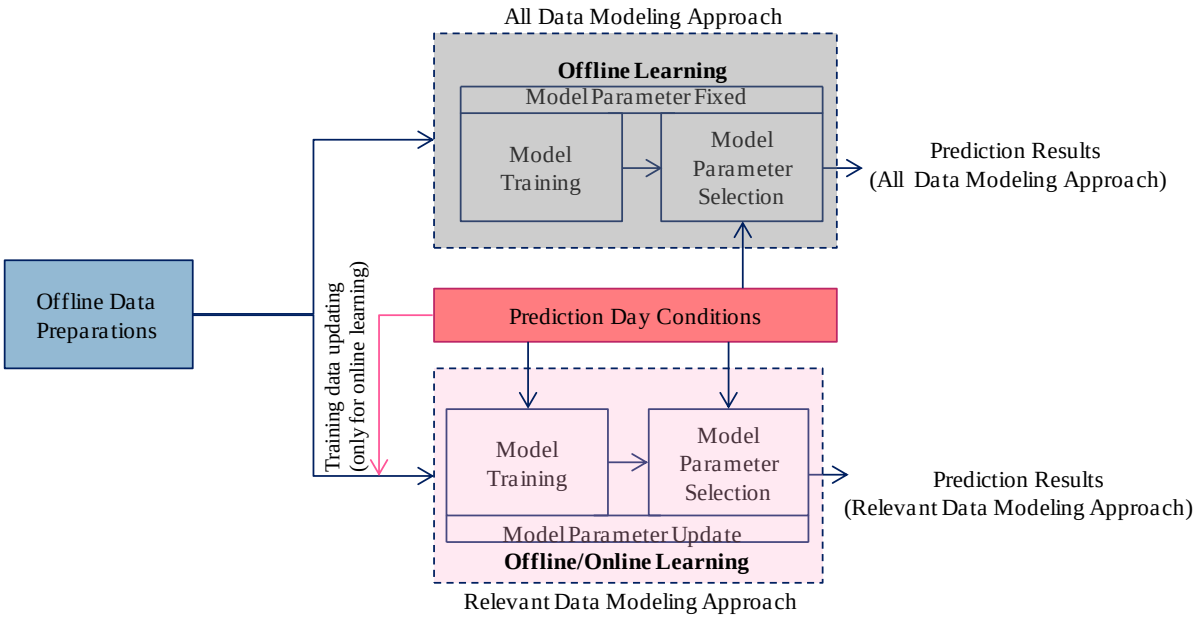

Figure 3.3: Whole framework of all data and relevant data modeling approach based on offline and online learning for energy load prediction

It can be observed that offline data are prepared for both types of modeling approaches: "**all data**" and "**relevant data**" to build a model. In the modeling approach "**all data**", offline data are used for model training by making the learning system offline resulting in a fixed model parameters selection for a building. Then the prediction day conditions (e.g., forecasted weather conditions, occupancy, building operating conditions etc.) are used to predict energy load from this fixed model parameters (see **definition 2.1**, **Chapter 2**). On the contrary, the approach "**relevant data**" uses both offline and online database updating: The offline learning (see **definition 2.9**, **Chapter 2**) uses the prediction day conditions to select few representative datasets for model training and consequently a specific model parameter selection is started for each prediction day conditions. On the other hand, online learning (see **definition 2.9**, **Chapter 2**) uses the database that is updated with new measurements (for an existing building or new numerical simulation) after the day to predict is happened and can use training data to build a model for each consecutive day. The individual block that is used in whole framework is described briefly in later cases.

3.2 Offline Data Preparations

The offline data preparations block represented in Figure (3.3) includes a data classification, an additional feature generation and the most significant feature determination for the two modeling approaches "**all data**" and "**relevant data**" and is shown in Figure (3.4).

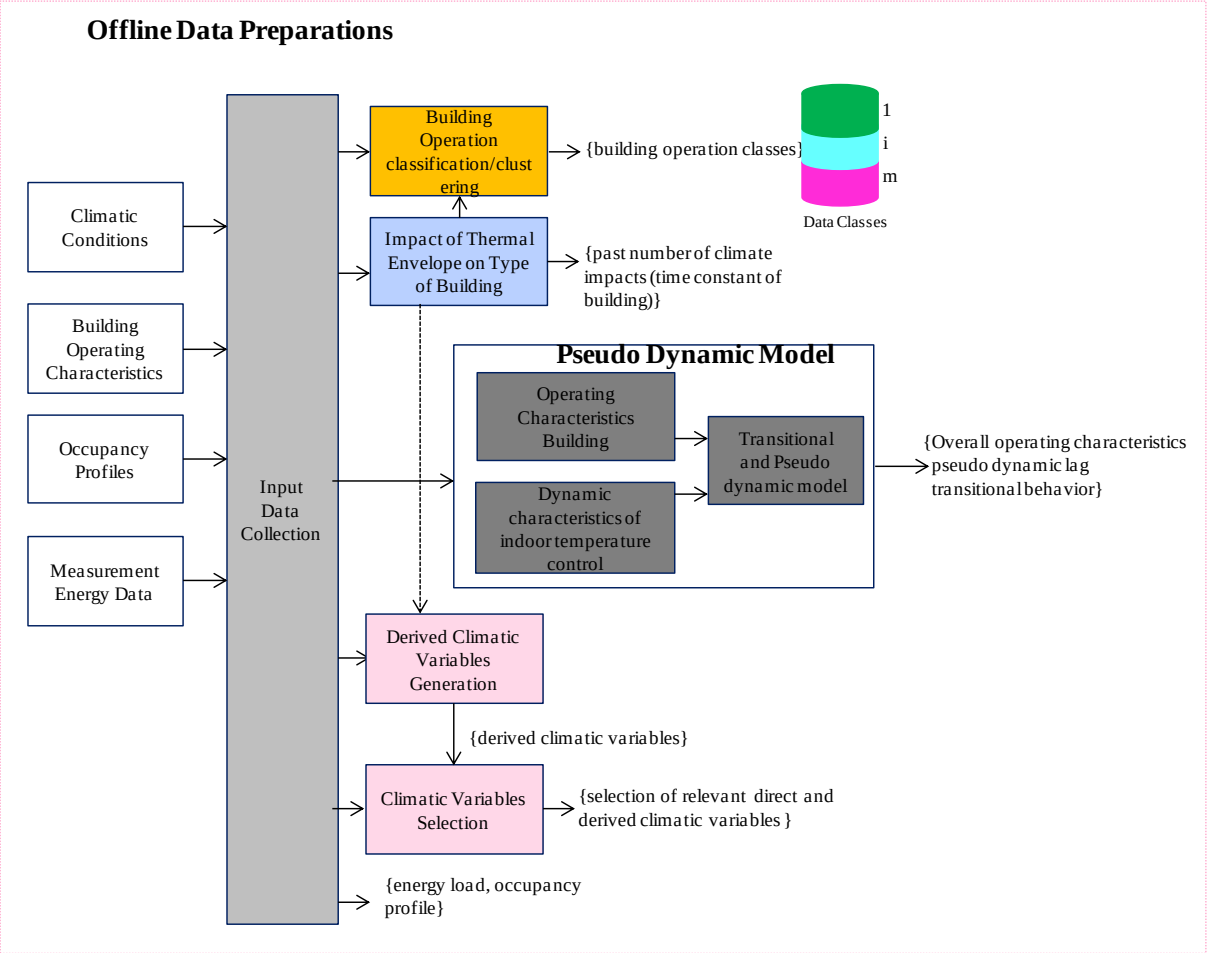

Figure 3.4: Preparations of offline data

The hourly input data are initially collected (from measurement or from numerical simulations). Some of the input data, particularly building operating conditions and occupancy profile can be even approximated. The building operations classes are determined in "**Building classification/clustering**" block and these represent the functioning profile of building during a week (detailed in **section 3.2.1**). Then, the impact of thermal envelope in building is evaluated based on simple physical understanding (e.g., time constant of building) in "**Impact of Thermal Envelope on Type of Building**" block to determine the number of past day climate impacts on the energy load of building. In addition, from the building operating conditions, occupancy profile[12] and the dynamic characteristics of the indoor temperature control in a building, "**Pseudo Dynamic Model**" (PDM) is developed to reflect the dynamics of occupancy and their interactions with building operating conditions. More details about PDM are outlined in **section 3.2.2**. Furthermore, the climatic variables, especially solar gain that directly impacts on building geometry are derived (for detail, see **section 3.2.3**). Finally, the climatic variable influence is only

[12] since the occupant's activities are modeled with fixed occupancy, the dynamic characteristics of occupant can be even stochastic in nature and this assumption might impact the accuracy in the prediction of energy load

evaluated through "**Climatic Variables Selection**" block since we have considered fixed schedules of the building operating conditions and the occupancy (see **remark 3.1**) as detailed in **section 3.2.4**.

3.2.1 Building Operation Classification/Clustering

The purpose of this subsection represented in Figure (3.4) is to identify the functioning classes of building operating profile of building during a week. These classes greatly affect the model accuracy while predicting energy load of a building. For instance, in an offices building, the energy load during normal working days is higher than the one during the weekend; while in a residential building, the energy load during working and weekend days can be similar. In addition, it also depends on the thermal envelope type of building. For example, the ratio window to wall in an offices building is significantly higher than one in a residential building.

There are various statistical methods applied to classify the data (in our case, the building energy load), for details see **remark 2.4, Chapter 2**. We used a statistical analysis based on canonical variate analysis (CVA) proposed by Li et al. [49] since this analysis transforms the input datasets into new axes with a visual representation and improves efficiently the decision to analyze the data. For details, see **Appendix C**.

The inputs of this classification is the heating or cooling energy consumption of a building and outputs of this classification represent building operation type in a week.

3.2.2 Pseudo Dynamic Model

The purpose of this model represented in Figure (3.4) is to introduce a priori knowledge on the dynamic behavior of the building. The inputs of this model are the occupancy (in terms of time patterns) and building operating conditions during a day and the dynamic characteristics of the indoor temperature control in a building. The outputs of this block are the derived features that represent dynamic behavior of the building.

In order to include the dynamic behavior in fixed occupancy patterns, we have proposed a novel PDM which includes a hidden transitional effect of occupancy by using time attributes when there is a change between the occupancy and the building operating conditions as a consequence of a set-point indoor temperature or the ventilation etc. This time attributes express a priori knowledge of the heat consumption dynamics.

The operating characteristics of building obtained from the operating conditions and the occupancy profile is shown in Figure (3.5) where x-axis represents operating period and y-axis represents magnitude of operating conditions. This magnitude is indirectly related to the power consumption or energy demand of the building. We represent the operating characteristics by a state and a transition where a state means a constant for a set-point indoor temperature and a transition means a change for this set-point temperature. The transition levels have a similar feature[13]. However, the energy demands required for a transition from point 2 to 3 or from point 6 to 7 are different. If the energy demand level of state 0, 1, 2, 3 and 4 in operating characteristics is represented by α_{ij}, then the energy demand required for transition from point j to point i is represented by β_{ij} in the transitional characteristics shown in Figure (3.6).

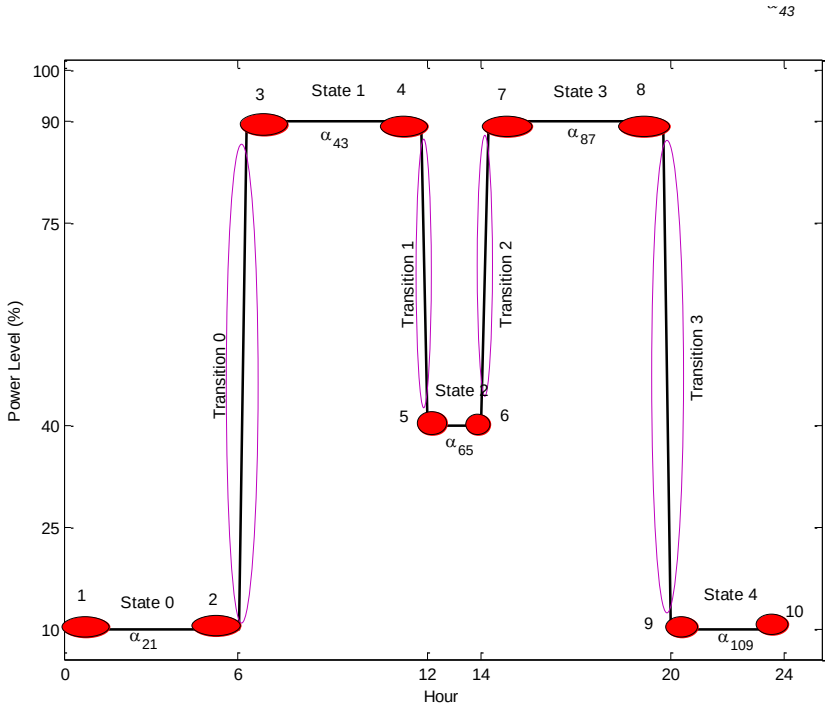

Figure 3.5: Overall operating characteristics of building (for a day)

The transitional characteristics corresponding to the operating conditions characteristics can be written as:

$$\beta_{ij} = \begin{matrix} \beta_{(i-2)(j-2)} + 2\Delta\beta \left|\alpha_{ij} - \alpha_{(i-2)(j-2)}\right|, & \forall i = 4,6,8 \ldots, j = 3,5,7 \\ \beta_0 & j = 1, i = 2 \end{matrix} \quad (3.1)$$

[13] To clearly explain this concept: we are in a theoretical situation. We are assuming that weather conditions are "fixed". One knows a priori the usual indoor temperature control. ESCOs have to provide enough energy to supply this demand of a building. So, just after an increasing of the set-point indoor temperature more energy is provided (in order to meet the demand and get steady state) than the energy requires for the steady state.

Where, β_0, $\Delta\beta$ and $|.|$ represent an initial energy demand level, a step size of a transition of the energy demand level and the absolute value respectively. Each level ($\beta_{21}, \beta_{43}, \beta_{65}, \beta_{87}$ and β_{109}) represents the transitional level and depends on the energy level of operational characteristics.

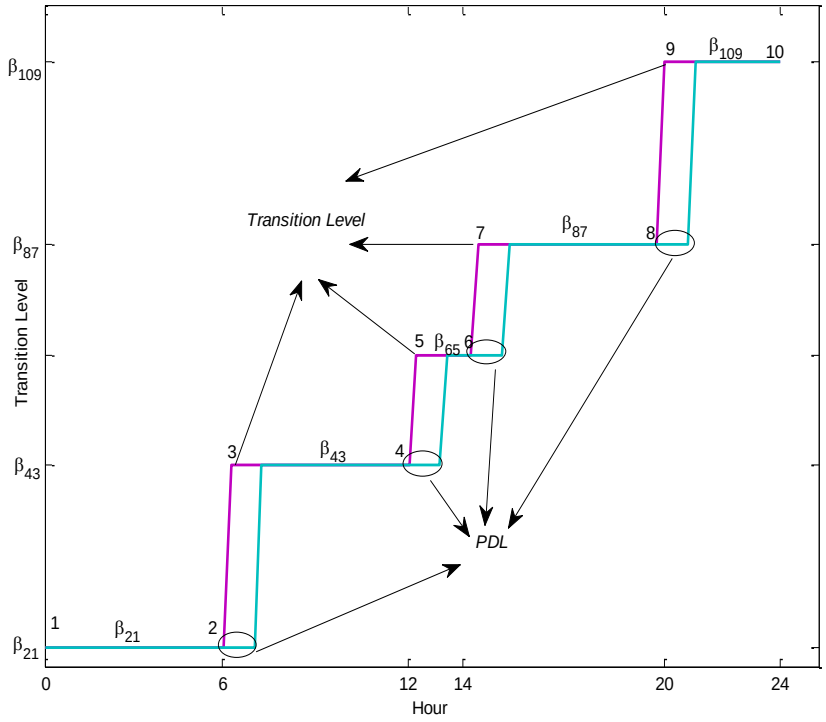

Figure 3.6: Transitional and pseudo dynamic characteristics (for a day)

The transitional characteristics describes the energy transition level of the operational characteristics (occupancy and building operating conditions), nonetheless, the dynamic transitional characteristics of the energy level attributes impact is still lacking.

The dynamic transition characteristics is modeled by a first order model of the indoor air temperature and the heating system (in an open loop) shown in Figure (3.7) where $\tau_{air,in}$ represents the time constant due to indoor thermal capacity. It can be seen that this time constant represents the time it takes to reach a new steady state for this indoor temperature. This time constant corresponds to the classical 63% of the new steady state. The steady state time $T_{steady,air,in}$ corresponds to the range $[3\tau_{air,in}, 6\tau_{air,in}]$. A **black-box** model must learn that for the same weather conditions (and occupancy), a change of the set-point temperature impacts the heating load on a period $[3\tau_{air,in}, 6\tau_{air,in}]$. We propose to indicate this time pattern by an array repeated with numerous time lags called pseudo dynamic lags (PDL), see Figure (3.6) with only one repetition or with multiple repetitions. We named this model as a "**pseudo dynamic**".

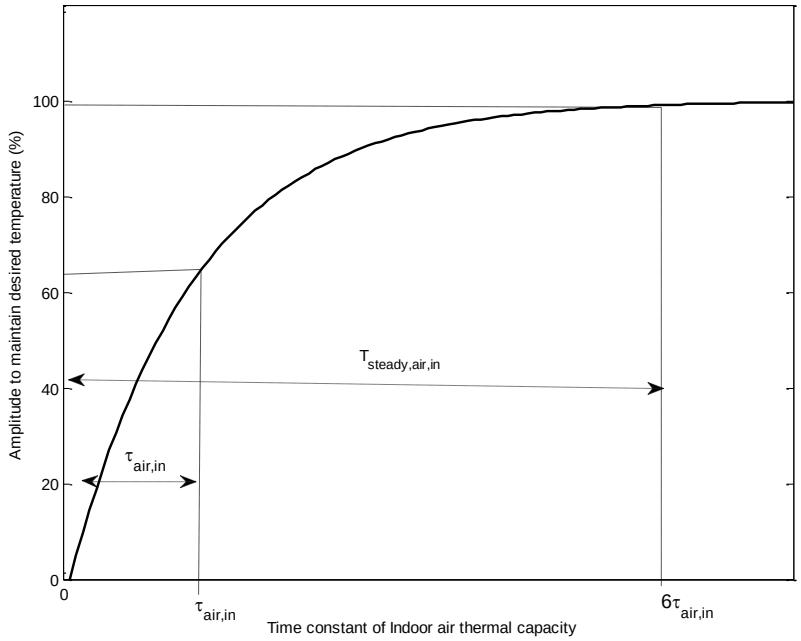

Figure 3.7: Dynamic characteristics of indoor temperature control in a building

The output of this PDM consist overall building operating conditions, transitional behavior and PDL.

3.2.3 Derived Climatic Variables Generation

The purpose of this block shown in Figure (3.4) is to generate derived climatic variables.

The energy demand requirement in LEB is largely depend on the solar gains in building thus solar gain on walls ($\emptyset_{Sint}$) and solar gain transmitted through windows ($\emptyset_{Sext}$) are derived from climatic variables. In addition, external heat transfer (by conduction/convection with the external temperature T_{ext} and by solar radiation on walls $\emptyset_{Sint}$) are stored by walls for a long period (for detail see **Section 3.1.1**) and in order to include these storage behavior, these derived climatic variables are further modified.

We proposed to consider this storage effects by introducing a temporal moving average window (T_{ext_TDM} for the external temperature and $\emptyset_{Sint\_TDM}$ for the solar gain on walls) depending on the past day climatic conditions dynamics.

Those numbers of past day climatic conditions are obtained in "**Impact of Thermal Envelope on Type of Building**" in "**Offline Data Preparations**" block in Figure (3.4). Thus, the output of this "**Derived Climatic Variables Generation**" block are the time patterns of solar gain on walls

($\emptyset_{Sint}$), the time patterns of solar gain transmitted through windows ($\emptyset_{Sext}$), the temporal moving average of external temperature (T_{ext_TDM}) and the temporal moving average of solar gain on walls ($\emptyset_{Sint\_TDM}$).

The horizon of those moving averages is one parameter for the case study.

3.2.4 Climatic Variables Selection

The purpose of this selection shown in Figure (3.4) is to determine which weather variables and their dynamics are relevant for the prediction of energy consumption and limit the variables during the learning phase.

The basic ideas for deriving the climatic variables are:

1. Under fixed conditions (building operations and occupancy profile), the climatic conditions are the only variables that impacts the building energy consumption.

2. Recalling that climatic variables selection is based on the selection of features which have a highest correlation or the features which provide a high accuracy in the prediction model or combination of features which gives high accuracy in the prediction model. Such a combination of features selection might be effective but it may not reflect the physical significance of importance of each feature. The autocorrelation method is not suitable because it make correlation of the future heating with its past heating and violates the proposed methodology of dynamical model to predict for several days ahead. The principal component analysis (PCA) has several drawbacks since it depends on input data distribution and sometimes it neglects some high relevant inputs which might increase the model performance.

Considering the above facts, we used a filter as a climate feature selection method (**Chapter 2**, **remark 2.5**) which is based on simple correlation indexes. Such a correlation measures the strength and the weakness of a linear relationship between two features (external temperature and heating energy consumption for example). We select the input features if its correlation indexes is higher than threshold values (φ_{th})[14] and discard the features if it less than this threshold values. We used Pearson coefficient of correlation to determine the relevance of the climatic conditions (e.g., solar radiation, external temperature) and the derived climatic conditions (e.g., solar gain on walls and solar gain transmitted through windows) and a sample cross correlation to determine the

[14] limiting value which provides a benchmark to compare with the features value

hourly dynamics of those selected climatic conditions (e.g., external temperature of last 1 to 4 hours from prediction hours).

Denoting the input features of the training data by x (e.g., external temperature, horizontal solar radiation) and the output features by y (e.g., heating energy consumption), a Pearson correlation coefficient is calculated by dividing the covariance of input and output features to their individual standard deviations as shown in Equation (3.2), where r represents the Pearson correlation coefficient, cov(xy) represents the strength of linear relationship between two features x and y, S_x and S_y are the standard deviations of the features x and y, n is the total number of training data and $\bar{x}$ and $\bar{y}$ are the sample mean of time series x and y.

$$r = \frac{cov(xy)}{S_x S_y} \tag{3.2}$$

$$cov(xy) = \frac{1}{n-1}\sum_{i=1}^{n}(x^i - \bar{x})(y^i - \bar{y})$$

The Pearson correlation coefficient varies in the range -1 to 1. Representing time series of an input feature x_t (e.g., external temperature) and an output feature y_t (e.g., heating energy consumption) at lags Φ=0,$\pm$1, $\pm$2 etc., a cross correlation function determines the time delay between this input and this output. When these two input and output are best aligned with maximum (or minimum if these are negatively correlated) in the same point, then it is regarded as in a good time dynamics accordingly. The cross-correlation between two input and output time series is given by Equation (3.3).

$$r_{xy}(\Phi) = \frac{cov(xy\Phi)}{S_x S_y} \tag{3. 3}$$

where,

$$cov(xy\Phi) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-\Phi}(x_t - \bar{x})(y_{t+\Phi} - \bar{y}); & \Phi = 0,1,2,\dots\dots \\ \frac{1}{n}\sum_{t=1}^{n+\Phi}(y_t - \bar{y})(x_{t-\Phi} - \bar{x}); & \Phi = 0,-1,-2,\dots\dots \end{cases}$$

In Equation (3.3), $r_{xy}(\Phi)$ represents cross-correlation between two time series x and y at Φ lag, $cov(xy\Phi)$ is covariance of two time series x and y at Φ lag, S_x and S_y are the sample standard deviations of the time series x and y, $\bar{x}$ and $\bar{y}$ are the sample means of time series x and y [100].

Eventually, the input of this selection are the climatic conditions variables (e.g., external temperature, solar radiation), derived climatic conditions variables (e.g., solar gain on walls, temporal moving average of solar gain on wall) and its past hour dynamics (e.g., external temperature of last 2 hours) and the output of this block are the selected direct and derived climatic feature and their past hour behaviors.

3.3 Prediction Day Conditions

The prediction day conditions block represented in Figure (3.3) provide information about the prediction day (or couple of days). The input of this block represents the forecast weather day (or couple of days), the expected conditions (occupancy profile and building operations) and information about the building operation classes, see Figure (3.8). The outputs of this block depend on the modeling approach, either "**all data**" or "**relevant data**".

- For "**all data**" modeling approach, the outputs contain forecast weather conditions, occupancy, building operating characteristics and information regarded to building operation classes. These are used for prediction of energy load from the parameter selected by **AI model**.
- For "**relevant data**" modeling approach, the outputs contain the forecast weather conditions, a 24 hours time patterns and the previous days and the building operation classes (which are used to select similar weather conditions). In addition, it also contains outputs similar to "**all data**" modeling approach which are used for prediction from the model parameter selected by **AI model**.

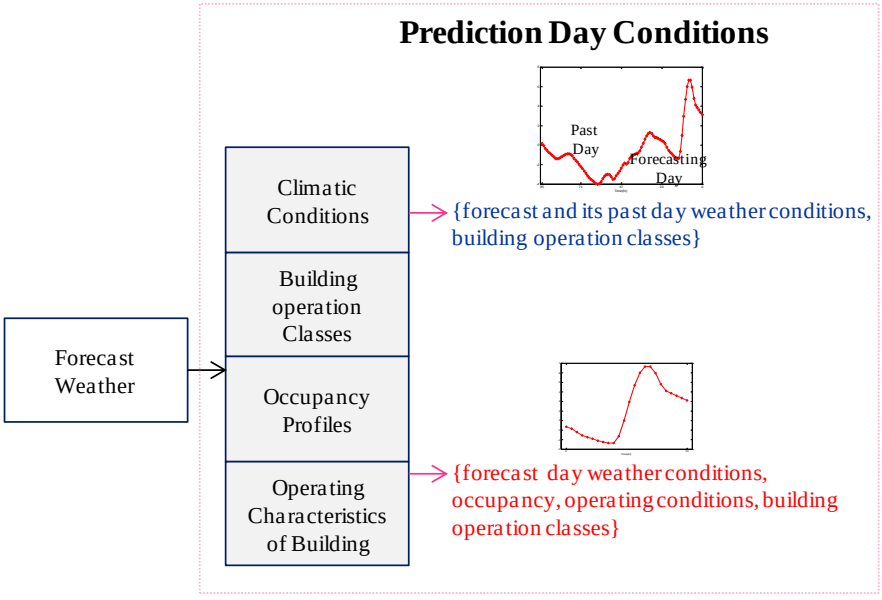

Figure 3.8: Prediction day conditions

3.4 All Data Modeling Approach

Definition 3.2: Cross validation

A cross-validation refers to the splitting of training data into multiple sets of validation.

The framework of "**all data**" modeling approach represented in Figure (3.3) based on offline learning for energy load prediction is shown in Figure (3.9). It can be seen that in the "**all data**" modeling approach, all offline training data obtained in "**Offline Data Preparations**" block in Figure (3.4) includes energy load, occupancy profile, building operating characteristics (building operating condition, transitional characteristics and pseudo dynamic lag) together with selected direct and indirect climatic variables and their dynamics for each building operation classes. These offline data used for model training are divided into each building operation classes and **AI model** are evaluated and learned accordingly using cross validation (see **definition 3.2**). Then, the parameters of **AI model** are identified for each building operation classes. Finally, forecast day weather conditions, occupancy, operating conditions obtained in "**Prediction Day Conditions**" block in Figure (3.8) are used for prediction of energy load from the identified parameters of learned **AI model** for each building operation classes. This concluded that "**all data**" modeling approach are based on offline learning because their parameters are not changed with new dataset availability in the future due to the inconvenience of model training CPU-time while updating all offline data.

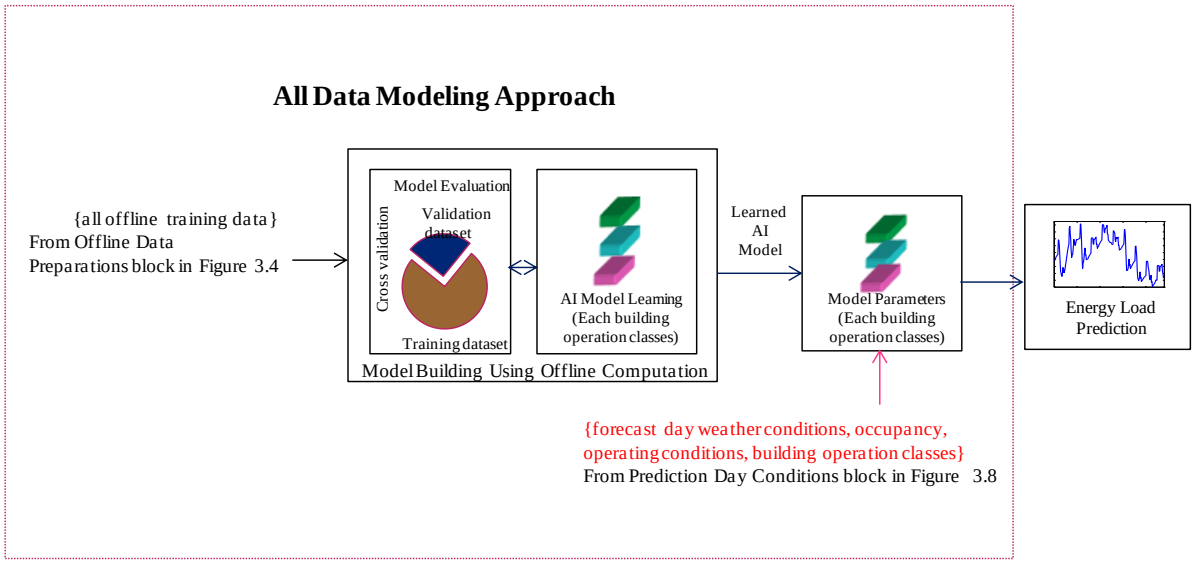

Figure 3.9: Framework of all data modeling approach based on offline learning for energy load prediction

3.5 Relevant Data Modeling Approach

The framework of the proposed "**relevant data**" modeling approach represented in Figure (3.3) uses relevant data selection method for each prediction day conditions and is shown in Figure (3.10). Relevant data selection basically use selection of similar day data based on similarity of climatic conditions, building operation classes (see, **Section 3.2.1**), impact of past day climate dynamics (obtained from "**Impact of Thermal Envelope on Type of Building**" block in Figure 3.4), number of relevant days data and weight of climatic conditions on energy load of building.

Research Question 1: How to "introduce" such kinds of dynamic behavior in AI model to predict the energy consumption couple of days for a LEB?

The relevant data selection is done in three main steps:

1. Considering the prediction day conditions shown in Figure (3.8), it is possible to calculate the variation of climatic conditions such as external temperature between the prediction day and a day in the training database based on **deviation criteria.** We select different **deviations criteria** based on simple physical understanding and pattern recognition methods:
    i. The physical methods are based on heating degree-day and modified heating degree day that includes variation of energy load weight effect at different time intervals.
    ii. The pattern recognition methods: Frechet distance and dynamic time warping are based on finding similarity patterns.

    These simple physical understanding and pattern recognition methods compared each prediction day and its past days with each training days and its past days climatic conditions. The more details about identification of similar climatic conditions are mentioned in **Section 3.5.1**.

    Consequently, for one prediction day and one specific day in the training database, the deviation criteria is a vector whose size is equal to the number of climatic conditions (external temperature, solar radiation, humidity etc.)

2. The purpose of the second step is to combine the **deviation criteria** vector in order to select a sub-database for relevant data modeling.
    i. A pre-calculation on the training database is first done independently to the prediction day. A wavelet analysis is performed for the determination of influence of climatic variables on building energy load. We used wavelet decomposition in order to reduce several climatic variables of the past days to transform into wavelet coefficients without losing the properties of several day behaviors. With daily average energy load and suitable wavelet coefficients of climatic conditions and their past days, weight of selected climatic conditions and their past days are determined by using support vector machine (SVM) based on linear kernel. The number of past day climatic conditions for each prediction day depends on the type of building that is obtained from "**Impact of Thermal Envelope on Type of Building**" block in Figure (3.4). The details about the influence of climatic variable in building load are provided in **Section 3.5.2**.
    ii. Knowing those weights, the previous **deviation criteria** vector is weighted by those wavelet coefficients so only one metric criterion is used to select suitable days from the training database.
3. The final step is the identification of number of days for the model training. The smaller weights obtained from step 2 means the more relevance of day for model training since the closest match of prediction day conditions with training day database will have smallest weights. Hence, depending on the number of relevant day (see **section 3.5.3**), the smallest weights are selected from the training database to find suitable day for model training for particular prediction day conditions.

Then the relevant datasets for each prediction day are used to build an **AI model** using cross validation. Other detail about machine learning based **AI model** and cross validation are detailed in **Section 3.6**. Finally, the parameters are identified from learnt **AI model** for each prediction day conditions. We therefore conclude that parameters are changed each day based on prediction day conditions due to fewer relevant datasets to represent whole behavior and its computational realization. During the prediction conditions, forecast day weather conditions, occupancy, operating conditions and building operation classes obtained in "**Prediction Day Conditions**" block in Figure (3.8) are used for prediction of energy load from identified parameter of **AI model** of a particular day.

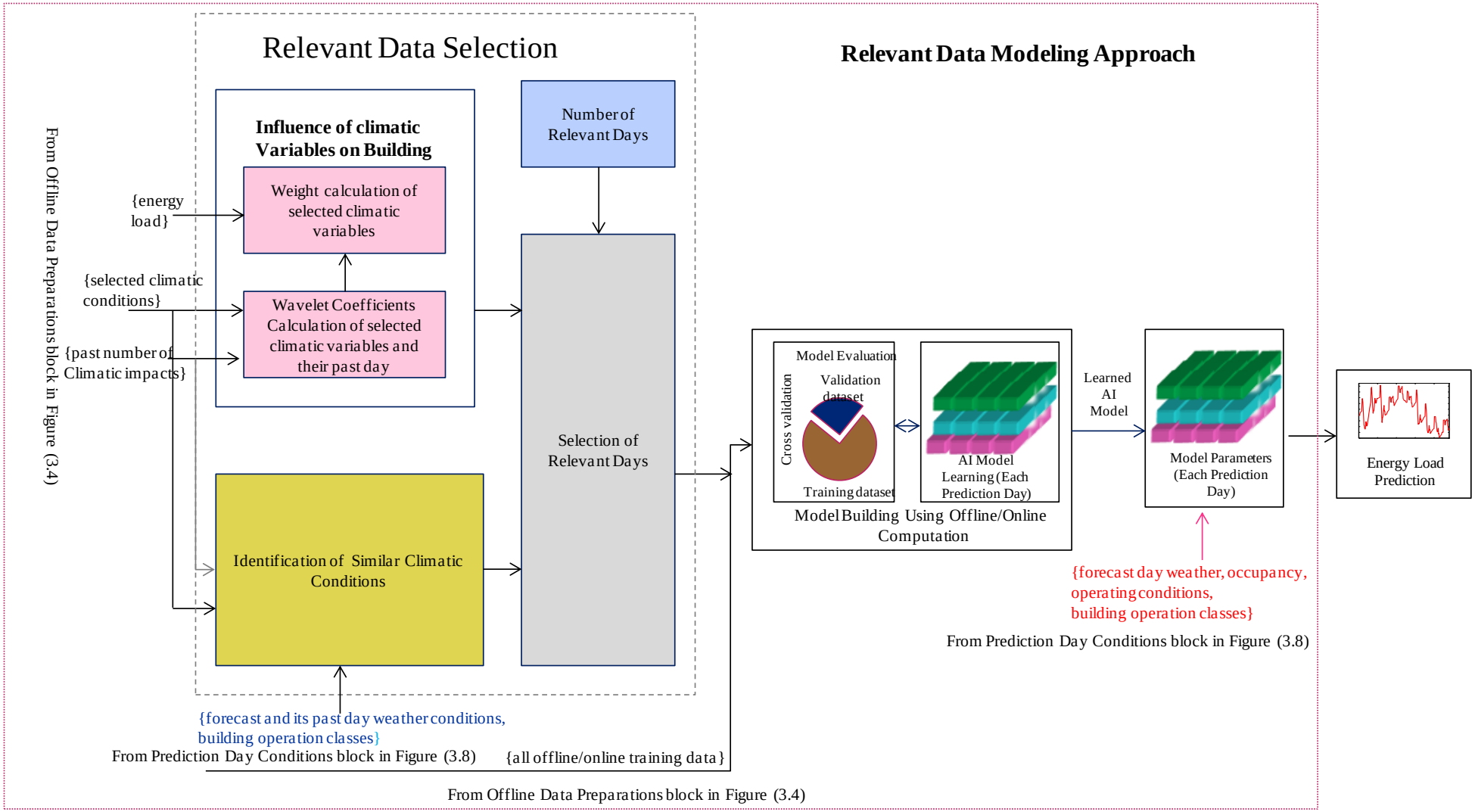

Figure 3.10: Framework of proposed relevant data modeling approach based on online/offline learning

3.5.1 Identification of Similar Climatic Conditions

The purpose of this identification is to select similar climatic conditions of prediction day and its past day.

We have identified the similar behavior of climatic conditions (e.g., external temperature) of prediction and its past days dynamics with training day data based on physical understanding and pattern recognition methods. The physical methods are based on Heating Degree Day (HDD) and **modified HDD** (proposed in this manuscript), and pattern recognition methods are based on Frechet Distance and Dynamic Time Warping.

Heating Degree Day

Representing HDD of database by a notation $\mathbf{HDD_N}$, see Equation (3.4); and forecast HDD of prediction day by $\widehat{HDD}$; the similar day from N number of HDD database are determined by comparing forecasted $\widehat{HDD}$ with $\mathbf{HDD_N}$ and is represented by $\mathrm{H}\left(\widehat{\mathrm{HDD}}, \mathbf{HDD_N}\right)$ shown in Equation (3.5).

$$\mathbf{HDD_N} = \begin{bmatrix} \mathrm{HDD}_1 \\ \mathrm{HDD}_2 \\ \vdots \\ \mathrm{HDD}_\mathrm{N} \end{bmatrix} \tag{3.4}$$

$$\mathrm{H}\left(\widehat{\mathrm{HDD}}, \mathbf{HDD_N}\right) = \left\| \widehat{HDD} - \mathbf{HDD_N} \right\| \tag{3.5}$$

Since LEB have large time constant, HDD of past days from prediction day also impact the building energy load. In order to avoid the weight effect of HDD from forecasted day with past day, normalization is performed in order to compare the weight of forecasted HDD and its past day. This normalization avoids the weight effect that might come from different range of HDD value of forecasted and its past day. Normalized similarity weight, i.e., $\mathbf{nHDD_N}$ correspondence to forecasted $\widehat{HDD}$ of prediction day from database is shown in Equation (3.6).

$$\mathbf{nHDD_N} = \frac{\mathrm{H}\left(\widehat{\mathrm{HDD}}, \mathbf{HDD_N}\right) - \min\left\{\mathrm{H}\left(\widehat{\mathrm{HDD}}, \mathbf{HDD_N}\right)\right\}}{\max\left\{\mathrm{H}\left(\widehat{\mathrm{HDD}}, \mathbf{HDD_N}\right)\right\} - \min\left\{\mathrm{H}\left(\widehat{\mathrm{HDD}}, \mathbf{HDD_N}\right)\right\}} \tag{3.6}$$

Assuming *u* number of past day impacts due to time constant of building obtained from "**Impact of Thermal Envelope on Type of Building**" block in Figure (3.4), all normalized weight for prediction day together with *u* number of past day for N number of training day is expressed as:

$$\mathbf{nW_{HDD,N}} = [\mathbf{nHDD_N}(1) \quad \mathbf{nHDD_N}(2) \ \ldots \quad \mathbf{nHDD_N}(u)] \tag{3.7}$$

Modified Heating Degree Day

The proposed **modified HDD** method considers the weighted factor of energy consumption at each time steps to the average energy consumption of training day to distinguish different weights to different time intervals during a day. It includes variation of climatic conditions day-to-day and indirectly includes internal gains and solar gains effect by introducing weighted energy consumption.

Denoting climatic variables (e.g., external temperature) of training database by a generic notation $\mathbf{X_N}$, see Equation (3.8) and its corresponding energy consumption and daily mean energy consumption of training database by $\mathbf{Y_N}$ and $\overline{\mathbf{Y}}_\mathbf{N}$ respectively, and forecasted weather of prediction day (e.g., external temperature) by $\hat{\mathrm{X}}$, see Equation (3.9); modified HDD determined the similarity by Equation (3.10) and are represented by $\mathrm{mH}\left(\hat{\mathrm{X}}, \mathbf{X_N}\right)$. In Equation (3.8) and (3.9), $\mathcal{M}$ is the sampling length of data of each day.

$$\begin{bmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \\ \vdots \\ \mathrm{X}_\mathrm{N} \end{bmatrix} = \begin{bmatrix} \mathrm{x}_1(1) & \mathrm{x}_2(1) & \cdots & \mathrm{x}_\mathcal{M}(1) \\ \mathrm{x}_1(2) & \mathrm{x}_2(2) & \cdots & \mathrm{x}_\mathcal{M}(2) \\ \vdots & \vdots & . & \vdots \\ \mathrm{x}_1(\mathrm{N}) & \mathrm{x}_2(\mathrm{N}) & \cdots & \mathrm{x}_\mathcal{M}(\mathrm{N}) \end{bmatrix} \tag{3.8}$$

$$\hat{\mathrm{X}} = [\hat{\mathrm{x}}_1 \quad \hat{\mathrm{x}}_2 \quad \ldots \quad \hat{\mathrm{x}}_\mathcal{M}] \tag{3.9}$$

$$\mathrm{mH}\left(\hat{\mathrm{X}}, \mathbf{X_N}\right) = \left(\sum_{\mathrm{i}=1}^{\mathcal{M},\mathrm{j}} \frac{\mathbf{Y}_{i,\mathrm{j}}}{\overline{\mathbf{Y}}_{i,\mathrm{j}}} \left(\hat{\mathrm{X}} - \mathbf{X}_{i,\mathrm{j}}\right)^2\right)^{1/2} = \begin{bmatrix} \mathrm{mH}_1 \\ \mathrm{mH}_2 \\ \vdots \\ \mathrm{mH}_\mathrm{N} \end{bmatrix} \quad \forall \mathrm{j} = 1 \text{ to } \mathrm{N} \tag{3.10}$$

Equation (3.10) considers weighted factor at different time intervals for N number of training day data by the coefficient $\sum_{\mathrm{i}=1}^{\mathcal{M}} \frac{\mathrm{Y_i}}{\overline{\mathrm{Y}}_\mathrm{i}}$. The lower value of $\mathrm{mH}\left(\hat{\mathrm{X}}, \mathbf{X_N}\right)$ means more similarity of prediction day weather $\hat{\mathrm{X}}$. Normalization is performed on the similarity weights of climatic conditions (e.g., external temperature) obtained from Equation (3.10) so that weight can be compared and effect of other climatic variable (e.g., horizontal solar radiation) does not affect each other. Normalized similarity weight, i.e., $\mathbf{nH_N}$ of climatic variable (e.g., external temperature) is shown in Equation (3.11).

$$\mathbf{nH_N} = \frac{\mathrm{mH}\left(\hat{\mathrm{X}}, \mathbf{X_N}\right) - \min\left\{\mathrm{mH}\left(\hat{\mathrm{X}}, \mathbf{X_N}\right)\right\}}{\max\left\{\mathrm{mH}\left(\hat{\mathrm{X}}, \mathbf{X_N}\right)\right\} - \min\left\{\mathrm{mH}\left(\hat{\mathrm{X}}, \mathbf{X_N}\right)\right\}} \tag{3.11}$$

Equation (3.12) represents normalized similarity weight for single climate variable. For other climatic variables and their past days dynamics (e.g., horizontal solar radiation), it follows Equation (3.8-3.11) to represent normalized similarity weight. Representing number of past day climatic variables by *u* (e.g., external temperature of the last two days obtained from "**Impact of Thermal Envelope on Type of Building**" block in Figure 3.4) and most significant climatic variables for particular building by *v* (e.g., external temperature, horizontal solar radiation etc. from "**Climatic Variables Selection**" block in Figure 3.4), the normalized similarity weight of all prediction days selected climatic variables and their past dynamics can be written in general form by $\mathbf{nw}_{v,\mathbf{N}}$ and Equation (3.11) is further modified to Equation (3.12). For example in Equation (3.12), normalized similarity weight vector of external temperature of prediction day $\mathbf{nH}_{\mathbf{1},\mathbf{N}}(1)$, the day before prediction day $\mathbf{nH}_{\mathbf{2},\mathbf{N}}(1)$ etc. are represented by $\mathbf{nw}_{\mathbf{1},\mathbf{N}}$; and normalized similarity weight vector of horizontal solar radiation of prediction day $\mathbf{nH}_{\mathbf{1},\mathbf{N}}(2)$, day before prediction day $\mathbf{nH}_{\mathbf{2},\mathbf{N}}(2)$ etc. are represented by $\mathbf{nw}_{\mathbf{2},\mathbf{N}}$.

$$\begin{bmatrix} \mathbf{nw}_{\mathbf{1},\mathbf{N}} \\ \mathbf{nw}_{\mathbf{2},\mathbf{N}} \\ \vdots \\ \mathbf{nw}_{v,\mathbf{N}} \end{bmatrix} = \begin{bmatrix} \mathbf{nH}_{\mathbf{1},\mathbf{N}}(1) & \mathbf{nH}_{\mathbf{2},\mathbf{N}}(1) & \cdots & \mathbf{nH}_{u+\mathbf{1},\mathbf{N}}(1) \\ \mathbf{nH}_{\mathbf{1},\mathbf{N}}(2) & \mathbf{nH}_{\mathbf{2},\mathbf{N}}(2) & \cdots & \mathbf{nH}_{u+\mathbf{1},\mathbf{N}}(2) \\ \vdots & \vdots & . & \vdots \\ \mathbf{nH}_{\mathbf{1},\mathbf{N}}(v) & \mathbf{nH}_{\mathbf{2},\mathbf{N}}(v) & \cdots & \mathbf{nH}_{\mathbf{u}+\mathbf{1},\mathbf{N}}(v) \end{bmatrix} \tag{3.12}$$

Dynamic Time Warping

Dynamic time warping (DTW) is a distance measure time series method which finds the similar patterns of signal between two time series though they are not aligned in time. DTW finds similarities based on acceleration-deceleration of signals within the time dimension. Because of a large time constant in LEB, the influence of climatic variables influence the building and vary according to time, hence, DTW is more suitable. This method has been used in pattern recognition to find similarities of building energy patterns [101]. It determines the similarities of climatic variables by calculating Euclidean distance of training days and predicted day climatic variable and their past days in different warping path.

