Agenzia nazionale per le nuove tecnologie,
l'energia e lo sviluppo economico sostenibile

MINISTERO DELLO SVILUPPO ECONOMICO

Ricerca di Sistema elettrico

Clustering di abitazioni per la scelta di modelli di disaggregazione di consumi elettrici realizzati tramite reti deep

C. La Riccia, V. Piccialli, A. Sudoso

Università di Roma Tor Vergata

Report RdS/PAR2018/023

CLUSTERING DI ABITAZIONI PER LA SCELTA DI MODELLI DI DISAGGREGAZIONE DI CONSUMI ELETTRICI REALIZZATI TRAMITE RETI DEEP

C. La Riccia, V. Piccialli, A. Sudoso (Università degli Studi di Roma Tor Vergata, DICII)

Dicembre 2019

Report Ricerca di Sistema Elettrico

Accordo di Programma Ministero dello Sviluppo Economico - ENEA
Piano Annuale di Realizzazione 2018
Area: Efficienza energetica e risparmio di energia negli usi finali elettrici e interazione con altri vettori energetici
Progetto: D.6 Sviluppo di un modello integrato di smart district urbano
Obiettivo: b Sistemi e servizi smart per edifici.
Responsabile del Progetto: Claudia Meloni, ENEA

Il presente documento descrive le attività di ricerca svolte all'interno dell'Accordo di collaborazione "*Clustering di tipologie di abitazione per scegliere modelli di disaggregazione di consumi elettrici realizzati tramite reti deep".*
Responsabile scientifico ENEA: Claudia Snels
Responsabile scientifico DICII Università degli Studi di Roma Tor Vergata: Veronica Piccialli

Indice

Sommario

L'idea di questo lavoro è quello di raggruppare le abitazioni delle 10 case di Centocelle per cui sono stati raccolti i dati della disaggregazione del consumo elettrico in cluster simili rispetto ai comportamenti di consumo dei residenti nell'abitazione.
Per far questo si sono individuate prima di tutto delle caratteristiche (features) che fossero significative ai fini della caratterizzazione del comportamento energetico dei residenti e che fossero estraibili dal consumo generale.
Utilizzando il main (potenza elettrica in Watt) di ogni abitazione, sono state individuate 10 caratteristiche (features) che permettono di catturare il profilo di consumo di ciascuna di esse.
Si è quindi applicato l'algoritmo k-means alle 10 abitazioni caratterizzate con le precedenti 10 features ottenendo quattro cluster, uno costituito da una sola casa C4, uno con due abitazioni, C3 e C5, uno con 3 abitazioni C2, C7, C9 e uno con 4 abitazioni, C1, C6, C8, e C10. C4 costituisce un cluster a sé, e questo corrisponde al fatto che è l'unica casa con un unico abitante e con consumi energetici molto molto bassi.
Per gli altri cluster è stata scelta la casa più vicina al centroide ed è stato addestrato un opportuno modello di machine learning su quella casa.
I dati del consumo elettrico sia generale che dei singoli elettrodomestici sono aggregati su intervalli di 15 minuti, risultando quindi in dataset di dimensioni non eccessive. Per questo motivo si sono utilizzate reti con architetture non troppo complesse. L'architettura è stata scelta tramite cross validation.
E' stata addestrata una rete per ogni elettrodomestico utilizzando come input la serie main e come target la serie dell'appliance corrispondente. Il test set corrisponde all'ultimo mese della coppia main-appliance e il validation set che è temporalmente consecutivo al training set viene utilizzato per scegliere l'architettura della rete (numero di strati e numero di neuroni in ciascun strato) e il numero ottimo di epoche tramite early stopping. L'errore viene calcolato sulla base dell'errore medio assoluto (MAE). La bontà del modello viene testata prima di tutto sul test set di quella coppia casa-appliance che corrisponde all'ultimo mese di dati disponibili. Dopo di che il modello viene testato sull'intero insieme di dati delle altre case nel centroide.
I risultati sono molto buoni in quanto si ha un'ottima aderenza del profilo di consumi.

1 Introduzione

Il problema della disaggregazione dei consumi elettrici è noto in letteratura con l'acronimo NILM (dall'inglese Non-intrusive Load Monitoring). Con il termine NILM [1] si fa riferimento ad un processo volto a stimare, sulla base delle variazioni del consumo generale, il consumo energetico di un dato elettrodomestico. I sistemi NILM permettono quindi di monitorare il consumo elettrico all'interno di una abitazione senza richiedere l'utilizzo di smart meter dedicati ai singoli elettrodomestici. Questo tipo di studio consente di fornire informazioni dettagliate sui consumi agli utenti, così da indurli a modificare le loro abitudini verso un uso più saggio dell'energia elettrica.

Obiettivo del presente progetto è quello di dimostrare come si possa utilizzare un unico modello di machine learning [2, 3] per predire il consumo del singolo elettrodomestico su case caratterizzate da comportamento energetico simile.

Lo studio ha coinvolto dieci case situate a Roma, per le quali sono stati registrati dati sul consumo elettrico sia generale che per singolo dispositivo nel corso dell'anno 2018. In particolare ci sono stati forniti dati per una durata massima di sette mesi e una durata minima di un mese e mezzo.

Il primo passo è stato quello di effettuare il clustering delle abitazioni sulla base di features opportune che potessero descrivere in modo soddisfacente il comportamento energetico degli abitanti dell'abitazione.

Per ogni cluster è stata scelta la casa più vicina al centroide ed è stato addestrato un opportuno modello di machine learning su quella casa.

I dati del consumo elettrico sia generale che dei singoli elettrodomestici sono aggregati su intervalli di 15 minuti, risultando quindi in dataset di dimensioni non eccessive. Per questo motivo sono state utilizzate reti con architetture non troppo complesse. L'architettura è stata scelta tramite cross validation. E' stata addestrata una rete per ogni elettrodomestico utilizzando come input la serie main e come target la serie dell'appliance corrispondente. Il test set corrisponde all'ultimo mese della coppia main-appliance e il validation set che è temporalmente consecutivo al training set viene utilizzato per scegliere l'architettura della rete (numero di strati e numero di neuroni in ciascun strato) e il numero ottimo di epoche tramite early stopping. L'errore viene calcolato sulla base dell'errore medio assoluto (MAE) o dell'errore quadratico medio (MSE). La bontà del modello viene testata prima di tutto sul test set di quella coppia casa-appliance che corrisponde all'ultimo mese di dati disponibili. Dopo di che il modello viene testato sull'intero insieme di dati delle altre case nel centroide.

I risultati sono molto buoni, e vengono riportati sia graficamente che sulla base di metriche opportune definite in letteratura [6].

2 Descrizione delle attività svolte e risultati

In questa sezione verranno descritte le attività che hanno caratterizzato il progetto oggetto di questo report.

2.1 Preprocessing dei dati

Il primo passo è stato quello di analizzare i dati, in modo da capire quali elettrodomestici fossero presenti in ogni casa e con quale mole di dati.
Per individuare facilmente i valori mancanti (NA), per ogni elettrodomestico e per ogni casa (potenza elettrica complessiva) i dati disponibili sono stati rappresentati graficamente tramite mappe di calore. Poiché i dati sono aggregati a 15 minuti, in un giorno ci sono al più 96 osservazioni (verde). Con questa rappresentazione grafica vediamo come sono distribuiti i valori mancanti su base giornaliera e capiamo dove intervenire in fase di pre-processing con tecniche sostituzione di un dato record con alternative coerenti. A tal proposito è stata utilizzata l'interpolazione lineare per sostituire valori mancanti corrispondenti a giorni sparsi del calendario e giorni per cui la frazione di valori mancanti è trascurabile (arancione), mentre per blocchi consecutivi da minimo di 3 a ad un massimo di 30 giorni (rosso) è stata utilizzata la media puntuale tra il periodo precedente e quello successivo al blocco. Valori mancanti all'inizio e alla fine del periodo di copertura dei dati forniti sono stati eliminati per evitare di introdurre rumore.
In figura 1 riportiamo la mappa di calore relativa al main della casa C6, e nelle figure 2 e 3 le mappe di colore relative ai due elettrodomestici (frigorifero e lavatrice) della stessa casa. Si nota come la distribuzione dei dati mancanti sia diversa nelle tre figure. In questo caso specifico, abbiamo rimosso il primo blocco di dati mancanti e sostituito quelli successivi con la tecnica descritta sopra.

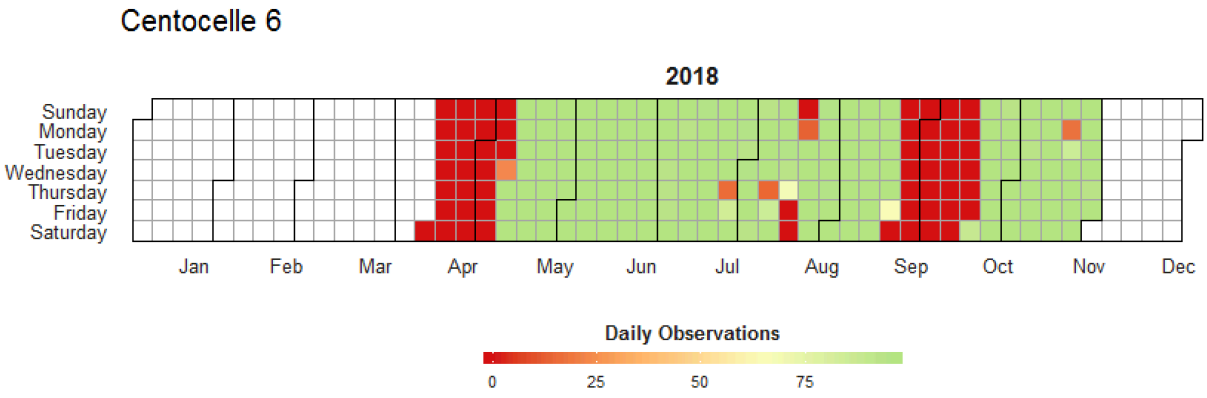

Figura 1: mappa di calore per individuare i dati mancanti relativa al main della casa C6

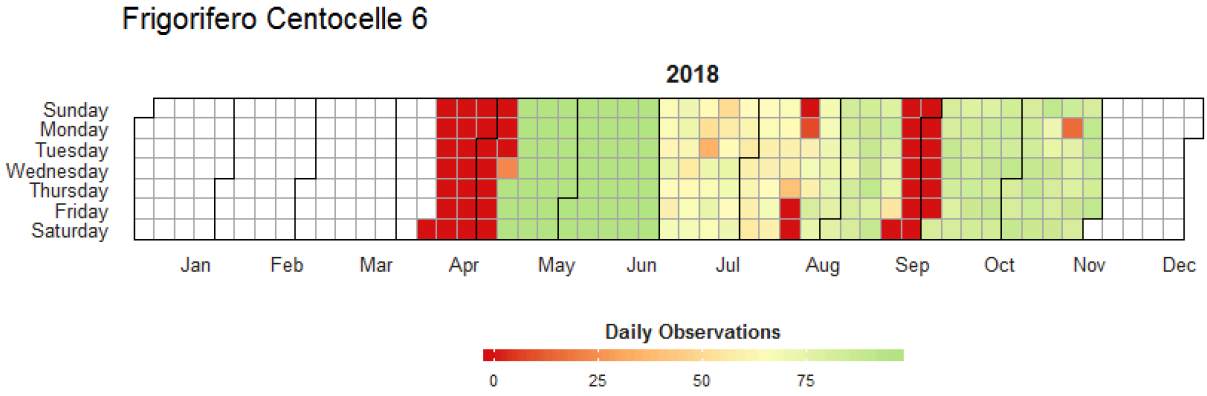

Figura 2: mappa di calore per individuare i dati mancanti relativa al frigorifero della casa C6

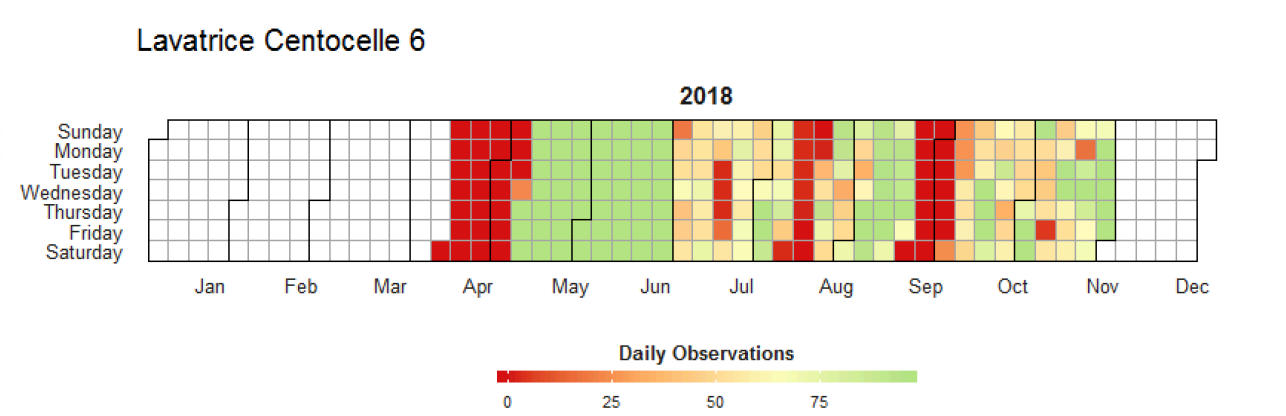

Figura 3: mappa di calore (per individuare i dati mancanti) relativi alla lavatrice della casa C6

A fronte della sostituzione dei dati mancanti, si ottiene un insieme di dati puliti la cui dimensione dipende dalla casa e dall'elettrodomestico.

