Università degli studi di Milano–Bicocca
Scuola di Economia e Statistica

Corso di Laurea in
Scienze Statistiche ed Economiche

Analisi dei consumi energetici degli edifici dell'Università degli studi di Milano-Bicocca

Relatore: Dott. Tommaso Rigon

Tesi di laurea di:
Marco Carrettoni
Matricola N. 851292

Anno Accademico 2021/2022

Indice

Ringraziamenti

Ringrazio Tommaso Rigon, relatore del presente elaborato, per la sua disponibilità e per il tempo dedicatomi durante l'intera stesura.

Capitolo 1

Introduzione

Il tema della gestione energetica, dell'analisi e della previsione dei consumi è di grande attualità. In questo contesto il loro monitoraggio porta ad evidenti vantaggi operativi nell'ambito aziendale.

L'energia elettrica e termica fanno sempre più parte del costo vivo sostenuto dalle aziende per poter produrre ed essere competitive sul mercato. Un'adeguata verifica dei consumi, e un'analisi oculata dell'efficienza energetica, portano ad avere maggior controllo su numerosi aspetti dell'azienda, in primis la produttività e la sicurezza.

Lo scopo del monitoraggio del consumo di energia all'interno di un'azienda è innanzitutto ottenere gli indicatori KPI (*Key Performance Indicator*) in modo da capire i consumi energetici di ogni utenza utilizzata, sia essa energia elettrica, gas metano o vapore. Da questa prima analisi si potranno sviluppare delle azioni correttive e rendere l'industria più competitiva. Questa attenzione ai consumi nel mondo aziendale consente di:

- effettuare un'ottimizzazione delle risorse disponibili;
- ricalibrare i turni di lavoro;
- scegliere i contratti di fornitura energetica più profittevoli.

Per raggiungere l'obiettivo della massima efficienza energetica, un'azienda deve conoscere i reali flussi energetici, individuare i consumi anomali e valutare interventi di riqualificazione. Il presente lavoro ha come scopo quello di svolgere queste analisi per gli edifici dell'Università degli studi di Milano-Bicocca. Oggi si è raggiunta la consapevolezza etico-sociale che l'energia deve essere misurata, contabilizzata e determinata economicamente come qualsiasi altra materia prima. Un buon controllo dell'efficienza energetica implica la capacità degli impianti di

raggiungere il risultato prefissato con il minor consumo di energia, riducendo i costi e l'impatto sull'ambiente.

L'obiettivo del lavoro svolto è individuare eventuali consumi anomali, da segnalare ai responsabili della gestione energetica dell'università; in secondo luogo si cerca di raggruppare gli edifici con andamento simile nei consumi, valutando le caratteristiche di ogni gruppo ottenuto, favorendo l'applicazione di politiche energetiche costruite specificatamente sulla base del gruppo di appartenenza dell'edificio preso in considerazione.

Il lavoro ha come scopo quello di fornire dati utili all' *energy manager*, al fine di attuare un efficientamento energetico, dove necessario.

Nello specifico, le analisi riguardano gli edifici U1, U2, U3, U4, U5, U6, U7, U8, U9, U14, U16, U17 per il periodo 2018 - 2021. Per l'U8 sono presenti solo gli anni 2019 - 2020 e per l'U17 solo il 2021.

In questo arco temporale è possibile vedere i cambiamenti nei consumi a causa dell'emergenza sanitaria; infatti, proprio a causa di questa e all'introduzione delle lezioni a distanza, ci sono state notevoli riduzioni nei consumi ed è possibile verificare come i differenti edifici abbiano reagito alle chiusure.

Le analisi svolte, riportate nel Capitolo 3, hanno successivamente portato a considerare uno sviluppo nel campo della regressione non parametrica: in particolare vengono utilizzate le *spline di regressione* per smussare l'andamento dei consumi nel tempo e, ottenuti i coefficienti della funzione smussata, questi vengono utilizzati per raggruppare gli edifici.

Una volta ottenuta la suddivisione con questo metodo, si svolge un confronto rispetto al precedente raggruppamento, con l'obiettivo di visualizzare le differenze nei gruppi ottenuti e quali possano essere le variabili discriminanti nel decidere l'appartenenza ad un insieme piuttosto che ad un altro.

Capitolo 2

Regressione non parametrica

I metodi parametrici utilizzati per risolvere problemi di carattere univariato o multivariato, hanno la limitazione di dover ricorrere all'introduzione di ipotesi piuttosto restrittive, oltre alle assunzioni per renderli applicabili. Nel momento in cui le ipotesi sottostanti vengono a mancare si utilizzano approcci non parametrici. Questa categoria di metodi ha la particolarità di non fare riferimento ad alcuna formulazione parametrica per f, cioè la funzione che esprime la relazione tra variabile risposta e covariate. L'obiettivo diventa quello di stimare f senza assumere che quest'ultima appartenga ad una specifica classe di funzioni parametriche, ma stabilendo solo alcune condizioni matematiche di regolarità.

L'approccio non parametrico risulta essere particolarmente efficace, soprattutto nel caso di grandi quantità di dati. Infatti, in questi contesti i modelli parametrici sono spesso poco efficaci dato che cercano di riassumere dati attraverso un numero limitato di parametri. Utilizzando metodi che offrono grande flessibilità questa difficoltà può essere superata.

2.1 Regressione Locale

2.1.1 Formulazione di base

Il fine è quello di esaminare la relazione che collega due quantità, rappresentate dalle variabili x e y, usando la formula del tipo:

$$y = f(x) + \varepsilon, \tag{2.1}$$

dove ε è il termine di errore casuale non osservabile. Senza perdita di generalità si può assumere che $\mathbb{E}[\varepsilon] = 0$ perché un possibile valore diverso da zero può essere incluso in $f(x)$.

Considerando un generico punto fissato x_0 appartenente ai numeri reali, si vuole stimare $f(x)$ nel punto x_0. Se la funzione $f(x)$ è derivabile, con derivata continua in x_0, allora, basandosi sullo sviluppo in serie di Taylor, $f(x)$ è localmente approssimata dalla linea passante per il punto $(x_0, f(x_0))$, data da:

$$f(x) = \underbrace{f(x_0)}_{\beta_0} + \underbrace{f'(x_0)}_{\beta_1}(x - x_0) + \text{resto},$$

dove il resto ha un ordine di grandezza minore di $|x - x_0|$. Si stima $f(x)$ in un vicinato di x_0 mediante un criterio che sfrutta questo fatto, utilizzando le n coppie (x_i, y_i) per $i = 1, ..., n$. Introducendo un criterio analogo alla stima dei minimi quadrati, con la differenza che le osservazioni sono pesate in base alla loro distanza dal punto x_0, si ottiene:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \{y_i - \beta_0 - \beta_1(x_i - x_0)\}^2 w_i, \tag{2.2}$$

dove i pesi w_i sono scelti in modo da essere più grandi quando $|x_i - x_0|$ è più piccolo. La formula 2.2 rappresenta una forma particolare del criterio dei *minimi quadrati pesati*, una generalizzazione dei minimi quadrati dove è presente una serie di pesi.
Seguendo questo criterio, le stime dei parametri $\beta = (\beta_0, \beta_1)^T$ sono:

$$\hat{\beta} = (X^T W X)^{-1} X^T W y,$$

dove X è una matrice $n \times 2$ con l'i-esima riga pari a $(1, (x_i - x_0))$, W è una matrice diagonale $n \times n$ con w_i come elementi diagonali. Siccome i pesi w_i sono costruiti con una prospettiva "locale" rispetto x_0, la stima risultante da questo metodo è chiamata *regressione locale*.
Un modo per selezionare i pesi è impostare

$$w_i = \frac{1}{h} w\Big(\frac{x_i - x_0}{h}\Big),$$

dove $w(\cdot)$ è una funzione di densità simmetrica attorno l'origine, che in questo contesto, è chiamata *kernel*, e h (con $h > 0$) che rappresenta una fattore di scala, detto *parametro di lisciamento*.
L'espressione 2.2 dipende dai pesi w_i, che a loro volta dipendono da h, $w(\cdot)$ e x_0. Anche con h e kernel $w(\cdot)$ fissati, il problema di minimizzazione dipende da x_0, e

stimare $f(x)$ per differenti scelte di x richiede molte operazioni di minimizzazione. Ripetere le minimizzazioni non è un problema dato che è possibile mostrare che la stima relativa ad un generico punto x può essere ottenuta dalla formula:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\{a_2(x;h) - a_1(x;h)(x_i - x\}w_i y_i}{a_2(x;h)a_0(x;h) - a_1(x;h)^2} \tag{2.3}$$

dove

$$a_r(x;h) = \left\{\sum (x_i - x)^r w_i\right\}/n \qquad \text{per r =0,1,2.}$$

Ora si tratta di una stima non iterativa e lineare in y_i e si può scrivere:

$$\hat{f}(x) = s_h^T y$$

per un vettore adatto $s_h \in \mathbb{R}$ che dipende da h e $x_1, ..., x_n$. Solitamente la stima di $f(x)$ è fatta su un insieme di m valori (generalmente equidistanti) che coprono l'intervallo di interesse per la variabile x. Ognuna delle m stime può essere calcolata attraverso un'unica operazione matriciale del tipo:

$$\hat{f}(x) = S_h y \tag{2.4}$$

dove S_h è una matrice $m \times n$, chiamata *matrice di lisciamento*. Se n è molto grande, è possibile ridurre le dimensioni della matrice S_h raggruppando le variabili in classi.
La scelta di approssimare una funzione $f(x)$ localmente attraverso una retta può essere rilassata adottando un polinomio. Alcune alternative possono essere polinomi di grado 0 o 2. Nel primo caso la stima per ogni punto risulterà essere una media pesata dei dati appartenenti al vicinato; una variante di questa procedura, chiamata *k-nearest-neighbor*, è solitamente un'opzione migliore. Il polinomio di grado 2 è una scelta appropriata nel momento in cui i dati mostrano picchi e cavi acuti, perché questa variante è più adatta a riprodurre curve ripide.

2.1.2 Scelta del parametro di lisciamento

Rimane il problema della scelta di h e $w(\cdot)$, sebbene l'aspetto veramente importante riguarda la scelta del parametro di lisciamento h. Un valore piccolo di h produce una curva stimata che si avvicina maggiormente al comportamento

locale dei dati ed è più volatile, dato che i pesi sono calcolati per una finestra piccola, rendendo la stima più sensibile alla volatilità locale dei dati. Nell'altro caso, con h grande, viene prodotto l'effetto opposto, con una curva più smussata. Infatti, guardando il caso in cui $var\{\varepsilon_i\} = \sigma^2$ è positiva e costante per tutte le osservazioni e assumendo che queste siano incorrelate, allora, sotto le adatte condizioni di regolarità per f, possiamo provare che per h sufficientemente vicino a 0 e n abbastanza grande, vale l'approssimazione:

$$\mathbb{E}\{\hat{f}(x)\} \approx f(x) + \frac{h^2}{2}\sigma_w^2 f''(x), \qquad var\{\hat{f}(x)\} \approx \frac{\sigma^2}{nh}\frac{\alpha(w)}{g(x)}, \tag{2.5}$$

in cui $\sigma_w^2 = \int z^2 w(z)dz$, $\alpha(w) = \int w(z)^2 dz$ e $g(x)$ indica la densità da cui sono state estratte le x_i. Queste equazioni mostrano che la distorsione è un multiplo di h^2 e la varianza è multipla di $1/nh$. Si dovrebbe scegliere $h \to 0$ per diminuire la distorsione, ma così facendo si farebbe divergere la stima della varianza. Nel caso opposto, in cui $h \to \infty$, si riduce la varianza ma diverge la distorsione. Le relazioni 2.5 sono valide in caso di ipotesi abbastanza restrittive, ma lo stesso tipo di risultati si ottengono in caso di ipotesi meno stringenti.
La soluzione da adottare è quella di minimizzare la somma della varianza e la radice quadrata della distorsione, ottenendo un bilanciamento tra le due componenti. La scelta migliore per h è asintoticamente pari a:

$$h_{opt} = \left(\frac{\alpha(w)}{\sigma_w^4 f''(x)^2 g(x) n}\right)^{1/5}. \tag{2.6}$$

Grazie a questa espressione si può notare che:

- h tende a 0 come $n^{1/5}$, dunque decresce molto lentamente.
- se venisse sostituita h_{opt} nell'espressione di media e varianza 2.5, si noterebbe che l'errore quadratico medio tende a 0 come $n^{-4/5}$; questo implica che la stima non parametrica è intrinsecamente meno efficiente della corrispondente parametrica che decresce ad una velocità pari a n^{-1}, nel caso in cui il modello parametrico sia adeguato.

Operativamente la scelta di h può essere fatta anche attraverso altre vie. Un'alternativa alla 2.6 è il metodo della *Convalida Incrociata* (CV) e del *Criterio di Informazione di Akaike* con una variazione (AIC_c):

$$AIC_c = \log(\hat{\sigma}^2) + 1 + \frac{2\{tr(S_h) + 1\}}{n - tr(S_h) - 2},$$

in questo caso:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (y_i - \hat{f}(x_i))^2 = \frac{1}{n} y^T (I_n - S_h)^T (I_n - S_h) y,$$

è la stima della varianza residua σ^2, e $tr(S_h)$ indica la traccia della matrice S_h, che costituisce una misura sostitutiva del numero di parametri coinvolti.

Per visualizzare un'applicazione di quanto introdotto, attraverso il software `R`, vengono simulate 300 osservazioni estratte da una *distribuzione uniforme* con estremi [-1,1]. Queste vengono inserite nella funzione

$$f(x) = -e^{-20x^2} + \frac{x}{3} + x^3,$$

alla quale viene aggiunto un disturbo casuale attraverso l'estrazione di valori pseudo-casuali da una *distribuzione normale* con media 0 e scarto quadratico medio 0.5, grazie al comando `rnorm(300, 0, 0.5)`.

Successivamente viene applicata la regressione locale con il comando `sm.regression` della libreria `sm`. Questo permette di creare una stima non parametrica per dati con una variabile risposta e fino a due esplicative. Inoltre, è possibile selezionare il metodo secondo cui deve essere scelto il parametro di lisciamento; nel presente caso sono stati confrontati i risultati di AIC_c e CV.

Visualizzando la Figura 2.1 si nota che attraverso l'utilizzo della Convalida incrociata si riesce ad ottenere un andamento (linea blu) più vicino al reale andamento della funzione generatrice dei dati (che solitamente non è nota). In particolare, il valore di `h` selezionato con CV è di 0.089 mentre con l'AIC si ottiene un valore di 0.222. Infatti, nel secondo caso si vede un andamento maggiormente smussato, in linea con quanto riportato dalla (2.5) che stabilisce che per valori più grandi di `h`, siccome questo termine appare al denominatore della formula della varianza, implicano variabilità minore.