The illustration of DTW to find similarity patterns is shown in Figure (3.11) where DTW calculates the Euclidean distance between climatic variables (e.g., external temperature) of training and prediction day in two warping path. The path that minimizes sum of Euclidean distance between two time series is chosen as optimal warping path. This process continues to calculate optimal warping path for each training day with prediction day climatic variables to determine similarity weights.

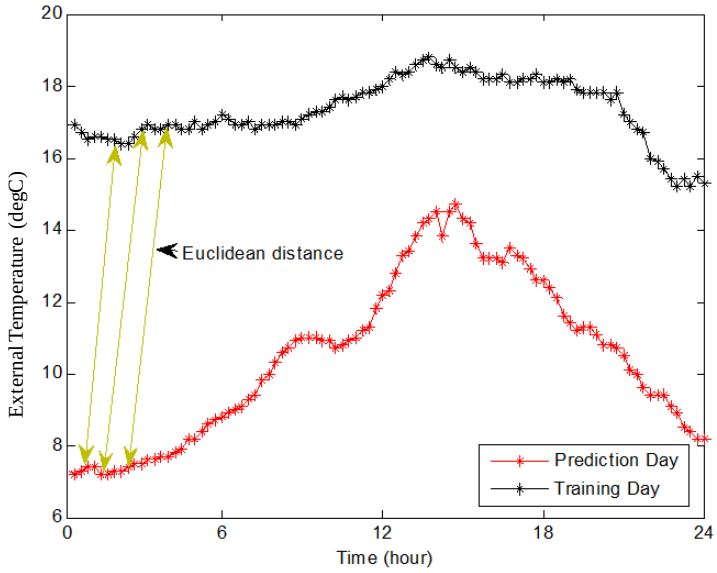

Figure 3.11: Illustration of dynamic time warping to select similar climatic variables (e.g., external temperature)

The similarity days of forecast weather of prediction day or days are determined by comparing $\hat{X}$ (shown in Equation 3.9) with the weather of database $\mathbf{X_N}$ **(**shown in Equation 3.8) by minimizing DTW and are represented by $D(\hat{X}, \mathbf{X_N})$ as shown in Equation (3.13), for details to calculate DTW see Keogh and Ratanamahatana [102].

$$D(\hat{X}, \mathbf{X_N}) = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{bmatrix} \tag{3.13}$$

In Equation (3.13), the lower value of $D(\hat{X}, \mathbf{X_N})$ means the similarity of prediction day weather $\hat{X}$ with the N corresponding training day. Similarly, normalization is performed on similarity weights (Equation 3.13) and then to other climatic conditions. So, it follows the same procedure mentioned in Equation (3.11-3.12) similar to modified HDD.

Frechet Distance

Frechet distance (FD) is a pattern recognition method that measures the similarity degree between two continuous curves. If the FD of two curves is small, then the curves are similar and if the FD is large, curves are said to be dissimilar.

Figure (3.12) shows illustrations of identification of similar climatic variables (e.g., external temperature) by FD. It can be seen that prediction day forecast weather is compared with training

day database and the FD method selects the training day that has smallest value. For instance, the Frechet Distance between prediction day and training days 1-3 of climatic variables (e.g., external temperature) shown in Figure (3.12) are 0, 0.24 and 0.26 and illustrated that training day 1 is more similar compare to other training days.

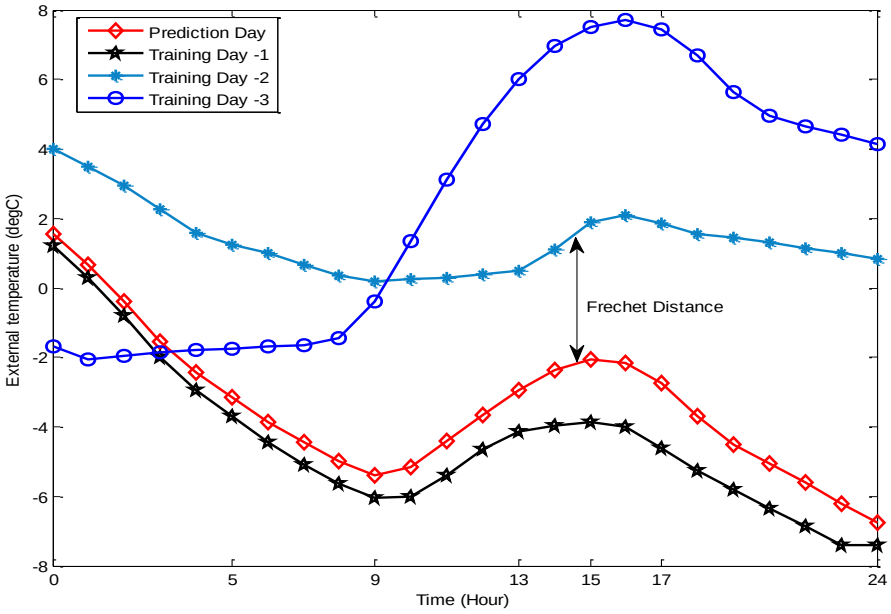

Figure 3.12: Illustration of Frechet distance to select similar climatic variables (e.g., external temperature)

Similar to DTW method, FD method determined the similarity days of forecast weather of prediction day by comparing $\hat{X}$ (shown in Equation 3.9) with the weather of database $\mathbf{X_N}$ (shown in Equation 3.8) and are represented by $\mathrm{FD}(\hat{X}, \mathbf{X_N})$ as shown in Equation (3.14), for details on calculation of FD, see Wylie and Zhu [103].

$$\mathrm{FD}(\hat{X}, \mathbf{X_N}) = \begin{bmatrix} \mathrm{FD}_1 \\ \mathrm{FD}_2 \\ \vdots \\ \mathrm{FD}_N \end{bmatrix} \tag{3.14}$$

The lower the value of $\mathrm{FD}(\hat{X}, \mathbf{X_N})$ in Equation (3.14) means the more similarity of prediction day weather $\hat{X}$ with N corresponding training day. Similarly, normalization is performed on similarity weights (Equation 3.14) and also performed to other climatic conditions, therefore it follows the Equation (3.11-3.12).

3.5.2 Influence of Climatic Variables on Building

The influences of climatic variables depend on building properties such as insulation, thermal mass and geometrical parameters. It also depends on the energy consumption type of building, e.g., external temperature is more dominant for heating energy consumption whereas solar gain is more dominant for cooling energy consumption. In case of heating energy consumption for LEB, solar gains is also equally important since the heat transfer by conduction/convection is absorbed by the walls for long period to maintain equilibrium indoor climate. Hence, it is essential to find the influence of each selected climatic variables on energy load of building.

Selected climatic variables obtained in "**Climatic Variables Selection**" block in Figure (3.4) are pre-processed using wavelet analysis shown in "**Relevant Data Selection**" block in Figure (3.10). The suitable decomposition level is obtained by observing the reconstruction of the original signal using approximation and detail coefficients. Consequently, depending on the type of climate variables, these are converted into wavelet low frequency and high frequency components shown in "**Wavelet coefficients calculation of selected climatic variables and their past day**" block in Figure (3.10). For instance, the heat energy is transfer by external temperature T_{ext} in the walls for a long period, so the decomposed signals of them are expressed by low-frequency and high-frequency coefficients. On the other hand, though heat energy transfer by solar gain on walls $\emptyset_{Sint}$ for a long period, their average behavior is sufficient to characterize daily average heating load so they are expressed by low frequency components. Moreover, solar gain transmitted through windows $\emptyset_{Sext}$ has fast impact in the indoor temperature and their responses to the heating load is at the same instant of time, hence these are considered only by low frequency coefficients. Horizontal solar radiation $\emptyset_{Sh}$ is integrated with solar gain on walls $\emptyset_{Sint}$ itself and their average behavior can easily characterize the daily average heating load thus expressed by low frequency coefficients. Details about wavelet decomposition are presented in Mallat [104]. In order to determine influence of climatic variables, i.e., weights of decomposed low and high frequency components of climatic variables on energy load of building, a daily average energy load is used. The weight of these selected variables is calculated by using **SVM based on linear kernel** and are represented by "**Weight Calculation of Selected Climatic Variables**" block in Figure (3.10). This block is represented as an intermediate model. Details about SVM and kernel are outlined in Section 3.6.2.

We have applied the wavelet decomposition to climatic variables (e.g., external temperature) day by day. By denoting low frequency coefficients at desired level of wavelet by *a* and high frequency coefficients by *d*, the weight of the low/high frequency coefficients of climatic variable is obtained from SVM based on linear kernel and is represented in Equation (3.15-3.16). In Equation (3.15-3.16), *z* is the decomposed length of $\mathcal{M}$ sample length data in a day.

$$a = [a_1\ a_2\ \dots\ a_z] \tag{3.15}$$

$$d = [d_1\ d_2\ \dots\ d_z] \tag{3.16}$$

Then, the total approximation and detail coefficient weight of climatic variable (e.g., external temperature) can be estimated from Equation (3.17-3.18):

$$a_T = \sqrt{\sum_{i=1}^{z} a_i^2} \tag{3.17}$$

$$d_T = \sqrt{\sum_{i=1}^{z} d_i^2} \tag{3.18}$$

Equation (3.17-3.18) which represents total approximation and detail coefficients are further converted into desired weight (wc) of particular climate variable (e.g., external temperature) and is represented by:

$$\mathrm{wc} = \sqrt{a_T^2 + d_T^2} \tag{3.19}$$

Equation (3.19) represents wavelet coefficient weight of a single climate variable and for other climatic variables and their past days dynamics (e.g., horizontal solar radiation) depending on low or high frequency requirements; it follows Equation (3.15-3.19). Then the normalized wavelet coefficient $\mathbf{nC}_v$ is calculated for most significant variables of building *v* and their past dynamics *u*, thus Equation (3.19) is further modified to Equation (3.20).

$$\mathbf{nC}_v = \frac{\mathrm{wc_{ij}} - \min\{\mathrm{wc_{ij}}\}}{\max\{\mathrm{wc_{ij}}\} - \min\{\mathrm{wc_{ij}}\}}\ \forall \mathrm{i} = 1{:}\,v; \mathrm{j} = 1{:}\,1+u \tag{3.20}$$

Equation (3.20) can also be simplified to Equation (3.21) to represent influence of each climate variable and their past days in terms of wavelet coefficients $\mathbf{nC}_v$ in each variable normalized

form. For example in Equation (3.21), the influence of climate variable external temperature T_{ext}, i.e., total influence $nc(1)$, prediction day $nc_1(1)$, the day before prediction day $nc_2(1)$ etc. on building load are represented by $nC_1$and the influence of horizontal solar radiation $\emptyset_{Sh}$, i.e., total influence $nc(2)$, prediction day $nc_1(2)$, day before prediction day $nc_2(2)$ etc. on building load are represented by nC_2.

$$\begin{bmatrix} nC_1 \\ nC_2 \\ \vdots \\ nC_v \end{bmatrix} = \begin{pmatrix} nc(1) \\ nc(2) \\ \vdots \\ nc(v) \end{pmatrix} * \begin{bmatrix} nc_1(1) & nc_2(1) & \cdots & nc_{u+1}(1) \\ nc_1(2) & nc_2(2) & \cdots & nc_{u+1}(2) \\ \vdots & \vdots & . & \vdots \\ nc_1(v) & nc_2(v) & \cdots & nc_{u+1}(v) \end{bmatrix} \tag{3.21}$$

The influence of climatic variables on building load using **SVM based on linear kernel** is also compared with **LSM based on regression model** (see **remark 2.7**, **Chapter 2** for LSM based on regression).

3.5.3 Selection of Relevant Days

The suitable choice of number of days for model training depends on the performance of prediction model. In case of modified HDD, the final weight of all training days $\mathbf{W_N}$ that depends on similarity of climatic variables of training and prediction day, and their building impacts is obtained by deducing Equation (3.21) and Equation (3.12).

$$\mathbf{W_N} = \mathbf{nC_v}.(\mathbf{nw_{v,N}})^{\mathbf{T}} = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_N \end{bmatrix} \tag{3.22}$$

Assuming suitable number of relevant training by *l* among N number of training days for model training, the smallest weight is selected from $\mathbf{W_N}$ as relevant days weight shown in Equation (3.23), where $\mathbf{W_l}$ represents weight of relevant days. Correspondingly, relevant day is determined from the relevant weights $\mathbf{W_l}$.

$$\mathbf{W_l} = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_l \end{bmatrix} \tag{3.23}$$

Similarly, final weight is identified by deducing Equation (3.21) and Equation (3.13) for DTW method and it follows Equation (3.22) and Equation (3.23) for the determination of relevant day. In case of FD method, final weight is identified by deducing Equation (3.21) and Equation (3.14) and for the calculation of relevant day; it follows Equation (3.22) and Equation (3.23).

For the determination of relevant days from HDD, the influence of external temperature T_{ext} is only considered since HDD methods are based on daily average external temperature to estimate the energy load of building. Equation (3.7) along with only influence of external temperature effect from Equation (3.21) are used to calculate the weight of all training days similar to Equation (3.22) and finally *l* number of smallest weights are selected as relevant days.

3.6 Artificial Intelligence Model

3.6.1 Artificial Neural Network

We have used three layered multi-layered perceptron (MLP) neural network (see **Appendix B.1.2** for MLP neural network) since it can be applied for both static and dynamic model. We have used single hidden layer based on the suggestion of Kolmogorov's theorem [105] since single hidden layer is sufficient to approximate any function with given suitable hidden neurons.

Denoting input variables of the featuring database, see **definition 2.8**, (e.g., selected climatic conditions, occupancy, building operating conditions etc.) by the input layer consisting x_i neurons where i varies from 0 to f; the hidden layer consisting z_j neurons where j varies from 1 to g; and output variable of training data (e.g., heating load) by the output layer consisting one signal neuron y; the neural network estimates the output given by:

$$y = \sum_{j=1}^{g} w_j \, f\left(\sum_{i=1}^{f+1} w_{ji} x_i\right) \tag{3.24}$$

where, g is the number of hidden neurons, f represents number of input features including bias thus represented by $f + 1$ and f(.) is the activation function (for detail on activation function, see **Appendix B.1.6**). The w is the weight connecting between each neuron which we are interested to identify since this weight provides regression function and correspondingly prediction of building energy load. We have used tangent hyperbolic activation function in the hidden and output layer.

The activation function is used to provide non-linearity in the estimation function approximated by the neural network. The minimization of training error approximated by neural network is calculated in Equation (3.26).

$$J(w) = \frac{1}{2n} \sum_{k=1}^{n} \left[y^{(k)} - y_a^{(k)}\right]^2 \tag{3.26}$$

Where, $J(w)$, n, y and y_a are training error functions, number of training data, estimation from the neural network (e.g. estimation of heating load of training data) and actual training output (e.g. actual heating load of training data) respectively. The purpose of Equation (3.26) is to provide input-output mapping by adjusting the initial weight to fixed weight to minimize the training error. The training process is done via batch learning (see **definition 2.6**, **Chapter 2**). There are many types of training algorithm used to update the model weight like gradient descent, gradient descent with momentum, Newton's method, etc. [106]. However, they are often slow to train and take more time to compute gradient with second derivatives namely hessian matrix. The algorithms like conjugate gradient, quasi-Newton and levenberg-marquardt provides faster optimization to adjust the weights. We have used levenberg-marquardt algorithm since it is widely used and takes the approximation of hessian matrix in the form of Newton's method which is quite fast and model weight update equation w_{t+1} is given as:

$$w_{t+1} = w_t - [[H^T H + \mu I]^{-1} H^T J(w)] \tag{3.27}$$

In Equation (3.27), hessian matrix is approximated as $[H^T H]$ and gradient is computed as $H^T J(w)$, where H is Jacobian matrix, $J(w)$ is vector of training error function, w_t is initial model weight, μ is suitable chosen scalar and I is identity matrix. Update model weight thus depends on the training error function and scalar value of μ called parameters should define before training. Based on the difference between estimated output by the network and actual training data shown in Equation (3.26), the weights are adjusted and these adjustments are according to the decrease in the training error. If this training error is greater than maximum desired goal (pg) given by Equation (3.28) (where ϑ is constant value parameter to be define by the readers), then this process is repeated until the errors propagating through the neural network are in desired tolerance level. When this error remains at the satisfactory level, the training is stopped and the network holds the constant weight. These constant weights were later used to identify and predict the energy load when the input is presented in the network.

$$pg = \vartheta \sum_{k=1}^{n} y_a^{(k)} \tag{3.28}$$

The other way we stopped the training is by checking the performance on each iteration (epoch). For this, the actual training data is divided into training and validation data (see **definition 2.5**, **Chapter 2**). We defined the stopping criteria based on the performance of validation data so that training can be stopped when the validation error goes on increasing.

Then we addressed the problems of under-fitting, over-fitting and local minima problems (see **remark 2.1-2.2**, **Chapter 2**) of neural network by proposing degree of freedom (DOF) adjustment. DOF of neural network model is the difference between number of training equations and number of model parameters in the network. It should be always $\gg 1$ and depends on the optimum size of hidden neurons.

$$\mathrm{DOF} = \mathrm{N_{Eqn,tr}} - \mathrm{N_{\theta}} \tag{3.29}$$

Where, the number of training Equations ($\mathrm{N_{Eqn,tr}}$) is further given by Equation (3.30). The number of model parameters ($\mathrm{N_{\theta}}$) for a single hidden layer neural network are given by the Equation (3.31).

$$\mathrm{N_{Eqn,tr}} = \mathrm{n\ x\ N_o} \tag{3.30}$$

$$\mathrm{N_{\theta}} = (\mathrm{N_i} + 1)\mathrm{N_h} + (\mathrm{N_h} + 1)\ \mathrm{N_o} \tag{3.31}$$

Where, $\mathrm{N_o}$ is the number of output neuron (e.g. if only heating load, then $\mathrm{N_o}$=1). $\mathrm{N_{\theta}}$, $\mathrm{N_i}$ and $\mathrm{N_h}$ represents number of model parameters, number of input neurons and number of hidden neurons respectively.

The performance goal in Equation (3.28) is adjusted according to degree of freedom. The modified performance goal ($\mathrm{pg_{md}}$) is further given by Equation (3.32).

$$\mathrm{pg_{md}} = \frac{\vartheta\ \mathrm{DOF}\ \sum_{k=1}^{n} y_a^{(k)}}{\mathrm{N_{Eqn,tr}}} \tag{3.32}$$

We also define maximum hidden neuron ($\mathrm{N_{h,max}}$) threshold values to avoid over-fitting and is given by Equation (3.33), where, δ represents the scalar constant value. These threshold values further depend on DOF.

$$\mathrm{N_{h,max}} \cong \frac{1}{\delta}\frac{(\mathrm{N_{\theta}} - \mathrm{N_o})}{(\mathrm{N_i} + \mathrm{N_o} + 1)} \tag{3.33}$$

Finally, in order to select the best parameters of model from validation data, we also split data based on k-fold cross validation and evaluates the performance on k number of validation folds, for detail on parameter selection see **Section 3.6.5**.

3.6.2 Support Vector Machine

There are many implementations of support vector machine (SVM) to build a model (for details on SVM and its available package, see **Appendix B.2**) and we have used LibSVM. SVM are used for classification and regression problems, and support vector regression (SVR) is used as an artificial intelligence model since the building energy consumption prediction is regression problem. Denoting the input variables of the featuring database, see **definition 2.8**, (e.g., selected climatic conditions, occupancy etc.) by x and output variables of featuring database (e.g. heating load) by y, SVR tries to find the hyperplane that maximizes the margin and the equation that separates the hyperplane is given by:

$$f(x) = \mathbf{w}\,\emptyset(x) + b \tag{3.34}$$

where, **w** and b are constant, $\emptyset(.)$ is the mapping function which will be used to map input vector x into higher dimensional called kernel space. Then the SVR finds the hyperplane by satisfying minimization of the quadratic problem to calculate **w** and b [107]:

$$\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{n}(\xi_k + \xi_k^*) \tag{3.35}$$

In Equation (3.35), C is the regularization parameter which determines the degree of training error and controls the trade-off between model complexity and fitting errors, n is the number of training data and ξ_k and ξ_k^* are slack variables which penalize the training error by Vapnik's ε- insensitive loss function (for detail on ε- insensitive loss function, see **Appendix B.2**). The $\frac{1}{2}\|\mathbf{w}\|^2$ term helps to improve the generalization of SVR by regulating the degree of model complexity and $C\sum_{k=1}^{n}(\xi_k + \xi_k{}^*)$ controls the training error.

Equation (3.35) is further transformed into new objective function with the introduction to kernel function and then SVR produces the regression function. The training is performed based via batch learning (see **definition 2.6**, **Chapter 2**). It can be noticed that SVR is sensitive with the choice of kernel functions and parameters C and ε (for detail about kernel function and other parameters, see **Appendix B.2.1-B.2.2**) and these parameters should be selected properly. We have used RBF as a kernel for building consumption prediction model and linear kernel to estimate the weights of climatic coefficient impacts on building load (see **Section 3.5.2**). In order to select the best parameter of SVM, we split the available training data based on k-fold cross

validation and select the parameters of model that has less error performance while averaging k number of validation folds, for details on data-splitting strategies see **Section 3.6.5**.

3.6.3 Boosted Ensemble Decision Tree

There are many techniques for developing decision tree model (for details on decision tree, see **Appendix B.3**) and we have used CART decision tree [108]. However, a decision tree model is itself unstable since it heavily depends on data and small effect of data may have large impact on model performance. In order to address this, ensemble method based on boosting has stabilized effect by averaging [109] and we have used boosted ensemble decision tree (for details on boosting, see **Appendix B.5.2**).

Assuming $\mathcal{L}$ be the leaf node for each input i (e.g., selected climatic conditions, occupancy etc.), the prediction (e.g., heating energy load) is further simplified so that $y_{\mathcal{L}1}\ldots y_{\mathcal{L}n}$ be the prediction of the training data in node $\mathcal{L}$, then the boosting decision tree models estimate the regression output given by Equation (3.36).

$$f(x) = \sum_{m} \beta_m \, y_{\mathcal{L},m} = \sum_{m} \beta_m \left(\frac{1}{n_m} \sum_{i=1}^{n_m} y_{\mathcal{L},i,m} \right) \tag{3.36}$$

where, $\beta_m (m = 1,2, \ldots M)$ are the weight coefficient of given node of tree m. The parameters β_m and number of trees (n_m) are estimated by minimizing error function. Then, we solved the loss function through the optimization problem based on Freidman [110]. For details on loss function minimization, see **Appendix B.5.2**.

3.6.4 Random Forest

We have used Random Forest (RF) proposed by Breiman [111] and its detail is given in **Appendix B.4**. Denoting input featuring database, see **definition 2.8**, (e.g., selected climatic conditions, occupancy etc.) by x with n number of sample size of training dataset T_n, the bootstrap sample is selected randomly from the n observations with replacement from T_n where the probability of each sample drawn is 1/n [111]. The more details on bootstrapping combined with aggregation also called bagging is highlighted in **Appendix B.5.1**. Then bagging method selects the bootstrap samples from the training dataset ($T_n^{b_1}, \ldots\ldots T_n^{b_B}$) where b represents bootstrap and B represents bootstrap size. Then, the CART decision tree algorithm is used to train the model from B number of bootstrap size. While constructing decision trees, we have considered 1/3 of random

features from total number of input features suggested by Breiman [111]. The process of building decision tree is continued until minimum number of leaf node of decision tree and maximum number of bootstrap size has been reached and estimation from each bootstrap is $\left(f\left(x, T_n^{b_1}\right), \ldots., f\left(x, T_n^{b_B}\right)\right)$. Then the output of random forest is obtained by combing the output from each decision tree and is given by:

$$f(x) = \frac{1}{B}\sum_{i=1}^{B} f\left(x, T_n^{b_i}\right) \tag{3.37}$$

3.6.5 Practical Aspects in AI

In AI techniques, there are different tasks to be understood and considered before the model training: normalization of input-output data as a pre-processing step for the data, data splitting strategies for best model parameter selection and the model evaluation for performance measures.

Some machine learning algorithm can have the problems because of bias due to different scales of features. For instance, if the input and output data are not normalized, then there is a chance of some features (e.g., external temperature) to be significant than other features (e.g., solar radiation), thus normalization makes the scaling/range of each variable similar. The more details on widely used normalization techniques are presented in **Appendix B.6.1**. We have used min-max normalization to a fixed range 0 to 1 since most of the "**Relevant Data Selection**" methods block in Figure (3.10) is also in same range of normalization. In addition to this, the neural network model that used non-linear activation functions particularly tangent hyperbolic function is also defined by the threshold values of 1.

Apart from normalizing the input and output data, the best parameters of model should be selected in order to avoid under-fitting or over-fitting (see **remark 2.2**, **Chapter 2**). In case of ANN, the over-fitting problems arise due to improper choice of hidden layers, hidden neurons and size of weights. In addition, length of training data influences the over-fitting of **AI models**. In case of SVM, over-fitting might arises due to the large value of C penalty parameter and the low ε-insensitive loss function. In case of decision tree and random forest, over-fitting might be due to large number of trees. In order to avoid this over-fitting, the general practice is to split the data into training and validation. We have used k-fold cross validation since it divides the training data into k equal parts of validation data for data splitting. Details about widely used data splitting strategies including k-fold cross validation are shown in **Appendix B.6.2**.

The performance of prediction model is evaluated based on coefficient of determination (R^2) and root mean square error (RMSE) shown in Equation (3.38-3.39) where y is the actual energy load, $\bar{y}$ is the mean of actual energy load and $\hat{y}$ is the predicted energy load of each day.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.38}$$

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)^{\frac{1}{2}} \tag{3.39}$$

3.7 Conclusion

This chapter provides methodological framework to predict the building energy load using machine learning artificial intelligence model. It details preparations of offline data for model training. For example, it first discusses the indirect climatic variables generation block. Later, it describes pseudo dynamic model to introduce a priori knowledge on the dynamic behavior of building. It then mentions on building operation classification/clustering to group the building operation according to functioning profile of building. Lastly, it discusses the pre-processing steps of climatic variable selection to select significant direct and derived climatic variables and their dynamics.

After that, it proposes two kinds of modeling approaches: "**all data**" and "**relevant data**" to select input featuring database for model training. It provides detailed depth on different kinds of "**relevant data**" modeling approaches. Finally, it discusses the four machine learning models: Neural Network, Support Vector Machine, Ensemble Boosted Decision Tree and Random Forest as an **AI model** for the two above described modeling approaches.

The next chapter will discuss about the application of methodology for simulated data generation from TRNSys building simulation tools for single-zone and multiple-zone CB to LEBs. In order to apply the methodology, several building parameters should be defined to make prediction of building energy load. For instance, number of past day climate impacts (u) parameter is required to depict time constant of building. In addition to this, initial energy load level (β_0), step size of transition of energy load ($\Delta\beta$) and steady state time constant due to indoor thermal capacity ($T_{steady,air,in}$) parameters are required to generate derived features that provide prior knowledge on the dynamic of the building. Moreover, threshold value (φ_{th}) and number of lags (Φ) parameters are necessary to determine most significant climatic variables.

In case of "**relevant data**" modeling approach, the parameters (C and ε), kernel function and number of cross-validation (k) are required to determine the influence of climatic variables on building using SVM based on linear kernel method. Furthermore, decomposition level (z) is required to transform the climatic conditions into suitable wavelet coefficients which determine the influence of climatic conditions on building. Also, number of relevant days (l is required to define number of days to build a model based on "**relevant data**" modeling approach. Finally, several parameters are required to build **AI models**. In case of ANN, activation function in the hidden and output layer, the parameters of training algorithm: μ and its increment and decrement factors, minimum hidden neurons, maximum hidden neurons set by δ value, parameters to stop training: number of iterations and performance goals defined by ϑ and number of cross-validation (k) for model selection should be defined. Similarly, in SVM, the σ parameter for kernel function, other parameters: C and ε, and number of cross-validation (k) for model selection should be defined. For boosted ensemble decision tree, number of tress (n_m), number of leaf in each tree (β_m) and learning parameter (υ) should be initialized. Consequently, for Random Forest, number of trees in forest (B), bootstrap sample drawn with replacement, number of randomly selected features in each split to grow trees and number of leaf in each tree needs to be initialized.

Then the methodologies follow several steps after available data (e.g., climatic conditions, occupancy, building operating conditions, thermal energy consumption etc.):

Step-1: Building operation classification/clustering
Step-2: Pseudo dynamic model
Step-3: Climatic variables selection
Setp-4: Sets of input features
Step-5: Analysis of climatic variables on the building load (in the "**relevant data**" modeling approach)
Step-6: Selection of sub-database (in the "**relevant data**" modeling approach)
Step-7: Heat load prediction

Chapter 4: Application to Building Simulation

The purpose of this chapter is to apply the methodology using the two modeling approaches: "**all data**" and "**relevant data**" to large buildings. Those large buildings are interesting for ESCOs.

- The case study (building geometry and materials, occupancy profile, building operating conditions etc.) has been done in collaboration with Veolia Research & Innovation (VERI) engineers.
- The heat demand databases are generated using TRNsys.
- Single-zone and multi-zone building models have been introduced to test the methodologies.
- Different kinds of occupancies (residential, office and commercial) have been studied too.

4.1 Buildings Characteristics

4.1.1 Buildings Description

The buildings are based on French standards and details about the CBs to LEBs are summarized in Table (4.1). The Case 1- Case 3 buildings are CBs with single-zone configuration where buildings are considered based on the year of construction. For example, Case-1, Case-2 and Case-3 are based on the U-value of walls for the standard construction of different periods: <1945, 1975-1982 and 1989-2000 respectively. The Case 4 - Case 6 are LEBs with both single and multi-zone configuration. For instance, the Case-4 building volume is divided into three zones where zone-1 consists of the floor level 1-2, zone-2 consists of the levels 3-4 and zone-3 consists of the levels 5-6. Similarly, Case-5 building volumes are divided into three zones where zone-1, zone-2 and zone-3 represents the floor levels 1-3, the levels 4-7 and the levels 8-10 respectively. For Case-6 multi-zone building model, the volumes are divided into two zones: zone-1 (North) and zone-2 (South). It can be further observed from Table (4.1) that the building types are varied according to the U-values for the walls, the roof and the glazing. For instance, the CBs (Case 1 – Case 3) have U-values of the walls, the roof and the floor in the range [0.5-2] in $W/m^2.K$ and U-value of glazing 2.95 $W/m^2.K$ whereas LEBs (Case 4 – Case 6) have U-values of the walls, the roof and the floor of 0.25 $W/m^2.K$ and U-value of glazing in the range [1.43-1.76] $W/m^2.K$. The glazing rates on the external walls are lower in Case 1- Case 4 buildings compare to Case 5- Case

6 buildings revealing the fact that buildings purposes (residential, offices and commercial) are different.

Descriptions	Case 1	Case 2	Case 3	Case 4[1*]	Case 5[1*]	Case 6[1*]
Floor Surface (m^2)	3333	3333	3333	3333	1372	10521
Number of floor	6	6	6	6	10	1
Total surface (m^2)	20000	20000	20000	20000	13720	10521
External wall South (m^2)	4000	4000	4000	4000	4450	330
External wall North (m^2)	4000	4000	4000	4000	4450	330
External wall West (m^2)	1250	1250	1250	1250	-	330
External Wall East (m^2)	1250	1250	1250	1250	-	330
Floor height (m)	3.2	3.2	3.2	3.2	3.2	3.2
U-value of walls, roofs and floors ($W/m^2.K$)	2	1	0.5	0.25	0.25	0.25
U-value of glazing $W/m^2.K$	2.95	2.95	2.95	1.76	1.43	1.43
Glazing rate on each external wall (%)	25	25	25	25	30	30
Building Type	Residential	Residential	Residential	Residential	Office	Commercial
Single/Multi-zone Type	Single	Single	Single	Single/Multi	Single/Multi	Single/Multi

Table 4.1: Description of buildings

*: Multi-zone configurations

The materials composition on the external walls, the roof and the floor for the different building cases is shown in Table (4.2). It can be seen that the external insulation thickness goes on increasing from CBs to LEBs whereas concrete thickness remains same for all kinds of building.

Materials	Case 1			Case 2			Case 3			Case 4 - Case 6		
	Walls	Roof	Floor	Walls	Roof	Floor	Walls	Roof	Floor	Walls	Roof	Floor
Concrete (mm)	200	200	200	200	200	200	200	200	200	200	200	200
Polystrene (mm)	10		10	30			65		65	140		140
Polyurethane (mm)		6			20	30		50			110	

Table 4.2: Description of materials use for buildings

4.1.2 Climatic Conditions

The climatic conditions variables: external temperature (T_{ext}), sky temperature (T_{sky}), horizontal solar radiation ($\emptyset_{Sh}$) and direct solar radiation ($\emptyset_{D}$) for four different climatic locations: Paris, Lille, Lyon and Clermont-Ferrand are generated from Meteonorm software[15]. The summary statistics in terms of minimum, maximum, mean and deviation of climatic variables for four different climatic locations is shown in Table (4.3). It can be seen that external temperature T_{ext} goes to about -8 ^{0}C with deviations of around 7 ^{0}C for different climatic locations. In case of sky

[15] http://www.meteonorm.com/en/

temperature T_{sky}, the minimum and maximum value is around -34 ^{0}C and 28.8 ^{0}C with deviation of around 9 ^{0}C. Similarly, it can also be seen that maximum horizontal solar radiation $\emptyset_{Sh}$ is about 1000 kW/m^2 with a deviation of around 203 kW/m^2 for different climatic locations.

Climatic Variables	Climatic Locations	Summary Statistics			
		Minimum	Maximum	Mean	Deviation
External Temperature (^{0}C)	Paris	-6.8	34.2	11.9	7.1
	Lille	-7.4	31.7	11.0	6.7
	Lyon	-7.2	35.6	12.9	8.0
	Clermont-Ferrand	-9.3	32.9	11.9	7.7
Temperature of Sky (^{0}C)	Paris	-30.1	28.7	3.6	8.7
	Lille	-24.6	25.5	2.8	8.1
	Lyon	-31.3	28.8	3.7	9.2
	Clermont-Ferrand	-34.0	28.2	2.1	9.8
Horizontal Solar Radiation (kW/m^2)	Paris	0	1008	118	190
	Lille	0	956	116	189
	Lyon	0	1021	139	217
	Clermont-Ferrand	0	1015	141	217
Direct Solar Radiation (kW/m^2)	Paris	0	894	48	121
	Lille	0	831	49	126
	Lyon	0	881	67	151
	Clermont-Ferrand	0	876	70	153

Table 4.3: Summary statistics of climatic conditions at different locations

4.1.3 Occupancy Profile

Different kinds of occupancies profiles are considered for different cases. The occupancy profile of single-zone Case 1- Case 4 building is shown in Figure (4.1).

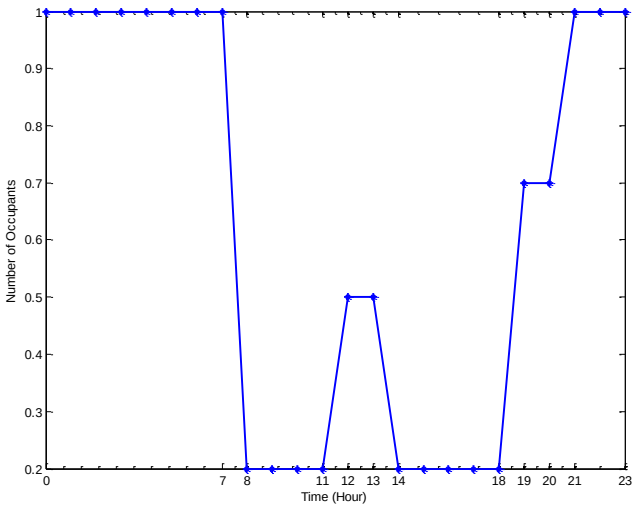

Figure 4.1: Occupancy profile of single-zone Case1 - Case4 building

This profile is similar for all days in a week. It can be seen that building is fully occupied from 21:00 to 7:00 hour and there is a transition in occupancy in the afternoon. This kind of occupancy profile represents similar behavior to residential building profile. For the multi-zone Case-4 building (Case-4*), the occupancy profile is similar to the one shown in Figure (4.1) for three zones (zone 1- zone 3). In both single-zone (Case 1- Case 4) and multi-zone (Case-4*) buildings, the occupancy rate is 0.05 per m^2 and internal gains per occupants are 75 W.

The occupancy profile of single-zone Case-5 building is shown in Figure (4.2) where building is occupied only during Monday to Friday.

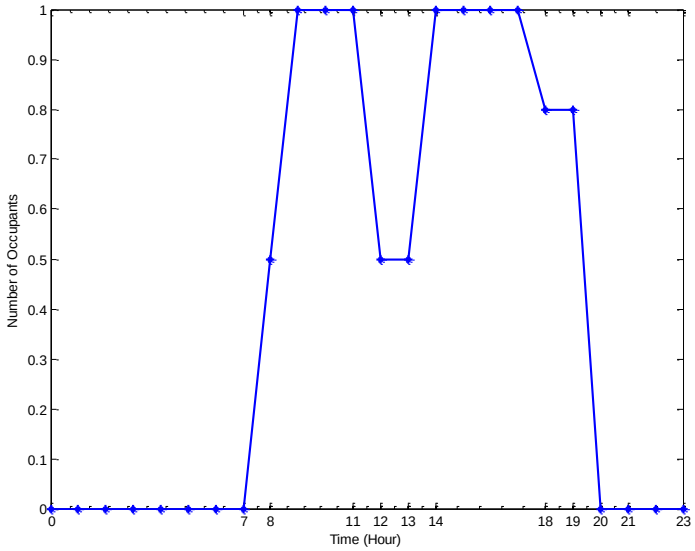

Figure 4.2: Occupancy profile of single-zone Case-5 building

For buildings with such a configuration (Figure 4.2), there is no occupancy during morning and night but there is a transition in occupancy during a day revealing that profile looks similar to the offices building. In case of multi-zone Case-5 building (Case-5*), the occupancy profile is similar to the one shown in Figure (4.2) where zone-1 is occupied during Monday to Friday, zone-2 is occupied during Monday to Saturday and zone-3 is occupied for all days in a week (Monday to Sunday). For this type of building, the occupancy rate is 0.1 per m^2 and internal gains per occupants are 75 W.

Similarly, the occupancy profile of multi-zone Case-6 building (Case-6*) is shown in Figure (4.3) where notation “zone-1” and “zone-2” represents the occupancies in two zones: zone 1 and zone 2 respectively. In zone 1, the building is occupied during Monday-Saturday whereas in the other zone, the building is occupied during the whole week (Monday-Sunday). In case of single-zone

building (Case-6), the occupancy reflects the notation "zone-1" and the building is occupied during Monday-Saturday only. For this type of building, the occupancy rate is 0.2 per m^2 and internal gains per occupants are 75 W.

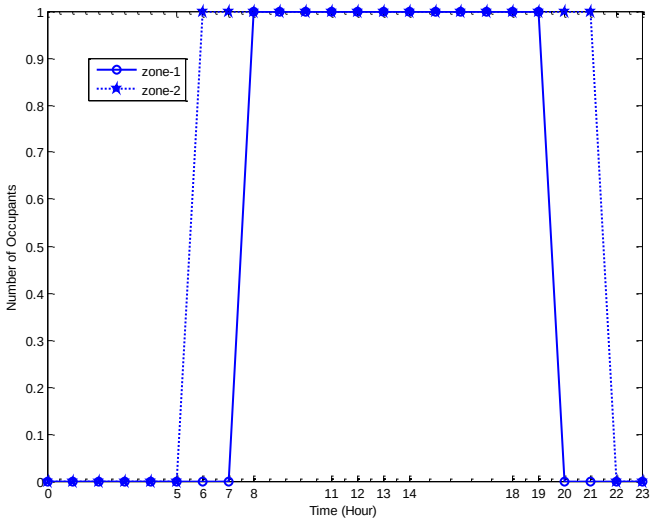

Figure 4.3: Occupancy profile of multi-zone Case-6 building

4.1.4 Building Operating Conditions

The operating conditions (set-point temperature, lighting and ventilation) vary for different types of building. Figure (4.4) shows the operating conditions for Case 1- Case 4 buildings in which the set-point temperature (represented by SP2) only varies. Whereas, the lighting (represented by L1) and the ventilation (represented by V1) are constant along the day at 0.3 W/m^2 and 1 W/m^3 respectively. For single-zone building model, the set-point temperature is represented by "SP1" signifying 21^0C during all days in a week. In case of multi-zone building model, the lighting and the ventilation are similar to the ones shown in Figure (4.4) for the different zones. However, the set-point temperature varies in the different zones, for instance, the set-point temperature is represented by "SP1" (21^0C all hours) and "SP2" (18^0C: 0-5h and 22-23h; 21^0C: 6-21 h) in zone-1 and zone-2 respectively for all days in a week. On the other hand, in other zone "zone-3", the set-point temperature is represented by "SP1" schedule during Monday-Friday and by "SP2" schedule during Saturday-Sunday.

