Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile

Ministero della Transizione Ecologica

Ricerca di Sistema elettrico

Analisi di dati energetici di illuminazione pubblica (PELL IP) attraverso metodi di analisi lineari e non lineari

Fabio Leccese, Mariagrazia Leccisi

ROMA TRE UNIVERSITÀ DEGLI STUDI

Report Rd/PTR(2021)/035

Analisi di dati energetici di illuminazione pubblica (PELL IP) attraverso metodi di analisi lineari e non lineari

Fabio Leccese, Mariagrazia Leccisi (Università degli Studi “Roma Tre”)

Dicembre 2021

Report Ricerca di Sistema Elettrico

Accordo di Programma Ministero della Transizione Ecologica-ENEA
Piano Triennale di Realizzazione 2019-2021 - III annualità
Obiettivo: Tecnologie
Progetto: Tecnologie per la penetrazione efficiente del vettore elettrico negli usi finali
Work package: Local Energy District
Linea di attività: 1.45 - Consolidamento del PELL in relazione agli edifici pubblici e analisi dei dati della Smart Road
Responsabile del Progetto: Claudia Meloni, ENEA
Responsabile del Work package: Claudia Meloni, ENEA

Il presente documento descrive le attività di ricerca svolte all’interno dell’Accordo di collaborazione “Smart Energy in Sistemi Pubblici: analisi di affidabilità e qualificazione dei dati per ridurre le incertezze di sistema”
Responsabile scientifico ENEA: Dott. Francesco Pieroni.

Responsabile scientifico Università degli Studi “Roma Tre”: Prof.re **Fabio Leccese.**

Indice

Sommario

Il presente documento descrive l'attività di ricerca inerente l'analisi di dati energetici provenienti dagli smart meters presenti sui quadri elettrici degli impianti di illuminazione pubblica distribuiti sul territorio e gestiti dal Sistema PELL-IP, utilizzando metodi di analisi di tipo lineare e non lineare al fine di studiare sia l'andamento dei consumi energetici di un impianto, sia la presenza di anomalie nello stesso.
A tal fine sono stati utilizzati i dati forniti da ENEA, come input per l'utilizzo di tecniche di classificazione lineari, in particolare la PCA (Principal Component Analysis), e non lineari con particolare riferimento alle reti neurali SOM (Self Organizing Map).
Lo scopo di tale analisi è quella di classificare i dati di input in gruppi di dati simili tra loro, per evidenziare sia la coerenza dell'andamento dei consumi, ma anche eventuali dati anomali. Attraverso il successivo studio dei dati anomali classificati dagli algoritmi, ed analizzati con modalità visuali come grafici specifici, è stato possibile evidenziare anomalie o problematiche presenti sull'impianto, che possono poi essere segnalate ai tecnici per le opportune verifiche.
Il risultato ottenuto è stato l'implementazione di due tecniche di analisi dati, una rete neurale (Self Organizing Map) e l'algoritmo dell'Analisi delle Componenti Principali, tramite MATLAB, che ha mostrato come questi metodi siano efficaci nel classificare questi dati, evidenziandone le caratteristiche attraverso modalità grafiche.
In particolare la rete neurale completa il risultato della PCA, che presenta alcune limitazioni, permettendo di dettagliare in modo più specifico eventuali problematiche evidenziate dalla classificazione ottenuta.

1 Introduzione

Il Laboratorio di Misure Elettriche ed Elettroniche del Dipartimento di Scienze dell'Università degli Studi "Roma Tre" è stato coinvolto da ENEA nell'Accordo di Programma tra Ministero dello Sviluppo Economico ed ENEA-Piano Triennale di Realizzazione 2019-2021-nel quale il Laboratorio è un Co-beneficiario. Il nostro Laboratorio è stato interessato per una attività di ricerca dal titolo "Smart Energy in Sistemi Pubblici: Analisi di Affidabilità e Qualificazione dei Dati per Ridurre le Incertezze di Sistema".

Per la nostra ricerca si è da subito palesata la necessità di dover gestire una disponibilità di dati enorme sia in senso orizzontale, per la grande quantità di grandezze elettriche a disposizione, sia in senso verticale, per la grande disponibilità di occorrenze che giornalmente vengono raccolte per ogni singola grandezza.

Lo scorso anno, attraverso l'analisi della letteratura in merito, si è lavorato all'individuazione di algoritmi e metodiche di elaborazione e studio che potessero essere utilizzate nell'ambito dei Big Data energetici.

Lo studio della letteratura fatta lo scorso anno e l'osservazione dei dati a nostra disposizione e forniti da ENEA, ci ha suggerito di focalizzare l'attenzione su un aspetto peculiare dei nostri dati che ci ha consentito di analizzare il nostro data set nel modo più accurato ed efficiente possibile: la loro granularità temporale. Infatti, la disponibilità dei dati sull'energia elettrica è di tipo quartorario. Questa risoluzione temporale non è facilmente rintracciabile in letteratura nella quale si trovano studi su dati aggregati con risoluzione tipicamente del giorno o, in casi particolari, dell'ora. Può anche accadere di avere disponibilità di dati con elevata risoluzione temporale, ma o si analizza una singola grandezza elettrica e non tutte le variabili di cui il nostro data set è composto, o l'analisi è limitata ad un piccolo arco temporale (ora od al massimo giorno).

La granularità quartoraria prevede campioni di tempo di 15 minuti ciascuno, che si traducono in 96 campioni al giorno per ciascuna grandezza; questo evidenzia una elevata complessità del sistema e delle conseguenti analisi. Le inevitabili difficoltà nella gestione di un data set così ricco, ci ha portato a valutare metodiche di analisi che permettessero una riduzione della complessità del sistema semplificando la sua analisi e studiandone la parte statisticamente più significativa.

Tra i vari approcci abbiamo considerato in primis la PCA (Principal Component Analysis) e successivamente le reti neurali ed in particolare quelle di tipo SOM (Self Organizing Map).

Entrambi i metodi permettono di classificare i vari dati in gruppi che supportano medesime caratteristiche; questo permette di far emergere le caratteristiche di ogni singolo gruppo. Tuttavia essi lavorano in modi profondamente differenti: la PCA è una metodologia di analisi tipo lineare che si basa sul calcolo della varianza, mentre le reti SOM sono una metodologia di analisi non lineare che si basa sulla autoorganizzazione di una struttura neuronale in base alla distanza euclidea tra i vari neuroni.

Entrambi i metodi sono risultati efficienti per la classificazione dei dati a nostra disposizione e sono state evidenziate alcune casistiche particolari che possono dare indicazioni circa la bontà dei dati e, indirettamente, dell'impianto monitorato.

La rete SOM completa i risultati della PCA, in quanto, oltre alla classificazione, fornisce anche una quantificazione della somiglianza tra i dati di più cluster, grazie anche ad informazioni grafiche estremamente chiare nella loro interpretazione.
Al fine di ampliare le casistiche di studio, abbiamo simulato circa 800 osservazioni, ottenendo una buona capacità di classificazione dei dati, riducendo la complessità del problema a pochi clusters da analizzare.

2 PELL-IP

Prima di dettagliare le attività specifiche relative a questa annualità, arrivati al terzo ed ultimo anno di questo progetto condiviso insieme ad ENEA, riteniamo sia utile riassumere brevemente quanto fatto in precedenza in modo da consentire al lettore di verificare come le attività impostate nei due anni precedenti siano state propedeutiche a quelle implementate quest'anno. Durante il primo anno ci siamo occupati dello studio del PELL-Public Energy Living Lab [1-3], una piattaforma per la raccolta, gestione, organizzazione e valutazione di dati strategici statici e dinamici relativi alle infrastrutture energivore urbane. Gran parte dell'attività lavorativa è stata volta a migliorare il sistema, al testing delle funzionalità e alla verifica e risoluzione di anomalie del sistema. A questo scopo, l'applicazione è stata anche oggetto di studio approfondito, una sorta di "reverse engineering" per conoscere, da un punto di vista tecnico, sia l'applicazione stessa che il suo database. Questo ci ha consentito di poter operare autonomamente sulla stessa valutandone la possibilità di migliorie effettivamente realizzate sia dal punto di vista della fruibilità del portale da parte dell'utente finale sia occupandoci della sua stabilità.

Durante la seconda annualità, abbiamo proseguito l'attività di test e verifica di nuove funzionalità e di quelle già in essere, ci siamo poi occupati della ricerca di algoritmi per l'efficienza energetica ed il monitoraggio degli impianti che è stata immaginata e sviluppata per soddisfare le specifiche esigenze del PELL, ed in particolare di quella sezione del lavoro che ricade sotto il nome di KPI (Key Performance Indicator). A tal fine abbiamo fatto un lavoro di analisi formale dei dati, in particolare quelli di consumo, presenti sulla piattaforma, cercando di capire quali algoritmi presenti in letteratura potessero essere utili all'analisi di questi dati. Ad esempio, attraverso l'analisi statistica dei dati di consumo, è possibile ragionare su eventuali predizioni del comportamento dell'impianto (o degli impianti) sotto indagine, in modo da procedere ad una progettazione/programmazione di interventi migliorativi.

Se la scorsa annualità ci ha visti lavorare sull'individuazione di algoritmi e metodologie presenti in letteratura che fossero adeguate ai nostri dati di consumo energetico, quest'anno abbiamo lavorato all'implementazione di alcuni di questi metodi ed allo studio dei risultati ottenuti

L'analisi dei dati energetici è in generale un argomento molto sviluppato in quanto propone sfide sia da un punto di vista tecnologico, che da un punto di vista scientifico; infatti la tendenza è nello sviluppo di tecnologie per il risparmio energetico, da applicare agli impianti (es. parzializzazione) e da utilizzare in sostituzione alle tecnologie obsolete ed ad elevato consumo (ad esempio lampade a LED in sostituzione di tecnologie di illuminazione classiche; le prime offrono efficienza molto più elevata e quindi, in tempi ragionevoli, permettono il recupero dell'investimento potendo contare su costi legati ai consumi energetici inferiori).

E' di fondamentale importanza capire "come" e "quanto" si consumi, e questo non può prescindere dalla conoscenza dei dati di consumo e dalla loro analisi.

In letteratura è molto discussa l'analisi energetica relativa ad edifici residenziali, che è un filone di studio molto rilevante anche grazie alla grande quantità di variabili che possono entrare il gioco durante il monitoraggio energetico di tali ambienti. Ad esempio, la numerosità degli abitanti di un edificio, il tempo e gli orari della loro permanenza, il numero di elettrodomestici utilizzati, ecc. possono dare indicazioni e fornire dati interessanti per l'analisi.

L'ambito dell'illuminazione pubblica è sicuramente stato affrontato in letteratura ma in modo limitato; la presenza di un impianto censito nella sua interezza, tramite l'applicazione PELL, fornisce un valore aggiunto all'analisi dei dati; infatti conoscere i dettagli dell' impianto, come la tipologia di sorgenti luminose, eventuali apparecchi presenti sui lampioni, i loro tempi di accensione e spegnimento, eventuali dati circa la parzializzazione o la riduzione del flusso luminoso (es. Crepuscolare), può fornire altre informazioni sullo stato dell'impianto o su eventuali problematiche presenti.

Non esiste un metodo che sia migliore di un altro, spesso i metodi vengono utilizzati in combinazione in quanto ciascuno fornisce informazioni specifiche sui dati. Ad esempio l'analisi delle serie temporali tramite

regressione fornisce informazioni sull'andamento temporale dei dati, oppure è utile per comprendere l'assenza di eventuali dati, mentre l'analisi degli stessi dati tramite algoritmi di clustering elimina di fatto il concetto di "tempo", in favore di una aggregazione secondo determinati parametri.

Ciò non vuol dire che la temporalità sia esclusa, bensì essa entra a far parte delle molteplici variabili che consentono una aggregazione che fornisca risultati indicativi.

Tecniche come la PCA, sulla quale ci siamo focalizzati particolarmente, permettono una certa riduzione della complessità del sistema attraverso la trasformazione dei dati da uno spazio dimensionale superiore ad uno inferiore. Questo vuol dire semplificare il sistema e semplificare l'analisi andandone a studiare la parte di dati più significativa.

Ci sono vari tipi di approcci alle tecniche di riduzione della dimensionalità, ad esempio l'MDA (Mixture discriminant analysis) o le reti neurali, oppure la PCA che è la tecnica più comune.

Avendo a disposizione dei dati quartorari, cioè un campione ogni quarto d'ora, ci è sembrato particolarmente interessante affrontare il problema utilizzando algoritmi di clustering, in modo da lavorare direttamente con l'elevato numero di variabili (96).

Di seguito verrà descritto:

- il set dei dati da analizzare
- le metodologie utilizzate per l'analisi, cioè:
    - serie temporali/regression
    - PCA (Principal Component Analysis)
    - reti neurali

Per ciascuna metodologia saranno evidenziati i risultati dell'analisi.

2.1 Il set dei dati

I dati di consumo provengono dagli smart meters presenti sui quadri elettrici appartenenti ad ogni impianto di illuminazione pubblica.

Per ogni POD (point of delivery), è possibile che siano presenti più quadri elettrici, mentre ad ogni quadro elettrico sono associati più apparecchi e punti luce.

Il gestore della rete recupera i dati provenienti dai misuratori e li invia in formato JSON verso un broker MQTT sui server ENEA che si occupa dell'acquisizione e del salvataggio su un server Hadoop appositamente configurato.

Al fine di garantire una standardizzazione dei dati che permetta di raccogliere e gestire i dati in modo più efficiente possibile, i dati inviati sulla piattaforma devono essere implementati secondo le "specifiche smart city platform" (SCPS-https://smartcityplatform.enea.it) realizzate nell'ambito del progetto di ricerca del Sistema Elettrico e sviluppate al fine di standardizzare la raccolta dei dati. SCP (Smart City Platform) è un ampio progetto nato per fornire a differenti categorie di utenti (cittadini, comuni, gestori di energia, ecc.) degli strumenti per raccogliere e valutare i dati urbani attraverso l'implementazione di modalità specifiche per favorire l'interoperabilità dei sistemi di gestione. Il PELL è una piattaforma ICT che fa parte di questo progetto e implementa delle specifiche condivise tra i vari utenti di questo sistema per la raccolta dei dati relativi agli impianti di illuminazione pubblica e ai loro consumi.

Nell'ambito del progetto PELL in capo ad ENEA, il nostro lavoro ha previsto il recupero dei dati di consumo (che implementano le specifiche disponibili al sito http://www.pell.enea.it/download) e la successiva loro

elaborazione per l'analisi. A tal fine, l'utilizzo di specifiche standardizzate permette di avere dati in un unico formato semplificandone la loro gestione anche da un punto di vista informatico.

I dati di consumo arrivano quindi in una struttura dati standardizzata che necessita però di alcune operazioni preliminari per poter rendere fruibili le informazioni in essa contenute. A tal proposito, quest'anno, abbiamo sviluppato delle funzioni che ci permettessero di accedere ai dati all'interno di questa struttura; per comodità del lettore, per comprendere meglio il lavoro fatto e le modalità operative attuate, di seguito si descrivono brevemente sia la struttura dei dati oggetto del nostro lavoro, sia le funzioni che abbiamo sviluppato per poterci lavorare.

Figura 1–L'"UrbanDataset" definisce la struttura dei documenti utilizzati per rappresentare i dati

La specifica prevede la definizione di un "UrbanDataset" che definisce la struttura dei documenti utilizzati per rappresentare i dati, il quale contiene 3 sezioni (Figura 1).

- Specification: contiene la definizione delle proprietà utilizzate nell'UrbanDataset da un punto di vista semantico (Figura 2)
- Context: contiene le informazioni relative al contesto del dataset, come i dati del produttore e le date di creazione (Figura 3)
- Values: contiene i dati espressi attraverso una lista di proprietà del tipo nome-valore (Figura 4)

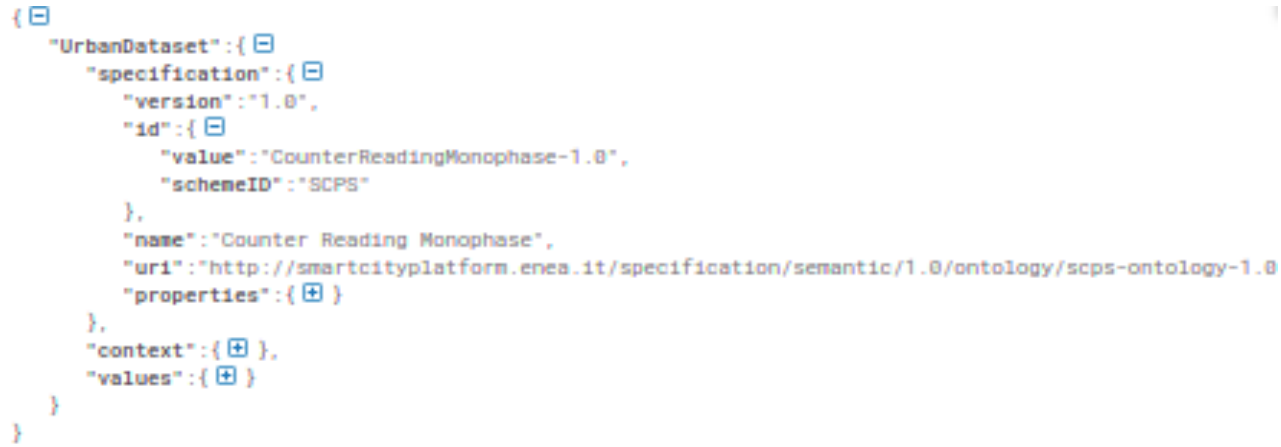


Figura 2-Sezione "Specification" dell'UrbanDataset

È bene specificare che l'UrbanDataset è una struttura generica che può essere utilizzata in vari ambiti proprio grazie alla possibilità di standardizzare i dati in formato condiviso; pertanto di seguito si farà riferimento solo all'UrbanDataset relativo al "CounterReading", cioè il formato specifico relativo ai dati di consumo provenienti dagli Smart Meters e da storicizzare su Hadoop.

```
{
   "UrbanDataset":{
      "specification":{ },
      "context":{
         "producer":{
            "id":"SL-SCPS-1",
            "schemeID":"SCPS"
         },
         "timeZone":"UTC+2",
         "timestamp":"2019-05-13T15:12:19",
         "coordinates":{
            "format":"WGS84-DD",
            "latitude":42.041292,
            "longitude":12.382046
         },
         "language":"IT"
      },
      "values":{ }
   }
}
```


Figura 3-sezione "context" dell'UrbanDataset

La Figura 4 mostra i dati di nostro interesse; infatti le prime due sezioni non contengono dati utili all'analisi mentre la sezione Values contiene i dati di consumo relativi a varie grandezze elettriche, nonché i dati relativi al codice POD (Point of Delivery) e Quadro Elettrico.

Ogni "id" rappresenta un dato quartorario, che comprende:

- le coordinate del POD (Point of Delivery)
- le date di inizio e fine validità del dato
- una lista di properties che includono i dati del POD e del Quadro elettrico. I dati sono indicati secondo la notazione name-value.

Esistono due differenti CounterReading, uno per utenze monofase ed uno per utenze trifase, che si differenziano per le properties, più numerose per le utenze trifase perché si riferiscono a tutte e 3 le fasi.

La struttura del JSON descritta viene salvata su Hadoop, e successivamente può essere recuperata tramite software di analisi dei dati, ad esempio Apache Spark.

La struttura del CounterReading, come già spiegato, è standardizzata al fine di favorire il recupero e lo studio dei dati, indipendentemente dalla loro provenienza e dal gestore che li invia al server.

Tuttavia tale struttura non è ottimale per l'analisi dei dati, in quanto nel file sono presenti molti dati trascurabili, e la lista di properties è molto scomoda da gestire quando viene effettuato il recupero dei dati, non consentendo spesso una gestione ottimale degli stessi.

```
"context":{ },
"values":{
   "line":[
      {
         "id":1,
         "coordinates":{
            "format":"WGS84-DD",
            "latitude":42.041292,
            "longitude":12.302046
         },
         "period":{
            "start_ts":"2019-05-12T00:00:00",
            "end_ts":"2019-05-12T00:15:00"
         },
         "property":[
            {
               "name":"ActiveEnergy",
               "val":"503023.4"
            },
            {
               "name":"PODID",
               "val":"IT120E12345678"
            },
            {
               "name":"ReactiveEnergy",
               "val":"31387.90"
            },
            {
               "name":"ActivePowerPhase",
               "val":"4.60"
            },
            {
               "name":"ApparentPowerPhase",
               "val":"0.87"
            },
            {
               "name":"CurrentLine",
               "val":"22.86"
            },
            {
               "name":"PowerFactorPhase",
               "val":"0.92"
            },
            {
               "name":"ReactivePowerPhase",
               "val":"0.25"
            },
            {
               "name":"VoltagePhase",
               "val":"222.8"
            }
         ]
      },
      {
         "id":2,
```


Figura 4-sezione "Value" dell'UrbanDataset

Per questo motivo, è bene trasformare il JSON originale in un formato più adatto all'analisi dati, effettuando alcune trasformazioni, appresso elencate:

- estrazione dei soli dati di interesse, cioè quelli presenti nella sezione "Value" dell'UrbanDataset e conseguente eliminazione dei dati ridondanti e non necessari al nostro scopo
- trasformazione di ogni "line" del JSON in una riga corrispondente al dato temporale quartorario
- trasformazione di una lista di coppie name/val per ciascun dato quartorario in una colonna differente. Le coppie name/val corrispondono alle property relative al dato quartorario, e corrispondono ai codici POD e quadro elettrico nonché alle varie grandezze elettriche misurate.

Il set di dati considerato in questo documento è stato esportato quindi in un file .csv (comma Separated value) in cui ogni record contiene i dati relativi ad un singolo campione temporale, e in cui le colonne siano le seguenti (per un CounterReading trifase):

- start _period
- end_period
- ActiveEnergy
- PODID
- ElectricalPanelID
- ActivePowerPhase1
- ActivePowerPhase2
- ActivePowerPhase3
- ApparentPowerPhase1
- ApparentPowerPhase2
- ApparentPowerPhase3
- CurrentLine1
- CurrentLine2
- CurrentLine3
- PowerFactorPhase1
- PowerFactorPhase2
- PowerFactorPhase3
- ReactiveEnergy
- ReactivePowerPhase1
- ReactivePowerPhase2
- ReactivePowerPhase3
- TotalActivePower
- TotalApparentPower
- TotalReactivePower
- VoltagePhase1
- VoltagePhase2
- VoltagePhase3

La Figura 5 mostra il CSV ricavato dalla trasformazione del JSON: ogni record costituisce un campione temporale per ogni grandezza elettrica misurata.

	start_period	end_period	ActiveE	PODID	Electric	ActiveP	ActiveP	ActiveP	Appare	Appare	Appare	Current	Current	Current	PowerF	PowerF	PowerF	Reactiv	Reactiv	Reactiv	Reactiv	TotalAc	TotalAp	TotalRe	Voltage	Voltage	Voltage
2	26/06/2019 00:00	26/06/2019 00:15	598.51	UVAX		769.17	769.56	833.42	0	0	0	3.47	3.6	3.79	1.0	0.98	1.0	0	2.84	163.42	74.38	0	0	0	225.71	224.52	224.89
3	26/06/2019 00:15	26/06/2019 00:30	590.79	UVAX		775.58	771.73	839.17	0	0	0	3.48	3.59	3.57	1.0	0.98	1.0	0	5.06	168.75	68.27	0	0	0	211.45	225.48	225.78
4	26/06/2019 00:30	26/06/2019 00:45	595.22	UVAX		772.61	768.76	836.2	0	0	0	3.48	3.59	3.79	1.0	0.98	1.0	0	4.78	170.77	70.94	0	0	0	226.13	224.97	225.34
5	26/06/2019 00:45	26/06/2019 01:00	592.3	UVAX		765.92	766.89	831.06	0	0	0	3.46	3.59	3.56	1.0	0.98	1.0	0	3.45	168.51	75.59	0	0	0	210.26	224.18	224.61
6	26/06/2019 01:00	26/06/2019 01:15	589.58	UVAX		720.79	772.19	839.83	0	0	0	3.47	3.6	3.8	1.0	0.98	1.0	0	2.59	165.61	67.12	0	0	0	211.21	210.23	225.65
7	26/06/2019 01:15	26/06/2019 01:30	594.22	UVAX		775.67	775.05	842.81	0	0	0	3.48	3.6	3.81	1.0	1.66	1.0	0	3.83	165.32	64.5	0	0	0	226.82	225.61	226.1
8	26/06/2019 01:30	26/06/2019 01:45	597.22	UVAX		782.83	779.76	849.43	0	0	0	3.49	3.6	3.82	1.0	0.91	1.69	0	9.06	163.52	58.67	0	0	0	227.83	226.65	227.1
9	26/06/2019 01:45	26/06/2019 02:00	601.47	UVAX		781.39	777.64	848.38	0	0	0	3.49	3.6	3.82	1.0	0.98	1.0	0	6.89	162.69	57.6	0	0	0	227.76	226.6	227.11
10	26/06/2019 02:00	26/06/2019 02:15	600.34	UVAX		776.24	775.51	846.28	0	0	0	3.48	3.6	3.82	1.0	0.98	1.0	0	7.64	167.04	64.96	0	0	0	227.14	225.97	226.5
11	26/06/2019 02:15	26/06/2019 02:30	598.61	UVAX		776.37	775.5	842.3	0	0	0	3.48	3.6	3.8	1.0	0.98	1.0	0	5.88	166.97	63.67	0	0	0	227.11	225.93	226.36
12	26/06/2019 02:30	26/06/2019 02:45	597.17	UVAX		774.57	772.61	841.82	0	0	0	3.47	3.6	3.8	1.0	1.66	1.0	0	9.15	170.91	66.45	0	0	0	226.79	225.58	226.03
13	26/06/2019 02:45	26/06/2019 03:00	595.17	UVAX		772.84	772.94	839.14	0	0	0	3.47	3.6	3.8	1.0	0.98	1.0	0	6.74	167.68	68.86	0	0	0	226.44	225.15	225.64
14	26/06/2019 03:00	26/06/2019 03:15	594.7	UVAX		774.03	772.3	838.83	0	0	0	3.48	3.6	3.8	1.0	0.98	1.0	0	2.94	166.52	69.9	0	0	0	226.59	225.33	225.87
15	26/06/2019 03:15	26/06/2019 03:30	594.52	UVAX		770.37	768.92	835.8	0	0	0	3.46	3.59	3.79	1.0	0.98	1.0	0	2.39	169.17	73.39	0	0	0	226.17	224.9	225.5
16	26/06/2019 03:30	26/06/2019 03:45	592.5	UVAX		772.7	772.9	840.45	0	0	0	3.47	3.59	3.8	1.0	0.98	1.0	0	3.41	162.46	67.87	0	0	0	226.52	225.26	225.8
17	26/06/2019 03:45	26/06/2019 04:00	594.64	UVAX		779.06	776.32	846.27	0	0	0	3.48	3.6	3.81	1.0	0.98	1.0	0	5.07	161.61	66.84	0	0	0	212.25	226.04	226.68
18	26/06/2019 04:00	26/06/2019 04:15	593.0	UVAX		774.9	773.12	839.6	0	0	0	3.48	3.59	3.8	1.0	0.98	1.0	0	3.4	162.5	70.03	0	0	0	226.71	225.4	225.95
19	26/06/2019 04:15	26/06/2019 04:30	595.6	UVAX		772.43	771.23	841.42	0	0	0	3.47	3.59	3.81	1.0	1.66	1.0	0	2.55	168.43	66.0	0	0	0	226.6	225.5	225.96
20	26/06/2019 04:30	26/06/2019 04:45	594.64	UVAX		774.41	767.8	837.77	0	0	0	3.48	3.58	3.79	1.0	1.66	1.0	0	4.33	174.75	71.91	0	0	0	226.58	225.29	225.81
21	26/06/2019 04:45	26/06/2019 05:00	593.56	UVAX		724.88	770.08	841.95	0	0	0	3.48	3.58	3.8	1.0	0.98	1.0	0	4.78	172.83	63.88	0	0	0	226.98	215.37	226.16
22	26/06/2019 05:00	26/06/2019 05:15	596.12	UVAX		776.42	771.64	841.36	0	0	0	3.48	3.59	3.8	1.0	0.98	1.0	0	4.2	167.97	66.61	0	0	0	226.94	225.62	226.1

Figura 5-Trasformazione del Json in un CSV contenente i dati di consumo per tutte le grandezze elettriche. Nello studio ci riferiremo prevalentemente all'Energia Attiva

Il set di dati comprende un periodo che va da luglio 2020 a dicembre 2020, per un unico POD.

È evidente che, come in caso di dati quartorari, per cui per ogni giorno sono disponibili 96 valori di energia attiva, corrente, tensione, energia reattiva ecc., per singolo POD, la quantità di dati possa potenzialmente essere molto consistente; un esempio reale è costituito dagli impianti di illuminazione pubblica di Genova che comprende circa 600 POD, e di La Spezia che ne comprende circa 250.

Ad ogni POD sono associati più carichi elettrici, pertanto i dati indicati fanno riferimento alla somma di tutti i valori relativi ad ogni dispositivo dell'impianto. Ad esempio, nella situazione in figura, un valore di energia attiva di 590,79 kW/h indica l'energia consumata nel periodo indicato da tutti gli apparecchi afferenti ad un POD e ad un quadro elettrico.

In tale scenario ci si ritrova in un ambito "Big Data", cioè grandi quantità di dati eterogenei che crescono quantitativamente e necessitano di essere gestiti.

Si è pensato dapprima di analizzare i dati dal punto di vista temporale (capitolo 2.3.1), per individuare problematiche macroscopiche, valutare l'andamento dei dati di consumo nel lungo periodo e valutare quali metodiche di analisi fossero più adeguate nel caso in oggetto.

L'analisi temporale è molto utile ad identificare problematiche e fattori legati all'andamento dei dati nel tempo (ad esempio picchi di consumi), che hanno un senso aggregando i dati per giorno oppure per mese; infatti, i valori quartorari anomali potrebbero essere trascurati in caso di analisi di grandi archi temporali (settimane, mesi, anni), ma si è preferito conservare la granularità quartoraria per presentare una analisi che offrisse una precisione quanto più elevata possibile.

Per questo motivo abbiamo applicato algoritmi di clustering, in particolare la PCA, per valutare i dati e creare "aggregati" o cluster di dati, che abbiano determinate caratteristiche di correlazione, in modo da ridurre lo studio dei dati a pochi aggregati degli stessi.

2.2 I software utilizzati

MATLAB (https://it.mathworks.com/product/matlab.html) è una piattaforma di programmazione e calcolo numerico utilizzata da anni per l'analisi dei dati, lo sviluppo di algoritmi e la creazione di modelli.

È un software a pagamento, molto utilizzato in ambito scientifico ed universitario per la sua elevata affidabilità e soprattutto per la presenza di tools per svariati ambiti di lavoro e ricerca.

Tra le funzionalità disponibili si evidenziano:

- Analisi dei dati
- Visualizzazione ed esplorazione dei dati

- Progettazione di algoritmi
- Creazioni di applicazioni
- Calcolo parallelo su larga scala
- Possibilità di utilizzo con vari linguaggi di programmazione (Python, C/C++,java, ecc.).

2.3 *Metodologie di analisi*

2.3.1 Dati come serie temporali

I dati in nostro possesso possono essere espressi come "serie temporali", cioè una serie di misurazioni che si susseguono l'un l'altra e che possono essere rappresentate su un piano in cui in ascissa è indicato l'istante temporale ed in ordinata il valore della grandezza elettrica misurata in quell'istante. Ad esempio, la Figura 6 mostra l'andamento temporale quartorario dei dati da luglio a dicembre per un singolo POD. Ogni singolo puntino blu indica l'energia attiva assorbita dal carico elettrico legato al POD per un intervallo di tempo quartorario.

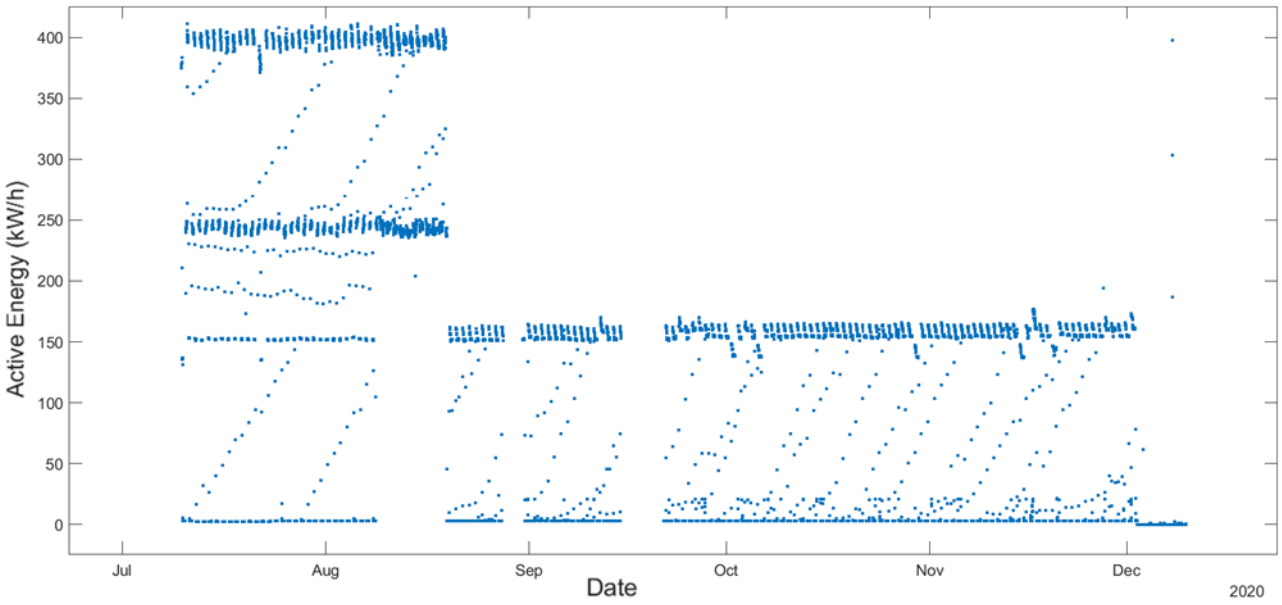


Figura 6 - l'andamento temporale quartorario dei dati da luglio a dicembre

Le funzioni che sono state utilizzate in MATLAB sono le seguenti:

t1=t(timerange('2020-01-01 00:00:00','2020-12-31 23:59:00'),:)

plot(t1.start_period,t1.ActiveEnergy,'.')

dove ***t*** è l'insieme dei dati in formato tabellare contente i dati di consumo, come evidenziato nei paragrafi precedenti; ***t1*** è un sottoinsieme tabellare dei dati di ***t***, in cui viene selezionato un timerange relativo all'anno 2020. Infatti, i dati originari presentano anche un intervallo di 14 giorni relativo all'anno 2019 ma, per motivi di visualizzazione, sono stati esclusi da questa prima analisi; ***plot*** è la funzione che grafica i dati di energia attiva (contenuti nella colonna "ActiveEnergy" di ***t1***) in funzione della "start date", cioè della data di inizio di validità del singolo campione.

La start_date è in formato YYYY-MM-DD HH:MM:SS.

La Figura 6, che ricordiamo fa riferimento ai dati di un singolo POD per un periodo di circa 6 mesi dell'anno 2020, mostra un andamento anomalo dei consumi per i mesi di luglio ed agosto, evidenziando un consumo molto elevato in alcuni giorni.

Se effettuiamo una aggregazione dei dati giornaliera, abbiamo l'andamento mostrato nella Figura 7 che conferma un consumo anomalo per i mesi di luglio e agosto.

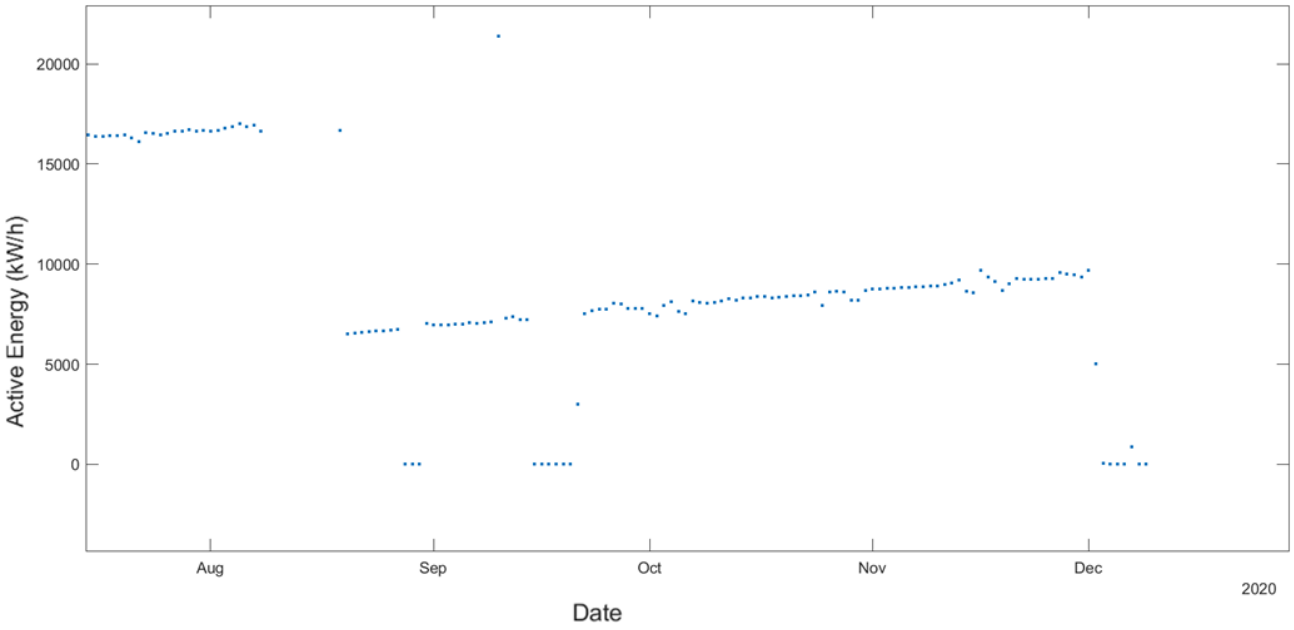


Figura 7-Aggregazione dati giornaliera

Le funzioni:

```
tEnergy=t1(:,'ActiveEnergy')
tEnergyDay=retime(tEnergy,'daily','sum')
plot(tEnergyDay.start_period,tEnergyDay.ActiveEnergy,'.')
```

provvedono a fornire i dati secondo l'aggregazione di nostro interesse; la prima funzione considera i dati presenti nella Colonna "ActiveEnergy" di ***t1***, mentre la funzione ***retime*** di MATLAB restituisce una nuova tabella che contiene le stesse variabili di ***tEnergy*** ma equi-spaziate temporalmente secondo un metodo specificato.

