Anomaly Detection in Electricity Consumption

Arianna Pera[1,2], Laura Rapino[1,3], Francesco Fustini[1,4]

Sommario
Il progetto si propone di studiare alcuni metodi di *anomaly detection* applicati a serie storiche relative ai consumi di elettricità in edifici universitari. L'obiettivo principale è identificare ed eventualmente spiegare la presenza di giornate caratterizzate da consumi anomali, con il fine ultimo di proporre un'integrazione del processo negli stessi sistemi di monitoraggio energetico permettendo un'analisi real time.
Partendo dai dati messi a disposizione dall'Università di Milano-Bicocca, relativi agli edifici U1 e U6, lo studio approfondisce alcune tecniche di *anomaly detection*: una decomposizione stagionale delle serie temporali, lo studio dei residui di modelli TBATS e SARIMA, l'implementazione di un modello Isolation Forest e la clusterizzazione con algoritmo K-means.
I risultati ottenuti sono analizzati singolarmente per edificio, riscontrando la presenza di anomalie soprattutto nei periodi di festività e nei periodi di chiusura o in giornate in cui si può ipotizzare l'avviamento di attività di manutenzione. Per quanto concerne i giorni anomali caratterizzati da consumi molto più elevati di quanto ci si aspetterebbe, si individuano delle date che rientrano nelle sessioni d'esame (Giugno-Luglio per un edificio e Febbraio per l'altro). Alcune giornate, inoltre, risultano etichettate come outliers per entrambi gli edifici, ad esempio nel caso del 6 Agosto 2018, dei giorni prossimi al Natale 2018, del 19 Giugno 2019, del 9 e 10 Luglio 2019 e del 6 Aprile 2020.
L'intero studio è svolto utilizzando il programma *R* (versione 4.0.4).

Tag
Elettricità — Serie Storiche — Outlier — anomaly detection — R

[1] *CdLM Data Science, Università degli Studi di Milano-Bicocca*
[2]matr. 827964, [3]matr. 831346, [4]matr. 830697

Indice

Introduzione

I vent'anni più caldi della storia sono stati registrati durante gli ultimi 22 anni. Il cambiamento climatico è una questione sempre più delicata e richiede azioni efficaci: circa l'80% dell'energia a livello globale proviene da combustibili fossili e circa il 40% dell'energia totale è impiegata nella generazione di elettricità ([1]).

La ricerca di fonti energetiche alternative (le cosiddette "green sources") è sicuramente un'azione positiva per frenare il fenomeno del *climate change* ma da sola non è sufficiente: occorre anche cercare di diminuire il consumo energetico complessivo. Combinando informazioni sui consumi di energia degli utenti e sistemi di *anomaly detection* è possibile assistere i consumatori nella scelta di un comportamento energetico più sostenibile. L'individuazione di picchi energetici anomali consente, inoltre, di agire in modo preventivo rispetto agli stessi in futuro. Le enormi quantità di dati provenienti da sensori intelligenti installati in uffici, edifici residenziali e luoghi pubblici rende possibile un'analisi approfondita e puntuale dei consumi di energia elettrica. L'utilizzo corretto e consapevole di tali dati permette di individuare consumi anomali e anche, eventualmente, di comprendere le cause di tali anomalie. Come conseguenza, risulta possibile sviluppare metodi e sistemi di *anomaly detection* in tempo reale, o addirittura in modalità preventiva, favorendo un miglior processo decisionale e aiutando a ridurre gli sprechi energetici.

La gestione dei consumi energetici è una tematica centrale per l'Università di Milano-Bicocca sia in ambito ambientale che economico. L'ottimizzazione dei consumi, la riduzione degli sprechi e la promozione di un comportamento energetico più consapevole fanno parte di una strategia volta alla redazione di un piano energetico più sostenibile. Negli ultimi anni, l'Università si è sempre più impegnata in attività di monitoraggio dei consumi energetici, sostituendo gli impianti di illuminazione e condizionamento in modo da renderli più efficienti e dotati di sensori intelligenti che ne permettono un'analisi continua.
A partire dai dati riguardanti il consumo di energia elettrica degli edifici U1 e U6 dell'Università di Milano-Bicocca negli dal 2018 al 2020, questa ricerca si propone di realizzare un'analisi e un confronto delle anomalie nelle serie storiche utilizzando alcuni tra i diversi metodi disponibili in questo settore. Il report si sviluppa nel seguente modo:

- Definizione dell'**obiettivo di ricerca** e individuazione del target di riferimento (sezione 1);
- Presentazione delle **metodologie** su cui si basa lo studio e brevi cenni teorici sulla loro definizione (sezione 2);
- **Overview sui dati** messi a disposizione dall'Università e loro caratteristiche, pregi e limiti per l'analisi (sezione 3);
- Fase di **lettura, preprocessing e analisi preliminari** sui dati, con identificazione di dati mancanti e necessaria trasformazione della variabile relativa al consumo di energia (sotto-sezione 4.1);
- Implementazione delle procedure di **anomaly detection**: decomposizione stagionale attraverso la mediana, studio dei residui di modelli TBATS e SARIMA, analisi dei punteggi *anomaly score* di modelli Isolation Forest, identificazione delle osservazioni anomale attraverso un'analisi delle distanze dai centroidi dei gruppi prodotti con clusterizzazione K-means (sotto-sezione 4.2);
- Discussione della **metodologia di comparazione** scelta e dunque della modalità di selezione definitiva delle giornate anomale (sotto-sezione 4.3);
- Presentazione dei **risultati** (sezione 5);
- **Conclusioni** ed eventuali sviluppi futuri della ricerca (sezione 6);

1. Obiettivo

L'obiettivo principale dello studio è quello di identificare la presenza di anomalie nelle serie storiche dei consumi di elettricità dell'Università di Milano-Bicocca. Il target di riferimento per l'analisi potrebbe essere l'Area Infrastrutture e Approvvigionamenti dell'Università, la quale si occupa di gestire utenze e consumi elettrici e di promuovere azioni strutturali e comportamentali atte al risparmio e all'efficienza energetica.
Diverse metodologie sono utilizzate e confrontate ai fini di fornire una panoramica sufficientemente ampia su alcune delle tecniche che possono essere impiegate in questo ambito. L'analisi si propone anche di cercare una spiegazione alle anomalie individuate relazionandole, ove possibile, a specifici eventi o situazioni. In questo senso, la validazione dei risultati da parte di un esperto di dominio e/o di un tecnico dell'Università permetterebbe di attribuire maggiore robustezza alle considerazioni riportate.
Gli approcci studiati sono confrontati in modo da poter osservare il ripetersi o il variare di giorni (o periodi dell'anno) etichettati come anomali rispetto ai consumi di energia elettrica negli edifici universitari di riferimento. Gli algoritmi specifici sono implementati tramite **R** e consistono in metodologie che spaziano dall'ambito descrittivo a quello di cluster analysis. I metodi usati sono la decomposizione stagionale delle serie attraverso

la mediana, i modelli TBATS e SARIMA, l'algoritmo Isolation Forest e la clusterizzazione tramite K-means.

2. Metodologie

Gli algoritmi di *anomaly detection* sono molteplici e spaziano dalla classificazione agli approcci neurali, passando per l'analisi dei residui.
I successivi sottoparagrafi descrivono i modelli e gli strumenti presi in considerazione per l'individuazione di anomalie nel contesto di questa ricerca.

2.1 TBATS

Il modello TBATS (Trigonometric, Box-cox transform, Arma errors, Trend and Seasonal components), introdotto nel 2011 in [2], estende il modello BATS aggiungendo una rappresentazione trigonometrica delle componenti stagionali attraverso la serie di Fourier.
Il modello permette di gestire stagionalità complesse (multi-stagionalità o stagionalità ad alta frequenza) usando una tecnica di *exponential smoothing*. I parametri del modello sono i seguenti:

- ω: parametro Box-Cox;
- p, q: parametri ARMA;
- ϕ: il parametro *damping* per il trend;
- $\{m_i, k_i\}$, con $i = 1, ..., T$: periodi stagionali (m_i) e numero di armoniche (k_i) necessarie per la componente stagionale i-esima.

