POLITECNICO DI TORINO

Corso di Laurea Magistrale
in
Ingegneria Elettrica

Tesi di Laurea Magistrale

Analisi e ricostruzione dei profili di carico di utenze elettriche

Relatore
Chiar.mo Prof. *Gianfranco Chicco*

Correlatore
Ing. *Andrea Mazza*

Candidato
Giovanna Bernardo
279363

Anno accademico 2020/2021

Sommario

Introduzione

Anche in uno studio ben progettato e controllato, l'acquisizione di un elevato numero di dati temporali può presentare delle irregolarità e pone le basi del problema dei dati errati o mancanti.

Per dato *errato* si intende un dato presente ma contrassegnato dal sistema di acquisizione dati come non attendibile, oppure un dato che viene dichiarato errato dopo aver eseguito una procedura di verifica delle caratteristiche dei dati misurati.

Per dato *mancante* (*missing data*) si intende il valore di una variabile che non viene memorizzato in un determinato periodo di osservazione. Il problema è abbastanza comune in tutte le ricerche e può avere un effetto significativo sulle conclusioni che si possono trarre, andando a ridurre il potere statistico di uno studio, producendo stime distorte e per ultimo, portando a conclusioni non valide.

Il presente lavoro ha come obiettivo l'analisi di serie temporali costituite da dati acquisiti da profili di carico di tipo residenziale. È pertanto necessario rendere questi valori completi per garantirne l'integrità, l'accuratezza e la coerenza nell'intervallo di tempo osservato. Nel caso studio analizzato tale l'integrità è necessaria per due motivi fondamentali:

- per previsioni future basandosi sui dati passati;
- per la costruzione della curva cumulativa con conseguenze sulla validità della stessa.

In particolare, il tema del presente lavoro riguarda la classificazione e la sostituzione dei mancanti nelle serie temporali delle curve di carico elettrico.

Per i carichi elettrici, soprattutto per aggregati di carichi, il valore minimo è maggiore di zero. Se i dati mancanti vengono ritenuti nulli, si può visualizzare la presenza di questi sia sulla curva di carico, sia sulla curva cumulativa (dove i dati nulli sono all'inizio o alla fine della cumulativa, secondo il tipo di cumulativa considerato). Se invece i dati mancanti corrispondono all'assenza di valori (ad esempio, ''not a number'', NaN) nel costruire la curva cumulativa si ottiene un andamento con valori dai quali in generale

non si riesce a rilevare direttamente la possibile mancanza di dati, se non si osserva che il numero totale di dati disponibili è inferiore a quelli che dovrebbero essere presenti nel periodo temporale considerato.
Differente è la situazione per i carichi termici, perché alcuni dati reali possono assumere un valore pari a zero e dall'andamento della curva cumulativa è difficile identificare dati mancanti, con il rischio di confrontare dati effettivamente con valore nullo con quelli non noti ma contraddistinti dallo stesso valore.

La prima fase del presente lavoro presenta gli studi concentrati sulla gestione dei dati mancanti e sui metodi per evitarli o minimizzarli. Esistono numerose tecniche e metodologie per affrontare tale problema, ma ogni situazione è diversa dalle altre e gli unici parametri in base alla quale prendere decisioni sono la struttura dei dati e la natura delle variabili. I risultati ottenuti dall'applicazione di un preciso metodo devono essere interpretati e verificati per valutare quanto effettivamente sono affidabili e reali. In letteratura è possibile distinguere quattro gruppi di metodi per il trattamento dei dati mancanti:

1. metodi basati sulle sole unità osservate;
2. procedure di ponderazione;
3. metodi basati sui modelli;
4. metodi d'imputazione.

È opportuno precisare che, in riferimento al caso analizzato nel seguente lavoro, la soluzione non è univoca perché le tecniche di ricostruzione possono differire in base alla dimensione del ''buco'' di dati e alla distanza tra due ''buchi'' successivi.
Per comprendere meglio il tutto è opportuno far riferimento a un semplice esempio pratico come può esserlo quello di un frigorifero domestico per il quale il *duty-cycle* dà indicazioni sullo stato (acceso/spento). Per esso si supponga che manchi un solo punto, in particolare l'ultimo prima di andare a zero. Una prima tecnica di risoluzione potrebbe essere data dalla media dei due valori adiacenti al dato mancante, 1 e 0, con un valore risultante di 0.5. Il punto trovato, però, non fornisce alcuna informazione utile

perché non rappresenta né lo stato di ON, né lo stato di OFF. Da ciò, si comprende che l'obiettivo è identificare un metodo che consenta di individuare un'irregolarità in un determinato punto in modo che andando a valutare l'andamento rispettivamente prima e dopo, si possa stimare il valore corretto da inserire.

È necessario "osservare" i dati e identificare la migliore metodologia che consenta di ricostruire l'irregolarità presente.

La successiva fase del lavoro si focalizza sui metodi matematici generalmente sviluppati per la ricostruzione dei dati mancanti, i quali si basano su determinate informazioni. Essi possono risultare insufficienti, portando a risultati che nella maggior parte dei casi si discostano da quanto accade realmente. È necessaria la conoscenza del dominio specifico di applicazione per poter ottenere risultati più accurati.
In particolare, nell'analisi delle curve di carico è importante inserire elementi tratti dalla conoscenza del tipo di carico e delle modalità di impiego dell'elettricità da parte degli utenti. In questo modo si possono migliorare i risultati che si ottengono dall'applicazione diretta di metodi di calcolo di uso generale. Ad esempio, l'applicazione della Trasformata Discreta di Fourier e, successivamente, della Trasformata Inversa di Fourier ha il fine di ricavare informazioni circa ciascuna utenza. Nello specifico, è possibile visualizzare e identificare la presenza di una certa periodicità giornaliera e settimanale e collocare ciascuna utenza in una precisa fascia di consumo sulla base del valore massimo e minimo raggiunto.

Le curve di carico degli utenti elettrici possono avere andamenti molto diversi. Per accorpare comportamenti elettrici simili, si procede con un raggruppamento delle utenze aventi lo stesso andamento in classi, denominate *cluster*.
I metodi di Clustering consentono di formare un insieme di giorni "tipo" che possono servire a classificare i giorni incompleti confrontandoli con i centroidi di ciascun cluster.

Tuttavia, si dimostra che questo modo di procedere, anche se è utile a raggruppare le curve di carico secondo criteri di similitudine, risulta limitato in quanto manca la conoscenza locale per la costruzione delle singole curve di carico. Per migliorare la ricostruzione è necessario individuare gli andamenti prevalenti e rappresentare i dati che circondano il gruppo di dati mancanti tramite curve che riproducono le caratteristiche principali della parte di dati nota. Ad esempio, l'impiego di informazioni date dalle ampiezze e dalle fasi di sinusoidi che corrispondono alle armoniche, ovvero sinusoidi con una frequenza multipla della frequenza fondamentale determinata come l'inverso della lunghezza del periodo di analisi, consente di rilevare eventuali regolarità. La presenza di armoniche significative pone le basi per poter effettuare un'analisi armonica in modo da migliorare e regolarizzare l'andamento ricostruito con la giusta frequenza. Altri tipi di regolarità vengono identificati con appositi strumenti matematici.

I risultati del presente lavoro intendono evidenziare come la definizione e applicazione di procedure che dipendono dalla conoscenza del dominio possano risultare migliori rispetto a procedure puramente basata su formulazioni matematiche generali utilizzate in qualsiasi contesto.

Capitolo 1

Dati mancanti

1.1 Tipologia di dati mancanti

L'incompletezza di una serie temporale di dati è dovuta a tre cause raggruppate nelle seguenti categorie:

1. mancata copertura;
2. mancate risposte totali;
3. mancate risposte parziali.

Data la popolazione obiettivo, per "mancata copertura" si intende l'esclusione dalla lista di campionamento di alcune unità; esse vengono escluse anche dai risultati avendo probabilità nulla di essere selezionate.

Nel caso di "mancate risposte totali" non si ha alcuna informazione per una determinata unità di interesse.

Per ultimo, la categoria delle "mancate risposte parziali" comprende situazioni in cui le informazioni rilevate sono ritenute accettabili per il caso in analisi, ma alcune mancano per motivi diversi.[1]

La categoria delle mancate risposte parziali è quella presa in osservazione in questo studio. Presupponendo dunque questa situazione di dati mancanti, è possibile specificare tre diversi tipi di struttura da un punto di vista teorico (Little & Rubin, 1987):

- MCAR, *missing completely at random,* se gli eventi che portano alla mancanza di un determinato tipo di dati avvengono in modo causale e indipendenti sia dalle variabili osservabili sia dai parametri d'interesse non osservabili;

[1] R. Pertile, "Aspetti critici negli studi clinici: problemi di metodo e applicazione", Tesi di dottorato, Università degli Studi di Padova, a.a. 2008, relatore P. Facchin

- MAR, *missing at random,* si presenta quando la mancanza non è casuale, ma può essere completamente spiegata da variabili che presentano informazioni complete;
- NMAR, *not missing at random,* in cui la variabile manca e il motivo è legato alla sua stessa mancanza.[2]

1.2 Metodi per il trattamento dei dati mancanti

Il metodo impiegato per trattare i dati mancanti dipende dal motivo per cui essi mancano, ovvero dall'appartenenza a una delle tre categorie presentate al paragrafo precedente.

Dopo di che, un altro elemento che influisce sulla scelta è il tipo di struttura di dati mancanti (MCAR, MAR, NMAR).

I metodi proposti in letteratura possono essere classificati in quattro gruppi:

1. metodi basati sulle sole unità osservate;
2. procedure di ponderazione;
3. metodi basati sui modelli;
4. metodi di imputazione.

1.2.1 Metodi basati sulle sole unità osservate

Il metodo prevede direttamente l'eliminazione completa o parziale di quei campi in cui mancano informazioni relative alle variabili di interesse. Come si evince, tale approccio risulta semplice da applicare ma comporta una diminuzione di precisione e un aumento della distorsione in particolare quando i dati non sono MCAR.

Si impiegano due diversi tipi di approccio:

- l'esclusione completa e definitiva (*listwise*) dell'unità di cui manca il dato su una o più variabili;

[2] Alvira Swalin (2018), "How to Handle Missing Data"
< https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>

- l'esclusione parziale (*pairwise*) della variabile di cui manca il valore; più precisamente, considerando analisi per analisi, se per una variabile manca un campo, essa sarà esclusa soltanto nell'analisi che richiama quel campo, mentre sarà considerata per altri impieghi.

1.2.1.1 *Eliminazione listwise*

L'eliminazione *listwise* è il metodo più utilizzato nella gestione dei dati mancanti, anche se comporta un'alta perdita di dati risultando pertanto inaffidabile e poco vantaggioso. Tuttavia, esso risulta un'ottima soluzione quando il campione è sufficientemente ampio e dunque l'eliminazione non costituisce un problema. Se la struttura di dati è MCAR, il metodo produce stime imparziali e risultati conservativi; in caso contrario l'eliminazione *listwise* può causare distorsioni nelle stime dei parametri.

1.2.1.2 *Eliminazione pairwise*

L'eliminazione *pairwise* elimina le informazioni solo quando manca un particolare dato nello studio in analisi. Se mancano dati altrove, i valori esistenti sono utilizzati per ricavare altre informazioni. Il vantaggio, rispetto alla eliminazione prima esaminata, è l'utilizzo di tutte le informazioni osservate, ma con lo svantaggio che i parametri del modello vengono impiegati in diversi campi con numeri diversi. Tale metodo produce risultati migliori per i dati MCAR o MAR, ma in presenza di molte osservazioni mancanti, l'analisi risulta scarsa.

1.2.2 Metodi di ponderazione

I dati mancanti nei casi di mancata copertura sono trattati con delle procedure di ponderazione basate su informazioni provenienti da una fonte di dati esterna. Alle unità osservate a fine rilevazione si assegna un fattore moltiplicativo andando a modificare i pesi assegnati e in aggiunta ottenere una rappresentazione anche di quelle non osservate. A fronte della difficoltà di ricavare informazioni ausiliarie necessarie

all'assegnazione di pesi coerenti e evitare di giungere a quantità pesate non affidabili, sicuramente il vantaggio di tale metodo deriva da una semplicità di applicazione.

1.2.3 Metodi basati su modelli

Questi metodi consistono nella determinazione di un modello parametrico, i cui parametri sono determinati utilizzando il metodo della massima verosimiglianza assumendo che i dati siano MAR e che i parametri determinati siano differenti da quelli proprio del generatore di dati mancanti.

1.2.3.1 *Metodo della massima verosimiglianza*

Con il metodo della massima verosimiglianza i parametri sono stati stimati utilizzando i dati disponibili, mentre i dati mancanti si ottengono dai parametri di quelli noti.
Quando sono presenti dati mancanti ma relativamente completi, essi possono essere ricavati utilizzando la distribuzione condizionata delle altre variabili.

1.2.3.2 *Expectation-Maximization (EM)*

Si avvicina al metodo di massima verosimiglianza ma crea un insieme di dati in cui i valori mancanti sono calcolati con i valori stimati con l'impiego dei metodi di massima verosimiglianza.
Questo approccio inizia con la fase di aspettativa in cui vengono calcolati determinati parametri come varianze, covarianze e medie. Tali stime vengono utilizzate per creare un'equazione di regressione per stimare i dati mancanti. Il passaggio iniziale di aspettativa viene ripetuto con i nuovi parametri, dove vengono determinate le nuove equazioni di regressione. Tale procedura viene ripetuta finché il sistema non si stabilizza e la matrice di covarianza per l'iterazione successiva è la medesima dell'iterazione precedente.
Ad ogni nuovo insieme di dati senza valori mancanti, viene incorporato un termine di disturbo associato a ciascun valore stimato con il fine di avere un'indicazione del grado di incertezza associato all'imputazione dell'*expectation-maximization.*

Quando i dati mancanti sono molti, il metodo può richiedere un tempo di calcolo e di convergenza elevato, risultando abbastanza complesso.[3]

1.2.4 Metodi di imputazione

I metodi di imputazione vengono usati nel caso di mancate risposte parziali e si basano sull'assegnazione di un valore sostitutivo al dato mancante producendo un insieme completo. Questi metodi risultano semplici e intuitivi e permettono di ricondursi a situazioni di dati completi senza scartare nessuno dei dati osservati.

I metodi di imputazione proposti in letteratura sono numerosi, ma possono essere distinti in tre classi:

- metodi *deduttivi* nei quali il valore stimato deriva da informazioni presenti;
- metodi *deterministici* nei quali alle variabili con le stesse caratteristiche vengono attribuite gli stessi valori;
- metodi *stocastici* nei quali le imputazioni ripetute per unità aventi le stesse caratteristiche possono produrre differenti valori imputati; presentano una componente aleatoria, detta *residuo*, corrispondente a uno schema probabilistico associato al particolare metodo di imputazione prescelto.

Il trattamento dei valori mancanti con tecniche d'imputazione richiede che le mancate risposte siano di tipo MCAR o MAR.

1.2.4.1 *Metodi di imputazione deduttivi*

Sfruttano le informazioni derivanti dai dati noti in modo da poter stimare il valore da sostituire al dato mancante. È necessario, tuttavia, la definizione di modelli di comportamento specifici del fenomeno in esame ed è necessario avere un buon grado di conoscenza dei dati e delle relazioni esistenti al loro interno.

[3] Huyn Kang (2013), "The prevention and handling of the missing data" <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/>

1.2.4.2 *Metodi di imputazione deterministici*

a) Imputazione con media ***(Mean imputation overall)***

Con questo metodo si sostituiscono tutte le mancate risposte nella variabile y con un unico valore, ovvero la media calcolata sul totale dei dati, dunque $\overline{y_r}$. È un metodo utilizzato solo per le variabili quantitative (per le variabili qualitative al posto del valore medio viene utilizzata la moda).

Per evitare di ottenere risultati incoerenti, è preferibile utilizzare tale metodo nei casi di un numero ridotto di dati mancanti e con valori da determinare che non sono casuali.

Alla semplicità di calcolo e di facile applicazione, si contrappone lo svantaggio di introdurre una distorsione non trascurabile nella distribuzione della variabile, creando un picco artificiale in corrispondenza del valore medio e giungendo a risultati non sempre buoni per la stima della varianza e per le relazioni tra le variabili.

b) Imputazione con media all'interno delle classi (*Mean imputation within classes)*

Si divide il campione totale in classi di imputazione in base ai valori assunti da prefissate variabili ausiliarie considerate esplicative di y e si calcola la media dei rispondenti della variabile y all'interno di ogni classe. Ciascuna media viene poi assegnata ai valori mancanti appartenenti alla stessa classe:

$$y_{mhi} = \overline{y_{rh}}$$

per l'*i*-esimo non rispondente della classe *h* (con *h = 1, 2...H*).

Tale metodo può essere utile nei casi in cui esistano poche relazioni tra le variabili e il risultato dell'analisi è coerente con la stima di medie e di aggregati.

Ha il vantaggio di ridurre le distorsioni generate dalle mancate risposte ed è semplice da applicare, ma introduce distorsioni nella distribuzione della variabile, creando una serie di picchi artificiali in corrispondenza della media di ciascuna classe. In particolare, i valori calcolati riflettono la variabilità tra le classi (*between*) ma non quella all'interno delle classi (*within*).

c) Imputazione con regressione ***(Predictive regression imputation)***

Tramite le informazioni delle variabili esistenti viene fatta una previsione e sostituito il valore ottenuto a quello mancante: i valori dei rispondenti vengono impiegati per stimare i parametri della regressione per la variabile di studio y sulla base di variabili ausiliarie considerate esplicative di y. I modelli di regressioni sono lineari quando la variabile y è quantitativa; sono invece modelli log-lineari o logistici nel caso in cui, invece, la variabile y è qualitativa.

Il principale vantaggio deriva dall'utilizzo di un numero elevato di variabili sia quantitative che qualitative con una conseguente riduzione delle distorsioni generate dall'assenza di dati. Tuttavia, la forte influenza di dati anomali e la presenza di distorsioni nella distribuzione della variabile, necessita di un modello diverso per ogni variabile in cui si ha un dato mancante, con possibili risultati non sempre veritieri.

d) Imputazione con matching della media ***(Predictive mean matching)***

In questo metodo, il valore da imputare ottenuto mediante il metodo dell'imputazione con regressione è confrontato con quelli di y disponibili per i casi completi. Il valore più vicino al riferimento ottenuto dalla regressione è quello scelto per l'imputazione.

Esso è adatto per le variabili quantitative e per un numero elevato di valori, comportando una riduzione delle distorsioni generate dall'assenza di dati.

Al contempo, esso genera distorsioni nella distribuzione della variabile, non conserva sufficientemente la variabilità delle variabili, contribuisce alle distorsioni nelle relazioni tra quelle non utilizzate nel modello ed è necessario ricorrere ad un modello differente per ogni variabile in cui manca il valore.

e) Imputazione dal più vicino donatore ***(Nearest-neighbour imputation)***

Ad ogni dato mancante viene attribuito il valore di quello noto più vicino, determinato tramite una funzione di distanza applicata alle variabili ausiliarie.

I passi per applicare tale metodo sono:

- applicazione di una opportuna funzione di distanza tra la variabile del campione con dato mancante e tutte le altre variabili note;
- determinazione del valore più vicino al valore di interesse;
- impiego del valore determinato per il dato mancante.

Possono essere impiegate differenti funzioni di distanza; quelle usate generalmente sono:

- La distanza *Euclidea*;
- La distanza *ponderata;*
- La distanza di *Mahalanobis*;
- La distanza *Minmax*.

Il metodo è particolarmente impiegato quando il numero di dati mancanti è ridotto, nei casi in cui il campione è ampio e trovare un valore da sostituire a molte variabili simultaneamente è semplice, ottenendo risultati accettabili dal punto di vista della qualità. È consigliabile evitare tale metodo quando si presenta un numero elevato di dati mancanti (principalmente se una stessa variabile manca in più situazioni), quando le informazioni sono di carattere quantitativo e le indagini sono di piccole dimensioni.

1.2.4.3 *Metodi di imputazione stocastici*

Imputazione stocastica singola

Definito un insieme di valori "plausibili" da prendere come riferimento, l'imputazione singola è relativamente semplice da utilizzare perché sostituisce ad ogni dato mancante un valore appartenente all'insieme definito in partenza. Al contempo, ottenuto l'insieme di dati completi, non tiene conto di eventuali incertezze perché considera tutti i dati, compresi quelli stimati, come se fossero effettivamente stati osservati.

a) Imputazione con donatore casuale *(Random donor imputation overall)*

Tale tecnica si basa sulla scelta random di un campione tra le unità osservate e si attribuisce tale valore al dato mancante. Solitamente, si considera un campione di

riferimento caratterizzato dall'assenza di ripetizioni per evitare che uno stesso valore venga utilizzato molte volte portando a una riduzione della variabilità ed effetti negativi sulle relazioni. È un metodo di facile applicazione il cui utilizzo è consigliato nei casi in cui il campione è costituito da poche variabili preferibilmente tra loro correlate e in cui mancano pochi dati. Si cerca di sostituire il dato mancate con un valore pressoché "reale", uno stesso donatore viene utilizzato una sola volta e i valori calcolati non generano gravi distorsioni nella distribuzione della variabile. Al contempo, si possono ottenere distorsioni non trascurabili nelle relazioni tra le variabili se si presentano quelle che hanno uguale probabilità di risposta e se le variabili con dati mancanti hanno caratteristiche simili a quelle con dati completi.

b) Imputazione con donatore casuale all'interno delle classi *(Random donor imputation within classes)*

Il punto di partenza di tale tecnica è innanzitutto la creazione di classi e all'interno di ciascuna si sostituiscono i dati che mancano con un valore scelto in modo random all'interno della classe di appartenenza. Questo metodo viene impiegato principalmente nei casi di campioni di grosse dimensioni in cui si possono creare più classi e quindi avere più valori di riferimento da attribuire, ma, al contempo, è preferibile che siano presenti poche variabili per ridurre le distorsioni nelle relazioni.

Innanzitutto, il valore sostituito viene considerato "reale", viene scelto in genere con le stesse caratteristiche e viene utilizzato una sola volta per ridurre la variabilità delle distribuzioni marginali. Infine, ad un numero di classi più elevato corrisponde la possibilità di selezionare il valore da sostituire da variabili vicine.

Può capitare di avere una perdita di dettaglio nella formazione delle classi di imputazione a causa del passaggio dei dati continui in gruppi discreti.

c) Imputazione tramite hot-deck sequenziale *(Sequential hot-deck imputation)*

Definite delle classi di imputazione, al dato mancante viene assegnato un valore selezionato casualmente nella classe oppure un valore considerato rappresentativo della medesima, come può esserlo il valore medio. I campi sono trattati in maniera sequenziale: se uno presenta una mancata risposta, questa viene sostituita dal valore iniziale assegnato alla sua classe di appartenenza; se invece il campo dispone della risposta, questa sostituisce il valore precedentemente memorizzato per la sua classe di imputazione.

Di solito questo metodo è utilizzato in indagini in cui il numero di casi che presentano dati mancanti può essere anche molto elevato.

Tale metodo è caratterizzato da una efficienza computazionale. Gli svantaggi, però, derivano dalla programmazione della procedura di imputazione, la quale può risultare complicata e dispendiosa in termini di tempo e vi sono problemi di ripetibilità dovuti all'impiego di uno stesso valore. Lo svantaggio principale si presenta nel caso in cui all'interno di una classe si susseguono più dati mancati e a questi viene sostituito l'ultimo valore noto comportando una riduzione della variabilità. Infine, nel caso che il numero di classi è alto, potrebbero esserci problemi circa la correttezza dei valori da prendere come riferimento nelle sostituzioni.

d) Imputazione tramite hot-deck gerarchico *(Hierarchical hot-deck imputation o Flexible matching imputation)*

Tale metodo prevede una prima fase di formazione di classe prendendo come riferimento gruppi dettagliati di variabili ausiliarie. Tale metodo si basa su una scala gerarchica tra i rispondenti e in modo che se non è possibile trovare un donatore adatto nell'iniziale classe di imputazione le classi collassano per passare a un livello più basso. Un campione sufficientemente ampio fa sì che le classi di imputazione abbiano abbastanza valori da utilizzare come donatori preferibilmente con le medesime caratteristiche e un valore sostituito che è considerato "reale".