2.1.3 Parametro di lisciamento variabile e *loess*

Ci sono molte possibili varianti del metodo della regressione locale. La variante più comune riguarda l'uso di un'ampiezza di banda non costante lungo l'asse delle x, in base al livello di sparsità dei punti osservati. Infatti, nel momento in cui le x_i sono disperse è ragionevole utilizzare un valore di h più elevato. Queste considerazioni intuitive sono confermate dall'espressione (2.6), dove la presenza di $g(x)$ al denominatore mostra che quando $g(x)$ è bassa, e lo è nel momento

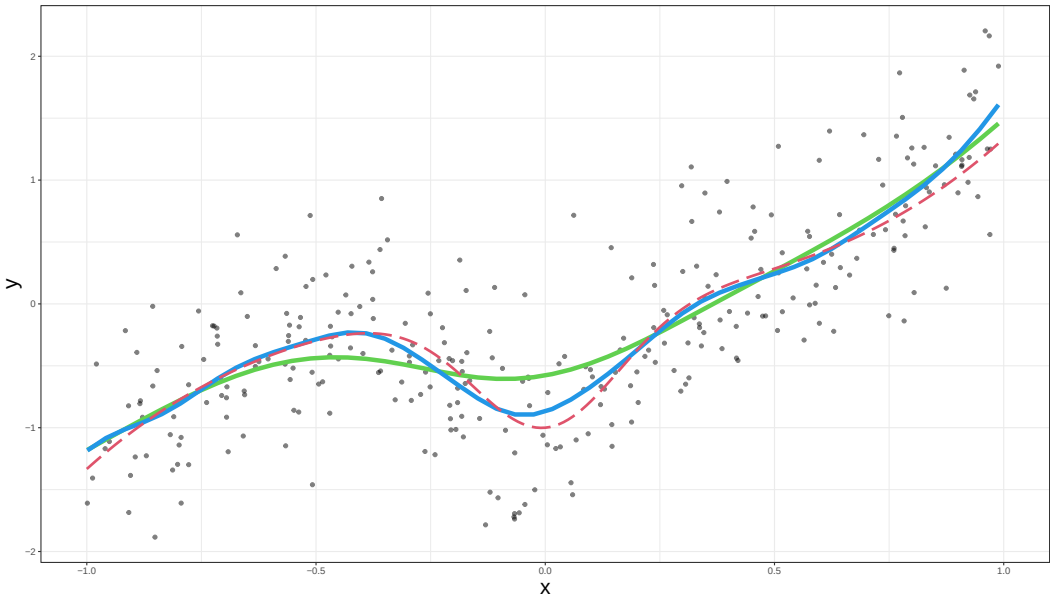

Figura 2.1: Regressione locale con in verde il metodo AIC e in blu CV. La linea rossa rappresenta la vera funzione generatrice dei dati.

in cui le x_i sono sparse, si deve utilizzare un valore di h ottimale maggiore per mantenere uguale la varianza della stima di f.

Una tecnica che utilizza questa intuizione è il *loess*, molto simile alla regressione locale. Una caratteristica distintiva del *loess* è che esprime il parametro di smussamento attraverso una proporzione delle effettive osservazioni che sono utilizzate per la stima di $f(x)$ in un certo punto delle ascisse. Questa proporzione viene mantenuta costante allargando o stringendo l'intervallo sull'asse x. Il grado di smussamento è quindi regolato dalla proporzione di osservazioni che contribuiscono alla stima di $f(x)$; questa frazione costituisce a tutti gli effetti il parametro di lisciamento nel *loess*. Un'altra peculiarità del *loess* è quella di combinare l'idea della regressione locale con la *stima robusta*, che implica la sostituzione della forma quadratica nella (2.2) con un'altra funzione in modo da limitare l'effetto delle osservazioni anomale. Per ottenere la *robustezza* della stima, il *loess* utilizza un kernel con supporto limitato (generalmente il tricubico) che ha il vantaggio di distinguere in modo chiaro quali osservazioni utilizzare e quali no all'interno della stima.

Facendo riferimento ai dati precedentemente simulati, viene utilizzato il comando `loess` che richiede in entrata la formula con la relazione esistente tra variabile risposta e covariate e lo span, cioè la proporzione di osservazioni che andranno di volta in volta a contribuire alla stima della f per un determinato punto. Visualizzando la Figura 2.2, si vede come al variare dello span cambi il

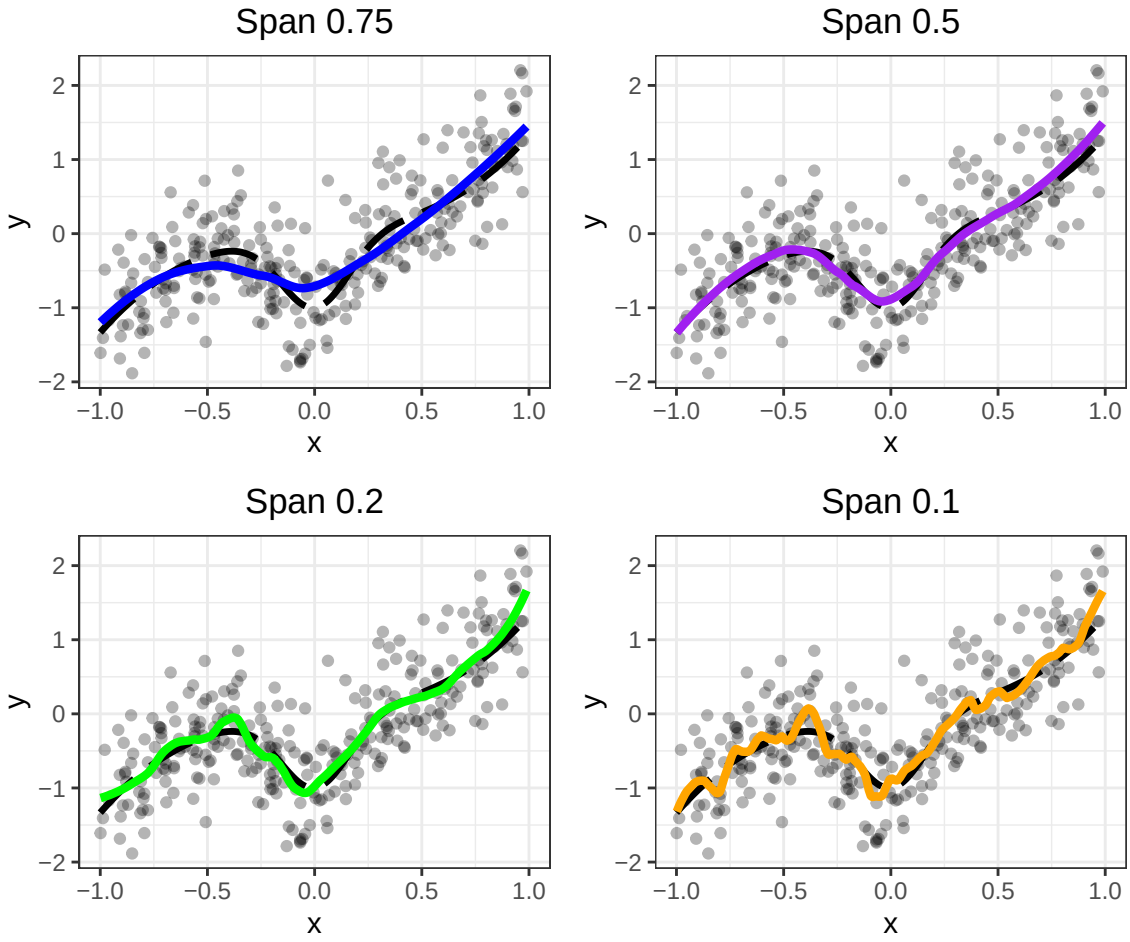

Figura 2.2: Utilizzo del comando `loess`. La linea nera rappresenta la funzione generatrice dei dati e le altre linee sono ottenute con differenti valori dello span.

livello di smussamento delle funzione stimata $\hat{f}(x)$. Con un valore più alto, ci sarà un maggior numero di punti a contribuire alla stima della curva, ottenendo un maggior smussamento; viceversa, nel caso di span basso, la curva stimata tende ad essere più variabile, seguendo maggiormente l'andamento locale dei punti. In questo caso, utilizzando una proporzione pari al 50%, viene generata una curva che si avvicina maggiormente alla reale funzione generatrice dei dati.

2.2 Spline

2.2.1 Funzioni di tipo Spline

Il termine *spline* è usato in matematica per indicare la costruzione di funzioni polinomiali a tratti, utilizzate per approssimare funzioni di cui sono noti solo alcuni punti. Vengono scelti K punti $\xi_1 < \xi_2 < ... < \xi_K$, detti *nodi*, lungo l'asse x. Si vuole costruire una funzione $f(x)$ in modo che passi esattamente attraverso i nodi e sia libera negli altri punti, con il vincolo che abbia una certa regolarità nell'andamento generale. La costruzione avviene nel seguente modo: tra due nodi successivi, nell'intervallo (ξ_i, ξ_{i+1}), la curva $f(x)$ coincide con un polinomio adatto, di grado prefissato pari a d, e queste sezioni di polinomi si incontrano nei punti $\xi_i (i = 2, ..., K-1)$, nel senso che la risultante funzione $f(x)$ ha derivata continua dal grado 0 al grado $d-1$ in ognuno dei nodi ξ_i. Il grado solitamente utilizzato è $d = 3$, quindi si parla di *spline cubiche*. Questa scelta è fatta perché le discontinuità nella derivata terza non si riescono a cogliere graficamente.
Le condizioni precedenti possono essere scritte come:

$$f(\xi_i) = y_i \quad \text{per} \quad i = 1, ..., K$$
$$f(\xi_i^-) = f(\xi_i^+), \quad f'(\xi_i^-) = f'(\xi_i^+), \quad f''(\xi_i^-) = f''(\xi_i^+) \quad \text{per} \quad i = 2, ..., K-1$$

dove $g(x^-)$ e $g(x^+)$ indicano il limite da sinistra e da destra della funzione $g(\cdot)$ nel punto x.

La struttura del problema richiede alcune condizioni: ognuna delle $K-1$ componenti necessita di 4 parametri; ci sono K vincoli del tipo $f(\xi_i) = y_i$ e $3(K-2)$ vincoli di continuità relativamente alla funzione e alle prime due derivate. É necessario introdurre due ulteriori vincoli siccome la differenza tra coefficienti e vincoli è di 2 unità e il sistema precedente non identifica univocamente una funzione.

Sono state fatte molte proposte riguardo le due condizioni aggiuntive, molte delle quali riguardano gli intervalli o i punti estremi della funzione. Una scelta particolarmente semplice consiste nel vincolare le derivate seconde dei polinomi nei due nodi esterni ad essere nulle, $f''(\xi_1) = f''(\xi_K) = 0$; da questo deriva che i due polinomi estremi sono localmente lineari al secondo ordine di approssimazione. La funzione risultante è chiamata *spline cubica naturale*.

2.2.2 Spline di regressione

Lo strumento precedente può avere diversi utilizzi in statistica nello studio delle relazioni tra una variabile esplicativa x e una risposta y attraverso l'utilizzo di n coppie di osservazioni (x_i, y_i) per $i = 1, ..., n$.

Applicando quanto introdotto alla regressione parametrica, con l'ipotesi che $f(x; \beta)$ sia una funzione spline, si suddivide l'asse delle ascisse in $K + 1$ intervalli separati da K nodi, $\xi_1, ..., \xi_K$, e vengono interpolati gli n punti con il criterio dei minimi quadrati, ottenendo dei coefficienti β che sono parametri non vincolati dei $K + 1$ polinomi costituenti.

Rispetto a quanto detto nel paragrafo precedente 2.2.1, c'è una certa differenza in quanto la selezione dei coefficienti della funzione spline non può più avvenire secondo vincoli del tipo $f(\xi_i) = y_i$, perché K e n sono slegate e $K << n$. Questo significa che è necessario utilizzare un criterio di regressione tra i dati e la funzione interpolante, come quello dei minimi quadrati o un altro simile.

Se venissero utilizzate le spline cubiche, il numero totale di parametri sarebbe di $4(K + 1)$ soggetti a $3K$ vincoli di continuità, cioè β deve avere $K + 4$ componenti. La soluzione al problema di minimo può essere scritta nella forma equivalente

$$f(x; \beta) = \sum_{j=1}^{K+4} \hat{\beta}_j h_j(x), \tag{2.7}$$

dove

$$\begin{aligned} h_j(x) &= x^{j-1} \qquad \text{per j = 1,...,4} \\ h_{j+4}(x) &= (x - \xi_j)^3_+, \quad \text{per j = 1,...,K} \end{aligned}$$

e $a_+ = max(a, 0)$. La soluzione è costituita da una opportuna combinazione lineare di una *base di funzioni* $\{h_j(x), j = 1, ..., K + 4\}$, costituita in parte da polinomi elementari e in parte da funzioni del tipo $max(0, (x - \xi)^3)$. É importante la scelta

del numero dei K nodi e la loro posizione lungo l'asse delle x. Infatti, K è visto come un parametro che regola la complessità del modello. Una volta scelto K, nel momento in cui non si ha informazione riguardo la forma della funzione da stimare, una scelta ragionevole per posizionare i nodi è quella di distribuirli uniformemente lungo l'asse delle x.

Applicando ai dati quanto visto, attraverso il comando `bs` della libreria `spline` può essere generata la matrice di covariate per la *regressione polinomiale spline*. La matrice risultante può essere inserita nel comando `lm` in modo da ottenere una stima della curva interpolante i dati. Nel presente caso sono stati utilizzati polinomi di terzo grado con differente numerosità di nodi. Come accennato precedentemente, la scelta del numero di nodi e la loro posizione ha un ruolo determinante nello stabilire la forma della curva finale. Infatti, visualizzando la Figura 2.3, in alto a sinistra si può vedere che, supponendo di non conoscere la reale funzione sottostante i dati, i nodi sono posizionati uniformemente lungo l'asse delle x. Subito a destra, invece, questi vengono posizionati in modo differente, seguendo le zone di picco e cavo nei dati. I grafici nella seconda riga riportano le stesse considerazioni utilizzando soltanto 3 nodi.
Da questo confronto risulta evidente l'importanza del numero e della posizione dei nodi, infatti, conoscendo la struttura di variabilità dei dati, basterebbero 3 nodi nella corretta posizione per approssimare in modo soddisfacente la reale funzione generatrice.

