POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Elettrica

Tesi di Laurea Magistrale

Indicatori di flessibilità e raggruppamento dei carichi residenziali per la partecipazione ai programmi di Demand Response

Relatore

Prof. Gianfranco Chicco

Candidato

Matteo Guerrieri

Anno accademico 2020/2021

Indice

Introduzione 3
1 Metodi di clustering per i sistemi elettrici 5
1.1 Classificazione dei metodi di clustering 5
1.2 Tecniche di clustering 7
1.3 Indici di validità dei risultati del clustering 20
1.4 Applicazione dei metodi di clustering per curve di carico residenziali 25
1.5 Bibliografia 27
2 Base di dati per l'analisi di curve di carico residenziali 28
2.1 Esempi di dataset disponibili 28
2.2 Pre-elaborazione dei dati 31
2.3 Bibliografia 35
3 Analisi sulla flessibilità della domanda aggregata 36
3.1 Introduzione ai programmi di Demand Response 37
3.2 Variazioni di domanda per curve di carico residenziali aggerate 38
3.3 Analisi probabilistica e statistica delle variazioni di carico aggregato 42
3.4 Indicatori di flessibilità per carichi residenziali aggregati 48
3.5 Caso studio 57
3.6 Bibliografia 74
4 Selezione degli utenti per la partecipazione ai programmi di Demand Response 75
4.1 Clustering per l'identificazione di utenti target 76
4.2 Caso studio 78
4.3 Bibliografia 88
Conclusione 89

Introduzione

Le reti elettriche sono strutture complesse e fortemente interconnesse, le quali richiedono processi di modernizzazione, al fine di adeguarsi alle nuove tecnologie disponibili e far rispettare le specifiche vigenti, come quelle legate all'ambiente, che impongono nuovi scenari di transizione energetica, raggiungibili con l'elettrificazione e la decarbonizzazione. Le moderne reti di distribuzione consentono di integrare risorse energetiche distribuite (*Distributed Energy Resources*: *DER*), di incrementare l'utilizzo di fonti di energia rinnovabile (*Renewable Energy Sources*: *RES*), di incentivare la diffusione di veicoli elettrici (*Electric Vehicles*: *EV*) e di fornire nuovi servizi al cliente. Al fine di abilitare strutture che permettono la transizione energetica, gli operatori del sistema di distribuzione devono essere in grado di garantire affidabilità, efficienza, qualità, sicurezza e resilienza nelle reti. La digitalizzazione gioca un ruolo importante in questi sistemi, in quanto, grazie ai contatori intelligenti (*smart meter*), che oltre a contribuire allo sviluppo delle "reti del futuro" (*smart grid*), consentono di utilizzare le informazioni dei consumi dei singoli utenti residenziali per realizzare nuove strategie al fine di aiutare il conseguimento delle nuove sfide. Infatti, tecniche di gestione della domanda come il *demand-side management* (*DSM*) permettono di raggiungere l'equilibrio tra produzione e domanda di energia, con lo scopo di ottenere benefici economici, di efficienza e di gestione del sistema, ad esempio la riduzione dei picchi di domanda. In particolare, la tecnica di *Demand Response* (*DR*) consiste nella partecipazione attiva degli utenti sulla gestione della domanda, i quali modificano i propri consumi su richiesta tramite incentivi o attraverso segnali basati sul prezzo, ottenendo una gestione della rete più efficiente ed economica. Tuttavia, tale tecnica è maggiormente utilizzata per utenti commerciali o industriali, i quali oltre ad avere un maggiore consumo rispetto a singoli utenti residenziali, presentano dei modelli di comportamento più regolari. La finalità di questo elaborato è di illustrare metodologie ed applicazioni utili a diversi soggetti del sistema elettrico, come gli operatori dei sistemi di distribuzione (*Distribution System Operators*: *DSO*) o gli aggregatori, per facilitare e sfruttare al meglio l'avviamento dei programmi di DR per gruppi di utenti residenziali.

Storicamente le informazioni sui consumi dei singoli utenti residenziali erano disponibili in modo cumulativo (trimestralmente) non fornendo informazioni sui consumi giornalieri o infragiornalieri. Le caratteristiche dei consumi degli utenti, al fine di raggrupparli con proprietà simili ed effettuare analisi specifiche, erano solitamente basate su ulteriori dati, come quelli sociodemografici, sulla tipologia di abitazione, sul numero di occupanti o su questionari riguardanti la loro occupazione e il loro stile di vita. Tuttavia, ora è possibile applicare tecniche avanzate di analisi dei dati (*data mining*) sulle misure effettuate dagli smart meter (aventi alte risoluzioni), le quali consentono di individuare e raggruppare utenti in maniera più innovativa e accurata, estraendo dai dati stessi le proprietà specifiche. I metodi di *clustering*, particolari tecniche di intelligenza artificiale (*Artificial Intelligence*: *AI*), vengono utilizzati in diversi ambiti scientifici, allo scopo di estrarre, in modo automatico, informazioni utili dai dati disponibili raggruppandoli per caratteristiche simili. Essi, ampiamente utilizzati in svariati ambiti scientifici, trovano applicazioni anche nel campo dei sistemi elettrici, come: il riconoscimento del comportamento degli utenti e l'identificazione dei loro modelli di consumo, il miglioramento dell'offerta tariffaria, la disaggregazione del carico in un'abitazione (conosciuta come *NILM*), la segmentazione della clientela, la manutenzione predittiva per componenti del sistema elettrico, ecc. Tuttavia, l'utilizzo di tali tecniche sui dati dei consumi di utenti residenziali, richiedono particolari attenzioni in quanto presentano un'alta variabilità nell'utilizzo di apparecchiature nei diversi giorni rendendo difficile l'identificazione di modelli di consumo appropriati. Differenti soluzioni a queste problematiche sono proposte nei successivi

capitoli. Nonostante ciò, le caratteristiche ottenute dalle informazioni disponibili sui consumi, consentono all'operatore del sistema di fornire strategie di gestione della domanda specifiche. Poiché la domanda sta svolgendo un ruolo sempre più importante nella gestione dei sistemi elettrici, indicazioni riguardo la potenziale flessibilità ottenibile da essa, sono necessarie per l'operatore del sistema che intende instaurare programmi di DR ai suoi clienti. In questo contesto, la flessibilità della domanda aggregata viene intesa come la disponibilità di far cambiare i consumi degli utenti, in intervalli di tempo specifici, valutando il comportamento collettivo di un gruppo di carichi. Lo studio di utenti residenziali a livello aggregato è motivato sia perché il singolo utente non fornisce un importante contributo nella gestione della domanda, e sia perché i carichi aggregati presentano curve di carico più regolari e quindi più facili da trattare. Dunque, l'individuazione dei periodi con maggiore flessibilità informa l'operatore della DR su quali intervalli della giornata conviene avviare i programmi. Inoltre, è conveniente sapere anche a quali gruppi di utenti rivolgersi per richiedere la partecipazione ai programmi specifici. A tale scopo, l'operatore potrebbe costruirsi un *portfolio* clienti, differenziando per gruppi gli utenti con uguali modelli di consumo, al fine di proporre programmi specifici. Tuttavia, la variabilità dei loro consumi nei diversi giorni deve essere studiata appositamente, in modo da distinguere gli utenti più stabili da quelli più variabili, con l'obiettivo di ottenere clienti *target*, ai quali l'operatore della DR, insieme all'informazione del modello dei loro consumi, potrebbe offrire, con priorità, gli incentivi per la partecipazione ai programmi specifici negli intervalli con maggiore flessibilità.

Il *Capitolo 1* presenta una panoramica sui più comuni metodi di clustering applicati nell'ambito dei sistemi elettrici, evidenziando le metriche che utilizzano, gli indici di validità dei risultati e le diverse metodologie di applicazione per gli utenti residenziali, i quali richiedono particolari attenzioni.

Nel *Capitolo 2*, dopo la proposta di diverse basi di dati di utenti residenziali, disponibili e facilmente accessibili, vengono eseguite su alcune di esse delle verifiche, eliminando dati anomali e ricostruendo quelli mancanti, al fine di renderli utilizzabili negli studi successivi.

Nel *Capitolo 3* viene eseguita un'analisi sulla flessibilità del comportamento degli utenti residenziali nella domanda aggregata, valutando l'effetto che, il periodo di campionamento dei dati e il livello di aggregazione degli utenti, presentano sulle variazioni di carico e quindi sugli indicatori di flessibilità definiti.

Il *Capitolo 4* illustra un'ulteriore applicazione utile per facilitare e sfruttare al meglio i programmi di DR; dopo una segmentazione delle curve di carico degli utenti in base ai loro modelli di consumo, attraverso un calcolo basato sul concetto di entropia viene effettuata una classificazione di essi in base al loro livello di stabilità dei consumi nei diversi giorni, e ciò consente di identificare clienti target a cui proporre con priorità la partecipazione ai programmi specifici.

Per l'implementazione dei differenti algoritmi e lo svolgimento delle analisi effettuate, si è utilizzato il software *Matlab 2020b*.

1 Metodi di clustering per i sistemi elettrici

Il *clustering* o *cluster analysis* è l'insieme delle attività che hanno lo scopo di raggruppare e classificare unità statistiche, da un set di dati, attraverso l'identificazione di caratteristiche comuni. A differenza della *classificazione*, che cerca di distinguere i vari oggetti e assegnarli in determinate classi definite a priori con le rispettive caratteristiche, il clustering è una tecnica di apprendimento automatico (*machine learning*) di tipo non supervisionato, in cui l'algoritmo deve riuscire a identificare gli oggetti che si somigliano e raggrupparli tra loro; le diverse misure di similarità o dissimilarità tra gli elementi, sono alla base dei diversi approcci di raggruppamento. La cluster analysis è ad oggi uno degli strumenti più potenti e utilizzati nel data mining e nelle tecniche comuni per l'analisi dei dati, utilizzata in molti campi come pattern recognition, image analysis, computer graphics, data compression e predictive maintenance. Questa analisi, dunque, trova applicazione in svariati ambiti dove il partizionamento dei dati disponibili è essenziale per la ricerca di modelli interpretativi della realtà. Alcuni esempi si riscontrano in biologia, medicina per l'imaging medico o analisi dell'attività antimicrobica, informatica, marketing per il market research, scienze sociali per l'analisi del crimine o data mining educativo, ecc.

1.1 Classificazione dei metodi di clustering

Il clustering non è un algoritmo specifico, ma può essere ottenuto da varie tecniche che differiscono in modo significativo sia per via della definizione stessa di *cluster* e sia per il procedimento che porta alla loro determinazione [1,2]. Non esiste un algoritmo di clustering "corretto", ma deve essere scelto in modo appropriato all'applicazione d'interesse. Un algoritmo progettato per un tipo di modello, generalmente fallisce se viene applicato su un insieme di dati che contiene un modello radicalmente diverso. I metodi di clustering sono quindi caratterizzati sia dal procedimento con il quale svolge la ricerca dei gruppi, e da una misura del grado di similarità tra coppie di elementi; modificando uno o l'altro fattore, si può ottenere un'innumerevole varietà di metodi.

Classificazione per modelli di cluster

Come specificato precedentemente, uno dei motivi dell'ampia varietà dei metodi, sta nel concetto stesso di gruppo; esistono infatti diversi modelli di cluster su cui si basano i vari algoritmi, trovando interesse a seconda dell'applicazione. Alcuni modelli tipici sono:

- Modelli basati sulla distanza (*distance models*): algoritmi basati su questi modelli sono molto popolari in letteratura, perché generalmente possono essere usati con qualunque tipologia di dati, purché si possa costruire un'adeguata funzione di distanza tra le varie unità o rispetto ad un valore centrale, detto centroide o medoide a seconda se appartiene o meno al set di dati.
- Modelli basati sulla densità (*density-based models*): questi modelli definiscono i cluster come aree di maggiore densità rispetto al resto del set di dati.
- Modelli basati sui gruppi (*group models*): sono modelli non molto raffinati, in cui gli algoritmi di questo tipo forniscono solo informazioni generali sul raggruppamento.

- Modelli basati su distribuzioni statistiche (*distribution models*): i cluster vengono modellati utilizzando distribuzioni statistiche.

- Modelli basati su sottospazi (*subspace models*): utilizzati in biclustering, ovvero tecniche che consentono il clustering simultaneo delle righe e colonne di una matrice di dati $M \times N$ formata da M oggetti e N dimensioni ovvero caratteristiche del singolo oggetto. Perciò l'algoritmo fornisce un sottoinsieme di righe che mostra un comportamento simile in un sottoinsieme di colonne o viceversa;

- Modelli basati su grafi (*graph-based models*): i cluster rappresentano una *clique*, ovvero un sottoinsieme delimitato di nodi, in un grafo non orientato.

- Modelli basati su grafi orientati (*signed graph models*): gli algoritmi basati su questo modello, fanno riferimento ai grafi che presentano un segno specifico ad ogni ramo.

- Modelli neurali (*neural models*): sono modelli basati su reti neurali artificiali, ispirati ad una semplificazione di una rete neurale biologica; essi presentano caratteristiche simili a uno o più modelli definiti precedentemente.

Classificazione per tipo di algoritmo

La classificazione più diffusa distingue gli algoritmi di clustering in *gerarchici* e *non gerarchici* (o *partizionali*). I primi presentano delle procedure iterative che considerano tutti i livelli di similarità tra i gruppi, ottenendo così un'ampia gerarchia di cluster dalla loro fusione o divisione su ogni livello. I metodi gerarchici vengono ulteriormente distinti in *agglomerativi* o *divisivi* a seconda se si adotta un approccio bottom-up o top-down. Negli algoritmi agglomerativi tutti gli elementi sono inizialmente considerati cluster a sé, e man mano che si procede con l'algoritmo, avviene l'unione dei cluster su diversi livelli, fino ad ottenere un unico gruppo contenente tutte le unità di partenza. Gli algoritmi divisivi, viceversa, suddividono gli elementi partendo da un unico cluster, ottenendo gruppi sempre più piccoli finché il numero dei cluster coincide con il numero delle unità. Tra i due approcci quello agglomerativo è più semplice da programmare e inoltre presenta un minore rischio di errore nell'allocazione degli elementi. I metodi non gerarchici, al contrario, producono un'unica suddivisione dell'insieme di partenza, considerata ottimale rispetto al criterio adottato. Questi metodi, in generale, partono inizialmente da una suddivisione provvisoria degli elementi ed effettuano una serie di riallocazioni finché non risulta soddisfatto un criterio di ottimo. I metodi non gerarchici si possono a loro volta distinguere in *metodi di suddivisione iterativa* che eseguono degli spostamenti effettivi degli elementi, assegnando ogni unità al gruppo più vicino e *metodi di programmazione matematica*, i quali si basano su spostamenti virtuali degli elementi secondo la soluzione di un problema di massimo o minimo vincolato. Il difetto dei metodi non gerarchici è che dipendono dalle scelte iniziali di partizione degli oggetti, dei nuclei e della funzione obiettivo, il che li rende poco affidabili a meno che non si tenti di superare quest'inconveniente. Un altro tipo di classificazione viene fatta fra algoritmi *esatti* ed *euristici*. I primi determinano una suddivisione ottimale degli elementi in gruppi, ossia viene ricercata la migliore soluzione tra le possibili partizioni. Gli algoritmi euristici, invece, forniscono una soluzione approssimativamente ottima, ma che si discosta dalla migliore possibile; tuttavia, ciò ha un vantaggio computazionale rispetto ai metodi esatti, i quali presentano una numerosità di operazioni elementari che cresce esponenzialmente con il numero di unità (dimensione dell'insieme di dati da analizzare).

Classificazione in base ai risultati prodotti

Oltre alle differenziazioni sul tipo di algoritmo, i metodi di clustering possono essere distinti in funzione ai risultati che essi forniscono. Definendo come funzione di appartenenza la relazione che associa ad ogni elemento il grado di appartenenza ad un gruppo, si possono distinguere due tipologie di metodi: i *metodi esclusivi* (o *classici*) e i *metodi non-esclusivi* (o *fuzzy*). I primi, chiamati anche *hard clustering*, sono caratterizzati da una funzione di appartenenza che può assumere solo i valori {0,1}, ovvero gli elementi appartengono o non appartengono ad un determinato gruppo. Al contrario, i secondi, chiamati anche *soft clustering*, presentano un intervallo di definizione della funzione tra 0 e 1, per cui ogni elemento può appartenere a ciascun cluster con un certo grado, come se fosse la probabilità di appartenenza. Poiché gli elementi non sono sempre associati con esattezza ad un gruppo, esistono i casi in cui un elemento possa essere assegnato in modo indifferente a più cluster. I metodi fuzzy non solo riescono ad associare gli elementi in diversi gruppi senza nessuna "forzatura", ma presentano anche un buon indice di non apparenza ad uno specifico gruppo. Questi metodi, dunque, non hanno la pretesa di fornire dei risultati esatti su come si aggregano i dati, ma, al contrario mettono in evidenza l'imprecisione intrinseca nei dati. Bisogna però prestare attenzione a non confondere i metodi fuzzy con le classificazioni che si ottengono a partire dai dati "sfocati" ma che presentano comunque una funzione di appartenenza a valori {0,1} e pertanto appartengono alla categoria dei metodi classici. Inoltre, per entrambe le categorie, se un elemento è associato a più cluster e la loro unione fornisce un valore superiore all'unità, si parla di *metodi sovrapposti*. Questi metodi vengono utilizzati quando sono presenti elementi con caratteristiche intermedie a due o più gruppi e quindi sarebbe meglio assegnarli ad entrambi; i gruppi che vengono creati sono chiamati solitamente *clump* e vengono analizzati con tecniche di *clumping*.

1.2 Tecniche di clustering

Le tecniche di clustering si basano su misure di somiglianza tra le diverse unità appartenenti ad un insieme di dati di partenza. Nei più comuni approcci, la similarità è vista come una distanza tra i vari elementi presenti in uno spazio multidimensionale [3,4]. Perciò, gli algoritmi basati su di esse, sono molto sensibili alla scelta della metrica con cui viene calcolata la distanza, influenzando in tal modo la bontà dell'analisi. La definizione di un'appropriata metrica di distanza dipende dalla natura delle caratteristiche che costituiscono il set di dati, i quali possono essere numerici, categorici, binari o misti. Riferendosi a un insieme di dati numerici, rappresentati tramite una matrice $\mathbf{X}$ di dimensione $M \times N$ con M il numero di elementi ed N la loro dimensione e considerando due generici dati $\mathbf{x}_r = (x_{r1}, x_{r2}, \ldots, x_{rN})$ e $\mathbf{x}_s = (x_{s1}, x_{s2}, \ldots, x_{sN})$, si possono distinguere alcune metriche [5]:

Distanza Euclidea: $$d_{rs} = \mathrm{dist}(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{\sum_{j=1}^{N} \left(x_{rj} - x_{sj}\right)^2} = \|\mathbf{x}_r - \mathbf{x}_s\|_2 \qquad (1.2.1)$$

Distanza cityblock (o *di Manhattan*): $$d_{rs} = \mathrm{dist}(\mathbf{x}_r, \mathbf{x}_s) = \sum_{j=1}^{N} \left|x_{rj} - x_{sj}\right| \qquad (1.2.2)$$

Distanza di Minkowski: $$d_{\mathrm{rs}} = \mathrm{dist}(\mathbf{x}_{\mathrm{r}}, \mathbf{x}_{\mathrm{s}}) = \sqrt[q]{\sum_{j=1}^{N} \left| x_{\mathrm{rj}} - x_{\mathrm{sj}} \right|^{q}} \tag{1.2.3}$$

Distanza di Mahalanobis: $$d_{\mathrm{rs}} = \mathrm{dist}(\mathbf{x}_{\mathrm{r}}, \mathbf{x}_{\mathrm{s}}) = \sqrt{(\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{s}})\mathbf{C^{-1}}(\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{s}})'} \tag{1.2.4}$$

dove $\mathbf{C^{-1}}$ è l'inverso della matrice di covarianza della coppia di vettori.

Queste espressioni però risentono di eventuali valori dominanti rispetto ad altri, e ciò potrebbe condizionare la formazione dei raggruppamenti da parte degli algoritmi di clustering. Per evitare questo sbilanciamento nella valutazione delle distanze, è indispensabile effettuare una *normalizzazione*[1] dei dati iniziali, tale da renderli comparabili in un intervallo di valori tra 0 e 1. È inoltre possibile rappresentare le distanze d_{rs} tra le varie coppie di *M* oggetti appartenenti ad un insieme **X**, con una matrice simmetrica **D** di dimensioni *M*×*M* :

$$\mathbf{D} = \begin{bmatrix} 0 & \mathrm{d}_{12} & \dots & \mathrm{d}_{1M} \\ \mathrm{d}_{21} & 0 & \dots & \mathrm{d}_{2M} \\ \vdots & \vdots & \ddots & \dots \\ \mathrm{d}_{\mathrm{M1}} & \dots & \dots & 0 \end{bmatrix} \tag{1.2.5}$$

Al fine di rendere più chiara la trattazione delle diverse tecniche di clustering, si è scelto come modello esemplificativo un'applicazione nell'ambito dei sistemi elettrici; in particolare si è fatto riferimento all'analisi dei dati provenienti da diverse curve di carico di *M utenti* caratterizzate da *N dimensioni* corrispondenti alla suddivisione temporale del periodo di osservazione. In questo caso, poiché si tratta di serie temporali, la normalizzazione viene eseguita per ogni *m*-esimo utente, rispetto al valore massimo della sua curva di carico, ottenendo in tal modo un insieme **X** di curve confrontabili nell'intervallo tra 0 e 1. Uno scopo del clustering potrebbe essere quello di andare a caratterizzare diverse tipologie di utenze, attraverso la creazione di gruppi con proprietà simili. In *Figura 1.2.1* sono mostrate le curve di carico normalizzate per ogni singolo utente, mettendo in evidenza il loro comportamento medio, calcolato ogni quarto d'ora.

[1] In generale, la *normalizzazione* (*min-max normalization*) viene eseguita per ogni *n*-esima caratteristica degli *M* oggetti appartenenti all'insieme di dati **X**, valutando il valore massimo e minimo che essa raggiunge tra i vari oggetti, secondo la seguente equazione:

$$x_{m,n}^{norm} = \frac{x_{m,n} - x_n^{min_m}}{x_n^{max_m} - x_n^{min_m}} \qquad \text{per } m = 1,..,M \quad e \quad \text{per } n = 1,..,N$$

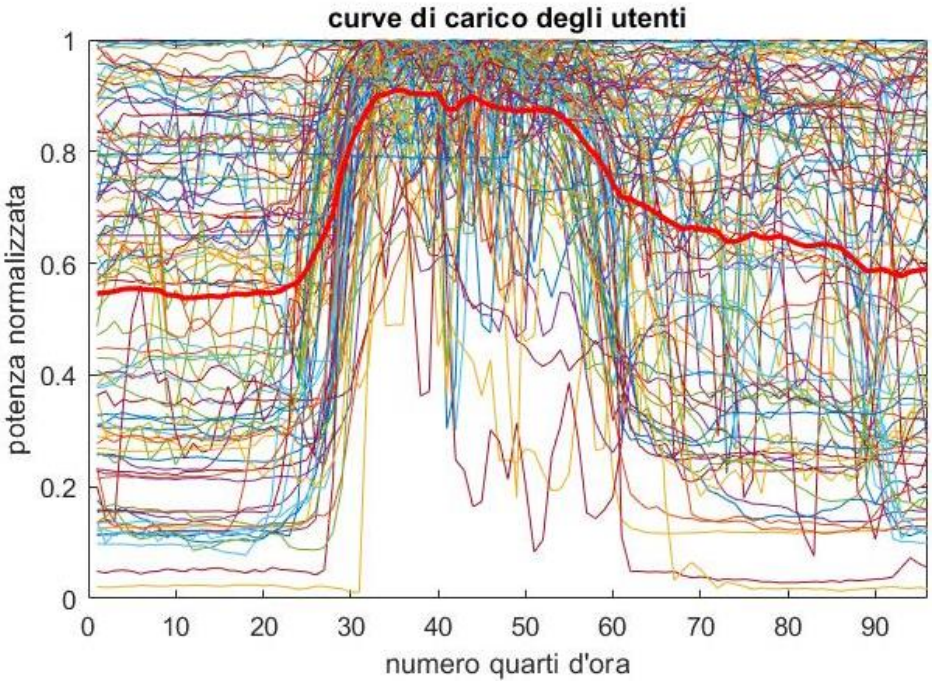


Figura 1.2.1- Curve di carico normalizzate di 100 utenti

Hierarchical Clustering

Le tecniche di hierarchical clustering sono algoritmi gerarchici basati sulla distanza, i quali trasformano la matrice delle distanze in una partizione che può essere rappresentata da un grafico ad albero chiamato dendrogramma, in cui l'ordinata rappresenta la distanza alla quale si uniscono i cluster, mentre nell'ascisse vengono posizionati gli oggetti. Tagliando il dendrogramma a diversi livelli, si ottengono diversi numeri di cluster e le loro partizioni. Nell'esempio considerato si è deciso di raggruppare gli elementi in 6 cluster, ottenendo così una classificazione delle differenti curve di carico di 100 utenti, rappresentate in *Figura 1.2.2*.

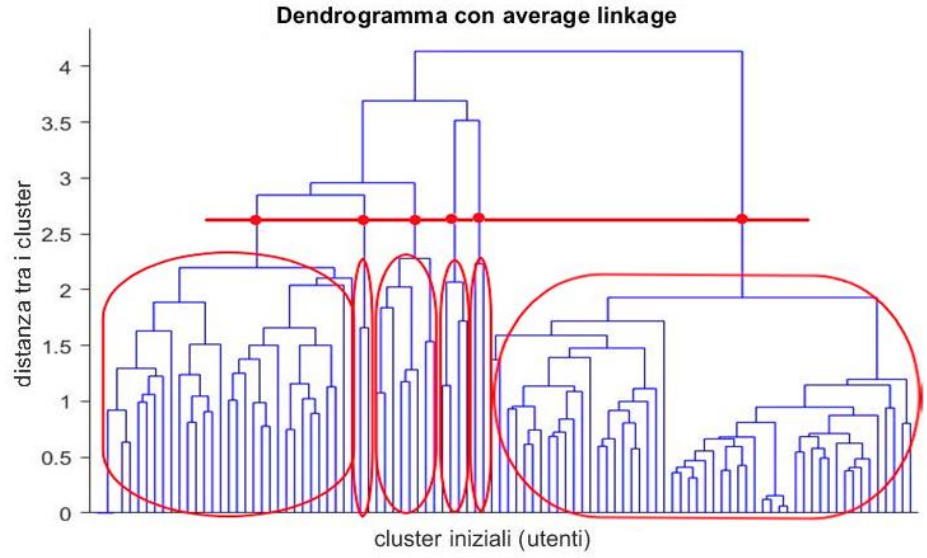


Figura 1.2 2 - Hierarchical clustering con average linkage, dendrogramma e risultati della suddivisione in sei cluster - Parte 1

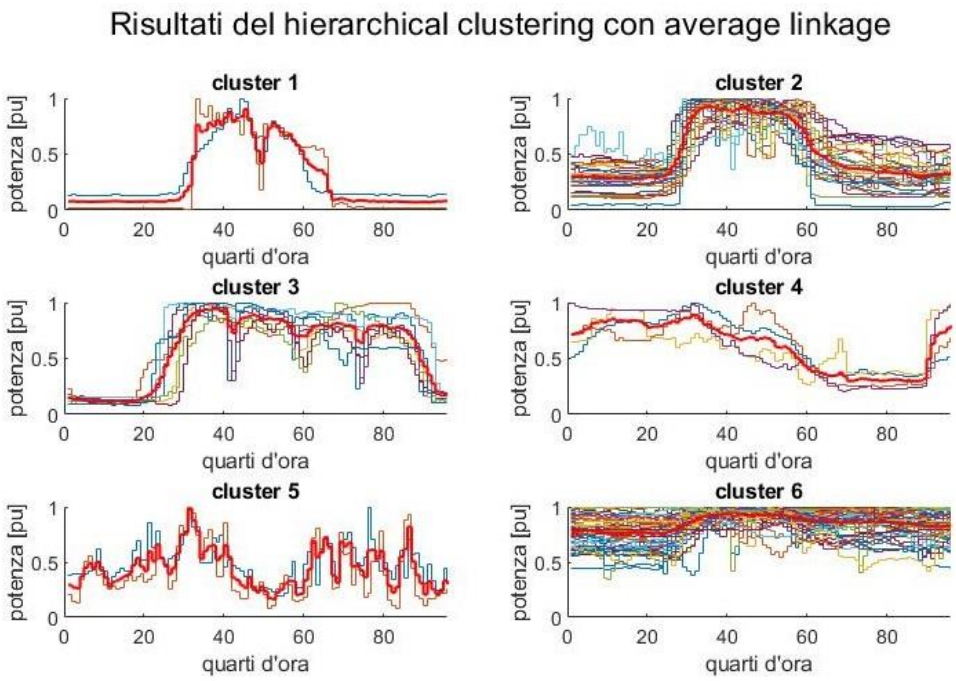


Figura 1.2.2 - Hierarchical clustering con average linkage, dendrogramma e risultati della suddivisione in sei cluster - Parte 2

Osservando la tipologia agglomerativa, l'algoritmo inizia considerando tutte le *M* unità come singoli cluster, poi gradualmente si ha la loro unione, fino a quando si ottiene un unico cluster, generando così il dendrogramma.

Step 1 → Inizia con *M* cluster, dove ogni oggetto ne rappresenta uno; si calcola la matrice delle distanze **D** per gli *M* cluster;

Step 2 → Ricerca la minima distanza nella matrice;

Step 3 → Aggrega i due cluster più vicini e aggiorna la matrice con le nuove distanze tra i nuovi cluster e gli altri;

Step 4 → Ripete lo step 2 e 3 finché gli oggetti si racchiudono in un unico cluster.

Per lo step 3 sono presenti una varietà di scelte su come avviene l'unione dei cluster, e quindi sulla definizione di distanza tra i due. Tra le più comuni si possono trovare:

- *Single linkage*: la distanza tra due gruppi è rappresentata considerando una singola coppia di elementi dei due cluster, più vicina. Considerando D_{pq} la distanza tra due cluster $\mathbf{X}_p$ e $\mathbf{X}_q$, e $d_{rs} = \text{dist}(\mathbf{x}_r, \mathbf{x}_s)$ la distanza tra due elementi appartenenti rispettivamente ai due gruppi, si ha:

$$D_{pq} = d(\mathbf{X}_p, \mathbf{X}_q) = \min_{\mathbf{x}_r \in \mathbf{X}_p} \min_{\mathbf{x}_s \in \mathbf{X}_q} d_{rs} \tag{1.2.6}$$

  Questo è un criterio estremo che porta in genere alla formazione di un grande cluster, insieme a molti cluster più piccoli.

- *Complete linkage*: calcola la distanza di due cluster come la distanza tra gli elementi più lontani all'interno dei due gruppi:

$$D_{\mathrm{pq}} = \mathrm{d}(\mathbf{X}_{\mathrm{p}}, \mathbf{X}_{\mathrm{q}}) = \max_{\mathbf{x}_{\mathrm{r}} \in \mathbf{X}_{\mathrm{p}}} \max_{\mathbf{x}_{\mathrm{s}} \in \mathbf{X}_{\mathrm{q}}} d_{\mathrm{rs}} \tag{1.2.7}$$

- *Average linkage*: la distanza viene calcolata come media aritmetica delle distanze tra i singoli elementi. Questo criterio tende a formare grandi cluster di elementi simili e piccoli gruppi per gli elementi molto diversi. Considerando M_p e M_q il numero di elementi presenti all'interno dei due cluster, la distanza tra essi viene calcolata come:

$$D_{\mathrm{pq}} = \mathrm{d}(\mathbf{X}_{\mathrm{p}}, \mathbf{X}_{\mathrm{q}}) = \frac{1}{M_{\mathrm{p}} M_{\mathrm{q}}} \sum_{\mathbf{x}_{\mathrm{r}} \in \mathbf{X}_{\mathrm{p}}}^{M_{\mathrm{p}}} \sum_{\mathbf{x}_{\mathrm{s}} \in \mathbf{X}_{\mathrm{q}}}^{M_{\mathrm{q}}} d_{\mathrm{rs}} \tag{1.2.8}$$

- *Centroid linkage*: usa la distanza Euclidea tra i centroidi dei cluster come misura della distanza tra essi; definendo il centroide $\bar{\mathbf{c}}_{\mathrm{p}}$ del cluster $\mathbf{X}_{\mathrm{p}}$ come il punto medio tra gli elementi del cluster:

$$\bar{\mathbf{c}}_{\mathrm{p}} = \frac{1}{M_{\mathrm{p}}} \sum_{\mathbf{x}_{\mathrm{r}} \in \mathbf{X}_{\mathrm{p}}}^{M_{\mathrm{p}}} \mathbf{x}_{\mathrm{r}} \tag{1.2.9}$$

  la distanza tra i cluster viene calcolata come:

$$D_{\mathrm{pq}} = \mathrm{d}(\mathbf{X}_{\mathrm{p}}, \mathbf{X}_{\mathrm{q}}) = \left\| \bar{\mathbf{c}}_{\mathrm{p}} - \bar{\mathbf{c}}_{\mathrm{q}} \right\|_2 \tag{1.2.10}$$

- *Ward's linkage*: questo criterio, chiamato anche di minima varianza, si basa sull'incremento della somma dei quadrati all'interno di un cluster. I gruppi sono formati in modo da minimizzare questo incremento dovuto alla loro ipotetica unione:

$$D_{\mathrm{pq}} = \mathrm{d}(\mathbf{X}_{\mathrm{p}}, \mathbf{X}_{\mathrm{q}}) = \sqrt{\frac{2 M_{\mathrm{p}} M_{\mathrm{q}}}{M_{\mathrm{p}} + M_q} \left\| \bar{\mathbf{c}}_{\mathrm{p}} - \bar{\mathbf{c}}_{\mathrm{q}} \right\|_2} \tag{1.2.11}$$

  dove $\|\cdot\|_2$ rappresenta la distanza Euclidea tra i centroidi dei cluster $\mathbf{X}_{\mathrm{p}}$ e $\mathbf{X}_{\mathrm{q}}$; il fattore 2 posto davanti al prodotto $M_{\mathrm{p}} M_{\mathrm{q}}$, come alcuni autori fanno, potrebbe essere omesso.
  Questo criterio fornisce, in genere, cluster più piccoli ma di grandezza molto simile.

In *Figura 1.2.3* sono presenti degli esempi di hierarchical clustering ottenuti con linkage differenti; in particolare, si può notare come il metodo che utilizza il single linkage è soggetto ad un fenomeno di concatenamento (*chaining*), ovvero i cluster più grandi sono inclini ad attirare a sé altri cluster più vicini, raggruppando così sempre un numero maggiore. Dunque, questo metodo non è molto utile per riassumere i dati, tuttavia, riesce a identificare facilmente i dati anomali (*outlier*), i quali saranno gli ultimi ad essere raggruppati. Al contrario nel complete linkage e Ward's linkage non sono particolarmente evidenti fenomeni di chaining, però, mentre

nel complete linkage si ha comunque una differenziazione tra taglie dei cluster, nel Ward's linkage si hanno raggruppamenti di taglia più simile tra loro e meno sensibili ai dati anomali.

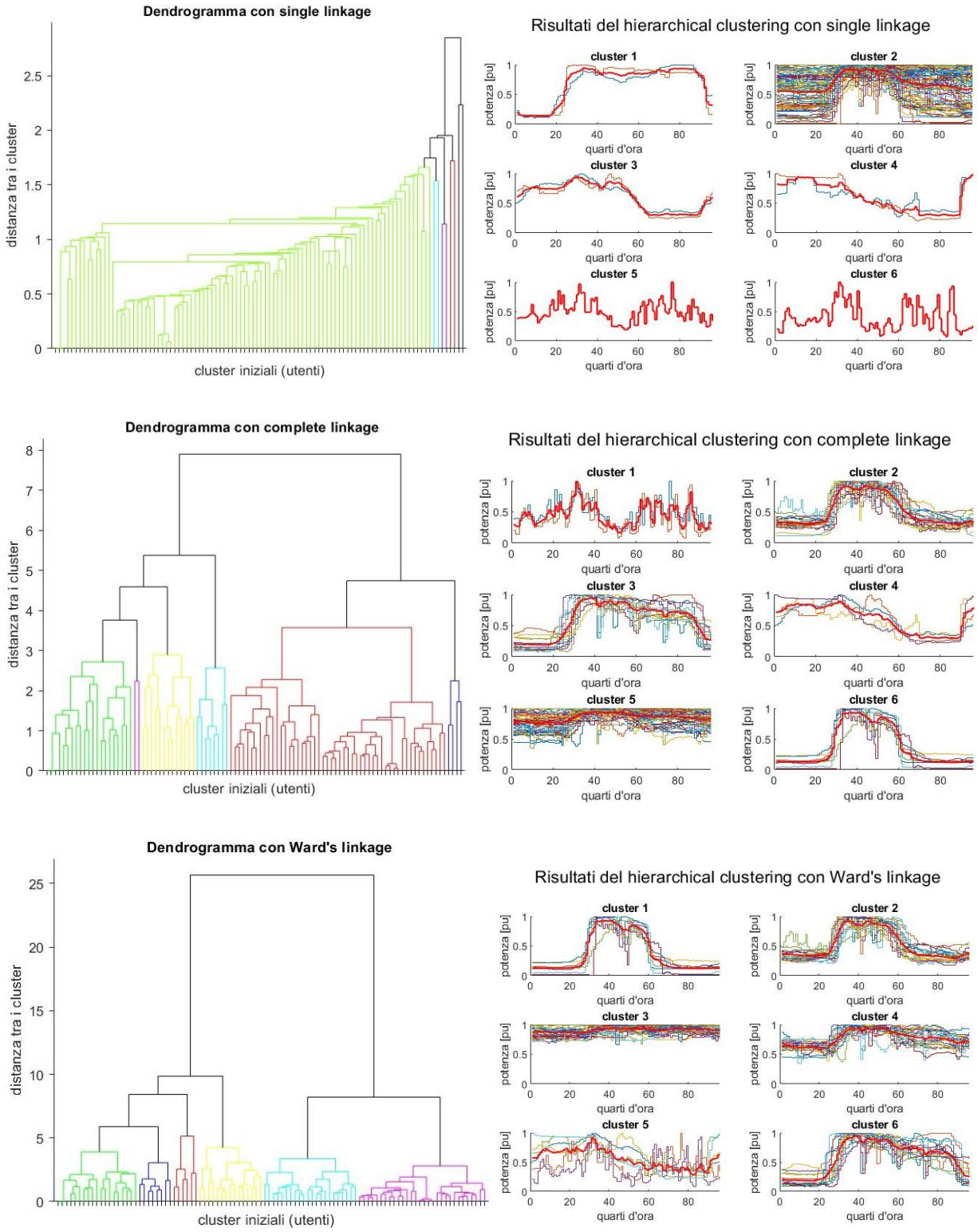


Figura 1.2.3 - Hierarchical clustering: dendrogramma e risultati utilizzando Single linkage, complete linkage e Ward's linkage

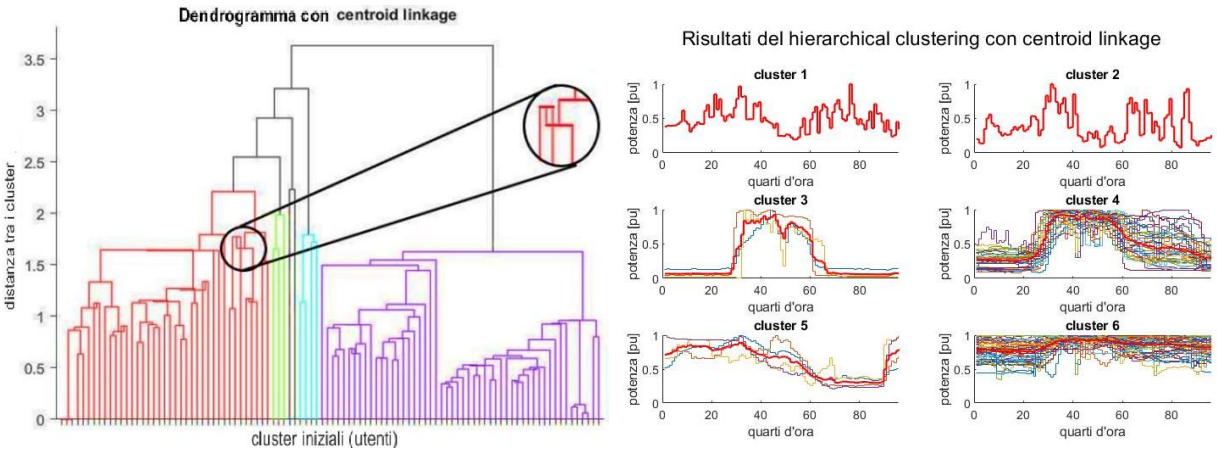


Figura 1.2.4 - Hierarchical clustering con centroid linkage

Un risultato singolare è visibile in *Figura 1.2.4*. Per il dataset usato nell'esempio, l'utilizzo dell'algoritmo gerarchico con centroid distance, risulta essere inappropriato, poiché la costruzione del dendrogramma avviene in maniera non monotona, ovvero, in alcuni passaggi si ottengono valori di distanza inferiori a quelli precedenti, contrapponendosi così al principio base dell'algoritmo. In *Figura 1.2.5* è riportato il caso dello hierarchical clustering con average linkage, in cui è stata adottata una definizione differente di distanza rispetto al caso di *Figura 1.2.2*. Nonostante si utilizzi lo stesso algoritmo, la tipologia di metrica influisce sui risultati, infatti si può notare che la distanza cityblock, in questo caso, ha consentito comunque la creazione di gruppi di piccola taglia, ma con un numero di elementi superiori rispetto alla distanza Euclidea.

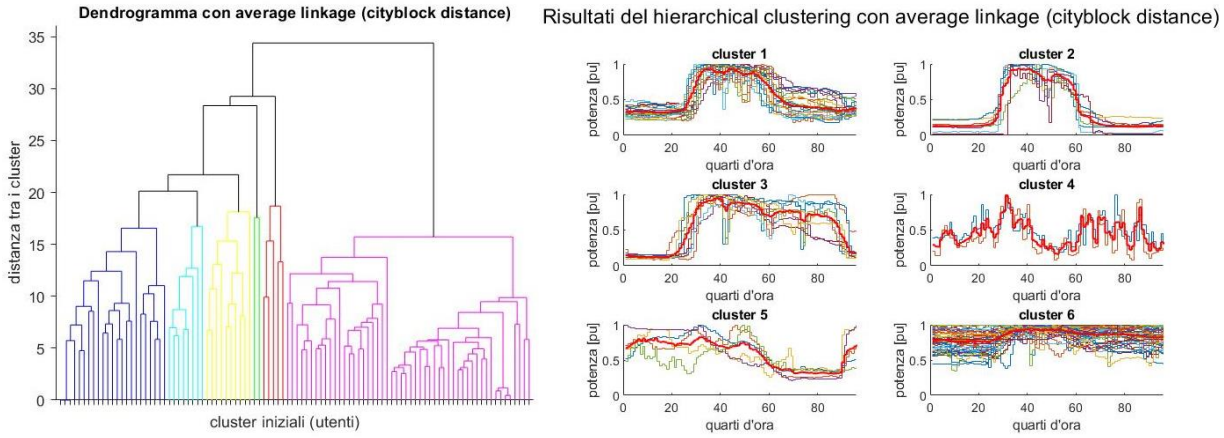


Figura 1.2.5 - Hierarchical clustering: dendrogramma e risultati utilizzando average linkage con distanza di cityblock

K-Means Clustering

Il k-means è la tecnica di clustering più utilizzata e semplice tra gli algoritmi partizionali, ed è basato sulla misura delle distanze tra centroidi, ossia tra i valori centrali dei gruppi. Queste tecniche raggruppano gli elementi molto simili tra loro, come si può notare in *Figura 1.2.6*, attraverso specifiche funzioni obiettivo e miglioramenti iterativi della qualità delle partizioni, richiedendo la scelta di un certo numero di parametri iniziali.

