Analisi dei momenti di elevato consumo energetico universitario

Autore 1: Giacomo Stoffa, g.stoffa1@campus.unimib.it
Autore 2: Tommaso Redaelli, t.redaelli7@campus.unimib.it
Autore 3: Leonardo Riva, l.riva37@campus.unimib.it
Autore 4: Riccardo Merlo, r.merlo2@campus.unimib.it

Sinossi. La ricerca ha lo scopo di definire ed analizzare i momenti di eccessivo consumo energetico nell'università degli studi di Milano-Bicocca tra il 2018 e il 2020, negli edifici U1 e U6. In particolare, sono stati confrontati e descritti a livello statistico i due edifici, individuate le anomalie e riconosciute tipologie di consumo giornaliero diverse per caratteristiche di andamento. Questa analisi permette di visualizzare quando e come si eccede nel consumo, e quindi consentire di proporre dei piani di contenimento e risparmio mirati. Con i risultati ottenuti abbiamo compreso e delineato, partendo dalla serie storica iniziale, le componenti principali di consumo con i rispettivi andamenti e stagionalità. Successivamente sono state evidenziate le principali differenze presenti tra edifici ed anni, e con un occhio ai residui generati, sono state messe in luce numerose anomalie, con differenze tra gli edifici, a volte legate a periodi estivi, altre a periodi di lezioni universitarie, altre ancora irregolari e non direttamente riconducibili ad eventi specifici. Infine, abbiamo ottenuto tramite clustering differenti prototipi di consumo giornaliero. Il posizionamento dei cluster, all'interno della serie temporale, permette di individuare ed associare a diversi periodi, diversi significati e criticità.

Parole chiave: *Consumo energetico; Serie storiche; TBATS; Anomaly detection; Self-Organizing-Map.*

Indice

1 Introduzione

L'energia elettrica nel 2019 in Bicocca è costata 7 milioni di euro, più che qualsiasi altro servizio tecnico-gestionale (*Università degli studi di Milano-Bicocca, 2019*), nonostante l'ateneo sia tra i più attenti ai consumi energetici (*GreenMetric, 2020*). Gli sprechi energetici, una delle principali cause del riscaldamento globale, sono un problema che sta molto a cuore all'ateneo, come ribadito dalla sua attiva partecipazione a molte iniziative a favore della sostenibilità ambientale (*Università degli studi di Milano-Bicocca, 2021*). Il nostro impegno è quello di aiutare a scovare dinamiche che comportano utilizzi irrazionali e sprechi in maniera puntuale all'interno dei dati che ci sono stati forniti riguardo il triennio 2018-20. Molte campagne, come M'illumino di meno, in favore del risparmio energetico e degli stili di vita sostenibili, vengono supportate da Bicocca come da molti altri istituti, ma questi eventi cercano di ridurre drasticamente i consumi solo per poche ore, rare volte nel corso dell'anno. Il nostro lavoro, al contrario, vorrebbe sensibilizzare sui consumi, individuando le istanze che portano ad

un così grande spreco, in maniera tale da ottenere risultati solidi e dal maggiore impatto.

2 Obiettivi

L'analisi dei consumi energetici universitari ha due obiettivi principali. Il primo riguarda il rilevamento delle anomalie, che consiste nello scovare eventuali rilevazioni che si discostano in maniera significativa dall'andamento generale e che quindi non vengono spiegate da trend e stagionalità. A monte di questa analisi vi è quindi la decomposizione della serie storica iniziale come somma di diverse componenti, in modo da modellarne tutti gli aspetti. Inoltre, un confronto tra edifici dal punto di vista statistico può essere di supporto per comprendere il quadro generico della situazione. L'analisi delle anomalie permette di ricercare pattern utili a descriverle con maggior dettaglio ed eventualmente ad abbozzare un piano di contenimento e correzione. Successivamente, ci si pone l'obiettivo di distinguere e descrivere diverse tipologie di consumo giornaliero di un edificio nell'arco di un anno, caratterizzando, per ogni giornata, il consumo elettrico osservato sotto differenti punti di vista. Le informazioni ottenute tramite questa analisi potrebbero essere poi utili per la figura di un energy manager, nel momento in cui esso si trovi a dover prendere decisioni di ottimizzazione.

3 Aspetti metodologici

Tra le tecniche utilizzate nel corso dell'intera analisi, quelle fondamentali sono TBATS, K-Means per serie storiche, Isolation Forest e SOM.

3.1 TBATS

TBATS (*De Livera et al., 2009*) è un metodo di previsione in grado di modellare serie storiche con stagionalità complesse e multiple, a differenza del più semplice ARIMA. TBATS è acronimo di Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components. L'algoritmo modella infatti le stagionalità con rappresentazioni trigonometriche (tramite serie di Fourier), con un exponential smoothing state space model e una Box-Cox transformation, in maniera completamente automatizzata. Un modello TBATS permette alle stagionalità di cambiare lentamente nel tempo ed è in grado di modellare delle stagionalità di lunghezza non intera: per esempio, data una serie con osservazioni giornaliere, è possibile tenere conto degli anni bisestili usando una lunghezza di 365,25. D'altro canto, è molto lento con serie storiche lunghe.

Per scegliere il modello finale, TBATS costruisce in realtà diversi modelli (con diverse combinazioni di parametri), andando a preferire quello con il miglior AIC (Akaike Information Criterion). TBATS è adatto alla natura dei dati dell'analisi in questione, in quanto ci sono chiaramente più di una stagionalità da rimuovere.

3.2 K-Means per serie storiche

K-means è un metodo di classificazione non supervisionata largamente conosciuto. In questo documento verrà utilizzato nelle serie storiche (*Dotis-Georgiou, 2018*), nel contesto di anomaly detection (*Lima et al., 2010*). I metodi di clustering risultano utili quando metodi più tradizionali che usano threshold, come IQR, risultano inefficaci.

3.3 Isolation Forest

Isolation forest (*Liu et al., 2008*) è un algoritmo per identificare anomalie usando isolamento: a differenza dei normali metodi di anomaly detection (che definiscono un'anomalia come istanze non conformi alla "norma"), esso isola esplicitamente le anomalie usando alberi binari, senza dover applicare un intenso processo di profiling di tutte le istanze per determinare cosa è normale e cosa no. Insieme al k-means, si presta bene allo scopo di individuare outlier dei nostri dati.

3.4 SOM e super-SOM

Le Self-Organizing Maps sono una tipologia di reti neurali allenate utilizzando apprendimento non supervisionato per produrre una rappresentazione (di solito una mappa

bidimensionale) dello spazio dei dati in input. Ciò rende le SOM un valido strumento per l'analisi dei dati: in particolare sono molto utili per visualizzare i dati, e ad analizzare tramite clustering le relazioni esistenti tra le variabili del dataset. Le super-SOM (*Sivakkumaran, 2020*) sono una variante dell'algoritmo classico, la quale permette alle features in input di essere raggruppate in livelli differenti. La procedura di clustering avviene tramite misure di similarità computate separatamente su ogni layer, ponderata per il peso assegnato ad ognuno di questi, al fine di identificare il neurone "vincente" sul quale posizionare l'osservazione.

4 I dati

I dati provengono dagli uffici interni di Bicocca. Rappresentano i consumi energetici degli edifici U1 e U6 nell'arco temporale compreso tra 2018 e 2020 inclusi.

I file sono suddivisi per mese e presentano vari errori, sia di ripetizione che di formattazione errata; sono stati quindi corretti (quasi completamente in Python) ed uniti. Le entry hanno una granularità a livello del quarto d'ora, con dei missing values in tutto giugno 2020 in U6 e il 31 luglio 2020 in entrambi gli edifici. La feature di interesse è il consumo energetico attivo, misurato in kW, campionato nei 15 minuti considerati. Altri dati disponibili presenti sono la potenza massima e il consumo reattivo induttivo, ma sono stati scartati poiché ritenuti non necessari ai fini della ricerca.

