Analisi dell'impatto pandemico e metereologico sul consumo energetico dell'Università di Milano-Bicocca

Autore 1: Beatrice Barzaghi, 831829, b.barzaghi3@campus.unimib.it
Autore 2: Stefano Bassis, 826368, s.bassis@campus.unimib.it
Autore 3: Andrea Maenza, 826568, a.maenza@campus.unimib.it
Autore 4: Guglielmo Muoio, 826029, g.muoio@campus.unimib.it

Sinossi

Il seguente articolo vuole andare ad illustrare come due principali fattori vadano ad influenzare il consumo di energia elettrica in due edifici dell'università Milano-Bicocca. I due fattori in questione sono il manifestarsi delle restrizioni dovute alla pandemia di COVID-19 e le condizioni meteo rilevate nella zona in cui si trova l'università. Al fine di effettuare queste analisi sono stati utilizzati modelli descrittivi e predittivi in modo tale da indagare gli andamenti e le relazioni tra diversi fenomeni e fornire possibili previsioni future che potrebbero risultare utili per ideare nuove strategie energetiche migliorative. Oltre all'utilizzo delle funzioni più comuni, come ad esempio modelli ARIMA ed ETS, verrà presentato un approccio di regressione lineare: l'interpolazione spline. I principali risultati ottenuti mostrano come le restrizioni legate alla pandemia di COVID-19 abbiano impattato notevolmente i livelli di energia consumati come da aspettative e come anche la temperatura esterna influisca sulla quantità di energia utilizzata. È stato possibile osservare come una temperatura più elevata porti a consumi maggiori, probabilmente dovuti all'accensione degli impianti di aria condizionata, mentre, durante tutto il resto dell'anno, questi crescono linearmente con la temperatura.

Indice

Parole chiave

Consumo energetico; COVID-19; Meteo; Serie storiche; Spline

1 Introduzione

L'energia elettrica viene definita come la quantità di energia disponibile grazie al flusso di cariche elettriche in un conduttore o grazie alle forze tra particelle cariche.
In particolare in questa analisi si tratterà di consumo di energia elettrica. Attraverso lo studio dei dati relativi ai consumi elettrici, è possibile cercare di ottimizzare questi ultimi adottando comportamenti più efficienti che potreb-

bero portare ad una diminuzione dei consumi. La gestione dei consumi energetici diventa quindi una tematica strategica anche per le università sia dal punto di vista economico che ambientale. Il fenomeno di diminuzione dei consumi energetici si è evidenziato notevolmente durante gli avvenimenti drammatici avvenuti nel 2020 a causa del *lockdown* dovuto all'emergenza sanitaria.

L'impostazione del *report* è la seguente: le prime due sezioni riguardano gli obiettivi dell'analisi e gli aspetti metodologici che verranno utilizzati per raggiungere i primi. Successivamente vengono presentati i dati e le rispettive fonti e caratteristiche. L'aspetto più pratico dello studio, riguardante il processo di analisi e trattamento dei dati, viene presentato nella quinta sezione. A seguito di tutte le osservazioni fatte e delle performance ottenute, nelle ultime due sezioni vengono presentati i risultati, le conclusioni e i possibili sviluppi futuri dello studio.

2 Obiettivi

La gestione dei consumi energetici, elettrici e termici, in termini sia ambientali che economici, è una tematica strategica per l'Università Milano-Bicocca. Monitorare l'andamento, identificare le inefficienze, ottimizzare la gestione, promuovere il risparmio energetico sono tutte azioni fondamentali per migliorare la sostenibilità dell'Università. Il continuo monitoraggio dei consumi energetici è utile al fine di individuare eventuali criticità ed escogitare piani di miglioramento, sia dal punto di vista strutturale che comportamentale, come ad esempio una superiore gestione degli impianti e la promozione di atteggiamenti eco-sostenibili da parte di personale e studenti [7]. In questo studio si è voluto evidenziare l'impatto che il COVID-19 ha avuto sulla quantità dell'energia consumata dalla Bicocca. Dal punto di vista della sostenibilità il COVID-19 ha avuto un impatto "positivo"? Come sono variati i consumi energetici con l'avvento della didattica a distanza? Il secondo obiettivo di ricerca riguarda la possibilità di realizzazione di un piano di gestione sostenibile dell'energia sulla base di informazioni meteorologiche. Si vuole quindi appurare se esiste una relazione tra i consumi e il meteo attraverso la realizzazione di un database dei consumi energetici integrato con informazioni meteorologiche.

3 Aspetti metodologici

3.1 Decomposizione

Le serie storiche permettono di raccontare la dinamica di un determinato fenomeno. In generale una serie storica può essere vista come la somma di tre componenti principali: un andamento di fondo (o generale), chiamato anche *trend*, un andamento stagionale e una componente casuale, nota anche come *random error*. Il processo di decomposizione, ovvero di identificazione di ognuna delle tre differenti componenti, diventa un tema fondamentale interpretativo di lettura dei fenomeni.
Più in particolare il *trend* rappresenta l'andamento delle osservazioni statistiche che si muove su un periodo relativamente lungo, riuscendo ad esprimere come la serie storica sta evolvendo. Per quanto riguarda le altre due componenti solitamente quella stagionale ha andamento periodico mentre la componente di rumore viene utilizzata per rappresentare l'incertezza del fenomeno.
Una serie storica x_t può quindi essere scomposta nel seguente modo:

$$x_t = m_t + s_t + \epsilon_t \tag{1}$$

dove m_t rappresenta il *trend*, s_t la stagionalità e ϵ_t la componente casuale.

Attraverso la conoscenza delle componenti di una serie storica è possibile trovare delle traiettorie future di quest'ultima più o meno probabili. La sola conoscenza della traiettoria passata non vincola la traiettoria futura ma se la serie storica presenta un *trend* significativo e non ha oscillazioni troppo forti è maggiormente possibile cogliere l'effetto futuro.