L'approccio *exponential smoothing* è utilizzato per produrre errori di previsione *one step ahead* e i parametri sono selezionati in modo da massimizzare la funzione di *conditional likelihood* risultante.
Il modello TBATS è utilizzato, nel contesto di questa ricerca, sia per selezionare la stagionalità delle serie temporali sia, attraverso l'analisi dei residui, per identificare le anomalie in presenza di stagionalità multipla (si vedrà che tale è il caso di U1). Il criterio considerato per la selezione della stagionalità si basa sul valore di AIC: l'Akaike Information Criterion misura la qualità relativa di un modello statistico su un determinato set di dati ed è definito come $AIC = -2ln(L) + 2k$, con L funzione di massima verosimiglianza del modello e k numero di parametri stimati. Il criterio basato sull'AIC prevede la selezione del modello che presenta il valore minore di tale misura e il suo utilizzo appare sensato solo se valutato per un insieme di modelli di una medesima classe (ad esempio, TBATS) ma non appare interpretabile se usato per confrontare modelli appartenenti a classi diverse.

2.2 SARIMA

I modelli Seasonal AutoRegressive Integrated Moving Average (SARIMA o Seasonal ARIMA, introdotti in [3]) sono un'estensione dei modelli ARIMA che permette la gestione di componenti stagionali. Rispetto al modello ARIMA classico, un modello SARIMA aggiunge parametri relativi all'autoregressione, al differencing, alla media mobile della componente stagionale e al periodo della stagionalità della serie.
Complessivamente, un modello SARIMA è specificato come

$$SARIMA(p,d,q)(P,D,Q)[m]$$

dove p è l'ordine autoregressivo del trend, d è l'ordine di differencing del trend, q è l'ordine di media mobile del trend, P è l'ordine autoregressivo della componente stagionale, D è l'ordine di differencing della componente stagionale, Q è l'ordine di media mobile della componente stagionale e m indica il numero di time steps per ogni periodo stagionale.
Il modello SARIMA è impiegato, nel contesto di questa ricerca, per identificare la presenza di anomalie attraverso lo studio dei residui nel caso di stagionalità singola.

2.3 Seasonal Decomposition of Time Series by Median

Un approccio considerato per l'individuazione di anomalie nei dati è quello descritto da [4] e consiste nella decomposizione di una serie temporale in componente stagionale, trend e componente residua. Quest'ultima, nello specifico, è calcolata come $R_X = X - S_X - \tilde{X}$, dove X rappresenta la serie temporale di interesse, S_X è la componente stagionale determinata con l'approccio LOESS (LOcally weighted polynomial regrESSion) e $\tilde{X}$ è la mediana della serie temporale, utilizzata per ricavare i residui al posto del trend. In seguito alla determinazione della componente residuale della serie, le anomalie possono essere inviduate sfruttando il test GESD (Generalized Extreme Studentized Deviate test) descritto nel paragrafo 2.4.
Un approccio alternativo si basa sulla decomposizione STL (Seasonal and Trend decomposition using Loess) la quale stima i residui come $R_X = X - T_X - S_X$, apparendo dunque simile all'approccio precedente tranne per il fatto che qui si considera l'eliminazione del trend T_X attraverso un filtro media mobile.
L'approccio di decomposizione stagionale attraverso la

mediana risulta preferibile, rispetto all'approccio STL, nel caso di componente stagionale più forte del trend.

2.4 Generalized Extreme Studentized Deviate

Il test Generalized Extreme Studentized Deviate, descritto in [5], è un metodo statistico spesso utilizzato per individuare la presenza di anomalie in un dataset. Esso richiede un upper bound per il numero di outliers che ci si aspetta di trovare nei dati (k) e conduce k test separati: un test per un outlier, un test per due outliers e così via. Il test si basa sulle seguenti ipotesi:

$$\begin{cases} H_0: & \text{non ci sono outliers nei dati} \\ H_1: & \text{ci sono fino a } k \text{ outliers nel dataset} \end{cases}$$

La statistica test di riferimento è la seguente:

$$C_i = \frac{max_i|x_i - \bar{x}|}{s} \tag{1}$$

con $\bar{x}$ valore medio dei dati e s deviazione standard.
Il test procede alla rimozione dell'osservazione che soddisfa $max_i|x_i - \bar{x}|$ e ricalcola il valore della statistica a partire da $n-1$ osservazioni fino a che non sono rimosse k osservazioni in totale. Si hanno quindi k valori delle statistiche test che devono essere confrontati con altrettanti valori critici, calcolati come segue:

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t^2_{p,n-i-1})(n-i+1)}} \tag{2}$$

con $i = 1,2,...,k$, n numero di osservazioni complessive, $t_{p,v}$ quantile $100p$ della distribuzione t di Student con v gradi di libertà e $p = 1 - \frac{\alpha}{n-i+1}$, con α livello di significatività.
Il numero complessivo di anomalie è definito considerando il valore maggiore di i tale da garantire $C_i > \lambda_i$.

2.5 Isolation Forest

Basandosi sull'idea che le anomalie in un dataset siano in quantità limitata e caratterizzate da valori degli attributi molto diversi da quelli assunti dalle osservazioni "normali", è possibile applicare l'approccio di *anomaly detection* proposto in [6].
L'algoritmo consiste nella generazione di una serie di Isolation Trees (binary trees) e identifica le anomalie come osservazioni che hanno una *average path length* breve. Infatti, le anomalie tendono a essere isolate più in prossimità della radice dell'albero mentre le osservazioni normali tendono a essere isolate più in profondità, richiedendo generalmente un numero maggiore di partizioni.
La procedura si basa su una fase di training, in cui sono costruiti diversi isolation trees usando dei subsamples del training set, e su una fase di evaluation, in cui si ottengono i cosiddetti *anomaly scores* per le istanze di test. Tale score è calcolato come:

$$s = 2^{-\frac{\bar{d}}{c(n)}} \tag{3}$$

dove $\bar{d}$ è il valore atteso di *path length* della collezione di alberi e $c(n)$ è la *average path length* nel caso di osservazioni uniformemente casuali. Lo score risulta essere normalizzato e permette di realizzare le seguenti considerazioni:

- se s è vicino a 1, l'osservazione in questione è sicuramente un'anomalia;
- se s è decisamente inferiore a 0.5, l'osservazione in questione può essere definita come "normale";
- se s è vicino a 0.5 per tutte le istanze, i dati non presentano anomalie distinte.

2.6 K-means Cluster Analysis

L'algoritmo K-means è un metodo di *cluster analysis* partizionale che cerca di determinare k gruppi di osservazioni minimizzando la varianza interna agli stessi. La procedura è iterativa: a ogni ciclo, gli elementi sono assegnati al gruppo il cui centroide risulta il più vicino e, di volta in volta, viene aggiornato il valore del centroide in base alle nuove disposizioni in clusters.
Nel caso dell'*anomaly detection* si può ragionare come suggerito nell'articolo [7]: se la distanza tra un'osservazione e il centroide più vicino alla stessa è maggiore di una certa *threshold*, si considera l'osservazione come un'anomalia. E' possibile decidere di ordinare in senso decrescente le distanze e procedere etichettando una determinata percentuale dei punti caratterizzati dalle distanze maggiori come outliers.
L'algoritmo K-means è abbastanza semplice ed efficiente ma richiede la determinazione del numero k di clusters da considerare. Per identificare il valore ottimale del parametro è possibile considerare due metodi:

- **Elbow method**: la procedura consiste nell'analizzare graficamente la varianza spiegata in funzione del numero di gruppi. Il numero ottimale k è determinato dal valore corrispondente al gomito della curva della varianza;

- **Silhouette method**: si basa sulla realizzazione di un grafico dove si riporta il valore di Silhouette rispetto al numero di gruppi valutati. Il valore di Silhouette misura quanto un elemento è simile al proprio cluster di appartenenza rispetto che agli altri clusters, assumendo valori compresi tra -1 (oggetto assegnato incorrettamente al cluster) e +1 (oggetto correttamente assegnato). Nello specifico, il valore di Silhouette per un oggetto i è definito come

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}, |C_i| > 1$$

dove $|C_i|$ è il numero di elementi del cluster C_i e definendo

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

con $d(i, j)$ distanza tra gli elementi i e j del cluster C_i e

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

con $d(i, j)$ distanza tra gli elementi i e j appartenenti, rispettivamente, al cluster C_i e C_k.
Si seleziona come numero ottimale di clusters k quello corrispondente a valori di Silhouette maggiori (prossimi a 1).