Ai dati forniti sono stati poi integrati dati meteorologici giornalieri (*Archivio ilMeteo.it, Chinnam A, 2020*) di temperatura e umidità.

5 Analisi processo e trattamento dei dati

L'esposizione delle analisi segue l'ordine con la quale sono state eseguite, quindi parte dalla decomposizione, seguita da analisi descrittive e confronti tra edifici; quindi, l'identificazione delle anomalie e infine clustering sui tipi di consumo giornaliero.

5.1 Decomposizione

Dal punto di vista operativo, per prima cosa è stata ridotta la granularità del dato, dal quarto d'ora all'ora, tramite media oraria; è il calcolo ottimale in quanto permette di mantenere l'andamento, ma allo stesso tempo riducendo la rumorosità della serie e la quantità di osservazioni, permettendo di velocizzare l'oneroso processo di modellazione. Successivamente sono state identificate le stagionalità nei dati. Sono state osservate le serie storiche e i grafici di autocorrelazione parziale (PACF). Dai grafici (fig. 1), e a rigor di logica, si intuisce come ci sia un ciclo giornaliero (di notte i consumi sono ridotti) e uno settimanale (nel weekend i consumi sono ridotti).

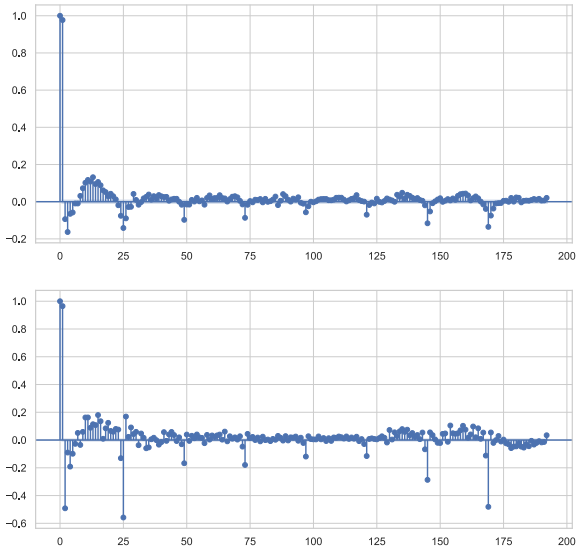

Fig. 1 – Grafici di autocorrelazione parziale, rispettivamente per U1 e U6.

Inoltre, osservando le serie storiche ad una granularità ancora inferiore (settimanale) si nota facilmente un andamento annuale (consumo estivo maggiore).

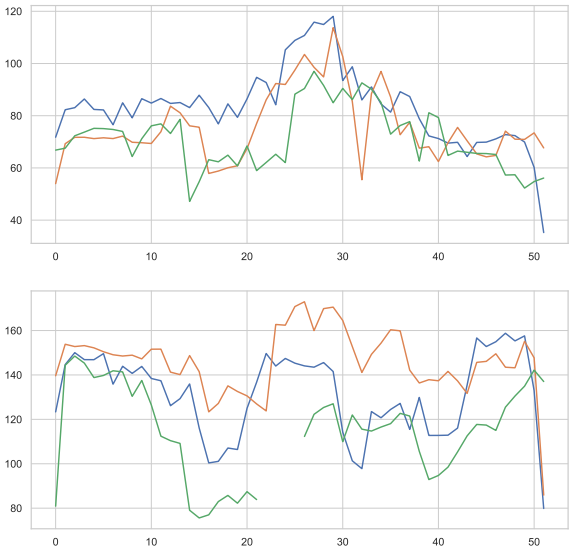

Fig. 2 – Serie storiche a livello settimanale, rispettivamente per u1 e u6.

Una volta identificate con certezza le stagionalità, è stata effettuata, con TBATS, una decomposizione su entrambe le serie, tramite il package R "forecast".

La decomposizione non è stata effettuata sull'intera serie storica, siccome ha chiaramente periodi con diverse strutture di stagionalità. La migliore suddivisione (che raggiunge un compromesso tra l'essere una possibile reale suddivisione dell'anno solare e il tentativo di avere la più possibile corretta modellizzazione) è quella per semestre universitario. Questa decisione è ottimale, in quanto i punti di stacco tra uno e l'altro (1° marzo e 1° settembre) riescono a suddividere periodi caldi e freddi (i quali hanno evidenti consumi differenti) e inoltre viene anche considerato fortuitamente in un nuovo semestre l'inizio della pandemia di Covid-19 del marzo 2020. Bisogna notare però che la serie inizia da gennaio 2018 e finisce a dicembre 2020, dunque i semestri corrispondenti saranno più brevi. Conseguenza di questa suddivisione è che non verrà presa in considerazione la stagionalità annuale, ma solo le restanti due.

Prima di eseguire l'algoritmo, c'è da considerare il problema dei missing values. Per i valori mancanti puntuali (le ore "mancanti" del giorno di cambio d'ora a marzo, in cui si passa dall'1:59 alle 3) viene effettuata una semplice interpolazione lineare, in quanto a quell'ora l'andamento è sempre costante; questo serve per fare in modo che tutti i periodi della stagionalità siano di numerosità 24 e 24*7 entry. Per quanto riguarda il mese di giugno 2020 nell'edificio U6, di cui mancano tutti i dati, è stata presa una media dei due anni precedenti, riadattata alla minore varianza del 2020, a cui è stato aggiunto infine un random noise. Ovviamente, essendo dei dati creati ad hoc, le anomalie che si presentano in questo mese sono da ignorare; il processo è utile solo per fare in modo che la modellazione trovi delle stagionalità coerenti e sensate.

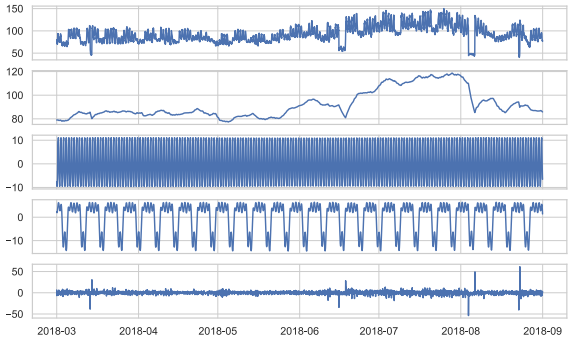

Fig. 3 – Esempio di decomposizione di una serie semestrale (prima in alto), con i relativi trend, due stagionalità e residui.

I risultati confermano che le stagionalità scelte sono quelle corrette: anche testando con valori differenti, con queste stagionalità l'AIC è maggiore.

Andando a vedere come è stata modellizzata la serie nel complesso, concatenando tutti i semestri, si nota come in U1 il trend sia discretamente pulito, tranne nel secondo semestre del 2019/2020, colpito maggiormente dal covid (dove quindi non c'è più una stagionalità fissa e, di conseguenza, l'andamento richiama molto le osservazioni iniziali).

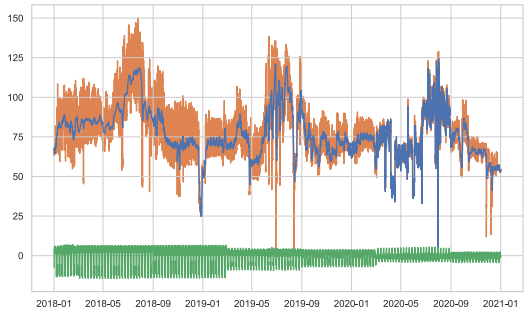

Fig. 4 – Confronto tra trend e serie iniziale (U1), con relativa stagionalità settimanale (in verde).

In U6 si ottengono risultati leggermente peggiori, dovuti probabilmente al fatto che la serie è molto instabile nel tempo, soprattutto nel 2018. Nel 2020 si ha lo stesso effetto precedentemente descritto in U1.