Nel nostro caso abbiamo effettuato la somma (***sum***) di tutti i campioni di ***tEnergy*** per ogni intervallo giornaliero (***daily***), e ne abbiamo graficato l'andamento attraverso la funzione ***plot.***

È possibile utilizzare altri metodi di aggregazione che comprendano, ad esempio, l'interpolazione, oppure la rimozione dei record duplicati o ancora, in caso di timerange irregolari è possibile utilizzare la funzione "***retime***" per rendere tali intervalli regolari.

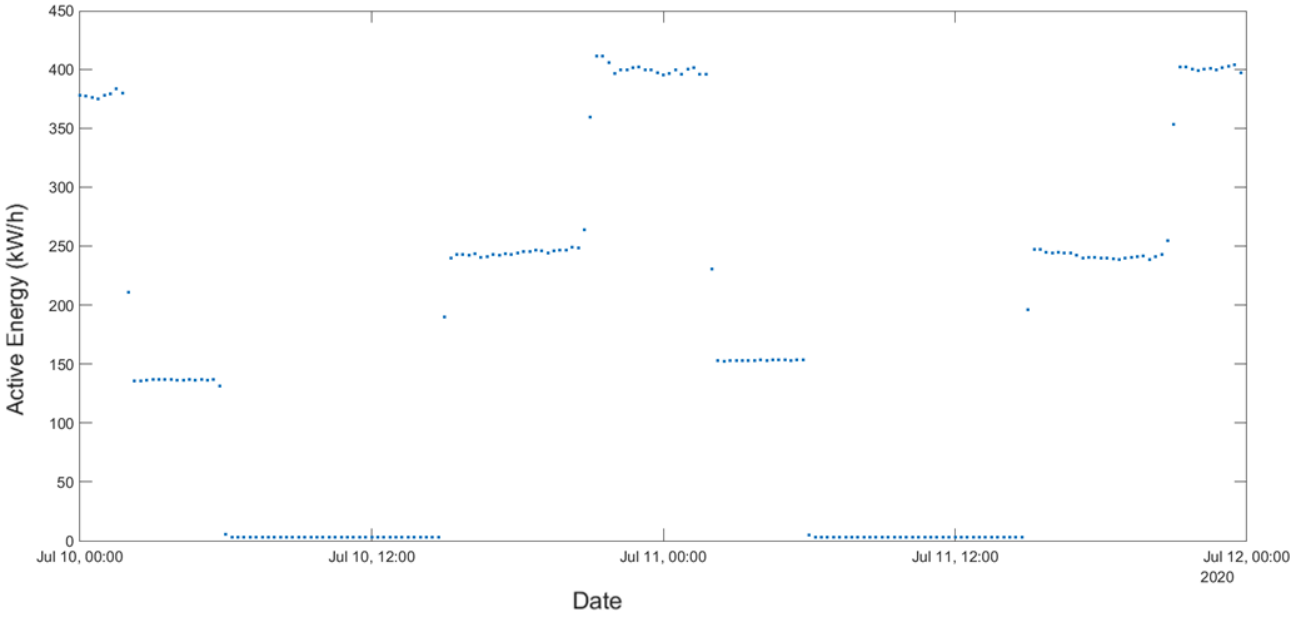


Figura 8 - dati quartorari relativi al 10-12 luglio

I dati giornalieri e quartorari danno informazioni differenti; il dato giornaliero semplifica la visualizzazione dei dati temporali nel grafico facendo risaltare picchi (sia positivi che negativi) di consumi anomali, ma non riesce a dare indicazione di variabilità di consumo di energia nella giornata (bassa risoluzione), cosa che invece il dato quartorario consente grazie ad una valutazione più dettagliata dell'andamento del consumo durante l'intero arco della giornata. Una risoluzione più alta permette di evidenziare anomalie legate all'orario che possano verificarsi più o meno occasionalmente a scapito però di una maggiore difficoltà nella gestione dei dati.

Questa differenza di risoluzione è evidenziata nelle Figura 6 (dati quartorari) eFigura 7 (dati aggregati per giorno) nelle quali è possibile notare una evidente anomalia legata ad un consumo giornaliero anomalo tra luglio ed agosto che non può essere giustificato perché, durante i mesi estivi, il consumo giornaliero dovrebbe essere inferiore rispetto ai mesi invernali. Dalla Figura 7 si evince l'anomalia, ma è solo dai dati quartorari (Figura 6) che si può vedere come i consumi siano distribuiti durante la giornata e siano più consistenti in alcuni orari.

Analizzando nel dettaglio i dati giornalieri dal 10 luglio, si evidenzia come per i giorni fino al 19 agosto vi sia un elevato consumo di energia che potrebbe avere varie motivazioni (malfunzionamenti, furto di energia, o l'introduzione di apparati elettronici quali telecamere o antenne montate sui lampioni, ecc.), mentre per i restanti giorni l'andamento risulta crescente compatibilmente con la stagionalità (verso l'inverno i consumi aumentano perché i lampioni sono accesi per più ore).

Se prendiamo la Figura 8 che mostra i dati quartorari relativi al 10-12 luglio, si evince un consumo elevato dalle 00:00 alle 02:15, e un consumo inferiore fino alle 06:00. In seguito, il consumo resta minimo fino alle 15:00 e poi risale nuovamente. L'andamento può essere giustificato dall'utilizzo di sistemi di risparmio di energia, che riducono il flusso luminoso.

Oltre al consumo elevato, viene evidenziato che l'impianto resta attivo per un periodo di tempo continuativo, cioè sembra non spegnersi mai. Questo dettaglio è molto importante perché il numero di ore di accensione degli impianti di illuminazione pubblica è normato da ARERA (autorità di regolazione di energia reti e ambiente) che, con la delibera n.52/04, definisce le modalità per l'attribuzione su base oraria dell'energia elettrica prelevata dagli impianti di illuminazione pubblica (https://www.arera.it/it/docs/04/052-04.htm).

La delibera propone una tabella che definisce gli orari di accensione degli impianti considerando ogni decade di ogni mese e la zona geografica di appartenenza (Figura 9), qualora il sistema mi indichi un consumo

dell'impianto, anche piccolo, oltre l'orario previsto dalla delibera, significa che vi è una qualche forma di dispersione di energia da studiare successivamente.

La Figura 9 mostra i dati relativi alla fascia geografica centrale (Abruzzo, Emilia Romagna, Friuli Venezia Giulia, Lazio, Marche, Sicilia, Toscana, Trentino Alto Adige, Umbria e Veneto); per la fascia geografica occidentale (Liguria, Lombardia, Piemonte, Sardegna, Valle d'Aosta) gli orari sono posticipati di 15 minuti mentre, per la fascia geografica orientale (Basilicata, Calabria, Campania, Molise, Puglia), gli orari sono anticipati di 15 minuti.

mese	decade	ora convenzionale di accensione	ora convenzionale di spegnimento
Gennaio	1	17.05	7.55
	2	17.15	7.50
	3	17.25	7.45
Febbraio	1	17.40	7.35
	2	17.55	7.20
	3	18.10	7.05
Marzo	1	18.20	6.50
	2	18.35	6.30
	3	18.50	6.10
Aprile	1	20.05	6.50
	2	20.15	6.30
	3	20.30	6.10
Maggio	1	20.45	5.55
	2	20.55	5.40
	3	21.10	5.30
Giugno	1	21.20	5.20
	2	21.25	5.20
	3	21.30	5.20
Luglio	1	21.30	5.30
	2	21.20	5.40
	3	21.10	5.45
Agosto	1	20.55	6.00
	2	20.40	6.15
	3	20.20	6.30
Settembre	1	20.00	6.45
	2	19.40	6.55
	3	19.20	7.10
Ottobre	1	19.00	7.20
	2	18.40	7.35
	3	18.25	7.45
Novembre	1	17.10	7.00
	2	16.55	7.15
	3	16.50	7.25
Dicembre	1	16.50	7.40
	2	16.50	7.45
	3	16.55	7.55

Figura 9-Tabella che definisce gli orari di accensione degli impianti, considerando ogni decade di ogni mese e la zona geografica

I tempi di accensione sono definiti da ARERA in modo convenzionale e sono dei punti di riferimento per l'analisi dei dati. Avendo a disposizione dei tempi convenzionali, è dunque possibile valutare i dati quartorari alla luce degli orari indicati da ARERA, ed evidenziare eventuali anomalie che dovessero verificarsi.

Ad esempio se si analizzano i dati temporali relativi al periodo dal 09/08 al 18/08 (Figura 10) nelle ore centrali del giorno non risulta alcuna riduzione di flusso luminoso, pertanto è presumibile che alcuni o tutti i lampioni afferenti al POD risultino sempre accesi.

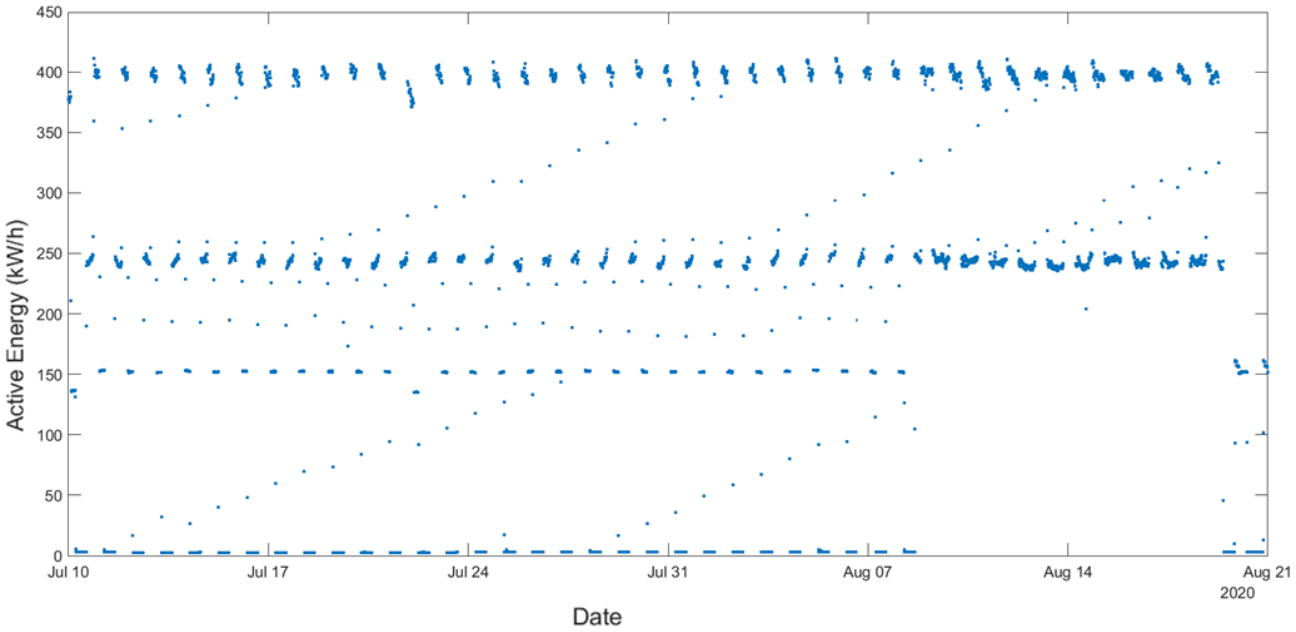


Figura 10 – nel periodo 09-18/08 non risulta riduzione dei consumi nelle ore centrali del giorno in cui dovrebbe essere nullo

Invece dal 20/08 (Figura 11) l'andamento dei consumi è crescente, compatibilmente con l'aumento delle ore di funzionamento dovuto alla stagionalità.

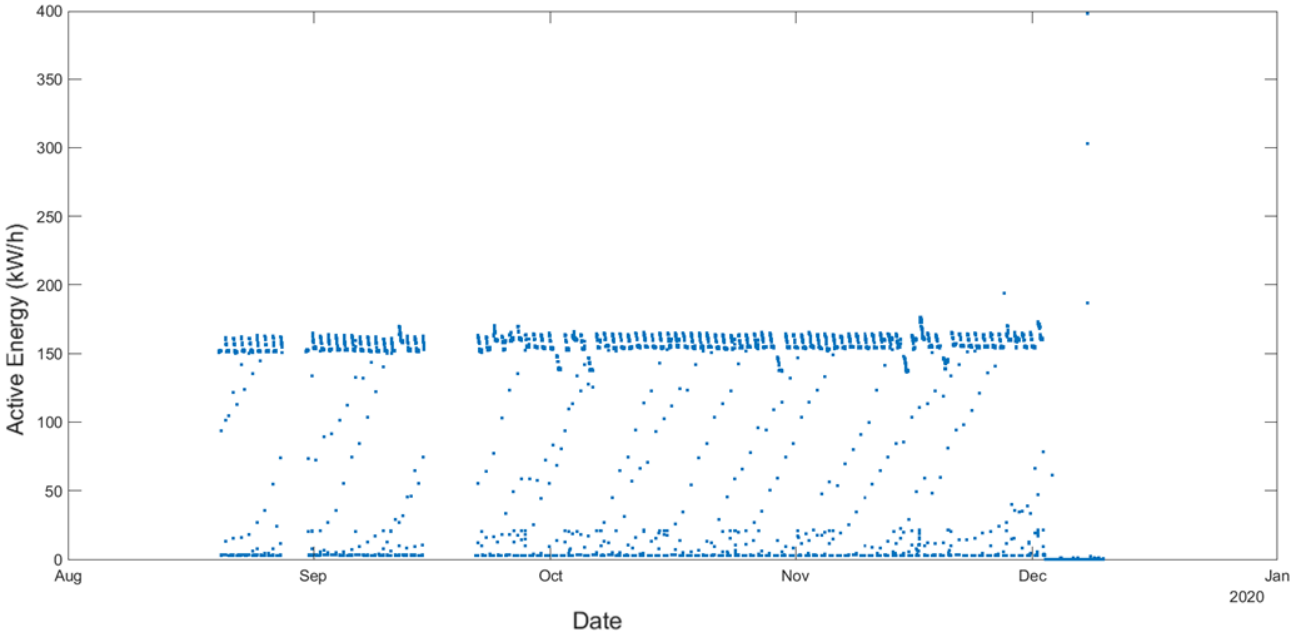


Figura 11 – andamento tendenzialmente crescente dei consumi peril periodo 09-12/2020

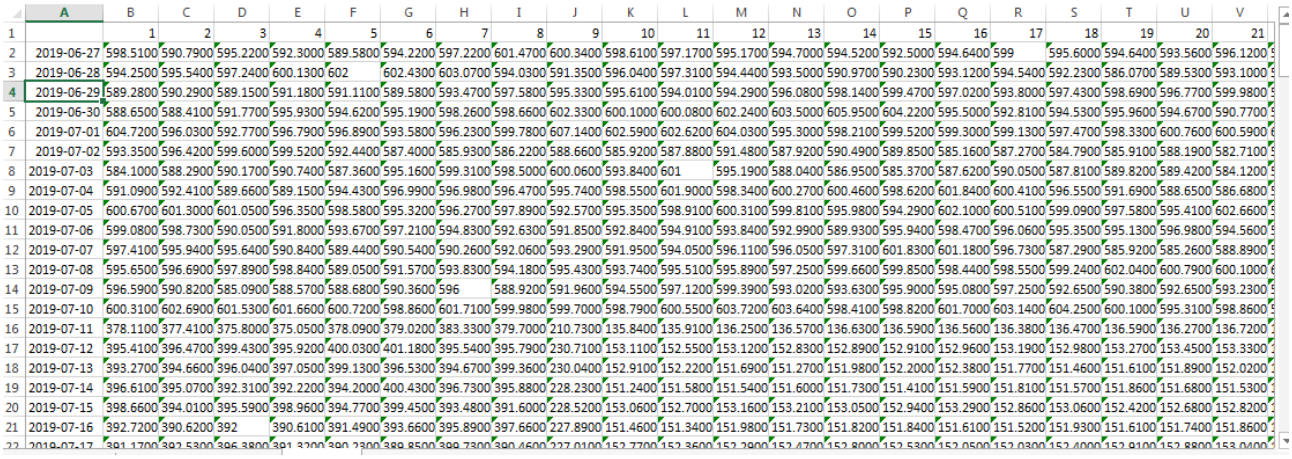
Le serie temporali danno molte informazioni che, come la letteratura insegna, possono essere analizzate con tecniche regressive [4-7]. Tuttavia, l'elevata numerosità delle variabili (96) rende poco pratico l'utilizzo di queste tecniche la cui complessità cresce pesantemente al crescere delle varabili di ingresso; inoltre, l'andamento temporale dei consumi quartorari evidenziato dai grafici precedenti, ci ha suggerito di puntare l'attenzione sulla ricerca di eccezioni che consentano di evidenziare specificità dei dati tali da far supporre che si possano raggruppare in casi da manuale.

Per questo motivo abbiamo optato per l'analisi dei dati attraverso altri metodi che permettano di ridurre la complessità del problema identificando legami tra le variabili, evidenziandone le similitudini e/o le differenze, in modo da interpretare meglio i fenomeni enucleati nelle variabili.

2.3.2 Analisi dati tramite PCA

La PCA (Principal Component Analysis) consiste in una metodica di analisi che prevede la semplificazione di problemi che presentano molte variabili.

Poiché il nostro dataset è costituito da dati quartorari corrispondenti ad ogni quarto d'ora di ogni giorno, possiamo considerare il nostro dataset come una tabella in cui sulle righe abbiamo il giorno e sulle colonne abbiamo i dati quartorari, per un totale di 96 variabili corrispondenti ai campioni quartorari.



	A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
2	2019-06-27	598.5100	590.7900	595.2200	592.3000	589.5800	594.2200	597.2200	601.4700	600.3400	598.6100	597.1700	595.1700	594.7000	594.5200	592.5000	594.6400	599	595.6000	594.6400	593.5600	596.1200
3	2019-06-28	594.2500	595.5400	597.2400	600.1300	602	602.4300	603.0700	594.0300	591.3500	596.0400	597.3100	594.4400	593.5000	590.9700	590.2300	593.1200	594.5400	592.2300	586.0700	589.5300	593.1000
4	2019-06-29	589.2800	590.2900	589.1500	591.1800	591.1100	589.5800	593.4700	597.5800	595.3300	595.6100	594.0100	594.2900	596.0800	598.1400	599.4700	597.0200	593.8000	597.4300	598.6900	596.7700	599.9800
5	2019-06-30	588.6500	588.4100	591.7700	595.9300	594.6200	595.1900	598.2600	598.6600	602.3300	600.1000	600.0800	602.2400	603.5000	605.9500	604.2200	595.5000	592.8100	594.5300	595.9600	594.6700	590.7700
6	2019-07-01	604.7200	596.0300	592.7700	596.7900	596.8900	593.5800	596.2300	599.7800	607.1400	602.5900	602.6200	604.0300	595.3000	598.2100	599.5200	599.3000	599.1300	597.4700	598.3300	600.7600	600.5900
7	2019-07-02	593.3500	596.4200	599.6000	599.5200	592.4400	587.4000	585.9300	586.2200	588.6600	585.9200	587.8800	591.4800	587.9200	590.4900	589.8500	585.1600	587.2700	584.7900	585.9100	588.1900	582.7100
8	2019-07-03	584.1000	588.2900	590.1700	590.7400	587.3600	595.1600	599.3100	598.5000	600.0600	593.8400	601	595.1900	588.0400	586.9500	585.3700	587.6200	590.0500	587.8100	589.8200	589.4200	584.1200
9	2019-07-04	591.0900	592.4100	589.6600	589.1500	594.4300	596.9900	596.9800	596.4700	595.7400	598.5500	601.9000	598.3400	600.2700	600.4600	598.6200	601.8400	600.4100	596.5500	591.6900	588.6500	586.6800
10	2019-07-05	600.6700	601.3000	601.0500	596.3500	598.5800	595.3200	596.2700	597.8900	592.5700	595.3500	598.9100	600.3100	599.8100	595.9800	594.2900	602.1000	600.5100	599.0900	597.5800	595.4100	602.6600
11	2019-07-06	599.0800	598.7300	590.0500	591.8000	593.6700	597.2100	594.8300	592.6300	591.8500	592.8400	594.9100	593.8400	592.9900	589.9300	595.9400	598.4700	596.0600	595.3500	595.1300	596.9800	594.5600
12	2019-07-07	597.4100	595.9400	595.6400	590.8400	589.4400	590.5400	590.2600	592.0600	593.2900	591.9500	594.0500	596.1100	596.0500	597.3100	601.8300	601.1800	596.7300	587.2900	585.9200	585.2600	588.8900
13	2019-07-08	595.6500	596.6900	597.8900	598.8400	589.0500	591.5700	593.8300	594.1800	595.4300	593.7400	595.5100	595.8900	597.2500	599.6600	599.8500	598.4400	598.5500	599.2400	602.0400	600.7900	600.1000
14	2019-07-09	596.5900	590.8200	585.0900	588.5700	588.6800	590.3600	596	588.9200	591.9600	594.5500	597.1200	599.3900	593.0200	593.6300	595.9000	595.0800	597.2500	592.6500	590.3800	592.6500	593.2300
15	2019-07-10	600.3100	602.6900	601.5300	601.6600	600.7200	598.8600	601.7100	599.9800	599.7000	598.7900	600.5500	603.7200	603.6400	598.4100	598.8200	601.7000	603.1400	604.2500	600.1000	595.3100	598.8600
16	2019-07-11	378.1100	377.4100	375.8000	375.0500	378.0900	379.0200	383.3300	379.7000	210.7300	135.8400	135.9100	136.2500	136.5700	136.6300	136.5900	136.5600	136.3800	136.4700	136.5900	136.2700	136.7200
17	2019-07-12	395.4100	396.4700	399.4300	395.9200	400.0300	401.1800	395.5400	395.7900	230.7100	153.1100	152.5500	153.1200	152.8300	152.8900	152.9100	152.9600	153.1900	152.9800	153.2700	153.4500	153.3300
18	2019-07-13	393.2700	394.6600	396.0400	397.0500	399.1300	396.5300	394.6700	399.3600	230.0400	152.9100	152.2200	151.6900	151.2700	151.9800	152.2000	152.3800	151.7700	151.4600	151.6100	151.8900	152.0200
19	2019-07-14	396.6100	395.0700	392.3100	392.2200	394.2000	400.4300	396.7300	395.8800	228.2300	151.2400	151.5800	151.5400	151.6000	151.7300	151.4100	151.5900	151.8100	151.5700	151.8600	151.6800	151.5300
20	2019-07-15	398.6600	394.0100	395.5900	398.9600	394.7700	399.4500	393.4800	391.6000	228.5200	153.0600	152.7000	153.1600	153.2100	153.0500	152.9400	153.2900	152.8600	153.0600	152.4200	152.6800	152.8200
21	2019-07-16	392.7200	390.6200	392	390.6100	391.4900	393.6600	395.8900	397.6600	227.8900	151.4600	151.3400	151.9800	151.7300	151.8200	151.8400	151.6100	151.5200	151.9300	151.6100	151.7400	151.8600
22	2019-07-17	391.1700	392.5300	396.3800	391.3200	390.2300	389.8500	389.7300	390.4600	227.0100	152.7700	152.3600	152.2900	152.4700	152.8000	152.5300	152.0500	152.0300	152.4000	152.9100	152.8800	153.0400

Figura 12-trasformazione del csv in una tabella in cui sulle righe abbiamo il giorno, sulle colonne abbiamo i dati quartorari

Grazie all'aggregazione dei dati realizzata dalla PCA, viene ridotto il numero di variabili. Ciò si traduce in una "trasformazione" delle variabili (definite componenti principali) che vengono ordinate in base al valore decrescente della loro varianza.

Questo significa che le prime componenti principali presentano una varianza maggiore, pertanto basta analizzare queste, riducendo di fatto la complessità del problema.

La PCA è una tecnica statistica molto utilizzata in vari ambiti, perché a differenza di altre metodiche compensa la perdita di informazione dei dati originali, tipica della semplificazione di un problema, attraverso la scelta del numero di componenti da analizzare.

Da un punto di vista geometrico, la PCA consiste in una trasformazione che avviene attraverso la proiezione delle variabili originarie in un nuovo sistema di riferimento n-dimensionale, individuato dalle componenti principali; su un asse viene rappresentata la componente che ha la maggior varianza del sistema, su un altro asse una componente che ha una varianza inferiore alla prima, e così via fino all'ultima componente.

Se consideriamo una matrice M con n righe e m colonne, possiamo considerare un nuovo spazio a m dimensioni in cui ogni variabile rappresenta un asse di coordinate. Ad esempio in caso di m = 3, il piano è

tridimensionale e gli assi sono x_1, x_2, x_3. Ogni osservazione può essere disposta in questo spazio, che risulterà in uno sciame di punti (Figura 13).

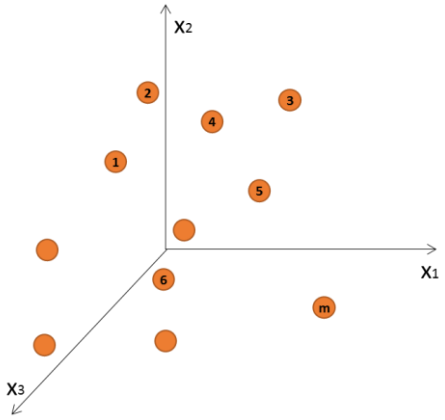


Figura 13-Lo spazio in cui gli assi sono rappresentati dalle n variabili, mentre le m osservazioni sono rappresentate da uno sciame di punti.

La PCA consiste nel creare un nuovo piano cartesiano, in cui un asse è la prima componente principale ed è la linea dello spazio x_1, x_2, x_3 che meglio approssima i dati nel senso dei minimi quadrati, cioè che massimizza la varianza. Ogni osservazione che venga proiettata su questo nuovo asse rappresenta un punteggio o "score" per la prima componente.

La seconda componente è una nuova linea nello spazio ortogonale rispetto alla prima componente e migliora l'approssimazione dei dati. Ogni osservazione viene proiettata su questo asse e otterrà un punteggio nuovo relativo a questa seconda componente la cui varianza è comunque inferiore a quella rappresentata dalla prima componente (Figura 14).

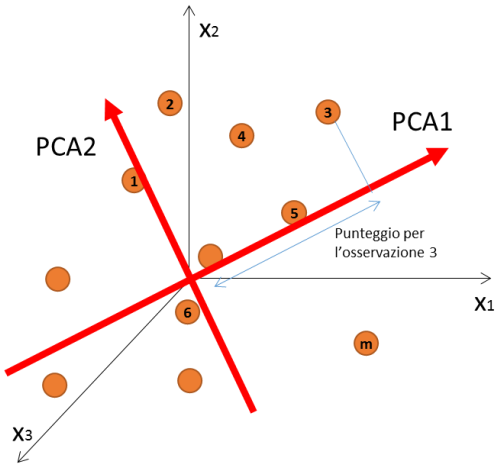


Figura 14-il nuovo piano PCA1-PCA2 e la proiezione della terza osservazione su PCA1

Il piano PCA1, PCA2 rappresenta il nuovo piano cartesiano sui cui assi vengono proiettate le osservazioni dello spazio. Esso è centrato nell'origine del piano originale, e questo avviene perché vengono inizialmente calcolate le medie delle variabili che rappresentano un punto disposto nel punto medio tra le osservazioni; quindi la media viene centrata nell'origine del piano x_1, x_2, x_3 attraverso una traslazione.

Da un punto di vista matematico, la PCA parte da un problema a n variabili

$$X_1, X_2, X_3, \ldots X_i, \ldots X_n$$

ed identifica altrettante n variabili differenti da quelle di partenza

$$Y_1, Y_2, Y_3, \dots Y_i, \dots Y_n$$

Tale che ciascuna sia ottenuta come combinazione lineare delle n variabili originarie.

Ciascuna variabile Y_i contiene informazioni sintetizzate relative ai dati di partenza, permettendo così di estrarre il numero più elevato possibile di informazioni che vengono concentrate in un set più ristretto di quello iniziale, anche se, ancorché composto ancora da un numero molto elevato di variabili.

A livello matriciale i dati di partenza possono essere riassunti in una matrice $\vec{X}$

$$\vec{X}= \begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{pmatrix}$$

in cui le m righe rappresentano le osservazioni e le n colonne sono le variabili considerate per il fenomeno.

Quello che si vuole ottenere è una matrice di dati le cui variabili sono combinazione lineare delle variabili di partenza:

$$\vec{Y}= \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{pmatrix}$$

In questa situazione la prima componente principale risulta essere:

$$Y_1= A_{11}X_{11}+ A_{12}X_{12}+\dots+A_{1n}X_{1n}$$

I coefficienti vengono calcolati cercando di massimizzare la varianza del vettore $\overrightarrow{Y_1}$ utilizzando il metodo dei minimi quadrati:

$$1 = A_{11}^2+ A_{12}^2+\dots+A_{1n}^2$$

Ripetendo l'operazione per tutti i vettori $\overrightarrow{Y_i}$ si individueranno m variabili:

$$\begin{cases} Y_1= A_{11}X_{11}+ A_{12}X_{12}+\dots+A_{1n}X_{1n} \\ \dots \\ Y_m= A_{m1}X_{m1}+ A_{m2}X_{m2}+\dots+A_{mn}X_{mn} \end{cases}$$

per le quali devono essere soddisfatte le relative condizioni qui elencate:

$$\begin{gathered} 1 = A_{11}^2+ A_{12}^2+\dots+A_{1n}^2 \\ \dots \quad \dots \quad \dots \\ 1 = A_{m1}^2+ A_{m2}^2+\dots+A_{mn}^2 \end{gathered}$$

La varianza del primo coefficiente deve essere maggiore della varianza del secondo coefficiente e così via, cioè:

$$Var(Y_1) > Var(Y_2) > \cdots > Var(Y_m)$$

Tutte le variabili $\overrightarrow{Y_i}$ sono incorrelate.

Il risultato della trasformazione consiste in una matrice di n componenti principali nella quale la prima presenta la varianza maggiore e le successive a decrescere.

Poiché il compito della PCA è quello di analizzare un numero inferiore di dati rispetto a quelli originari, al fine di effettuare l'analisi vengono considerate solo le componenti che rappresentano l'80-90% della variabilità complessiva, trascurando le componenti a minore varianza.

Se si volesse conservare maggiore informazione per minimizzare la perdita di dati è si potrebbe sempre considerare anche le componenti a minore varianza.

Da un punto di vista geometrico, invece, le componenti principali individuano un nuovo sistema di riferimento, in modo da rappresentare su un asse la componente che rappresenta la maggior variabilità del sistema, mentre sull'altro asse una componente che rappresenta una variabilità inferiore alla prima.

Un piccolo appunto sulla definizione di componente principale riguarda il suo significato statistico; infatti in un set di dati in cui alcune variabili sono parzialmente correlate tra loro, le direzioni di massima varianza sono quelle che garantiscono la massima correlazione, e corrispondono alle prime componenti principali. Le componenti i cui autovalori hanno varianza minore rappresentano dati con scarsa o nulla correlazione. Poiché il rumore di una variabile è non correlato con le altre variabili, escludere tali componenti vuol dire filtrare il rumore non correlato.

È possibile che all'interno del rumore vi sia comunque dell'informazione, tuttavia l'analisi delle componenti principali permette di decidere se considerare anche le componenti a varianza minore, riducendo l'errore dovuto alla loro assenza, oppure escluderle dall'analisi.

La PCA fornisce dunque una rappresentazione adeguata dello spazio in cui viene conservato il carattere statistico dell'insieme dei dati. Nel nostro caso un pattern multidimensionale (96 variabili) viene ridotto in uno spazio di 2 dimensioni raggruppando i dati in clusters o classi.

Una specifica di questa tecnica richiede la standardizzazione delle variabili. Infatti, se i dati non sono omogenei (ad esempio per unità di misura o dimensioni di diversi ordini di grandezza) le componenti principali potrebbero dare risultati errati e, dunque, una errata interpretazione.

Un esempio è l'utilizzo di diverse unità di misura, ad esempio parte dei dati possono essere espresse in W/h e la restante parte in kW/h, per cui le componenti principali possono essere più o meno impattate. A tale scopo è importante standardizzare i dati.

2.3.3 Applicazione della PCA

Al fine di effettuare l'analisi, la PCA è stata implementata in MATLAB. Riteniamo utile descrivere brevemente le operazioni che devono essere effettuate in MATLAB per gestire i dati ed anche per consentire, ad eventuali terzi, di replicare le nostre attività.

Il primo step è stato quello di esportare i dati di consumo JSON salvati su Hadoop in formato csv per cui:

- Ogni rigo rappresenta un campione quartorario
- Ogni colonna rappresenta una proprietà ricavata dal JSON, in particolare:
  - name
  - timestamp
  - lat
  - lon
  - start _period
  - end_period
  - ActiveEnergy
  - PODID
  - ElectricalPanelID
  - ActivePowerPhase1

- ActivePowerPhase2
- ActivePowerPhase3
- ApparentPowerPhase1
- ApparentPowerPhase2
- ApparentPowerPhase3
- CurrentLine1
- CurrentLine2
- CurrentLine3
- PowerFactorPhase1
- PowerFactorPhase2
- PowerFactorPhase3
- ReactiveEnergy
- ReactivePowerPhase1
- ReactivePowerPhase2
- ReactivePowerPhase3
- TotalActivePower
- TotalApparentPower
- TotalReactivePower
- VoltagePhase1
- VoltagePhase2
- VoltagePhase3

L'informazione relativa al tempo è descritta nel campo "start_period", mentre il campione misurato è descritto nel campo "ActiveEnergy".

I campi "PODID" ed "ElectricalPanelID" rappresentano rispettivamente gli identificatori del Point of Delivery e del quadro elettrico.

Il secondo step è stato quello di importare tali dati in MATLAB in un formato che fosse utile al nostro scopo. Il formato a noi gradito prevede che:

- Ogni rigo rappresenti un intero giorno
- Ogni colonna rappresenti il campione quartorario relativo all'energia attiva

Un formato così presentato prende il nome, all'interno di MATLAB, di Cell Array.

I dati disponibili, fanno riferimento ad un POD e ad un quadro elettrico. Ricordiamo che:

- ad ogni quadro elettrico sono associati più lampioni (costituiti da apparecchi, sorgenti luminose, eventuali strutture che consumino energia elettrica come fotocamere, antenne ecc..), quindi il dato relativo ai consumi è riferito a tutti i lampioni afferenti al singolo quadro;
- nella nostra analisi non stiamo raggruppando i lampioni per quadro elettrico ma stiamo considerando i dati relativi ai lampioni afferenti al singolo POD.

Per effettuare la trasformazione dei dati al fine di applicare la PCA, è stata creata una funzione (Figura 15) che, presi i dati di input del Cell Array, restituisca un vettore contenenti i giorni (le osservazioni) e una matrice contenente le 96 variabili per ogni osservazione.

```
function [giorno,M]=separaGiorni(data)
C = strsplit(data{1,5},' ');
old_date=C{1};
M=[];
giorno=cell(1);
giorno{1}=old_date;
counter=1;
counterj=1;
for i =1:length(data)
    C = strsplit(data{i,5},' ');
    new_date=C{1};
    if strcmp(new_date,old_date)
        M(counter,counterj)=data{i,7};
        counterj=counterj+1;
    else
        counter=counter+1;
        counterj=1;
        M(counter,counterj)=data{i,7};
        giorno{counter}=new_date;
        counterj=counterj+1;
    end
    old_date=new_date;
end
```

Figura 15-Funzione che prende i dati di input del Cell Array e restituisca un vettore contenente i giorni (le osservazioni) e una matrice contenente le 96 variabili per ogni osservazione.

In seguito all'applicazione di questa funzione, è stato possibile finalmente applicare la PCA ai dati contenuti nella matrice M.

M è una matrice di dimensioni 158 x 96, cioè abbiamo trasformato i dati estratti da Hadoop in una struttura che presenta 158 osservazioni per 96 variabili.

È importante notare come, a questo punto, tutti i valori quartorari rappresentino le variabili del nostro sistema, indipendentemente dal loro significato; da un punto di vista squisitamente matematico, la PCA non fa altro che cercare di ridurre il problema di 96 variabili, in un problema di 4 o 5 variabili, raggruppando i dati in cluster e semplificando il problema.

MATLAB presenta già una funzione specifica che permette l'analisi delle componenti principali dei dati grezzi, sollevandoci dalla necessità di implementare tale funzione manualmente.

La funzione ***pca(M)*** è descritta nel seguente modo:

[coeff, score, latent, tsquared, explained]=pca(M)

Dove :

- coeff è la matrice n x n dei coefficienti delle componenti principali, definiti anche "loadings", per la matrice m x n M. Ogni colonna di coeff contiene coefficienti per un componente principale e le colonne sono in ordine decrescente di varianza del componente.

- score contiene i punteggi delle componenti principali, cioè le rappresentazioni di M nello spazio delle componenti principali. È una matrice di dimensioni 158 x 96, come la matrice M
- latent contiene le varianze delle componenti principali, rappresentati dagli autovalori della matrice di covarianza di M. Ha dimensione 96 x 1
- tsquared è la statistica T-squared di Hotelling per ogni osservazione
- explained contiene la percentuale della varianza totale per ogni componente principale.

Analizzando il dato contenuto in "explained", risulta la seguente tabella di 96 righe rappresentanti le componenti principali in cui, in ordine decrescente, è indicata la percentuale della varianza totale per ciascuna componente (Figura 16).

96x1 double

	1	2
1	65.6308	
2	18.7287	
3	10.7051	
4	2.1273	
5	0.9583	
6	0.4942	
7	0.3174	
8	0.1870	
9	0.1798	
10	0.1680	
11	0.1036	
12	0.0795	
13	0.0586	
14	0.0429	
15	0.0409	
16	0.0254	
17	0.0242	
18	0.0211	
19	0.0182	
20	0.0167	

Figura 16-tabella della percentuale della varianza totale per ciascuna componente.