2.2.3 Spline di lisciamento

Un altro modo per utilizzare le funzioni di tipo *spline* nello studio della relazione tra variabili è per introdurre un approccio alla stima non parametrica, alternativa alla regressione locale. Considerando il criterio dei minimi quadrati penalizzati

$$D(f, \lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_{-\infty}^{+\infty} \{f''(t)\}^2 dt \tag{2.8}$$

dove λ è un parametro positivo di penalizzazione del grado di irregolarità della curva f, quantificato dall'integrale di $f''(x)^2$, cioè agisce come parametro di lisciamento.

Se $\lambda \to 0$ non c'è penalizzazione per l'irregolarità di $f(x)$, quindi il criterio precedente non è influenzato da $f(x)$ fuori dalle ascisse $x_1, ..., x_n$ e la soluzione

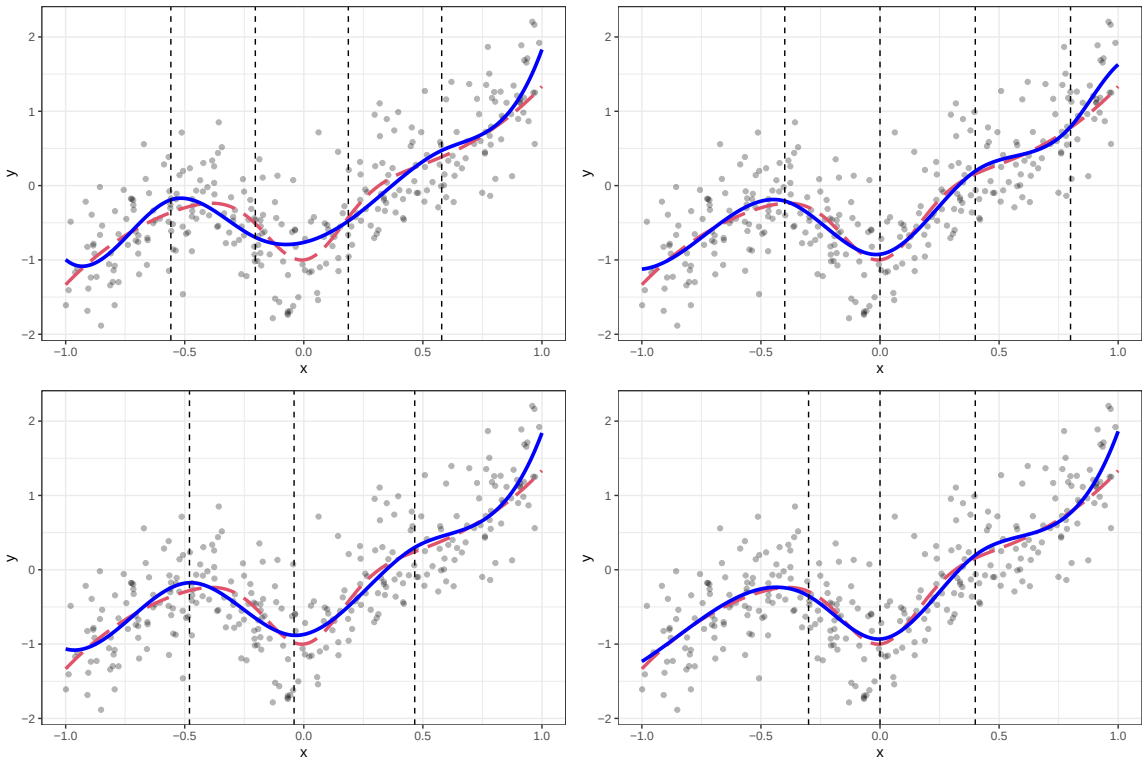

Figura 2.3: Spline di regressione utilizzando polinomi di terzo grado, nella prima riga sono utilizzati quattro nodi, nella seconda solo tre. La linea rossa tratteggiata rappresenta la reale funzione generatrice, la linea blu la funzione stimata e le linee tratteggiate la posizione dei nodi.

ottima $\hat{f}(x_i)$ è la media aritmetica delle y_i corrispondenti a quella data ascissa, per ciascuna x_i osservata ma non è determinata per altri valori di x.

Se $\lambda \to \infty$ la penalità è massima e comporta adattare una retta, siccome viene imposto $f''(x) \equiv 0$, il risultato è la retta dei minimi quadrati. Quindi λ ha lo stesso ruolo che aveva h nel caso della regressione locale.

Un risultato significativo afferma che la soluzione del problema di minimizzazione è costituito da una funzione del tipo *spline cubica naturale*, i cui nodi sono i punti x_i distinti. La soluzione può essere scritta come

$$\hat{f}(x) = \sum_{j=1}^{n_0} \theta_j N_j(x)$$

dove n_0 è il numero di x_i distinti e gli $N_j(x)$ sono basi delle 'spline cubiche naturali'. In questo caso il numero di parametri e nodi corrisponde siccome i vincoli sono connessi al fatto che ci sia una spline naturale. Si può riscrivere

$$D(f, \lambda) = (y - N\theta)^T (y - N\theta) + \lambda \theta^T \Omega \theta$$

dove N indica la matrice in cui la j-esima colonna contiene i valori di N_j in corrispondenza agli n_0 valori distinti di x, e la matrice Ω il cui generico elemento è $\int N_j''(t) N_k''(t) dt$. La soluzione del problema di ottimizzazione è data da

$$\hat{\theta} = (N^T N + \lambda \Omega)^{-1} N^T y \tag{2.9}$$

che dipende dalla scelta del parametro di lisciamento λ.

Sostituendo questa espressione di $\hat{\theta}$ nella formula di $f(x)$, si ottiene $\hat{y} = \tilde{S}_\lambda y$ per una certa matrice $\tilde{S}_\lambda$ di dimensione $n_0 \times n_0$; si tratta quindi di un altro *lisciatore lineare* e si parla di *spline di lisciamento*.

Dal punto di vista computazionale non viene utilizzata la 2.9 che coinvolgerebbe una matrice di ordine n_0 ma vengono utilizzati degli algoritmi più efficienti Inoltre, quando la quantità di dati è molto elevata, si può ridurre il numero di nodi utilizzati, senza perdita di accuratezza, come avveniva nel caso della regressione locale.

Utilizzando i dati simulati, con il comando `smooth.spline` è possibile applicare le spline di lisciamento ai vettori x e y, con l'opzione di selezionare il parametro λ attraverso la Convalida Incrociata. Per visualizzare l'effetto della scelta di λ

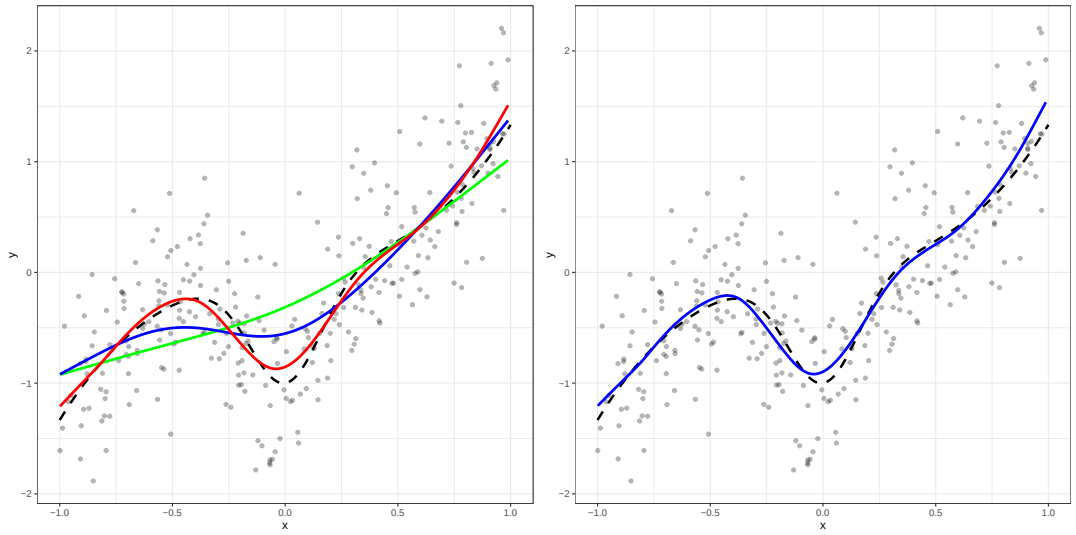

Figura 2.4: Spline di lisciamento; nel grafico a sinistra si vede come varia lo smussamento della funzione al variare del parametro λ. Il grafico a destra rappresenta la funzione stimata con λ scelto tramite CV (pari a 0.00106).

sono stati scelti tre valori differenti e questi sono poi confrontati con il valore ottenuto grazie al CV.

I valori di λ utilizzati sono: 0.002, 0.05, 0.8; nella Figura 2.4 la linea rossa del grafico a sinistra rappresenta la funzione che si ottiene con il valore di λ pari a 0.002, 0.05 in blu e 0.8 in verde. La prima curva è meno smussata rispetto le altre due opzioni, in particolar modo se confrontata al caso $\lambda = 0.8$, dove la funzione si avvicina molto ad una retta (caso $\lambda \to \infty$). Questo risultato è in linea con la teoria precedentemente introdotta, siccome il parametro λ rappresenta la penalizzazione per il grado di irregolarità della curva stimata e, di conseguenza, all'aumentare del suo valore aumenterà il livello di lisciamento. Nel grafico a destra, invece, si vede che il parametro selezionato attraverso CV ($\lambda = 0.00106$), genera una funzione molto vicina alla reale struttura sottostante ai dati (linea nera tratteggiata).

Si veda ad esempio Azzalini & Scarpa (2012) per una descrizione dettagliata di queste tecniche non parametriche.

Capitolo 3

Analisi consumi elettrici

L'università degli studi di Milano-Bicocca venne istituita il 10 giugno 1998. Il campus Bicocca sorge nell'omonimo quartiere milanese un tempo sede di grandi industrie come Pirelli e Breda. Il nuovo Ateneo si inserisce in un ampio progetto avviato intorno al 1986 e coordinato dall'architetto Vittorio Gregotti. L'ateneo dispone di 21 edifici localizzati nel quartiere Bicocca ad eccezione della sede della facoltà di Medicina e Chirurgia che si trova a Monza (Centrale, 2022).

L'analisi dei consumi elettrici degli edifici è inserita in un contesto storico in cui risulta più che mai essenziale diminuire gli sprechi e rendere efficiente il consumo, con il fine di ridurre l'impatto ambientale e i costi in termini economici. I dati a disposizione sono rilevati ogni 15 minuti attraverso dei contatori in ogni edificio. Per tutti gli edifici a disposizione è presente un dataset iniziale con le seguenti variabili di interesse:

- `POD`, contenente il codice identificativo del contatore presente nell'edificio;
- `data`, giorno di rilevazione del consumo;
- `ora`, quarto d'ora a cui fanno riferimento i consumi rilevati. Questa variabile può assumere valori 0, 1500, 3000, ..., 234500 dove, ad esempio, 0 indica il consumo dalle 00:00:00 fino alle 00:15:00;
- `consumo attiva prelevata`, energia che viene prelevata e consumata dall'impianto (da un minimo di $0,54$ kW ad un massimo di $262,35$ kW);
- `consumo reattiva induttiva prelevata`, energia che viene assorbita dai macchinari elettrici ma che non viene impiegata in lavoro (assume valori simili alla variabile precedente).

Questi dati sono stati integrati con le informazioni riguardo la `superficie`, `numero di aule` (per gli edifici in cui questo dato è disponibile), `numero di studenti`, `temperatura media esterna` e dal sito dell'università sono state reperite le informazioni riguardo le grandi attrezzature presenti negli edifici.

Tutte le analisi sono state svolte sulla variabile `consumo attiva prelevata`, rispetto a data, ora ed edificio di rilevazione. Gli edifici presi in considerazione nelle successive elaborazioni sono: U1, U2, U3, U4, U5, U6, U7, U8 (unico del Polo di Monza), U9, U14, U16 e U17.

Come anticipato, il periodo di riferimento parte dal primo gennaio 2018 e arriva al 31 dicembre 2021, con rilevazioni ogni 15 minuti espresse in kW. Nel caso di U8 e U17 sono presenti solo alcuni anni, motivo per cui questi edifici verranno esclusi dalle analisi di raggruppamento.

Dopo una prima pulizia del dato da ripetizioni ed errori di rilevazione, sono stati imputati i mesi mancanti, dove possibile. Per effettuare questa operazione sono state utilizzate diverse regressioni lineari: dato un giorno e un orario mancante, vengono individuati lo stesso giorno e orario negli altri anni. In questo modo vengono trovati 3 punti a cui viene interpolata una retta attraverso una regressione lineare e questa è utilizzata per fare previsione. Così facendo, il valore imputato tiene conto della tendenza del consumo negli anni, senza utilizzare tecniche particolarmente complesse. I passaggi appena svolti vengono replicati per ogni osservazione mancante, con il fine di rendere completa la serie dei consumi. Gli edifici con dati mancanti per cui è stata effettuata questa imputazione sono U1, U2, U4, U9, U16.