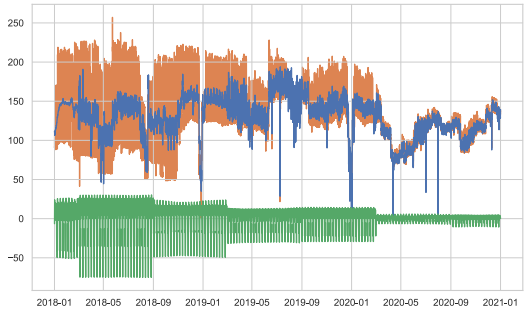

Fig. 5 – Confronto tra trend e serie iniziale (U6), con relativa stagionalità settimanale (in verde).

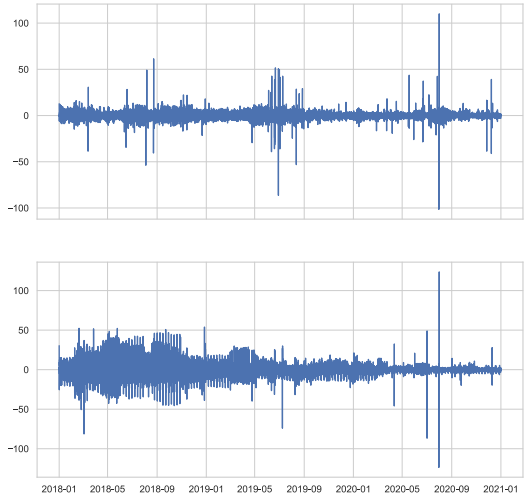

Fig. 6 – Residui totali, rispettivamente di U1 e U6.

I residui del consumo energetico relativo all'edificio U6 sono significativamente differenti di anno in anno. Nel 2018 presentano un'ampia varianza, mentre è inferiore nel 2019 e ancora meno nel 2020. Il trend e la stagionalità del 2018 non riescono a spiegare tutta l'informazione per via della presenza elevata di anomalie, che quindi ricadono nei residui, mentre nel 2019 il rumore diventa costante ed entra a far parte della stagionalità con un conseguente innalzamento della baseline (visibile nella serie storica). Questa differenza potrebbe suggerire che sono stati implementati dei metodi di consumo efficiente nel corso di quegli anni, che hanno permesso una diminuzione generale dei consumi. Al contrario i residui di U1 sono regolari ad eccezione del primo semestre del 2020.

5.2 Confronto edifici

Una volta compresa la struttura e l'andamento dei consumi nei diversi anni delle due diverse strutture, è stato deciso di procedere con l'analisi dei dati confrontando i due edifici, descrivendo delle statistiche descrittive.

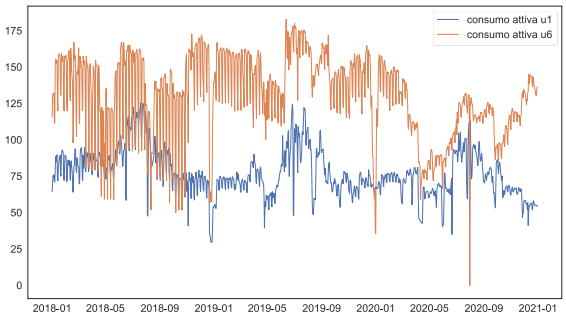

Fig. 7 – Serie storiche a media giornaliera

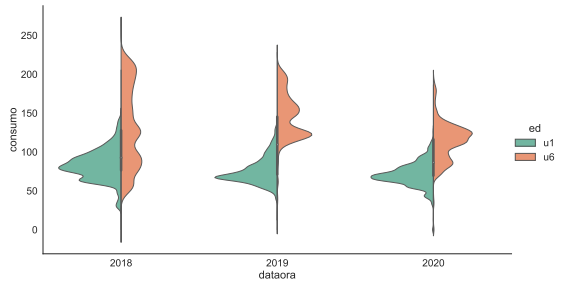

Fig. 8 – Violin plot che mostrano min-max-iqr-distribuzione.

Di primo impatto si nota che in U1 si ha un andamento più costante col passare degli anni.

Al contrario, in U6 la distribuzione è irregolare ed ha una media che varia nel tempo. Considerando poi le medie dei consumi secondo le partizioni scandita da ora legale e ora solare, si nota come i due edifici si comportano in maniera differente.

Edificio	Ora solare (media)	Ora legale (media)
U1	70.4529	81.3735
U6	138.4659	124.5742

Per confermare ciò, sono stati ricampionati i valori settimanalmente: le settimane con il consumo più elevato all'interno di ogni anno si verificano proprio nei periodi mostrati in precedenza; avviene conseguentemente anche per le settimane con consumi minori. Andando ad osservare la serie ad una diversa granularità (media mensile), è possibile visualizzare in maniera più marcata le differenze negli andamenti del consumo all'interno di un anno.

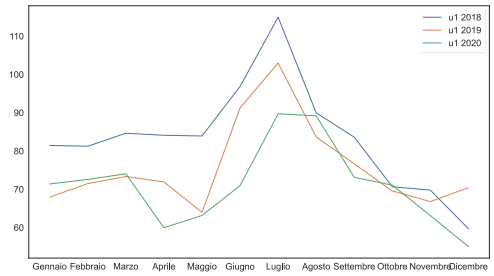

Fig. 9 – Consumo energetico U1 a media mensile.

In U1 (fig. 9) si conferma il picco di consumo prevalentemente nei mesi caldi, come luglio e agosto, andando a diminuire verso la fine dell'anno.

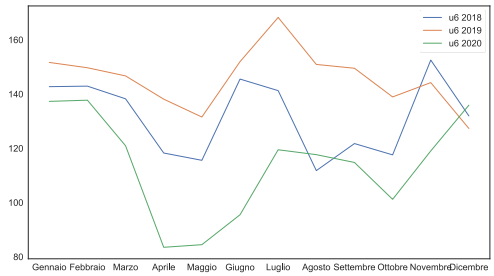

Fig. 10 – Consumo energetico U6 a media mensile.

Diversamente, nell'edificio U6 l'andamento non è facilmente definibile, in quanto i valori sono particolarmente influenzati dalla presenza delle sessioni d'esame come gennaio/febbraio, giugno/luglio e novembre.

È possibile, inoltre, verificare la presenza di correlazioni, in diversi anni. Considerando la media giornaliera per ogni anno divisa per stabile, viene calcolato il correlogramma (scatter matrix).

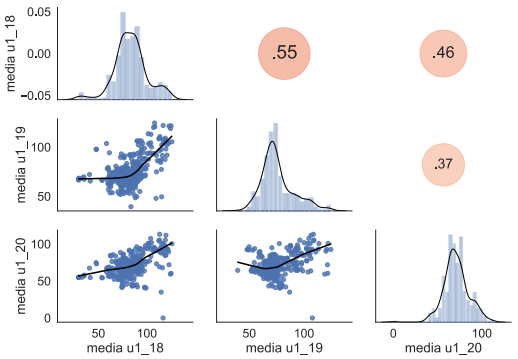

Fig. 11 – Correlogramma annuale per U1.

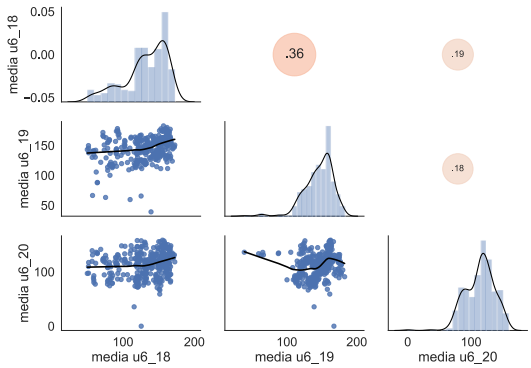

Fig. 12 – Correlogramma annuale per U6.

In nessuno degli edifici ci sono forti correlazioni; nonostante ciò, in U1 si verificano andamenti più costanti nel tempo, confermando le osservazioni sui violin plot.