Pertanto, da quanto si evince dai dati, le prime 4 componenti rappresentano il 97,2% della varianza totale, mentre le percentuali relative alle componenti successive risultano trascurabili rispetto ad esse.

Un grafico che descrive l'andamento della variabile "***explained***" per ogni componente principale è mostrato in Figura 17.

Esso si può ottenere con MATLAB attraverso la funzione:

pareto(explained)

La funzione ***"pareto"*** genera un grafico in cui i dati relativi alla percentuale di varianza sono mostrati come barre in ordine decrescente (non devono esserci valori negativi oppure NaN-not a number); per definizione

il grafico risultante mostra le prime dieci barre oppure le barre che rappresentano almeno il 95% della distribuzione cumulativa.

Il grafico in figura mostra solo le prime 3 componenti, perché esse rappresentano più del 95% della varianza globale.

Rispetto al dato numerico in Figura 16, nel grafico di Pareto è possibile notare in modo molto evidente il salto tra la prima componente e la seconda; infatti il primo componente, da solo, rappresenta più del 65% della varianza complessiva, mentre la seconda e la terza componente rappresentano rispettivamente poco più del 18% e del 10 % della varianza. Questo ci può aiutare a definire quante e quali componenti considerare nell'analisi, riducendo le dimensioni del problema e quantificare in qualche modo la perdita di informazione dovuta all'esclusione delle altre componenti nell'analisi.

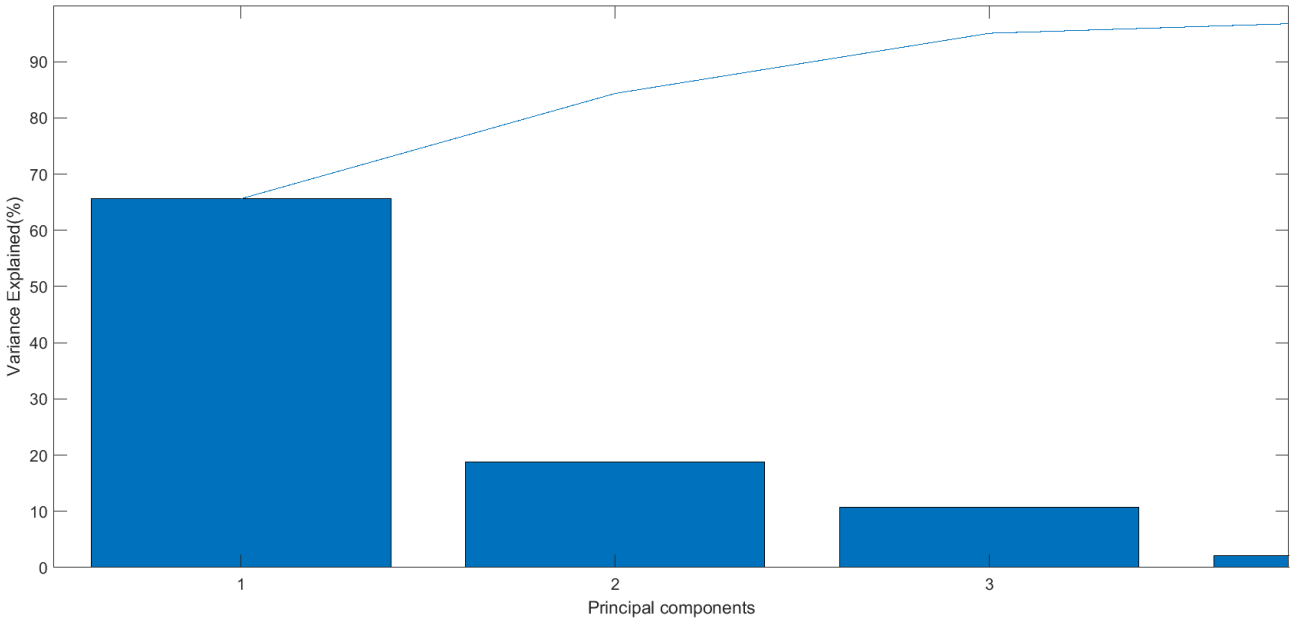


Figura 17- un grafico in cui i dati relativi alla percentuale di varianza sono mostrati come barre in ordine decrescente

Lo studio delle componenti a maggior varianza permette di ridurre considerevolmente la variabilità dei dati, mantenendo una buona quantità di informazione; se si desidera migliorare l'informazione riducendo l'incertezza è sempre possibile aumentare il numero di componenti da studiare

2.3.4 Numero di componenti principali da considerare

Come già evidenziato, scopo della PCA è quello di ridurre la complessità del sistema, ridimensionando il numero di variabili da studiare.
Le componenti che costituiscono la maggiore varianza sono recuperabili graficamente dal ***grafico di pareto*** e numericamente dalla variabile ***"explained"***.
Al fine di scegliere il numero di componenti ideali si possono utilizzare i seguenti criteri:

- Considerare le componenti che rappresentano l'80/90% della varianza
- Prendere le variabili che abbiano varianza maggiore della varianza media (la media di tutte le varianze di tutte le componenti), cioè la ***"Regola di kaiser"***

Nel caso in oggetto sono state considerate le prime 3 componenti principali, e attraverso i grafici bidimensionali a dispersione, sono stati rappresentate le combinazioni tra le prime 3 componenti, al fine di evidenziare eventuali raggruppamenti che potessero essere interessanti.

2.3.5 Risultati

Si è scelto dunque di effettuare il grafico di scattering, relativo alle prime 2 componenti PCA1 e PCA2. La funzione utilizzata è la seguente:

scatter(score(:,1),score(:,2))

da cui il grafico in Figura 18.

Il grafico delle prime due componenti principali mostra una clusterizzazione che riduce la complessità a pochi cluster da studiare.

Si evidenziano 6 clusters ed alcuni dati sparsi che devono essere analizzati perché non inclusi in alcun cluster.

Come si evince dal grafico, non vi è alcuna informazione temporale su giorni ed ore relative ai dati, infatti, i 96 dati quartorari sono stati trattati come variabili indipendenti l'una dall'altra.

Al fine di studiare la tipologia di dato descritta da ogni cluster è stato necessario, per ciascuno di essi, recuperare le osservazioni (quindi i giorni) a cui le componenti fanno riferimento.

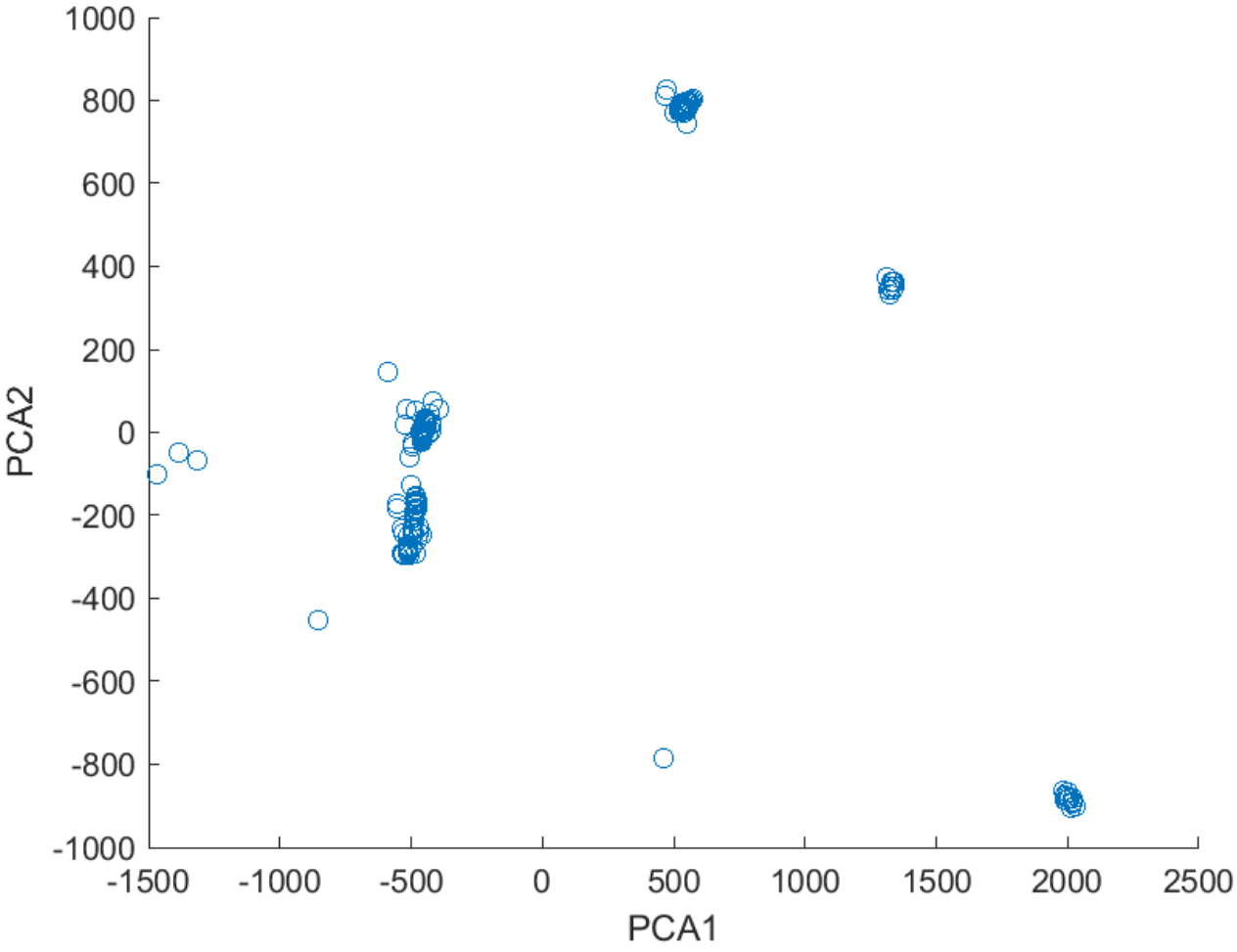


Figura 18-grafico delle prime due componenti principali

Ogni punto sul piano cartesiano è individuato da un valore di PCA1 e un valore di PCA2, pertanto si è pensato di isolare ogni cluster attraverso una finestra che permettesse di individuare tutti i punti ad esso associati.

Una volta identificati i punti di minimo e di massimo per PCA1 e PCA2 è stata applicata la funzione ***getClusterIndex*** (riportata in Figura 19), che restituisce un vettore contenente gli indici relativi alle osservazioni, quindi ai giorni.

```
function [V]=getClusterIndex(xmin,xmax,ymin,ymax,score)
V=[];
%per ciascun intervallo di x e y recupero gli indici
for i =1:length(score)
    valx = score(i,1);
    valy = score(i,2);
    %if(val < 0)
    %    val = -val;
    %end
    if(valx >= xmin && valx <= xmax && valy >= ymin && valy <= ymax )
        V(length(V)+1)= i;
        i=i+1;
    else
        i=i+1;
    end
end
```

Figura 19- funzione prende in input i valori minimi e massimi per le componenti PCA1 e PCA2, nonché la tabella "score" che contiene le componenti principali, e per ogni record per cui le componenti sono comprese negli intervalli passati come input, viene recuperato l'indice ed inserito nel vettore

La funzione ***getClusterIndex*** prende in input i valori minimi e massimi per le componenti PCA1 e PCA2, nonché la tabella "score" che contiene le componenti principali, e per ogni record per cui le componenti sono comprese negli intervalli passati come input, viene recuperato l'indice ed inserito nel vettore.

Il vettore risultante V contiene tutti gli indici associati a quel cluster specifico.

Il passo successivo ha previsto l'analisi dei dati attraverso la visualizzazione grafica degli stessi per ogni cluster.

Attraverso la funzione:

plot(M(V,:)')

è stato possibile effettuare il grafico dell'andamento delle 96 variabili per i giorni relativi al cluster; il grafico risultante presenta in ascissa le 96 variabili, in ordinata i valori di energia attiva per ogni campione.

Si fa presente che, nel caso specifico di 96 variabili che corrispondono ai 96 valori quartorari, il grafico risultante coincide con l'andamento temporale, per cui il significato del grafico è quello di visualizzare la sovrapposizione di tutti i campioni quartorari di energia attiva, per tutti i giorni appartenenti al cluster analizzato.

Ovviamente il grafico assume questo significato in virtù del fatto che le 96 variabili corrispondono all'intero giorno; infatti, potremmo pensare di incrementare il numero di variabili in modo da aggiungere informazioni interessanti come, ad esempio, la stagionalità.

Di seguito l'analisi dettagliata di ogni cluster e dei dati in essi compresi.

2.3.6 Applicazioni della PCA al set di dati

Il primo cluster che si vuole analizzare è quello, ingrandito, in Figura 20, che presenta valori di:

- PCA1 compreso tra 1950 e 2050
- PCA2 compreso tra -860 e -910.

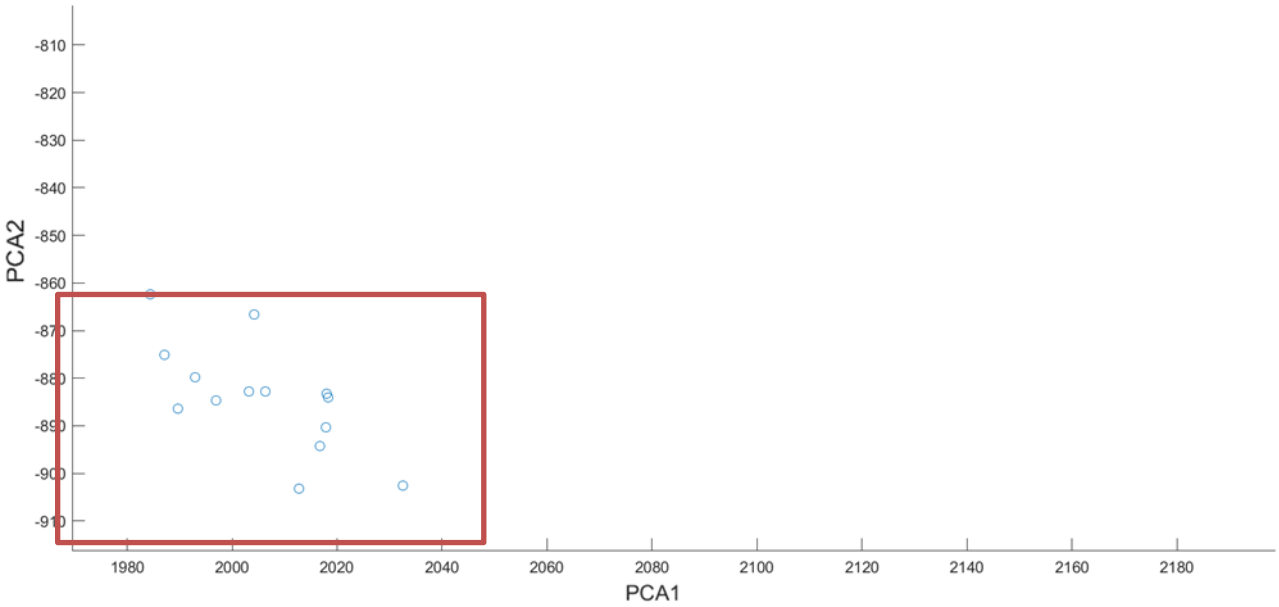


Figura 20-dettaglio del cluster con PCA1 compresa tra 1950 e 2050 e PCA2 compreso tra -860 e -910

Applicando la funzione apposita, il vettore V è un array di 14 elementi con indici compresi tra 1 e 14.

Come atteso, ogni cluster è costituito prevalentemente da osservazioni con indici vicini tra loro; ad esempio, per questo cluster, i dati corrispondono agli unici 14 giorni del 2019 presenti nel file.

Il secondo cluster analizzato (Figura 21) è costituito da valori di:

- PCA1 tra1200 e 1400
- PCA2 tra 300 e 400

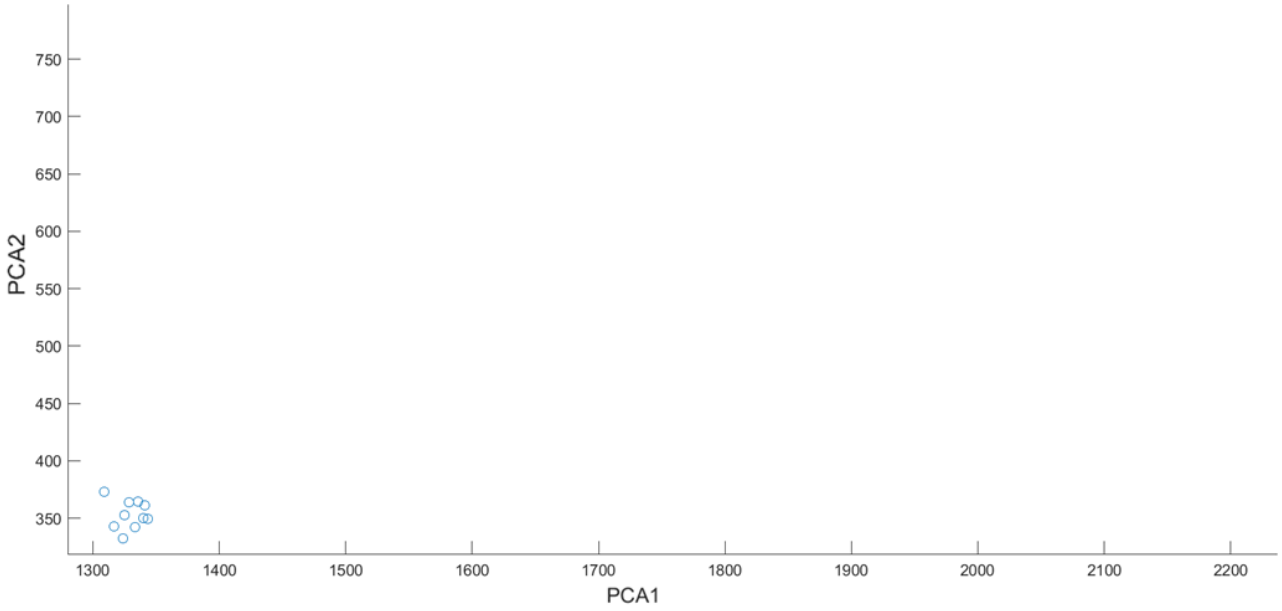


Figura 21-dettaglio del cluster con PCA1 tra1200 e 1400 e PCA2 tra 300 e 400

In questo caso il vettore V degli indici delle osservazioni comprende 10 elementi con indici tra 45 e 54, che corrispondono ai giorni dal 09/08/2020 al 18/08/2020 per cui risulta un consumo continuo per tutto il giorno come mostrato nella figura seguente (Figura 22):

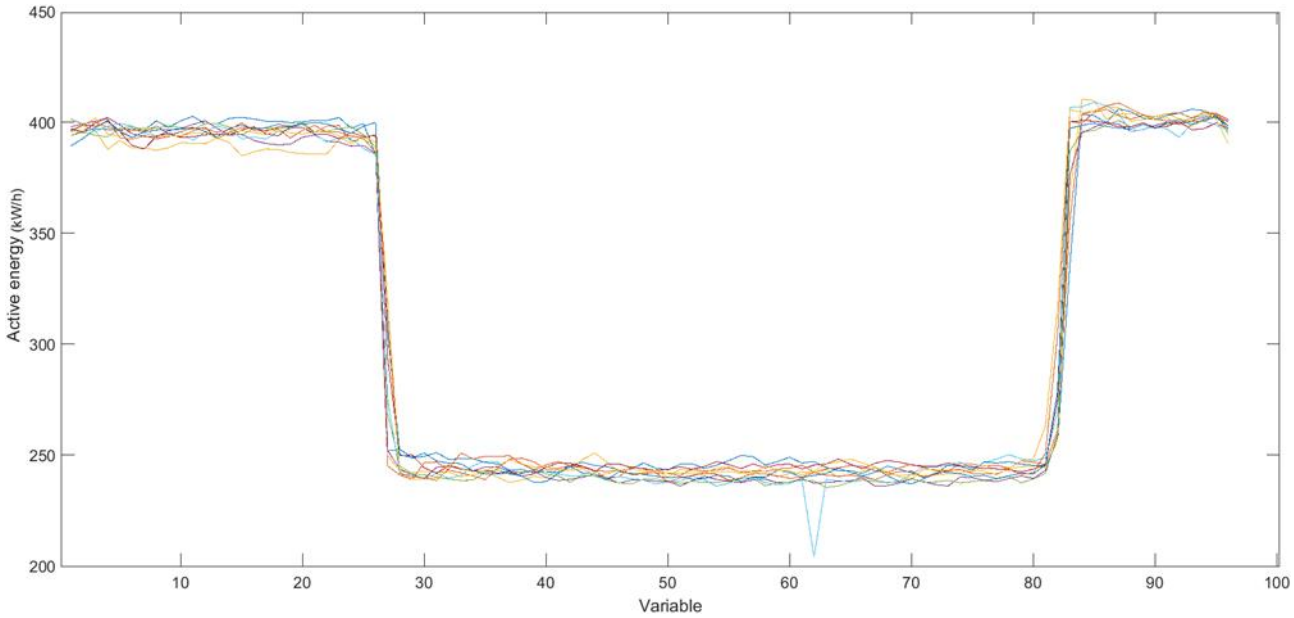


Figura 22-grafico delle osservazioni relative al periodo dal 09/08/2020 al 18/08/2020, in cui risulta un consumo continuo per tutto il giorno

La figura mostra chiaramente l'andamento dell'energia attiva durante il giorno, evidenziando come, seppur diminuendo intorno ai 250 kW/h verso le 7:00 del mattino, in realtà il consumo di energia attiva sia comunque elevato per gran parte del giorno in cui, invece, il consumo dovrebbe essere ridotto.

Una possibile motivazione può essere, ad esempio, un errato funzionamento del crepuscolare in caso di giornata nuvolosa.

Il terzo cluster analizzato è costituito da valori di:

- PCA1 450 e 600

- PCA2 tra 750 e 850

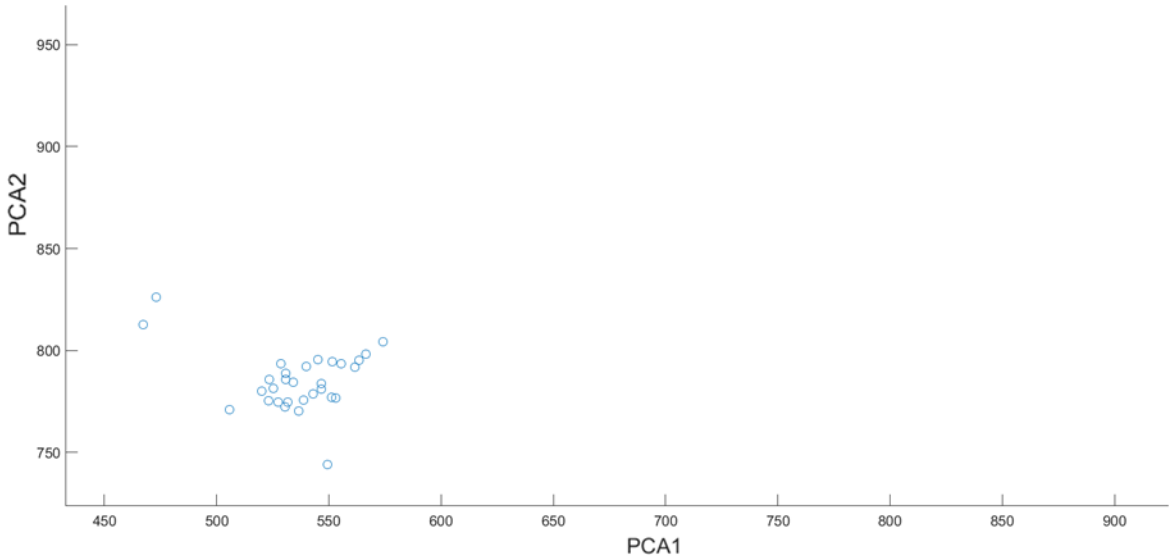


Figura 23-dettaglio del cluster con PCA1 tra 450 e 600 e PCA2 tra 750 e 850

In tal caso il vettore V comprende gli indici da 15 a 44, sebbene il 15 (10/07/2021), il 27 (22/07/2021) e il 44 (08/08/2021) siano leggermente fuori dal cluster.

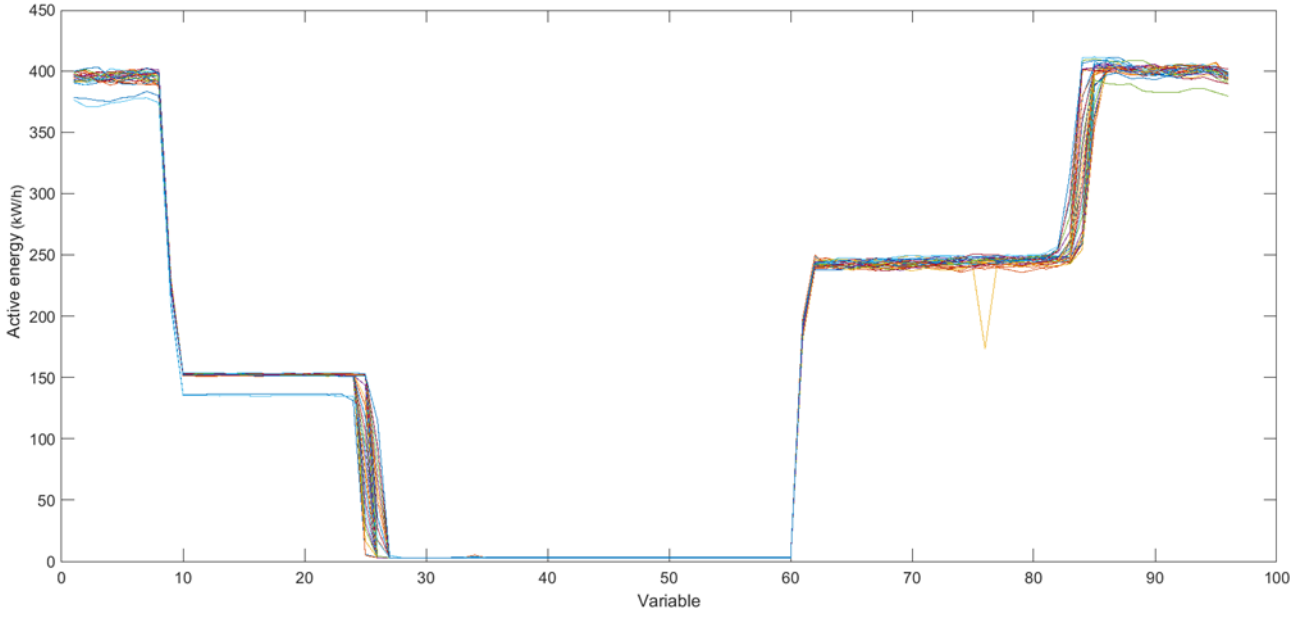


Figura 24- grafico delle osservazioni relative al periodo dal 10/07/2020 allo 08/08/2020, in cui risulta una possibile anomalia di funzionamento

La Figura 24 mostra l'andamento dell'energia attiva per tutti i 96 campioni quartorari; i primi campioni (0-8 corrispondenti circa dalle 00:00 alle 02:00), mostrano un consumo di circa 400 kW/h che diminuisce intorno a 150 kW/h per i successivi campioni (9-24 corrispondenti circa dalle 02:00 alle 06:00) fino allo spegnimento durante le ore centrali del giorno (25-60 corrispondenti circa dalle 06:00 alle 15:00).

In seguito il consumo di energia attiva aumenta nuovamente intorno ai 250 kW /h fino alle 20:00 circa, per poi raggiungere i 400 kW/h.

Poiché il periodo indicato dai campioni è quello che va dal 10/07/2020 allo 08/08/2020, potrebbe esserci una anomalia in quanto l'orario di accensione risulterebbe troppo prolungato rispetto alle necessità stagionali.

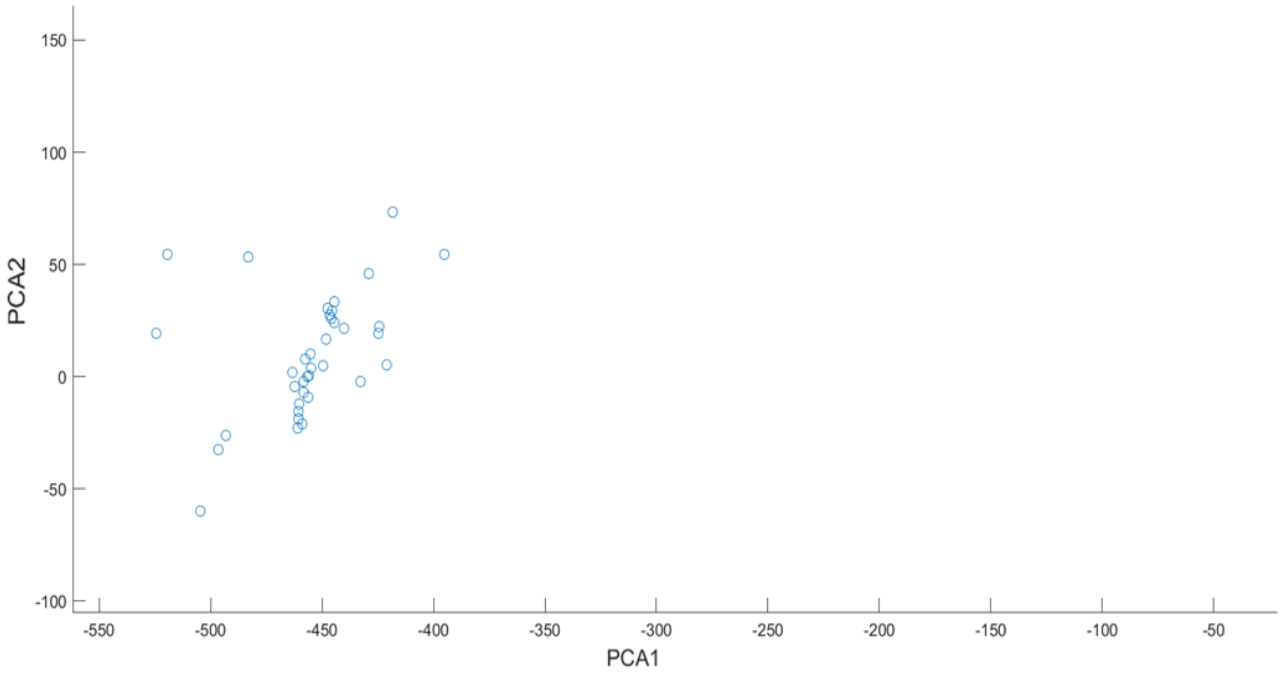
Il quarto cluster (Figura 25) analizzato è costituito da valori di:

- PCA1 tra -550 e -350
- PCA2 tra -100 e 100

Il cluster sembra abbastanza ampio, il vettore V comprende 37 valori con indici che vanno da 114 a 150, corrispondenti al periodo dal 29/10/2020 al 10/12/2020.

La Figura 26, mostra un andamento specifico dell'energia attiva durante il giorno, e da essa è possibile ricavare alcune informazioni interessanti.

- Il valore dell'energia attiva durante le ore di funzionamento dell'impianto, relativamente al POD studiato, risulta significativamente più basso rispetto al valore dell'energia attiva visto nei cluster descritti in precedenza. Infatti essi presentavano valori massimi di 400 kW/h, mentre per questo cluster stiamo sull'ordine dei 160 kW/h.
- La sovrapposizione dei dati giornalieri mostra alcuni dati che, a parte l'andamento medio, presentano anomalie rispetto ad esso. Questo giustifica l'ampiezza del cluster e può porre interessanti spunti di analisi in merito ai giorni che presentano le anomalie.



- **Figura 25-dettaglio del cluster con PCA1 tra -550 e -350e PCA2 tra-100 e 100**

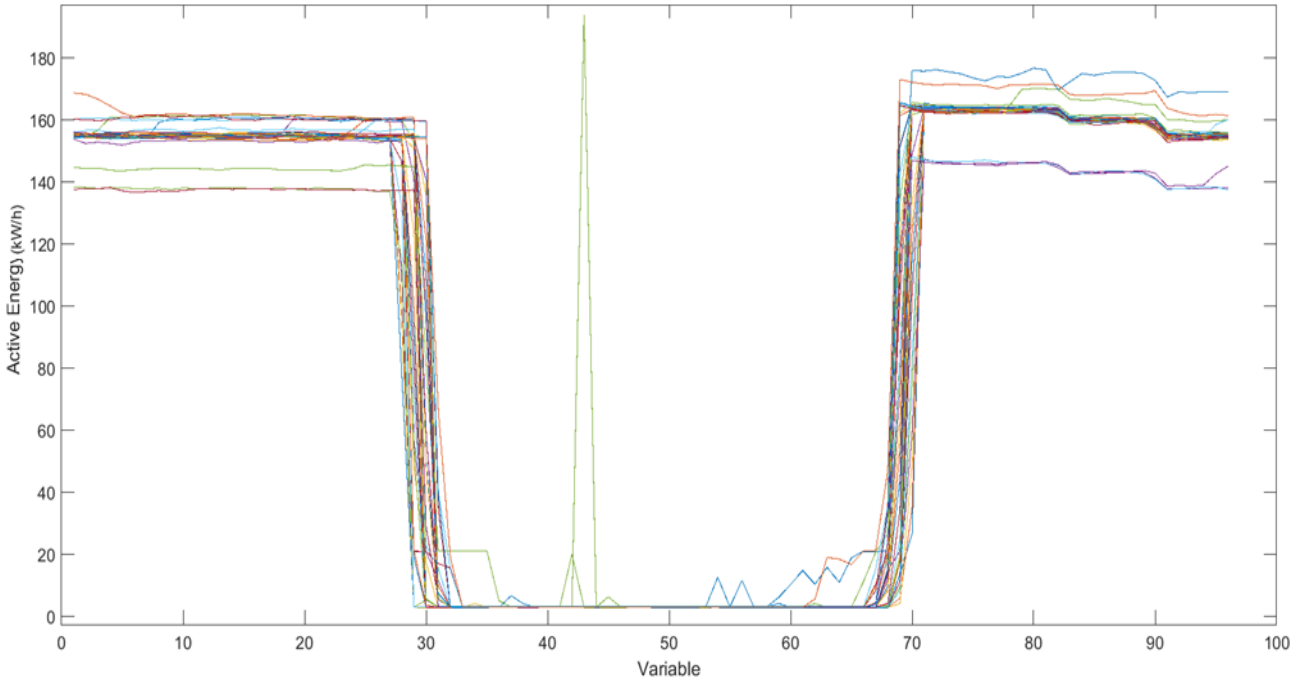


Figura 26-grafico delle osservazioni relative al periodo dal 29/10/2020 al 10/12/2020

Il giorno con indice 146, ad esempio, presenta un picco al 43esimo campione della giornata, cioè le 10:45, evidenziando un consumo anomalo temporaneo.

Anche altri giorni del cluster presentano andamenti anomali, ma se si considera la media in Figura 27, si può vedere come tali anomalie, nel contesto globale dell'andamento dell'energia, possano considerarsi trascurabili.