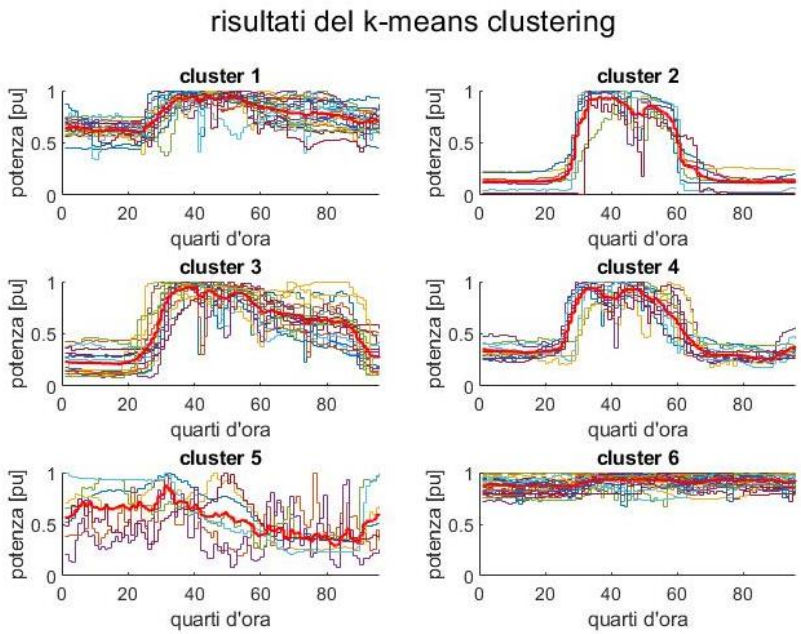


Figura 1.2.6 - Risultati del k-means clustering

L'algoritmo crea *K* partizioni dei dati e assegna, ad ognuna di esse, i valori iniziali dei centroidi, in maniera casuale o tramite alcune informazioni euristiche; l'algoritmo procede con la minimizzazione della varianza intra-gruppo totale (per tutti i *K* gruppi), in modo da trovare la migliore partizione degli oggetti nei cluster a loro più vicini:

$$\min_p \sum_{p=1}^{K} \sum_{\mathbf{x_r} \in \mathbf{X}_p} \| \mathbf{x}_\mathrm{r} - \bar{\mathbf{c}}_p \|_2^{\,2} \tag{1.2.12}$$

dove $\mathbf{x}_\mathrm{r}$ è un oggetto assegnato ad un cluster p con centroide $\bar{\mathbf{c}}_p$ ad un generico passaggio dell'algoritmo. Il processo viene iterato finché l'algoritmo non converge, ricalcolando i centroidi per i nuovi cluster come la media tra gli elementi contenuti:

$$\bar{\mathbf{c}}_\mathrm{p}{}' = \frac{1}{M'_\mathrm{p}} \sum_{\mathbf{x_r} \in \mathbf{X}'_\mathrm{p}}^{M'_\mathrm{p}} \mathbf{x}_\mathrm{r} \tag{1.2.13}$$

Step 1 → Inializzazione dei *K* punti come centroidi iniziali;

Step 2 → Assegna ogni elemento al centroide più vicino, secondo un calcolo di distanza;

Step 3 → Ricalcola e aggiorna i centroidi dei *K* cluster cambiati;

Step 4 → Ripete lo step 2 e 3 fino ad un criterio di arresto o fino a quando i cluster non variano.

Il k-means è un algoritmo che converge molto rapidamente; tuttavia, la scelta dei centroidi iniziali e la stima del numero di cluster *K* impattano la qualità delle soluzioni finali, non consentendo il raggiungimento di un ottimo globale. Perciò l'algoritmo è comunemente eseguito più volte con differenti inizializzazioni casuali, in modo da scegliere la soluzione più soddisfacente da quelle

prodotte. Esistono diverse varianti del k-means, alcune si basano su un'inizializzazione differente, come ad esempio il *k-means++* che sceglie in modo meno casuale i centri iniziali, altre sul processo di assegnazione dei singoli elementi ai rispettivi cluster, tra cui:

- *k-medians*:
  Questo metodo anziché utilizzare la media come nel k-means classico, usa la mediana per determinare il centroide di ciascun gruppo. Ciò, insieme all'utilizzo della distanza cityblock, comporta una riduzione dell'errore per ogni cluster rendendo il raggruppamento più compatto. La funzione obiettivo che viene utilizzata è la seguente:

$$\min_p \sum_{p=1}^{K} \sum_{\mathbf{x}_r \in \mathbf{X}_p}^{M_p} |\mathbf{x}_r - \bar{\bar{\mathbf{c}}}_p| \qquad \text{dove} \qquad \bar{\bar{\mathbf{c}}}_p = \boldsymbol{med}(\mathbf{X}_p) \tag{1.2.14}$$

  Il centroide di un generico cluster $\mathbf{X}_p$ viene rappresentato dal vettore $\bar{\bar{\mathbf{c}}}_p$, contenente, per ogni *j*-esima colonna (componente del vettore $\mathbf{x}_r \in \mathbf{X}_p$), la mediana della *j*-esima caratteristica tra tutti i dati presenti nel *p*-esimo cluster, ovvero un valore "reale" compreso nel *set* di dati; anche se ciò non implica che il vettore di mediana finale appartenga all'insieme di dati. Al contrario, l'utilizzo della media aritmetica oppure della mediana in combinazione con la distanza euclidea (norma quadratica), non produrrà necessariamente un vettore con caratteristiche appartenenti all'insieme di dati, e ciò renderebbe meno affidabile tale algoritmo, per insiemi di dati discreti o binari.

- *K-medoids o Partitioning Around Medoids* (*PAM*):
  Questi algoritmi, a differenza di quelli precedenti, scelgono come centro del cluster un punto appartenente al set di dati, chiamato medoide, e consentono l'utilizzo di un qualsiasi tipo di distanza. L'algoritmo è di tipo euristico, per cui partendo dai medoidi iniziali esegue una ricerca veloce non esaustiva, trovando delle soluzioni non proprio ottimali:

  Step 1 → Seleziona in modo casuale *K* punti, dall'insieme dei dati **X**, come medoidi iniziali;

  Step 2 → Assegna ogni elemento al medoide più simile secondo una funzione di costo, definita dal calcolo di distanze;

  Step 3 → Seleziona casualmente un elemento non medoide $\mathbf{x}_r$;

  Step 4 → Calcola la funzione di costo totale S_t come la somma dei costi dei singoli elementi rispetto ai propri medoidi e la confronta con una funzione di costo totale S_t' calcolata come se $\mathbf{x}_r$ fosse il nuovo medoide del cluster $\mathbf{X}_p$ in cui esso appartiene: $S = S_t' - S_t$;

  Step 5 → Se $S < 0$, $\mathbf{x}_r$ rappresenterà il nuovo medoide del cluster $\mathbf{X}_p$ e si aggiorna l'insieme dei medoidi;

  Step 6 → Ripete dallo step 2 fino ad un criterio di arresto o fino a quando l'insieme dei medoidi non subisce cambiamenti.

Generalmente, rispetto al k-means, sono metodi più robusti al rumore e ai dati anomali.

- *Fuzzy c-means* (*FCM*):
  Questa tecnica è una versione fuzzy del k-means, in cui ogni elemento può essere assegnato con un certo grado di appartenenza a più cluster. Fuzzy c-means è la denominazione classica con cui il metodo è stato presentato; tuttavia, in letteratura si trova anche con il nome di *fuzzy k-means*. Pertanto, i dati più anomali possono trovarsi nei gruppi in misura minore, quindi con gradi di appartenenza inferiori, rispetto ai dati centrali dei cluster. Il grado di appartenenza $\mathbf{w}_{\mathrm{r,p}}$ di un elemento $\mathbf{x}_{\mathrm{r}}$ ad un generico cluster $\mathbf{X}_{\mathrm{p}}$ rappresentato dal suo centroide $\tilde{\mathbf{c}}_{\mathrm{p}}$, viene calcolato come:

$$\mathbf{w}_{\mathrm{r,p}} = \frac{1}{\sum_{p=1}^{K} \left(\frac{\mathbf{x}_{\mathrm{r}} - \tilde{\mathbf{c}}_{\mathrm{p}}}{\mathbf{x}_{\mathrm{r}} - \tilde{\mathbf{c}}_{\mathrm{p}}} \right)^{\frac{2}{\beta-1}}} \tag{1.2.15}$$

  dove β è un parametro che determina il livello di *fuzziness* di un cluster; maggiore sarà il suo valore, più alto è il grado di "sfocatura" che assumerà il cluster finale. In assenza di sperimentazione è comunemente impostato a 2.

  L'algoritmo è molto simile al k-means, presentando il seguente problema di ottimizzazione, che viene iterato con successivi aggiornamenti di $\mathbf{w}_{\mathrm{r,p}}$ e $\tilde{\mathbf{c}}_{\mathrm{p}}$, fino a convergenza del metodo.

$$\min \sum_{p=1}^{K} \sum_{\mathbf{x}_{\mathrm{r}} \in \mathbf{X}_{\mathrm{p}}}^{M_p} \mathbf{w}_{\mathrm{r,p}}^{\beta} \left\| \mathbf{x}_{\mathrm{r}} - \tilde{\mathbf{c}}_{\mathrm{p}} \right\|_2^2 \qquad \text{dove} \quad \tilde{\mathbf{c}}_{\mathrm{p}} = \frac{\sum_{\mathbf{x}_{\mathrm{r}} \in \mathbf{X}_{\mathrm{p}}}^{M_p} \mathbf{w}_{\mathrm{r,p}}^{\beta} \mathbf{x}_{\mathrm{r}}}{\sum_{\mathbf{x}_{\mathrm{r}} \in \mathbf{X}_{\mathrm{p}}}^{M_p} \mathbf{w}_{\mathrm{r,p}}^{\beta}} \tag{1.2.16}$$

  Come il k-means, questo metodo è sensibile ai valori anomali e le soluzioni finali ottenute corrisponderanno ad un minimo locale della funzione obiettivo, essendo dipendenti dalla scelta iniziale dei pesi.

In *Figura 1.2.7* sono riportati i risultati delle varianti k-medoids e fuzzy c-means. In particolare, sono stati messi in risalto, per ogni cluster, i centroidi ai quali i singoli elementi si avvicinano; per il k-medoids, essi rappresentano esattamente delle curve di carico appartenenti al set di dati. Per il fuzzy c-means, poiché ogni elemento appartiene a più cluster secondo un determinato grado di appartenenza, si è deciso di rappresentare ogni elemento al cluster con il quale si ha il grado più alto.

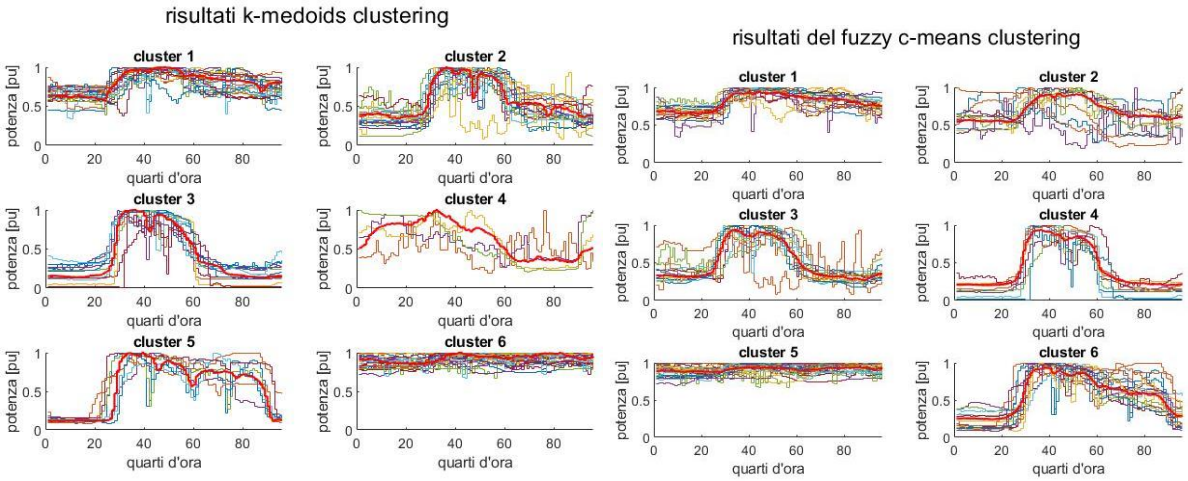


Figura 1.2.7 - Risultati del k-medoids e fuzzy c-means clustering

Le tecniche di clustering di tipo partizionale, dunque, hanno la possibilità di migliorare gradualmente la qualità dei risultati ottenuti, attraverso l'iterazione di un processo di ottimizzazione. Ciò non è possibile per gli algoritmi gerarchici, i quali, una volta ottenuto il dendrogramma, non permettono di modificare le ripartizioni già ottenute ai diversi livelli. Ulteriori vantaggi dei metodi di k-means sono la semplicità di implementazione e l'efficienza dal punto di vista computazionale rispetto ai metodi gerarchici. Tuttavia, le tecniche di hierarchical clustering presentano il vantaggio di fornire una rappresentazione grafica dei raggruppamenti, che potrebbe essere molto utile all'utente finale durante la valutazione dei risultati. Inoltre, gli algoritmi gerarchici sono di tipo deterministico, per cui la qualità delle soluzioni finali potrebbe essere migliore rispetto ai k-means che, al contrario, sono algoritmi euristici. Però per via dell'aumento della memoria necessaria per contenere i termini della matrice delle distanze (pur memorizzando soltanto i termini della parte triangolare superiore per motivi di simmetria della matrice) e del tempo di esecuzione, il clustering gerarchico è limitato solo a problemi di piccola scala; diverse soluzioni a queste problematiche sono presenti in letteratura, ottenute anche con una combinazione tra le due tecniche. Tuttavia, per la loro semplicità e facilità d'uso, le due tecniche di clustering proposte sono utilizzate in molte applicazioni, oltre ad essere la base di nuovi algoritmi di clustering, i quali sono ancora un'area attiva di ricerca.

Follow-The-Leader algorithm

L'algoritmo *follow-the-leader*, proposto negli articoli [6,7] in ambito dei sistemi elettrici, è un'ulteriore tecnica utilizzata per effettuare una ripartizione automatica dei dati, con il vantaggio di non richiedere a priori né il numero di cluster e né i rispettivi valori dei loro centroidi. Il metodo, segue una procedura deterministica, partendo dall'assegnazione di una soglia di distanza (*distance threshold* ρ) la quale ha il compito di stabilire, tramite un confronto con la distanza tra l'oggetto in esame e il centroide più vicino, se è necessario procedere alla creazione di un nuovo cluster o aggiornare quelli presenti. Dunque, il numero di gruppi che verranno creati dipenderà dal valore della soglia scelto, come si vede in *Figura 1.2.8*. In particolare, abbassando la soglia si otterranno sempre più gruppi differenti, migliorando l'accuratezza della ripartizione, fino ad arrivare al caso ideale in cui il numero di gruppi coincide con il numero di elementi analizzati, il quale, però, non ha alcun senso in un processo di raggruppamento. Al contrario elevati valori

della soglia, indicano che vengono accettati all'interno di un gruppo anche soggetti con minore similarità. Un giusto valore può essere determinato anche in funzione al numero di cluster desiderato; grazie alla rapidità di tale algoritmo, ciò è possibile dalla ripetizione della procedura, con diversi valori di soglia, fino ad ottenere il numero di cluster voluto, seguendo un approccio *trial-error*.

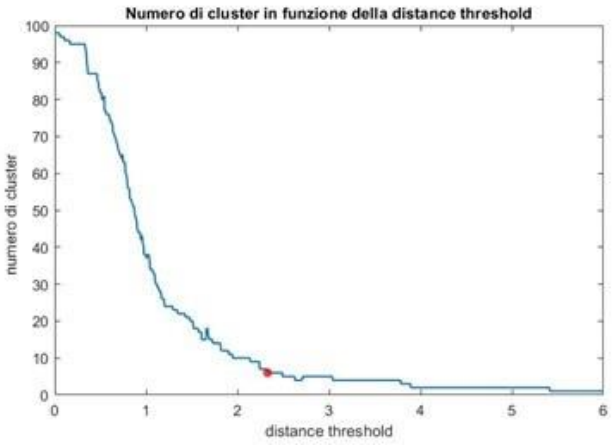


Figura 1.2.8 - Numero di gruppi in funzione della soglia, e valore di soglia per la formazione di 6 cluster

L'algoritmo è composto da diversi cicli, durante i quali vengono creati e riaggiornati i diversi cluster; l'arresto avviene quando non si hanno più cambiamenti del numero di elementi all'interno dei gruppi per un intero ciclo oppure quando si supera il numero massimo di iterazioni.

Step 1 → *Inizio primo ciclo*. Assegna al primo cluster il primo oggetto preso in maniera deterministica da una lista degli *M* oggetti dell'insieme **X**, considerandolo come centroide di esso;

Step 2 → Valuta, per l'*m*-esimo elemento successivo della lista, le distanze tra l'oggetto e i centroidi dei vari gruppi presenti, e trova il cluster più vicino;

Step 3 → Confronta la minima distanza con il valore di soglia ρ: se tale distanza supera la soglia, genera un nuovo cluster, associando l'oggetto come suo centro. Altrimenti assegna l'elemento al gruppo con distanza minima, e ricalcola il valore del centroide di tale cluster;

Step 4 → Ripete lo step 2 e lo step 3 per ogni oggetto appartenente all'insieme di partenza, concludendo il primo ciclo dell'algoritmo con tutti gli elementi associati ad un gruppo dei *K* creati;

Step 5 → *Inizio nuovo ciclo*. Per ogni oggetto della lista, in maniera consecutiva, rivaluta la minima distanza tra l'*m*-esimo elemento e i diversi centroidi dei *K* gruppi formati. Riassegna tale oggetto al cluster più vicino, aggiornando il centroide del gruppo di provenienza e del nuovo cluster formato. Ripete per ogni oggetto;

Step 6 → Ripete lo step 5 fino al raggiungimento del numero massimo di cicli o fino a quando non ci sono riassegnazioni degli oggetti all'interno dell'intero ciclo.

La versione base dell'algoritmo prevede l'utilizzo della distanza Euclidea, come misura della distanza tra un oggetto e il centro di un determinato cluster. Tuttavia, grazie al processo di normalizzazione dell'insieme dei dati di partenza, per ogni *n*-esima caratteristica, è possibile attribuire un rispettivo peso, il quale ha lo scopo di considerare la natura dispersiva dei dati, esaltando le caratteristiche che presentano una maggiore varianza tra i valori degli *M* oggetti appartenenti ad esso. In tal modo, viene introdotta una versione modificata dell'algoritmo, *Modified Follow-The-Leader*, la quale utilizza una distanza Euclidea ponderata per il calcolo

della distanza tra l'oggetto $\mathbf{x}_m = (x_{r1}, x_{r2}, \ldots, x_{rN})$ appartenente all'insieme $\mathbf{X}$ e il centroide $\bar{\mathbf{c}}_k^{(i)}$ del k-esimo cluster, corrispondente all'i-esimo ciclo dell'algoritmo:

$$\mathrm{dist}\left(\mathbf{x}_m, \bar{\mathbf{c}}_k^{(i)}\right) = \sqrt{\sum_{n=1}^{N} \frac{\sigma_n^2}{\bar{\sigma}^2}\left(x_{mn} - \bar{c}_{kn}^{(i)}\right)^2} \qquad (1.2.17)$$

avendo come pesi, per ogni $n = 1, \ldots, N$, il rapporto tra la varianza dell'n-esima caratteristica di tutti gli M oggetti dell'insieme σ_n^2, e la media di tali varianze per tutte le N caratteristiche $\bar{\sigma}^2$.
Si considerano, all'i-esimo ciclo, due gruppi $\mathbf{X}_p^{(i)}$ e $\mathbf{X}_q^{(i)}$, tra i K cluster presenti, con rispettivamente $N_p^{(i)}$ e $N_q^{(i)}$elementi. Ogni qualvolta che si ha una riassegnazione dell'elemento $\mathbf{x}_m$ da un gruppo p al gruppo q, l'algoritmo riaggiorna il valore dei centri $\bar{\mathbf{c}}_p^{(i)}$ e $\bar{\mathbf{c}}_q^{(i)}$ e il numero degli elementi appartenenti ai cluster, come segue:

$$\bar{\mathbf{c}}_q^{(i)^*} = \frac{N_q^{(i)}\bar{\mathbf{c}}_q^{(i)} + \mathbf{x}_m}{N_q^{(i)^*}} \qquad dove \quad N_q^{(i)^*} = N_q^{(i)} + 1 \qquad (1.2.18)$$

$$\bar{\mathbf{c}}_p^{(i)^*} = \frac{N_p^{(i)}\bar{\mathbf{c}}_p^{(i)} - \mathbf{x}_m}{N_p^{(i)^*}} \qquad dove \quad N_p^{(i)^*} = N_p^{(i-1)} - 1 \qquad (1.2.19)$$

In *Figura 1.2.9* sono riportati i risultati del processo di clustering eseguito per l'insieme dati preso di riferimento. Si può notare come, esso, riesce a ripartire in classi ben distinte gli elementi di partenza, risultando molto sensibile ai dati anomali. Tuttavia, questo tipo di algoritmo, essendo di tipo deterministico, potrebbe fornire risultati differenti a seconda dell'ordine in cui vengono presentati i dati di partenza (lista degli oggetti).

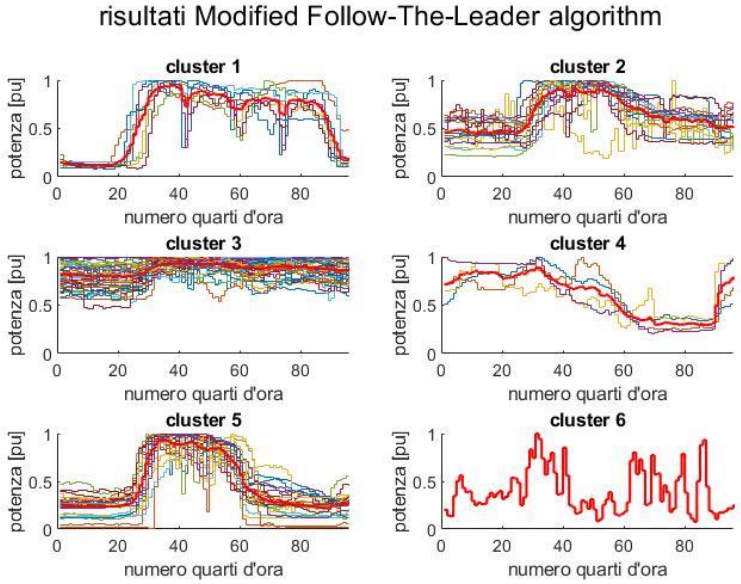


Figura 1.2.9 - Risultati del Modified Follow-The-Leader algorithm

1.3 Indici di validità dei risultati del clustering

Le diverse tecniche di clustering portano, in genere, a soluzioni differenti; dunque, risulta necessario adottare alcune metodologie per misurare la qualità dei risultati di clustering e confortare le prestazioni tra i diversi tipi di algoritmi. Gli indici di validità o *validity indices* sono comunemente usati in letteratura per selezionare il miglior metodo di clustering adatto all'applicazione e per determinare il numero ottimale di cluster prodotti dall'insieme di dati studiato [3,4]. Questi indicatori si suddividono in due categorie: gli *indici interni,* che sono calcolati solamente dai risultati forniti dal clustering, e gli *indici esterni*, che confrontano i cluster generati con informazioni esterne all'algoritmo. Una comune tecnica per valutare quanto bene un oggetto $\mathbf{x}_r$ sia assegnato al proprio cluster $\mathbf{X}_p$, rispetto agli oggetti negli altri cluster, è quella di calcolare il *coefficiente di Silhouette*:

$$s(\mathbf{x}_r) = \frac{b(\mathbf{x}_r) - a(\mathbf{x}_r)}{\max\{b(\mathbf{x}_r), a(\mathbf{x}_r)\}} \tag{1.3.1}$$

$$\text{dove} \qquad a(\mathbf{x}_r) = \frac{1}{M_p - 1} \sum_{\mathbf{x}_s \in \mathbf{X}_p,\ \mathbf{x}_s \neq \mathbf{x}_r}^{M_p} d_{rs} \ ; \qquad b(\mathbf{x}_r) = \min_{q, q \neq p} \frac{1}{M_q} \sum_{\mathbf{x}_s \in \mathbf{X}_q}^{M_q} d_{rs} \tag{1.3.2}$$

in cui $a(\mathbf{x}_r)$ rappresenta la media delle distanze tra l'elemento di un gruppo e tutti gli altri che gli appartengono, mentre $b(\mathbf{x}_r)$ è la minima media delle distanze tra l'oggetto e tutti gli elementi appartenenti ad ogni altro gruppo; d_{rs} è una qualunque metrica di distanza.

Il valore del coefficiente è definito tra -1 e 1, e i valori alti indicano che l'oggetto in questione è ben assegnato al proprio cluster e scarsamente abbinato agli altri. Se la maggior parte degli elementi ha un valore di silhouette elevato, i risultati del metodo di clustering sono affidabili, contrariamente la presenza consistente di valori bassi o negativi indica che la soluzione trovata è inappropriata e la causa potrebbe essere un numero di cluster inadatto. Infatti, oltre ad essere un indice di validità, il coefficiente di Silhouette è un buon metodo per la determinazione del numero di cluster ottimale da utilizzare nei metodi di clustering.

In *Figura 1.3.1* è riportato il risultato grafico del coefficiente di Silhouette calcolato sui metodi hierarchical clustering. I risultati mostrano, tutto sommato, dei buoni valori del coefficiente; un caso particolare presenta, però, l'algoritmo con single linkage, in quanto il cluster 2 ha un significativo numero di elementi che non sono ben rappresentati in quel gruppo, e questo è dovuto alla dimensione del cluster per via dell'effetto di chaining. In *Figura 1.3.2* sono invece presenti i valori del coefficiente di Silhouette riferiti ai metodi partizionali, i quali, in questo esempio, presentano dei raggruppamenti più "coerenti" rispetto a quelli gerarchici, poiché sia il numero, che il valore di elementi negativi, sono inferiori. Un caso a parte è il cluster 2 nel fuzzy c-means, il quale, presenta più elementi con un coefficiente negativo che positivo, e ciò è riscontrabile anche in *Figura 1.2.7*, dove le curve sono molto differenti tra loro. Buoni risultati sono ottenuti anche con l'algoritmo Follow-the-leader in *Figura 1.3.3*.

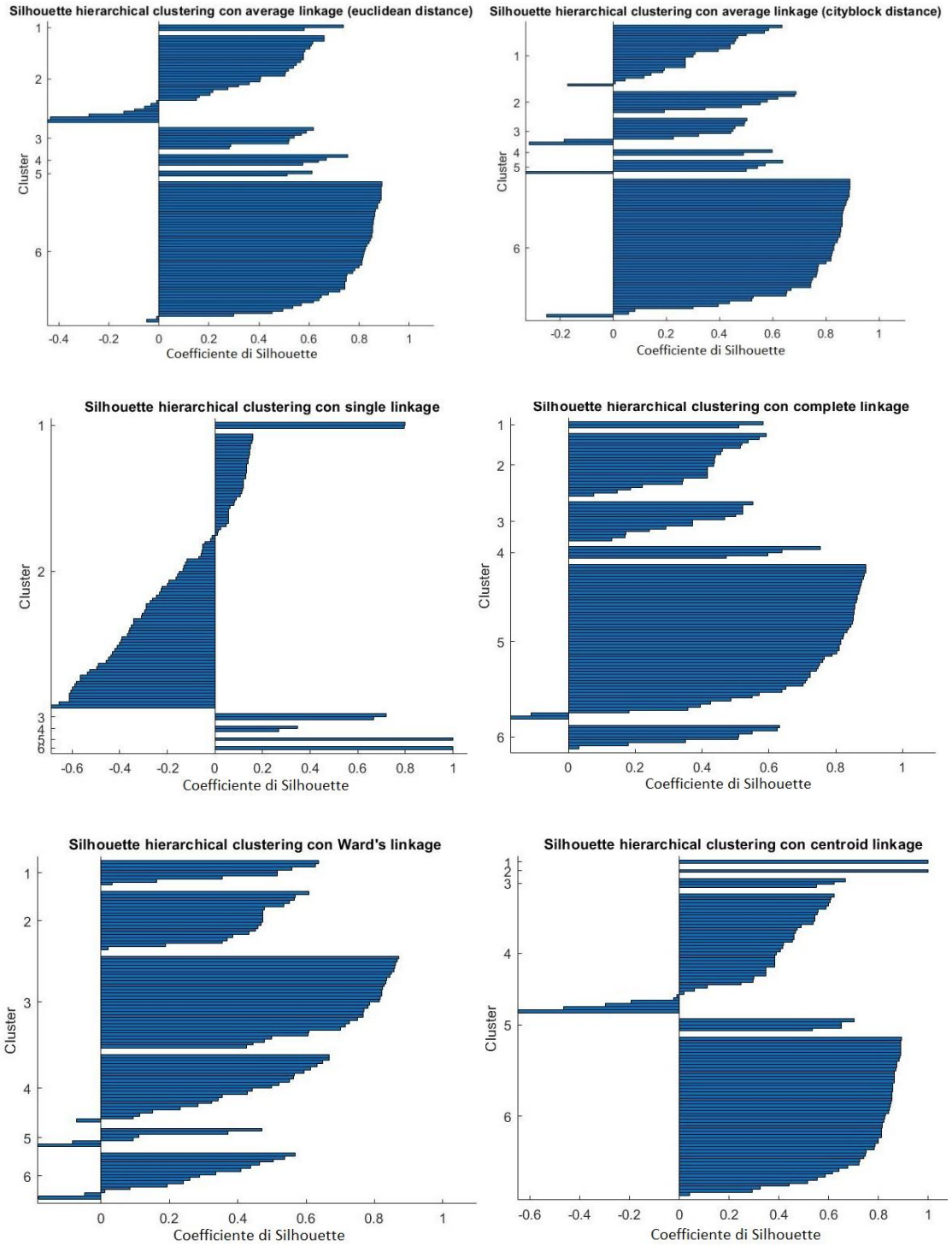


Figura 1.3.1 - Coefficiente di Silhouette su metodi di hierarchical clustering

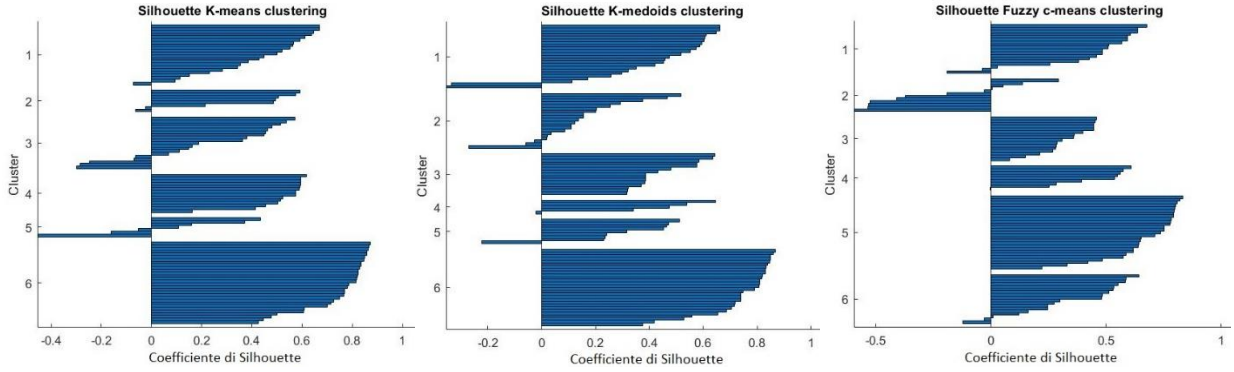


Figura 1.3. 2 - Coefficiente di Silhouette su metodi di k-means clustering e varianti

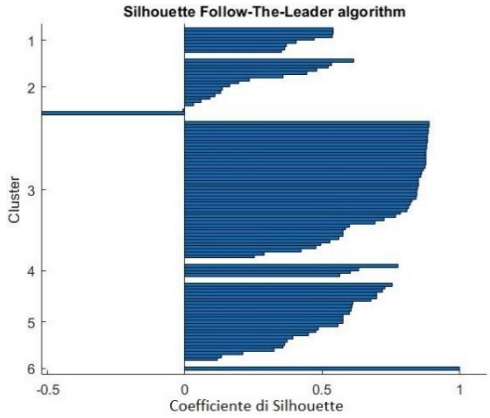


Figura 1.3.3 - Coefficiente di Silhouette sul metodo follow-the-leader

Poiché gli indici di validità che vengono utilizzati maggiormente dipendono dalle specifiche applicazioni, in seguito si fa riferimento ad ulteriori indici interni utili per problemi di clustering i cui dati provengono da serie temporali. Per la loro definizione è necessario fornire alcune specializzazioni di distanze basate sulla tipologia Euclidea:

Distanza tra vettori: $$\mathrm{d}(\mathbf{y},\mathbf{x}) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - x_j)^2} \qquad (1.3.3)$$

dove **x** e **y** sono due vettori di dimensione *N*.

Distanza vettore-set: $$\mathrm{d}(\mathbf{y},\mathbf{X}) = \sqrt{\frac{1}{M}\sum_{\mathbf{x}_\mathrm{r}\epsilon\mathbf{X}}^{M} d^2(\mathbf{y},\mathbf{x}_\mathrm{r})} \qquad (1.3.4)$$

dove **y** è il vettore del quale si calcola la distanza rispetto ad un insieme **X** di *M* elementi e $\mathbf{x}_\mathrm{r}$ è un generico elemento interno all'insieme.

Distanza media tra set: $$\mathrm{d}(\mathbf{X},\mathbf{Y}) = \frac{1}{M_\mathrm{X}M_\mathrm{Y}}\sum_{\mathbf{x}_\mathrm{r}\in\mathbf{X}}^{M_\mathrm{X}}\sum_{\mathbf{y}_\mathrm{s}\in\mathbf{Y}}^{M_\mathrm{Y}} d(\mathbf{y}_\mathrm{s},\mathbf{x}_\mathrm{r}) \qquad (1.3.5)$$

dove **X** e **Y** sono due insiemi di M_X e M_Y elementi.

Distanza intraset: $$\hat{\mathrm{d}}(\mathbf{X}) = \sqrt{\frac{1}{2M}\sum_{\mathbf{x_s}\in\mathbf{X}}^{M} d^2(\mathbf{x_s}, \mathbf{X})} \tag{1.3.6}$$

dove $\mathbf{X}$ è un insieme di M elementi e $\mathbf{x}_s$ è un generico elemento all'interno.

Si possono ora definire i seguenti indici di validità, considerando $\mathbf{X}_p$ (o $\mathbf{X}_q$) un generico cluster tra i K presenti, $\mathbf{C}$ la matrice rappresentativa dei loro centroidi $\mathbf{c}_p$ (o $\mathbf{c}_q$) e $\mathbf{x}_r$ (o $\mathbf{x}_s$) un generico elemento all'interno dell'insieme $\mathbf{X}$ composto da M elementi:

- *Clustering Dispersion Index* (CDI):

$$\mathrm{CDI} = \frac{1}{\hat{d}(\mathbf{C})}\sqrt{\frac{1}{K}\sum_{p=1}^{K}\hat{\mathrm{d}}^2\left(\mathbf{X}_p\right)} \tag{1.3.7}$$

- *Modified Dunn Index* (MDI):

$$\mathrm{MDI} = \frac{\max\limits_{1\le k\le K}\{\hat{\mathrm{d}}(\mathbf{X}_k)\}}{\min\limits_{p\neq q}\{\mathrm{d}(\mathbf{c}_p,\mathbf{c}_q)\}} \tag{1.3.8}$$

- *Davies-Bouldin Index* (DBI):

$$\mathrm{DBI} = \frac{1}{K}\sum_{k=1}^{K}\max_{p\neq q}\left\{\frac{\hat{\mathrm{d}}(\mathbf{X}_p)+\hat{\mathrm{d}}(\mathbf{X}_q)}{\mathrm{d}(\mathbf{c}_p,\mathbf{c}_q)}\right\} \tag{1.3.9}$$

- *Scatter index* (SI):

$$\mathrm{SI} = \frac{\sum_{r=1}^{M}\mathrm{d}^2(\mathbf{x}_r,\mathbf{p})}{\sum_{q=1}^{K}\mathrm{d}^2(\mathbf{c}_q,\mathbf{p})} \tag{1.3.10}$$

dove $\mathbf{p}$ è chiamato *pooled scatter* ed è calcolato come:

$$\mathbf{p} = \frac{1}{M}\sum_{s=1}^{M}\mathbf{x}_s$$

- *Mean Index Adequacy* (MIA):

$$\mathrm{MIA} = \sqrt{\frac{1}{K}\sum_{p=1}^{K}\mathrm{d}^2\left(\mathbf{c}_p, \mathbf{X}_p\right)} \tag{1.3.11}$$

In *Figura 1.3.4* sono riportati gli indici di validità calcolati sulla base dei risultati ottenuti da alcuni dei metodi di clustering visti precedentemente. Questi indici sono stati definiti in modo che valori inferiori definiscono una migliore distribuzione degli elementi all'interno dei cluster.

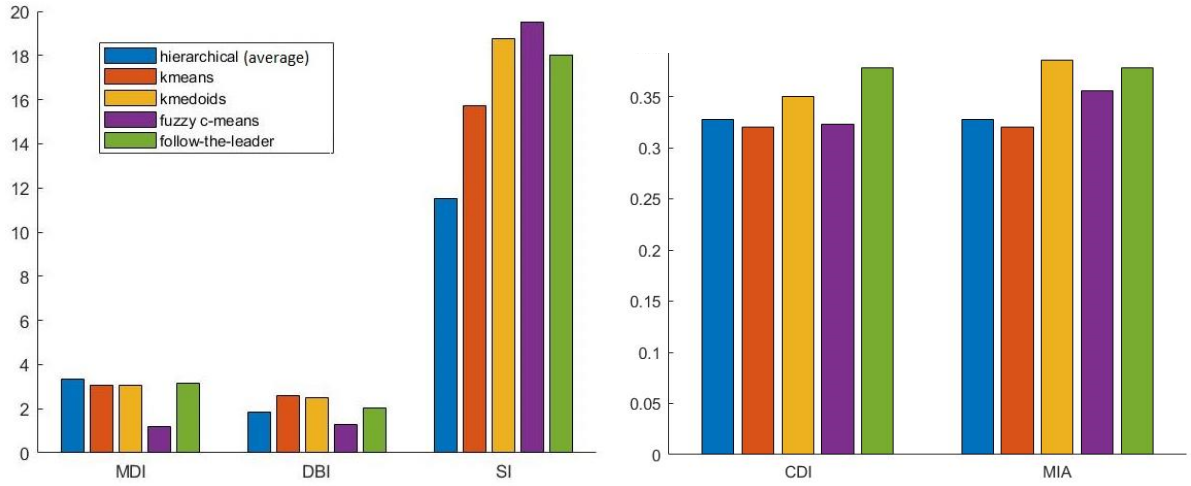


Figura 1.3.4 - Indici di validità dei risultati del clustering dei differenti metodi

L'utilizzo di questi indici, per il confronto delle diverse tecniche di clustering, deve essere effettuato sulla stessa base di dati e con un numero di cluster finali uguale per tutti. È possibile valutare inoltre, nel caso in cui non è stabilito a priori, il numero di cluster che fornisce una migliore partizione, osservando l'andamento di tali indici al variare del numero di cluster ottenibili e scegliendo quello ottimale in corrispondenza del cambio di pendenza di tali andamenti decrescenti (migliori risultati all'aumentare dei gruppi generati). Ciò permette inoltre di confrontare le diverse tecniche di clustering e scegliere quella che presenta risultati soddisfacienti per il dataset considerato. Ulteriori indici, utili nella valutazione del numero di cluster ottimale, sono ottenuti delle estensioni del coefficiente di Silhouette riferite all'intero insieme di dati **X** composto da *M* elementi raggruppati in *K* cluster:

- *Average Silhouette Coefficient* ($\mathrm{Avg_{sc}}$):

$$\mathrm{Avg_{sc}} = \frac{1}{M}\sum_{\mathbf{x}_r \in \mathbf{X}}^{M} \mathrm{s}(\mathbf{x}_r) \qquad (1.3.12)$$

- *Global Silhouette* (GS):

$$\mathrm{GS} = \frac{1}{K}\sum_{\mathrm{p}=1}^{K} \mathrm{s}_{\mathbf{X}_p} \qquad (1.3.13)$$

  dove $\mathrm{s}_{\mathbf{X}_p}$ rappresenta il coefficiente di Silhouette medio all'interno del generico cluster $\mathbf{X}_p$.

Nella *Figura 1.3.5* sono riportati i valori del GS al variare del numero di cluster per diverse tecniche di clustering applicate sull'insieme dati considerato. Il valore ottimale del numero di cluster risulta essere 20 in quanto, per questo indicatore, corrisponde al valore per cui l'andamento inizia la crescita. Inoltre, si evidenziano le migliori prestazioni, per questo insieme di dati, del Follow-the leader e dello hierarchical clustering con average linkage rispetto agli altri metodi, in quanto presentano valori più elevati, che per questo indice rappresenta un miglior raggruppamento. Inoltre, è opportuno evidenziare che per i metodi non deterministici sono state

ripetute una centinaia di volte i risultati per ogni valore del numero di cluster, riportando i valori medi ottenuti.