Per effettuare la decomposizione sono stati utilizzati modelli ARIMA e ETS.
Una serie storica si dice stazionaria nel momento in cui non è presente la componente di *trend*, ovvero quando la media e la varianza del processo risultano essere costanti nel tempo e la sua covarianza si presenta come invariante per traslazioni. In questi casi i modelli utili in funzione di previsione sono i modelli ARMA, *autoregressive moving average*, ovvero modelli più

realistici sulla quale si basano i modelli ARIMA, in cui è presente la componente di *trend*. All'acronimo viene aggiunta la lettera "I" a rappresentare la parola *Integrated*: questo tipo di modello va a lavorare, in fase iniziale, sul processo per eliminare la non stazionarietà presente nella media. Al fine di stimare il *trend* si è deciso di utilizzare un approccio a media mobile: si tratta di un indicatore utilizzato per stimare il valore di un punto del dominio basandosi sulla media calcolata nei punti che appartengono ad un determinato intorno definito a piacere dall'utilizzatore. I modelli ETS, *exponential smoothing models*, sono modelli più moderni ingegneristici che risultano essere maggiormente interpretabili. Al fine di effettuare previsioni, l'algoritmo di ETS calcola una media ponderata su tutte le osservazioni nei dati della serie temporale in input. I pesi scelti risultano essere esponenzialmente in diminuzione nel corso del tempo, in particolare si è scelto di stimare il *trend* utilizzando un'estensione del modello ETS chiamata TBATS, *exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components*, che riesce a tenere conto della multistagionalità del processo [4, 8].

3.2 Modellazione effetto COVID-19

Per stimare l'impatto di una variabile su un fenomeno è stato deciso di utilizzare un modello lineare che regredisce un target di tipo serie-storica su una covariata di carattere binario. La covariata in questione rappresenta la presenza o assenza di restrizioni dovute alla pandemia di COVID-19 mentre il target è la serie storica del consumo energetico con granularità settimanale. Questo livello di granularità è stato scelto per ottenere risultati più significativi. Il risultato d'interesse di questo modello è il coefficiente relativo alla variabile rappresentante il COVID-19 e il p-value ad esso associato. Infatti non si vuole costruire un modello che descriva la variabile target bensì un modello che evidenzi l'impatto di una variabile esplicativa sul target. Affinché il modello sia robusto è importante che rispetti le ipotesi alla base dei modelli lineari. Le ipotesi in questione sono:

- **collinearità**: le variabili esplicative non devono essere perfettamente correlate tra loro altrimenti la matrice dei coefficienti beta non è invertibile e non è possibile determinare i parametri del modello. In caso di correlazione imperfetta invece aumenta la stima della varianza dei coefficienti con il rischio di valutare come non significative variabili che in realtà lo sono;
- **omoschedasticità**: la varianza dei residui deve essere costante altrimenti lo stimatore ai minimi quadrati del modello non è più efficiente e non è più B.L.U.E., ovvero *best linear unbiased estimator*;
- **normalità**: in caso i residui del modello non si distribuiscano normalmente le stime ottenute tramite i minimi quadrati non equivalgono a quelle ottenute tramite massima verosimiglianza. Conseguentemente non è possibile fare correttamente inferenza sui parametri; per gli stimatori dei minimi quadrati non vale più il teorema di Cramer-Rao;
- **non-autocorrelazione**: in caso di serie spaziali o temporali è possibile che i residui siano correlati tra loro in base alla loro posizione nello spazio o nel tempo, in questo caso si parla di autocorrelazione spaziale o temporale. Se questa ipotesi non viene rispettata lo stimatore ai minimi quadrati non è più efficiente e non è più B.L.U.E.

3.3 Modellazione dell'impatto del meteo

Al fine di valutare l'impatto del meteo si è costruito un modello predittivo il cui scopo è quello di prevedere il consumo energetico in base ai valori assunti da variabili di carattere meteorologico e dalla variabile relativa al COVID-19 precedentemente introdotta. Per costruire una previsione adeguata è stato necessario dividere il *dataset* di lavoro in due *subset*: uno di *training*, contenente circa il 70% delle osservazioni e uno di *test* con le restanti. Il modello viene costruito sul primo sottoinsieme di osservazioni, dopodiché viene utilizzato il *test set* al fine di valutare la bontà dei risultati ottenuti. Le osservazioni associate ai differenti *subset*, training e test, sono state assegnate in modo casuale senza seguire l'ordine temporale. Così facendo le osservazioni con valore 1 sulla variabile relativa al COVID-19 risultano essere distribuite equamente tra i differenti nuovi *dataset*.

Seguendo questa metodologia si rinuncia a considerare il valore aggiunto dato dalla storicità della serie per tenere in considerazione l'effetto del COVID-19. Questa scelta è dovuta al fatto che per ottenere le migliori previsioni possibili è necessario considerare le variabili che impattano significativamente sul consumo energetico tra cui il COVID-19. Essendo che il consumo energetico nel periodo analizzato è fortemente condizionato dalla presenza di restrizioni dovute alla pandemia, è stato ritenuto fondamentale considerare questa variabile per costruire un modello sui dati a disposizione. Per questo motivo è stato deciso di rinunciare a tenere in considerazione la temporalità presente nei dati preferendo considerare l'impatto della pandemia e migliorare la robustezza del modello. Anche in questo caso è necessario rispettare le ipotesi alla base dei modelli lineari dichiarate al paragrafo precedente.

Per modellare questa relazione è stata utilizzata una funzione spline [3]. Le spline sono funzioni di regressione con la caratteristica di essere calcolate solo su una parte del dominio. È infatti possibile segmentare il dominio di una variabile in più parti grazie all'utilizzo di nodi. Una funzione di regressione viene calcolata in ogni sezione del dominio con l'aggiunta del vincolo che la funzione di regressione così ottenuta sia continua su tutto il dominio. Inoltre, se il grado selezionato per la funzione di regressione è maggiore o uguale a 2, si aggiunge il vincolo che la funzione di regressione sia derivabile su tutto il dominio. Questa tipologia di regressione consente maggior flessibilità nella modellazione, soprattutto nei casi in cui vi sia una variazione dell'andamento della funzione in una sezione specifica del dominio. Nello studio in questione si ritiene che l'accensione dell'aria condizionata possa modificare la relazione tra temperatura ed energia consumata, motivo per cui si è valutato adatto l'utilizzo di un modello spline. Anche in questo modello è necessario che le ipotesi alla base dei modelli lineari vengano soddisfatte. Infine, una volta sviluppato il modello, per valutare la bontà dei risultati si è svolta una previsione del consumo energetico basandosi sulle variabili presenti nel *test set* e confrontandole con i valori osservati. In questo confronto, si è utilizzata una correlazione lineare il cui valore determina la bontà delle previsioni fornite dal modello e, di conseguenza, la validità di quest'ultimo.

4 I dati

Al fine di perseguire gli obiettivi di ricerca sono stati reperiti i dati da due fonti differenti:

1. Università Milano-Bicocca, per quanto riguarda i dati energetici;
2. Arpa, per quanto riguarda i dati meteorologici.

4.1 Università Milano-Bicocca

I dati riguardanti l'energia consumata dai due edifici in analisi, ovvero gli edifici U1 e U6 dell'Università Milano-Bicocca, sono memorizzati in differenti file, in formato csv, xltx o xlsx. In particolare per ogni edificio e per ogni anno, 2018, 2019 e 2020, sono presenti differenti file corrispondenti a ciascun mese dell'anno. Si è potuto notare che alcuni mesi presentano più file in formato diverso, come ad esempio il documento relativo all'edificio U6 riguardante i dati di marzo 2020, e altri mesi risultano essere suddivisi in più di un file.