È possibile una perdita di dettaglio nella formazione delle classi di imputazione a seguito della conversione di dati continui in gruppi discreti, ma vengono conservate le relazioni tra le sole variabili usate per costruire le classi.

e) Imputazione con regressione casuale *(Random regression imputation)*

Con questa tecnica i valori imputati sono sempre stimati con una equazione di regressione al termine della quale si forma un termine residuale.

Tale metodo prevede l'impiego di un gran numero di variabili, sia quantitative sia qualitative, in modo da ridurre, maggiormente rispetto ad altri metodi, le distorsioni generate dall'assenza di dati e i valori stimati non generano distorsioni nella distribuzione della variabile. Tuttavia, provoca distorsioni nelle relazioni tra le variabili non utilizzate nel modello ed è pertanto necessario valutare un metodo differente per ogni variabile in cui manca il dato. Per concludere, i valori stimati possono discostarsi fortemente dalla realtà e la presenza di dati anomali è un fattore che può influenzarlo in modo determinante.

Imputazione multipla

L'imputazione multipla si può riassumere in alcuni passaggi:

1. imputare i valori mancanti usando un appropriato modello d'imputazione che incorpori l'imputazione casuale (es. data augmentation), ripetendo l'operazione *M* volte;
2. produrre l'analisi d'interesse per ognuno degli *M dataset* risultanti utilizzando le analisi statistiche standard;
3. combinare le stime degli *M dataset* seguendo le regole di Rubin (Rubin, 1987) per produrre risultati soddisfacenti.

L'imputazione multipla permette di effettuare analisi in presenza di dati mancanti basandosi su diversi set di dati completi.[4]

[4] R. Pertile, "Aspetti critici negli studi clinici: problemi di metodo e applicazione", cit., p.4

I metodi per la generazione delle imputazioni dipendono dal *pattern* (modello) di dati mancanti.

- Nel caso di *pattern* monotoni si possono applicare metodi parametrici come i modelli di regressione (Rubin, 1987) e le stime di massima verosimiglianza. In particolare, sono stati determinati dei fattori che possono essere applicati alla funzione di verosimiglianza per presentarla in forma esplicita. Si possono anche applicare metodi non parametrici come il *propensity scores* (Rubin, 1987).
- Per *dataset* con *pattern* di dati mancanti qualsiasi si ricorre a metodi iterativi. La principale tecnica per modelli parametrici con dati incompleti si basa su stime di massima verosimiglianza tramite l'algoritmo EM *Expectation-Maximization* (Dempster, Laird, Rubin, 1977), o in alternativa si ricorre a metodi MCMC *Monte Carlo Markov Chain* (*Gibbs Sampling*, *Data Argumentation*, l'algoritmo di *Metropolis-Hastings*) (Schafer 1997).
- Un'ulteriore alternativa è di ricondurre il campione con i dati mancanti che si basa su *pattern* qualsiasi a un *pattern* monotono con l'ausilio dei metodi MCMC e poi applicare ai dati ottenuti i metodi propri dei pattern monotoni richiamati precedentemente. [5]

L'imputazione multipla ha il vantaggio di ottenere una stima della varianza relativamente semplice e flessibile, applicabile a qualsiasi tipo di valore calcolato ed è utilizzabile per completare valori mancanti in indagini multivariate di dati mancanti.
Come è stato osservato, esistono differenti metodi per realizzare un'imputazione multipla. I metodi MCMC, e soprattutto algoritmi che utilizzano il *Data Argumentation*, potrebbero risultare costosi dal punto di vista computazionale e la convergenza potrebbe risultare difficile da determinare.
Da ciò deriva l'esigenza di ricercare vie alternative, puntando sui metodi di imputazione semiparametrici o non parametrici con una richiesta di assunzioni distribuzionali più deboli sulla variabile da determinare o addirittura non richiesta. Un modo per realizzare questo tipo di imputazione multipla è con il metodo ABB (*Approximate Bayesian*

[5] Prof.ssa L. Neri, Capitolo 7, "La qualità dei dati: l'imputazione dei dati mancanti"

Bootstrap) che viene considerato un approccio non parametrico per l'imputazione multipla.[6]

1.3 Procedura di ricostruzione dei dati di misura applicata da Terna

1.3.1 Cause di mancanza dati

La mancanza di dati di misura rilevati dal contatore è associabile o a un errore nel sistema di trasmissione del dato di misura oppure all'assenza del dato sul contatore installato in campo.

1.3.2 Ricostruzione dei dati di misura

La ricostruzione può essere effettuata ricorrendo a differenti modalità di seguito esposte:

a) Ricostruzione con ricorso a misure alternative

Tale metodologia si applica in locale quando il punto di rilievo è dotato di contatore di riserva o di un contatore di riscontro.

b) Ricostruzione mediante riferimento a bilancio di sito

Nel caso di una porzione di rete elettrica (o impianto) in cui le linee siano tutte dotate di Apparecchiature di Misura, la misura mancante può essere determinata tramite l'impiego della somma algebrica dei valori degli altri contatori che contribuiscono al bilancio energetico.

Nel calcolo di quest'ultimo si considera unicamente l'energia attiva.

c) Ricostruzione polinomiale di intervalli inferiori a un'ora

In caso di assenza di dati per intervalli inferiori o uguali a 4 quarti d'ora (4 campioni della curva di carico), si applica il seguente criterio:

[6] R. Pertile, "Aspetti critici negli studi clinici: problemi di metodo e applicazione", cit., p.4

Interpolazione polinomiale contigua

Si stimano i campioni mancanti utilizzando una interpolazione definita di ordine 2N+2 considerando N+1 valori prima del buco di dati e N+1 valori dopo e quindi di grado 2N+1 o in alternativa, di ordine N+1 utilizzando N+1 valori tra prima e dopo con la condizione che i campioni di valori siano adiacenti escludendo la parte di dati mancanti, dunque di grado N.

d) Ricostruzione mediante elaborazione di misure alterative

Si ricorre all'impiego di misure alternative quando risulta difficile applicare una delle metodologie precedentemente esposte. In questo caso i valori sono ottenuti con una ricostruzione cha si basa o sulle misure rilevate sul medesimo montante oppure ricorrendo all'applicazione di un algoritmo matematico che tenga conto dei valori letti in altri punti di misura.

e) Ricostruzione mediante stime

Si applica quando il punto di misura è stimato in base ai suoi profili storici oppure i dati mancanti sono relativi a intervalli superiori a un'ora.

Elaborazione di più punti di misura dei quali almeno uno è stimato in base ai profili storici

Il punto di misura per questa applicazione deve avere un comportamento regolare nel tempo.

In questo caso, si utilizza un algoritmo di ricostruzione ottenuto definendo opportuni coefficienti percentuali sulla base dei punti noti. Se dovesse mancare una delle misure, si applica il coefficiente medio calcolato sui periodi precedenti.

Ricostruzione dati di misura per periodi >1h mediante analisi dei profili storici

Per la ricostruzione dei dati di misura per periodi maggiori di 1 ora mediante analisi di profili storici occorre fare una distinzione in funzione della tipologia di punto di scambio tra:

- punti di scambio con flussi di energia periodici nel tempo;
- punti di scambio senza storicità.

A seconda del caso si hanno due differenti metodi di ricostruzione:

1. *Interpolazione su medie di curve*

È utilizzabile solo nei casi in cui l'andamento dell'energia assorbita in funzione del tempo sia periodico.

Note le curve di carico giornaliere o settimanali, si stimano i dati mancanti calcolando la media dei campioni conosciuti:

- 2N curve di carico utilizzando andamenti simili di N curve di carico precedenti prima del "buco", e di N curve di carico successive al buco, con $N \geq 1$ e definito in di volta in volta in base al caso analizzato;
- come alternative, prendendo N curve di carico in totale e utilizzandone N-M precedenti e M successive al buco e con $N \geq M \geq 1$ e scelti sempre in funzione dell'analisi corrente;

2. *Ricostruzione con altri dati*

A differenza del caso precedente è impiegato per andamenti dell'energia non regolari nel tempo. Possono presentarsi due differenti soluzioni:

- Tramite il misuratore UTF in cui, partendo dalle ultime rilevazioni del contatore UTF presente, si stima per differenza l'energia rilevata dal contatore con quella delle misure note. Il valore ottenuto sarà ridistribuito equamente sui quarti d'ora privi di misura;
- Nel caso in cui non fosse applicabile nessuno dei metodi esposti precedentemente, occorrerebbe verificare in locale se sono presenti Apparecchiature di Misure alternative.

f) Gestione casi anomali

Per la gestione di casi anomali si rientra nelle misure etichettate come "*Power failure parziale*", con mancanza dovuta a una momentanea disalimentazione del contatore o

a un abbassamento temporaneo di tensione sullo stesso. In questo caso il dato viene inteso come valido e non si deve procedere alla ricostruzione. [7]

[7] Terna (2006) "Procedure operative per la gestione delle informazioni e dei dati nell'ambito del sistema di misura" <http://collaudo.download.terna.it/terna/0000/0105/65.pdf>

Capitolo 2
Analisi matematica applicata alle curve di carico di utenze residenziali

2.1 Introduzione al caso studio

Nel presente lavoro, sono stati presi in considerazione 150 profili di tipo residenziale. Per ciascuno di essi, considerando un periodo di osservazione di un anno solare discretizzato in quarti d'ora, viene fornito il valore dell'energia assorbita in kWh.

Talvolta, nell'acquisizione dei dati, alcuni valori possono essere errati o non disponibili, andando a formare il cosiddetto "buco" di dati. La presenza di quest'ultimo comporta problemi nelle previsioni future sulla base di quelle passate o nella costruzione della curva cumulativa con una possibile perdita della sua validità.

Assunto che possono esserci pochi "buchi" ma lunghi, oppure tanti ma brevi e vicini, l'obiettivo è indentificare la migliore metodologia che consenta di ricostruire l'irregolarità presente.

Nel caso studio in esame è stata considerata l'utenza 1 come quella di riferimento ed è stata effettuata un'analisi su un periodo temporale di 55 giorni corrispondente a quello più lungo privo di dati mancanti o errati.

2.2 La Trasformata Discreta di Fourier

Una funzione periodica di periodo T può essere ricostruita come una combinazione lineare infinita di funzioni seno e coseno. Maggiore è il numero di termini sommati nella serie, più l'approssimazione data dalla somma delle funzioni trigonometriche si avvicina al valore esatto.

Campionare un segnale $x(t)$ significa "estrarre" dal segnale stesso i valori che assume a istanti temporali equi-spaziati.

Il segnale $x(t)$ può essere ricostruito completamente partendo dai suoi campioni discreti $x_{discr}(n)$ presi con passo di campionamento T_S se e soltanto se il contenuto in frequenza del segnale trasformato è uguale a zero per tutte le frequenze f tali che $2\pi f \geq \frac{\pi}{T_S}$.[8]

La minima frequenza, detta *frequenza di Nyquist*, è il valore di frequenza necessario a campionare un segnale analogico senza perdere informazioni. Nello specifico, il *Teorema del campionamento di Nyquist-Shannon* afferma che data una funzione la cui trasformata di Fourier sia nulla al di fuori di un certo intervallo di frequenze, nella sua conversione analogica-digitale la minima frequenza di campionamento necessaria per evitare l'aliasing e perdita di informazione nella ricostruzione del segnale analogico originario deve essere maggiore del doppio della sua frequenza massima.[9]

La *Trasformata Discreta di Fourier* consente di effettuare l'analisi in frequenza dei segnali a tempo discreto. Il calcolo diretto della *DFT* è piuttosto costoso in quanto richiede $O(N^2)$ operazioni di prodotto, dove N è la dimensione dello spazio su cui si applica la trasformata. Pertanto, si passa ad algoritmi *FFT (Fast Fourier Transform)* che calcolano la *DFT* riducendo a $O(NlogN)$ il numero di operazioni.

L'applicazione della *FFT* ad un segnale campionato con N punti produce in uscita un vettore con N punti complessi e il piano complesso è diviso in N direzioni, in cui ogni direzione corrisponde a una delle componenti della *FFT*.[10]

L'ampiezza delle componenti nel dominio della frequenza è calcolata tenendo conto delle proprietà di simmetria dell'uscita della *FFT.* L'asse delle frequenze è ricostruito considerando il massimo ordine armonico di interesse e il periodo fondamentale, generando il vettore delle frequenze e il vettore dell'ordine armonico.

Dunque, applicando la *DFT*, oppure la versione ottimizzata *FFT* ad N campioni otteniamo come risultato un numero discreto di oscillazioni sinusoidali.

[8] N. Lombardi, "L'Analisi di Fourier e la DFT", Tesi triennale, a.a. 2019-2020, relatore G. Rodriguez

[9] "Teorema del campionamento di Nyquist-Shannon" <https://it.wikipedia.org/wiki/Teorema_del_campionamento_di_Nyquist-Shannon>

[10] G. Chicco, "Serie di Fourier e campionamento", Distribuzione e utilizzazione dell'energia elettrica

Se il numero dei dati campionati è un numero pari *N*, l'algoritmo che calcola la Trasformata di Fourier fornisce un numero di oscillazioni distinte pari a $M = \frac{N}{2} + 1$.
Queste oscillazioni hanno frequenze da 0 (a cui corrisponde la componente continua) fino a metà della frequenza di campionamento dei dati f_c (a cui corrisponde la frequenza di Nyquist).
Se il numero di campioni è un numero dispari *N*, l'algoritmo che calcola la trasformata di Fourier fornisce un numero di oscillazioni $M = \frac{N-1}{2} + 1$.
Anche in questo caso le oscillazioni hanno frequenze da 0 fino a quasi metà della frequenza di campionamento dei dati.
Pertanto, come detto precedentemente, il limite superiore alle frequenze visualizzabili è dato dalla frequenza di campionamento. Nel caso di *N* pari esisterà un campione delle frequenze con esattamente tale frequenza. Lo spettro è costituito dalle componenti sinusoidali da frequenza 0 fino a $\frac{f_c}{2}$, mentre il numero di componenti sinusoidali è pari a $\frac{N}{2}$. La distanza tra una componente e la successiva risulta essere pari a $\frac{1}{NT}$ dove *T* è il tempo di campionamento e tale distanza è inversamente proporzionale alla "risoluzione in frequenza".
Essendo *NT* la grandezza temporale della finestra di dati, si ottiene che la risoluzione è inversamente proporzionale alla finestra di osservazione. Una "falsa" risoluzione frequenziale può essere ottenuta aumentando artificiosamente il numero di campioni di una finestra aggiungendo degli zeri in coda. Tale metodo è noto come *zero padding.* Questa operazione consente di avere campioni in frequenza ravvicinati ma non di distinguere due componenti vicine; è soltanto una rappresentazione più continua. [11]
Applicando infine al vettore ottenuto dalla *FFT* la *Inverse Fast Fourier Transform* si ottiene il segnale originale. Nello specifico, essa consente di ricavare, a partire da un segnale definito nel dominio della frequenza, quel segnale nel dominio del tempo la cui trasformata è il primo segnale.

[11] N. Lombardi, "L'Analisi di Fourier e la DFT", cit., p.23

2.3 Analisi sperimentale della DFT

L'applicazione della *Fast Fourier Transform* e successivamente della *Inverse Fast Fourier Transform* alle prime 50 utenze oggetto di studio, consente di visualizzare l'andamento dell'energia in funzione del tempo con l'obiettivo di visualizzare chiaramente le informazioni fornite dai dati per ciascuna utenza.

Innanzitutto, l'impiego delle due funzioni matematiche sopracitate permette di individuare una certa periodicità giornaliera o settimanale in modo da prevedere un certo andamento futuro sulla base di quello passato. Inoltre, per ciascun utente, si individua il valore massimo e il valore minimo di un certo andamento, andando poi a classificarli nella rispettiva fascia energetica e associare un prezzo diverso a seconda dell'orario e del giorno in cui viene utilizzata.

Per poter fare un'analisi accurata di quanto appena detto e con la premessa che le 50 utenze presentano dei dati mancanti, si individua per ciascuna il periodo temporale completo più lungo. Successivamente, per avere una maggiore facilità di elaborazione, si riduce il numero di utenze raggruppando quelle che presentano il medesimo periodo temporale. Risulta un numero di utenze ridotte pari a 21.

Ricordando che la *DFT* ha modulo con simmetria pari intorno alla frequenza di Nyquist, nella sua applicazione a ciascuna delle utenze restanti, si prende in considerazione soltanto la metà e, dunque, la restante parte è formata dai complessi coniugati che non danno informazioni aggiuntive. Definito uno specifico numero di armoniche prevalenti che vada a ottimizzare la rappresentazione della curva e che minimizzi il rumore ad essa associato, si precisa che le ampiezze dei termini armonici in genere diminuiscono all'aumentare dell'ordine dell'armonica. Pertanto, è evidente che è logico arrestare lo sviluppo in serie dopo un ragionevole numero di termini, in quanto il contributo dei successivi diventa sempre meno significativo, andando, anzi, a generare dei disturbi.

La successiva applicazione della *IFFT* consente di ottenere il segnale originale.

Vengono mostrati di seguito dei precisi andamenti specifici che danno prova di quanto appena detto.

La *Figura 2.1* mostra una certa periodicità settimanale che si ripete nell'arco delle 8 settimane esaminate. Si nota, inoltre, il maggior assorbimento di energia nei giorni lavorativi rispetto al sabato e alla domenica.

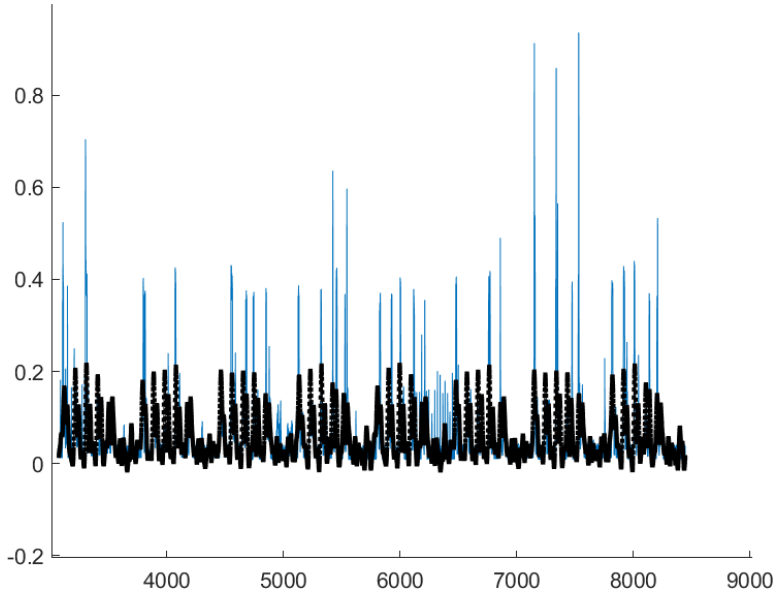

Figura 2.1- Andamento IFFT utenza 1

La *Figura 2.2* è un caso particolare perché nel periodo temporale più lungo privo di dati mancanti, corrispondente a 32 giorni, non è possibile tener contro della ripetibilità dei dati: l'utenza si trova in un determinato range di energia in alcune settimane e, in uno di valore minore in altre. Come evidenziato in figura, in determinati istanti di tempo l'andamento della curva va sotto al minimo e, nella realtà, non è possibile avere energia assorbita negativa. Per ovviare il problema si dovrebbe procedere con una compressione di ciascun valore anomalo, ovvero si fa in modo di avvicinarlo al minimo tramite un fattore moltiplicativo. Facendo in questo modo si cerca di non andare al di sotto del minimo e con un miglioramento anche dell'errore rispetto al vero andamento. Per comprendere quanto appena detto, in *Figura 2.3* si riporta l'andamento ottenuto dopo aver effettuato la compressione al periodo temporale indicato.

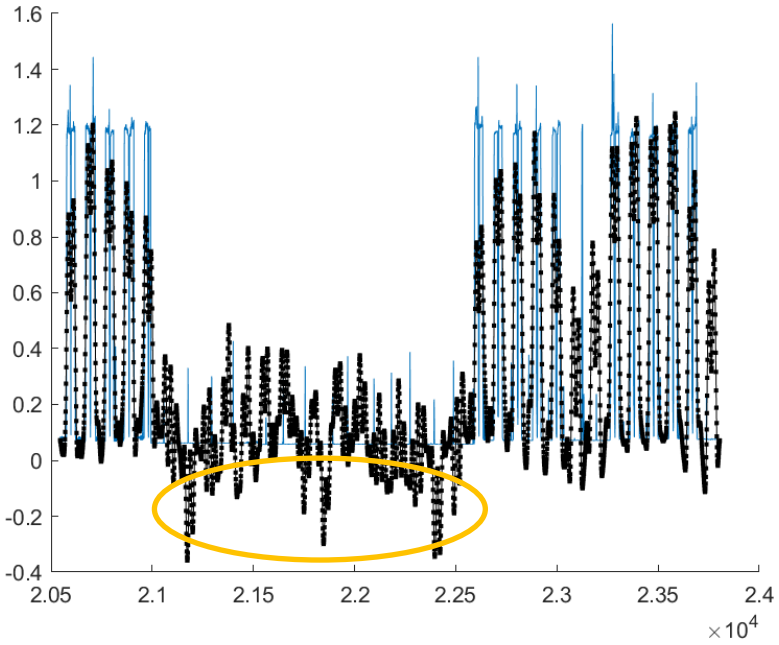

Figura 2.2 – Andamento IFFT utenza 4

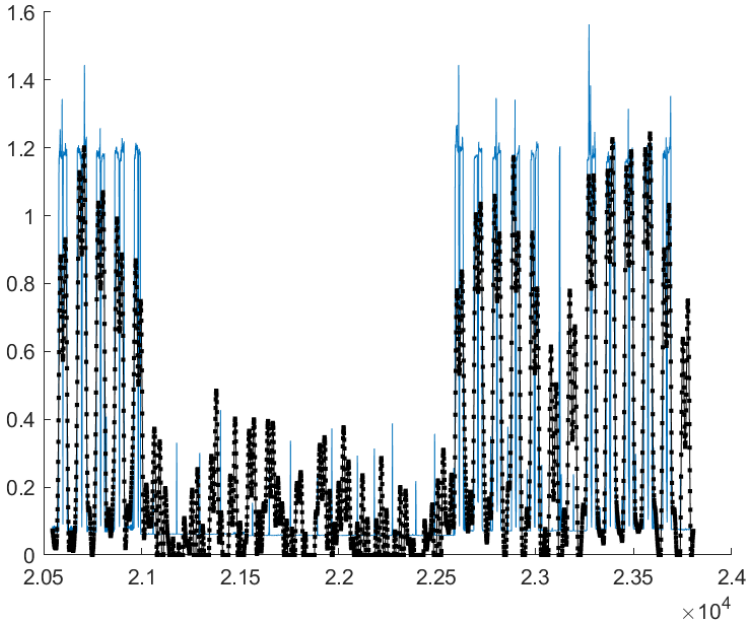

Figura 2.3 – Applicazione compressione utenza 4

La *Figura 2.4* invece è un caso particolare perché non permette facilmente di poter identificare la ripetibilità dei dati essendo questa limitata ad uno o al più due giorni. Tra gli istanti 2,13 e 2,26, si nota che l'energia in alcuni istanti assume valori negativi: è necessario applicare una compressione per evitare che nei medesimi istanti i valori non abbiano senso. Tuttavia, nello stesso intervallo la differenza tra il valore massimo e il

valore minimo è ridotta. Applicando la compressione la differenza va ulteriormente a ridursi e l'andamento si avvicina ancora meglio a quello piatto effettivo. Talvolta laddove si presenta un valore piatto, la ricostruzione tramite la *IFFT* presenta dei picchi e l'impiego della compressione avvicina ancora di più l'andamento a quello reale.