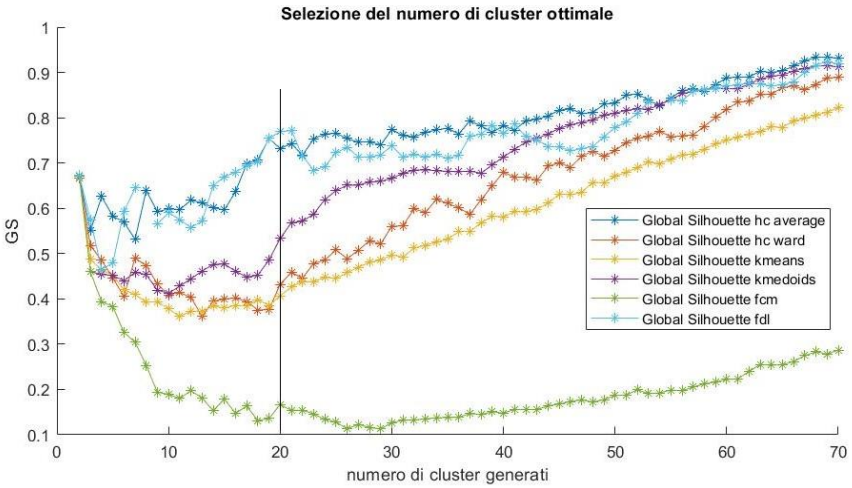


Figura 1.3.5 - Selezione del numero ottimale di cluster e confronto tra diverse tecniche di clustering con Global Silhouette

1.4 Applicazione dei metodi di clustering per curve di carico residenziali

Solitamente, l'applicazione dei metodi di clustering su curve di carico di utenti industriali o commerciali, prevede l'identificazione delle diverse condizioni di carico (come giorni festivi e feriali definiti in diverse stagioni), al fine di definire modelli di curve di carico tipiche in tali condizioni, chiamati *representative load patterns* (*RLP*) come presentato nell'articolo [8]. Essi, dunque sono ottenuti dalla media (o mediana) di tutte le curve di carico, del singolo utente, in una determinata condizione di carico (es. due settimane di giorni feriali in estate) e da una normalizzazione rispetto alla potenza media o di picco in quel periodo considerato. Tale processo applicato alle curve di carico disponibili, consente di livellare le variazioni occasionali che potrebbero verificarsi in uno specifico giorno e ottenere curve tipiche per ogni utente, nelle specifiche condizioni. Successivamente, si possono applicare direttamente ai RLP, i diversi metodi di clustering prendendo come caratteristiche del clustering direttamente i vari intervalli di tempo delle curve, come descritto nei paragrafi precedenti. Tuttavia, questo approccio risulta essere utile per utenti non residenziali, i quali presentano dei modelli di carico più regolari, rendendo tali procedure più affidabili. Contrariamente, la caratterizzazione delle curve di carico dei singoli utenti residenziali, richiede un'analisi statistica basata su diversi aspetti che vanno ad influenzare l'utilizzo dell'energia elettrica all'interno dell'abitazione, come ad esempio il numero di persone presenti nel nucleo familiare, le attività che essi svolgono, la loro età o il loro stile di vita, oltre ai diversi dispositivi che essi utilizzano. Inoltre, l'uso irregolare di tali apparecchiature e la presenza di pochi carichi che assorbono grandi potenze in un breve periodo di tempo, in maniera poco prevedibile, rendono difficile effettuare una rappresentazione dettagliata delle curve di carico, le quali risultano differenti da un giorno all'altro. Ciò significa che utenti con stesse caratteristiche e stesse apparecchiature risulterebbero comunque differenti, rendendo così inappropriati i metodi di clustering che utilizzano la distanza Euclidea come misura di similarità tra due serie temporali. Una soluzione a queste problematiche è l'utilizzo di una metrica basata

sulle forme, chiamata *Dynamic Time Warping* [9]. Essa infatti permette di deformare, in modo non lineare, gli assi dei tempi di due curve di carico prese in considerazione, tramite compressioni e allungamenti, in modo tale da trovare il miglior allineamento tra le due curve e poterle considerare simili pur presentando una diversità nel periodo di utilizzo di una stessa apparecchiatura. Tuttavia, l'utilizzo di tale distanza nell'algoritmo k-means, ha presentato qualche problema di inaccuratezza [10], che però non è stato riscontrato con l'utilizzo del k-medoids. Un'altra soluzione, per poter applicare i tipici metodi di clustering anche su curve di carico di tipo residenziale è quella di definire, a partire dalle serie temporali, un insieme di caratteristiche scelte opportunamente, come ad esempio fattori di forma [8] oppure di estrapolarle per mezzo di tecniche come: *discrete Fourier transform* [11] per definire coefficienti armonici, *discrete wavelet transform* [12] e *frequency domain analysis* [13]. La scelta delle caratteristiche da utilizzare come variabili d'ingresso ai metodi di clustering è molto importante, anche più della tecnica utilizzata. È spesso preferibile utilizzare un numero limitato di caratteristiche per ridurre l'onere computazionale e rendere i risultati del clustering ottimali. L'utilizzo di un numero ridotto di caratteristiche, richiede però, di selezionare le migliori proprietà che contengono maggiori informazioni possibili riguardo le serie temporali iniziali. Una proposta è stata sviluppata nell'articolo [14], nel quale particolari caratteristiche basate su dati probabilistici ricavati a partire dalle curve di carico dei singoli utenti vengono prese come riferimento per il clustering. A partire da una finestra temporale (es. tre settimane) del periodo di osservazione che si vuole studiare, prima si procede con la normalizzazione degli andamenti dei consumi di tutti gli utenti rispetto alle loro potenze di riferimento, e poi si ricava per ognuno le curve di durata (*Load Duration Curves*: *DLC*), disponendo in ordine decrescente i valori delle loro potenze. Partendo dai primi punti delle curve di durata, viene costruita una curva cumulativa delle variazioni medie di due successivi valori di tali curve tra tutti gli utenti, sommando di volta in volta i rispettivi valori. Dividendo tale curva per l'ultimo valore ottenuto, si costruisce una curva che viene interpretata come una funzione di distribuzione di probabilità (*CDF* dall'inglese *Cumulative Distribution Function*) dalla quale vengono individuati nove punti di taglio dell'asse orizzontale, a partire dagli ultimi decili di tale curva. I valori medi delle curve di durata di ogni utente all'interno degli ultimi nove intervalli, rappresentano le caratteristiche distintive di ogni utente, alle quali applicare i vari metodi di clustering. Questa metodologia si presta bene per la categorizzazione di utenti residenziali basata sulle intere curve di carico, raggruppando insieme utenti che presentano apparecchiature simili anche se utilizzate in periodi di tempo differenti. Ulteriori tecniche, più specifiche agli scopi che si vuole raggiungere, sono riportate nei capitoli successivi.

1.5 Bibliografia

[1] *Aggarwal C.C., Reddy C.K.: 'Data Clustering: Algorithms and Applications', 1st Edition, Chapman and Hall/CRC, 2014.*

[2] *Abu-Jamous B., Fa R., Nandi A. K.: 'Integrative Cluster Analysis in Bioinformatics', John Wiley & Sons, Incorporated, 2015.*

[3] *Chicco G., Napoli R., and F. Piglione: 'Comparisons among clustering techniques for electricity customer classification'. IEEE Trans. Power Syst., 21(2), pp. 933-940, 2006.*

[4] *Chicco G.: 'Overview and Performance Assessment of the Clustering Methods for Electrical Load Pattern Grouping'. Energy 42(1), pp. 68-80, 2012*

[5] *MathWorks, MATLAB Documentation.*

[6] *Chicco G., Napoli R., Postolache P., Scutariu M., and Toader C.: 'Customer Characterization Options for Improving the Tariff Offer'. IEEE Transactions on Power Systems, vol 18 (1), pp. 381-87, 2003.*

[7] *Chicco G., Napoli R., F. Piglione, Postolache P., Scutariu M., and Toader C.: 'Load Pattern-based Classification of Electricity Customers', IEEE Transactions on Power Systems, vol 19 (2), pp. 1232-239, 2004.*

[8] *Chicco G., Napoli R., F. Piglione, Scutariu M., Postolache P., and Toader C.: 'Emergent electricity customer classification'. IEE Proc Gener Transm Distrib. vol. 152, pp. 164–172, 2005.*

[9] *Teeraratkul T., O'Neill D., and Lall S.: 'Shape-Based Approach to Household Electric Load Curve Clustering and Prediction'. IEEE Transactions on Smart Grid, vol. 9, pp. 5196-5206, 2017.*

[10] *Niennattrakul V., and Ratanamahatana C. A.: 'On clustering multimedia time series data using k-means and dynamic time warping'. IEEE International Conference on Multimedia and Ubiquitous Engineering, pp. 733-738, 2007.*

[11] *Carpaneto E., Chicco G., Napoli R., and Scutari M.: 'Electricity customer classification using frequency–domain load pattern data', Int. J. Electr. Power & Energy Syst. vol. 28, pp. 13-20, 2006.*

[12] *Xiao Y., Yang J., Que H., Li M.J., and Gao Q.: 'Application of wavelet-based clustering approach to load profiling on AMI measurements', IEEE International Conference on Electricity Distribution (CICED), China, pp. 1537–1540, 2014.*

[13] *Zhong S., and Tam K.S.: 'Hierarchical classification of load profiles based on their characteristic attributes in frequency domain'. IEEE Trans. Power Syst., vol. 30, pp. 2434 - 2441, 2015.*

[14] *Cerquitelli T., Chicco G., Di Corso E., Ventura F., Montesano G., Del Pizzo A., Mateo González A., and Martin Sobrino E.: 'Discovering electricity consumption over time for residential consumers through cluster analysis', Proc. 14th DAS, Suceava, Romania, 2018.*

2 Base di dati per l'analisi di curve di carico residenziali

Per poter fare delle ricerche, sviluppare nuovi algoritmi o effettuare delle analisi, è necessario predisporsi di un insieme di dati specifici per lo scopo. Esistono diversi istituti accademici, governativi, scientifici, aziendali ecc., che presentano delle librerie (*data repository*) ovvero raccolte di dati di qualsiasi natura utili per la ricerca. Tuttavia, molti di essi non sono pubblici e/o gratuiti oppure risultano molto complessi per via della vastità di dati presenti all'interno, inoltre possono risultare inappropriati per gli scopi specifici della ricerca. Risulta perciò conveniente effettuare un'indagine mirata al genere di analisi che si vuole svolgere, cercando la tipologia, la dimensione e la provenienza di essi più consona per lo studio. *UCI Machine Learning Repository*, una raccolta di più di 400 dataset suddivisi per campo di applicazione, numero e tipologia di dati, è stata di grande aiuto per lo sviluppo della ricerca nell'ambito della data classification, del clustering e della regression. Tuttavia, ancora al giorno d'oggi presenta una ridotta varietà di scelta dei dataset nell'ambito dei sistemi elettrici. Un tentativo di costruzione di un *repository* per i dati riguardanti i consumi elettrici a livello residenziale è stato svolto nell'articolo [1] con lo scopo di raggruppare e confrontare diverse raccolte di dati, al fine di fornire un aiuto nella selezione del *dataset* per l'analisi che si sta svolgendo. I dati per i sistemi elettrici possono provenire da misurazioni reali, sondaggi oppure da simulazioni, e possono essere reperiti attraverso articoli di gruppi di ricerca, i quali forniscono maggiori dettagli ed esempi di applicazione. Grazie alle moderne tecnologie digitali come gli *smart meter*, i dati provenienti da essi oggi risultano più dettagliati, affidabili e accurati oltre ad essere ottenuti automaticamente, il che rende questa tipologia di dati favorita in molti studi. Tuttavia, per ragioni di *privacy* non tutti i dati possono essere disponibili o resi noti senza effettuare delle restrizioni. Alcune tipologie di dati possono essere ricavate dalle risposte a sondaggi sottoposti generalmente a consumatori o aziende. Solitamente, le informazioni che si ottengono, riguardano caratteristiche demografiche, comportamenti e profili dei consumatori, i loro livelli di confort oppure il loro approccio con le questioni ambientali, fornendo un'ottica più ampia, ovvero estesa a un grande numero di consumatori distribuiti geograficamente, rispetto ai dati ottenuti da misurazioni. In merito ai dati di consumi energetici di grandi città o paesi si può fare affidamento ad organizzazioni nazionali o internazionali come *International Energy Agency (IEA), Enerdata, data.gov.au, U.S. Energy Information Administration (EIA).* Un'altra valida alternativa, economica, rapida e conveniente, è quella di usare dei simulatori per generare i dati che più si addicono allo studio che si vuole svolgere. Il vantaggio è proprio quello di poter impostare e modificare facilmente, i parametri dei modelli utilizzati nella simulazione, ottenendo così lo scenario voluto. Tuttavia, seppur molto attendibili, non riflettono perfettamente i dati reali e richiedono informazioni dettagliate sui parametri fisici, ambientali e comportamentali dei modelli utilizzati, influenzando così l'accuratezza dei dati ottenuti.

2.1 Esempi di dataset disponibili

Sono stati riportati alcuni dei disponibili dataset utili per l'analisi di curve di carico di singoli utenti residenziali, al fine di fornire informazioni più dettagliate riguardo le caratteristiche, comuni o singolari, che si possono riscontare. Essi si presentano come serie temporali e possono variare in termini di frequenza (o periodo di campionamento) con la quale sono stati ottenuti

oppure in base alle grandezze che rappresentano (ad esempio potenze attive o reattive, tensioni, correnti), al numero di abitazioni ed eventualmente al numero dei loro occupanti o il livello di dettaglio dell'indagine ovvero se fanno riferimento all'intera abitazione oppure ai diversi circuiti o apparecchi al loro interno.

Reference Energy Disaggregation Data Set (REDD)

REDD [2] è una raccolta di dati, disponibile e gratuita, contenente informazioni reali, di potenze assorbite da diverse abitazioni, derivate da apparecchiature e circuiti a loro interno. L'obiettivo dei ricercatori di tale dataset è quello di incentivare la ricerca sull'energy disaggregation, in modo da eseguire facilmente algoritmi che determinano il contributo energetico di ogni singolo componente, a partire dalla conoscenza di un aggregato di apparecchiature. Il dataset presenta due tipi di dati in funzione della frequenza con cui sono stati raccolti:

- *low frequency power data*, i quali contengono misure, registrate con frequenza di circa 1 Hz, della potenza media sia delle reti principali, che dei singoli circuiti e/o apparecchi all'interno di sei abitazioni.
- *high frequency waveform data*, che presentano dati delle forme d'onda della corrente e della tensione in una fase delle reti principali nelle varie abitazioni.
- *high frequency raw data,* contiene dati non elaborati e senza compressioni di forme d'onda di tensioni e correnti, utili per verificare diversi metodi di compressione e filtraggio.

Il primo dei tre dataset rappresenta quello più utile per eseguire gli algoritmi di clustering sia per il vasto numero di misure ottenute da più appartamenti e sia per la possibilità di realizzare curve di carico giornaliere con il periodo di media voluto.

Pecan Street Dataport

Pecan Street [3] fornisce l'accesso, a collaboratori del settore e ricercatori universitari, ad una innumerevole quantità di dati riguardanti l'energia in ambito residenziale, raccolti da più di mille case situate nel Texas, Colorado, California, New York e in altri stati degli USA. I dati presenti sono organizzati in serie temporali riguardanti informazioni sull'energia ogni secondo, minuto o quarto d'ora, sull'acqua, sul gas naturale e dati ISO (Independent System Operator). Inoltre, sono presenti anche altre tipologie di dati come quelli relativi ai sondaggi sociodemografici, audit energetici e dati metereologici. In qualità di membro universitario di dataport, Pecan Street ha fornito l'accesso al set di dati relativi ai consumi di energia di apparecchiature e circuiti presenti in diverse abitazioni situate nelle seguenti regioni d'interesse:

- *New York*, con 6 mesi di dati relativi a 25 singole abitazioni e registrati per ogni secondo, minuto e quarto d'ora.
- *California*, con dati registrati ogni minuto e ogni 15 minuti, derivati da 23 abitazioni.
- *Austin (Texas),* in cui i dati si riferiscono a 25 case, registrati per ogni secondo, minuto e quarto d'ora.

Tracebase dataset

La raccolta dati *Tracebase* [4] è una collezione di consumi energetici di diverse apparecchiature elettriche, raccolti nelle città di Darmstadt e Sydney. I dati presentano informazioni sul giorno, orario e consumo misurato rispettivamente su uno e su otto secondi e sono raggruppati, con

frequenza di circa un secondo, nelle seguenti directory in base al processo di raccolta dei diversi dispositivi installati:
- *Complete,* contiene una raccolta dati completa, registrata nelle 24 ore giornaliere.
- *Incomplete,* presenta un insieme di misure nel quale si ha la perdita di qualche dato a causa di problemi relativi alla loro raccolta o trasmissione, risultando incompleti.
- *Synthetic,* raccoglie dati da apparecchiature reali solo durante il loro utilizzo, mediante dispositivi di misurazione. Mentre, per la realizzazione delle curve giornaliere, sono state aggiunte delle letture a consumo zero in quegli intervalli in cui non è stato utilizzato lo strumento di misura.
- *Australia,* sono presenti i dati raccolti a Sydney, i quali essendo di dimensioni ridotte, non sono state raggruppate nelle directory precedenti.

ECO data set (Electricity Consumption & Occupancy)

ECO [5] è un set di dati utile per diversi campi di ricerca e per eseguire algoritmi *NILM* (*Non-Intrusive Load Monitoring*) per l'energy disaggregation. I dati sono stati raccolti con frequenza di 1 Hz, per un periodo di otto mesi, da sei abitazioni in Svizzera e sono suddivisi in:
- dati sul consumo energetico di aggregati riguardanti potenze attive, correnti e tensioni e sfasamenti su tutte e tre le linee del sistema trifase.
- misure su diverse apparecchiature selezionate all'interno delle abitazioni.
- dati che includono informazioni sull'*occupancy* dell'abitazione.

Smart dataset*

Smart* [6] è un'ampia varietà sia di dati elettrici (consumo e generazione) che ambientali (temperatura e umidità) e operativi (eventi dell'interruttore). L'obiettivo di questo progetto è quello di ottimizzare i consumi energetici all'interno delle cosiddette "case intelligenti" le quali presentano strumentazioni avanzate per raccogliere dati di vita reale e per sperimentare nuove tecniche e algoritmi il per miglioramento dell'efficienza energetica domestica. In esso sono presenti due dataset con caratteristiche differenti:
- *Smart* Home Dataset* (versione del 2017), un set di dati di consumi totali (potenza media attiva) da più di 100 abitazioni, con risoluzione di 15 min per gli anni 2014 e 2015, e risoluzione di 1 min per l'anno 2016.
- *Smart* Microgrid Data Set,* un set di dati di consumi totali in un intero giorno ottenuti da più di 400 abitazioni con frequenza di un minuto.

Load Profile Generator

Load Profile Generator [7] è un generatore di profili di carico residenziali con la possibilità attribuire specifici comportamenti alle famiglie abitanti negli edifici presi in considerazione, influenzando così i loro consumi. È possibile dunque utilizzare modelli già predefiniti sia di singole case che di appartamenti comprendenti i comportamenti familiari, modificando il numero di persone, le loro attitudini al consumo e comportamenti giornalieri, oltre alla scelta di varie tipologie di dispositivi presenti all'interno dell'edificio, e alla presenza o meno di moduli fotovoltaici e stazioni di ricarica per veicoli elettrici, definendo così i percorsi che vengono abitualmente compiuti dai componenti della famiglia. Il software ha la possibilità, inoltre, di

modificare le impostazioni di calcolo in funzione dell'analisi che si sta effettuando, come il periodo di osservazione, il tempo di risoluzione, la temperatura e la localizzazione geografica.

2.2 Pre-elaborazione dei dati

Prima di effettuare una qualsiasi analisi è necessario fare alcune considerazioni sul dataset che si vuole analizzare, scegliendo opportunamente il periodo totale di osservazione per l'analisi (es. annuale, mensile, giornaliero), il numero di utenti sottomisura e l'intervallo di campionamento dei dati, cioè la precisione degli strumenti di misura (es. minuti, quarti d'ora, ore) che risulta essere limitato dalla quantità di dati raccolti dal sistema di misura. Una volta scelta la base di dati che si vuole analizzare, è opportuno effettuare delle verifiche su di essi, al fine di eliminare eventuali errori che potrebbero compromettere lo studio. L'insieme di dati scelto viene, perciò, sottoposto a procedure di *bad data detection* e *data cleaning,* effettuate manualmente come verrà visto in seguito, o tramite procedure automatiche [8,9]. Inoltre se si dispone di ulteriori informazioni come la misura dell'ampiezza della tensione oltre ai dati relativi ai consumi, si potrebbe migliorare il rilevamento delle interruzioni e buchi di tensione. Un successivo passo è quello dell'identificazione dei giorni anomali dovuti a festività, in quanto presentano comportamenti differenti da quelli standard. A tale scopo, si possono adottare differenti soluzioni come l'utilizzo di reti neurali artificiali (*Artificial Neural Network ANN*), nate principalmente per la previsione del carico (*short-term load forecasting STLF*) [10] ma che trovano applicazione in svariati ambiti come *function fitting*, *pattern recognition*, *data clustering* e *time-series analysis*. Essi rappresentano un modello computazionale basato su un insieme di unità connesse tra loro, chiamate neuroni artificiali, che simulano quelle del cervello umano. Tali unità comunicano tramite delle connessioni pesate, trasportando segnali di attivazione. Esistono diverse architetture di rete con cui vengono organizzati i neuroni, la più comune è quella basata su *multilayer perceptron MLP*, in cui essi, sono organizzati in diversi strati ognuno dei quali presenta neuroni non connessi tra loro, ma che potrebbero condividere segnali d'ingresso derivanti dagli stessi nodi iniziali o dagli stessi neuroni, in caso di *feed-forward architecture*. Ogni unità presenta, segnali d'ingresso con i relativi pesi (*weights*), e un parametro costante caratteristico dell'unità chiamato *bias*, entrambi utili per generare un segnale d'uscita, tramite una funzione di attivazione avente in entrata la combinazione lineare di tali ingressi. La stima di questi parametri, chiamata *training* della rete, può essere effettuata mediante differenti metodi di ottimizzazione il più comune è il *back-propagation*. Le reti neurali artificiali presentano il vantaggio di approssimare numericamente qualsiasi funzione continua con un'accuratezza desiderata, inoltre, essendo un metodo *data-driven*, sono in grado di individuare ed apprendere automaticamente delle relazioni ingresso-uscita, e immagazzinare tali relazioni nei parametri della rete, senza la necessità da parte del ricercatore di stabilire modelli e loro parametri a priori, ma fornendo solo vettori di ingresso e uscita. L'articolo [11] utilizza le reti neurali artificiali per prevedere periodi in cui si hanno condizioni di carico anomalo (come scioperi, vacanze, weekend lunghi ecc.) da dati sui consumi giornalieri. La procedura che viene utilizzata è un approccio combinato tra uno stadio non-supervisionato che fornisce una classificazione preventiva dei dati storici per mezzo della *Self-Organising Map* (*SOM*), e un secondo stadio supervisionato della procedura realizzato dalla ANN, la quale fornisce migliori risultati sulla previsione dei giorni anomali, grazie al primo stadio che assiste l'esperto umano nel trovare il training test per la rete. La SOM sviluppata da *Kohonen* è una

ANN composta da una rete predefinita di unità, di solito bidimensionali, che formano uno strato competitivo, in cui solo un'unità è attivata ad ogni ingresso. Essa, infatti, proietta un insieme di dati di *N* dimensioni in ingresso rappresentati con vettori pesati, in uno spazio dimensionalmente ridotto, in cui tali dati vengono descritti dalle loro posizioni in tale rete. L'algoritmo di apprendimento, richiede l'implementazione di un criterio di discriminazione, generalmente la distanza Euclidea tra gli ingressi e l'uscita, selezionando, per ogni dato d'ingresso, l'unità "vincitrice" con la minima distanza, e quelle vicine ad essa tramite una funzione (*neighbourhood function*). I pesi dei vettori di tali unità selezionate verranno continuamente aggiornate durante il training, generando aree separate (*bubbles of activity*) in espansione per tutta la mappa. Grazie alla conservazione della topologia dei dati iniziali (non presente nei classici clustering), la distanza tra queste aree è proporzionale a quella dei dati, fornendo una rappresentazione visiva delle caratteristiche del dataset di partenza che permette di riconoscere adeguatamente gruppi di dati simili ed eventuali outlier. Nell'articolo [12] sono state confrontate le prestazioni della SOM con una versione modificata dell'algoritmo classico Follow-The-Leader, al fine di identificare i giorni anomali in specifici mesi dell'anno. Un'altra metodologia presente in letteratura è l'applicazione di alcuni algoritmi di clustering per distinguere periodi in condizioni non standard da quelli standard. Gli autori dell'articolo [13] seguono una procedura a più stadi, partendo da sette sottomatrici riferite ad ogni giorno della settimana, per un dato anno di misure, e ad ognuna di esse, per ogni giorno, sono ricavate alcune caratteristiche, come la potenza media giornaliera, in modo tale da poter applicare un primo clustering in grado di separare ogni sottomatrice in sottogruppi stagionali. Alle curve di carico di tali sottogruppi sono applicati a loro volta altri algoritmi di clustering in grado di identificare gli outlier da cluster con minor numero di curve (una o due) e poi salvate in una lista, mentre i gruppi più grandi vengono mediati e classificati come curve di carico "standard". Un successivo stadio, rifinisce tale attività di identificazione, classificando ogni outlier in tre differenti classi: giorni lavorativi, sabati, e giorni festivi o domeniche. Un successivo clustering viene applicato ad ogni sottogruppo, stabilendo in tal modo se confermare il singolo outlier come giorno "non standard" oppure se inserirlo in una lista a parte. Gli algoritmi che gli autori hanno impiegato in questa trattazione utilizzano un'imitazione dell'attrazione gravitazionale (*gravitational attraction*) ed un algoritmo *Min-Max neuro-fuzzy* modificato.

Ricostruzione dei dati di *Smart* dataset* ed *ECO data set*

Per i seguenti studi si è scelto di utilizzare *Smart* dataset*, in quanto presenta un'elevata numerosità di abitazioni, ed *ECO data set*, nel quale si hanno differenti informazioni con una buona risoluzione temporale, utile per effettuare diverse analisi. Una volta portati in ambiente Matlab, i dataset sono stati verificati, osservando e correggendo manualmente ogni serie temporale, senza ricorrere a logiche automatiche di filtraggio e/o check dei dati, al contrario, necessarie per grandi quantità di dati. Le prime verifiche che sono state effettuate riguardano la completezza dei dati nelle serie temporali, in corrispondenza alla loro frequenza di acquisizione e l'individuazione di dati mancanti o persi, spesso identificati con *valori NaN*. In particolare, in *ECO data set*, per entrambe le raccolte di dati riguardanti le potenze attive (aggregati e singole apparecchiature), i dati persi, qualora risultassero maggiori di dieci consecutivi, sono stati distinti segnandoli con il valore -1. Oltre alla ricerca dei dati mancanti, è opportuno controllare se vi è la presenza di *spike*, cioè dati che assumono, in modo imprevedibile (non ripetitivo), valori superiori anche centinaia di volte quelli precedenti e privi di alcun significato fisico, ma

che possono incidere sul calcolo dei valori massimi. Inoltre, risulta ragionevole effettuare una ricerca di serie temporali completamente nulle, con lo scopo di eliminarle dal dataset e ridurre la numerosità dei dati da trattare. In seguito, sono state riportate, per i diversi dataset in esame, alcune serie temporali contenenti dati ingiustificati, i quali sono stati trattati in maniera differente, a seconda del dataset considerato, in quanto ognuno presenta caratteristiche differenti.

Il dataset *Smart* Microgrid* presenta una quantità considerevole di abitazioni contenenti informazioni nulle all'interno, a queste sono state aggiunte altre serie temporali come quelle riportate in *Figura 2.2.1*, le quali presentano esclusivamente degli *spike* privi di senso all'interno. Viceversa, in *Smart* Home Dataset* per l'anno 2016, sono stati riscontrati più di un centinaio di *spike* con valori molto elevati, come si può notare in *Figura 2.2.2*, i quali stati sostituiti con valori intermedi tra il punto precedente e quello successivo poiché, la maggior parte degli *spike* presenti in questo dataset sono seguiti da un buco notevole con elevata variazione negativa, risultando un comportamento anomalo rispetto agli andamenti precedenti, i quali presentano delle variazioni passando per punti intermedi.

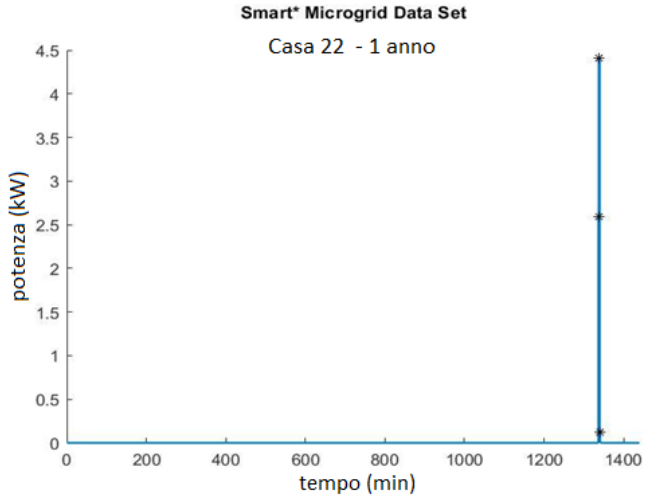


Figura 2.2.1 - Presenza di un dato anomalo in Smart Microgrid dataset – Casa 22*

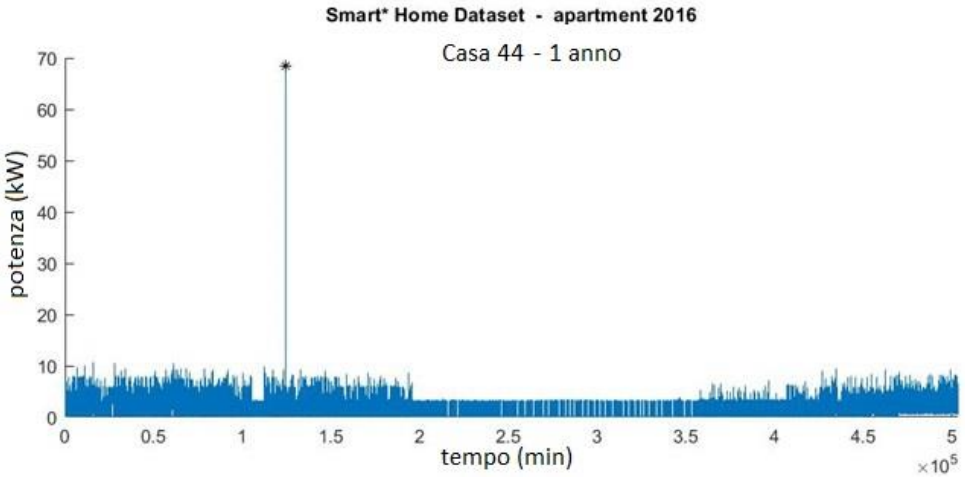


Figura 2.2.2 - Presenza di un dato anomalo in Smart Home Dataset –apartament 2016 in Casa 44*

A differenza dei dataset precedenti in *ECO dataset,* non sono stati rilevati *spike* all'interno delle raccolte di dati, ma è stata rilevata una quantità innumerevole di dati mancanti, per cui sono state applicate differenti soluzioni. Qualora il numero consecutivo di dati persi è risultato elevato, come mostrato in *Figura 2.2.3*, si è scelto di non considerare l'intera serie temporale, classificandola come una serie di dati nulla e quindi scartandola dal dataset; al contrario, per quelle serie temporali in cui il numero di dati persi consecutivi è risultato accettabile, al fine di non perdere troppe informazioni, è stato scelto di sostituire i dati anomali con valori che replicano un andamento simile nel resto del dataset, ovvero in presenza di una certa regolarità negli andamenti, è stata valutata la media delle variazioni in situazioni simili e applicata la stessa variazione per ricostruire un andamento simile, come è visibile in *Figura 2.2.4*.

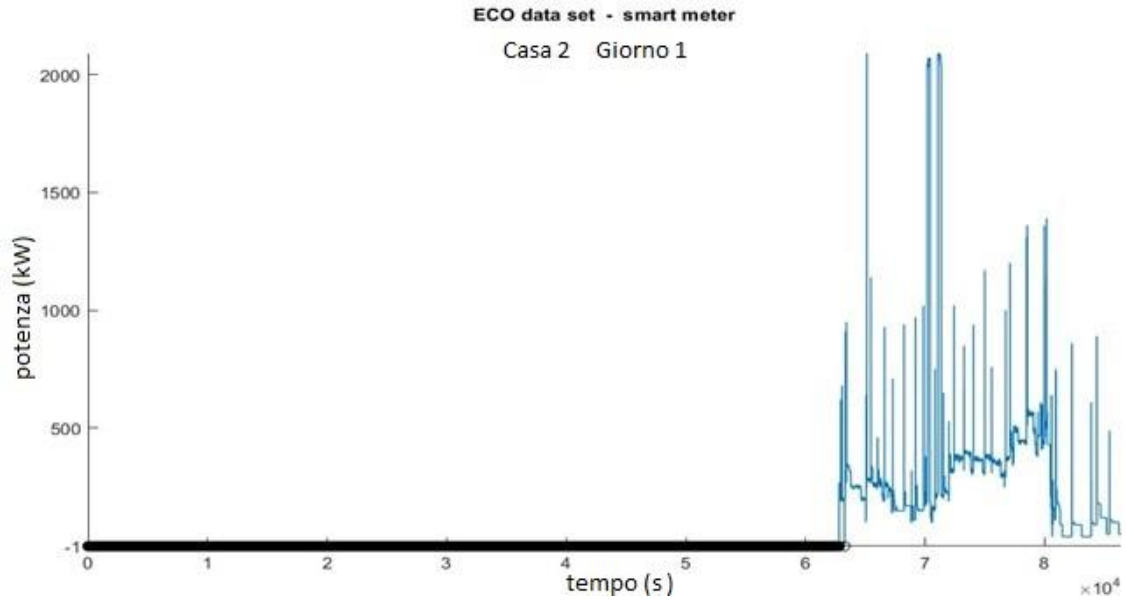


Figura 2.2.3 - Dati persi in ECO dataset _ smart meter - Casa 2 e Giorno 1

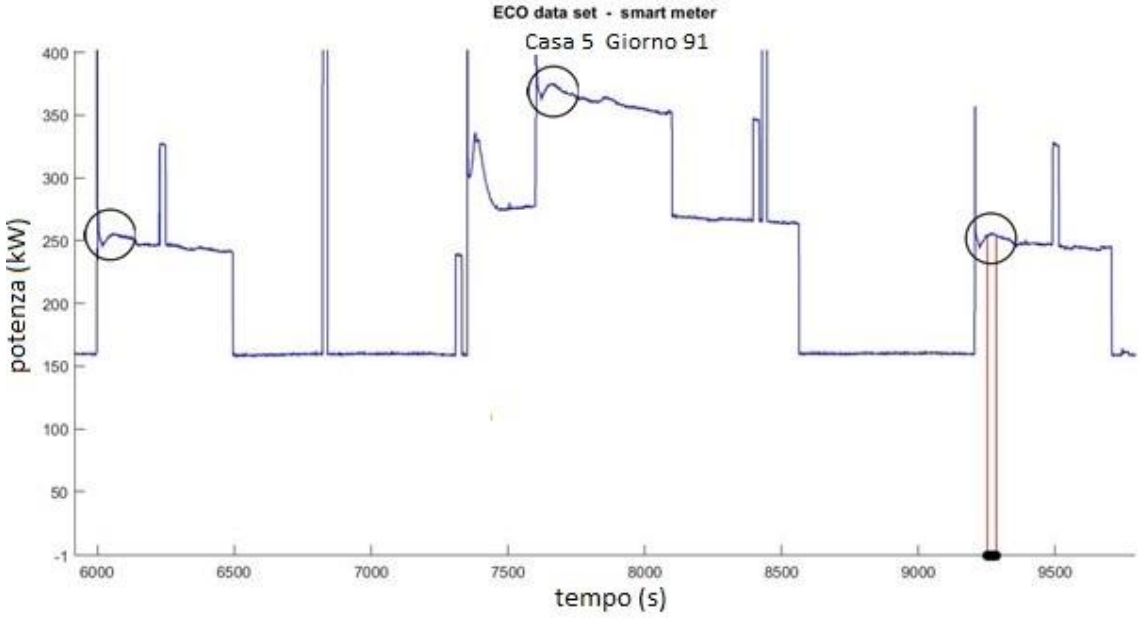


Figura 2.2.4 - Dati persi e loro sostituzione - ECO dataset _ smart meter in House 5 e Giorno 91

2.3 Bibliografia

[1] *Babaei T., Abdi H., Lim C.P., and Nahavandi S.: 'A Study and a Directory of Energy Consumption Data Sets of Buildings', Energy and Buildings, pp. 91-99, 2015.*

[2] *Kolter J.Z. and Johnson M.J.: 'REDD: A Public Data Set for Energy Disaggregation Research'. Proceedings of Workshop on Data Mining Applications in Sustainability (SIGKDD), 2011.*

[3] *Holcomb C., Pecan Street Inc.: 'a test-bed for NILM', International Workshop on Non-Intrusive Load Monitoring, Pittsburgh, PA, USA, 2012.*

[4] *Reinhardt A., Baumann P., Burgstahler D., Hollick M., Chonov H., Werner M., and Steinmetz R.: 'On the Accuracy of Appliance Identification Based on Distributed Load Metering Data'. Proceedings of the 2nd IFIP Conference on Sustainable Internet and ICT for Sustainability (SustainIT), 2012.*

[5] *Becke C. l, Kleiminger W., Cicchetti R., Staake T., and Santini S.: 'The ECO Data Set and the Performance of Non-intrusive Load Monitoring Algorithms'. Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys), 2014.*

[6] *Barker S., Mishra A., Irwin D., Cecchet E., Shenoy P., and Albrecht J.: 'Smart*: An Open Data Set and Tools for Enabling Research in Sustainable Homes'. Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD), 2012.*

[7] *Pflugradt N., Teuscher J., Platzer B., Schufft W.: 'Analysing Low-Voltage Grids using a Behaviour Based Load Profile Generator'. International Conference on Renewable Energies and Power Quality, Bilbao, vol. 1, 2013.*

[8] *Monticelli, A.: 'Electric power system state estimation', Proc. IEEE, vol. 88 (2), pp. 262–282, 2000.*

[9] *Zhang X, and Sun C.: 'Dynamic intelligent cleaning model of dirty electric load data'. Energy Conversion Management, Vol.49 (4), p.564-569, 2008.*

[10] *Hippert, H.S., Pereira, C.E., and Souza, R.C.: 'Neural networks for short-term load forecasting: a review and evaluation', IEEE Trans. Power Syst., vol. 16 (1), pp. 44–55, 2001.*

[11] *Lamedica, R., Prudenzi, A., Sforna, M., Caciotta, M., and Orsolini Cencelli, V.: 'A neural network-based technique for short-term load forecasting of anomalous load periods', IEEE Trans. Power Syst., vol. 11 (4), pp. 1749–1756, 1996.*

[12] *Chicco, G., Napoli, R., and Piglione, F.: 'Load pattern clustering for short-term load forecasting of anomalous days', Proc. IEEE Porto Power Tech 2001, Porto, Portugal, 10–13 September 2001, paper AIT2-377.*

[13] *Lamedica, R., Santolamazza, L., Fracassi, G., Martinelli, G., and Prudenzi, A.: 'A novel methodology based on clustering techniques for automatic processing of MV feeder daily load patterns', Proc. IEEE PES Summer Meeting 2000, Seattle, WA, 16–20 July 2000, Vol. 1, pp. 96–101.*

3 Analisi sulla flessibilità della domanda aggregata

Uno dei principali requisisti che deve soddisfare il sistema elettrico è quello di garantire il giusto equilibrio tra generazione e domanda. Tuttavia, entrambe sono molto incerte, soprattutto negli ultimi anni in cui si è vista un'importante integrazione delle sorgenti di energia rinnovabile che rendono il sistema ancora più complesso, con flussi di energia ad alta volatilità e bassa prevedibilità. Diverse soluzioni sono state proposte e attuate nel corso degli anni, integrando nella rete sistemi di accumulo (*Energy Storage Systems: ESS*), oppure agendo direttamente sia sul lato della generazione, con la disponibilità e l'intervento di impianti di generazione convenzionali, sia sulla domanda, con programmi specifici che rendono più efficiente la gestione dal lato dell'utente. Una delle più importanti caratteristiche che bisogna valutare per tali soluzioni è la flessibilità delle risorse nei vari settori del sistema elettrico. Con il termine *flessibilità*, nei sistemi elettrici, si intende la possibilità di sviluppare risorse disponibili per rispondere in modo adeguato e affidabile alle variazioni di carico e generazione nel tempo a costi accettabili. Ciò può essere raggiunto utilizzando componenti che possono adattare e muovere la produzione e il consumo a differenti intervalli di tempo. Esso è un tema generale che viene definito in base al contesto in cui si riferisce [1]:

- Sul lato della generazione, la flessibilità, viene intesa come la possibilità di modulare con una certa "rapidità" la potenza immessa dalle unità di produzione elettrica e viene quantificata con un livello di *flessibilità tecnica* basato sulla massima capacità di generazione e sulla presa in servizio sia di generatori individuali che dell'intero sistema di generazione.

- Sul lato rete, la flessibilità è definita come la capacità di una rete elettrica di distribuire le proprie risorse al fine di coprire le variazioni di potenza nel sistema. In questo caso sono presi in considerazione i carichi, le sorgenti e le riserve presenti nel sistema elettrico ottenendo in tal modo un indice di *flessibilità operativa* a partire dalla fornitura di potenza presente e dalla rapidità con la quale viene distribuita.

- Sul lato domanda, il concetto di flessibilità è applicato in modo differente a seconda se vengono considerate le singole apparecchiature oppure aggregati di carichi. Per i primi, gli indici di flessibilità sono molto influenzati dallo stile di vita e dalle preferenze degli utenti e vengono valutati effettuando sondaggi e questionari. Tra i quali, in letteratura, si possono trovare: *ADT* (*consumers' Acceptable Delay Time*), che indica il massimo periodo di tempo al quale si può posticipare il funzionamento di un apparecchiatura senza peggiorare il confort dell'utente, e *AFI* (*Appliance Flexibility Index*) il quale misura l'intervallo di tempo regolabile delle apparecchiature. Mentre, per gli aggregati di carichi, ci sono diversi approcci che permettono di definire la flessibilità, alcuni dei quali fanno uso di *sensitivity function* considerando le probabilità degli utenti a spostare, per un determinato tempo, l'uso delle apparecchiature in seguito ad una remunerazione nel nuovo periodo di utilizzo, altri utilizzano approcci *agent-based* fondati su algoritmi Q-learning che permettono di ottenere fattori di flessibilità utili per la simulazione dell'elasticità della domanda, oppure altri criteri che consistono nel suddividere le diverse tipologie di carico in "*sheddable"*, "*controllable"* o "*acceptable"*, definendo un diverso livello di flessibilità. L'approccio che viene seguito in questa trattazione è basato su statistiche di variazioni della domanda aggregata tra due intervalli di tempo successivi, le quali

quantificano la flessibilità tenendo conto dell'incertezza presente nei modelli delle curve di carico ottenute da gruppi di utenti residenziali. Per questa tipologia di carichi, vengono definiti i seguenti indicatori: *Flexibility Index of Aggregate Demand* (*FIAD*) che esprime l'andamento collettivo degli aggregati di carichi, valutando per ogni intervallo di tempo in cui è definito, il grado di flessibilità compreso tra 0 e 1, in cui la domanda aggregata risulta meno rigida e *Percentage Flexibility Level* (*PFL*) che indica la percentuale di domanda aggregata che è possibile aumentare o ridurre senza che si influenzi il cambiamento medio del gruppo di utenti, nella domanda.