5.3 Anomaly detection

A seguito delle osservazioni tratte sugli edifici, si può iniziare ad individuare le anomalie. Esse risiedono nei dati che non sono già spiegati da trend e stagionalità, ovvero nei residui, i quali vengono calcolati tramite la decomposizione

TBATS. Su questi dati, presi in input, sono stati utilizzati tre diversi metodi per la rilevazione delle anomalie: il range interquartile (vengono presi in considerazione le osservazioni fuori dal range 5-95, anziché il classico 25-75, poiché si assume di avere un numero inferiore di dati anomali) e due metodi di clustering, K-Means e Isolation Forest. Anche i record individuati da questi ultimi due metodi sono stati limitati, ipotizzando che nei dati fosse presente l'1% di anomalie.

Inoltre, al fine di ottenere risultati migliori, non sono stati presi i risultati singoli di ogni metodo, ma sono stati selezionati i record tramite diverse tecniche: intersezione, intersezione almeno 2 su 3 e unione. Dopo un confronto, la scelta finale è ricaduta sull'intersezione, in quanto più conservativa: se un'anomalia compare in tutte e tre i metodi, significa che si ha più confidenza che lo sia effettivamente. Infine, sono state selezionate solo le anomalie con valore del residuo positivo, ovvero quelle che indicano momenti di eccessivo consumo piuttosto che di risparmio.

Gli algoritmi dei due metodi di clustering sono stati presi dalla libreria *sklearn* per Python. I risultati ottenuti, con relativa discussione sono presentati nel *capitolo 6.2*.

5.4 Clustering sui consumi giornalieri

Il processo descritto in seguito è relativo all'analisi dei consumi nel 2018 nell'edificio U1.

5.4.1 Preparazione dei dati

Dalla serie storica, vengono ricavati i valori di consumo raggruppati per giorno e per ognuno di questi si considera la fascia oraria 7-19, essendo l'indagine orientata all'osservazione dei consumi nelle ore di maggiore attività universitaria. Inoltre, dei 48 valori interni all'arco temporale considerato (uno per quarto d'ora), vengono estratti i 12 punti corrispondenti al quarto d'ora 00, in modo da poter effettuare un'analisi non eccessivamente granulare e rendere quindi più semplice l'interpretazione finale. Sono stati poi raccolti i valori giornalieri di temperatura media ed umidità, in quanto si vuole osservare se a simili andamenti di consumo si associano anche simili condizioni meteorologiche. Al fine di giustificare l'utilizzo di queste due variabili è stata verificata a priori l'esistenza di un nesso con il consumo totale giornaliero, tramite una semplice regressione lineare, i cui risultati sono mostrati in *appendice*.

Ogni giornata viene quindi rappresentata attraverso i seguenti attributi:

1. *CO*: Caratterizza il livello di consumo campionato dalle ore 7 alle ore 19; è una lista di 12 valori: $[\, co_7,\ co_{8,}\ \ldots,\ co_{19}\,]$
2. *COP*: Rappresenta l'impatto percentuale di consumo per ogni ora, dalle 7 alle 19; è una lista di 12 valori: $\left[\ \frac{co_7}{sum\ (CO\,)}\ ,\ \ldots\ ,\ \frac{co_{19}}{sum\ (CO\,)}\ \right]$
3. *COD*: Rappresenta la differenza di andamento più o meno crescente tra le ore, dalle 7 alle 19; è una lista di 12 valori: $[\, co_7 - co_6\ ,\ \ldots\ ,\ co_{19} - co_{18}\,]$
4. *T*: Temperatura media
5. *U*: Umidità

5.4.2 Costruzione SOM

A causa della diversa dimensionalità tra attributi, si è scelto di utilizzare come algoritmo di clustering la Super-SOM. La prima operazione consiste nella definizione dei layer: ai tre attributi vettoriali CO, COP e COD è assegnato un layer ciascuno; un quarto layer è poi ottenuto accorpando i dati meteo, ovvero T e U. Successivamente, viene definita la struttura della mappa: data la scarsa numerosità di osservazioni in input (365 giorni), si è scelto, dopo alcuni tentativi, di impostarla a 5x5, con celle esagonali e topologia toroidale.

Per quanto riguarda l'addestramento della rete, sempre procedendo tramite trial-and-error, è stato scelto il numero di epoche e i valori dei pesi associati ai layer. Il numero di epoche scelto è

pari a 300: a causa della poca numerosità di osservazioni, proseguire oltre le 300 iterazioni non porta a miglioramenti degni di nota. La scelta dei pesi legati ai livelli si è basata su un compromesso tra importanza di significato del layer e distribuzione del peso associato a questo sulle features che contiene. È importante specificare che, assegnando un peso ad un layer, esso viene distribuito in uguali porzioni sulle feature in esso contenute. Per questo motivo al layer che riguarda il meteo, contenente sole due features, è stato assegnato il minor peso, pari al 10%. Al layer CO è assegnato un peso del 20%, in quanto la disposizione delle osservazioni su questo layer converge più rapidamente e facilmente rispetto agli altri; infine, ad entrambi gli attributi COP e COD, ritenuti i più carichi di informazioni utili alla distinzione dell'andamento di consumo, è stato assegnato un peso del 35% (quasi il 3% per ciascuna delle 12 feature interne ad ognuno di questi).

5.4.3 Cluster ottenuti e interpretazione dei risultati

Il package *kohonen.r* (*Buydens e Wehrens, 2007*), permette di visualizzare la mappa ottenuta, offrendo una visualizzazione degli attributi delle osservazioni interne ad ogni neurone, per ognuno dei layer specificati. Da queste, è possibile eseguire delle prime interpretazioni e raggruppare zone della rete in cluster differenti. Sempre a causa della numerosità delle osservazioni e della conseguente dimensione della mappa, si è scelto un numero di cluster non elevato, pari a 3. Come metodo di cluster sui vettori dei codebook della mappa viene utilizzato lo hierarchical clustering, i cui risultati, a confronto con quelli ottenuti tramite k-means, sono stati ritenuti leggermente migliori, ad eccezione dell'analisi in U1 nel 2019, dove invece è stato utilzzato quest'ultimo.

Di seguito vengono commentati i risultati ottenuti relativamente all'anno 2018 in U1. Le mappe vengono rappresentate dividendo già i nodi per cluster di appartenenza. Le giornate si dispongono in modo abbastanza omogeneo sui nodi.

I cluster sono i seguenti:

1. *Cl. 1*, blu, rappresenta il 32.3 % dei giorni
2. *Cl. 2*, verde, rappresenta il 45.7 % dei giorni
3. *Cl. 3*, rosso, il restante 22%.

Legenda layer 1, 2 e 3	Layer 4
h.7 h. 8 h.9 h.10 h.11 h.12 h.13 h.14 h.15 h.16 h.17 h.18 h.19	temperatura umidità

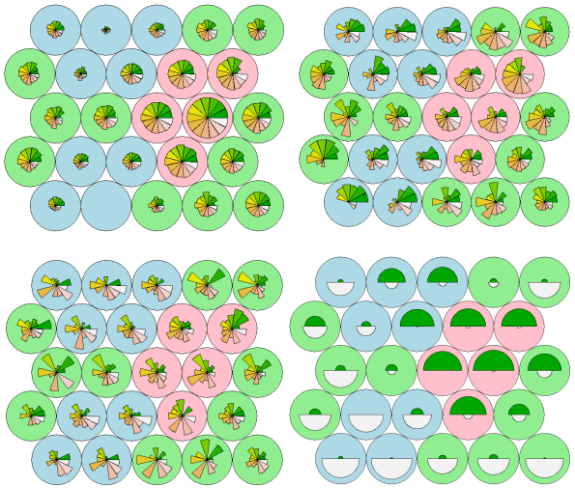

Fig. 13 – Mappe ottenute dai quattro layer, in ordine dall'alto in basso e da sinistra a destra: CO, COP, COD e T.

Il layer CO (1) spiega come il livello di consumo orario varia tra diverse giornate. Sebbene tra livelli differenti, il consumo si mantiene costante all'interno degli stessi nodi e le feature hanno grandezze simili tra loro.