Sommando i dati degli edifici di cui sono presenti i dati per tutti e quattro gli anni (vengono esclusi U8 e U17), si vede che il consumo generale dell'università ha una tendenza decrescente ma ci sono alcune particolarità: la variazione nel consumo tra 2018 e 2019 è stata del $-1,76\%$ mentre nel caso 2019-2020 è del $-14,31\%$; questa notevole differenza è dovuta alle chiusure a causa della pandemia. Nel 2020, infatti, la maggior parte delle attività e della vita universitaria è stata svolta a distanza, determinando una notevole diminuzione dei consumi negli edifici e in particolar modo in quelli del settore amministrativo. L'unica attività che era possibile svolgere in presenza era la ricerca e questo fatto è evidente anche dai dati: gli edifici con maggior numero di grandi attrezzature per la ricerca hanno avuto dei cali molto meno pronunciati di quanto non sia avvenuto in molti atri edifici nel periodo di marzo-aprile 2020. Questa distinzione tra edifici "amministrativi" e di "ricerca" sarà studiata in seguito nelle analisi

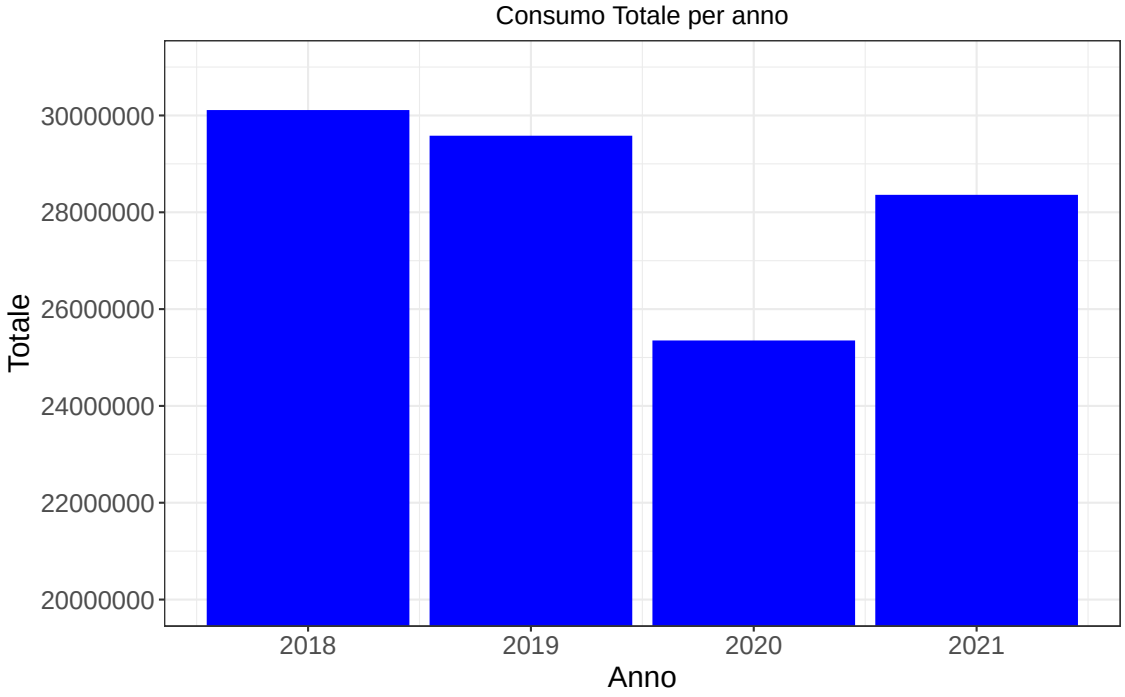

Figura 3.1: Totale consumo suddiviso per anno

dei raggruppamenti, dove risulterà essere un possibile fattore discriminante per l'appartenenza ad un insieme piuttosto che ad un altro.

Proprio a causa della brusca riduzione dei consumi nel 2020 dovuta alla pandemia, la variazione 2020-2021 è del $11,88\%$. Confrontando però il consumo del 2021 rispetto quello del 2019 si nota che c'è una riduzione del $-4,13\%$, confermando la generale tendenza dei consumi a diminuire come si vede dalla Figura 3.1.

Il generale andamento decrescente dei consumi non è dovuto sicuramente alla diminuzione degli studenti, che sono in continuo aumento (infatti dal grafico visualizzabile dal sito dell'università (programmazione e controllo, 2022a), nell'anno accademico 2020-2021 gli studenti erano 35.872, nel 2021-2022 sono aumentati a 37.012 iscritti). Questo risultato potrebbe rispecchiare la maggiore sensibilità dell'università rispetto agli sprechi e la continua ricerca volta a migliorare l'efficienza energetica.

In un secondo momento sono state create diverse aggregazioni: giornaliera, settimanale e mensile. In questo modo è possibile visualizzare diverse granularità del dato: nell'aggregazione mensile sono visibili le tendenze, togliendo il "disturbo" dato dalla volatilità del dato; andando nello specifico, con l'aggregazione giornaliera, si può studiare più a fondo gli eventuali valori anomali.

Analizzando il campione sulla base dell'aggregazione mensile, si vede che l'edificio maggiormente energivoro risulta essere l'U6, con un consumo medio mensile di 386.394, 248 kWh e un totale di 18.546.924 kWh consumati nei quattro anni di riferimento (Tablella 3.1). Questo edificio, da quanto riportato dal sito riguardante i dati delle infrastrutture dell'Università (programmazione e controllo, 2022b), risulta essere quello con più aule (46), con la maggior capienza (6193 posti) e contiene rettorato e biblioteca. Quanto ottenuto è abbastanza scontato: essendo uno degli edifici più grandi, è plausibile che sia anche quello con il consumo medio maggiore.

La controparte con il consumo elettrico minore è l'U17 con valore medio mensile di 4.020, 864 kWh e un totale di 48.250, 37 kWh nell'unico anno disponibile. Di questo edificio non sono riportate le aule e il numero di posti perché si tratta della segreteria studenti che, con una superficie di 2650 m^2, risulta essere l'immobile più piccolo tra quelli analizzati; come detto prima, il risultato ottenuto è abbastanza in linea con le aspettative. Infatti, come è logico pensare e come riscontrato dai dati, la superficie è un fattore determinante nello stabilire il livello di consumo medio mensile di un edificio; la metratura e il consumo medio mensile hanno in effetti un coefficiente di correlazione lineare pari a 0, 646, indice del fatto che all'aumentare di una variabile, aumenta anche l'altra nella stessa direzione.

Oltre a visualizzare i consumi medi e totali è interessante studiare come questi si distribuiscano nel tempo.

Calcolando la correlazione tra superficie e scarto quadratico medio del consumo sembrerebbe esistere la stessa relazione che si aveva nel caso della media: si ottiene un coefficiente di 0.662, che indica una relazione positiva tra superficie e radice della varianza. L'associazione non è rispettata rigorosamente. Infatti, guardando gli edifici con la variabilità maggiore si ottiene che l'U9 ha il consumo mensile più variabile in assoluto, superiore del 17, 1% rispetto allo scarto quadratico medio dell'U6 (che è il secondo edificio in termini di volatilità dei consumi), nonostante abbia una superficie pari al 38% di quest'ultimo. Questa "discrepanza" rispetto le attese potrebbe indicare la presenza di consumi anomali nell'U9 o l'utilizzo di macchinari particolari all'interno dell'edificio che necessitano di picchi di energia per svolgere determinate funzioni, rendendo il consumo più volatile.

EDIFICIO	MEDIA	SQM	MIN	MAX	TOTALE	SUPERFICIE (m²)
U1	215	40	145	342	10338	8076
U2	372	48	307	496	17844	23665
U3	323	48	234	449	15489	15646
U4	256	24	213	326	12309	17556
U5	206	35	138	287	9867	14410
U6	386	58	205	501	18547	69183
U7	253	51	143	346	12127	47716
U8	301	41	236	359	7216	15900
U9	250	67	125	427	11976	26371
U14	65	17	44	113	3100	6236
U16	36	6	23	52	1722	7134
U17	4	1	3	7	48	2650

Tabella 3.1: Statistiche principali consumo mensile per edificio espresse in MWh, la superficie riportata è espressa in m². Il totale tiene conto anche dei valori imputati.

Negli edifici con il consumo più stabile (e quindi variabilità minore), la relazione variabilità-superficie sembra essere rispettata, infatti l'U17 che ha la superficie più piccola è anche l'edificio con scarto quadratico medio minore.

Tralasciando i casi estremi, questi rapporti superficie-consumo medio e superficie-scarto quadratico medio sono rispettati ma non con estrema precisione: infatti, se un edificio ha una superficie maggiore rispetto ad un altro, questa non è una condizione sufficiente per stabilire che avrà un consumo medio e una variabilità maggiore. Un esempio è il caso di U3 e U4 che, pur avendo il primo una superficie inferiore di quasi 2000 m² rispetto al secondo, ha un consumo medio e uno scarto quadratico medio superiore (rispettivamente del $25,83\%$ e del $102,63\%$).

Questo fatto indica che la superficie non è l'unico fattore a determinare il livello e l'andamento del consumo ma ci sono altre variabili che possono condizionare queste misure. Vedendo l'influenza della metratura dell'edificio nel determinare quale sia quello con il consumo maggiore e minore, è stata effettuata una normalizzazione dei consumi rispetto a questa variabile. Così facendo, si ottiene il consumo per metro quadro di ogni edificio e i valori sono maggiormente confrontabili. La soluzione adottata è utile per evidenziare quale

EDIFICIO	MEDIA	SQM	MINIMO	MASSIMO	TOTALE
U1	26.668	4.920	17.964	42.360	1280.080
U2	15.709	2.047	12.964	20.962	754.031
U3	20.624	3.088	14.932	28.675	989.943
U4	14.607	1.358	12.118	18.570	701.127
U5	14.265	2.448	9.596	19.882	684.724
U6	5.585	0.831	2.959	7.245	268.085
U7	5.295	1.072	3.003	7.255	254.156
U8	18.910	2.567	14.854	22.590	453.852
U9	9.461	2.553	4.731	16.174	454.140
U14	10.357	2.656	7.110	18.190	497.112
U16	5.029	0.836	3.232	7.342	241.378
U17	1.517	0.416	1.019	2.543	18.208

Tabella 3.2: Statistiche principali consumo mensile espresso in kWh per metro quadro per ogni edificio. Le statistiche sono calcolate per il periodo 2018-2021 includendo anche i valori imputati.

sia effettivamente l'immobile con il dispendio energetico maggiore non solo a causa della grandezza superiore rispetto gli altri, ma per le attività che sono svolte al suo interno.

Utilizzando i dati normalizzati per metro quadro, visualizzabili nella Tabella 3.2, si ottiene che l'edificio con i consumi medi mensili più elevati risulta essere l'U1 con $26,7$ kWh/m² ed è anche quello con la variabilità maggiore, pari a $4,92$ kWh/m². L'immobile con il consumo medio più basso resta l'U17 con $1,5$ kWh/m² e uno scarto quadratico medio di $0,416$ kWh/m². Riguardando l'U6, che precedentemente era l'edificio con il consumo medio maggiore, ora risulta essere tra gli edifici con il consumo minore per metro quadro con un valore di 5.585 kWh/m². La stessa considerazione vale per la variabilità, infatti l'U9 che era l'edificio con maggior volatilità del consumo in termini assoluti, ora risulta essere il quinto in fatto di scarto quadratico medio. Questi risultati confermano ancora una volta quanto la superficie possa influenzare il consumo elettrico di un edificio e ulteriore conferma è data da uno studio (Corgnati et al., 2010) riguardo il consumo energetico di edifici ad uso scolastico svolto dall'*Agenzia Nazionale per le Nuove Tecnologie, l'Energia e lo Sviluppo Economico Sostenibile* (ENEA), secondo cui

i fattori che influenzano l'utilizzo di energia all'interno di un immobile possono essere raggruppati in 6 macro-categorie:

- Clima esterno;
- Caratteristiche geometriche e termo fisiche dell'edificio;
- Sistemi energetici e impiantistici a servizio dell'edificio;
- Aspetti gestionali e manutentivi;
- Richieste di qualità dell'ambiente interno;
- Comportamento dell'utente.

Di queste categorie, nel caso in esame, sono note solo il clima esterno e rispetto le caratteristiche geometriche è nota solo la superficie dell'edificio. Negli altri casi non sono presenti informazioni specifiche se non il fatto che gli edifici siano dotati di impianto di condizionamento.

Dall'analisi svolta da questo ente (ENEA) risulta che tra le variabili con più alto coefficiente di correlazione con i consumi ci sia anche la superficie, fatto riscontrato nel presente caso. Viene poi riportato che per avere un inquadramento generale della struttura è utile conoscere:

- Dati identificativi dell'edificio;
- Localizzazione;
- Indirizzo;
- Dati climatici;
- Ore di apertura dell'edificio;
- Condizionamento d'aria;
- Servizi per il ristoro;
- Aule.

Questi dati risulteranno utili per le successive analisi, in particolar modo per la sezione 3.2.

3.1 Distribuzione consumi nel tempo

Utilizzando le aggregazioni precedentemente create, in questa sezione si cerca di analizzare come i consumi si distribuiscano nelle ore del giorno e nei mesi. L'obiettivo è quello di verificare se sia presente una certa stagionalità nel consumo e analizzare la possibile presenza di picchi e eventuali consumi anomali rispetto quello che è l'andamento medio degli edifici e della domanda di energia elettrica in Italia.

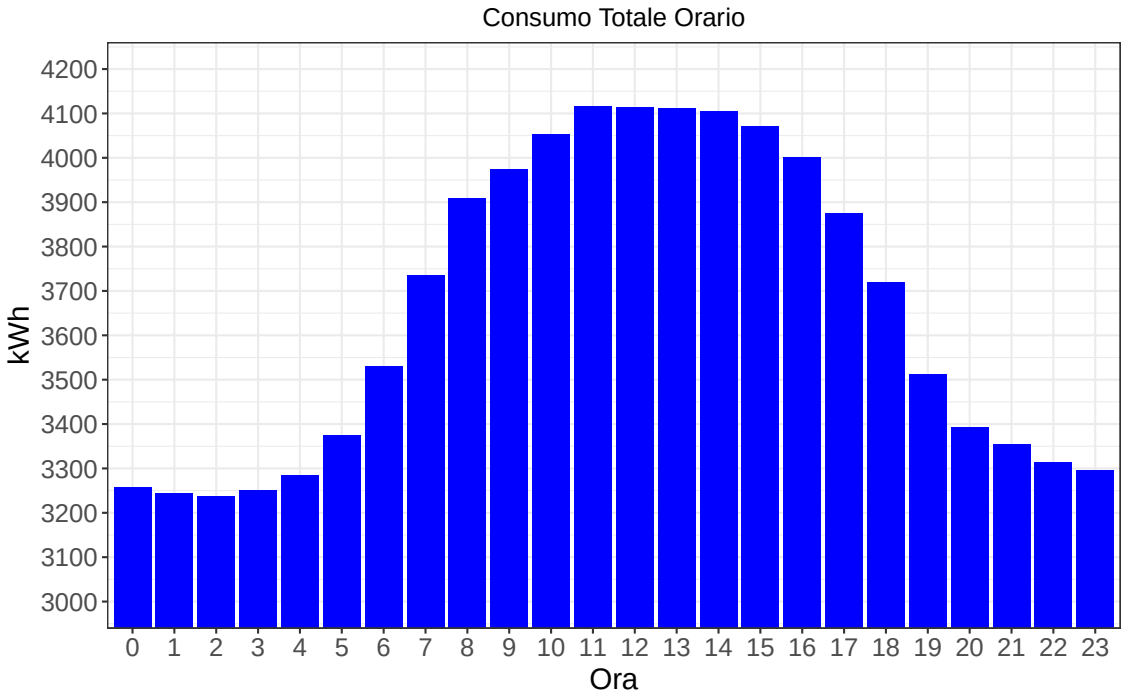

Figura 3.2: Totale consumo medio suddiviso per ora

3.1.1 Andamento orario dei consumi

Aggregando i dati su base oraria e sommando il consumo medio per ogni edificio si ottiene l'andamento mostrato nella Figura 3.2.