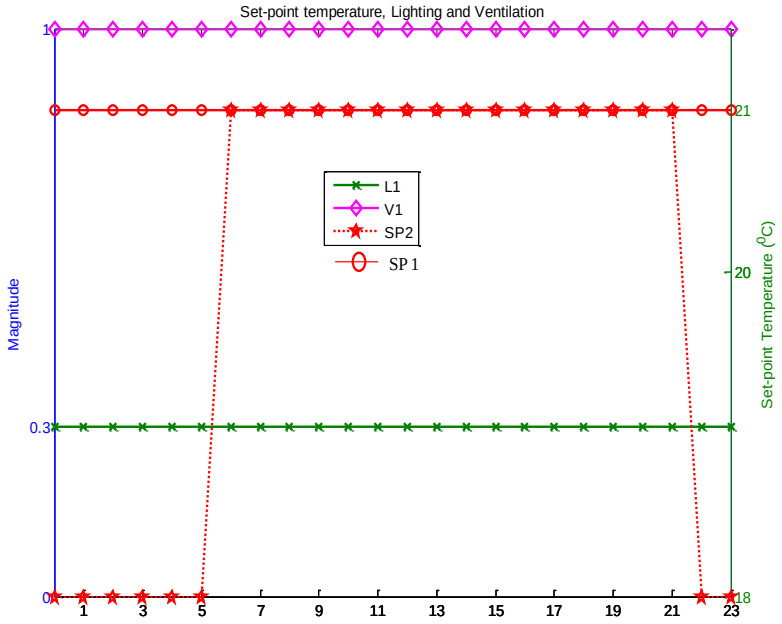

Figure 4.4: Operating conditions of Case 1-Case 4 building

For Case-5 building model, the set-point temperature, the lighting, the ventilation and the appliances profiles are shown in Figure (4.5). In case of single-zone building model, the set-point temperature during Monday-Friday is represented by "SP1" notation (16^0C: 0-5 h and 22-23h; 21^0C: 5-22h) and Saturday-Sunday is represented by "SP2" notation (16^0C: 0-23h). Similarly, the lighting, the ventilation and the appliances profiles during Monday-Friday are represented by "L1" "V1"and "A1" notation respectively whereas "L2", "V2"and "A2" notations are represented for Saturday-Sunday. For multi-zone building model, the set-point temperature "SP1", the lighting "L1", the ventilation "V1"and appliances "A1" are scheduled in zone-1 during Monday-Friday, zone-2 during Monday-Saturday and zone-3 during all days in a week. On the other hand, the set-point temperature "SP2", the lighting "L2", the ventilation "V2" and appliances "A2" are scheduled in zone-1 during "Saturday-Sunday" and zone-2 during Sunday.

The set-point temperature, the lighting and the ventilation for Case-6 building model are shown in Figure (4.6). In case of single-zone building model, the set-point temperature, the lighting and the ventilation are represented by notation "SP1" (16^0C: 0-7h and 20-23h; 21^0C: 8-19h), "L1" (0.05 W/m^2: 0-7h and 20-23h; 1.0 W/m^2: 8-19h) and "V2" (0.5 W/m^3: 0-7h and 20-23h; 1.0 W/m^3: 8-19h) respectively during Monday-Saturday. The set-point temperature in the Sunday is represented by notation "SP2". For multi-zone building model, all the operating conditions of building in one zone (zone-1) are similar to single-zone building model. For other zone, the set-

point temperature, the lighting and the ventilation schedules are represented by notation “SP3”, “L2” and “V2” for all days in a week (Monday-Sunday).

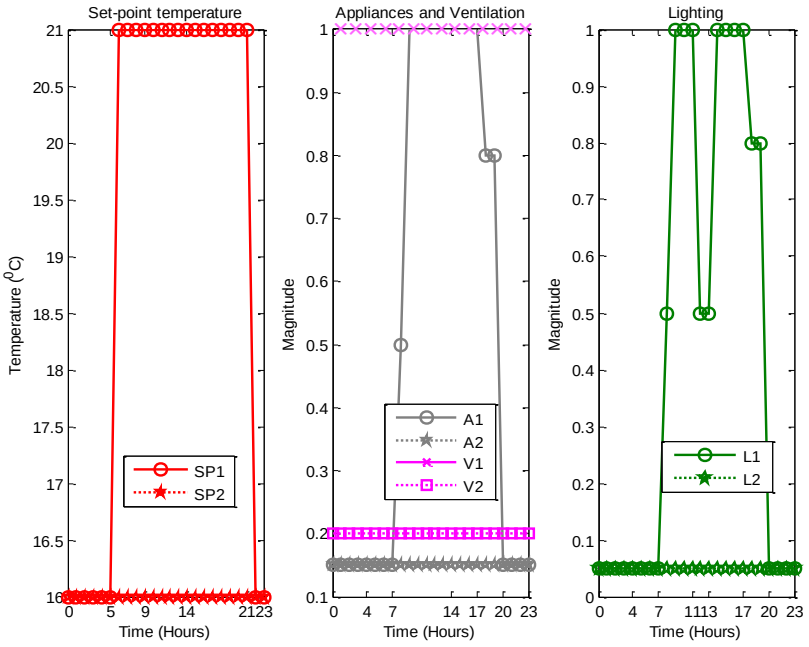

Figure 4.5: Operating conditions of Case-5 building

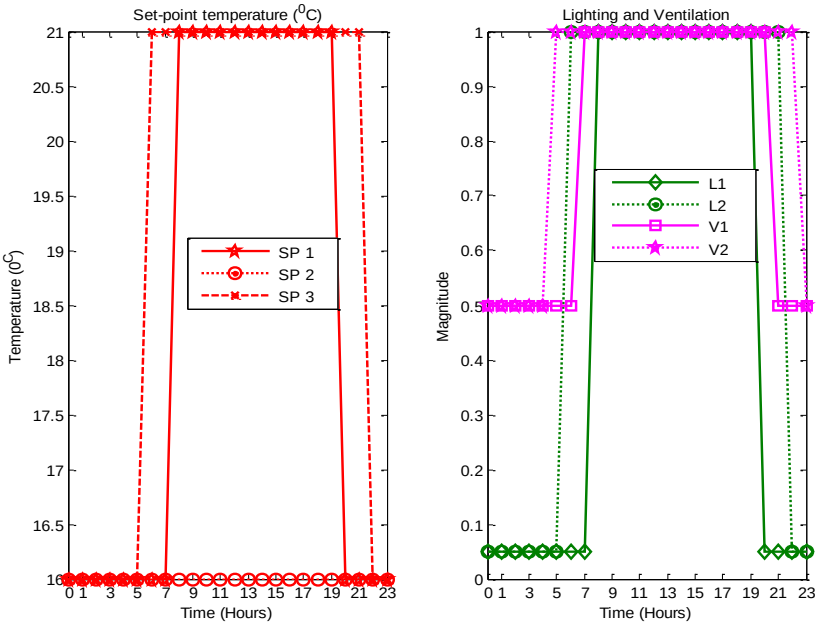

Figure 4.6: Operating conditions of Case-6 building

4.2 Simulation Data Generation

The hourly heating load of building is generated with the building simulation tool (TRNsys). In TRNsys, a single and multi-zone building can be modeled using lumped-capacitance analogy. We have used TRNsys version 17 single zone TYPE 56 model.

The derived climatic variables (the solar gain transmitted through the windows $\emptyset_{Sext}$ and the solar gain on the walls $\emptyset_{Sint}$) for the different locations and the different buildings are obtained from TRNsys and is shown in Figure (4.7-4.8). It can be seen that solar gain transmitted through the windows $\emptyset_{Sext}$ in CBs (Case1 – Case 3) is relatively higher than LEBs for different climatic locations. On the contrary, the solar gain on the walls $\emptyset_{Sint}$ in LEB (Case-4) is relatively higher than CBs (Case1 –Case 3) for different climatic locations but the solar gain on the walls $\emptyset_{Sint}$ in other LEBs (Case 5- Case 6) is lower than CBs. It is more noticeable that the solar gain transmitted through the windows $\emptyset_{Sext}$ and the solar gain on the walls $\emptyset_{Sint}$ in Case-6 building are relatively lower than other types of buildings.

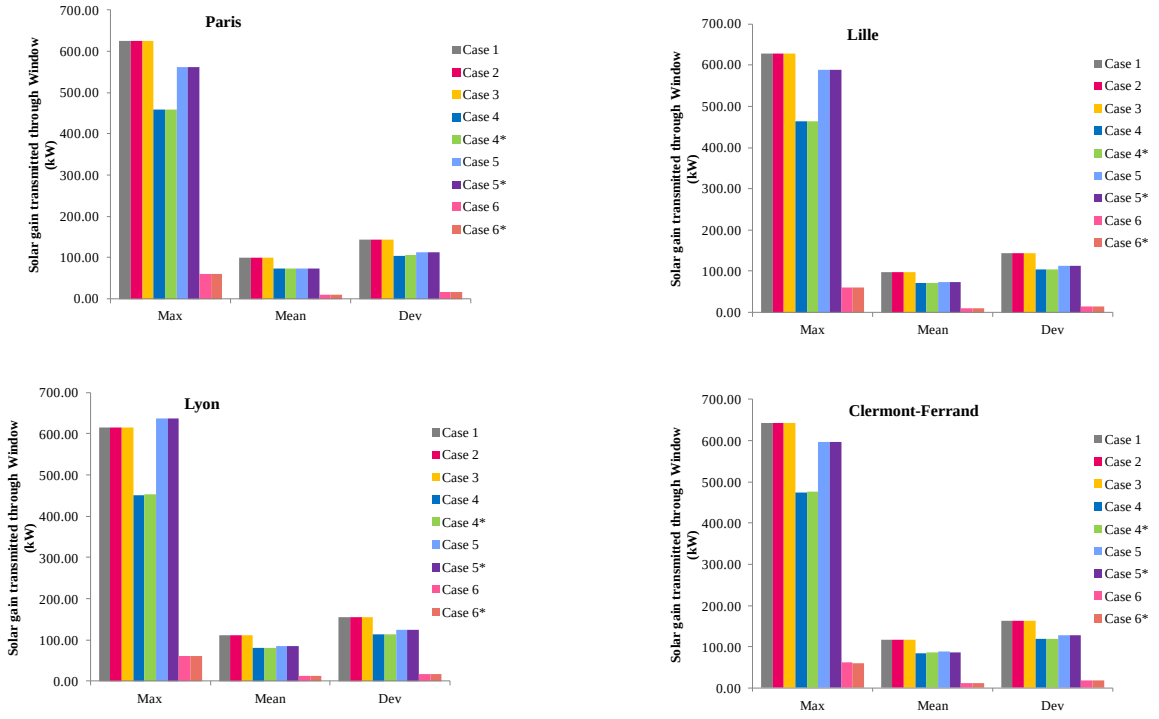

Figure 4.7: Summary statistics of the solar gain transmitted through the windows for four climatic locations (Paris, Lille, Lyon and Clermont-Ferrand)

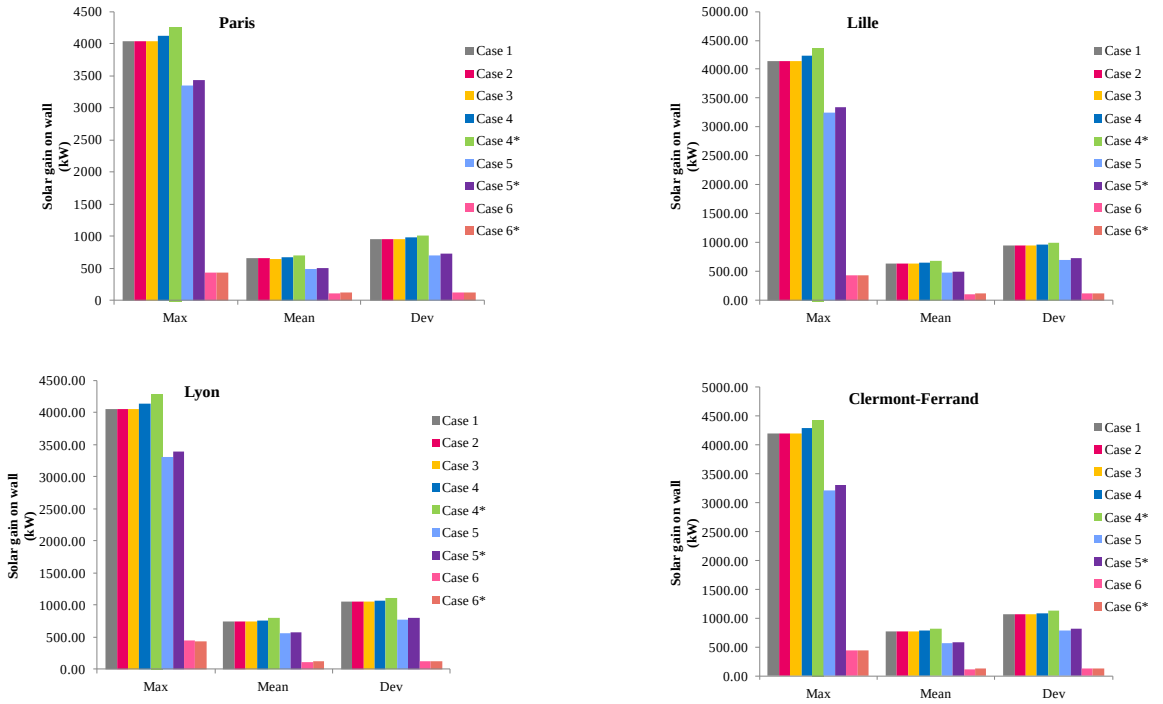

Figure 4.8: Summary statistics of the solar gain on the walls for four climatic locations (Paris, Lille, Lyon and Clermont-Ferrand)

The final annual energy demand for CBs and LEBs is shown in Figure (4.9) and it is noticed that the final energy demand varies according to climatic locations. The final heating energy demand varies from 36 to 82 kWh/m^2.yr for CBs whereas LEBs varies from 21 to 32.8 kWh/m^2.yr.

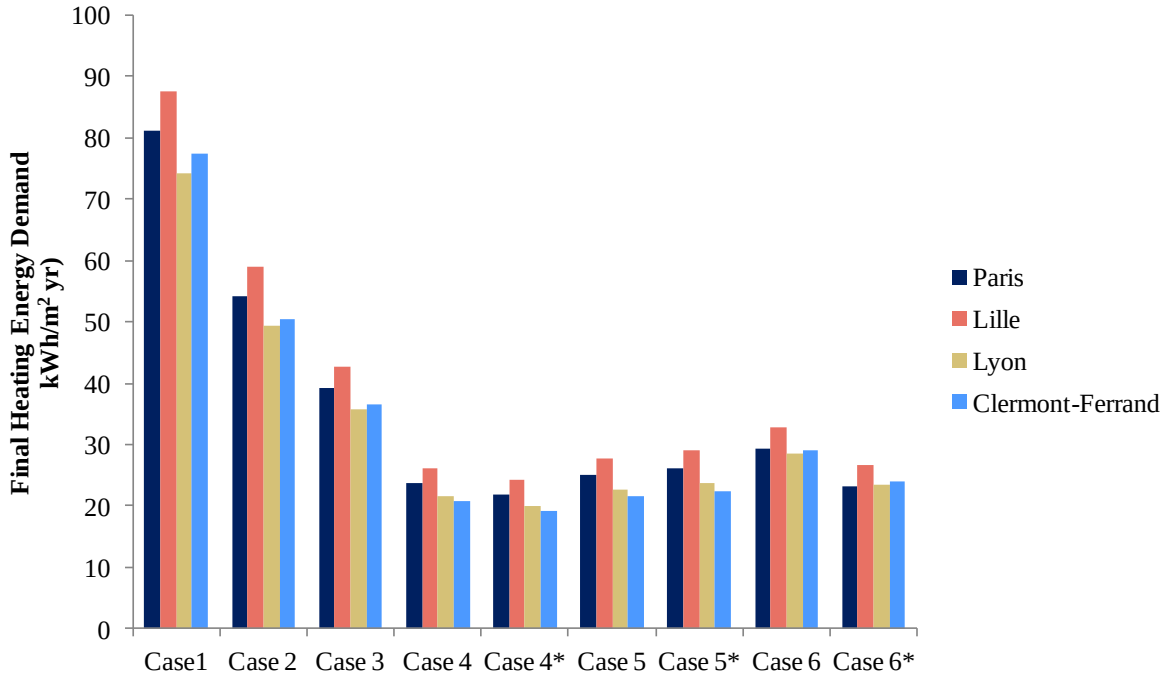

Figure 4.9: Final energy demand for CBs and LEBs

The dynamic response of building, i.e., time constant is evaluated at 20 cm concrete from TRNsys by maintaining constant indoor temperature. It was observed by turning off the heater and the internal gains. By assuming a first order dynamic behavior, this time constant of building models are calculated as shown in Table (4.4).

Building Types	Time Constant (Hours)
Case 1	30
Case 2	53
Case 3	76
Case 4	119
Case 4*	170
Case 5	210
Case 5*	217
Case 6	219
Case 6*	219

Table 4.4: Summary of time constant for different building types

4.3 Application of AI Modeling Methodology for CB to LEB

4.3.1 Introduction

The case study is applied to CBs (Case 1- Case 3) and LEB (Case-4). The shape factor of those buildings is calculated by using Equation (2.3) in **Chapter 2**. This factor is 0.22 for all types of buildings which is quite realistic for LEB. According to specification of LEBs criteria based on shape factor shown in Table (2.1) and Equation (2.1) in **Chapter 2**, the final energy demand is 23.6 kWh/m^2.yr which is quite convenient to the range of Case-4 building shown in Figure (4.9).

The time constant of CBs and LEBs shown in Table (4.4) represents the non-linear dynamics of building. But steady state time of building is sufficient to characterize non-linear dynamics and almost corresponds to 63% of time constant of 30 hours, 53 hours, 76 hours and 119 hours ≈ 1 day, 1.4 day, 2 days and 3 days respectively. However, for the sake of convenience, the number of past day climate impacts u in Equation (3.7, 3.12, 3.20-3.21) in **Chapter 3** corresponds to 1 for Case-1 to Case-2 building, 2 for Case-3 building and 3 for Case-4 building.

4.3.2 Recommendation for Applying the Methodology "Step by Step"

Step 1: Building Operation Classification/Clustering

The classification of building operations in "**Offline Data Preparations**" in Figure (3.4) in **Chapter 3** is shown in Figure (4.10) for Case-4 building model as an example. It can be seen that all days are represented by a single cluster of data from canonical variate (CV) analysis. Therefore, there is only one building operation class (this is similar for all the cases of this Section). Figure (4.11) shows the average heat load profile of each day of a week for Case-4 building. It is clear from Figure (4.10) and Figure (4.11) that the operating conditions of building remains the same during a week (this is similar for all the cases).

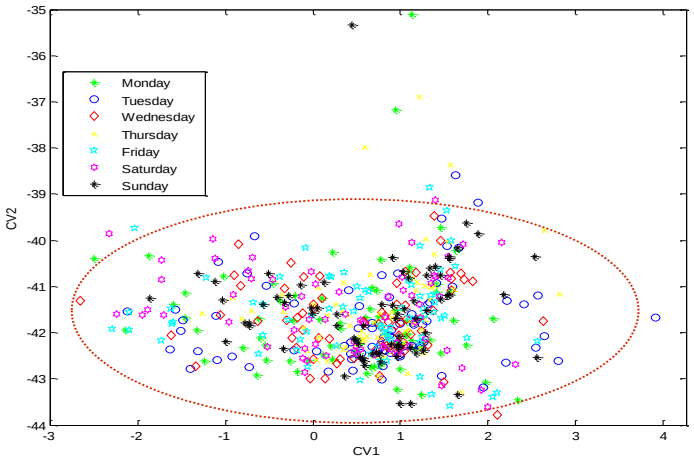

Figure 4.10: Classification of building operation classes (Case-4)

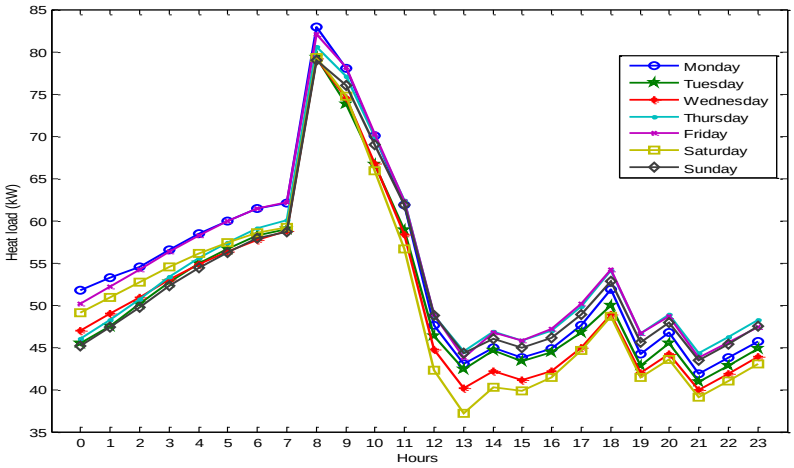

Figure 4.11: Functioning profile of building (Case-4)

Step 2: Pseudo Dynamic Model

The dynamic characteristic to control the indoor temperature of the building[16] (represented by steady state $T_{steady,air,in}$) is around 1-2 hours. It can be noticed that the operating conditions of the building (set-point temperature, ventilation and lighting shown in Figure 4.4) are constant during a day, so this schedule does not contain significant information for an **AI model**. However, the occupancy profile (shown in Figure 4.1) changes at periods 7-8, 11-12, 13-14, 18-19 and 20-21 hour; therefore the PDM directly depends on the these changes period.

A transitional and pseudo dynamic characteristics with 2 lags (due to steady state time) during a day are shown in Figure (4.12) where "Trans" represents transitional characteristics, "PDL-1" represents the pseudo dynamic lag at past 1 hour and "PDL-2" represents the pseudo dynamic lag at past 2 hours.

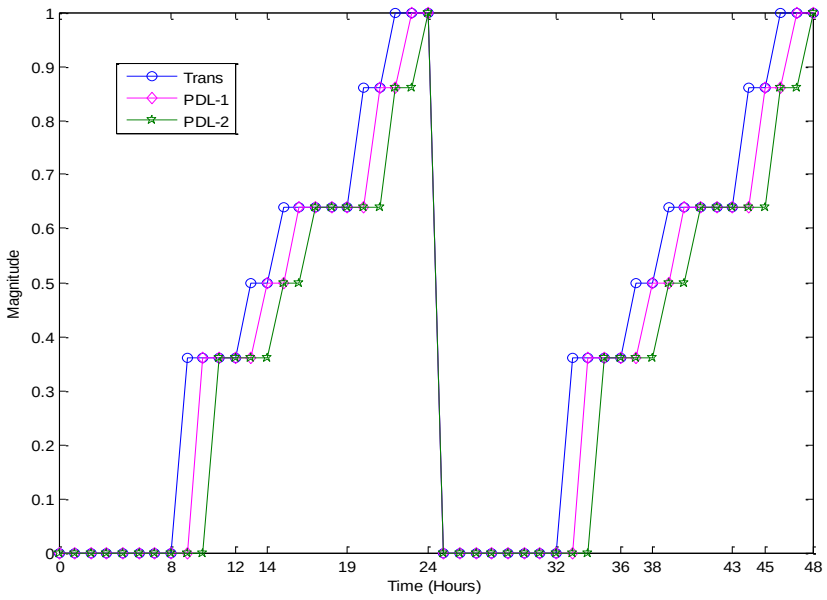

Figure 4.12: Transitional and pseudo dynamic characteristics during two consecutive day

The transitional levels in Figure (4.12) are calculated using the Equation (3.1) in **Chapter 3**. For this study, we assume β_0 to be zero and $\Delta\beta$ with an increment of 0.5. Furthermore, the effects of the transitional and pseudo dynamic effects on heating load can be understood from Figure (4.13) where each transitional level (represented by Trans) and pseudo dynamic lag (represented by PDL) correspond to the changes in heating load from one period to another. It is clear that

[16] The time constant of the indoor air is very different than the time constant of the building envelopes

dynamic behavior characteristics arising from occupants can be illustrated using transition and PDL (Trans and PDL shown in Figure 4.12).

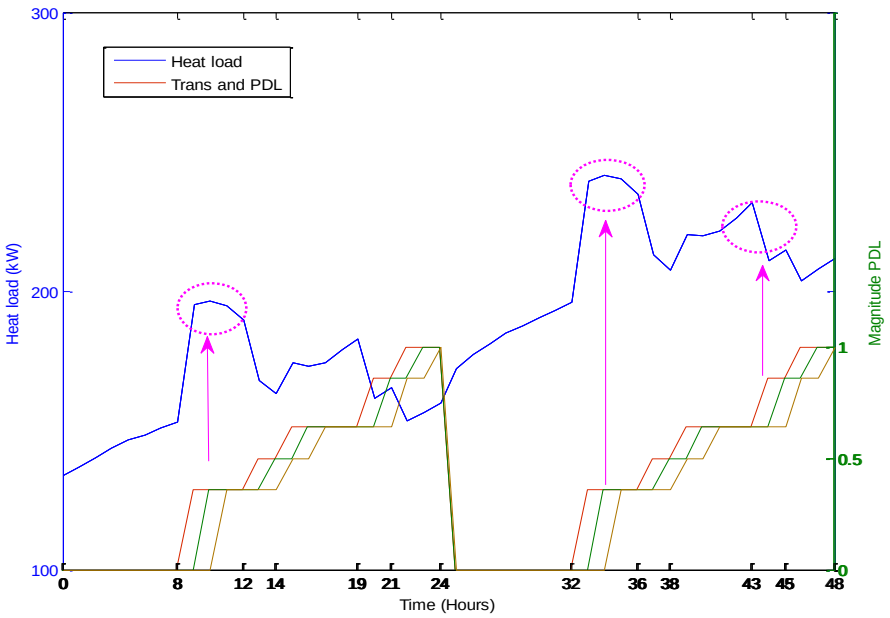

Figure 4.13: Pseudo dynamic transitional effects on heating load during two consecutive days

Step 3: Climatic Variables Selection

Research Question 2: What are the most significant features for different types of building?

The climatic variables considered for relevance determination are:

- Direct Climatic Data:
  - External temperature (T_{ext}), Temperature of sky (T_{sky})
  - Horizontal solar radiation ($\emptyset_{Sh}$), Direct solar radiation ($\emptyset_{D}$)
- Derived Climatic Data (depending on window of number of past day climate impacts *u* e.g., *u* in Case-1 to Case-2: 1 day, Case-3: 2 days and Case-4: 3 days)
  - Solar gain transmitted through windows ($\emptyset_{Sext}$), Solar gain on walls ($\emptyset_{Sint}$)
  - Temporal moving average of external temperature (T_{ext_TDM}), solar gain on walls ($\emptyset_{Sint\_TDM}$)

The correlation indexes (r) of all climatic variables to select important features represented by "**Climatic Variables Selection**" block (**Chapter 3** in Figure 3.4) for different buildings by applying Equation (3.2) is shown in Figure (4.14).

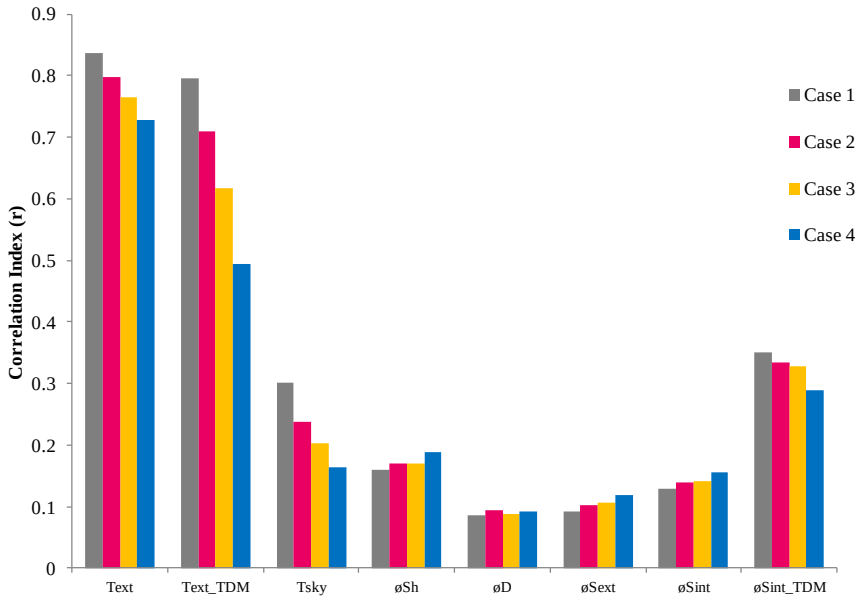

Figure 4.14: Correlation indexes on climatic conditions for CBs to LEB

In all cases, it is observed that the external temperature is more significant than the solar gains. It is also noticed that T_{ext_TDM} and $\emptyset_{Sint\_TDM}$ have higher correlation coefficients which provide further justification of thermal storage impacts from past day. The threshold value (φ_{th}) of 0.07 is chosen to determine the relevance of those variables since most of the climatic conditions have their correlation indexes above 0.07. If the threshold value is greater than 0.10, then their important characteristics especially solar gains (the r for solar gain transmitted through the windows $\emptyset_{Sext}$ is below 0.10 for most of the cases) in determining the heat load are missed. With consideration of φ_{th} with 0.07 value, the climatic condition T_{ext}, T_{sky}, $\emptyset_{Sh}$, $\emptyset_{D}$, $\emptyset_{Sint}$, $\emptyset_{Sext}$, T_{ext_TDM} and $\emptyset_{Sint\_TDM}$ are significant.

However, T_{sky} and T_{ext}, $\emptyset_{Sh}$ and $\emptyset_{D}$ have mutual cross-correlation effects and influences the **black-box** model. Therefore T_{ext} and $\emptyset_{Sh}$ are only selected because of their highest correlation compared to their mutual correlating variables. Thus, it can be concluded that that external temperature T_{ext}, horizontal solar radiation $\emptyset_{Sh}$, solar gain on walls $\emptyset_{Sint}$, solar gains transmitted through windows $\emptyset_{Sext}$, temporal moving average of external temperature T_{ext_TDM} and temporal moving average of solar gain on walls $\emptyset_{Sint\_TDM}$ are significant variables for different types of buildings. Since the climatic conditions T_{ext_TDM} and $\emptyset_{Sint\_TDM}$ are taken into account from T_{ext}

and $\emptyset_{Sint}$, only T_{ext}, $\emptyset_{Sh}$, $\emptyset_{Sint}$ and $\emptyset_{Sext}$ variables are used as a selected weather variables in finding similar patterns so *v* in Equation (3.12) and Equation (3.20- 3.22) in **Chapter 3** represents 4.

The cross-correlation indexes (r_{xy}) are performed by applying Equation (3.3) in **Chapter 3** at lags (Φ) 23 hours that provides time dynamics of selected weather variables. The time dynamics of the external temperature T_{ext} for last 23 hours to represent the thermal storage effects in different cases are shown in Figure (4.15). It can be seen that the cross-correlation indexes r_{xy} for external temperature T_{ext} reach maximum value at past 1-2 hours and decrease the r_{xy} value slowly for all the cases. This further illustrates that the time dynamics of T_{ext} is 1 hour with ± 1 hour deviations for all cases. Similarly, cross-correlation indexes r_{xy} are applied to other climatic variables: $\emptyset_{Sh}$, $\emptyset_{Sext}$ and $\emptyset_{Sint}$. As a result, their time dynamics ranges from 2 hours with ± 1 hour deviations for all cases. Thus, all the selected direct and derived climatic variables (T_{ext}, T_{ext_TDM}, $\emptyset_{Sh}$, $\emptyset_{Sext}$, $\emptyset_{Sint}$ and $\emptyset_{Sint\_TDM}$) and their dynamics ($\emptyset_{Sh}$, $\emptyset_{Sext}$ and $\emptyset_{Sint}$ at past 2 hours; and T_{ext} at past 1 hour) are represented by output of "**Climatic Variables Selection**" block in Figure (3.4) in **Chapter 3**.

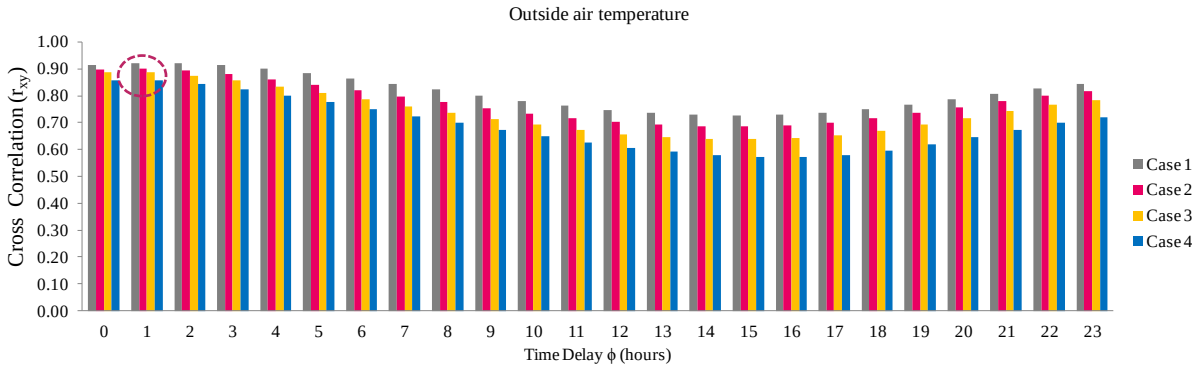

Figure 4.15: Cross-correlation indexes to select external temperature dynamics for CBs to LEB

Step 4: Sets of Input Features

Eight scenarios are studied to understand the physical fundamental of input features (see **definition 2.2**, **Chapter 2**) and summary of selected input and output variables of different scenarios are shown in Table (4.5). It can be observed that scenarios S1 to S3 are used to understand the envelope effects. The scenario S4 includes the behavior of occupancy and the scenarios S5- S7 include the building operating conditions, the transitional attributes information and the pseudo dynamic lag (till past 2 hours). Finally, the scenario S8 take into account all the

input features scenario of S7 and the temporal moving average window (external temperature T_{ext_TDM} and solar gain on walls $\emptyset_{Sint\_TDM}$). It can also be noticed from Table (4.5) that the input and output variables are shown in t where t varies from 1 to 24 hour since the prediction model relies on 1 day ahead (the parameters of **AI model** changes each day) and $\mathcal{M}$ in Equation (3.8-3.9) corresponds to 24.

Name		Description	Input Features Scenarios							
			S1	S2	S3	S4	S5	S6	S7	S8
Outputs	P (t)	Heat Load (kW)	×	×	×	×	×	×	×	×
Inputs	Text(t)	External temperature (^{0}C)	×	×	×	×	×	×	×	×
	Text (t-1)	External temperature at 1 hour time delay (^{0}C)	×	×	×	×	×	×	×	×
	øSh(t)	Horizontal solar radiation (kW)	×	×	×	×	×	×	×	×
	øSh(t-1)	Horizontal solar radiation at 1 hours delay (kW)	×	×	×	×	×	×	×	×
	øSh(t-2)	Horizontal solar radiation at 2 hours delay (kW)			×	×	×	×	×	×
	øSext(t)	Solar gain transmitted through window (kW)		×	×	×	×	×	×	×
	øSext(t-1)	Solar gain transmitted through window at 1 hour delay (kW)		×	×	×	×	×	×	×
	øSext(t-2)	Solar gain transmitted through window at 2 hours delay (kW)			×	×	×	×	×	×
	øSint(t)	Solar gain on wall (kW)		×	×	×	×	×	×	×
	øSint(t-1)	Solar gain on wall at 1 hour delay (kW)		×	×	×	×	×	×	×
	øSint(t-2)	Solar gain on wall at 2 hours delay (kW)			×	×	×	×	×	×
	occup	Occupancy profile [0 1]				×	×	×	×	×
	trans	Transitional attributes [0.2 1]					×	×	×	×
	PDL-1	Pseudo dynamic lag 1 [0.2 1]						×	×	×
	PDL-2	Pseudo dynamic lag 2 [0.2 1]							×	×
	Text_TDM	Temporal moving average of external temperature (^{0}C)								×
	øSint_TDM	Temporal moving average of solar gain on wall (kW)								×

Table 4.5: Summary of input and output variables of different scenarios

Step 5: Analysis of Climatic Variables on the Building Load

Research Question 3: How does the number of past days climatic variables influences the prediction accuracy of energy consumption of buildings?

In order to determine the influences of the past days climatic variables on the daily average heating load, signal analysis is initially performed using a wavelet analysis shown by "**Wavelet Coefficient Calculation of Selected Climatic Variables and their Past Day**" block in Figure (3.10) in **Chapter 3** or in **Section 3.5.2**. For this analysis, we have used Daubechies wavelet and the climatic variables are decomposed at 5 levels, thus the 24-hourly samples are re-sampled into 32 (2^5) samples. For the 32 samples of data in a day, the decomposition coefficients *z* equals to 5 in Equation (3.15-3.18). The decomposed signals of them are expressed by low-frequency and high-frequency coefficients in order to describe the fast and low effects for the heat stored in the

walls. The input-output parameters of an intermediate SVM model based on linear kernel are shown in Table (4.6).

Name	Descriptions
Input (depending on building type)	Wavelet coefficients: $T_{ext}(t)$, $T_{ext}(t-24)$,….. ,$T_{ext}(t-72)$, $ø_{Sh}(t)$,$ø_{Sext}(t)$,$ø_{Sint}(t)$,$ø_{Sint}(t-24)$,...,$ø_{Sint}(t-72)$
Output	Daily average heating load
C	$\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$
ε	{0.001, 0.01, 0.1, 0.2, 0.5}
Kernel function	Linear
Model selection	5-fold cross validation
Normalization	min-max
Datasets	Training and Validation: Lyon-1 year and Clermont-Ferrand-1 year Testing: Lille-1 year

Table 4.6: Parameters of SVM used for weight calculation

From Table (4.6), it is clear that the differences in the influences of climatic variables for the different cases lie to the set of inputs. For instance, Case-1 building has a past number of climatic impacts (u) of 1 day so its relevant input wavelet coefficients are: $T_{ext}(t)$, $T_{ext}(t-24)$, $\emptyset_{Sh}(t)$, $\emptyset_{Sext}(t)$, $\emptyset_{Sint}(t)$ and $\emptyset_{Sint}(t-24)$. Similarly, Case-4 building has a past climatic conditions impacts of 3 days so its relevant input wavelet coefficients are: $T_{ext}(t)$, $T_{ext}(t-24)$, $T_{ext}(t-48)$, $T_{ext}(t-72)$, $\emptyset_{Sh}(t)$, $\emptyset_{Sext}(t)$, $\emptyset_{Sint}(t)$, $\emptyset_{Sint}(t-24)$, $\emptyset_{Sint}(t-48)$ and $\emptyset_{Sint}(t-72)$. The output of this intermediate model that used wavelet decomposition (see **Section 3.5.2**) is the daily average heating load. The normalization is performed in the range of 0 to 1 using min-max normalization and model selections are based on k-fold cross validation where k equals to 5. In order to determine the influence of these wavelet coefficients, Lyon and Clermont-Ferrand wavelet climatic conditions are chosen as training and validation to fit the model, and Lille is used to test the model.

The influence of the number of the past climatic conditions up to past 5 days for different cases based on median[17] and overall RMSE and R^2 is shown in Figure (4.16) for the Case 1 to 4. It can be noticed that model performance is higher (higher median and overall R^2 values or lower median and overall RMSE values) at past day 1 for Case-1 building whereas its performance is

[17] The reason to choose median value is due to its robustness with the changes in performance data and is less affected by outliers compare to mean value

higher (higher: median and overall R^2 values or lower median and overall RMSE values) at past 1-2 days, past 2 days and past 3 days for Case-2, Case-3 and Case-4 respectively.

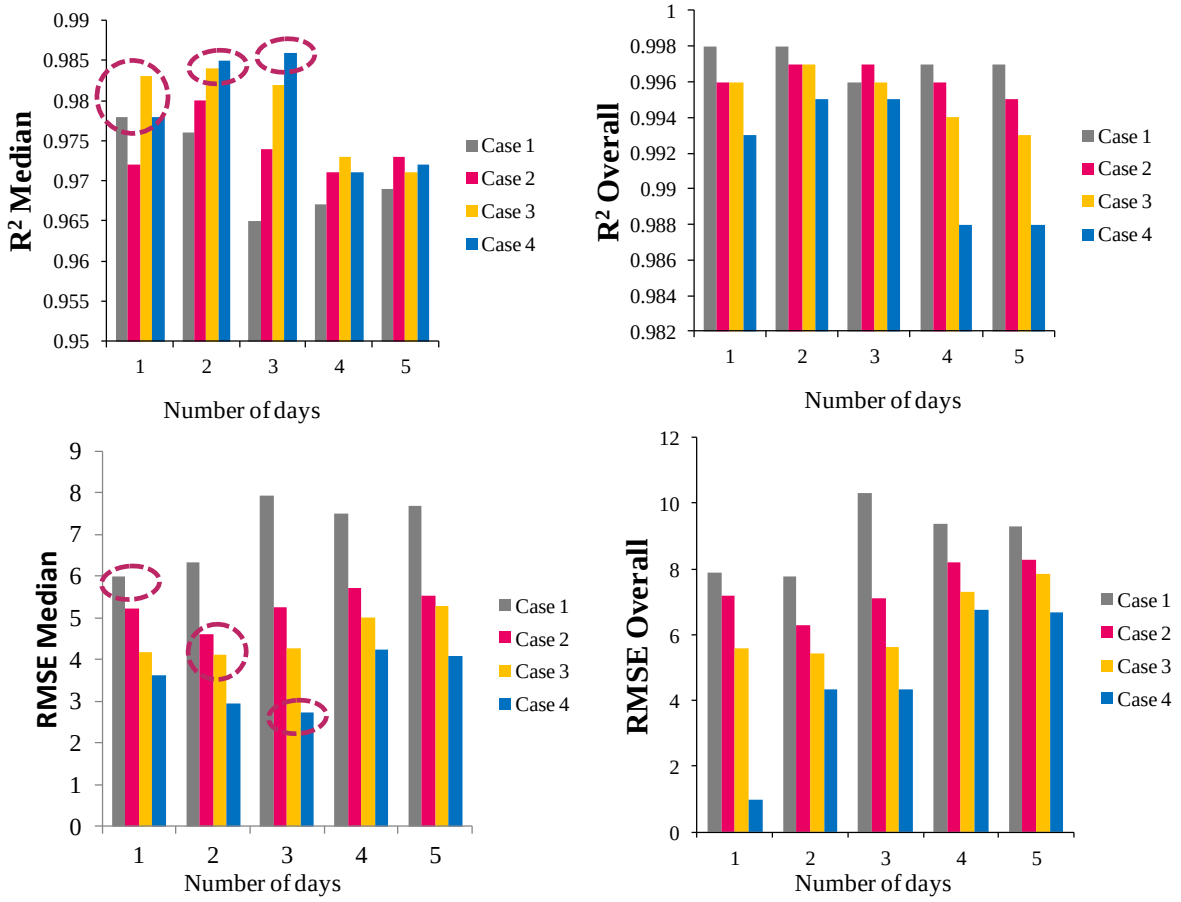

Figure 4.16: Influence of the number of the past climatic conditions days selection on different building cases

The weight factors affecting in selection of external temperature T_{ext} and solar gain on walls $\emptyset_{Sint}$ at past 1-5 days as an example for Case-4 building is shown in Figure (4.17) where pie diagram represents the total share of weights between T_{ext} and $\emptyset_{Sint}$, and bar diagram represents predicted and past day behavior weights of T_{ext} and $\emptyset_{Sint}$. In Figure (4.17), time t corresponds to prediction day, t-24 corresponds to day before prediction day and t-48 corresponds to last two days before prediction day and so on. It is clear from Figure (4.17) that prediction day external temperature T_{ext} is more dominant compared to previous days (t-24, t-48, t-72, t-96 and t-120) for daily average heat load of building. However, for the solar gain on the walls $\emptyset_{Sint}$, the prediction day has less weight compared to previous days (t-24, t-48, t-72, t-96 and t-120) on determining daily average heat load of building. Moreover, it is also seen that with the increasing number of selection days from 3, the weight effect of climatic variables: external temperature T_{ext} and solar

gain on walls $\emptyset_{Sint}$ are almost similar in days 3 to 5 and illustrates that 3 days seem quite significant to determine thermal dynamic response for Case-4 building.