Nella seguente tabella sono rappresentati i periodi effettivamente utilizzabili del consumo complessivo e dei singoli elettrodomestici per ogni abitazione:

	MAIN	FRIDGE	WASHING MACHINE	TV	DISHWASHER
C1	14/04 – 30/11	14/04 – 30/11	14/04 – 30/11	14/04 – 30/11	14/04 – 30/11
C2	14/04 – 30/11	14/04 – 30/11	14/04 – 30/11		
C3	14/04 – 30/08	14/04 – 30/09	14/04 – 30/06		
C4	14/04 – 30/11		14/04 – 30/06	14/04 – 30/06	
C5	14/04 – 30/11		14/04 – 30/11		14/04 – 30/11
C6	10/05 – 30/11	10/05 – 30/11	10/05 – 30/11		
C7	14/04 – 21/11	14/04 – 21/11	14/04 – 21/11		14/04 – 30/09
C8	09/05 – 31/07		09/05 – 30/06	09/05 – 30/06	
C9	09/05 – 30/11	09/05 – 30/11	09/05 – 07/07		09/05 – 30/11
C10	14/04 – 14/07	14/04 – 14/07	14/04 – 14/07	14/04 – 14/07	14/04 – 14/07

Si vede quindi che non tutte le abitazioni hanno gli stessi elettrodomestici e gli stessi intervalli temporali sui dati. Ci siamo concentrati su frigorifero, lavatrice e lavastoviglie. Una volta ottenuti dei dati puliti, il passo successivo è quello di scegliere gli elementi rappresentativi del comportamento di consumi energetico delle abitazioni su cui fare clustering.

2.2 Clustering

Il primo passo è stato quello di individuare delle caratteristiche (features) che fossero significative ai fini della caratterizzazione del comportamento energetico dei residenti e che fossero estraibili dal consumo generale. L'idea di fare clustering non è completamente nuova. In [4] viene proposta una tecnica chiamata Neighborhood NILM che sfrutta i dati delle case 'vicine' per disaggregare l'energia data una sola lettura di energia al mese. L'intuizione chiave del loro approccio è che le case "simili" hanno un consumo energetico "simile" per ogni singolo apparecchio. Neighborhood NILM abbina ogni casa con un insieme di "vicini" che hanno un'infrastruttura di sottomisurazione diretta, cioè contatori di potenza su singoli circuiti o carichi. Molte di queste case esistono già. Poi, stima che il consumo energetico a livello di elettrodomestici della casa target sia la media dei suoi vicini K. Le features usate in [4] per fare clustering di abitazioni sono le seguenti:

1. Consumo energetico aggregato di 12 mesi

2. Varianza del consumo energetico su 12 mesi
3. Consumo massimo/consumo minimo nei 12 mesi
4. Consumo massimo – consumo minimo nei 12 mesi
5. Area dell'abitazione
6. Numero di abitanti
7. Numero di stanze

Nel nostro caso, utilizzando il main (potenza elettrica in Watt) di ogni abitazione, sono state individuate 10 caratteristiche che permettono di catturare il profilo di consumo di ciascuna abitazione. A questo scopo dividiamo le 24h di un giorno in tre fasce orarie di consumo, per avere una migliore descrizione del comportamento energetico, ciascuna da 8h:

1) 08:00:00 – 16:00:00
2) 16:00:00 – 00:00:00
3) 00:00:00 – 08:00:00

Per una particolare abitazione, P_i è la potenza media nella fascia oraria $i = 1, 2, 3$ con deviazione standard S_i. P_{tot} è la potenza media su tutto il periodo disponibile con deviazione standard S_{tot}. Infine P_i^{WE} è la potenza media nel weekend in ciascuna fascia oraria e P_i^{WD} è la potenza media nei rimanenti giorni della settimana. Usando questa notazione, il profilo di consumo di ogni abitazione viene calcolato utilizzando le seguenti nove caratteristiche:

8. - **Potenza media relativa:** $P_i^R = \frac{P_i}{P_{tot}}, i = 1, 2, 3$
9. - **Deviazione standard relativa:** $S_i^R = \frac{S_i}{S_{tot}}, i = 1, 2, 3$
10. **- Potenza media Weekend vs. Weekday:** $W_i = \frac{|P_i^{WE} - P_i^{WD}|}{P_{tot}}, i = 1, 2, 3$
11. Infine si è scelto di includere il **numero di utenti nell'abitazione**

Si è quindi applicato l'algoritmo k-means alle 10 abitazioni caratterizzate con le precedenti 10 features. L'algoritmo k-means [5] è basato sul principio di minimizzare la dissimilarità nei cluster intesa come la somma delle distanze euclidee tra i vettori del cluster e un centroide (che deve essere determinato) del cluster. Per la scelta del numero di cluster si sono effettuati diversi tentativi, e si è scelto alla fine la suddivisione in 4 clusters, il cui risultato è rappresentato in Figura 1.

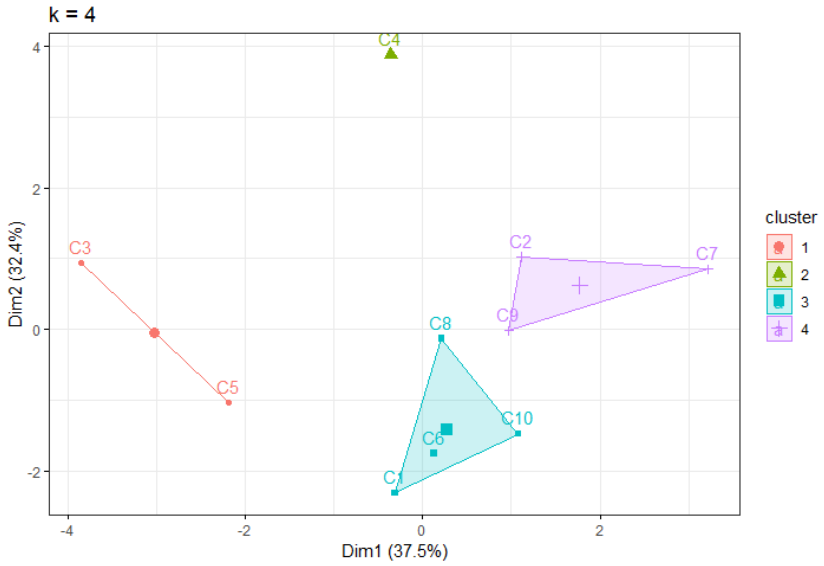

Figura 4: risultato di K-Means utilizzando le 10 case Centocelle (Aprile – Novembre)

Per la visualizzazione dei cluster il set delle variabili iniziali è stato trasformato in un nuovo set di variabili attraverso l'analisi delle componenti principali (PCA). L'algoritmo di riduzione della dimensionalità opera su 10 variabili e restituisce 2 nuove variabili (Dim1 e Dim2). Ogni dimensione rappresenta una certa quantità di varianza (o informazione) che è contenuta nel set di dati originale.
Si ottengono un cluster da 4 case, un cluster da 3 case, un cluster da 2 case, mentre C4 costituisce un cluster a sé, e questo corrisponde al fatto che è l'unica casa con un unico abitante e con consumi energetici molto molto bassi.
Riportiamo infatti l'andamento del main della casa C4:

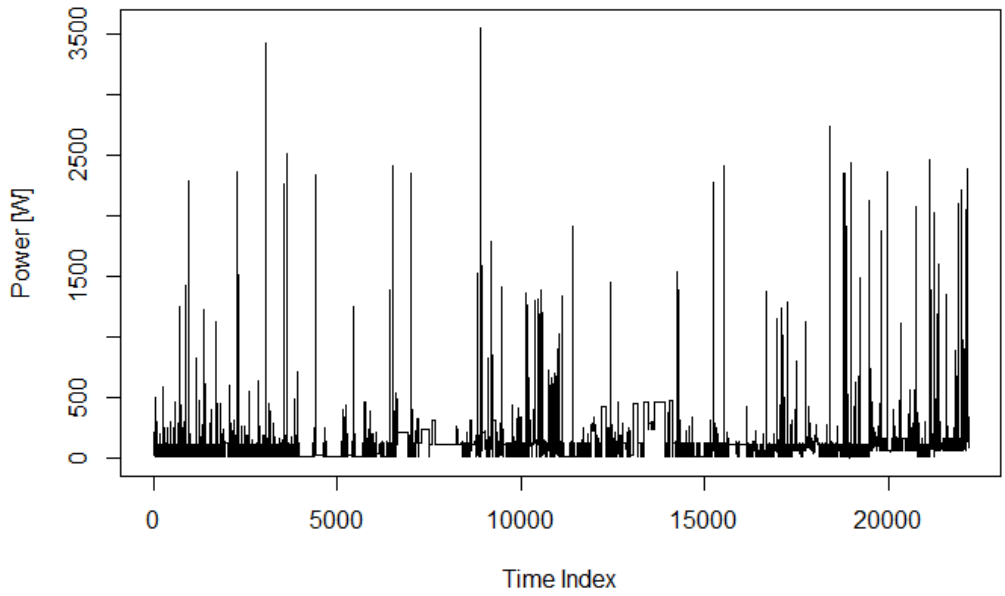

Figura 5: main della casa C4 sull'intervallo di dati disponibili

A differenza di quanto fatto in [4], la nostra idea è quella di costruire un modello di machine learning per ogni elettrodomestico per ogni cluster. Più in dettaglio, all'interno di ogni cluster, per ogni elettrodomestico, si è scelta la casa più vicina al centroide che presentasse dati di quell'elettrodomestico, per costruire il modello di riferimento di quell'elettrodomestico per quel cluster. Riportiamo nella seguente tabella le case scelte per costruire il modello per ogni elettrodomestico

	FRIDGE	DISHWASHER	WASHING MACHINE
CLUSTER C1,C6,C8,C10	C6	C10	C6
CLUSTER C2,C7,C9	C2	C9	C2
CLUSTER C3,C5	C3	-	C5
CLUSTER C4	-	-	-

Tabella 2: Serie storiche scelte per costruire il modello degli elettrodomestici in ogni cluster

Notiamo che su C4 non ci sono dati sufficienti per costruire un modello visto che ci sono solo un mese e mezzo di dati per la washing machine ma con pochissime attivazioni della stessa, come si vede in Figura 6.