3. Dati

I dati utilizzati sono forniti dall'*Università degli Studi Milano - Bicocca* e riguardano la potenza rilevata (kW) negli edifici universitari U1 e U6 durante gli anni 2018, 2019 e 2020. Il livello di dettaglio delle misure è di 15 minuti.
I dati fanno riferimento ai consumi di energia elettrica nei due edifici, in termini di luce e impianto di raffrescamento (solo nel caso di U1, in quanto U6 utilizza energia termica per il raffrescamento).
I dati originali si presentano come una collezione di file (in formato .xls, .xlsx, e .csv) contenenti le informazioni riguardo i consumi mensili di ciascuno dei due edifici. In totale, si hanno 84 file mensili, di cui alcuni ripetuti (si nota spesso, infatti, che lo stesso mese è presente sia come .csv che come .xlsx). La gestione dei file, tuttavia, non appare difficoltosa dal punto di vista computazionale in quanto la dimensione risulta contenuta. In seguito a un lavoro di integrazione e pulizia abbastanza dispendioso si ottengono due dataset contenenti 1052016 osservazioni e 8 attributi ciascuno, tra i quali vi sono:

1. POD: identificativo del contatore di energia;
2. DATA: giorno di rilevazione;
3. ORA: orario della rilevazione, espressa in formato *hhmmss*;
4. FL_ORA_LEGALE: flag relativo alla presenza di ora legale ("2") o solare ("1");
5. CONSUMO_ATTIVA_PRELEVATA: valore dei kW (potenza) nei 15 minuti considerati;
6. CONSUMO_REATTIVA_INDUTTIVA_PRELEVATA;
7. POTENZA_MASSIMA;
8. TIPO_DATO: tale attributo assume lo stesso valore, "E", per tutti i record e fa probabilmente riferimento alla presenza di un dato di tipo energetico.

Le variabili effettivamente valutate per l'analisi sono DATA, ORA e CONSUMO_ATTIVA_PRELEVATA.
Il maggior ostacolo affrontato nella fase di integrazione e pulizia dei dataset è causato dalla disomogeneità dei formati dei file di partenza: in alcuni casi gli attributi nella prima riga (*header*) sono separati da "," mentre nel resto del file il separatore è ";". Alcuni mesi, inoltre, sono caratterizzati da un duplice file che li identifica suddividendo i giorni di riferimento. Ciò accade, ad esempio, per Agosto 2018 e Aprile 2019 per U6. Un altro problema individuato riguarda il fatto che i consumi siano espressi in termini di potenza e non di energia effettivamente consumata: tale fatto rende necessaria una trasformazione prima di poter procedere all'aggregazione per ora, giorno o mese.
I dati risultano adatti rispetto all'obiettivo dell'analisi, permettendo uno studio a carattere generale delle anomalie nei consumi, sebbene appaiano in parte limitanti dal punto di vista dei metodi che possono essere applicati. Non essendo infatti presente alcun tipo di etichetta relativa alla non normalità di una determinata osservazione, risulta impossibile applicare metodi di apprendimento supervisionato che potrebbero fornire risultati quantitativi interessanti. Tale limitazione potrebbe comunque essere superata attraverso la richiesta di un supporto di un esperto di dominio.

4. Analisi

4.1 Lettura e preprocessing

La *data ingestion*, realizzata su R come il resto delle analisi, risulta complessa a causa della disomogeneità dei dati di partenza. Attraverso una serie di controlli manuali sui singoli file, è stato possibile costruire una pipeline di lettura che tiene conto degli specifici formati di definizione di ciascun documento.
In seguito a una serie di analisi preliminari sui dati, si procede alla fase di preprocessing con la trasformazione della variabile di consumo di energia e l'aggregazione dei dati.

4.1.1 Analisi preliminari

Considerando gli ID dei contatori di energia elettrica si osservano alcuni valori mancanti per U1 nel 2019: si tratta di record che presentano unicamente l'informazione relativa al consumo energetico e assumono valori molto elevati risultando, presumibilmente, un'aggregazione dei valori per giorno o mese. Si decide di eliminare tali record in quanto non utili ai fini dell'analisi. Si riscontrano anche altri casi di valori mancanti per U1 nel 2020 e per U6 nel 2019 e 2020 in cui, però, risulta non presente unicamente il dato relativo alla potenza massima che non appare fondamentale per l'analisi proposta. Si controlla che non vi siano istanze duplicate, eliminandole nel caso di identificazione.
Procedendo con l'esplorazione dei dati, si nota per entrambe le serie storiche un'incongruenza in corrispondenza del passaggio da ora legale a ora solare nel giorno 25/10/2020: alcuni orari di riferimento dei consumi appaiono duplicati e riportano in un caso il consumo effettivo, nell'altro un consumo nullo. Si sceglie di eliminare i record corrispondenti a quest'ultima situazione.
Sempre durante una fase di analisi preliminare si nota che ci sono diversi giorni caratterizzati da orari con valore di potenza pari a zero: nel caso si tratti di qualche ora si ipotizza la presenza di blackout prolungati. Un giorno in particolare, però, presenta una conformazione anomala: si tratta del 31/07/2020, caratterizzato da un consumo nullo di energia durante l'intera giornata in entrambi gli edifici. Non avendo a disposizione un esperto di dominio a cui chiedere una spiegazione rispetto a tale fenomeno e non volendo procedere all'inputazione dei dati attraverso una previsione che potrebbe influenzare eccessivamente le analisi, si sceglie di non trattare in alcun modo la giornata in questione ritenendo che sarà comunque successivamente etichettata come anomalia dai diversi metodi considerati.

L'ultimo problema da affrontare consiste nel fatto che il mese di Giugno 2020 in U6 risulta essere una copia del medesimo mese in U1, probabilmente a causa di un errore di inputazione. Tale situazione è evidente osservando la Figura 1 in cui l'asse delle y riporta i consumi in kWh (energia consumata), il cui calcolo sarà spiegato in seguito. Valutando che sia mantenere il mese di Giugno 2020 come copia di U1 per l'edificio U6 sia procedere con una sua previsione potrebbero influenzare negativamente gli algoritmi proposti, si sceglie di eliminare le osservazioni successive al 31/05/2021 per l'edificio U6 e dunque analizzare una serie temporale ridotta rispetto a quella di U1.
Uno sguardo approfondito al grafico in Figura 1 permette anche di operare un'analisi descrittiva dei due edifici. Ad esempio, si nota come la serie di U1 presenti dei picchi evidenti in corrispondenza dei mesi estivi, probabilmente a causa dell'impatto dell'impianto di raffrescamento elettrico. Si osserva anche una differenza in termini di entità del consumo dei due edifici: U6, decisamente più grande in termini di metratura e numero di persone ospitate, presenta dei consumi di elettricità tendenzialmente superiori a quelli di U1.

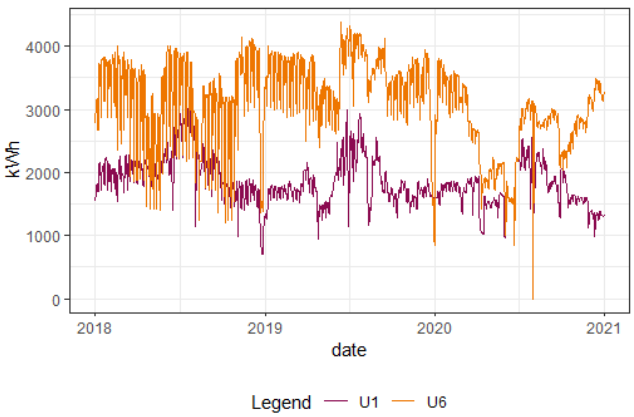

Figura 1. Energia consumata U1 vs. U6

4.1.2 Preprocessing

La potenza, espressa in kW, è una misura istantanea e come tale non può essere aggregata. Per ottenere l'energia consumata corrispondente a un certo valore di potenza occorre moltiplicare il valore di kW per 0.25: i dati, infatti, propongono il valore di potenza elettrica ogni 15 minuti (0.25 ore) e dunque per passare da potenza a energia consumata è necessaria una trasformazione che consideri tale fattore moltiplicativo.
Nell'analisi si ritiene interessante valutare le anomalie giornaliere di consumo di energia: i dati proposti con

granularità più fine sono dunque aggregati in modo da dare origine a un'unica misura dell'energia complessiva consumata in un giorno.

4.2 Anomaly detection

Il primo step dell'analisi consiste nello studio della stagionalità delle serie temporali considerate: l'approccio seguito valuta il fitting di un modello TBATS con stagionalità settimanale, mensile e annuale e ogni altra possibile combinazione di tali livelli. La stagionalità scelta per la serie storica considerata è quella che riporta i valori inferiori di AIC tra tutti quelli corrispondenti ai modelli confrontati. A questo punto, si propone per i dati un modello TBATS nel caso in cui si identifichi una stagionalità multipla o un modello SARIMA (selezionato con il metodo *auto.arima* di *R*) nel caso di stagionalità singola.
I metodi di *anomaly detection* considerati sono quattro: decomposizione della stagionalità attraverso la mediana, approccio TBATS/SARIMA, algoritmo Isolation Forest e cluster analysis con il metodo K-means.