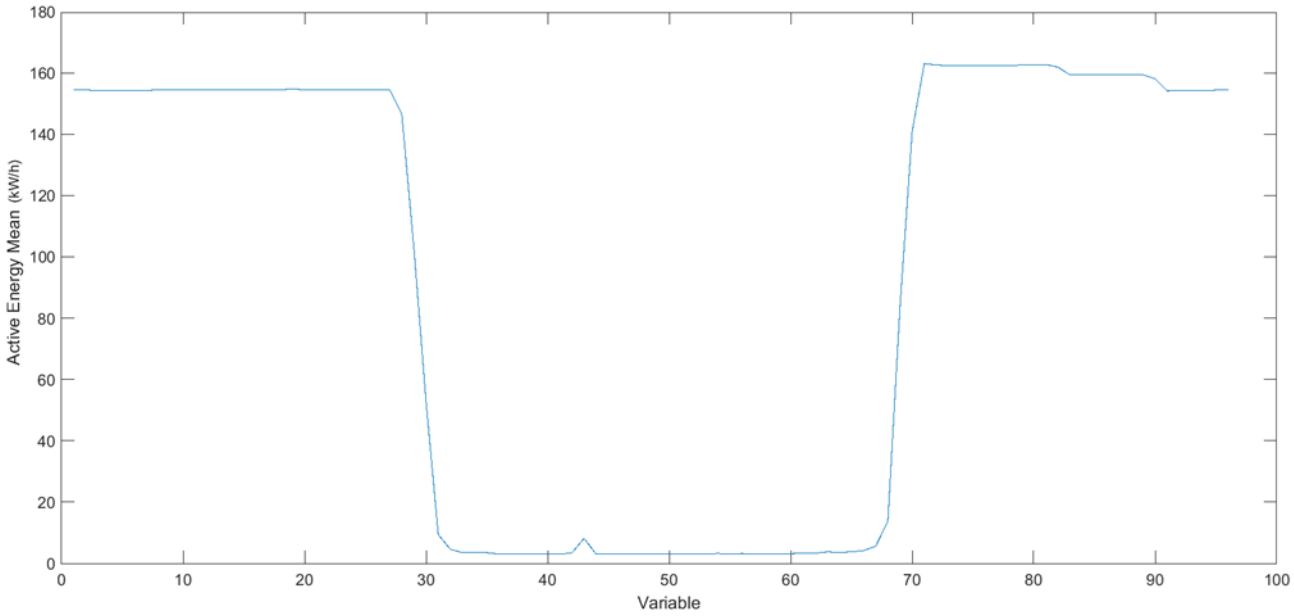


Figura 27-media delle osservazioni relative al periodo dal 29/10/2020 al 10/12/2020

Il quinto cluster (Figura 28) è costituito da 56 valori, per i quali:

- PCA1 tra -555 e -450
- PCA2 tra -300 e -120

Che corrispondono agli indici dal 56 al 112 escluso il 79

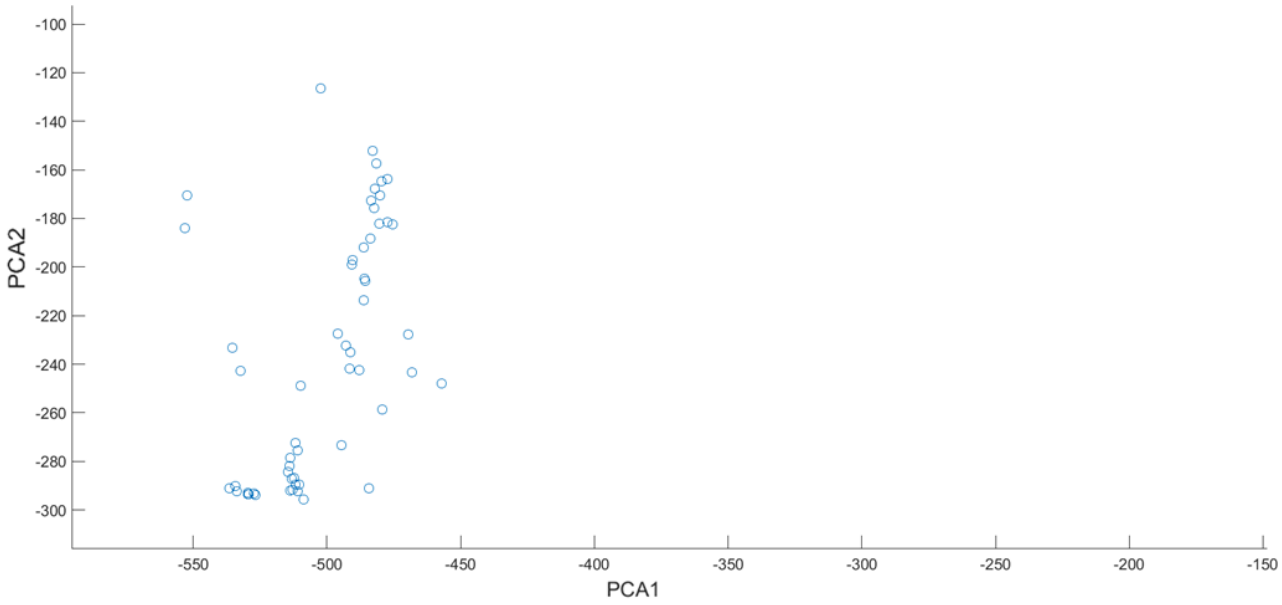


Figura 28-dettaglio del cluster con PCA1 tra --555 e -450 e PCA2 tra -300 e -120

Anche questo cluster è abbastanza ampio e comprende dati dal 20/08/2020 al 24/10/2020. (Figura 29 e Figura 30)

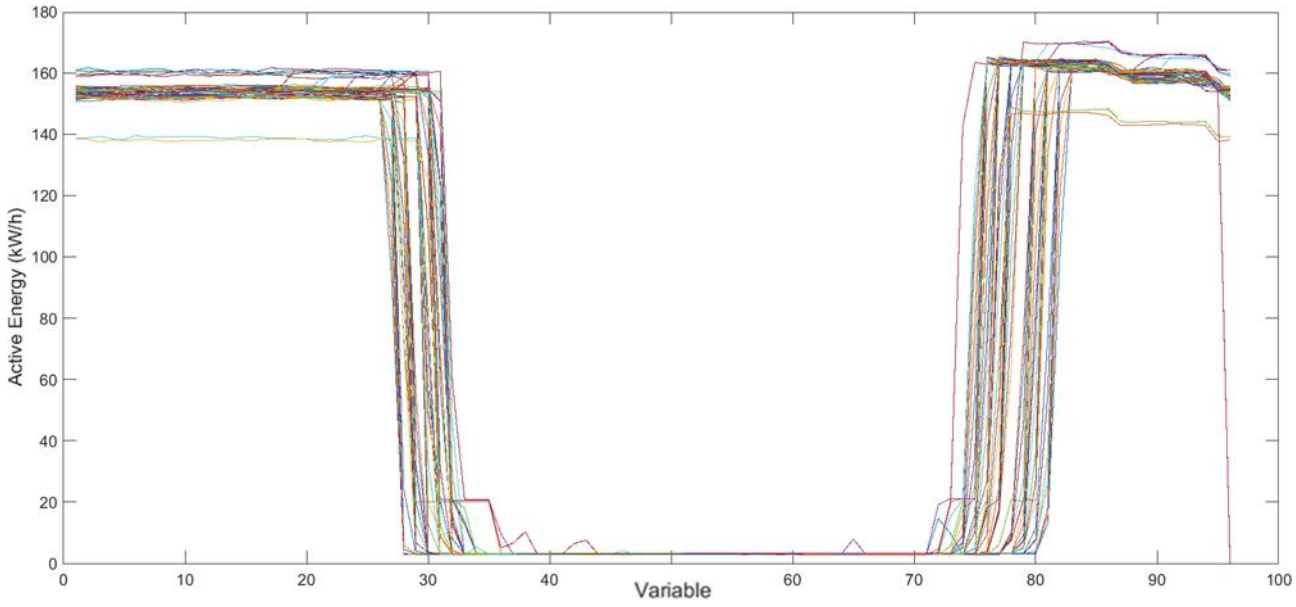


Figura 29-grafico delle osservazioni relative al periodo dal 20/08/2020 al 24/10/2020

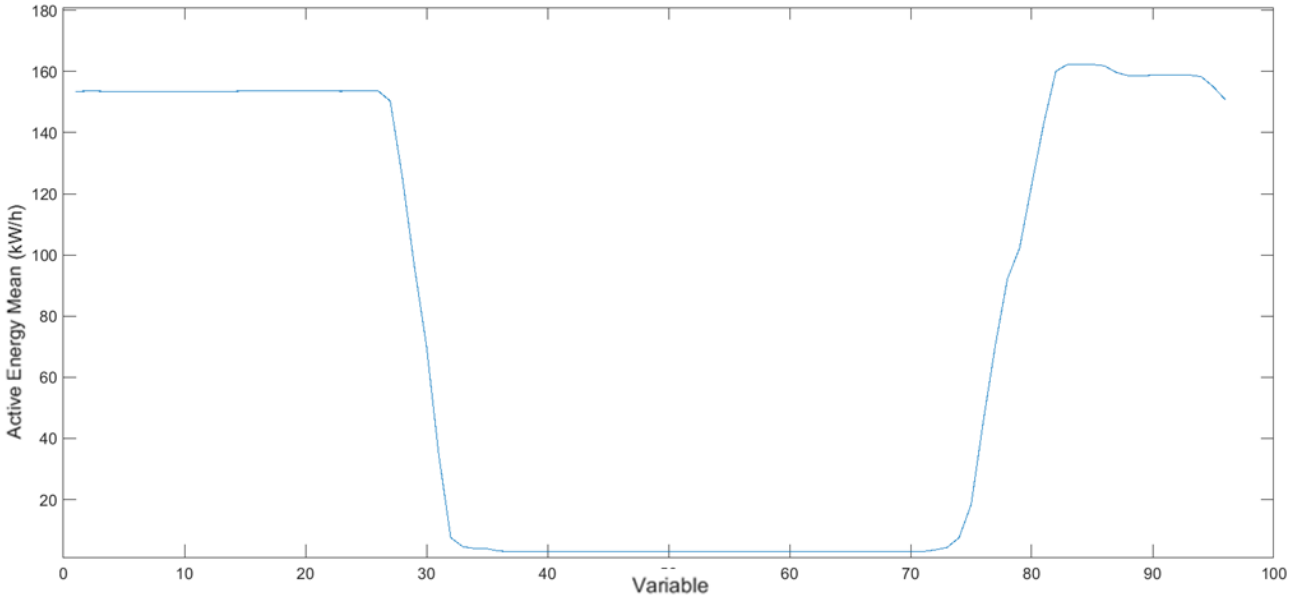


Figura 30-media delle osservazioni relative al periodo dal 20/08/2020 al 24/10/2020

A differenza del cluster precedentemente analizzato, si può notare che in quest'ultimo l'orario di accensione pomeridiano dell'impianto è successivo rispetto al precedente, infatti, questo è spiegabile considerando che il precedente cluster comprendeva un periodo autunno/inverno, mentre questo cluster comprende un periodo estivo che prevede una accensione posticipata.

Un ultimo cluster rilevato dall'analisi delle componenti principali, è quello per cui:

- PCA1 tra -1470 e -1468
- PCA2 tra -102 e -100

Che corrispondono agli indici 152, 153, 154, 155, 157 e 158, che corrispondono al periodo dal 04/12/2020 al 09/12/2020 (Figura 31 e Figura 32).

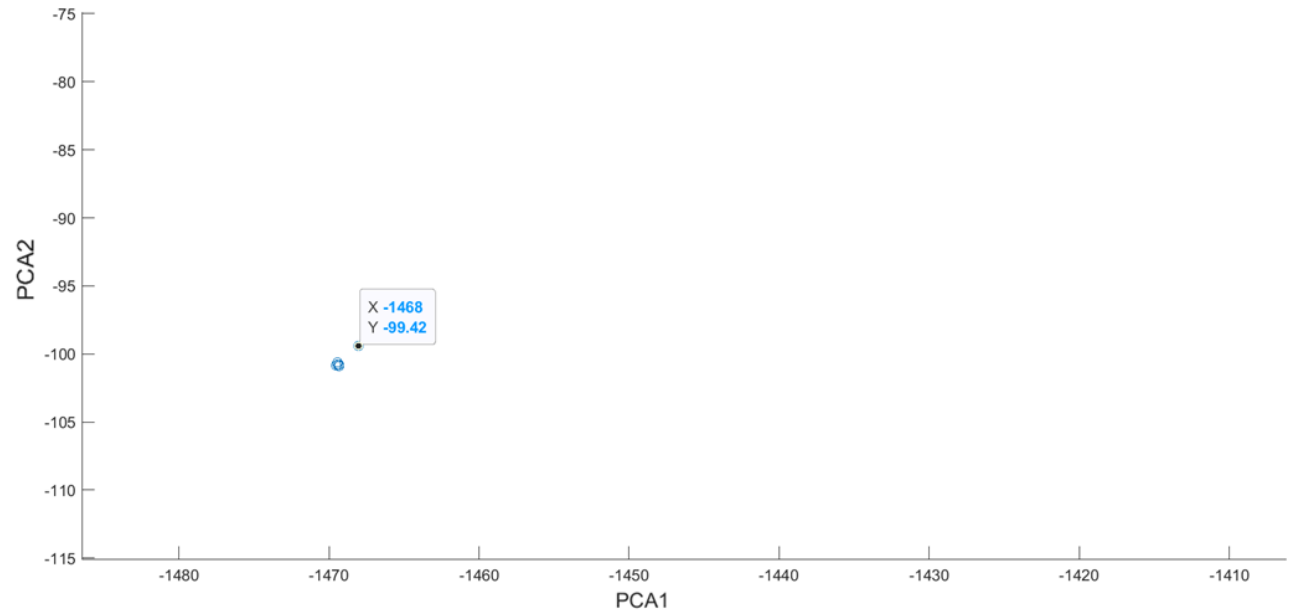


Figura 31-dettaglio del cluster con PCA1 tra -1470 e-1468 e PCA2 tra -102 e -100

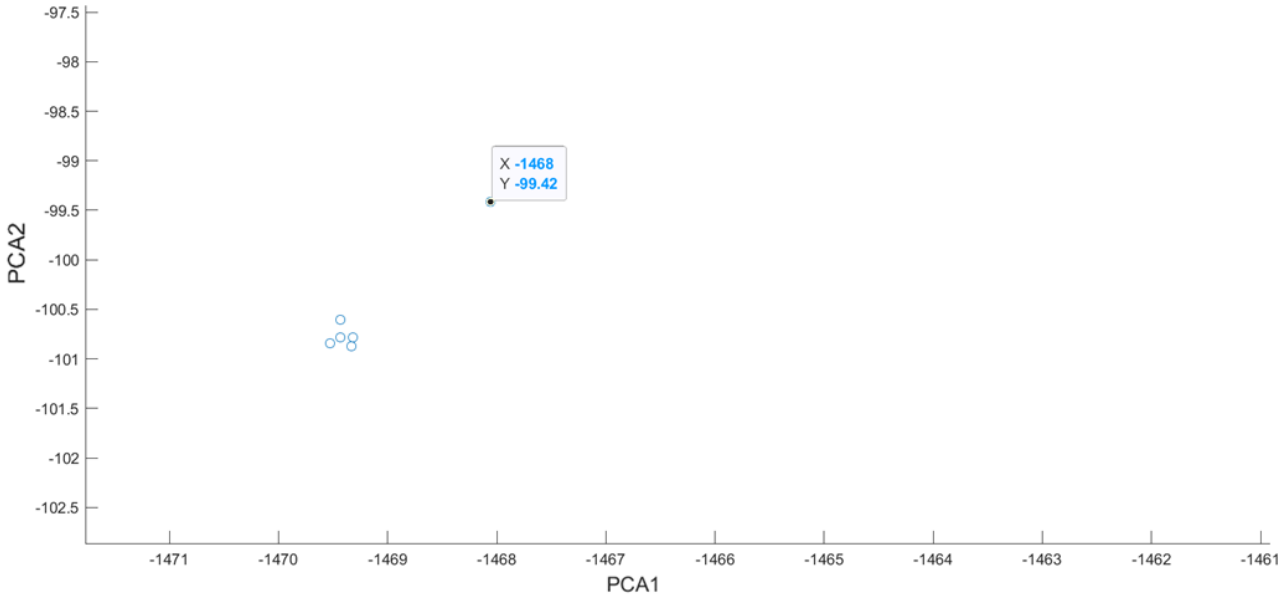


Figura 32-dettaglio del cluster con PCA1 tra -1470 e-1468 e PCA2 tra -102 e -100 in cui si vede la vicinanza molto stretta tra le varie osservazioni.

Analizzando l'andamento dell'energia attiva, risultano evidenti anomalie nei dati di questo cluster; in particolare si rileva l'assenza di molti dati per i giorni relativi, che evidenziano un possibile malfunzionamento nel periodo oggetto di analisi.

La Figura 33 mostra i dati che non sembrano appartenere ad alcuno dei cluster analizzati, pertanto rappresentano delle eccezioni che possono essere studiate.

È da specificare che l'analisi dei cluster non è sempre semplice, infatti se consideriamo il quarto ed il quinto cluster, essi sono molto vicini tra loro e possono essere analizzati come unico cluster oppure separatamente come clusters distinti. In quest'ultimo caso può non essere facile associare i dati ad un cluster piuttosto che ad un altro, in ogni caso ci è sembrato utile studiare i due clusters in modo separato per evidenziarne le differenze, in particolare per quanto riguarda l'orario pomeridiano di accensione dell'impianto.

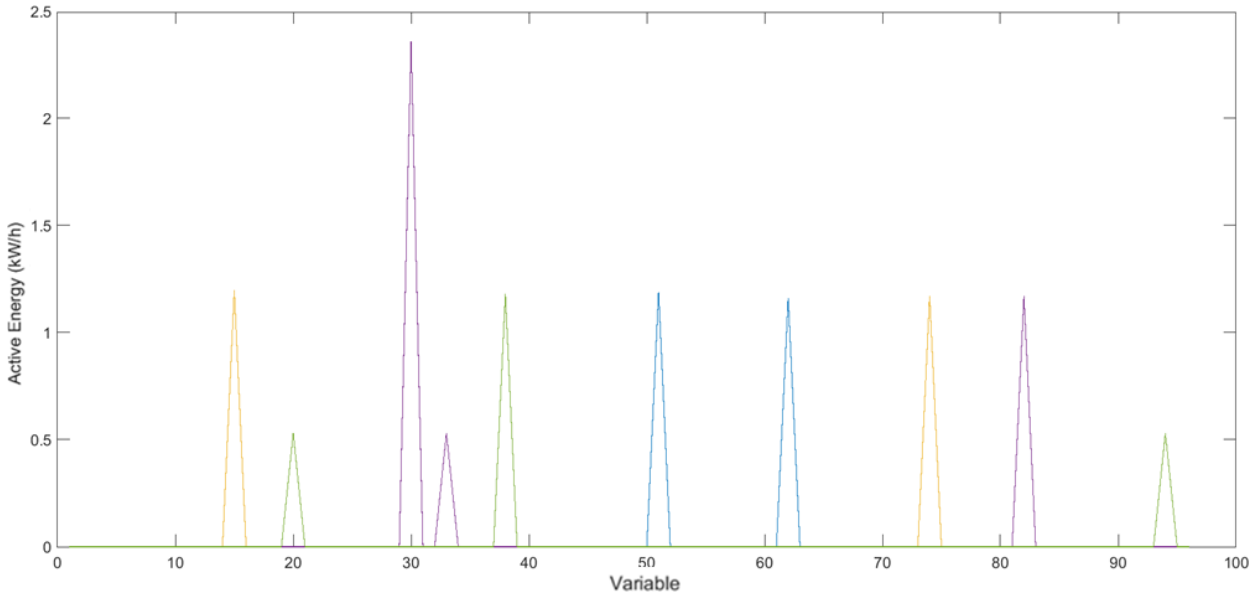


Figura 33 –grafico delle ossevazioni elative ai dati sparsi che non sembrano appartenere ad alcuno dei cluster analizzati, pertanto rappresentano delle eccezioni

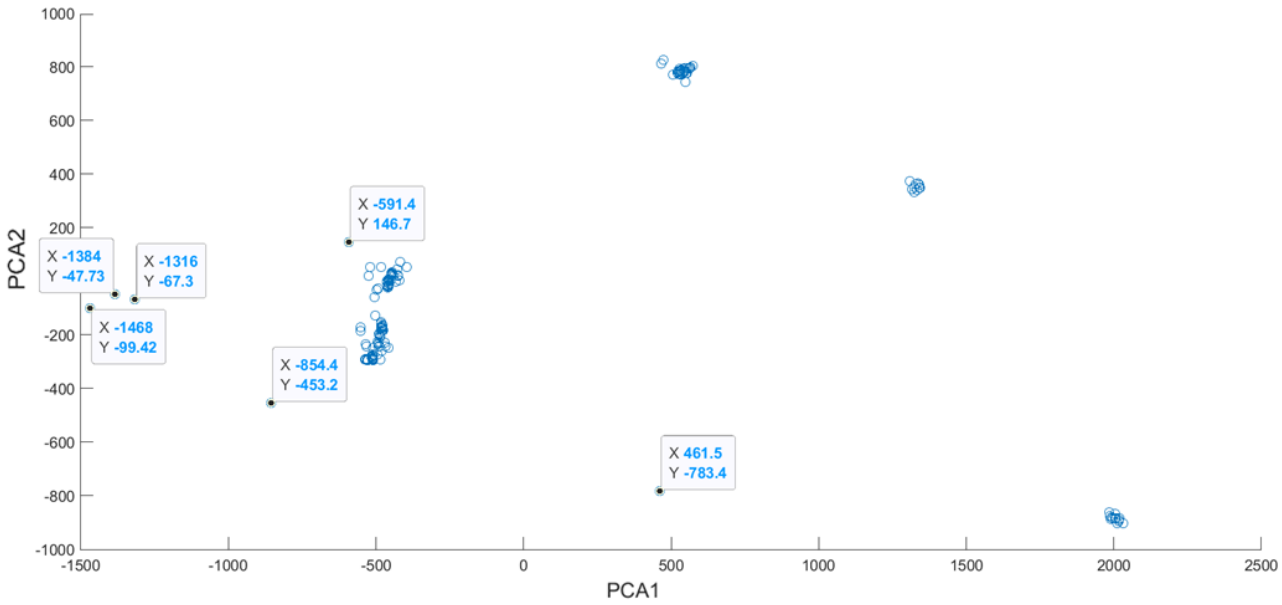


Figura 34-visulizzazione dei dati sparsi e relative componenti

I dati "sparsi" non associati ad alcun cluster (Figura 34) sono pochi, e corrispondono a record giornalieri in cui più campioni sono assenti.

Può essere utile riepilogare, attraverso la Figura 35, i dati dei cluster analizzati fino ad ora. In ogni caso si fa presente che i periodi possono non comprendere tutti i giorni, in quanto, in alcuni casi, sono totalmente assenti i dati giornalieri. La PCA non considera eventuali giorni mancanti, ma considera esclusivamente l'aggregazione dei dati in clusters indipendentemente dalla loro numerosità e dalla temporalità.

Indici	Periodo	Num. campioni
1-14	26/06/2019-09/07/2019	14
15-44	10/07/2020-08/08/2020	30
45-54	09/08/2020-18/08/2020	10
56-78;80-112	20/08/2020-14/09/2020-22/09/2020-24/10/2020	56
114-150	26/10/2020-01/12/2020	37
152-155;157-158	03/12/2020-06/12/2020-08/12/2020-09/12/2020	6
79,156,151,113,55		5

Figura 35-riepilogo dati dei cluster analizzati

Come già espresso in precedenza, l'utilizzo della PCA, attraverso una trasformazione lineare delle variabili di partenza, offre l'opportunità di analizzare l'eventuale presenza di aggregazioni in uno spazio trasformato.

Questi potenziali nuovi raggruppamenti potrebbero risultare interessanti perché basati su nuove componenti a varianza elevata. Infatti, le prime tre componenti rappresentano il 97% della varianza totale e allora, oltre ad analizzare solo le prime due componenti, è risultata utile anche l'analisi della prima e della terza componente, il cui grafico è riportato in Figura 36.

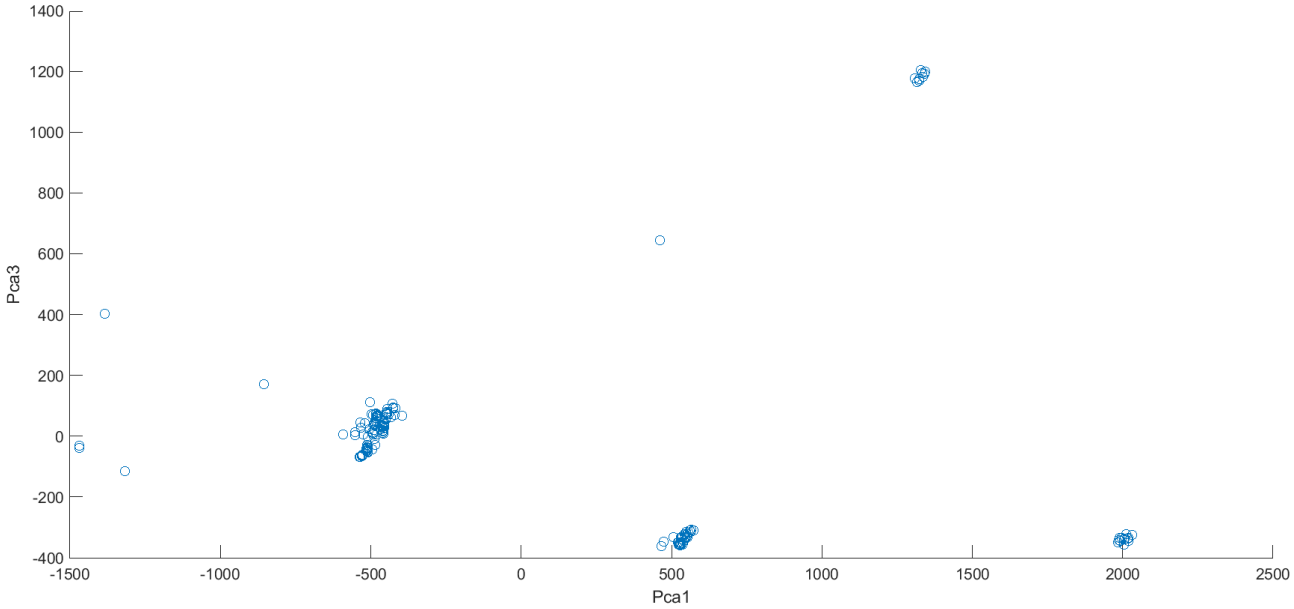


Figura 36-grafico relativo alla prima e alla terza componente

Analizzando la prima e la terza componente si evidenzia una clusterizzazione più marcata rispetto al precedente caso che vedeva coinvolte le prime due componenti.

A tal fine, si è modificata la funzione in Figura 37 che restituisce gli indici relativi al cluster, per utilizzare genericamente una componente per l'asse y passata come parametro della funzione.

```
function [V]=getClusterIndex1_3(xmin,xmax,ymin,ymax,score,yind)
V=[];
%per ciascun intervallo di x e y recupero gli indici
for i =1:length(score)
    valx = score(i,1);
    valy = score(i,yind);
    %if(val < 0)
    %    val = -val;
    %end
    if(valx >= xmin && valx <= xmax && valy >= ymin && valy <= ymax )
        V(length(V)+1)= i;
        i=i+1;
    else
        i=i+1;
    end
end
```

Figura 37-funzione che restituisce gli indici relativi al cluster, per utilizzare genericamente una componente per l'asse y passata come parametro della funzione

Analizzando i vari cluster come già fatto per le prime due componenti (l'analisi dettagliata non viene riportata per motivi pratici) si evince la situazione indicata nella Figura 38:

Indici	Periodo	Num. campioni
1-14	26/06/2019-09/07/2019	14
15-44	10/07/2020-08/08/2020	30
45-54	09/08/2020-18/08/2020	10
56-78; 80-150	20/08/2020-01/12/2020	94
152-155;157-158	03/12/2020-06/12/2020-08/12/2020-09/12/2020	6
79,156,151 ,55		4

Figura 38-riepilogo dati dei cluster analizzati per le componenti 1 e 3

Che non evidenzia nuovi cluster rispetto alla situazione precedente.

Tuttavia il quarto cluster, quello più numeroso, raggruppa sia il quarto che il quinto cluster relativi all'analisi PCA1-PCA2, evidenziando come questi ultimi siano trattati come cluster unico in relazione alle componenti PCA1-PCA3.

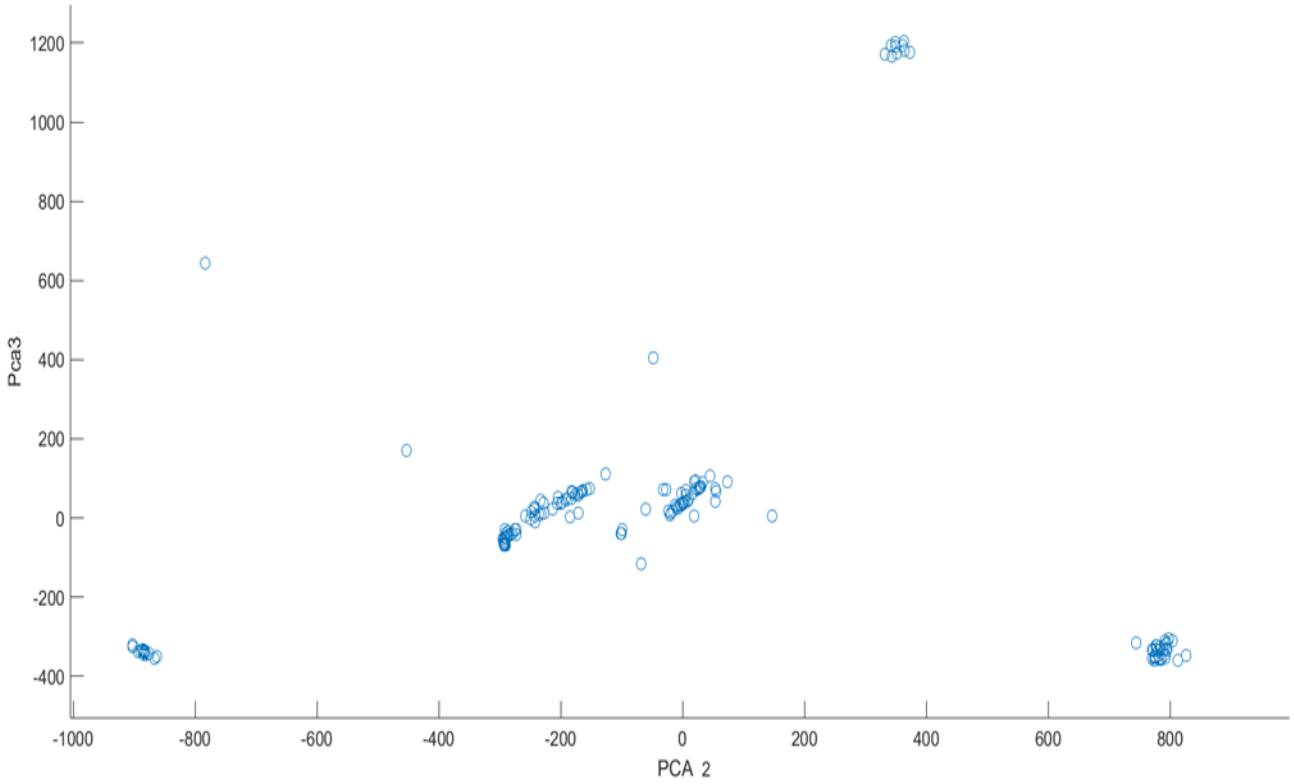


Figura 39-grafico relativo alla terza e alla seconda componente

Per completezza si è proceduto al grafico delle componenti PCA2-PCA3 per verificare l'eventuale presenza di cluster non ancora identificati e, dall'analisi, non risultano nuovi scenari rispetto a quelli precedentemente indicati (Figura 39).

2.3.7 Analisi dati tramite rete neurale

Al fine di effettuare ulteriori valutazioni e verificare la presenza di nuovi cluster, si è scelto di utilizzare una rete neurale di tipo SOM (Self Organizing Map) utilizzando in ingresso il dataset già disponibile.

MATLAB dispone di un tool specifico per la gestione delle reti neurali, NNCT (Neural Network clustering tool) che si propone di risolvere un problema di clustering mediante l'uso di una rete di tipo SOM attraverso un procedimento grafico.

In una rete di questo tipo, è previsto in ingresso un dataset (lo stesso già utilizzato per la PCA), e la creazione di una mappa di dimensioni variabili (ad es, 5x5,7x7…) che viene addestrata per un certo numero di epoche.

Si ricorda che il dataset è costituito dunque da 158 osservazioni, cioè i giorni che presentano i dati di misura, e da 96 variabili di input che sono i campioni quartorari.

Il tool di MATLAB è molto utile per progettare la rete neurale, perché prevede degli step che evidenziano le attività senza dover mettere mano alle singole funzioni, ed è pertanto utile anche a chiarire il funzionamento di una rete neurale.

Di seguito viene descritto brevemente il significato della rete neurale di tipo SOM, il tool di MATLAB che ci permette di comprendere gli step di gestione della rete neurale, e poi le funzioni utilizzate; in seguito presentiamo l'analisi dei dati prodotti dalle rete neurali e li confrontiamo con quanto ottenuto dalla PCA.

2.3.8 SOM –Self Organizing Map

Similmente alla Principal Component Analysis, le reti di tipo SOM, anche dette Mappe di Kohonen, sono utilizzate per ridurre la complessità in termini di dimensioni di problemi con elevati numeri di variabili, ma contrariamente alla PCA, mediante un approccio non lineare.

La mappa prevede una "griglia" di neuroni artificiali i cui "pesi" vengono definiti in fase di apprendimento (non supervisionato). Tali pesi vengono aggiornati in modo continuo ed automatico quando vengono presentati nuovi dati di ingresso attraverso un processo di "competizione" in cui il nodo che presenti un vettore dei pesi più vicino ad un certo input risulta vincitore. Ogni nuovo vettore di input opera una competizione, in cui il neurone vincitore (a cui quindi è associata l'osservazione di ingresso) è quello per cui il vettore dei pesi risulti a minor distanza dal vettore in ingresso, posizionando dunque oggetti simili più vicini ed oggetti diversi più distanti.

Il processo avviene in modo automatico, pertanto la rete si auto organizza attraverso la competizioni tra neuroni, specializzandosi qualora vengano aggiunti dati in ingresso.

2.3.9 Neural network Clustering Tool

Il tool di MATLAB per la gestione di reti neurali autorganizzanti (es reti di Kohonen) permette di configurare ed istruire la rete attraverso una rappresentazione grafica e, quindi, semplice per gli utenti.

Esso fornisce anche una serie di grafici che permettono di valutare visivamente i risultati e le performances della rete, e fornisce la possibilità di personalizzare le caratteristiche della rete stessa in modo semplice e rapido (Figura 40).

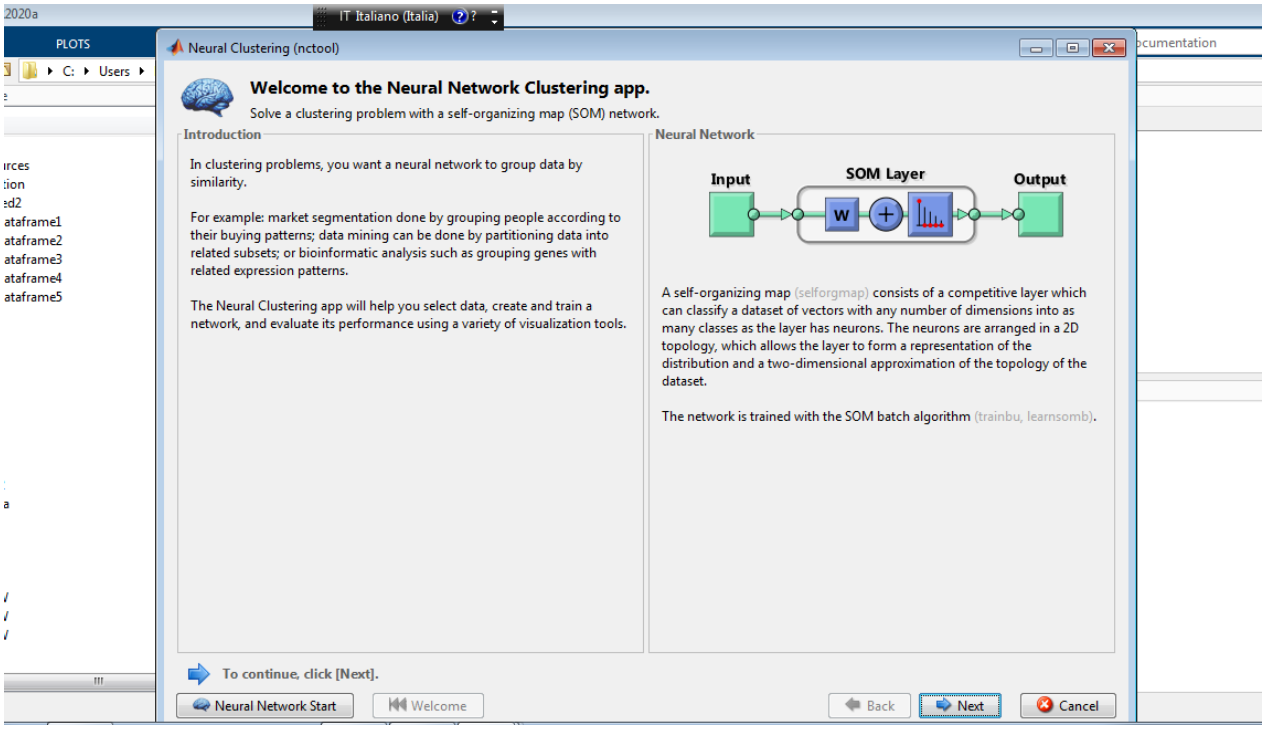


Figura 40-tool di MATLAB per la gestione di reti neurali autorganizzanti

La Figura 41mostra la sezione di inserimento dei dati da analizzare e con i quali viene inizialmente istruita la rete. Qui deve essere selezionata la variabile con le osservazioni e i dati di input, stabilendo se i campioni sono relativi alle righe o alle colonne. Poiché la nostra matrice è costituita da 158 osservazioni e 96 variabili e le osservazioni sono sulle righe, deve essere selezionato “Matrix rows”.

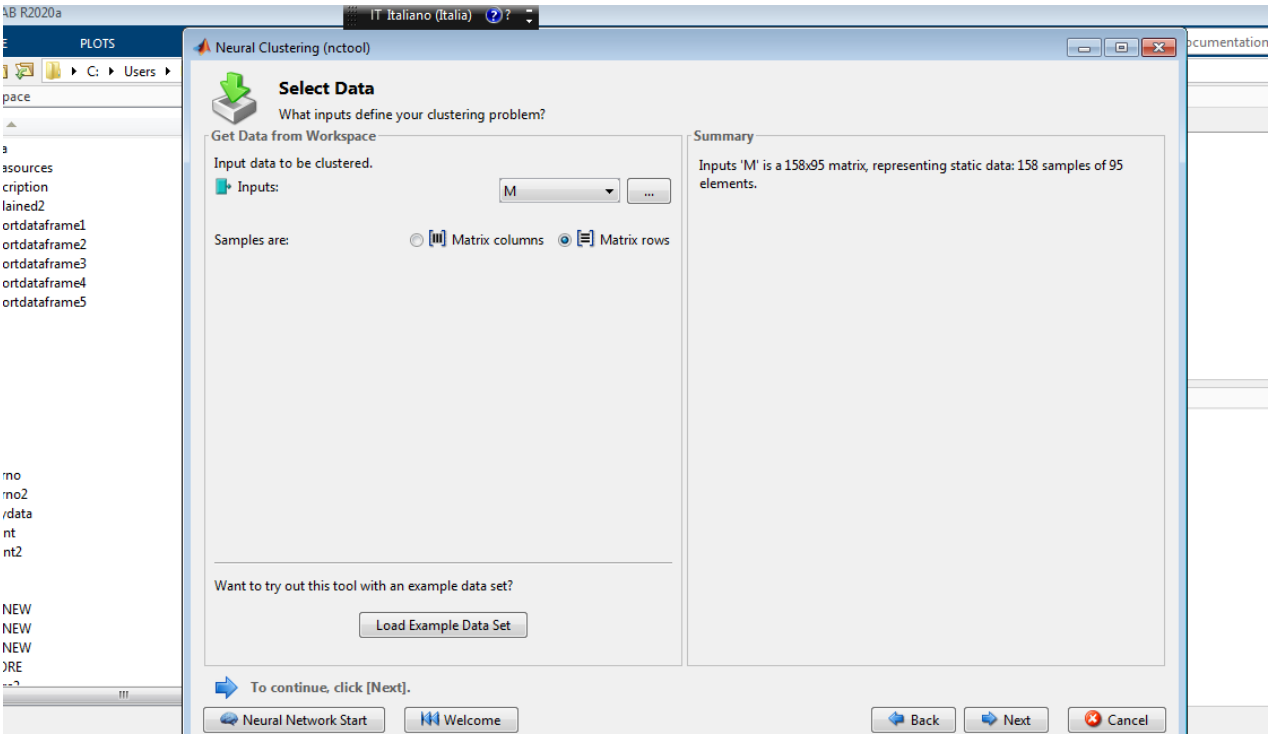


Figura 41-sezione di inserimento dei dati da analizzare e con i quali viene inizialmente istruita la rete

La sezione successiva (Figura 42) permette di definire la struttura della rete. La prima operazione è la definizione della dimensione della mappa bidimensionale che, in prima battuta, si è deciso avesse 25 neuroni con una disposizione 5x5.
Graficamente la sezione "Neural Network" mostra la struttura delle rete in cui sono presenti 96 variabili di input, un unico strato di tipo SOM costituito da 25 neuroni in una matrice 5x5, e 25 dati di output.