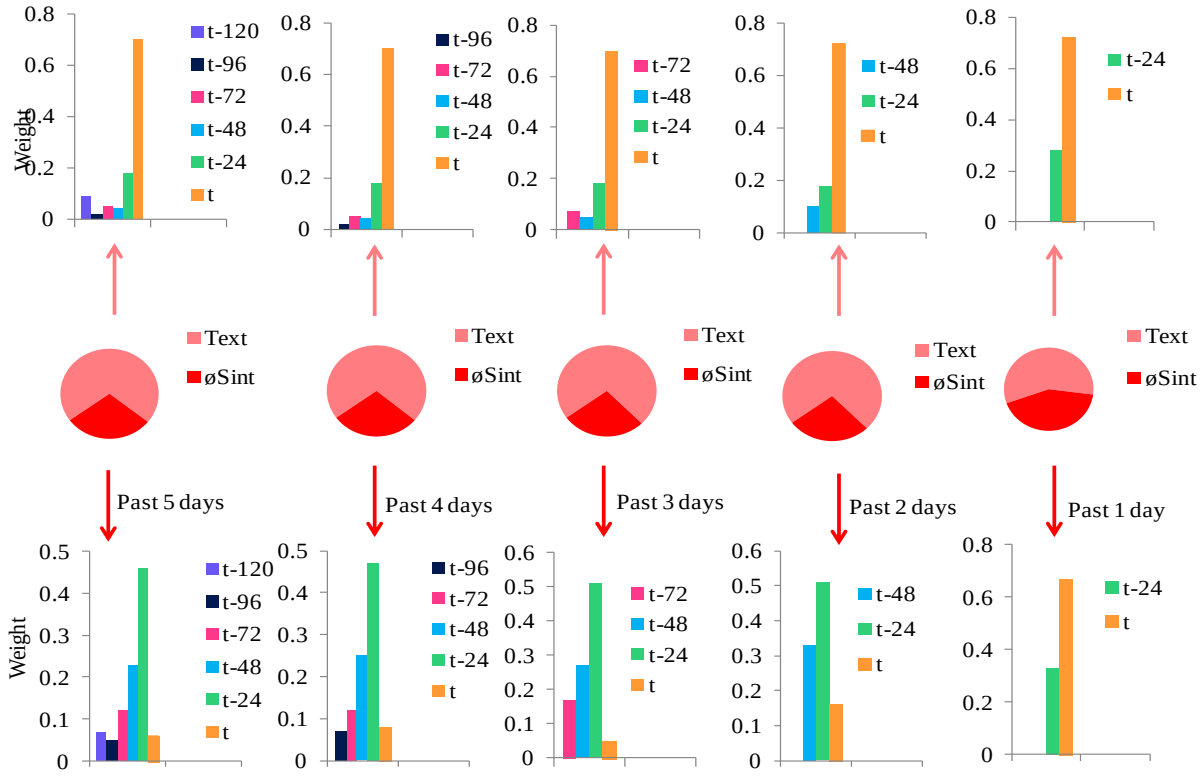

Figure 4.17: Influence of past day external temperature T_{ext} and solar gain on walls $\emptyset_{Sint}$ from prediction day on daily average heating load for Case-4 building

Intermediate Recommendations: The reader can consider the past 3 day's climatic conditions of external temperature T_{ext} and solar gain on walls $\emptyset_{Sint}$ for LEB. In case of CB, the past 1-2 days of these climatic conditions are significant. Therefore, number of past day climatic impacts *u* for CB and LEB are 1-2 and 3 respectively.

Comparison between SVM and Least Square Method (LSM)

The comparison between **SVM based on linear kernel** and a **LSM based on regression model** is performed by fitting the wavelet coefficients features shown in Table (4.6). The performance accuracy based on SVM and LSM is shown in Figure (4.18). It is shown that **SVM based on linear kernel** is better than **LSM based on regression model** due to its lower RMSE for all the given cases. For instance, the performance of **SVM based on linear kernel** is higher (lower RMSE=142) compared to **LSM based on regression model** (higher RMSE=184) for Case-4.

Therefore, this reveals that SVM based on linear kernel is better than **LSM based on regression model** to determine the influence of climatic variables on the building load.

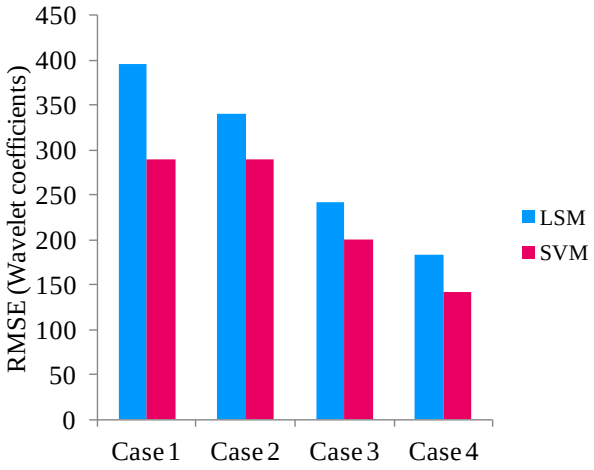

Figure 4.18: Performance while fitting wavelet coefficients using LSM based on regression and SVM based on linear kernel

The influence of climatic conditions weight using **SVM based on linear kernel** and **LSM based on regression model** is shown in Figure (4.19).

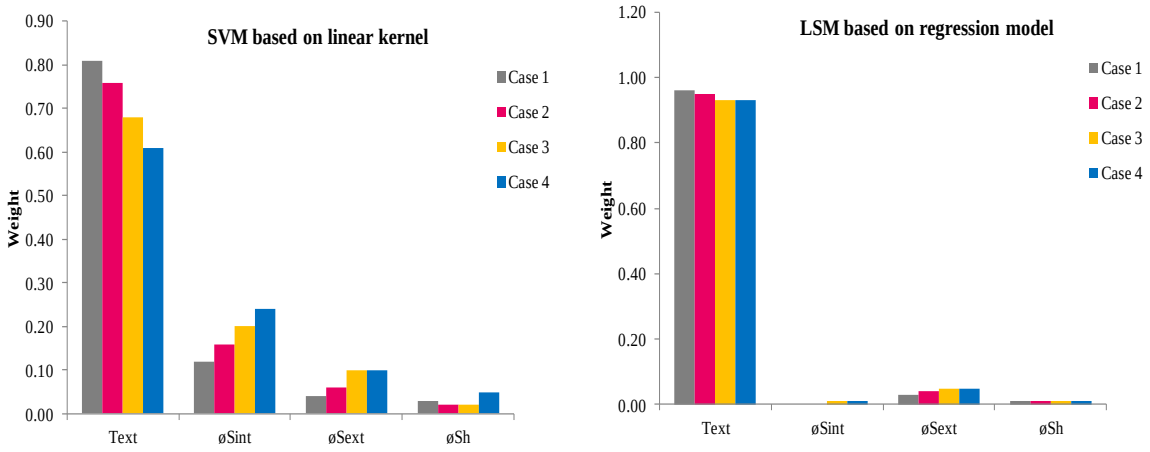

Figure 4.19: Influence of climatic conditions on different buildings using SVM based on linear kernel and LSM based on regression

As an example, with **SVM based on linear kernel** it is noticed that the impact of external temperature T_{ext} goes on decreasing and solar gains go on increasing while the building is migrating from CBs (Case-1 to Case-3) to LEB (Case-4). On the other hand, the influence of climatic conditions on daily average building load is different in **LSM based on regression model**. For instance, while using **LSM based on regression model**, it is observed that all the

buildings are dominated by external temperature T_{ext}, for example, in Case-4 the external temperature T_{ext} is 93% dominant on heating load which is followed by the solar gain transmitted through the windows $\emptyset_{Sext}$ (5%), the solar gain on the walls $\emptyset_{Sint}$ (1%) and the horizontal solar radiation $\emptyset_{Sh}$ (1%). Furthermore, LSM results reveal that influence of building load is highly dominated by T_{ext} and further signifies that $\emptyset_{Sint}$ and $\emptyset_{Sext}$ has less influences on heating load.

Intermediate Recommendations: The comparison study suggests using SVM based on linear kernel rather than LSM based on regression to determine the influence of climatic variables on LEB.

The individual prediction and past day behaviors of climatic conditions for different cases using SVM based on linear kernel is shown in Figure (4.20).

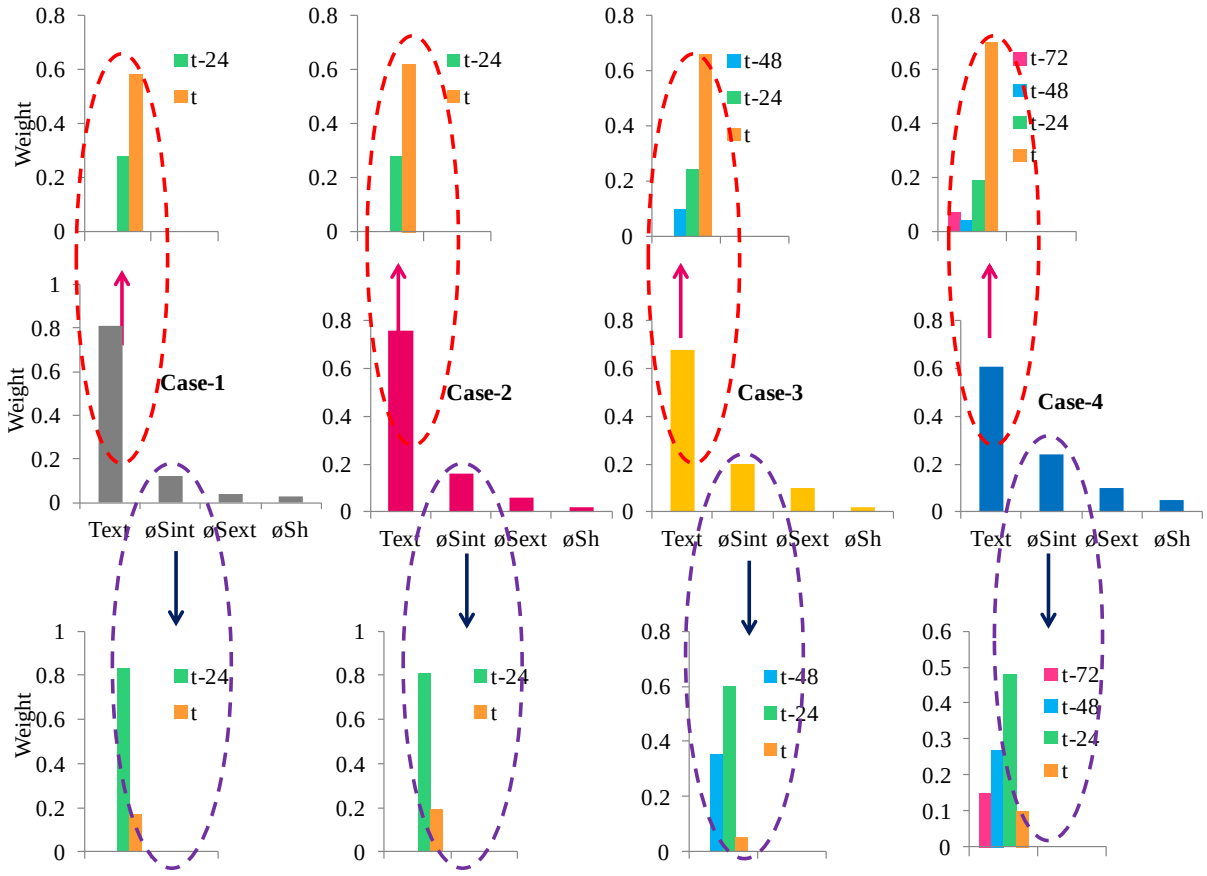

Figure 4.20: Individual weight distribution of prediction and past day climatic conditions for different building types (using SVM based on linear kernel)

As an example, it is observed that in Case-1 building, the external temperature T_{ext} is more dominant (81%) which is followed by the solar gain on the walls $\emptyset_{Sint}$ (12%), the solar gain

transmitted through the windows $\emptyset_{Sext}$ (4%) and the horizontal solar radiation $\emptyset_{Sh}$ (3%) whereas prediction day (corresponds to time t) the external temperature T_{ext} is more dominant (58%) compared to the previous days external temperature t-24 (42%) in determining heating load. In case of solar gain on the walls $\emptyset_{Sint}$, the prediction day has less weight (17%) whereas the previous day from prediction day at time t-24 has highest weight (83%). On the other hand, in Case-4 building, the external temperature T_{ext} is more dominant (61%) which is followed by the solar gain on the walls $\emptyset_{Sint}$ (23%), the solar gain transmitted through the windows $\emptyset_{Sext}$ (12%) and the horizontal solar radiation $\emptyset_{Sh}$ (4%) whereas the prediction day (corresponds to time t) external temperature T_{ext} is more dominant (70%) compared to the previous days external temperature (t-24, t-48 and t-72) in determining the heating load. In case of the solar gain on the walls $\emptyset_{Sint}$ for such LEB (Case-4), the prediction day has less weight (4%) whereas the previous day from the prediction day at time t-24 has highest weight (52%).

The summary of individual normalized weight matrix represented in Figure (4.20) obtained from Equation (3.21) in **Chapter 3** is given below for different types of buildings where coeff_T_{ext}, coeff_$\emptyset_{Sh}$, coeff_$\emptyset_{Sext}$ and coeff_$\emptyset_{Sint}$ represents weight vector matrix of T_{ext}, $\emptyset_{Sh}$, $\emptyset_{Sext}$ and $\emptyset_{Sint}$ respectively. The right hand side of first scalar contains the influence of climatic variables T_{ext}, $\emptyset_{Sh}$, $\emptyset_{Sext}$ and $\emptyset_{Sint}$ on building load. Right hand side of second matrix contains the influence of respective climatic variables of prediction day, last day before prediction, last two days before prediction day and last three days before prediction day.

$$\begin{bmatrix} coeff_{T_{ext}} \\ coeff_{\emptyset_{Sint}} \\ coeff_{\emptyset_{Sext}} \\ coeff_{\emptyset_{Sh}} \end{bmatrix}_{Case-1} = \begin{pmatrix} 0.81 \\ 0.12 \\ 0.04 \\ 0.03 \end{pmatrix} * \begin{bmatrix} 0.58 & 0.42 & 0 & 0 \\ 0.17 & 0.83 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} coeff_{T_{ext}} \\ coeff_{\emptyset_{Sint}} \\ coeff_{\emptyset_{Sext}} \\ coeff_{\emptyset_{Sh}} \end{bmatrix}_{Case-2} = \begin{pmatrix} 0.76 \\ 0.16 \\ 0.06 \\ 0.02 \end{pmatrix} * \begin{bmatrix} 0.62 & 0.38 & 0 & 0 \\ 0.19 & 0.81 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} coeff_{T_{ext}} \\ coeff_{\emptyset_{Sint}} \\ coeff_{\emptyset_{Sext}} \\ coeff_{\emptyset_{Sh}} \end{bmatrix}_{Case-3} = \begin{pmatrix} 0.68 \\ 0.20 \\ 0.10 \\ 0.02 \end{pmatrix} * \begin{bmatrix} 0.66 & 0.24 & 0.10 & 0 \\ 0.05 & 0.60 & 0.35 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \text{coeff}_{T_{ext}} \\ \text{coeff}_{\emptyset_{Sint}} \\ \text{coeff}_{\emptyset_{Sext}} \\ \text{coeff}_{\emptyset_{Sh}} \end{bmatrix}_{Case-4} = \begin{pmatrix} 0.61 \\ 0.24 \\ 0.10 \\ 0.05 \end{pmatrix} * \begin{bmatrix} 0.70 & 0.19 & 0.04 & 0.07 \\ 0.10 & 0.48 & 0.27 & 0.15 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Step 6: Selection of the Sub-Database

The selection of the sub-database are based on the influence of climatic variable weights obtained from step-5 and after the identification of similar climatic variables shown by "**Identification of Similar Climatic Conditions**" block in Figure (3.10) in **Chapter 3**. The HDD is calculated for prediction day and its number of past day climate impacts *u* (e.g., *u* in Case-1 to Case-2: 1 day, Case-3: 2 days and Case-4: 3 days) using Equation (3.7) in **Chapter 3**. Similarly, modified HDD (mHDD) similarity weights of external temperature T_{ext} and solar gain on walls $\emptyset_{Sint}$ are determined using prediction day and its number of past day climate impacts using Equation (3.12) in **Chapter 3**. In addition, similarity weight of horizontal solar radiation $\emptyset_{Sh}$ and solar gain transmitted through windows $\emptyset_{Sext}$ are calculated only for prediction day using Equation (3.12) in **Chapter 3**. Furthermore, similarity weights of T_{ext} and $\emptyset_{Sint}$ are determined by comparing prediction day and its number of past day climate impacts *u* with the training database based on DTW using Equation (3.13) in **Chapter 3** and based on FD using Equation (3.14) in **Chapter 3**. However, similarity weight of $\emptyset_{Sh}$ and $\emptyset_{Sext}$ are determined only by comparing prediction day with training day climatic behavior for both methods based on FD and DTW.

Finally, the final weights of all training days are calculated based on Equation (3.22) and then the 12 relevant days (*l* in Equation 3.23 in **Chapter 3** corresponds to 12) sub-databases are selected. The sensitivity analysis on number of relevant days with the prediction performance is carried in step-7.

Step 7: Heating Load Prediction

Let us remember that the TRNsys results have been generated using a single-zone model for the description of this step.

Initially, the model is evaluated using ANN based on DTW and then studied comparison of different **AI models** and "**relevant data**" modeling approaches. The different input features scenarios are considered for the analysis of ANN model based on Table (4.5). For each of the input features scenarios, the cost function in Equation (3.26) in **Chapter 3** is calculated by iterating up to 1000 times for each of the minimum and maximum hidden neurons. The maximum

hidden neurons is calculated using Equation (3.33) in **Chapter 3**, where δ is chosen 8 as it gives the flexibility in the degree of model parameters and minimum hidden neuron is chosen 1. The model parameters are updated based on Equation (3.27) in **Chapter 3** where training parameters are chosen to converge slowly due to the use of faster Levenberg-Marquardt training algorithm. In order to converge slowly, we chose relatively larger value of μ to be 1 where its value is increased with a factor of 1.5 and decreased with a factor of 0.8. The neural network model training is stopped when the iterations reached to 1000 and performance goal reached to the value given by the Equation (3.32) in **Chapter 3** where ϑ corresponds to 0.01. The summary of parameters of ANN model is shown in Table (4.7). The performance of different input scenarios are evaluated considering 1 year test data at Paris location for different cases and their performances are shown in Table (4.8).

Name	Descriptions
Input and output of model	S1 to S8 shown in Table (4.5)
Activation function	Hyperbolic tangent (hidden and output layer)
Hidden neurons	1 to maximum define in Equation (3.33)
Training algorithm	Levenberg-Marquardt
Stopping criteria	Number of iteration:1000 and performance goal define in Equation (3.32)
Model selection	5-fold cross validation
Normalization	min-max
Datasets	Training and Validation: Lyon-1 year, Clermont-Ferrand-1 year and Lille-1 year Testing: Paris-1 year

Table 4.7: Summary of ANN parameters

Models	Case 1				Case 2				Case 3				Case 4			
	Median		Overall		Median		Overall		Median		Overall		Median		Overall	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
S1	0.74	20.5	0.98	23.4	0.71	17.7	0.96	23.4	0.67	16.2	0.96	18.3	0.69	13.2	0.93	16.6
S2	0.75	20.1	0.98	23.0	0.69	17.5	0.97	20.5	0.69	16.1	0.96	18.1	0.68	14.1	0.94	15.5
S3	0.77	19.6	0.98	22.8	0.69	17.3	0.97	20.7	0.70	15.9	0.96	18.1	0.70	13.6	0.94	15.5
S4	0.88	14.7	0.99	18.5	0.88	12.3	0.98	16.2	0.90	10.4	0.98	13.6	0.93	7.2	0.97	9.7
S5	0.89	14.2	0.99	17.8	0.88	12.2	0.98	16.6	0.91	10.1	0.98	13	0.94	7.3	0.98	9.3
S6	0.88	14.5	0.99	19.5	0.86	13.1	0.98	17.2	0.90	10.1	0.98	13.4	0.93	7.2	0.97	10.6
S7	0.89	14.1	0.99	17.8	0.88	12.7	0.98	16.2	0.91	10.0	0.98	13.6	0.93	6.9	0.98	8.9
S8	0.93	10.9	0.99	14.1	0.92	9.5	0.99	13.6	0.96	6.0	0.99	8.5	0.97	3.9	0.99	6.0

Table 4.8: Comparison of different input features scenarios for different cases using DTW relevant data modeling approach based on ANN

The performances of different input features scenarios are evaluated based on median[18] and overall[19] values. It can be seen that the input feature scenario S1 that relies on the external temperature T_{ext} and the horizontal solar radiation $\emptyset_{Sh}$ is not fully sufficient to learn the behavior of heat load. It is noticed that CB (Case-1) that is more temperature dependent has higher performance (Median R^2=0.74; Overall R^2=0.98) compared to LEB (Case-4) that depends on solar gain (Median R^2=0.69; Overall R^2=0.93) which is the fact that the external temperature features are not sufficient to characterize for Case-4. It can be seen that median values performance gave better comparison than overall performance due to the evaluation of model performance each day. It is also observed that the input feature scenario S3 that relies on climatic conditions (external temperature and solar gain) including delay storage of the climatic conditions has more stable results in comparison to the input feature scenarios S1-S2 for all the cases.

In addition, it can be observed that occupancy profile has a major impact for all the cases shown by the input feature scenario S4 and their performance has been increased compared to scenario S3. For instance, the input feature scenario S4 has better performance (Median: R^2=0.88, RMSE=14.7; Overall: R^2=0.99, RMSE=18.5) compared to the input feature scenario S3 that depends only on climatic conditions (Median: R^2=0.77, RMSE=19.6; Overall: R^2=0.98, RMSE=22.8) for Case-1 building. With the introduction of transitional and pseudo dynamic lag, the performance has been slightly increased as well. It is clear that PDM with 2 hours lag (the input feature scenario S7) is sufficient to characterize the dynamics of indoor air rather than only 1 hour lag (the input feature scenario S6) and the best simulation result is obtained from the scenario S7 while comparing scenarios S1-S6 for all the cases. Furthermore, the most interesting result is given when we consider the temporal moving average behavior in scenario S8 compared to the best simulation results (scenario S7). Thus, it can be concluded that though scenario S7 has more consistent results compared to scenarios S1-S6, scenario S8 is even better due to the moving average of thermal storage effects behaviors.

Figure (4.21) shows the performance of scenarios S7 and S8 for different heating months (January –April and October- December) for different types of buildings (Case 1-4). It can be observed that the performance of scenario S8 is higher compared to scenario S7 for different heating months (January – April and October – December) for all the cases due to the inclusion of temporal moving average behavior of T_{ext} and $\emptyset_{Sint}$. But during April, both scenarios S7 and S8 have less performance compared to other months. This is due to the intermediate season where heating

[18] Evaluates the performance using median values of each prediction day from all testing condition of 1 year
[19] Evaluates the performance from all testing condition of 1 year

loads are mostly dominated by T_{ext}. During this intermediate season, heating demand is more or less similar in range of low peak demand thus it is not essential for ESCOs to have a good heating demand prediction. Therefore, Figure (4.21) illustrates that scenario S8 is better for all types of building (Case 1- Case 4) to predict heating load.

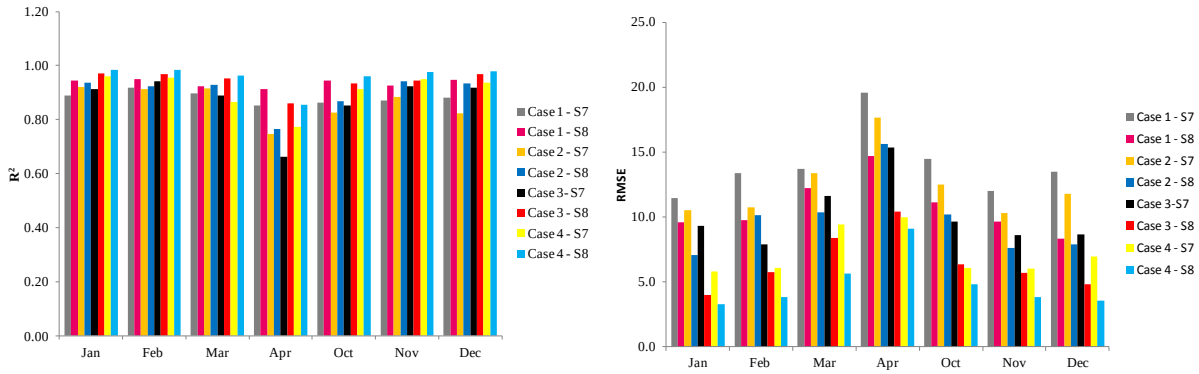

Figure 4.21: Performance of scenarios S7 and S8 for different heating months

As an example, Figure (4.22-4.23) shows the heating load prediction from scenarios S7 and S8 for some random days in months (January- March) for Case-3 and Case-4 building.

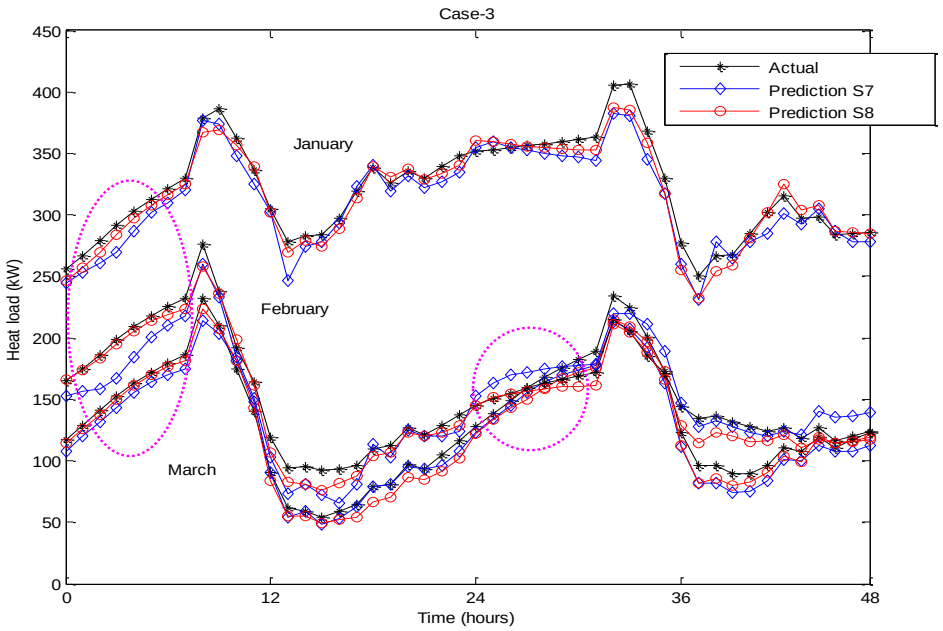

Figure 4.22: Prediction of heating load from input scenarios S7 and S8 for some random days in different months for Case-3 building

It can be noticed that for both types of buildings, most of the errors are accumulated during first hours of heating period (0-7) hour in morning, in particular to scenario

S8. This is because though the input feature scenario S7 considers building dynamics behavior, it does not include small transition of thermal energy storage in walls from past dynamics of climates: T_{ext} and $\emptyset_{Sint}$ which are the dominant variables for heating load. In contrast, the input feature scenario S8 comprises transition of thermal energy storage in wall from T_{ext} and $\emptyset_{Sint}$ by introducing temporal moving average window at past 3 days for Case-4 and at past 2 days for Case-3 building.

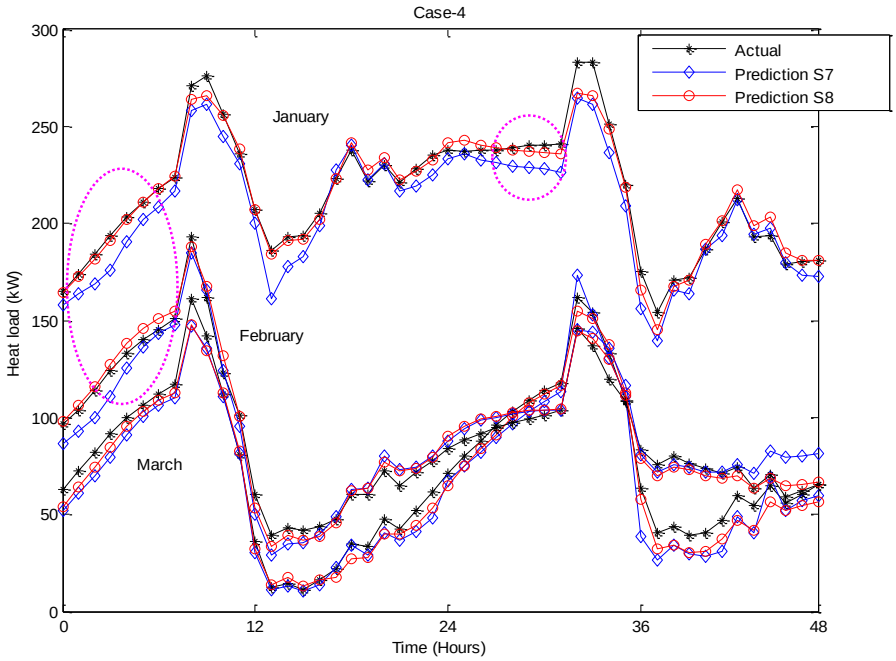

Figure 4.23: Prediction of heating load from input scenarios S7 and S8 for some random days in different months of Case-4 building

Intermediate Recommendations: The readers are suggested to use the input feature scenario S8 as a reference for all the cases at later use for the given building.

Sensibility Study: Influence of the Number of Relevant Days in the Prediction Performance

Research Question 4: How does number of data used for model training influences the performance of machine learning AI model?

Generally, the number of days used for model training depends on ten times the number of features (see **remark 2.6** in **Chapter 2**). The numbers of features are around 4-17 for different scenarios. Thus, number of training data should be around 170 hours equivalent to 7 days.

Considering this fact to determine stable and reliable performance values, the model is evaluated by varying the number of training day data from 5 days to 45 days shown in Figure (4.24). The performance of the model is evaluated by considering different test days. The median of their normalized RMSE performance is shown in Figure (4.24). It can be observed that the performance of the model decreases (RMSE increases) when the number of days is lower than 7 days and greater than 14 days.

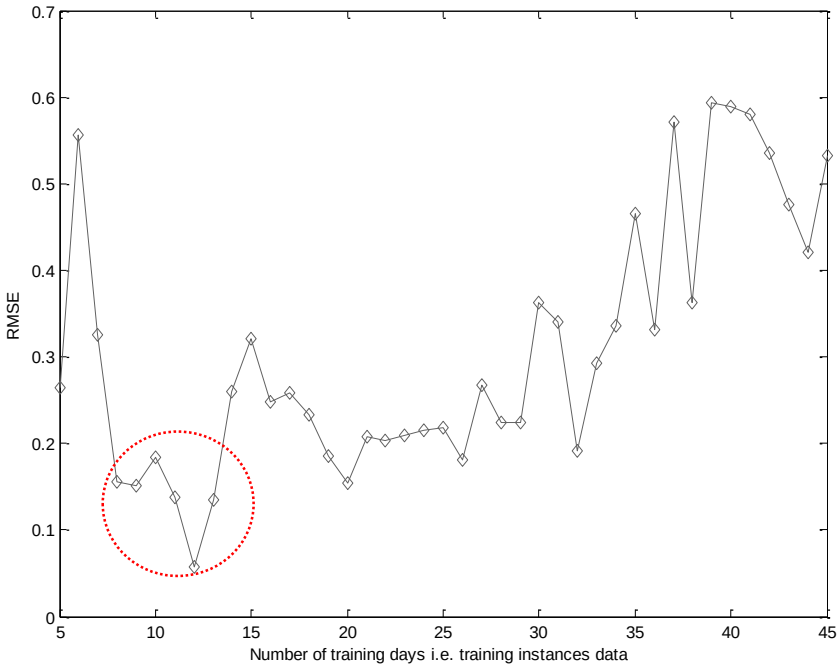

Figure 4.24: Influence of number of days data in accuracy of prediction model

Intermediate Recommendations: From the remark 2.6, the readers can consider the ratio of training days to be 10 times the number of features has a good recommendation since the performance has less error and stable values between 7 and 14 days. However, the readers are encouraged to use 12 days as relevant days as a general rule of thumb since it has higher performance.

Selection on AI Models

The choice of **AI models**: ANN, SVM, BEDT and RF depend on the choice of the relevant data selection methods (HDD, Modified HDD, FD and DTW).

The comparisons between different **AI models** are evaluated using input features of scenario S8 based on different "**relevant data**" modeling approaches. In case of SVM, the parameters C shown in Equation (3.35) in **Chapter 3** is tuned at $\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$ and ε is searched at $\{0.001, 0.01, 0.1, 0.2, 0.5\}$. The σ parameter shown in Equation (B.15) in **Appendix B** of RBF kernel is tuned at $\{2^{-15}, 2^{-14}, \ldots, 2^{15}\}$. Similarly, for BEDT, the number of trees n_m shown in Equation (3.36) in **Chapter 3** is searched from $\{25, 50, 100, 150, 200\}$ at increment of 25, the number of leaf in each tree β_m shown in Equation (3.36) is chosen 5 and the learning parameter υ in Equation (B.26) in **Appendix B** varies from [0.1 0.25 0.5 0.75 1]. Finally, for RF, the number of trees of forest B in Equation (3.37) is searched from $\{25, 50, 100, 150, 200\}$ at increment of 25, the random sample with replacement, i.e., bootstrap sample is 1, the number of randomly selected features in each split to grow trees is one-third of the number of features of scenario S8 and number of leaf in each tree is varied from [1 5 10 20 50]. The computation time is evaluated in 2.5 GHz CPU with 128 GB RAM. The parameters of ANN are similar to Table (4.7) and summary of parameters of SVM, BEDT and RF are shown in Table (4.9).

Support Vector Machine (SVM)		Boosted ensemble decision tree (BEDT)		Random Forest (RF)	
Input and output of model	S8 shown in Table (4.5)	Input and output of model	S8 shown in Table (4.5)	Input and output of model	S8 shown in Table (4.5)
Kernel Function	RBF shown in Equation B.15	Number of trees	[25 50 75 100... 200]	Number of trees in forest	[25 50 75 100... 200]
C	$\{2^{-5},2^{-4},...,2^{5}\}$	Number of leaf in each tree	5	Number of leaf in each tree	[1 5 10 20 50]
σ	$\{2^{-15},2^{-14},...,2^{15}\}$	learning rate	[0.1 0.25 0.5 0.75 1]	Number of random features	1/3(number of features of M10 model)
ε	[0.001, 0.01, 0.1,0.2,0.5]			Bootstrap sample	1
Normalization	min-max				
Model selection: 5-fold cross validation					
Datasets Training and Validation: Lyon-1 year, Clermont-Ferrand-1 year and Lille-1 year; Testing: Paris-1 year					

Table 4.9: Summary of SVM, BEDT and RF parameters

The comparison of different **AI models** using different "**relevant data**" modeling approaches for heat load prediction is shown in Table (4.10). It can be observed that SVM model has better performance compared to other **AI models** for all the cases whereas ANN and RF are also suitable for Case 3-4 building. On the other hand, BEDT performance is worst for all the cases. This might be because BEDT require large number of data for model training. In addition, it can be noticed that a modified HDD and pattern recognition method (DTW and FD) has higher performance compared to HDD method for all kinds of **AI models**. The poor performance of HDD method might be due to solar gains and internal gains effects are not included while selecting relevant data selections. On the other hand, the modified HDD method has better performance for Case 1-2 building while its performance decreases by small values for Case-4 building. The decrease in performance in Case-4 building is due to the fact that it relies on weight effect of energy demand during a day and this LEB has zero energy demand in many periods

resulting into daily mean energy consumption zero and violating selection methods principle. The more noticeable result is DTW "**relevant data**" modeling approach has comparable results with the modified HDD method and can be regarded as best "**relevant data**" modeling approach.

Models	Relevant Data Training Selection Methods	Case 1				Case 2				Case 3				Case 4			
		Median		Overall		Median		Overall		Median		Overall		Median		Overall	
		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
ANN	HDD	0.86	15.8	0.987	18.5	0.81	14.3	0.978	18	0.91	8.4	0.987	10.5	0.81	11.1	0.945	14.3
	Modified HDD	0.93	11.2	0.993	13.9	0.91	10.2	0.986	14.4	0.96	5.9	0.991	9	0.95	6	0.977	9.3
	Frechet Distance	0.93	10.9	0.99	14.1	0.92	9.5	0.99	13.6	0.96	6.1	0.99	8.6	0.93	6.4	0.971	10.3
	DTW	0.95	10.1	0.995	12.2	0.94	8	0.99	11.5	0.96	6	0.99	8.5	0.97	3.9	0.99	6
SVM	HDD	0.86	16.1	0.987	18.6	0.82	14	0.978	17.6	0.94	7.5	0.988	9.8	0.87	9.1	0.963	11.7
	Modified HDD	0.97	7.4	0.996	10.1	0.96	6.4	0.994	9.3	0.96	5.8	0.992	8.2	0.97	4.4	0.989	6.1
	Frechet Distance	0.96	7.2	0.995	10.6	0.95	6.2	0.994	9.6	0.98	5.4	0.994	7	0.98	3.6	0.99	5.2
	DTW	0.97	7.5	0.996	10.4	0.96	6.4	0.994	9.4	0.98	5.1	0.994	7	0.98	3.3	0.993	5.1
BEDT	HDD	0.78	20.1	0.982	22	0.71	17.1	0.972	20.1	0.84	11.7	0.976	14.4	0.81	10.6	0.944	14.4
	Modified HDD	0.89	13.4	0.991	15.4	0.85	12.2	0.984	15.2	0.92	8.9	0.983	12.1	0.94	6.6	0.983	7.9
	Frechet Distance	0.87	13.6	0.99	15.7	0.84	12.4	0.981	16.4	0.91	9.4	0.983	12.2	0.92	6.7	0.974	9.6
	DTW	0.89	13.2	0.992	15.2	0.85	12.2	0.984	15.1	0.91	9.3	0.985	11.2	0.93	6.6	0.98	8.6
RF	HDD	0.86	16.2	0.986	18.4	0.82	13.9	0.978	17.7	0.85	11.8	0.976	14.4	0.87	9.3	0.962	11.9
	Modified HDD	0.94	9.9	0.994	11.9	0.92	8.9	0.991	11.4	0.93	7.8	0.99	9.3	0.96	5.5	0.987	6.8
	Frechet Distance	0.94	10.1	0.992	12.4	0.92	9.3	0.986	14.4	0.94	7.5	0.989	9.7	0.94	5.9	0.978	9.1
	DTW	0.94	9.8	0.995	11.6	0.93	9.2	0.991	11.6	0.94	7.6	0.989	9.3	0.95	5.7	0.986	7.2

Table 4.10: Performance of AI models using different relevant data modeling approaches for different cases

The model training CPU-time from different AI models using DTW "**relevant data**" modeling approach for particular single prediction day as a reference considering Case-4 building is shown in Figure (4.25). It can be seen that model training CPU-time in SVM is quite faster than other AI models. The higher model training CPU-time in ANN might be due to the requirement of large number of model parameters and BEDT might be due to the necessity of tuning best learning rate for different types of trees. On the other hand, RF took large model training CPU-time compared to rest of the model. This large time might be due to the fine tuning of number of leaf in each decision tree.

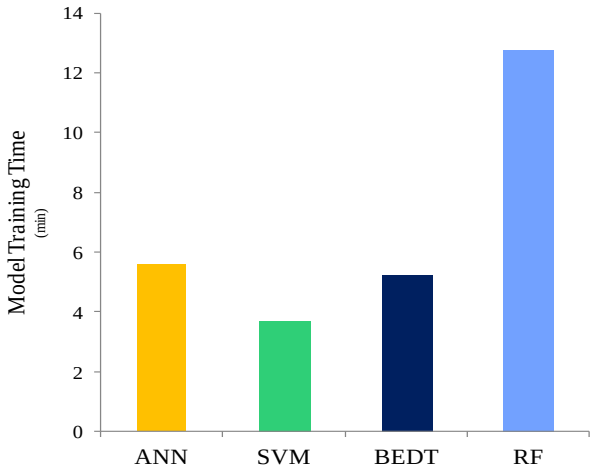

Figure 4.25: Model training CPU-time from different AI models

The model training CPU-time for different relevant data selection methods: HDD, mHDD, FD and DTW using SVM for a particular single day prediction is shown in Figure (4.26). It can be seen that FD method requires more training CPU-time (≈10 min) compared to other methods. This might be due to the fact that FD method defines number of paths in discrete form to follow the pattern recognition. In addition, this large time might be due to the large iterations it takes to find the smallest path for optimization problem solving. It is also seen that DTW and modified HDD methods training CPU-time are quite faster and reveal that these methods are useful for ESCOs and/or BEMS for optimal control applications.

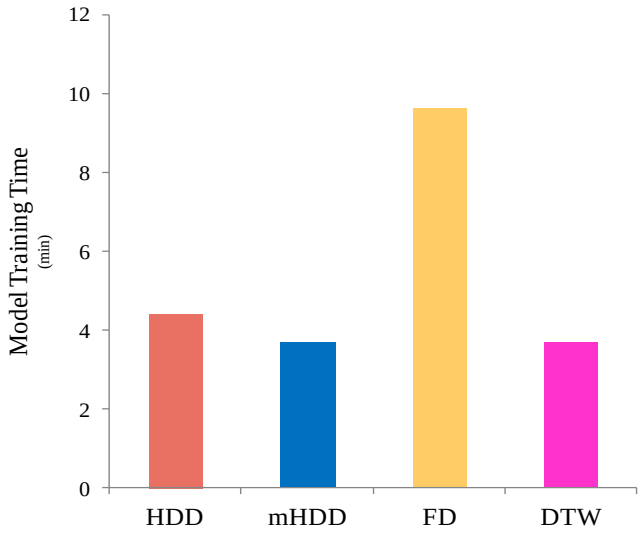

Figure 4.26: Model training CPU-time from different relevant data modeling approaches

Intermediate Recommendations: The readers are suggested to choose modified HDD or DTW method based on SVM as a reference due to their higher performance and faster model training CPU-time to predict heat load. It is more preferable to use modified HDD "**relevant data**" modeling approach for CBs and DTW for LEBs using SVM.

Comparison between the Modeling Approaches: "All Data" and "Relevant Data"

The DTW "**relevant data**" modeling approach using SVM is compared with "**all data**" approach using ANN, SVM, BEDT and RF considering input features of scenario S8. In case of BEDT and RF, the parameters of model are defined similar to Table (4.9) except that in both of the cases number of trees are searched from [25, 50, 75,….,500] at increment of 25. The parameters of ANN are defined similar to Table (4.7) and parameters of SVM are defined similar to Table (4.9). The models are compared using the Case-4 building due to the requirement of large model

training CPU-time in "**all data**" modeling approaches. The comparison between "**all data**" and "**relevant data**" modeling approaches are shown in Table (4.11). It can be seen that R^2 median performance in "**relevant data**" modeling approach is 0.09, 0.05, 0.12 and 0.02 times higher than "**all data**" modeling approach using ANN, SVM, BEDT and RF respectively. In addition, there is a significant reduction in RMSE value in "**relevant data**" approach compared to "**all data**" approach. The performance of "**all data**" modeling using RF is higher than "**all data**" modeling using SVM. Nevertheless, "**all data**" modeling approach using RF performance is lower than DTW "**relevant data**" modeling approach using SVM. Moreover, the model training CPU-time in "**relevant data**" modeling approach is 3 min 40 sec for a single day prediction while for "**all data**" modeling approach using ANN is 184 hours 43 min 6 sec, SVM is 75 hour 43 min 12 sec, BEDT is 1 hour 37 min 11 sec and 15 hour 42 min 18 sec for RF.

Performances	DTW based Relevant Data Modeling Approach		All Data Modeling Approach							
			ANN		SVM		BEDT		RF	
	Median	Overall	Median	Overall	Median	Overall	Median	Overall	Median	Overall
R^2	0.98	0.993	0.89	0.971	0.93	0.978	0.86	0.981	0.96	0.993
RMSE	3.3	5.1	8.4	10.3	7.1	9.1	8.9	8.5	4.4	4.9
Model Training Time	3 min 40 sec		184 hours 43 min 6 sec		75 hour 43 min 12 sec		1 hour 37 min 11 sec		15 hour 42 min 18 sec	

Table 4.11: Comparison of model performance of DTW based relevant data modeling approach using SVM with all data modeling approach using ANN, SVM, BEDT and RF

Because of the comparable results of prediction of heat load from "**all data**" modeling approach using RF, the sensitivity on size of training data is evaluated. The sizes of training data are varied from 3 years (Lyon, Clermont- Ferrand and Lille), 2 years (Lyon and Clermont-Ferrrand) and 1 year (only Lille) to test on Paris location shown in Figure (4.27).