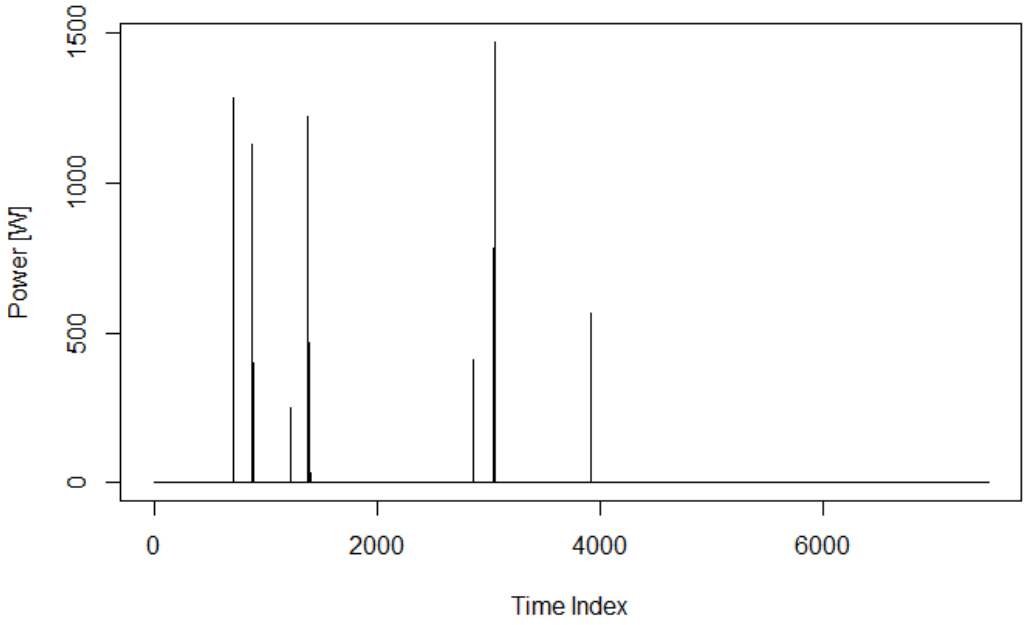

Figura 6: andamento della lavatrice della casa C4

2.3 Modelli di Machine Learning

Poiché i dati del consumo elettrico sia generale che dei singoli elettrodomestici sono aggregati su intervalli di 15 minuti, i dataset a disposizione per costruire i modelli sono di dimensioni contenute. Per questo motivo si sono utilizzate reti con architetture non troppo complesse, con un numero di strati tra uno e tre e numero di neuroni negli strati nascosti variabile nell'insieme {16, 32, 64, 128} e con funzioni di attivazione ReLU o sigmoide. Come funzione di loss per l'addestramento si è scelto in generale il MAE (Mean Absolute Error), ma per alcuni elettrodomestici per cui si avevano consumi molto bassi o poche attivazioni si è scelto il MSE (Mean Square Error), le cui espressioni vengono qui riportati

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t|\,, MSE = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2$$

dove $\hat{y}_t$ rappresenta la stima del consumo dell'elettrodomestico predetta dal modello e y_t il consumo reale all'istante di tempo t.
L'architettura è stata scelta tramite cross validation. E' stata addestrata una rete per ogni elettrodomestico utilizzando come input la serie main e come target la serie dell'appliance corrispondente.

Il test set corrisponde all'ultimo mese della coppia main-appliance e il validation set che è temporalmente consecutivo al training set viene utilizzato per scegliere l'architettura della rete (numero di strati e numero di neuroni in ciascun strato). Per quel che riguarda il numero ottimo di epoche viene scelto tramite early stopping per frigorifero e lavastoviglie, mentre per la lavatrice si guarda solo alla loss sul training set perché è difficile costruire un validation set rappresentativo visto il numero bassissimo di attivazioni di questo elettrodomestico.
La bontà del modello viene testata prima di tutto sul test set di quella coppia casa-appliance che corrisponde all'ultimo mese di dati disponibili.
Dopo di che il modello viene testato sull'intero insieme di dati delle altre case nel centroide.

Nelle Tabelle 3-5 sono riportate le architetture scelte per ogni elettrodomestico in ogni cluster:

CLUSTER C1-C6-C8-C10	SEQUENCE	LAYERS	LOSS
Fridge C6	4	Dense(units=32, activation="relu") Dense(units=4, activation="linear")	MAE
Washing Machine C6	8	Dense(units=64, activation='relu') Dropout(rate=0.1) Dense(units=32, activation="relu") Dropout(rate=0.1) Dense(units=64, activation="relu") Dense(units=8, activation="linear")	MSE
Dishwasher C10	4	Dense(units=64, activation="relu") Dropout(rate=0.1) Dense(units=32, activation="relu") Dropout(rate=0.1) Dense(units=64, activation="relu") Dense(units=4, activation="linear")	MAE

Tabella 3: Architetture scelte per gli elettrodomestici del cluster C1-C6-C8-C10

CLUSTER C3-C5	SEQUENCE	LAYERS	LOSS
Fridge C3	4	Dense(units=64, activation="relu") Dense(units=64, activation="relu") Dense(units=4, activation="linear")	MAE
Washing Machine C5	8	Dense(units=128, activation="relu") Dropout(rate=0.1) Dense(units=64, activation="relu") Dropout(rate=0.1) Dense(units=128, activation="relu") Dense(units=8, activation="linear")	MSE
Dishwasher C5	4	Dense(units=64, activation="relu") Dropout(rate=0.1) Dense(units=32, activation="relu") Dropout(rate=0.1) Dense(units=64, activation="relu") Dense(units=4, activation="linear")	MAE

Tabella 4: Architetture scelte per gli elettrodomestici del cluster C3-C5

CLUSTER C2-C7-C9	SEQUENCE	LAYERS	LOSS
Fridge C2	4	Dense(units=64, activation="relu") Dense(units=64, activation="relu") Dense(units=64, activation="relu") Dense(units=4, activation="linear")	MAE
Washing Machine C2	8	Dense(units=128, activation="relu") Dropout(rate=0.1) Dense(units=64, activation="relu") Dropout(rate=0.1) Dense(units=128, activation="relu") Dense(units=8, activation="linear")	MSE
Dishwasher C2	4	Dense(units=64, activation="relu") Dropout(rate=0.1) Dense(units=32, activation="relu") Dropout(rate=0.1) Dense(units=64, activation="relu") Dense(units=4, activation="linear")	MAE

Tabella 5: Architetture scelte per gli elettrodomestici del cluster C2-C7-C9

La colonna sequence si riferisce al fatto che per tenere traccia dell'andamento temporale della serie l'input alla rete viene fornito sotto forma di mini batch di 16 sequenze sovrapposte di una certa lunghezza. In figura 7 viene riportato un esempio di costruzione di input nel caso di sequenze di lunghezza 4 e batch di dimensione 4

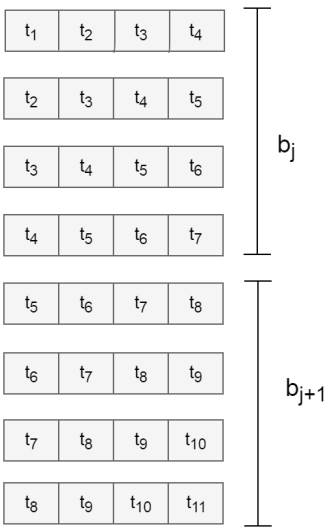

Figura 7: esempio di due mini batch da 4 con sequenze di lunghezza 4

La lunghezza della sequenza deve essere tale da permettere di catturare le attivazioni di quell'elettrodomestico e questo spiega la variabilità nelle tabelle, in quanto una sequenza di lunghezza 4 corrisponde a considerare le misurazioni in 1 ora (più che sufficiente per frigo e lavastoviglie), mentre una da 8 corrisponde a 2 ore che è la durata media di un ciclo di lavatrice.
I dati utilizzati per addestrare le varie reti sono costituiti dai dati aggregati a 15 minuti, ma per valutare la bontà dei risultati in modo che risulti fruibile al consumatore guardiamo alla predizione della percentuale di consumo dell'elettrodomestico rispetto al main complessivo nelle tre fasce orarie per giorno, in modo da avere una misura dei consumi aggregata facilmente interpretabile.

2.4 Risultati

In questa sezione riportiamo i risultati ottenuti discutendo prima quale sia il modo corretto di rappresentarli in modo facile da capire.

2.4.1 Metriche

Uno dei problemi del NILM è quello di valutare la bontà dei risultati, perché le metriche classiche utilizzate per analisi di serie storiche non sono facilmente interpretabili. Quando si può dire che un risultato effettivamente segue l'andamento della serie originale? Ad esempio non è facile stabilire quale sia un valore accettabile di MAE o di MSE.
Di recente questo problema è stato affrontato in [6] dove vengono definite le seguenti metriche di interesse:

Siano N il numero di osservazioni, P_t la potenza del meter all'istante t, $\widehat{P}_t$ la potenza nilm predetta all'istante t, $\Delta P_t = (\widehat{P}_t - P_t)$ l'errore tra la potenza nilm e quella del meter all'istante t.

RELATIVE ERROR (RE) = $\frac{\sum_{i=1}^{N} P_i - \sum_{i=1}^{N} \widehat{P}_i}{\sum_{i=1}^{N} P_i}$

AVERAGE ERROR (AE) = $\frac{1}{N}\sum_{i=1}^{N} \Delta P_i$

STANDARD DEVIATION OF ERROR (SDE) = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta P_i - AE)^2}$

MEAN ABSOLUTE ERROR (MAE) = $\frac{1}{N}\sum_{i=1}^{N} |\Delta P_i|$

MEAN SQUARED ERROR (MSE) = $\frac{1}{N}\sum_{i=1}^{N}(\Delta P_i)^2$

MATCH RATE (MR) = $\frac{\sum_{i=1}^{N} \min(P_i, \widehat{P}_i)}{\sum_{i=1}^{N} \max(P_i, \widehat{P}_i)}$

Il match rate in particolare è un sistema metrico in cui la valutazione si basa sul tasso di sovrapposizione tra le due serie di energia reale e di energia stimata. Varia tra 0 e 1. Se il valore è vicino a 1, la metrica indica una forte corrispondenza tra l'energia stimata e l'energia reale. Al contrario, un valore tendente a 0 indica una scarsa corrispondenza. Un valore di zero è solo possibile se l'energia vera e stimata sono entrambe pari a zero. Questa è la metrica che ha dimostrato le migliori prestazioni complessive [6] e per questo è quella che abbiamo adottato per valutare le prestazioni del modello.

Per questo motivo riportiamo prima i risultati in modo grafico, riportando la percentuale reale e quella predetta di consumo del singolo elettrodomestico nelle tre fasce orarie nei singoli giorni a disposizione. Questa costituisce una misura di interesse per il consumatore, in quanto permette di capire quale elettrodomestico sia più energivoro e in quale fase della giornata.

2.4.2 Risultati nei vari cluster

Consideriamo il cluster (C1, C6, C8, C10), e riportiamo nella figura 8 l'andamento del frigo sul suo test, in Figura 9 e 10 invece l'andamento del modello costruito su C6 sull'insieme di dati rispettivamente di casa C1 e C10.

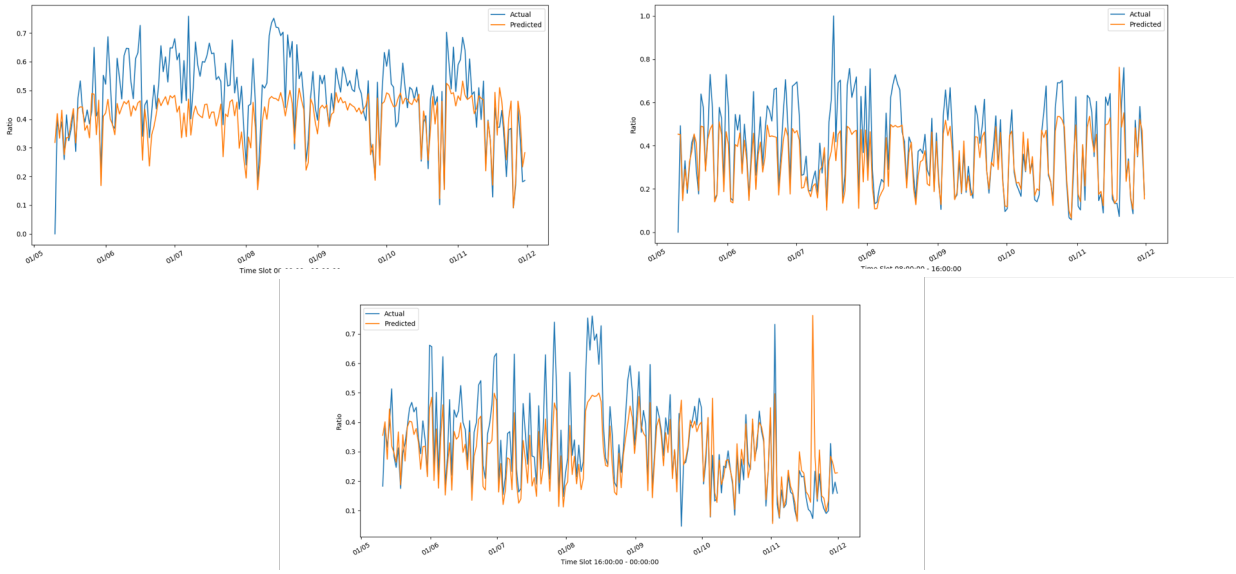

Figura 8: Andamento reale e stimato del frigorifero di C6 con il modello di C6 nelle tre fasce orarie