4.2.1 Decomposizione Stagionale

La decomposizione stagionale della serie temporale è svolta sfruttando la libreria *anomalize*, la quale fornisce l'accesso alla funzione *time_decompose* attraverso due metodi:

- *STL* (Seasonal Trend Decomposition con LOESS): approccio che funziona molto bene nel caso di trend a lungo termine ma performa in modo meno efficiente se la componente stagionale è più forte del trend;
- *Twitter*: metodo in cui i residui sono ottenuti rimuovendo la mediana dei dati al posto del trend e che risulta più efficiente quando la componente stagionale è preponderante.

Considerando la natura dei dati in analisi e la stagionalità identificata in precedenza per le serie temporali, si sceglie di applicare il metodo *Twitter*.
La funzione *time_decompose* si occupa della vera e propria fase di decomposizione della serie temporale. Tra i parametri di input, oltre al metodo prescelto per la decomposizione, è necessario indicare la frequenza temporale di riferimento, definita come il numero di osservazioni racchiuse in un ciclo all'interno dei dati. Nel caso di stagionalità singola appare immediata la definizione della frequenza. Nel caso di stagionalità multipla, invece, occorre scegliere il livello con maggior importanza per definire il ciclo di frequenza; nel caso in esame per U1, considerando che il modello TBATS costruito con singole stagionalità settimanali e annuali restituisce valori minori di AIC nel caso della stagionalità settimanale, si sceglie la frequenza *1 week*. Il parametro relativo al trend, che nel metodo considerato controlla l'ampiezza del periodo di riferimento per la mediana, è stabilito automaticamente dalla funzione attraverso l'analisi dei dati.
In seguito alla decomposizione e rimozione del trend attraverso la mediana, l'approccio seguito prosegue con la funzione *anomalize* e la rispettiva identificazione di outliers dalla distribuzione priva di trend e stagionalità, sfruttando il test GESD (descritto al paragrafo 2.4). Un altro metodo che può essere impiegato per l'identificazione delle anomalie è l'IQR, il quale sfrutta il range interquartile per stabilire una distribuzione di riferimento attorno alla mediana e selezionare così gli outliers, più efficiente computazionalmente ma meno accurato del test GESD. La funzione *anomalize* permette anche di definire *alpha*, parametro caratterizzante l'ampiezza delle bande di normalità per le osservazioni. Nel caso specifico di analisi, in seguito a esperimenti empirici, si sceglie di fissare $\alpha = \mathbf{0.05}$ sia per la serie relativa a U1 che per quella relativa a U6.
Infine, la funzione *time_recompose* permette di ricomporre le bande di normalità separando i valori "normali" da quelli anomali.

4.2.2 TBATS/SARIMA

Utilizzando un modello TBATS con stagionalità multipla settimanale e annuale si ricavano i valori fitted di U1 per confrontarli con le osservazioni effettive. In questo modo, sfruttando i residui del modello, è possibile valutare quali siano le osservazioni caratterizzate da residui particolarmente ampi e identificarle dunque come anomalie. Per quanto riguarda U6, valutando la presenza di una stagionalità settimanale, si sceglie di introdurre un modello SARIMA selezionandolo attraverso la funzione *auto.arima*, ricavando i valori fitted e procedendo attraverso i residui come nel caso di TBATS per U1.
I residui di entrambi i modelli sono analizzati in modo da verificare l'assenza di autocorrelazione tra gli stessi. Una volta ottenuti i residui, si valutano due approcci di lavoro:

1. Selezionare come outliers le osservazioni i cui residui si trovano nel top $x\%$ della distribuzione, $x \in \{2, 5, 10\}$;

2. Utilizzare il test GESD per identificare le anomalie nei dati.

Il secondo metodo viene implementato con il supporto della funzione *gesdTest* della libreria *PMCMRplus*; l'unico parametro richiesto in input, oltre al vettore dei residui del modello, è il massimo numero di outliers da testare nei dati. Per determinare tale valore si sceglie di fare riferimento al numero di anomalie identificate nel top $x\%$ della distribuzione dei residui dal primo approccio valutato, in modo da disporre di risultati facilmente comparabili.

4.2.3 Isolation Forest

Sfruttando la libreria *isotree* è possibile costruire un modello di Isolation Forest sui dati relativi alle serie temporali a disposizione. Nello specifico, la funzione *isolation.forest* consente di apprendere il modello, il quale viene successivamente riapplicato ai dati con il metodo *predict* in modo da valutare lo score di outlierness attribuito a ciascuna osservazione. La soglia di score che permette di identificare un'anomalia è legata alla definizione stessa del punteggio, compreso tra 0 e 1: valori attorno a 0.5 indicano anomalie "nella media", valori inferiori a 0.5 indicano osservazioni normali e valori prossimi a 1 identificano veri e propri outliers. Si sceglie di fissare la soglia di riferimento a **0.6**, in modo da cogliere sfumature medio-alte di anomalie nei dati.

4.2.4 K-means

L'ultimo approccio di *anomaly detection* valutato riguarda un algoritmo molto noto di cluster analysis: l'algoritmo K-means.
La funzione *kmeans* della libreria *stats* permette di applicare il metodo in modo abbastanza semplice, richiedendo in input il numero di centroidi (e dunque di clusters) da valutare nella procedura. La determinazione di tale parametro è un argomento molto caldo nella cluster analysis e influenza in modo anche molto evidente i risultati finali delle analisi. Per avere una panoramica più completa e valutare in modo più preciso, si sceglie di considerare sia il metodo Elbow che l'approccio relativo al punteggio di Silhouette. In base ai risultati ottenuti dallo studio di un numero di clusters compreso tra 2 e 10, si identifica come numero ottimale **5** nel caso di U1 e **3** nel caso di U6. Volendo garantire la riproducibilità dei risultati, si decide di fissare il seed **123** attraverso la funzione *set.seed*.
L'approccio di cluster analysis valuta come anomalie quelle osservazioni caratterizzate da una distanza molto elevata dal centroide loro più vicino. Per definire una soglia di riferimento relativa a tale distanza, si sceglie di valutare il top $x\%$, con $x \in \{2,5,10\}$, della distribuzione delle distanze dei punti dal centroide del gruppo di appartenenza.

4.3 Comparazione dei metodi

Essendo il task di *anomaly detection* condotto sui dati in esame di tipo *unsupervised*, non è possibile procedere attraverso la validazione dei risultati comparando gli outliers restituiti dai metodi con quelli effettivamente osservati. Una possibilità che mira a trasformare il task non supervisionato in supervisionato è quella di etichettare manualmente alcune delle osservazioni come anomale e di procedere con fasi di training e testing come nei metodi classici di Machine Learning. Non essendo tuttavia semplice etichettare manualmente i dati senza il supporto e il controllo di un esperto di dominio, si opta per un metodo alternativo. Considerando che, complessivamente, sono disponibili i risultati relativi a cinque differenti approcci (un risultato per ciascuno dei metodi considerati con un doppio valore nel caso dei due approcci di valutazione dei residui dei modelli TBATS/SARIMA), si decide di considerare la concordanza tra i metodi come criterio di definizione della "forza" delle anomalie proposte.
Data la presenza di metodi che considerano valori basati sulla percentuale top $x\%$, con $x \in \{2,5,10\}$, della distribuzione dei residui (TBATS/SARIMA) o delle distanze (K-means), si sceglie di realizzare una comparazione dei risultati a blocchi. Si creano dunque tre blocchi di risultati (blocco 2%. blocco 5% e blocco 10%), ciascuno contenente quelli ottenuti con una delle percentuali valutate avendo fissato gli outcome degli approcci che non sono influenzati da tali percentuali e appaiono dunque presenti in modo identico in tutti i blocchi. Per ciascun blocco di cinque risultati, si considera un'osservazione come anomalia sufficientemente forte se almeno due metodi su cinque la identificano come outlier. A seconda della percentuale di metodi che valutano un'anomalia come tale è possibile quindi definire la "forza" di ogni risultato.

5. Risultati

Lo studio iniziale della stagionalità delle serie temporali verte all'identificazione dei periodi di stagionalità che minimizzano i valori di AIC del modello TBATS. Si

riportano i risultati ottenuti dai modelli, in termini di AIC, nella Tabella 1.