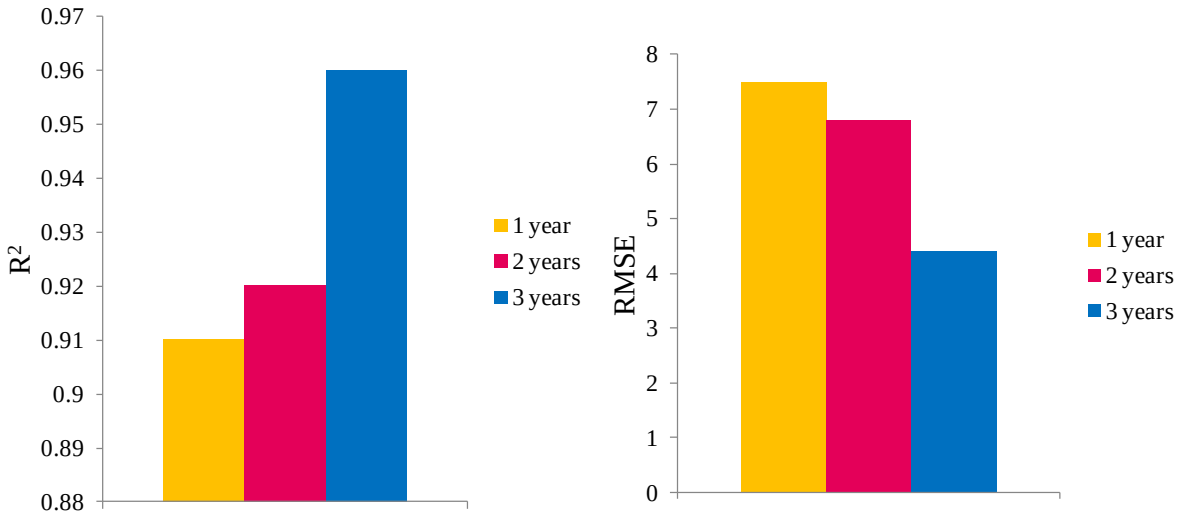

Figure 4.27: Influence of training size data using all data modeling approach using RF

It is clear from Figure (4.27) that performance of RF model goes on decreasing while decreasing the training data and reveals that RF is sensitive to the size of training data. One therefore can conclude that RF requires more training data to generalize the predictive model thus they are unsuitable for prediction of heat load with fewer training data.

As an example, the prediction from "**all data**" modeling approach using SVM and DTW "**relevant data**" modeling approach using SVM for some random days in January using scenario S8 is shown in Figure (4.28). It is clear that "**all data**" modeling approach have similar problems to scenario S7 using "**relevant data**" modeling approach for learning initial period (0-7) hour in the morning. This is explained by the fact that "**all data**" modeling approach uses a single model parameter (C, σ and ε) from all given training data due to the building operation classes are similar during the weeks. Therefore, "**all data**" modeling approach fails to generalize for each prediction day conditions. On the contrary, "**relevant data**" modeling approach changes model parameters for each prediction day and generalizes the specific conditions of prediction day though there is little problem in initial hour.

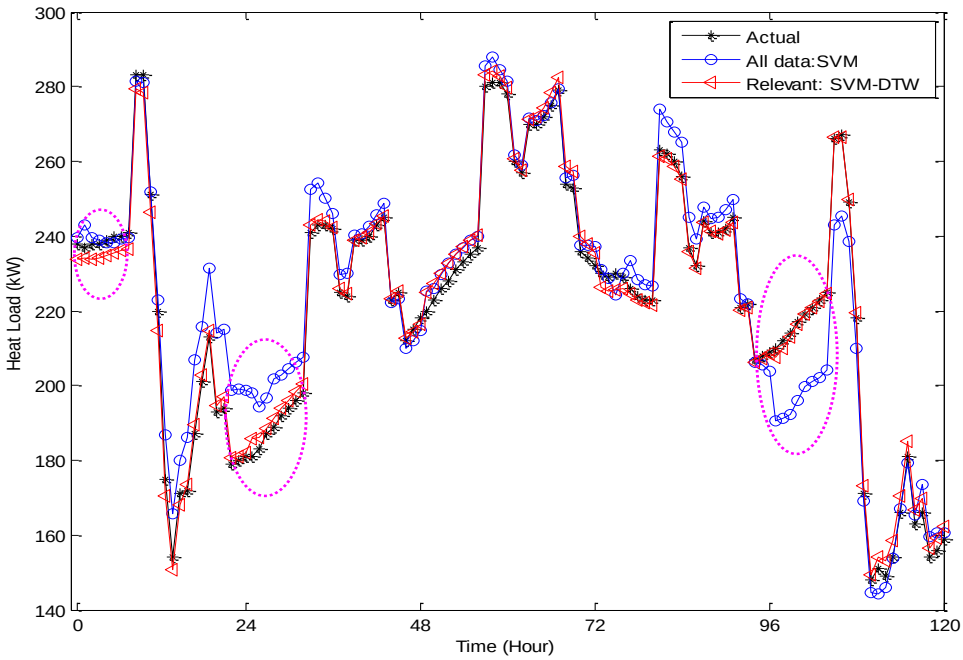

Figure 4.28: Prediction of heating load based on DTW relevant data modeling approach using SVM with all data modeling approach using SVM for some random days

We can therefore summarize that the major differences between "**all data**" and "**relevant data**" modeling approach lie on generalizing the prediction day behaviors. For example, if one has 365 days of data needs to be predicted, then there are 365 models in "**relevant data**" modeling

approach while there is a single model in "**all data**" modeling approach that generalizes each prediction condition (due to building operation classes is same during a week). It can be concluded that "**relevant data**" modeling approach includes the behavior of prediction conditions based on physical understanding by representing small data with significant model training CPU-time. Consequently, optimal model parameters (hidden neurons, performance goal in ANN; C, σ and ε in SVM; B and number of leaf in RF) are varied in "**relevant data**" modeling approach each day based on prediction conditions. Whereas in "**all data**" modeling approach, the optimal parameters are always constant (e.g., fixed initial defined parameters and 30 hidden neurons in ANN; C, γ and ε are 256, 1 and 0.01 in SVM; B and number of leaf are 375 and 1).

Intermediate Recommendations: The readers are suggested to use DTW relevant data modeling approach using SVM for heating load prediction. In case of large data available, all data modeling approach using RF is also suggested for application in heat load prediction.

Effects of Occupancy

The TRNsys results have been generated using a single-zone model for the description of this step. The application of methodology is applied to DTW relevant data modeling approach using SVM (suggested from intermediate recommendation) to Case 5-6 building with different occupancies (Figure 4.2-4.3). The shape factor and final energy demand of building are calculated similar to **Section 4.3.1**. It is found that shape factors of buildings are 0.23 and 0.35 for Case-5 and Case-6 building respectively. The final energy demand of Case-5 building is 24.1 kWh/m^2.yr whereas that of Case-6 building is 29.5 kWh/m^2.yr. Both of the cases have similar range of energy demand to that of Figure (4.9) for different climatic locations. Based on the intermediate recommendation for LEBs, the number of past day climatic conditions impacts u is chosen 3 for both cases.

Step 1: Building Classification/Clustering

The classification of building operation is calculated similar to **Section 4.3.2**. Figure (4.29) shows the building operation classification and it can be seen that buildings have three kinds of functioning profiles which are summarized in Table (4.12). The Monday CV has different behaviors than other days CV in both cases. In Case-5 building, the building operation day "Tuesday-Friday" has similar behavior whereas CV of "Tuesday-Saturday" in Case-6 building is identical. In building operation classes-3, there is no occupancy during Saturday-Sunday in Case-

5 building and during Sunday in Case-6 building (only few CVs) thus there is no requirement of heating load.

Building Operation Classes	Building Types	
	Case-5	Case-6
Classes-1	Monday	Monday
Classes-2	Tuesday-Friday	Tuesday-Saturday
Classes-3	Saturday-Sunday	Sunday

Table 4.12: Building operation classes for Case-5 and Case-6 building

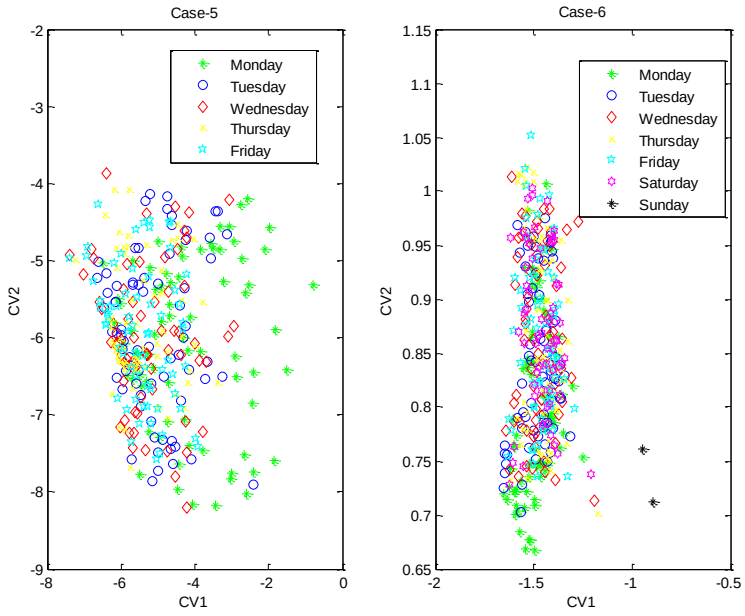

Figure 4.29: Classification of building operation classes (Case-5 and Case-6 buildings)

Step 2: Pseudo Dynamic Model

The dynamic indoor characteristics of building to control the indoor temperature represented by steady state ($T_{steady,air,in}$) are similar to **Section 4.3.2**. In Case-5 building, the operating conditions particularly set-point temperature is changed at period 5-6 and 21-22 hours; lighting is changed at period 7-8, 11-12, 13-14, 17-18 and 19-20 hours (shown in Figure 4.5). On the other hand, the set-point temperature and lighting operation of building are changed at period 7-8 and 19-20 hours; and ventilation operation is changed at period 6-7 and 20-21 hours in Case-6 building (shown in Figure 4.6). Based on these changing periods, the operational characteristics are formulated. Similarly, the occupancy profiles are changed at period similar to the behavior of lighting operation for both of the cases and accordingly transitional and pseudo dynamic model are

calculated at these changing periods with the same initial energy load level β_0 and step size of transition of energy load $\Delta\beta$ values shown in **Section 4.3.2**.

Step 3: Climatic Variables Selection

The correlation indexes (r) of direct and derived climatic variables: external temperature T_{ext}, temporal moving average of external temperature T_{ext_TDM}, sky temperature T_{sky}, horizontal solar radiation $\emptyset_{Sh}$, temporal moving average of horizontal solar radiation $\emptyset_{Sh\_TDM}$, direct solar radiation $\emptyset_{D}$, solar gain transmitted through windows $\emptyset_{Sext}$, solar gain on walls $\emptyset_{Sint}$ and temporal moving average of solar gain on walls $\emptyset_{Sint\_TDM}$ are calculated similar to **Section 4.3.2** and is shown in Figure (4.30) for Case-5 and Case-6 buildings. It can be noticed that for both of the cases, correlation indexes (r) of direct solar radiation $\emptyset_{D}$ are relatively lower than other climatic variables. In addition, it is noticed that Case-6 building external temperature T_{ext} is higher (r=0.72) than Case-5 building (r=0.52). The threshold value (φ_{th}) of 0.07 is chosen to determine the relevance of variables similar to **Section 4.3.2** and it is found that climatic variables: T_{ext}, T_{ext_TDM}, T_{sky}, $\emptyset_{Sh}$, $\emptyset_{Sh\_TDM}$, $\emptyset_{Sext}$, $\emptyset_{Sint}$ and $\emptyset_{Sint\_TDM}$ are significant for Case-5 building. It is also interestingly noticed that for Case-6 building, the solar gain on the walls $\emptyset_{Sint}$ and the solar gain transmitted through the windows $\emptyset_{Sext}$ are less significant while considering threshold value (φ_{th}) 0.07. In addition, due to mutual cross-correlation of external temperature T_{ext} and sky temperature T_{sky}, sky temperature T_{sky} is not considered due to its less correlation indexes compared to external temperature T_{ext} in both cases.

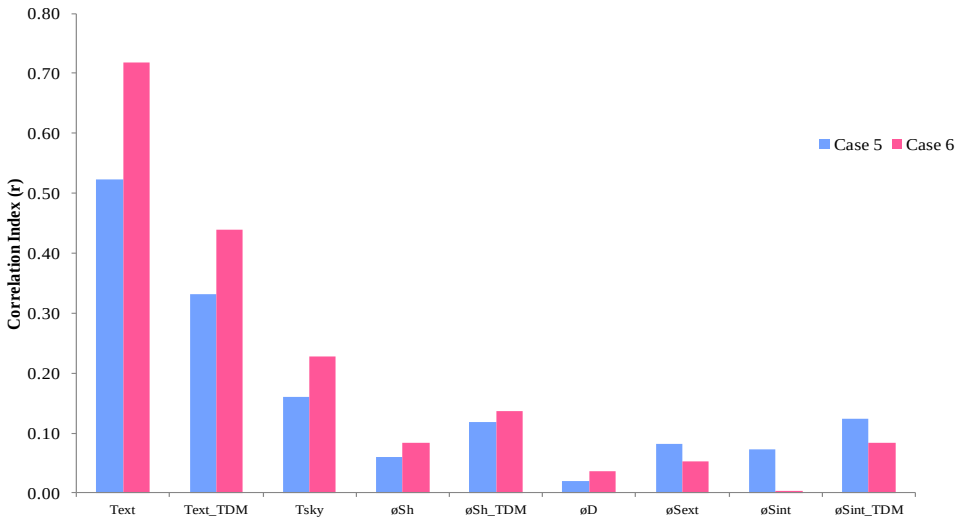

Figure 4.30: Correlation indexes of direct and derived climatic variables of Case-5 and Case-6 buildings

The cross-correlation (r_{xy}) is calculated similar to **Section 4.3.2** for lags (Φ) at 23 hours and it is found that external temperature T_{ext}, horizontal solar radiation $\emptyset_{Sh}$, solar gain transmitted through windows $\emptyset_{Sext}$ and solar gain on walls $\emptyset_{Sint}$ has past dynamics at 2 hours for both cases (Case-5 and Case-6). In Case-6 building, the temporal moving average behaviors of solar gain on walls $\emptyset_{Sint\_TDM}$ have higher correlation index (r=0.44) and due to solar gain on walls has past climatic effects at 2 hours, the effects of solar gain are also considered for further analysis though it is discarded from threshold value φ_{th}. Therefore, main significant direct and derived climatic variables are: external temperature T_{ext}, temporal moving average of external temperature T_{ext_TDM}, horizontal solar radiation $\emptyset_{Sh}$, solar gain transmitted through windows $\emptyset_{Sext}$, solar gain on walls $\emptyset_{Sint}$ and temporal moving average behavior of solar gain on walls $\emptyset_{Sint\_TDM}$; the past 1 hour dynamics of external temperature T_{ext} and past 2 hours dynamics of all solar radiations.

Step 4: Sets of Input Features

In order to understand the behavior of different input features, five input scenarios are considered and summary of input and output variables are shown in Table (4.13).

Name		Description	Scenarios				
			S1	S2	S3	S4	S5
Outputs	P (t)	Heat Load (kW)	×	×	×	×	×
Inputs	Text(t)	External temperature (^{0}C)	×	×	×	×	×
	Text (t-1)	External temperature at 1 hour time delay (^{0}C)	×	×	×	×	×
	øSh(t)	Horizontal solar radiation (kW)	×	×	×	×	×
	øSh(t-1)	Horizontal solar radiation at 1 hour time delay (kW)	×	×	×	×	×
	øSh(t-2)	Horizontal solar radiation at 2 hours time delay (kW)	×	×	×	×	×
	øSext(t)	Solar gain transmitted through window (kW)				×	×
	øSext(t-1)	Solar gain transmitted through window at 1 hour delay (kW)				×	×
	øSext(t-2)	Solar gain transmitted through windows at 2 hours delay (kW)				×	×
	øSint(t)	Solar gain on wall (kW)				×	×
	øSint(t-1)	Solar gain on wall at 1 hour delay (kW)				×	×
	øSint(t-2)	Solar gain on wall at 2 hours delay (kW)				×	×
	occup	Occupancy profile [0 1]	×	×	×	×	×
	oper	Operational characteristics [0 1]		×	×	×	×
	trans	Transitional attributes [0.2 1]		×	×	×	×
	PDL-1	Pseudo dynamic lag 1 [0.2 1]		×	×	×	×
	PDL-2	Pseudo dynamic lag 2 [0.2 1]		×	×	×	×
	Text_TDM	Temporal moving average of external temperature (^{0}C)			×		×
	øSh_TDM	Temporal moving average of horizontal solar radiation (kW)			×		
	øSint_TDM	Temporal moving average of solar gain on wall (kW)					×

Table 4.13: Summary of input and output variables of different scenarios

The scenario S1 consist external temperature and horizontal solar radiation with occupancy; scenario S2 includes transitional and pseudo dynamics effects in scenario S1; scenario S3 takes

into account temporal moving average behaviors of external temperature and horizontal solar radiation. Lastly, scenarios S4-S5 includes the derived climatic variables: solar gain transmitted through windows $\emptyset_{Sext}$ and solar gain on walls $\emptyset_{Sint}$.

Step 5: Analysis of Climatic Variables on the Building

By applying the **Section 4.3.2**, it is identified that the external temperature T_{ext} is highly dominant (85%) compared to solar gain (15%) for Case-5 building whereas external temperature is also highly dominant (94%) compared to solar gains (6%) in Case-6 building neglecting the impact of solar gain from derived variables. The impacts of solar gains are less important compared to external temperature T_{ext} in both cases and it might be because there is no requirement of heating load during early morning and night during working days and in the weekend (Saturday-Sunday in Case-5 and Sunday in Case-6 building). This further concludes that solar gain variables are less significant in determining thermal storage in walls.

Step 6: Selection of the Sub-Database

The selection of the sub-database are based on the influence of climatic variable weights obtained from step-5 and after the identification of similar climatic variables using DTW which follows similar steps in **Section 4.3.2**. The number of days for model training is based on the intermediate recommendations in **Section 4.3.2** and the number of days (*l*) is chosen 12.

Step 7: Heating Load Prediction

The model is evaluated using DTW "**relevant data**" modeling approach using SVM and the parameters of SVM are similar to that defined in Table (4.9). The performance comparison for different input features scenarios in both Case-5 and Case-6 buildings are shown in Table (4.14). It is seen that pseudo dynamic model (scenario S2) results in greater accuracy while comparing with the scenario S1 in both cases. There is a little improvement in overall performance in scenarios S4-S5 compared to scenario S2. In addition, there is no any improvement while using temporal moving average behaviors of past climatic conditions shown by results of scenarios S4-S5. It is most interestingly noticed that though the correlation indexes (r) due to solar gain transmitted through windows $\emptyset_{Sext}$ and solar gain on walls $\emptyset_{Sint}$ are lower in Case-6, there is a little improvement in scenario S4 while compared with scenario S2.

Input Features Scenarios	Case-5				Case-6			
	Median		Overall		Median		Overall	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
S1	0.220	101.6	0.410	108.4	0.947	22.7	0.954	27.8
S2	0.982	17.4	0.973	23.3	0.987	11.1	0.986	15.7
S3	0.976	18.1	0.971	24.1	0.983	12.1	0.984	16.5
S4	0.978	16.8	0.976	21.8	0.992	9.0	0.990	13.2
S5	0.978	16.6	0.975	22.6	0.991	9.0	0.988	14.2

Table 4.14: Prediction performance of different scenarios for Case-5 and Case-6 building based on DTW relevant data modeling approach using SVM

The prediction of some days in January using scenarios S2 and S4 is shown in Figure (4.31-4.32) for Case-5 and Case-6 building. It can be seen that both of the scenarios (S2 and S4) have ability to predict for the given load in both of the cases. This further reveals that for these types of buildings, with climatic variables: external temperature T_{ext} and horizontal solar radiations $\emptyset_{Sh}$ have good results (scenario S2) similar to that of using derived solar gains (scenario S4).

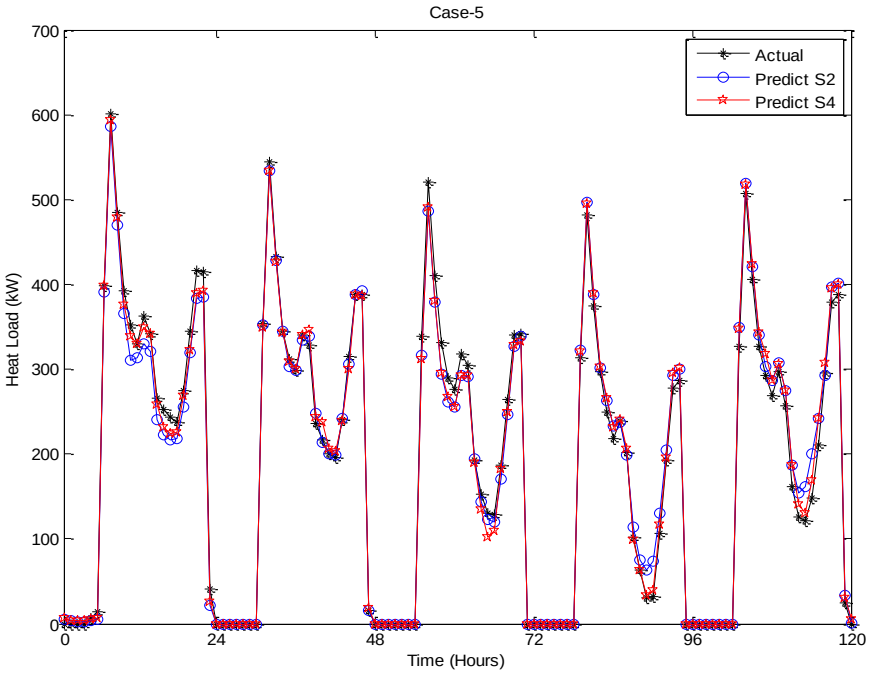

Figure 4.31: Prediction of heating load using scenarios S2 and S4 of Case-5 building (some random days in January)

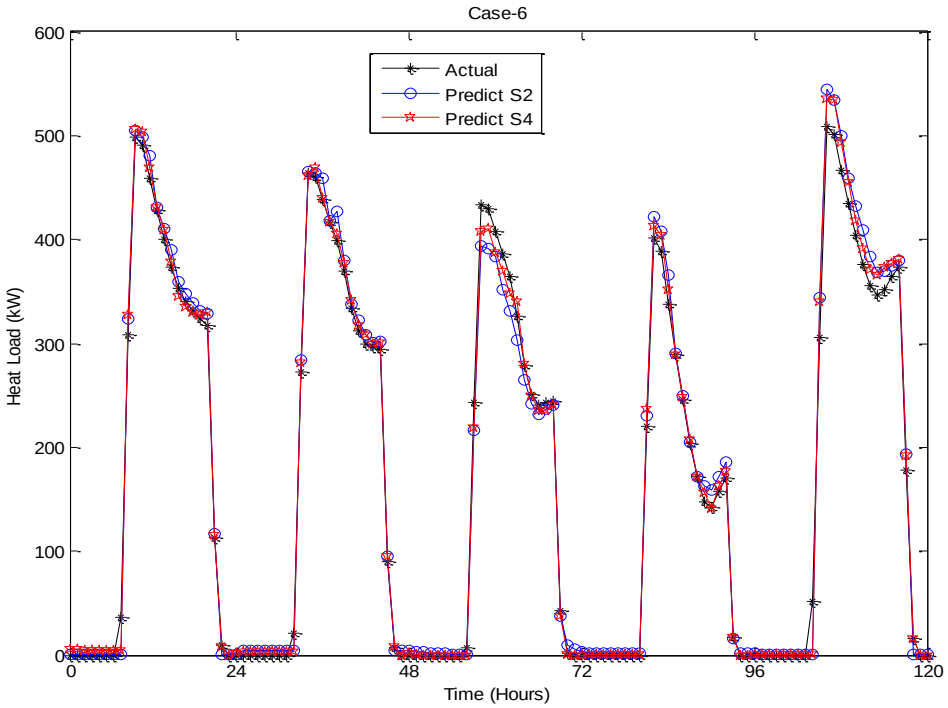

Figure 4.32: Prediction of heating load using scenarios S2 and S4 of Case-6 building (some random days in January)

Intermediate Recommendations: The methodology has proven to predict heat load with high accuracy for different occupancy single-zone building models based on DTW relevant data modeling approach using SVM. The readers have the choice to select the input features S2 or S4 depending on the availability and calculation steps on derived climatic variables: solar gain transmitted through the windows $\emptyset_{Sext}$ and solar gain on walls $\emptyset_{Sint}$. The building operating conditions and pseudo dynamic model have greater effects in prediction performance.

Effects due to Multi-zone Model

The TRNsys have been generated using multi-zone model for the description of this step. The main question is whether the methodology will be working in case of complex building model?

In Figure (4.9), the readers can see the difference in the final heating energy demand calculation with the single-zone (Case 4-6) to multi-zone (Case 4*-6*) building model. The same methodology mentioned in **Section 4.3.2** is applied to multi-zone building model according to the intermediate recommendations.

Step 1: Building Classification/Clustering

The building operations classifications for different multi-zone models are shown in Figure (4.33). It can be observed that for Case-4* building, all the CVs are clustered in one form thus can be regarded as single operation classes. Similarly, in Case-5* building, Saturday, Sunday, Monday has distinct profiles than other days, thus four kinds of building operation classes can be considered: Monday, Tuesday-Friday, Saturday and Sunday. In Case-6* building, Monday and Sunday have a distinct profiles but other days look similar thus can be regarded as three kinds of operation classes.

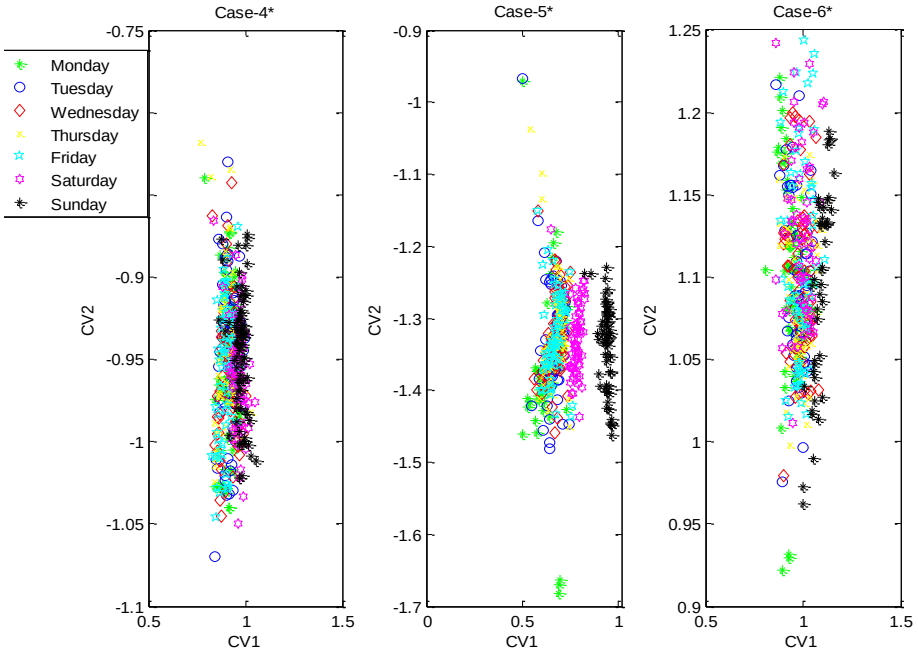

Figure 4.33: Classification of building operation classes (Case-4*, Case-5* and Case-6*)

Step 2: Pseudo Dynamic Model

The pseudo dynamic models are based on the **Section 4.3.2** described for single-zone building model. The only differences in multi-zone building model are the operating characteristics and pseudo dynamic models are developed based on the changing period at different zones in order to reflect differences in different zones. For instance, in Case-4* building, the occupancy profile is changed at periods 7-8, 11-12, 13-14, 18-19 and 20-21 hour (shown in Figure 4.1) and the building operating conditions remains constant for zone-1 (shown in Figure 4.4). However, in zone-2, building operating conditions are changed at period 5-6 and 21- 22 hour. Therefore, PDM

are developed based on these aggregated changing period (5-6, 7-8, 11-12, 13-14, 18-19 and 21-22 hour).

Step 3: Climatic Variables Selection

The correlation indexes (r) of direct and derived climatic variables are shown in Figure (4.34) and it can be seen that external temperature T_{ext} has higher correlation in Case-4* and Case-6* building compared to Case-5* building. In addition, it is also noticed that solar gains are more important in Case-4* building compared to Case-5* and Case-6* building. Other reason of this solar gain important in Case-4* building might be due to the occupancy profile behaviors in different zones (For instance, there is always same occupancy in all zones in Case-4* building). By using the threshold value (φ_{th}) defined similar to **Section 4.3.2**, the significant variables are T_{ext}, T_{ext_TDM}, T_{sky}, $\emptyset_{Sh}$, $\emptyset_{Sh\_TDM}$, $\emptyset_{Sext}$, $\emptyset_{Sint}$ and $\emptyset_{Sint\_TDM}$ by neglecting mutual cross-correlation indexes.

The cross-correlation indexes (r_{xy}) is similar to single-zone building model described in **Section 4.3.2**.

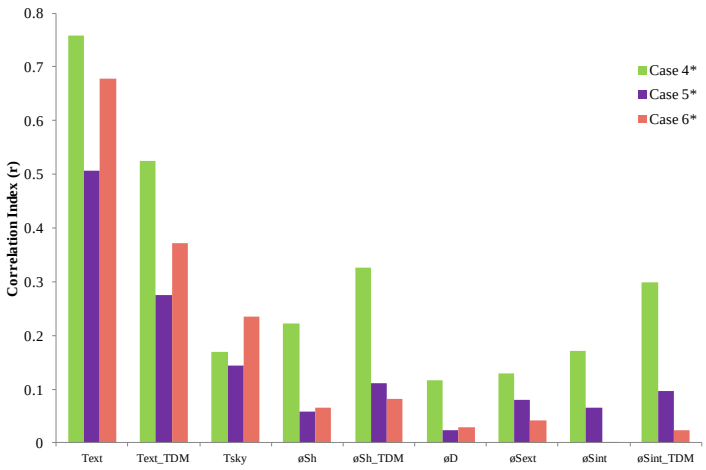

Figure 4.34: Correlation indexes of direct and derived climatic variables of Case-4*, Case-5* and Case-6* building

Step 4: Sets of Input Features

The input features considered for analysis are similar to shown in Table (4.13) due to the similar climatic variables selection.

Step 5: Analysis of Climatic Variables on the Building

The influence of climatic variables, in particular, solar gain is more foreseen in Case-4* building compared to Case-5* and Case-6* building. In Case-4* building, the influence of external temperature T_{ext} is 57% dominant which is followed by solar gain on walls $\emptyset_{Sint}$ (25%), solar gain transmitted through windows $\emptyset_{Sext}$ (10%) and horizontal solar radiation $\emptyset_{Sh}$ (5%). For Case-5* building, the influence of external temperature T_{ext} is only 56% which is followed by solar gain on walls $\emptyset_{Sint}$ (20%), solar gain transmitted through windows $\emptyset_{Sext}$ (18%) and horizontal solar radiation $\emptyset_{Sh}$ (6%). However, by neglecting the effects of derived climatic variables, the influence of external temperature T_{ext} is more dominant (86%) than solar gain (14%). Similarly, by neglecting the effects of derived climatic variables in Case-6* building, the external temperature T_{ext} weight is higher (94%) followed by solar gain (6%).

Step 6: Selection of the Sub-Database

The selections of sub-database are done after the step-5. DTW relevant data selection is performed to identify similar climatic conditions according to suggestion from intermediate recommendation. The influence of number of days for model training is considered 12 days.

Step 7: Heating Load Prediction

Similarly, mentioned in effects of different occupancies, SVM model and their parameters are defined in same way. The prediction of heating load for different scenarios and cases based on DTW "**relevant data**" modeling approach using SVM is shown in Table (4.15). It can be seen that scenarios S4 and S5 are suitable for all the cases. It can be noticeably seen that scenario S2 that uses direct climatic variables has little differences compared to scenarios S4 and S5.

Input Features	Case-4*				Case-5*				Case-6*			
	Median		Overall		Median		Overall		Median		Overall	
Scenarios	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
S1	0.730	25.1	0.842	28.7	0.245	72.8	0.452	84.1	0.900	19.0	0.922	26.7
S2	0.979	7.2	0.975	11.5	0.982	11.6	0.975	17.8	0.971	9.9	0.980	13.4
S3	0.980	7.5	0.977	11	0.976	13.4	0.974	18.3	0.970	10.3	0.978	14.2
S4	0.981	7.2	0.975	11.3	0.985	11.0	0.979	16.5	0.988	6.0	0.991	9.0
S5	0.985	7.1	0.978	10.7	0.980	10.6	0.980	16.2	0.979	8.3	0.986	11.6

Table 4.15: Prediction performance of heating load for different scenarios and cases based on DTW relevant data modeling approach using SVM

The prediction performance for some days in January is compared between scenarios S2 and S4 for different cases shown in Figure (4.35-4.37).

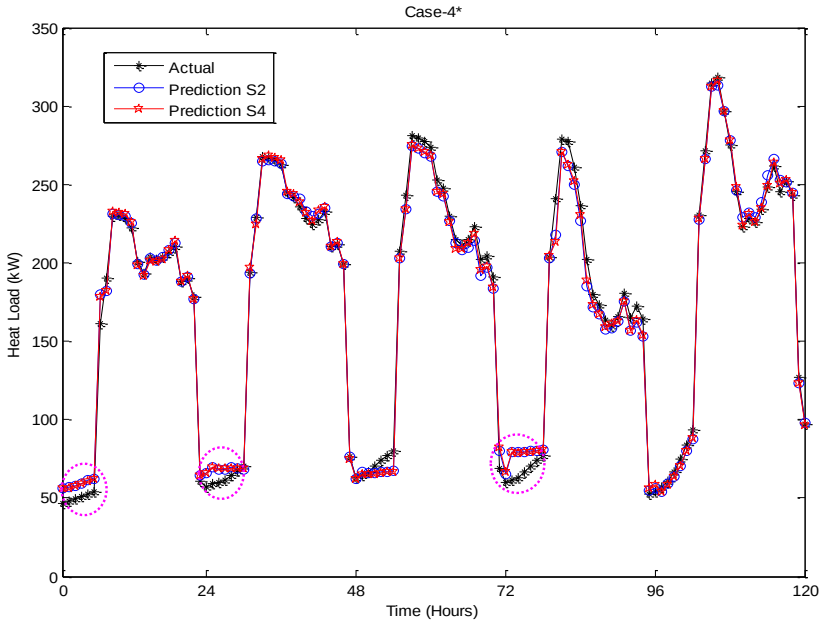

Figure 4.35: Prediction of heating load using scenarios S2 and S4 of Case-4* building (some random days in January)

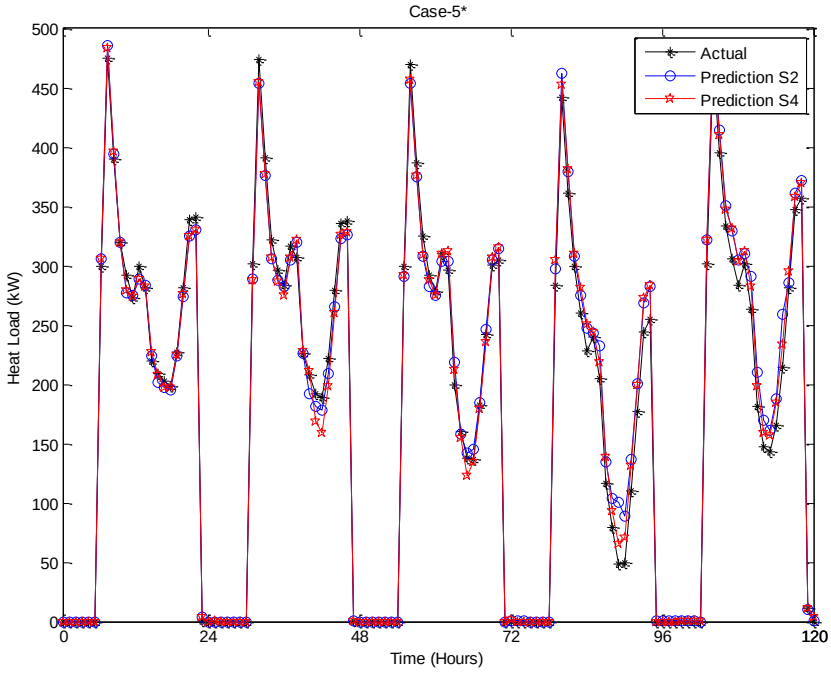

Figure 4.36: Prediction of heating load using scenario S2 and S4 of Case-5* building (some random days in January)

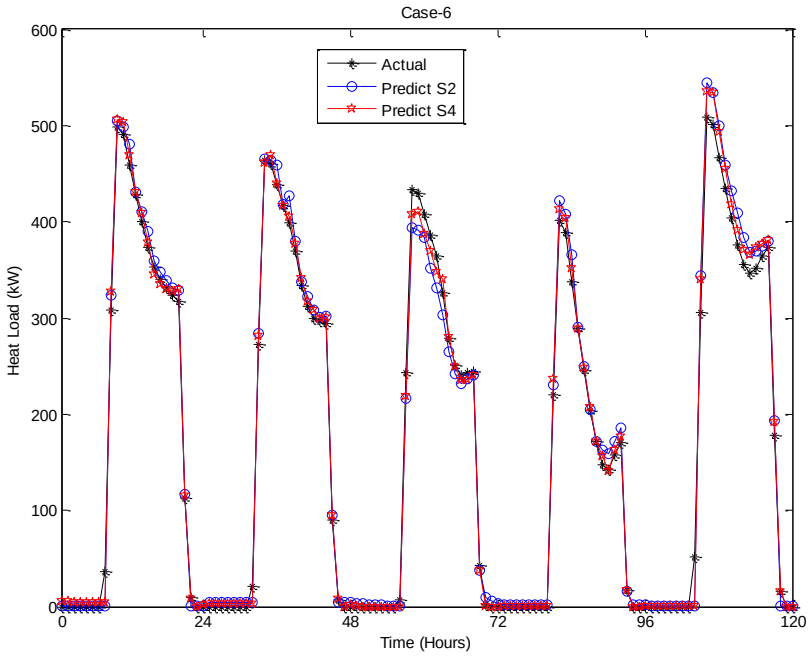

Figure 4.37: Prediction of heating load using scenarios S2 and S4 of Case-6* building (some random days in January)

It can be seen that for Case-4* building, it has little initial problem similar to "**all data**" modeling approach in Case-4 for both scenarios S2 and S4. In the other hand, there is no much difference in Case-5* and Case-6* buildings in both of the scenarios while comparing with the single-zone model (Case-5 shown in Figure 4.31 and Case-6 shown in Figure 4.32).

Intermediate Recommendations: The methodology has proven to predict heat load with high accuracy in case of complex building that uses multi-zone model based on DTW "**relevant data**" modeling approach using SVM. The readers are suggested to choose the input feature scenario S4 or S5 to make good prediction model. However, readers are also suggested to use scenario S2. This is due to scenario S2 avoids using derived climatic variables (solar gain on walls $\emptyset_{\mathrm{Sint}}$ and solar gain transmitted through windows $\emptyset_{\mathrm{Sext}}$). Finally, the readers are suggested to aggregate the changing period of occupancy and building operating conditions of multi-zones into one zone to develop pseudo dynamic model.

4.4 Conclusion

This chapter provides application of proposed methodology to predict heating load for CBs to LEBs. It first provides detail step-by-step procedure to single-zone CBs to LEBs and provides intermediate recommendations using "**relevant data**" modeling approach. It also provides comparison between different AI models and relevant data selections method and identified that DTW "**relevant data**" modeling approach using SVM has better performance compared to other models for all the cases.

Then it provides comparison study on two kinds of modeling approaches: "**relevant data**" and "**all data**". It is found that "**relevant data**" modeling approach has higher performance and faster model building CPU-time compared to "**all data**" approach revealing the benefit to use for ESCOs and/or BEMS in control and forecasting purposes for a longer period. In addition, it also provides study on different kinds of occupancies. The results reveal that the proposed approach is suitable for different kinds of occupancies.

Finally, the methodology is applied to multi-zone building model by aggregating the heat load with modification in pseudo dynamic model. The results showed that the proposed method provides higher prediction performance for complex multi-zone building model though a little performance decreases compared to single-zone building model.

The next chapter provides application of methodology to real buildings.

Chapter 5: Application- Real Building

5.1 Buildings Characteristics

5.1.1 Building Description

Ecole des Mines de Nantes (EMN) building located in Nantes, France is used as a real building. The building belongs to the mixed conventional and LEB type. The building has a total floor area of 25,000 m^2. It consist 900 students and 200 employees. It consists of 120 research and administrative rooms, 30 class rooms, 3 laboratories and 8 seminar halls. The area of class room is different from each other but each class room can be occupied by 18 to 28 students. It has also 2 big and 6 seminar halls which can accommodate up to 250 and 80 students respectively.

5.1.2 Data Collection

The building heating load and climatic conditions data are obtained from data acquisition system for 7 months (14/10/2012 – 28/02/2013) and (24/02/2014 – 02/05/2014) during the heating season period with 5 minutes sampling time. However, since BEMS are generally managed at 15 minutes sampling time, 5 minutes data samples are averaged at 15 minutes. The first period of data (14/10/2012 – 28/02/2013) belongs to CB and the second period of data (24/02/2014 – 02/05/2014) consists CB and LEB due to the construction of a new LEB which has been operated only during the second period.

The climatic conditions: external temperature (T_{ext}) and horizontal solar radiation ($\emptyset_{Sh}$) database are available, however, horizontal solar radiation $\emptyset_{Sh}$ has many missing data and outliers, thus it is not taken into consideration. The external temperature T_{ext} has summary statistics of minimum, average and maximum temperature of -1.5^{0}C, 11.4 ^{0}C and 21.5 ^{0}C respectively.

5.1.3 Occupancy Profile

The simplified/theoretical occupancy profile is shown in Figure (5.1). It can be seen that the building is only occupied during working day (Monday-Friday) from 8:00-17:45 hour which

corresponds to an office use time. It can be also seen that there are few occupancy at 12-13:30 hour because of lunch time. There is no occupancy during weekend (Saturday-Sunday).

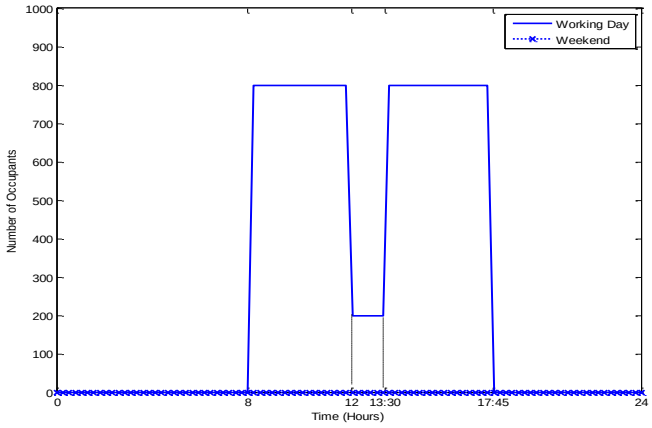

Figure 5.1: Occupancy profile for working days and weekend

5.1.4 Building Operating Conditions

The operating condition of building (only approximated set-point temperature) is known from the information provided by the building operator. Figure (5.2) shows the set-point temperature operation during working day. It can be seen that set-point temperature is maintained constant at 21 ^{0}C before entering and leaving the occupants.