3.1 Introduzione ai programmi di Demand Response

L'informazione sulla flessibilità può aiutare a risolvere diversi problemi nei sistemi elettrici, avviando alcune tecniche di gestione della domanda come i programmi di *Demand Response* (*DR*) [2]. Essi, infatti, sono uno strumento importante nei sistemi di distribuzione, in quanto permettono l'integrazione delle RES, la riduzione di emissioni e migliorano l'efficienza sia energetica che economica. La loro principale funzione è di variare i consumi degli utenti, in risposta a specifici programmi, applicando alcuni principi base come il *peak shaving*, che permette di ridurre i picchi di domanda, il *valley filling*, per aumentare il consumo di energia nelle ore in cui la domanda è minore, e il *load shifting*, con l'obiettivo di spostare i consumi da ore di punta a ore non di punta. Generalmente si suddividono in programmi basati su incentivi (*incentive-based programmes*) e programmi basati sul prezzo (*price-based programmes*). I primi, forniscono agli utenti dei pagamenti a seguito del servizio svolto durante un evento in cui il sistema lo necessita (ad esempio la riduzione della domanda nelle ore di punta), mentre i secondi, offrono dei prezzi dell'elettricità variabili nel tempo, con l'obiettivo di far spostare i consumi nei periodi in cui è più conveniente, facendo risparmiare sulla bolletta. I programmi di DR trovano maggiore applicazione su grandi utenti, come quelli nel settore commerciale e industriale. Gli utenti residenziali, nonostante il grosso contributo nella domanda dell'energia elettrica, trovano diverse problematiche nella gestione, per via dell'alta variabilità e poca prevedibilità dei loro consumi, oltre alla dipendenza di svariati fattori esterni. Tuttavia, grazie al recente sviluppo degli *smart meter* che forniscono all'operatore del sistema informazioni utili riguardo i consumi dei singoli utenti, i programmi di DR per carichi residenziali sono diventati un'attività di ricerca. Non a tutti gli utenti residenziali è possibile far avviare questi programmi. È necessario isolare gli utenti che presentano andamenti non utili a tali scopi e identificare gruppi di possibili candidati per questi programmi tramite tecniche di clustering specifiche, per esempio, basate sul tipico comportamento dei carichi residenziali nei rispettivi periodi temporali specifici della giornata. Nell'articolo [3], dopo aver effettuato alcune analisi statistiche sui dati, gli autori identificano quattro periodi temporali rilevanti in una giornata, valutando la frequenza delle mezz'ore (intervallo minimo con cui sono stati analizzati i dati all'interno della giornata) che superano una soglia di consumo in tutto il dataset, ovvero per tutti gli utenti e tutti i giorni dell'intero anno, distinguendo le stagioni. Dalle distribuzioni dei picchi di domanda che si presentano durante la giornata, per diverse taglie di consumi (analisi ripetuta su diverse soglie), sono stati messi in evidenza quattro periodi, che coincidono con i periodi tipici di attività familiari: *overnight*, *breakfast*, *daytime*, ed *evening*.

Sulla base di questi, sono state definite, per ogni utente, sette caratteristiche usate come ingresso al clustering:

- quattro potenze medie relative, come il rapporto tra la potenza media di tutte le giornate dell'anno riferite al periodo della giornata considerato e la potenza media dell'intera giornata;
- deviazione standard relativa media sull'intero anno, come la media dei rapporti tra le deviazioni standard e le potenze medie in ogni periodo della giornata;
- un indice stagionale, come la somma dei rapporti tra i valori assoluti delle differenze tra le potenze medie dei giorni invernali e di quelli estivi riferite nel periodo della giornata considerato e le potenze in tali periodi rispettivamente, mediate in tutto l'anno;
- un indice di differenza tra giorni feriali e fine settimana, come l'indice precedente ma con riferimento alla differenza tra potenze medie dei giorni feriali e quelle dei fine settimana.

Tali caratteristiche oltre ad analizzare i diversi periodi della giornata, mettono in evidenza le differenze di domanda anche in funzione della stagione e dei giorni della settimana, oltre ad una misura di irregolarità di un utente tramite la deviazione standard. Definite le caratteristiche, gli autori utilizzano il *finite mixture model* (*FMM*) come tecnica di clustering, abbastanza comune ma poco utilizzata nelle applicazioni di sistemi elettrici, la quale in confronto al k-means risulta essere più versatile, in quanto permette di lavorare anche con tipologie di dati continui o categoriali. Altre soluzioni di applicazione dei metodi di clustering, al fine di raggruppare utenti residenziali adatti alla partecipazione di programmi di DR specifici, sono riportate nel capitolo successivo.

3.2 Variazioni di domanda per curve di carico residenziali aggerate

La presenza di flessibilità nel comportamento degli utenti, risulta essere una soluzione interessante per ottenere un maggiore risparmio energetico nell'instaurare programmi specifici di DR. A tale scopo è necessario conoscere al meglio i modelli dei consumi degli utenti che andranno poi a rappresentare il reale andamento della domanda. Per via della maggiore disponibilità di dati reperibili nell'arco di un'intera giornata e per diversi anni, i carichi residenziali vengono spesso considerati in vari studi. Tuttavia, l'alta variabilità e la scarsa prevedibilità di tali tipologie di carico rendono difficile la caratterizzazione delle curve di carico dei singoli utenti. Il DSO o l'aggregatore, ovvero una singola entità nel sistema elettrico che raggruppa diversi agenti (consumatori, produttori, *prosumer* o loro combinazioni), gestiscono la domanda di un certo numero di curve di carico dei diversi utenti, usando un'ottica più sistemistica tramite un modello aggregato dei consumi, il quale risulta avere delle variazioni nel tempo più regolari rispetto ai consumi dei singoli utenti. A partire dal diverso comportamento dei singoli utenti, per un certo numero di utenti residenziali connessi alla stessa dorsale o cabina elettrica, è possibile, dunque, ricavare un andamento del carico totale che è prevedibile in modo relativamente semplice; la regolarità in queste curve dipende però dal numero di utenti aggregati. In particolare, se esso risulta essere piccolo (ad esempio inferiore a 20), l'evoluzione temporale del carico potrebbe risultare più fluttuante con variazioni significative a causa dell'utilizzo irregolare delle singole apparecchiature di grandi potenze presenti nelle abitazioni degli utenti. Ciò è stato dimostrato nell'articolo [4], nel quale, per un numero di utenti extraurbani è stata valutata la variazione di valor medio e deviazione standard della domanda aggregata.

L'analisi svolta in questo capitolo fa riferimento a dati dello *Smart* Home Dataset* per l'anno 2016, dopo aver svolto relative tecniche di clustering e aver selezionato un potenziale gruppo di utenti adatti ad eventuali programmi specifici di DR, come viene esposto nel paragrafo 3.5. In particolare, sono state ricavate, con il metodo Monte Carlo, 100 osservazioni di curve di carico aggregate, ognuna delle quali è stata realizzata a partire da un certo numero di curve di carico dei singoli utenti presi casualmente dal dataset selezionato, riferito a giorni feriali durante la stagione in cui si ha basso carico. In *Figura 3.2.1,* sono stati rappresentati due scenari delle 100 curve di carico aggregate, realizzate rispettivamente con 20 e 300 carichi aggregati; è possibile notare che bassi valori di utenti aggregati presentano maggiori incertezze sul valore medio.

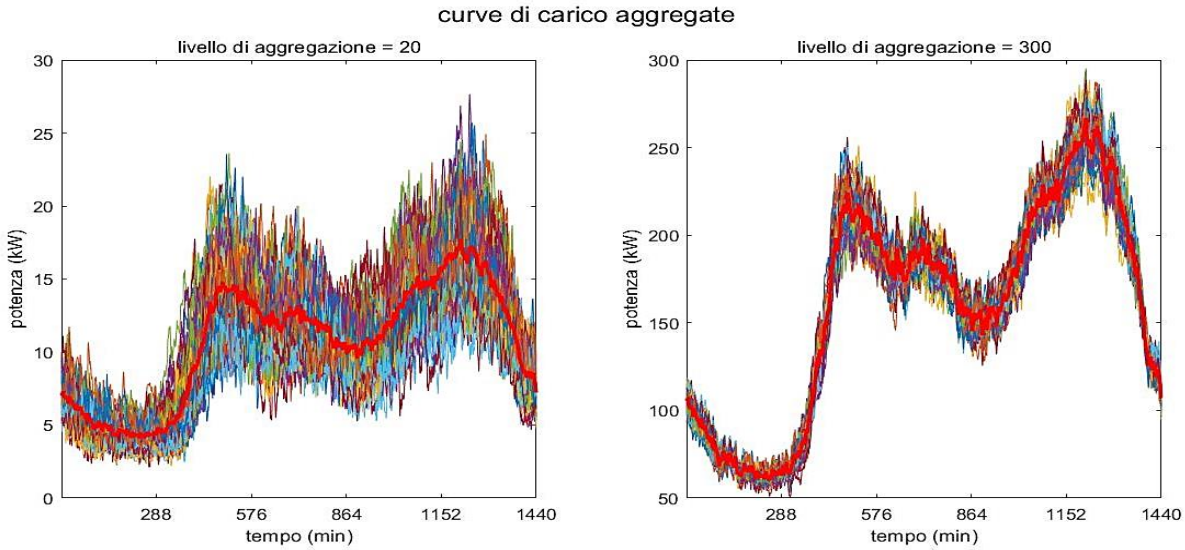


Figura 3.2.1 - Curve di carico aggregate per diversi livelli di aggregazione

Siano *a* il *livello di aggregazione*, ovvero il numero di utenti di cui è composto l'aggregato, e *s* il *periodo di campionamento*, cioè l'intervallo di tempo considerato tra due variazioni della domanda, due parametri ai quali si riferiscono tali variazioni e che hanno degli effetti importanti su di esse e quindi sul calcolo della flessibilità. Le diverse curve di carico aggregate sono rappresentate matematicamente attraverso una matrice $\mathbf{P}^{(a,s)}$ contenente K righe, pari al numero di osservazioni di esse, e n_s colonne, corrispondenti al numero di intervalli di tempo in cui è suddiviso il periodo di osservazione rispetto all'intervallo di campionamento considerato.

$$\mathbf{P}^{(a,s)} = \begin{bmatrix} p_{1,1\Delta t_s}^{(a,s)} & \cdots & p_{1,n_s\Delta t_s}^{(a,s)} \\ \vdots & \ddots & \vdots \\ p_{K,1\Delta t_s}^{(a,s)} & \cdots & p_{K,n_s\Delta t_s}^{(a,s)} \end{bmatrix} \tag{3.2.1}$$

L'analisi che viene svolta considera le variazioni di carico riferendosi ad un incremento o decremento tra due intervalli di tempo successivi della curva di carico aggregata, per un dato periodo di campionamento e livello di aggregazione, in modo tale da incorporare solo le informazioni sui cambiamenti della domanda, utili nella determinazione della flessibilità in certi periodi di tempo.

Sia $\Delta p_{k,x\Delta t_s}^{(a,s)}$ un generico valore della variazione di carico del *k-esimo* andamento temporale tra l'istante *x* e *x*-1 calcolato come:

$$\Delta p_{k,x\Delta t_s}^{(a,s)} = p_{k,x\Delta t_s}^{(a,s)} - p_{k,(x-1)\Delta t_s}^{(a,s)} \qquad \text{per } x = 2,3,\dots,K \tag{3.2.2}$$

Si costruisce la seguente matrice $\Delta\mathbf{P}^{(a,s)}$ delle variazioni di carico di dimensioni $K\times(n_s\text{-}1)$:

$$\Delta\mathbf{P}^{(a,s)} = \begin{bmatrix} \Delta p_{1,2\Delta t_s}^{(a,s)} & \cdots & \Delta p_{1,n_s\Delta t_s}^{(a,s)} \\ \vdots & \ddots & \vdots \\ \Delta p_{K,2\Delta t_s}^{(a,s)} & \cdots & \Delta p_{K,n_s\Delta t_s}^{(a,s)} \end{bmatrix} \tag{3.2.3}$$

Inoltre, risulta comodo rappresentare le matrici della domanda e delle sue variazioni attraverso vettori colonna, tenendo presente che ognuno di essi rappresenta un vettore delle *K* osservazioni nell'intervallo di tempo considerato:

$$\mathbf{P}^{(a,s)} = \left[\boldsymbol{p}_{1\Delta t_s}^{(a,s)},\dots,\boldsymbol{p}_{n_s\Delta t_s}^{(a,s)}\right] \tag{3.2.4}$$

$$\Delta\mathbf{P}^{(a,s)} = \left[\Delta\boldsymbol{p}_{2\Delta t_s}^{(a,s)},\dots,\Delta\boldsymbol{p}_{n_s\Delta t_s}^{(a,s)}\right] \tag{3.2.5}$$

In *Figura 3.2.2* e *Figura 3.2.3* sono riportate le diverse curve di carico aggregate giornaliere e le rispettive variazioni riferite alle stesse condizioni esplicitate precedentemente, ma con un livello di aggregazione pari a 50 e un periodo di campionamento di 30 minuti, allo scopo di rendere maggiormente visibile gli indicatori successivamente analizzati, oltre a fornire un ulteriore scenario a quelli studiati nel paragrafo 3.5.

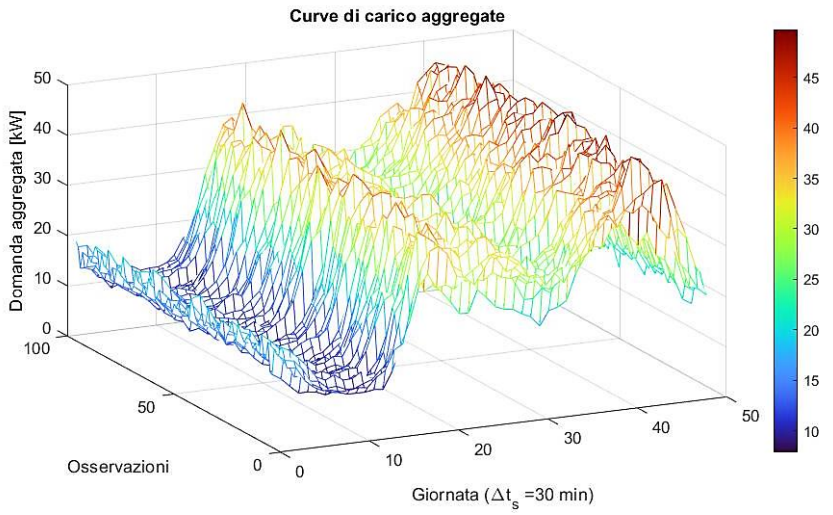


Figura 3.2. 2 - Curve di carico aggregate per un giorno feriale, con a=50 e s=1 min

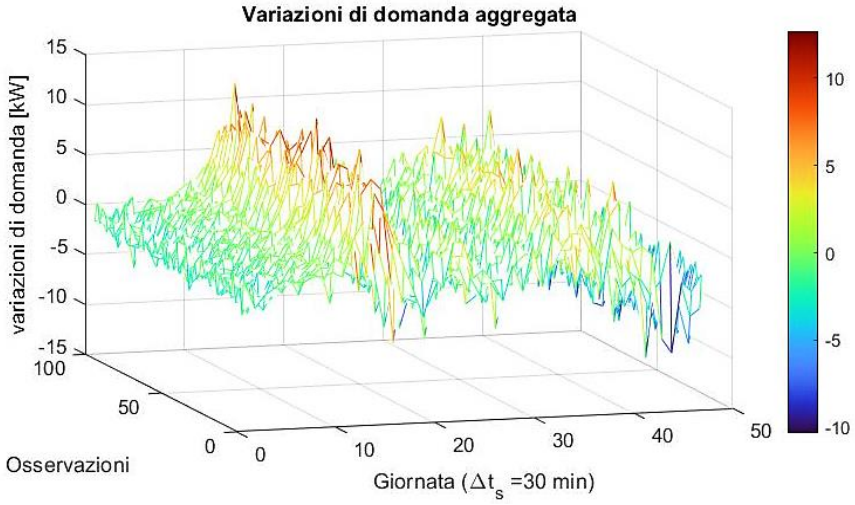


Figura 3.2.3 - Variazioni delle curve di carico aggregate per un giorno feriale, con a=50 e s=1 min

In diversi articoli presenti in letteratura sono state effettuate delle analisi statistiche per valutare l'effetto che il livello di aggregazione e il periodo di campionamento hanno sulle variazioni di carico [5, 6]. Anche esse sono state svolte su dataset di utenti residenziali extraurbani, e seppur i risultati non sono del tutto generalizzabili a causa di una possibile dipendenza da particolari situazioni topologiche e demografiche, risultano comunque indicativi di un possibile andamento delle variazioni di carico nel tempo. Sulle curve aggregate di tali carichi sono stati effettuati alcuni test statistici di tipo non parametrico (*Two-sample Kolmogorov-Smirnov test* e *Wilcoxon rank sum test*) al fine di valutare la similarità delle distribuzioni probabilistiche dei dati tra le diverse osservazioni. I vari test sono stati convalidati con ulteriori parametri che forniscono una relazione matematica e permettono di valutare le differenze tra i diversi dataset aventi differenti intervalli di campionamento e livelli di aggregazione: *ARSD* (*percentage relative standard deviation*), un parametro che indica la casualità nelle variazioni di carico (in valore assoluto) e *NLV%* (*percentage normalized load variations*) che misura l'andamento medio giornaliero delle

ampiezze delle variazioni di carico basato su un giorno tipico. I risultati che sono stati ottenuti dimostrano che per valori elevati dell'intervallo di campionamento tutte le osservazioni, approssimativamente, appartengono alla stessa distribuzione dei dati, perdendo così l'informazione dei diversi consumi degli utenti e quindi l'impatto che essi hanno sul sistema aggregato, allontanandoci così dalla reale dinamica del sistema. Anche con l'aumento del livello di aggregazione, l'evoluzione temporale dell'aggerato tende a diventare sempre più simile e quindi più liscia, ottenendo sempre meno informazioni sui cambiamenti dei carichi che invece potrebbero essere introdotte nelle curve aggregate, rinunciando così alla possibilità di sfruttare la flessibilità da tale andamento. Tali indici sono stati valutati per il dataset che si sta analizzando, ottenendo i risultati mostrati in *Figura 3.2.4*.

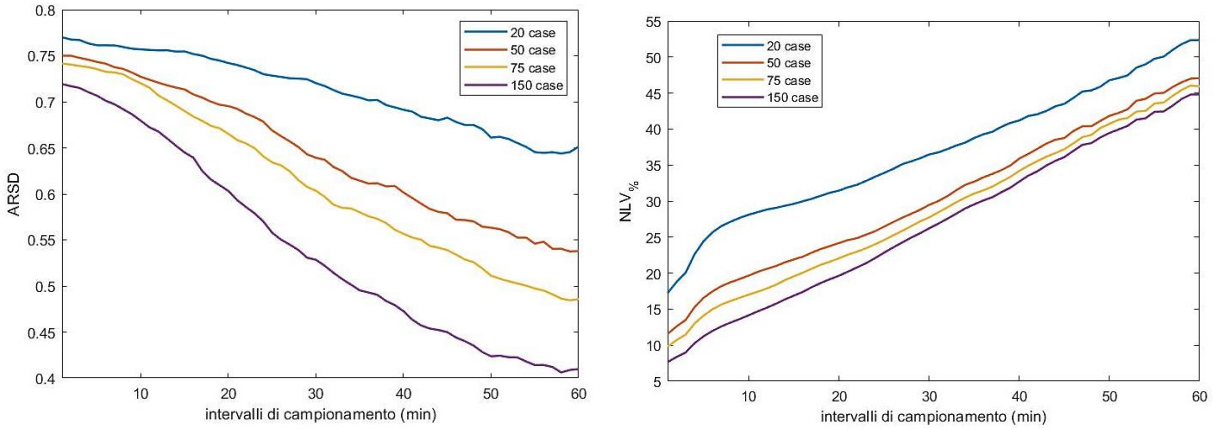


Figura 3.2.4 - Confronto di ARSD e NLV% per utenti residenziali aggregati con diverso livello di aggregazione e intervalli di campionamento

Essi verificano quanto detto, ovvero che per elevati valori del periodo di campionamento e del livello di aggregazione, tutte le osservazioni seguono lo stesso andamento, come indicato dal valore piccolo di *ARSD*. Mentre i valori di *NLV*% mostrano che l'ampiezza delle variazioni di carico aumenta con il crescere del periodo di campionamento ma decresce con l'aumentare con il numero di utenti aggregati. Dunque, più lungo è l'intervallo di campionamento meno possibilità c'è nel seguire la dinamica delle variazioni riducendo in tal modo la possibilità di prendere decisioni utili a influenzare la flessibilità della domanda. Per poter utilizzare in modo efficace le proprietà delle curve aggregate nella valutazione della flessibilità, è importante effettuare una giusta selezione dell'intervallo di campionamento e del livello di aggregazione. In particolare, se si vuole aumentare il numero di utenti aggregati, è necessario ridurre ancor più il tempo di campionamento per ottenere una rappresentazione accurata.

3.3 Analisi probabilistica e statistica delle variazioni di carico aggregato

Per estrarre l'andamento di crescita e decrescita della domanda aggregata e rendere i risultati più generali e facili da confrontare, viene utilizzato un metodo probabilistico basato su modelli di dati categoriali [7]. L'analisi dei dati categoriali o *categorical data analysis* è un tipo di analisi statistica, spesso utilizzata in svariati ambiti scientifici e tecnologici, utile per studiare quella

tipologia di dati costituiti da variabili discrete, o convertiti in tale forma, utilizzata per organizzare le osservazioni in gruppi di comuni caratteristiche. I dati categoriali possono derivare da osservazioni fatte su dati qualitativi, raggruppati in conteggi o tabelle di contingenza, o su dati quantitativi raggruppati entro determinati intervalli. Una variabile categoriale che può assumere solo due valori definiti, è chiamata variabile binaria o dicotomica, diversamente si parla di variabile *politomica*. Un caso importante è la variabile aleatoria di Bernoulli *X*, la quale può assumere solo i valori 0 e 1, presentando una funzione di massa (o densità) di probabilità del tipo:

$$\begin{cases} prob(X = 1) = p \\ prob(X = 0) = 1 - p \end{cases} \tag{3.3.1}$$

dove p rappresenta la probabilità che a seguito di una prova, o esperimento, l'esito sia un "successo" ($X = 1$); ovviamente dovrà essere $0 \leq p \leq 1$.

Esistono tre distribuzioni di probabilità basilari utilizzate per questa tipologia di analisi: la distribuzione binomiale, la distribuzione di Poisson e quella multinomiale. In particolare, la prima è ottenuta realizzando *n* ripetizioni indipendenti della prova di Bernoulli, ciascuna delle quali può concludersi con probabilità di successo p o di insuccesso $1 - p$. Questo tipo di distribuzione risulta essere adatta per rappresentare i cambiamenti della domanda, utilizzando diverse categorie per i diversi stati di incremento o decremento del carico, fornendo le informazioni di probabilità delle occorrenze di ogni categoria nello specifico intervallo di tempo. In realtà, potrebbe esistere una terza casistica in cui non ci sia nessun cambiamento del carico tra due intervalli di tempo consecutivi, significando che le possibili variazioni di domanda tra i due intervalli, risultano essere più piccole della risoluzione in ampiezza, ovvero il livello di accuratezza, dello strumento di misura. Tuttavia, con l'utilizzo di intervalli di tempo di ampiezza più elevata, tale categoria risulta essere trascurabile. Al fine di garantire la correttezza del modello binomiale, essa viene associata alle variazioni negative considerando solo due possibili risultati come risposta delle variabili bernoulliane:

1. Incremento della domanda
2. Non incremento della domanda

In particolare, per un livello di aggregazione *a* e un periodo di campionamento *s*, ad ogni istante di tempo $x\Delta t_s$ vengono eseguite *K* osservazioni, pari al numero di curve aggregate prese in esame (numero di righe di $\Delta\mathbf{P}^{(a,s)}$), ognuna delle quali corrisponde a una prova di Bernoulli con risultato:

$$u_{k,x\Delta t_s}^{(a,s)} = \begin{cases} 1, \ \Delta p_{k,x\Delta t_s}^{(a,s)} > 0 \\ 0, \ altrimenti \end{cases} \tag{3.3.2}$$

Perciò, ad ogni colonna della matrice delle variazioni di carico è associata una variabile aleatoria binomiale $U_{x\Delta t_s}^{(a,s)} \sim Bin\left(K, \ \omega_{x\Delta t_s}^{(a,s)}\right)$ definita con i due parametri caratteristici della distribuzione: *K* che indica il numero di ripetizioni e $\omega_{x\Delta t_s}^{(a,s)}$ la probabilità che un singolo evento della variabile aleatoria bernoulliana registri un successo, ovvero un aumento di domanda.

$$U_{x\Delta t_s}^{(a,s)} = \sum_{k=1}^{K} u_{k,x\Delta t_s}^{(a,s)} \tag{3.3.3}$$

Si noti che la variabile aleatoria binomiale può assumere qualsiasi valore tra 0 e *K*, poiché denota il numero totale di successi della variabile aleatoria bernoulliana che si verificano nelle *K* ripetizioni indipendenti dell'esperimento.

Estendendo le precedenti considerazioni per tutti gli intervalli di tempo del periodo di osservazione delle curve aggregate di carico, si può rappresentare in forma vettoriale (con numero di elementi pari al numero di variazioni di carico $n_s - 1$) i risultati delle variabili aleatorie binomiali $\boldsymbol{U}^{(a,s)}$ contenenti l'informazione sul numero di curve che incrementano la loro domanda e le loro probabilità $\boldsymbol{\omega}^{(a,s)}$, nei rispettivi intervalli di tempo.

$$\boldsymbol{U}^{(a,s)} = \left[U_{2\Delta t_s}^{(a,s)}, \ldots, U_{n_s\Delta t_s}^{(a,s)}\right] \tag{3.3.4}$$

$$\boldsymbol{\omega}^{(a,s)} = \left[\omega_{2\Delta t_s}^{(a,s)}, \ldots, \omega_{n_s\Delta t_s}^{(a,s)}\right] \tag{3.3.5}$$

Finora si è proceduto con una metodologia probabilistica per cui le distribuzioni e quindi i parametri caratteristici delle variabili aleatorie considerate sono note. In realtà, in statistica, il problema centrale è quello di fare dell'inferenza, ovvero di stimare i valori dei parametri incogniti utilizzando i dati osservati. Per cui risulta necessario utilizzare alcune tecniche di stima per poter determinare il valore della probabilità $\omega_{x\Delta t_s}^{(a,s)}$ al momento sconosciuta. Vi è una classe particolare di stimatori, detti di massima verosimiglianza[2] (*MLE*, dall'inglese *maximum likelihood estimator*) che è largamente utilizzata in statistica e in particolare nella stima puntuale dei parametri delle variabili aleatorie binomiali.

[2] Uno stimatore di massima verosimiglianza $\hat{\theta}$ è definito come il valore di θ che rende massima la funzione $f(x_1, x_2, \ldots, x_n|\theta)$, quando i valori osservati sono $x_1, x_2, \ldots, x_n$. La funzione f è detta funzione di *likelihood* e rappresenta la funzione di massa (o densità) congiunta dei valori osservati, mentre θ risulta essere l'incognita del quale si vuole ottenere la stima.

Sia $\widehat{\omega}_{x\Delta t_s}^{(a,s)}$ lo stimatore di massima verosimiglianza del parametro $\omega_{x\Delta t_s}^{(a,s)}$ di una distribuzione binomiale, esso coinciderà con la frazione di prove che hanno avuto successo:[3]

$$\widehat{\omega}_{x\Delta t_s}^{(a,s)} = \frac{1}{\mathrm{K}}\sum_{k=1}^{K} u_{k,x\Delta t_s}^{(a,s)} \;=\; \frac{U_{x\Delta t_s}^{(a,s)}}{\mathrm{K}} \tag{3.3.6}$$

Lo stimatore della probabilità di crescita della domanda $\widehat{\omega}_{x\Delta t_s}^{(a,s)}$, facilmente ricavabile dall'equazione precedente, risulta essere uno stimatore *corretto* poiché il suo valore atteso coincide con il parametro stimato $\mathrm{E}\left(\widehat{\omega}_{x\Delta t_s}^{(a,s)}\right) = \omega_{x\Delta t_s}^{(a,s)}$, e presenta una varianza pari a $Var\left(\widehat{\omega}_{x\Delta t_s}^{(a,s)}\right) = \frac{\omega_{x\Delta t_s}^{(a,s)}\left(1-\omega_{x\Delta t_s}^{(a,s)}\right)}{K}$.

Usando la notazione vettoriale si può riscrivere per tutti gli intervalli della serie temporale:

$$\widehat{\boldsymbol{\omega}}^{(a,s)} = \left[\widehat{\omega}_{2\Delta t_s}^{(a,s)}, \ldots, \widehat{\omega}_{n_s\Delta t_s}^{(a,s)}\right] \tag{3.3.7}$$

In *Figura 3.3.1* è stata riportata tale probabilità, insieme alla sua complementare, riferite al caso che è stato preso come esempio.

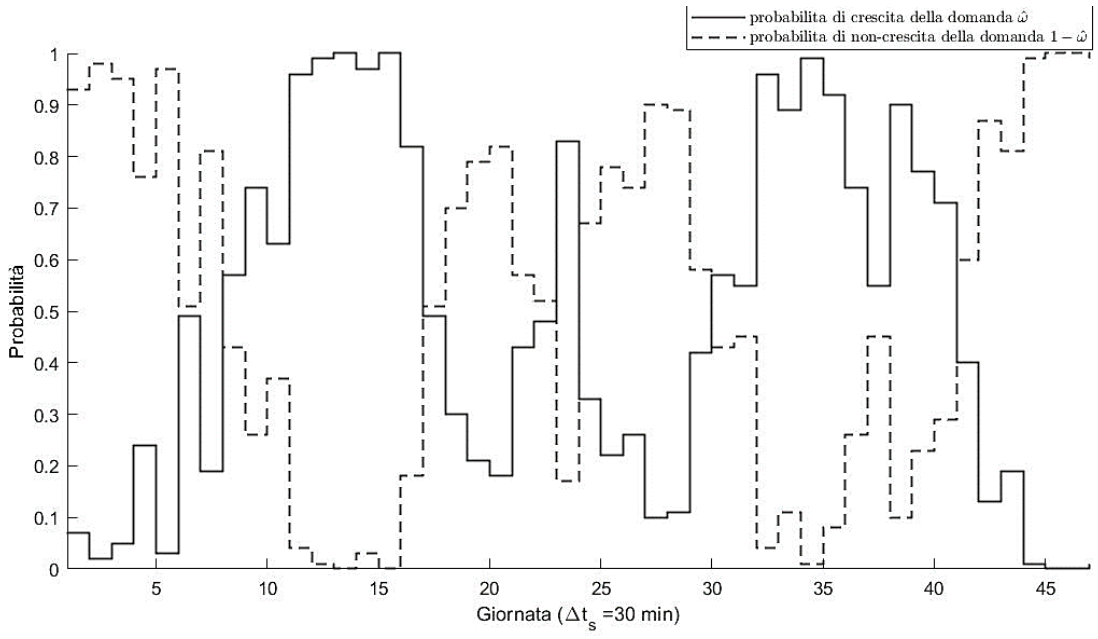


Figura 3.3.1 - Probabilità di crescita e non-crescita della domanda

[3] Per una variabile aleatoria di Bernoulli X con parametro incognito la probabilità di ottenere un successo p, la funzione di likelihood per una serie di campioni ottenuti da n prove indipendenti risulta essere:

$$f(x_1, x_2, \ldots, x_n|\mathrm{p}) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}, \qquad i = 1,2,\ldots,n \qquad con \qquad x_i = \{0,1\}$$

Il valore di p che massimizza tale funzione è: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Il valore della probabilità stimata $\widehat{\omega}_{x\Delta t_s}^{(a,s)}$ dipende dai risultati ottenuti dalle diverse prove effettuate su una determinata popolazione di dati, ed è uguale per ogni *k*-esima prova indipendente effettuata, in quanto associata alla variabile di Bernoulli dell'*x*-esimo intervallo. Tuttavia, se si dovesse effettuare lo stesso ragionamento su una popolazione (o scenario) differente, il risultato potrebbe essere differente, ciò significa che non ci si può aspettare che lo stimatore sia esattamente uguale al parametro effettivo della distribuzione considerata, ma solo che le sarà "vicino". Per valutare il grado di variabilità di una stima puntuale, invece di identificare un solo punto (stimatore puntuale), si preferisce identificare un intervallo (stima intervallare) a cui il parametro da stimare vi appartenga con un certo livello di fiducia. Sia $\hat{\theta}$ lo stimatore puntuale del parametro θ di una variabile aleatoria *X*, e $\left[\, \underline{\theta}, \overline{\theta} \,\right]$ l'intervallo di confidenza che si vuole stimare, gli estremi di tale intervallo sono anch'essi variabili aleatorie in funzione dello stimatore $\underline{\theta}\,(\hat{\theta})$, $\overline{\theta}\,(\hat{\theta})$ tale che:

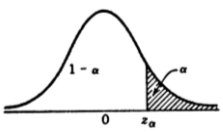
$$prob\left(\theta \in \left[\, \underline{\theta}, \overline{\theta} \,\right]\right) = 1 - \alpha$$

dove $1 - \alpha$ è il livello di confidenza assegnato con $\alpha \in (0,1)$; di solito si utilizza $\alpha = 0.05$ che corrisponde ad un intervallo di confidenza al 95%. Per ricavare il valore degli estremi dell'intervallo bisogna conoscere la distribuzione di probabilità di probabilità di *X* al fine di ricavare la funzione quantile[4] $Z_{\alpha/2} = \mathrm{Q}_X\left(1 - \frac{\alpha}{2}\right)$ per poter imporre che:

$$prob\left(-Z_{\alpha/2} < X < Z_{\alpha/2}\right) = 1 - \alpha$$

Ci sono molti metodi in letteratura per calcolare gli intervalli di confidenza per i parametri della variabile aleatoria binomiale. Il metodo più semplice consiste nell'approssimare la distribuzione di probabilità della variabile aleatoria che si sta considerando in una distribuzione normale, secondo il teorema del limite centrale. Però ciò risulta essere meno accurato quando il numero di prove effettuate (grandezza della popolazione) è troppo basso oppure quando, nel caso di distribuzione binomiale, la probabilità di successi è molto vicina a 0 o 1. Un miglioramento rispetto al metodo classico di approssimazione alla distribuzione nomale, è il *Wilson Score Interval*, utilizzato in molte pubblicazioni di ricerca, in quanto presenta diversi vantaggi, come quello di garantire un'approssimazione dell'intervallo di confidenza uguale a quello normale anche per piccoli valori del numero di prove effettuate e valori di probabilità prossimi agli estremi.

[4] Si definisce *funzione quantile* $Q_X(1-\alpha) = Z_\alpha$ di una distribuzione della variabile aleatoria *X* ad un determinato livello di confidenza $1 - \alpha$, la funzione inversa della *funzione di distribuzione* (o *CDF* dall'inglese *Cumulative Distribution Function*) $F_X(x) = prob(X \leq x)$, tale che: $Q_X(1-\alpha) = F_X^{-1}(1-\alpha) = Z_\alpha$
da cui si ricava $F_X(Z_\alpha) = prob(X \leq Z_\alpha) = 1 - \alpha$; ciò significa che la probabilità che la variabile aleatoria *X* assuma un valore maggiore di Z_α sia esattamente α. Nella seguente figura è rappresentato in forma grafica la definizione di quantile per una distribuzione normale standard.



In seguito, si è utilizzata una versione modificata di Wilson Score Interval come descritto in "*Interval Estimation for a Binomial Proportion"* da parte di L.D. Brown, T.T. Cai and A. DasGupta, ottenendo i seguenti estremi dell'intervallo di confidenza:

$$\left(\overline{\widehat{\omega}^{(a,s)}_{x\Delta t_s}},\ \underline{\widehat{\omega}^{(a,s)}_{x\Delta t_s}} \right) = \widehat{\omega}'^{(a,s)}_{x\Delta t_s} \pm Z\alpha_{/2}\sigma'^{(a,s)}_{x\Delta t_s} \tag{3.3.8}$$

dove:

$$\widehat{\omega}'^{(a,s)}_{x\Delta t_s} = \frac{\widehat{\omega}^{(a,s)}_{x\Delta t_s} + \frac{Z^2_{\alpha/2}}{2K}}{\left(1 + \frac{Z^2_{\alpha/2}}{K}\right)} \tag{3.3.9}$$

in cui

$$\sigma'^{(a,s)}_{x\Delta t_s} = \frac{\sqrt{\frac{\widehat{\omega}^{(a,s)}_{x\Delta t_s}\left(1 - \widehat{\omega}^{(a,s)}_{x\Delta t_s}\right)}{K} + \frac{Z^2_{\alpha/2}}{4K^2}}}{\left(1 + \frac{Z^2_{\alpha/2}}{K}\right)} \tag{3.3.10}$$

che corrispondono rispettivamente ai nuovi valori di media e deviazione standard riassegnati per il Wilson Score Interval, in una distribuzione approssimata normale, alla quale il valore di $Z\alpha_{/2}$ corrisponderà a 1.96 per un livello di confidenza al 95%, secondo le tabelle caratteristiche di tale distribuzione.

Per ogni Δt_s si rappresentano i seguenti vettori di dimensioni $n_s - 1$:

$$\widehat{\boldsymbol{\omega}}'^{(a,s)} = \left[\widehat{\omega}'^{(a,s)}_{2\Delta t_s}, \ldots, \widehat{\omega}'^{(a,s)}_{n_s\Delta t_s} \right] \tag{3.3.11}$$

$$\overline{\widehat{\boldsymbol{\omega}}'^{(a,s)}} = \left[\overline{\widehat{\omega}'^{(a,s)}_{2\Delta t_s}}, \ldots, \overline{\widehat{\omega}'^{(a,s)}_{n_s\Delta t_s}} \right] \tag{3.3.12}$$

$$\underline{\widehat{\boldsymbol{\omega}}'^{(a,s)}} = \left[\underline{\widehat{\omega}'^{(a,s)}_{2\Delta t_s}}, \ldots, \underline{\widehat{\omega}'^{(a,s)}_{n_s\Delta t_s}} \right] \tag{3.3.13}$$

Gli intervalli di confidenza sono molto utili in quanto indicano il livello di incertezza o variabilità dell'incremento o abbassamento del carico. Infatti, valori inferiori dell'ampiezza dell'intervallo suggeriscono che si ha un andamento più regolare sulla crescita e decrescita del carico nel periodo considerato.

3.4 Indicatori di flessibilità per carichi residenziali aggregati

Per l'operatore del sistema, è importante avere la possibilità di cambiare l'andamento dei consumi in un gruppo di utenti, al fine di poter gestire al meglio le curve di carico aggregato all'interno di un sistema, ma ciò dipende dal livello di flessibilità dei singoli utenti a collaborare in programmi specifici. A tale scopo vengono proposti in letteratura diversi indicatori di flessibilità della domanda aggregata [1, 8, 9].

Considerando i valori medi $\bar{p}_{x\Delta t_s}^{(a,s)}$ e $\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ dei vettori colonna di $\boldsymbol{p}_{x\Delta t_s}^{(a,s)}$ e $\Delta\boldsymbol{p}_{x\Delta t_s}^{(a,s)}$ per ogni intervallo, riportati in *Figura 3.4.1*, si può scrivere, a partire dalla definizione di variazione di carico vista nell'equazione (3.2.2), la seguente relazione tra i valori medi delle grandezze:

$$\bar{p}_{x\Delta t_s}^{(a,s)} = \bar{p}_{(x-1)\Delta t_s}^{(a,s)} + \overline{\Delta p}_{x\Delta t_s}^{(a,s)} \qquad per\ x = 2, \dots, n_s \tag{3.4.1}$$

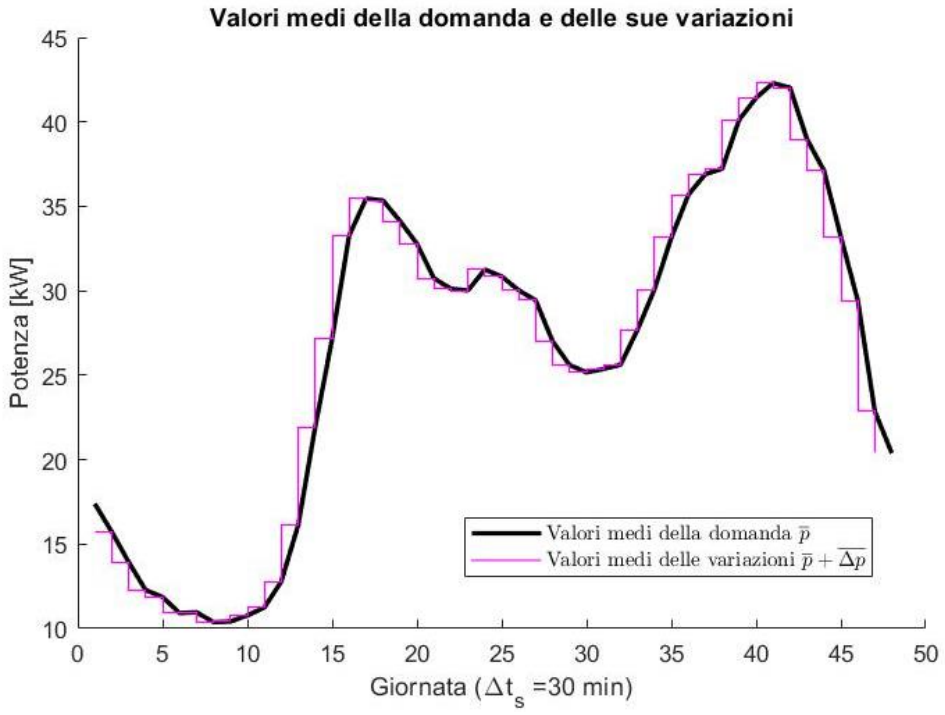


Figura 3.4.1 – Valori medi della domanda aggregata e delle sue variazioni

Se anziché considerare tutte le variazioni di potenza delle K curve in ogni intervallo $x\Delta t_s$, si considerano in maniera separata le variazioni positive da una parte e quelle non positive dall'altra, ricavando rispettivamente i loro valori medi ${}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ e ${}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$, è possibile esprimere $\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ in termini di crescita e non-crescita del carico, tramite la definizione di media aritmetica pesata:

$$\overline{\Delta p}_{x\Delta t_s}^{(a,s)} = \left(\frac{U_{x\Delta t_s}^{(a,s)}}{K}\right) {}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)} + \left(\frac{1 - U_{x\Delta t_s}^{(a,s)}}{K}\right) {}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)} \tag{3.4.2}$$

dove ai valori medi delle variazioni di carico quando la domanda è crescente e quando non lo è, si attribuiscono i pesi delle rispettive frequenze nelle K osservazioni. Essi corrispondono anche agli

stimatori di massima verosimiglianza, per le distribuzioni binomiali, delle probabilità che avvenga o meno l'incremento del carico secondo quanto visto nell'equazione (3.3.8) e nella loro riassegnazione per gli intervalli di Wilson Score Interval in (3.3.9).
In questi termini, è possibile riscrivere l'equazione come segue:

$$\overline{\Delta p}_{x\Delta t_s}^{(a,s)} = \left(\widehat{\omega}'^{(a,s)}_{x\Delta t_s}\right) {}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)} + \left(1 - \widehat{\omega}'^{(a,s)}_{x\Delta t_s}\right) {}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)} \tag{3.4.3}$$

Un'importante ipotesi potrebbe essere introdotta assumendo che il comportamento medio delle variazioni di domanda positive e negative non cambia con il livello di aggregazione e che quindi, un aumento o un abbassamento delle probabilità binomiali di crescita e decrescita della domanda non comporta variazioni in ${}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ e ${}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$. Ciò significa che $\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ dipende solo da $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$, e che ad un eventuale aumento di utenti che incrementano il carico segue un aumento della probabilità $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$, ma non varia il valore medio delle variazioni positive ${}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$.