Tabella 1. Valori AIC modelli TBATS

Seasonality	AIC U1	AIC U6
W	19090.62	15797.05
M	19242.76	16440.06
Y	19210.62	16446.53
WM	19106.92	15809.85
WY	19066.69	15808.14
MY	19224.81	16415.36
WMY	19093.54	15822.21

In base alle performance ottenute, si seleziona una stagionalità multipla di tipo settimanale e annuale per la serie temporale di U1 e una stagionalità settimanale per la serie di U6.
La tesi è ulteriormente confermata dall'osservazione della decomposizione delle serie definite con stagionalità multipla annuale, mensile e settimanale: è evidente come la componente mensile assuma un ruolo decisamente meno rilevante rispetto alle altre due. La Figura 2 mostra tale decomposizione per l'edificio U1, mentre la Figura 3 riporta la medesima visualizzazione relativa a U6.

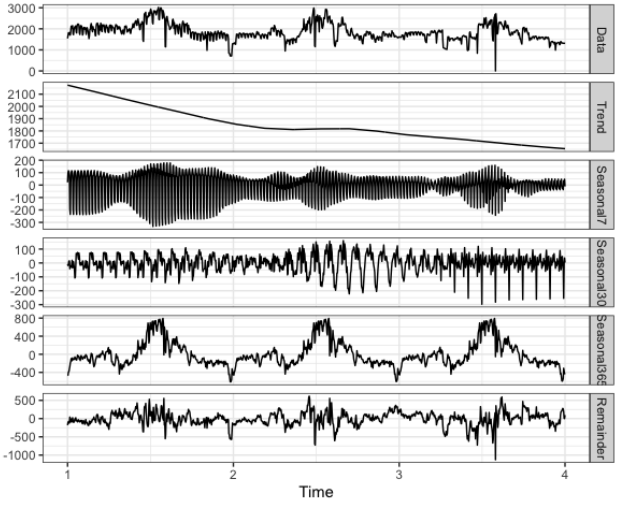

Figura 2. Decomposizione U1

In base al tipo di stagionalità rilevata è possibile costruire un modello TBATS con stagionalità settimanale e annuale per U1 e un modello SARIMA (selezionato con la funzione *auto.arima*) con stagionalità settimanale per U6. Tali modelli saranno successivamente utilizzati per la realizzazione di uno degli approcci di *anomaly detection* analizzati. Il test Ljung-Box per i residui di TBATS per U1 permette di accettare l'ipotesi nulla di

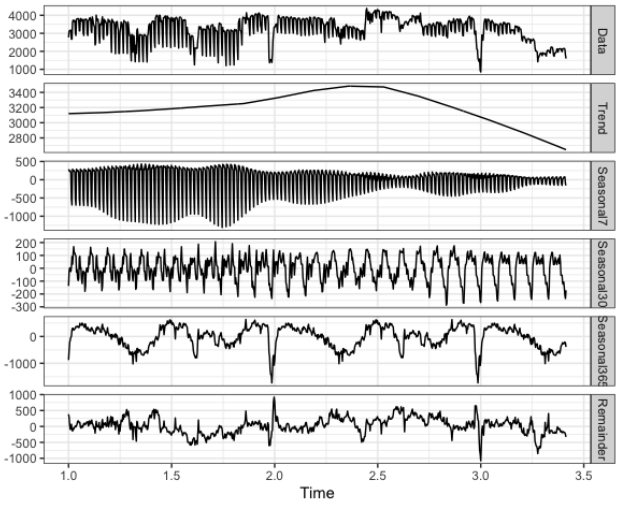

Figura 3. Decomposizione U6

non correlazione dei residui con un p-value pari a 0.11 e anche i grafici ACF e PACF, riportati in Appendice, confermano tale risultato. Similmente, il test svolto sui residui di SARIMA per U6 permette l'accettazione dell'ipotesi nulla con p-value pari a 0.27 e anche i grafici ACF e PACF in Appendice risultano in accordo con tale conclusione.
Un primo metodo per l'identificazione di anomalie è quello di decomposizione stagionale che permette di individuare, complessivamente, 29 anomalie per U1. Si ricorda che, in questo caso, si seleziona un parametro di $\alpha = 0.05$ per l'ampiezza delle bande di normalità. Il grafico in Figura 4 evidenzia in rosso i giorni anomali selezionati: si osservano valori lontani dalla normalità nell'estate 2018, a fine 2018, a metà 2019 e nell'estate 2020. Per quanto concerne U6, definendo sempre un parametro $\alpha = 0.05$, si individuano le 26 anomalie riportate in rosso nella visualizzazione in Figura 5: vi sono outliers nella seconda metà del 2018, a fine 2019 e nell'inverno 2020.

Il secondo approccio valutato per l'*anomaly detection* sfrutta un modello TBATS con stagionalità multipla settimanale e annuale per U1 (modello $TBATS(1,\{3,1\},-,\{<7,3>,<365,5>\})$), individuando *fitted values* che sono successivamente confrontati con i valori effettivi per selezionare le osservazioni anomale. Il grafico in Figura 6 permette un confronto tra il modello stimato e la serie temporale osservata di U1.

Per quanto concerne U6, data la stagionalità settimanale, si sfrutta il metodo *auto.arima* che seleziona un modello SARIMA $(1,0,2)(0,1,1)[7]$, i cui fitted values

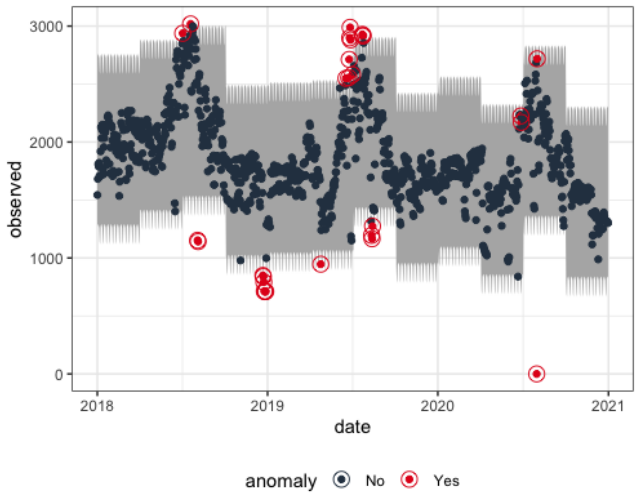

Figura 4. Anomalie con Decomposizione Stagionale U1

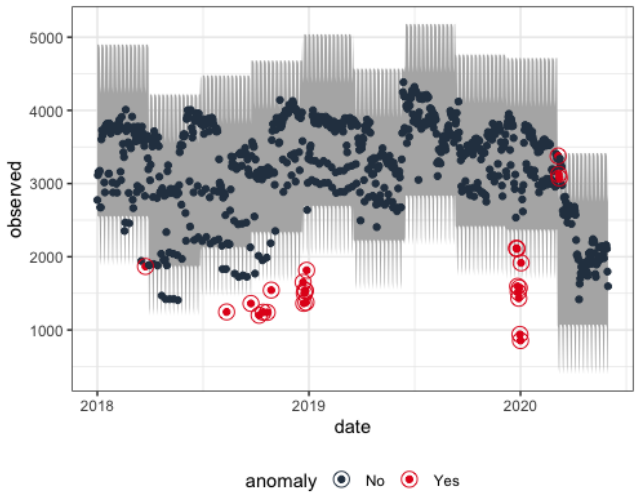

Figura 5. Anomalie con Decomposizione Stagionale U6

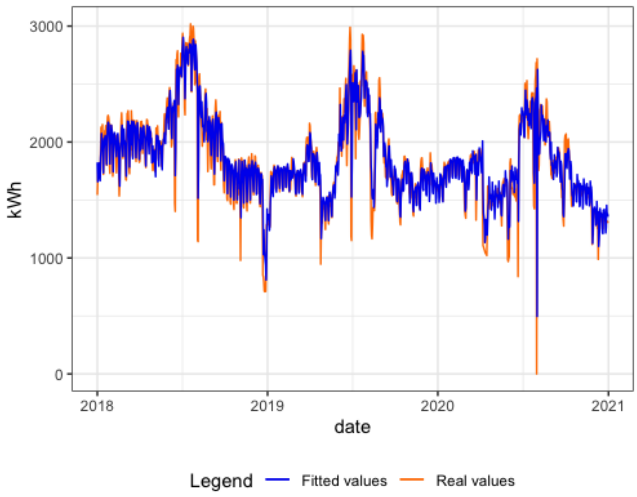

Figura 6. TBATS U1

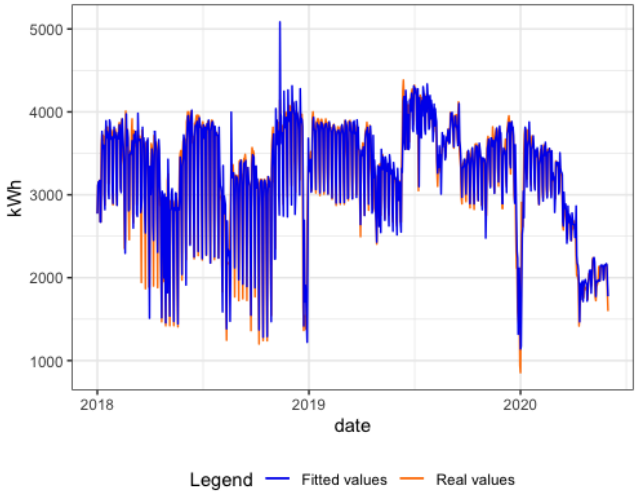

Figura 7. SARIMA U6

sono confrontati con i valori osservati della serie temporale. Una rappresentazione grafica del modello stimato a confronto con i dati osservati è riportata in Figura 7.