In ciascun documento sono presenti le seguenti variabili:

- **POD**: identificativo del contatore di energia;
- **data**: giorno in cui avviene la rilevazione;
- **ora**: orario di rilevazione, i dati vengono acquisiti ogni 15 minuti;
- **fl_ora_legale**: rappresenta se la rilevazione è stata effettuata in un periodo caratterizzato da ora legale o da ora solare;
- **consumo_attiva_prelevata**: valore dei kW assorbiti nei 15 minuti considerati (potenza);
- **consumo_reattiva_induttiva_prelevata**: consumo energetico prelevato ma non effettivamente impiegato;
- **potenza massima**: limite di potenza massima rilevabile dal contatore e prelevabile dalla rete elettrica.

4.2 Arpa

L'Agenzia Regionale per la Protezione dell'Ambiente si occupa di monitorare la situazione ambientale in Lombardia in tutti i suoi aspetti, come ad esempio quello meteorologico [2]. In particolare è possibile richiedere accesso alle schede degli indicatori ambientali e scaricare i relativi file. Gli indicatori ambientali sono uno strumento per rappresentare in modo sintetico e standardizzato le informazioni e offrire una visione con un passo temporale costante, così da poter rendere conto dell'evoluzione nel tempo dei fenomeni che si vogliono descrivere.

- **temperatura**: temperatura media oraria espressa in gradi Celsius;
- **umidità**: umidità media oraria espressa in termini percentuali;
- **radiazione_globale**: radiazione solare globale media espressa in $\frac{W}{m^2}$;
- **velocità_vento**: velocità media oraria del vento espressa in m/s;
- **precipitazioni**: precipitazioni cumulate espresse in millimetri.

4.3 Qualità dati

La qualità dei dati può essere definita come un insieme di caratteristiche, o dimensioni, che influiscono sulla capacità di soddisfare le esigenze e le aspettative esplicite o implicite degli utenti che usufruiscono dei dati. In particolare tra le diverse dimensioni sono state considerate l'attendibilità delle fonti, la consistenza, la completezza e l'unicità dei dati [1].

- **Attendibilità delle fonti**: le fonti sopracitate godono di un'elevata affidabilità, difatti i dati relativi al consumo energetico derivano da rilevazioni dirette effettuate mediante appositi contatori presenti negli edifici U1 e U6. L'altra fonte da cui provengono i dati, ovvero Arpa, opera nel settore della ricerca e della protezione ambientale ormai da diversi anni (a partire dal 1994) e, in quanto ente della pubblica amministrazione italiana, viene ritenuta una fonte affidabile.
- **Consistenza**: per valutare la consistenza dei dati sono state considerate le sue due principali definizioni.
  1. Consistenza dei dati con i vincoli di integrità: un esempio riguarda la data, la quale deve risultare consistente con la tipologia di orario, a seconda della presenza di ora legale o solare. Nei due *dataset* tutte le date comprese tra l'ultima domenica di marzo e l'ultima domenica di ottobre sono correttamente contraddistinte da orario legale, mentre quelle restanti da orario solare. In generale non sono stati riscontrati problemi di consistenza dei dati con i vincoli di integrità.
  2. Consistenza dei formati delle stesse variabili in differenti *dataset*: in entrambi i *dataset*, la data di rilevazione è riportata nel formato anno/mese/giorno e perciò risulta essere consistente. Anche le altre variabili non riportano problemi di consistenza del formato.
- **Completezza**: la completezza di un insieme di dati indica la copertura con la quale il fenomeno osservato è rappresentato nell'insieme di dati e può essere calcolata come il numero di valori non nulli sul numero totale di valori dell'insieme di dati considerato. Per i dati a nostra disposizione sono stati riscontrati alcuni problemi di completezza, di cui verrà data un'esauriente spiegazione nel paragrafo 5.1. L'esempio più evidente è quello del mese di giugno 2020: a causa di un errore le rilevazioni risultano essere uguali per entrambi gli edifici. In particolare le rilevazioni di U6 sono una copia di quelle di U1 e, di conseguenza, è stato deciso di eliminarle e di imputare i conseguenti valori mancanti.
- **Unicità dei dati**: i dati possono essere definiti unici se non viene riscontrata la presenza di duplicati. In particolare nei *dataset* di analisi solo due giorni riportano dei duplicati: il 30 aprile 2019, per l'edificio U1, e il 25 ottobre 2020 per entrambi gli edifici.

Punti di debolezza dei dati

- **Aggregazione per quarto d'ora**: i dati riportano una granularità eccessivamente

elevata e poco utile ai fini dell'analisi. Per questo motivo, durante la ricerca, i dati sono stati aggregati a livello giornaliero o settimanale.

- **Missing values per le osservazioni dell'edificio U6 di giugno 2020**: i dati relativi all'edificio U6 per il mese di giugno 2020 sono mancanti a seguito dell'errore precedentemente descritto.

- **Errori nel passaggio da orario legale a orario solare**: per quanto riguarda il 2020, nei giorni in cui si verifica il cambio d'ora, vengono create alcune osservazioni con valori mancanti.

Punti di forza dei dati

- **Buona qualità dei dati iniziali**: in generale la consistenza, la completezza e l'unicità dei dati sono proprietà quasi sempre garantite. Fanno eccezione alcuni casi precedentemente descritti.

- **Elevato numero di osservazioni**: la mole di dati a disposizione permette di svolgere analisi più robuste che portano, di conseguenza, a dei risultati più solidi.

5 Processo di trattamento dei dati

5.1 Data Preprocessing

Dopo aver accuratamente letto ed aggregato i dati riferiti ai due edifici in ciascun anno, si è passati ad effettuare un'attenta fase di *preprocessing* al fine di individuare, valutare e correggere eventuali incongruenze e/o errori presenti all'interno dei diversi *dataset*.