Il layer COP (2) mostra quali sono le ore che impattano maggiormente sul consumo giornaliero totale. Si notano distinzioni tra i nodi, dove prevalgono i consumi mattutini o pomeridiani o solamente quelli a inizio e fine giornata.

Il layer COD (3) mostra i maggiori incrementi e decrementi di consumo. Nel cluster 1 prevalgono incrementi a fine attività pomeridiana, il secondo ha un andamento più alternato, mentre il terzo ha pochi, ma grossi incrementi.

Il layer T+U (4) mostra le due caratteristiche meteo delle diverse giornate e la loro disposizione nei cluster. Si nota il terzo cluster molto più coeso degli altri due, i quali sembrano essere più sovrapposti.

Un'interpretazione più approfondita dei risultati viene discussa nel *capitolo 6.3*.

6 Risultati

6.1 Risultati confronto edifici – Test

Di seguito vengono mostrati i risultati di test statistici sul confronto dei consumi negli edifici e negli anni. I risultati sono di supporto ai grafici mostrati nel *capitolo 5.2*. I test sono stati realizzati tramite il package *stats* della libreria python *SciPy* (*SciPy – v. 1.7.0*). I risultati dei test, con relativi livelli di confidenza osservati, sono mostrati in *appendice*.

6.1.1 Kolmogorov-Smirnov

Kolmogorov-Smirnov è un test non parametrico per la verifica della forma delle distribuzioni campionarie. L'ipotesi nulla H_0 indica che le due distribuzioni sono "sufficientemente simili". Sono state svolte tutte le diverse combinazioni di parametri, cioè sulle componenti di trend e stagionalità di ciascun edificio, nei tre diversi anni, e nessuna di esse ha fornito un p-value tale da indicare un non rifiuto; al contrario, tutte le ipotesi vengono fortemente rifiutate con p-value=0.

6.1.2 Chi quadro per l'indipendenza

Il test del chi quadro definisce come ipotesi nulla H_0 il fatto che i campioni in analisi sono indipendenti. Sono stati preparati i dati, per ogni singolo anno partendo da un singolo edificio, campionati per ora. Vengono create cinque fasce di valori di egual dimensione e vengono contati quanti record cadono in ciascuna fascia, dividendo la serie in trimestri. I test eseguiti hanno portato ad un forte rifiuto dell'ipotesi nulla, indicando quindi una forte dipendenza tra i quattro trimestri all'interno di ciascun anno ed edificio.

6.1.3 t-test per l'uguaglianza delle medie nei diversi anni

Il t-test è usato per determinare se due gruppi hanno ugual media. La statistica t segue una distribuzione t di student con $n_1 + n_2 - 2$ gradi di libertà. Confrontando le medie di consumi orarie per ogni combinazione di tutti gli anni all'interno dei singoli edifici, ed anche tra i due edifici fissato l'anno (quindi come nel test precedente), si è concluso che non ci sono differenze ad alcun livello di significatività.

6.2 Risultati anomaly detection

Le anomalie riscontrate nei consumi energetici dell'edificio U1 (fig. 14) sono presenti maggiormente nel 2018, mentre negli anni successivi sono ridotte, in particolare circa del 30% nel 2019 e del 50% nel 2020. Inoltre, le anomalie si distribuiscono soprattutto nei mesi caldi, da maggio ad agosto, periodo nel quale vengono accesi i condizionatori e quindi si ha un maggiore dispendio energetico irregolare (presumibilmente dettato dalla temperatura), e nelle ore di apertura dell'edificio, quando appunto si ha il maggiore flusso di persone. Il giorno della settimana che presenta più anomalie è il lunedì, seguito dai successivi in maniera decrescente fino alla domenica.

Fig. 14 – Anomalie sulla serie storica iniziale in U1.

A differenza di U1, le anomalie riscontrate nei consumi energetici dell'edificio U6 (fig. 15) sono

presenti quasi esclusivamente nel 2018, con una piccolissima frazione nel 2019 e 2020. Inoltre, sono distribuite quasi normalmente rispetto ai mesi, con maggio in pole position con più di 35. Da notare però, a differenza di U1, non sono presenti impianti di raffrescamento ad energia elettrica, e quindi le anomalie non sono facilmente spiegabili. In questo edificio, il giorno della settimana dove le anomalie sono più frequenti è il sabato mentre gli altri giorni si equidistribuiscono, poichè, a differenza di altri edifici, U6 rimane aperto fino a metà giornata. Infine, sono maggiormente presenti durante la mattina presto, ossia orari in cui generalmente c'è un'alta affluenza oppure in cui vengono attivati determinati servizi temporizzati, infatti il picco delle 4 di mattina è direttamente preceduto e seguito da momenti di inattività.

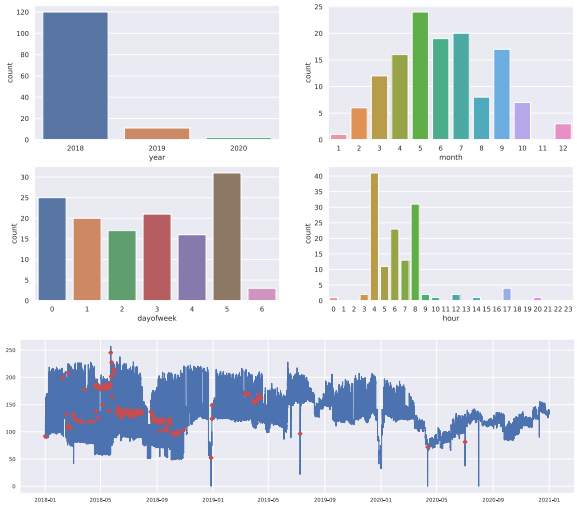

Fig. 15 – Anomalie sulla serie storica iniziale in U6.

Andando a filtrare per le modalità più frequenti di ciascuna categoria (anno, mese, giorno della settimana, ora) precedentemente descritta si possono fare osservazioni su diversi aspetti.

In U1:

- Sebbene il 2018 sia l'anno con più anomalie, analizzando il mese con la frequenza più elevata, ovvero **giugno**, scopriamo che queste compaiono maggiormente nel 2019, nello specifico il lunedì e il mercoledì costantemente durante tutto l'orario lavorativo.
- Le anomalie rilevate nel **2018** sono distribuite irregolarmente nei mesi e avvengono praticamente solo in settimana, quasi mai nei weekend, molto spesso alle 8 del mattino.
- Il **lunedì** presenta anomalie in tutti e tre gli anni, maggiormente nei mesi estivi. Inoltre, avvengono sempre di giorno, spesso di mattina.
- Le anomalie registrate alle **ore 8** di mattina sorgono spesso di lunedì e mercoledì. Inoltre, sono quasi esclusivamente nel 2018 e molto frequenti nei mesi di settembre e novembre.

In U6:

- Alle ore 8 di **maggio** 2018 sono presenti moltissime anomalie (in relazione al numero totale di anomalie riscontrate), distribuite in tutti i giorni della settimana tranne la domenica.
- Alle **ore 4** di mattina dei mesi estivi, soprattutto luglio e settembre, sono presenti molte anomalie, distribuite costantemente nei giorni lavorativi della settimana.
- Nel 2018, i **sabati** alle ore 6 presentano molte di anomalie, distribuite in maniera costante per tutti i mesi dell'anno tranne gennaio e novembre.

Inoltre, plottando le anomalie sulla serie storica a media giornaliera, si nota che in U1 le anomalie influiscono molto sui consumi giornalieri. Invece, in U6 questa cosa non sembra succedere: dove si presentano delle anomalie, la serie rimane sempre allo stesso livello dei periodi di quando non ce ne sono.