Dall'equazione (3.4.3) si possono fare alcune considerazioni riguardo il margine di flessibilità della domanda $\bar{p}_{x\Delta t_s}^{(a,s)}$ attraverso i valori che può assumere $\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ al variare del valore di probabilità di incremento $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$, per ogni intervallo di tempo considerato. Infatti, si nota che il massimo valore positivo corrisponde alla condizione ideale per cui $\overline{\Delta p}_{x\Delta t_s}^{(a,s)} = {}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$, in corrispondenza di un valore unitario della probabilità $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$ (certezza che tutti gli utenti aumentano il carico), poiché il secondo termine dell'equazione (3.4.3) può assumere esclusivamente valori negativi o nulli. Analogamente, il valore minimo negativo di $\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ si ha quando la probabilità di non crescita della domanda è unitaria (cioè per $\widehat{\omega}'^{(a,s)}_{x\Delta t_s} = 0$) e corrisponderà al valore di ${}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$.

Per ricavare la quantità di domanda aggregata flessibile, si può valutare la differenza, in un x-esimo intervallo di tempo, tra la domanda media effettiva $\bar{p}_{x\Delta t_s}^{(a,s)}$ e un altro valore di domanda media $\bar{p}_{x\Delta t_s}^{(a,s)^*}$ avente una probabilità di incremento pari a $\widehat{\omega}'^{(a,s)^*}_{x\Delta t_s}$, dall'equazione (3.4.3) sostituita in (3.4.1) :

$$\bar{p}_{x\Delta t_s}^{(a,s)} - \bar{p}_{x\Delta t_s}^{(a,s)^*} = \left(\widehat{\omega}'^{(a,s)}_{x\Delta t_s} - \widehat{\omega}'^{(a,s)^*}_{x\Delta t_s}\right) {}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)} + \left(\widehat{\omega}'^{(a,s)^*}_{x\Delta t_s} - \widehat{\omega}'^{(a,s)}_{x\Delta t_s}\right) {}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)} \tag{3.4.4}$$

dalla quale si può notare che qualunque variazione di $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$ determina un cambiamento opposto nelle variazioni medie di crescita ${}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ e non crescita ${}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$della domanda, contribuendo ad un raddoppio nel cambiamento della domanda aggregata, in quanto i due termini dell'equazione assumono lo stesso segno per qualsiasi valore di $\widehat{\omega}'^{(a,s)^*}_{x\Delta t_s}$.

Flexibility Index Aggregate Demand (FIAD)

Per poter definire il primo indicatore di flessibilità di domanda residenziale aggregata, si calcola il minimo tra le probabilità binomiali di crescita o non crescita della domanda, in ogni intervallo di tempo del periodo di osservazione dell'andamento:

$$\boldsymbol{\pi}^{(a,s)} = \min_{\forall \hat{\omega}'^{(a,s)}_{x\Delta t_s}} \left(\hat{\boldsymbol{\omega}}'^{(a,s)}, 1 - \hat{\boldsymbol{\omega}}'^{(a,s)} \right) \tag{3.4.5}$$

Ogni x-esimo elemento del vettore $\boldsymbol{\pi}^{(a,s)}$ è definito in un intervallo compreso tra 0 e 0.5, in quanto, essendo calcolato con il minimo tra due grandezze complementari definite tra 0 e 1, esso assume il valore massimo quando entrambe le probabilità sono pari a 0.5 e il minimo quando una delle due è nulla. In *Figura 3.4.2* sono messe a confronto le varie probabilità, che risultano simmetriche rispetto al valore 0.5, confermando ciò che è stato detto precedentemente, ovvero che al cambiamento della probabilità $\hat{\omega}'^{(a,s)}_{x\Delta t_s}$ consegue il cambiamento di $1 - \hat{\omega}'^{(a,s)}_{x\Delta t_s}$, fornendo un doppio contributo sulle variazioni della domanda aggregata secondo l'equazione (3.4.4). Si può inoltre notare che $\pi^{(a,s)}_{x\Delta t_s}$ fornisce un'informazione sullo scarto di probabilità necessario per arrivare alla condizione ottimale di $\hat{\omega}'^{(a,s)}_{x\Delta t_s}$ uguale a 0 o 1, nella quale si raggiunge il margine di flessibilità della domanda $\bar{p}^{(a,s)}_{x\Delta t_s}$. Il motivo per cui si è scelto di considerare il minimo tra le probabilità, è che risulta più semplice azzerare quella con valore inferiore, avendo uno scarto minore per essere annullata e di conseguenza portare al valore massimo la sua complementare con più facilità.

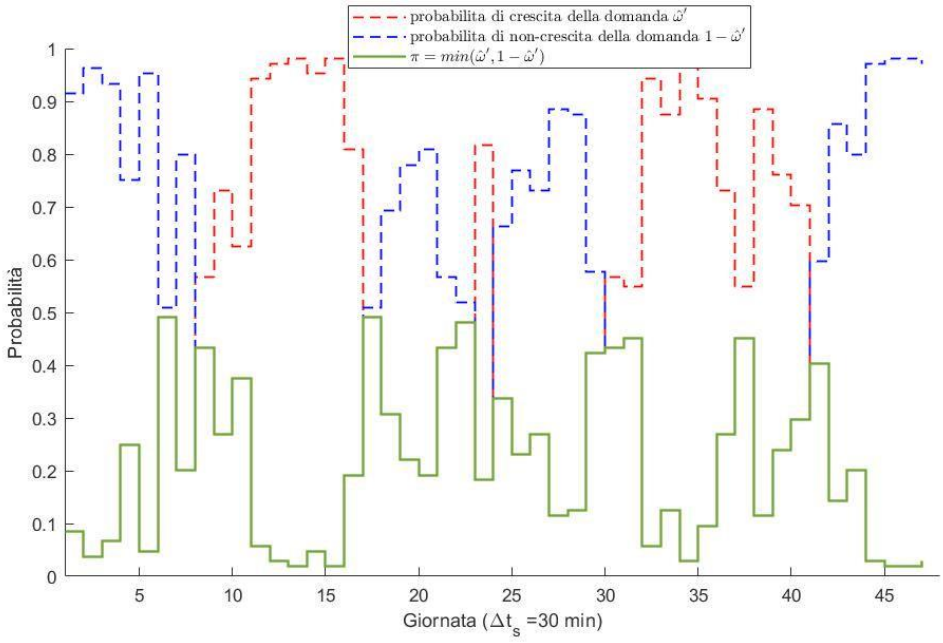


Figura 3.4.2 - Probabilità di crescita e non-crescita della domanda e indicatore $\pi^{(a,s)}_{x\Delta t_s}$

Per tenere conto del doppio contributo di $\boldsymbol{\pi}^{(a,s)}$ nella flessibilità della domanda, e per rendere la definizione dell'indicatore più intuitiva, assumendo valori compresi tra 0 e 1, si preferisce

utilizzare un fattore 2 (uguale al numero di categorie della distribuzione di probabilità binomiale) come segue:

$$FIAD \;=\; \boldsymbol{\varphi}^{(a,s)} \;=\; 2 \text{ x } \boldsymbol{\pi}^{(a,s)} \tag{3.4.6}$$

L'indicatore di flessibilità della domanda aggregata per carichi residenziali denominato *FIAD* (*Flexibility Index of Aggregate Demand*) è definito per ogni elemento $\varphi_{x\Delta t_s}^{(a,s)} \in [0,1]$ del periodo di osservazione delle curve di carico prese in esame. Per questo indicatore valgono le stesse considerazioni fatte per $\boldsymbol{\pi}^{(a,s)}$, ovvero quando $\varphi_{x\Delta t_s}^{(a,s)}$ è prossimo al suo valore massimo, significa che si ha circa la stessa probabilità che ci possa essere un incremento o abbassamento della domanda, e ciò indica che, per quell'intervallo di tempo considerato, i singoli utenti si comportano in maniera molto differente, evidenziando un comportamento complessivo meno rigido e più casuale, cosicché non emerge nessun andamento collettivo e il singolo utente si potrebbe trovare più propenso ad accettare cambiamenti nei suoi consumi. Contrariamente, valori bassi di $\varphi_{x\Delta t_s}^{(a,s)}$ indicano un minore livello di flessibilità, in quanto uno dei due valori di probabilità tende all'unità, per cui si ha un comportamento comune degli utenti che porta ad un irrigidimento della domanda aggregata, limitando così la possibilità di accettare cambiamenti da parte degli utenti. Dunque, la flessibilità valutata con questo indicatore è determinata in termini di probabilità a cambiare la domanda aggregata dovuta al comportamento dei singoli utenti.

Le relazioni finora introdotte fanno riferimento ai valori puntuali delle probabilità binomiali $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$ e $1 - \widehat{\omega}'^{(a,s)}_{x\Delta t_s}$, ad esse però sono associati gli intervalli di confidenza calcolati nel paragrafo precedente. Di seguito sono riportati gli estremi superiori e inferiori di tali intervalli, per gli indici $\boldsymbol{\pi}^{(a,s)}$e $\boldsymbol{\varphi}^{(a,s)}$.

A seconda del minimo selezionato in ogni *x*-esimo elemento del vettore $\boldsymbol{\pi}^{(a,s)}$, si ha:

$$\overline{\pi_{x\Delta t_s}^{(a,s)}} = \begin{cases} \overline{\widehat{\omega}_{x\Delta t_s}^{(a,s)}} & se\ \ \pi_{x\Delta t_s}^{(a,s)} = \widehat{\omega}'^{(a,s)}_{x\Delta t_s} \\ 1 - \overline{\widehat{\omega}_{x\Delta t_s}^{(a,s)}} & se\ \pi_{x\Delta t_s}^{(a,s)} = 1 - \widehat{\omega}'^{(a,s)}_{x\Delta t_s} \end{cases} \tag{3.4.7}$$

$$\underline{\pi_{x\Delta t_s}^{(a,s)}} = \begin{cases} \underline{\widehat{\omega}_{x\Delta t_s}^{(a,s)}} & se\ \ \pi_{x\Delta t_s}^{(a,s)} = \widehat{\omega}'^{(a,s)}_{x\Delta t_s} \\ 1 - \underline{\widehat{\omega}_{x\Delta t_s}^{(a,s)}} & se\ \pi_{x\Delta t_s}^{(a,s)} = 1 - \widehat{\omega}'^{(a,s)}_{x\Delta t_s} \end{cases} \tag{3.4.8}$$

Si noti che l'estremo superiore potrebbe assumere valori superiori a 0.5, infatti nonostante $\pi_{x\Delta t_s}^{(a,s)}$ selezioni un valore di probabilità massimo pari a 0.5, non è detto che il bordo superiore della probabilità selezionata sia limitato a tale valore, poiché definito nell'intervallo tra 0 e 1, come si può vedere in *Figura 3.4.3.* È necessario, dunque, per la definizione degli intervalli di confidenza

dell'indicatore *FIAD*, considerare come massimo valore del bordo superiore 0.5, in maniera tale da non far eccedere l'unità:

$$\overline{\boldsymbol{\varphi}^{(a,s)}} = 2 \text{ x } min\left(0.5, \overline{\boldsymbol{\pi}^{(a,s)}}\right) \tag{3.4.9}$$

$$\underline{\boldsymbol{\varphi}^{(a,s)}} = 2 \text{ x } \underline{\boldsymbol{\pi}^{(a,s)}} \tag{3.4.10}$$

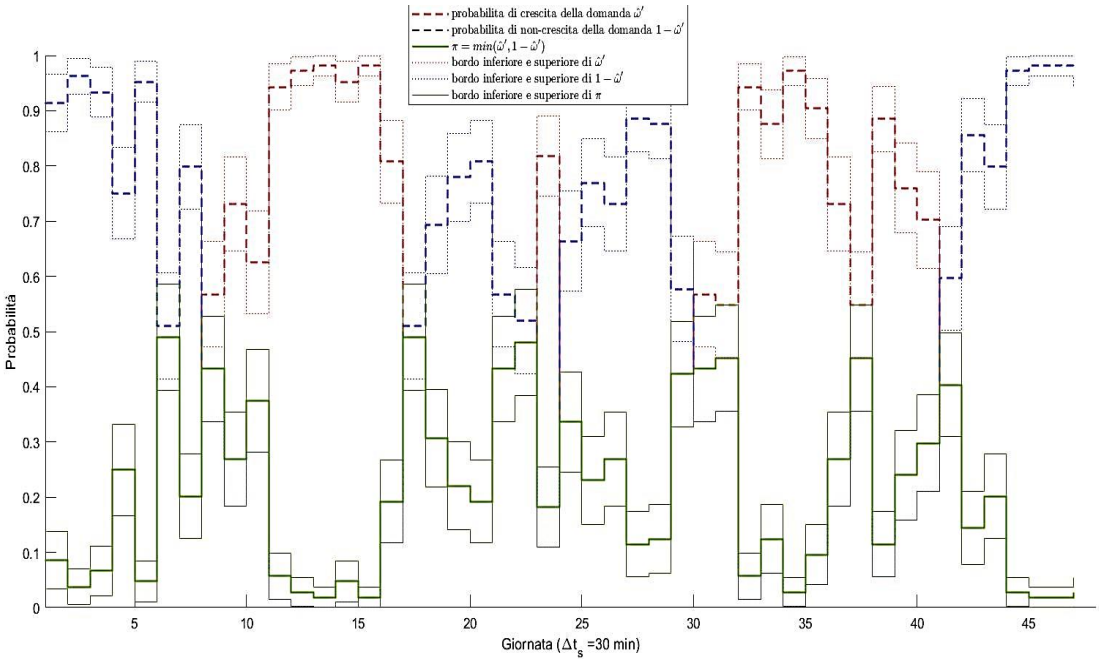


Figura 3.4.3 - Probabilità di crescita e non-crescita della domanda e indicatore $\pi_{x\Delta t_s}^{(a,s)}$ e i rispettivi intervalli di confidenza

In *Figura 3.4.4* è riportato l'indicatore FIAD con i propri intervalli di confidenza.

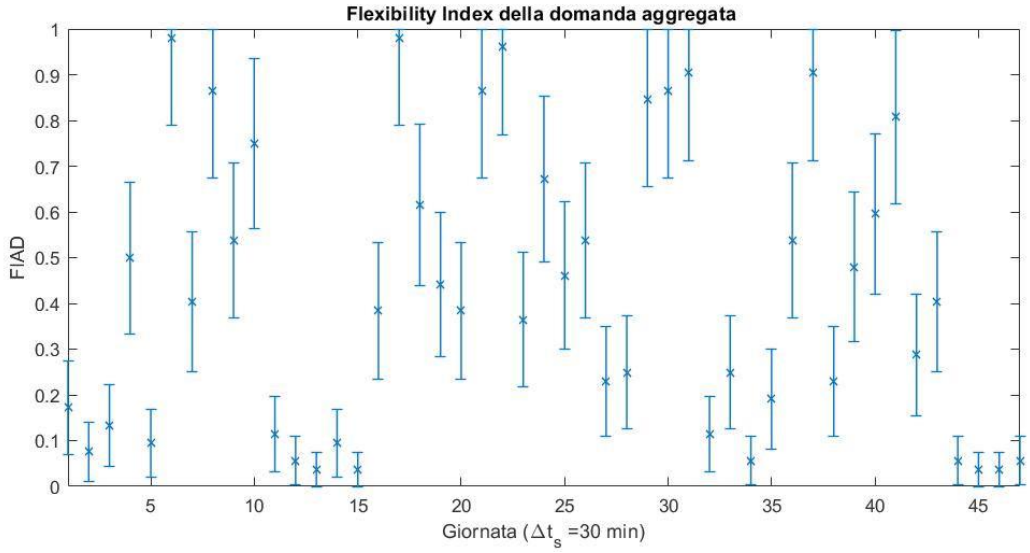


Figura 3.4.4 – FIAD con i rispettivi intervalli di confidenza

Avere elevati valori dell'indicatore *FIAD* non sempre garantisce la possibilità di effettuare cambiamenti importanti sulla curva di carico aggregata. Infatti, supponendo di avere un valore di $\varphi_{x\Delta t_s}^{(a,s)}$ prossimo al suo valore massimo, ovvero quando le probabilità binomiali delle due categorie sono simili, se si hanno piccole variazioni di ${}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ e di ${}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$, la quantità di domanda flessibile $\bar{p}_{x\Delta t_s}^{(a,s)} - \bar{p}_{x\Delta t_s}^{(a,s)^*}$ risulta piccola per poter sfruttare al meglio i programmi di DR. Al contrario, in presenza di notevoli variazioni medie della domanda aggregata, avere un valore di $\varphi_{x\Delta t_s}^{(a,s)}$ che tende all'unità, aumenta la possibilità di trarre benefici ragionevoli da tali programmi. Dunque, questo indicatore fornisce un'interpretazione probabilistica sul comportamento collettivo del carico aggerato, assegnando per ogni intervallo dell'andamento temporale un indice della possibilità di crescita o non crescita della domanda, accettando le dovute incertezze tramite gli intervalli di confidenza, ma non valuta la flessibilità in termini quantitativi della domanda.

Per avere un'idea sulla massima flessibilità ottenibile, e rendere più utile l'indicatore *FIAD* nello sviluppo dei programmi di DR da parte dell'operatore o dell'aggregatore, si introduce una versione modificata di tale indicatore, con l'aggiunta del termine $\boldsymbol{\varepsilon}^{(a,s)}$, il quale fornisce l'informazione sulle variazioni di carico positive o negative, per ogni intervallo di tempo $x\Delta t_s$, a seconda della seguente relazione:

$$MFIAD = \boldsymbol{\phi}^{(a,s)} = \boldsymbol{\varphi}^{(a,s)} \mathrm{x}\ \boldsymbol{\varepsilon}^{(a,s)} \tag{3.4.11}$$

con

$$\boldsymbol{\varepsilon}^{(a,s)} = \max_{\forall \widehat{\omega}'^{(a,s)}_{x\Delta t_s}} \left(\frac{{}^{+}\overline{\Delta \mathbf{p}}^{(a,s)}}{{}^{+}\overline{\Delta \mathbf{p}}^{(a,s)} - {}^{-}\overline{\Delta \mathbf{p}}^{(a,s)}}, \frac{{}^{-}\overline{\Delta \mathbf{p}}^{(a,s)}}{{}^{+}\overline{\Delta \mathbf{p}}^{(a,s)} - {}^{-}\overline{\Delta \mathbf{p}}^{(a,s)}} \right) \tag{3.4.12}$$

dove ${}^{+}\overline{\Delta \mathbf{p}}^{(a,s)}$, ${}^{-}\overline{\Delta \mathbf{p}}^{(a,s)}$ sono i vettori dei valori medi variazioni di crescita e non crescita della domanda, di dimensioni pari a $n_s - 1$, per tutto l'andamento temporale.

Il contributo di $\boldsymbol{\varepsilon}^{(a,s)}$, ovvero del massimo valore tra le due variazioni di carico, nell'equazione del *FIAD*, permette di annullare l'ipotesi fatta inizialmente riguardo i valori medi delle variazioni di carico positive e negative, le quali sono state assunte costanti, indipendentemente dalle probabilità. L'indicatore $MFIAD$ (*Modified Flexibility Index of Aggregate Demand*) è reso applicabile in modo più generale, considerando la possibile dipendenza di ${}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ e di ${}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ dalla probabilità di crescita (e non crescita) della domanda $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$ da specifici gruppi di utenti. Si nota che, in questo caso, il massimo valore dell'indicatore, anche esso definito tra 0 e 1, è raggiungibile se entrambi i termini sono unitari. In particolare, il termine aggiuntivo $\varepsilon_{x\Delta t_s}^{(a,s)}$ assume il valore unitario, nella condizione ideale in cui si hanno solamente variazioni positive oppure solo quelle negative, per cui tutti gli utenti hanno un comportamento crescente o decrescente, nell'intervallo di tempo considerato. In contrapposizione, come è stato già fatto presente, l'indicatore $\varphi_{x\Delta t_s}^{(a,s)}$ otterrà il massimo valore quando le due probabilità complementari di crescita e non crescita della domanda, sarà pari a 0.5, ovvero quando la metà degli utenti sta aumentando o diminuendo i loro consumi. Risulta, perciò, molto difficile che l'indicatore *MFIAD* raggiunga il valore unitario, però potrebbe

andarci vicino se il *FIAD* è prossimo al suo massimo valore e l'ampiezza delle variazioni di domanda ${}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ (oppure ${}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$) tende a valori nulli (possibile se il periodo di campionamento Δt_s è molto piccolo), ovvero quando gli utenti che si comportano in maniera molto casuale presentano dei valori medi delle variazioni negative dominanti su quelle positive (o viceversa).

In *Figura 3.4.5* sono confrontati i valori di *FIAD* e *MFIAD* in ogni intervallo di tempo, risultando il secondo sempre inferiore rispetto al primo.

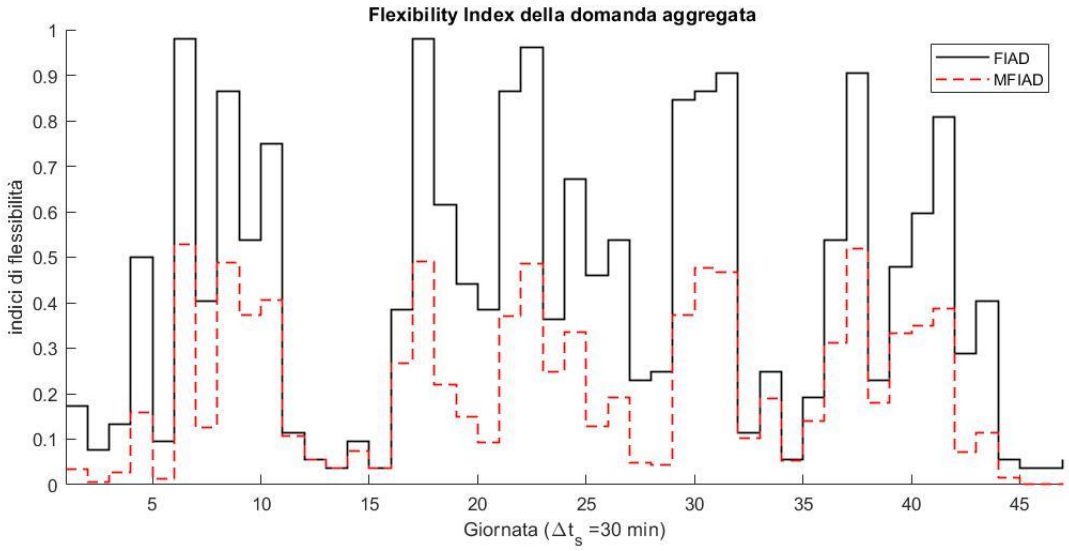


Figura 3.4. 5 – Confronto tra FIAD e MFIAD

Percentage Flexibility Level (PFL)

Per poter valutare la quantità di domanda flessibile, si utilizza l'indicatore *PFL* (*Percentage Flexibility Level*), il quale esprime la percentuale di tale domanda, associata al valore di $\varphi_{x\Delta t_s}^{(a,s)}$ corrispondente[5]:

$$PFL = \boldsymbol{\psi}^{(a,s)} = \frac{{}^{+}\overline{\boldsymbol{\Delta p}}^{(a,s)} - {}^{-}\overline{\boldsymbol{\Delta p}}^{(a,s)}}{\overline{\boldsymbol{p}}^{(a,s)}} \left(\frac{\boldsymbol{\varphi}^{(a,s)}}{2} \right) \text{x } 100 \qquad (3.4.13)$$

Esso, infatti, indica la percentuale di domanda aggregata che è possibile aumentare o ridurre senza che si influenzi il cambiamento medio del gruppo di utenti nella domanda. Per effettuare una variazione della domanda da $\bar{p}_{x\Delta t_s}^{(a,s)}$ a $\bar{p}_{x\Delta t_s}^{(a,s)^*}$ senza modificare il valore di ${}^{+}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$ e di ${}^{-}\overline{\Delta p}_{x\Delta t_s}^{(a,s)}$,

[5] Una versione modificata di *PFL* può essere realizzata sostituendo l'indicatore *FIAD* $\boldsymbol{\varphi}^{(a,s)}$ con la sua versione generalizzata *MFIAD* $\boldsymbol{\phi}^{(a,s)}$:

$$PFL_{MFIAD} = \frac{{}^{+}\overline{\boldsymbol{\Delta p}}^{(a,s)} - {}^{-}\overline{\boldsymbol{\Delta p}}^{(a,s)}}{\overline{\boldsymbol{p}}^{(a,s)}} \left(\frac{\boldsymbol{\phi}^{(a,s)}}{2} \right) \text{x } 100$$

bisogna far cambiare il comportamento degli utenti, in modo tale da variare il valore di $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$ in $\widehat{\omega}'^{(a,s)^*}_{x\Delta t_s}$ e come conseguenza anche i loro complementari. In particolare, valori elevati di *PFL* si ottengono se tutti gli utenti cambiassero singolarmente il loro comportamento dall'aumentare la domanda nel diminuirla e viceversa, ciò è possibile solo quando la curva aggregata non risulta molto rigida, in cui si hanno elevati valori di $\varphi^{(a,s)}_{x\Delta t_s}$ e quindi le probabilità sono lontane dai loro valori estremi.

Il *PFL* tiene conto di tutto il margine di flessibilità, considerando sia la possibile crescita che decrescita nella domanda. Per ottenere informazioni separate, si considerano i seguenti indicatori, i quali rappresentano la massima percentuale del livello di flessibilità nelle variazioni rispettivamente di crescita e di decrescita nella domanda, ottenuta nei casi ideali in cui si hanno tutte le variazioni positive o tutte negative.

$$^{+}\boldsymbol{\psi}^{(a,s)} = \frac{^{+}\overline{\boldsymbol{\Delta \mathbf{p}}}^{(a,s)} - {}^{-}\overline{\boldsymbol{\Delta \mathbf{p}}}^{(a,s)}}{\overline{\boldsymbol{p}}^{(a,s)}} \left(1 - \widehat{\boldsymbol{\omega}}'^{(a,s)}\right) \text{ x } 100 \qquad (3.4.14)$$

$$^{-}\boldsymbol{\psi}^{(a,s)} = \frac{^{+}\overline{\boldsymbol{\Delta \mathbf{p}}}^{(a,s)} - {}^{-}\overline{\boldsymbol{\Delta \mathbf{p}}}^{(a,s)}}{\overline{\boldsymbol{p}}^{(a,s)}} \left(\widehat{\boldsymbol{\omega}}'^{(a,s)}\right) \text{ x } 100 \qquad (3.4.15)$$

Come è stato discusso precedentemente per l'equazione (3.4.3), il massimo margine positivo di flessibilità, e quindi del valore di $^{+}\psi^{(a,s)}_{x\Delta t_s}$, si ottiene nella condizione ideale in cui $\widehat{\omega}'^{(a,s)^*}_{x\Delta t_s} = 1$ richiedendo una variazione della probabilità $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$ pari a $1 - \widehat{\omega}'^{(a,s)}_{x\Delta t_s}$. Nell'equazione (3.4.13), viene sostituito il termine $\left(\frac{\boldsymbol{\varphi}^{(a,s)}}{2}\right)$ con tale cambiamento richiesto alla probabilità, in modo da ottenere il massimo livello di flessibilità nell'equazione (3.4.14). Dunque, la massima percentuale di domanda che può essere incrementata si ha quando la probabilità che ci sia un incremento $\widehat{\omega}'^{(a,s)}_{x\Delta t_s}$ è nulla, avendo in tal caso il massimo margine di variazione positivo e quindi di cambiamento dalla decrescita alla crescita della domanda. Le stesse considerazioni possono essere estese per l'indicatore riferito alla variazione negativa, tanto più grande è la probabilità che il carico stia aumentando, in quell'intervallo di tempo considerato, tanto maggiore sarà la quota di domanda che si potrà ridurre. Tuttavia, questi due indicatori forniscono indicazioni sulla variabilità del carico piuttosto che sulla flessibilità che può essere ottenuta dal carico aggregato, il quale è soggetto al comportamento collettivo degli utenti.

In *Figura 3.4.6* sono messi a confronto l'indicatori $\boldsymbol{\psi}^{(a,s)}$, ${}^{-}\boldsymbol{\psi}^{(a,s)}$ e ${}^{+}\boldsymbol{\psi}^{(a,s)}$.

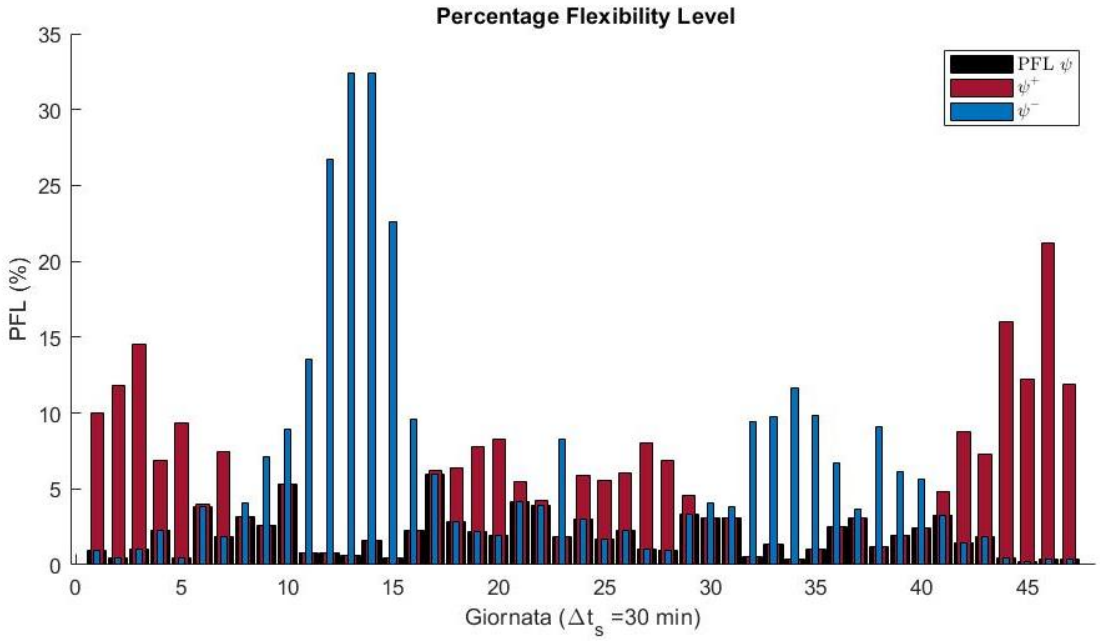


Figura 3.4 6 – PFL, ${}^{-}\boldsymbol{\psi}^{(a,s)}$ e ${}^{+}\boldsymbol{\psi}^{(a,s)}$

In *Figura 3.4.7* e in *Figura 3.4.8* sono riportate rispettivamente la quantità di domanda flessibile e i valori medi della domanda aggregata senza e con l'aggiunta della quota flessibile.

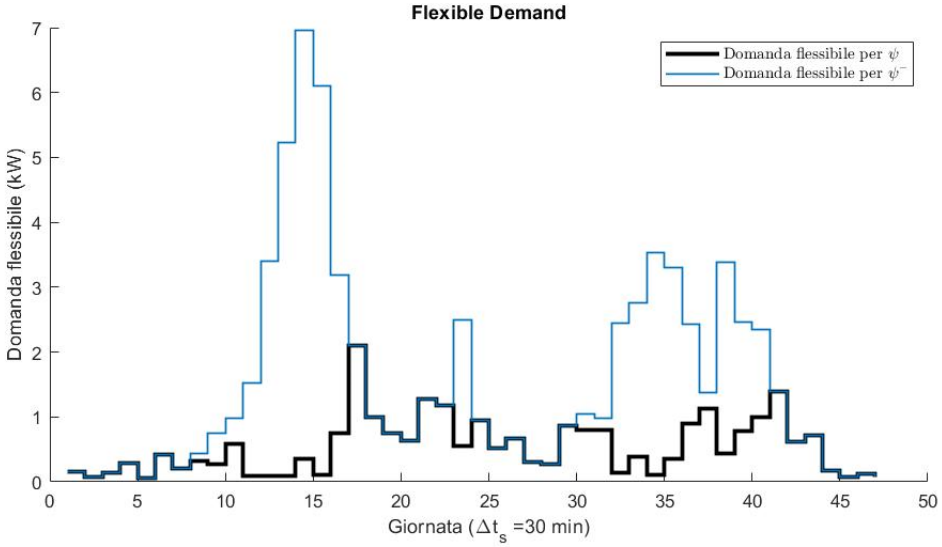


Figura 3.4.7 – Quantità di domanda flessibile con l'indicatore PLF e ${}^{-}\boldsymbol{\psi}^{(a,s)}$

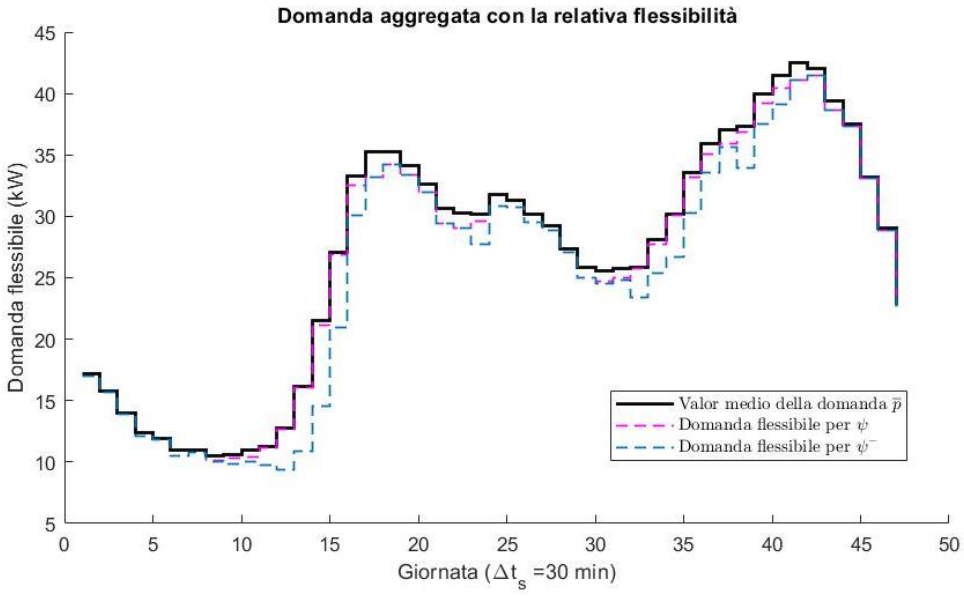


Figura 3.4 8– Curva di carico aggregato media, con e senza la quantità di domanda flessibile

Anche per gli indicatori $\boldsymbol{\psi}^{(a,s)}$, ${}^{-}\boldsymbol{\psi}^{(a,s)}$ e ${}^{+}\boldsymbol{\psi}^{(a,s)}$, possono essere calcolati i rispettivi intervalli di confidenza sostituendo alle loro definizioni (3.4.13), (3.4.14), (3.4.15), rispettivamente i valori dei bordi superiori e inferiori di $\boldsymbol{\varphi}^{(a,s)}$ e di $\widehat{\boldsymbol{\omega}}'^{(a,s)}$ definiti nelle equazioni (3.4.9), (3.4.10), (3.3.10) e (3.3.11).

Il *PFL* è un indicatore molto utile all'aggregatore o all'operatore del sistema elettrico per osservare i limiti di flessibilità in termini percentuali. Valori elevati di *PFL*, in alcuni intervalli di tempo considerati, suggeriscono che c'è una possibilità notevole di ridurre o aumentare la domanda aggregata in tali intervalli, in quanto gli utenti sono liberi ad accettare cambiamenti nei loro consumi. Contrariamente, valori di percentuale inferiore, indicano che la curva di carico aggregata è abbastanza rigida da impedire cambiamenti sui consumi dei singoli utenti. Inoltre, se si conosce il valore del prezzo dell'energia elettrica nel periodo di tempo osservato, è possibile valutare l'efficacia dei programmi di DR, qualora si avesse un'elevata flessibilità per far spostare i consumi da periodi in cui il prezzo è maggiore in quelli in cui è minore.

3.5 Caso studio

A partire dai dataset pre-elaborati nel capitolo 2, per svolgere le analisi di flessibilità sulle curve di carico, è stato scelto lo *Smart* Home Dataset* per l'anno 2016, in quanto presenta un buon intervallo di campionamento pari a un minuto, un numero abbastanza elevato di abitazioni (114 case) con misure che si estendono per quasi un intero anno (347 giorni completi), permettendo, in tal modo, di svolgere analisi sulla flessibilità in diversi minuti della giornata, diverse stagioni o in giorni anomali. Per effettuare un'analisi statistica corretta e poter valutare l'incertezza e la casualità negli andamenti delle variazioni di carico, è necessario costruire un certo numero di curve di carico aggregate a partire dallo stesso dataset preso in esame. Tuttavia, se si utilizzassero i dati senza nessuna elaborazione, a causa dei differenti comportamenti tra tutti gli utenti, non si riuscirebbero a osservare delle tendenze rilevanti per i carichi in periodi di tempo coerenti. È

necessario, dunque, isolare tali carichi e identificare i gruppi di candidati per i programmi di DR. Una volta effettuate alcune considerazioni sui dati, eseguita una cluster analysis e individuati i gruppi di utenti adatti per la DR, sono state costruite le curve di carico aggregate su tali gruppi, al fine di andare a valutare la flessibilità della domanda. L'analisi è stata svolta per diversi livelli di aggregazione e intervalli temporali di campionamento, valutando le differenze che si hanno sia sulle variazioni di carico che sugli indicatori di flessibilità.

Interpretazione e organizzazione dei dati

Dopo aver svolto le classiche procedure di pulizia e ricostruzione dei dati viste nel capitolo precedente, è opportuno osservare e interpretare i dati che si stanno analizzando. In *Figura 3.5.1* sono riportati gli andamenti annuali di tutti gli utenti, evidenziano il loro comportamento medio.

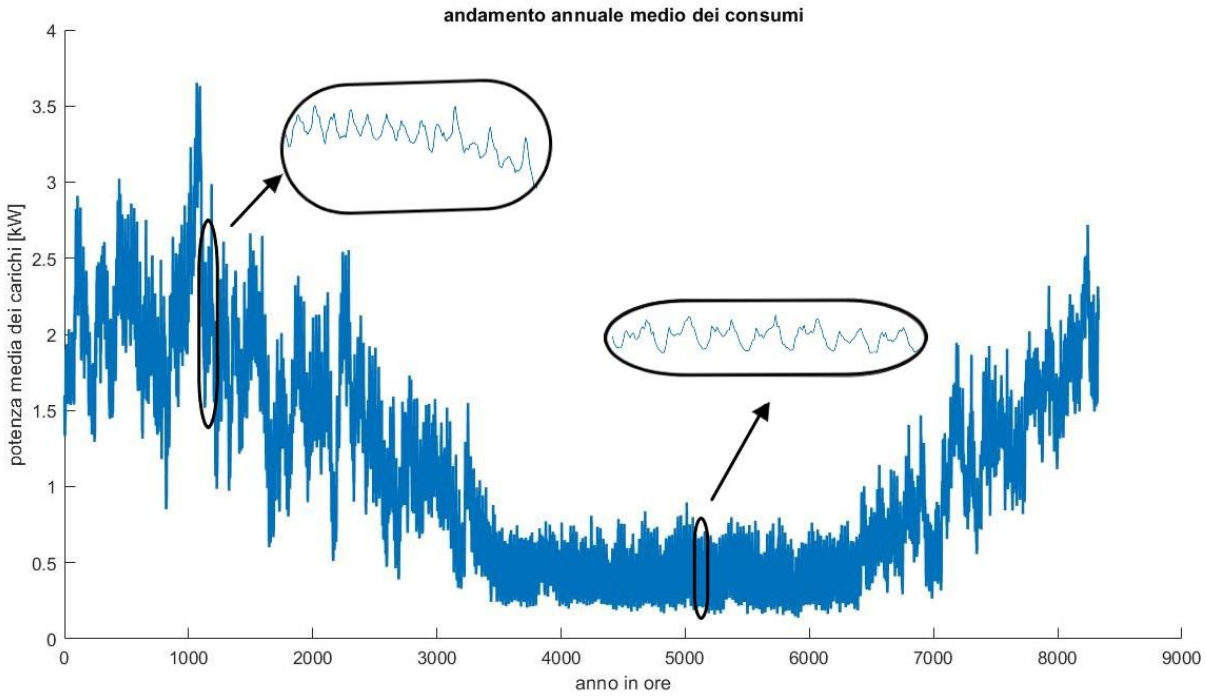


Figura 3.5.1 – Andamento medio annuale dei carichi residenziali

Si può notare fin da subito la dipendenza dei consumi di elettricità al variare delle stagioni, in particolare, durante i primi e gli ultimi mesi dell'anno si ha un maggiore consumo medio, probabilmente per via dell'elevata necessità di un riscaldamento o raffreddamento nel caso in cui tale area geografica si trovi nell'emisfero australe (ciò non viene specificato nel dataset considerato). Situazione molto differente nei mesi intermedi, in cui i consumi si riducono drasticamente (a differenza di un clima italiano) presentando inoltre una maggiore regolarità durante tale periodo, motivo per cui è stato scelto di proseguire l'analisi considerando le curve di carico degli utenti in queste specifiche condizioni. Al fine di individuare la presenza di periodi di tempo rilevanti nell'arco della giornata, corrispondenti a periodi tipici di attività comuni tra le abitazioni studiate, sono state calcolate, in tutto il dataset, il numero di mezz'ore per cui si superano i valori 1 kW, 3 kW e 6 kW per tutte le curve di carico giornaliere presenti nel dataset, divise per stagioni in base ai loro consumi medi. In *Figura 3.5.2* si vede che, considerando come soglia minima 1 kW, durante la stagione 1 e la stagione 4, la differenza tra i periodi della giornata è minima, in quanto i consumi in queste condizioni sono mediamente superiori a tale soglia. Nella

stagione 3, invece, si vedono i picchi di domanda rilevanti durante le ore mattutine ed in quelle serali, poiché già 1 kW discrimina molte curve di carico con consumi inferiori durante le ore notturne.