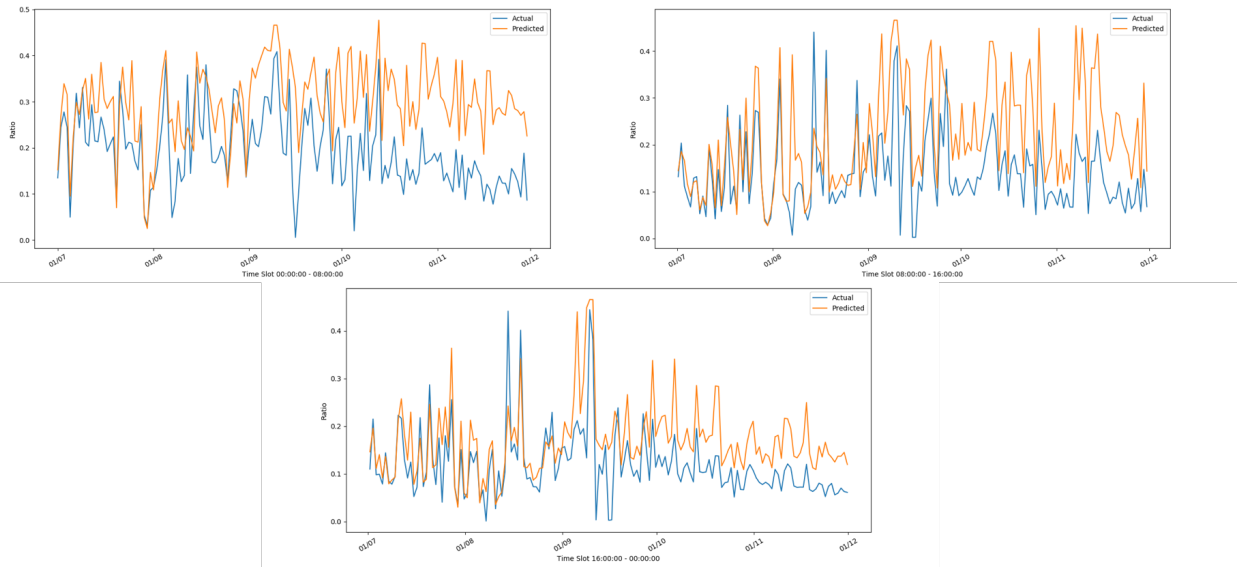

Figura 9: Andamento reale e stimato del frigorifero di C1 con il modello di C6 nelle tre fasce orarie

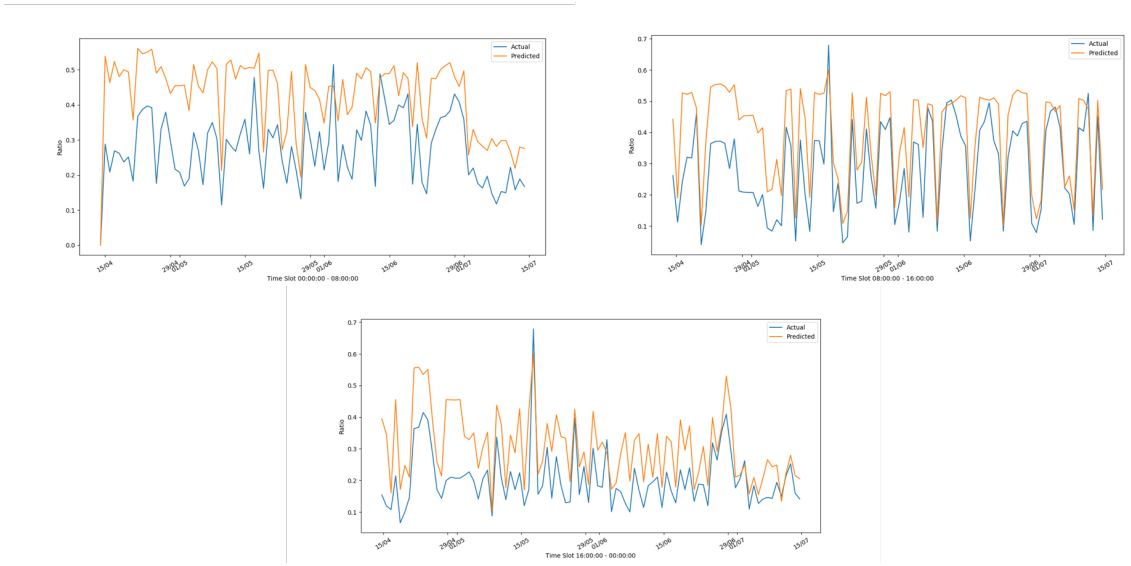

Figura 10: Andamento reale e stimato del frigorifero di C10 con il modello di C6 nelle tre fasce orarie

Come si può vedere dalle figure c'è una buona aderenza tra serie reale e serie predetta. Risultati simili si hanno sulla lavastoviglie il cui modello è costruito utilizzando l'insieme di dati della casa C10, in quanto C6 non ha dati sulla lavastoviglie. Nelle figure 11-13 riportiamo i risultati sulla lavastoviglie (che è presente solo nelle case C10 e C1).

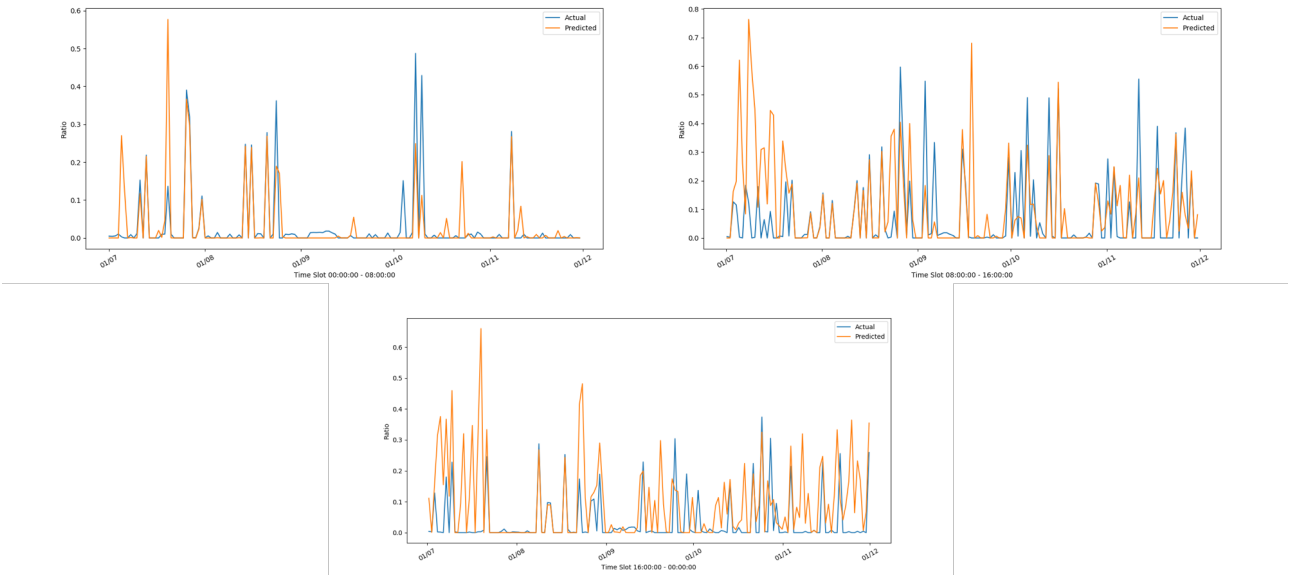

Figura 11: Andamento reale e stimato della lavastoviglie di C10 con il modello di C10 nelle tre fasce orarie

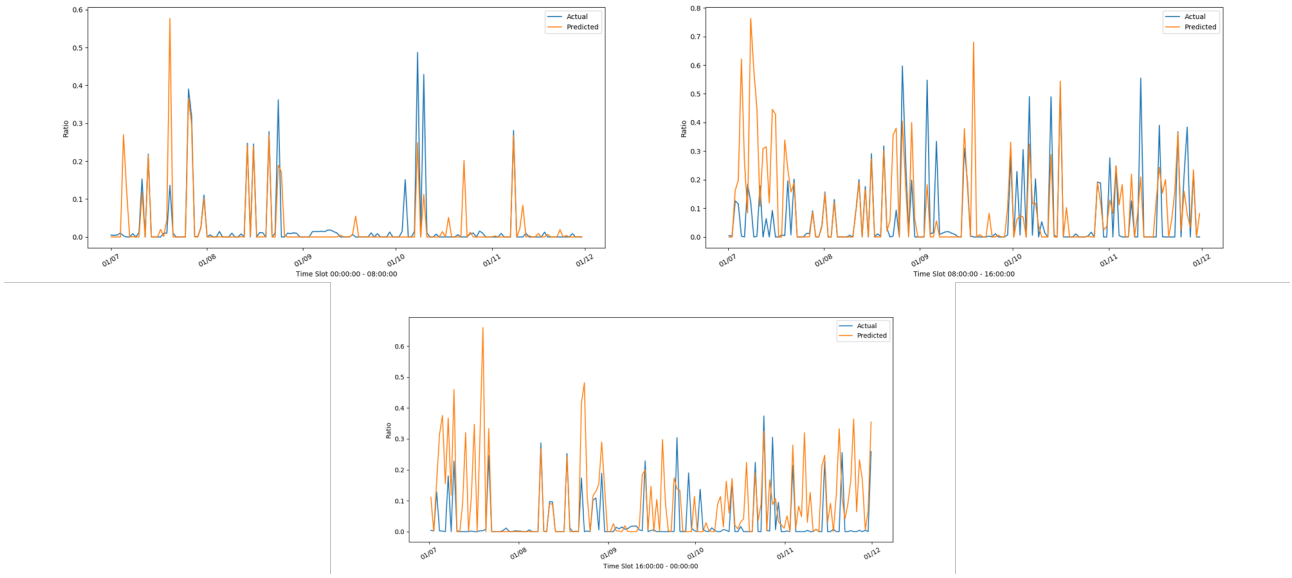

Figura 12: Andamento reale e stimato della lavastoviglie di C1 con il modello di C10 nelle tre fasce orarie

L'elettrodomestico più difficile da stimare è la lavatrice a causa della numerosità di attivazioni molto più bassa. Questo si vede guardando all'andamento reale e predetto nelle figure

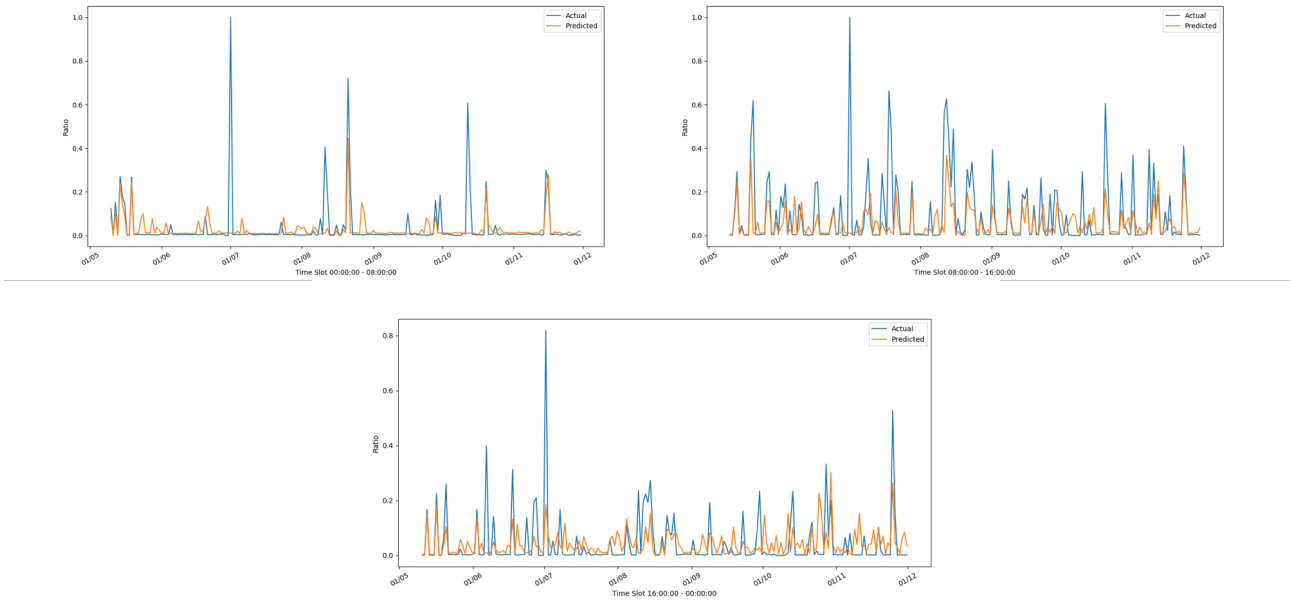

Figura 13: Andamento reale e stimato della lavatrice di C6 con il modello di C6 nelle tre fasce orarie