Come è possibile vedere dal grafico, il consumo si concentra nelle ore diurne, in particolare le ore di picco sono nella fascia 11-14. Le medie aggregate per tutti gli edifici riportano che non vi è un eccessivo scostamento tra consumo medio giornaliero e notturno. Infatti, considerando giorno il tempo di apertura della maggior parte degli edifici, cioè dalle 7:30 fino alle 20:30, si ottiene che il dispendio energetico notturno è inferiore soltanto del $15,02\%$ rispetto al consumo diurno. Questo dato potrebbe indicare la presenza di molti macchinari a ciclo continuo, cioè che necessitano una costante fornitura di energia elettrica.

Cercando quali siano gli edifici con il consumo notturno maggiore in proporzione alla propria media, si vede che quelli con la minor differenza giorno-notte sono:

- U2 con una variazione del $7,40\%$ tra giorno e notte;
- U17 con una variazione del $9,08\%$;
- U3 con una variazione del $10,89\%$.

Nel caso dell'U17, questa differenza minima tra notte e giorno è dovuta ad un generale consumo ridotto anche durante il giorno. Questo edificio, come anticipato, è sede della segreteria studenti con orari di ricevimento che vanno dalle 9:30 alle 12:30 e non ha picchi elevati nei consumi diurni (massimo alle 9 con 5,97 kWh). Una possibile spiegazione per la così bassa differenza nel consumo notte - giorno per U2 e U3 è la presenza di molti macchinari di grandi dimensioni con costante richiesta di energia. Consultando il sito dell'Università, nella sezione "Grandi Attrezzature", è possibile vedere quali e quanti di questi macchinari siano effettivamente presenti in ogni edificio.

3.1.2 Settimana weekend

Altra interessante analisi riguarda la differenza nel consumo tra giorni lavorativi e festivi. Questo sviluppo può essere utile per individuare il livello di consumo minimo per ogni edificio, visualizzabile nel momento in cui questo dovrebbe essere chiuso al pubblico. Inoltre è possibile ricercare se ci sono valori particolarmente elevati quando non dovrebbe. È importante tenere a mente che U6 dalle 8:00 alle 14:00, U7 dalle 8:00 alle 18:30, e U8 dalle 8:00 alle 12:30 restano aperti anche il sabato. Gli edifici con il minor scarto tra consumo medio settimanale e feriale sono:

- U2 con una differenza del 5,05% (pari a 626,2 kWh),
- U1 con 7,97% (pari a 577,1 kWh),
- U3 con 8,45% (pari a 917,7 kWh),
- U4 con 9,16% (pari a 792,3 kWh).

Questi edifici sono particolari perché architettonicamente identici tra loro e questa somiglianza nella distribuzione dei consumi porta a pensare che anche il loro utilizzo sia molto simile. Uno stacco particolare si ha nel caso dell'U2 con una variazione minima tra giorni lavorativi e non, risultato che non stupisce dato che questo immobile risultava anche tra quelli con il consumo più simile tra notte e giorno. Vedendo queste peculiarità, e sapendo della presenza di molti macchinari all'interno di questi edifici, si spiega il motivo per cui il consumo sia distribuito in questo modo. Le attrezzature all'interno di questi edifici, che potremmo definire come adibiti a ricerca, necessitano di una costante alimentazione, implicando un consumo che si mantiene elevato anche nel momento in cui non dovrebbero essere accessibili al pubblico.

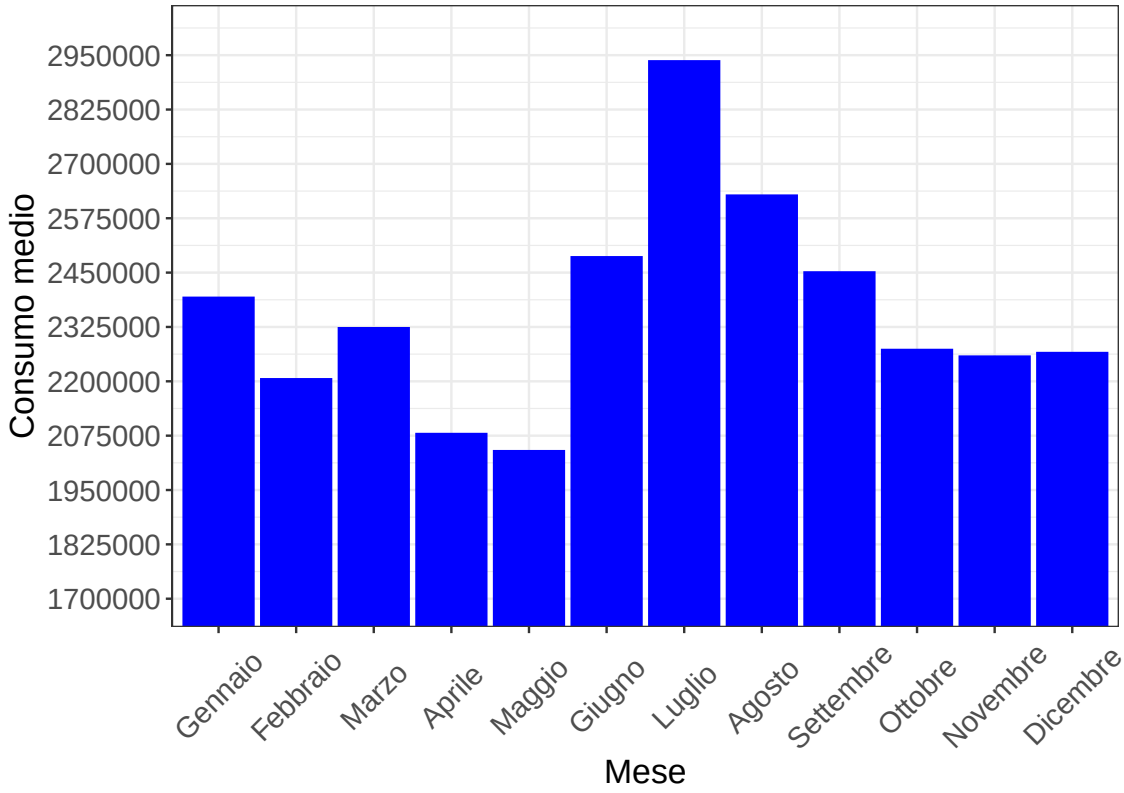

Figura 3.3: Consumo medio per mese in kWh.

3.1.3 Andamento mensile dei consumi

Utilizzando l'aggregazione degli edifici su base mensile si calcola il consumo medio, distintamente per mese. Come visibile nella Figura 3.3, il mese a maggior consumo medio risulta essere luglio con una media di 2.935.853 kWh, seguito da agosto con 2.627.026 kWh e giugno con 2.485.277 kWh. Negli edifici presi singolarmente il mese a maggior consumo medio è luglio, ad eccezione dell'U17 dove risulta essere giugno. I mesi estivi, infatti, sono caratterizzati da picchi a causa degli impianti di condizionamento che necessitano di una grande quantità di energia, caratterizzando la serie dei consumi con una forte componente stagionale. Il comportamento individuato non è una peculiarità degli edifici universitari ma replica l'andamento della domanda di energia in Italia (Terna, 2020).

In alcuni edifici questa componente stagionale è mascherata da un consumo elevato anche durante i mesi invernali, come nel caso dell'U4 che, avendo un consumo stabile ma abbastanza elevato, ha dei picchi estivi meno evidenti rispetto ad un edificio come l'U9, dove il mese di luglio ha uno scostamento del 45, 72% rispetto la media (Figura 3.4).

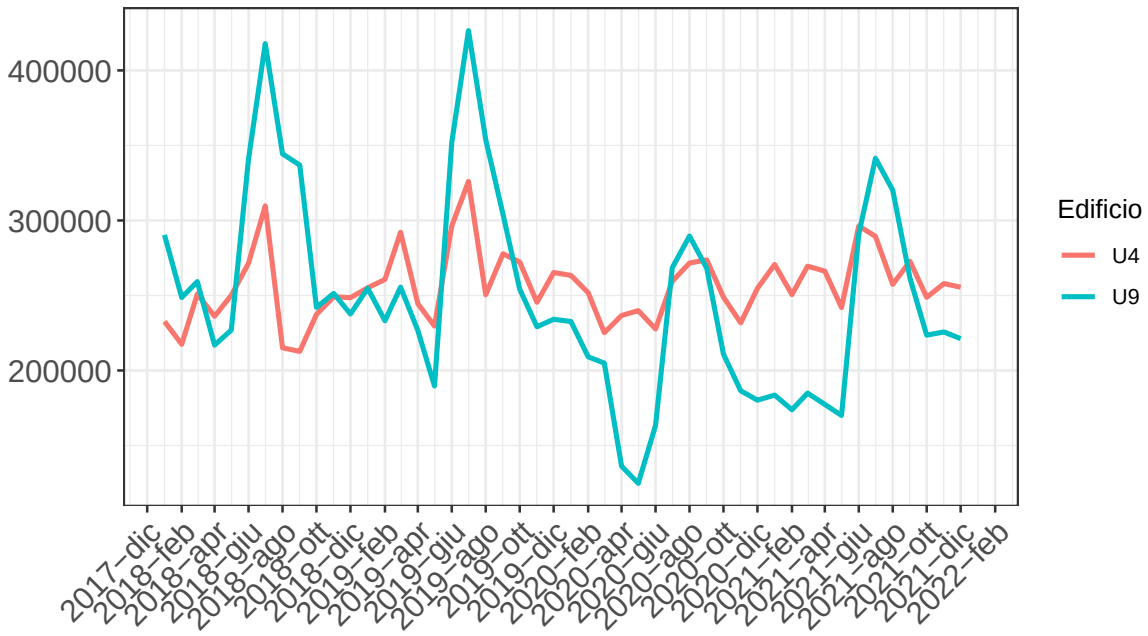

Figura 3.4: Andamento consumo mensile degli edifici U9 e U4

Andando più a fondo si riscontra una caratteristica particolare nell'U6 dove il secondo mese con consumo medio maggiore è gennaio seguito da novembre e dicembre. Un motivo per questo risultato potrebbe essere l'utilizzo del riscaldamento elettrico invece che a gas ma non sono disponibili notizie a riguardo. Date le informazioni a disposizione, la causa di consumi così alti nel periodo invernale risulta sconosciuta e potrebbe essere dovuta a qualche anomalia o a particolari utilizzi dell'edificio in questione che non sono riportati nel sito ufficiale dell'università.

3.2 Analisi di raggruppamento

In questa sezione gli edifici sono utilizzati come osservazioni da raggruppare al fine di valutare l'esistenza di gruppi di immobili con caratteristiche simili dal punto di vista del consumo. Tipicamente non si ha conoscenza sul numero e sulla natura dei gruppi, per questo si utilizzano metodi che, a partire dalle osservazioni disponibili, trovino la struttura migliore e a posteriori si cerca di interpretare i risultanti.

Il raggruppamento rientra nei metodi non supervisionati: infatti, le osservazioni oggetto di studio, non hanno una classe di appartenenza. Questa tecnica permette di trovare relazioni tra i dati sulla base delle caratteristiche

delle osservazioni stesse. L'obbiettivo è creare gruppi che siano molto simili all'interno e diversi all'esterno.

Per poter attuare questo studio è stato applicato il *raggruppamento gerarchico* che può essere suddiviso in due tipologie: *agglomerativo* e *divisivo*. Nel caso oggetto di studio è stato utilizzato un tipo agglomerativo.

Questo parte da uno stato in cui ogni osservazione costituisce un gruppo a sè; successivamente si calcola la matrice di distanze e, sulla base di questa, vengono aggregati i due gruppi (le due osservazioni nel primo passo) che risultano essere più simili. La sequenza di operazioni appena descritta prosegue fino a quando non si ottiene un unico gruppo contenente tutte le osservazioni.

Nel calcolare la distanza tra gruppi possono essere utilizzati diversi legami, in inglese *linkage*, sulla base dei quali si misura la distanza tra due gruppi di osservazioni. I *linkage* maggiormente noti sono:

- *singolo* in cui la distanza tra due gruppi è rappresentata dalla distanza minima tra due punti appartenenti ai due gruppi. Peculiarità di questo legame è il cosiddetto *effetto catena*: consente di trovare gruppi dalla forma particolare ma, al tempo stesso, rischia di legare osservazioni che non appartengono allo stesso insieme.

- *completo* in cui la distanza tra due gruppi è pari alla distanza tra i due punti più lontani tra i due gruppi. Caratteristica di questo legame è il fatto di trovare gruppi molto compatti al loro interno ma dalla forma circolare, con il pericolo di non cogliere gruppi dalla forma irregolare;

- *medio* in cui la distanza tra due gruppi è data dalla media delle distanze di tutte le combinazioni di punti tra i due insiemi.

Nel caso presente è stato utilizzato il legame medio siccome restituisce il valore maggiore di *silhouette media*. Questa metrica è utile a valutare la bontà del raggruppamento, maggiore è il suo valore, minore sarà la distanza all'interno dello gruppo e maggiore quella tra i gruppi. In particolare, per ogni osservazione i, si ha:

- A, gruppo di appartenenza di i;

- k numero di gruppi;

- C qualsiasi gruppo differente da A;

- $b(i) = \min_{C \neq A} d(i, C)$, distanza minima dell'osservazione *i* rispetto tutte le osservazioni degli altri gruppi;
- $a(i) = \frac{1}{n_k} \sum_{l:l \in A} d(i, l)$, distanza media di *i* dagli altri punti appartenenti a A.

Dati gli elementi elencati, si calcola la silhouette per l'unità u_{i^*} come:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i))\}}. \tag{3.1}$$

Questo coefficiente viene poi calcolato per tutte le osservazioni e per avere un valore riassuntivo della bontà del raggruppamento trovato, se ne calcola la media, trovando la silhouette media. Questa può avere valore compreso tra $[-1, 1]$, dove -1 è il caso non desiderabile (Kaufmann & Rousseeuw, 2012).

La matrice di distanze, che identifica quanto un immobile sia "vicino" ad un altro rispetto all'andamento del consumo, nel caso corrente è sviluppata attraverso la distanza di Canberra, calcolata nel seguente modo:

$$d(x_i, x_j) = \sum_{j=1}^{p} \frac{|x_i - x_j|}{|x_i| + |x_j|}, \tag{3.2}$$

dove i termini con denominatore pari a 0 sono esclusi. Da quanto si può vedere nell'equazione (3.2), questa distanza può essere interpretata, in un certo modo, come una distanza in termini relativi tra due punti; infatti, al numeratore si trova la differenza tra i due in valore assoluto che viene rapportata alla somma di essi.

Queste scelte dal punto di vista metodologico sono state utilizzate per fare raggruppamento sia sui dati espressi in kWh che in kWh/m².