Una prima analisi riguarda il controllo del numero di giorni contenuti all'interno di ciascun *dataset*, di norma 365 tranne nel caso di anno bisestile, come ad esempio l'anno 2020, in cui dovrebbero essere presenti 366 giorni. Parallelamente si è verificato se il numero di osservazioni per ciascun giorno fosse corretto: essendo presente infatti una granularità a livello di quarto d'ora, per ciascun giorno si dovrebbero avere 96 osservazioni eccezion fatta per i giorni del cambio d'ora in cui si dovrebbero presentare 92 osservazioni nella giornata di passaggio da ora solare a ora legale e 100 nel caso opposto. Oltre a ciò, si è controllato l'eventuale presenza di valori mancanti e di possibili duplicati.
Nello specifico per ciascuna coppia edificio-anno, si sono riscontrati i seguenti problemi:

- **U1-2018**: nessun problema;

- **U1-2019**: numero di osservazioni sbagliate, maggiori del previsto. Verificata la presenza di valori mancanti e rimossi questi, si hanno solo 2 osservazioni in meno che, a seguito di un controllo veloce, risultano essere valori aggregati senza alcun tipo di contesto; per tale motivo si scartano. Il numero di giorni è giusto, ma un giorno ha duplicati: il 30 aprile; di conseguenza vengono rimossi;

- **U1-2020**: 17760 osservazioni non hanno la potenza massima ma fortunatamente è un campo non rilevante ai fini dello studio. Il giorno 25 ottobre ci sono il doppio delle osservazioni (192) mentre dovrebbero essere 100 dato il cambio d'ora da legale a solare. Esplorando i dati si notano osservazioni con valori nulli che sono probabilmente dovute al cambio d'ora che ha generato osservazioni nulle, anche queste vengono rimosse;

- **U6-2018**: nessun problema;

- **U6-2019**: ci sono problemi di valori mancanti esclusivamente relativi ai dati di potenza massima;

- **U6-2020**: sono emerse le stesse difficoltà dell'altro edificio. Presenza di *missing values* per la potenza massima e presenza di duplicati per il giorno di cambio d'ora legale-solare verificatosi il 25 ottobre.

Prima di salvare i dataset, si è voluto trasformare e uniformare i tipi delle variabili, perlopiù numeriche ad eccezione del *Datetime*, in modo tale da permettere una facile ed immediata aggregazione, avvenuta poi in un secondo momento.

5.2 Esplorazione dei dati

In primo luogo sono stati concatenati i *dataset* creati rispettivamente per i diversi edifici. Inoltre è stato necessario occuparsi della creazione

di alcune variabili temporali, come ad esempio un *timestamp*, al fine di facilitare le operazioni di visualizzazione.
Viene riportato in Figura 1 un primo grafico rappresentante il consumo di energia rispetto ai differenti anni 2018, 2019 e 2020.

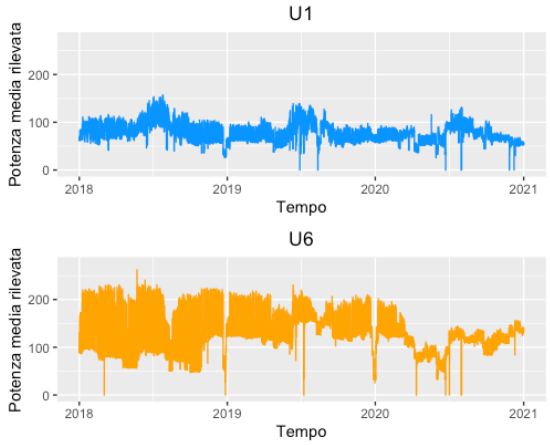

Figura 1: Potenza media rilevata

Si procede successivamente alla creazione dei *dataset* nominati rispettivamente U1_daily e U6_daily, in cui i dati risultano essere aggregati a livello giornaliero, poichè precedentemente questi presentavano una granularità di 15 minuti. Tuttavia non è possibile aggregare dati di potenza media senza effettuare prima una trasformazione della variabile. Infatti la potenza elettrica si misura in kW e non si tratta di una grandezza cumulativa al contrario dell'energia, solitamente misurata in kWh. Di conseguenza si è creata una nuova variabile, energia_consumata, attraverso la seguente formula:

$$Energia = \sum_{i=1}^{n} Potenza \cdot Tempo \quad (2)$$

dove $Tempo = 0.25$ poiché ci si riferisce a quarti d'ora e n varia a seconda del livello di aggregazione. Ad esempio:

- $n = 96$ a livello giornaliero, ad eccezione dei giorni caratterizzati da cambio d'ora
- $n = 672$ a livello settimanale

In seguito alla creazione dei due nuovi *dataframe* vengono effettuate alcune prime analisi esplorative come ad esempio delle *summary* e dei grafici, in particolare *ggplot* come ad esempio boxplot in grado di evidenziare la presenza di *outliers*. In particolare si osserva che per quanto riguarda i dati relativi all'edificio U1, sono presenti 33 valori anomali che risultano essere sparsi su tutto il mese di luglio dell'anno 2018, *outliers* a valori alti per la fine di giugno 2019 e infine l'anno 2020 ne presenta solo due bassi rispettivamente il 31 luglio, in cui si è verificato un blackout in alcuni quartieri di Milano, e il 21 giugno. Attraverso l'utilizzo del boxplot in Figura 2 si è potuto osservare che, nel caso dell'edificio U6, sono presenti 19 *outliers* di cui è interessante notare che appartengono ad osservazioni quasi tutte riguardanti giorni festivi, come ad esempio le domeniche.

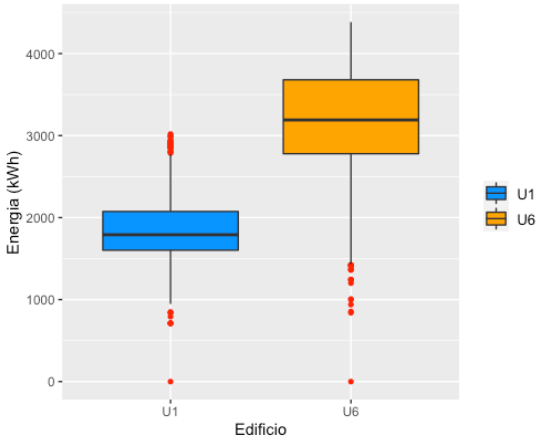

Figura 2: Boxplot energia giornaliera

Sempre attraverso il boxplot di Figura 2 è possibile notare come i dati riguardanti l'energia consumata per l'edificio U6 risultano essere più elevati rispetto a quelli dell'edificio U1. Questo potrebbe essere dovuto alla differente superficie e differenti peculiarità dei due edifici: l'edificio U1 risulta essere infatti più piccolo rispetto a U6, nel primo risultano essere presenti 13 aule, mentre nel secondo ben 46. Data l'ampia metratura dell'edificio U6 sono presenti un maggior numero di ascensori, scale mobili ed una mensa. Inoltre riuscire a raffreddare l'ambiente attraverso l'utilizzo di aria condizionata, la quale ha un grande impatto sul consumo di energia, potrebbe risultare più faticoso.