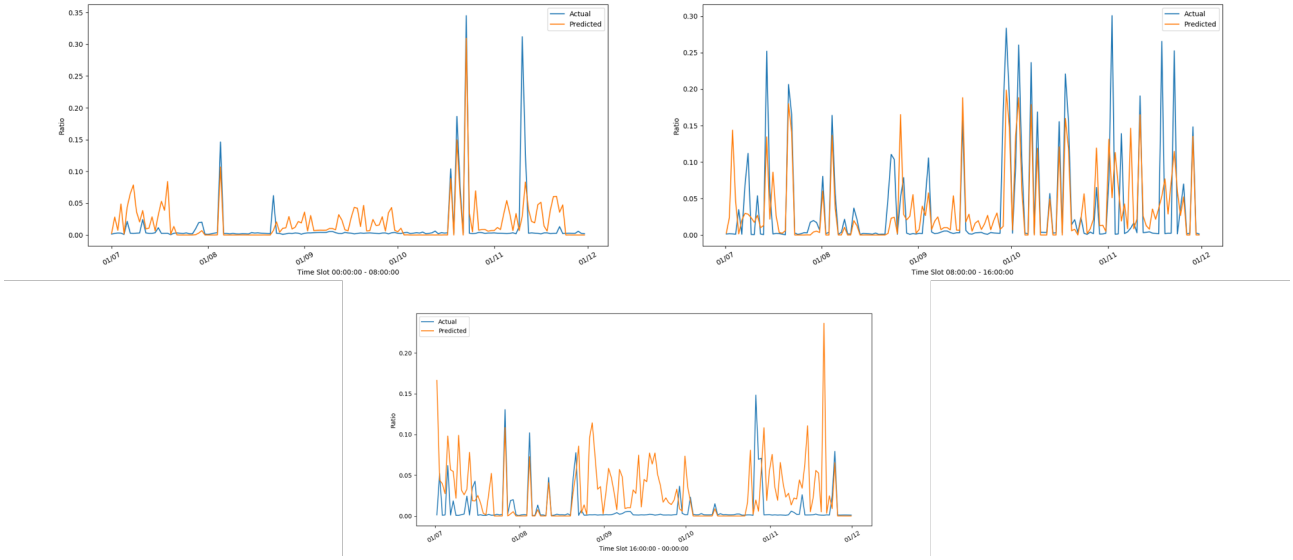

Figura 14: Andamento reale e stimato della lavatrice di C1 con il modello di C6 nelle tre fasce orarie

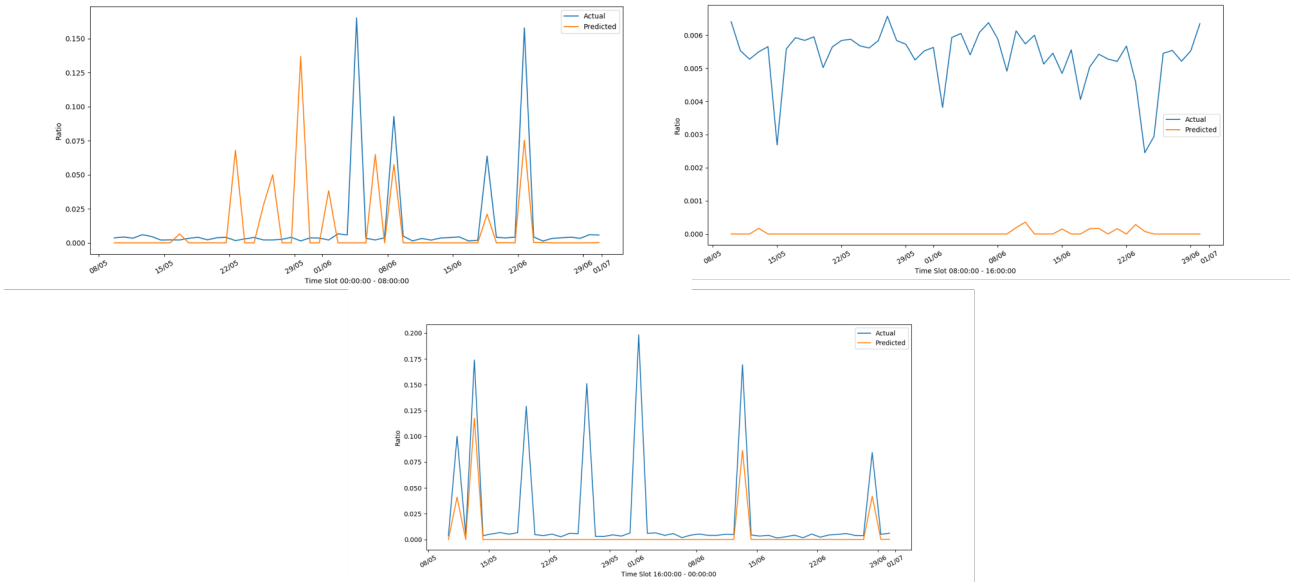

Figura 15: Andamento reale e stimato della lavatrice di C8 con il modello di C6 nelle tre fasce orarie

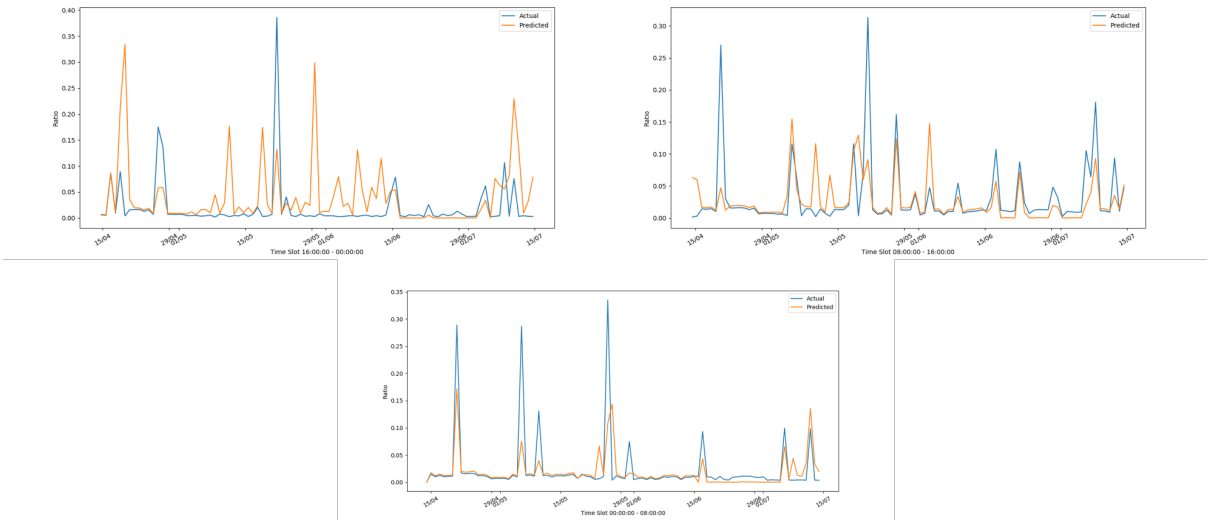

Figura 16: Andamento reale e stimato della lavatrice di C10 con il modello di C6 nelle tre fasce orarie

L'intuizione derivante dalle figure è confermata dalla tabella 6 in cui è riportato il match rate corrispondente alle figure 8-16:

CLUSTER C1-C6-C8-C10	MATCH RATE		
	Slot 1	Slot 2	Slot 3
Washing Machine C1 (C6)	0.315	0.452	0.179
Washing Machine C6 (C6)	0.321	0.355	0.281
Washing Machine C8 (C6)	0.164	0.062	0.238
Washing Machine C10 (C6)	0.460	0.479	0.253
Fridge C1 (C6)	0.628	0.604	0.659
Fridge C6 (C6)	0.809	0.793	0.803
Fridge C10 (C6)	0.642	0.709	0.659
Dishwasher C1 (C10)	0.485	0.401	0.266
Dishwasher C10 (C10)	0.573	0.767	0.833

Tabella 6: Match Rate per gli elettrodomestici e le case del cluster (C1, C6, C8, C10)

La tabella conferma la bontà dei modelli, con valori del match rate elevati soprattutto per frigo e lavastoviglie.

Passiamo adesso al cluster (C2, C7, C9). In questo caso i modelli del frigo e della lavastoviglie sono addestrati tramite i dati di C2, mentre la lavastoviglie è addestrata su C9. Nelle figure seguenti riportiamo la percentuale di consumo giornaliero predetti e reali dei vari elettrodomestici.

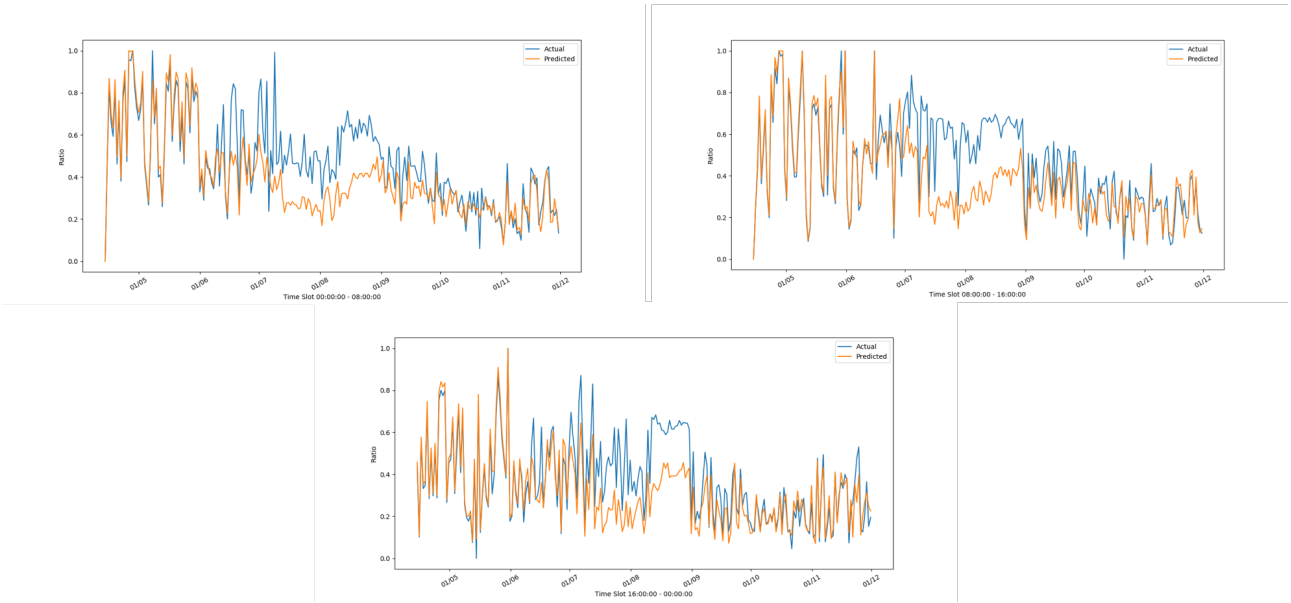

Figura 17: Andamento reale e stimato del frigorifero di C2 con il modello di C2 nelle tre fasce orarie