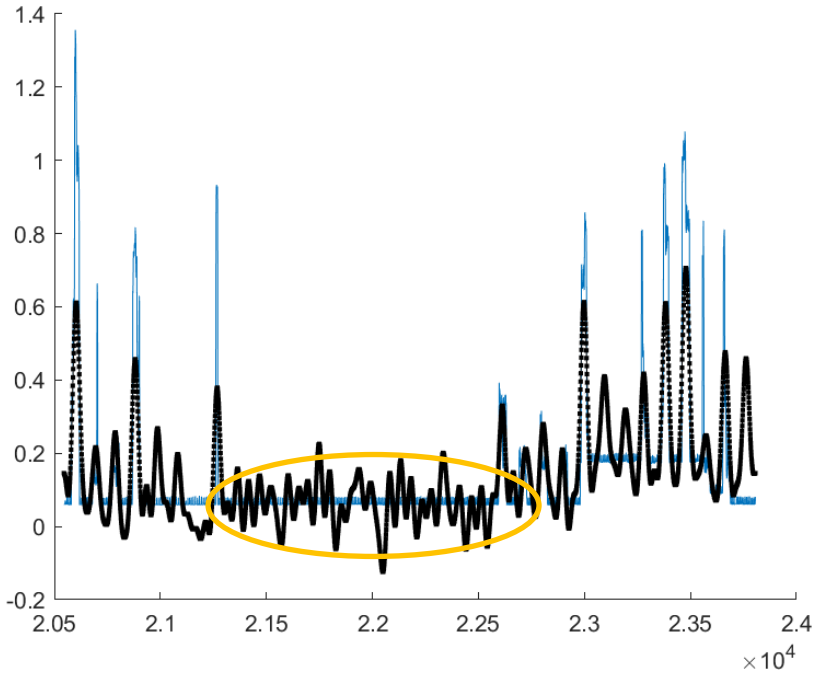

Figura 2.4 - Andamento IFFT utenza 27

Per ultimo, la *Figura 2.5*, presenta una periodicità settimanale con differenti andamenti e valori di energia il sabato e la domenica, facendo sì che questi giorni possano essere valutati con un ragionamento differente.

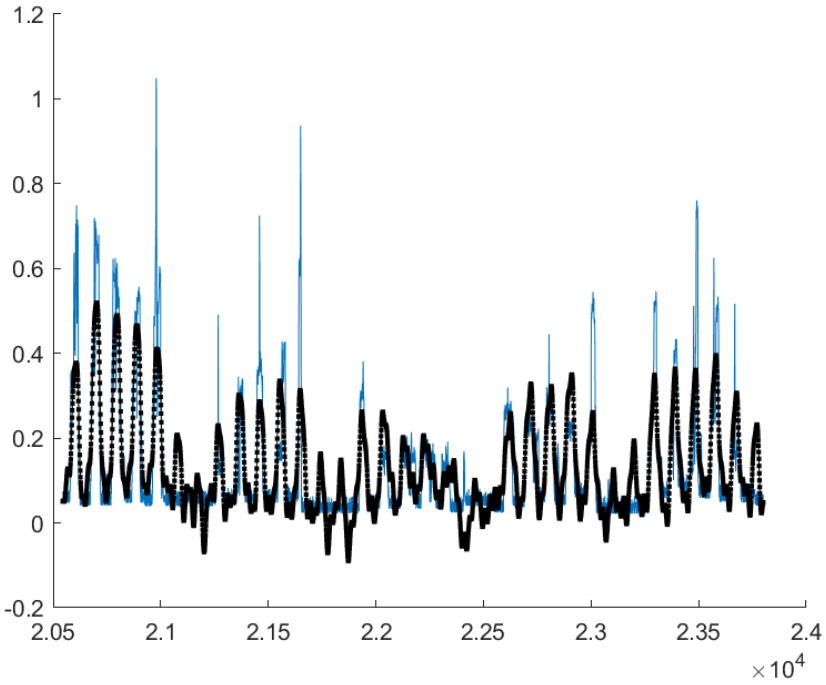

Figura 2.5 - Andamento IFFT utenza 35

Delle considerazioni separate meritano le figure di seguito mostrate per il fatto che richiedono per la maggior parte del tempo un valore di energia basso, con la conclusione di non avere informazioni precise circa l'utenza in analisi, precludendo la possibilità di ricostruire i dati mancanti.

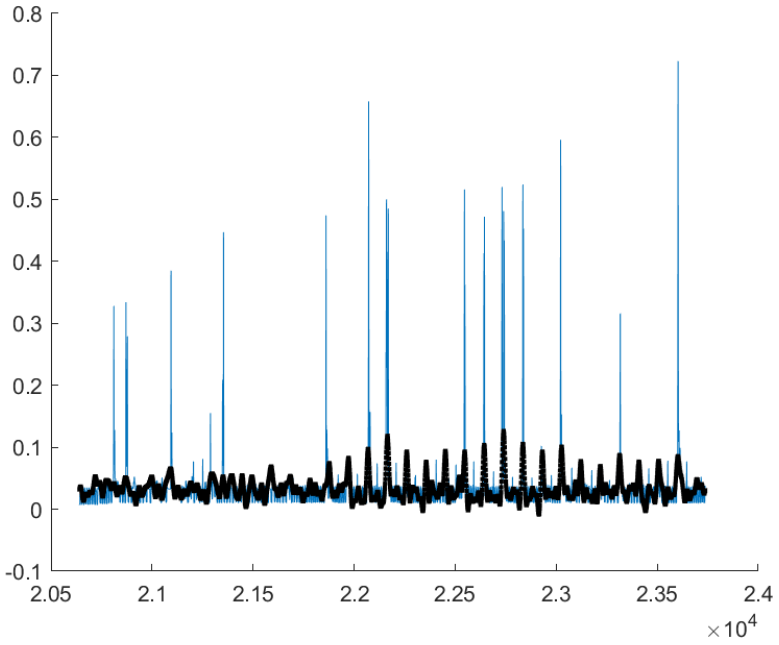

Figura 2.6 - Andamento IFFT utenza 3

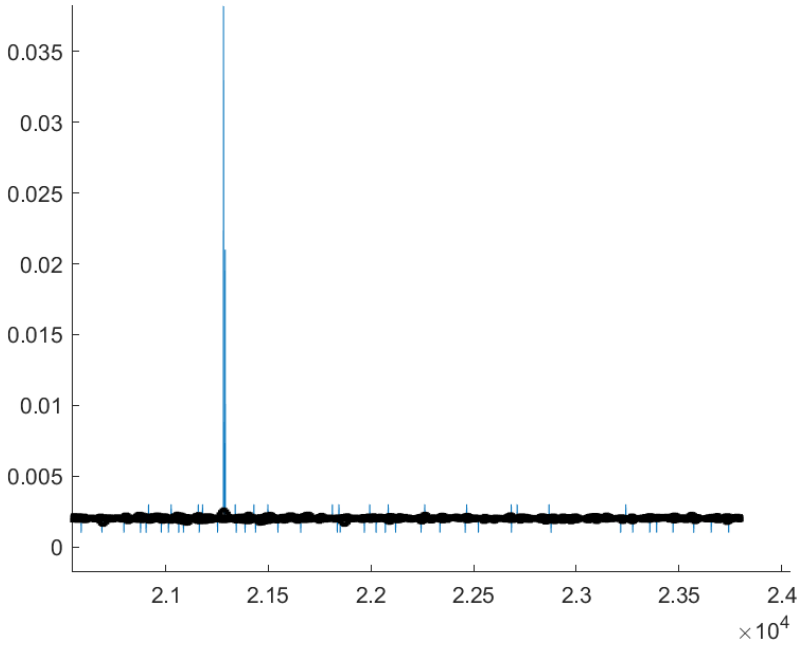

Figura 2.7 - Andamento IFFT utenza 14

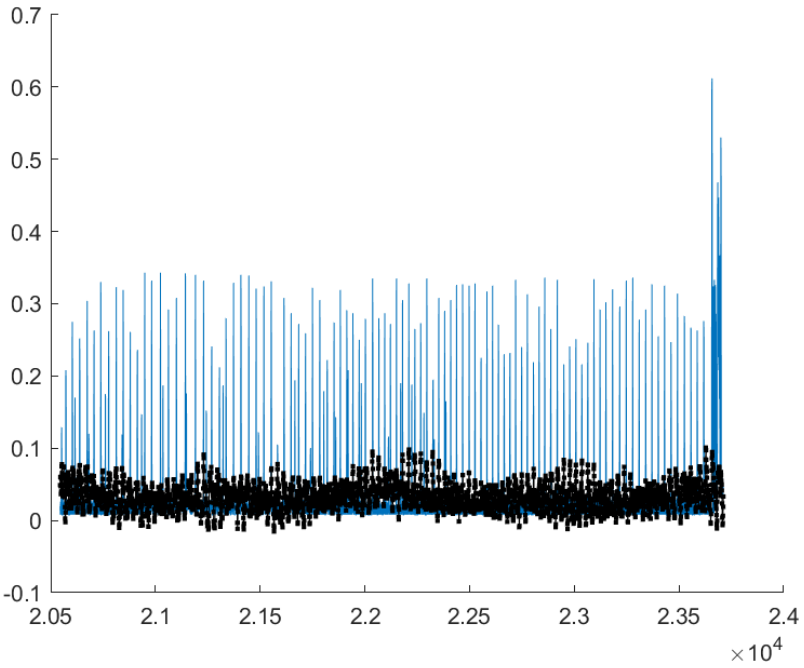

Figura 2.8 - Andamento IFFT utenza 32

Da questi ultimi casi, per poter fare una selezione delle utenze residenziali in relazione all'ordine di grandezza dell'energia richiesta, si ricorre alla curva di durata del carico definita come una curva che mette in relazione il carico e il tempo. Lungo l'asse delle

ordinate viene rappresentato il valore dell'energia in kWh assorbita dal carico in ordine decrescente, mentre lungo l'asse delle ascisse il tempo. [12]

Si è scelto di non tener conto di quelle utenze in cui al 5% del tempo corrisponde un valore di energia inferiore a 100 kWh.

Vengono mostrate le curve di carico di alcune utenze con questa caratteristica.

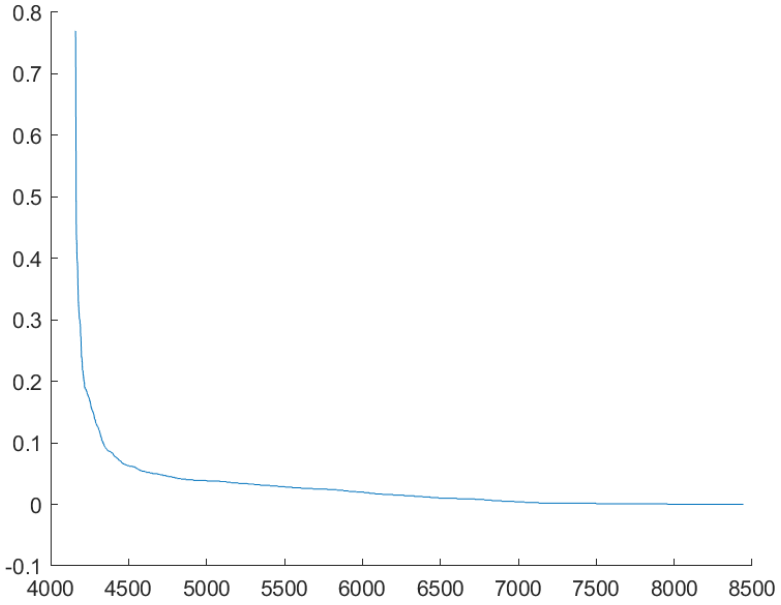

Figura 2.9 – Curva di carico utenza 6

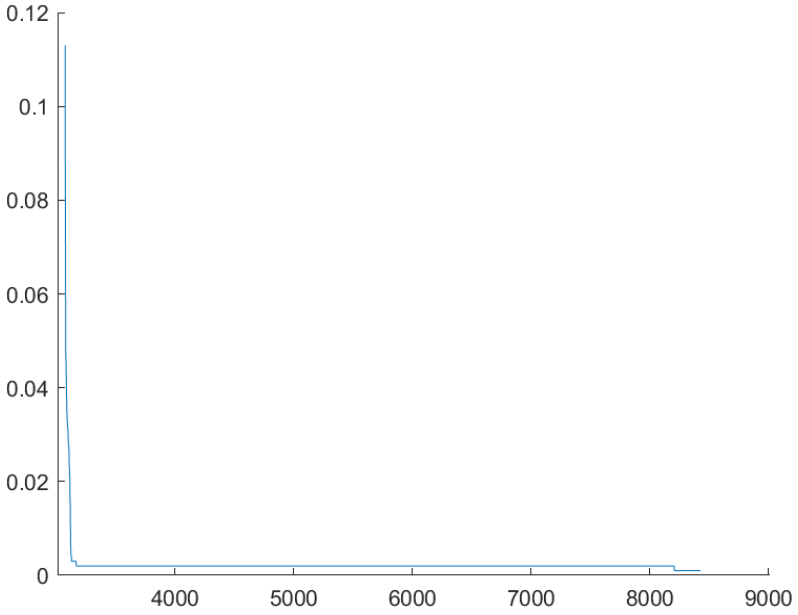

Figura 2.10 – Curva di carico utenza 28

[12] "Load duration curve" <https://illustrationprize.com/it/519-load-duration-curve.html>

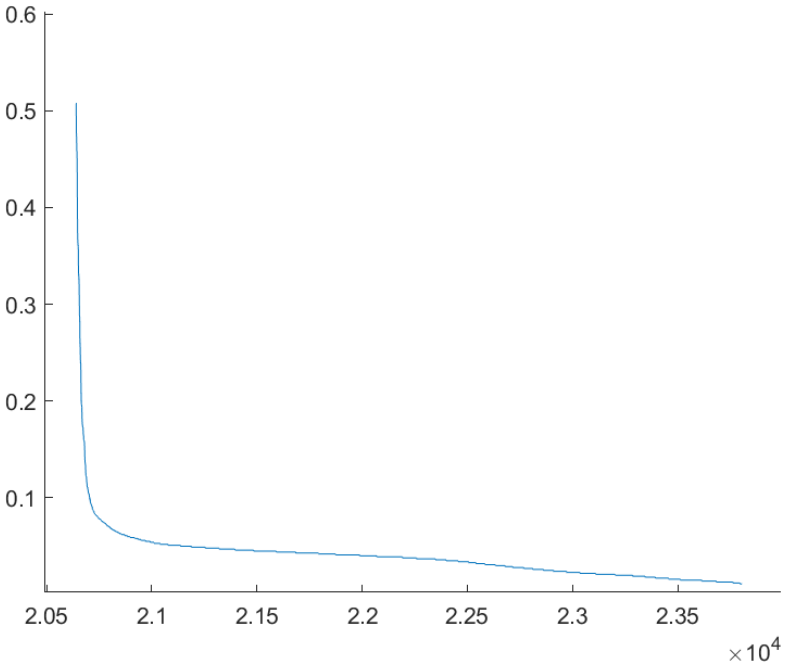

Figura 2.11 – Curva di carico utenza 36

2.4 Processing Time Series Data con MATLAB: *Fill Missing Values*

Matlab mette a disposizione delle funzioni che matematicamente vanno a ricostruire i dati mancanti basandosi sulle informazioni a disposizione.

La funzione *Fill Missing Values* riempie i dati mancanti con i valori calcolati dai punti vicini in base al metodo specifico indicato.

Nel caso studio analizzato è stato applicato tale toolbox di Matlab alle utenze con caratteristiche differenti, ottenendo risultati diversi anche in relazione al metodo specificato in input.

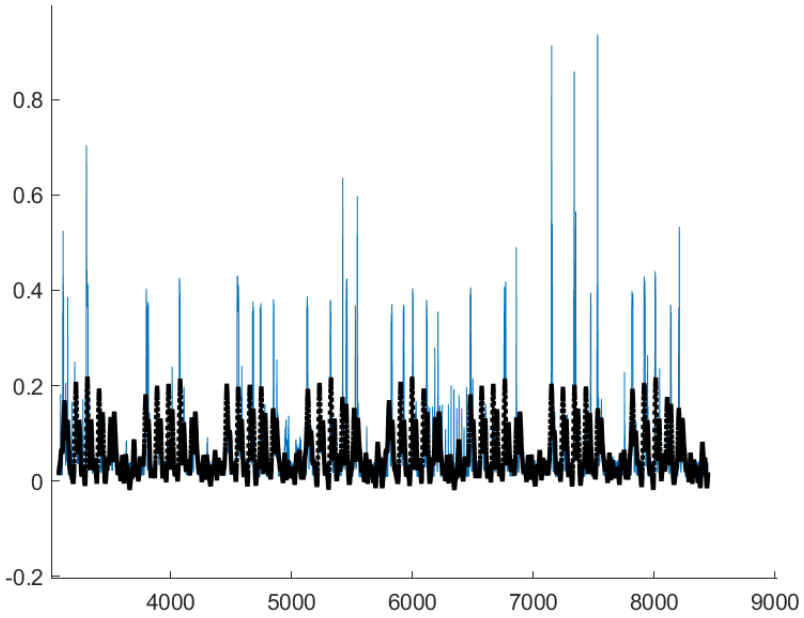

Figura 2.12 - Andamento IFFT utenza 1

Prendendo in esame l'utenza raffigurata, si nota che essa presenta una certa ripetibilità settimanale. Supponendo sia presente un buco fittizio negli ultimi giorni del periodo in questione, si applica la funzione sopracitata sulla base dei dati precedenti.

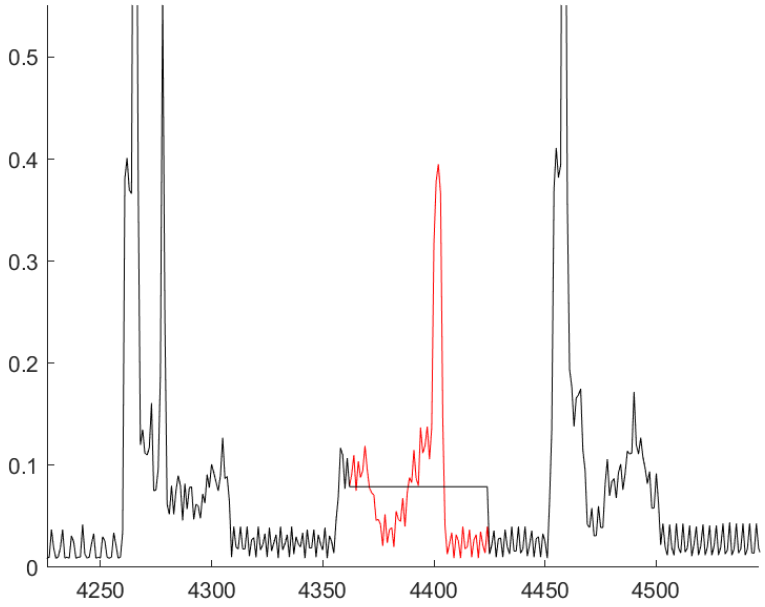

Figura 2.13 – Fill Missing Values 'constant' utenza 1

Dal risultato ottenuto in *Figura 2.13* è evidente che l'andamento vero (in rosso) si discosta in modo non trascurabile dalla curva ricostruita, andando addirittura a sostituire i dati mancanti con valori costanti.

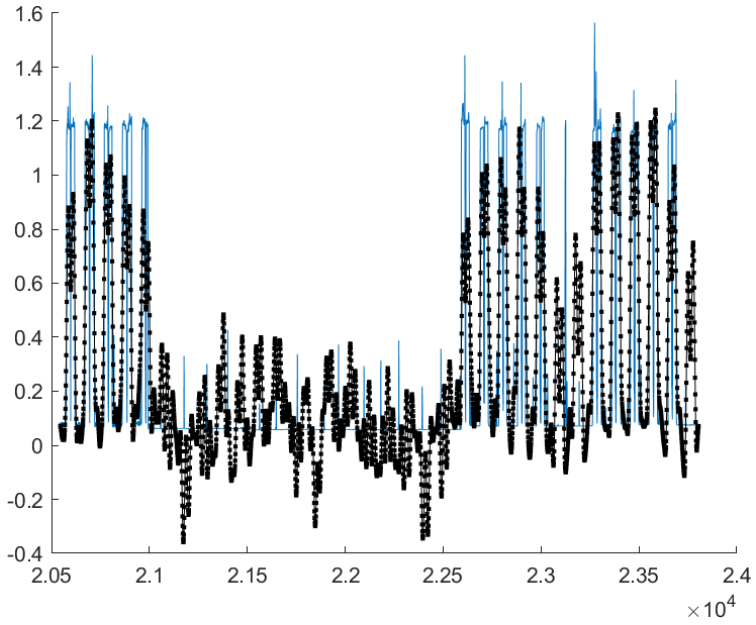

Figura 2.14 - Andamento IFFT utenza 4

Si applica ora la funzione *Fill Missing Values* all'utenza in *Figura 2.14* focalizzando l'attenzione sul cambio di livello tra la terza e la quarta settimana.
Variando i metodi applicativi, di seguito si possono visualizzare i risultati ottenuti.

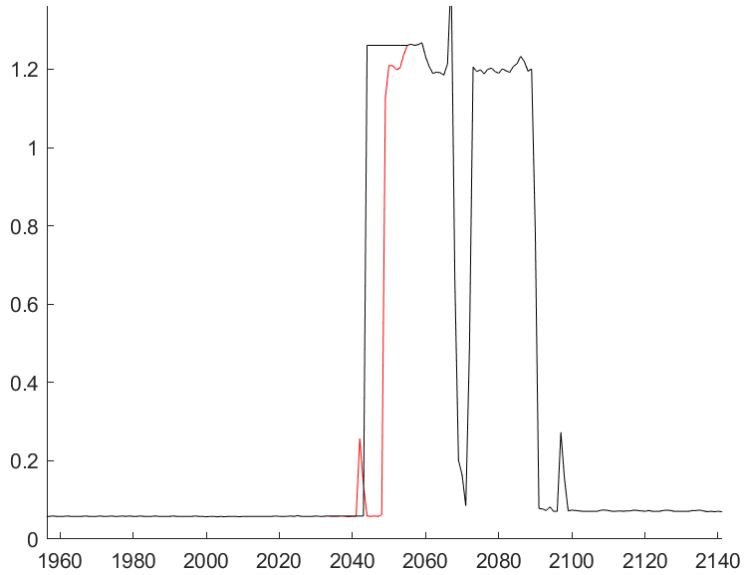

Figura 2.15 - Fill Missing Values 'nearest' utenza 4

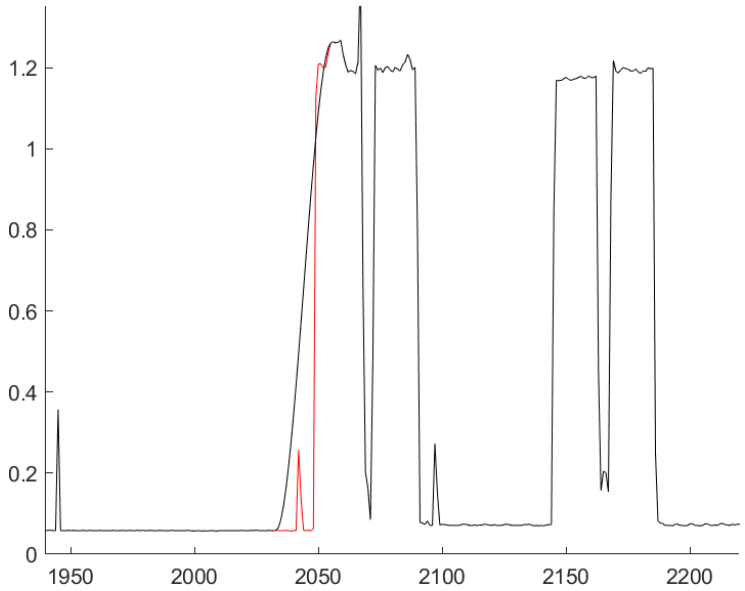

Figura 2.16 - Fill Missing Values 'pchip' utenza 4

In *Figura 2.15* è stato usato il metodo *'nearest* ' dando come risultato una spezzata che si allontana dal vero andamento.