Un primo metodo valutato per l'individuazione delle anomalie data l'applicazione di modelli TBATS/SARIMA consiste nel selezionare la top $x\%$ della distribuzione dei residui dei modelli proposti, con $x \in \{2, 5, 10\}$. Si riportano, nella Tabella 2, i quantili dei residui corrispondenti alle diverse percentuali considerate per U1 e in Tabella 3 i rispettivi valori per U6.

Un secondo approccio di selezione basato sui residui dei modelli TBATS e SARIMA sfrutta il test GESD e utilizza come upper bound per il numero di outliers da considerare nei dati le quantità di osservazioni anomale individuate con il metodo precedente basato sui quantili.

Tabella 2. Quantili dei residui anomalie U1

x (top x%)	% oss. norm.	Q. res. (kWh)	n. outliers
2	98%	417.79	22
5	95%	291.05	55
10	90%	196.30	110

Tabella 3. Quantili dei residui anomalie U6

x (top x%)	% oss. norm.	Q. res. (kWh)	n. outliers
2	98%	716.46	18
5	95%	500.58	45
10	90%	292.57	89

Il terzo approccio di identificazione di anomalie si basa sull'algoritmo di *Isolation Forest*: etichettando come

outliers quelle osservazioni cui è associato un valore di *anomaly score* maggiore o uguale a 0.6 è possibile ottenere la situazione riassunta nella Figura 8 per U1 e nella Figura 9 per U6.

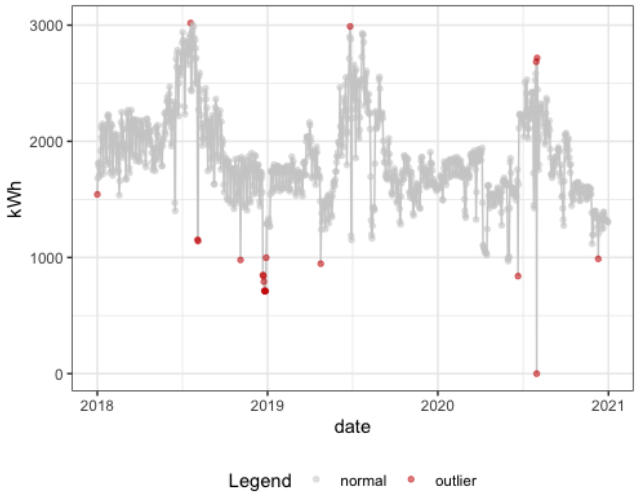

Figura 8. Isolation Forest U1

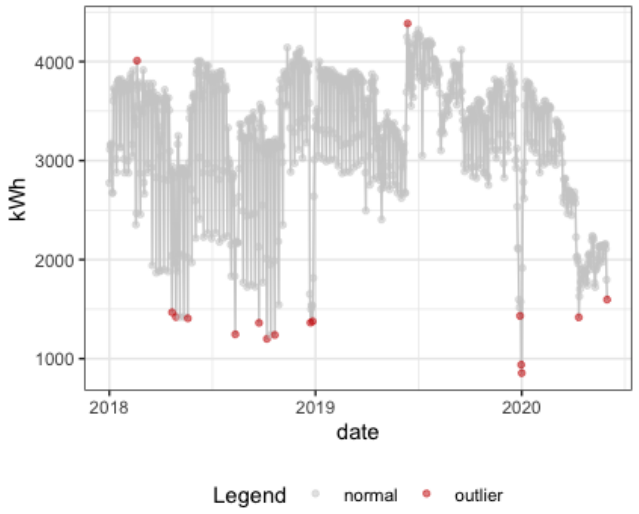

Figura 9. Isolation Forest U6

L'ultimo metodo considerato nell'analisi segue un approccio di clusterizzazione e considera l'algoritmo K-means per l'individuazione di anomalie. Un primo passaggio, fondamentale per gli sviluppi successivi, consiste nell'identificazione del numero ottimale k di clusters da utilizzare per raggruppare le osservazioni. Si sceglie di utilizzare un approccio che considera il criterio Elbow in unione al metodo Silhouette. I grafici riportati nelle Figure 10 e 11, rispettivamente per U1 e U6, permettono di selezionare attraverso un bilanciamento dei due criteri un valore di $k = 5$ per U1 e $k = 3$ per U6. In quest'ultimo caso, essendo la selezione con il metodo Elbow più complessa rispetto a quanto osservato per U1, si sceglie di far prevalere il criterio di Silhouette.

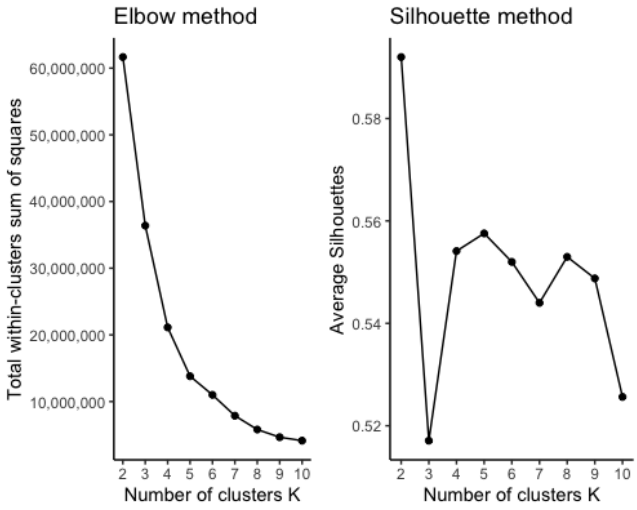

Figura 10. K Selection U1

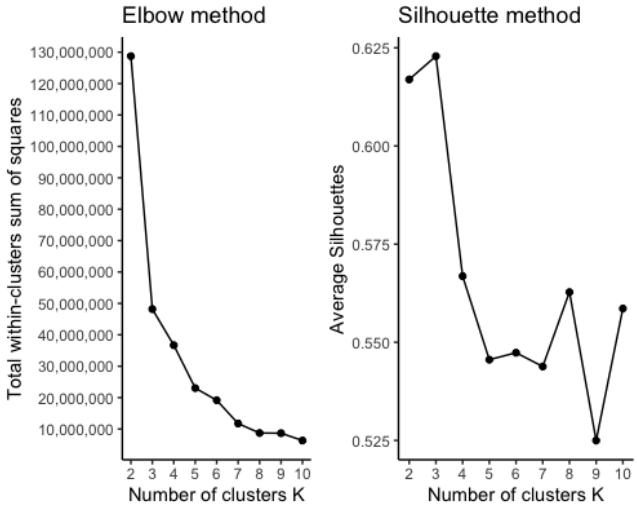

Figura 11. K Selection U6

Il metodo proposto prevede di valutare come anomalie quelle osservazioni che sono particolarmente distanti dal centroide del cluster di appartenenza. Sfruttando l'approccio basato sulla selezione delle anomalie nel top $x\%$ della distribuzione delle distanze dai centroidi, con $x \in \{2, 5, 10\}$, si ottengono i risultati riassunti in Tabella 4.

Scegliendo di definire come anomalie quelle osservazioni che rientrano nel top 5% della distribuzione delle

Tabella 4. Anomalie con K-means

x (top x%)	n. outliers U1	n. outliers U6
2	22	18
5	55	44
10	110	88

distanze dai centroidi, si osservano i risultati riportati nei grafici in Figura 12 per U1 e in Figura 13 per U6.