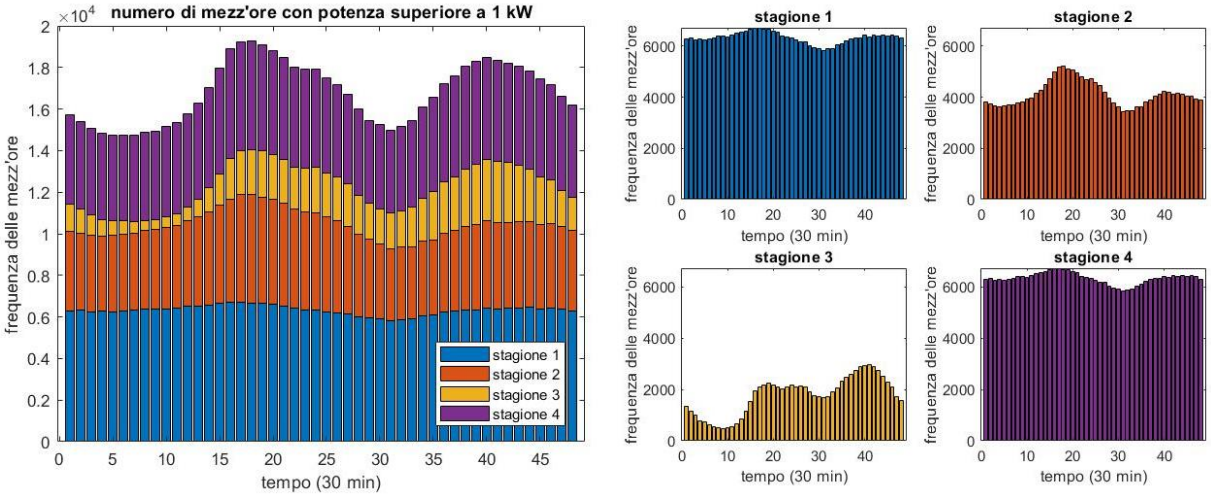


Figura 3.5.2 – Numero di mezz'ore che superano 1 kW di potenza in tutto il dataset, distinto per stagioni

In *Figura 3.5.3*, viene mostrata la frequenza delle mezz'ore che superano i 3 kW nell'arco della giornata; numeri ovviamente inferiori rispetto a quelli precedenti in tutte le stagioni, ma in questo caso si vede maggiormente la differenza tra le diverse ore della giornata, anche nelle stagioni in cui si hanno consumi maggiori. In tali stagioni un maggior numero di curve aventi consumi elevati, presentano più frequentemente dei picchi durante le ore mattutine piuttosto che in quelle serali, nonostante ciò, i periodi in cui essi si verificano rimangono indifferenti, anche tra le varie stagioni.

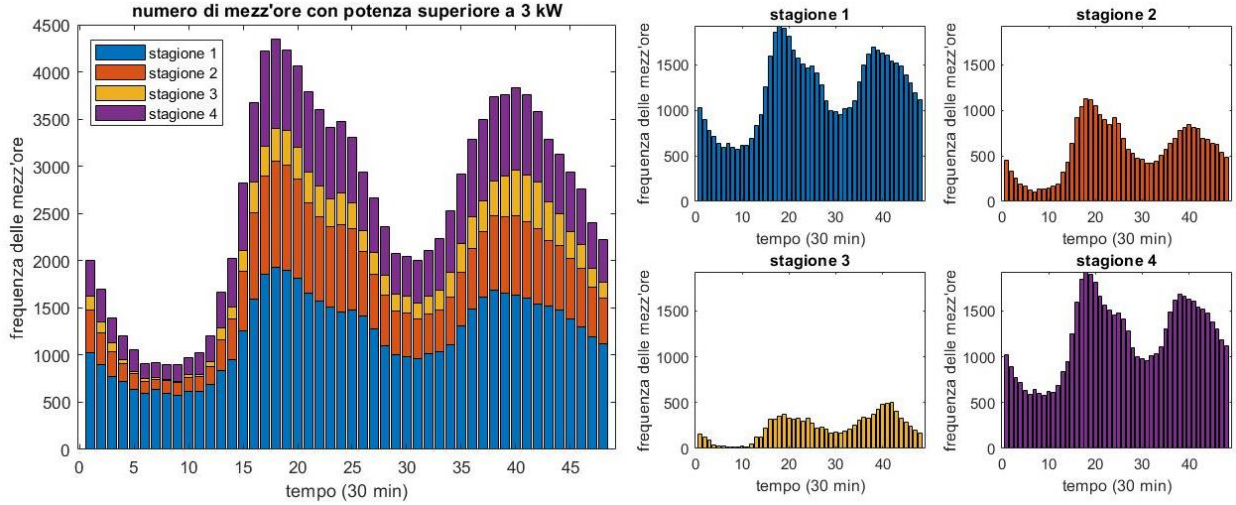


Figura 3.5.3 – Numero di mezz'ore che superano 3 kW di potenza in tutto il dataset, distinto per stagioni

Le stesse considerazioni possono essere effettuate osservando le frequenze di picchi di domanda superiori a 6 kW, riportati in *Figura 3.5.4,* in cui i valori sono però notevolmente ridotti.

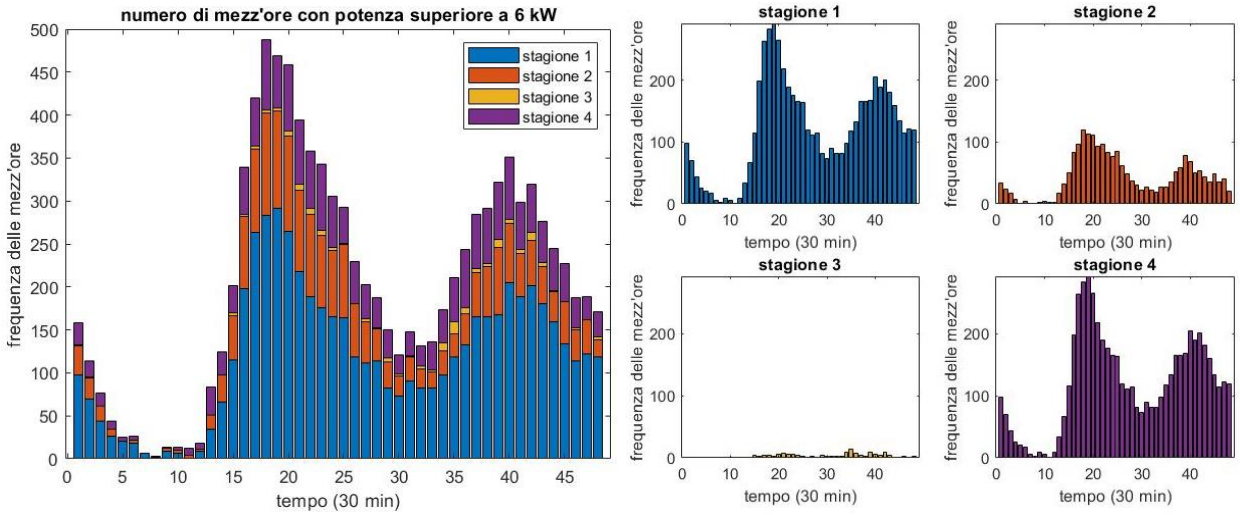


Figura 3.5.4 – Numero di mezz'ore che superano 6 kW di potenza in tutto il dataset, distinto per stagioni

Per confermare queste considerazioni, si è deciso di effettuare un diagramma a scatola e baffi (*box and whiskers*), ovvero una rappresentazione grafica utile per descrivere la distribuzione di un campione di dati tramite dei rettangoli (per ogni caratteristica) delimitati dal primo e terzo quartile, i quali sono divisi al loro interno dalla mediana, e dei segmenti delimitati dai valori estremi. La *Figura 3.5.5* mostra i risultati relativi ai giorni feriali divisi per stagione, scegliendo di non rappresentare gli *outlier* presenti nella distribuzione.

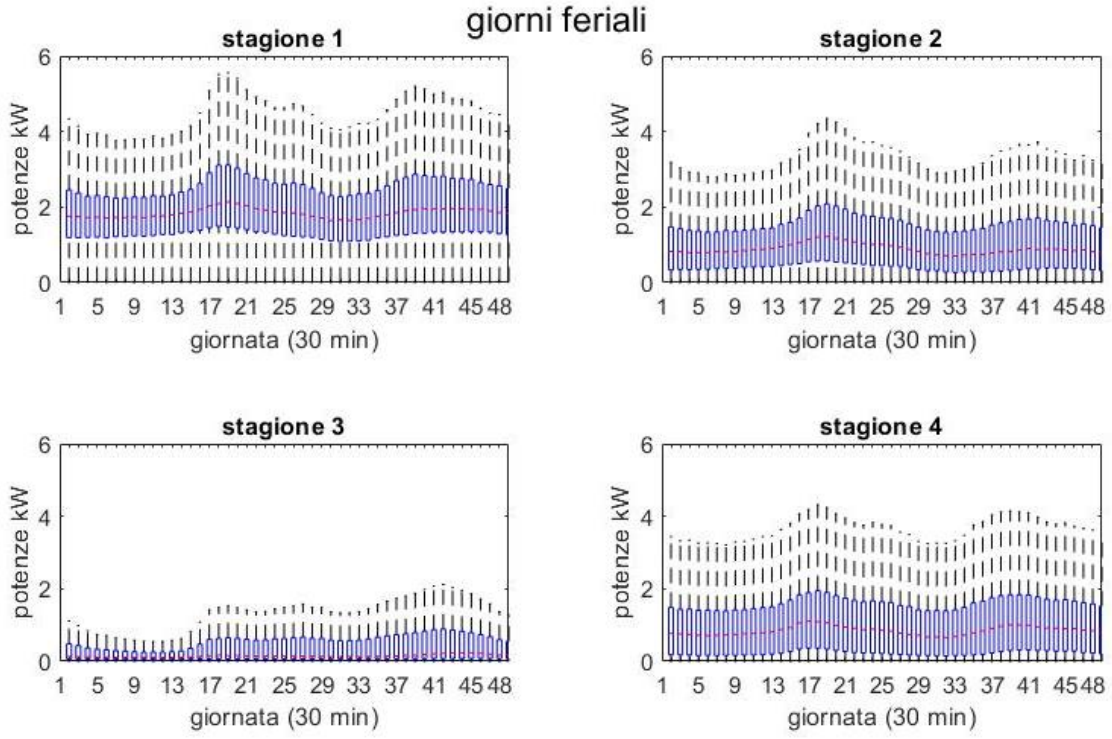


Figura 3.5.5 – Boxplot delle curve di carico di tutti gli utenti in tutto l'anno diviso in stagioni, per giorni feriali

Si osserva l'alta variabilità delle curve di carico, soprattutto per le stagioni 1 e 4 relative ai primi e ultimi mesi dell'anno. Nonostante ciò, si riesce ad apprezzare la differenza dei consumi tra le stagioni e la presenza di alcuni picchi che si verificano circa nelle stesse ore della giornata, quasi indipendentemente dalla stagione. In *Figura 3.5.6* si possono vedere i risultati riferiti ai fine settimana, in cui si nota che, a differenza dei giorni lavorativi, le ore di punta della mattina sono leggermente ritardate, coerente con un'abituale comportamento dei residenti, in tali giorni, i quali si alzano più tardi. Inoltre, anche i consumi durante la giornata sembrerebbero essere superiori, dovuti ad un maggior numero di ore in cui gli utenti risiedono nelle loro abitazioni. Durante le ore notturne non si verificano grossi cambiamenti.

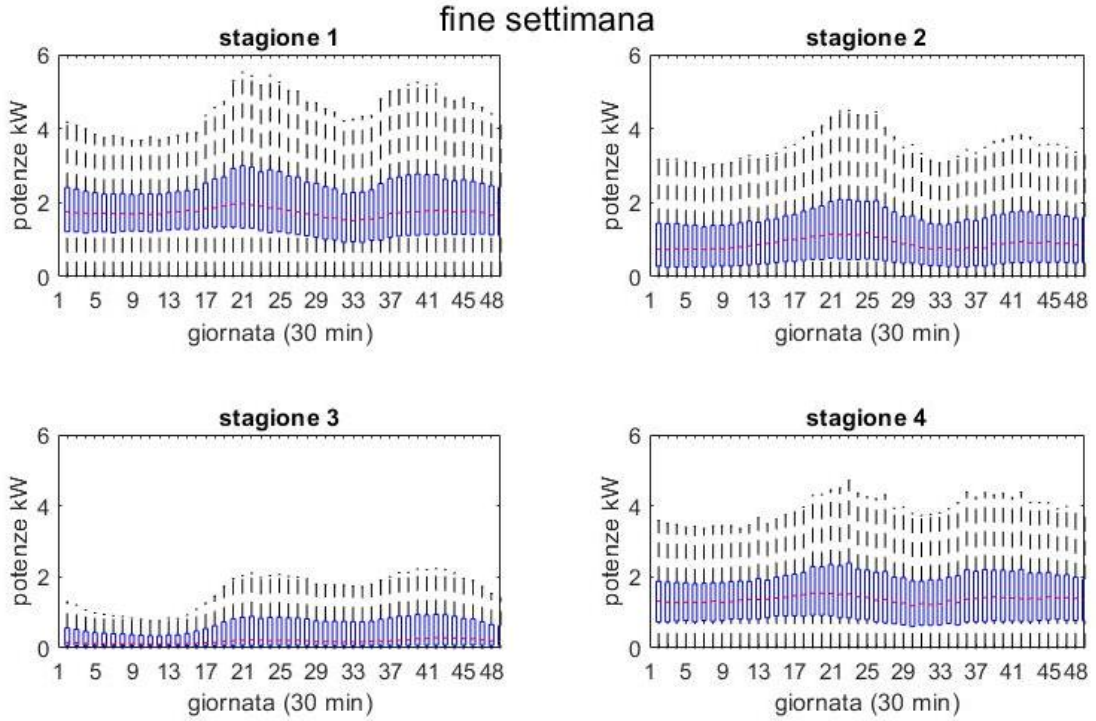


Figura 3.5.6 – Boxplot delle curve di carico di tutti gli utenti in tutto l'anno diviso in stagioni, per i fine settimana

In seguito sono stati riportati i quattro specifici periodi della giornata, nei quali si identificano le tipiche attività svolte dai vari utenti nei giorni feriali:

- *Periodo 1 (durante la notte)*: dalle 23:00 alle 6:00
  (dalla 1° alla 12° mezz'ora e dalla 46° alla 48° mezz'ora)
- *Periodo 2 (durante la mattina)*: dalle 6:00 alle 9:30
  (dalla 12° alla 19° mezz'ora)
- *Periodo 3 (durante il giorno)*: dalle 9:30 alle 15:30
  (dalla 19° alla 31° mezz'ora)
- *Periodo 4 (durante la sera)*: dalle 15:30 alle 23:00
  (dalla 31° alla 46° mezz'ora)

mentre per i fine settimana si identificano i seguenti periodi:

- *Periodo 1 (durante la notte)*: dalle 23:30 alle 7:00
  (dalla 1° alla 14° mezz'ora e dalla 47° alla 48° mezz'ora)
- *Periodo 2 (durante la mattina)*: dalle 7:00 alle 11:30
  (dalla 14° alla 23° mezz'ora)
- *Periodo 3 (durante il giorno)*: dalle 11:30 alle 16:30
  (dalla 23° alla 33° mezz'ora)
- *Periodo 4 (durante la sera)*: dalle 16:30 alle 23:30
  (dalla 33° alla 47° mezz'ora)

Il successivo passo è quello di suddividere tale insieme di dati e isolare gruppi di utenti non adatti per i programmi di DR, ciò comporterebbe, però, a ridurre drasticamente il numero di curve con cui costruire i carichi aggregati, producendo delle variazioni di carico aggregato più pronunciate nei rispettivi intervalli e delle osservazioni di tali carichi più simili tra loro, falsificando l'analisi sulla flessibilità, in quanto presentano un comportamento comune dovuto alla ripetizione della stessa abitazione. Perciò, non disponendo di un dataset con elevato numero di utenti, al fine di non alterare tale analisi, l'insieme delle curve di carico dei vari utenti viene esteso, considerando come ulteriori abitazioni anche i diversi giorni feriali, in quanto ogni famiglia comunque presenta un comportamento differente per ogni giorno della settimana. In tal modo l'insieme dati di partenza, con cui proseguire l'analisi, corrisponderà ad un insieme formato da 570 curve differenti (considerando solo i giorni feriali) riferite alla stagione 3.

Cluster analysis

Lo scopo di questa trattazione è di andare a valutare la flessibilità delle curve di carico aggregate, al fine di individuare periodi della giornata in cui il DSO o l'aggregatore possa avere più successo nell'incentivare un gruppo selezionato di utenti residenziali per apportare dei cambiamenti nella domanda. La probabilità di effettuare tali cambiamenti e la percentuale di domanda che potrebbe essere variata, in intervalli di tempo specifici, viene valutata seguendo un'analisi di flessibilità. Tuttavia, utenti che hanno consumi quasi costanti in tutta la giornata, nelle condizioni di carico considerate, o che presentano dei comportamenti singolari, come grossi consumi solamente durante le ore notturne, dovranno essere eliminati dal dataset in esame, in quanto andrebbero a impattare sulle curve aggregate, non consentendo di effettuare considerazioni sul tipico comportamento dei carichi nei rispettivi periodi temporali specifici della giornata. Per poter individuare, quali gruppi di carichi isolare e quindi identificare quelli utili per la DR, è stata eseguita una cluster analysis, basata su caratteristiche definite dai periodi specifici della giornata. Sono state ricavate 5 caratteristiche, per ogni utente, quattro delle quali rappresentano le potenze medie relative nei rispettivi periodi temporali (Eq. 3.5.1), mentre l'ultima è una deviazione standard media nella giornata (Eq. 3.5.2), per tenere in considerazione della variabilità dei consumi di ogni utente.

Dunque, per ogni utente *m*-esimo, le 5 caratteristiche sono state calcolate nel seguente modo:

$$1-4] \qquad p_{rel,m}^{i} = \frac{\bar{p}_m^i}{\bar{p}_m} \qquad per\ \ i = 1,2,3,4 \qquad (3.5.1)$$

$$5] \qquad \bar{\sigma}_m = \frac{1}{4}\sum_{i=1}^{4}\frac{\bar{\sigma}_m^i}{\bar{p}_m^i} \qquad (3.5.2)$$

dove $\bar{p}_m^i$ e $\bar{\sigma}_m^i$ sono la potenza media e la deviazione standard media riferite all'*i*-esimo periodo temporale caratteristico della giornata, calcolate durante tutto l'intervallo di tempo di osservazione (es. durante la stagione 3), mentre $\bar{p}_m$ è la potenza media giornaliera nel periodo di osservazione.

Al fine di scegliere un numero di cluster adatto a rappresentare al meglio la partizione dei differenti gruppi, è stato calcolato l'indice *Global Silhouette* (*GS*) al variare del numero di cluster prodotti da differenti algoritmi. Dal loro confronto, per il dataset che si sta analizzando, l'algoritmo gerarchico con average linkage, presenta soluzioni migliori. Il valore ottimale del numero di cluster prodotti è stato scelto, come il valore con cui l'andamento medio del GS al variare del numero di gruppi, inizia a salire, come mostrato in *Figura 3.5.7*. Tuttavia, per confermare queste ipotesi, e misurare ulteriormente la qualità dei risultati del clustering, sono stati calcolati altri indici di validità, mostrati in *Figura 3.5.8*. Per i metodi non deterministici (k-means e k-medoids) sono stati riportati i valori medi degli indici ottenuti da una ripetizione di 100 volte dell'algoritmo, per ogni numero di cluster.

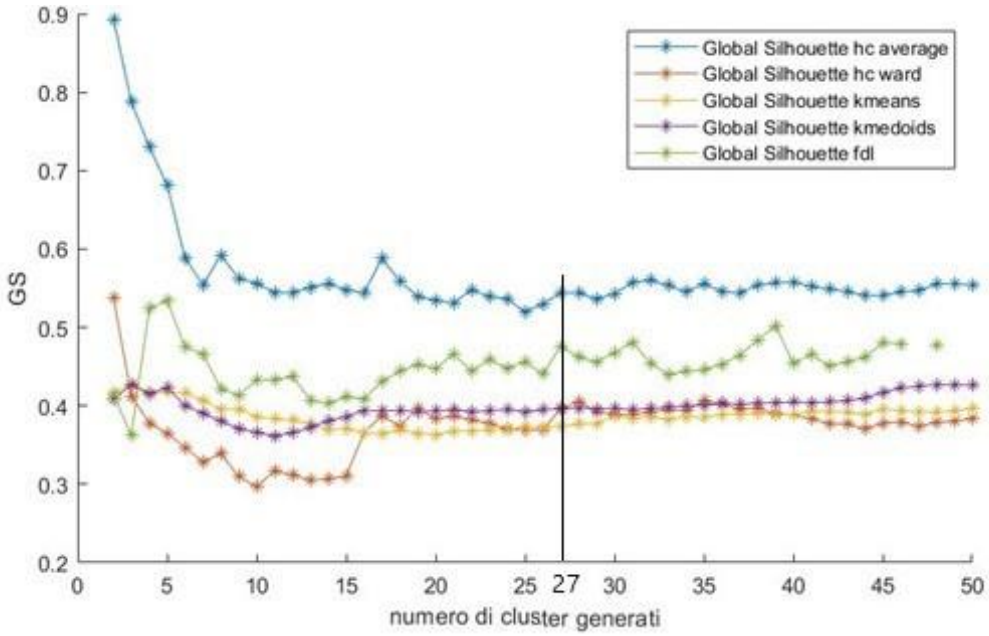


Figura 3.5.7 – Global Silhouette coefficient per diversi metodi di clustering al variare del numero di cluster

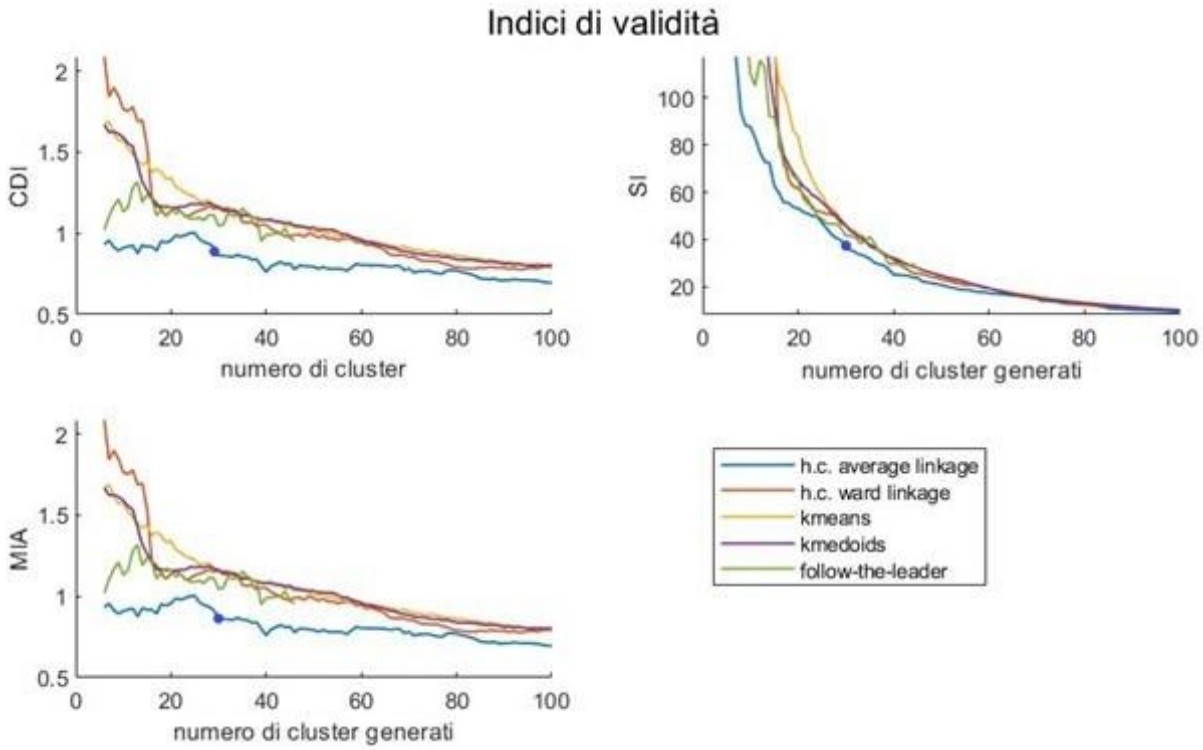


Figura 3.5.8 – Indici di validità per diversi metodi di clustering al variare del numero di cluster

Tali risultati, ottenuti dai diversi metodi di clustering, confermano la scelta di utilizzare lo hierarchical clustering con average linkage, in quanto presenta prestazioni migliori, nel caso specifico. La scelta del valore ottimale del numero di gruppi da generare è confermata anche da tali indici, che in questo caso, è rappresentato dal punto in cui cambia la pendenza delle curve. Per tali motivi è stato deciso di effettuare una ripartizione in 27 gruppi utilizzando lo hierarchical clustering con average linkage, ed ottenendo i risultati in *Figura 3.5.9*. È stato deciso di rappresentare i risultati per i diversi gruppi, tramite i diagrammi a scatola e baffi, in modo tale da vedere le differenze tra i valori medi delle caratteristiche, le variazioni da tali valori e gli eventuali *outlier*. Si può notare che i gruppi contenenti singoli utenti aventi caratteristiche singolari, come i cluster 18, 19 e 27, presentano scarsi consumi durante la giornata e solamente dei picchi durante un solo periodo specifico; tali comportamenti risultano abbastanza sistematici durante il periodo di tempo osservato, in quanto presentano bassi valori della deviazione standard media (quinta caratteristica). Particolarmente distintivi sono anche i cluster 5 e 6, che presentano un insieme di utenti che mediamente mostrano dei consumi costanti in tutto l'arco della giornata tipica, e che quindi non possono essere presi in considerazione allo scopo dell'analisi svolta. Gruppi come il 21, 22 e 23 sono stati scartati in quanto presentano un grosso contributo durante le ore notturne rispetto al resto della giornata, comportamento alquanto differente rispetto ai restanti cluster. Per avere un confronto con le curve di carico presenti in ogni gruppo, in *Figura 3.5.10* sono riportate le curve di carico corrispondenti agli utenti appartenenti ai diversi cluster, evidenziando i modelli rappresentativi delle curve di carico di ogni gruppo; da cui si può notare che i primi e gli ultimi cluster presentano un piccolo contributo sui consumi e perciò non utili nei programmi di DR. Una volta scartati i cluster che non sono validi per effettuare le considerazioni sulla flessibilità della domanda aggregata, si è ricavata una matrice contenente 295 andamenti temporali corrispondenti ai possibili candidati per il DR.

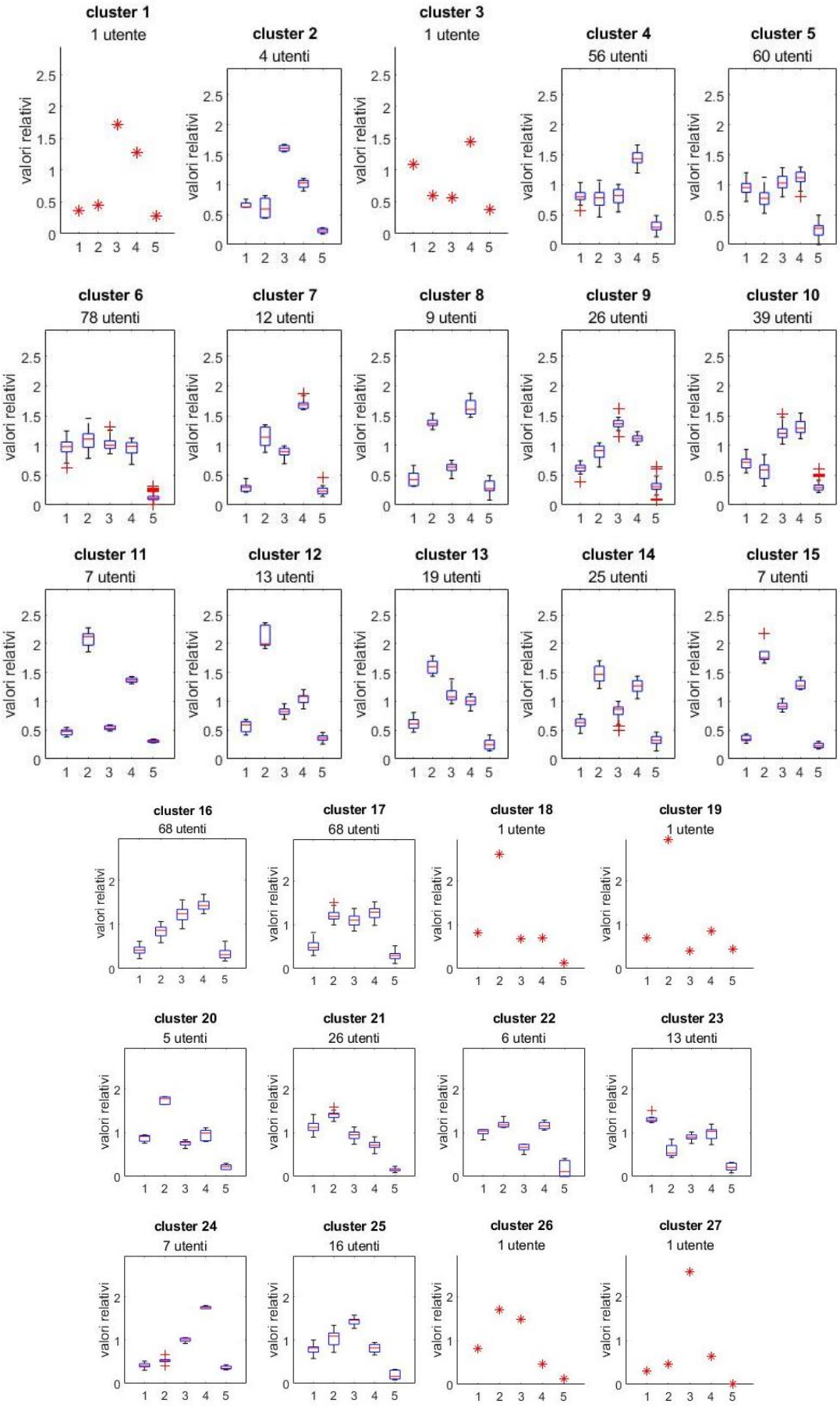


Figura 3.5.9 – Risultati del hierarchical clustering con average linkage effettuato sulle 5 caratteristiche

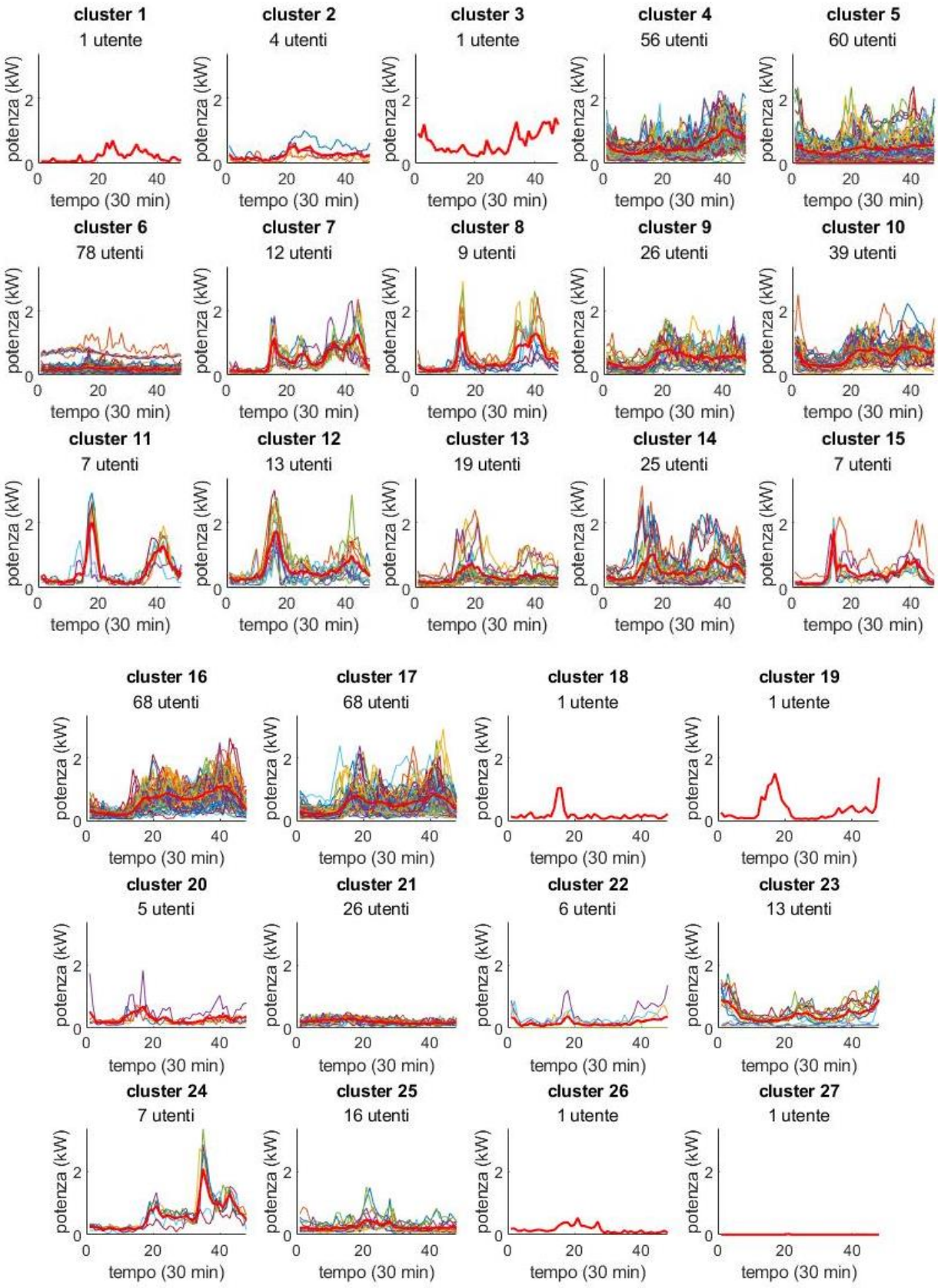


Figura 3.5.10 – Curve di carico corrispondenti agli utenti appartenenti ai diversi cluster

Curve di carico aggregate

Dall'insieme degli andamenti giornalieri precedentemente selezionati, sono state costruite 100 curve aggregate, ognuna delle quali è stata realizzata tramite il metodo Monte Carlo, estraendo casualmente un numero di carichi, riferiti ad un intervallo di campionamento specifico, pari al livello di aggregazione considerato. Sono stati presi in considerazione i seguenti scenari riferiti a giorni feriali della stagione 3, e mostrati in *Figura 3.5.11*.

- *Scenario 1*: livello di aggregazione pari a 20 case e intervallo di campionamento di 15 minuti;
- *Scenario 2*: livello di aggregazione pari a 150 case e intervallo di campionamento di 15 minuti;
- *Scenario 3*: livello di aggregazione pari a 75 case e intervallo di campionamento di 5 minuti;
- *Scenario 4*: livello di aggregazione pari a 150 case e intervallo di campionamento di 5 minuti.

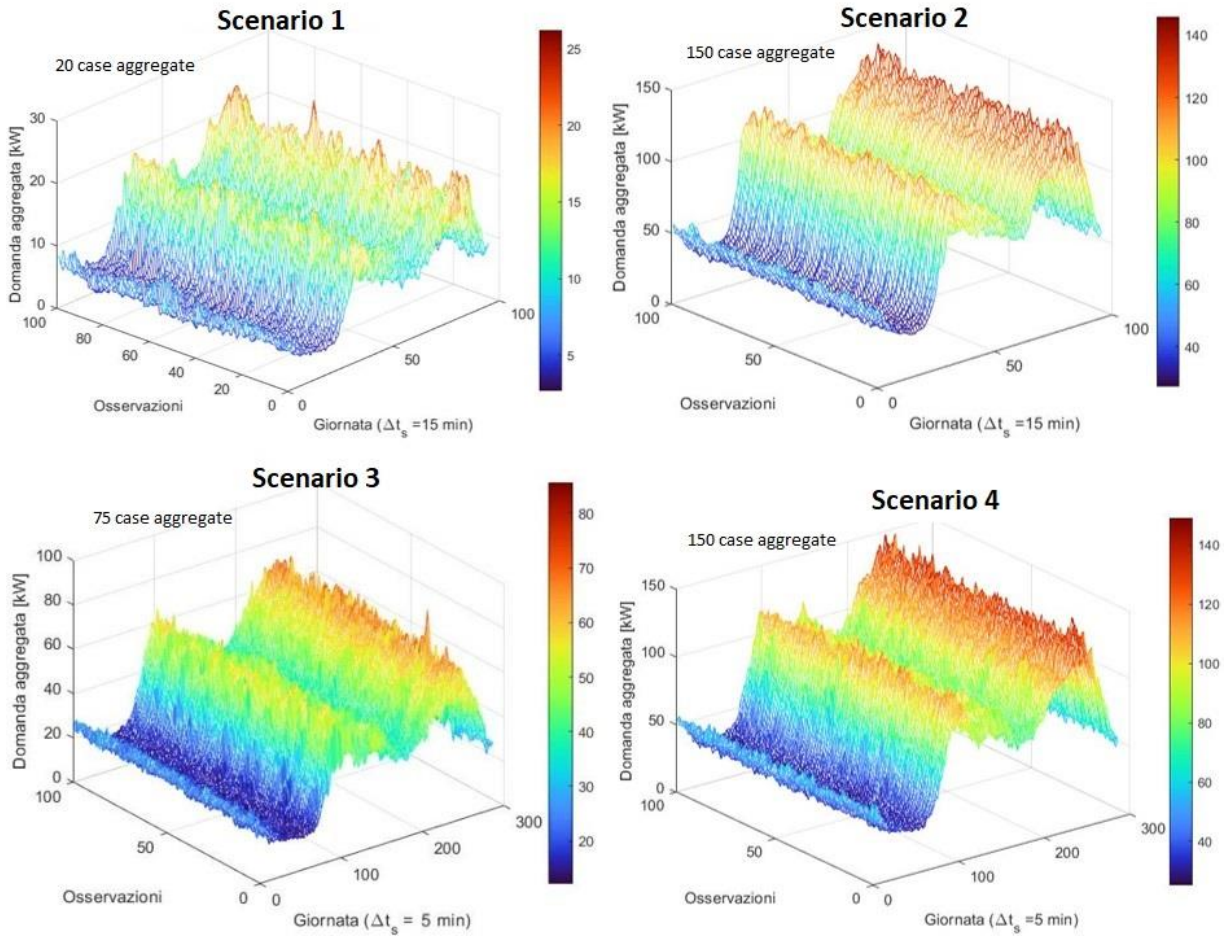


Figura 3.5.11 – Curve di carico aggregate corrispondenti ai 4 scenari

Si può notare, oltre ovviamente alla differenza di potenza in base al numero di carichi aggregati, la variabilità tra le diverse curve aggregate presenti in ogni casistica. Infatti, per livelli di aggregazione più bassi, le curve sono molto più differenti tra loro per via della casualità con cui sono costruite, contrariamente per gli scenari 2 e 4, esse sono state realizzate estraendo con più probabilità gli stessi carichi, ciò in accordo con gli andamenti dell'indice ASRD considerati in

Figura 3.2.4. Stesse considerazioni si sono ottenute all'aumentare dell'intervallo di campionamento adottato, in quanto perdendo la vera dinamica del carico le curve risultano più piatte.

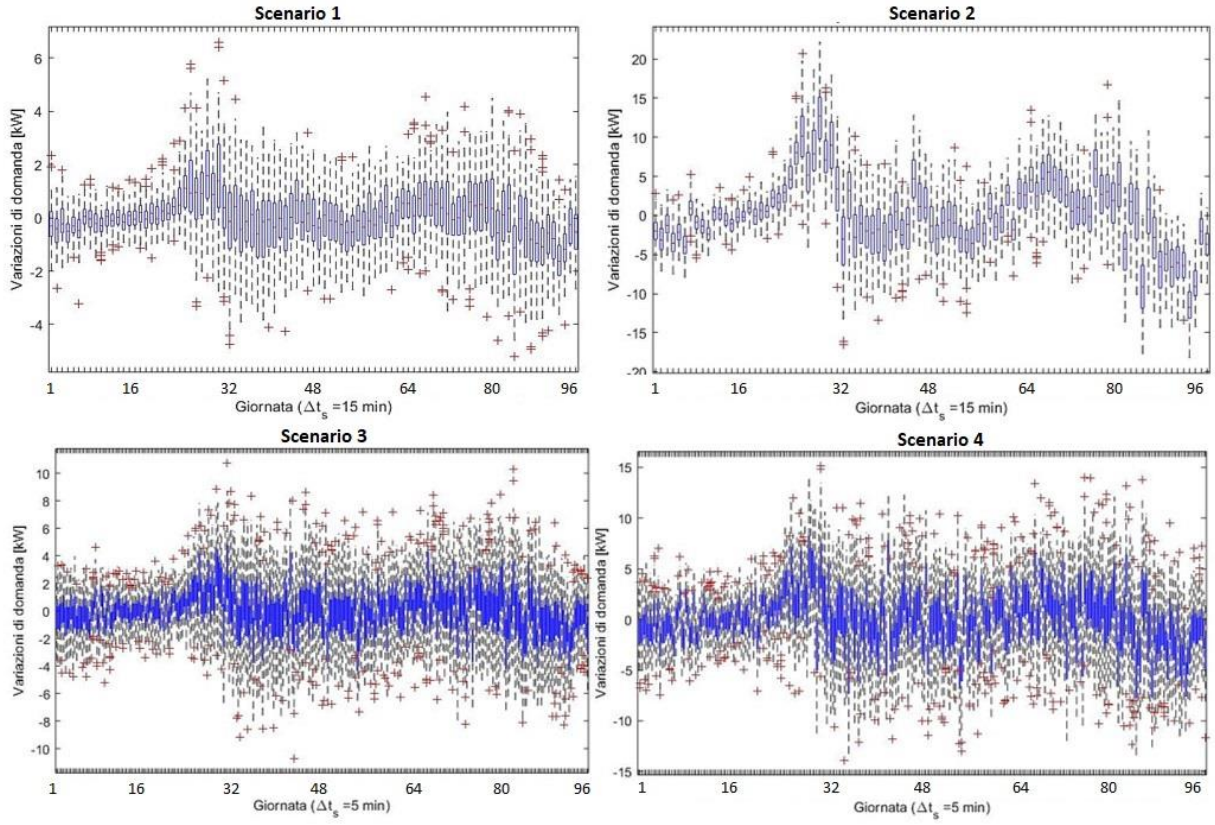


Figura 3.5.12 – Boxplot delle variazioni di carico corrispondenti ai 4 scenari

In *Figura 3.5.12*, sono rappresentati i diagrammi a scatola e baffi delle variazioni della domanda, rispettivamente ai casi in studio. Tali diagrammi dimostrano che gli scenari realizzati con un numero maggiore di carichi aggregati, presentano una maggiore differenza delle variazioni di carico in alcuni periodi della giornata (ad esempio il picco di domanda mattutino) rispetto ad altri, evidenziando un comportamento più rigido essendo ottenute da un numero maggiore di curve simili. In *Figura 3.5.13* sono riportati gli andamenti medi delle potenze nei vari scenari e le rispettive variazioni.