Invece, in *Figura 2.16* il metodo *'pchip* ' effettua una interpolazione cubica a tratti che preserva la forma con un incremento dell'energia in modo graduale, e dunque, con evidente trascuratezza del cambio di livello.

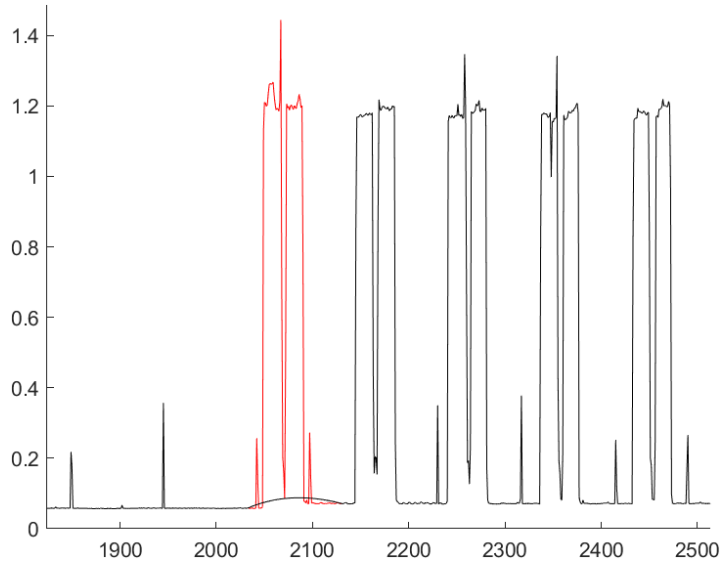

Figura 2.17 - Fill Missing Values 'spline' utenza 4

Incrementando la dimensione del buco di dati fittizio, in *Figura 2.17* si osserva il risultato ottenuto dall'utilizzo del metodo *'spline'* il quale effettua sempre una interpolazione cubica a tratti. In questo caso viene completamente trascurato l'aumento dell'energia assorbita, effettuando semplicemente un raccordo tra il primo e l'ultimo punto noti prima del buco.

Come ultimo esempio, si esamina una utenza che presenta un andamento regolare, come quello mostrato in *Figura 2.18*.

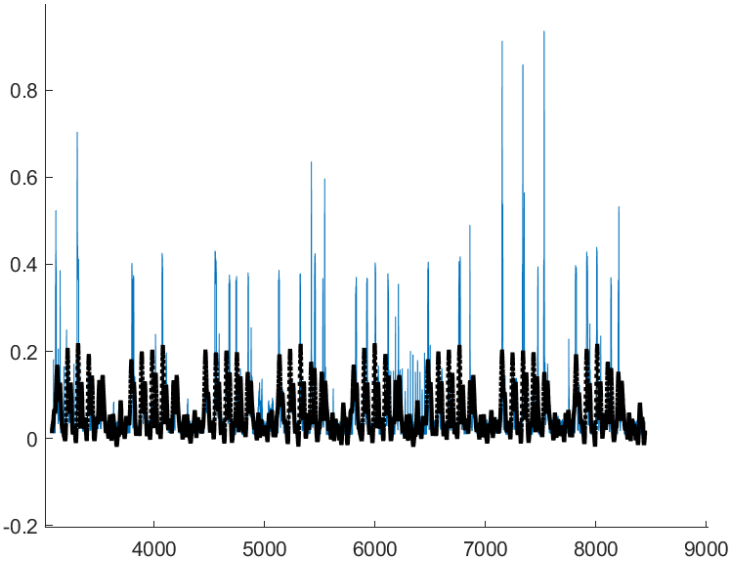

Figura 2.18 - Andamento IFFT utenza 21

Applicando la funzione di Matlab a un piccolo buco di sei dati mancanti e definendo sempre specifici metodi applicativi, si mostrano i risultati ottenuti.

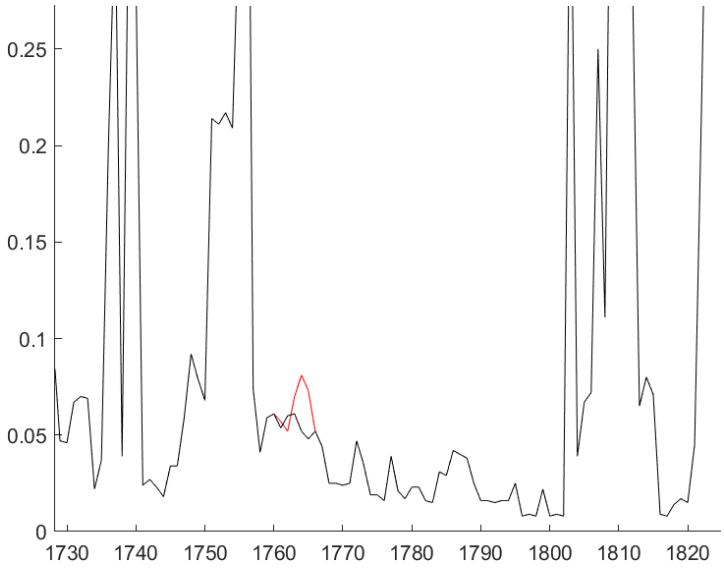

Figura 2.19 - Fill Missing Values 'movmean' utenza 21

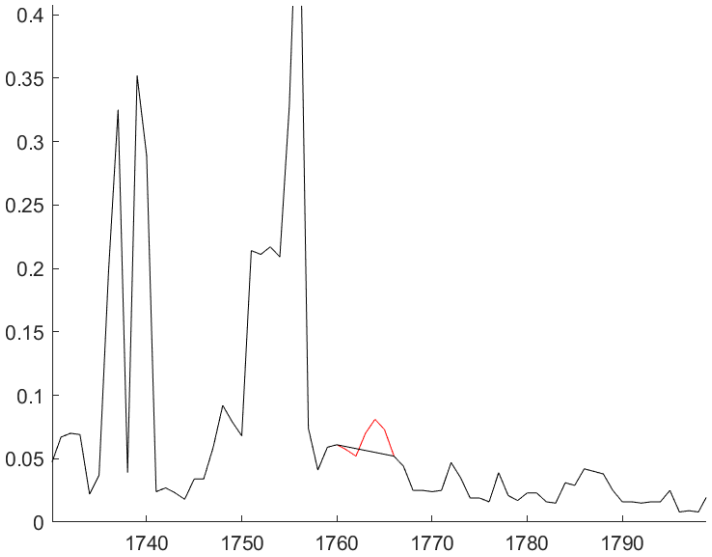

Figura 2.20 - Fill Missing Values 'linear' utenza 21

In *Figura 2.19* è stata applicata la *'movmean'* la quale applica la media mobile alla finestra temporale in questione.

In *Figura 2.20*, invece, è stato utilizzato il metodo *'linear'* che effettua l'interpolazione lineare di valori adiacenti non mancanti.

Considerando un buco di dati fittizio di dimensione maggiore presente nella stessa utenza si ottengono i risultati sotto riportati.

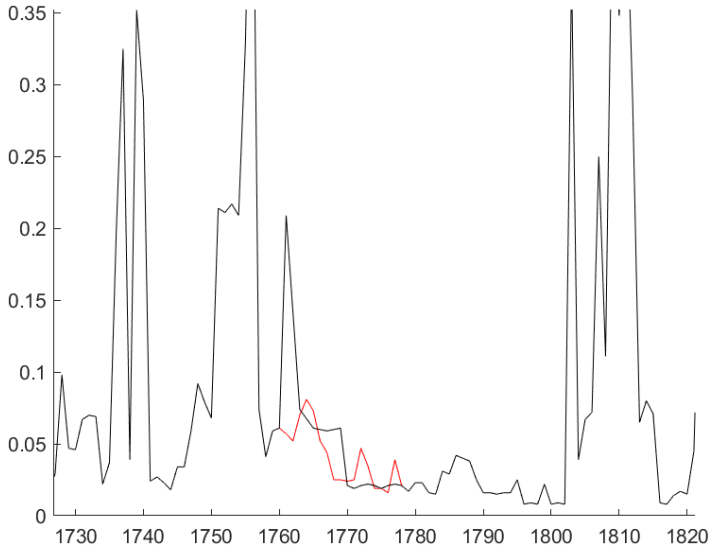

Figura 2.21 - Fill Missing Values 'movmedian' utenza 21

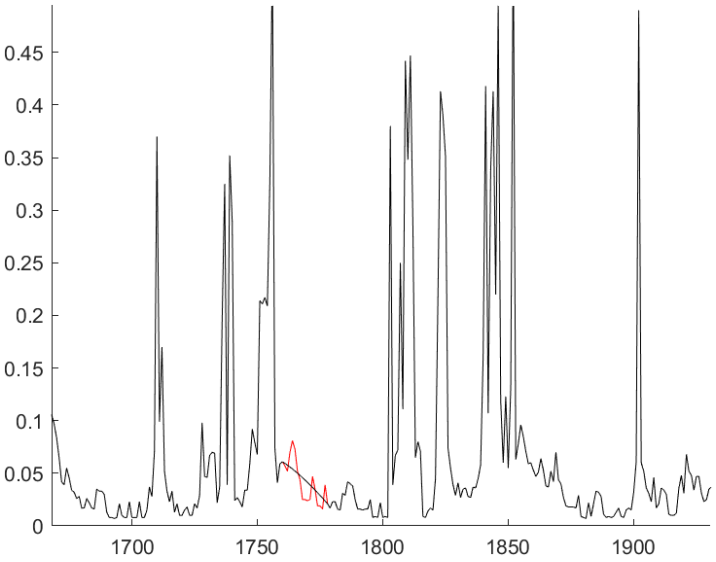

Figura 2.22 - Fill Missing Values 'makima' utenza 21

Il metodo impiegato in *Figura 2.21* è la *'movmedian'* che effettua la mediana mobile alla finestra temporale in questione.

Per ultimo, in *Figura 2.22* è stata applicata *'makima'*, la quale esegue una interpolazione cubica di Hermite Akima modificata.

Dalle figure presentate e dai risultati rilevati si può concludere che c'è una netta differenza tra quello che accade applicando i metodi di analisi matematica e quello che accade con la conoscenza del dominio.

2.5 Ulteriori metodi di analisi

MATLAB offre dei metodi di analisi aggiuntivi che possono essere impiegati per effettuare la ricostruzione di una porzione di dati mancanti.

Preso il giorno 47 come riferimento, il cui andamento è mostrato in *Figura 2.23*, per ciascuna metodologia esposta, viene riportata la ricostruzione sia nel caso che la porzione di dati mancanti si trovi in una zona regolare e sia quando è presente un cambio di livello.

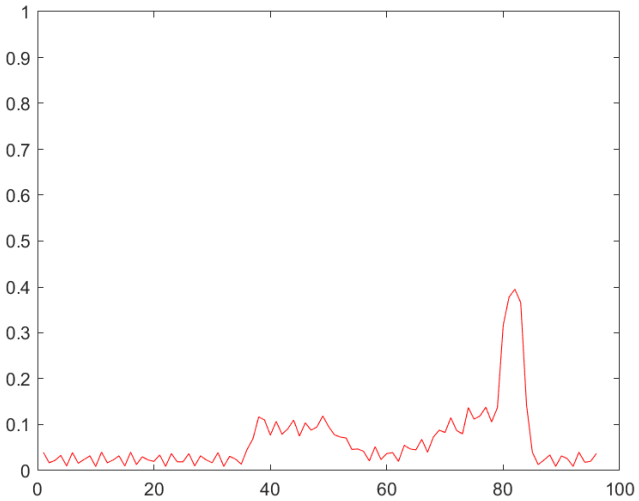

Figura 2.23 – Andamento utenza 47

2.5.1 Interpolazione polinomiale

L'interpolazione polinomiale è l'interpolazione di una serie di valori con una funzione polinomiale che passa per i punti dati. In particolare, un qualsiasi insieme di $n + 1$ punti

distinti può essere sempre interpolato da un polinomio di grado n che assume esattamente il valore dato in corrispondenza dei punti iniziali.[13]

Questo metodo viene utilizzato per l'approssimazione di funzioni andando a determinare una funzione g , appartenente a una classe prescelta di funzioni, che meglio approssima una data funzione f. Nel caso discreto la funzione f è nota su un insieme di $n+1$ nodi $x_0, x_1, \dots, x_n$ appartenenti a un intervallo *[a,b]*.

La funzione g con cui si vuole approssimare la funzione f è un polinomio di grado opportuno, i cui coefficienti vengono determinati in modo che l'approssimazione sia la migliore possibile, compatibilmente con i dati che si hanno a disposizione. Affinché venga rispettato questo requisito è necessario che la distanza tra la funzione f e la funzione g sia minima. [14]

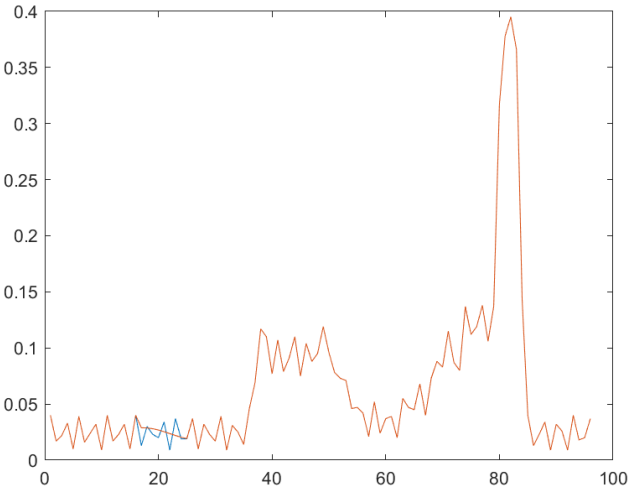

Figura 2.24 - Ricostruzione interpolazione buco regolare

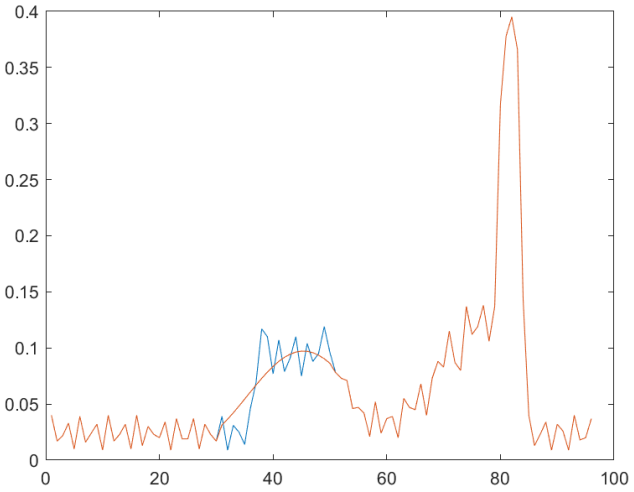

Figura 2.25 - Ricostruzione interpolazione cambio di livello

2.5.2 Support Vector Machine Regression

Il *Support Vector Machine Regression* rientra nella categoria dell'intelligenza artificiale. Esso fa uso di una prima fase di allenamento in cui apprende da esempi, e una

[13] "Interpolazione polinomiale" <https://it.wikipedia.org/wiki/Interpolazione_polinomiale>

[14] Unipi, Capitolo 5, "Interpolazione polinomiale"

successiva fase, detta di funzionamento o generalizzazione, la quale permette di generalizzare e gestire nuovi dati nello stesso dominio applicativo. In poche parole, è un sistema che impara dagli esempi a migliorare la sua conoscenza sulla gestione dei dati in modo migliore ed efficiente.

La linea retta necessaria per l'adattamento dei dati viene definita iperpiano. L'obiettivo è trovare un iperpiano in uno spazio n-dimensionale che classifichi distintamente i punti dati; quelli su entrambi i lati dell'iperpiano più vicini a quest'ultimo sono chiamati *vettori di supporto.* Essi influenzano la posizione e l'orientamento dell'iperpiano e aiutano a costruire il *Support Vector Machine.*

In particolare, gli *iperpiani* sono limiti decisionali utilizzati per prevedere l'output continuo. I vettori di supporto sono usati per tracciare la linea richiesta che mostra l'output previsto dell'algoritmo.

Un *kernel* è un insieme di funzioni matematiche che prende i dati come input e li trasforma nella forma richiesta. Questi sono generalmente usati per trovare un iperpiano nello spazio dimensionale superiore. I kernel utilizzati includono *Linear, Non-Linear, Polynomial, Radial Basis Function (RBF)* e *Sigmoid.*

Per ultimo, le *linee di confine* sono le due linee che vengono designate attorno all'iperpiano per creare un margine tra i punti dati.

Il *Support Vector Machine Regression* è un algoritmo di apprendimento supervisionato utilizzato per prevedere i valori discreti. L'idea alla base è trovare la linea più adatta, ovvero l'iperpiano che ha il numero massimo di punti e adattare la linea migliore all'interno di un valore di soglia. Quest'ultimo è la distanza tra l'iperpiano e la linea di confine.

Il modello prodotto dall'algoritmo dipende solo da un sottoinsieme dei dati di addestramento, perché la funzione di costo ignora i campioni la cui previsione è vicina al loro obiettivo.

Il principio di funzionamento prevede di attribuire gli esempi alle categorie in modo che per ogni categoria ci sia un numero notevole di esempi. Per l'addestramento crea un modello che assegna i nuovi esempi a una tra tutte le categorie. L'idea principale è

quella di minimizzare l'errore. Nella regressione, dovendo prevedere un valore continuo, si decide un margine di tolleranza dell'errore. La particolarità di questo metodo è la possibilità di utilizzare una funzione non lineare per minimizzare l'errore, ovvero una funzione curva che meglio approssima i dati.

Il *Support Vector Machine Regression* presenta alcuni vantaggi:

1. è robusto verso gli outlier;
2. il modello decisionale può essere facilmente aggiornato;
3. ha un'eccellente capacità di generalizzazione, con un'elevata precisione di previsione;
4. la sua implementazione è facile. [15]

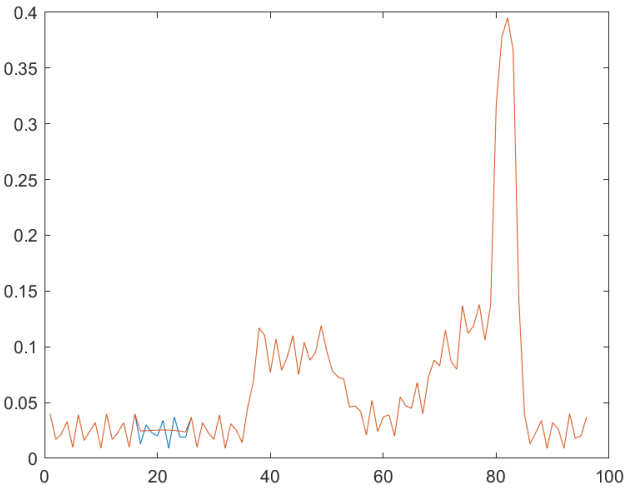

Figura 2.26 – Ricostruzione SVM buco regolare

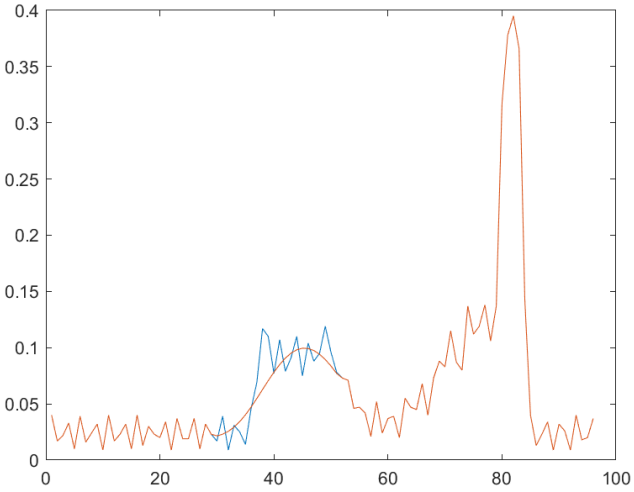

Figura 2.27 – Ricostruzione SVM cambio di livello

2.5.3 Missing Data in MATLAB

La presenza di dati che contengono valori mancanti, comporta nel tracciamento del grafico temporale degli spazi vuoti.

È possibile utilizzare il comando *misdata* per stimare i valori mancanti. Questo comando interpola linearmente i valori mancanti per stimare il primo modello.

[15] Ashwin Raj (Ottobre 2020), "Unlocking the True Power of Support Vector Regression" <https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0>

Successivamente, utilizza questo modello per stimare i dati mancanti come parametri riducendo al minimo gli errori di previsione dell'output ottenuti dai dati ricostruiti.
È possibile specificare la struttura del modello che si desidera utilizzare nell'argomento del *misdata.*
Per un dato modello, i dati mancanti sono stimati come parametri in modo da minimizzare gli errori di previsione dell'output ottenuti dai dati ricostruiti. [16]

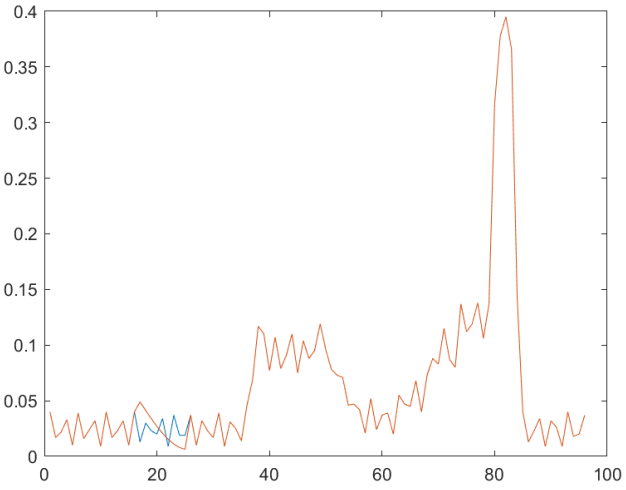

Figura 2.28 – Ricostruzione misdata buco regolare

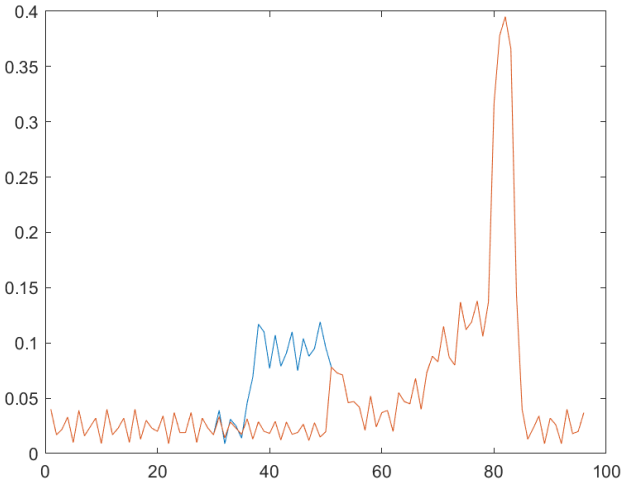

Figura 2.29 – Ricostruzione misdata cambio di livello

2.5.4 Reti neurali

Le reti neurali sono composte da neuroni e archi che collegano i neuroni tra di loro.
Esse sono solitamente composte da più livelli ognuno dei quali è a loro volta composto da uno o più neuroni. Nel modello più semplice di una rete neurale sono presenti un livello di input e un livello di output e tutti i neuroni del primo livello sono connessi, tramite degli archi, a tutti quelli del secondo livello. I valori di input passano attraverso gli archi pesati e mutano di conseguenza; quando i neuroni del livello di output ricevono i valori applicano una funzione che genera un nuovo valore. Il modello della rete può essere complicato aggiungendo dei livelli intermedi chiamati livelli nascosti, che a loro volta avranno un certo numero di neuroni e saranno collegati ai neuroni del

[16] MathWorks, “misdata”
< https://it.mathworks.com/help/ident/ref/misdata.html>

livello precedente e a quelli del livello successivo tramite degli archi. Quando i neuroni del livello nascosto hanno ottenuto un valore, lo trasformano con una funzione chiamata ''funzione di attivazione''. I valori ottenuti saranno la previsione che la rete neurale ha prodotto per i valori in input.