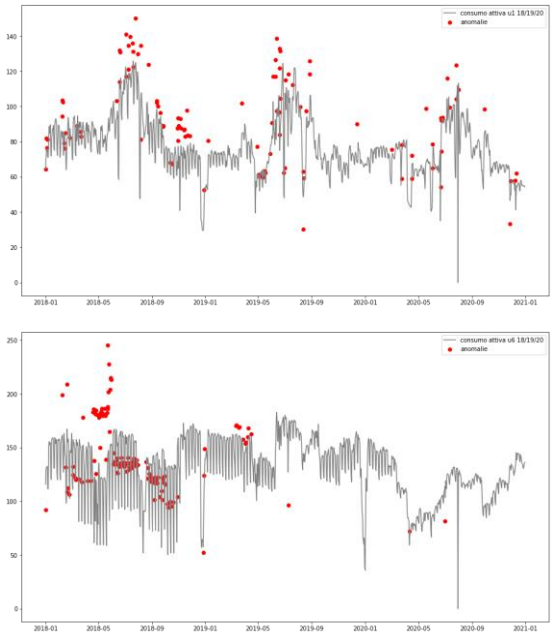

Fig. 16 – Anomalie sulle serie storiche sulle medie giornaliere, in U1 e U6 rispettivamente.

6.3 Risultati SOM

Dalla SOM ottenuta, per ogni nodo si ricavano i vettori di features dei relativi codebook, che verranno utilizzati per calcolare il prototipo di ogni cluster nei diversi layer. Al loro interno, per ogni cluster, tramite i codebook dei suoi nodi si computa una media delle features, ponderata per numerosità di osservazioni appartenenti al nodo. I risultati ottenuti mostrano le differenze tra caratteristiche tipiche dei consumi nei cluster in modo più interpretabile, permettendo di distinguere i diversi aspetti sopra descritti tramite la lettura delle mappe. Di seguito vengono mostrati i prototipi dei due edifici U1 e U6 negli anni 2018 e 2019, paragonandone man mano le caratteristiche. L'analisi è stata condotta solamente su questi due anni a causa della mancanza di dati riguardanti temperatura e umidità nel 2020 e inoltre perché orientata alla classificazione dei giorni nel periodo di attività universitaria. Essendo il 2020 stravolto dal covid, un paragone con gli altri periodi risulterebbe ambiguo; di conseguenza, non sarebbe stato opportuno porre sullo stesso piano gli eventuali risultati ottenuti con quelli degli anni precedenti.

In *appendice* sono mostrati i prototipi denormalizzati, i quali permettono di associare alle differenze qui mostrate, anche dei valori reali.

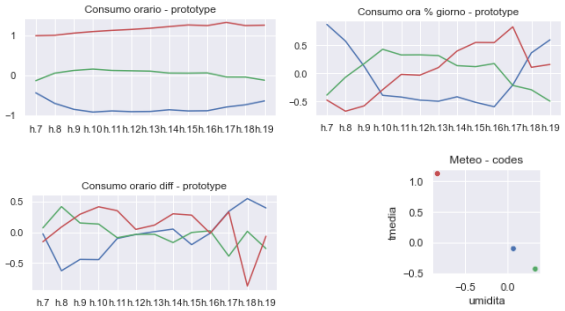

Fig. 17 – Prototipi U1 2018

Il consumo orario, non varia all'interno dei cluster, le tre tipologie sono ben separate a livello di consumo. Nel cluster 1, contenente le giornate di consumo minore, l'impatto maggiore è infatti dato dagli orari di inizio e fine giornata. Il cluster 2 è quello delle giornate di consumo medio complessivo, l'impatto maggiore è dato dalla fascia centrale della giornata, in particolare in mattinata. Il cluster 3 è invece quello dai consumi maggiori, l'impatto orario in queste giornate è scarso la mattina e cresce col passare della giornata, con un picco nel pomeriggio intorno alle 17. Questo cluster è anche quello che si differenzia maggiormente per i due valori metereologici, con alte temperature e bassi valori umidità.

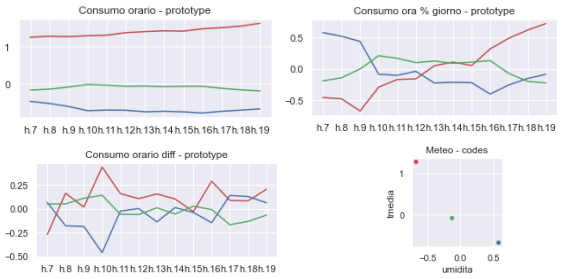

Fig. 18 – Prototipi U1 2019

Come l'anno precedente, si rilevano 3 fasce distinte di consumo, ma a differenza del 2018, il cluster di consumo minimo e intermedio sembrano essere più simili tra loro. Si vede come per il cluster 2 il consumo orario, assoluto e

percentuale, sia minore rispetto al 2018, anche negli orari mattutini. Rimane invece il terzo cluster, legato a giorni con maggiore temperatura e con consumi elevati, in particolare in fascia pomeridiana.

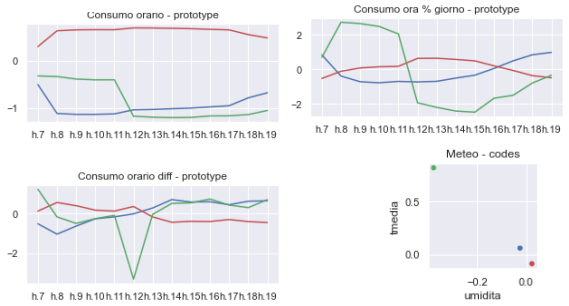

Fig. 19 – Prototipi U6 2018

A differenza di U1, nel 2018 in U6 i gruppi ben distinguibili, in consumo assoluto, sono due, il cluster 1 e 3. I due cluster rappresentano rispettivamente i consumi minori e maggiori, questi poi si mantengono abbastanza costanti tra le ore, come si può vedere dai prototipi di consumo percentuale e differenziale. Il cluster 2 rappresenta invece i giorni con un consumo mediocre la mattina, in particolare fino alle ore 12, dopo le quali si ha un consumo anche inferiore a quello del cluster 1. Sempre a differenza di U1, il consumo dell'edificio U6 non sembra essere troppo influenzato dalle alte temperature: il cluster 3 infatti non presenta particolari valori di temperatura o umidità.

Fig. 20 – Prototipi U6 2019

In U6 nel 2019 tutti e tre i cluster si distinguono bene in quanto a fascia di consumo. Il primo cluster rappresenta sempre le giornate dal minor consumo, dove i grossi costi si hanno a inizio e fine giornata. Il secondo cluster rappresenta consumi medio-elevati dove l'impatto si ha un impatto costante ed elevato la mattina e il primo pomeriggio, con particolari picchi alle 8 e le 12. Il terzo cluster rappresenta i consumi ancor più elevati, dove le ore di maggior consumo iniziano più tardi (fine mattinata intorno alle 10) ma si protraggono almeno fino alle 19. Infine, a differenza dell'anno precedente, il cluster con i consumi maggiori è anche quello che spicca maggiormente per le alte temperature.

Una volta distinte le proprietà dei cluster in entrambi gli edifici e anni, si può ora osservare dove queste differenti tipologie di giornata cadono maggiormente. Di seguito vengono mostrate alcune distribuzioni dei cluster paragonate tra edifici e anni.

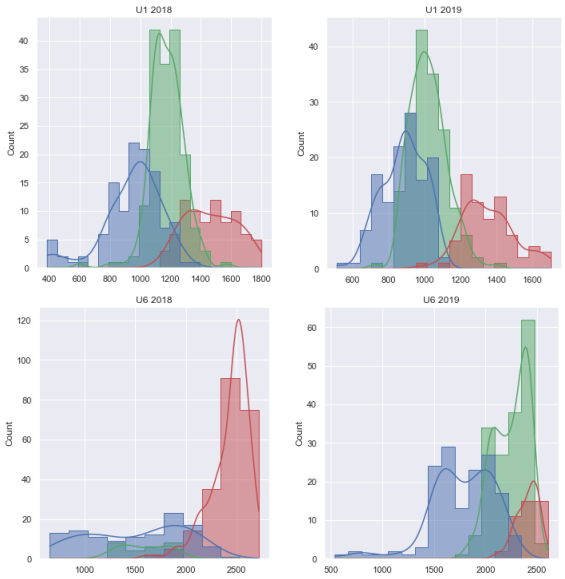

Fig. 21 – Distribuzione cluster per consumo giornaliero totale

In U1, in entrambi gli anni, i cluster si distribuiscono similmente sul consumo, ma nel 2019 si ha un passaggio di alcuni giorni dal cluster 2 al cluster 3, di consumo minore. In U6, nel 2018 si ha invece un cospicuo numero di giorni di consumo particolarmente elevato, mentre nel 2019 si ha un miglioramento, con un aumento delle giornate appartenenti al cluster 1 e 2, pur essendo quest'ultimo caratterizzato anch'esso prevalentemente da consumi elevati.