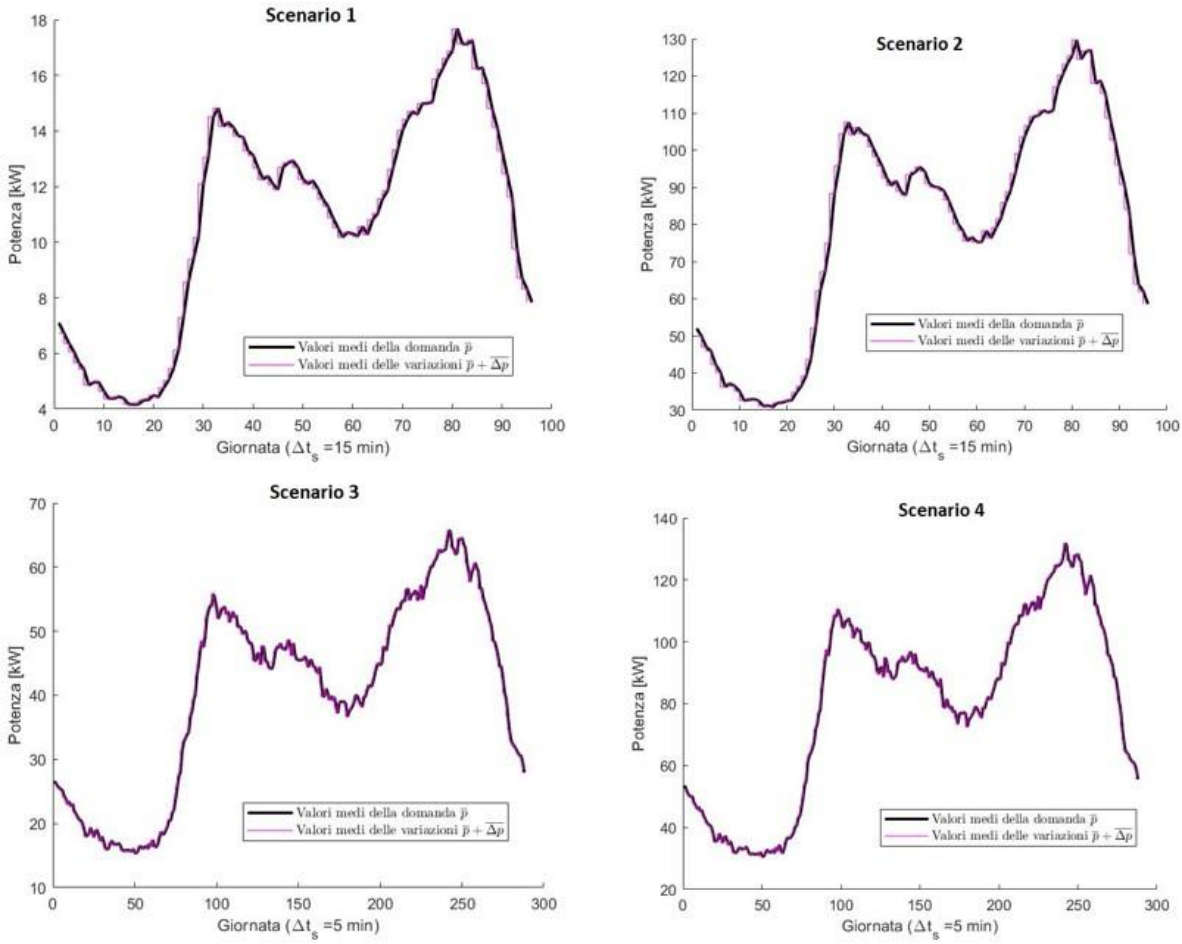


Figura 3.5.13 – Andamenti medi della domanda e le rispettive variazioni, corrispondenti ai 4 scenari

Analisi della flessibilità

Sulla base delle considerazioni esposte nel paragrafo precedente, sono stati calcolati gli indici di flessibilità per le curve di carico aggregate, nei vari scenari. L'obiettivo di tale analisi è di individuare intervalli di tempo che presentano una potenziale disponibilità da parte degli utenti a variare i propri consumi, al fine, ad esempio, di ridurre la domanda aggregata tramite programmi di DR. In *Figura 3.5.14* sono stati riportati gli indici *FIAD* e *MFIAD* che identificano in termini probabilistici la flessibilità a cambiare il comportamento collettivo degli utenti aggregati. Si osserva che, sia durante la mattina che durante la sera la flessibilità è molto bassa rispetto al resto della giornata, per via di un andamento molto più rigido della curva aggregata, dovuto al comune comportamento degli utenti in tali periodi. Si conferma ciò che è stato dedotto dalle variazioni di carico, ovvero che con l'aumentare del livello di aggregazione, in tali periodi della giornata, il comportamento collettivo è ancora più incisivo. Tuttavia, se si considerano intervalli di campionamento minori, si riescono ad ottenere maggiori informazioni sulla dinamica del carico, individuando anche in tali periodi della giornata, intervalli di tempo con flessibilità notevoli. Il *MFIAD*, in confronto al *FIAD*, fornisce un margine quantitativo (massimo cambiamento nella domanda) oltre alla probabilità di incremento o decremento della domanda, accettando variazioni (positive o negative) in essa. Si nota infatti che, diminuendo il livello di aggregazione o l'intervallo di campionamento, i valori di tali indici risultano essere superiori, riuscendo a sfruttare al meglio la casualità nei comportamenti dei diversi utenti aggregati.

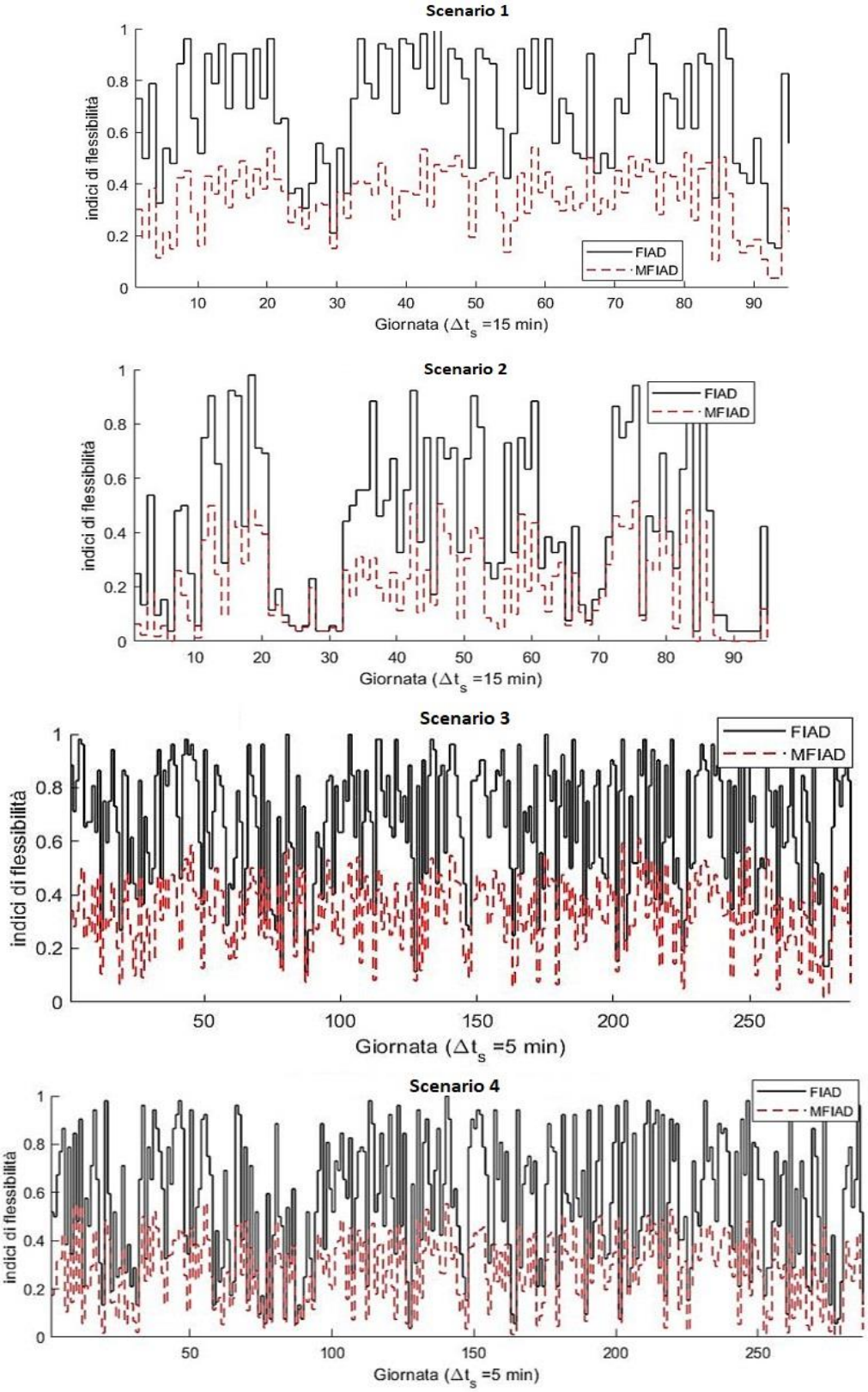


Figura 3.5.14 – Indici FIAD e MFIAD corrispondenti ai 4 scenari

Per ottenere informazioni sui valori percentuali di domanda flessibile in ogni istante di tempo in cui sono definiti questi indicatori, sono stati calcolati gli indici PFL_{FIAD} e PFL_{MFIAD} riportati in *Figura 3.5.15*; mentre in *Figura 3.5.16*, sono stati confrontati i valori di *PFL* (ottenuto con il *FIAD*) e $\bar{\boldsymbol{\psi}}$. Si nota che, anche in questo caso, durante i periodi della giornata visti in precedenza, ovvero in corrispondenza della rampa mattutina e quella serale, gli indicatori di flessibilità presentano

valori inferiori, a causa della rigidezza della domanda aggregata. Tuttavia, l'indicatore $\bar{\boldsymbol{\psi}}$, rispetto a gli altri, assume valori decisamente più elevati, che si discostano maggiormente in questi periodi della giornata, in quanto esso non tiene in considerazione il comportamento collettivo degli utenti, bensì il massimo margine di variazione della domanda ottenibile nel caso ideale in cui tutti gli utenti cambiano il loro comportamento. Dalle 5:30 alle 8:00 del mattino (dal 22° al 32° quarto d'ora), dalle 16.30 fino 18:00 (dal 66° al 72° quarto d'ora) e durante la sera dalle 19:00 alle 20:00 (dal 76° al 80° quarto d'ora) si osservano valori elevati di $\bar{\boldsymbol{\psi}}$, indicando una buona possibilità nel far diminuire la domanda, per via del comportamento comune degli utenti nell'aumentare i consumi; tuttavia, questa tendenza indica una scarsa probabilità a far cambiare tale comportamento, riducendo la flessibilità nella domanda, come indicato dal *FIAD* e dal *PFL*. Osservando la percentuale di domanda flessibile, si riesce a vedere la riduzione che si ha all'aumentare del numero di carichi aggregati che, oltre a presentare un comportamento più rigido in determinati periodi della giornata, si riesce a ridurre meno la percentuale di domanda rispetto a gli altri casi per via delle minori variazioni di carico in tali scenari. A parità di livello di aggregazione inoltre si nota che, il diminuire dell'intervallo di campionamento consente di sfruttare meglio la dinamica del carico, individuando intervalli flessibili anche durante periodi con maggiore rigidità, ciò nonostante, la percentuale di variazioni di carico inferiore osservabile dai valori di $\bar{\boldsymbol{\psi}}$ durante le ore mattutine. Particolarmente evidente è l'effetto che hanno durante le ore notturne, sia l'aumento del livello di aggregazione che l'intervallo di campionamento, in cui piccoli consumi di apparecchiature come i frigoriferi operano in maniera non sincrona tra le diverse abitazioni. Infatti, lo scenario 2 presenta valori inferiori degli indici in tali periodi, per via dell'effetto di tale mediazione, che rende meno influente il fenomeno di questi particolari consumi, ma ciò riduce notevolmente la flessibilità della domanda aggregata.

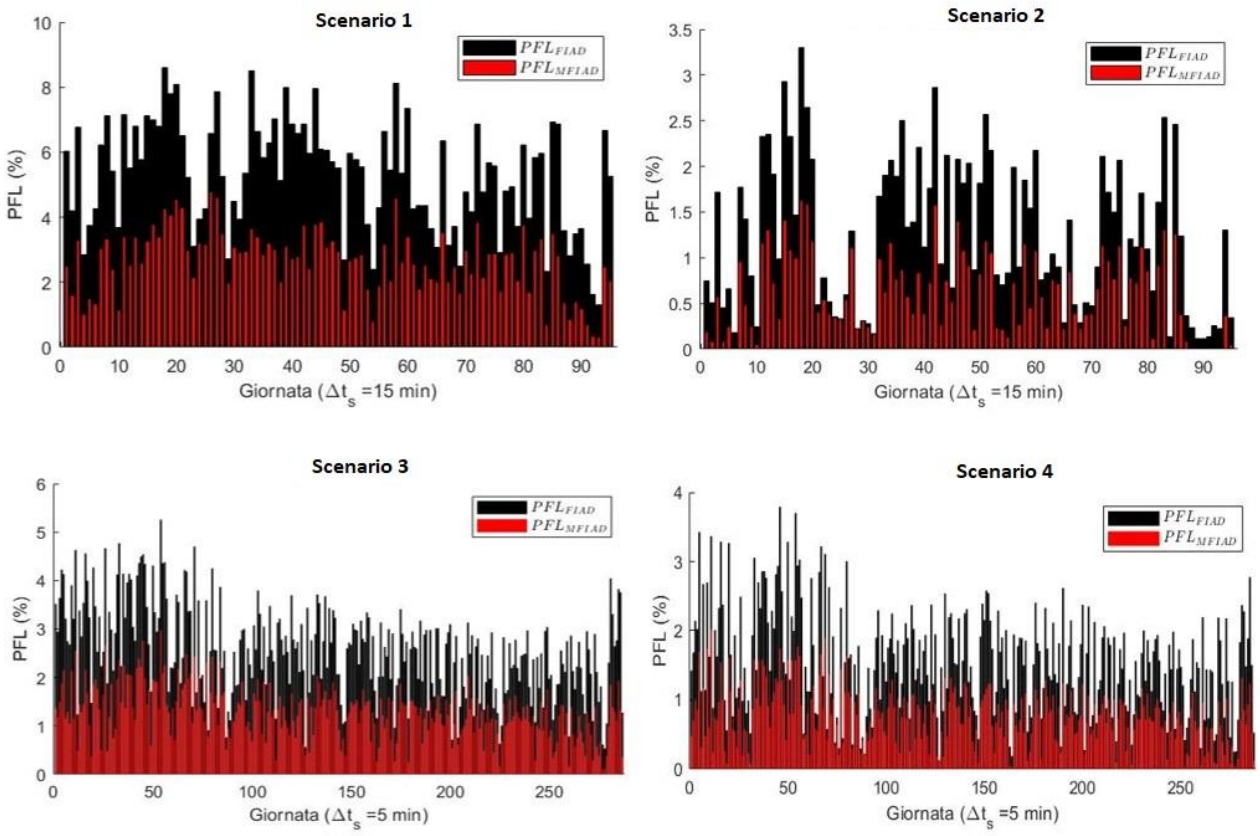


Figura 3.5.15 – Indici PFL_{FIAD} e PFL_{MFIAD} corrispondenti ai 4 scenari

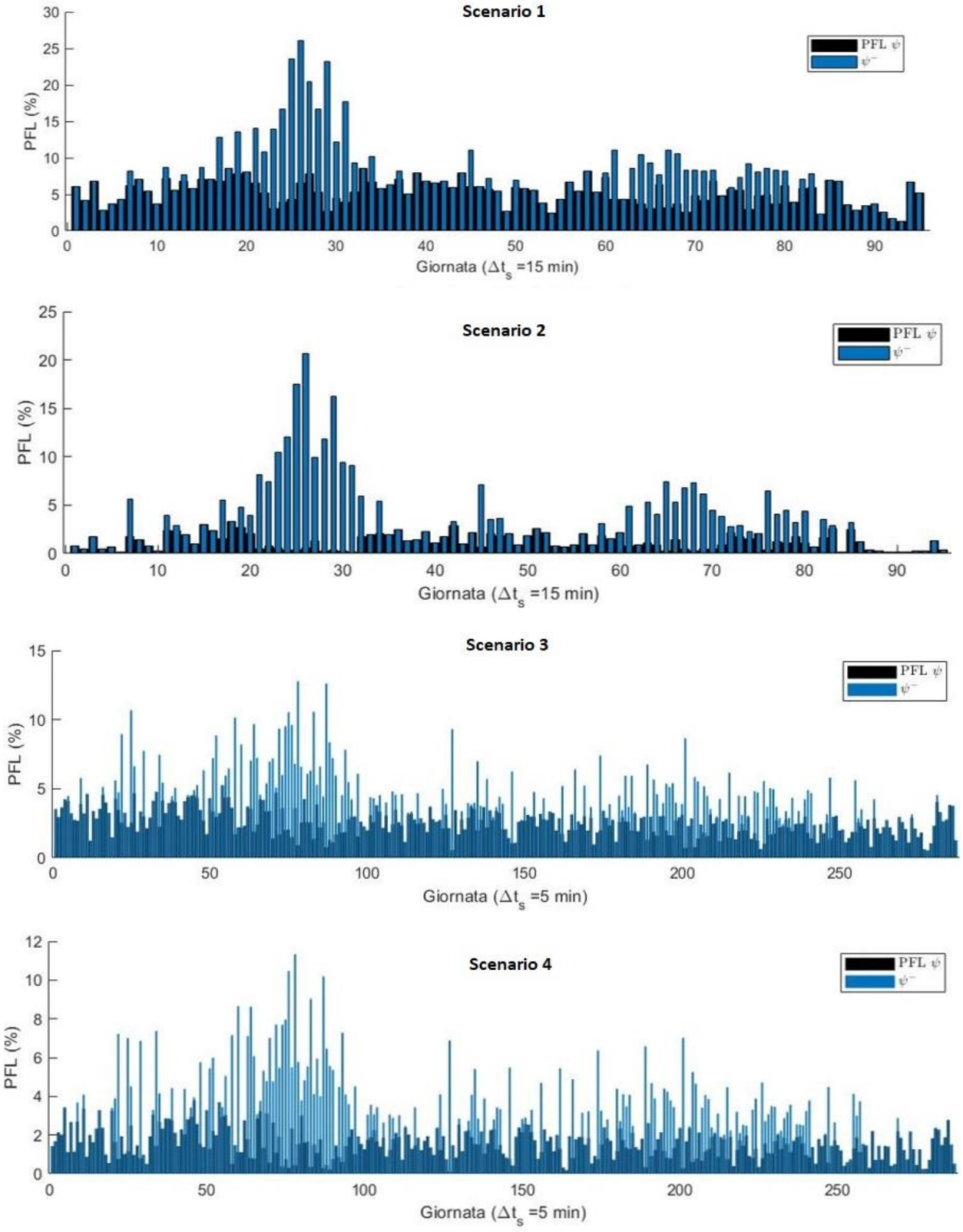


Figura 3.5.16 – Indici PFL_{FIAD} e $\bar{\psi}$. corrispondenti ai 4 scenari

Gli indicatori utilizzati ricavano informazioni dalle variazioni di domanda e sono utili all'operatore del sistema, per selezionare intervalli di tempo adatti ad avviare programmi di DR. Infatti, valori elevati indicano la possibilità di trovare nel gruppo di utenti selezionati, più candidati disposti ad accettare tali cambiamenti. In *Figura 3.5.17* sono riportati gli effetti che si otterrebbero sulla domanda aggregata, se si riuscisse a far cambiare il comportamento degli utenti secondo tali indici.

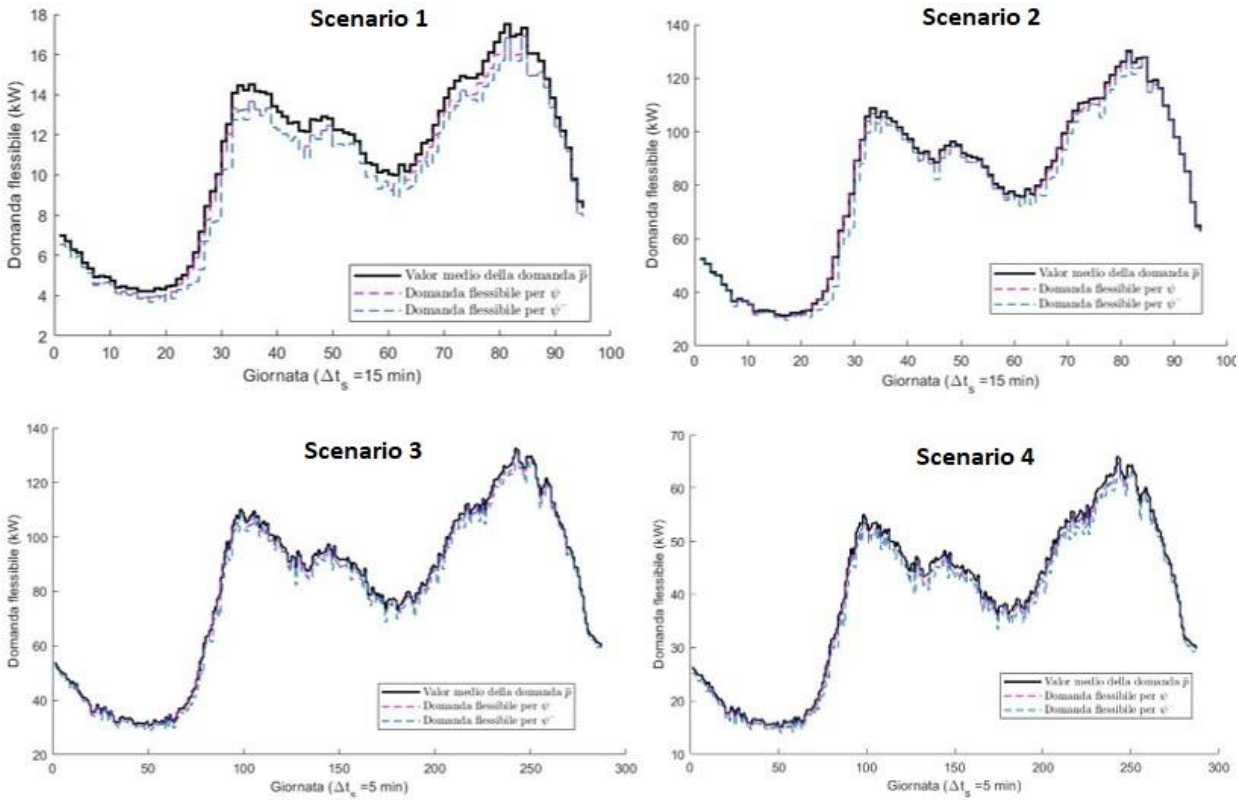


Figura 3.5.17– Domanda aggregata con la relativa flessibilità, corrispondente ai 4 scenari

3.6 Bibliografia

[1] *Sajjad, I.A., Chicco, G., and Napoli, R.: 'Definitions of demand flexibility for aggregate residential loads', IEEE Transactions on Smart Grid, vol. 7 (6), pp. 2633–2643, 2016.*

[2] *Waseem M., Sajjad I.A., Haroon S.S., Amin S., Farooq H., Martirano L., and Napoli, R.: 'Electrical Demand and Its Flexibility in Different Energy Sectors', Electric Power Components and Systems, vol. 48 (12-13), pp.1339-1361, 2020.*

[3] *Haben, S., Singleton, C., and Grindrod, P.: 'Analysis and clustering of residential customers energy behavioral demand using smart meter data', IEEE Transactions on Smart Grid, vol. 7 (1), pp. 136-144, 2016.*

[4] *Cagni, A. Carpaneto, E., Chicco, G., and Napoli, R.: 'Characterisation of the aggregated load patterns for extraurban residential customer groups', Proceedings of the 12th IEEE Mediterranean Electrotechnical Conference (MELECON), pp. 951–954, 2004.*

[5] *Sajjad, I.A., Chicco, G., and Napoli, R.: 'A statistical analysis of sampling time and load variations for residential load aggregations', IEEE 11th International Multi-Conference on Systems, Signals and Devices, SSD 2014, no. April, pp. 1–6, 2014*

[6] *Sajjad, I.A., Chicco, G., and Napoli, R.: 'Effect of aggregation level and sampling time on load variation profile-A statistical analysis', 17th IEEE Mediterranean Electrotechnical Conference (MELECON), pp. 208–212, 2014.*

[7] *Sajjad, I.A., Chicco, G., and Napoli, R.: 'A probabilistic approach to study the load variations in aggregated residential load patterns', Power Systems Computation Conference (PSCC), pp. 1–7, 2014.*

[8] *Waseem, M., Sajjad, I.A., Martirano, L., and Manganelli, M.: 'Flexibility assessment indicator for aggregate residential demand', IEEE International Conference on Environment and Electrical Engineering and Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), 2017.*

[9] *Waseem, M., Sajjad, I.A., Napoli, R., and Chicco, G.: 'Seasonal effect on the flexibility assessment of electrical demand', 2018 3rd International Universities Power Engineering Conference (UPEC), pp. 1-6, Sept. 2018.*

4 Selezione degli utenti per la partecipazione ai programmi di Demand Response

Una delle più grandi sfide negli ultimi decenni è di sfruttare al meglio le nuove tecnologie come gli *smart meter* attraverso tecniche avanzate di analisi dei dati, al fine di raggiungere obiettivi sostenibili a livello mondiale, come la riduzione del consumo di energia, l'aumento della capacità e della stabilità nelle reti elettriche ridotte a causa dell'incremento di fonti rinnovabili. La flessibilità sul lato della domanda rappresenta un'importante risorsa da poter sfruttare, e grazie alle nuove tecnologie è possibile sviluppare strategie che permettono la gestione della domanda con la promozione di programmi di *Demand Response* (*DR*) anche per utenti residenziali, i quali occupano una grossa parte dei consumi totali. L'impatto del singolo utente residenziale, tuttavia, non è molto significativo, dunque, il ruolo dell'aggregatore è cruciale, in quanto permette di collezionare la flessibilità del gruppo di carichi, al fine di ottenere un maggiore contributo sulla gestione della domanda. Aumentare il numero della partecipazione[6] degli utenti ai programmi di DR significa dunque aumentare la possibilità di sfruttare la potenziale flessibilità, ottenendo maggiori vantaggi dalla DR. Tuttavia, non sempre avere molti partecipanti comporta benefici, infatti l'articolo [1] dimostra che un numero ben selezionato di utenti presenta maggiori vantaggi nella riduzione della domanda, rispetto a quello ottenibile con tutti gli utenti disponibili. L'operatore della DR dovrebbe essere in grado sia di valutare la potenziale flessibilità della domanda aggregata individuando periodi di tempo in cui essa è maggiore e nei quali si ha la disponibilità a fornire diversi servizi rispetto alle differenti richieste (in genere dal gestore del sistema elettrico), e sia di selezionare gruppi opportuni di utenti ai quali è possibile richiedere con priorità la partecipazione ai programmi di DR opportuni per sfruttare al meglio i benefici. Nel capitolo precedente è stata analizzata una metodologia che, oltre a fornire all'operatore il livello ottimale di aggregazione degli utenti e il periodo di campionamento per i sistemi di misura da utilizzare per il rilevamento dei loro consumi, consente di valutare la possibile flessibilità nella domanda aggregata di un gruppo di utenti residenziali in diversi periodi della giornata. Gli indicatori definiti estraggono informazioni dalle variazioni di domanda, che potrebbero essere utili nel processo decisionale dell'operatore per stabilire quali intervalli presentano una maggiore flessibilità del carico residenziale aggregato e quindi maggiormente convenienti per avviare programmi di DR. Infatti, da un gruppo di utenti in esame, in base ai valori degli indicatori, è possibile valutare una probabile risposta dell'utente a segnali di prezzo o incentivi per i diversi periodi. Negli intervalli di tempo in cui si hanno indici di flessibilità inferiori, le azioni proposte per rimodellare la domanda aggregata potrebbero essere poco efficienti per via dell'indisponibilità di molti utenti a cambiare le proprie abitudini. Tuttavia, nei periodi di tempo in cui si dispone un'elevata flessibilità, potrebbe essere più utile per l'operatore del sistema o l'aggregatore proporre incentivi agli utenti residenziali, in quanto un maggior numero di candidati ai programmi di DR potrebbero accettare cambiamenti nei loro consumi. Al fine di dare

[6] La partecipazione degli utenti ai programmi di DR non implica il raggiungimento degli obiettivi previsti, in quanto non tutti gli interessati seguono le richieste. Per presentare un programma di DR è necessario effettuare una valutazione dei benefici ottenibili, includendo la risposta dei singoli utenti agli stimoli forniti, che dipende però, da condizioni metereologiche, comfort e sensibilità al consumo energetico non facili da modellizzare. Gli articoli [2,3] propongono alcune metodologie per simulare tali relazioni, per ottenere gradi di affidabilità degli utenti e per stimare il beneficio della DR a partire dalla simulazione dei carichi degli utenti, aventi apparecchiature controllabili come i carichi termostatici (TLC). Inoltre è opportuno considerare un probabile comportamento strategico da parte dei partecipanti, i quali per ottenere maggiori guadagni, modificano i propri consumi poco prima dell'evento di DR.

un notevole contributo sulla gestione della domanda, e quindi assicurare il successo di tali programmi, un'autorità centrale seleziona un gruppo opportuno di utenti tra quelli disposti a partecipare. Nei programmi di DR, gli utenti che partecipano vengono remunerati con pagamenti diretti o riduzioni delle loro bollette. Tuttavia, il reclutamento di tali utenti prevede dei costi dovuti a diverse attività come il marketing, l'educazione e il supporto dei clienti a tali programmi, e le tecnologie necessarie al funzionamento. È necessario per l'operatore, incoraggiare la partecipazione ai programmi specifici di DR, con una certa priorità, gruppi di utenti più adatti a tali scopi, al fine di non incorrere in costi considerevoli dovuti alla non corretta selezione dei clienti *target*, oltre alla poca capacità di DR dovuta all'indisponibilità di molti utenti. Infatti, un'ulteriore inefficienza dei programmi di DR che potrebbe insorgere a causa della non corretta identificazione di opportuni utenti, è di incentivare e remunerare utenti che inizialmente sono dichiarati partecipanti ai programmi, ma che in seguito, quando sono chiamati a cambiare i loro consumi (ad esempio il giorno prima) risultano non disponibili. La selezione degli utenti target, da parte dell'aggregatore o del DSO, viene effettuata sulla base dei modelli di consumo degli utenti, suddividendo i propri clienti in gruppi con modelli di carico simili e identificando, sulla base dei programmi di DR che si vuole avviare, il gruppo di utenti che soddisfi al meglio le caratteristiche richieste. Tuttavia, questa selezione dovrà essere accompagnata da una classificazione dei propri clienti in funzione della stabilità dei loro consumi nei diversi giorni (a partire dai dati storici), al fine di assicurarsi che gli utenti selezionati rispettino al meglio le caratteristiche dei consumi richieste durante l'evento della DR, poiché, un uso casuale dell'energia elettrica di un gruppo di utenti potrebbe impedire il raggiungimento degli obiettivi previsti dall'operatore.

4.1 Clustering per l'identificazione di utenti target

Per assicurare il successo della DR, l'aggregatore o l'operatore del sistema seleziona un gruppo di clienti *target*, ovvero di potenziali utenti che potrebbero rispondere agli stimoli dei programmi, al fine di incoraggiare la riduzione dei consumi e portare benefici all'intero sistema. L'identificazione di tali clienti viene normalmente effettuata, prima tramite una segmentazione tra tutti gli utenti dell'operatore in base alla similarità tra i modelli dei loro consumi, e successivamente viene scelto a quali gruppi rivolgersi per soddisfare i requisiti richiesti dai specifici programmi di DR. Dunque, l'operatore si costruisce un *portfolio* clienti con diversi gruppi all'interno, ai quali può fornire delle opportunità differenziate: per esempio, se viene prevista un'ondata di calore durante il pomeriggio, l'operatore potrebbe proporre di ridurre i consumi incentivando per prima gli utenti che presentano i picchi di consumo in tale periodo, al fine di trarre maggiori benefici. Tuttavia, gli utenti residenziali presentano bassi consumi ed elevate variabilità tra i diversi giorni, il che rende più difficoltosa l'efficacia della segmentazione di tali clienti. Per ovviare a queste problematiche, in letteratura sono state proposte differenti soluzioni come la selezione di utenti che presentano maggiori valori di stabilità dei consumi oppure l'aggiunta nel processo di segmentazione sia di fattori che riguardano la taglia dei consumi degli utenti che le informazioni sulla posizione dei picchi di domanda. La corretta selezione degli utenti permette di ottenere, dalla partecipazione di utenti che presentano un maggior potenziale di DR, un vantaggio più significativo dei programmi. Gli autori dell'articolo [4] propongono una metodologia basata sulla valutazione della *consistenza* tra i modelli dei

consumi di un utente per diversi giorni consecutivi. Tale misura permette di scartare utenti con alte variabilità nei consumi essendo poco prevedibili e successivamente di selezionare gli utenti che presentano un picco di domanda in un periodo di tempo specifico per i programmi di DR. Tuttavia, molto frequente è l'applicazione di tecniche avanzate di *data mining* sui modelli dei consumi, al fine di estrarre caratteristiche comportamentali di differenti utenti. In particolare, l'utilizzo dei metodi di *clustering* su curve di carico giornaliere di un utente, permette di raggruppare curve simili in un unico cluster, identificando un andamento tipico dei consumi. Tali metodi, quindi, possono essere utili nella segmentazione degli utenti creando differenti gruppi con comportamento simile, e successivamente, dai risultati di tale processo, è possibile la selezione di utenti adatti ai programmi di DR. Gli autori dell'articolo [1] propongono una metodologia a due stadi, ovvero dopo aver eseguito le solite elaborazioni dei dati (data preprocessing, data selection e data cleansing), e aver individuato tramite un'analisi di correlazione le caratteristiche utili per rappresentare le curve di carico (in particolare l'ora di picco e il consumo giornaliero), normalizzano con *min-max normalization* tali caratteristiche ed effettuano il primo stadio di clustering utilizzandole come variabili d'ingresso. Questo stadio è utile a raggruppare gli utenti sia in base alla taglia dei loro consumi, e quindi al contributo che possono fornire nei programmi di DR, e sia in base a quando si verifica il picco di domanda, informando così se gli utenti si trovano nelle loro case e quindi sapere quando avviare tali programmi. Successivamente viene applicato, su ogni gruppo formato precedentemente, un secondo stadio di clustering direttamente sulle curve di carico normalizzate degli utenti appartenenti, al fine di migliorare la separazione dei diversi utenti in base ai modelli dei loro consumi e migliorare i risultati conseguibili dall'avviamento di specifici programmi di DR su gruppi opportunamente selezionati. Gli autori confrontano i risultati ottenuti con differenti metodi di clustering come k-means, SOM e fuzzy k-means (FCM) sia per metodologie a singolo che a doppio stadio, ottenendo risultati migliori con i secondi, in quanto consentono di selezionare gruppi di utenti più propensi a sfruttare al meglio i programmi utili per la riduzione della domanda. La metodologia che è stata scelta di utilizzare per l'identificazione di utenti adatti a programmi specifici di DR è una soluzione utile per trattare curve di carico residenziali, in quanto consente di trasformarle in rappresentazioni simboliche attraverso una versione modificata della tecnica *symbolic aggregate approximation* (*SAX*), a cui è possibile applicare metodi di clustering adatti a gestire anche dati categoriali come lo *hierarchical*, il *k-modes* o il *DBSCAN*, metodo basato sulla densità. Tale tecnica presenta diversi vantaggi, come la possibilità di suddividere l'asse tempi in periodi specifici nell'arco della giornata dai quali è possibile ottenere caratteristiche d'interesse, e la possibilità di trattare una serie limitata di simboli piuttosto che lavorare con valori numerici dei consumi; ciò rende tale tecnica adatta per analizzare un insieme di curve di carico residenziali anche di grandi dimensioni. Nell'articolo [5] gli autori, dopo aver normalizzato le curve di carico giornaliere di tutti gli utenti e aver individuato le condizioni di carico, determinano degli intervalli di tempo durante il giorno secondo delle analisi statistiche, ed effettuano una trasformazione intermedia dei dati con una tecnica usata per ridurre la dimensione dei dati chiamata *piecewise aggregate approximation* (*PAA*), facendo corrispondere ad ogni intervallo precedentemente definito il valore medio dei dati che gli appartengono. Dopodiché effettuano la trasformazione dei dati nella rappresentazione simbolica SAX, discretizzando l'asse delle ordinate delle curve di carico normalizzate, in intervalli definiti da una serie di *breakpoint* determinati sulla base dei quantili dell'intero dataset. Ad ogni intervallo definito da tali punti, viene associato un simbolo (ad esempio una lettera dell'alfabeto),

generando in tal modo, per ogni curva di carico, una stringa di elementi (una *parola*) pari al numero di intervalli temporali definiti nella giornata. Il calcolo della distanza che viene utilizzato per valutare la similarità tra due parole, tiene conto sia della possibilità di avere diverse lunghezze dei periodi temporali che della presenza di punti massimi e minimi nelle serie temporali. I risultati del clustering, basati sul calcolo di tali distanze, in accordo con le necessità del DSO o dell'aggregatore, sono utili per assistere i programmi di DR residenziali dalla suddivisione delle diverse curve di carico di tutti gli utenti in classi di modelli di consumo. Tuttavia, per tenere in considerazione la stabilità dei consumi degli utenti nei diversi giorni successivi, è opportuno classificarli secondo delle metriche basate su concetti di *entropia*, distinguendo in tal modo utenti con maggiore entropia, ovvero con curve di carico molto variabili da un giorno all'altro, e quindi adatti per programmi di DR basati su incentivi, dagli utenti con maggiore stabilità nel tempo, quindi con livelli bassi di entropia, adatti per programmi di DR basati sul prezzo. Questa metodologia, dunque, permette di superare la problematica che incorre nell'implementazione dei programmi di DR per utenti residenziali, dovuta alla scarsa conoscenza del comportamento degli utenti sia nella singola giornata che nei differenti giorni, offrendo all'operatore differenti gruppi di utenti suddivisi in base ai modelli di consumo e al livello di stabilità, al fine di incentivare con priorità utenti target per gli scopi specifici.

4.2 Caso studio

Nel capitolo precedente è stata effettuata un'analisi della flessibilità su curve di carico aggregate realizzate a partire da un insieme di 570 "utenti" selezionato per 15 giorni feriali nella terza stagione, ed ottenuto a partire dal dataset *Smart* Home Dataset* per l'anno 2016. Lo stesso insieme di dati selezionato, con periodo di campionamento di 15 minuti[7], è stato utilizzato per fornire un'applicazione dei metodi di clustering su dati in rappresentazione SAX, al fine di identificare caratteristiche peculiari delle curve di carico degli utenti e successivamente effettuare una *entropy analysis* per classificare gli utenti in base alla loro stabilità dei loro consumi nei vari giorni, utile all'operatore dei programmi di DR per ottenere clienti target.

Preparazione e rappresentazione SAX dei dati

Una volta selezionato il dataset da analizzare, prima di effettuare la normalizzazione delle curve di carico, è stato applicato un filtro per poter scartare utenti che, per almeno 5 giorni, presentano consumi inferiori a 2.5 kWh giornalieri. Infatti se ci dovesse essere un utente che per alcuni giorni non si trova nell'abitazione, l'effetto della normalizzazione delle curve di carico farebbe perdere l'informazione sulle taglie dei consumi (che per tali giorni risulterebbero bassi), classificando le curve di carico dell'utente come simili nei diversi giorni e l'utente verrebbe considerato stabile. Successivamente è stata eseguita, per ogni utente, la normalizzazione di tutte le curve di carico giornaliere (*Daily Load Curves DLC*) rispetto al valore massimo della potenza registrata nella giornata, al fine di mettere in evidenza le informazioni sugli andamenti e le variazioni del carico

[7] La scelta di utilizzare dati al quarto d'ora rispetto ai dati forniti al minuto, oltre ad avere una migliore rappresentazione grafica delle curve di carico giornaliere, permette di ottenere, come verrà visto in seguito, valori dei *breakpoint* più elevati in quanto la CDF, seppur realizzata con i valori in rappresentazione PAA, risulta essere meno ripida, per via della normalizzazione iniziale sulle curve di carico, che viene effettuata su valori mediati a 15 min, e quindi meno influenti agli alti picchi improvvisi dei dati al minuto.

nell'arco della giornata. Sulla base dei periodi specifici della giornata definiti nel capitolo precedente, è stata eseguita la tecnica *PAA* per ridurre la dimensione dei dati da fornire in ingresso nelle analisi successive. Per questa applicazione il periodo durante le ore notturne è stato suddiviso in *periodo 1* e *periodo 5*:

- *Periodo 1*: da 00:00 a 6:00 (dal 1° al 24° quarto d'ora)
- *Periodo 2*: dalle 6:00 alle 9:30 (dal 25° al 38° quarto d'ora)
- *Periodo 3*: dalle 9:30 alle 15:30 (dal 39° al 62° quarto d'ora)
- *Periodo 4*: dalle 15:30 alle 23:00 (dal 63° al 92° quarto d'ora)
- *Periodo 5*: dalle 23:00 a 00:00 (dal 93° al 96° quarto d'ora)

Normalmente la tecnica *PAA* trasforma i dati dal dominio del tempo originario, in uno spazio ridotto suddiviso in intervalli di tempo uguali, sostituendo i valori che appartengono ad ogni intervallo nei loro valori medi. In questo caso, ogni curva di carico giornaliera di ogni utente rappresentata da un vettore $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ di 96 elementi (valori al quarto d'ora) viene trasformata in un vettore $\mathbf{z} = (z_1, z_2, \ldots, z_H)$ di 5 valori (con $N < H$), corrispondenti ai valori medi dei valori appartenenti in ognuno dei 5 periodi temporali di diversa ampiezza, secondo la seguente relazione:

$$z_h = \frac{1}{k_h} \sum_{y_n \in T_h} y_n \tag{4.2.1}$$

dove z_h è il valore dell'h-esimo elemento di $\mathbf{z}$, corrispondente a T_h ovvero l'h-esimo periodo della giornata di ampiezza k_h (numero di elementi del vettore $\mathbf{y}$ contenuti nell'intervallo h-esimo).

Una volta ottenuto il dataset ridotto $M \times H$ di tutte le curve di carico di ogni utente, i dati sono stati trasformati nella rappresentazione simbolica con la tecnica *SAX*, suddividendo l'asse delle ampiezze delle DLC normalizzate e associando una lettera dell'alfabeto ad ogni intervallo dell'asse. La suddivisone in Q intervalli viene determinata dal calcolo dei $Q - 1$ quantili corrispondenti alla distribuzione statistica dell'intero dataset in rappresentazione PAA; ovvero, una volta costruita la funzione di distribuzione delle probabilità (*CDF*) dell'intero dataset, l'asse delle ordinate viene ripartito in Q regioni con stessa probabilità, a cui corrispondono in ascisse i quantili utilizzati come *breakpoint* per la suddivisione dell'asse delle ampiezze delle DLC, come è mostrato in *Figura 4.2.1.* Ad ogni intervallo definito dai breakpoint (quantili della CDF) dell'asse delle ampiezze delle DLC, vengono associate Q lettere dell'alfabeto, come rappresentazione simbolica. Ogni elemento delle DLC in rappresentazione SAX, è ottenuto a partire dai valori delle curve di carico giornaliere rappresentate da una serie temporale di dimensioni ridotte (dati PAA), associando ad ogni valore la lettera corrispondente all'intervallo in cui esso appartiene, trasformando così le serie temporali in *parole* di H lettere.
In *Figura 4.2.2*, è stata riportata una curva di carico di un utente, in un generico giorno del periodo considerato, rappresentato in serie temporale originaria (96 quarti d'ora) e nella sua

rappresentazione PAA, evidenziando gli intervalli dell'asse delle ampiezze corrispondenti alle lettere: "a, b, c, d, e, f, g, h". La rappresentazione *SAX* di tale curva di carico risulta: "b h e g b".