Le reti neurali possono essere utilizzate per la regressione ponendo un unico neurone nel livello di output; in tal modo è possibile allenare una rete neurale a predire valori continui. Nell'ambito della regressione si utilizza l'apprendimento supervisionato. Per farlo è necessario un dataset etichettato dove l'etichetta è rappresentato dal valore di output e cioè dalla variabile di cui si vogliono dare previsioni. Per misurare l'errore della rete si utilizza lo scarto quadratico medio con l'obiettivo di minimizzarlo in modo da ridurre al minimo l'errore.

Nello specifico, il *Deep neural network* è un tipo specifico di machine learning che propone reti neurali multilivello costituite da un livello di input, uno di output e da due o più livelli nascosti organizzati gerarchicamente. [17]

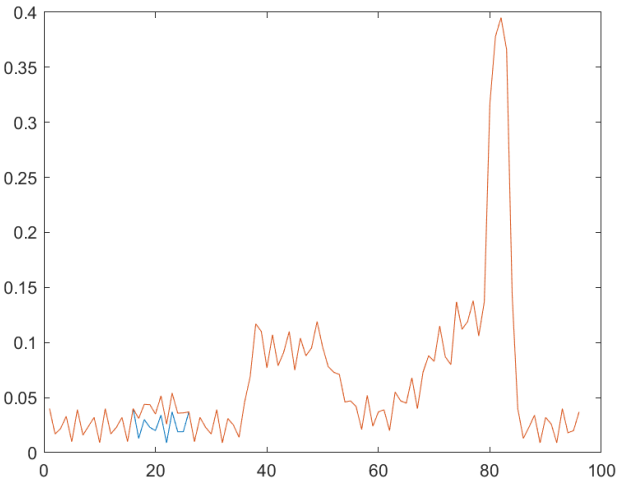

Figura 2.30 - Ricostruzione reti neurali buco regolare

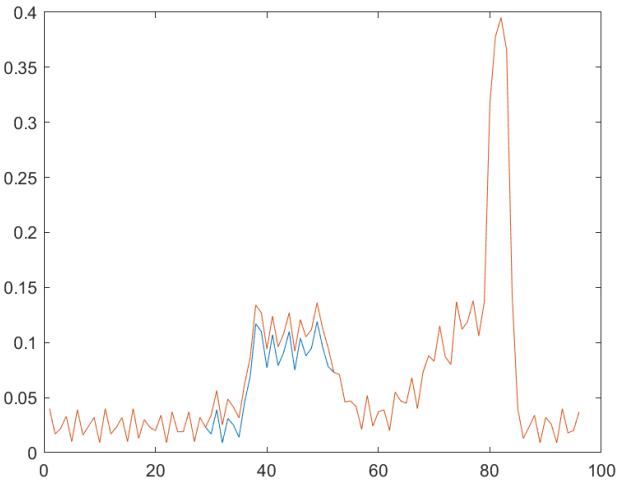

Figura 2.31 – Ricostruzione reti neurali cambio di livello

[17] S. Righini, "Applicazione del Machine Learning a scenari di inquinamento ambientale", Tesi di laurea, Università di Bologna, a.a.2017-18, relatore V. Ghini

2.5.5 Regressione

I modelli di regressione descrivono la relazione tra una variabile di risposta y (output) e una o più variabili predittive x (input). Il Machine Learning Toolbox consente di adattare i modelli di regressione lineare, lineare generalizzato e non lineare. Una volta adattato è possibile utilizzare il modello per prevedere o simulare le risposte.

Nel modello di regressione lineare l'obiettivo è trovare la *retta di regressione*, ovvero quella retta che passa attraverso i punti del diagramma a dispersione e che descrive la relazione lineare tra le due variabili. L'equazione della retta consente di predire il valore della variabile risposta e di trovare il tasso di variazione della variabile di output da quella di input.

La variabile dipendente è una funzione delle variabili indipendenti più un termine d'errore. Quest'ultimo è una variabile causale e rappresenta una variazione non controllabile e imprevedibile nella variabile dipendente. I parametri sono stimati in modo da descrivere al meglio i dati. Il metodo più comunemente usato è il metodo dei *minimi quadrati*. Poiché in un diagramma a dispersione è possibile tracciare molte rette, questo metodo consente di trovare la retta che rende più piccole le distanze tra i punti dei dati e la retta di regressione: trovare la retta per la quale sia minima la somma di tutti i quadrati di queste distanze.

Avente a disposizione la retta di regressione, si possono determinare i punti della retta che corrispondono a valori specificati di x. Questi punti sono detti *previsioni*. Le previsioni sono affidabili solo per valori di x che cadono nell'intervallo dei valori osservati.

Per capire il livello di adattamento della retta di regressione ai dati, si valutano i residui che misurano la dispersione dei punti al di sopra e al di sotto della retta di regressione.

Nell'impiego della regressione lineare devono essere soddisfatte alcune assunzioni:

1. per ogni valore di x esiste una popolazione di possibili valori di y la cui media giace sulla retta di regressione;
2. per ogni valore di x la distribuzione dei possibili valori di y è normale;
3. la varianza dei valori di y è la stessa per tutti i valori di x;

4. per ogni valore di x le misure di y rappresentano un campione casuale estratto dalla popolazione.

Nella regressione non lineare, invece, si assume che la relazione tra x e y sia non lineare, ma si adatti a un'altra funzione come ad esempio le curve quadratiche. [18]

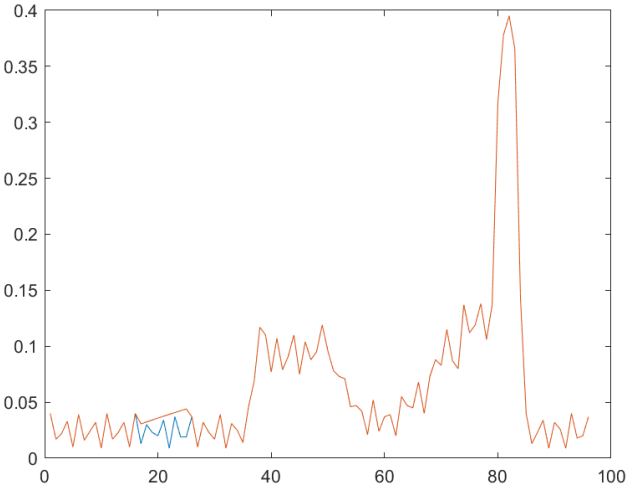

Figura 2.32 – Ricostruzione regressione buco regolare

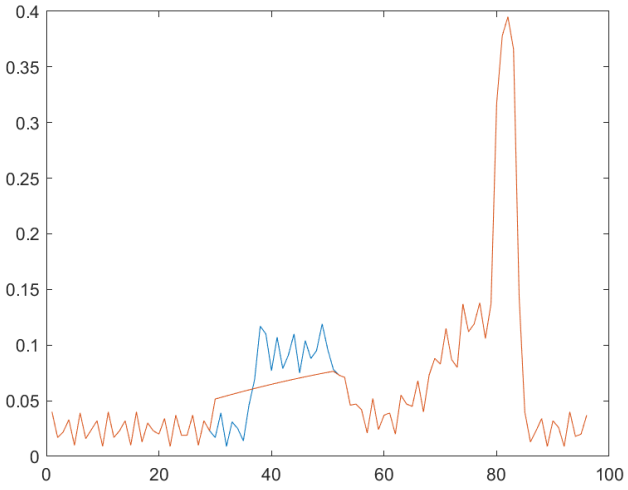

Figura 2.33 – Ricostruzione regressione cambio di livello

[18] C. Capiluppi (2007-08), "La regressione lineare", Facoltà di Scienze della Formazione"

Capitolo 3
Metodi di ricostruzione dei Missing Data

3.1 Metodi di Clustering per creare gruppi di utenti

3.1.1 Introduzione alla Cluster Analysis

Essendo i metodi matematici messi a disposizione da Matlab non sufficienti a ricostruire i buchi di dati, si procede con un raggruppamento delle utenze in classi che hanno lo stesso comportamento elettrico.

Il *Clustering* è un insieme di tecniche di analisi dei dati volta alla selezione e al raggruppamento di elementi omogenei in un insieme di dati sulla base della somiglianza tra gli elementi. Gli algoritmi di *Clustering* tengono conto della distanza reciproca tra gli elementi, e quindi l'appartenenza o meno a un insieme dipende da quanto l'elemento preso in esame è distinto dall'insieme stesso.

La procedura per raggruppare le utenze prevede specifici passi:

1. misurare i dati del modello di carico in una specifica condizione di carico;
2. selezionare ed eliminare i dati errati;
3. selezionare *H* caratteristiche rappresentative;
4. elaborare il modello di carico per costruire *MxH* set di dati di input per la specifica condizione di carico;
5. eseguire la procedura di classificazione per formare gruppi di cluster;
6. determinare i centroidi derivanti dai gruppi delle utenze e dai dati del modello di carico nel dominio del tempo;
7. calcolare gli indici di validità del clustering;
8. calcolare i profili di carico finali per ogni classe di cluster in una specifica condizione di carico.[19]

[19] G. Chicco, "Electrical load representation", Distribuzione e utilizzazione dell'energia elettrica

3.1.2 Applicazione del *k-means*

Matlab mette a disposizione delle funzioni che eseguono il clustering dei dati in input. Come già affermato precedentemente, nel presente lavoro di tesi, si prende in esame l'utenza 1, la quale è la prima che presenta un periodo temporale più lungo privo di dati mancanti pari a 55 giorni, precisamente avente come giorno iniziale il 32-esimo e giorno finale l'87-esimo. Si precisa che il punto di partenza è la creazione di buchi ''fittizi'' che vengano poi ricostruiti con specifici metodi in modo da valutare quanto l'andamento reale dell'energia nel tempo si discosta da quello ricostruito.

Si definisce una matrice 55x96 contenente lungo le righe i singoli giorni del periodo temporale preso in esame e lungo le colonne i 96 quarti d'ora di cui è composta una giornata.

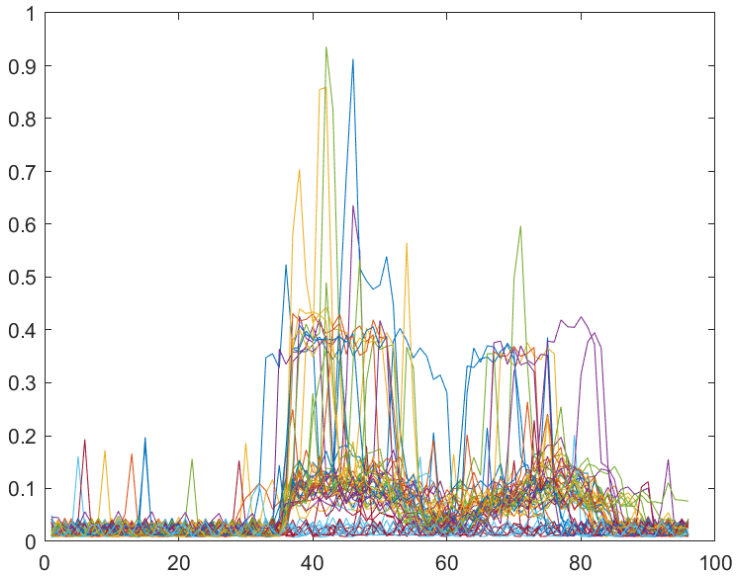

Figura 3.1 – k-means: suddivisione giornaliera

In *Figura 3.1* sono rappresenti i 55 giorni con dati ad ogni quarto d'ora connessi tra loro tramite delle linee.

La funzione *'k-means'*, dopo aver definito il numero *k* di cluster, ripartisce le osservazioni nei *k* cluster e restituisce un vettore *n* x *1* con gli indici di appartenenza a ciascun cluster.

Scelto *k* pari a 5, vengono riportati di seguito l'andamento del modello di carico di ciascun giorno al cluster di appartenenza.

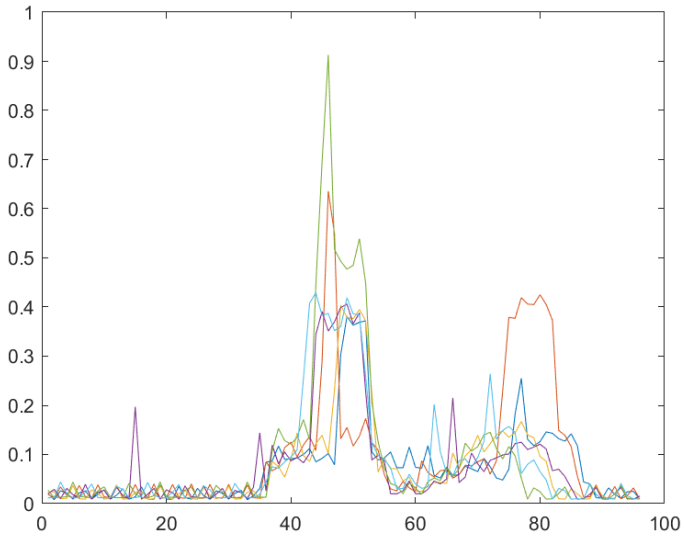

Figura 3.2 – Cluster 1

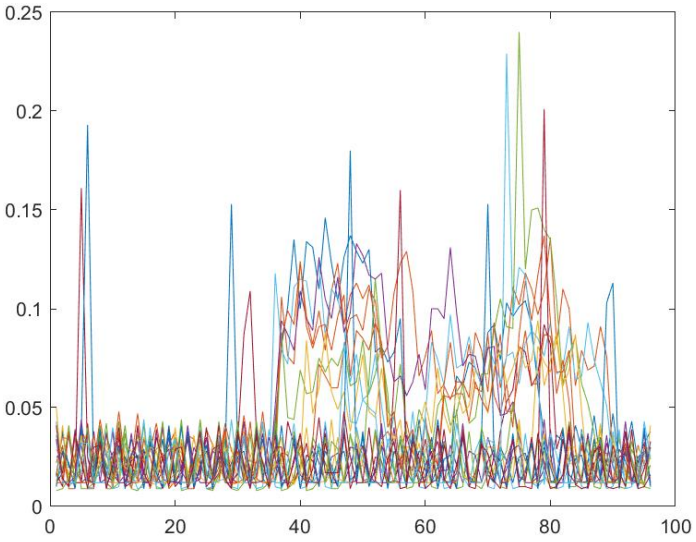

Figura 3.3 – Cluster 2

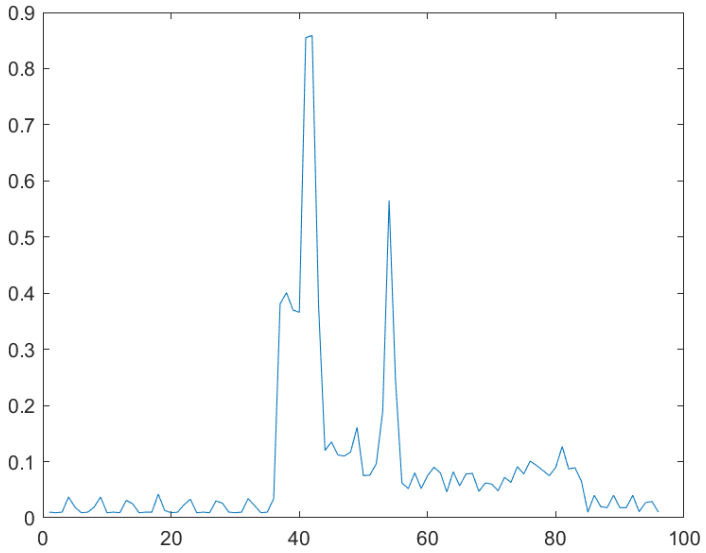

Figura 3.4 – Cluster 3

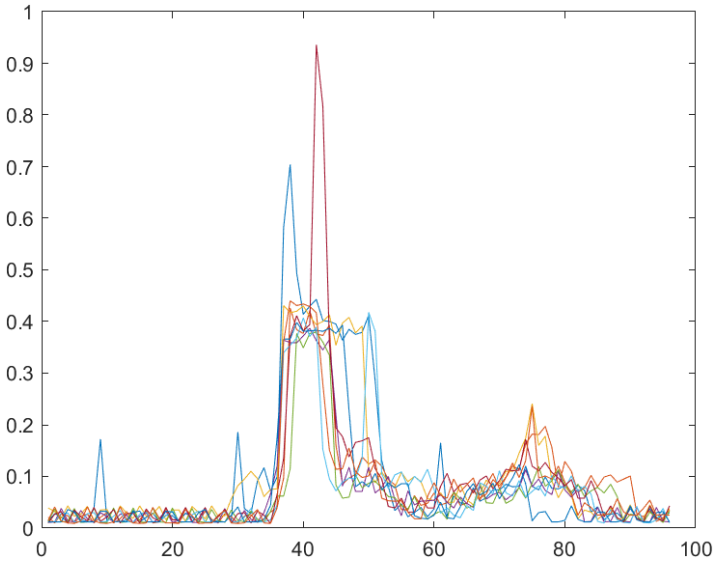

Figura 3.5 – Cluster 4

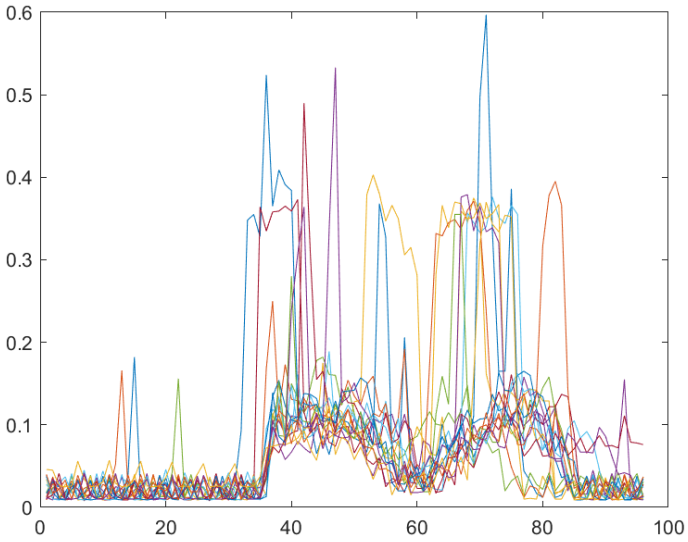

Figura 3.6 – Cluster 5

Per concludere, si determina il centroide di ciascun cluster con l'obiettivo di formare un insieme di giorni ''tipo'' che possano poi servire per classificare i giorni con i dati mancanti.

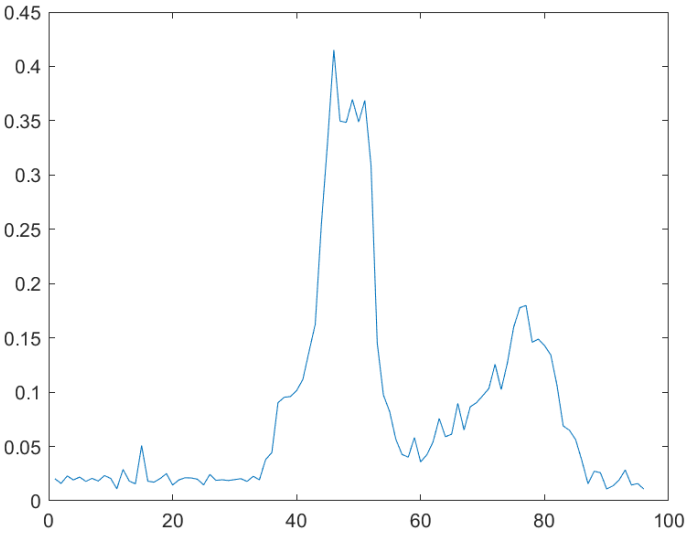

Figura 3.7 – Centroide cluster 1

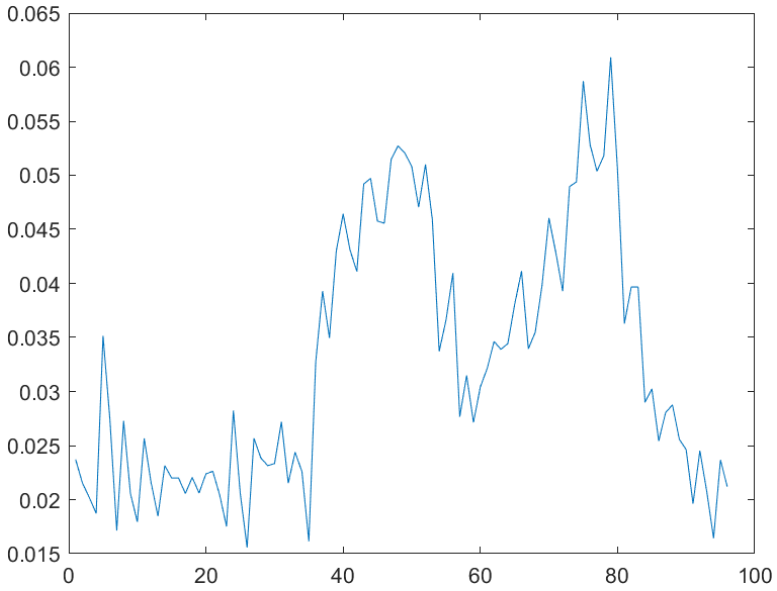

Figura 3.8 – Centroide cluster 2

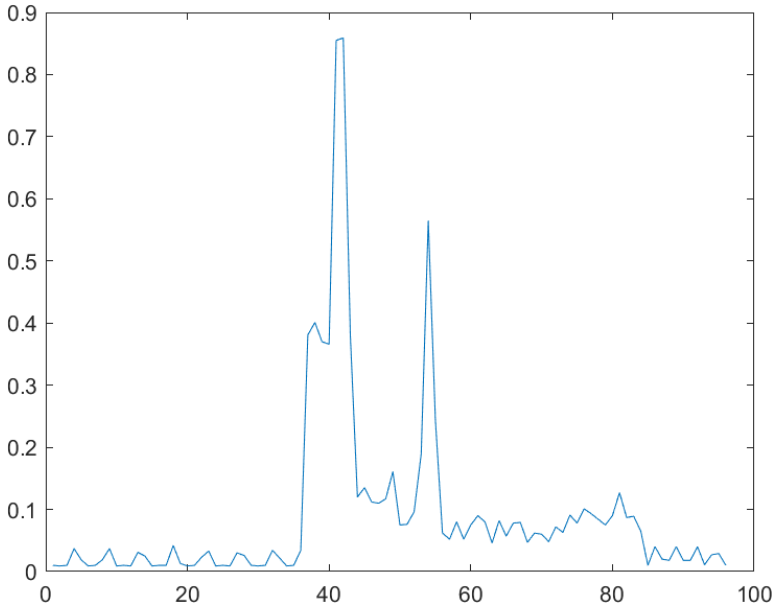

Figura 3.9 – Centroide cluster 3

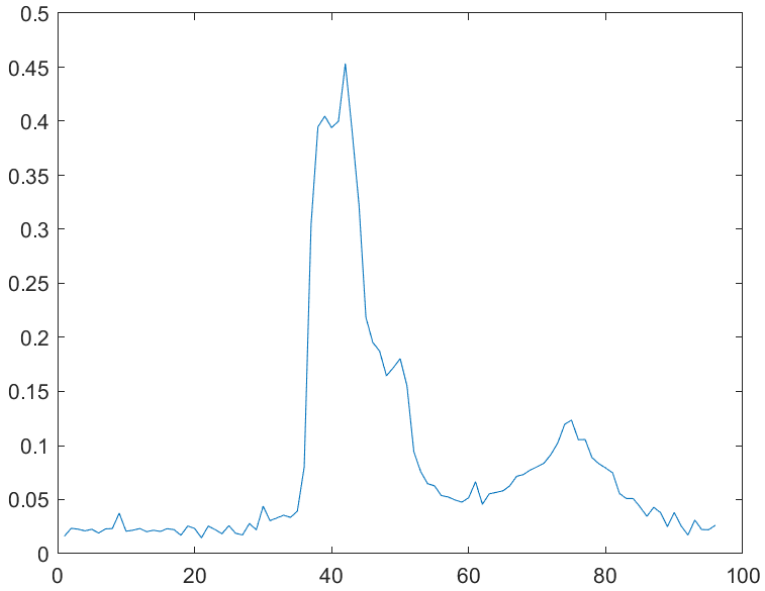

Figura 3.10 – Centroide cluster 4

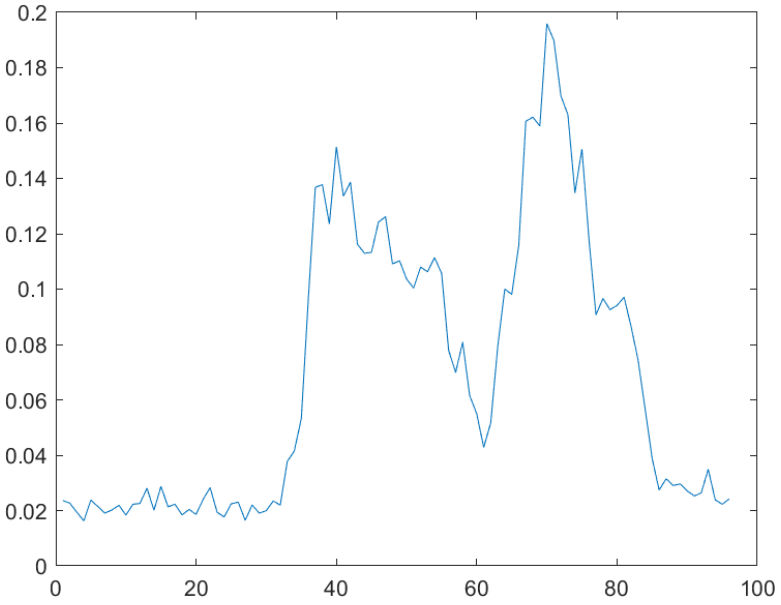

Figura 3.11 – Centroide cluster 5

3.2 Ricostruzione dei dati mancanti per similitudine

Dall'applicazione del metodo del *Clustering* ai 55 giorni presi come riferimento dell'utenza 1, si ottengono come risultato i 5 andamenti prima mostrati che consentono di classificare i giorni con i dati mancanti, attraverso un'associazione al tipo di andamento che presenta una maggiore somiglianza con il giorno preso in considerazione.