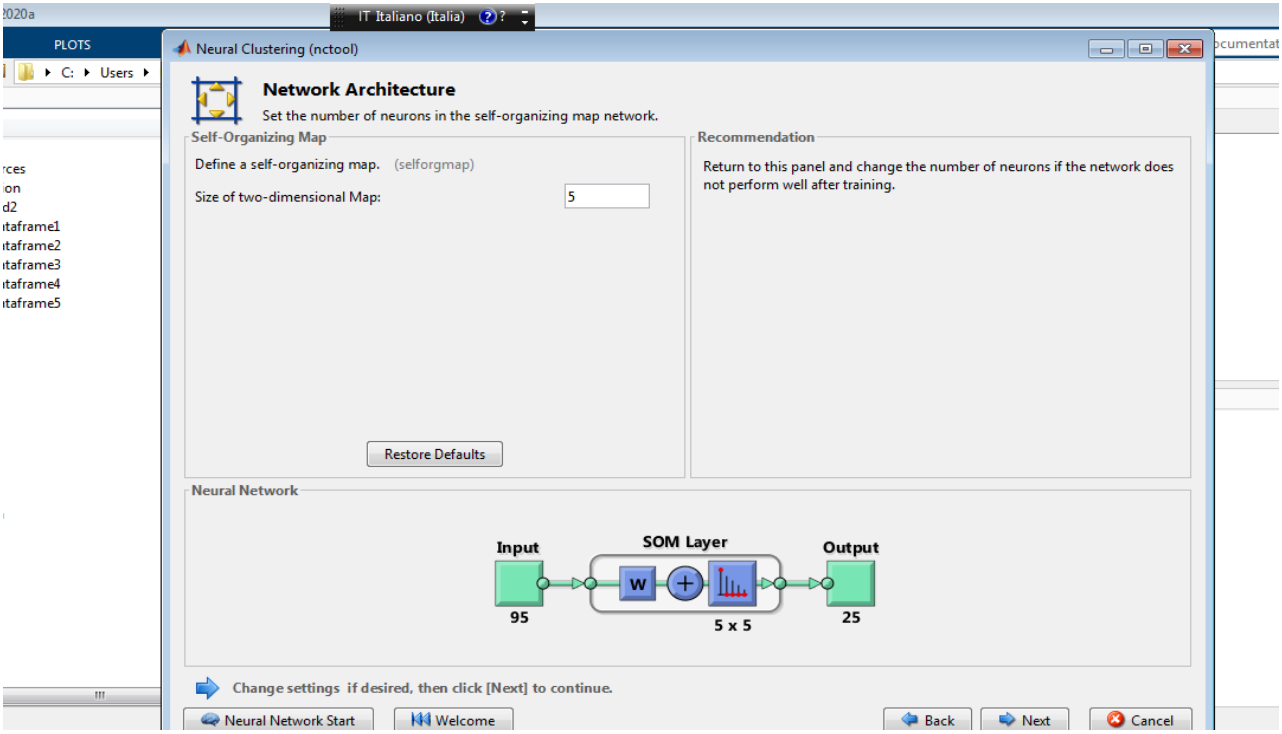


Figura 42-definizione della struttura della rete

Definita l'architettura della rete è necessario istruirla (Figura 43) per fare in modo che i pesi si auto configurino in base ai dati che sono stati passati come input. In questa fase è bene possedere un buon numero di dati.

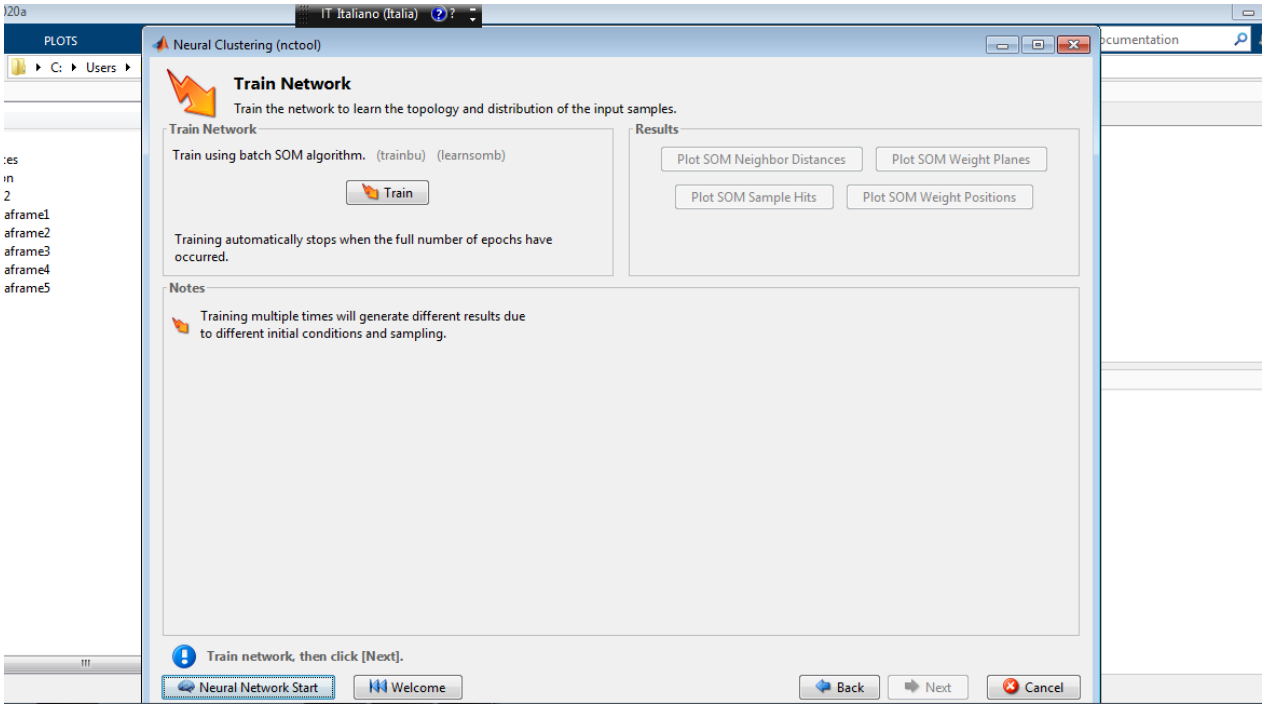


Figura 43- fase di training della rete neurale

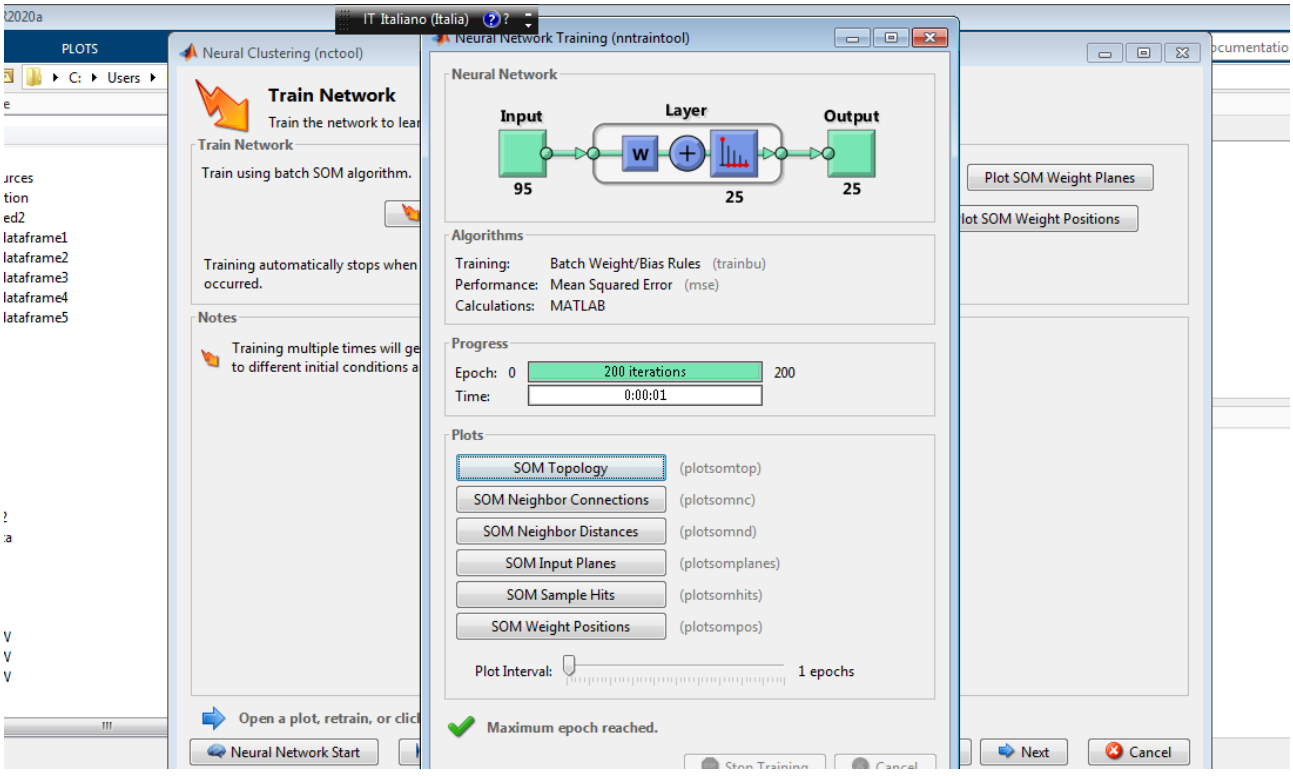


Figura 44-dettagli tecnici della rete neurale generata ed addestrata

Il pannello in Figura 44 mostra:

- La struttura delle rete da un punto di vista grafico
- La progressione del processo di istruzione per un numero di epoche definite (pari a 200)

- I grafici relativi alla SOM (distribuzione dei pesi, distanze, osservazioni classificate con i neuroni ecc.)

Descriviamo di seguito alcuni dei grafici utili all'analisi dei dati.
Il grafico "SOM Topology" (Figura 45) descrive la topologia della mappa, cioà la distribuzione dei neuroni nello strato interno della rete.

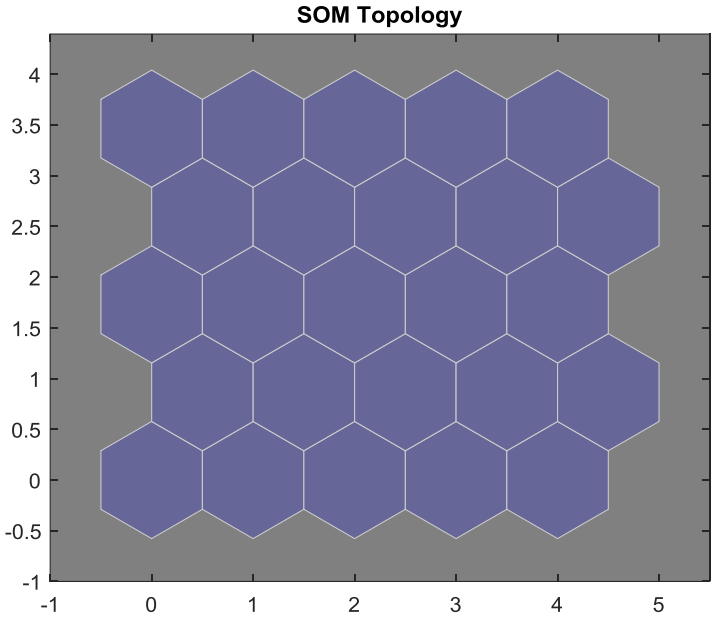


Figura 45- topologia dei neuroni per una rete mappa 5x5

Il grafico "SOM Neighbor connections" (Figura 46) mostra le connessioni tra i neuroni.

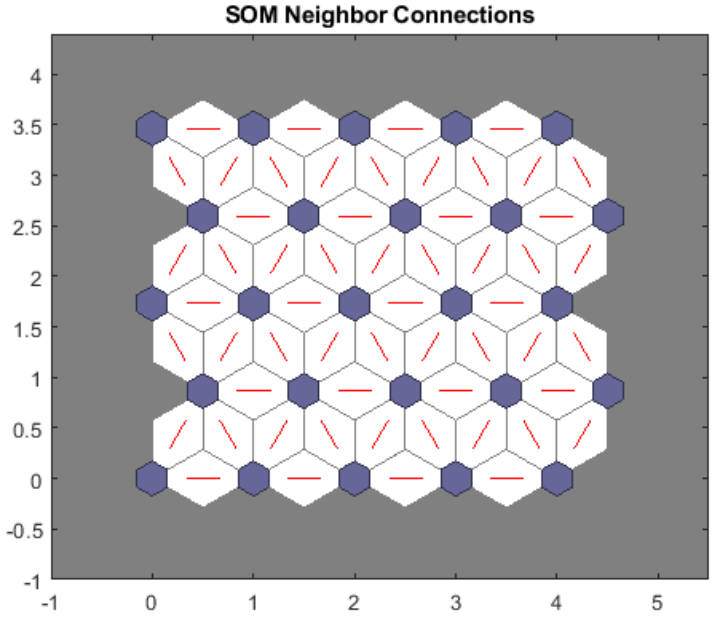


Figura 46 grafico SOM Neighbor connections che mostra le connessioni tra i neuroni

Si può notare come i neuroni siano disposti su una griglia esagonale anche se non è l'unica disposizione possibile.

I neuroni possono essere disposti su vari tipi di strutture ma quelle più comuni sono **l'esagonale** e la **rettangolare**; entrambe sono soggette ad un problema legato al fatto che i neuroni sui confini della rete hanno meno vicini, pertanto hanno una ridotta possibilità di aggiornamento e per cui i neuroni centrali sono "più istruiti" di quelli ai bordi. In questo la struttura esagonale risente di meno di questo effetto.

Il grafico "SOM Neighbor distances" permette di visualizzare graficamente la distanza dei vari neuroni, mostrando una prima clusterizzazione per aree.

Le connessioni di colore nero indicano neuroni distanti tra loro, mentre le connessioni di colore giallo indicano neuroni vicini.

In Figura 47 si può vedere una divisione in aree che rappresenta già una indicazione di "similarità" tra i dati.

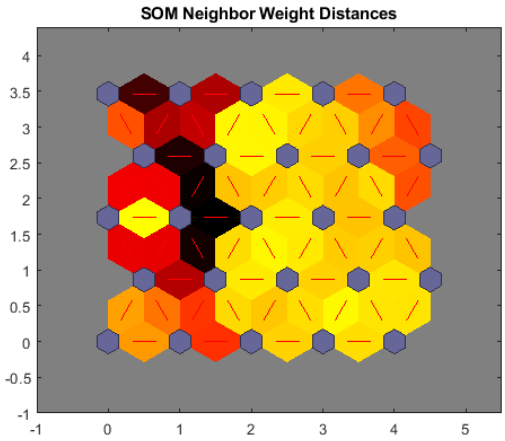


Figura 47-divisione in aree che rappresentano una indicazione di "similarità" tra i dati

Il grafico in Figura 48 prende il nome di "SOM sample hits" e mostra, per ogni neurone nel piano, il numero di osservazioni che lo caratterizza.

Il primo neurone è quello in basso a sinistra, caratterizzato da 30 osservazioni, mentre il neurone 25 è quello in alto a destra e caratterizzato da 1 sola osservazione. La numerazione dei neuroni va da sinistra verso destra.

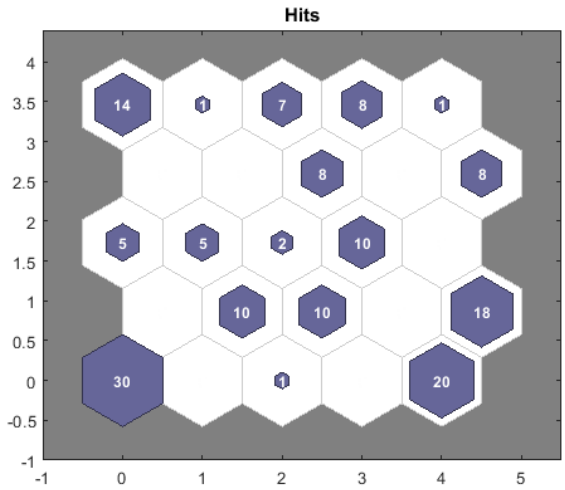


Figura 48-grafico SOM sample hits mostra per ogni neurone nel piano, il numero di osservazioni che lo caratterizza

Una volta effettuata l'analisi, qualora non fosse soddisfacente, è possibile istruire nuovamente la rete attraverso il tasto "retrain" (Figura 49).

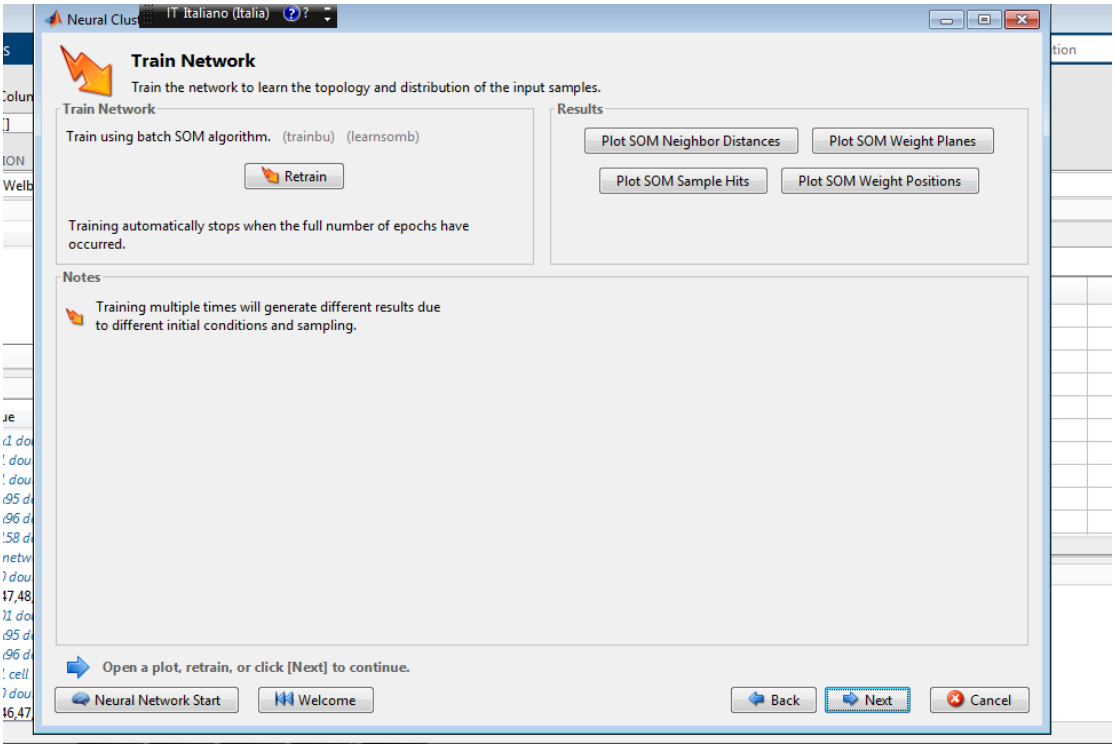


Figura 49- qualora il primo trainining non fosse soddisfacente, è possibile istruire nuovamente la rete

La sezione successiva (Figura 51) permette di valutare la rete neurale ed eventualmente aggiungere nuovi dati di input.

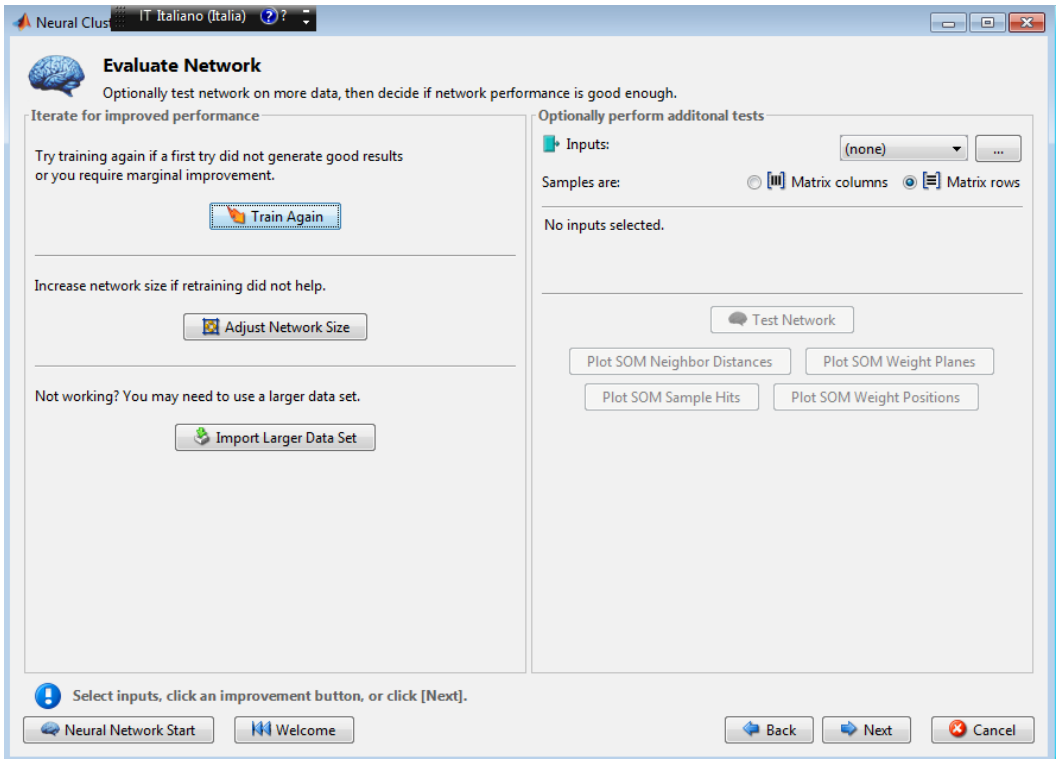


Figura 50-schermata di valutazioen della rete, con la possibilità di istruirla nuovamente aggiungendo ulteriori dati

Ogni training successivo porterà risultati differenti in quanto la mappa si riorganizzerà in modo randomico e differente dalla volta precedente.

L'aggiunta di ulteriori dati è molto importante per la specializzazione della rete neurale e per perfezionarne le caratteristiche, pertanto è possibile aggiungere dati per dare modo alla rete neurale di organizzarsi nuovamente specializzandosi sempre di più.

È inoltre possibile incrementare le dimensioni della mappa (tasto "Adjust network size") aumentando il numero di neuroni. Non esiste un numero "giusto" di neuroni in una rete, il numero dipende dalle dimensioni del set in ingresso e dall'analisi dei risultati che ogni mappa offre; si procede per tentativi per verificare quale sia la dimensione che possa garantire la migliore clusterizzazione.

La schermata successiva (Figura 52) offre la possibilità di:

- Generare una versione che è possibile utilizzare standalone su altri sistemi
- Generare il codice per permettere di studiare in modo puntuale le funzioni utilizzate dal tool, ed esportare le singole funzioni direttamente nella console di MATLAB, utilizzandole tramite riga di comando
- Generare lo schema che descrive la rete neurale

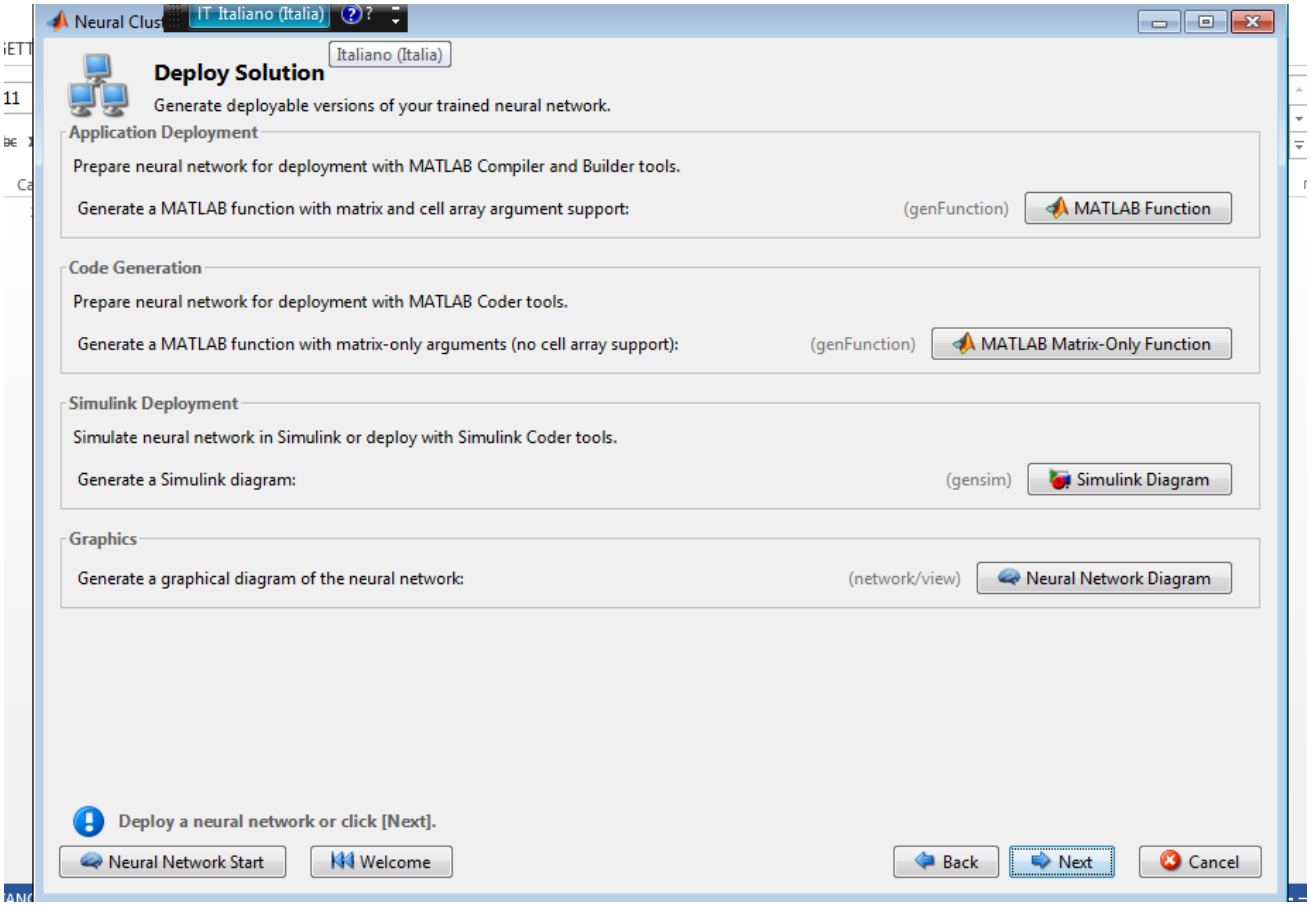


Figura 51-schermata di gestione della rete neurale in cui è possibile effettare il deploy della rete oppure esportare il codice direttamente nella console di MATLAB

Infine il tool offre la possibilità di salvare gli script e riutilizzare la rete appena generata e tutto il lavoro effettuato (Figura 52).

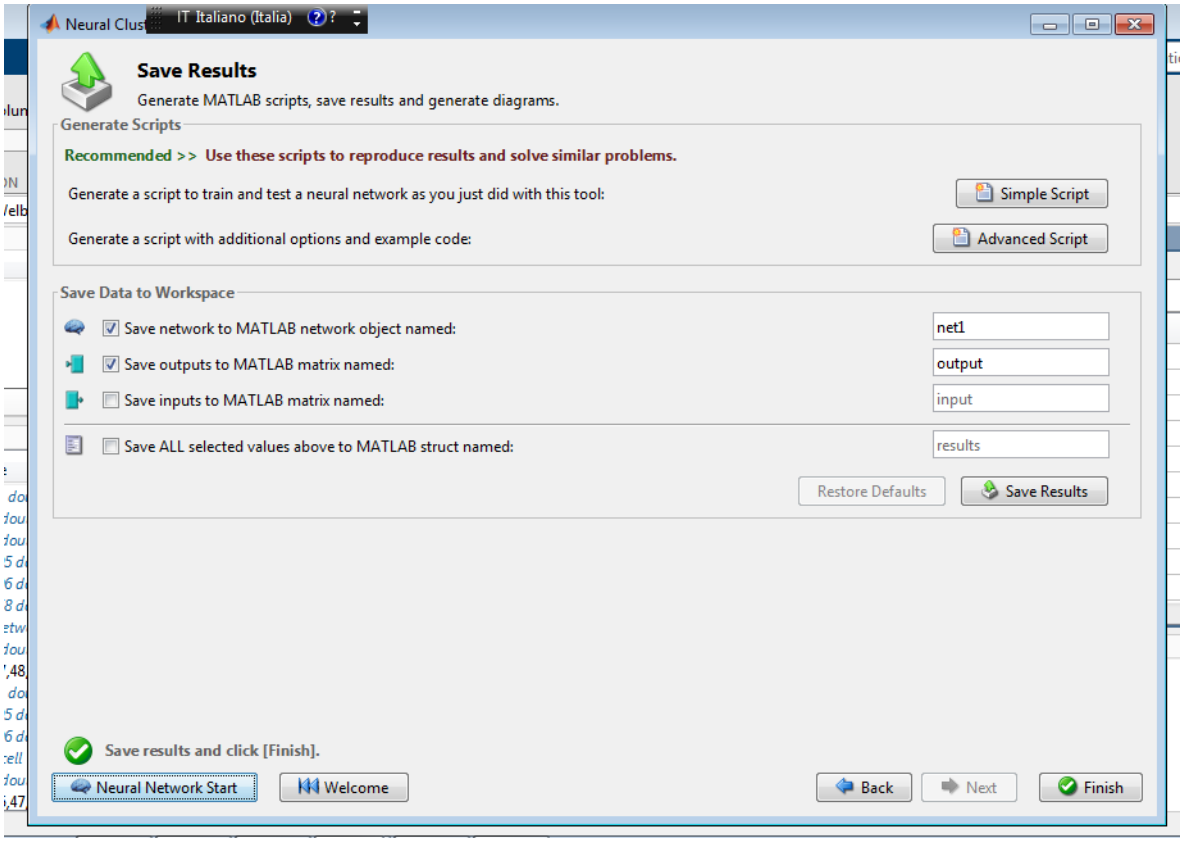


Figura 52-Schermata di salvataggio della rete neurale creata

2.3.10 Analisi dei dati tramite rete neurale SOM

Per l'analisi dei dati si è proceduto a valutare dapprima le dimensioni della mappa, e dunque il numero di neuroni, che meglio evidenzino eventuali situazioni particolari, e ad analizzare la clusterizzazione ottenuta per verificare la presenza di nuove aggregazioni interessanti ottenute.

Abbiamo considerato per prima cosa una rete neurale con una mappa di dimensioni 5x5, comprendente pertanto 25 neuroni

In ingresso sono considerate 96 variabili per 158 osservazioni, come già indicato nei paragrafi precedenti.

La rete è descritta graficamente in Figura 53.

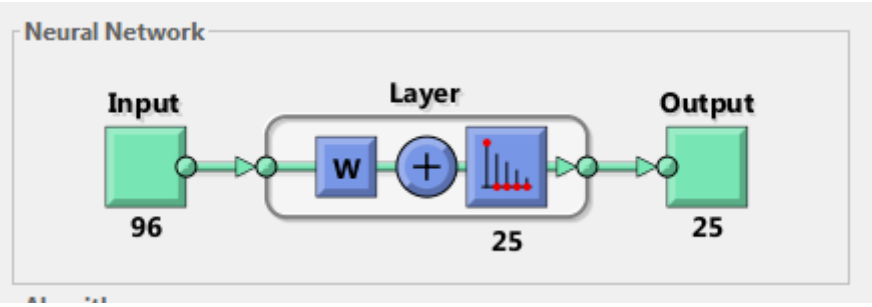


Figura 53 –descrizione sintetica della rete

Utilizzando il set di dati di input, il grafico relativo alla distribuzione delle osservazioni per i 25 neuroni nella rete, è descritto in Figura 54.

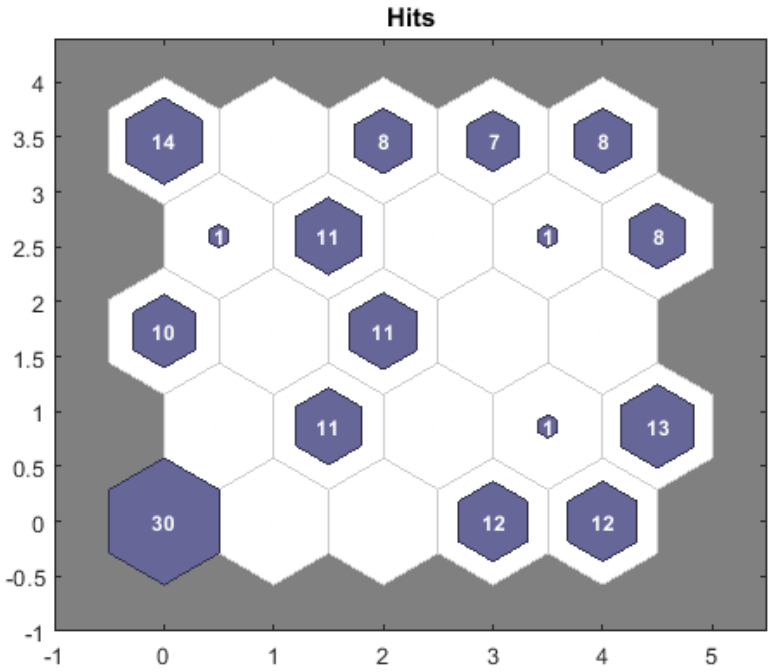


Figura 54-grafico relativo alla distribuzione delle osservazioni per i 25 neuroni nella rete

Essa mostra i 25 neuroni e il numero indicato in ogni neurone indica il numero di osservazioni presenti in ciascuno di essi.
È interessante notare anche le relazioni tra i centri dei cluster attraverso il grafico della matrice distanza/peso, in cui i colori rappresentano la distanza tra i pesi dei neuroni vicini. Il colore nero indica una distanza elevata, il giallo una distanza minore (Figura 55).

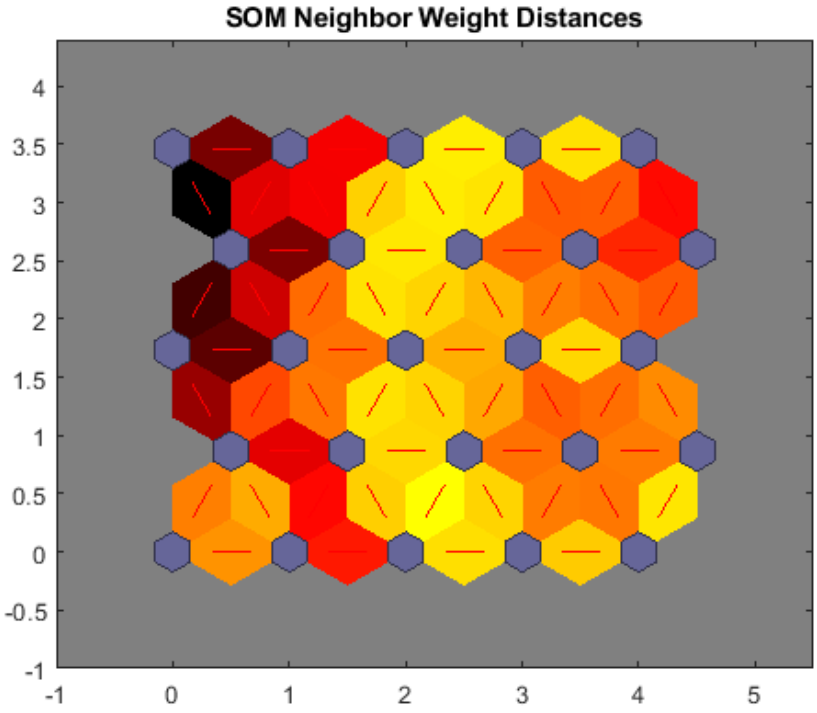


Figura 55-grafico della matrice distanza/peso, in cui i colori rappresentano la distanza tra i pesi dei neuroni vicini

Il primo neurone associa 30 osservazioni. Per analizzare i dati, e capire le 30 osservazioni a quale periodo dei dati si riferiscono, è necessario utilizzare alcune specifiche funzioni. La funzione che serve a tale scopo **vec2ind** converte gli indici in vettori. Nel nostro caso dato ogni input (matrice M) si può recuperare l'informazione che esprime da quale neurone è classificato l'input stesso.