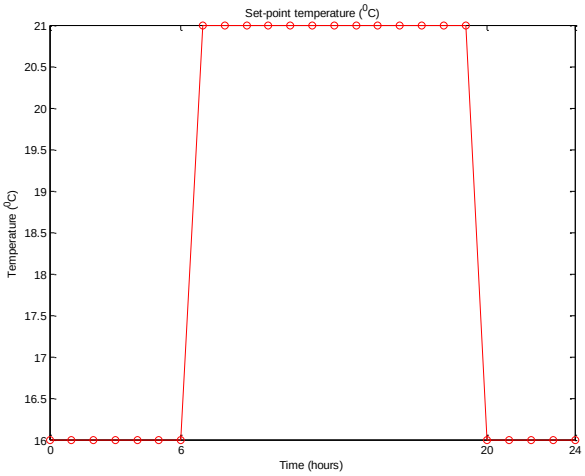

Figure 5.2: Operating conditions of building for working day

5.2 Application of AI Modeling Methodology

5.2.1 Introduction

The dynamic response of building, i.e., the time constant represented by "**Impact of Thermal Envelope on Type of Building**" block in Figure (3.4) in **Chapter 3** and represented by Equation (2.6) in **Chapter 2** is assumed to be 1 day since this building belongs to conventional categories. Therefore, number of past day climate impacts *u* in Equation (3.7, 3.12, 3.20-3.21) in **Chapter 3** corresponds to 1.

The heat demand and the occupancy profile during working day is shown in Figure (5.3) and it can be noticed that only occupancy profile does not precisely gives information about the heat demand characteristics.

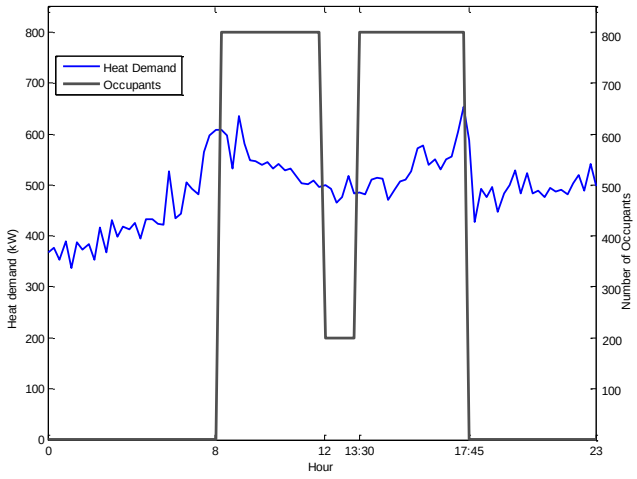

Figure 5.3: Heating power demand and occupancy profile during working days

5.2.2 Recommendation for Applying the Methodology "Step by Step"

Step 1: Building Operation Classification/Clustering

The classification of building operation shown in "**Building operation classification/clustering**" block in Figure (3.4) in **Chapter 3** is shown in Figure (5.4) and CV analysis show two kinds of cluster: weekend and working day.

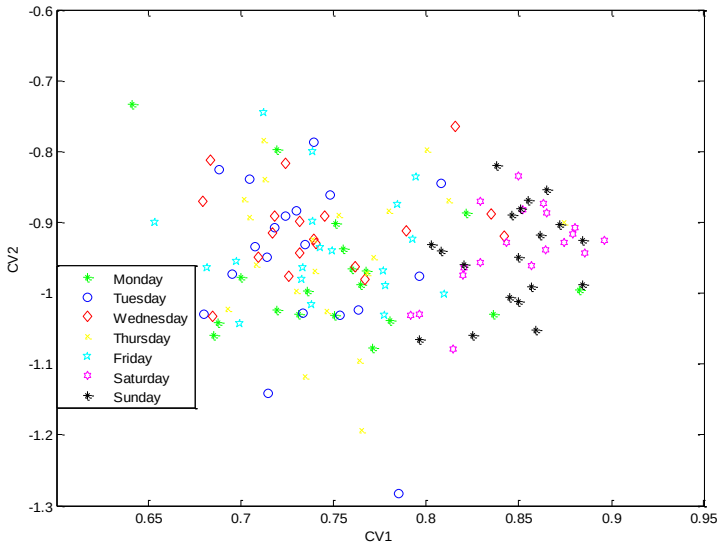

Figure 5.4: Classification of building operation classes

It can be observed that CV analysis distinguish two kinds of building operation: working day (Monday- Friday) and weekend (Saturday-Sunday) though some of the working days belong to the cluster of weekends. Figure (5.5) shows the average of heat load profile of each day of a week and it is clear that building operating conditions can be categorized into two forms: working (Monday-Friday) and weekend (Saturday-Sunday).

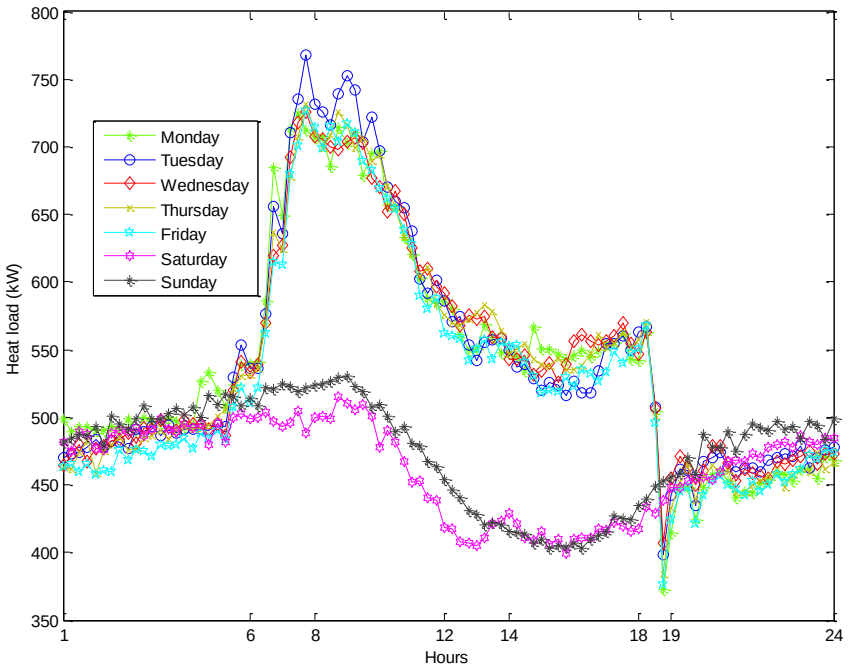

Figure 5.5: Functioning profile of building

Step 2: Pseudo Dynamic Model

The dynamic indoor air characteristic of building represented by steady state ($T_{steady,air,in}$ in **Section 3.2.2** in **Chapter 3**) is around few hours. The set-point temperature of the building is changing at period 6 and 20 hours all the days (shown in Figure 5.3) and the occupancy profile is changing at period 8, 12, 13:30 and 17:45 hour during working days (shown in Figure 5.2). The PDM thus depends on both changing period of occupancy and building operating conditions.

The transitional and pseudo dynamic characteristics are calculated using Equation (3.1) in **Chapter 3** assuming β_0 to be zero and $\Delta\beta$ with an increment of 0.03. Figure (5.6) shows the transitional and pseudo dynamic characteristics during working days for illustrations. It can be seen that four pseudo dynamic lag (PDL) is used since the sampling time of data is 15 minutes and the steady state time $T_{steady,air,in}$ corresponds to 1 hour. In the Figure (5.6), "Trans" represents the transitional characteristics and "PDL-4" represents the pseudo dynamic lag at past 1 hour. However, to understand the phenomena of pseudo dynamic lag, PDL is varied from 1 to 4. Therefore, "PDL-1" means a pseudo dynamic model with transition lag 1 (at past 15 minutes), "PDL-2" means pseudo dynamic model with transition lag 2 (at past 30 minutes) and so on.

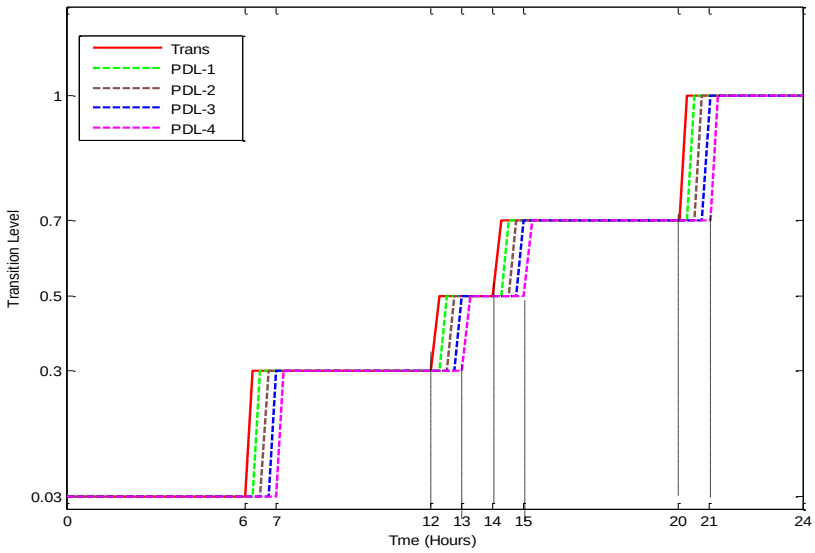

Figure 5.6: Transitional and pseudo dynamic characteristics during a day

The effect of transitional and pseudo dynamic characteristics on the heating demand is illustrated in Figure (5.7). It can be seen that the transition in energy demand at different time periods are clearly represented by transitional and pseudo dynamic characteristics. It can also be seen that for

couple of days (more than 24 hours), pseudo dynamic model shows the transition of heat demand variation.

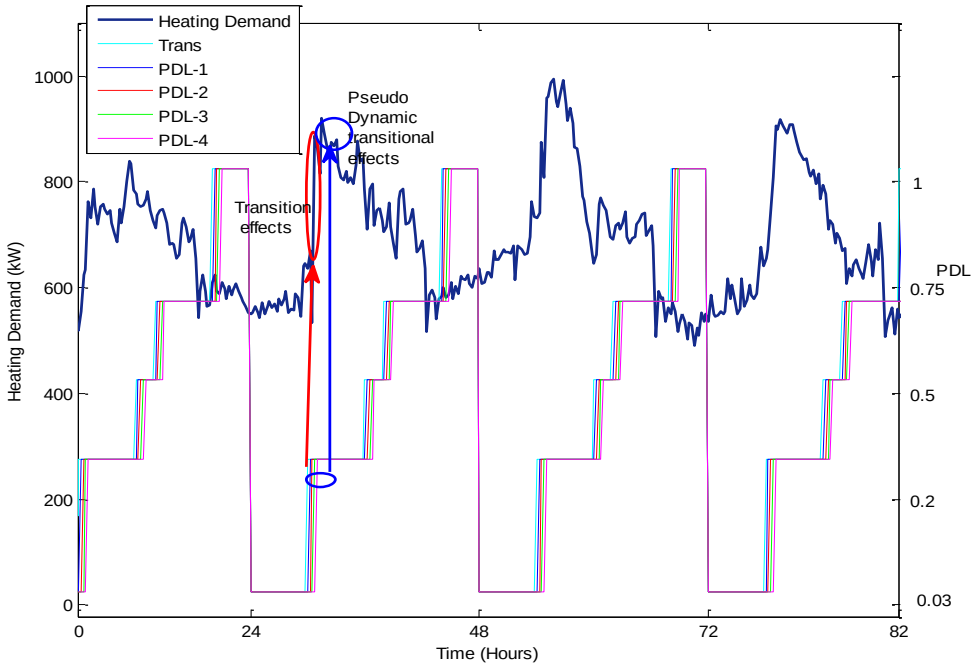

Figure 5.7: Pseudo dynamic transitional effects on heating load

Step 3: Climatic Variables Selection

The climatic variables: external temperature T_{ext} and the temporal moving average of external temperature T_{ext_TDM} are used for relevance determination. The correlation indexes (r) of external temperature T_{ext} and temporal moving average of external temperature T_{ext_TDM} by applying Equation (3.2) are: 0.60 and 0.59 and are represented by "**Climatic Variables Selection** block in Figure (3.4) in **Chapter 3**. This further shows that both features: external temperature T_{ext} and temporal moving average of external temperature T_{ext_TDM} are significant.

The cross-correlation indexes (r_{xy}) of external temperature T_{ext} is performed by applying Equation (3.3) in **Chapter 3** at lags (Φ) 96 equivalent to 24 hours and is shown in Figure (5.8). It is clear that external temperature T_{ext} has maximum cross-correlation indexes r_{xy} at past 1-2 samples and then decreases its cross-correlation indexes after. Thus, the external temperature T_{ext} and its past 2 samples delay and T_{ext_TDM} on window of past 1 day are represented by output of "**Climatic Variables Selection**" block in **Chapter 3** in Figure (3.4).

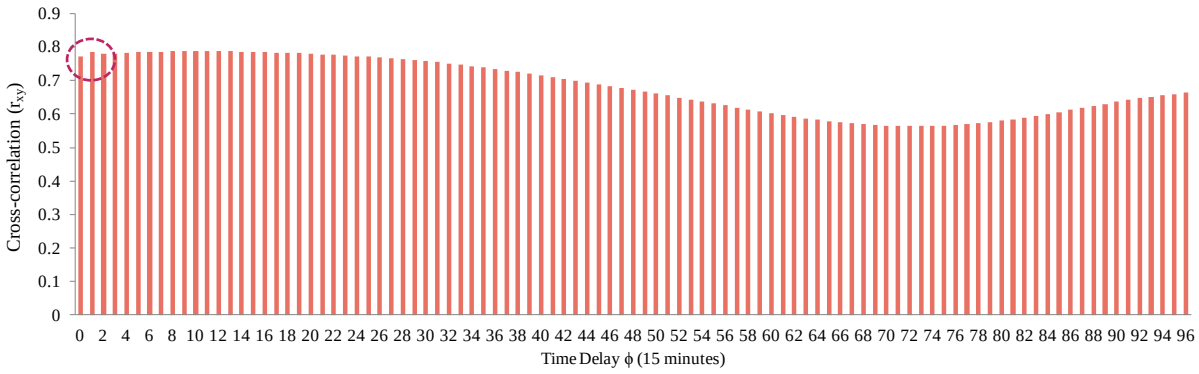

Figure 5.8: Cross-correlation indexes to select external temperature dynamics

Step 4: Sets of Input Features

Different input features are considered to understand the physical significance of each features and Table (5.1) shows the summary of input and output variables of different scenarios. It can be seen that scenario S1 includes the effect of external temperature T_{ext} and its past 30 minutes sample dynamics obtained from step-3. The scenario S2 considers the behavior of occupancy, scenario S3 take into account an operational characteristics and scenario S4 includes the transitional behavior. Similarly, scenarios S5-S8 consider the pseudo dynamic lag (PDL) at 4 lags. Finally, scenario S9 includes the behavior of temporal moving average window of external temperature T_{ext}. It can be observed from Table (5.1) that input and output variables of different scenarios are shown in t where t varies from 1 to 96 hour due to the realization of 1 day ahead prediction and $\mathcal{M}$ in Equation (3.8-3.9) corresponds to 96.

Name		Description	Scenarios								
			S1	S2	S3	S4	S5	S6	S7	S8	S9
Outputs	P(t)	Heat Load (kW)	×	×	×	×	×	×	×	×	×
Inputs	Text(t)	External temperature (^{0}C)	×	×	×	×	×	×	×	×	×
	Text (t-1)	External temperature at past 15 min delay (^{0}C)	×	×	×	×	×	×	×	×	×
	Text (t-2)	External temperature at past 30 min delay (^{0}C)	×	×	×	×	×	×	×	×	×
	occup	Occupancy profile [0 to 1]		×	×	×	×	×	×	×	×
	oper	Operational characteristics [0 1]			×	×	×	×	×	×	×
	trans	Transitional characteristics [0.2 1]				×	×	×	×	×	×
	PDL-1	Pseudo dynamic lag 1 [0.2 1]					×	×	×	×	×
	PDL-2	Pseudo dynamic lag 2 [0.2 1]						×	×	×	×
	PDL-3	Pseudo dynamic lag 1 [0.2 1]							×	×	×
	PDL-4	Pseudo dynamic lag 2 [0.2 1]								×	×
	Text_TDM	Temporal moving average of external temperature (^{0}C)									×

Table 5.1: Summary of input and output variables of different scenarios

Step 5: Analysis of Climatic Variables on the Building Load

The influence of past days climatic variables on daily average heating load is determined using wavelet analysis represented by "**Wavelet Coefficient Calculation of Selected Climatic Variables and their Past Day**" block in Figure (3.10) in **Chapter 3**. The Daubechies wavelet analysis is considered for the study at 7 levels in order to represent 96 samples of data (2^7) thus decomposition level z is 7 in Equation (3.15-3.18) in **Chapter 3**. The heat energy is transfer by the external temperature T_{ext} in the walls for a long period, so the decomposition of it is expressed by low-frequency and high-frequency coefficient.

The parameters of an intermediate model: SVM based on linear kernel are summarized in Table (5.2). It can be seen that inputs of model are the wavelet coefficients of external temperature of prediction day $T_{ext}(t)$ and past 1 day $T_{ext}(t-24)$; and output of model is the daily average heating load. In addition, normalization is carried out with min-max normalization and parameters of model are selected using k-fold cross validation where k equals to 5. The training data (October-December: 2012; January: 2013) are used to determine the weight coefficients.

Name	Descriptions
Input	Wavelet coefficients: $T_{ext}(t)$, $T_{ext}(t-24)$
Output	Daily average heating load
C	$\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$
ε	{0.001, 0.01, 0.1, 0.2, 0.5}
Kernel function	Linear
Model selection	5-fold cross validation
Normalization	min-max
Datasets	Training and Validation: October-December: 2012, January:2013

Table 5.2: Parameters of SVM used for weight calculation

It is found that for this type of building, the influence of external temperature T_{ext} is 74% and that of previous day is 26%.

Comparison between SVM and Least Square Method (LSM)

The comparison between **SVM based on linear kernel** and a **LSM based on regression model** is performed to calculate the influence of external temperature T_{ext}. The influence of wavelet external temperature coefficients T_{ext} on daily average heating load from both methods is shown in Figure (5.9). It can be noticed that for such CB, the influence of both methods are similar. For

instance, with SVM method, the effect of external temperature T_{ext} has 74% effect on prediction day compare to 26% on past day from prediction day (t-24) on daily average heating load. Similarly, with LSM method, the prediction day external temperature T_{ext} has 68% effect on daily average heating load compared to past day from prediction day (t-24) that has 32% influence.

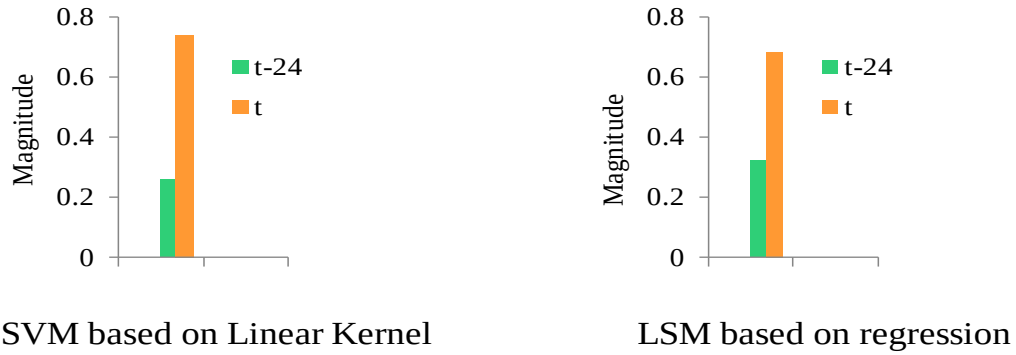

Figure 5.9: Influence of past 1 day of external temperature T_{ext} on daily average heating load using SVM based on linear kernel and LSM based on regression

Intermediate Recommendations: The readers are suggested to choose both types of method: SVM based on linear kernel and LSM based on regression model for weight determination for CBs because of their similar influence of climatic conditions on building load.

Step 6: Selection of Sub-Database

The selection of sub-databases are based on the weight determination of climatic variables from step-5 and after the identification of similar climatic variables shown by "**Identification of Similar Climatic Conditions**" block in Figure (3.10) in **Chapter 3**. The HDD is calculated for prediction day and its past 1 day using Equation (3.7) in **Chapter 3**. Similarly, modified HDD similarity weights of external temperature T_{ext} is determined using prediction day and its past 1 day using Equation (3.12) in **Chapter 3**. Finally, external temperature T_{ext} similarity weight is compared between prediction day and its past 1 day based on DTW using Equation (3.13) in **Chapter 3** and based on FD using Equation (3.14) in **Chapter 3**.

Then, the final weights of the entire database are calculated from Equation (3.22) and 12 relevant days (l in Equation 3.23 in **Chapter 3** corresponds to 12) are selected as a sub-database for model training. Then, later we performed the sensibility analysis on the number of days for model training.

Step 7: Heating Load Prediction

The **AI model** is initially evaluated based on DTW "**relevant data**" modeling approach using ANN. Hence, the comparison between different AI models and "**relevant data**" modeling approaches are performed. The input and output variables of ANN model are based on Table (5.1) and others parameters: activation function, hidden neurons, training algorithm, stopping criteria, model selection and normalization are based on Table (4.7) in **Chapter 4**. The datasets of October-December, 2012 and January 2013 are used for training and validation; and datasets of February, 2013 and February-April, 2014 are used for testing. Similarly, the cost function, hidden neurons and performance goal are calculated similar to "Step-7: Heating Load Prediction" in **Chapter 4.**

The prediction performance for different scenarios are shown in Table (5.3) and the performances of model are evaluated based on median and overall values similar to model performance in **Chapter 4**.

Models	Median		Overall	
	R^2	RMSE	R^2	RMSE
S1	0.14	101.6	0.53	97.0
S2	0.44	77.3	0.67	82.2
S3	0.71	55	0.67	81.5
S4	0.72	57.9	0.73	73.6
S5	0.72	54.3	0.74	71.4
S6	0.73	52	0.83	63.7
S7	0.76	50.8	0.80	58.8
S8	0.77	48.9	0.85	54.5
S9	0.77	48.9	0.85	54.0

Table 5.3: Comparison of different scenarios based on DTW relevant data modeling approach using ANN

It can be observed that scenario S1 that relies only on external temperature T_{ext} has very poor performance (Median: R^2=0.14, RMSE=101.6; Overall: R^2=0.53, RMSE=97). With the introduction of occupancy profile in scenario S2, the performance is increased (Median: R^2=0.44, RMSE=77.3; Overall: R^2=0.67, RMSE=82) compared to scenario S1. In addition, by using operational characteristics in scenario S3, the performance is increased a lot in terms of median values compared to scenario S2. This further concludes that operational characteristics have a

strong effect in prediction of heat load. Furthermore, with the introduction of transitional and pseudo dynamic lag in scenario S4-S8, the performance is increased slightly and the higher accuracy is achieved in scenario S8 at PDL 1 hour. Finally, with the introduction of temporal moving average of past 1 day in scenario S9, the performance is slightly increased in overall RMSE compared to scenario S8. Thus, scenario S8 or S9 is chosen as reference model for later cases.

Intermediate Recommendations: The readers are suggested to use the input feature scenario S8 or S9 as a reference for all the cases at later use for the considered building.

Sensibility Study: Influence of the Number of Relevant Days in the Prediction Performance

As recommended in the **Remark 2.6** in **Chapter 2**, the selection of relevant days for model training should be ten times the number of features (in this real application, features equal to 4-11), thus number of relevant days data require is about 110 samples equivalent to ≈ 2 days. However, since the sampling data is at 15 minutes resolution, the data of 2 days are not sufficient to divide the datasets into training and validation. Thus relevant data are varied between 5 days to 20 days for the model training to study the sensitivity analysis shown in Figure (5.10).

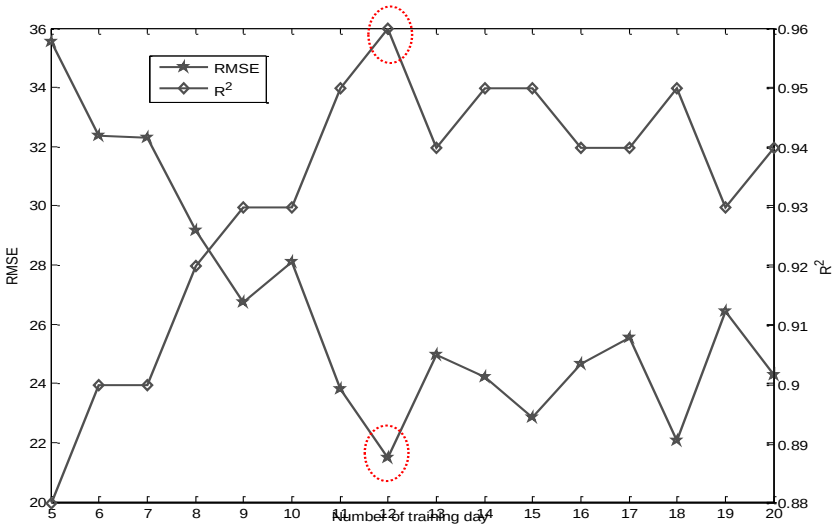

Figure 5.10: Influence of relevant days data on the accuracy of prediction model

It can be observed that the performance of model decreases (RMSE higher and R^2 lower) for few numbers of training days data (≤ 6 days) and high number of training days (≥ 13 days) with some

fluctuations. However, the performance increases between 7 and 12 days but more noticeably higher performance is achieved at 12 training days. Therefore, in this study 12 days (*l* in Equation 3.23 in **Chapter 3** corresponds to 12) is used as relevant days for model training shown in block "**Number of Relevant Days**" in Figure (3.10) in **Chapter 3**.

Intermediate Recommendations: From the **Remark 2.6**, the ratio of training days to be 10 times the number of features does not validate for the real building. The readers are thus encouraged to perform the sensibility analysis for their given cases. However, the performance has a noticeable higher performance between 7 and 12 days, thus readers are suggested to use 12 days as relevant days for model training.

Selection on AI Models

The choice of AI model (ANN, SVM, BEDT and RF) depends on the choice of the relevant data selection methods (HDD, modified HDD, FD and DTW). The performances of model are evaluated using the scenario S8 suggested from intermediate recommendations.

The parameters of ANN are similar to Table (4.7) in **Chapter 4** and the parameters of SVM, BEDT and RF are similar to Table (4.9) except that training and testing datasets are different. In order to evaluate the model, datasets of October-December, 2012 and January 2013 are used for training and validation; and datasets of February, 2013 and February-April, 2014 are used for testing. The performances of different AI models using different relevant data selection method are shown in Table (5.4).

Models	HDD				Modified HDD				Frechet Distance				DTW			
	Median		Overall		Median		Overall		Median		Overall		Median		Overall	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
ANN	0.753	47.4	0.850	55	0.751	49.4	0.853	54.4	0.751	50.8	0.843	56.2	0.77	48.9	0.850	54.5
SVM	0.772	48.8	0.862	52.7	0.781	46.8	0.862	52.9	0.749	51.3	0.84	56.5	0.751	50.1	0.856	53.7
BEDT	0.653	48.7	0.833	58	0.707	53.8	0.822	59.9	0.729	53.8	0.833	58	0.652	51.5	0.819	60.5
RF	0.704	49.5	0.834	57.9	0.735	53.6	0.839	56.9	0.675	49.5	0.837	57.4	0.704	54.5	0.836	57.6

Table 5.4: Performance of different AI models using HDD, modified HDD, FD and DTW relevant data selection method

It can be seen that performance of ANN and SVM are higher compared to BEDT and RF illustrating that both BEDT and RF are sensitive with the training data. The most interesting result for this type of building are that the HDD method based on SVM has better performance (Median:

R^2=0.772, RMSE=48.8; Overall: R^2=0.862, RMSE=52.7) than the FD method. It is observed that simplified physical methods (modified HDD) has higher accuracy noticeably in median performance (Median: R^2=0.781, RMSE=46.5; Overall: R^2=0.862, RMSE=52.9) compared to other relevant data selection methods. This might be due to the weight effect introduced during a whole day that differentiates the degree of energy consumption profile. On the other hand, DTW relevant data selection method also provides reasonable accuracy compared to modified HDD method.

The prediction of some random days based on modified HDD and DTW "**relevant data**" modeling approach using SVM are shown in Figure (5.11). It is seen that both of the methods have similar performance except that error occured during initial hour.

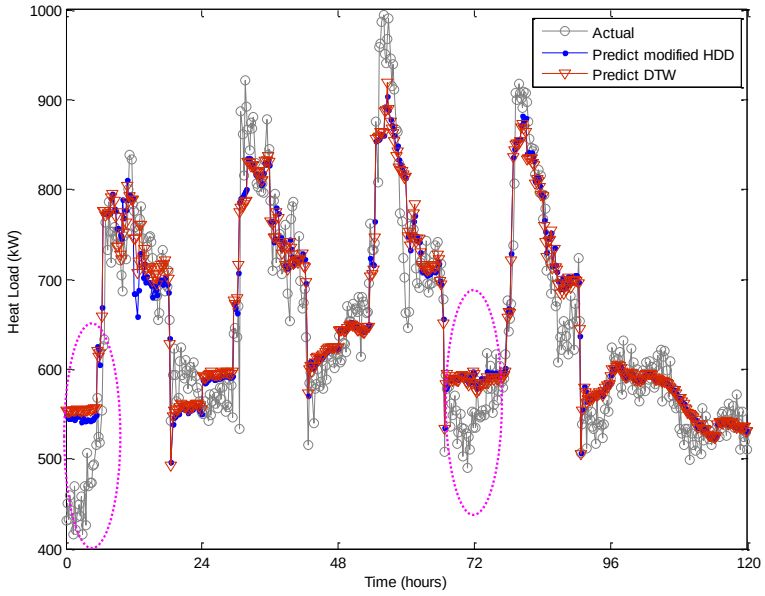

Figure 5.11: Prediction of heating load using different relevant data selections based on SVM (for some random days)

The model training CPU-time using different AI models for each prediction day using DTW relevant data selection is shown in Figure (5.12). It can be seen that the fastest model training methods are SVM and BEDT, whereas ANN and RF requires large model training CPU-time. The long model training process is due to the sampling time of the data and the time required for the optimization to find the parameters of model.

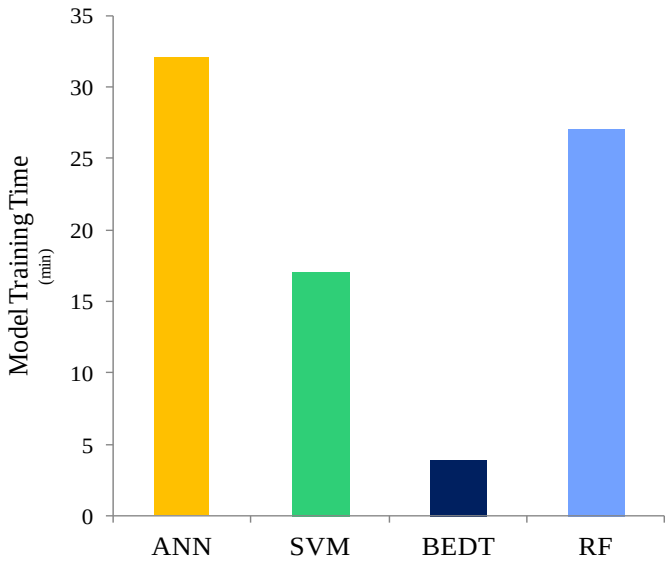

Figure 5.12: Model training CPU-time using different AI models using DTW

The model training CPU-time requirement from different "**relevant data**" modeling approach using SVM for a random prediction day is shown in Figure (5.13). It can be seen that all the "**relevant data**" modeling approaches requires similar model training CPU-time except than the HDD method. The few extra minutes requirement for HDD method might be due to the selection of relevant day for model training is different than other methods resulting in extra time for SVM to solve the optimization problem to find model parameters.

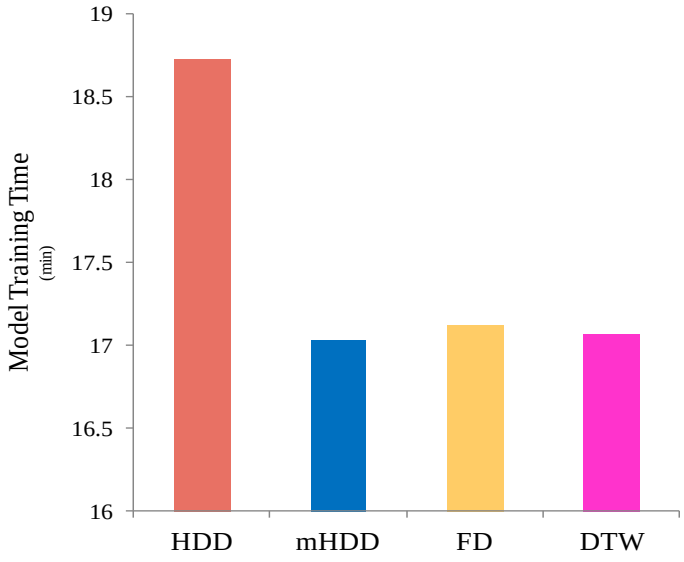

Figure 5.13: Model training CPU-time using different relevant data modeling approach using SVM

Intermediate Recommendations: The readers are suggested to use modified HDD or DTW method based on SVM or ANN as a reference model due to their higher performance. However, SVM method is more preferable due to faster model training CPU-time to predict heat load compared to ANN method.

Comparison between the Modeling Approaches: "All Data" and "Relevant Data"

The comparison between DTW "**relevant data**" modeling approach using SVM is performed with "**all data**" modeling approaches using ANN, SVM, BEDT and RF considering input features scenario S8 shown in Table (5.1). The parameters of ANN are defined similar to Table (4.7) in and SVM similar to Table (4.9) in **Chapter 4**. Similarly, the parameters of model in both BEDT and RF are defined similar to Table (4.9) in **Chapter 4** except that in both cases, number of trees are searched from [25, 50, 75,….,500] at increment of 25. The model comparisons of all methods for working day and weekend are shown in Table (5.5).

Building Functioning Type	Performances	DTW Relevant Data Modeling Approaches using SVM		All Data Modeling Approaches							
				ANN		SVM		BEDT		RF	
		Median	Overall	Median	Overall	Median	Overall	Median	Overall	Median	Overall
Working Day	R^2	0.80	0.86	0.55	0.84	0.56	0.85	0.54	0.83	0.53	0.78
	RMSE	51.5	57.6	91.2	60	90	58.9	92.7	63	93	71.8
	Model Training Time	15 min 3 sec		14 hour 18 min 5 sec		3 hour 15 min 45 sec		42 min 27 sec		6 hour 5 min 45 sec	
Weekend	R^2	0.63	0.82	0.41	0.75	0.48	0.81	0.33	0.66	0.39	0.70
	RMSE	36.3	42.7	53	50.1	44.1	43.9	50.1	59.2	52.7	55.3
	Model Training Time	12 min 48 sec		2 hour 57 min 12 sec		22 min 27 sec		26 min 14 sec		2 hour 29 min 9 sec	

Table 5.5: Comparison of model performance of DTW relevant data modeling approach using SVM with all data modeling approach using ANN, SVM, BEDT and RF

It is clearly seen that DTW "**relevant data**" modeling approach using SVM is superior (Working Day- Median: R^2=0.80, RMSE=51.5; Overall: R^2=0.86, RMSE=57.6; Weekends- Median: R^2=0.63, RMSE=36.3; Overall: R^2=0.82, RMSE=42.7) than "**all data**" modeling approaches. It is noticeably seen that weekend performances are relatively lower than the working day. This might be due to the very few data belonging to the weekend for model training. It is also seen that computation CPU-time in "**relevant data**" modeling approach is faster than "**all data**" modeling approaches. Thus, "**relevant data**" modeling approach are suitable for ESCOs and/or BEMS for control applications.

As an example, the prediction of some random test days for working days and weekend based on "**all data**" and DTW "**relevant data**" modeling approach using SVM is shown in Figure (5.14-5.15).

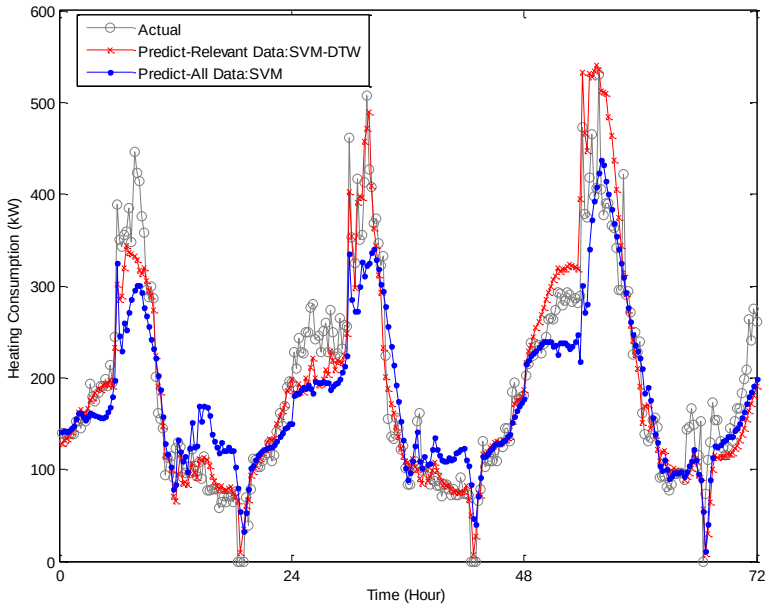

Figure 5.14: Prediction of heating load based on all data and DTW relevant data modeling approach using SVM for working days

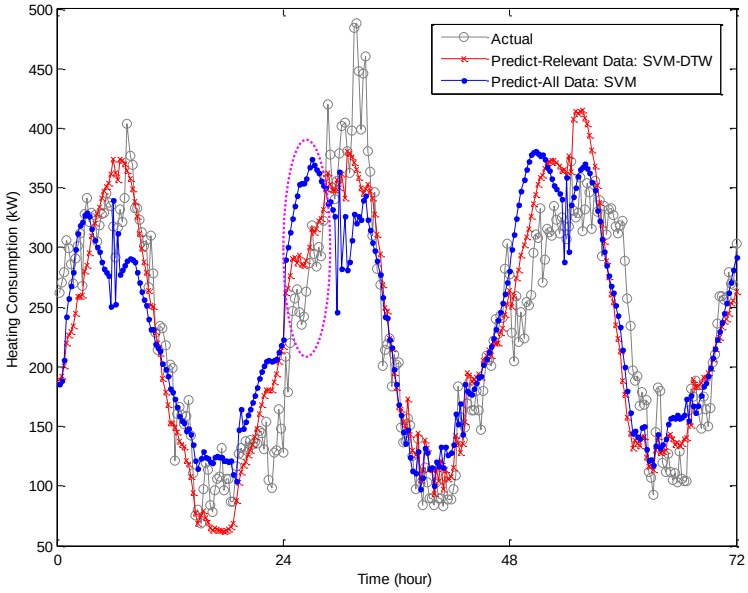

Figure 5.15: Prediction of heating load based on all data and DTW relevant data modeling approach using SVM for weekend

Figure (5.14-5.15) illustrated that "**relevant data**" modeling approach has higher performance compared to "**all data**" modeling approach for both working days and weekend. It is also noticed that both modeling approaches generalize quite well approximately after 12 till 24 hours for each prediction day, however, "**all data**" modeling approach fails to generalize quite well during an initial period (0-9) hours for each prediction day. This might be because "**all data**" modeling approaches focuses to generalize the model in terms of overall data and lacks the generality for specific hour prediction conditions, for example, during initial period (0-9) hours. In contrast, "**relevant data**" modeling approaches have almost learnt the heating energy consumption behavior during initial period since it considers selection of relevant days based on dynamic behavior of external temperature.

Intermediate Recommendations: The readers are suggested to use "**relevant data**" modeling approach using SVM compared to "**all data**" modeling approach for heating load prediction.

5.3 Conclusion

This chapter provides an application of methodology to mixed conventional and low energy office building for heating load prediction using "**all data**" and "**relevant data**" modeling approaches. The comparative analysis on different relevant data selections: HDD, modified HDD, FD and DTW and different AI models: ANN, SVM, BEDT and RF were also performed. It is found that modified HDD "**relevant data**" modeling approach using SVM performs better accuracy revealing the fact that these methods consider the weight effect to differentiate degree of energy consumption during a day. In addition, it is also noticed that HDD "**relevant data**" modeling approach using SVM is also suitable for the given studied building.

Furthermore, the model has been built with pseudo dynamic model to include dynamics of indoor air characteristics and results show greater accuracy while using such input features.

The next chapter will focus on summary of overall manuscript and the perspectives in the future.

Chapter 6: Summary and Future Works

6.1 Summary

The building energy load prediction is important for ESCOs and/or BEMS to manage the thermal energy demand for control and planning purposes. It not only helps to manage the energy consumption in buildings but also provides to reduce the green-house emission. This manuscript provides detail on energy demand modeling approach for LEBs.

Initially, we introduced the thermal energy performance measures of LEBs and its comparison with CBs. We highlighted several building characteristics: insulation, time constant, window/glazing etc. We then summarized the evolution of LEBs in Europe and noticed that most of the European countries are focusing to migrate towards VLEBs or PEBs. After that, we presented three kinds of building energy models: **white-box**, **gray-box** and **black-box** models (statistical linear regression and machine learning AI particularly ANN, SVM, BEDT and RF) to estimate and predict thermal energy demand. Then we made comparison of different models based on several factors: input data, modeler experience, simplicity of calibration (in terms of input use), training data, model training CPU-time, requirement of building physical information, accuracy etc. These review work drawn a conclusion that both **white-box** and **gray-box** model require many input data of building since they are based on physical principles. Moreover, sometimes all these physical information of building are not applicable for ESCOs and/or BEMS during the operation phase of building. This further justified that machine learning based **AI black-box** models seem more suitable because of their requirement of few input information and their capacity to adapt the model in the future unknown environment.

Then, we went through deep understanding of different AI models and found that two kinds of modeling approaches: "**all data**" and "**relevant data**" exists. We found that "**all data**" approach which uses all available data for model training has been numerously applied in literatures. However, "**all data**" modeling approach has several drawbacks due to the redundancy of input data, complexities in model building CPU-time and to update the model parameters in future environment. On the other hand, we found that the "**relevant data**" modeling approach uses few representative day data to build a model. Noticeably, "**relevant data**" modeling approach has been extensively applied in electrical load but not applied in thermal load for LEBs. Although the

methodologies that applied for electrical load have advantages due to small representative data selection. Nevertheless, "**relevant data**" modeling approach have still some limitations to consider solar gain and past day climatic conditions influences. Also, the previous studies focus on daily average energy load of prediction day or previous day to select representative day data and these are not adapted for LEBs. If the prediction model is not only for a day ahead but for longer periods, then the prediction methods will rely on previously predicted daily averaged energy load to select database and errors might be accumulated.

We thus addressed the complexity in considering few representative day data for model training in LEBs by considering deviation criteria between prediction day and training database depending on their past day climatic conditions influences. These deviation criteria are based simple physical understanding: HDD and modified HDD, and pattern recognition methods: FD and DTW. Then we determined the wavelet coefficients of climatic conditions to calculate their influences on building load. After that, these wavelet coefficients are combined with the deviation criteria to get one metric criterion to select representative days for model training. Before developing such metric criteria, we developed several sub-modules as a pre-processing step to build an **AI model**. Firstly, we built "**Building Operation Classification/Clustering**" module to identify the functioning classes of building operation profile (during a week for example). Secondly, we developed novel "**Pseudo Dynamic Model**" to introduce a priori knowledge on the dynamic behavior of the building. Thirdly, we generated derived climatic variables to consider the thermal storage effects on the walls. Lastly, we built "**Climatic Variables Selection**" module to determine significant direct and derived climatic variables and their dynamics.