Mediante la *summary* del dataset U6_daily si è potuto notare la presenza di due differenti POD, ovvero di due diversi identificativi del contatore di energia. Questo è dovuto ad un errore nella raccolta delle osservazioni: è possibile notare che i dati relativi all'edificio U6 di giugno 2020 sono identici ai dati rispettivi dell'edificio U1. Al fine di sistemare questo problema si potrebbe decidere di rimuovere i dati errati, rim-

piazzarli con valori mancanti oppure cercare un metodo di imputazione dei dati. L'imputazione dei dati purtroppo risulta essere difficile poiché è mancante tutto il mese di giugno; non si tratta di dati isolati ma di giorni consecutivi. Per effettuare una prima interpolazione dei dati viene utilizzato il pacchetto *imputeTS* in cui una retta va a descrivere l'andamento dei dati mancanti, congiungendo l'ultimo dato corretto, ovvero il 31 maggio 2020, con il prossimo dato corretto, ovvero l'1 luglio.

In seguito si è deciso di provare a correggere i dati attraverso decomposizione stagionale, la quale risultava essere quasi identica al primo metodo di interpolazione. L'ultimo approccio prevede la suddivisione della serie storica in stagioni e successivamente l'imputazione dei dati, stagionalmente separata, attraverso utilizzo di un algoritmo, ovvero viene sfruttato uno split stagionale, rappresentato in Figura 3 [5]. Una volta "corretti" i dati si ripropongono delle analisi esplorative [6].

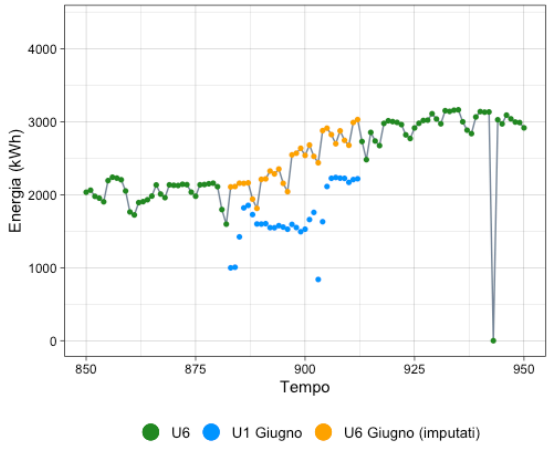

Figura 3: Imputazione tramite split stagionale

Per quanto riguarda i dati scaricati dal portale Arpa si è deciso di effettuare una prima analisi delle correlazioni presenti tra le differenti variabili rappresentanti ognuna diverse grandezze meteorologiche, oltre che alla variabile principale rappresentante l'energia.

Come si evince dalla Figura 4, è possibile osservare che nessuna delle variabili presenta distribuzione normale e come le due variabili che risultano essere più correlate all'energia siano temperatura e radiazione, le quali risultano a loro volta fortemente correlate tra loro.

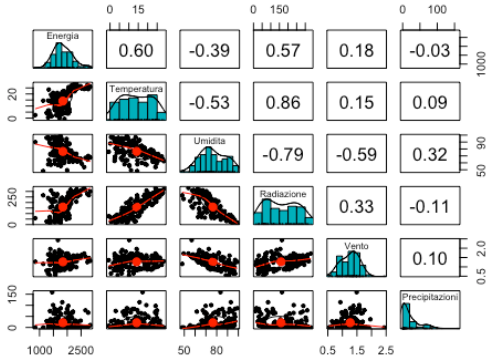

Figura 4: Distribuzioni e correlazioni tra variabili meteo

5.3 Decomposizione

Una delle prime analisi svolte sui dati preprocessati è stata quella di stimare le componenti della serie storica a disposizione: *trend*, stagionalità e rumore bianco, un tipo particolare di errore casuale. Per stimare il *trend* del consumo energetico sono state calcolate finestre di ampiezza differente con lo scopo di implementare un "lisciamento" della serie tramite media mobile. Considerando che la frequenza delle osservazioni è di un quarto d'ora e che il totale delle osservazioni è 105216, la finestra utilizzata è di 1000 osservazioni. Maggiore è la dimensione della finestra più liscio sarà il *trend*, al contrario più piccola è la finestra più il *trend* seguirà gli andamenti locali. La stagionalità della serie storica è stata calcolata su vari livelli poiché si sono notati diversi andamenti ciclici. In particolare una stagionalità a livello giornaliero, settimanale e annuale. Queste ultime sono state calcolate utilizzando modelli ARIMA e modelli ETS. Infine il rumore bianco è stato calcolato in modo analogo alla stagionalità. Queste analisi sono state svolte sia per l'edificio U1 che per l'edificio U6.

5.4 Modellazione effetto COVID-19

Uno degli scopi dello studio è valutare l'impatto avuto dal COVID-19 sull'erogazione di energia nel solo edificio U6. L'obiettivo è studiare questo impatto tramite un modello descrittivo che sia in grado di esprimere un coefficiente significativo sulla variabile binaria: presenza o assenza di limitazioni dovute alla diffusione del virus. Inoltre questo modello vuole essere ro-

busto nelle ipotesi, in particolare di normalità e non autocorrelazione dei residui. La scelta di un modello descrittivo viene motivata dal fatto che si è preferito descrivere l'impatto del COVID-19 nel passato piuttosto che svolgere previsioni sul futuro a causa dell'unicità dell'evento. Per sviluppare questo modello si è rivelato efficace aggregare i dati a livello settimanale; in questo modo è stato possibile limitare l'impatto di osservazioni anomale sulla serie storica e quindi rendere più solido il modello. Quest'ultimo è stato calcolato tramite una funzione di *time series linear model* che consente di calcolare modelli di regressione lineare utilizzando come variabile esplicativa una serie storica. Un'ulteriore azione per limitare l'impatto delle osservazioni anomale è applicare un "lisciamento" della serie storica tramite una rete neurale con un singolo *hidden layer* implementato dalla funzione *nnetar*. Inoltre è stata svolta una trasformazione di Box-Cox per rendere più simile a una distribuzione normale la variabile target. Questa trasformazione ha però il difetto di rendere più difficile l'interpretazione dei coefficienti stimati.

5.5 Modellazione dell'impatto del meteo

L'ultimo obiettivo è la valutazione dell'impatto del meteo sulla variabile target relativamente all'edificio U6. Per fare questo si sono sfruttati i dati meteorologici di Arpa, introdotti precedentemente. In questo caso lo scopo del modello è predittivo. L'obiettivo è quello di stimare l'erogazione energetica in base al valore assunto dalle variabili di carattere meteorologico. I dati sono stati aggregati a livello settimanale per poter inserire la variabile dicotomica relativa al COVID-19 nel modello, poiché ritenuta fondamentale al fine di spiegare il consumo energetico.