3.2.1 Raggruppamento consumi kWh

Come anticipato tutte le analisi di raggruppamento sono state svolte escludendo gli edifici U8 e U17 siccome avevano dati presenti solo parzialmente negli anni. Utilizzando i consumi espressi in kWh, con la distanza di Canberra e il legame medio, si ottiene che il numero di gruppi che massimizza la silhouette (con un valore di 0,74) è 2. Visualizzando il *dendrogramma* nella Figura 3.5 è possibile notare l'ordine con cui sono aggregati gli edifici. Utilizzando due gruppi si vede che U14 e U16 sono a parte. Dai dati sulle superfici sappiamo che questi due edifici sono quelli con la metratura più ridotta e, come già detto, questa variabile è determinante nello stabilire il livello di consumo di un immobile. Il

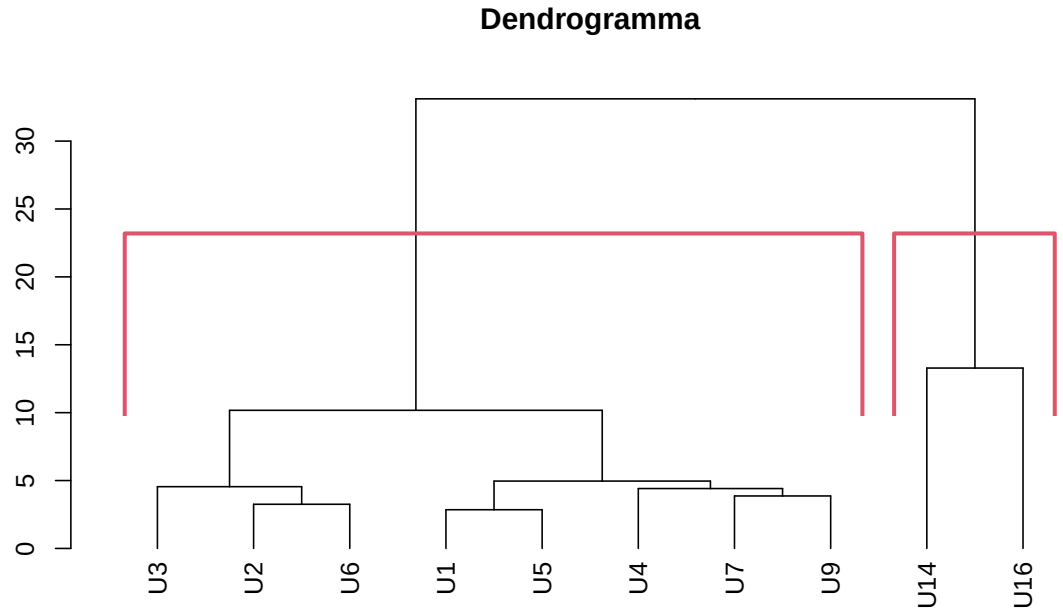

Figura 3.5: Dendrogramma raggruppamento gerarchico agglomerativo utilizzando distanza di Canberra e legame medio.

raggruppamento basato sui consumi in kWh sembra rilevare due gruppi sulla base del livello di consumo. Ne vengono individuati due perché non avendo scalato i dati per la superficie, la differenze nel consumo dei due edifici rimasti separati rispetto agli altri è significativa. Infatti, calcolando le statistiche principali sulla base del raggruppamento ottenuto si vede che il gruppo contenente U14 e U16 ha un consumo medio mensile di 50.229 kWh e uno scarto quadratico medio di 19.013 kWh mentre l'altro gruppo ha dei rispettivi valori di 282.544 kWh e 80.420 kWh. In questo caso, un fattore determinante nel classificare un edificio in un insieme piuttosto che in un altro è appunto la superficie.

Andando più a fondo in questo raggruppamento, se si dovesse suddividere il gruppo più popoloso si otterrebbe una suddivisione del tipo: U2, U3, U6 da una parte e U1, U5, U4, U7, U9 dall'altra. Il primo gruppo ha una media mensile di 360.275 kWh consumati con uno scarto quadratico medio di 58.080 kWh, il secondo avrebbe una media di 235.905 kWh con una variabilità simile pari a 50.209 kWh. Viene nuovamente fatta una suddivisione basata sul livello medio di consumo. In questo caso, però, la superficie non è più l'elemento discriminante

perché, ad esempio, U3 che ha una superficie abbastanza ridotta, viene classificato con U6 che è l'edificio con la maggior superficie in assoluto.

Un elemento distintivo del gruppo a consumo minore tra i due è la presenza di una chiara componente stagionale, visibile anche nel caso di U2, U3 e U6 ma è meno evidente a causa di ulteriori picchi durante l'anno. Riguardo questa caratteristica l'U4 sembrerebbe dover appartenere all'altro gruppo siccome non ha picchi stagionali particolarmente evidenti ma il fatto di avere un consumo medio relativamente ridotto se confrontato agli edifici dell'altro gruppo porta alla classificazione ottenuta.

Un dettaglio interessante consiste nel capire perché gli edifici di Piazza della Scienza (U1, U2, U3, U4) siano in due gruppi distinti pur avendo la stessa struttura. Visualizzando le statistiche riassuntive degli edifici in questione (Tabella 3.1) risulta chiaro: U1 e U4 (che sono inseriti nel gruppo a consumo inferiore) hanno effettivamente un utilizzo di energia elettrica decisamente più basso rispetto agli altri; la stessa conclusione vale per la variabilità dei consumi che si stabilizzano su livelli decisamente differenti (U2 e U3 hanno uno scarto quadratico medio di circa 48.300 kWh mentre U1 di 39.700 e U4 di 23.800).

3.2.2 Raggruppamento consumi kWh/m²

Utilizzando i dati del consumo riscalati rispetto alla superficie si ottiene che il numero ottimale di gruppi (in termini di massimizzazione della silhouette) risulta essere 4. Anche in questo caso gli edifici sono divisi sulla base dei livelli di consumo ma sono depurati dall'influenza della superficie. Vengono individuati quattro differenti livelli:

- nella fascia di consumo alta si trovano U1 e U3 con un consumo medio mensile di $26,65$ kWh/m²;
- nella fascia medio-alta sono presenti U2, U3 e U5 con consumo medio mensile di $14,86$ kWh/m²;
- nella fascia medio-bassa con U9, U14 e livello di consumo di $9,91$ kWh/m²;
- nella fascia bassa con U6, U7 e U16 con un consumo di $5,30$ kWh/m².

Il fatto che la superficie non risulti più discriminante nel raggruppare i dati è evidente dal fatto che U16, secondo edificio più piccolo, è classificato negli edifici a consumo basso con U6 e U7 che sono gli edifici più grandi.

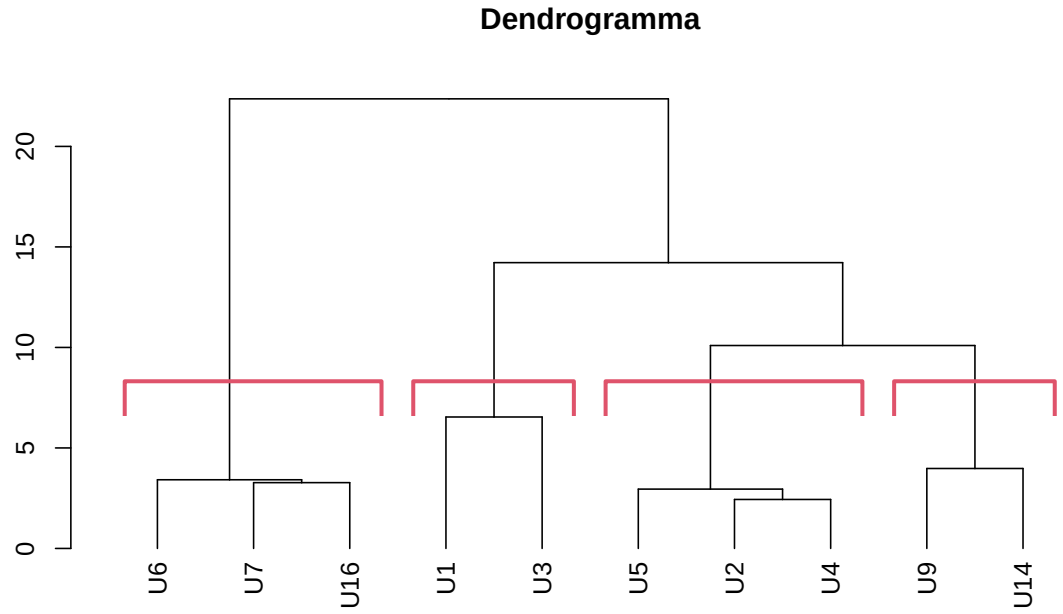

Figura 3.6: Dendrogramma raggruppamento gerarchico agglomerativo utilizzando distanza di Canberra e legame medio su dati in kWh/m².

I raggruppamenti trovati non possono essere utilizzati per trarre conclusioni sull'efficienza del consumo di un gruppo rispetto ad un altro, perché ogni edificio può avere differenti utilizzi con conseguenti livelli di consumo. Nei gruppi a consumo alto e medio-alto sono riportati edifici con la maggior concentrazione di grandi attrezzature. Questi potrebbero essere definiti come edifici di "ricerca", nel senso che al loro interno sono svolte operazioni che necessitano di un determinato tipo di attrezzature, con un conseguente livello di consumo elettrico elevato. Gli altri due gruppi contengono edifici in cui sono presenti macchinari dal dispendio energetico inferiore (almeno in linea teorica) e in cui la principale parte della superficie è adibita ad aule o uffici, sono la classe di edifici "amministrativi". Si potrebbe dire che il differente livello di consumo trovato con questi dati identifica le differenti attività svolte all'interno degli immobili.

Ulteriore analisi per confermare quanto appena affermato è quella di confrontare l'effetto delle chiusure a causa della pandemia sui gruppi di edifici trovati. Sapendo che la ricerca universitaria ha continuato a svolgersi anche durante le chiusure, potremmo individuare gli edifici adibiti a ricerca o comunque quelli in cui questa attività è svolta in modo consistente dal punto di vista del dispendio

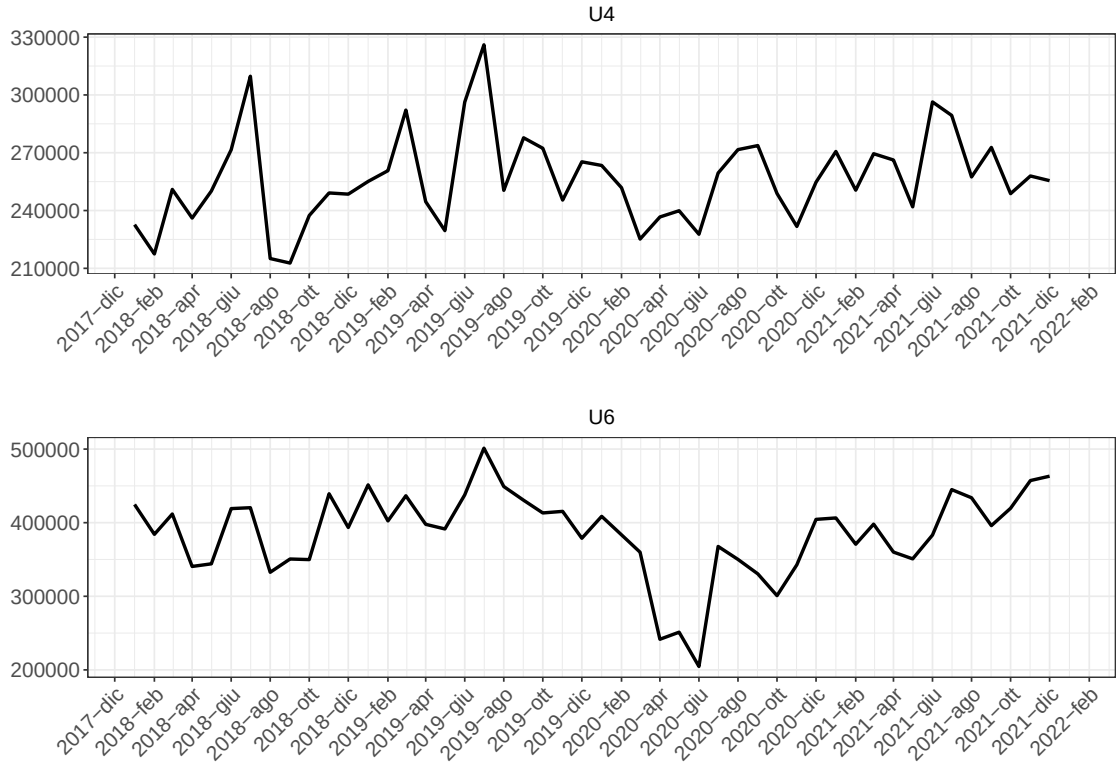

Figura 3.7: Andamento serie mensile degli edifici U4 e U6.

energetico. I fabbricati nelle fasce di consumo alta e medio-alta sono quelli per cui il cavo di marzo-aprile 2020 è decisamente meno evidente. Nelle altre due fasce di consumo, invece, i cavi sono maggiormente pronunciati, ad indicare che in questi edifici le principali funzioni sono quelle di uffici e aule per le lezioni, motivo per cui hanno risentito maggiormente delle chiusure. Due edifici esemplari di questa differenza sono l'U4 (appartenente alla fascia di consumo medio-alta) e l'U6 (fascia medio-bassa). Vedendo la Figura 3.7 è evidente il differente effetto delle chiusure sui due edifici. Infatti, U6 nel mese di giugno 2020 ha raggiunto il punto di minimo assoluto nei consumi mentre l'U4 ha avuto una riduzione che potremmo definire in linea con l'andamento generale della sua serie storica.

3.3 Meteo

Riprendendo quando detto nella ricerca svolta dall'ente Enea sui consumi elettrici di edifici scolastici (Corgnati et al., 2010), uno dei fattori influenzanti il consumo è la temperatura esterna. Volendo approfondire questo aspetto, sono stati reperiti i dati delle temperature nella zona dell'Università degli studi di Milano Bicocca

tramite il sito di Arpa Lombardia (Arpa, 2022). Grazie a questi dati è stato svolto uno studio sulle correlazioni tra consumi e temperature esterne, rispetto le varie aggregazioni temporali, ed è stato riscontrato che, tra i gruppi precedentemente individuati, quello con maggior correlazione tra le due variabili è la fascia medio-bassa contenente U9 e U14 con dei valori rispettivamente di 0.57 e 0.72. L'U14 è l'edificio con la correlazione più alta in assoluto tra consumo medio mensile e temperatura media. Al contrario, il gruppo di edifici con andamento più scollegato rispetto le temperature è quello della fascia medio-bassa contenente U6, U7 e U16. I coefficienti di correlazione per questi edifici rispetto al calore esterno sono rispettivamente di $-0,13$, $0,01$ e $0,28$. In generale, la correlazione tende a diminuire leggermente all'aumentare della granularità del dato, in risposta alla maggiore volatilità.