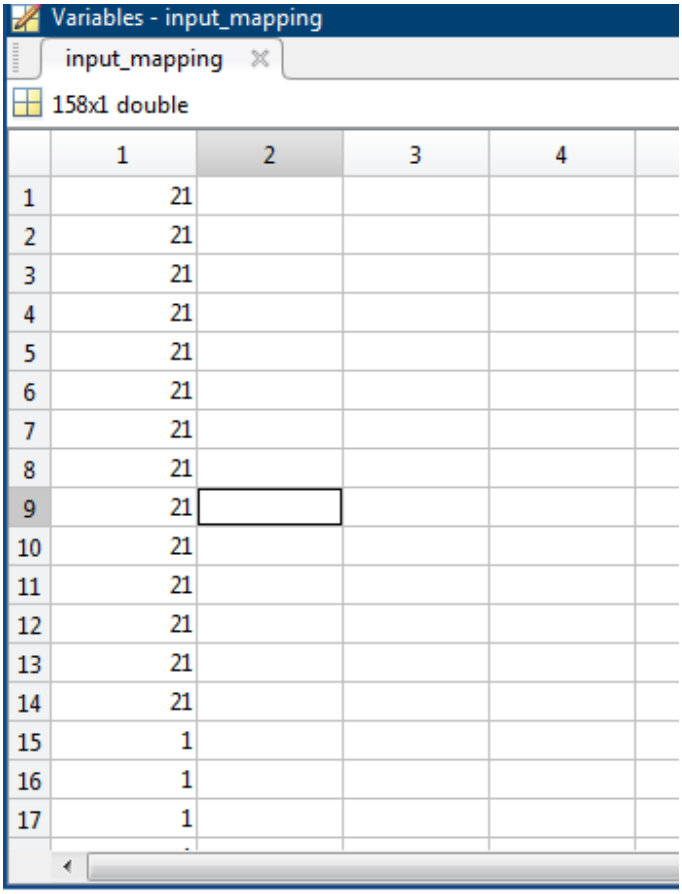
La funzione è la seguente:

in_map=vec2ind(net(M))'

dove **net** è la variabile che rappresenta la rete neurale, M è la matrice di input, e **in_map** è un vettore tale che il valore di ogni input corrisponde al neurone in base al quale l'input è stato classificato.

La Figura 56 mostra le prime 17 righe del vettore **in_map**, che corrispondono alle osservazioni con indici da 1 a 17.

Le osservazioni con indici da 1 a 14 sono classificate dal neurone 21, mentre le osservazioni con indici da 15 a 17 sono classificate dal neurone 1. Il vettore contiene la classificazione per ciascuna delle 158 osservazioni caricate in input.



	1	2	3	4
1	21			
2	21			
3	21			
4	21			
5	21			
6	21			
7	21			
8	21			
9	21			
10	21			
11	21			
12	21			
13	21			
14	21			
15	1			
16	1			
17	1			

Figura 56-vettore contenente la classificazione per ciascuna delle 158 osservazioni caricate in input.

L'ordine dei numeri comincia in basso a sinistra e termina in alto a destra procedendo per righe. Il neurone 1 è quello posizionato in basso a sinistra, mentre il neurone 21 è in alto a sinistra; in Figura 57 è possibile vedere la disposizione dei neuroni.

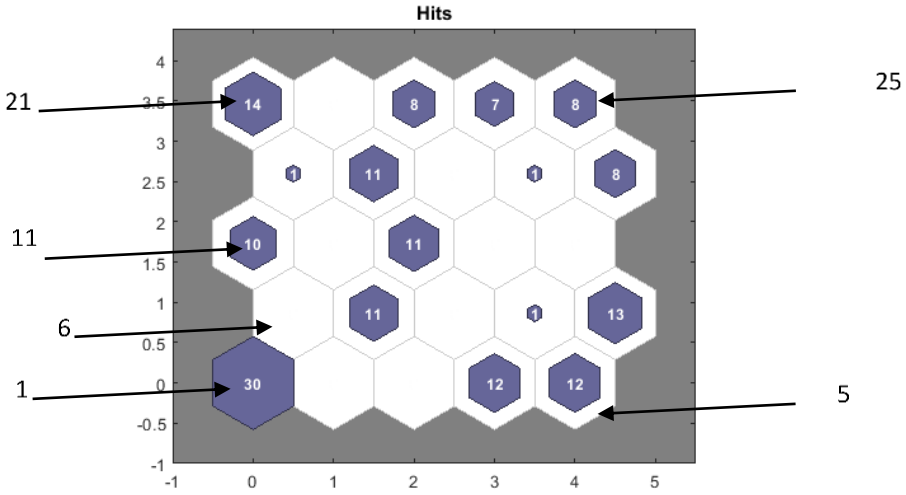


Figura 57-Il neurone 1 è quello posizionato in basso a sinistra, mentre il neurone 21 è in alto a sinistra

A questo punto devono essere recuperati gli input che sono classificati da ciascun neurone, e questo può essere ricavato recuperando tutti gli indici di **in_map** associati a ciascun neurone.
A tal fine si utilizza la funzione di MATLAB **find** nel seguente modo:

```
neuron21_index = find(in_map==21)
```

dove neuron21_index è il vettore che contiene gli indici, e la funzione cerca tutti gli indici associati al 21esimo elemento del vettore dei neuroni **in_map,** ed è mostrato in Figura 58

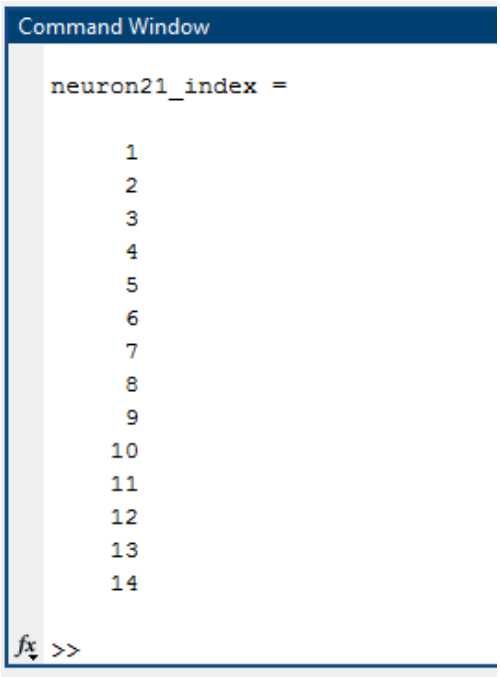


Figura 58-vettore che contiene gli indici associati al 21esimo elemento del vettore dei neuroni

Come si può vedere vengono restituiti i primi 14 indici. A questo punto dobbiamo capire a quali giorni si riferiscono tali indici, e per fare questo basta passare al vettore che contiene tutti i giorni relativi alle osservazioni tutto il vettore degli indici:

```
neuron21_day=giorno2(neuron21_index)
```

neuron21_day è un vettore che contiene tutti i giorni classificati dal neurone 21 (Figura 59).

```
Command Window

>> neuron21_value=giorno2(neuron21_index)

neuron21_value =

  1×14 cell array

  Columns 1 through 6

    {'2019-06-26'}    {'2019-06-27'}    {'2019-06-28'}    {'2019-06-29'}    {'2019-06-30'}    {'2019-07-01'}

  Columns 7 through 12

    {'2019-07-02'}    {'2019-07-03'}    {'2019-07-04'}    {'2019-07-05'}    {'2019-07-06'}    {'2019-07-07'}

  Columns 13 through 14

    {'2019-07-08'}    {'2019-07-09'}

fx >>
```

Figura 59-Il vettore contiene tutti i giorni classificati dal neurone 21.

Se vediamo il grafico dei pesi (Figura 60), si può vedere come i pesi relativi a tali dati, che ricordiamo essere relativi all'anno 2019, siano relativamente lontani dai pesi dei neuroni vicini ad indicare una "diversità" dei dati.

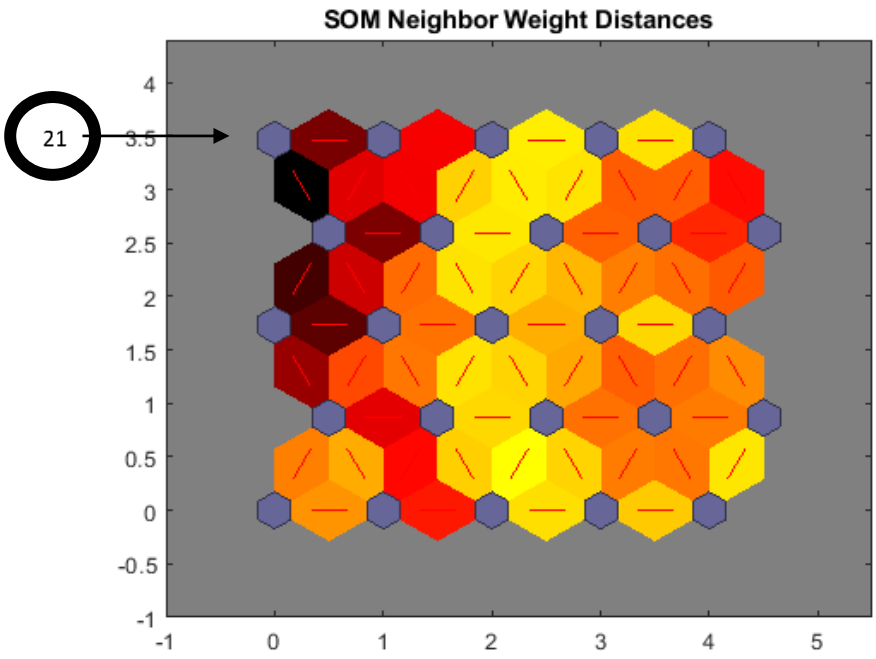


Figura 60-grafico dei pesi in cui si evidenzia, per il neurone 21, la distanza dagli altri neuroni, cioè una differenza sostanziale nei dati

L'area centrale è caratterizzata da neuroni i cui pesi sono molto vicini tra loro (in varie tonalità di giallo), che corrispondono a cluster di dati "simili".

In Figura 61 è mostrato il dettaglio di alcuni cluster evidenziati dalla rete neurale.

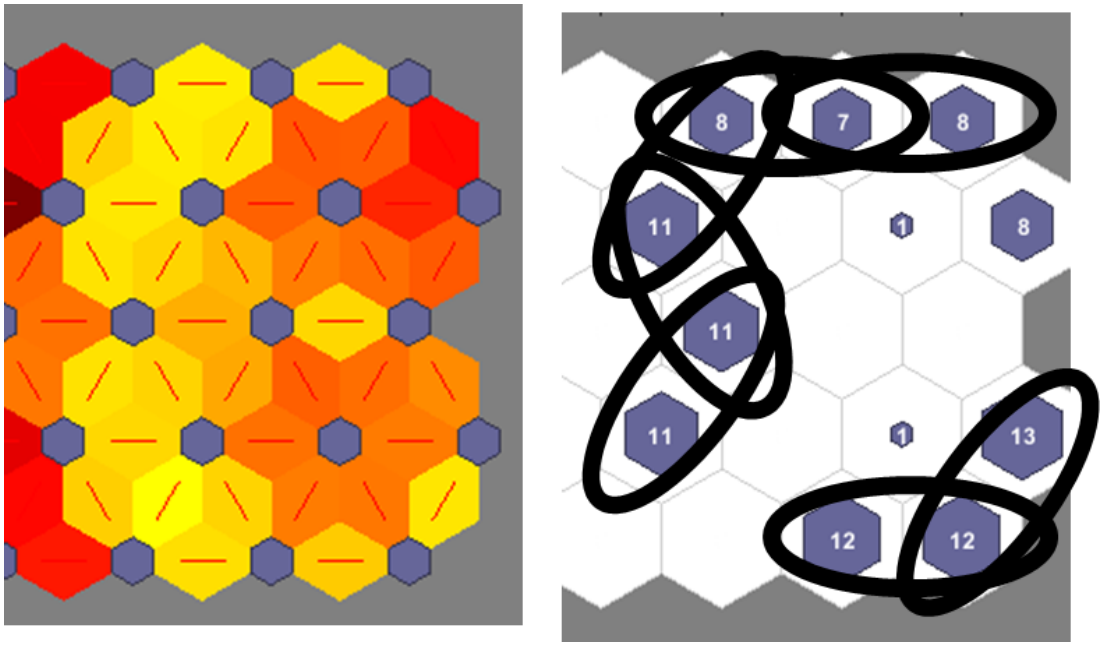


Figura 61-dettaglio di alcuni cluster evidenziati dalla rete neurale e relative osservazioni

Il cluster più grande comprende 6 neuroni: il 7, 13, 17, 23, 24, 25.

Attraverso le funzioni di recupero degli indici, riutilizzata per tutti i neuroni, si ottiene un cluster di 56 elementi che corrisponde al quinto cluster già evidenziato tramite PCA (Figura 29)

Il secondo cluster è costituito da 3 neuroni con indici 4, 5 e 10.

Effettuando il recupero degli indici delle osservazioni, si ottiene un cluster di 37 elementi, che corrisponde al quarto cluster già evidenziato tramite PCA.

Gli altri neuroni che classificano osservazioni, presentano pesi con distanze più elevate, e ciò indica che le rappresentazioni dei dati sono tanto più differenti tra ogni cluster vicino, quanto più scuro è il colore che identifica la distanza tra i loro pesi.

Abbiamo già visto che il cluster associato al neurone 21 rappresenta i 14 dati relativi al 2019

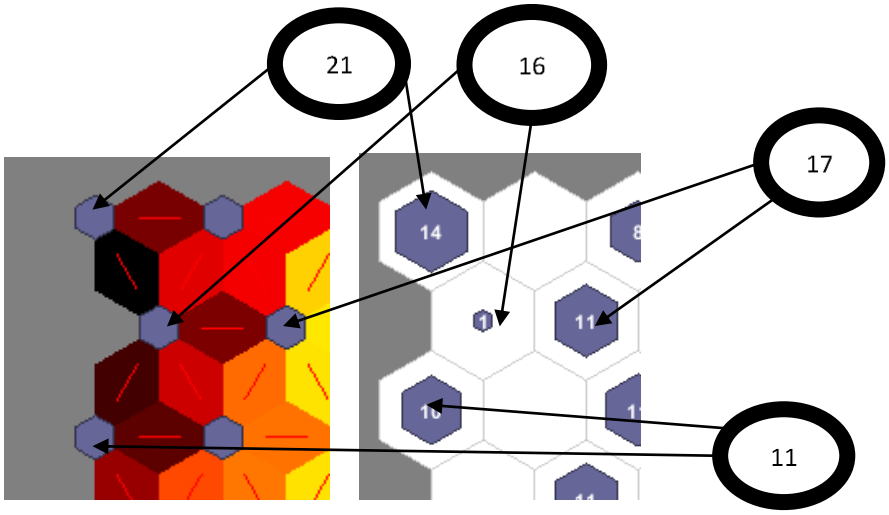


Figura 62-rappresentazione dei cluster associati ai neuroni 21, 16,17,11

Il neurone 21 è vicino al neurone 16, tuttavia i loro pesi sono molto distanti tra loro, infatti il neurone 16 contiene una sola osservazione relativa al 19/08/2020 (Figura 62).

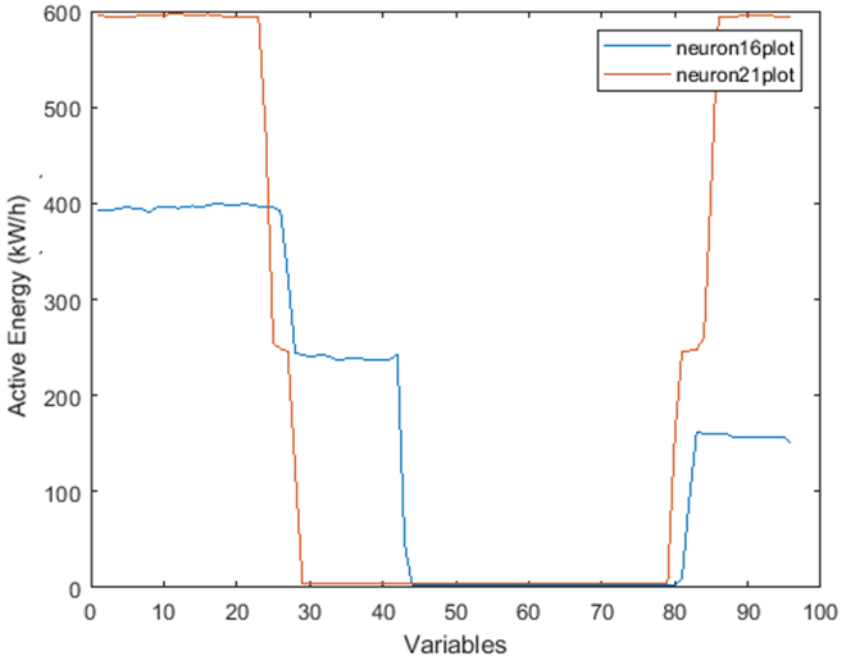


Figura 63-media degli andamenti delle osservazioni per i neuroni 21 e 16 in cui si vede che gli andamenti dell'energia attiva sono molto differenti come atteso.

Se osserviamo la Figura 63, che rappresenta la media degli andamenti delle osservazioni per ciascun neurone considerato (21 e 16) ci rendiamo conto che, effettivamente, gli andamenti dell'energia attiva sono molto differenti come atteso.

Se osserviamo anche l'andamento dei dati classificati dal neurone 11, che è vicino al neurone 16 ma con distanza abbastanza elevata da esso, è chiaro che si tratti di andamenti totalmente differenti tra loro, pertanto la classificazione sembra avere un senso.

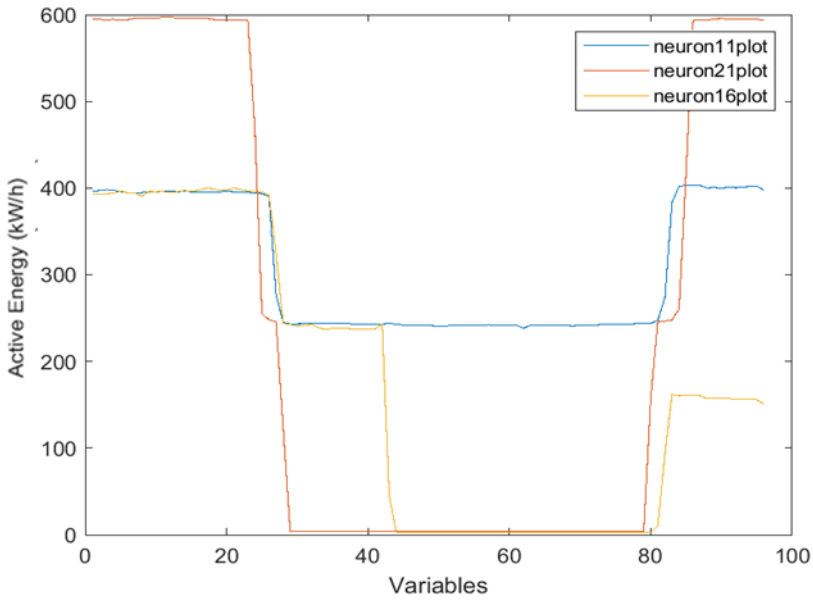


Figura 64-il neurone 11 classifica dati che sono visibilmente lontani rispetto ai dati classificati dal neurone 16

Il cluster relativo al neurone 11, comprende 10 osservazioni per il periodo che va dal 09/08 al 18/08, che corrisponde al secondo cluster evidenziato dalla PCA.
È da notare come il neurone 16, che ha classificato una sola osservazione, rappresenti uno dei dati sparsi analizzati tramite la PCA e non inclusi in alcun cluster. La rete neurale, in modo più semplice rispetto alla PCA, ci dà anche l'indicazione di "quanto" tale osservazione sia distante dalle altre osservazioni (Figura 64).
Il neurone 17, è vicino al neurone 16, ed appartiene al primo cluster che è stato analizzato e che comprende i neuroni 7,13,17,23,24,25.
La distanza tra i due neuroni è descritta da un colore rosso che indica una distanza tra i pesi dei neuroni 16 e 17, inferiore alla distanza tra i pesi dei neuroni 16 e 11. Questo vuol dire che l'osservazione relativa al neurone 16 è più simile alle osservazioni del cluster classificato dal neurone 17 che alle osservazioni classificate dal neurone 11.
Questo è confermato dalla Figura 65 che mostra la media dell'andamento delle osservazioni per il cluster 1, per il neurone 11, per il neurone 16 e per il neurone 17, in cui si vede che la media dei dati relativi alle osservazioni classificate dal neurone 16 hanno una somiglianza "migliore" con quella relativo al cluster 1 piuttosto che con quella relativa al cluster 11.

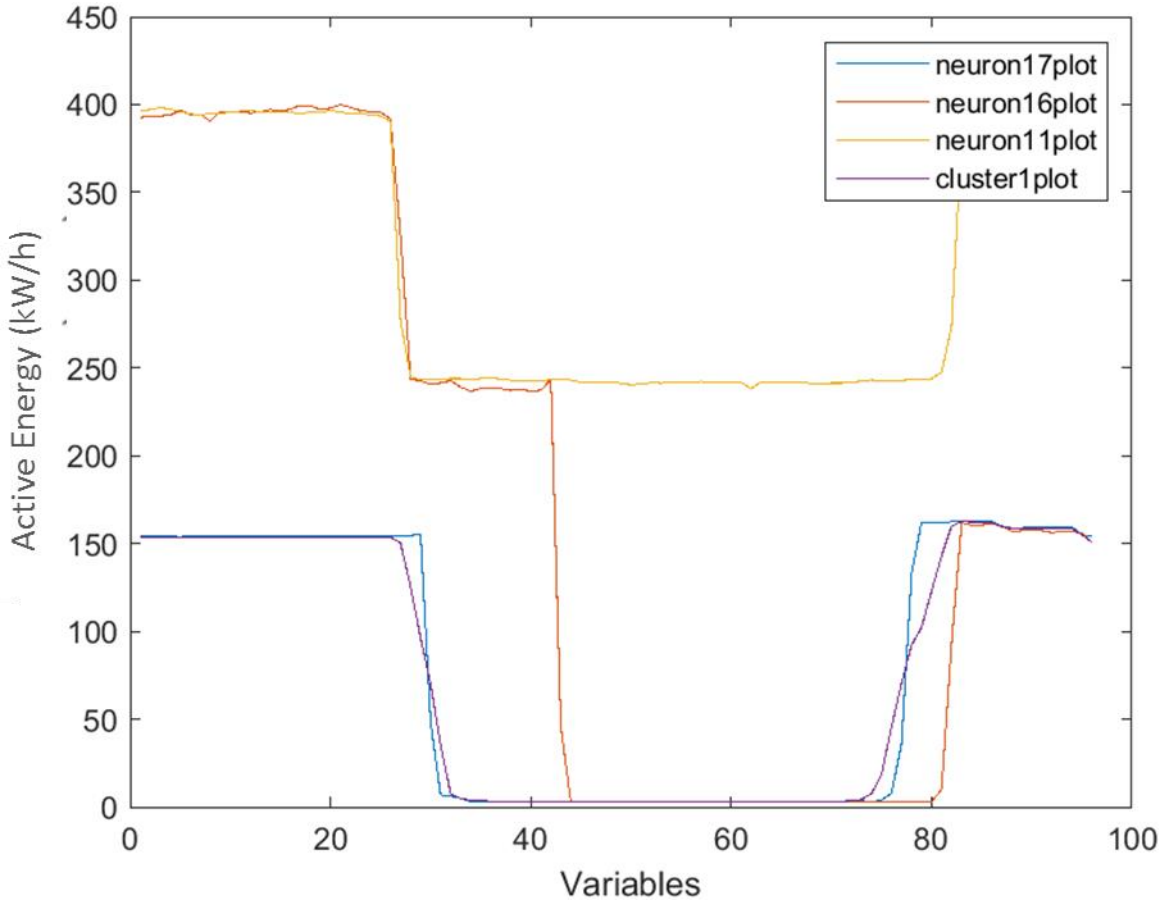


Figura 65-media dell'andamento delle osservazioni per il cluster 1, per il neurone 11, per il neurone 16 e per il neurone 17; la media dei dati relativi alle osservazioni classificate dal neurone 16 ha una somiglianza "migliore" con quella relativa al cluster 1 piuttosto che con quella relativa al cluster 11

Il neurone 1 classifica 30 osservazioni dal 10/07 al 08/08 che corrisponde al terzo cluster identificato dalla PCA.

La Figura 66 mostra gli ultimi cluster da analizzare.

I restanti 3 neuroni, il 19, il 20 e il 9, rappresentano dei dati anomali, e rispecchiano gli ultimi 2 clusters esaminati durante l'analisi della PCA.

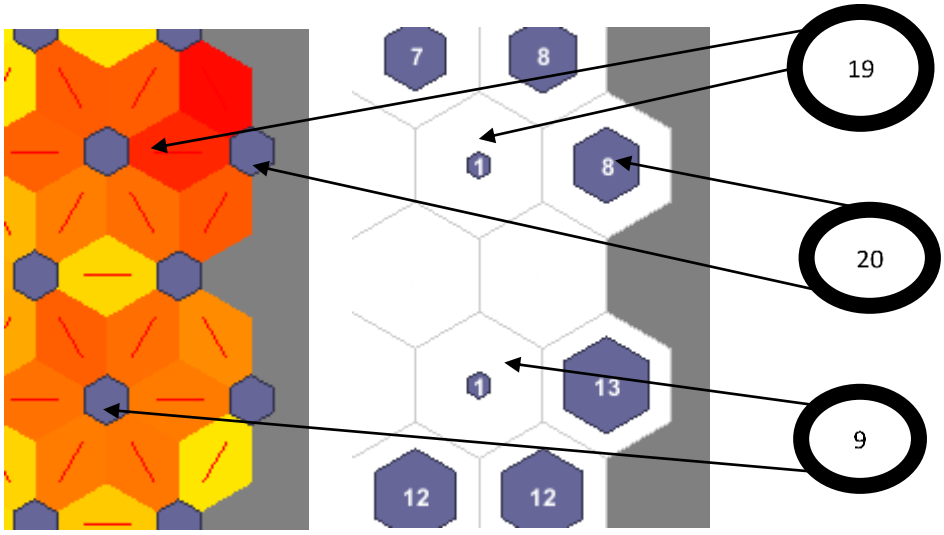


Figura 66-clusters relativi a dati probabilmente anomali

Ancora una volta, tramite la rete Neurale, è possibile valutare quantitativamente quanto simili siano i dati classificati da diversi neuroni.
Il Neurone 9 classifica una osservazione che è molto simile alla media delle osservazioni classificate dal cluster 2, costituito dai neuroni 4, 5 e 10 (Figura 67).

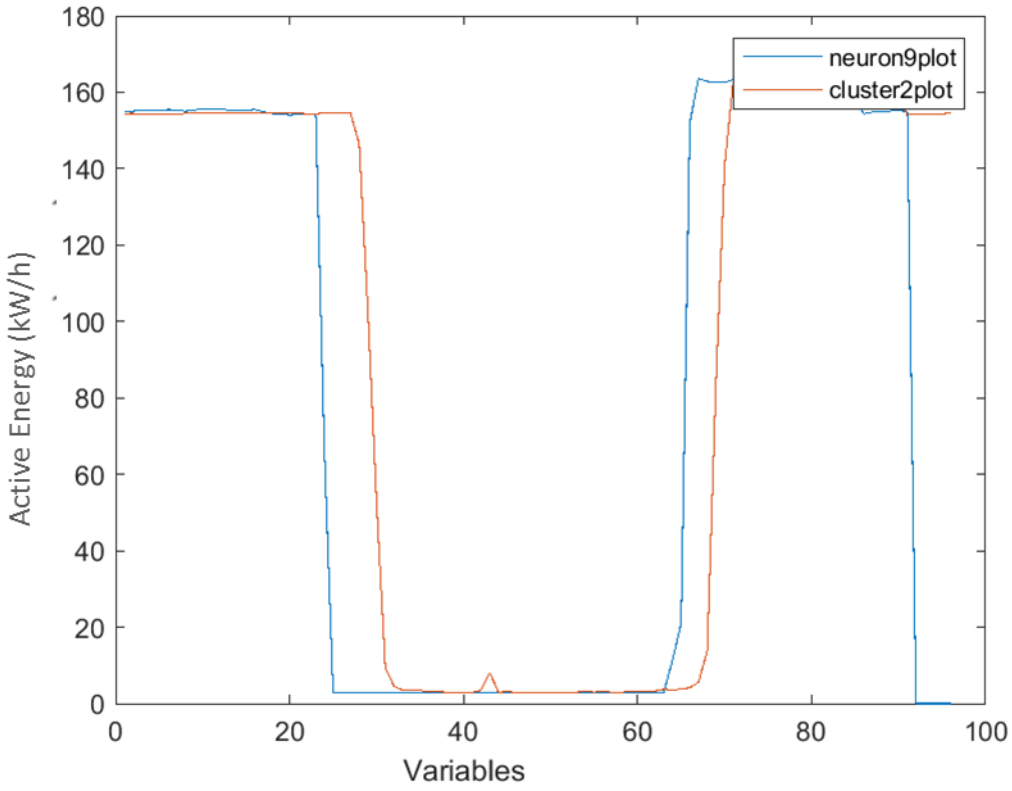


Figura 67 –i dati del Neurone 9 classifica sono molto simili alla media delle osservazioni classificate dal cluster 2, costituito dai neuroni 4,5 e 10

Infine la Figura 68 mostra la media dei dati relativi alle osservazioni classificate dai neuroni 19 e 20 e dal cluster 1.

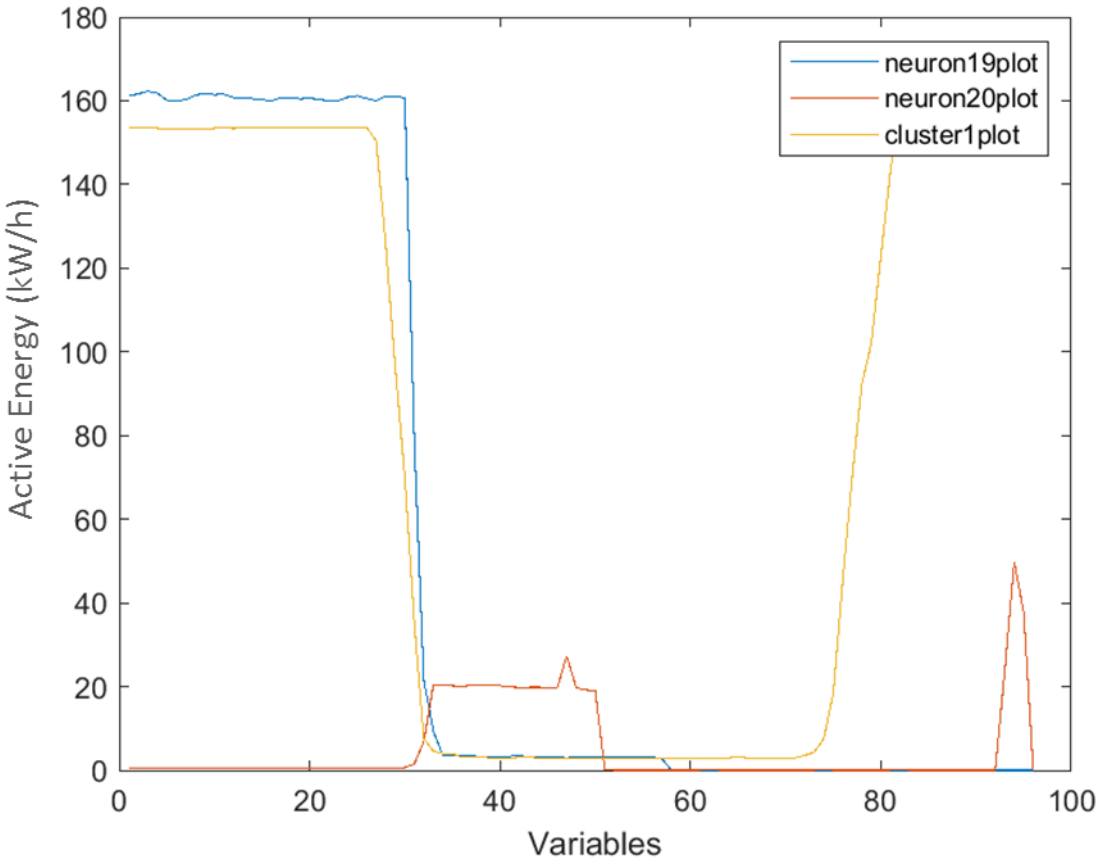


Figura 68- media dei dati relativi alle osservazioni classificate dai neuroni 19 e 20 e dal cluster 1.

Anche in quest'ultimo caso il grafico rispecchia la distanza dei pesi dei 3 clusters: i dati classificati dal cluster 19 sono "simili" a quelli relativi al cluster 1, mentre quelli relativi al cluster 20 sono differenti e rappresentano i veri dati "anomali" rispetto agli altri.

2.3.11 Analisi

Si è cercato di valutare la clusterizzazione tramite rete neurale anche incrementando il numero di neuroni. Infatti, maggiore è il numero di neuroni migliore dovrebbe essere la specializzazione di ciascun neurone.
È stata quindi valutata una rete neurale la cui mappa ha dimensioni 7X7 (Figura 69)

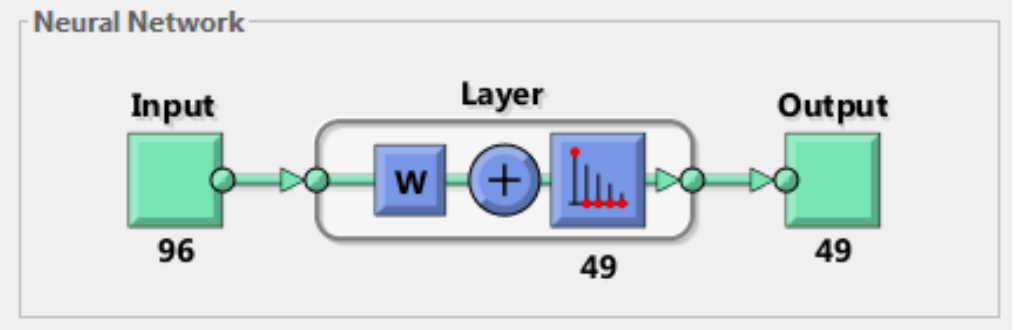


Figura 69-schema sintetico di una SOM 7x7

La topologia è descritta in Figura 70.

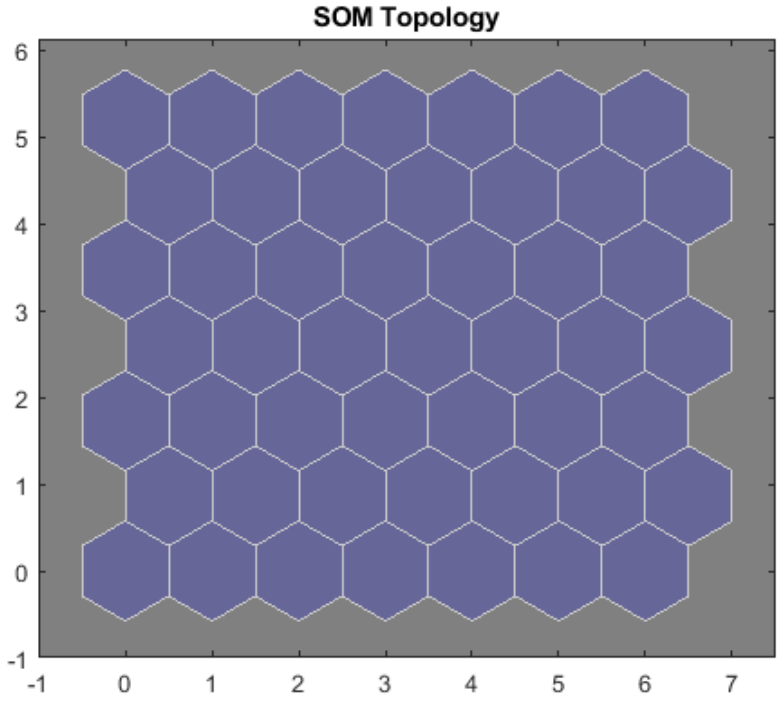


Figura 70-topologia di una rete 7x7

Le Figura 71 e Figura 72 mostrano invece rispettivamente la classificazione delle osservazioni per ogni neurone della mappa e la distanza tra i pesi associati ad ogni singolo neurone.

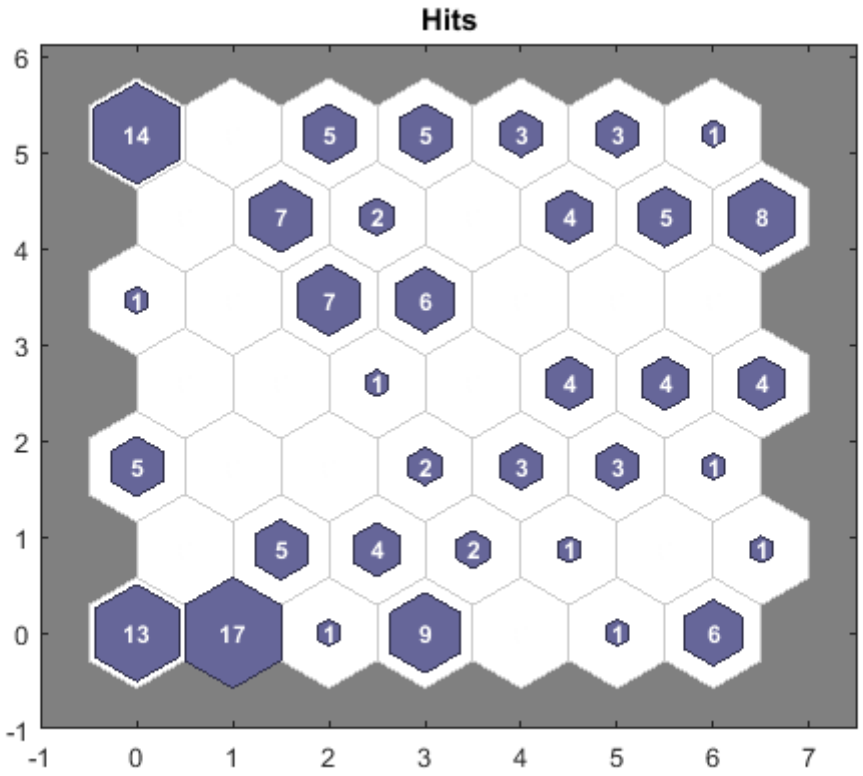


Figura 71-classificazione delle osservazioni per ogni neurone della mappa

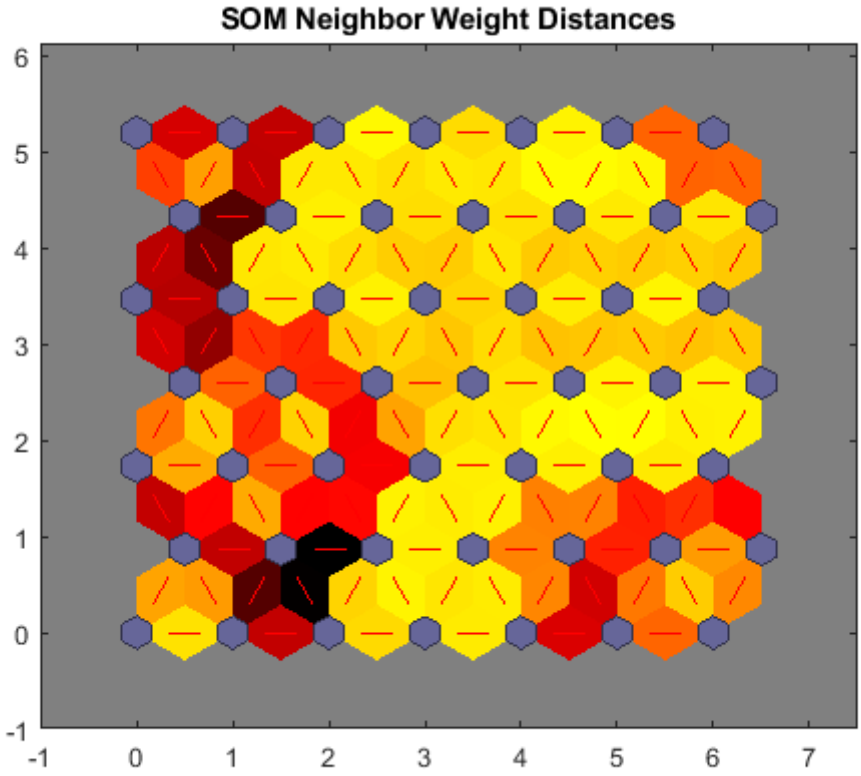


Figura 72-distanza tra i pesi associati ad ogni singolo neurone.