In primo luogo è stata svolta una selezione delle variabili da inserire nel modello. Dato il possibile legame delle variabili tra loro, si è svolta una fase di *feature selection* sfruttando il criterio di Akaike in direzione *backward*. Le variabili selezionate sono: temperatura, temperatura al quadrato, radiazione globale e COVID-19. Tra le variabili scelte è stata inserita la temperatura al quadrato poiché lo scatterplot tra la variabile target e la temperatura suggerisce la presenza della relazione quadratica, come si evince dalla Figura 5.

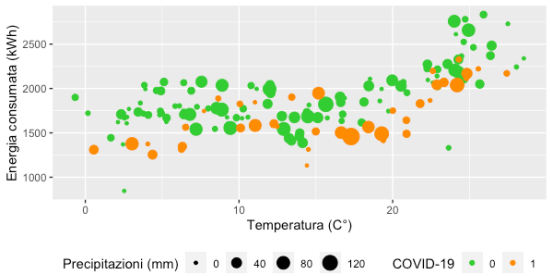

Figura 5: Scatterplot energia consumata vs temperatura

Successivamente si sono svolti controlli sulle ipotesi di non collinearità, omoschedasticità, assenza di outlier e normalità dei residui. Svolgendo i rispettivi test d'ipotesi sono state mantenute come variabili esplicative solo la temperatura al quadrato e il COVID-19, ed inoltre sono state eliminate due osservazioni ritenute influenti. Questo processo ha consentito di mantenere in analisi solo una variabile esplicativa suggerendo la possibilità di regredire la potenza erogata su una beta-spline della temperatura invece che sulla temperatura al quadrato. Questa scelta viene giustificata dal fatto che l'aumento dell'erogazione energetica per valori alti di temperatura sia causata dall'accensione dell'aria condizionata. Si è ritenuto adatto modellare la variabile temperatura tramite beta-spline sulla base dell'idea che gli andamenti siano differenti a seconda che l'aria condizionata sia spenta o accesa. In particolare, si è deciso di implementare una beta-spline lineare con nodo posizionato a 19.3 °C di temperatura. Questo modello si è dimostrato essere più performante rispetto al precedente. Infine sono state svolte le previsioni sui dati di test basandosi sul modello con implementazione della beta-spline.

6 Risultati

6.1 Decomposizione

Il primo risultato ottenuto tramite decomposizione delle serie storica è il trend relativo ai due edifici.
Dal grafico in Figura 6, è possibile notare la differenza tra i due andamenti. In particolare, nell'edificio U1, si registrano dei picchi nel

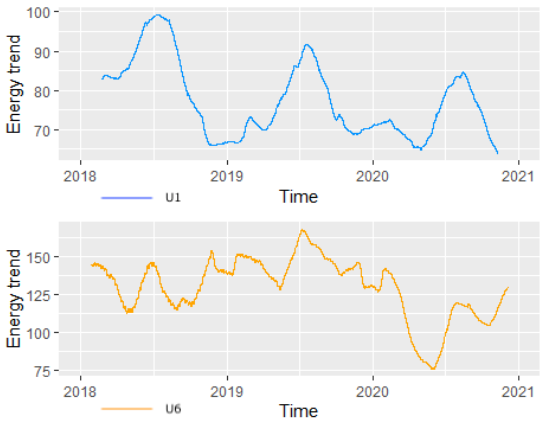

Figura 6: Trend

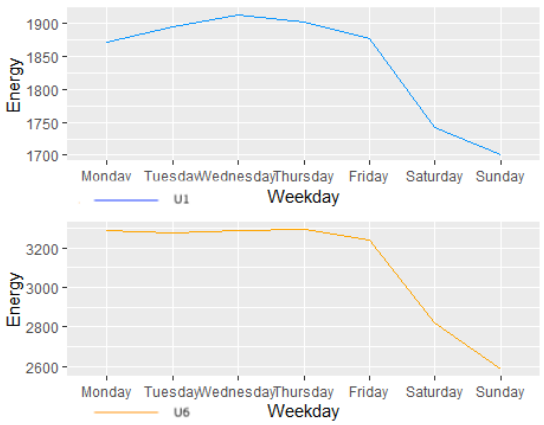

Figura 8: Andamento settimanale

periodo estivo. Riguardo l'edificio U6, invece, si notano picchi meno definiti e un notevole calo relativo al periodo del *lockdown* di marzo e aprile 2020.

Successivamente si riportano le analisi di stagionalità. In questo caso, per semplicità di lettura dei grafici, si è preferito riportare uno spaccato degli andamenti stagionali registrati.
Il primo spaccato presentato è relativo all'andamento orario.

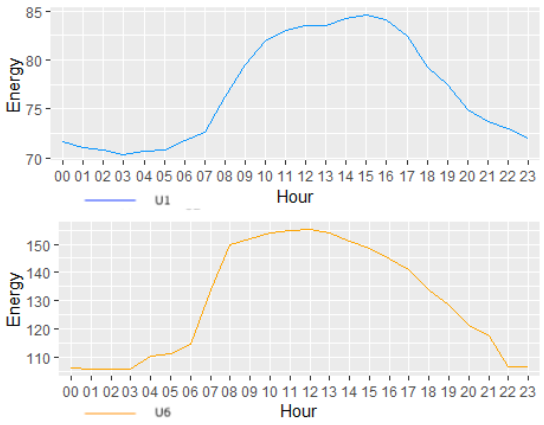

Figura 7: Andamento orario

L'andamento tra i due edifici è simile e si nota un rapido incremento del consumo energetico attorno alle 8 del mattino, un picco nelle ore centrali della giornata e una diminuzione costante fino alle ore serali.

Il successivo livello di stagionalità analizzato è relativo all'andamento giornaliero di cui vengono riportati gli spaccati in Figura 8 e 9.
Anche in questo caso l'andamento registrato è simile tra i due edifici seppur il consumo in U6 sia pressoché costante durante i giorni feriali mentre in U1 si nota un picco il mercoledì. In entrambi i casi si ha un netto e prevedibile calo dei consumi nei giorni festivi.

L'ultimo livello di stagionalità analizzato è relativo all'andamento mensile.