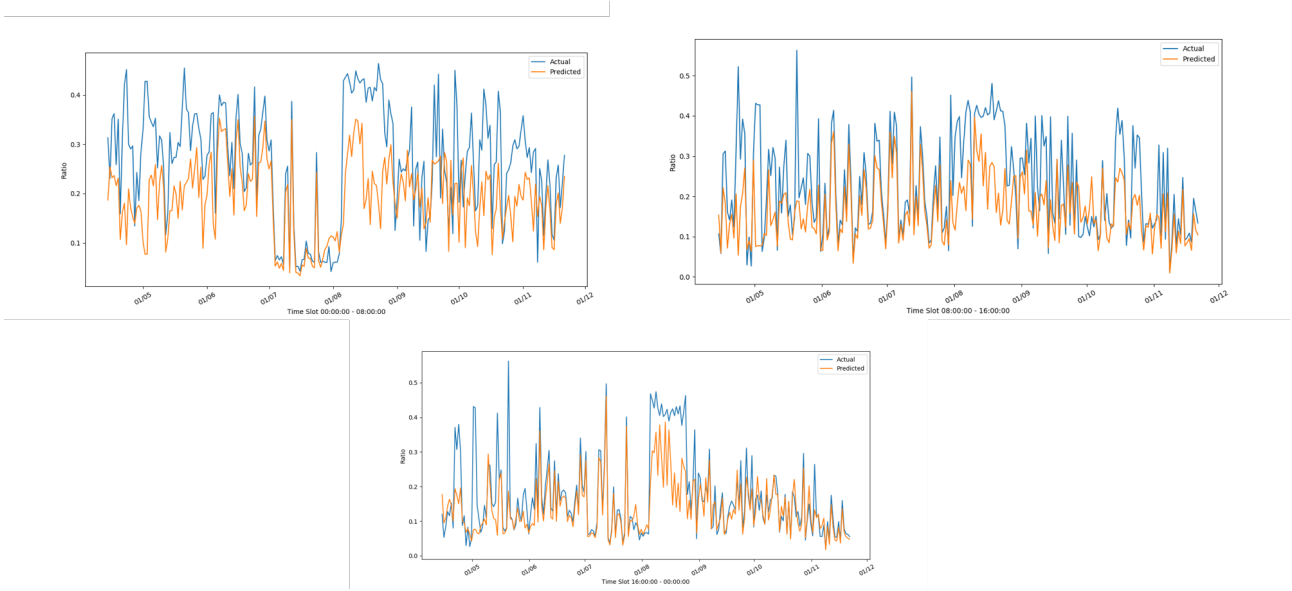

Figura 18: Andamento reale e stimato del frigorifero di C7 con il modello di C2 nelle tre fasce orarie

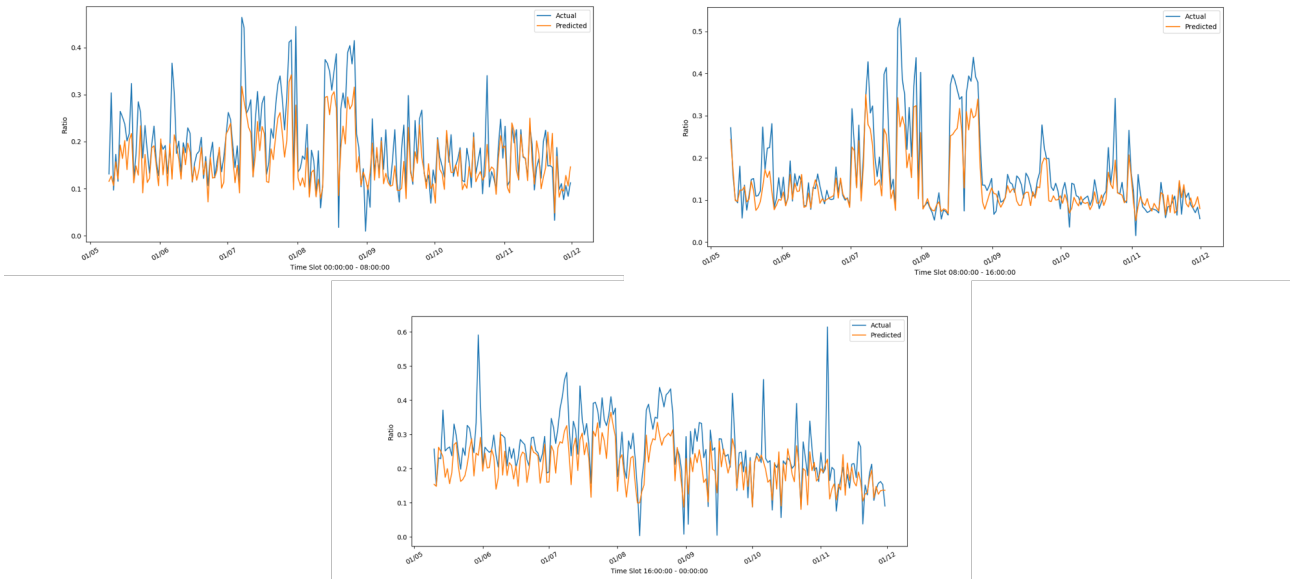

Figura 19: Andamento reale e stimato del frigorifero di C9 con il modello di C2 nelle tre fasce orarie

Come si vede c'è un'elevata corrispondenza tra le due serie.

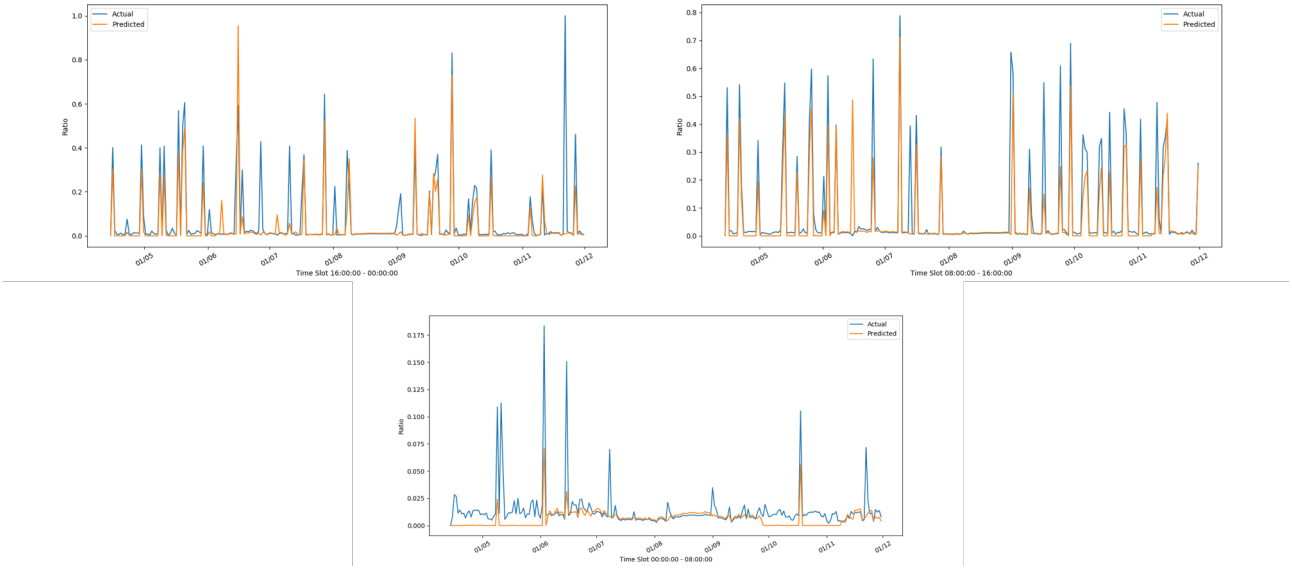

Figura 21: Andamento reale e stimato della lavatrice di C2 con il modello di C2 nelle tre fasce orarie

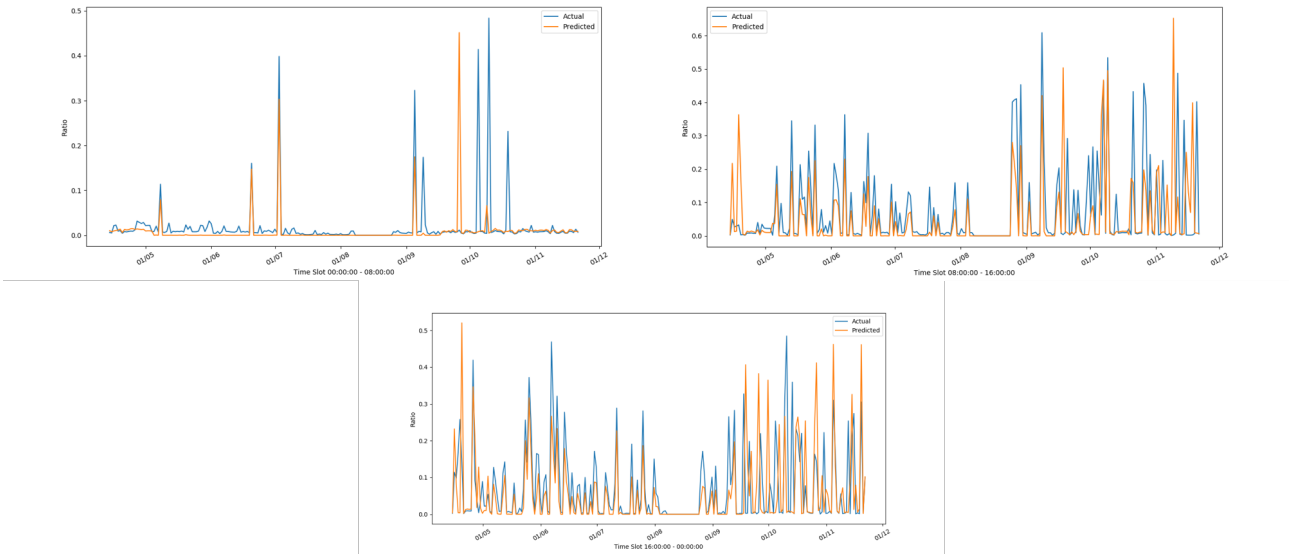

Figura 22: Andamento reale e stimato della lavatrice di C7 con il modello di C2 nelle tre fasce orarie

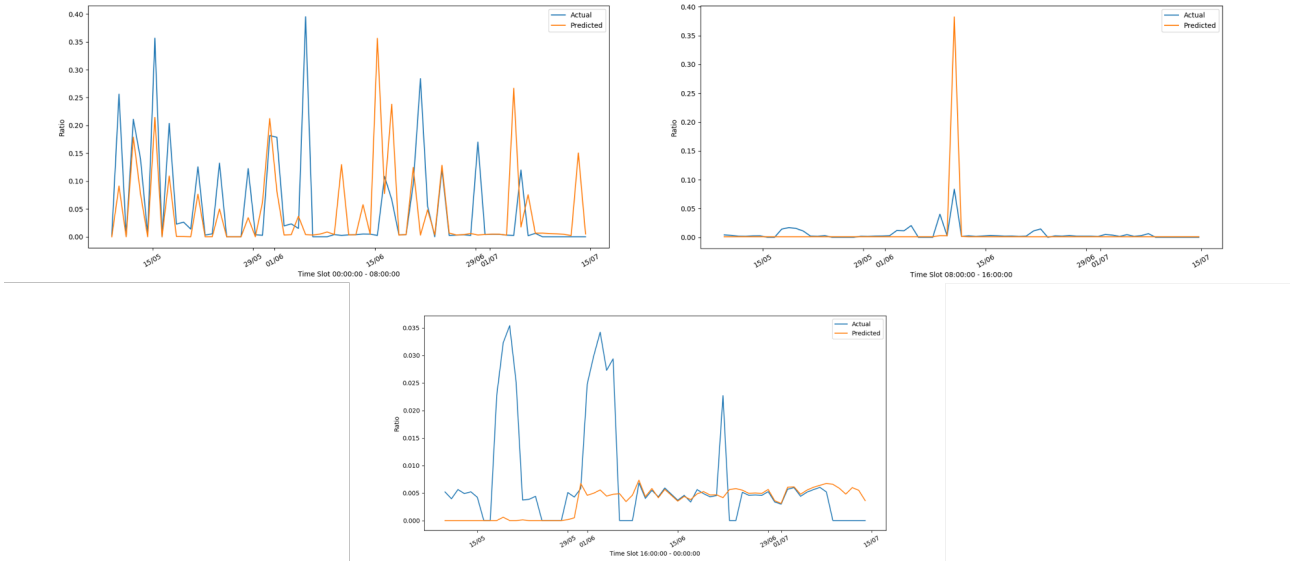

Figura 23: Andamento reale e stimato della lavatrice di C9 con il modello di C2 nelle tre fasce orarie

Come sempre la lavatrice è l'elettrodomestico più difficile da predire soprattutto nelle fasce orarie in cui ci sono poche attivazioni.
Infine riportiamo i grafici relativi alla lavastoviglie