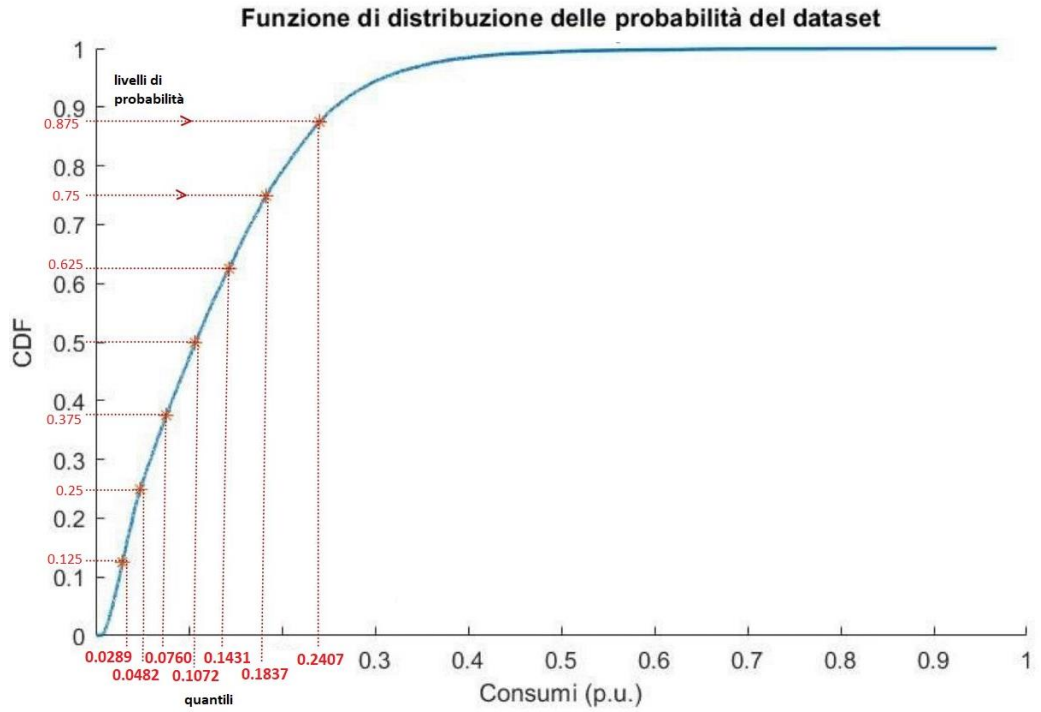


Figura 4.2.1 - CDF del dataset in rappresentazione PAA e i relativi 7 quantili corrispondenti ai livelli di probabilità

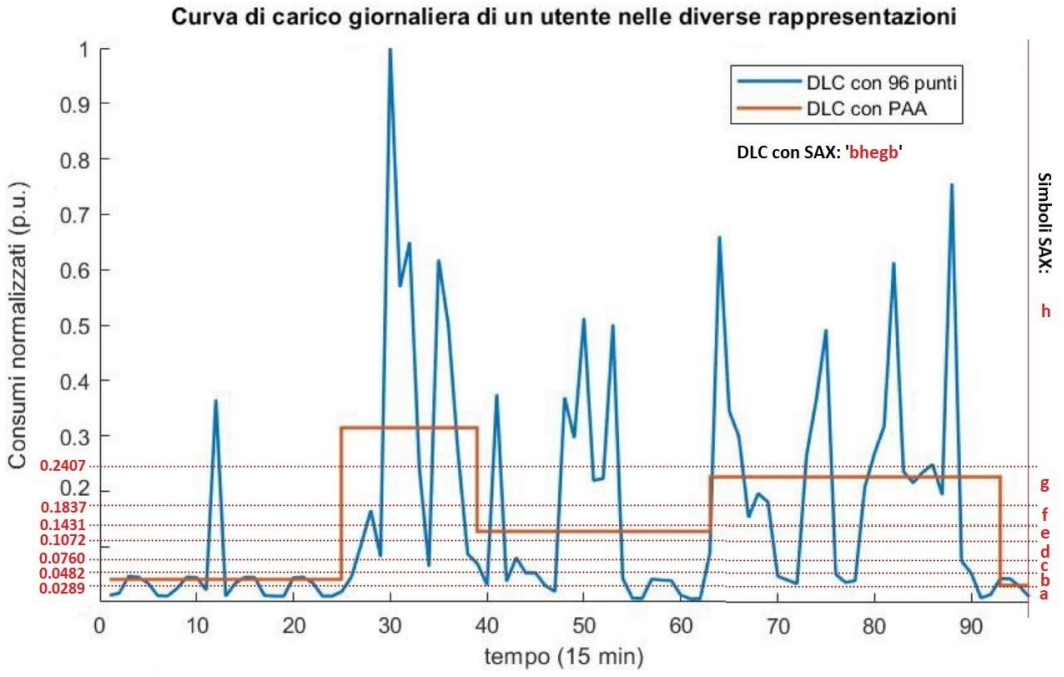


Figura 4.2.2 - DLC dell'utente 17 al 14° giorno in rappresentazione originale e in PAA e intervalli definiti dai breakpoint

Clustering per la segmentazione delle curve di carico residenziali

Una volta ottenuta la rappresentazione *SAX* del dataset, al quale ogni curva di carico di ciascun utente è associata una parola, è opportuno definire un nuovo metodo per la valutazione della distanza tra i diversi oggetti caratterizzati da dati categoriali. La distanza tra i diversi simboli può essere valutata associando ad ogni q-esimo intervallo dell'asse delle ampiezze definito precedentemente, un indicatore ind_q:

$$ind_q = \frac{z_{q-1} - z_q}{2} \tag{4.2.2}$$

dove z_q e z_{q-1} sono gli estremi superiore e inferiore dell'q-esimo intervallo dell'asse delle ampiezze considerato per la rappresentazione simbolica. La distanza tra due generiche lettere α e β, associate ciascuna ad un rispettivo intervallo (q_α e q_β) è definita come:

$$d_{\alpha\beta} = \mathrm{dist}(\alpha, \beta) = \left| ind_{\mathrm{q}_\alpha} - ind_{\mathrm{q}_\beta} \right| \tag{4.2.3}$$

La distanza tra due curve di carico giornaliere r e s, rappresentate rispettivamente dalle *parole* $W_r = (\alpha_1, \ldots, \alpha_H)$ e $W_s = (\beta_1, \ldots, \beta_H)$, viene calcolata dalla seguente espressione[8]:

$$MINIDIST(W_r, W_s) = \sqrt{\sum_{h=1}^{H} k_h \mathrm{d}^2_{\alpha_h \beta_h}} \tag{4.2.4}$$

dove k_h è l'ampiezza dell'h-esimo periodo temporale delle DLC in rappresentazione PAA.

Le distanze tra le varie coppie delle M curve di carico giornaliere sono state calcolate, al fine di realizzare la matrice delle distanze (o *matrice di similarità*) di dimensioni $M \times M$, utilizzata come ingresso agli algoritmi di clustering. L'analisi dei gruppi è stata realizzata tramite lo hierarchical clustering con average linkage, in quanto il metodo gerarchico è utilizzabile per dati categoriali ed inoltre permette di utilizzare come parametro d'ingresso la matrice delle distanze, ricavata a partire dai dati SAX ottenuti dagli 8 simboli possibili. Il numero ottimale di cluster da generare è valutato dal calcolo degli indici di validità MIA e CDI, ed è pari a 52 cluster, come mostrato in *Figura 4.2.3*. I risultati ottenuti dal clustering per il dataset considerato sono presentati in *Figura 4.2.4*; per valutare la forma e i modelli dei consumi per ogni gruppo, sono stati riportati i centri di ciascun cluster, ottenuti dalla media delle DLC appartenenti all'interno.

[8] La distanza *MINIDIST* è generalmente applicata su una suddivisione equispaziata dell'asse tempi, e la definizione della distanza tra due lettere viene definita in modo differente [5]. In questa trattazione sono apportate alcune modifiche alla definizione di tale distanza, al fine di renderla adatta alle differenti ampiezze dei periodi di tempo a cui è suddiviso l'asse delle ascisse e al fine di ottenere migliori risultati del clustering.

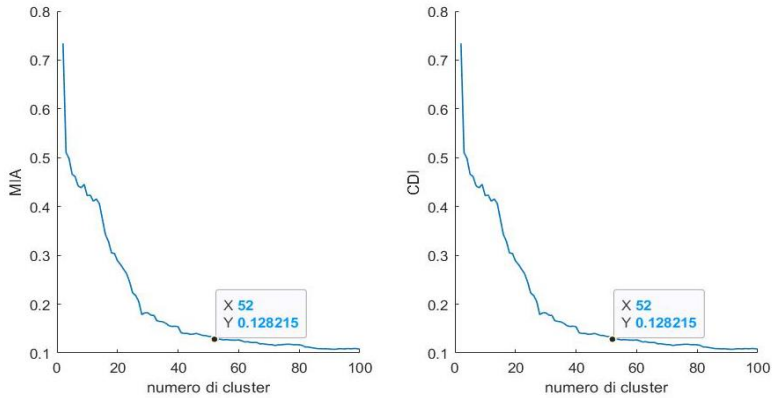


Figura 4.2.3 - Indici di validità dei risultati del clustering al variare del numero di cluster generati

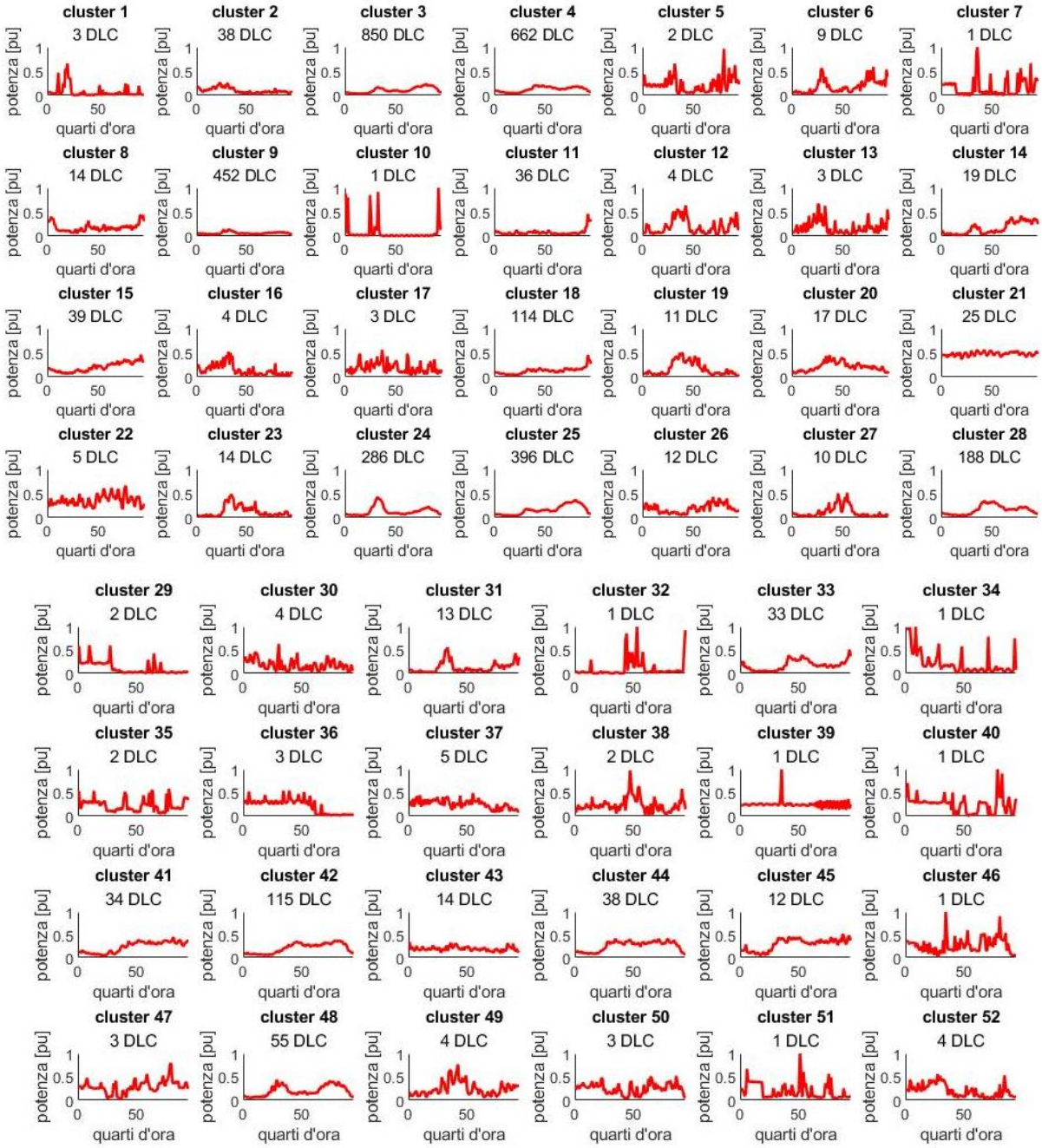


Figura 4.2.4 - Centri dei cluster ottenuti dallo hierarchical clustering con average linkage

Escludendo i gruppi con una sola curva di carico all'interno, come i cluster 7, 10, 32, 34, 39, 40, 46 e 51 si riescono a identificare i diversi modelli dei consumi in base alla forma e al periodo in cui si presentano i picchi di domanda, distinguendo le seguenti categorie:

- *Picco mattutino*: in questa categoria fanno parte le DLC che presentano maggiori consumi nelle prime ore della giornata (periodo 2) come i cluster 2, 9, 23 e 24. In particolare il gruppo 24 mostra un picco più definito e di valore elevato, in quanto, probabilmente, molte apparecchiature di tali DLC sono utilizzate sempre negli stessi intervalli di tempo, a differenza dei gruppi 2 e 23 in cui il consumo è più distribuito nella mattinata.

- *Picco di metà giornata*: i gruppi come 19, 20, 27 e 28 fanno parte di questa categoria, poiché i loro picchi di consumo, generalmente più estesi, si verificano nel primo pomeriggio oppure con ritardo rispetto alle ore mattutine (periodo 3 della giornata).

- *Picco serale*: i modelli dei consumi dei cluster 14 e 25 presentano dei picchi nel periodo serale (periodo 4). In particolare il gruppo 25 mostra un significante e decisivo aumento dei consumi in tale periodo, rispetto al resto della giornata.

- *Picco notturno*: in questa categoria fanno parte i cluster 8, 11, 15 e 18 in cui i picchi di consumo risultano essere intorno la mezzanotte o oltre (periodo 1 e 5 della giornata). Il gruppo 15 presenta una lenta crescita dei consumi dalla mattina fino alla notte, contrariamente ai gruppi 8, 11 e 18 che mostrano un picco decisivo nelle ore notturne.

- *Doppi picchi*: i gruppi che presentano un duplice picco in periodi differenti della giornata sono stati classificati in questa categoria. I cluster 3, 6 e 48 presentano dei picchi nelle ore serali e durante la mattinata, il gruppo 4 ha un picco serale e uno a metà giornata, il cluster 33 presenta i picchi a metà giornata e durante la notte, mentre, il cluster 31 li ha durante la mattinata e nelle ore notturne, e il gruppo 26, invece, li presenta in serata e durante la notte.

- *Consumi costanti*: in questa categoria sono stati raggruppati tutti i cluster che presentano dei modelli di consumo costanti nell'intera giornata o in porzioni di essa (ore diurne), in cui non si evidenziano picchi di consumo rilevanti. A tale categoria appartengono i cluster 21 e 43 che hanno dei profili completamente piatti lungo tutta la giornata, e i cluster 41, 42, 44 e 45, i quali presentano consumi relativamente costanti nelle ore diurne e nelle prime ore notturne.

I cluster 1, 5, 12, 13, 16, 17, 22, 29, 30, 35, 36, 37, 47, 49, 50 e 52 non sono stati presi in considerazione poiché presentano un numero di elementi piccolo per cui è anche difficile classificarli correttamente nelle categorie sopra elencate, in quanto molto variabili. Inoltre, si può notare che il gruppo con maggior numero di curve di carico risulta essere il cluster 3 con 850 DLC presenti (circa il 23,8% delle totali), il cui modello dei consumi risulta essere coerente con quello ottenuto nel capitolo precedente; infatti, evidenzia un comportamento generale dei consumi degli utenti nei giorni feriali durante il periodo della stagione 3 presa in esame.

I risultati del clustering sono utili per aiutare il DSO o l'aggregatore che intende avviare programmi specifici di DR, in quanto forniscono informazioni sui modelli di consumo degli utenti. Tuttavia, dalle sole considerazioni effettuate sui gruppi generati non è possibile stabilire quali utenti è meglio incentivare prima di altri per far modificare i loro carichi in specifici periodo di tempo, poiché le DLC dei differenti giorni per ogni utente potrebbero essere distribuite in più cluster, il che evidenzia un comportamento meno stabile dell'utente nei diversi giorni, e quindi poco affidabile a rispettare i requisiti richiesti per programmi di DR specifici. Un'ulteriore analisi dei risultati ottenuti dal clustering è necessaria per classificare al meglio gli utenti in base alla loro stabilità nei consumi.

DR targeting: classificazione degli utenti sulla base dell'entropia

Per poter identificare e suddividere i diversi utenti e classificarli come buoni candidati per programmi specifici di DR, si utilizzano i risultati ottenuti dal clustering precedente e si valutano le appartenenze delle curve di carico, di ogni singolo utente, nei rispettivi gruppi. Un utente che presenta le proprie curve di carico dei diversi giorni in molti cluster differenti, indica un comportamento poco stabile e irregolare per cui non adatto a programmi di DR basati sul prezzo. Contrariamente, un utente che si presenta in pochi cluster, indica un comportamento comune nei diversi giorni, mostrando una maggiore stabilità dei propri consumi e ciò, insieme all'informazione del suo modello di consumi simile a quello di altri utenti, consente di applicare i programmi di DR specifici. Al fine di classificare gli utenti sulla base della loro variabilità del comportamento nei rispettivi giorni del periodo considerato, viene adottato il concetto di *entropia* secondo la teoria dell'informazione. La misura dell'entropia, chiamata anche *entropia di Shannon*, viene definita come il valore atteso dell'autoinformazione[9], ovvero l'informazione media contenuta in ciascun evento x_c degli C emessi da una sorgente X; nel caso di variabile aleatoria discreta, essa è definita come media pesata, con le rispettive probabilità $\mathcal{P}(x_c)$, di ogni evento:

$$\mathcal{H}(\mathrm{X}) = -\sum_{c=1}^{C} \mathcal{P}(x_c)\ \log_{\mathrm{b}}(\mathcal{P}(x_c)) \qquad (4.2.5)$$

Per valutare il grado di entropia di un utente avente una distribuzione delle sue DLC nelle C classi corrispondenti ai C cluster generati dal clustering, viene utilizzata l'equazione (4.2.5), considerando $\mathcal{P}(x_i)$ la probabilità che una DLC dell'utente appartenga alla c-esima classe (cluster). Tali probabilità rappresentano le frequenze relative di ogni classe, ovvero per ogni cluster c vengono contate il numero di DLC dell'utente presenti in esso e diviso per il numero di giorni considerati nel periodo di osservazione del dataset (15 gironi per il dataset in studio), corrispondente al numero totale di DLC dell'utente. Agli utenti che presentano tutte le proprie curve di carico in un unico cluster, corrisponde un valore di entropia nullo, in quanto la frequenza relativa, è unitaria per quel cluster, e nulla negli altri. Se il numero di classi è maggiore o uguale

[9] Nella teoria dell'informazione, si definisce *autoinformazione* di un evento *x,* l'informazione sulla quantità d'incertezza associata all'evento e calcolata come:

$$\mathcal{I}(x) = -\log_b(\mathcal{P}(x))$$

dove $\mathcal{P}(x)$ è la probabilità che l'evento *x* accada e *b* è la base del sistema di numerazione posizionale (con *b*=2 l'unità di misura è il bit, relativo al sistema binario). Si nota che quanto più probabile è un evento, tanto minore sarà l'autoinformazione.

al numero totale di DLC dell'utente, il valore massimo dell'entropia, corrisponde alla condizione per cui la probabilità di appartenenza ai rispettivi cluster è la stessa e pari all'inverso del numero totale di curve di carico dell'utente. Nella *Figura 4.4.5* è riportata la tabella relativa ai valori dell'entropia per ogni utente del dataset in studio, ordinata da utenti con minor entropia a utenti con valori elevati. Inoltre sono stati riportati, per ogni utente, il numero di classi in cui sono contenute tutte le DLC dei diversi giorni dell'utente, la classe a cui corrisponde il maggior numero di curve di carico e la frequenza relativa, espressa in percentuale, delle DLC dell'utente appartenenti alla classe con maggiore numero di curve, rispetto alle DLC totali dell'utente.

Utente	Entropia	Tot. classi in cui l'utente appartiene	Classe con maggiore DLC	Percentuale di appartenenza nella classe con maggiore DLC
69	0.3927	2	3	86.67
108	0.4851	3	3	86.67
118	0.7648	3	3	73.33
177	0.8572	4	24	73.33
228	0.8572	4	9	73.33
111	0.8609	3	24	66.67
127	0.9496	5	24	73.33
181	0.9496	5	9	73.33
199	0.9496	5	3	73.33
173	0.9503	3	4	60.00
145	0.9533	4	4	66.67
179	0.9533	4	3	66.67
31	0.9882	4	24	66.67
78	0.9882	4	25	66.67
165	0.9882	4	3	66.67
196	1.0200	4	3	60.00
226	1.0438	3	9	46.67
238	1.0438	3	3	46.67
44	1.0625	4	3	53.33
182	1.0625	4	3	53.33
231	1.0625	4	9	53.33
150	1.0776	4	3	60.00
104	1.0833	4	3	46.67
113	1.0833	4	25	46.67
227	1.0833	4	3	46.67
229	1.0833	4	4	46.67
33	1.1369	4	3	53.33
63	1.1369	4	3	53.33
65	1.1369	4	4	53.33
114	1.1369	4	4	53.33
134	1.1369	4	9	53.33
26	1.1711	4	4	46.67
107	1.1945	4	3	53.33
36	1.2049	5	9	60.00
130	1.2049	5	3	60.00
225	1.2049	5	3	60.00
82	1.2293	5	3	53.33
161	1.2293	5	3	53.33
105	1.2351	4	3	40.00
142	1.2520	4	9	40.00
186	1.2520	4	3	40.00
93	1.2635	5	24	46.67
135	1.2635	5	4	46.67
139	1.2654	4	3	33.33
56	1.2700	4	3	40.00
110	1.2869	5	25	53.33
122	1.2869	5	4	53.33
132	1.2869	5	3	53.33
162	1.2869	5	4	53.33
183	1.2869	5	9	53.33
190	1.2869	5	24	53.33
1	1.2973	6	24	60.00
9	1.2973	6	3	60.00
55	1.2973	6	9	60.00
131	1.2973	6	9	60.00
207	1.2973	6	25	60.00
223	1.2973	6	24	60.00
72	1.3218	5	4	53.33
208	1.3218	5	24	53.33
59	1.3379	5	3	46.67
201	1.3379	5	3	46.67
58	1.3605	5	3	46.67
62	1.3605	5	24	46.67
53	1.3624	5	4	40.00
141	1.3624	5	3	40.00
211	1.3624	5	4	40.00
71	1.3793	6	4	53.33
156	1.3793	6	3	53.33
159	1.3793	6	25	53.33
176	1.3954	5	3	46.67
60	1.4019	5	4	40.00
61	1.4019	5	42	40.00
67	1.4019	5	3	40.00
80	1.4019	5	24	40.00
92	1.4019	5	3	40.00
219	1.4019	5	4	40.00
97	1.4142	6	44	53.33
148	1.4142	6	3	53.33
155	1.4142	6	4	53.33
214	1.4142	6	3	53.33
50	1.4154	5	3	33.33
73	1.4154	5	9	33.33
129	1.4303	6	4	46.67
191	1.4303	6	3	46.67
209	1.4303	6	3	46.67
40	1.4322	5	3	33.33
81	1.4322	5	9	33.33
37	1.4368	5	4	40.00
125	1.4368	5	3	40.00
235	1.4368	5	3	40.00
41	1.4549	6	3	40.00
99	1.4549	6	3	40.00
202	1.4549	6	4	40.00
14	1.4878	6	3	46.67
68	1.4878	6	4	46.67
112	1.4878	6	9	46.67
102	1.4898	5	3	33.33
117	1.4898	5	4	33.33
138	1.4898	5	4	33.33
204	1.4898	5	3	33.33
160	1.5066	7	9	53.33
221	1.5066	7	3	53.33
57	1.5227	6	4	46.67
237	1.5227	6	25	46.67
27	1.5292	5	25	26.67
29	1.5292	6	25	40.00
180	1.5292	6	3	40.00
187	1.5292	6	3	40.00
212	1.5292	6	4	40.00
43	1.5427	6	3	33.33
48	1.5427	6	24	33.33
64	1.5427	6	3	33.33

Figura 4.2.5 - Classifica degli utenti in base ai valori di entropia - Parte 1

94	1.5519	6	9	40.00
195	1.5519	6	3	40.00
2	1.5822	6	3	33.33
121	1.5822	6	9	33.33
154	1.5822	6	3	33.33
164	1.5822	6	4	33.33
188	1.5822	6	3	33.33
189	1.5822	6	4	33.33
205	1.5822	6	3	33.33
6	1.5868	6	3	40.00
86	1.5868	6	9	40.00
116	1.5868	6	3	40.00
175	1.5868	6	25	40.00
30	1.6151	7	3	46.67
126	1.6151	7	3	46.67
174	1.6151	7	28	46.67
15	1.6171	6	3	33.33
18	1.6217	7	3	40.00
35	1.6351	7	3	33.33
77	1.6397	6	25	33.33
16	1.6566	6	3	26.67
79	1.6566	6	3	26.67
89	1.6566	6	3	26.67
171	1.6566	6	3	26.67
45	1.6746	6	3	33.33
3	1.6792	7	25	40.00
106	1.6792	7	4	40.00
123	1.6792	7	4	40.00
144	1.6792	7	9	40.00
197	1.6792	7	25	40.00
151	1.6914	6	4	26.67
192	1.6914	6	3	26.67
75	1.7095	7	9	33.33
84	1.7095	7	3	33.33
172	1.7095	7	9	33.33
185	1.7095	7	4	33.33
210	1.7095	7	9	33.33
213	1.7141	6	28	26.67
87	1.7141	7	4	40.00
163	1.7141	7	3	40.00
10	1.7321	7	4	33.33
28	1.7321	7	4	33.33
203	1.7321	7	28	33.33
146	1.7367	6	4	20.00
88	1.7490	7	9	26.67
178	1.7490	7	3	26.67
19	1.7670	7	4	33.33
49	1.7670	7	3	33.33
100	1.7670	7	25	33.33
153	1.7670	7	25	33.33
95	1.7839	7	3	26.67
170	1.7839	7	3	26.67
236	1.7839	7	3	26.67
133	1.8000	9	9	46.67
147	1.8019	7	48	33.33
24	1.8019	8	3	33.33
234	1.8019	8	4	33.33
7	1.8065	7	4	26.67
54	1.8065	7	4	26.67
220	1.8065	7	4	26.67
218	1.8065	8	4	40.00
11	1.8414	7	18	26.67
21	1.8414	7	24	26.67
22	1.8414	7	4	26.67
124	1.8414	7	4	26.67
143	1.8414	7	3	26.67
224	1.8414	7	4	26.67
232	1.8414	7	4	26.67
38	1.8594	8	9	33.33
83	1.8594	8	9	33.33
101	1.8594	8	4	33.33
140	1.8594	8	25	33.33
167	1.8594	8	15	33.33
184	1.8594	8	4	33.33
91	1.8763	7	25	26.67
103	1.8763	7	4	26.67
42	1.8763	8	3	26.67
76	1.8763	8	3	26.67
96	1.8763	8	25	26.67
119	1.8763	8	3	26.67
136	1.8763	8	9	26.67
168	1.8763	8	2	26.67
5	1.8943	8	3	33.33
13	1.8943	8	4	33.33
23	1.8943	8	4	33.33
157	1.8943	8	18	33.33
169	1.8943	8	4	33.33
233	1.8943	8	9	33.33
120	1.8989	8	4	26.67
206	1.8989	8	4	26.67
230	1.8989	9	9	40.00
17	1.9338	8	24	26.67
20	1.9338	8	3	26.67
51	1.9338	8	3	26.67
52	1.9338	8	28	26.67
90	1.9338	8	4	26.67
198	1.9338	8	25	26.67
32	1.9518	9	4	33.33
194	1.9518	9	48	33.33
46	1.9565	8	4	20.00
222	1.9565	8	4	20.00
47	1.9687	8	28	26.67
137	1.9687	8	4	26.67
34	1.9687	9	9	26.67
8	1.9867	9	28	33.33
152	1.9867	9	9	33.33
193	1.9867	9	25	33.33
74	1.9913	8	3	20.00
217	1.9913	8	4	20.00
166	2.0489	9	9	20.00
66	2.0611	9	9	26.67
85	2.0611	9	9	26.67
128	2.0611	9	9	26.67
200	2.0611	9	4	26.67
12	2.0838	9	3	20.00
25	2.0838	9	4	20.00
98	2.0838	9	3	20.00
115	2.1186	9	4	20.00
158	2.1186	9	25	20.00
109	2.1762	10	3	20.00
215	2.1762	10	24	20.00
216	2.2111	10	25	20.00
149	2.2460	11	3	26.67
4	2.3035	11	25	20.00
39	2.3035	11	28	20.00
70	2.3959	12	3	20.00

Figura 4.2.6 - Classifica degli utenti in base ai valori di entropia - Parte 2

Si può notare che, generalmente, i valori dell'entropia si riducono al diminuire del numero delle classi a cui appartengono le curve di carico degli utenti; tuttavia, esso non può essere utilizzato come misura della stabilità dell'utente nei diversi giorni, poiché esistono condizioni in cui diversi utenti vengono classificati con uguale stabilità, pur avendo distribuzioni di frequenze differenti all'interno di uno stesso numero di classi. Un esempio è l'utente 56 che presenta una distribuzione di frequenza 6-5-2-2 in quattro classi differenti, e l'utente 177 che ha le sue DLC distribuite nelle quattro classi con frequenza 2-1-11-1. Pur essendo distribuite nello stesso numero di classi, l'utente 177 risulta più stabile, come evidenziato anche dal valore inferiore di entropia, e meglio

rappresentato dalla classe 24, contenente 11 delle sue curve di carico. Inoltre si può notare che tale utente potrebbe essere un buon candidato per i programmi di DR che intendono ridurre i consumi durante le prime ore della giornata, poiché, oltre ad avere una buona stabilità nei consumi, può essere associato alla classe 24 che corrisponde alla categoria "*picco mattutino"*. In *Figura 4.4.6* sono stati messi a confronto due utenti che presentano diversi livelli di entropia: le curve di carico giornaliere dell'utente con basso livello di entropia (utente 69), mostrano un andamento abbastanza stabile nei diversi giorni, il cui modello dei consumi è ben rappresentato dalla classe 3 avente un picco mattutino e uno serale, mentre le DLC dell'utente con alto grado di entropia (utente 70), si mostrano abbastanza instabili, variando da un giorno all'altro e non permettono di classificarli con un modello di consumi attendibile.

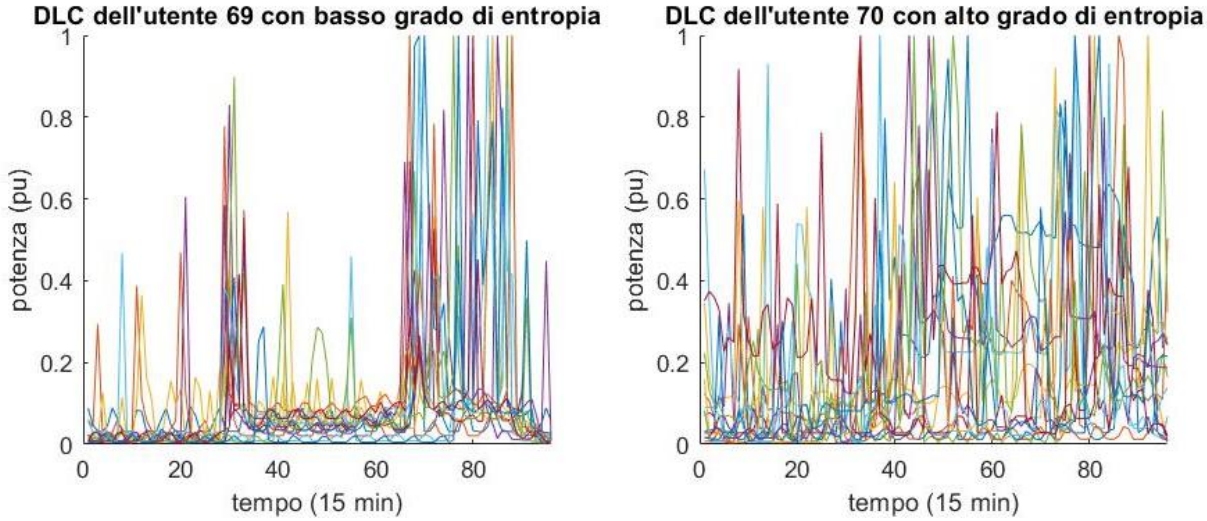


Figura 4.2.7 - Curve di carico nei diversi giorni di due utenti con basso e alto grado di entropia

Dunque la misura dell'entropia permette di classificare gli utenti sulla base della loro variabilità nei differenti giorni, ed insieme all'informazione sui modelli di consumi ottenuti dal clustering delle curve di carico giornaliere, consente di ottenere utenti target ai quali l'operatore di DR può offrire programmi (per esempio, *TOU: time of use*) che forniscono tariffe personalizzate per il tempo di utilizzo e quindi basate sui modelli dei consumi degli utenti, al fine di incentivare la riduzione del carico in intervalli di tempo specifici. Infatti utenti che presentano alta stabilità nei consumi e che appartengono allo stesso gruppo di modelli di consumo, producono dei picchi elevati negli stessi periodi della giornata, offrendo un significativo contributo sulla domanda aggregata che potrebbe essere ridotto, se nell'intervallo risulta esserci abbastanza flessibilità, tramite programmi di DR basati sul prezzo, incentivando tali utenti target a partecipare. Tuttavia, anche gli utenti classificati come poco stabili nei consumi possono essere utili per instaurare programmi di DR basati su incentivi, ad esempio programmi *DLC* (*Direct Load Control*), che permettono di controllare direttamente i carichi degli utenti senza influenzare troppo le loro abitudini e i loro confort, in quanto si tratta di utenti con comportamenti poco abitudinari e che quindi garantiscono un maggiore successo sulla richiesta di partecipazione.

4.3 Bibliografia

[1] *Lee E., Kim J., and Jang D.: 'Load Profile Segmentation for Effective Residential Demand Response Program: Method and Evidence from Korean Pilot Study'. Energies, vol. 13 (6), p. 1348, 2020.*

[2] *V. Rasouli, Á. Gomes and C. H. Antunes, 'Characterization of Aggregated Demand-side Flexibility of Small Consumers', 2020 International Conference on Smart Energy Systems and Technologies (SEST), pp. 1-6, 2020.*

[3] *G. Comodi, A.Bartolini, F. Carducci, M. Botticelli, 'Implementazione di un simulatore per demand response di uno smart district', Report RdS/PAR2016/007, (ENEA), 2017.*

[4] *Rashid H., Singh P., and Ramamritham K.: 'Revisiting Selection of Residential Consumers for Demand Response Programs'. Proceedings of the 4th ACM International Conference on Systems for Energy-efficient Built Environments, pp. 1-4, 2017.*

[5] *Rajabi A., Eskandari M., Jabbari Ghadi M., Ghavidel S., Li L., Zhang J., and Siano P.: 'A Pattern Recognition Methodology for Analyzing Residential Customers Load Data and Targeting Demand Response Applications'. Energy and Buildings, vol. 203, p. 109455, 2019.*

Conclusione

I moderni sistemi di distribuzione prevedono l'utilizzo di innovativi dispositivi di monitoraggio, come gli *smart meter*, i quali consentono di collezionare grandi quantità di dati dei consumi elettrici con elevate risoluzioni. Le tecniche di clustering risultano essere necessarie per poter gestire i dati in modo semplice ed efficiente, al fine di poter estrarre caratteristiche utili all'operatore del sistema facilitandolo in svariate mansioni. In questo elaborato sono state presentate differenti metodologie applicate nei sistemi elettrici, valutando i rispettivi processi di raggruppamento e fornendo i più comuni indici di validità dei risultati del clustering, con lo scopo di confrontare i risultati ottenuti dalle diverse tecniche e scegliere quello più efficiente per l'insieme di dati analizzato. Inoltre, questi indici si prestano utili anche nella scelta del numero ottimale di cluster da generare per ottenere migliori risultati. Tuttavia, per i dati riguardanti i consumi di utenti residenziali, i quali presentano alte variabilità nell'utilizzo delle loro apparecchiature, sono state illustrate differenti soluzioni; in particolare, per l'insieme di dati utilizzato in questo elaborato (*Smart* Home Dataset* per l'anno 2016), dopo i processi di "pulizia" e "ricostruzione "dei dati, sono state ricavate alcune caratteristiche riguardanti dei specifici periodi della giornata, corrispondenti ad attività comuni nel comportamento degli utenti nei giorni feriali e in condizioni di carico specifiche. Su tali caratteristiche, l'applicazione dei metodi di clustering come lo hierarchical clustering con average linkage, il quale ha presentato risultati migliori, consente di suddividere al meglio i gruppi di utenti residenziali ed isolare quelli che non utili alla partecipazione di programmi di DR. L'informazione sulla possibilità di far cambiare l'andamento dei consumi di un gruppo di utenti, al fine di poter gestire meglio la domanda, è di rilevante importanza per l'operatore del sistema; è dunque indispensabile effettuare una valutazione di flessibilità del comportamento degli utenti. La metodologia utilizzata nell'elaborato permette di definire degli indicatori, come il *FIAD* e il *MFIAD* che determinano in termini probabilistici la flessibilità a cambiare il comportamento collettivo degli utenti aggregati, e il *PFL* che permette di ottenere informazioni sui valori percentuali di domanda flessibile in ogni istante di tempo in cui esso è definito. Tali indicatori ricavano le informazioni dalle variazioni di domanda e sono utili all'operatore del sistema, per selezionare intervalli di tempo adatti ad avviare programmi di DR. Questi concetti sono stati applicati sui dati in studio valutando, per quattro scenari realizzati con differenti valori del livello di aggregazione e del periodo di campionamento, l'effetto che producono tali parametri sulle variazioni di carico delle curve aggregate. Si è visto che più lungo è l'intervallo di campionamento meno possibilità c'è nel seguire la dinamica delle variazioni riducendo in tal modo la possibilità di prendere decisioni utili a influenzare la domanda. Anche con l'aumento del livello di aggregazione, l'andamento del carico tende a diventare sempre più simile e quindi più liscio, ottenendo sempre meno informazioni sui cambiamenti dei carichi che invece potrebbero essere introdotte nelle curve aggregate. Per poter utilizzare in modo efficace le proprietà delle curve aggregate nella valutazione della flessibilità, è importante effettuare una giusta selezione dell'intervallo di campionamento e del livello di aggregazione. In particolare, se si vuole aumentare il numero di utenti aggregati, è necessario ridurre ancor più il tempo di campionamento per ottenere una rappresentazione accurata. Questo approccio può aiutare l'operatore del sistema nella selezione di una struttura di misura da cui ottenere i dati con una risoluzione ottimale e quindi trovare il giusto compromesso tra il livello di aggregazione e il periodo di campionamento. Inoltre, si è osservato che sia durante la mattina che durante la sera la flessibilità è molto bassa rispetto al resto della giornata, per via di un andamento molto più rigido della curva aggregata, dovuto al comune comportamento degli utenti in tali periodi, e ciò comporterebbe ad una probabile indisponibilità degli utenti a cambiare i propri consumi, rendendo

inefficienti le azioni proposte dell'operatore di DR ai clienti. Tuttavia, se si considerano intervalli di campionamento minori, si riescono ad ottenere maggiori informazioni sulla dinamica del carico, anche in periodi della giornata meno flessibili, individuando brevi intervalli di tempo in cui è possibile applicare programmi specifici di breve durata, utili ad esempio in condizioni di emergenza del sistema. I periodi di tempo in cui si dispone un'elevata flessibilità, tuttavia, potrebbero essere più utili all'operatore della DR, in quanto indicano un maggior numero di possibili candidati ai programmi che potrebbero accettare facilmente i cambiamenti nei loro consumi. Oltre all'individuazione di questi periodi, è vantaggioso per l'operatore incoraggiare alla partecipazione dei programmi, con una certa priorità, gruppi di utenti più adatti a tali scopi, al fine di non incorrere in costi dovuti alla non corretta selezione di clienti *target*, oltre alla poca capacità di DR dovuta all'indisponibilità di molti utenti. Avere a disposizione un portfolio di clienti raggruppati per categorie di modelli di consumo, insieme all'informazione sulla loro stabilità, potrebbe assicurare il successo dei programmi, fornendo un notevole contributo sulla gestione della domanda. Una segmentazione di tutte le curve di carico degli utenti è stata effettuata attraverso il clustering su una rappresentazione SAX dei dati, la quale presenta il vantaggio di trattare una serie limitata di simboli piuttosto che lavorare con valori numerici dei consumi, rendendo tale tecnica adatta ad analizzare un insieme di curve di carico residenziali anche di grandi dimensioni. I risultati ottenuti hanno permesso di distinguere modelli di consumo in base alla forma e al periodo in cui si presentano i picchi di domanda, definendo categorie specifiche che possono essere utilizzate dall'operatore per proporre offerte differenziate. Inoltre la classificazione degli utenti in base alla loro stabilità nei consumi, ottenuta dal calcolo dell'entropia, ha permesso di distinguere utenti maggiormente stabili nei diversi giorni, i quali risultano adatti ai programmi di DR basati sul prezzo, ed utenti il cui comportamento non è abitudinario, i quali si prestano meglio per programmi basati sugli incentivi. Questa metodica, dunque, permette di superare la problematica che incorre sull'implementazione di tali programmi per utenti residenziali, dovuta alla scarsa prevedibilità del comportamento degli utenti sia nella singola giornata che nei differenti giorni, offrendo all'operatore differenti gruppi di utenti suddivisi in base ai modelli di consumo e al livello di stabilità, al fine di incentivare con priorità utenti target per gli scopi specifici. La valutazione della potenziale flessibilità della domanda aggregata, individuando periodi di tempo in cui essa è maggiore, e la selezione di gruppi opportuni di utenti ai quali è possibile richiedere con priorità la partecipazione ai programmi di DR specifici, sono strumenti utili all'operatore per facilitarlo nell'avviamento di tali tecniche cercando di trarre anche maggiori benefici da esse. Tuttavia, per presentare un programma di DR sono necessari ulteriori informazioni, come la risposta effettiva dei singoli utenti agli stimoli forniti dall'operatore, i quali non tutti sono intenzionati a partecipare ai programmi e quelli che lo sono non sempre eseguono le specifiche richieste, inoltre alcuni potrebbero adottare strategie comportamentali al fine di aumentare i loro guadagni. Avere l'informazione sull'effettiva partecipazione degli utenti e sulla risposta che forniscono ad un evento di DR, permettono di valutare effettivamente il beneficio di tali programmi. Tuttavia, lo scopo della tesi è quello di illustrare metodologie ed applicazioni utili agli operatori di DR per agevolare l'avviamento di tali programmi su utenti residenziali; un effettivo beneficio sul sistema potrà essere valutato con studi successivi.