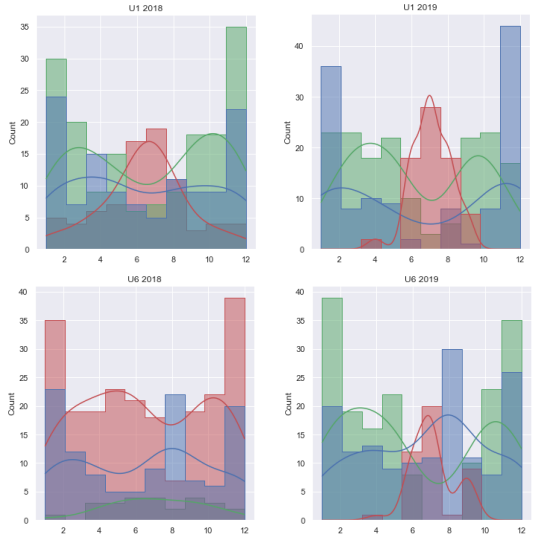

Fig. 22 – Distribuzione cluster per mese

In U1, il cluster 3 dei consumi elevati, cattura la maggioranza delle giornate estive, mentre gli altri due cluster si distribuiscono in modo simile tra mesi. In particolare, si nota però come nel passaggio tra 2018 e 2019, il cluster 3 sia più ristretto al periodo estivo, provocando quindi un aumento del numero di appartenenti al cluster 2. Nel 2018 in U6 la presenza del cluster 3 ad eccezione di agosto, è dominante e costante nei mesi. Nel 2019 la distribuzione è simile a quella di U1, dove il gruppo di maggiore consumo cade in estate, ma a differenza di questa si ha una maggiore sovrapposizione tra i cluster 1 e 2.

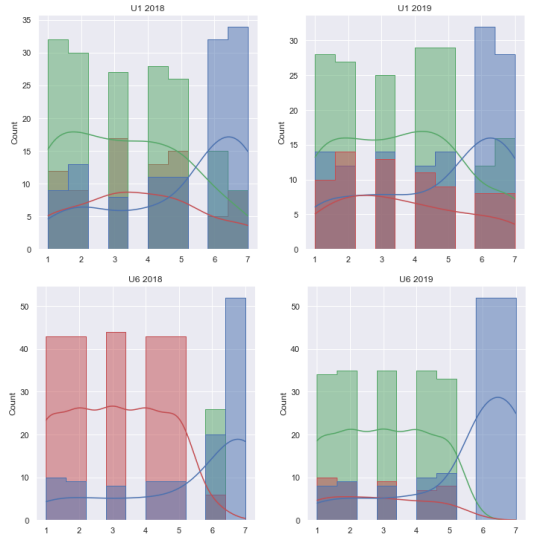

Fig. 23 – Distribuzione cluster nella settimana

In U1 nel 2018 e 2019 la situazione è simile, la maggioranza dei giorni dei giorni lavorativi appartiene al cluster di consumo medio, si nota anche qui, una diminuzione di giorni del cluster 3 tra 2018 e 2019. I weekend in entrambi gli anni sono rappresentati maggiormente dal cluster 1. In U6 nel 2018, i giorni di cluster 3 sono di gran lunga i più presenti (e in modo costante) nelle giornate lavorative. Si vede come il cluster 2 catturi interamente ed esclusivamente i sabati, mentre al cluster 1 di consumo minimo appartengono in grande maggioranza le domeniche. Nel 2019 si ha invece un miglioramento, dove i consumi più elevati vengono ridotti, a favore di consumi leggermente minori (cluster 2). Il cluster di consumo minimo cattura questa volta l'intero weekend e come nel 2018, mantiene un minimo di presenza (al pari dei giorni di maggiore consumo) anche nei giorni lavorativi.

7 Conclusioni, limiti e possibili sviluppi

7.1 TBATS

Limitazione della decomposizione tramite TBATS è il fatto che la modellazione su dati rumorosi e mancanti come quelli dei consumi energetici è problematica e ha dovuto subire dei workaround importanti, come la ricostruzione di giugno 2020. Ipoteticamente, sarebbe ideale implementare una decomposizione più performante su questo tipo di dati.

7.2 Anomalie

Le anomalie osservate hanno mostrato degli andamenti comuni che possono dare input ad analisi più approfondite alla ricerca di informazioni più specifiche.

In U1 le anomalie si presentano con ricorrenza nei mesi estivi, in concomitanza con le alte temperature, prevalentemente nelle ore di maggior affluenza. Il tutto riporta quindi a cause di natura metereologica dovute al caldo. In U6 le anomalie sono state rilevate quasi solo nel 2018,

hanno caratterizzato da un ampio intervallo di consumo, nelle prime ore della giornata. Una possibile spiegazione può essere, data la grossa dimensione dell'edificio, la presenza di accensioni temporizzate o controlli (come back-up, pulizia, etc…).

I metodi utilizzati si sono dimostrati, nel complesso, utili e hanno generato risultati esplicativi e sensati. L'utilizzo di tecniche più complesse, mediante Deep Learning, potrebbe essere un ulteriore sviluppo al fine di migliorare ulteriormente i risultati ottenuti. Uno dei problemi maggiori è determinato principalmente dai dati in input alle varie tecniche di anomaly detection. Infatti, le anomalie trovate derivano direttamente dai residui ottenuti tramite decomposizione, ed è quindi attribuibile a quest'ultima qualsiasi tipo di imprecisione successiva.

Un'ulteriore analisi, che potrebbe aiutare nella comprensione dei dati rilevati, è il confronto con una base di conoscenza degli eventi (es. hackathon, conferenze, fiere, …), che generano dei carichi anormali rispetto al generale andamento dei consumi.

7.3 SOM

Attraverso i cluster e i relativi prototipi ricavati sono stati analizzati i diversi pattern di consumo nelle giornate e osservati due principali aspetti:

1. I valori dei prototipi stessi, i quali permettono di distinguere in che modalità avviene il consumo nell'arco della giornata, individuando così all'interno di questa le principali fasce orarie critiche in cui tipicamente avvengono i consumi maggiori.
2. La distribuzione dei cluster su diversi periodi temporali, al fine di poter notare quali possono essere i periodi in cui si ha con più probabilità un certo tipo di andamento di consumo piuttosto che un altro.

I valori ottenuti e mostrati riguardanti l'analisi dei consumi sui due edifici permettono di distinguere giorni caratterizzati da andamenti diversi del consumo, sia per quanto riguarda il livello generale, sia per valori di decremento e incremento e sia per l'impatto di ogni ora nel complesso della giornata. Ciò nonostante, è possibile che la scelta degli attributi a monte dell'analisi non sia adeguata al fine di osservare differenze più nascoste; ad esempio l'analisi del consumo in U6 nel 2018 individua poche differenze tra mesi e tra giorni (lavorativi) della settimana. Tuttavia sarebbe comunque possibile svolgere analisi più specifiche e mirate sulla distribuzione dei cluster nei periodi temporali, ad esempio distribuzione dei cluster per ogni settimana all'interno dei mesi o ancora distribuzione dei cluster per i 7 giorni settimanali all'interno di ogni mese.