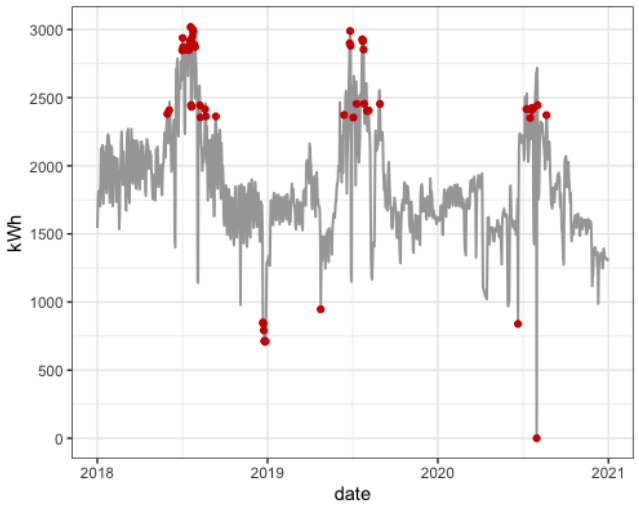

Figura 12. K-means U1

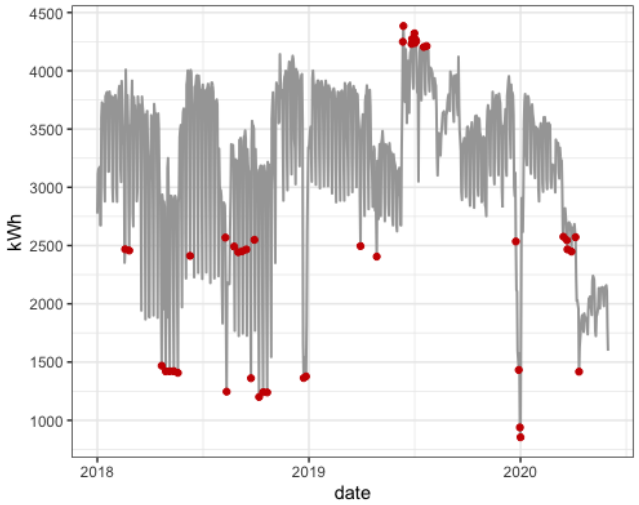

Figura 13. K-means U6

Considerando come riferimento definitivo il top 5% (top 5% della distribuzione dei residui nel caso di TBATS/SARIMA e top 5% della distribuzione delle distanze dai centroidi nel caso di K-means, tenendo fissi i risultati dei metodi non influenzati dalle percentuali) per la valutazione finale, si osserva che per U1 tutti i metodi considerati identificano come anomalia il giorno 31/07/2020. Tale risultato era atteso, in quanto la giornata in questione è caratterizzata da un consumo energetico nullo durante le 24 ore e rappresenta una delle problematiche rilevate durante la fase di analisi preliminare. Anche le giornate 22/12/2018, 24/12/2018, 25/12/2018, 26/12/2018, 25/04/2019 sono identificate come outliers da tutti i metodi, mentre quattro approcci su cinque identificano come anomalie i giorni 4/08/2018, 5/08/2018, 23/07/2019, 21/06/2020 e 1/08/2020. Nei casi sopra citati, le anomalie riguardano valori di consumo più bassi di quelli che ci si aspetterebbe tranne che per i giorni 23/07/2019 e 1/08/2020. L'identificazione del 23/07/2019 come anomalia caratterizzata da consumi molto elevati potrebbe essere in parte spiegata dalla sessione estiva di esami e dalla conseguente necessità di utilizzo del raffrescamento elettrico. Considerando il 1/08/2020 come giorno di chiusura dell'Università (si tratta del primo Sabato del mese tipico delle ferie estive), la presenza di consumi molto elevati risulta anomala e dovrebbe essere punto di partenza per indagini più approfondite. Per quanto concerne i casi di anomalia in cui le osservazioni sono caratterizzate da consumi più bassi di quanto atteso, si tratta perlopiù di giorni di vacanza o festività: i due giorni di Agosto 2018 corrispondono al primo weekend del mese, fine delle attività didattiche e di esami e inizio delle ferie per i docenti, a Dicembre 2018 si rilevano anomalie nei giorni a ridosso del Natale e il 25/04 corrisponde all'Anniversario della Liberazione. Relativamente al 21/06/2020, una Domenica, non si rilevano festività particolari ma si ipotizza un'attività di manutenzione al di fuori degli orari di apertura che dovrebbe essere confermata da un esperto di dominio. Il grafico in Figura 14 rappresenta una sintesi di quanto appena esposto, riportando le osservazioni anomale identificate da almeno tre metodi su cinque per la serie temporale di U1.

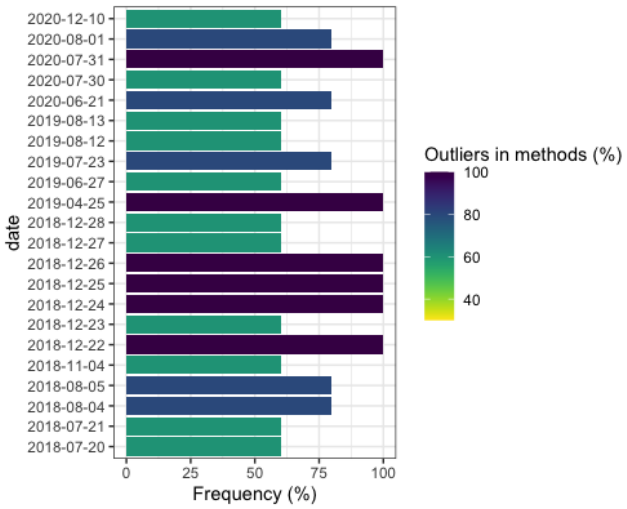

Figura 14. Anomalies U1

In Figura 15 è possibile osservare le anomalie identificate da almeno tre approcci per U6: tutti i metodi valutati selezionano il 31/12/2019 come anomalo. Si tratta della giornata relativa al Capodanno 2020 e l'anomalia può apparire dunque in parte spiegata dalla festività.

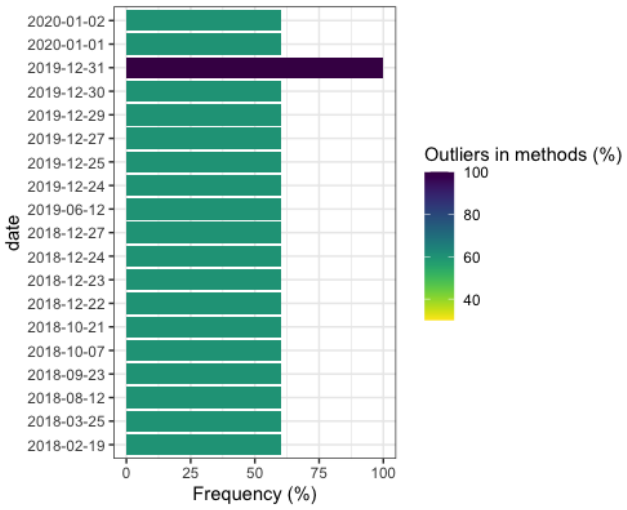

Figura 15. Anomalies U6

Volendo indagare la corrispondenza di anomalie tra U1 e U6 per comprendere se vi siano condizioni generali che vadano oltre l'atomicità dei singoli edifici, si valuta quali siano le date identificate come anomale da almeno due metodi su cinque per entrambe le serie storiche:

- 6/08/2018: corrisponde al primo Lunedì del mese, potrebbero esserci state attività di manutenzione sugli impianti;
- 22/12/2018-23/12/2018-24/12/2018-27/12/2018: le anomalie sono in parte dovute alle festività natalizie. Il fatto che tale situazione di non normalità non sia però ripetuta in anni successivi richiede una spiegazione e analisi ulteriore;
- 19/06/2019: si tratta di un Mercoledì, forse sfruttato dai tecnici per effettuare attività di manutenzione;
- 9/07/2019-10/07/2019: giornate di Martedì e Mercoledì, forse sfruttate dai tecnici per effettuare attività di manutenzione;
- 6/04/2020: primo Lunedì del mese, possibili attività di manutenzione o inizio pausa didattica per alcuni corsi di laurea.

Considerando come target l'Area Infrastrutture e Approvvigionamenti di Milano-Bicocca, appare interessante non solo comprendere quali siano le anomalie a livello generale ma anche identificare dei picchi di consumo di energia particolarmente elevati. Infatti, sebbene bassi consumi possano essere facilmente associati a periodi di festività, consumi più alti del normale potrebbero essere indice di malfunzionamento o programmazione non ottimale degli impianti. Ad esempio, osservando i risultati relativi alla serie storica di U1, ci si rende conto che più di un metodo analizzato segnala la presenza di consumi energetici molto più elevati del normale, oltre che nelle giornate già evidenziate del 23/07/2019 e 1/08/2020, il 20/07/2018, il 27/06/2019 e a inizio Agosto 2020: sarebbe interessante riuscire a comprendere la causa di tale anomalia, forse in parte spiegabile a partire dalla presenza della sessione estiva e di una necessità di raffrescamento prolungato a causa degli esami svolti in aula. Nel caso di U6, è il 19/02/2018 ad essere identificato da più di un metodo come anomalia caratterizzata da altissimi valori di consumo di elettricità. Anche in questo caso, la sessione di esami potrebbe in parte spiegare l'accaduto ma occorrerebbe la conoscenza di un tecnico per ottenere una spiegazione più accurata.