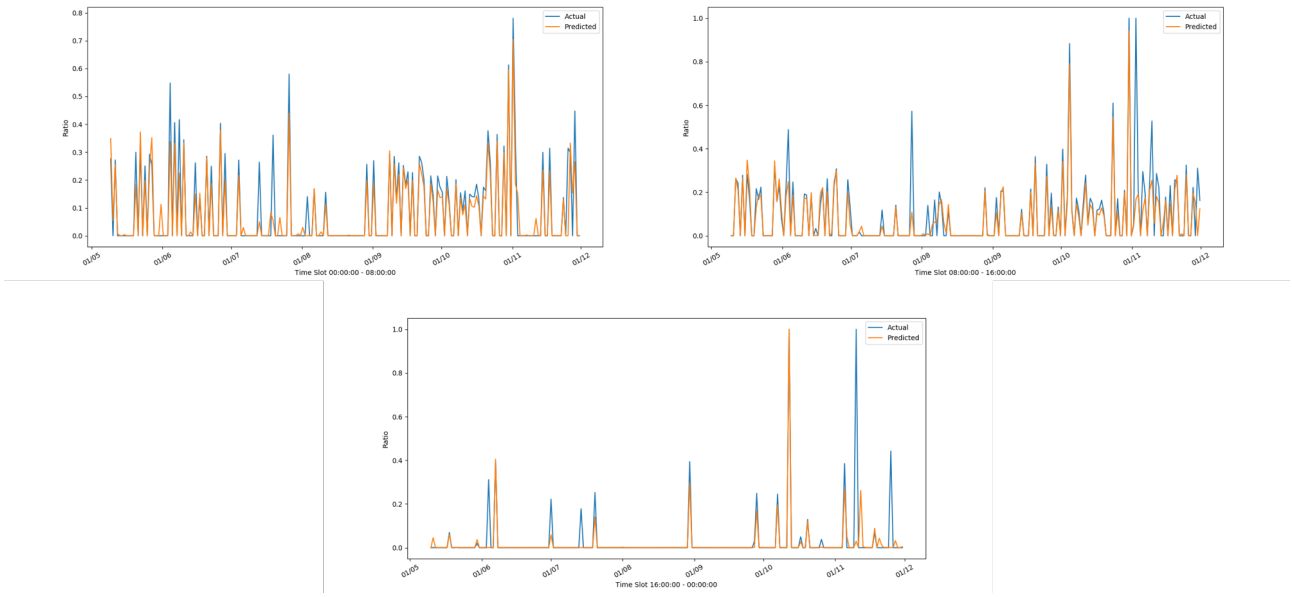

Figura 24: Andamento reale e stimato della lavastoviglie di C9 con il modello di C9 nelle tre fasce orarie

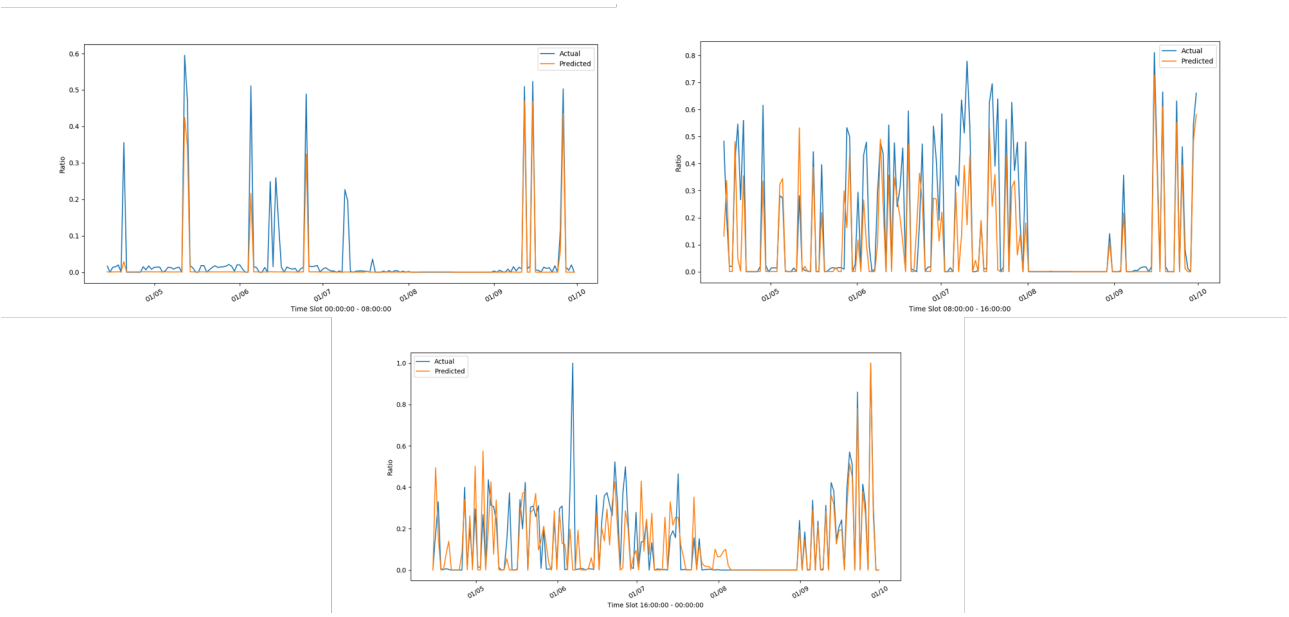

Figura 25: Andamento reale e stimato della lavastoviglie di C7 con il modello di C9 nelle tre fasce orarie

Di nuovo la bontà dei risultati può essere esaminata tramite i valori numerici del match rate, riportati in tabella 7.

CLUSTER C2-C7-C9	**MATCH RATE**		
	Slot 1	Slot 2	Slot 3
Washing Machine C2 (C2)	0.383	0.601	0.537
Washing Machine C7 (C2)	0.294	0.376	0.413
Washing Machine C9 (C2)	0.329	0.204	0.306
Fridge C2 (C2)	0.799	0.777	0.785
Fridge C7 (C2)	0.686	0.695	0.745
Fridge C9 (C2)	0.793	0.789	0.782
Dishwasher C7 (C9)	0.459	0.571	0.602
Dishwasher C9 (C9)	0.742	0.723	0.765

Tabella 7: Match Rate per gli elettrodomestici e le case del cluster (C2, C7, C9)

Infine riportiamo i risultati per il cluster (C3,C5) che sono comunque meno rappresentativi in quanto l'unico elettrodomestico presente in entrambe le case è la lavatrice.

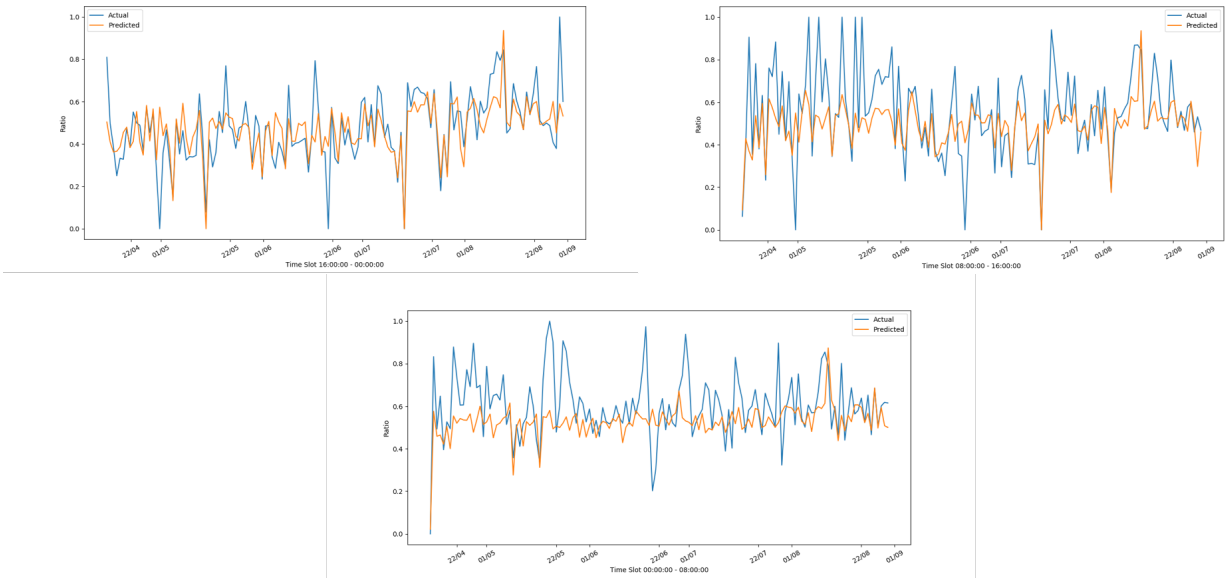

Figura 26: Andamento reale e stimato del frigorifero di C3 con il modello di C3 nelle tre fasce orarie

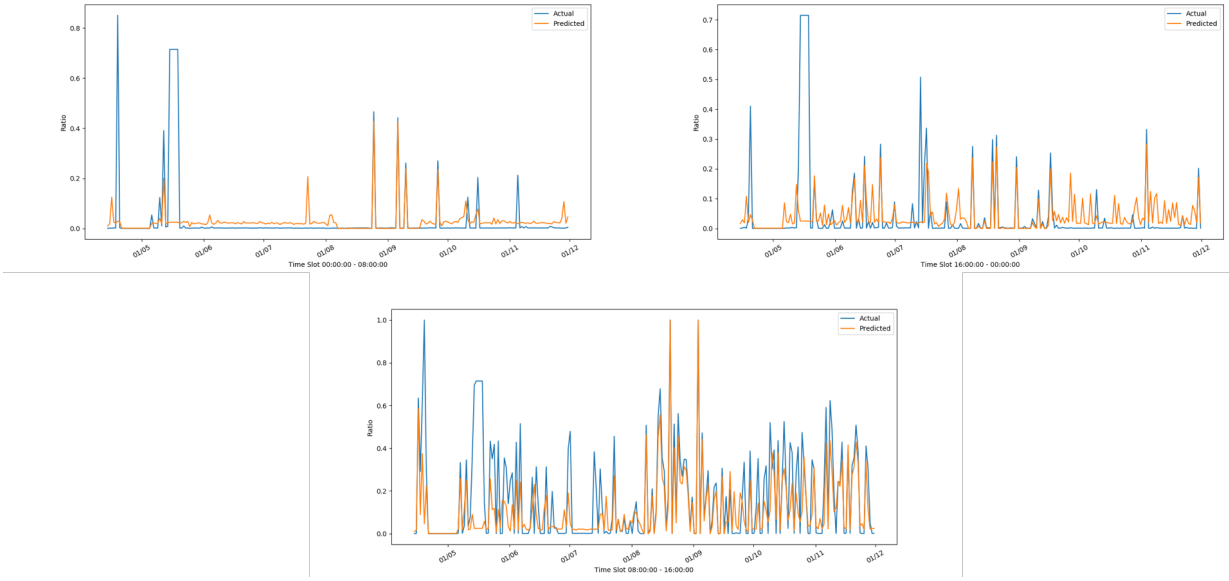

Figura 27: Andamento reale e stimato della lavatrice di C5 con il modello di C5 nelle tre fasce orarie

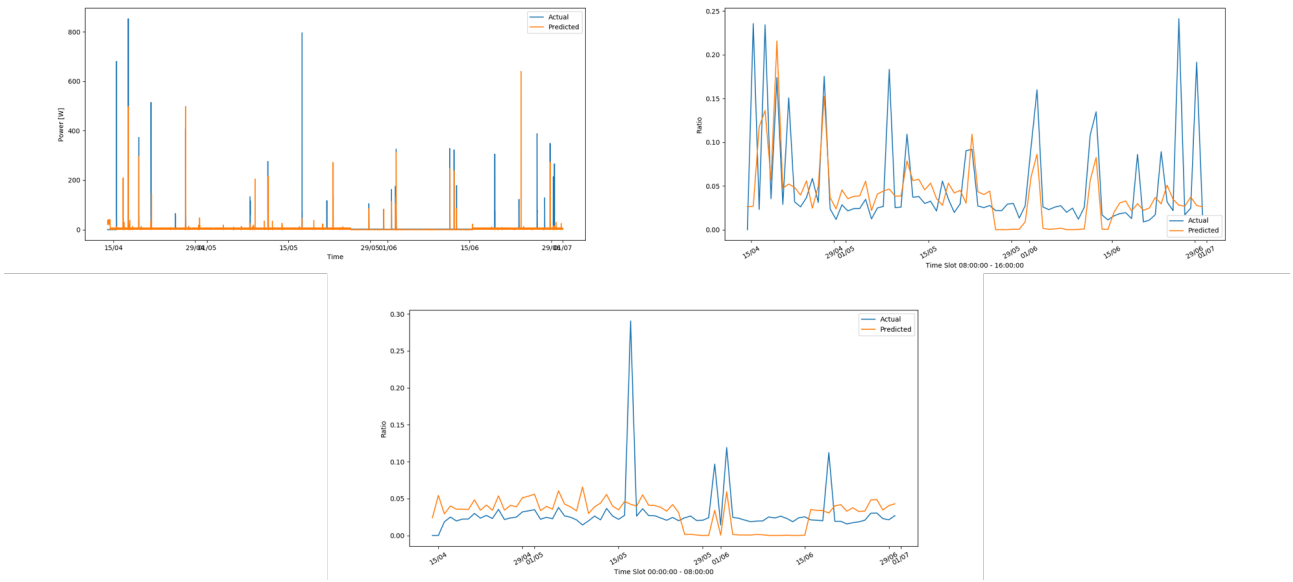

Figura 28: Andamento reale e stimato della lavatrice di C3 con il modello di C5 nelle tre fasce orarie