Anche in questo caso, come per le anomalie, risulterebbe utile associare le diverse attività universitarie ai cluster ottenuti. Questo potrebbe essere vantaggioso per un esperto di dominio, il quale potrebbe esprimere sia un parere su quali siano i migliori giorni e/o orari per effettuare diversi tipi di attività, diminuendo quindi il rischio di provocare grossi picchi di consumo energetico, sia individuare gli eventi o gli ambiti che provocano i maggiori sforzi energetici, permettendo di conseguenza un intervento più accurato.

8 Riferimenti bibliografici

1. Buydens, Lutgarde M. C. and Wehrens, Ron (2007). *Self-and Super-organizing Maps in R: The kohonen Package.* Disponibile su http://dx.doi.org/10.18637/jss.v021.i05
2. Chinnam, A. (2020). *Milan Weather Data [Dataset]*. Zenodo. Disponibile su https://doi.org/10.5281/ZENODO.3992354.
3. De Livera, Alysha M. and Hyndman, Rob J. and Snyder, Ralph (2009). *Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing.* Disponibile su https://doi.org/10.1198/jasa.2011.tm09771.
4. Dotis-Georgiou, Anais (2018) *Why Use K-Means for Time Series Data? (Part One).* Disponibile su https://www.influxdata.com/blog/why-use-k-means-for-time-series-data-part-one/. Visualizzato il 02/06/2020.
5. GreenMetric (2020). *Ranking by Country 2020 – Italy*. Disponibile su https://greenmetric.ui.ac.id/rankings/ranking-by-country-2020/Italy. Visualizzato il 20/05/2021.
6. ilMeteo s.r.l. (2021). *Archivio Meteo Milano*. Disponibile su https://www.ilmeteo.it/portale/archivio-meteo/Milano. Visualizzato il 23/05/2021.
7. Lima, Moisés F. and Zarpelão, Bruno B. and Sampaio, Lucas D. H. and Rodrigues, Joel J. P. C. and Abrão, Taufik and Proença, Mario Lemes (2010). *"Anomaly detection using baseline and K-means clustering"*, SoftCOM 2010, 18th International Conference on Software, Telecommunications and Computer Networks, pag. 305-309. Disponibile su https://ieeexplore.ieee.org/abstract/document/5623690.
8. Liu, F. T. and Ting, K. M. and Zhou, Z.-H. (2008). *Isolation Forest*. 2008 Eighth IEEE International Conference on Data Mining. 2008 Eighth IEEE International Conference on Data Mining (ICDM), pag. 413-422. Disponibile su https://doi.org/10.1109/ICDM.2008.17.
9. Sivakkumaran, Lakshminarayanan (2020). *Application of Self-Organizing Maps on Time Series Data for identifying interpretable Driving Manoeuvres.* Disponibile su https://etrr.springeropen.com/track/pdf/10.1186/s12544-020-00421-x.pdf.
10. SciPy v. 1.7.0. (2020). *Statistical functions (scipy.stats)*. Disponibile su https://docs.scipy.org/doc/scipy/reference/stats.html. Visualizzato il 18/06/2021.
11. Università degli Studi Milano-Bicocca (2019). *Budget Unico di Ateneo - Esercizio 2019.* Disponibile su https://www.unimib.it/sites/default/files/bilancio_unico_di_ateneo_di_previsione_annuale_con_allegati_esercizio_2019_0.pdf. Visualizzato il 20/05/2021.
12. Università degli studi di Milano-Bicocca (2019). *L'Università di Milano – Bicocca sottoscrive la Global Climate Emergency Letter*. Disponibile su https://www.unimib.it/news/luniversita-milano-bicocca-sottoscrive-global-climate-emergency-letter. Visualizzato il 20/05/2021.
13. Università degli studi di Milano-Bicocca (2021). *Bicocca: un'Università in cammino verso la sostenibilità.* Disponibile su https://www.unimib.it/ateneo/bicocca-cammino-verso-sostenibilita. Visualizzato il 20/05/2021

9 Appendice

9.1 Regressione lineare temperatura-umidità

Regressione lineare consumo giornaliero U1 2018 con temperatura media e umidità, autocorrelazione corretta tramite metodo GLS.

Significatività Modello (Pr > F)	< 0.0001
R^2 corretto	0.82
Valore Durbin-Watson autocorrelazione	1.64
Coefficiente per temperatura media	79.05
Coefficiente per umidità	1.53

Dal modello si vede come la temperatura sia legata al consumo, meno l'umidità, anche se viene comunque mantenuta essendo un fattore importante nella percezione del clima e della temperatura.

9.2 Test Kolmogorov-Smirnov per il confronto delle distribuzioni

		test	p-value
U1-U6	2018	-111.11	0
	2019	-250.31	0
	2020	-170.81	0
U1	2018-19	52.4	0
	2019-20	25.36	5.83E-137
	2018-20	69.25	0
U6	2018-19	-35.23	2.09E-254
	2019-20	87.61	0
	2018-20	34.56	1.85E-245

9.3 Test Chi-Quadro per l'indipendenza

		test	p-value
U1	2018	3986.84	0
	2019	2597.92	0
	2020	3143.34	0
U6	2018	491.45	1.46E-97
	2019	1179.81	3.86E-245
	2020	4062.26	0

9.4 Prototipi *"denormalizzati"*

Per ogni nodo, media delle features dei suoi appartenenti, e per ognuno dei tre cluster, media delle features dei nodi, ponderata per numerosità delle osservazioni di questi.

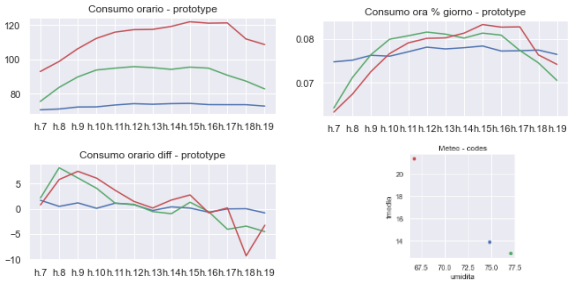

Fig. 24 – Prototipi U1 2018

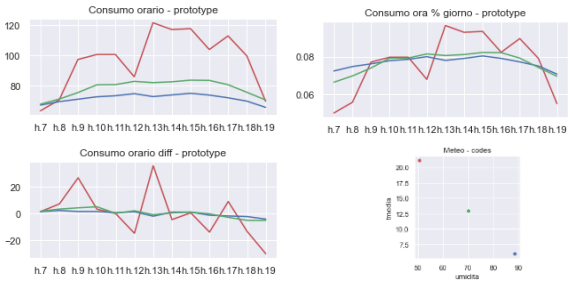

Fig. 25 – Prototipi U1 2019

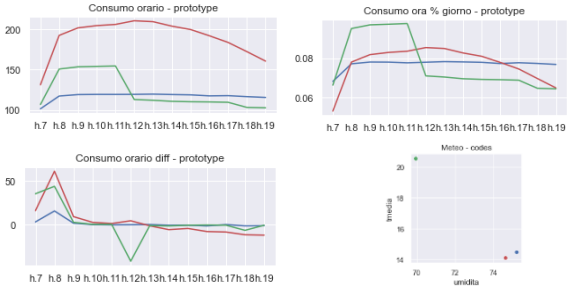

Fig. 26 – Prototipi U6 2018

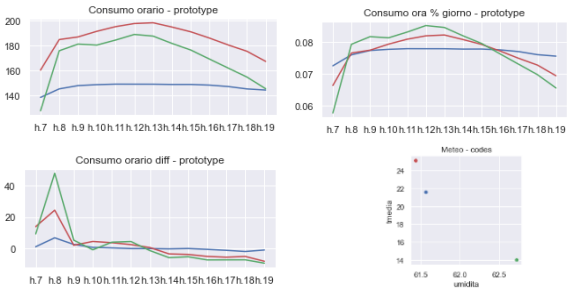

Fig. 27 – Prototipi U6 2019