6. Conclusione e possibili sviluppi

L'analisi svolta si propone di studiare diversi metodi che consentano l'identificazione di anomalie in serie storiche relative ai consumi energetici di edifici universitari. La capacità di individuare tempestivamente situazioni anomale consente di agire in modo puntuale ed evitare danni laddove possibile. Lo studio propone come target

di interesse l'Area Infrastrutture e Approvvigionamenti dell'Università di Milano-Bicocca, la quale si occupa del monitoraggio e della gestione dei consumi di energia nei diversi edifici.

Attraverso l'applicazione di metodi di apprendimento non supervisionato quali decomposizione stagionale e analisi dei residui supportata dal test GESD, valutazione dei residui di un modello TBATS/SARIMA, sviluppo di un modello Isolation Forest e definizione di un sistema di raggruppamento tramite K-means, è possibile individuare delle osservazioni giornaliere distanti dal comportamento generale. L'approccio seguito si basa su una comparazione empirica dei risultati ottenuti con i diversi metodi e permette l'individuazione di alcune giornate come outliers. Nello specifico, si nota che durante le festività natalizie del 2018 si sono verificate alcune anomalie sia nell'edificio U1 che nell'U6 e che si sono riscontrate situazioni al di fuori della norma anche nei giorni 6/08/2018, 19/06/2019, 9-10/07/2019 e 6/04/2020 in entrambi gli edifici. Sebbene tale risultato possa essere in parte spiegato dalla frequenza ridotta di studenti e personale negli edifici o da attività di manutenzione, risulterebbe comunque utile la consultazione di un tecnico che permetta di verificare se non vi siano stati altri fenomeni particolari. Volendo valutare la presenza di giornate caratterizzate da consumi più alti del normale, si sono selezionati i giorni 20/07/2018, 27/06/2019, 23/07/2019 e il periodo a inizio Agosto 2020 per U1 e il giorno 19/02/2018 per U6.

Lo studio, pur mettendo a disposizione una panoramica abbastanza ampia sulle situazioni anomale in termini di consumi elettrici degli edifici valutati, risulta limitato dall'impossibilità di un riscontro sull'effettiva presenza di anomalie nelle giornate selezionate. Sarebbe infatti interessante poter disporre di un dataset etichettato o, in alternativa, poter ricorrere a un esperto che valuti i risultati ottenuti dall'applicazione dei metodi unsupervised. Si riconosce, inoltre, la possibilità di espansione futura dello studio con metodologie più avanzate, ad esempio attraverso l'utilizzo di approcci neurali per le serie temporali come i modelli LSTM. Un ulteriore e interessante sviluppo futuro, infine, potrebbe riguardare l'applicazione del metodo su scala più ampia a livello di edifici universitari o di granularità più fine (ad esempio considerando singole ore invece di giorni) e un'integrazione diretta nel sistema di rilevazione dei consumi in modo da poter accedere a insight relativi alla presenza di anomalie in modalità real-time.

Appendice

Si riportano i grafici ACF e PACF dei residui dei modelli TBATS e SARIMA per U1 e U6.

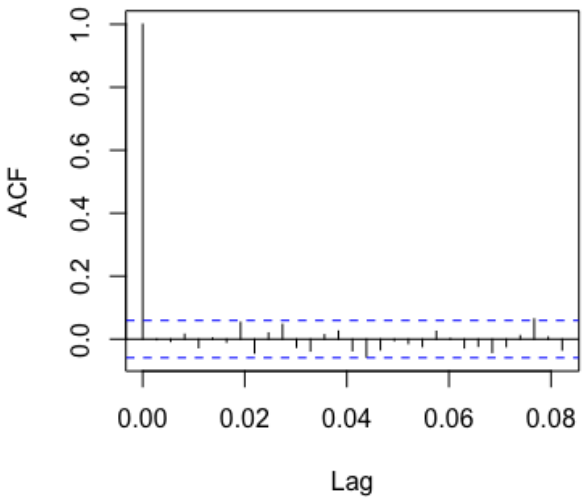

Figura 16. ACF residui TBATS U1

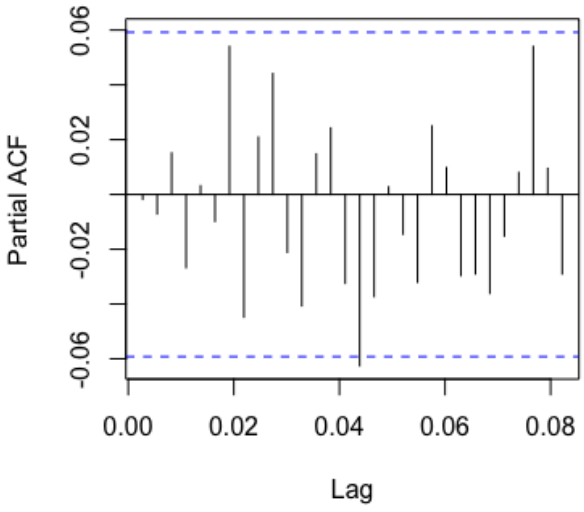

Figura 17. PACF residui TBATS U1

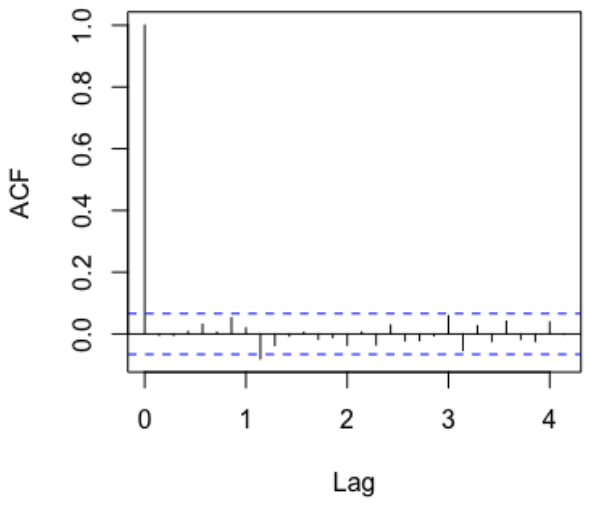

Figura 18. ACF residui SARIMA U6

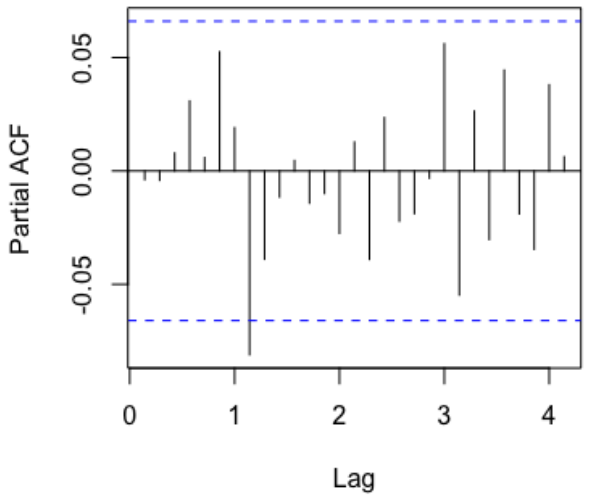

Figura 19. PACF residui SARIMA U6

Codice

Il codice utilizzato per l'analisi, sia in fase di lettura e preprocessing che in fase di *anomaly detection* vera e propria, è disponibile al link GitHub.

Riferimenti bibliografici

[1] Yassine Himeur, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Applied Energy*, 287:116601, 2021.

[2] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American statistical association*, 106(496):1513–1527, 2011.

[3] George.E.P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[4] Jordan Hochenbaum, Owen S Vallis, and Arun Kejariwal. Automatic anomaly detection in the cloud via statistical learning. *arXiv preprint arXiv:1704.07706*, 2017.

[5] Bernard Rosner. Percentage points for a generalized esd many-outlier procedure. *Technometrics*, 25(2):165–172, 1983.

[6] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.

[7] Bora Kizil. Introduction to anomaly detection in time-series data and k-means clustering, Oct 2020.