Innanzitutto, per poter dimostrare quanto appena detto, si sceglie un giorno arbitrario tra quelli prima citati e si crea un buco di dati mancanti ''falso''. A tal punto, si procede secondo specifici passi di seguito elencati:

1. si costruisce una versione ''ridotta'' della curva in esame eliminando i dati in corrispondenza del periodo temporale in cui è definito il buco;
2. si costruisce una versione ''ridotta'' dei centroidi eliminando i valori in corrispondenza dello stesso periodo temporale;
3. si calcola la distanza tra le due curve ''ridotte'';
4. si assegna la curva in esame al cluster il cui centroide ha una distanza minore rispetto alla curva posta in analisi;
5. si utilizzano le serie temporali complete delle curve per ricostruire i dati mancanti.

Si precisa, che in prima analisi vengono creati buchi appartenenti a zone regolari, in quanto se questo si presentasse in una zona al confine tra due comportamenti diversi, non ci sarebbe modo di capire se l'aumento (o la diminuzione) di energia inizierebbe prima o dopo la parte con i dati mancanti.

Applicando i passi precedentemente elencati a buchi arbitrari di dimensioni variabili per definiti giorni del periodo temporale in questione, si ottengono le ricostruzioni mostrate nelle figure sottostanti. Per ciascun giorno vengono rappresentate sia le 24 ore complete discretizzate in quarti d'ora, sia l'intervallo temporale in cui è presente il buco. Inoltre, al giorno completo (in rosso) si sovrappone, in corrispondenza dei dati mancanti, l'andamento ricostruito.

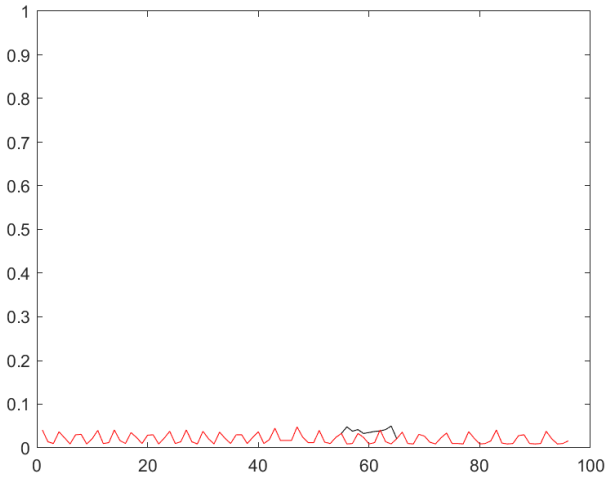

Figura 3.12 - Andamento ricostrutito giorno 14 per similitudine buco di dimesione maggiore

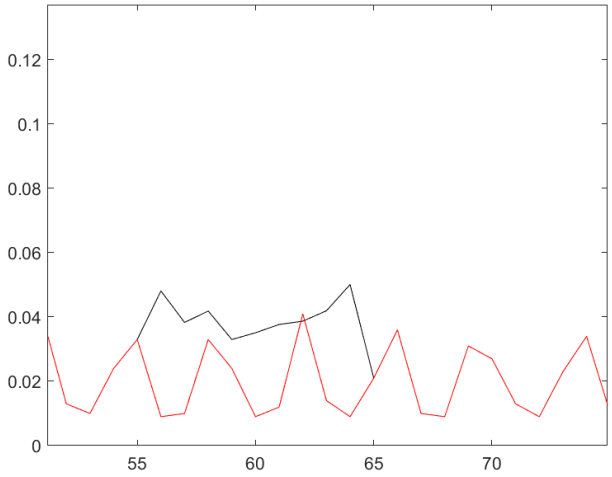

Figura 3.13 – Zoom andamento ricostruito giorno 14 buco di dimensione maggiore

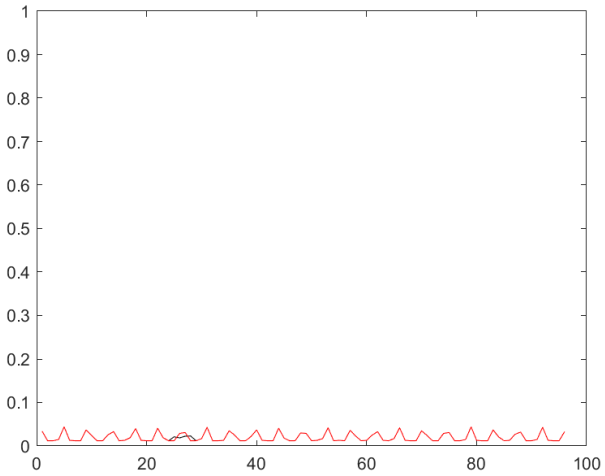

Figura 3.14 – Andamento ricostrutito giorno 14 per similitudine buco di dimesione minore

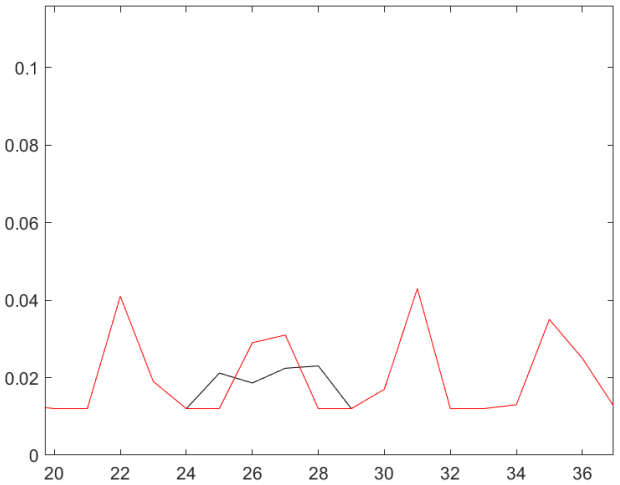

Figura 3.15 – Zoom andamento ricostruito giorno 14 buco di dimensione minore

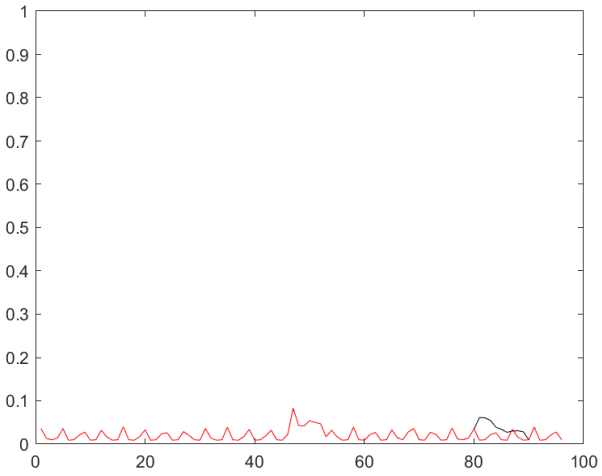

Figura 3.16 - Andamento ricostrutito giorno 27 per similitudine

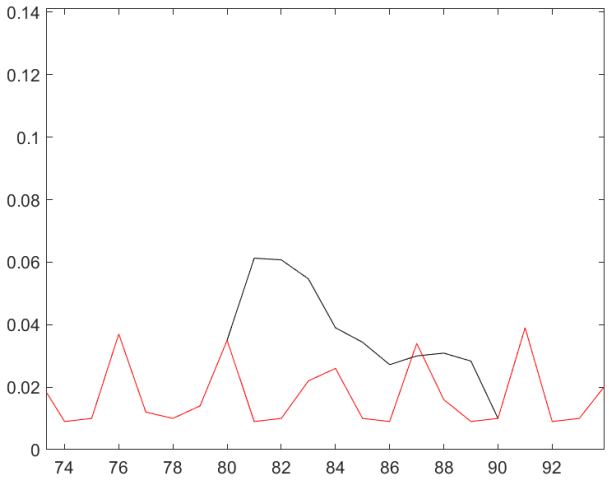

Figura 3.17 - Zoom andamento ricostruito giorno 27

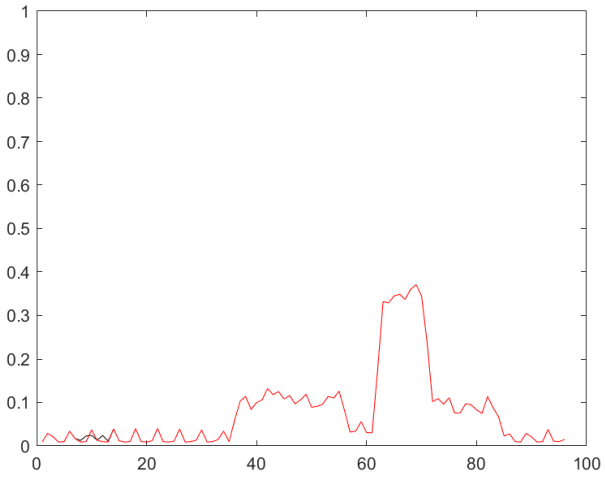

Figura 3.18 - Andamento ricostrutito giorno 32 per similitudine

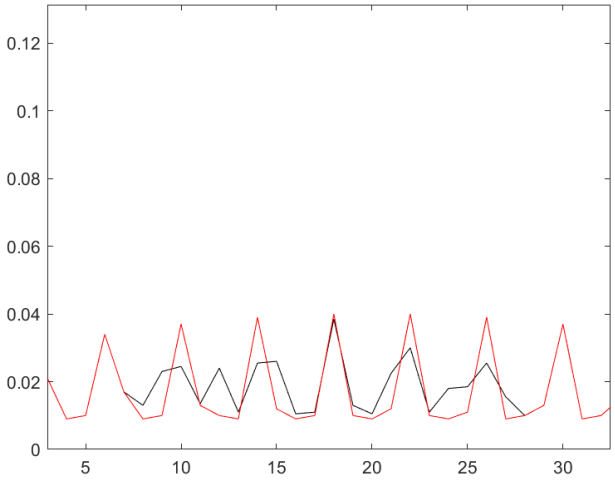

Figura 3.19 – Zoom andamento ricostruito giorno 32

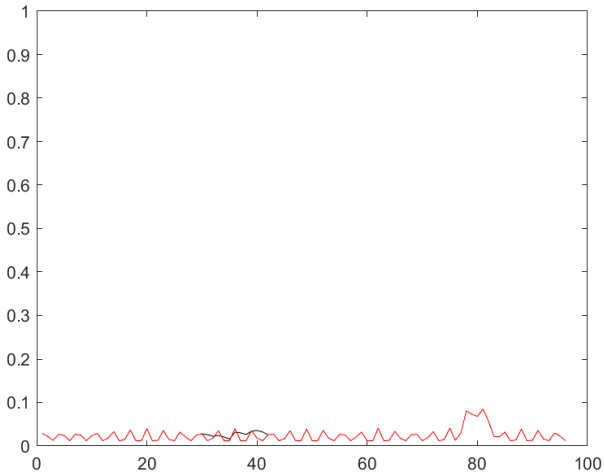

Figura 3.20 - Andamento ricostrutito giorno 6 per similitudine

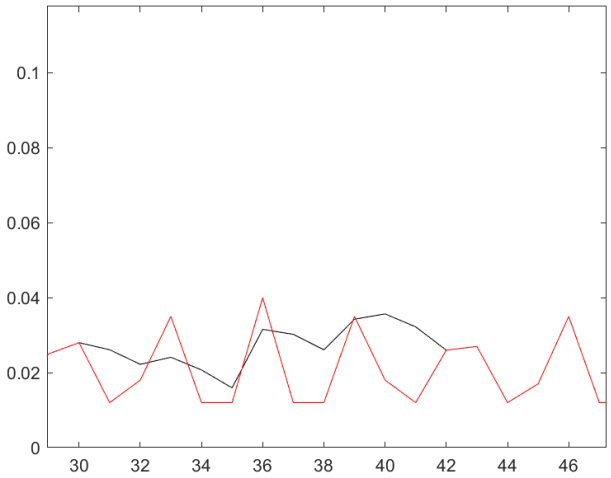

Figura 3.21 – Zoom andamento ricostruito giorno 6

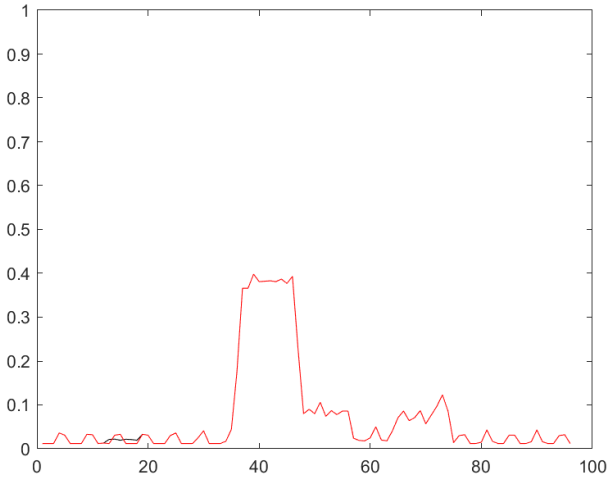

Figura 3.22 - Andamento ricostrutito giorno 45 per similitudine

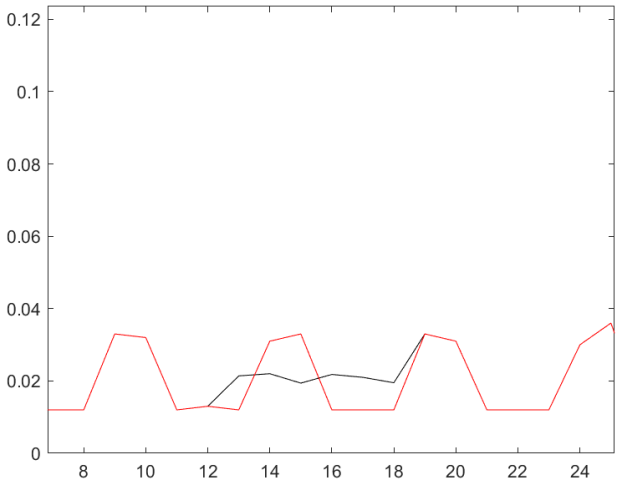

Figura 3.23 – Zoom andamento ricostruito giorno 45

Dai risultati ottenuti si nota che il metodo di ricostruzione applicato risulta ancora insufficiente, essendo l'errore tra i valori esatti e quelli stimati non trascurabile.
Quello che manca è la conoscenza locale; è possibile ottenere ciò con l'analisi armonica, ovvero una ricostruzione effettuata con l'ausilio delle armoniche derivanti degli andamenti circostanti il buco di dati in modo da riprendere la regolarità che manca.

3.3 Analisi armonica e interpretazione della *FFT*

Lo sviluppo in *Serie di Fourier* afferma che un segnale periodico può essere sintetizzato mediante sinusoidi legate armonicamente tra loro, ovvero sinusoidi con frequenza multipla di una frequenza comune, detta frequenza fondamentale e pari all'inverso del periodo del segnale. Pertanto, i segnali periodici possono essere ricostruiti tramite i segnali sinusoidali scegliendo in modo opportuno la frequenza fondamentale (la quale determina il periodo del segnale), e le ampiezze e le fasi delle sinusoidi che determinano la forma del segnale. [20]
Da quanto appena detto, l'obiettivo è quello di ricostruire un andamento applicando la *Fast Fourier Transform* e, successivamente, l'*Inverse Fast Fourier Transform* ai dati successivi al buco, in modo da effettuare una ricostruzione all'indietro con le stesse caratteristiche a livello armonico. Anche in questo caso, come il metodo presentato al paragrafo precedente, l'attenzione si focalizza su un intervallo temporale in cui i dati successivi al buco presentano un andamento regolare.
Con tale metodo applicativo l'attenzione si focalizza sulle oscillazioni locali. Si precisa che la conoscenza locale dà ottimi risultati soltanto quando ha senso utilizzarla, ovvero quando ci sono le armoniche prevalenti. Ciò può essere valutato osservando lo spettro armonico sempre dello stesso intervallo a cui precedentemente è stata applicata la *FFT* e la *IFFT*.

[20] L. Verdoliva, "Sviluppo in Serie di Fourier", Appunti di Teoria dei Segnali, a.a. 2010-2011

Vengono mostrati di seguito, a titolo di esempio, gli andamenti di due giorni che presentano una certa regolarità.

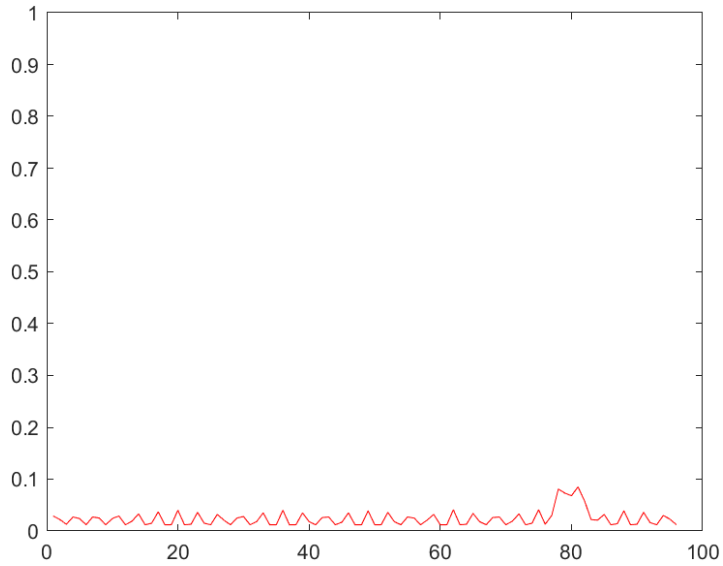

Figura 3.24 – Profilo di carico giorno 6

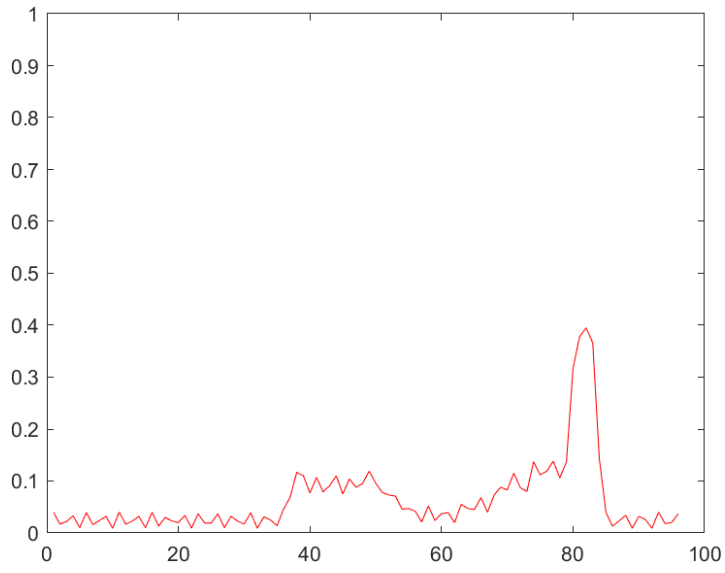

Figura 3.25 – Profilo di carico giorno 46

Il primo andamento rappresentato in *Figura 3.24* risulta essere quasi totalmente regolare, il secondo invece soltanto nelle prime ore della giornata.

Per il primo si consideri ora un buco ''fittizio'' tra gli i punti 50 e 58 sull'asse del tempo (ascisse) e si prendano come riferimento per l'analisi armonica i successivi 16 valori. Si ottiene lo spettro armonico raffigurato in figura:

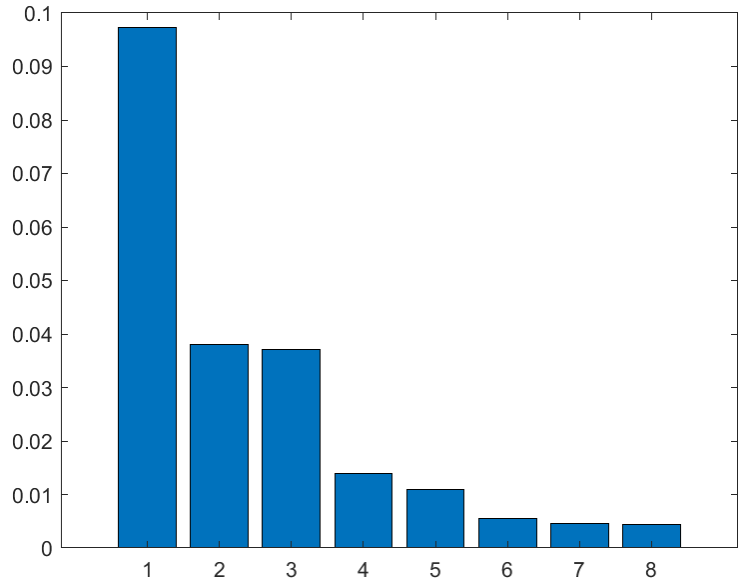

Figura 3.26 – Spettro armonico giorno 6

Ripetendo il ragionamento per il giorno considerato in *Figura 3.25*, in cui il buco si fissa tra i punti 15 e 20 e si prendono in analisi i 23 valori successivi per l'analisi armonica, includendo, pertanto, anche il cambio di livello. Lo spettro armonico è il seguente:

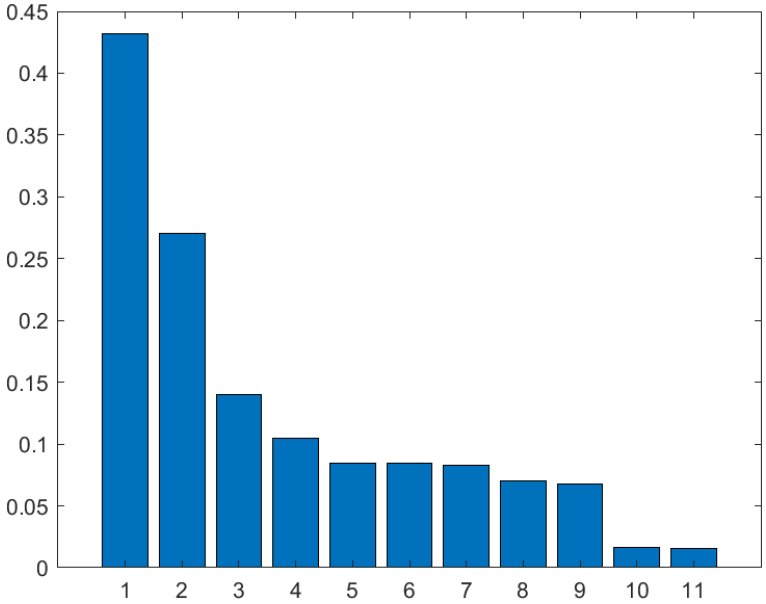

Figura 3.27 – Spettro armonico giorno 46

Dagli spettri è stata prima di tutto esclusa la prima armonica, ovvero la componente continua in modo da individuare le armoniche non trascurabili.