Questi dati non devono portare a pensare che l'edificio con consumo maggiormente correlato alla temperatura esterna sia meno efficiente in termini di isolamento termico.

L'elevata correlazione rispetto alla temperatura esterna potrebbe essere dovuta ad un sistema di raffreddamento che è particolarmente dispendioso e/o inefficiente dal punto di vista energetico, causando picchi nei consumi elettrici. In ogni caso, soprattutto nel periodo estivo, è norma che i consumi aumentino a causa della temperatura esterna, come già accennato facendo riferimento alla domanda di energia elettrica italiana (Terna, 2020). Infatti, procedendo in questa direzione, si può vedere che la correlazione tra consumi elettrici giornalieri e temperatura, calcolata esclusivamente nei mesi da maggio a settembre è particolarmente elevata, anche per edifici che nel resto dell'anno non consumano in modo particolarmente legato alla temperatura esterna, come è possibile vedere dalla Tabella 3.3.

U3 è l'edificio con consumo estivo maggiormente correlato alla temperatura con un valore di 0.734. Cercando di vedere quanta parte del consumo elettrico di questo edificio sia effettivamente dovuta all'impianto di condizionamento, vedendo la serie dei consumi giornalieri con sovrapposto l'andamento delle temperature, si nota che intorno al 22 giugno 2020 è presente un picco piuttosto repentino nei consumi. Confrontando con gli altri edifici si nota che questi picchi sono sempre collocati più meno negli stessi giorni. Si potrebbe ipotizzare che nell'intorno trovato sia stato attivato l'impianto di condizionamento con il conseguente picco nei consumi. Studiando il dato puntualmente e ipotizzando che l'impianto di condizionamento abbia effettivamente iniziato a funzionare il

Edificio	Correlazione Anno	Correlazione Estiva
U1	0.526	0.585
U2	0.245	0.603
U3	0.506	0.734
U4	0.23	0.455
U5	0.59	0.674
U6	−0.112	0.322
U7	0.001	0.415
U9	0.507	0.655
U14	0.621	0.684
U16	0.177	0.554

Tabella 3.3: Correlazioni calcolate tra consumi giornalieri e temperatura per tutto l'anno e solo per i mesi da maggio a settembre.

23 giugno dell'anno in questione (non ci sono informazioni sull'effettiva data) si vede un passaggio da 8.940 kWh consumati il 22 giugno a 11.279 kWh il 23; si tratta di un aumento del 26%. Dato che l'aumento potrebbe essere dovuto anche ad altri fattori, per togliere questa eventuale distorsione si è deciso di confrontare il consumo medio della settimana precedente (tenendo conto solo dei giorni infrasettimanali) e quella successiva il 23 giugno. La settimana precedente si ha un consumo medio giornaliero di 8.903 kWh, la successiva di 12.540 kWh, corrispondente ad un aumento del $40,85\%$. Sotto le opportune ipotesi, questa differenza spiegherebbe la stagionalità presente nei mesi estivi.

3.4 Shiny

Gran parte delle analisi svolte sono state raccolte per costruire un'applicazione interattiva disponibile al link `https://bicocca-datalab.shinyapps.io/consumi-edifici/`. Questa è stata realizzata utilizzando la libreria `Shiny` di `R`. L'applicazione è stata creata per generare uno strumento utile per i responsabili della gestione energetica dell'università. Al suo interno troviamo la suddivisione in diverse schermate:

- nella sezione *Grafici* è possibile selezionare l'aggregazione (mensile, settimanale, giornaliera), il tipo di dato (kWh o kWh/m^2), l'edificio di interesse, il mese iniziale e finale da visualizzare. Sulla base della selezione dell'utente

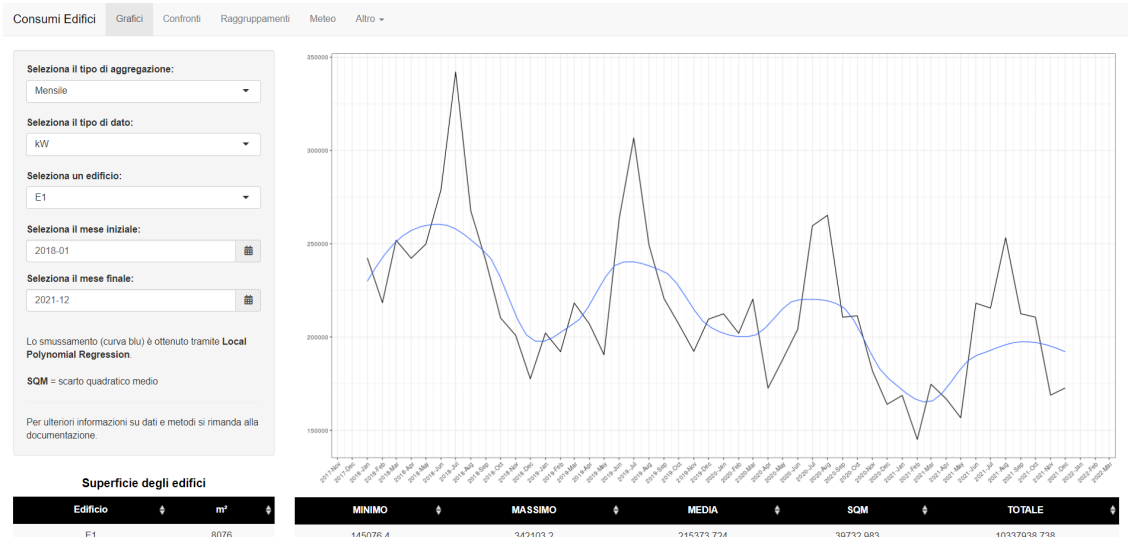

Figura 3.8: Sezione Grafici applicazione in cui è presente il pannello laterale con le possibili scelte e la visualizzazione della serie storica dei consumi.

vengono riportati i grafici con l'andamento della serie storica dei consumi, le statistiche principali per edificio, la superficie di tutti gli edifici e i dati puntuali dei consumi.

- nella sezione *Confronti* viene selezionato il tipo di dato, gli edifici da confrontare, l'immobile rispetto al quale verranno calcolate le variazioni nelle statistiche principali, mese iniziale e finale da visualizzare. In questa sezione viene prodotto un grafico con le serie degli edifici selezionati sovrapposte, le statistiche principali ogni edifici con la variazione rispetto a quello selezionato come base di confronto ed è riportata la tabella con le superfici degli edifici.

- nella sezione *Raggruppamenti* viene selezionato il tipo di dato e il numero di gruppi da visualizzare. Il valore prestabilito per quest'ultimo parametro è quello che massimizza la silhouette media. Viene riportato il dendrogramma, la tabella delle superfici e le statistiche principali calcolate sulla base del raggruppamento ottenuto.

- nella sezione *Meteo* le opzioni da selezionare sono le stesse della schermata Grafici. In questo caso viene riporta la tabella con le superfici, il grafico della serie storica dei consumi con sovrapposto l'andamento della temperatura esterna e una tabella riportante i suoi valori puntuali.

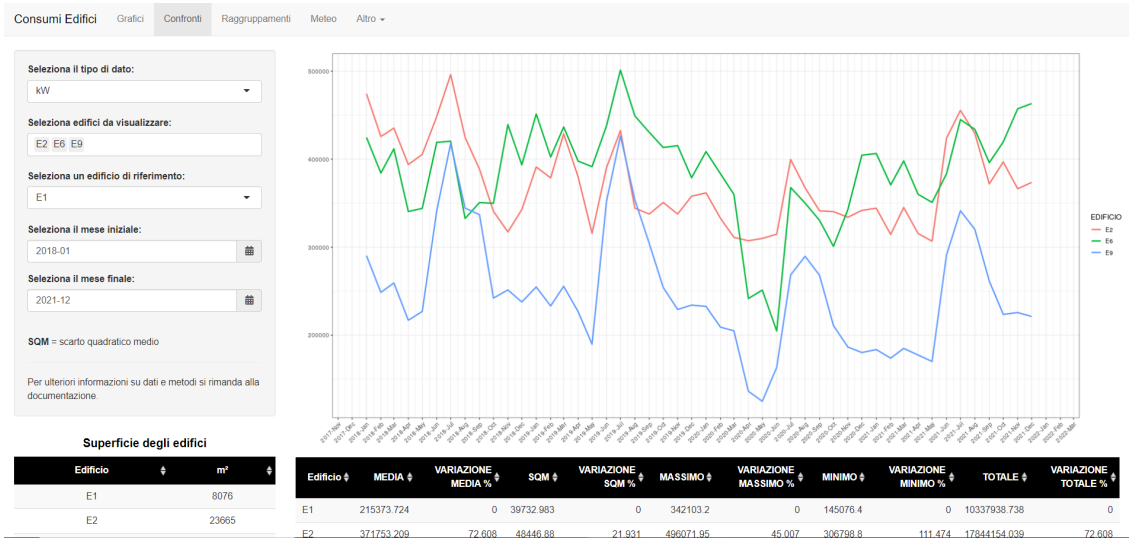

Figura 3.9: Schermata Confronti con grafici sovrapposti e statistiche principali confrontate rispetto ad un edificio.

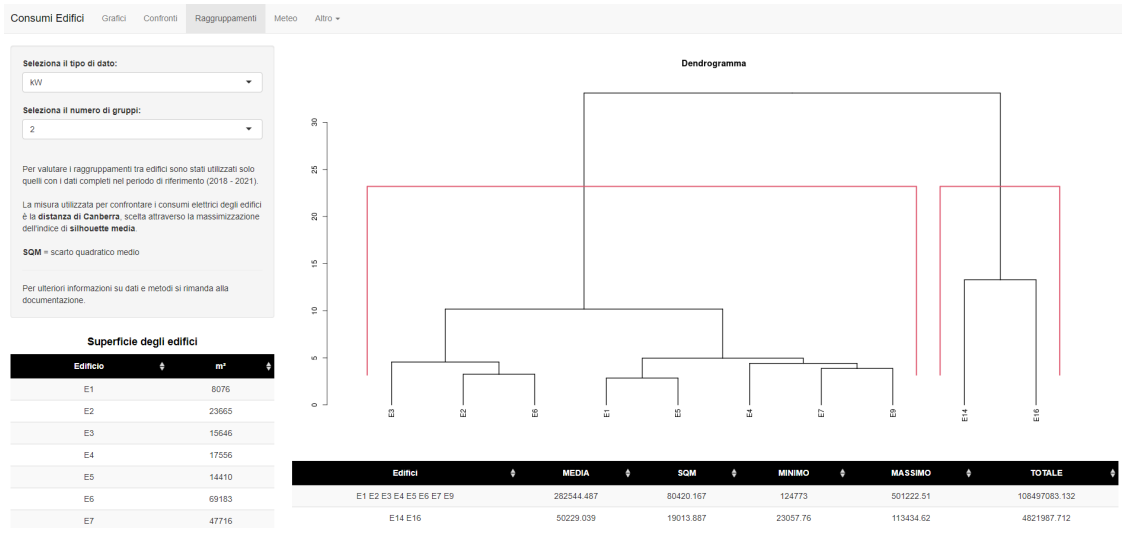

Figura 3.10: Schermata Raggruppamenti rappresentante il dendrogramma dell'aggregazione ottenuta e le statistiche calcolate distintamente per gruppo.

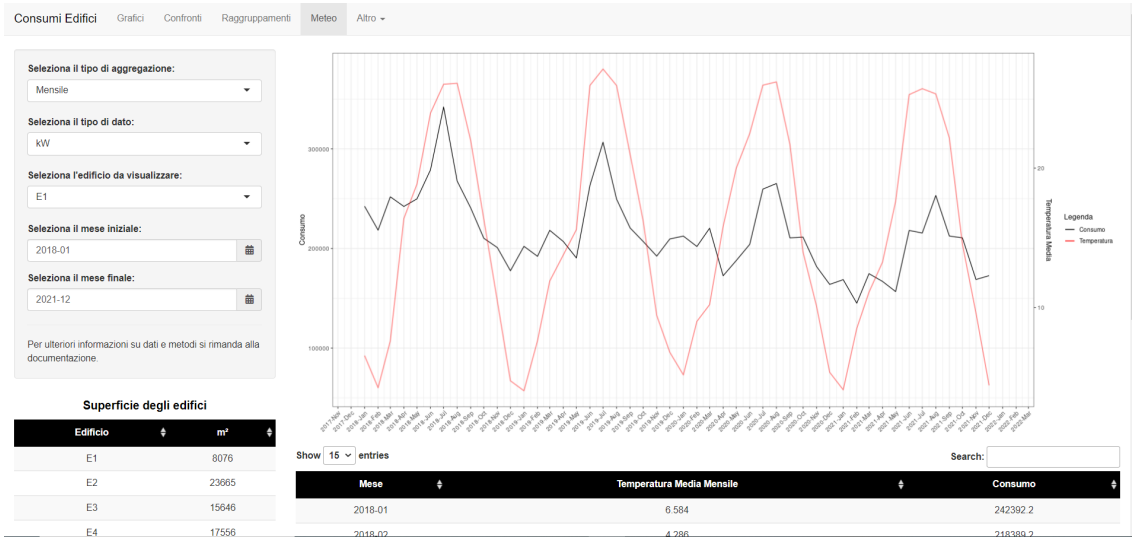

Figura 3.11: Schermata Meteo con andamento dei consumi con sovrapposta la variazione della temperatura esterna.

- l'opzione *Altro* è suddivisa in *Guida all'utilizzo* in cui viene spiegato cosa si trova in ogni sezione precedentemente esposta e una *Documentazione* in cui sono spiegati i metodi utilizzati e le scelte fatte durante lo sviluppo dell'applicazione.

Questa applicazione è reperibile online al link riportato precedentemente e in bibliografia (Carrettoni & Mannarino, 2022).