We then applied the methodology using two kinds of modeling approaches: "**all data**" and "**relevant data**" to large simulated CBs and LEBs. The building data were generated from simulation tools TRNsys. The methodology is tested step-by-step. Firstly, we applied our methodology to single-zone CB and LEB model. We identified that past climatic conditions: external temperature and solar gains on walls have significant impact for all types of building. In case of CB, the past 1-2 days of these climatic conditions are important and past 3 days are significant for LEBs. Secondly, we performed a comparison on relevance of "**SVM based on Linear Kernel**" and "**LSM based on Regression**" for analysis of the climatic variables influence on building load. We found that "**SVM based on Linear Kernel**" is more suitable for LEBs. Thirdly, we investigated several input feature scenarios and identified that occupancy, pseudo dynamic transitional effects and derived climatic features has a greater influence in different buildings. We also found that impact of solar gain is increasing when the building is migrating from CB to LEBs. Fourthly, we performed the sensibility analysis on the influence of the number

of days for model training and identified that good performance can be achieved between 7 and 14 days. Fifthly, we performed comparison on choice of AI models (ANN, SVM, BEDT and RF) and choice of relevant data selection method (HDD, modified HDD, Frechet Distance and DTW) for model training. We identified that modified HDD or DTW method based on SVM can be taken as a reference AI model to predict thermal energy demand because of their very high prediction accuracy and faster model training CPU-time. It is also noticed that modified HDD method are more preferable for CBs and DTW method for LEBs. Sixthly, we made a comparisons between two kinds of modeling approaches: "**all data**" and "**relevant data**" and it is observed that DTW relevant data selection using SVM is better performances than "**all data**" approach for different AI models. Seventhly, we tested the performance of prediction model with different occupancy and found that proposed methodology can guaranteed very high prediction accuracy. We recognized that building operating conditions and pseudo dynamic model has greater effects in the prediction performance. Finally, the multi-zone building model is examined and the methodology has guaranteed high accuracy to predict heat load for multi-zone using DTW relevant data selection using SVM. The major difference while using multi-zone model compared to single-zone model is that we have to consider the changing period of occupancy and building operating conditions of multi-zone into aggregated one-zone to formulate the transition in the pseudo dynamic model.

After that, we applied the methodology in real mixed CB and LEB at Ecole des Mines de Nantes. We evaluated the methodology step-by-step similar to simulation building. Firstly, we made a comparison between "SVM based on Linear Kernel" and "LSM based on Regression Model" for weight determination and found that they have similar weights. Secondly, we tested on different input feature scenarios and found that occupancy and pseudo dynamic model has a significant effect in the model performance. Thirdly, we performed the sensibility analysis on the influence of the number of days for model training and observed that the performance of model is higher between 7 and 12 days. Fifthly, we compared different AI models (ANN, SVM, BEDT and RF) and relevant data selection methods (HDD, modified HDD, Frechet Distance and DTW) and noticed that modified HDD or DTW using SVM and ANN have higher performance. However, SVM method is more preferable due to its faster model training CPU-time to predict thermal heat load. Finally, we compared "**all data**" and "**relevant data**" modeling approaches and identified that DTW "**relevant data**" modeling approach using SVM has higher performance to "**all data**" modeling approaches for heating load prediction.

6.2 Future Works

There are also several research problems in energy demand or consumption prediction as a future steps and these are summarized below:

- Develop a criterion to combine different relevant data selections methods and AI models. For example, combination of modified HDD and DTW relevant data selection methods and combination of AI models: ANN and SVM.

- Develop an automatic feature selection method to identify important features during the operation phase of the building.

- Develop a methodology to learn the behavior in one building and apply it to unknown buildings that have different physical and geometrical parameters. In this research, the methodology is aim on known building where physical and geometrical properties are provided and make prediction under various climatic conditions.

Bibliography

[1] J. Laustsen, "Energy efficiency requirements in building codes-Policies for new buildings," International Energy Agency, OECD/IEA, 2008.

[2] International Energy Agency, "Energy effiiciency market report: Market trends and medium-term prospects," 2015.

[3] C. Hopfe and R. Mcleod, The passivehaus designer's manual: a technical guide to low and zero energy buildings, Taylor & Francis, 2015.

[4] Service de l'observation et des statistiques, "Chiffres clés de l'energie," Edition 2014, 2014.

[5] European Commission, Low energy buildings in Europe: Current state of play, definitions and best practice, Brussels, 2009.

[6] D. Mumovic and M. Santamouris, A handbook of sustainable building design and engineering: An integrated approach to energy, health and operational performance, Taylor & Francis, 2009.

[7] K. Thomsen et K. Wittchen, European national strategies to move towards very low energy buildings, Danish Building Research Institute, Aalborg University, 2008.

[8] A. Hermelink, S. Schimschar, T. Boermans, L. Pagliano, P. Zangheri, R. Armani, K. Voss and E. Musall, Towards nearly zero-energy buildings: Defination of common priniciples under the EPBD., Final Report, 2012.

[9] P. Torcellini, S. Pless and M. Deru, Zero energy buildings: A criticial look at the definition, Pacific Grove, California: ACEEE Summer Study, 2006.

[10] W. Pessenlehner and A. Mahdavi, "Building morphology, transparence and energy performance," in *Eighth International IBPSA Conference*, Eindhoven, Netherlands, 2003.

[11] S. Sadineni, S. Madala and R. Boehm, "Passive building energy savings: A review of building envelope components," *Renewable and Sustainable Energy Reviews*, vol. 15, pp. 3617-3631, 2011.

[12] E. Giueseppe, Nearly zero energy buildings and proliferation of micro-organisms: A current issue for highly insulated and airtight building envelopes, Springer Briefs in Applied Science and Technology, 2013.

[13] R. Mcmullan, Environmental science in building, 7th Edition, Palgrave Macmillan, 2012.

[14] D. Li, L. Yang and J. Lam, "Zero energy buildings and sustainable development implications - A review," *Energy*, vol. 54, pp. 1-10, 2013.

[15] M.-L. Persson, A. Roos and M. Wall, "Influence of window size on the energy balance of low energy houses," *Energy and Buildings*, vol. 38, pp. 181-188, 2006.

[16] A. Pecourt, "Inertie dans le batiements passifs - constante de temps," 2015. [Online]. Available: http://www.energelio.fr/documentation/Pause_cafe/IsolationxInertiecste_temps_Passibat_2014.pdf. [Accessed 12/ 10/ 2015].

[17] L. Swan and V. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques," *Renewable and Sustainable Energy Reviews*, vol. 13, pp. 1819-1835, 2009.

[18] H. Zhao and F. Magoules, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews* , vol. 16, pp. 3586-3592, 2012.

[19] Z. Li, Y. Han and P. Xu, "Methods for benchmarking building energy consumption against its past or intended performance: An overview," *Applied Energy*, vol. 124, pp. 325-334, 2014.

[20] A. Foucquier, S. Robert, F. Suard, L. Stéphan and A. Jay, "State of art in building modeling and energy performances prediction: A review," *Renewable and Sustainable Energy Reviews*, vol. 23, pp. 272-288, 2013.

[21] X. Li and J. Wen, "Review of building energy modeling for control and operation," *Renewable and Sustainable Energy Reviews*, vol. 37, pp. 517-537, 2014.

[22] F. Amara, K. Agbossou, A. Carnenas, Y. Dubé and S. Kelouwani, "Comparison and simulation of building thermal models for effective energy management," *Smart Grid and Renewable Energy*, vol. 6, pp. 95-112, 2015.

[23] D. Coakley, P. Raftery and M. Keane, "A review of methods to match building energy simulation models to measured data," *Renewable and Sustainable Energy Reviews*, vol. 37, pp. 123-141, 2014.

[24] T. Berthou, "Développement de modèles de batiment pour la prévision de charge de climatisation et l'élaboration de stratégies d'optimisation énergetique et d'effacement," in *Thèse soutenue* , Doctorat Paris Tech, 2013.

[25] M. Al-Homoud, "Computer aided building energy analysis techniques," *Building and Environment*, vol. 36, pp. 421-433, 2001.

[26] P. Ma and N. Guo, "Modeling of thermal mass in a small commercial building and potential improvement by applying TABS," *American Journal of Mechanical Engineering,* vol. 3, no. 2, pp. 55-62, 2015.

[27] Y. Cengel and A. Ghajar, Heat and mass transfer: Fundamentals and applications, Fifth Edition, McGraw Hill Education, 2015.

[28] D. Crawley, L. Lawrie, F. Winkelmann, W. Buhl, Y. Huang, C. Pedersen, R. Strand, R. Liesen, D. Fisher, M. Witte and J. Glazer, "EnergyPlus: creating a new generation building energy simulation programs," *Energy and Buildings,* vol. 33, pp. 319-331, 2001.

[29] S. Citherlet, "Towards the holistic assessment of building performance based on integrated simulation approach," in *PhD Thesis,* Swiss Federal Institute of Technology, 2001.

[30] A. Kalagasidis, P. Weitzmann, T. Nielsen, R. Peuhkuri, C.-E. Hagentoft and C. Rode, "The international building physics toolbox in simulink," *Energy and Buildings,* vol. 39, pp. 665-674, 2007.

[31] A. Husaunndee, R. Lahrech, H. Vaezi-Nejad and J. Visier, "SIMBAD: A simulation toolbox for the design and test of HVAC control systems," in *Proceedings of the 5th IBPSA Conference*, 1997.

[32] D. Crawley, J. Hand, M. Kummert and B. Griffith, "Contrasting the capabilities of building energy performance simulation programs," *Building and Environment,* vol. 43, pp. 661-673, 2008.

[33] S. Wang and Y. Chen, "Transient heat flow calculation for multilayer constructions using a frequency-domain regression method," *Building and Environment,* vol. 38, pp. 45-61, 2003.

[34] P. Bacher, H. Madsen, H. Nielsen and B. Perers, "Short-term heat load forecasting for single family houses," *Energy and Buildings,* vol. 65, pp. 101-112, 2013.

[35] O. Ogunsola and L. Song, "Application of a simplified thermal network model for real-time thermal load estimation," *Energy and Buildings,* vol. 96, pp. 309-318, 2015.

[36] S. Wang and X. Xu, "Parameter estimation of internal thermal mass of building dynamic models using genetic algorithm," *Energy Conversion and Management,* vol. 47, pp. 1927-1941, 2006.

[37] T. Berthou, P. Stabat, R. Salvazet and D. Marchio, "Development and validation of a gray box model to predict thermal behavior of occupied office buildings," *Energy and Buildings,* vol. 74, pp. 91-100, 2014.

[38] X. Lu, T. Lu, C. Kibert and M. Viljanen, "Modeling and forecasting energy consumption for heterogeneous building using a physical-statistical approach," *Applied Energy,* vol. 144, pp. 261-275, 2015.

[39] J. Lam, K. Wan, S. Wong and T. Lam, "Principal component analysis and long-term building energy simulation correlation," *Energy Conversion and Management,* vol. 51, pp. 135-139, 2010.

[40] T. Olofsson, S. Andersson and R. Ostin, "A method of predicting the annual building heating demand based on limited performance data," *Energy and Buildings,* vol. 28, pp. 101-108, 1998.

[41] T. Olofsson and S. Andersson, "Long-term energy demand predictions based on short-term measured data," *Energy and Buildings,* vol. 33, pp. 309-318, 2001.

[42] H. Chaowen and W. Dong, "Prediction of hourly cooling load of buildings based on neural networks," *International Journal of Smart Home,* vol. 9, pp. 35-32, 2015.

[43] K. Wan, D. Li, D. Liu and J. Lam, "Future trends of building heating and cooling loads and energy consumption in different climates," *Building and Environment,* vol. 46, pp. 223-234, 2011.

[44] X. Li, D. Lixing, L. Jinhu, X. Gang and L. Jibin, "A novel hybrid approach of KPCA and SVM for building cooling load prediction," in *IEEE Third International Conference on Knowledge Discovery and Data Mining*, Phuket, Thailand, 2010.

[45] R. Yokoyama, T. Wakui and R. Satake, "Prediction of energy demands using neural network with model identification by global optimization," *Energy Conversion and Management,* vol. 50, pp. 319-327, 2009.

[46] C. Deb, L. Eang, J. Yang and M. Santamouris, "Forecasting diurnal cooling energy load for institutional buildings using artificial neural networks," *Energy and Buildings,* vol. xx, p. xx, 2015.

[47] J. Lam, K. Wan, K. Cheung and L. Yang, "Principal component analysis of electricity use in office buildings," *Energy and Buildings,* vol. 40, pp. 828-836, 2008.

[48] N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou and P. Patargias, "Using principal component and cluster analysis in the heating evaluation of the school building sector," *Applied Energy,* vol. 87, pp. 2079-2086, 2010.

[49] X. Li, C. Bowers and T. Schnier, "Classification of energy consumption in buildings with outlier detection," *IEEE Transactions on Industrial Electronics,* vol. 57, pp. 3636-3644, 2010.

[50] X. Gao and A. Malkawi, "A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm," *Energy and Buildings,* vol. 84, pp. 607-616, 2014.

[51] M. Santamouris, G. Mihalakakou, P. Patargias, N. Gaitani, K. Sfakianaki, M. Papaglastra, C. Pavlou, P. Doukas, E. Primikiri, V. Geros, M. Assimakopoulos, R. Mitoula and S. Zerefos, "Using intelligent clustering techniques to classify the energy performance of school buildings," *Energy and Buildings,* vol. 39, pp. 45-51, 2007.

[52] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[53] R. May, G. Dandy and H. Maier, "Artificial neural networks - Methodological advances and biomedical applications," in *Chapter-2 Review of input variable selection methods for artifical neural networks*, 2011.

[54] H. Zhao and F. Magoules, "Feature selection for predicting building energy consumption based on statistical learning method," *Journal of Algorithms & Computational Technology,* vol. 6, pp. 59-78, 2012.

[55] A. Kusiak, M. Li and Z. Zhang, "A data-driven approach for steam load prediction in buildings," *Applied Energy,* vol. 87, pp. 925-933, 2010.

[56] R. Jovanovic, A. Stretenovic and B. Zivkovic, "Ensemble of various neural networks for prediction of heating energy consumption," *Energy and Buildings,* vol. 94, pp. 189-199, 2015.

[57] Y.-M. Zhang and W.-G. Qi, "Interval forecasting for heating load using support vector regression and error correcting markov chains," in *IEEE Eighth International Conference on Machine Learning and Cybernetics,* Baoding, China, 2009.

[58] S.-H. Cho, W.-T. Kim, C.-S. Tae and M. Zaheeruddin, "Effect of length of measurement period on accuracy of predicted annual heating energy consumption of buildings," *Energy Conversion and Management,* vol. 45, pp. 2867-2878, 2004.

[59] T. Catalina, J. Virgone and E. Blanco, "Development and validation of regression models to predict monthly heating demand for residential buildings," *Energy and Buildings,* vol. 40, pp. 1825-1832, 2008.

[60] K. Yun, R. Luck, P. Mago and H. Cho, "Building hourly thermal load prediction using an indexed ARX model," *Energy and Buildings,* vol. 54, pp. 225-233, 2012.

[61] A. Ben-Nakhi and M. Mahmoud, "Cooling load prediction for buildings using general

regression neural networks," *Energy Conversion and Management,* vol. 45, pp. 2127-2141, 2004.

[62] S. Kalogirou, G. Florides, C. Neocleous and C. Schizas, "Estimation of daily heating and cooling loads using artificial neural network," in *World Congress*, Napoli, 2001.

[63] C.-W. Yan and J. Yao, "Application of ANN for the prediction of building energy consumption at different climate zones with HDD and CDD," in *IEEE Second International Conference on Future Computer and Communication*, Wuhan, China, 2010.

[64] S. Naji, A. Keivani, S. Shamshirband, U. Alengaram, M. Jumaat, Z. Mansor and M. Lee, "Estimating building energy consumption using extreme machine learning methods," *Energy*, vol. 97, pp. 506-516, 2016.

[65] S. Kalogirou and M. Bojic, "Artificial neural networks for the prediction of the energy consumption of a passive solar building," *Energy*, vol. 25, pp. 479-491, 2000.

[66] S. Kwok and E. Lee, "A study of the importance of occupancy to building cooling load prediction by intelligent approach," *Energy Conversion and Management,* vol. 52, pp. 2555-2564, 2011.

[67] G. Tso and K. Yau, "Predicting electricity energy consumption: A comparison of regression, decision tree and neural networks," *Energy*, vol. 32, pp. 1761-1768, 2007.

[68] B. Ekici and U. Aksoy, "Prediction of building energy consumption by using artificial neural networks," *Advances in Engineering Software*, vol. 40, pp. 356-362, 2009.

[69] A. Neto and F. Fiorelli, "Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption," *Energy and Buildings,* vol. 40, pp. 2169-2176, 2008.

[70] Q. Li, Q. Meng, J. Cai, H. Yoshino and H. Mochida, "Applying support vector machine to predict hourly cooling load in the building," *Applied Energy,* vol. 86, pp. 2249-2256, 2009.

[71] Z. Yu, F. Haghighat, B. Fung and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy and Buildings,* vol. 10, pp. 1637-1646, 2010.

[72] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings,* vol. 49, pp. 560-567, 2012.

[73] J.-S. Chou and D.-K. Bui, "Modeling heating and cooling loads by artificial intelligence for energy-efficient building design," *Energy and Buildings,* vol. 82, pp. 437-446, 2014.

[74] C. Fan, F. Xiao and S. Wang, "Development of prediction models for next-day building

energy consumption and peak power demand using data mining techniques," *Applied Energy*, vol. 127, pp. 1-10, 2014.

[75] L. Xuemei, D. Lixing, S. Ming, X. Gang and L. Jibin, "A novel air-conditioning load prediction based on ARMA and BPNN model," in *IEEE Asia Pacific Conference on Information Processing*, Shenzhen, China, 2009.

[76] K. Li, H. Su and J. Chu, "Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study," *Energy and Buildings*, vol. 43, pp. 2893-2899, 2011.

[77] X. Wang and M. Meng, "A hybrid neural network and ARIMA model for energy consumption forecasting," *Journal of Computers*, vol. 7, pp. 1184-1190, 2012.

[78] B. Withdrow and M. Kamenetsky, "Statistical efficiency of adaptive algorithms," *Neural Networks*, vol. 16, pp. 735-744, 2003.

[79] Y. Chen, P. Luh, C. Guan, Y. Zhao, L. Michel, M. Coolbeth, P. Friedland and S. Rourke, "Short-term load forecasting: Similar day based wavelet neural networks," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 322-330, 2010.

[80] C. Sun, J. Song, L. Li and P. Ju, "Implementation of hybrid short-term load forecasting system with analysis of temperature sensitivities," *Soft Computing*, vol. 12, no. 7, pp. 633-638, 2008.

[81] T. Senjyu, H. Sakihara and Y. Tamaki, "Next day load curve forecasting using neural network based on similarity," *Electric Power Components and System*, vol. 29, no. 10, pp. 939-948, 2001.

[82] A. Jain, P. Singh and K. Singh, Short-term load forecasting using fuzzy interference and ant colony optimization, Swarm, Evolutionary and Memetic Computing, Lecture Notes in Computer Science, Springer, 2011.

[83] Q. Mu, Y. Wu, X. Pan, L. Huang and X. Li, "Short-term load forecasting using improved similar days method," in *IEEE Asia Pacific Power and Energy Conference*, Chengdu, China, 2010.

[84] P. Mandal, T. Senjyu, N. Urasaki and T. Funabashi, "A neural network based several hour ahead electrical load forecasting using similar days approach," *Electrical Power and Energy Systems*, vol. 28, pp. 367-373, 2006.

[85] Y.-J. He, Y.-C. Zhu, J.-C. Gu and C.-Q. Yin, "Similar day selecting based neural network model and its application in short term load forecasting," in *Proceedings of the IEEE*

Fourth International Conference on Machine Learning and Cybernetics , Guangzhou, China, 2005.

[86] C. Roldan-Blay, G. Escriva-Escriva, C. Alvarez-Bel and C. Roldan-Porta, "Upgrade of an artifical neural network prediction method for electrical consumption of forecasting using an hourly temperature curve model," *Energy and Buildings*, vol. 60, pp. 38-46, 2013.

[87] A. Jain and B. Satish, "Clustering based short-term load forecasting using support vector machines," in *IEEE Burcharest Power Tech Conference*, Bucharest, Romania, 2009.

[88] A. Ghanbari, S. Ghaderi and M. Azadeh, "A clustering based genetic fuzzy expert system for electrical energy demand prediction," in *Second International Conference on Computer and Automation Engineering*, Singapore, 2010.

[89] F. Pasila, "Multivariate inputs for electrical load forecasting on hybrid neuro-fuzzy and fuzzy c-means forecaster," in *IEEE World Congress on Computational Intelligence*, Hong Kong, 2008.

[90] V. Yadav and D. Srinivasan, "Autocorrelation based weighing strategy for short-term load forecasting with the self-organizing map," in *Second International Conference on Computer and Automation Engineering*, Singapore, 2010.

[91] W. Sun, "Ant colony based feedforward NN short-term load forecasting model with input selection and DA clustering," in *Fourth IEEE International Conference on Natural Computation*, Jinan, China, 2008.

[92] D.-X. Duan, "Short-term load prediction based on ant colony clustering elman neural network model," in *Second International Workshop on Computer Science and Engineering*, Qingdao, China, 2009.

[93] F. Marin, F. Garcia-Lagos, G. Joya and F. Sandoval, "Global model for short-term load forecasting using artificial neural networks," in *IEEE Proceeding on Generation, Transmission and Distribution*, 2002.

[94] M. Grzenda and B. Macukow, "Demand prediction with multi-stage neural processing," *Advances in Natural Computation and Data Mining*, pp. 131-141, 2006.

[95] S. Tafreshi and M. Farhadi, "Improved SOM based method for short-term load forecast of Iran power network," in *8th International Power Engineering Conference*, Singapore, 2007.

[96] Y. Sun, Y. Wang and F. Xiao, "Development and validation of a simplified online cooling load prediction strategy for a super high-rise building in Hong Kong," *Energy Conversion and Management*, vol. 68, pp. 20-27, 2013.

[97] P. Gonzalez and J. Zamarreno, "Prediction of hourly energy consumption in buildings based on a feedback artificial neural network," *Energy and Buildings,* vol. 37, pp. 595-601, 2005.

[98] J. Yang, H. Rivard and R. Zmeureanu, "Online building energy prediction using adaptive artificial neural networks," *Energy and Buildings,* vol. 37, pp. 1250-1259, 2005.

[99] J. Nazarko, A. Jurczuk and W. Zalewski, "ARIMA models in load modeling with clustering approach," in *IEEE Power Tech Conference*, St. Petersburg, Russia, 2005.

[100] G. Box, G. Jenkins and G. Reinsel, Time series analysis: Forecasting and Control, 4th Edition, John Wiley & Sons, 2008.

[101] F. Iglesias and W. Kastner, "Analysis of similarity measures in time series clustering for the discovery of building energy patterns," *Energies,* vol. 6, pp. 579-597, 2013.

[102] E. Keogh and C. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems,* vol. 7, pp. 358-386, 2005.

[103] T. Wylie and B. Zhu, "Following a curve with the discrete Frechet distance," *Theoretical Computer Science,* vol. 556, pp. 34-44, 2014.

[104] S. Mallat, "A theory of multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 11, no. 7, pp. 674-693, 1989.

[105] W. Yu, H. He and N. Zhang, Advances in neural networks- ISSN, in: Fourth international symposium on neural networks, New York: Springer-Berlin , 2009.

[106] S. Haykin, Neural Network: A comprehensive Foundation, Prentice Hall, 1999.

[107] K.-L. Du and M. Swamy, Neural networks and statistical learning, Springer, 2014.

[108] L. Breiman, J. Friedman, C. Stone and R. Olshen, Classification and regression trees, New York: Chapman & Hall, 1984.

[109] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: Data mining, inference and prediction, Second Edition, Springer, 2013.

[110] J. Freidman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics,* vol. 29, no. 5, pp. 1189-1232, 2001.

[111] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[112] O. Dombayci, "The prediction of heating energy consumption in a model house by using artificial neural networks in Denizli-Turkey," *Advances in Engineering Software,* vol. 41, pp. 141-147, 2010.

[113] M. Kolokotroni, M. Davies, B. Croxford, S. Bhuiyan and A. Mavrogianni, "A validated methodology for the prediction of heating and cooling energy demand for buildings within the Urban Heat Island: Case-study of London," *Solar Energy*, vol. 84, pp. 2246-2255, 2010.

[114] G. Krese, M. Prek and V. Butala, "Analysis of building electricity consumption data using an improved cooling degree day method," *Journal of Mechanical Engineering*, vol. 58, pp. 107-114, 2012.

[115] A. Zilouchian and M. Jamshidi, Intelligent control system using soft computing methodologies, CRC Press, 2001.

[116] G.-B. Huang, "Learning capability and storage capacity of two hidden-layer feedforward networks," *IEEE Transactions on Neural Networks*, vol. 14, no. 2, pp. 274-281, 2003.

[117] V. Vapnik, The nature of statistical learning theory, Springer, 1995.

[118] B. Scholkopf, C. Burges and A. Smola, Advances in kernel methods: support vector learning, MIT Press, 1999.

[119] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, 2001.

[120] K. Pelckmans, J. Suykens, T. Vangestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor and J. Vandewalle, "LS SVMlab: a MATLAB/c toolbox for least square support vector machines," Tutorial, KU Leuven-ESAT, Leuven, 2002.

[121] T. Joachims, "Making large-scale SVM learning practical," *Advances in kernel methods: support vector learning*, pp. 169-184, 1999.

[122] B. Scholkopf, A. Smola, R. Williamson and P. Barlett, "New support vector algorithm," *Natural Computation*, vol. 12, pp. 1207-1245, 2000.

[123] M. van Wezel, M. Kagie and R. Potharst, "Boosting the accuracy of pricing models," http://repub.eur.nl/pub/7145/ei2005-50.pdf, 2005.

[124] J. Quinlan, "Introduction of decision trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.

[125] J. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann, 1993.

[126] Z.-H. Zhou, Ensemble methods: Foundations and algorithms, Chapman & Hall/CRC, Machine learning & pattern recognition series , 2012.

[127] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[128] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an

application to boosting," *Journal of Computer and System Sciences,* vol. 55, no. 1, pp. 119-139, 1997.

[129] E. B., "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics,* vol. 7, no. 1, pp. 1-26, 1979.

[130] L. Breiman, "Heuristics of instability and stabilization in model selection," *The Annals of Statistics,* vol. 24, no. 6, pp. 2350-2383, 1996.

[131] P. Buhlmann and B. Yu, "Analyzing bagging," *The Annals of Statistics,* vol. 30, no. 4, pp. 927-961, 2002.

[132] 1. Breiman, "Using adaptive bagging to debias regressions," Technical Report no-547, University of California, Berkeley, 1999.

[133] R. Schapire, "The strength of weak learnability," *Machine learning,* vol. 5, pp. 197-227, 1990.

[134] J. Freidman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis - Non-linear methods and data mining,* vol. 38, no. 14, pp. 367-378, 2002.

[135] P. Buhlmann and T. Hothorn, "Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science,* vol. 22, no. 4, pp. 477-505, 2007.

[136] D. Saha, P. Alluri and A. Gan, "Accidental analysis and preventation," *Boosted trees for ecological modeling and prediction,* vol. 88, no. 1, pp. 243-251, 2007.

[137] M. Van Wezel, M. Kagie and R. Potharst, "Boosting the accuracy of pricing models," http://repub.eur.nl/pub/7145/ei2005-50.pdf, 2005.

[138] K. Priddy and P. Keller, Artificial neural networks: An introduction, SPIE, 2005.

[139] P. Burman, "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika,* vol. 76, no. 3, pp. 503-514, 1989.

[140] A. Molinaro, R. Simon and R. Pfeiffer, "Prediction error estimation: a comparison of re-sampling methods," *Bioinformatics,* vol. 21, no. 15, pp. 3301-3307, 2005.

Appendix A- Steady State Model

The steady state model can be classified into two categories:

- Degree-hour or day method
- Bin method

The degree-hour or day method considers an energy requirement of building is due to the difference between external temperature and base temperature of building. This method assumes that average heat gains from solar radiation and internal gains is balanced by heat loss due to fixed mean daily external temperature [25]. Figure (A.1) shows the base temperature of building and it can be seen that if the base temperature of building is higher than the external temperature then there is necessity of heating energy consumption. On the contrary, if the base temperature is below the external temperature then there is necessity of cooling energy consumption.

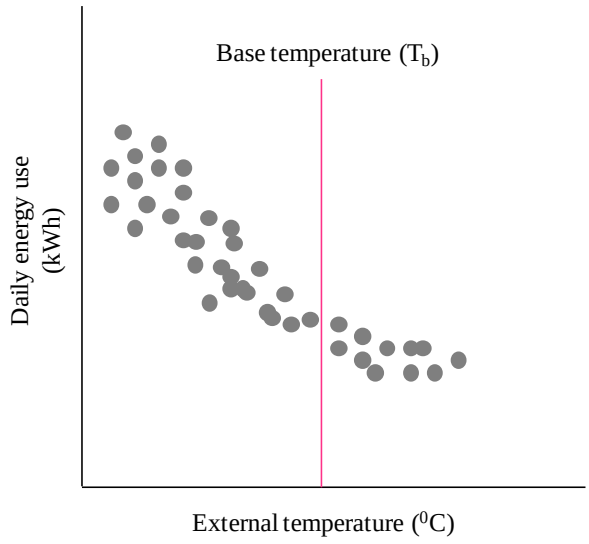

Fig. A.1: Illustration of base temperature

The degree-hour method is given by [112]:

$$DD_h(t) = (1\ \text{day}) \sum_{1}^{24} \left(T_{fb,buil} - T_{air,e}(t)\right)^+ \quad (A.1)$$

$$DD_h(t) = (1\ \text{h}) \left(T_{fb,buil} - T_{air,e}(t)\right)^+ \quad (A.2)$$

$$DD_c(t) = (1\ \text{h}) \left(T_{air,e}(t) - T_{fb,buil}\right)^+ \quad (A.3)$$

Where, $DD_h(t)$ is heating degree-hour (^{0}C/h), $DD_c(t)$ is cooling degree-hour (^{0}C/h), $T_{fb,buil}$ is fixed base temperature of building (^{0}C) (usually 18^{0}C in Central Europe) and $T_{air,e}(t)$ is external temperature (^{0}C). In Equation (A.1 –A.3), + indicates that calculation is valid when $\left(T_{b,buil} - T_{air,e}(t)\right)$ difference is positive or zero. Then, the heating energy consumption is given by Equation (A.4) [112] and cooling energy consumption is given by Equation (A.5) [113]:

$$E_h(t) = \frac{DD_h(t) * UA_{buil}}{\eta_h} \quad (A.4)$$

$$E_c(t) = \frac{\dot{m}_{air,in} \text{ x } c_{p,air,in} \text{ x } DD_c(t)}{COP_c} \quad (A.5)$$

Where, $E_h(t)$ is heating energy consumption of building (kWh), UA_{buil} is overall building heating loss coefficient (W/K), η_h is overall seasonal average efficiency of heating equipment (%), $E_c(t)$ is cooling energy consumption of building (kWh), $\dot{m}_{air,in}$ is mass flow rate of environmental indoor air (kg/s), $c_{p,air,in}$ is specific heat capacity of indoor air (kJ/kg) and COP_c is the overall coefficient of performance of cooling system (%). The overall heat loss coefficient of building (UA_{buil}) further given by [113]:

$$UA_{buil} = \frac{A_{buil} \text{ x } U_{buil} + \frac{1}{3} \text{x } N_{r,inf} \text{ x } V_{buil}}{1000} \quad (A.6)$$

Where, A_{buil} is area of building (m^2), U_{buil} is total U-value of building envelope components (W/m^{2}K) $N_{r,inf}$ is air infiltration rate changes per hour (h^{-1}) and V_{buil} is volume of building (m^3). In the Equation (A.6), the numerical values 1/3 represent typical values of density and specific heat of indoor air and conversion to air changes per hour.

However, degree-hour or day method that uses fixed base temperature of building does not consider internal heat gains and does not include variation of temperature on the performance of equipment. By considering fixed based temperature, it lags physical fundamental principle which depends on several factors such as insulation level, materials composition, internal and solar heat gains, desired set-point temperature and occupant's behavior ([25];[114]).

To overcome the limitation of degree-hour or day-method, variable degree-hour or day method exist in literature and it assumes heat gain or heat loss of building is balanced by variable base

temperature instead of fixed base temperature. The variable base temperature of building $T_{vb,buil}(t)$ in ^0C is given by [25]:

$$T_{vb,buil}(t) = T_{set-point}(t) - \frac{\dot{Q}_{sol}(t) + \dot{Q}_{int}(t)}{UA_{buil}} \quad (A.7)$$

Where, $T_{set-point}(t)$ is set-point temperature of building (^0C), $\dot{Q}_{sol}(t)$ is solar heat gain in the building (W) and $\dot{Q}_{int}(t)$ is internal heat gain in the building (W). Nevertheless, these methods are also not precise for building dominated by internal gains and low U-value of envelope components.

On the other hand, bin method is similar to variable base temperature but it rely functions of bins in terms of closeness using external temperature to estimate total building heating and cooling energy consumption. The average values of external temperature bins is used for energy load prediction and heating energy consumption $E_h(t)$ is further given by [25].

$$E_h(t) = \frac{UA_{buil} \sum_{j=1}^{bn} N_{bin,j} \left(T_{b,buld,j}(t) - T_{air,e,j}(t)\right)}{\eta_h} \quad (A.8)$$

Where, bn is number of bins, $N_{bin,j}$ is number of hours of occurrence of the jth bin, $T_{air,e,j}$ (t) is the external temperature at the jth bin and $T_{b,buld,j}(t)$ is the base temperature of jth bin for building. These bin methods are more accurate than degree-day method since it considers hourly weather data unlike daily average values, nevertheless, these methods has several drawbacks since it can neglects the extreme high or low climatic conditions and thermal mass effects of building.

Appendix B – Machine Learning based Artificial Intelligence

Machine learning techniques have been widely used in various applications of science and engineering. This machine learning based artificial intelligence techniques is helpful for linear and non-linear solving problems of higher dimensional and big data. This appendix details four types of machine learning based artificial intelligence techniques: artificial neural network, support vector machine, decision tree, random forest and concept of ensemble methods as a building energy modeling tools and practical aspects before training the model.

B.1 Artificial Neural Network

B.1.1 McCulloch Pitts Model

The biological neuron is the foundation of neural network concept forwarded by McCulloch and Pitts to solve the linear and non-linear complex problems. The neuron is the basic element of the nervous system including brain and is composed of three components: the cell body (soma), the axon and the dendrites. The structure of biological neurons is shown in Figure (B.1).

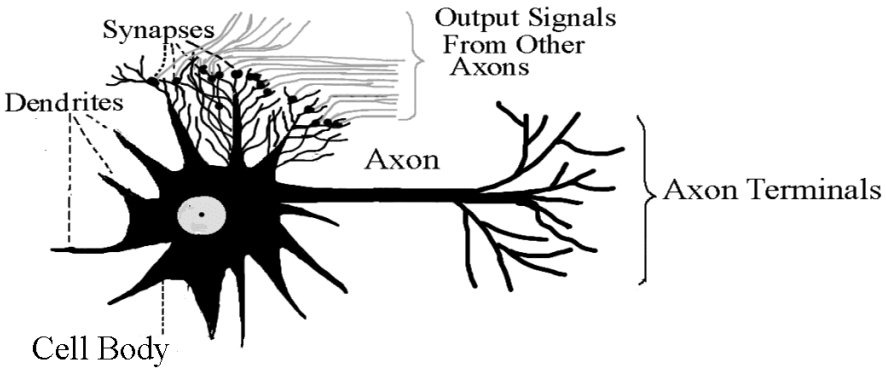

Fig. B.1: Structure of biological neurons [115]

As shown in Figure (B.1), the neuron receives the chemical input from other neurons through dendrites. The axon of single neurons forms the synaptic connection with many neurons and is the output channel to the other neurons. The connecting junction of neurons is called a synapse. The neuron thus receives signal from other neurons through cell body and dendrites, and integrates the stimulations. If this stimulates is higher, it increases the polarization of receiving nerve cell, and if this excitation is beyond the threshold value, then the neuron excites its own impulse to other neurons and sends a spike or output signal to other neurons through its axon. If these stimulates is below than the threshold values, then input will decay and they will not generate any action. The schematic of neural network is shown in Figure (B.2) and comparison of biological neuron with artificial neuron is shown in Table (B.1).

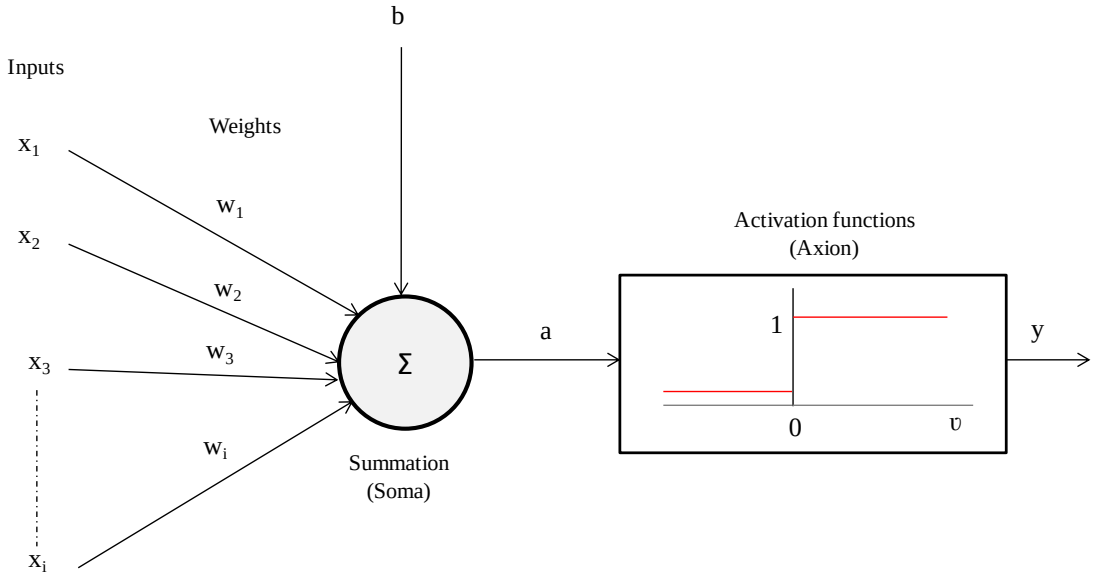

Fig. B.2: Schematic of artificial neuron network

Table B.1: Comparison of biological neuron and artificial neuron

Biological neuron	Artificial neuron
Cell body	Neuron
Dendrite	Input
Synapse	Weights
Axon	Output

In the Figure (B.2), neural network consist input represented by $x_1, x_2, \ldots\ldots, x_i$ and its corresponding weight are $w_1, w_2, \ldots\ldots, w_i$ i varies from 1 to u and u is the number of inputs; and b represents offset values called bias. Total signal (a) from input neurons is given by:

$$a = \sum_{i=1}^{u} x_i w_i + b \qquad (B.1)$$

Then, all the connecting weights are summed with input signals and compared with the threshold value ϑ. Finally, the output y is given by $y = f(a)$, where f is an activation function and also called transfer function. The McCulloh-Pitts perceptron model thus is defined by:

$$y = \begin{cases} 1 & \text{if } a \geq \vartheta \\ 0 & \text{if } a < \vartheta \end{cases} \qquad (B.2)$$

Equation (B.2) illustrates that the output of the neural network model will be high if the combined input is higher than the threshold values and zero if the combined input is below the threshold values.

B.1.2 Multi Layer Perceptron

Multi-layer perceptron (MLP) neural network are widely neural network and Figure (B.3) shows the MLP neural network which consists: input, hidden and output layers.

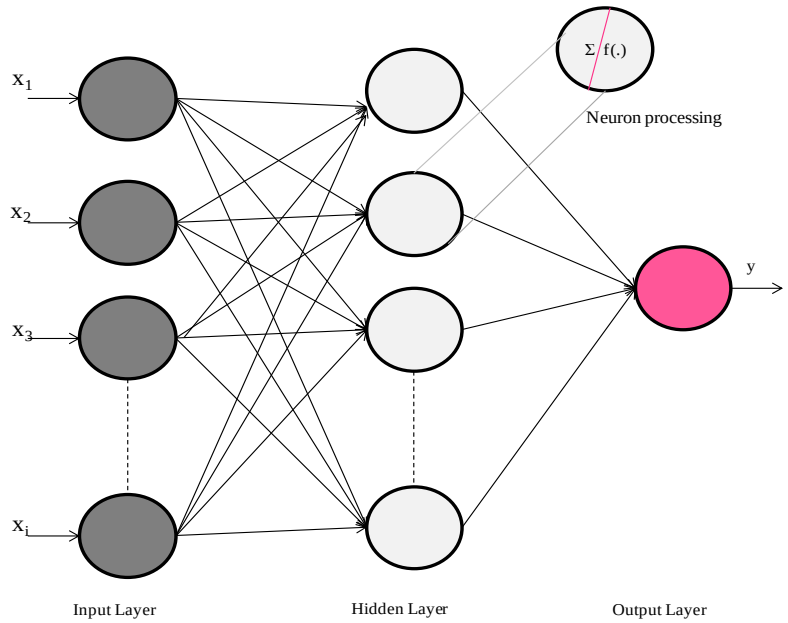

Figure B.3: Feed-forward multi layer perceptron neural network

The input layer receives the input data where each neuron correspond to each feature element (e.g., selected climatic conditions, occupancy profile etc.). The activation function for the input neurons is usually f(x)=x. The input data from input layer is further sends to the hidden layer.

The hidden layer processes the data and sends it to the next layer. There can exist more than one hidden layer and there is no any mathematical method which can be used to find the number of hidden layers. Generally, adding extra hidden layer improves the performance of model, however model training time is high due to additional complex structure. The output layer (e.g., heating load) is the last layer which receives the input signal from the last hidden layer. In each layer, the neurons receive the input signal from the previous layer and proceed to the next layer without feedback, thus this type of MLP neural network is also called feed-forward neural network. In the Figure, x_i is the input neurons, y is the output neuron and f(.) is the activation function which provides mapping of hidden layer to output layer. In order to find the weights and bias of the neurons, the training is performed and there are many algorithm and back-propagation algorithm is widely used method. This method initially calculates the training error by comparing network output obtained while flowing from input via hidden layer to output layer. After that, errors are back-propagated to the hidden layer and then the input layer so that the weight and bias are adjusted accordingly to minimize the error. Such process is repeated many times until the error propagating will be smaller.

B.1.3 Recurrent Neural Network

The recurrent neural network is similar to MLP feed-forward neural network except that neurons in the hidden layer are connected by the time delay (D^{-1}) which provides information of the past and is shown in Figure (B.4). It means output neurons is feedback to the previous neurons and thus signal flows both in forward and backward direction. This kind of learning is based on past experiences and are also called dynamic neural network.