Per valutare il numero di armoniche prevalenti che risultano soddisfacenti a rappresentare l'andamento, si imposta il tutto come una distribuzione di probabilità e si calcola il numero di armoniche necessarie a coprire il 50% dell'area totale. Ciò è ottenuto sommando, di volta in volta, l'ampiezza dello spettro in corrispondenza di ciascuna frequenza, per la somma di tutte le ampiezze costituenti il medesimo. Quando tale somma giunge al 50% dell'area totale, la posizione corrispondente indica il numero di armoniche prevalenti.

In *Figura 3.26* il numero di armoniche prevalenti, ad esempio, è pari a 2. Invece, in *Figura 3.27,* avendo considerato anche il cambio di livello, si può osservare innanzitutto un numero maggiore di armoniche di ampiezza minore e, prima di arrivare al limite imposto, si prosegue verso un valore più elevato.

Esiste un'ulteriore tecnica di analisi, la *media mobile*, che dà un'indicazione sintetica dell'andamento dell'energia in un determinato giorno. Dal punto di vista matematico, questo indicatore effettua una ''media'' di una definita quantità di dati, rispetto ad una finestra temporale ''mobile'', perché considera soltanto le ultime *N* rilevazioni in ordine di tempo. Nel calcolo di una media mobile, infatti, il numero di elementi è fisso, mentre la finestra temporale considerata avanza; il dato vecchio della serie viene sostituito con quello nuovo.[21]

Considerando innanzitutto come giorni di riferimento gli stessi analizzati in *Figura 3.24* e *Figura 3.25*, vengono poste di seguito i risultati ottenuti dall'applicazione della media mobile.

In *Figura 3.28* si osserva che fino al punto 42 l'andamento è regolare, relativamente stabile lungo il periodo temporale e con una certa periodicità: questo permette di capire quanti e quali possono essere i dati da considerare nell'analisi armonica.

[21] Ufficio Studi Money.it, (2019), “Le medie mobili: cosa sono, come si calcolano e quali sono le più usate”, Money.it <https://www.money.it/Le-medie-mobili-cosa-sono-come-si>

Inoltre, dalla rappresentazione della banda del valore massimo e del valore minimo, si ha un'ulteriore conferma che i valori si mantengono regolari restando all'interno della fascia indicata. Differente è il ragionamento se si considerassero anche i dati successivi, in quanto, come si nota, non sarebbe più rispettata la banda e di conseguenza, la ricostruzione tramite la *FFT* si discosta da quanto accade realmente. In *Figura 3.29* viene mostrata la ricostruzione escludendo il cambio di livello; la *Figura 3.30* invece è un esempio dell'andamento che si ottiene se si includesse anche l'intervallo in cui i dati aumentano di valore. La conclusione che si può estrarre da questi due esempi è che ci sono delle porzioni in cui si può lavorare con le armoniche e porzioni in cui non si può lavorare.

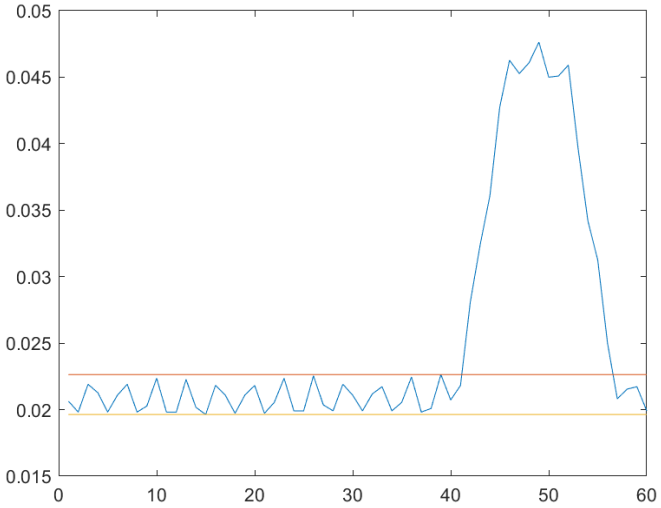

Figura 3.28 – Media mobile giorno 6

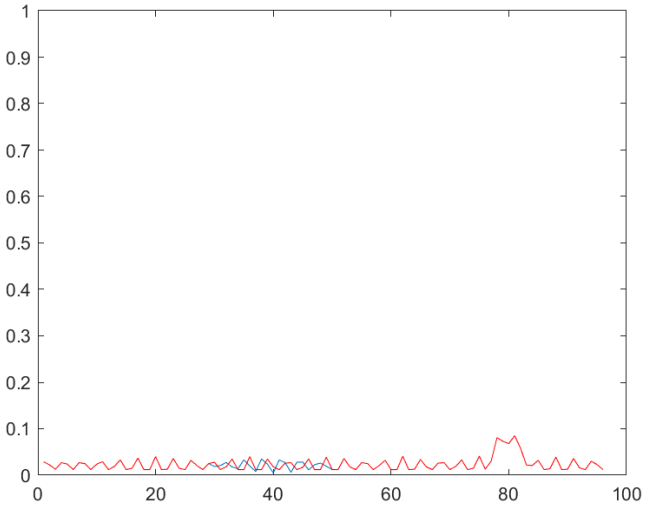

Figura 3.29 – Andamento ricostruito giorno 6 buco regolare

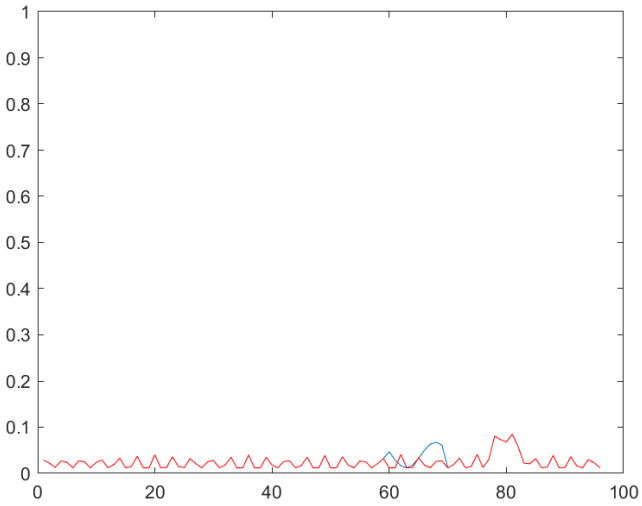

Figura 3.30 – Andamento ricostruito giorno 6 cambio di livello

In *Figura 3.31* invece si nota chiaramente la variabilità presente, la quale presenta una certa similitudine con l'andamento osservato in *Figura 3.25.* Ciò consente di dedurre che quando c'è una grande differenza la *Fast Fourier Transform* non ha senso perché si ricostruiscono armoniche di ordine basso e con andamenti aventi comportamenti troppo differenti. Anche il limite di banda del valore minimo e del valore massimo ha senso soltanto fino all'istante 42. In *Figura 3.32* viene mostrata la ricostruzione tramite la *FFT* nel caso in esame, ottenendo un risultato che si discosta totalmente dal vero andamento.

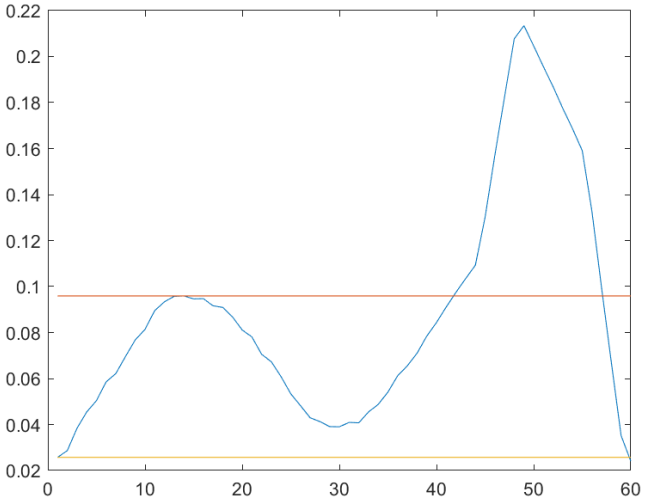

Figura 3.31- Media mobile giorno 46

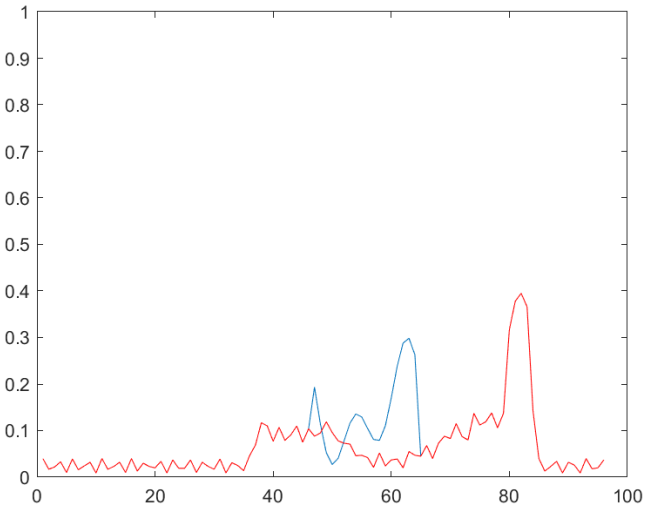

Figura 3.32 – Andamento ricostruito giorno 46 cambio di livello ampio

Dalle considerazioni appena fatte, nel presente lavoro di tesi, l'attenzione si focalizza sulla ricostruzione dei dati che presentano per un periodo sufficientemente ampio una certa regolarità. Il metodo impiegato, tramite l'*Analisi di Fourier*, dà risultati soddisfacenti per le ipotesi fatte; in particolare, esso non solo si basa su ciò che accade nel passato (come i metodi classici noti), ma tiene conto anche del futuro. La peculiarità di quest'ultimo è la conoscenza del dominio, la quale consente di effettuare una suddivisione settimanale e/o giornaliera dei giorni simili.
Di seguito si riportano gli andamenti ricostruiti di specifici giorni che rispettano le ipotesi fatte.

Tuttavia, nell'ultimo esempio mostrato si può fare un'ulteriore osservazione.
Applicando il metodo di ricostruzione fino al punto 42 della *Figura 3.33*, i valori restano nella banda indicata e i risultati ottenuti con le armoniche principali di Fourier sono soddisfacenti in termini di ampiezza, andando a rispettare il limite per le armoniche prevalenti esposto a inizio paragrafo.

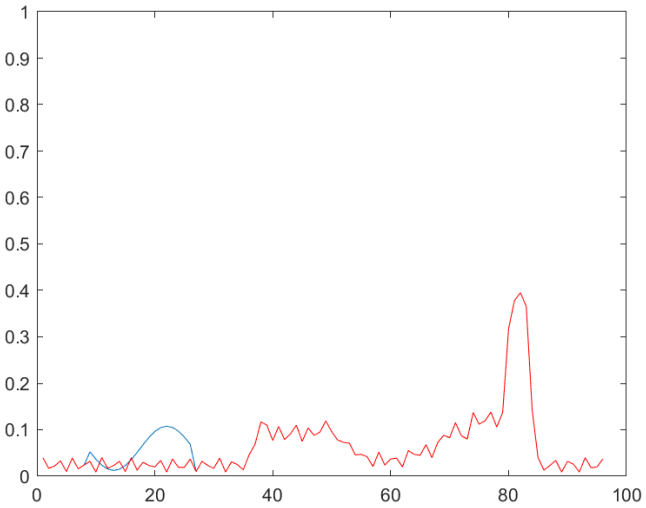

Figura 3.33 – Andamento ricostruito giorno 46 cambio di livello

Da questi due esempi particolari mostrati si afferma che il metodo di analisi basato su Fourier è idoneo soltanto in casi particolari di sostituzione di dati e dà una ricostruzione accettabile di quegli andamenti che con altri metodi classici è difficile ottenere limitandosi alla conoscenza passata senza tener presente di regolarità o eventuali cambi di livello.

Tuttavia, ci sono casi in cui esso non può essere applicato: la conoscenza del dominio dà informazioni circa l'andamento, ma non è sufficiente alla corretta ricostruzione.

Di seguito sono riportati altri casi particolari di applicazione, andando a precisare quando i risultati ottenuti sono accettabili e quando non lo sono.

In *Figura 3.34* e in *Figura 3.35* è mostrata la media mobile del giorno 3 la quale, come si evince non risulta regolare, con una differenziazione della banda massimo/minimo per due differenti intervalli di tempo.

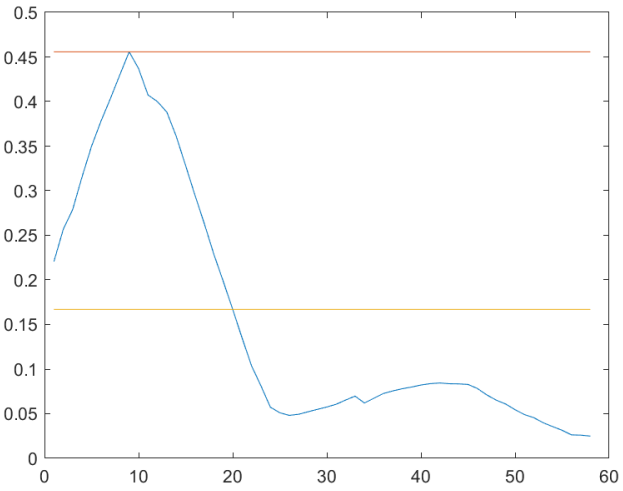

Figura 3.34- Media mobile giorno 3 prima fascia

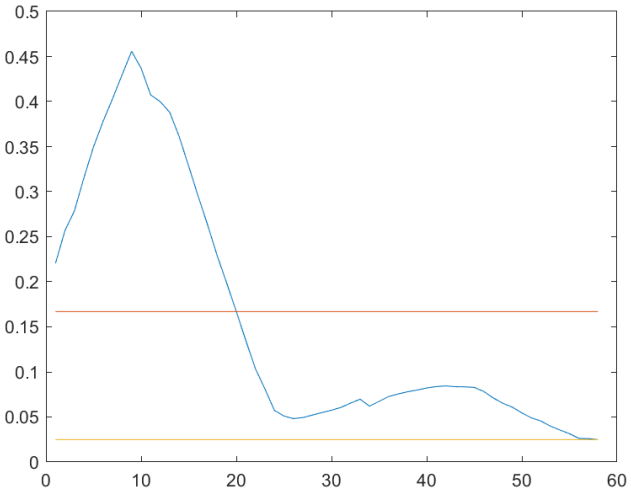

Figura 3.35 - Media mobile giorno 3 seconda fascia

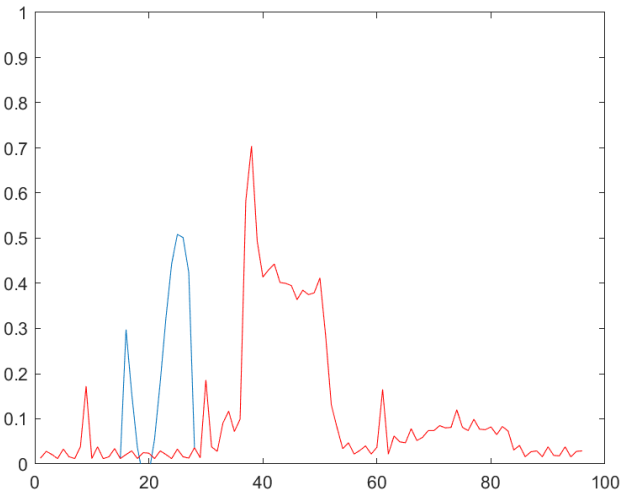

Figura 3.36 – Andamento ricostruo gorno 3 cambio di livello ampio

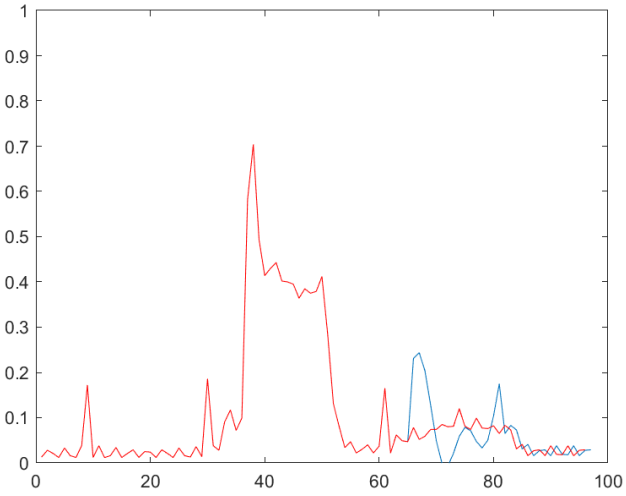

Figura 3.37 – Andamento ricostrutito giorno 3 riferimento indietro

In entrambe le situazioni essendo la variazione frequente e differente in ampiezza, il metodo di ricostruzione non è valido, portando a risultati con un elevato margine di errore.

Si precisa che in *Figura 3.37* è stato applicato il metodo di Fourier prendendo come riferimento per l'applicazione un intervallo temporale passato anziché futuro.

Si consideri di seguito il giorno 8.

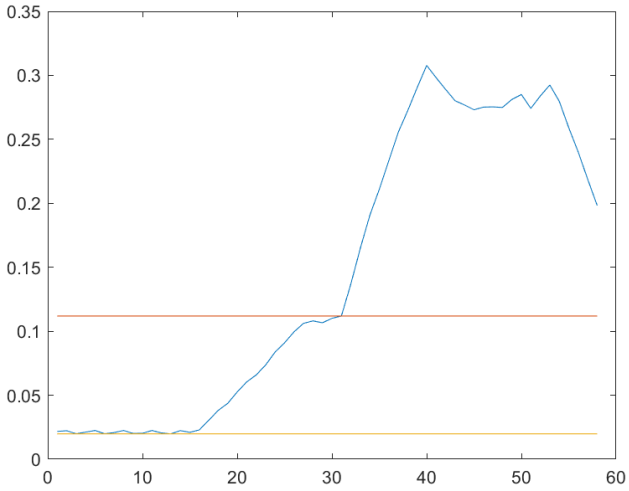

Figura 3.38 - Media mobile giorno 8 prima fascia

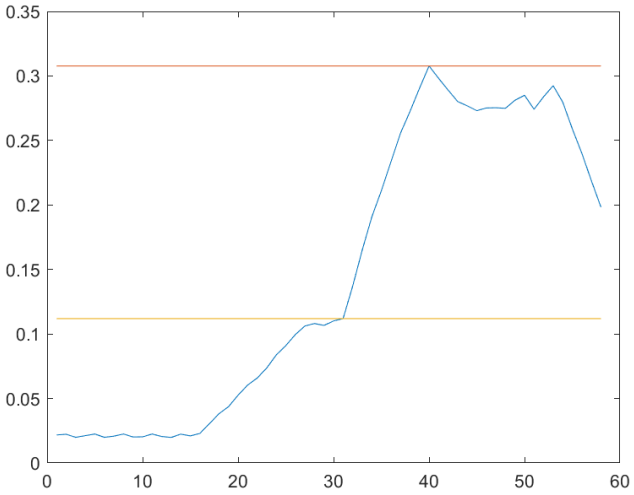

Figura 3.39 - Media mobile giorno 8 seconda fascia

Le considerazioni circa la media mobile sono le medesime fatte per il caso precedente.

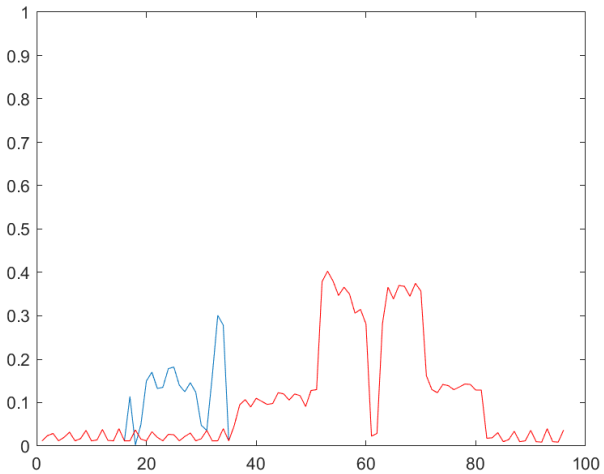

Figura 3.40 – Andamento ricostruo gorno 3 cambio di livello ampio

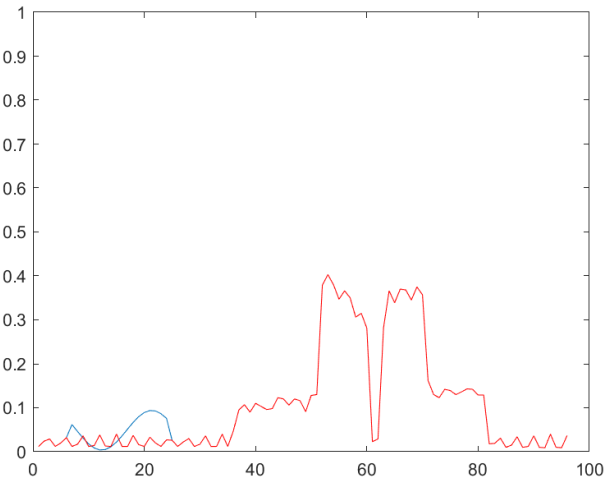

Figura 3.41 – Andamento ricostruito buco regolare

Il primo andamento ricostruito, andando ad incorporare nella *FFT* anche il periodo con un cambio di livello tra gli istanti 42 e 60, non è considerato accettabile, in quanto la ricostruzione non è utile. Differente invece è ciò mostrato in *Figura 3.41* poiché, anche se si tiene contro di una variazione ridotta tra i punti 37 e 42 e anche se la ricostruzione non è perfetta, l'errore è basso e le armoniche prevalenti sono sufficienti a rappresentare l'andamento ricostruito.

Infine, viene presentato il giorno 17 come ultimo caso di esempio.