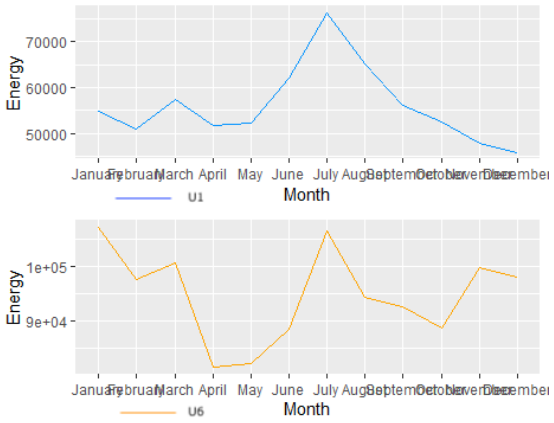

Figura 9: Andamento mensile

In questo caso la differenza di andamento tra i due edifici è notevole. A fronte di un comune picco nei mesi estivi, nell'edificio U6 si nota un notevole calo nei consumi nei mesi autunnali e primaverili per ottenere un altro picco nei mesi invernali. Nell'edificio U1 invece si nota un calo costante da luglio fino a gennaio con un leggero picco nel mese di marzo.

L'ultimo elemento utile alla decomposizione della serie storica è il rumore bianco che viene riportato nella Figura 10 per i due edifici. In particolare nell'edificio U6 i valori riportati sono maggiori in valore assoluto poiché il consumo energetico in quell'edificio risulta essere

maggiore.

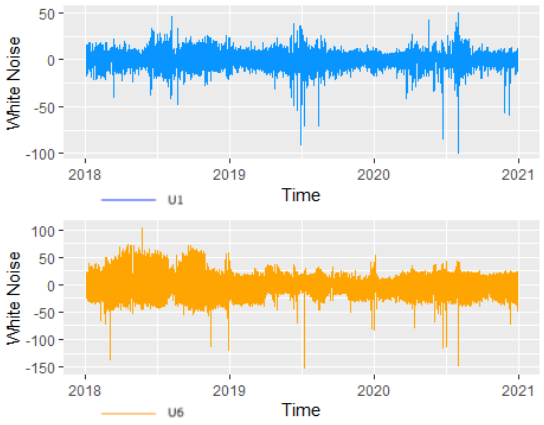

Figura 10: Rumore bianco

6.2 Modellazione effetto COVID-19

Il modello ottenuto riporta, relativamente alla variabile COVID-19, un p-value sulla significatività pressoché pari a 0, rispettando quindi un livello α standard del 5%. Il coefficiente relativo alla presenza di restrizioni dovute al COVID-19 è -194.41, indicando che, in caso di restrizioni, il consumo energetico medio per l'edificio U6 diminuisce di 194.41 kWh, a parità delle altre variabili esplicative.

Successivamente si valuta la robustezza del modello valutando le ipotesi che ne stanno alla base con i relativi grafici diagnostici. Dai grafici si nota che sono rispettate le ipotesi di autocorrelazione parziale dei residui, come testimonia il grafico in Figura 11. I principali problemi del modello sono relativi alla normalità dei residui. Come si può notare dall'istogramma e dal qqplot, la normalità dei residui non è rispettata. Questo è il principale problema del modello e si ritiene che sia dovuto alla poca normalità dei dati. Non vengono inseriti nel modello i dati trasformati con Box-Cox poiché il miglioramento nella normalità dei residui non risultava essere notevole a fronte di una difficile interpretazione dei risultati. Per valutare ulteriormente la normalità è stato applicato il test di Shapiro-Wilk che restituisce un p-value pressoché uguale a 0. Questo porta al rifiuto dell'ipotesi di normalità dei residui. In conclusione, seppur ci siano dei problemi di robustezza del modello, l'impatto del COVID-19 sembra essere significativo sul consumo energetico.

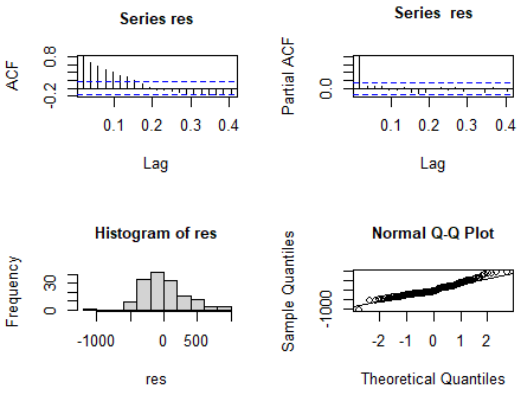

Figura 11: Grafici diagnostici relativi alla modellazione dell'effetto COVID-19

6.3 Modellazione dell'impatto del meteo

Il modello finale scelto per valutare l'impatto del meteo sul consumo energetico è una regressione sulla variabile COVID-19 precedentemente introdotta e una regressione spline di primo grado con nodo a 19.3 °C sulla variabile temperatura. In primo luogo viene mostrato come la spline interpola i dati di temperatura rispetto ai valori di energia consumata all'interno della settimana.

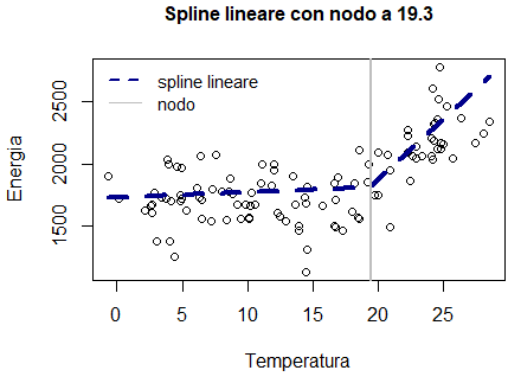

Figura 12: spline lineare

Si nota un cambio di trend in corrispondenza del nodo della spline. Questo potrebbe essere dovuto all'accensione dell'aria condizionata e porta a pensare che la spline così costruita possa ben rappresentare la relazione tra le due variabili. Infatti il coefficiente angolare relativo al primo tratto di retta è 82.01, indicante un aumento dei consumi di 82.01 kWh per l'aumento unitario di temperatura, al netto delle altre variabili. Il coefficiente angolare relativo al se-

	R^2	MAE	MAPE	Corr$(y, \hat{y})$
Training	0.67	122.18	6.74%	0.89
Test	0.49	178.45	9.87%	0.84

Tabella 1: Risultati ottenuti

condo tratto di retta invece è 763.75 a testimoniare il rapido aumento dei consumi per valori alti di temperatura. In questo caso, il modello è predittivo per cui il risultato principale è la correlazione tra le previsioni sul *dataset* di test fornite dal modello sviluppato sul *dataset* di train. Questa correlazione ha valore 0.84 e il grafico in Figura 13 mostra la relazione tra valori osservati e valori previsti. La linea arancione rappresenta la bisettrice del primo quadrante sulla quale si disporrebbero tutte le osservazioni in caso di previsione perfetta.