Capitolo 4

Regressione spline e clustering funzionale

In questa sezione viene approfondito il raggruppamento degli edifici cercando di aggregare gli immobili sulla base dell'andamento del consumo e non solo rispetto al livello medio. Per poter ottenere quanto anticipato, l'idea è quella di standardizzare la serie storica dei consumi di ogni edificio, cioè renderla a media nulla e varianza unitaria. Successivamente viene attuata una *regressione spline* per ogni serie grazie al comando `bs` nella libreria `splines`. Questo comando permette di creare una base di funzioni che saranno utilizzare come regressori nel modello per spiegare i consumi. In relazione al numero di basi spline scelte si otterrà una matrice $10 \times M$, dove M è appunto il numero di basi utilizzate e svolge la funzione di *parametro di lisciamento*: maggiore sarà il numero di basi utilizzate, minore sarà lo smussamento ottenuto nella serie e viceversa. La matrice $10 \times M$ è poi utilizzata come regressore nella funzione `lm` attraverso la quale si ottiene la matrice di coefficienti. Su quest'ultima si calcolano le distanze e viene attuato il raggruppamento. Facendo in questo modo, due edifici risulteranno simili nel momento in cui i loro coefficienti hanno valori vicini.

Andando nel dettaglio, in questo caso applicando la distanza di Canberra (3.2) o la classica euclidea non ci sono differenze nel raggruppamento ottenuto, quindi è stata scelta la seconda calcolata come:

$$d^2(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \qquad \text{con} \quad x_i = (x_{i1}, ..., x_{ip})^T. \tag{4.1}$$

Per la selezione del numero di basi M da utilizzare è stato svolto uno studio a livello grafico con l'ausilio del *Criterio di informazione Bayesiano* (BIC) per valutare il modello lineare. Questo criterio è utilizzato per la selezione del miglior numero di parametri. Infatti, nei principali metodi statistici applicati per la

stima di parametri ignoti si utilizza la massimizzazione della *log-verosimiglianza*. Nel massimizzare questa misura bisogna tener conto del numero di parametri utilizzati, introducendo una penalità adeguata per i regressori aggiuntivi. Questa è introdotta perché all'aumentare del numero di parametri, la log-verosimiglianza aumenta o al limite rimane invariata; con la penalizzazione per i regressori aggiuntivi si evita di utilizzare modelli eccessivamente complicati. La formula del BIC è:

$$\mathrm{BIC} = -2 \log \mathrm{L}(\hat{\theta}) + p \log n, \tag{4.2}$$

dove $\mathrm{L}(\hat{\theta})$ è il valore massimo della verosimiglianza ottenuto in corrispondenza del parametro $\hat{\theta}$ e p il numero di parametri utilizzati nel modello.
Nel caso in esame sono stati testati i gradi di libertà da 3 a 30 ed è risultato che il valore minimo del BIC $(123,64)$ si ottiene nel caso di 11 gradi di libertà. Insieme allo studio grafico si è valutato che il valore ottimale è pari a 12, dove 3 sono per il grado dei polinomi utilizzati e 9 sono i nodi. L'aggiunta di un grado di libertà rispetto al caso ottimale non porta ad un eccessivo aumento del BIC, che arriva ad una valore di $127,97$. Se invece il numero di parametri fosse stato scelto solo sulla base della massimizzazione della log-verosimiglianza i gradi di liberà sarebbero stati 28. In questo modello però la maggior parte dei parametri non risulterebbe significativa nello spiegare l'andamento del consumo nel tempo.

Il legame utilizzato per calcolare le distanze tra gruppi è quello completo perché genera insiemi maggiormente interpretabili. Il numero ottimale di gruppi è 3, valore che massimizza la silhouette media, anche se il valore ottenuto è basso $(0,27)$. Il dendrogramma risultante è visualizzabile nella Figura 4.1. In questo caso i gruppi ottenuti non sono influenzati dal livello di consumo medio e tanto meno dalla superficie (basta vedere, ad esempio, che U2 e U16 sono nello stesso gruppo nonostante il primo abbia una superficie più di tre volte maggiore rispetto al secondo). Visualizzando l'andamento delle serie nel tempo, distintamente per gruppo (Figura 4.2), si può vedere che il primo, contenente U1, U2, U9 e U16, è caratterizzato maggiormente da una tendenza decrescente nel periodo di riferimento (2018-2021) ed è particolarmente evidente la stagionalità del consumo nei mesi estivi. La tendenza decrescente è visibile anche grazie al fatto che questi edifici sono tra quelli con il consumo elevato nel 2018, come è possibile vedere dal grafico. Nel primo gruppo è presente l'U1 che sappiamo essere l'edificio pilota per un progetto volto a rendere efficiente il consumo energetico degli edifici dell'Università. Il fatto che appartenga al primo gruppo, in cui il trend

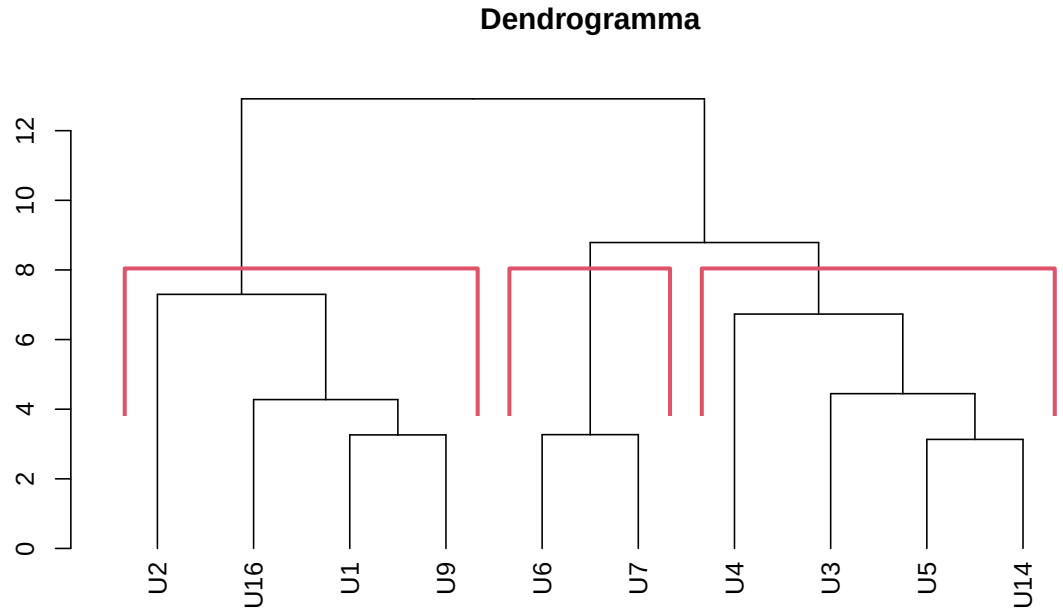

Figura 4.1: Dendrogramma raggruppamento gerarchico agglomerativo utilizzando distanza euclidea e legame completo sulla matrice dei coefficienti.

decrescente è maggiormente marcato, conferma che le misure attuate sembrano portare a riscontri concreti nella riduzione del consumo. Il secondo gruppo, che comprende U3, U4, U5 e U14, ha un andamento stabile, non si nota una tendenza particolare e persiste la stagionalità nei consumi che però è leggermente mascherata nei mesi estivi del 2020 a causa di due edifici in cui non si verifica una significativa riduzione dei consumi nella stagione invernale, come avviene solitamente. I due edifici citati sono U3 e U4 ma non sono disponibili dati che possano giustificare un dispendio energetico così elevato nel periodo invernale. Il terzo gruppo, composto da U6 e U7 è quello con andamento decisamente differente rispetto ai precedenti: i consumi di questi due edifici sono quelli che hanno subito la maggior riduzione nel periodo di aprile-maggio 2020 a causa delle chiusure dovute alla pandemia. Inoltre nell'andamento di questo gruppo non si nota una stagionalità così marcata che, anzi, sembrerebbe essere quasi del tutto assente. Si vede in modo chiaro il momento in cui sono state ridotte le chiusure perché nei consumi si verifica una crescita piuttosto repentina, in un primo momento, nei mesi estivi del 2020, probabilmente per l'accensione degli impianti di condizionamento, successivamente con la ripresa delle lezioni

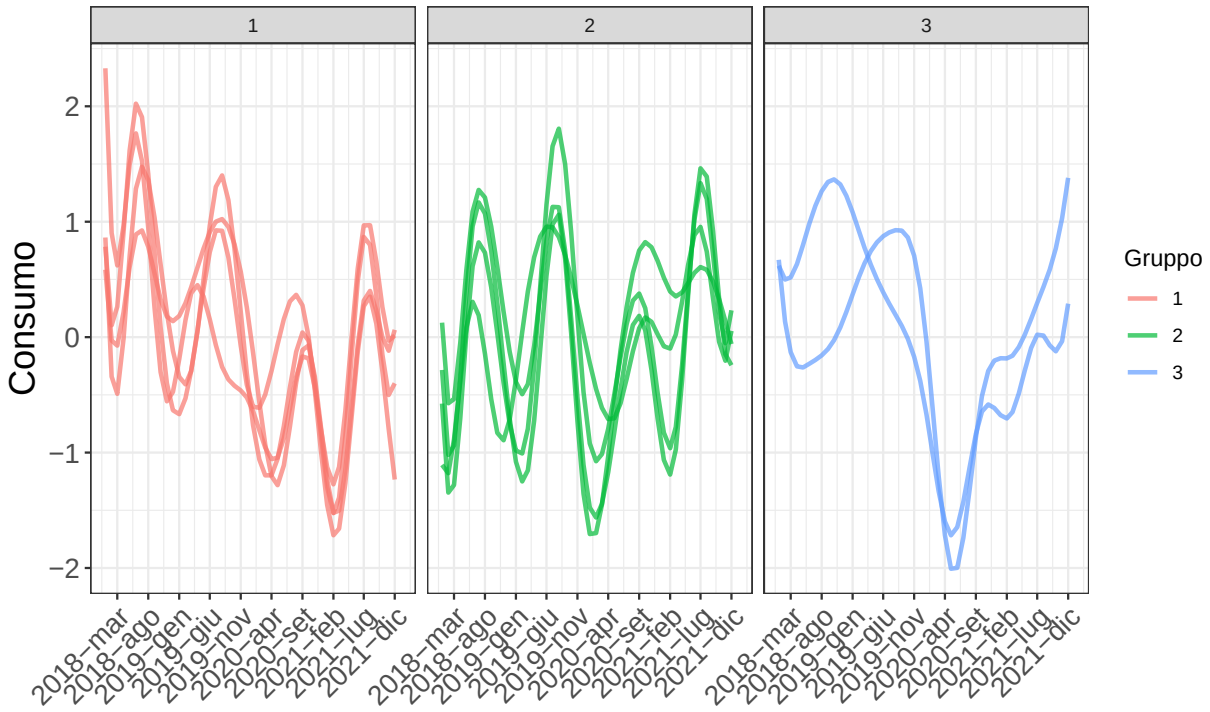

Figura 4.2: Andamento dei consumi nel tempo distintamente per gruppo ottenuto facendo raggruppamento sui coefficienti ottenuti con regressione spline.

in presenza per le matricole ad aprile 2021. U6 e U7 hanno subito maggiormente il calo dei consumi a causa delle chiusure perché sono gli edifici con il maggior sviluppo di attività amministrative al loro interno e hanno anche la mensa, la cui chiusura potrebbe implicare un'ulteriore riduzione dei consumi.

In conclusione, l'algoritmo utilizzato sembra essere in grado di raggruppare gli edifici sulla base dell'andamento del consumo nel tempo, identificando in modo preciso le differenze negli andamenti come la differenza nei picchi stagionali tra i gruppi e delle discrepanze nella riduzione dei consumi dovuta alle chiusure.

Capitolo 5

Conclusioni

In conclusione al lavoro svolto è possibile notare che differenti tecniche di raggruppamento possono essere utilizzate sui dati riguardanti i consumi degli edifici dell'Università degli studi di Milano Bicocca.

Sulla base delle scelte fatte per raggruppare gli immobili è possibile dividere gli edifici rispetto a differenti caratteristiche: nel caso del raggruppamento basato sui dati espressi in kWh, utilizzando la distanza di Canberra e legame medio si ottiene una divisione in due gruppi in cui la variabile discriminante risulta essere la superficie dell'edificio con un gruppo contenente gli edifici di più grandi dimensioni e l'altro con i due edifici più piccoli. Nel caso dei dati espressi in kWh/m², vengono individuati 4 gruppi in base al livello di consumo medio. Questo consumo medio non ha più l'influenza della superficie al suo interno siccome ogni valore è riportato rispetto al metro quadro. Il raggruppamento ottenuto individua i differenti utilizzi degli edifici con la suddivisione in immobili prevalentemente ad uso amministrativo e immobili di ricerca. Nell'ultima tecnica utilizzata, normalizzando i dati, cioè rendendo le serie a media nulla e varianza unitaria, viene fatto il raggruppamento sui coefficienti della regressione spline attuata per ogni edificio. In questo modo vengono individuati gruppi di edifici (in particolar modo 3) che hanno un andamento simile del consumo nel tempo, riuscendo ad intercettare i picchi stagionali e le differenti risposte dei consumi rispetto alle chiusure avvenute nel periodo di riferimento (2018-2021).

Grazie a questi raggruppamenti è possibile vedere quali edifici siano simili e in base a quale caratteristica. Questa conoscenza può essere utile al fine di applicare specifiche politiche di contenimento e monitoraggio dei consumi su specifici gruppi di immobili.

Bibliografia

Arpa (2022). Dati sulle temperature. https://www.arpalombardia.it/Pages/Meteorologia/Previsioni-e-Bollettini.aspx#/topPagina.

Azzalini, A. & Scarpa, B. (2012). *Data analysis and data mining: An introduction*. OUP USA.

Carrettoni & Mannarino (2022). Shiny app: Consumi edifici. https://bicocca-datalab.shinyapps.io/consumi-edifici/.

Centrale, R. (2022). La storia. https://www.unimib.it/ateneo/storia.

Corgnati, S. P. et al. (2010). Edifici tipo, indici di benchmark di consumo per tipologie di edificio, ad uso scolastico (medie superiori e istituti tecnici) applicabilità di tecnologie innovative nei diversi climi italiani. https://www.enea.it/it/Ricerca_sviluppo/documenti/ricerca-di-sistema-elettrico/fabbisogni-consumi-energetici/7-polito.pdf.

Kaufmann & Rousseeuw (2012). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

programmazione e controllo, S. (2022a). Dati sugli studenti. https://trasparenza.unimib.it/amministrazione-trasparente/altri-contenuti/ateneo-cifre/dati-sugli-studenti#:~:text=laurea%20nell'a.-,a.,il%20suo%20apice%20nell'a.

programmazione e controllo, S. (2022b). Dati sulle infrastrutture. https://trasparenza.unimib.it/amministrazione-trasparente/altri-contenuti/ateneo-cifre/dati-sulle-infrastrutture.

Terna (2020). L'evoluzione del mercato elettrico: tutti i dati. https://www.terna.it/it/sistema-elettrico/statistiche/evoluzione-mercato-elettrico.