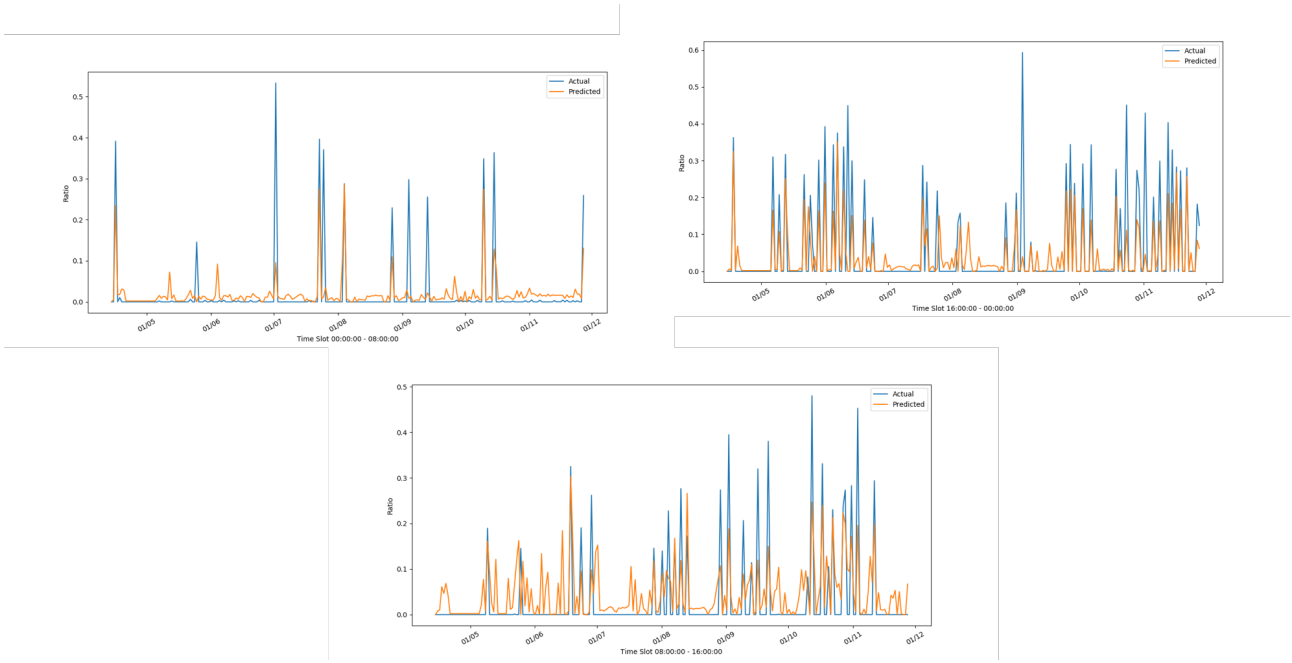

Figura 29: Andamento reale e stimato della lavastoviglie di C5 con il modello di C5 nelle tre fasce orarie

Di nuovo riportiamo in tabella i match rate corrispondenti:

CLUSTER C3-C5	**MATCH RATE**		
	Slot 1	Slot 2	Slot 3
Washing Machine C3 (C5)	0.451	0.471	0.452
Washing Machine C5 (C5)	0.195	0.528	0.381
Fridge C3 (C3)	0.818	0.806	0.848
Dishwasher C5 (C5)	0.358	0.306	0.443

Tabella 8: Match Rate per gli elettrodomestici e le case del cluster (C2, C7, C9)

A ulteriore riprova della bontà dei cluster costruiti si riporta per ogni elettrodomestico il comportamento dei modelli all'esterno del proprio cluster nelle tabelle 9-11.

HOUSE ID	MATCH RATE								
	Fridge C2			Fridge C3			Fridge C6		
	Slot 1	Slot 2	Slot 3	Slot 1	Slot 2	Slot 3	Slot 1	Slot 2	Slot 3
C1	0.613	0.600	0.649	0.595	0.573	0.660	**0.628**	**0.604**	**0.659**
C2	**0.799**	**0.777**	**0.785**	0.718	0.746	0.755	0.766	0.766	0.763
C3	0.695	0.710	0.711	**0.818**	**0.806**	**0.848**	0.802	0.762	0.796
C6	0.582	0.583	0.557	0.801	0.788	0.798	**0.809**	**0.793**	**0.803**
C7	**0.686**	**0.695**	**0.745**	0.502	0.566	0.579	0.597	0.616	0.618
C9	**0.793**	**0.789**	**0.782**	0.654	0.699	0.626	0.672	0.694	0.707
C10	0.549	0.421	0.638	0.509	0.536	0.604	**0.642**	**0.709**	**0.659**

Tabella 9: Match Rate relativo al frigorifero per ogni casa per ogni modello costruito

HOUSE ID	MATCH RATE								
	Washing Machine C2			Washing Machine C5			Washing Machine C6		
	Slot 1	Slot 2	Slot 3	Slot 1	Slot 2	Slot 3	Slot 1	Slot 2	Slot 3
C1	0.326	0.274	0.107	0.220	0.305	0.144	**0.315**	**0.452**	**0.179**
C2	**0.383**	**0.601**	**0.537**	0.312	0.452	0.410	0.517	0.281	0.294
C3	0.349	0.213	0.313	**0.451**	**0.471**	**0.452**	0.483	0.410	0.464
C5	0.120	0.305	0.144	**0.195**	**0.528**	**0.381**	0.101	0.328	0.253
C6	0.167	0.183	0.138	0.149	0.229	0.167	**0.321**	**0.355**	**0.281**
C7	**0.294**	**0.376**	**0.413**	0.195	0.267	0.286	0.276	0.223	0.331
C8	0.070	0.233	0.196	0.072	0.118	0.201	**0.164**	**0.062**	**0.238**
C9	**0.329**	**0.204**	**0.306**	0.199	0.075	0.192	0.292	0.113	0.201
C10	0.321	0.297	0.192	0.360	0.307	0.170	**0.460**	**0.479**	**0.253**

Tabella 10: Match Rate relativo alla lavatrice per ogni casa per ogni modello costruito

HOUSE ID	MATCH RATE								
	Dishwasher C5			Dishwasher C9			Dishwasher C10		
	Slot 1	Slot 2	Slot 3	Slot 1	Slot 2	Slot 3	Slot 1	Slot 2	Slot 3
C1	0.249	0.358	0.275	0.304	0.398	0.384	**0.485**	**0.401**	**0.266**
C5	**0.358**	**0.306**	**0.443**	0.259	0.211	0.351	0.229	0.232	0.256
C7	0.332	0.463	0.382	**0.459**	**0.571**	**0.602**	0.446	0.520	0.382
C9	0.464	0.420	0.398	**0.742**	**0.723**	**0.765**	0.538	0.518	0.344
C10	0.333	0.550	0.454	0.344	0.577	0.490	**0.573**	**0.767**	**0.833**

Tabella 11: Match Rate relativo alla lavastoviglie per ogni casa per ogni modello costruito

Le tabelle confermano la bontà dei cluster scelti in quanto i modelli costruiti su case dello stesso cluster funzionano sempre meglio (hanno match rate più alto) rispetto a quelli esterni al cluster. Alcune rare eccezioni si hanno nelle fasce orarie dove l'elettrodomestico viene utilizzato meno.

3 Conclusioni

In questo lavoro si è definito un algoritmo di clustering per individuare abitazioni con consumi elettrici simili I risultati ottenuti dai modelli di machine learning sono molto buoni in relazione alle metriche di performance impiegate in letteratura per questa classe di problemi. Questi risultati permettono di validare i cluster in termini delle caratteristiche utilizzate per individuare profili comuni di consumo tra le abitazioni. Come ulteriore prova della bontà dei cluster, i risultati delle case testate su modelli che non fanno parte del proprio cluster di appartenenza evidenziano che la disaggregazione è più precisa se si utilizzano modelli addestrati su case con profili di consumo simili. Infine, lo scenario tipico che segue la messa in produzione dei modelli è quello di estrarre le caratteristiche del consumo complessivo di un'abitazione, calcolare la distanza euclidea tra la casa in oggetto e i centroidi dei cluster individuati, assegnare la casa al cluster per cui la distanza è minima e disaggregare il consumo energetico utilizzando il modello dell'elettrodomestico per il cluster di appartenenza.

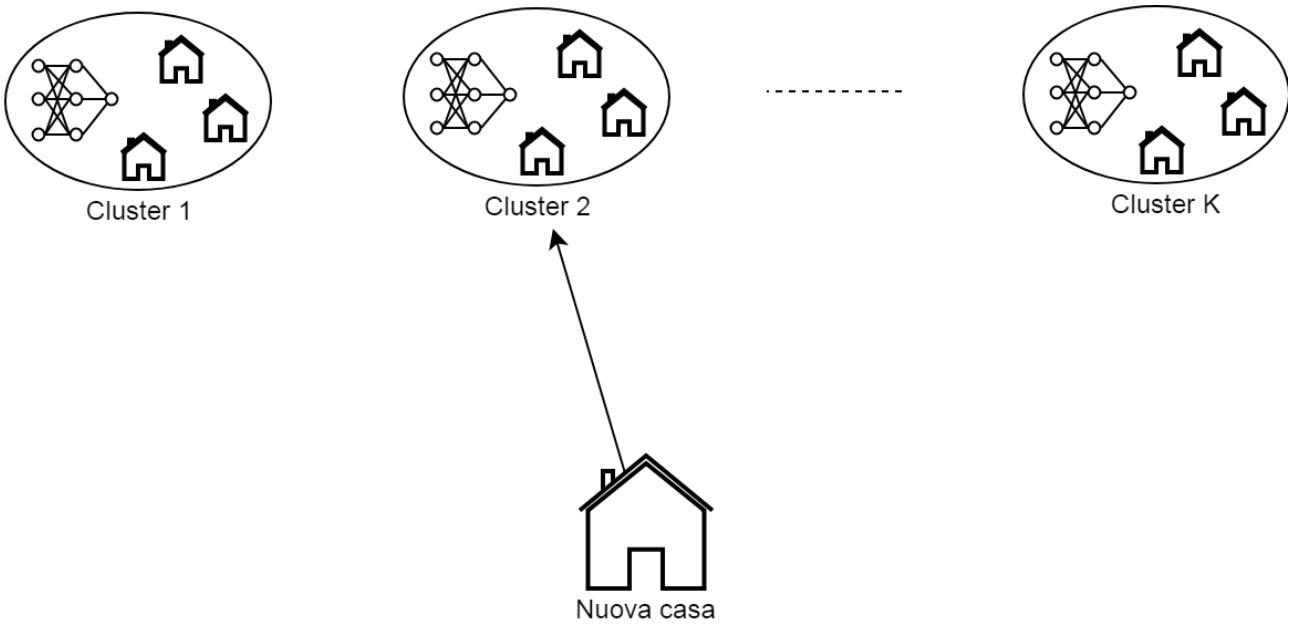

Figura 30: strategia di clustering per la disaggregazione del consumo elettrico

4 Riferimenti bibliografici

1. M. Figueiredo, A. de Almeida, B. Ribeiro, "Home electrical signal disaggregation for non-intrusive load monitoring (NILM) systems", Neurocomputing, 96, 2012, 66-73.
2. W. He, Y. Chai, "An Empirical Study on Energy Disaggregation via Deep Learning", *Advances in Intelligent Systems Research*, *133, 2016*.
3. J. Kelly, W. Knottenbelt "Neural nilm: Deep neural networks applied to energy disaggregation" In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments* (pp. 55-64). ACM.
4. Batra, N., Singh, A., & Whitehouse, K. (2015). Neighbourhood nilm: A big-data approach to household energy disaggregation. *arXiv preprint arXiv:1511.02900*.
5. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.
6. Pereira, Lucas, and Nuno Nunes. "Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2018): e1265.

5 Abbreviazioni ed acronimi

NILM: Nonintrusive Load Monitoring

MAE: mean absolute error

MSE: mean square error

MR: match rate