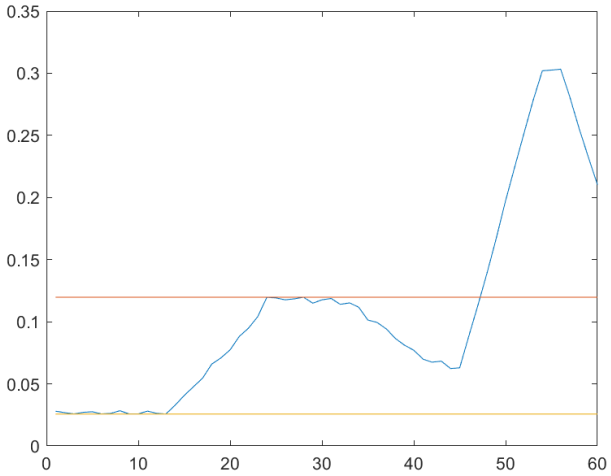

Figura 3.42 - Media mobile giorno 17 prima fascia

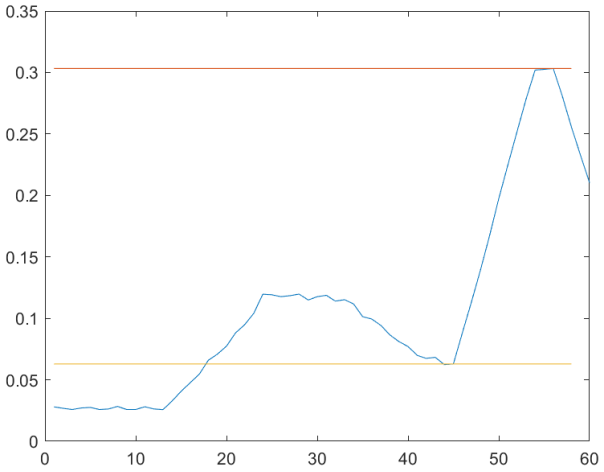

Figura 3.43 - Media mobile giorno 17 seconda fascia

Anche in questa situazione è possibile distinguere due periodi temporali differenti con due differenti bande per il massimo e il minimo.

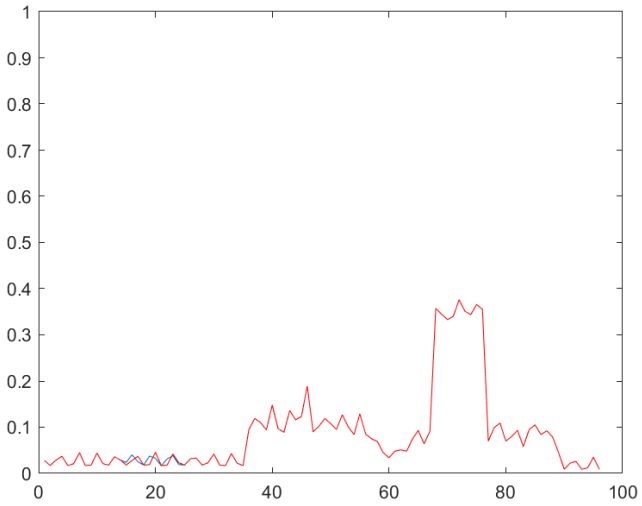

Figura 3.44 – Andamento ricostruito giorno 17 buco regolare

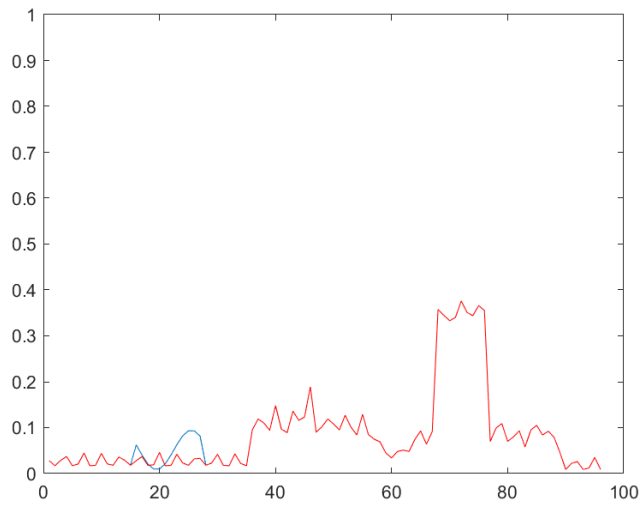

Figura 3.45– Andamento ricostruito giorno 17 cambio di livello

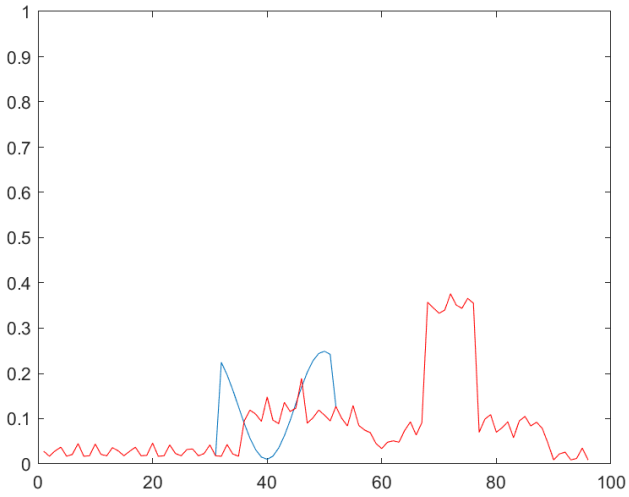

Figura 3.46 –Secondo andamento ricostruito giorno 17 cambio di livello

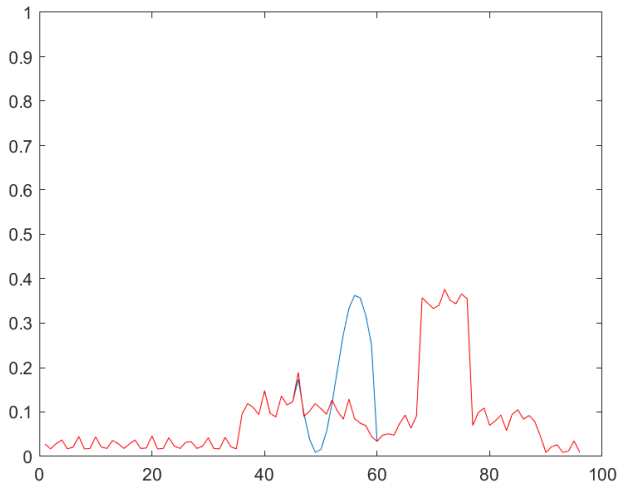

Figura 3.47 - Andamento ricostruito giorno 17 cambio di livello ampio

Le prime due figure danno risultati che sono considerati corretti e, nello specifico, la *Figura 3.44* considera nell'applicazione del metodo un intervallo di riferimento stabile; come si nota la ricostruzione è quasi perfetta.

Nelle ultime si nota che l'andamento presenta una grande variazione rispetto alla realtà e ciò avvalora la teoria secondo cui l'analisi armonica si applica soltanto a determinate porzioni che presentano una regolarità.

3.4 Verso la ricostruzione completa

Come dimostrato, l'analisi armonica è applicabile a specifiche e particolari porzioni di profili di carico, ma ciò non basta per ricostruire l'andamento completo.

Tramite l'ausilio della media mobile è possibile riconoscere una certa regolarità e servendosi dell'analisi armonica questa regolarità viene impiegata nel metodo di ricostruzione. Nelle porzioni temporali in cui si presenta un cambio di livello sarebbe opportuno togliere il valore medio in modo che le oscillazioni iniziali si avvicinano ad un andamento sinusoidale che oscilla intorno lo zero: l'obiettivo è trasformare l'andamento iniziale in modo da isolare il valore medio rispetto alle oscillazioni. Le ultime non fuoriescono dalla fascia di minimo/massimo e la ricostruzione non si discosta molto dal valore effettivo. Per fare ciò è necessaria una stima della media

mobile partendo da una valutazione preliminare delle frequenze tramite l'analisi di Fourier in modo da eliminare quelle più elevate che possono distogliere l'andamento dalla sua regolarità. A quelle restanti si sottrae la media mobile con un passo adeguato che possa raccogliere queste frequenze ottenendo delle oscillazioni pressoché stabili. Si effettua la ricostruzione considerando solo quest'ultime con un risultato nettamente migliore.

Per dimostrare quanto appena detto si consideri il giorno 46 già analizzato nel paragrafo precedente e nel cui seguito, in *Figura 3.48,* è riportato l'andamento completo.

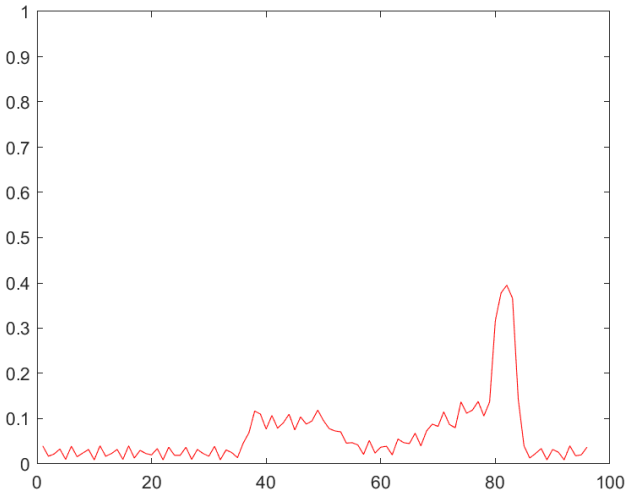

Figura 3.48 - Profilo di carico giorno 46

Posto un buco tra gli istanti 37 e 43 (ovvero in cui si ha una variazione rispetto all'andamento iniziale), applicando l'analisi armonica alla porzione di dati precedente il buco si ha una determinata oscillazione e lo stesso per la parte successiva. Le oscillazioni ottenute hanno armoniche simili nelle due porzioni considerate con i valori minimi e i valori massimi che non si discostano dalla fascia prefissata. Dall'analisi della media mobile scaturisce che passando dalla prima alla seconda parte c'è stato un cambio di livello e lo scopo è quello di raccordare queste due, anche se non si hanno informazioni dettagliate su ciò che accade nella porzione intermedia sconosciuta.

Per ricostruire i buchi con dati mancanti si combinano le informazioni derivanti dalla media mobile, dall'analisi di Fourier e dal toolbox *Fill Missing Values* analizzato nel

presente lavoro, ottenendo un andamento che si avvicina a quello esatto. Si riporta in *Figura 3.49* ciò che si ottiene applicando soltanto l'analisi armonica a un intervallo di tempo non periodico che, come dimostrato, dà un risultato errato e privo di senso. Mentre in *Figura 3.50* si mostra ciò che è ottenuto applicando la conoscenza del dominio di applicazione con il metodo classico citato.

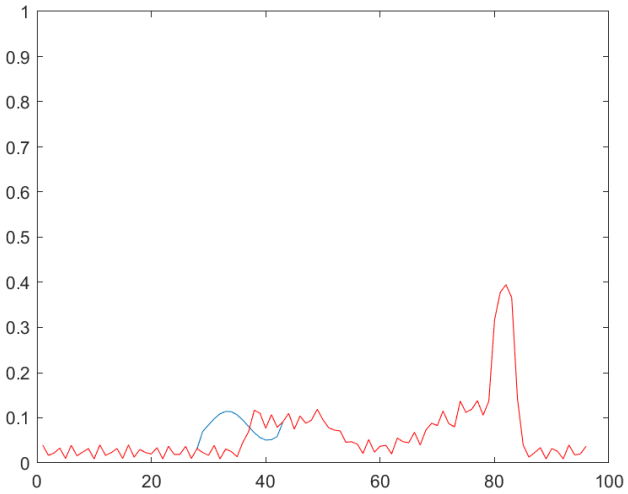

Figura 3.49 – Andamento ricostruito giorno 46 tramite analisi armonica

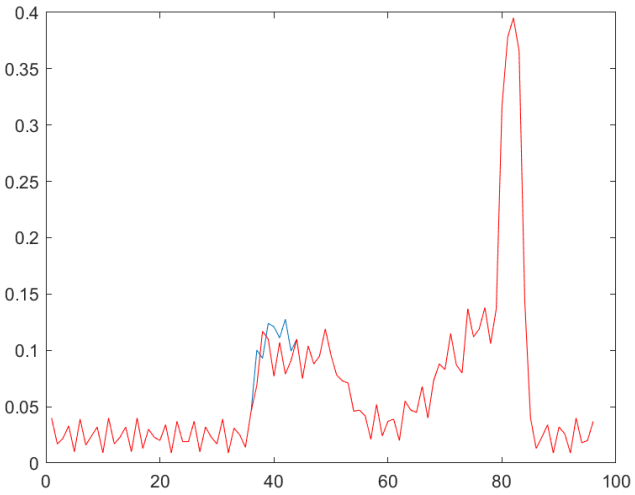

Figura 3.50 – Ricostruzione completa giorno 46

Riprendendo il giorno 17 esposto al paragrafo precedente con un buco tra gli istanti 31 e 45, si riporta in *Figura 3.51* la ricostruzione ottenuta considerando soltanto l'analisi di Fourier applicata ad un periodo non regolare e pertanto errata e in *Figura 3.52* l'andamento ricostruito combinando l'analisi di Fourier con l'ausilio delle reti neurali.

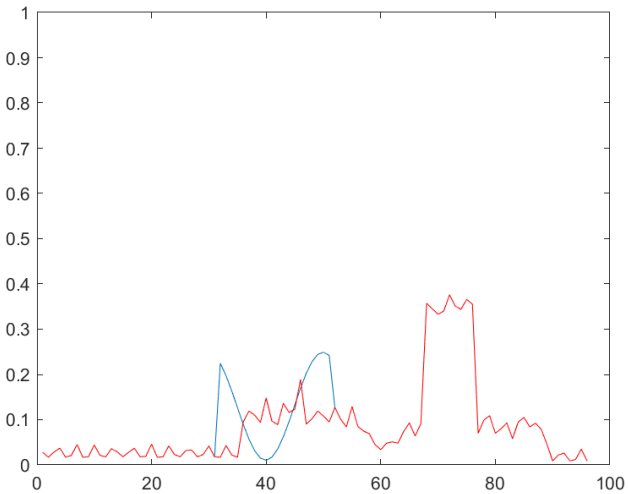

Figura 3.51 – Andamento ricostruito giorno 17 tramite analisi armonica

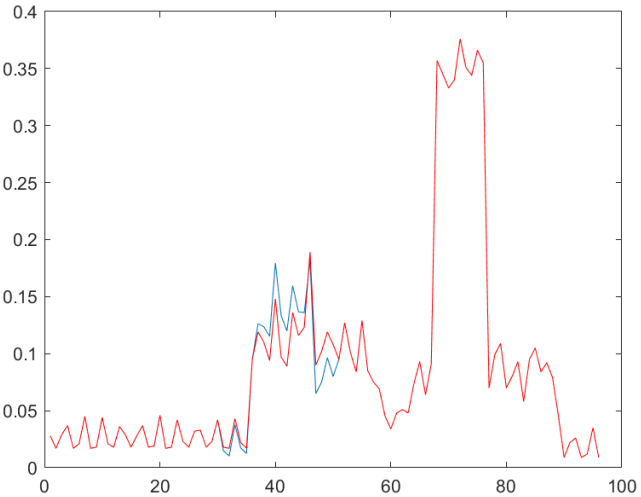

Figura 3.52 – Ricostruzione completa giorno 17

Il medesimo ragionamento è applicato al buco di dati presente tra gli istanti 45 e 60 del giorno 17 come mostrato nelle figure sottostanti.

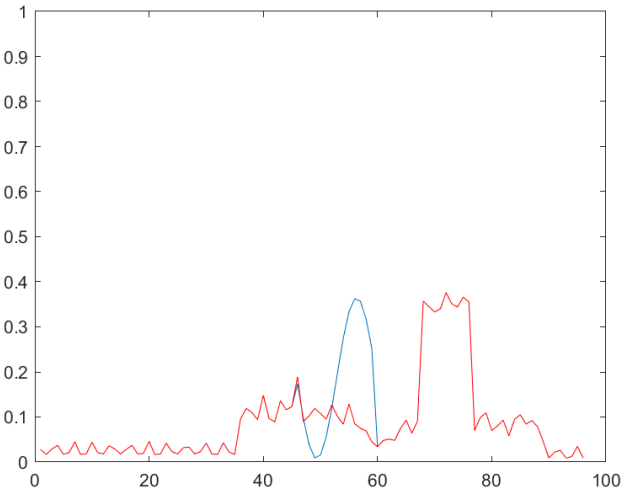

Figura 3.53 – Secondo andamento ricostruito giorno 17 Tramite analisi armonica

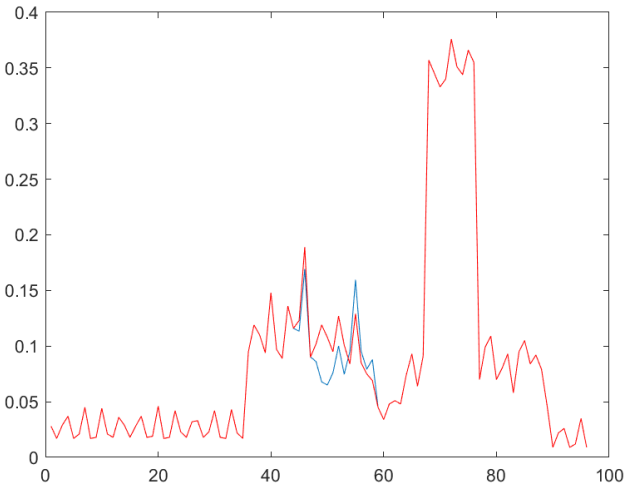

Figura 3.54 – Secondo andamento ricostruito completo giorno 17

Dai risultati ottenuti si giunge alla conclusione che per ottenere una ricostruzione che possa considerarsi accettabile anche negli intervalli non regolari, è necessario combinare la conoscenza del dominio con i metodi di analisi noti.

Conclusioni

Nel seguente lavoro di tesi si è cercato di dare una nuova visione al trattamento dei dati mancanti. La perdita di dati comporta delle difficoltà in fase di analisi producendo stime distorte, riducendo l'efficienza e l'affidabilità di uno studio e giungendo a conclusioni talvolta errate.

Le varie metodologie di gestione dei *missing data* possono portare a risultati differenti per uno stesso campione e con valori che si discostano radicalmente da quanto accade nella realtà. Si è dimostrato che l'applicazione di procedure che si basano sulla conoscenza locale portano a risultati migliori rispetto a formulazioni matematiche generiche.

La migliore metodologia è ottenuta analizzando singolarmente le differenti situazioni in quanto la soluzione non è univoca: è stato necessario individuare gli andamenti prevalenti e ricostruire i dati mancanti servendosi delle caratteristiche delle curve di carico circostanti il buco di dati. Le informazioni derivanti dall'*Analisi di Fourier* e dalla *media mobile* hanno consentito di individuare eventuali regolarità.

Nel momento in cui ci si trova dinanzi a situazioni isolate con un improvviso cambio di livello nell'assorbimento di energia elettrica da parte delle utenze residenziali, si è giunti alla conclusione che un giusto compromesso tra la conoscenza del dominio di applicazione e i metodi matematici noti porta a una ricostruzione affidabile e che possa considerarsi accettabile in termini di errore rispetto al vero andamento.

Fonti

R. Pertile, ''Aspetti critici negli studi clinici: problemi di metodo e applicazione'', Tesi di dottorato, Università degli Studi di Padova, a.a. 2008, relatore P. Facchin;

Alvira Swalin (2018), ''How to Handle Missing Data''
< https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>;

Huyn Kang (2013), ''The prevention and handling of the missing data''
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/>;

Prof.ssa L. Neri (2020), Capitolo 7, ''La qualità dei dati: l'imputazione dei dati mancanti'';

Terna (2006) ''Procedure operative per la gestione delle informazioni e dei dati nell'ambito del sistema di misura''
<http://collaudo.download.terna.it/terna/0000/0105/65.pdf>;

N. Lombardi, ''L'Analisi di Fourier e la DFT'', Tesi triennale, Università degli studi di Cagliari, a.a. 2019-2020, relatore G. Rodriguez;

''Teorema del campionamento di Nyquist-Shannon''
<https://it.wikipedia.org/wiki/Teorema_del_campionamento_di_Nyquist-Shannon>;

G. Chicco, ''Serie di Fourier e campionamento'', Distribuzione e utilizzazione dell'energia elettrica, Politecnico di Torino, a.a. 2020-2021;

''Load duration curve''
<https://illustrationprize.com/it/519-load-duration-curve.html>;

G. Chicco, ''Electrical load representation'', Distribuzione e utilizzazione dell'energia elettrica, Politecnico di Torino, a.a. 2020-2021;

L. Verdoliva, ''Sviluppo in Serie di Fourier'', Appunti di Teoria dei Segnali, Università degli Studi di Napoli Federico II, a.a. 2010-2011;

Ufficio Studi Money.it, (2019), ''Le medie mobili: cosa sono, come si calcolano e quali sono le più usate'', Money.it
<https://www.money.it/Le-medie-mobili-cosa-sono-come-si>;

''Interpolazione polinomiale''
< https://it.wikipedia.org/wiki/Interpolazione_polinomiale>;

P. Ghelardoni, Capitolo 5, ''Interpolazione polinomiale'', Unipi, a.a. 2016-2017;

Ashwin Raj (Ottobre 2020), ''Unlocking the True Power of Support Vector Regression''
<https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0>;

MathWorks, ''misdata''
< https://it.mathworks.com/help/ident/ref/misdata.html>;

S. Righini, ''Applicazione del Machine Learning a scenari di inquinamento ambientale'', Tesi di laurea, Università di Bologna, a.a.2017-18, relatore V. Ghini;

C. Capiluppi, ''La regressione lineare'', Facoltà di Scienze della Formazione', Università di Verona, a.a. 2007-2008.

Ringraziamenti

Ringrazio il Professore Gianfranco Chicco, relatore di questa tesi di laurea, per avermi seguito in questo percorso e per essere stato sempre disponibile per chiarimenti e per consigli con grande professionalità. Lo ringrazio per avermi supportata costantemente, trasmettendomi grande conoscenza circa il tema trattato; l'entusiasmo e l'impegno che ho mantenuto durante l'intero lavoro sono dovuti alla profonda esperienza del mio relatore.

Ringrazio il mio correlatore, l'Ingegnere Andrea Mazza, per il suo aiuto e per la sua disponibilità dimostratemi durante l'intero periodo. Lo ringrazio per avermi chiarito dubbi e incertezze ogni qualvolta ne avessi bisogno e per avermi dato anche consigli costruttivi.

Ringrazio i miei genitori per avermi appoggiato sia moralmente che economicamente in qualsiasi decisione. Mi hanno permesso di arrivare fin qui, al termine del mio percorso di studi. Mi hanno incoraggiata nei momenti di difficoltà e sono stati sempre disponibili nel momento del bisogno.

Ringrazio mia sorella Carmela che con la sua esperienza mi ha dato consigli formativi, mi ha sempre fornito un aiuto e mi ha rimproverato quando ce ne fosse bisogno.
Ringrazio Erika, per avermi sostituito in alcuni momenti come sorella maggiore, per avermi sopportato i giorni antecedenti gli esami, per avermi strappato un sorriso, per avermi fatto sentire la sua presenza nei momenti difficili e per essersi messa sempre disposizione.

Ringrazio Daniele, il mio ragazzo, per essermi stata vicino costantemente, per aver sempre creduto in me, per aver speso parole preziose nei miei confronti e per aver mostrato ammirazione e fiducia. Senza la sua presenza il mio percorso di studio e di

vita non sarebbe stato lo stesso, mi ha aiutato a crescere, ad affrontare i momenti difficili e a riprendere fiducia in me stessa.
Lo ringrazio per essersi messo al secondo posto quando ne avessi bisogno, per non aver preteso mai nulla e per aver sempre appoggiato le mie scelte.
Grazie per avermi dato la tranquillità di cui avevo bisogno e per avermi reso felice; grazie anche a te oggi ho raggiunto questo traguardo.

Per ultimo, devo ringraziare me stessa, per aver avuto la determinazione e la volontà di provare sempre e la forza di non mollare mai. Ho capito che non bisogna aspettare il momento giusto per iniziare a fare qualcosa, perché il momento giusto non arriverà mai e quando si presenta una occasione bisogna coglierla al volo. Ho raggiunto la consapevolezza che dinanzi ad una sconfitta non bisogna mai rinunciare e che siamo noi a decidere chi vogliamo essere e non gli eventi esterni.
Al di termine di questo percorso, sono soddisfatta della persona che sono e di ogni attimo che ho vissuto nella mia vita.