L'analisi condotta è similare a quella effettuata per la mappa 5X5, cioè si sono analizzati i vari clusters costituiti a volte da più neuroni, in base alle distanze tra i pesi di ciascun neurone o gruppo di neuroni.

Come nel caso precedente, l'area centrale è costituita da un'ampia area di osservazioni simili; è possibile distinguere all'interno di essa due aree separate (Figura 73):

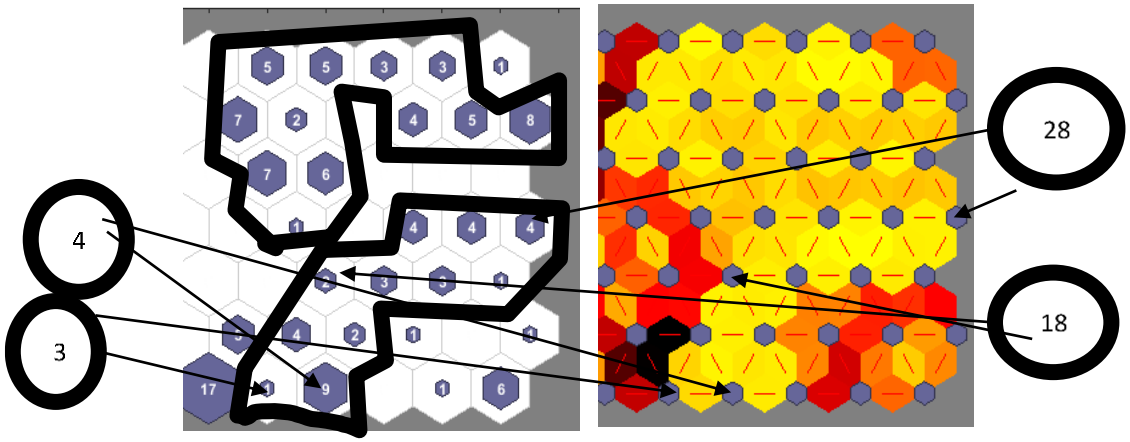


Figura 73-il grafico mostra l'area centrale della rete, in cui sono presenti osservazioni simili; è possibile distinguere all'interno di essa due aree separate

L'area inferiore comprende i neuroni 3, 4, 10, 11, 18, 19, 20, 21, 26, 27, 28, per un totale di 37 osservazioni ed periodo che va dal 26/10 al 01/12, che corrisponde ad un cluster identificato dalla PCA.

L'area superiore comprende i neuroni 24, 31, 32, 37, 38, 40, 41, 42, 45, 46, 47, 48, per un totale di 56 osservazioni ed un periodo che va dal 20/08 al 24/10 e che corrisponde ad un cluster identificato dalla PCA.

Il grafico in Figura 74 mostra la parte laterale della rete in cui sono presenti alcuni cluster lontani tra loro

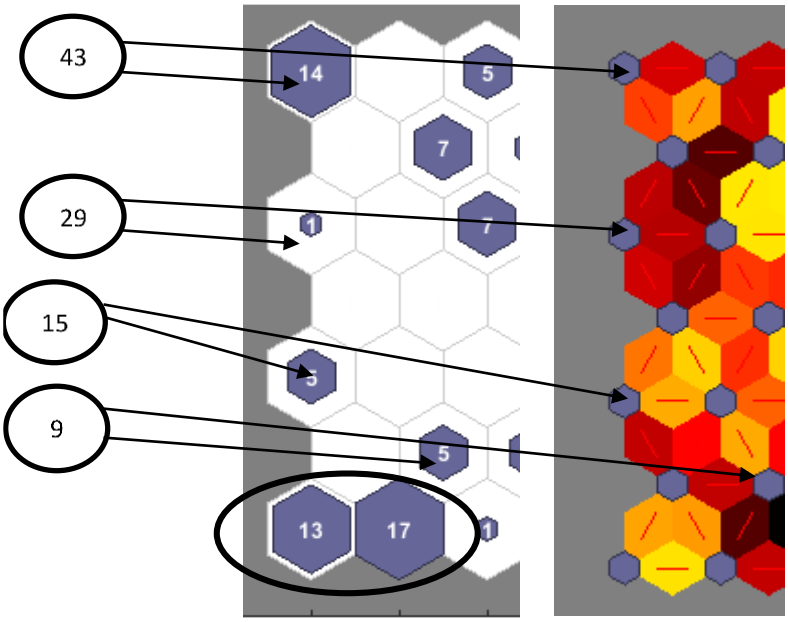


Figura 74-il grafico mostra la parte laterale della rete in cui sono presenti alcuni cluster lontani tra loro

Il neurone 43 comprende i dati relativi al 2019.
Il grafico in Figura 75 rappresenta l'andamento della media delle osservazioni per i neuroni 9, 29 e 15

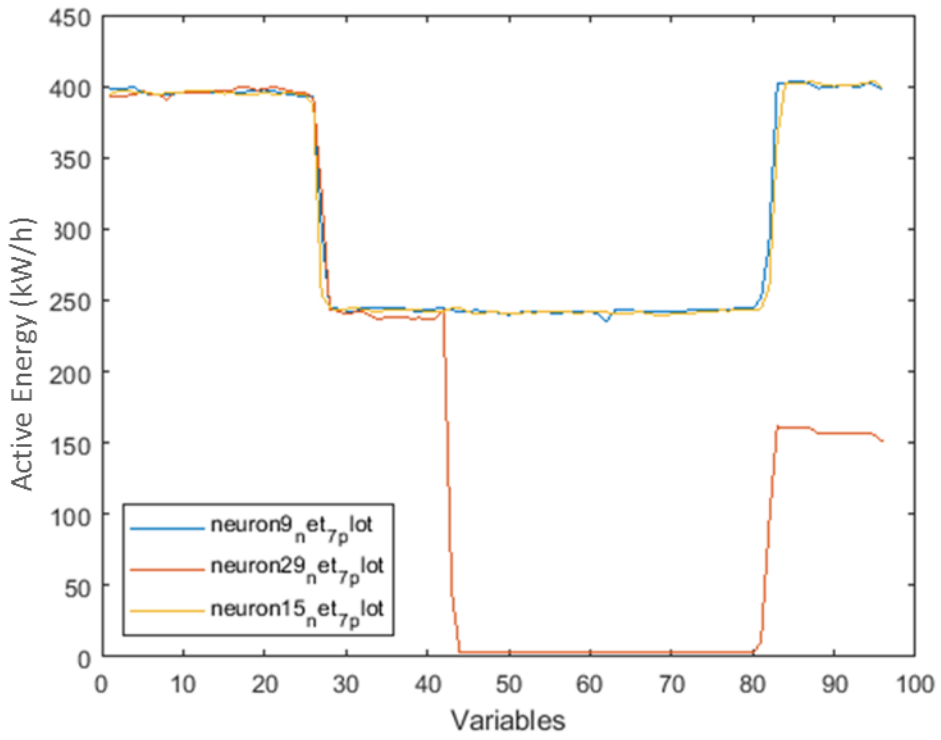


Figura 75-l'andamento della media delle osservazioni per i neuroni 9, 29 e 15

L' osservazione associata al neurone 29 (19/08) ha un andamento profondamente differente dalle altre, come atteso dai grafici e come previsto dalla PCA (per cui risultava tra i dati sparsi).

Le osservazioni relative ai neuroni 15 e 9, sono invece relative al periodo che va dal 09/08 al 18/08, evidenziando una clusterizzazione non prevista in caso di matrice 5X5 e neanche in caso di PCA. I dati relativi ai due neuroni sono in cluster separati tra loro, evidenziando una probabile migliore specializzazione della rete neurale nella classificazione di quel tipo di dato.
I neuroni 1 e 2 sono vicini, e rappresentano un cluster a parte che comprende 30 osservazioni per un periodo che va dal 10/07 all'08/08. Anche in tal caso si tratta di una clusterizzazione già prevista.

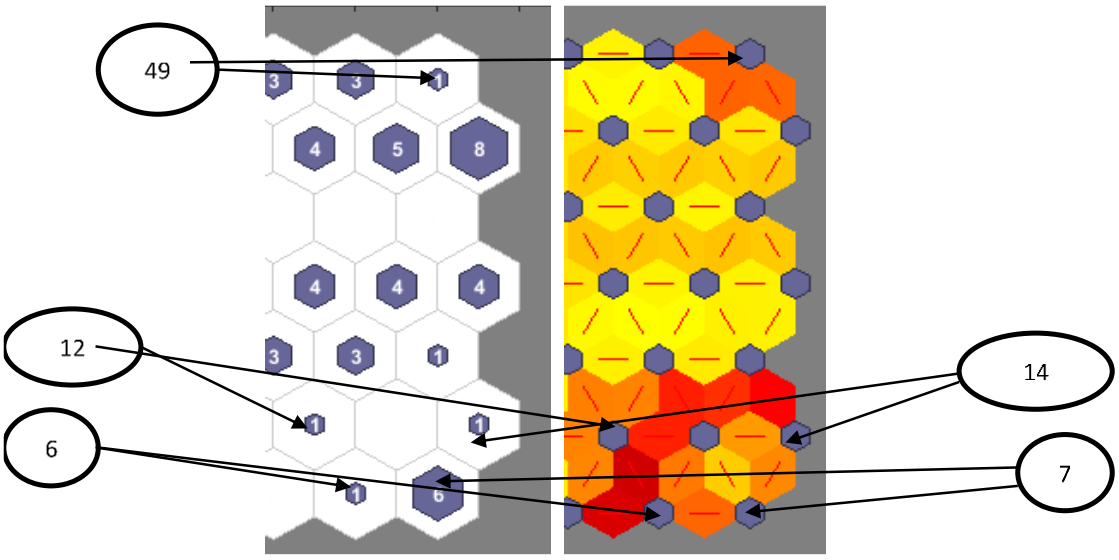


Figura 76-neuroni che risultano come cluster separati, presentando pesi distanti tra neuroni vicini.

La Figura 76 mostra i neuroni che risultano come cluster separati, presentando pesi distanti tra neuroni vicini. Il neurone 49 classifica il giorno 02/12/2021 (Figura 77), mentre il neurone 12 classifica il giorno il 25/10 (Figura 78)

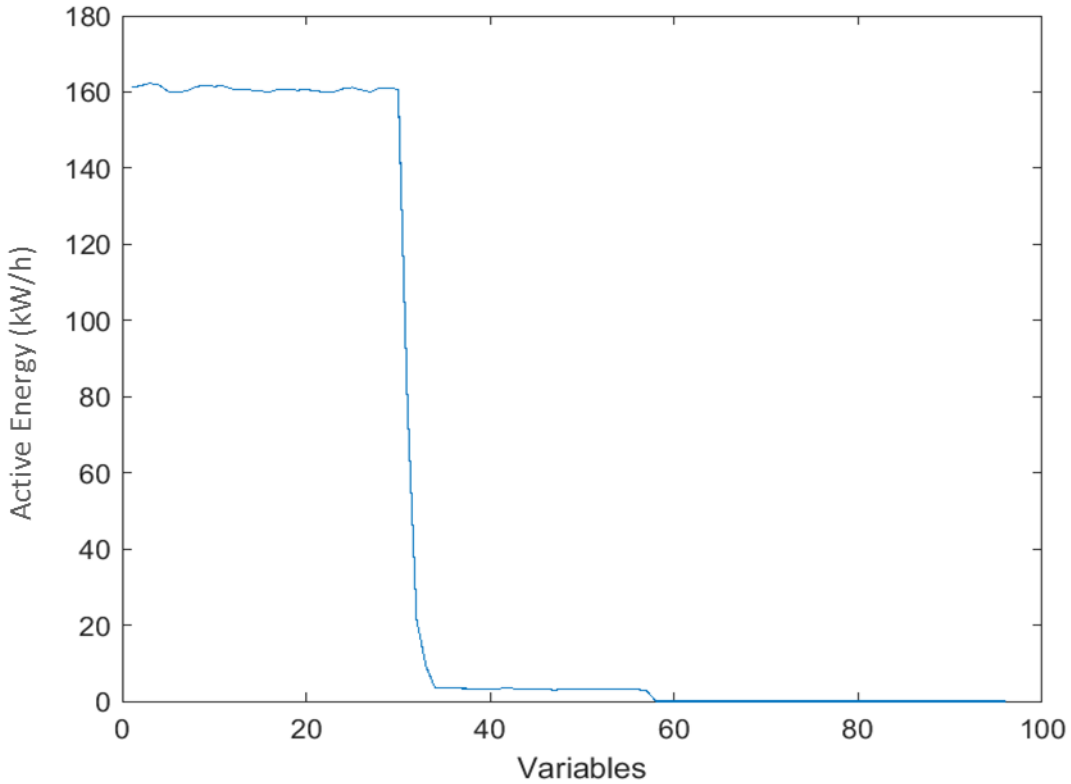


Figura 77-andamento dei dati classificati dal neurone 49, cioè il giorno 02/12/2020

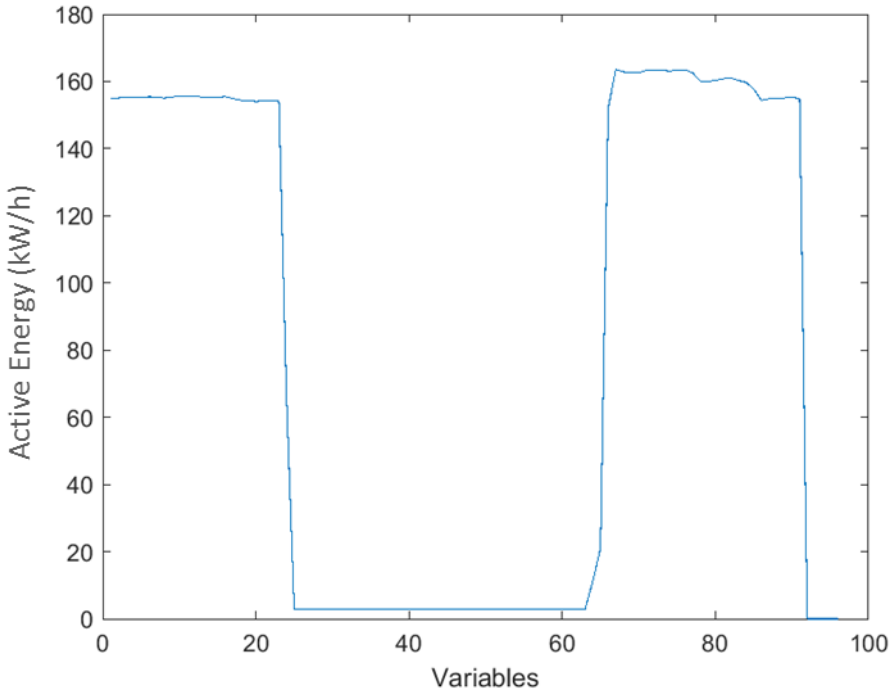


Figura 78-andamento dei dati classificati dal neurone 12, cioè il giorno 25/10/2020

La Figura 79 mostra l'andamento dell'energia attiva per i neuroni 6(21/09), 7 e 14 (dal 03/12 al 09/12, stessa clusterizzazione ottenuta con la PCA). È interessante ancora una volta notare come la rete neurale dia l'indicazione quantitativa della "somiglianza" tra gli andamenti delle osservazioni.
Questi ultimi dati rappresentano chiaramente dei dati anomali, presentando assenza di valori di Energia Attiva.

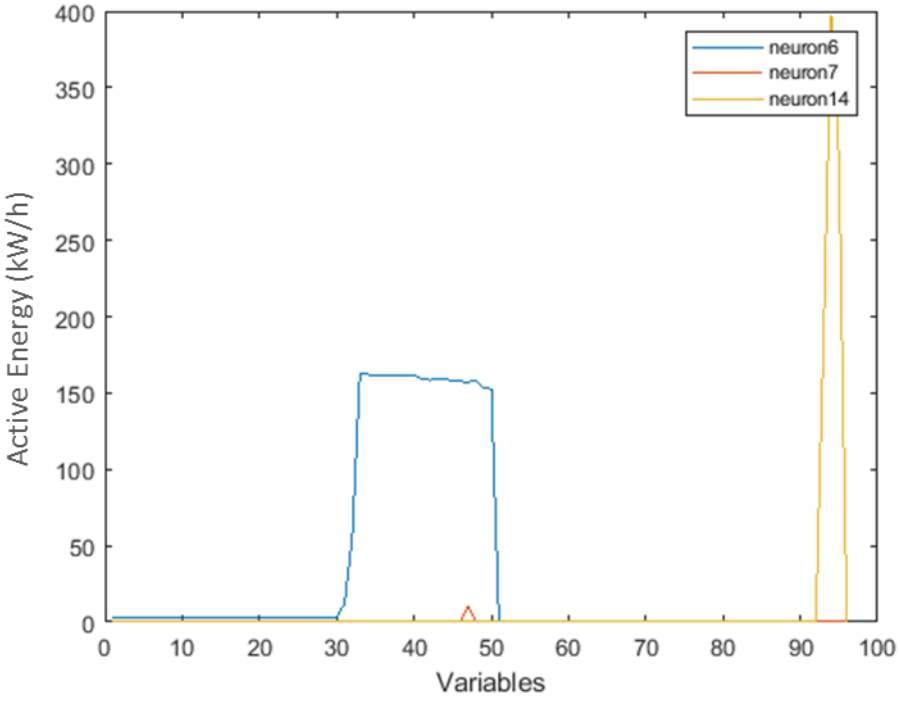


Figura 79- l'andamento dell'energia attiva per i neuroni 6(21/09), 7 e 14 (dal 03/12 al 09/12

A titolo di esempio, possiamo analizzare il grafico temporale (Figura 80), in cui in ascissa è indicato il tempo ed in ordinata il valore di energia attiva, possiamo vedere i dati relativi ai giorni classificati dai neuroni, da cui si vede che molti dati sono pari a zero.

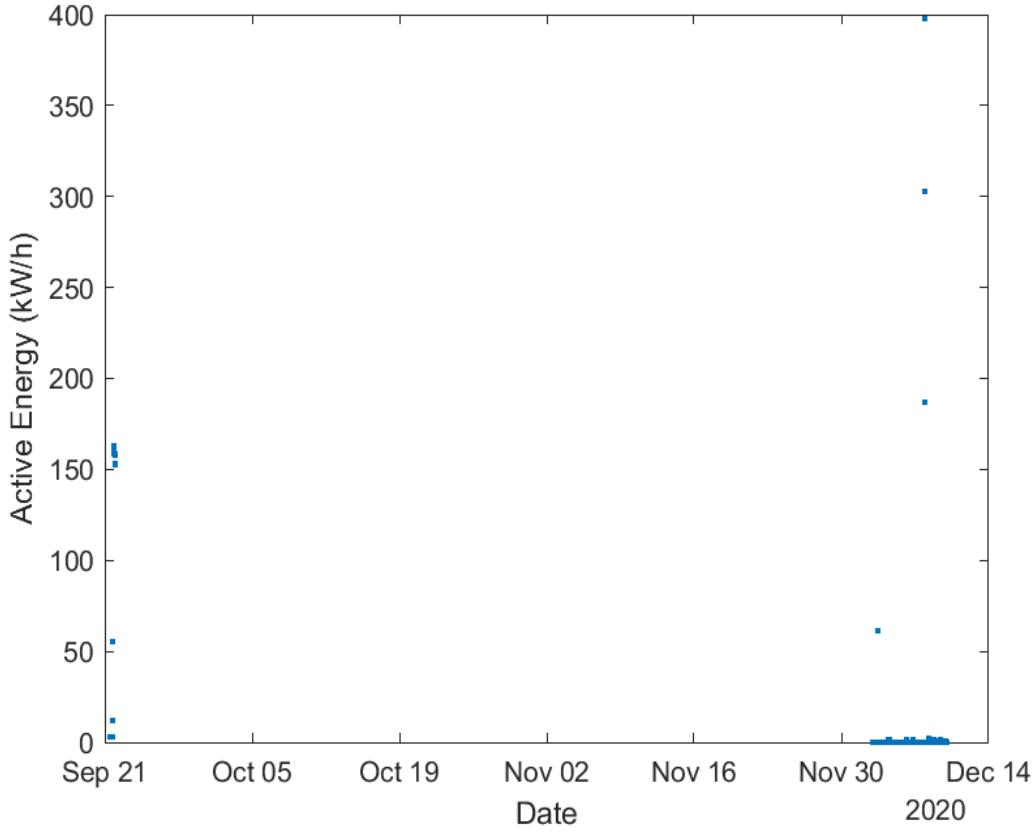


Figura 80-grafico temporale dei giorni che presentano valori di energia attiva assenti o pari a zero

Come già evidenziato la rete neurale dà molte indicazioni non solo sui vari cluster, ad esempio in quest'ultimo caso ha valorizzato una ulteriore classificazione per i neuroni 15 e 9 relativi al periodo che va dal 09/08 al 18/08, ma anche quanto i vari cluster siano collegati tra loro.

2.3.12 Incremento dei dati

Siccome i dati reali prelevati dai POD non hanno mostrato alcuni tipi di anomalie, segno di corretto funzionamento degli impianti, si è proceduto ad utilizzare dei dati autogenerati per ampliare le casistiche di studio (circa 800 osservazioni).

A tal motivo è stata anche implementata una matrice SOM 15x15 comprendente dunque 255 neuroni.

Il grafico dei pesi è rappresentato nella Figura 81, nella quale si rende più evidente la suddivisione delle aree e dei vari cluster, separati dai collegamenti di colore più scuro, che separano i neuroni con distanze maggiori e rappresentano i bordi tra i clusters.

La grafica ci permette di identificare visivamente i vari clusters (Figura 82 e Figura 83), ma essendo una metodologia indiretta, seppur immediata, non è sufficiente a garantire l'assenza di errori di valutazione.

Come già indicato, la visualizzazione permette di raggruppare visivamente i dati fornendo informazioni, tuttavia è bene utilizzare tecniche aggiuntive al fine di cercare di identificare i clusters in modo più accurato, evidenziando maggiormente eventuali problematiche sui dati.

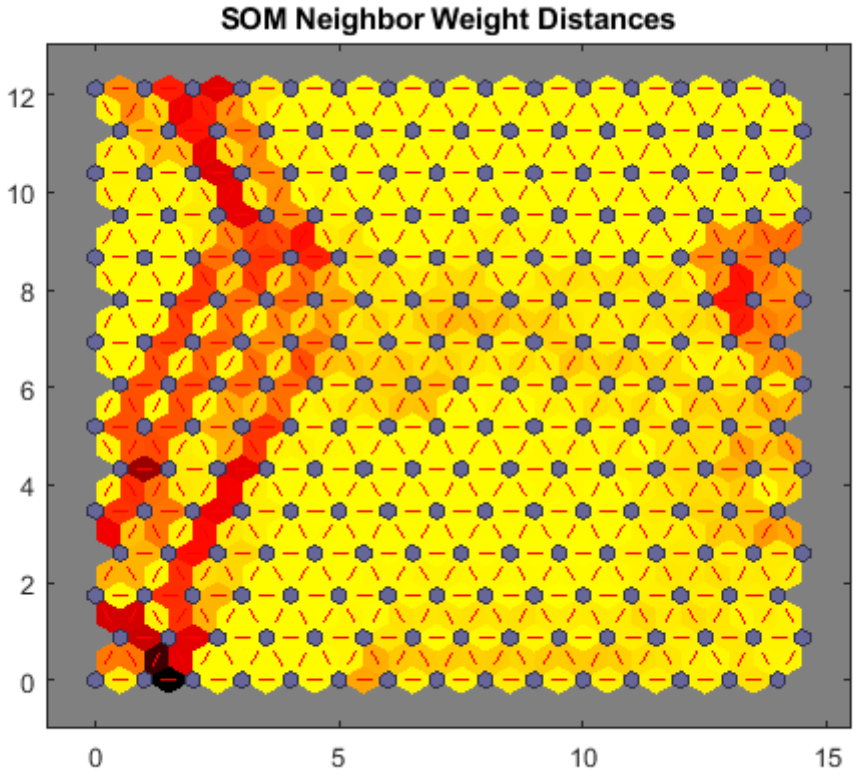


Figura 81-Grafico dei pesi per una matrice SOM 15x15 comprendente 255 neuroni

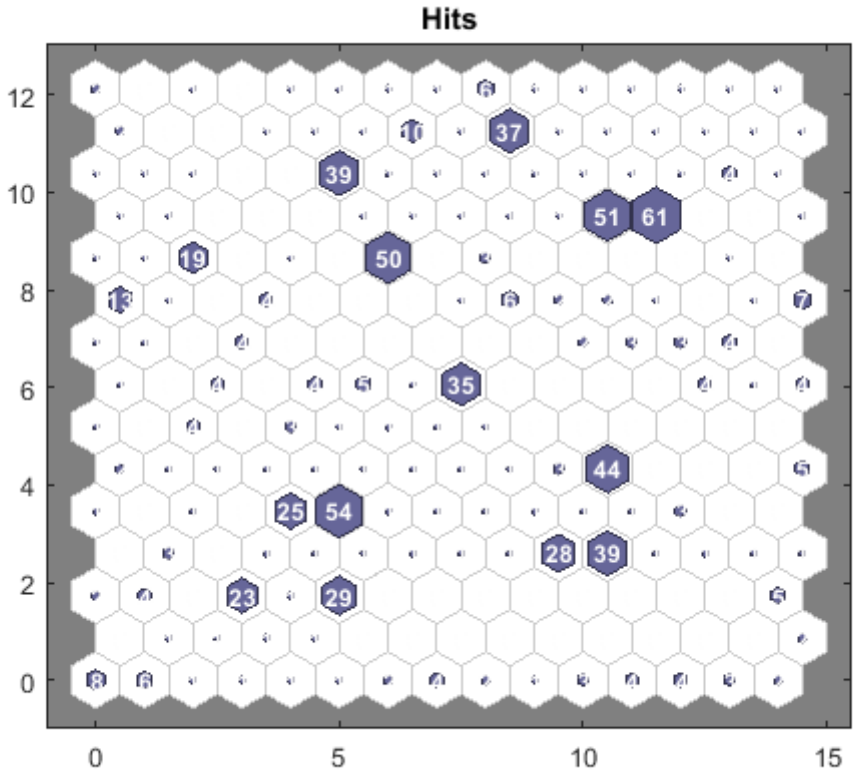


Figura 82-Grafico delle osservazioni per una matrice SOM 15x15 comprendente 255 neuroni

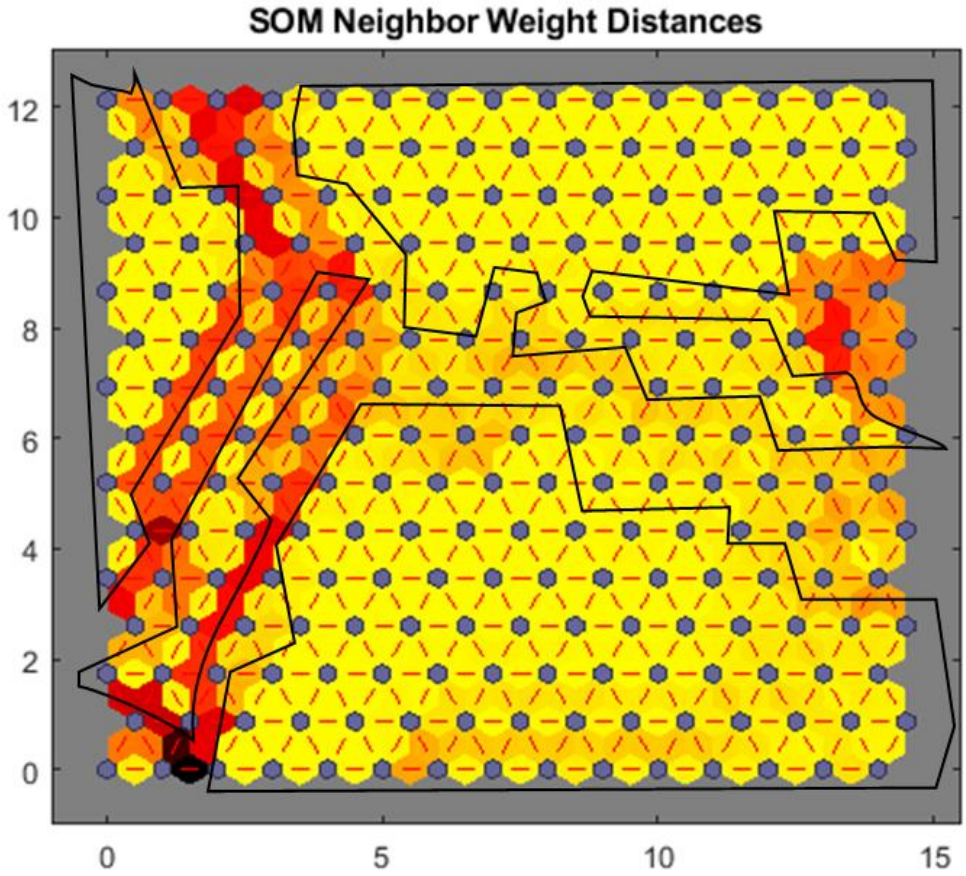


Figura 83-Cluster definiti per una matrice SOM 15x15

Per fare questo abbiamo tentato di applicare una ulteriore tecnica di clustering ai dati, in particolare l'algoritmo K-means e l'indice di Davies Bouldin che, combinati insieme, danno una idea del numero di cluster ottimale in base ai dati in possesso.

L'indice di Davies-Bouldin si basa sul rapporto tra la somma delle distanze medie tra gli elementi e il centro di due clusters e la distanza dal centro tra i due cluster.

Questo rapporto viene sommato per ogni coppia di cluster, ed infine deve essere minimizzato per evidenziare i cluster meno distanti tra loro.

A tal fine è stata utilizzata la funzione

evalcluster(M,'kmeans't 'DaviesBouldin','KList','[1,10]')

in cui M è la matrice di input, cioè le osservazioni, **clust** è l'algoritmo di clustering utilizzato (K-means nel nostro caso), **Name** e **Value** è una coppia di argomenti che nel nostro caso è **KList** -> [1,10] che indica il numero di cluster che vogliamo testare tramite l'algoritmo K-means (10 clusters) per cercare il numero ottimale di cluster.

La funzione applicata ai dati (800 osservazioni), restituisce i dati in Figura 84 che consistono in un numero di cluster ottimali pari a 4.

```
>> E=evalclusters(dati,'kmeans','DaviesBouldin','KList',[1:10])

E =

  DaviesBouldinEvaluation with properties:

    NumObservations: 800
         InspectedK: [1 2 3 4 5 6 7 8 9 10]
    CriterionValues: [1×10 double]
           OptimalK: 4
```

Figura 84-numero ottimale di cluster secondo la regola di Davies Bouldin

La Figura 85 mostra il valore dell'indice di Davis-Bouldin al variare del numero di cluster, da cui si evince che il valore più basso dell'indice corrisponde ad un numero di cluster pari a 4.

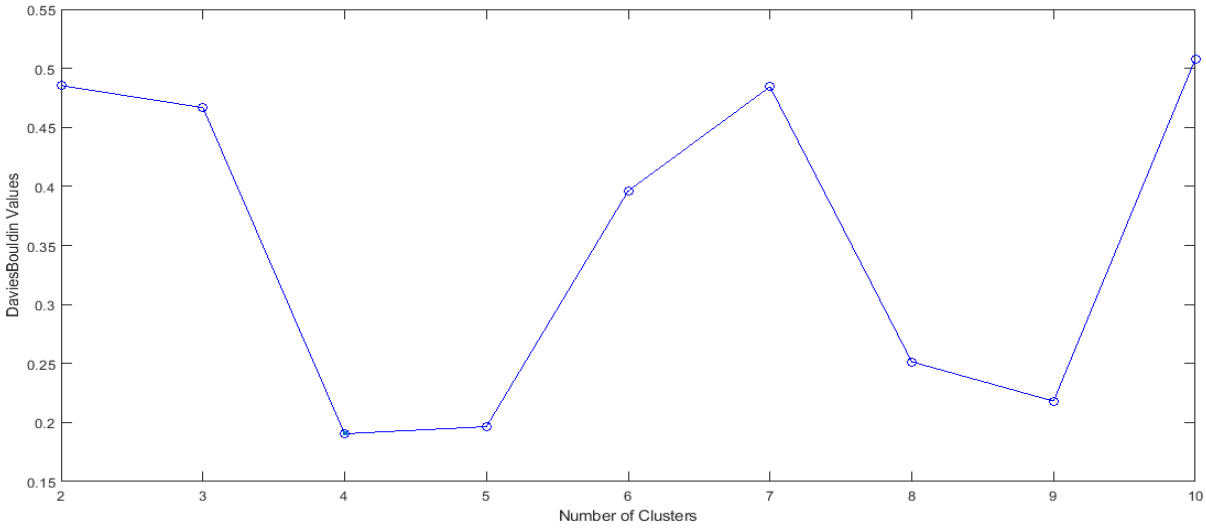


Figura 85- valore dell'indice di Davis-Bouldin al variare del numero di cluster

Il risultato è coerente con quanto ottenuto dalla mappa. Ovviamente maggiore è il numero di osservazioni migliore è la classificazione che si ottiene.

Restano parzialmente fuori dai clusters alcune osservazioni che possono essere considerate dati anomali oppure errati (assimilabili a rumore), che vengono brevemente descritti di seguito.

I neuroni 1 e 2 rappresentano i 14 record del 2019, già analizzati in precedenza.

Il neurone 90 rappresenta 5 osservazioni relative a dicembre 2020 (02/12/2020 e dal 10/12/2020 al 13/12/2020) il cui andamento medio è rappresentato in Figura 86.

Esso rappresenta sicuramente un andamento dei consumi anomalo, presentando valori molto bassi di energia attiva anche durante le re serali in cui è prevista invece una accensione.
Anche il neurone 150 presenta un cluster anomalo già evidenziato (Figura 87).

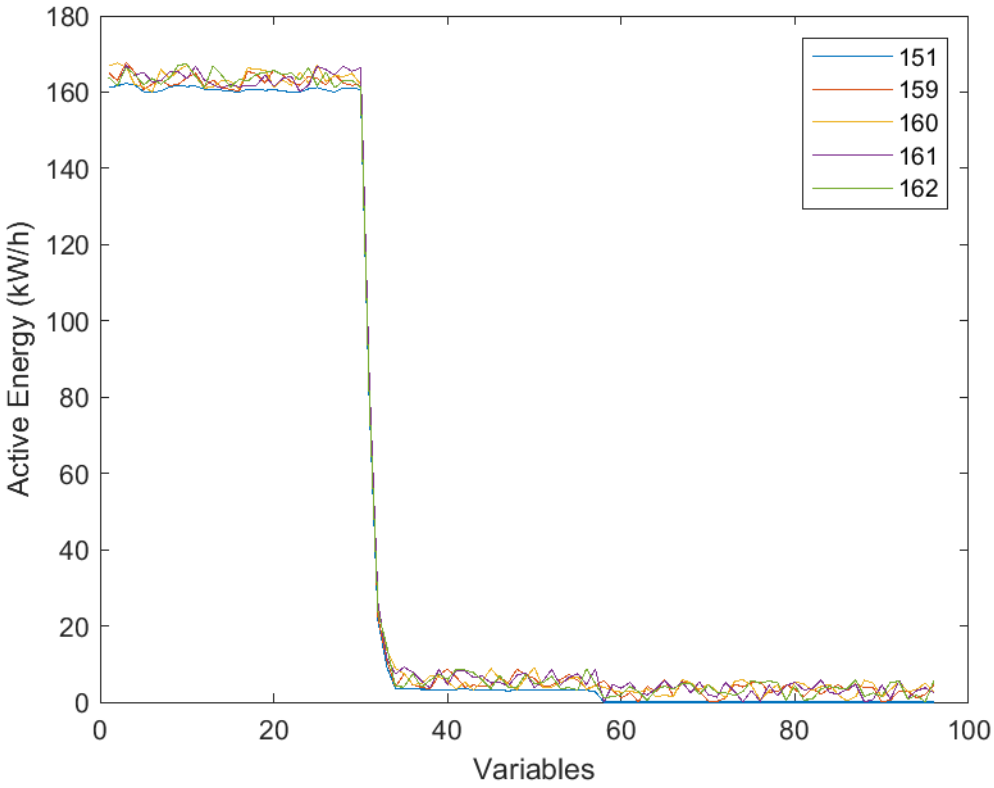


Figura 86-andamento delle osservazioni relative a dicembre 2020 (neurone 90)

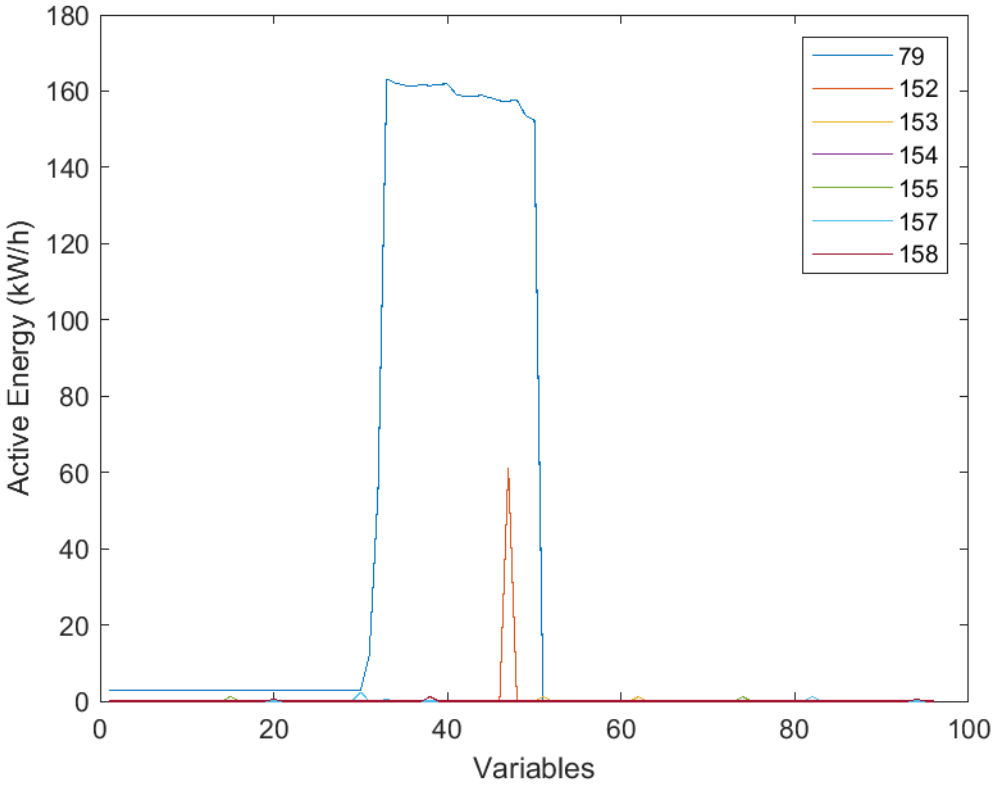


Figura 87-cluster relativo al neurone 150, comprendente 7 osservazioni(in legenda sono indicati gli indici) che corrispondono al 21/09/2020, dal 01/12/2020 al 06/12/2020 e dallo 08/12/2020 allo 09/12/2020

Il cluster 149 è rappresentato in Figura 88 e comprende un'unica osservazione che presenta solo 2 valori di energia attiva.