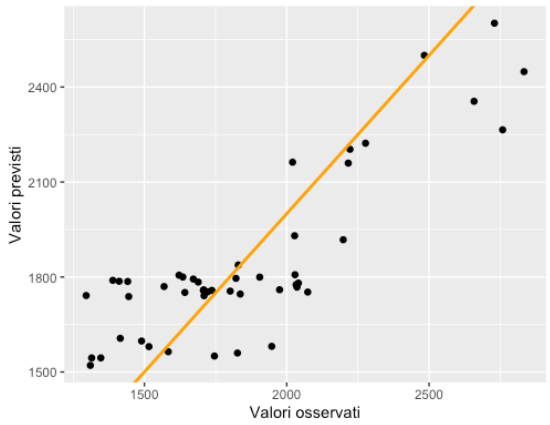

Figura 13: Relazione tra valori osservati e valori previsti dal modello meteo

Si notano le quattro osservazioni in alto a destra rispetto alla bisettrice del primo quadrante. Questo indica che il modello sottostima valori di consumo energetico particolarmente elevati. Un altro indice d'interesse è il MAE che sui dati di test ha valore 178.45 kWh, inteso come media dei valori assoluti dei residui. Infine si valuta la robustezza del modello attraverso i grafici presenti in Figura 14.

Lo scatterplot dei valori previsti sui residui indica l'omoschedasticità dei dati che è confermata dal test di Breusch-Pagan che riporta p-value 0.99. Il qqplot indica la normalità dei residui che è a sua volta confermata dai test di Shapiro-Wilk e Kolmogorov-Smirnov i quali riportano p-value rispettivamente di 0.06 e 0.22. Lo Scale-Location plot indica che non ci sono osservazioni con residuo standardizzato eccessivamente elevato per cui non si evidenziano problemi di outlier. Infine lo scatterplot dei residui sul leverage conferma quanto indicato dal grafico precedente mostrando l'assenza di osservazioni con leverage particolarmente elevato. In conclusione il modello risulta essere robusto e con buone performance sui dati di test.

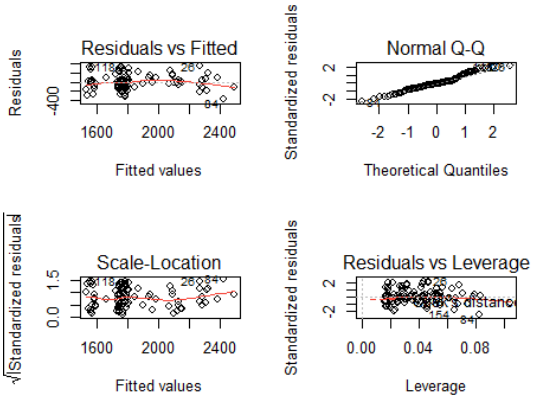

Figura 14: Grafici relativi alle ipotesi del modello

7 Conclusioni e possibili sviluppi

Il primo risultato ottenuto riguarda le decomposizioni delle serie storiche relative ai due edifici, attraverso le quali si è potuto notare come il trend non risulti essere stazionario. Inoltre sono stati riscontrati diversi livelli di stagionalità: oraria, giornaliera e mensile su entrambi gli edifici. I risultati ottenuti presentano, come da aspettative, delle similitudini tra i due edifici ma anche alcune differenze notevoli, come ad esempio l'andamento relativo alla stagionalità mensile, come è possibile osservare attraverso la Figura 9. Questo mostra come la differente affluenza agli edifici, dovuta ai diversi scopi di questi, possa influenzare in modo notevole il consumo energetico. Successivamente è stato analizzato l'effetto della pandemia di COVID-19 sull'utilizzo di energia da parte di un edificio. In particolare questa analisi è stata svolta sull'edificio U6, poiché ritenuto il principale del campus. L'obiettivo è quantificare l'ovvia riduzione dei consumi. Tramite il modello descritto al paragrafo 5.4, è stato possibile quantificare che mediamente, durante le restrizioni dovute alla pandemia, il consumo energetico si riduce, in media e a parità di altri fattori, di una quantità pari a 194.41 kWh. Infine, è stato studiato l'impatto di alcune variabili relative al meteo

sul consumo energetico. Dopo una fase di selezione iniziale l'unica variabile influente risulta essere la temperatura. La relazione tra la temperatura e il consumo energetico è stata modellata tramite spline poiché si è notato, a partire da una temperatura di 19.3 °C, un notevole aumento dell'utilizzo di energia probabilmente dovuto all'accensione dell'aria condizionata. Questo modello è stato costruito con fini predittivi ottenendo una correlazione tra valori osservati e valori previsti di 0.84.
Rimane aperta la questione relativa alla modellazione dell'effetto COVID-19 e dell'impatto meteo riguardante l'edificio U1, dato che per queste sezioni ci si è voluti concentrare esclusivamente sull'edificio principale dell'Università, ovvero U6. In particolare sarebbe interessante avere a disposizione i dati di molteplici edifici dell'Università di Milano-Bicocca, o anche di altri atenei, per creare un modello che, sulla base di parametri variabili relativi al singolo edificio e alle previsioni meteorologiche, possa con un certo grado di precisione prevedere il consumo energetico al fine di ottimizzare le risorse da impiegare. Infine un possibile ulteriore sviluppo potrebbe essere quello di monitorare l'impatto che il COVID-19 avrà nei prossimi anni, in un'ottica di cambiamento a lungo termine per quanto riguarda le abitudini di consumo energetico per gli edifici pubblici.

Riferimenti bibliografici

[1] Rezzani A. *Big Data: Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati*. Maggioli Editore, 2013.

[2] ARPA. Dati metereologici, 2021. `https://www.arpalombardia.it/`.

[3] De Boor C. A practical guide to Splines (rev. edition). 2001.

[4] Fattore M. Fundamentals of time series analysis for the working data scientist. 2020.

[5] CRAN R Project. Package "imputeTS", 2021. `https://cran.r-project.org/web/packages/imputeTS/imputeTS.pdf`.

[6] Variawa R. Energy forecasting with time series analysis, 2019. RPubs by RStudio. `https://rpubs.com/Ryder/555526`.

[7] UNIMIB. Energy management, 2021. `https://www.unimib.it/ateneo/energy-management`.

[8] Holmes E. E.; M. D. Scheuerell; E. J. Ward. Applied time series analysis for fisheries and environmental sciences, 2021. NOAA Fisheries, Northwest Fisheries Science Center. `https://nwfsc-timeseries.github.io/atsa-labs/`.