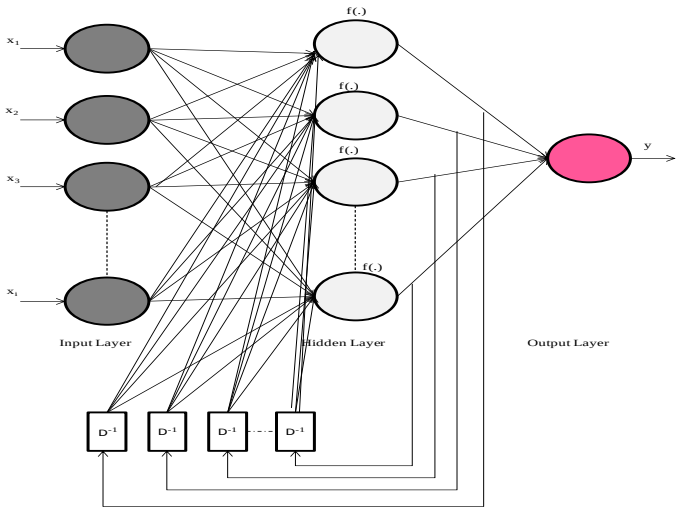

Fig. B.4: Recurrent neural network

B.1.4 Radial Basis Function Neural Network

The radial basis function (RBF) neural network differs from MLP neural network based on the activation function and is shown in Figure (B.5). Unlike MLP neural network which determined hidden layer by the weighted sum at the activation function, RBF neural network uses radial basis function at the hidden layer. This activation function at the hidden layer is given by:

$$f_j = f_j(x_i) = \frac{\|x_i - m_j\|}{\sigma_j} \tag{B.3}$$

Where, $f_j(.)$ is the jth radial basis function, x is input vector, m_j is the jth center point and σ_j is the width of the RBF. Typically, $f_j(.)$ is Gaussian function and is further given as:

$$f_j(x) = \exp^{\frac{\|x_i - m_j\|^2}{2\sigma_j^2}} \tag{B.4}$$

The output of radial basis function is given by:

$$y = \sum_j w_j f_j(x_i) \tag{B.5}$$

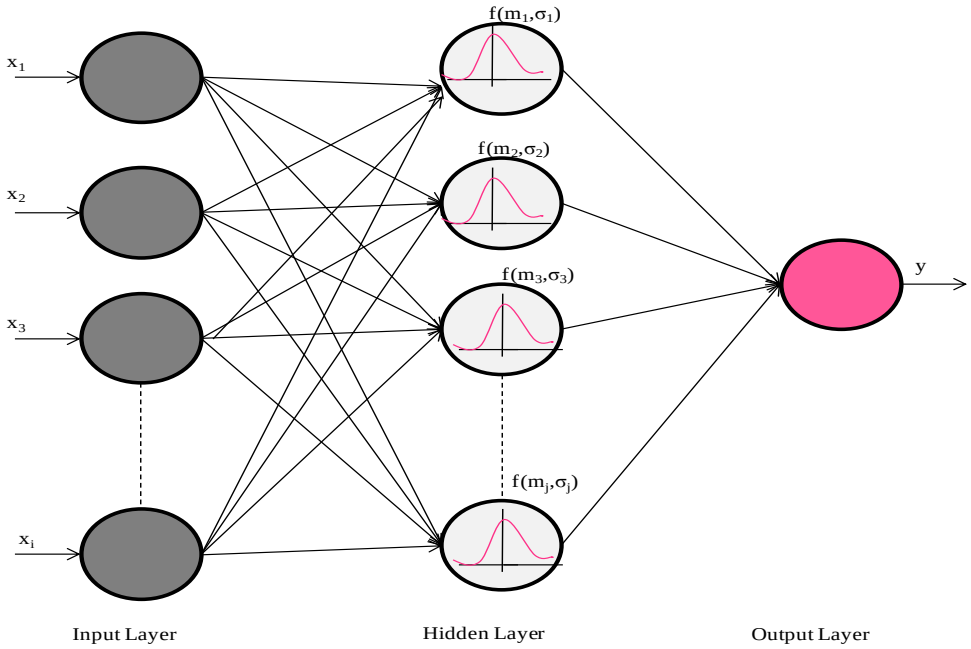

Fig. B.5: Radial basis function neural network

B.1.5 Hidden Neurons

The hidden neurons affect the performance of neural network. If the neural network has small number of hidden neurons, there is a chance that neural network not to capture the representation

of data leading to under-fitting of model. With the increase of hidden neurons, the performance of model can also be increased, but there are chances of network to be over-learned. Optimal choice of hidden neurons is thus necessary and there is no any robust rule in the determination of hidden neurons for the neural network. There are few literature focus to determine the number of hidden neurons. Kalogirou and Bojic [65] recommended hidden neurons based on input and output neurons and total number of training data and is given by Equation (B.6).

$$N_h = \frac{1}{2}(N_i + N_o) + \sqrt{N_s} \quad (B.6)$$

Where, N_h, N_i, N_o and N_s are number of hidden neurons, input neurons, output neurons and total number of data points in the training data. For Huang [116], the estimation of hidden neurons depends only on output neurons and number of training data and their estimation of hidden neurons for two hidden layer feed-forward network is given by the followings:

$$N_h = 2\sqrt{(N_o + 2)N_s} \quad (B.7)$$

The number of hidden neurons in the first layer (N_{1h}) is given by:

$$N_{1h} = 2\sqrt{(N_o + 2)N_s} + 2\sqrt{\frac{N_s}{(N_o + 2)}} \quad (B.8)$$

The number of hidden neurons in the second layer (N_{2h}) is given by:

$$N_{2h} = N_o\sqrt{\frac{N_s}{(N_o + 2)}} \quad (B.9)$$

B.1.6 Activation Functions

The activation function plays a significant role in mapping the non-linear functions. Typically, there are five types of activation functions widely used: linear, binary, piecewise linear, sigmoidal (s-shaped) and tangent hyperbolic shown in Figure (B.6) and their mapping function is shown in Equation (B.10).

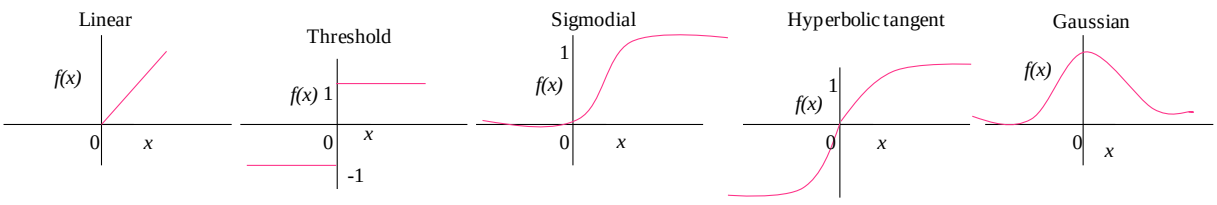

Fig. B.6: Activation functions used in the neural network

$$f(x) = \begin{cases} \text{Linear} & \zeta x \\ \text{Threshold} & \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases} \\ \text{Sigmoidal} & \dfrac{1}{1+e^{-\alpha x}} \\ \text{Hyperbolic tangent} & \dfrac{e^{\alpha x}-e^{-\alpha x}}{e^{\alpha x}+e^{-\alpha x}} \\ \text{Gaussian} & \exp\left(-\dfrac{x^2}{2\sigma^2}\right) \end{cases} \quad \text{(B. 10)}$$

In Equation (B.10), ζ is slope for linear activation function; α is the shape parameter for widely used sigmoidal and hyperbolic tangent activation; σ is control parameter for the width of the Gaussian activation function.

B.2 Support Vector Machine

Support vector machine (SVM) are built upon the state-of-the-art in kernel and are widely used in science and engineering. They are based on the statistical learning theory and have possibility to utilize kernel based methods to map the input features into higher dimensional plane to solve the complex non-linear problems. The non-linear function approximation with different kernels is the main strength of SVM but they are equally good in solving linear problems too. They provide advantages like noise robustness, maximum-margin etc. in comparison to simple regression model. SVM was originally utilized by Vapnik [117] for binary classification problem. After that, it was continuously followed by Vapnik, Drucker, Burges, Kaufman and Smola [118]. There are many libraries providing the implementation of SVM like LibSVM [119], LS-SVMlab [120] and SVMlight [121] etc.

These are widely used for classification problem and Figure (B.7) shows the classification problem where SVM tries to separate the data with the introduction of maximum margin by hyperplane, which is the common interest of SVM. The training data that are closest to the hyperplane are called support vectors and the distance between support vectors of different classes is called margin.

SVM is also used for regression problems where training sets are non-linearly separable, It finds the solution by suitable kernel function to map the non-linear input space into higher dimensional feature space where the separation hyperplane is found shown in Figure (B.8). Generally, there are two kinds of SVM for regression based on the controlling of training errors and support vectors: ε-SVR and υ-SVR [107]. There is no such difference between these two kinds of SVM,

but only the differences lies in the parameter. In this manuscript, we have used ε-SVM since this is widely used and more detail on υ-SVM is found in Scholkopf et al. [122].

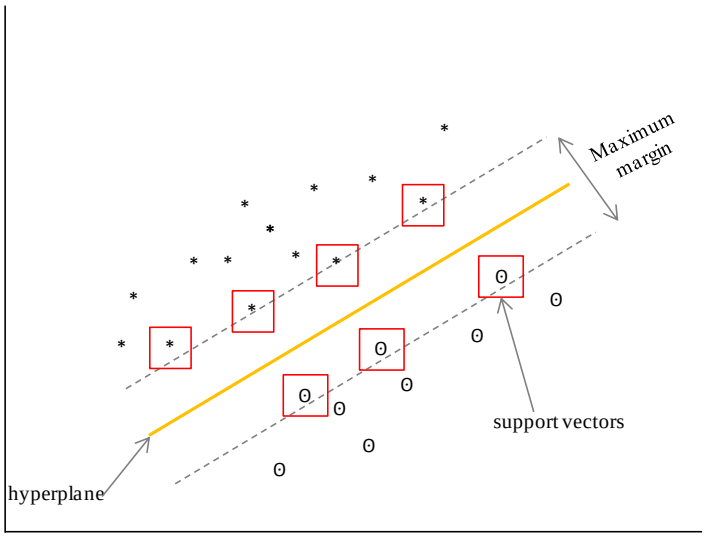

Fig. B.7: Separation hyperplane in support vector machine

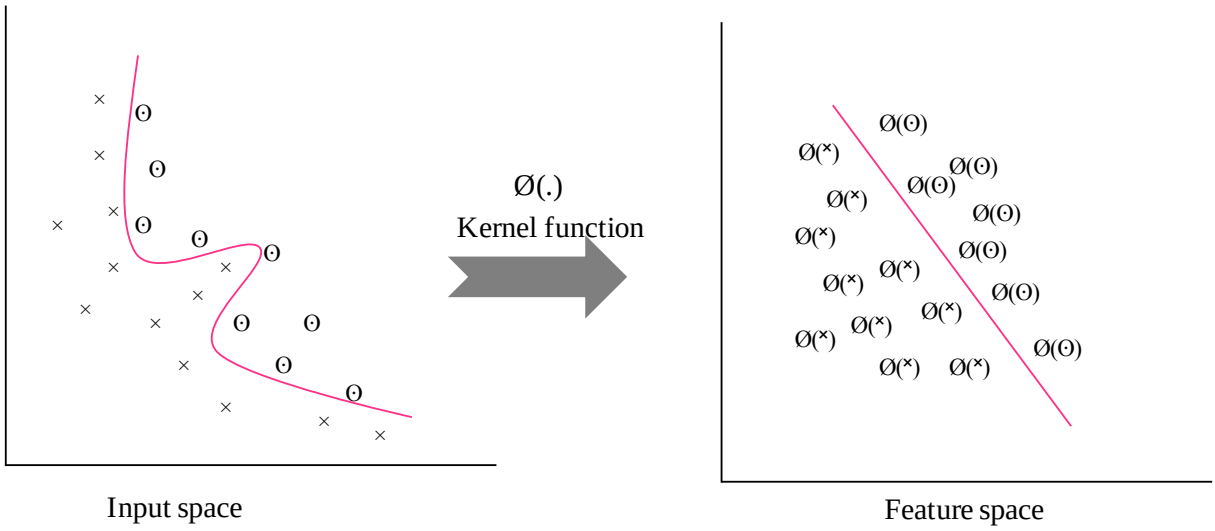

Fig. B.8: Transformation of input space into feature space

ε-SVM can be defined as follows: A original training data consists $\{(x_i, y_i)\ i = 1,2 \ldots n\}$ where x_i is the predictor variable and y_i is the response variable; the SVM tries to find the hyperplane that maximizes the margin and the equation of hyperplane is given by:

$$f(x) = \mathbf{w}\ \emptyset(x) + b \qquad (B.11)$$

where, **w** and b are constant, $\emptyset(.)$ is the mapping function which will be used to map input vector x into higher dimensional called kernel space. Then the SVM finds hyperplane by minimizing of quadratic problem:

$$\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{n}(\xi_k + \xi_k^*) \tag{B.12}$$

subject to

$$\begin{pmatrix}\text{minimize}\\ w,b,\xi,\xi^*\end{pmatrix}\begin{cases}(w^T x_k + b) - y_k \le \varepsilon + \xi_k\\ y_k - (w^T x_k + b) \le \varepsilon + \xi_k^*\\ \xi_k\xi_k^* \ge 0\end{cases} \qquad k=1, 2, 3 \ldots..N$$

By applying Lagrangian multipliers (α, α^*), Equation (B.12) is further formulated in order to minimize the dual quadratic problem.

$$\underset{\alpha,\alpha^*}{\text{minimize}}\begin{cases}\frac{1}{2}\sum_{k,i=1}^{n}(\alpha_k - \alpha_k^*)(\alpha_i - \alpha_i^*)\,K(x_k, x_i)\\ \varepsilon\sum_{k=1}^{n}(\alpha_k + \alpha_k^*) + \sum_{k=1}^{N} y_k(\alpha_k - \alpha_k^*)\end{cases} \tag{B.13}$$

subject to

$$\begin{cases}\sum_{k=1}^{n}(\alpha_k - \alpha_k^*) \le 0\\ 0 \le \alpha_k, \alpha_k^* \le C\end{cases} \qquad k=1, 2, 3 \ldots..N$$

Then the SVM output generates the regression and is represented in the following form:

$$f(x) = \sum_{k=1}^{n}(\alpha_k - \alpha_k^*)\,K(x, x_k) + b \tag{B.14}$$

The vectors with $(\alpha_k - \alpha_k^*) \neq 0$ in Equation (B.14) are support vectors; n is the number of training data; K is the kernel function; and b is solved using boundary conditions.

B.2.1 Kernel Functions

Kernel function plays a significant role to allow the training process simple by mapping the input data space which are non-separable to a separable data in higher dimensional space. There are four types of kernel functions widely used: linear, polynomial, radial basis function (RBF) and sigmoidal; and their representation is given as:

$$K(x_i, x_j) = \begin{cases} \text{Linear} & (a.\, x_i^T x_j + b) \\ \text{Polynomial with degree d} & (x_i^T x_j + 1)^d \\ \text{Radial} & \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \\ \text{Sigmoidal} & \tanh(a.\, x_i^T x_j + b) \end{cases} \quad \text{(B. 15)}$$

Where, x_i and x_j are input feature space (e.g. external temperature and occupancy profile); a and b are constant ; d is degree of polynomial and σ is tuning parameter of Gaussian radial basis function which control the width of kernel functions. In fact, this σ^2 determines the influences of input training transferred into kernel and if this value is small, then the input feature space are closer. If σ^2 is higher, then the input feature space are very far.

B.2.2 Parameter C and ε

The parameter C accounts for training errors and control the strength of penalty factor for error allowed during the training. Higher values of C will produce larger relative penalties and lead to problem of over-fitting. This further means large number of support vector will be required for the optimization problem. However, lower values of C will under fit the training data too. The parameter ε control the width of margin error, i.e., ε-insensitive loss. Higher value of ε produces simpler models and lead to problem of under-fitting. This further results in the solution to be sparse since it selects the fewer number of support vector. Lower value of ε, on the contrary, produces over-fitting of model.

B.3 Decision Tree

Decision tree is a statistical model widely used for classification and regression problems. Typically, trees are grown with binary recursive partition through series of splits or nodes shown in Figure (B.9). Initially, the root node is partitioned into two split nodes: left and right. This splitting of nodes continuously grows until fulfillment of stopping criteria is achieved. The node where tree stops growing is called leaf and all these values are averaged at the leaf node for the final prediction.

It is defined as follows [123] : Assuming two features x_1 and x_2 shown in Figure (B.9), the tree splits the features into five non-overlapping regions $R_1 \ldots R_5$ and known as leaf node. It can be seen that tree $x_i \leq c$ splits the regions into sub regions (left in the tree $x_i \leq c$ and right in the tree $x_i > c$). Assuming R be the leaf nodes with sets of input possible feature $x_1 \ldots x_j$ with j distinct

and non-overlapped regions, the goal of decision tree is to find the regions $R_1 \ldots R_j$ that minimize the sum of squares of regions.

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R,j}\right)^2 \tag{B.16}$$

Where, $\hat{y}_{R,j}$ is mean output value of the training data within the jth box and y_i is the actual output of training data.

Therefore, the main decision is the criteria for the choice of feature to be used in each node, how to calculate split from the node and how to decide that node is leaf. In order to address the above issues, various decision tree algorithms exist: ID3 Iterative Dichotomiser 3 [124], C4.5 successor of ID3 [125], CART classification and regression trees [108] etc.

In ID3 algorithm, the information gain criterion is used for split selection and with given training data set T, the entropy of T is given by ([126])

$$\mathrm{Ent}(T) = -\sum p(x|T) \log p(x|T) \tag{B.17}$$

Where, x is input data set and Ent is entropy. With the division of training data into subset $T_1, \ldots, T_k$, the entropy is reduced and the amount of information gain is given by [126]:

$$G(T; T_1, \ldots, T_k) = \mathrm{Ent}(T) - \sum_{i=1}^{k} \frac{|T_k|}{|T|} \mathrm{Ent}(T_k) \tag{B.18}$$

The features that have lowest information gain is used to select the split. However, the features that are selected only by information gain could have better fitting with the training data but cannot generalize for prediction or unseen condition. In order to address this, Quinlan proposes gain ratio given by:

$$p(T; T_1, \ldots, T_k) = G(T; T_1, \ldots, T_k). \left(- \sum_{i=1}^{k} \frac{|T_k|}{|T|} \log \frac{|T_k|}{|T|} \right)^{-1} \tag{B.19}$$

Equation (B.19) therefore considers variation of information gain by normalizing with the number of features. The features that have higher gain ratio is used to select the best split.

Similarly, CART another decision tree algorithm that uses Gini index was proposed by Breiman et al. [108] to select the split that maximizes the Gini index and is given by:

$$G_{gini}(T; T_1, \dots, T_k) = I(T) - \sum_{i=1}^{k} \frac{|T_k|}{|T|} I(T_k) \quad (B.20)$$

Where,

$$I(T) = 1 - \sum p(x|T)^2$$

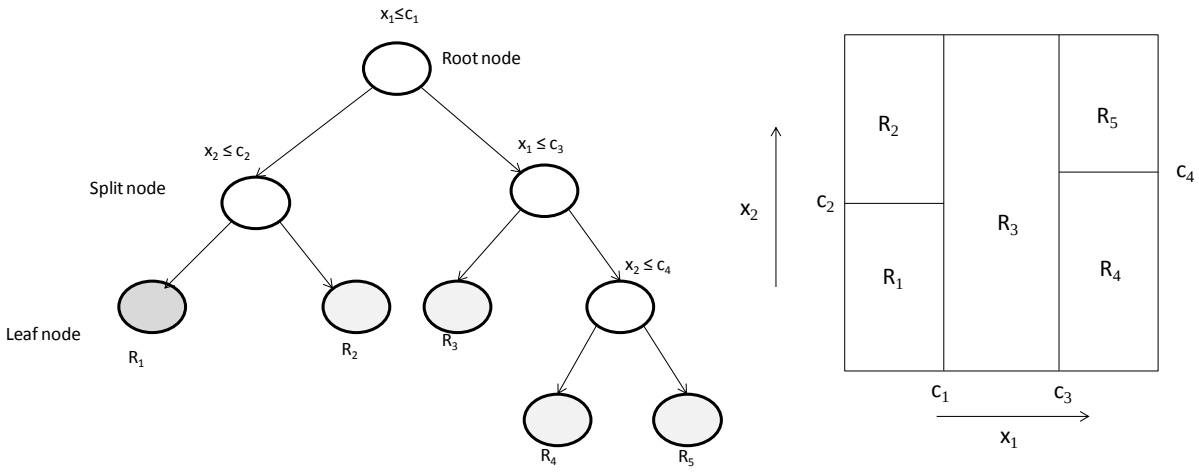

Fig. B.9: A simple illustration of decision tree [123]

B.4 Random Forest

One of the main problems of decision tree is the chances of over-fitting while fitting the training data by maximizing the depth of the trees. Breiman [111] proposed random forest to address this issue by using several decision trees. He proposed combination of regression trees with bagging (bootstrap aggregation) which uses random sampling with replacement from the original training data to build several bootstrapped datasets. Furthermore, it constructs the number of regression trees model by splitting the node with randomly selected subsets of features. Then, each splitting nodes, the conventional selection was performed to group into two proceeding nodes and the best split was selected. This process continues to split the nodes until the leaf nodes was met and the trees are build with maximum sizes. Finally, the output of all decision tree is aggregated. The overview of prediction from random forest is shown in Figure (B.10) where it can be seen that initial training data are divided into B number of bootstrap sample data sets for B number of trees and decision tree is build in each bootstrap and finally these are aggregated to make prediction.

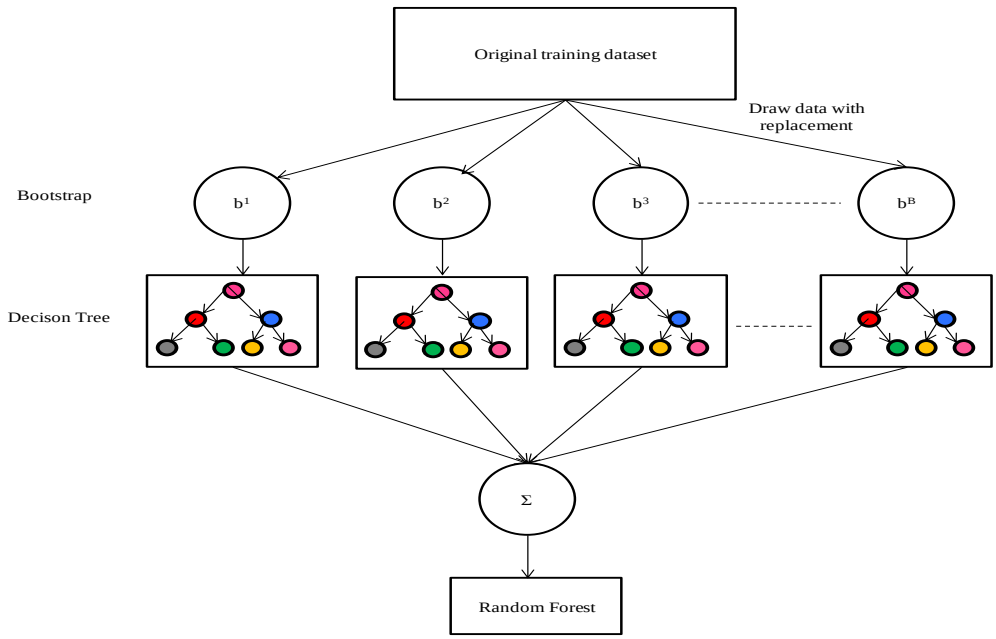

Fig. B.10: Overview of random forest

One of the characteristics of RF is "out-of-bag (OOB)" where it uses 2/3 of training data to build the model and 1/3 of training data is drawn from original training data set and are not involved in the construction of best-split decision tree and are called "out-of-bag" samples [111].

The main parameter that governs the random forest are: bootstrap size, number of trees, number of random possible variable at each splitting nodes and minimum number of leaf nodes of the trees. Random Forest are less sensitive to the parameters since the increase of bootstrap do not create a problem in over-fitting due to averaging effect of the trees in the ensemble [111].

B.5 Ensemble

Ensemble is the combination of multiple trained models to produce the prediction. The most popular methods are: bagging [127] and boosting [128]. Ensemble method performs better results than individual model [111].

Representing training set by (x_i, y_i) for M number of sub-model, the ensemble model takes the following form:

$$f(x) = \sum_{i=1}^{M} \beta_i \hat{f}_i(x) \qquad (B.21)$$

where, β_i=1,2..B are the weights of the ith sub-model and $\hat{f}_i(x)$ are trained model for the ith training set.

B.5.1 Bagging

Bagging is one of the widely used ensemble algorithm for classification and regression. It uses different training samples randomly with replacement from the original training data to represent bootstrap sample (for detail on bootstrap, see [129]) and then separate model are build with each bootstrap sample. The output of the model is obtained by aggregating the average from different model and thus reduces the generalization error [130].

It can be defined as follows: The original training data consists $\{(x_i, y_i)\ i = 1,2 \ldots n\}$ where x_i is the predictor variable and y_i is the response variable; then m number of bootstrapped training data is obtained from sample with replacement from the original training data. Then the boostrapped training data is trained with chosen training algorithm to get predicted $\hat{f}_m(x)$ for m model. Finally the prediction is obtained by averaging all the trained model:

$$f(x) = \frac{1}{B}\sum_{m=1}^{B} \hat{f}_m(x) \qquad (B.22)$$

More details about bagging is found in [131]; [132].

B.5.2 Boosting

Boosting is another ensemble machine learning algorithm and was proposed for the classification problem but it has also been widely used to solve the regression problem. It was first introduced by Schapire and named AdaBoost [133] and then it was improved by gradient method introduced by Freidman [110];[134] to build gradient boosted regression tree model.

It is noticed that bagging involves creation of multiple bootstrap sample and model are obtained by fitting into the separate training data of each bootstrap and final model is obtained by averaging the results to create a single predictive model. The boosting work in similar way to bagging except that it fits the trees sequentially and the basic concept is to prioritize for poorly fitted training data based on the results from the previous tree alone instead of all previously fitted trees [135]. It first fit the training data and residuals are calculated. The training data points corresponding to the higher residuals are assigned more weight in order to fit the next tree and this process is continuously forwarded until existing trees are remained unchanged.

It can be defined as follows ([136]; [137]): A original training data consists $\{(x_i, y_i)\ i = 1,2 \ldots n\}$ where x_i is the predictor variable and y_i is the response variable. Then the model can be approximated as a function f(x) for the response variable y and boosting decision tree algorithm estimate the basis function $b(x;\ \gamma_m)$as:

$$f(x) = \sum_m f_m(x) = \sum_m \beta_m\, b(x;\ \gamma_m) \tag{B.23}$$

where, $\beta_m (m = 1,2, \ldots M)$ are the expansion coefficients and $b(x;\gamma_m)$ are regression trees with the parameter γ_m represents the split variable. The coefficient β_m represents the weight of given nodes at each tree and it determines how the prediction from given trees are combined. The parameter β_m and γ_m are estimated by minimizing the loss function $L(y, f(x))$ which further indicate the prediction performance.

This loss function can be solved through optimization problem and Freidman [110] approximates the loss based on steepest descent. The methods follows by initializing the model $f_0(x)$ with constant value and grow the number of trees (m=1 to M). Then, the residuals for each training data are calculated as given:

$$\gamma_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{B.24}$$

Then the regression tree is fitted to γ_{im} to estimate γ_m of $b(x;\gamma_m)$. After that, parameters β_m are estimated by minimizing $L(y, f_{m-1}(x_i)) + \beta b(x;\gamma_m)$. Finally, the update f_m is obtained from Equation (B.25).

$$f_m(x) = f_{m-1}(x_i) + \beta_m b(x;\ \gamma_m) \tag{B.25}$$

The important aspect of gradient boosting is the regularization by shrinkage to control the learning rate and reduces the risk of over-fitting and Equation (B.25) is further modified as:

$$f_m(x) = f_{m-1}(x_i) + \upsilon\, \beta_m b(x;\gamma_m) \tag{B.26}$$

Where, υ is learning rate and is $0 < \upsilon \leq 1$. If the learning rate is <1, the ensemble requires more learning iterations. Finally, the predicted response is:

$$f(x) = \sum_m f_m(x) \tag{B.27}$$

More details about boosting are found [110], [134] and [109].

B.6 Practical Aspects in Artificial Intelligence

In artificial intelligence model, there are different task to be considered before training the model. Brief introduction of these practical aspects is explained below:

B.6.1 Normalization of Input and Output Data

Normalization is the process to make the magnitude of each variable similar so that there is no risk of slower convergence. In addition, it also helps to speed up training time and removes the outliers in the data and more details about it is found in Priddy and Keller [138]. The widely used normalization are discussed below:

Min-Max Normalization

It normalizes the input and output data to a fixed range usually from 0 to 1 or from -1 to 1. The min-max normalization is given by:

$$x_{i,n} = \frac{x_i - x_{min}}{x_{max} - x_{min}}\left(max_{target} - min_{target}\right) + min_{target} \quad (B.28)$$

Where, x_{min}, x_{max} are the minimum and maximum values of the input data; min_{target} and max_{target} are the minimum and maximum target values; $x_{i,n}$ is the normalized input data. Similarly, the normalization is performed for the output data.

Z-Score Normalization

It normalizes the training data of each feature by using mean (μ) and standard deviation (σ) of each feature of training data. The z-score normalization is given by:

$$x_{i,n} = \frac{x_i - \mu_i}{\sigma_i} \quad (B.29)$$

Sigmodial Normalization

Sigmodial normalization performs transformation into non-linear form using sigmodial functions: logistic or hyerpbolic function. The sigmodial normalization is given by:

$$x_{i,n} = \frac{1}{1 + e^{-\frac{(x_i - \mu_i)}{\sigma_i}}} \quad (B.30)$$

$$x_{i,n} = \frac{1 - e^{-\frac{(x_i - \mu_i)}{\sigma_i}}}{1 + e^{-\frac{(x_i - \mu_i)}{\sigma_i}}} \quad (B.31)$$

Equation (B.30) represents the logistic sigmoidal and normalizes the data in the range between 0 to 1. Equation (B.31) represents the normalization with hyperbolic tangent and normalizes the data in the range between -1 and 1.

B.6.2 Data Splitting

The data-driven models are prone to either under-fitting or over-fitting because of too many degrees of freedom in model. Generally, the overtraining/over-fitting can be observed by evaluating the model in validation set. The complexity of the model due to training and validation phases is shown in Figure (B.11). We can see that the more complex model can fit better than simple model with high variance and low bias during training phase. In case of validation phase, it is seen that less complex model have high prediction error with low variance and high bias, for details see Hastie et al. [109].

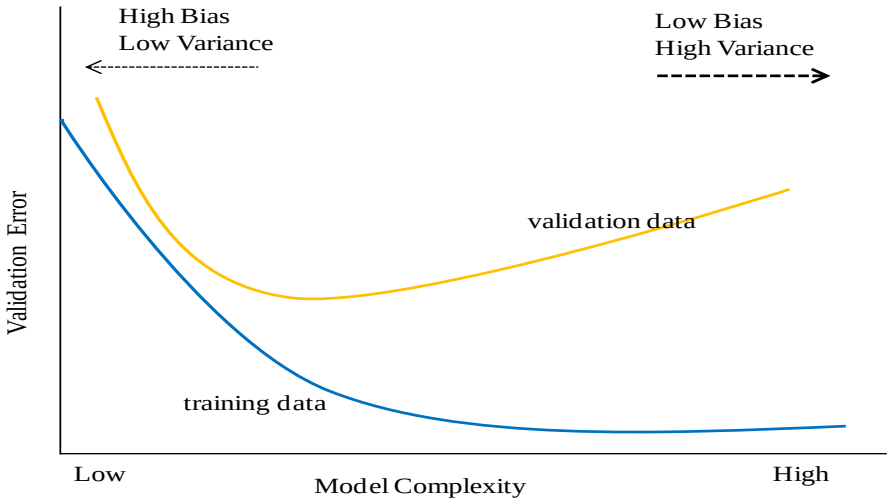

Fig. B.11: Influence of model complexity during training and testing phase [109]

The selection of model order thus depends on the complexity of model and these choices are further limited by bias and variance tradeoffs. If the model order is high, the complexity of model will increase and this has to be trade off by loss in approximation accuracy. If the model order is low, the complexity decreases and model error is dominated by approximation error due to insufficient fitting or capturing of non-linear data. This disadvantage of over-fitting and under-fitting of the model are reduced by splitting the data. A detail study on various splitting data

methods is found in (Burman [139]; Molinaro et al. [140]) and the most widely used techniques are described below:

Hold-out method

This is a simplest kind of data-splitting techniques where total number of training data sets are divided into training and validation shown in Figure (B.12). It holds certain amount of data for validation (about 1/3 of data sets) and remaining data are used for training sets. The advantage of this method is that it requires less training time.

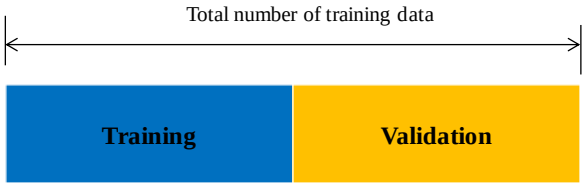

Fig. B.12: Hold-out method

However, this method has several drawbacks. If the small amount of training data is used, then the variance of the model will be larger. On the other hand, if large amount of training data is used, then small validation set might result in poor performance to select best parameters of model.

Random Sub-Sampling

Random sub-sampling is another hold out method where whole training data is randomly split into subsets shown in Figure (B.13). For each of the number of data splits, the model is trained and error is evaluated. The final model is evaluated by averaging the error estimate from individual data splits.

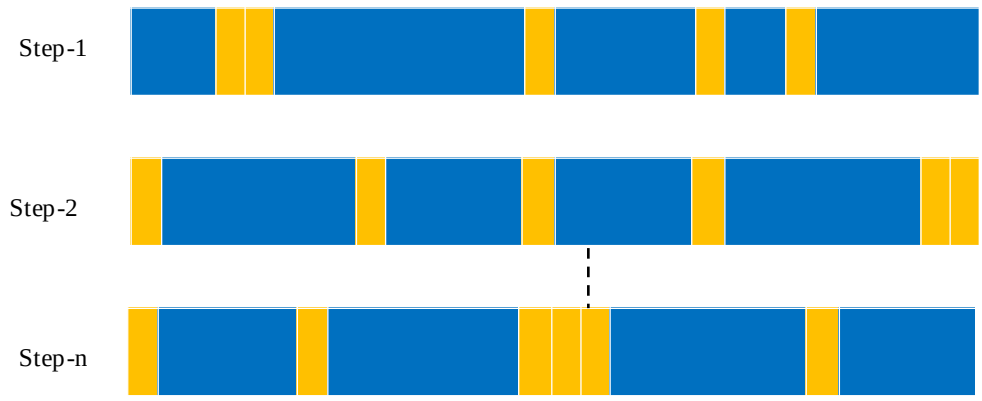

Fig. B.13: Random sub-sampling

K-fold Cross-Validation

K-fold cross-validation is the most popular data splitting method where data sets are divided into k-number of equal parts shown in Figure (B.14). This concept originates since the selection of single validation data split might not be representative of the training data sets. In this method, one fold is used for evaluation of model and remaining k-1 folds are used for training. This method is similar to repeated hold-out method and has advantages of using all the training data sets for evaluating and learning the model. The result of the final model is obtained as the average of the k-fold results to find the best parameters of model.

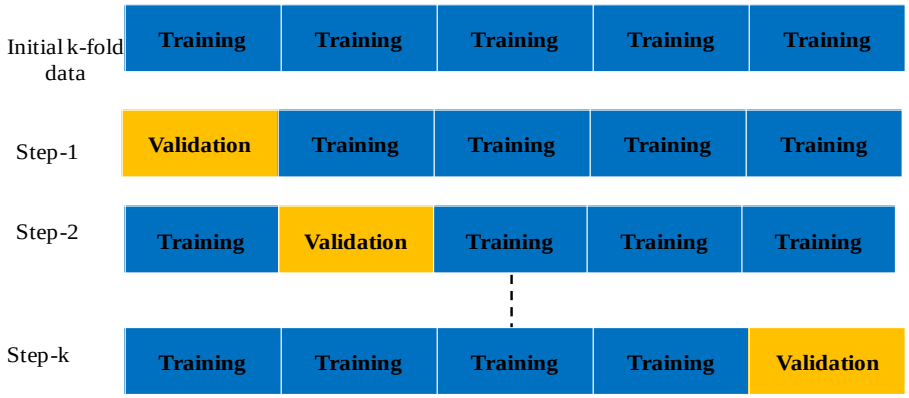

Fig. B.14: K-fold Cross Validation

Leave-one-out cross validation

Leave-one-out cross validation is a special case of k-fold cross validation where k=n and n represents the number of training data sample shown in Figure (B.15). It uses one sample of data for evaluating the model and rest of the training data for learning the model. Because of using n repeated sample of data for evaluating the model, the model training time is too high. This method is computationally expensive.

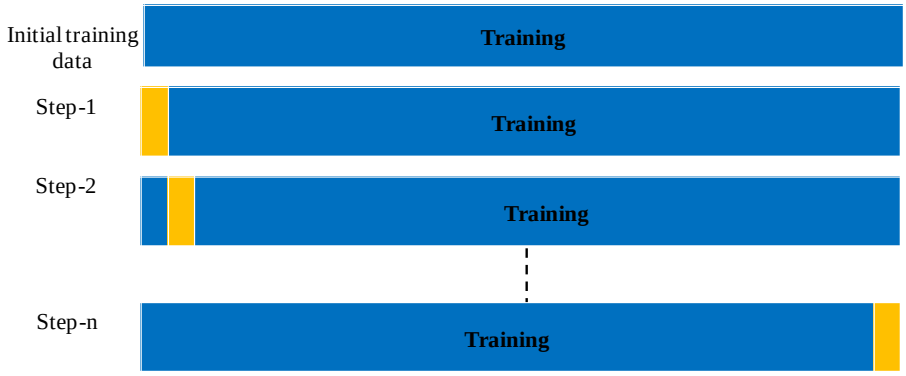

Fig. B.15: Leave-one-out cross validation

Bootstrap

The basic concept of bootstrap is to randomly select with replacement from the training data set shown in Figure (B.16). While selecting the bootstrap sample, the sample may be chosen again from the original data set more than once. This is considered best way if the training data samples are smaller. This process is repeated for the specified number of bootstrap B. Then the model is evaluated on each bootstrap sample and final model is selected by averaging these B estimates.

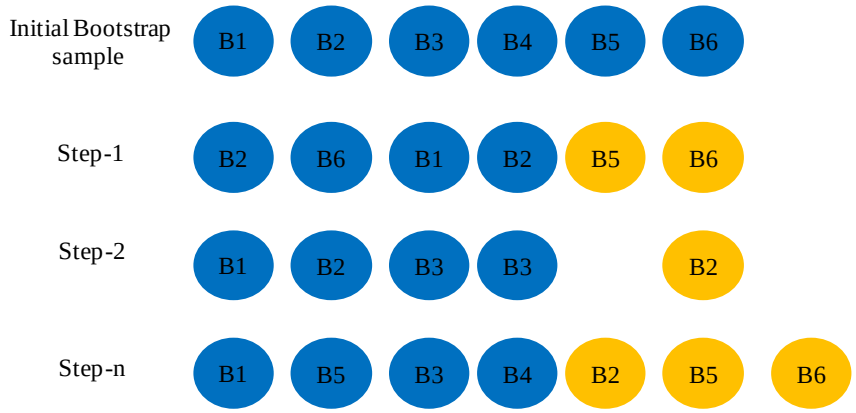

Fig. B.16: Bootstrap

Appendix C- Building Operation Classification/Clustering

Cannonical Variate Analysis (CVA) is a multivariate discriminate tool, which is based on covariance matrix of variables and used to show the correlation degree between input vector data sets. It transform the original input vector data sets into new axes called canonical variable without losing relevant information from the input data sets and make this component even better than input data set. Basic idea of this transformation into new axes is first extraction of statistical features and dynamical change of daily energy consumption of building and later used this feature to analyze from CVA whether it is easier for classification or not. Main statistical features of daily time series include daily average energy load (P_m) and maximum daily energy load (P_{max}). In order to reduce the seasonal variance from model, minimum value of energy load on particular day is reduce from daily energy load [49]. Then, the auto-regression model is applied for the dynamical change of energy load of building for each day and is shown in Equation (C.1). Equation (C.1) assumes that current sample $y(t)$ can be predicted from the linear weight of the sum of p sample values i.e. $y(t-1), y(t-2) \dots y(t-p)$, where p is the model order, q^i is the ith coefficient of the pth model order, ε is the noise parameter and y^0 is the initial value of the p model order. With given the order of p for auto regression model, the parameters q^m and ε can be estimated and burg algorithm is used for determination of coefficients in this study, for details see Li et al. [49]. Therefore, the statistical feature includes daily average energy load, maximum daily energy load and auto-regression model coefficients which are then input to the CVA to transform into canonical variables.

$$y(t) = y^0 + \sum_{m=1}^{p} q^m \, y(t-m) + \varepsilon \qquad (C.1)$$

The input statistical features including auto-regression model coefficients of each day is divided into K group i.e. $X_{N*M} = \{x^{ij}\}$, where N is the number of training days data sets, M is the number of statistical features, K is the operating profile of building in which each operating profile contains $(i = 1 \dots K)$ samples. For example, if X is the statistical features of one year data with model order of four, then,$X = [P_m \; P_{max} \; q^1 \; q^2 \; q^3 \; q^4]$, K contains different kinds of operating profile from Monday to Sunday i.e. K=7, N=365 and M=6. Group covariance matrix (C_w) and

between the group covariance matrix (C_B) is calculated in equation (C.2 –C.3), where, x^{ij} is the jth sample in the ith group, $\bar{x}^i$ is the mean vector in the ith group and $\bar{x}$ is the mean vector in overall database.

$$C_w = \frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{M}(x^{ij}-\bar{x}^i)\,(x^{ij}-\bar{x}^i)^T \qquad (C.2)$$

$$C_B = \frac{1}{K-1}\sum_{i=1}^{K} n^i(\bar{x}^i-\bar{x})\;\;(\bar{x}^i-\bar{x})^T \qquad (C.3)$$

To determine canonical variables from Equation (C.2–C.3), a direction **W** needs to be determined which satisfies the condition in equation (C.4), where, λ is Eigen values and **W** is Eigen vectors.

$$C_B\mathbf{W} = \lambda\, C_w\mathbf{W} \qquad (C.4)$$

Thus, canonical variables (CV) is determined as:

$$CV = X\,\mathbf{W} \qquad (C.5)$$

Thèse de Doctorat

Subodh PAUDEL

Méthodologie pour estimer la consommation d'énergie dans les bâtiments en utilisant des techniques d'intelligence artificielle

Methodology to Estimate Building Energy Consumption Using Artificial Intelligence

Résumé

Les normes de construction pour des bâtiments de plus en plus économes en énergie (BBC) nécessitent une attention particulière. Ces normes reposent sur l'amélioration des performances thermiques de l'enveloppe du bâtiment associé à un effet capacitif des murs augmentant la constante de temps du bâtiment. La prévision de la demande en énergie de bâtiments BBC est plutôt complexe. Ce travail aborde cette question par la mise en œuvre d'intelligence artificielle (IA). Deux approches de mise en œuvre ont été proposées : « all data » et « relevant data ». L'approche « all data » utilise la totalité de la base de données. L'approche « relevant data » consiste à extraire de la base de données un jeu de données représentant le mieux possible les prévisions météorologiques en incluant les phénomènes inertiels. Pour cette extraction, quatre modes de sélection ont été étudiés : le degré jour (HDD), une modification du degré jour (mHDD) et des techniques de reconnaissance de chemin : distance de Fréchet (FD) et déformation temporelle dynamique (DTW). Quatre techniques IA sont mises en œuvre : réseau de neurones (ANN), machine à support de vecteurs (SVM), arbre de décision (DT) et technique de forêt aléatoire (RF). Dans un premier temps, six bâtiments ont été numériquement simulés (de consommation entre 86 kWh/m².an à 25 kWh/m².an) : l'approche « relevant data » reposant sur le couple (DTW, SVM) donne les prévisions avec le moins d'erreur. L'approche « relevant data » (DTW, SVM) sur les mesures du bâtiment de l'Ecole des Mines de Nantes reste performante.

Mots clés
Consommation d'énergie dans le bâtiment, Prévision, Bâtiment basse consommation, Intelligence artificielle, Jeu de données représentatives, Apprentissage en ligne et hors ligne

Abstract

High-energy efficiency building standards (as Low energy building LEB) to improve building consumption have drawn significant attention. Building standards is basically focused on improving thermal performance of envelope and high heat capacity thus creating a higher thermal inertia. However, LEB concept introduces a large time constant as well as large heat capacity resulting in a slower rate of heat transfer between interior of building and outdoor environment. Therefore, it is challenging to estimate and predict thermal energy demand for such LEBs. This work focuses on artificial intelligence (AI) models to predict energy consumption of LEBs. We consider two kinds of AI modeling approaches: "all data" and "relevant data". The "all data" uses all available data and "relevant data" uses a small representative day dataset and addresses the complexity of building non-linear dynamics by introducing past day climatic impacts behavior. This extraction is based on either simple physical understanding: Heating Degree Day (HDD), modified HDD or pattern recognition methods: Frechet Distance and Dynamic Time Warping (DTW). Four AI techniques have been considered: Artificial Neural Network (ANN), Support Vector Machine (SVM), Boosted Ensemble Decision Tree (BEDT) and Random forest (RF). In a first part, numerical simulations for six buildings (heat demand in the range [25 – 85 kWh/m².yr]) have been performed. The approach "relevant data" with (DTW, SVM) shows the best results. Real data of the building "Ecole des Mines de Nantes" proves the approach is still relevant.

Key Words
Building Energy Consumption, Prediction, Low Energy Building, Machine Learning, Small representative data, Online and Offline Learning