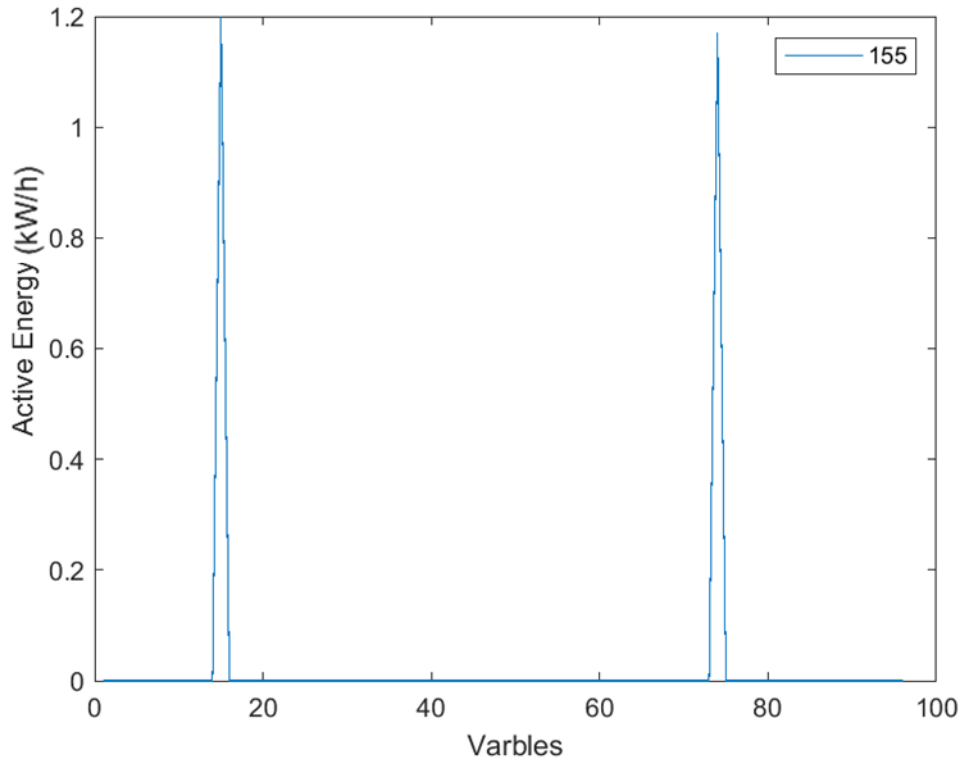


Figura 88-cluster relativo al neurone 149 comprendente una singola osservazione

Anche le osservazioni classificate dai neuroni 164 e 213, rappresentano anch'essi dei dati errati (Figura 89).

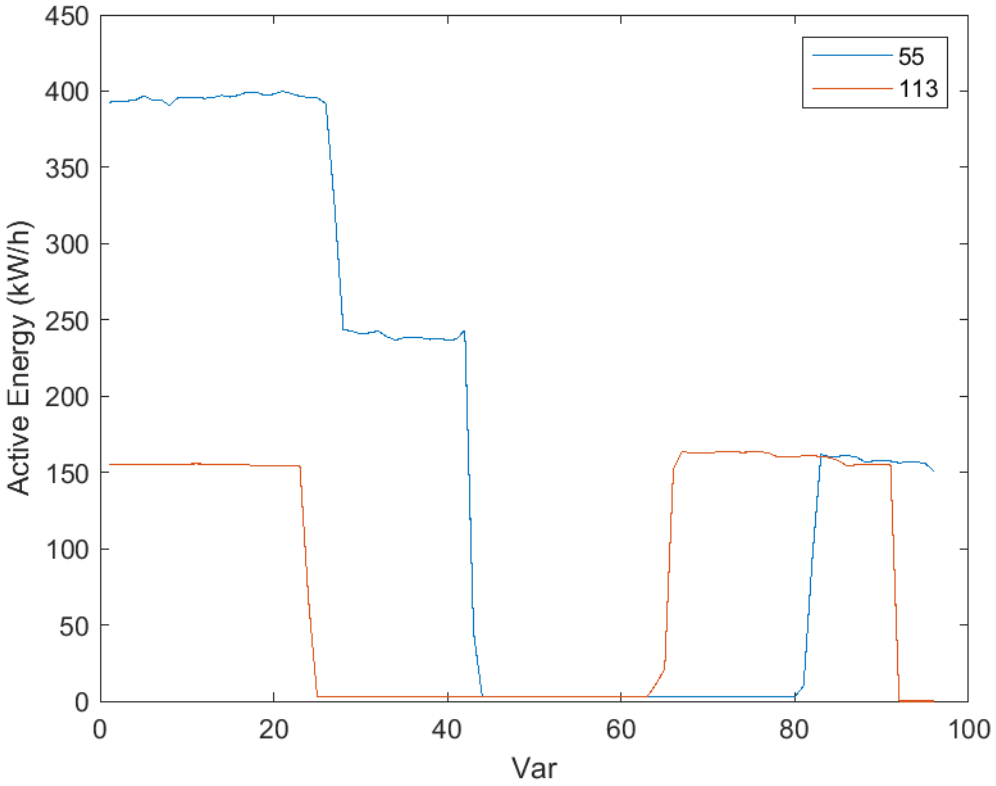


Figura 89-cluster relativi ai neuroni 164 e 213

Analizzati i dati sparsi, caratterizzati da distanze elevate tra i pesi, vediamo ora i clusters più grandi che corrispondono alla maggior parte delle osservazioni.

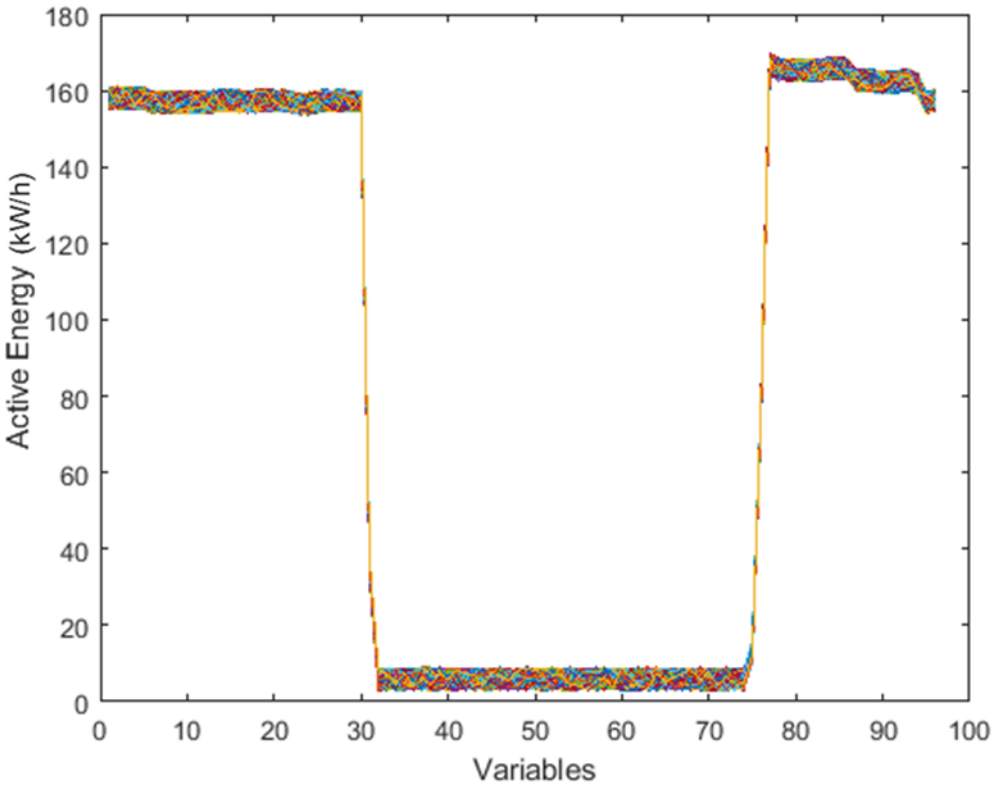


Figura 90-andamento delle osservazioni per il cluster comprendente circa 350 osservazioni (il 40% del totale)

La Figura 90 e la Figura 91Figura 92 mostrano il cluster più grande comprendente circa 350 osservazioni, che rappresenta più del 40% delle osservazioni totali.
L'andamento è quello tipico di un impianto di illuminazione pubblica che, durante il giorno, a partire dalle 8:00 circa si spegne e si accende nuovamente verso le 17:45.

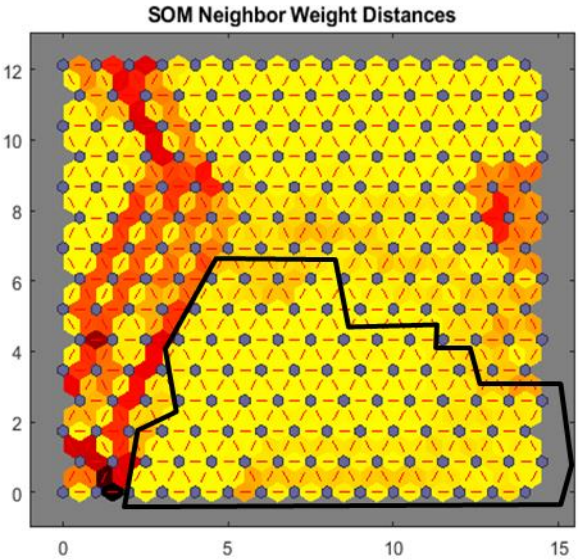


Figura 91-il cluster comprendente circa 350 osservazioni (il 40% del totale)

La Figura 92 mostra il secondo cluster di dimensioni grandi, che comprende circa 250 dati, cioè circa il 30% delle osservazioni totali.

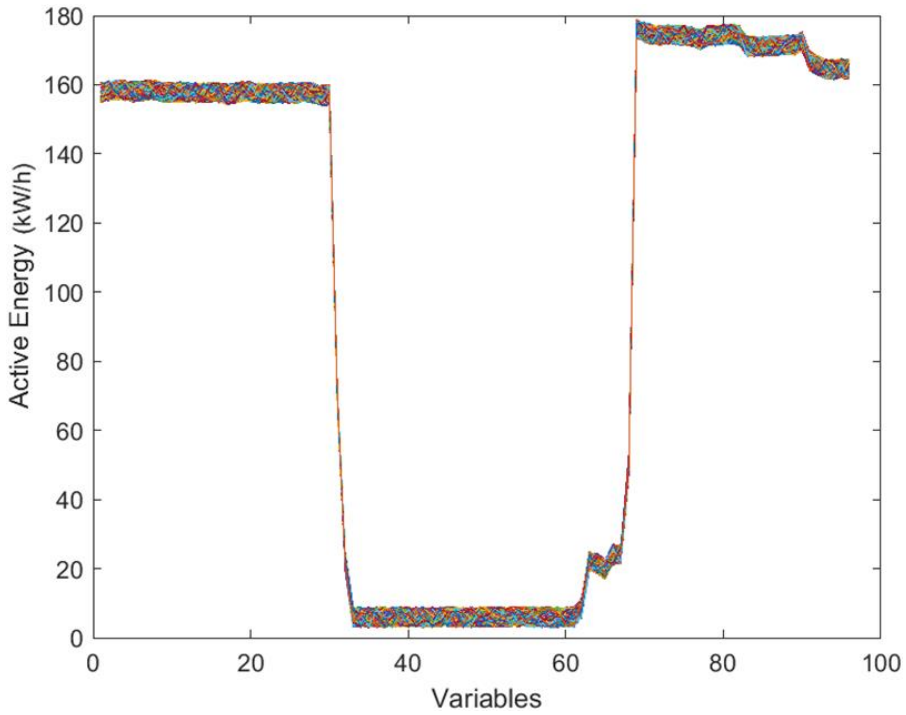


Figura 92-andamento delle osservazioni per il cluster comprendente circa 250 osservazioni (il 30% del totale)

Essa mostra un andamento particolare di un impianto che intorno alle 07:15 si spegne per poi accendersi nuovamente intorno alle 15:15 ma ha un andamento crescente fino alle 17:00 circa, orario dopo il quale l'impianto va a regime.
La spiegazione dell'andamento anomalo tra le 15:00 e le 17:00 è dovuto al fatto che i lampioni collegati al POD considerato non si accendono tutti allo stesso momento, oppure che vi siano, ad esempio, meccanismi di regolazione automatica dell'illuminazione per cui, in alcune strade, l'illuminazione prevede un maggior dispendio di energia rispetto ad altre, quindi alcuni lampioni consumano di più e altri di meno.

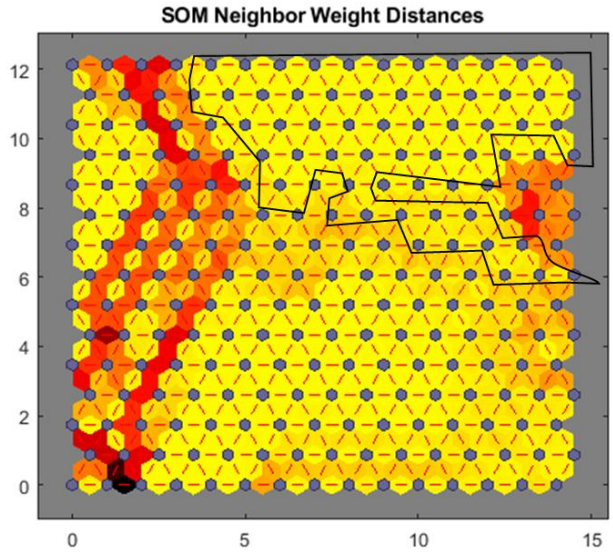


Figura 93- il cluster comprendente circa 250 osservazioni (il 30% del totale)

La Figura 94 e la Figura 95 mostrano un ulteriore cluster in cui il consumo risulta essere più elevato rispetto a quello rilevato negli altri cluster.
Ciò significa che per un certo periodo di tempo l'impianto presenta un consumo più che doppio rispetto alla norma, e questo può essere segnale di un problema.
Anche il periodo in cui l'impianto dovrebbe essere spento o comunque presentare un basso assorbimento, presenta un consumo di circa 250 kW/h che risulta molto elevato.
Un temporaneo incremento dei consumi è indicazione di un assorbimento non previsto dovuto, ad esempio, a modifiche nell'impianto (aggiunta di un carico come antenne, videocamere, altre attrezzature che possono essere montate sui lampioni, ecc.) oppure a cause di tipo fraudolento (furto di energia).

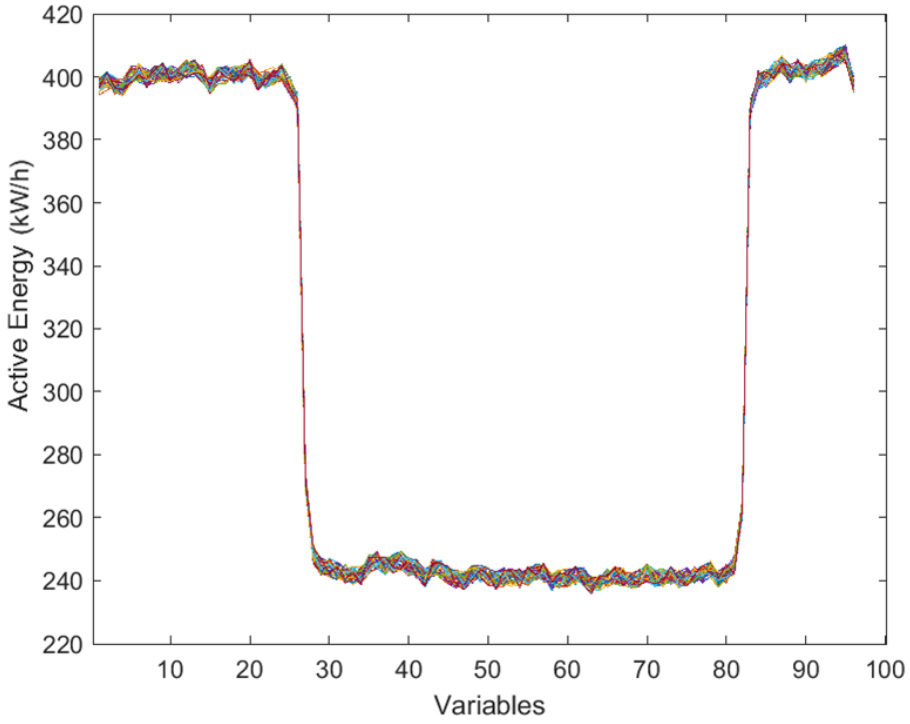


Figura 94-andamento delle osservazioni in cui si evidenzia un consumo più elevato

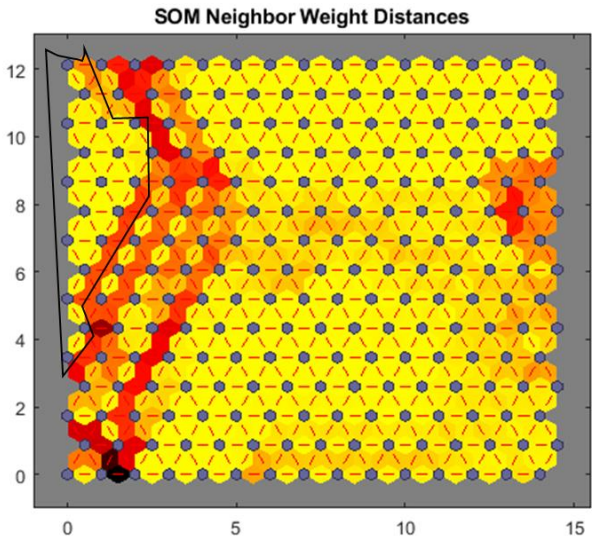


Figura 95-cluster che classifica osservazioni in cui si evidenzia un consumo più elevato

L'ultimo cluster (Figura 96 e Figura 97) che abbiamo analizzato, presenta in andamento particolare dell'energia attiva. Infatti il consumo di energia attiva diminuisce progressivamente fino allo spegnimento durante le ore centrali del giorno, aumentando nuovamente in 2 fasi. Questo è dovuto ad una rimodulazione del flusso luminoso durante la notte e durante il pomeriggio, per abbassare il consumo di energia

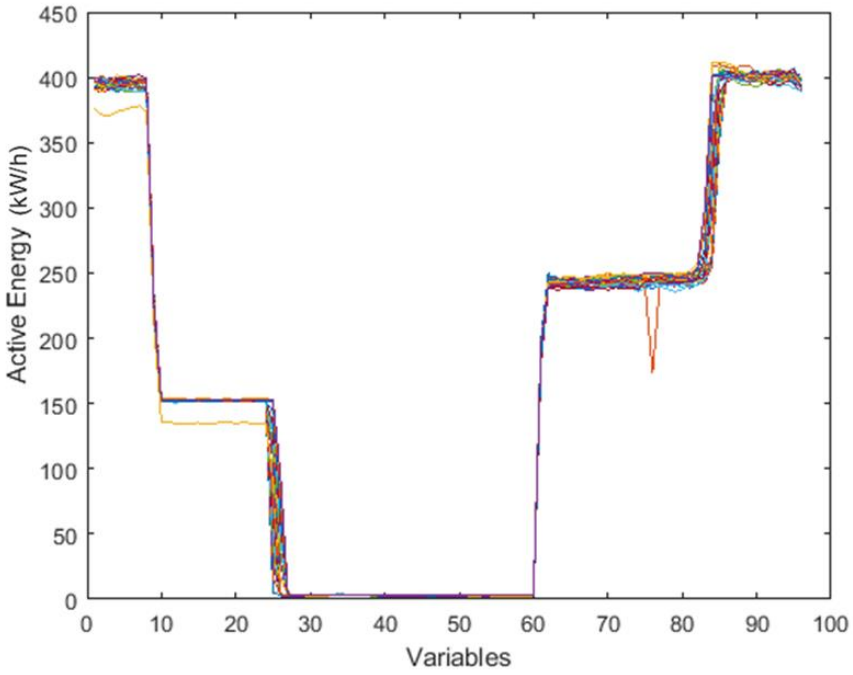


Figura 96-andamento del consumo di energia attiva, che diminuisce progressivamente fino allo spegnimento durante le ore centrali del giorno, aumentando nuovamente in 2 fasi

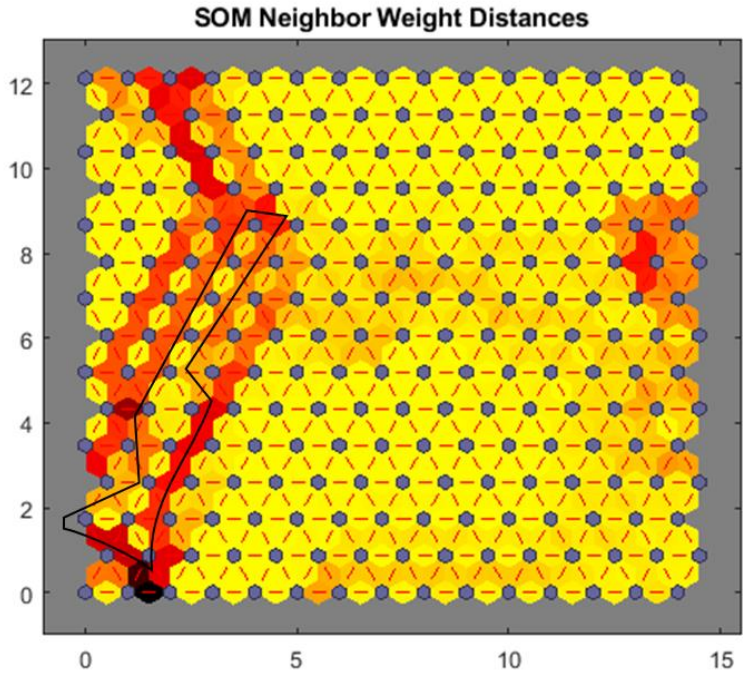


Figura 97-cluster relativo ad osservazioni che presentano un andamento del consumo di energia attiva, che diminuisce progressivamente fino allo spegnimento durante le ore centrali del giorno, aumentando nuovamente in 2 fasi

2.3.13 Risultati

Dall'analisi sopra riportata è possibile effettuare alcune considerazioni interessanti.

Entrambi i metodi di analisi, la PCA e la rete SOM, offrono la possibilità di "classificare" i dati di ingresso suddividendoli in gruppi ciascuno supportato da determinate caratteristiche, riducendo dunque la complessità del problema.

La PCA è una metodologia di tipo lineare, che si basa sul calcolo della varianza, mentre la rete SOM è una metodologia di analisi dei dati non lineare che si basa sulla autoorganizzazione di una struttura neuronale in base alla distanza euclidea tra i vari neuroni; dall'analisi, entrambi i metodi mostrano risultati comparabili, ed una buona clusterizzazione, tuttavia la rete neurale fornisce maggiori informazioni che permettono di valutare meglio a quale cluster appartenga una determinata osservazione.

Si può dire che la rete SOM, in qualche modo completi la PCA fornendo quelle informazioni che permettono di definire quanto una osservazione sia simile a quelle vicine, anche da un punto di vista grafico.

Una rete neurale può superare quelli che sono i limiti della PCA, essendo applicabile praticamente a ogni set di dati, senza vincoli o ipotesi iniziali.

La visualizzazione grafica dei collegamenti tra i vari neuroni permette di quantificare la distanza tra essi e, dunque, tra i vari dati.

Attraverso la simulazione di nuovi dati, riconducibili a casistiche di vario tipo, è stata provata una rete neurale di dimensioni superiori che ha evidenziato comunque una buona capacità di classificazione dei dati, riducendo la complessità del problema a pochi clusters da analizzare.

Durante il training della rete neurale, i pesi dei neuroni vengono aggiornati e alcuni neuroni che classificano i dati risultano più distanti oppure isolati dagli altri. Questi evidenziano anomalie casuali e rappresentano una minima parte dei casi che possono presentarsi (ad esempio assenza di campioni). In tal caso è possibile che il misuratore possa avere dei problemi che impediscono la corretta gestione dei dati, oppure che l'impianto, in determinati momenti, presenti in effetti delle anomalie.

I clusters più grandi, anche da un punto di vista numerico, rappresentano la gran parte delle osservazioni che comunque devono essere analizzate caso per caso per verificarne il significato.

Sebbene entrambe le tecniche di analisi possano prendere in input varie tipologie di dati e un grande numero di variabili, riducendo la complessità del sistema, la rete Neurale è più efficiente nel determinare anomalie o non linearità tra i dati rispetto alla PCA.

Un altro vantaggio della rete neurale è che durante la fase di training o addestramento della rete, i pesi vengono attribuiti in modo da ridurre il rumore attraverso il calcolo della media dei vari punti. Infatti, ogni dato che si aggiunge ad un cluster permette una riorganizzazione efficiente della rete, evidenziando anomalie reali e riducendo invece eventuali dati considerati come rumore.

Lo svantaggio principale della rete Neurale SOM è stabilire la dimensionalità corretta della rete, in modo che sia il più possibile efficace.

Sebbene ci siano suggerimenti in merito, tuttavia tale dimensione deve essere verificata sperimentalmente attraverso varie prove e verificando la sua dimensione ottimale. Dato il numero limitato di dati, al fine di verificare la classificazione degli stessi attraverso una rete SOM, siamo partiti da reti neurali di dimensioni limitate, ma il numero di neuroni può arrivare anche a 5 volte il numero di osservazioni, pur non essendoci una regola specifica.

3 Conclusioni

Nell'ambito del piano triennale della ricerca 2019-2021 per il sistema elettrico nazionale, per il quale l'ENEA ha predisposto il piano triennale di realizzazione, PTR 2019-2021, il Dipartimento di Scienze dell'Università degli Studi "Roma Tre" è stato interessato per una attività di ricerca dal titolo "Smart Energy in Sistemi Pubblici: Analisi di Affidabilità e Qualificazione dei Dati per Ridurre le Incertezze di Sistema".

L'università, partendo dallo studio della letteratura in merito all'analisi di Big Data energetici che ha interessato principalmente la scorsa annualità e basandoci sui dati fornitoci da ENEA, durante questo terzo anno (2021), si è occupata di implementare alcune specifiche metodologie di analisi cercando di sfruttare, in modo quanto più accurato ed efficiente possibile, la granularità quartoraria dei dati di consumo.

A tal fine sono stati elaborati e trasformati i dati a nostra disposizione perché potessero essere studiati attraverso strumenti informatici appositi (in particolare MATLAB), e successivamente sono state applicate due metodologie che permettono la riduzione della complessità del sistema, semplificandone l'analisi, e permettendone di studiare la parte statisticamente più significativa.

Tra i vari approcci disponibili abbiamo utilizzato la PCA (Principal Component Analysis) e le reti neurali di tipo SOM (Self Organizing Map) perché supportati da una robusta letteratura che li indicava come potenziali tecniche elettive per l'analisi del nostro data set.

Le due metodologie lavorano classificando i dati in gruppi o cluster. Ciascun cluster raggruppa i dati sulla base di specifici criteri supportati dagli stessi dati. L'uso di queste metodologie consente di ridurre la complessità di un problema comunque articolato all'analisi di pochi raggruppamenti di dati, permettendo di valutare le caratteristiche di ogni singolo gruppo, che non emergerebbero con altrettanta chiarezza analizzando esclusivamente il trend temporale dei dati.

Sebbene effettuino la stessa operazione di classificazione dei dati, la modalità di lavoro delle due metodologie è differente; infatti, mentre la PCA è una metodologia di analisi tipo lineare, che si basa sul calcolo della varianza, la rete SOM è una metodologia di analisi non lineare che si basa sulla autoorganizzazione di una struttura neuronale in base alla distanza euclidea tra i vari neuroni.

Entrambi i metodi sono risultati molto efficienti per la classificazione dei dati a nostra disposizione, permettendo l'individuazione di cluster specifici.

Successivamente ogni cluster è stato analizzato valutando le caratteristiche delle osservazioni da essi classificate, ed evidenziando sia alcune casistiche di lavoro dell'impianto, sia alcune anomalie legate a dati mancanti o visibilmente errati.

Questo fornisce una indicazione della qualità dei dati a disposizione, e attraverso lo studio delle varie casistiche, anche della bontà dell'impianto monitorato.

Se da un punto di vista della classificazione entrambi i metodi sono risultati efficaci, da un punto di vista analitico la rete SOM ha mostrato una migliore "quantificazione" della similitudine tra i dati di più cluster, grazie anche a informazioni grafiche dalle quali è più facile estrarre risultati sintetici.

I dati fornitici da ENEA hanno presentato alcune casistiche interessanti ma non esaustive; abbiamo quindi pensato di ampliare i casi di studio attraverso la simulazione di circa 800 nuove osservazioni, ottenendo ancora una volta una buona capacità di classificazione dei dati, e riducendo la complessità del problema a pochi clusters da analizzare.

Il lavoro condotto dal Laboratorio di Misure Elettriche ed Elettroniche di Roma Tre per quest'anno ha quindi dimostrato che l'applicazione di metodologie specifiche di analisi lineare e non lineare permette di individuare informazioni in un dominio trasformato che sintetizzano e fanno emergere aspetti salienti di un sistema comunque complesso consentendone una sua più immediata gestione operativa.

Nel nostro caso, il sistema complesso è rappresentato dal database del sistema PELL-IP oggetto di parte dell'attività ENEA. L'enorme quantità di dati reali di consumo provenienti da impianti di illuminazione

pubblica e immagazzinati nel database lo classifica come un progetto che ricade nel topic scientifico dei “Big Data”. La classificazione dei dati conseguenti all’applicazione della PCA e delle reti SOM al database rappresenta una novità importante in quanto, a nostra conoscenza, l’applicazione di queste tecniche ad un database così vasto e particolare non è un argomento facilmente rintracciabile nella letteratura scientifica. I risultati ottenuti sono estremamente confortanti e lasciano intravedere uno sviluppo dell’attività di ricerca. L’uso di queste metodologie consente di studiare agevolmente dati così particolari, e ciò apre a nuove modalità di analisi con il fine di gestire in modo ancor più efficiente il monitoraggio degli impianti di illuminazione pubblica.

4 Riferimenti bibliografici

I riferimenti bibliografici devono essere richiamati nel testo con numeri progressivi tra parentesi quadre e riportati a fine testo con il seguente formato:

1. ENEA "*Progetto Lumiere".* Disponibile on line al sito: http://www.progettolumiere.enea.it/
2. ENEA "*PELL-Lumière & Public Energy Living Lab (PELL) per una gestione efficiente della Pubblica Illuminazione*" Disponibile al sito: http://www.enea.it/it/comunicare-la-ricerca/events/pell_18mag16/ENEA-Roma.
3. Specifiche di contenuto di riferimento PELL-illuminazione pubblica https://geodati.gov.it/geoportale/images/Specifica-PELL-IP_ver-1.0_20180723.pdf, 23-07-2018
4. D. Liu, Q. Chen and K. Mori, "Time series forecasting method of building energy consumption using support vector regression," 2015 IEEE International Conference on Information and Automation, 2015, pp. 1628-1632, doi: 10.1109/ICInfA.2015.7279546.
5. A. Aranda, G. Ferreira and M. D. Mainar-Toledo, "Multiple regression models to predict the annual energy consumption in the Spanish banking sector", Energy and Buildings, vol. 49, pp. 380-387, March 2012.
6. U. Marikkar, A. S. Jameel Hassan, M. S. Maithripala, R. I. Godaliyadda, P. B. Ekanayake and J. B. Ekanayake, "Modified Auto Regressive Technique for Univariate Time Series Prediction of Solar Irradiance," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 22-27, doi: 10.1109/ICIIS51140.2020.9342694.
7. J. Xiao, Haiyan Sun, Yi Hu and Y. Xiao, "GMDH based auto-regressive model for China's energy consumption prediction," 2015 International Conference on Logistics, Informatics and Service Sciences (LISS), 2015, pp. 1-6, doi: 10.1109/LISS.2015.7369754.

5 Abbreviazioni ed acronimi

PELL: Progetto Public Energy Living Lab

KPI: Key Performance Indicator

LED: Light Emitting Diode

kW: kilowatt

IP: illuminazione pubblica

PCA: Principal Component Analysis

MDA: Mixture discriminant analysis

JSON: Javascript object notation

MQTT: Message Queue Telemetric Transport

ENEA: Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile

POD: Point Of Delivery

SCPS: Smart City Platform Specification

SCP: Smart City Platform

CSV: Comma Separated Value

ARERA: Autorità di regolazione per energia, reti e ambiente

NNCT: Neural Network clustering tool

SOM: Self Organizing Map

Appendice: Laboratorio di Misure Elettriche ed Elettroniche dell'Università degli Studi "Roma Tre": Curriculum Scientifico

Responsabile: Dott. Ing. Ph.D. RTI Fabio Leccese

Collaboratori: Dott. Enrico Petritoli (Assegnista di Ricerca), Dott.sa Mariagrazia Leccisi (Borsista)

Il laboratorio fa parte del Gruppo Nazionale delle Misure Elettriche ed Elettroniche (GMEE) i cui scopi principali sono lo studio delle misure o "metrologia", l'analisi di qualità fisiche e la realizzazione di campioni di misura con particolare attenzione allo studio dell'incertezza di misura.

In questo quadro generale, il nostro laboratorio segue da anni diverse linee di ricerca tra le quali la qualità dell'energia (power quality-dal 2004), l'analisi informativa dei segnali (dal 2002), i controlli di apparati locali e remoti ed in particolare di sistemi di risparmio energetico applicati ad illuminazione e riscaldamento (dal 2008), la sensoristica distribuita incluse le Wireless Sensor Network (dal 2008) e le analisi affidabilistiche di sistemi complessi (dal 2013) trovano ampia utilità e complementarità con le attività svolte in ENEA dal gruppo del Laboratorio Smart Cities and Communities

Ciascuna linea presenta peculiarità proprie che coinvolgono non solo il campo specifico delle misure, ma anche settori ad esso correlati quali l'elettronica, l'elettrotecnica, le telecomunicazioni, l'informatica e l'automazione. Il Laboratorio progetta e sviluppa sistemi di misura avvalendosi dei software più moderni come Orcad o Protel e programmando microcontrollori di varie famiglie come Microchip o Siemens, processori ARM, avendo confidenza anche con la progettazione di FPGA. I linguaggi di programmazione più usati sono il C, la piattaforma .NET e vari linguaggi "WEB oriented".

Nel campo della Didattica abbiamo nel tempo sviluppato dei percorsi all'interno dei nostri Dipartimenti rivolti al mondo dell'ambiente e dell'energia con materie come Qualità Ambientale, Qualità dell'Energia, Elementi di Misure per l'Analisi Ambientale, Alimentazione da Fonti Rinnovabili e Strumentazione Avanzata di Misura che, nel tempo, sono state apprezzate da un numero crescente di studenti.

A testimonianza dell'esperienza maturata, il nostro lavoro ha portato ad oltre 200 pubblicazioni scientifiche per la maggior parte presentate su Riviste Scientifiche Internazionali o su Congressi Scientifici Internazionali, come riscontrabile dalle 143 riportate sul Database Scopus. Di queste, 7 sono state scritte insieme ad ENEA e sempre con ENEA sono stati affrontati insieme 7 progetti di ricerca.

Si segnalano i premi nazionali/internazionali vinti dal Laboratorio

1) 2016-PUBLONS-The sentinels of Science Awards 2016 The top 10 percent of reviewers- Certified Sentinel of Science award recipient: As one of the top 10 per cent of researchers contributing to the peer review of the field of Chemistry

2) 2018-II Forum Nazionale Delle Misure-Sezione GMEE-Padova, 17-19 Settembre 2018: Miglior Poster per l'articolo: "Measurements of Q factor in microwave resonators: relevance of the calibration" a cura di K. Torokhtii, A. Alimenti, N. Pompeo, F. Leccese, F. Orsini, A. Scorza, S.A. Sciuto, E. Silva.

3) 2018-IEEE International Workshop on Metrology for the Sea, October 08-10, Bari, Italy: Miglior Demo per il drone di nuova concezione con movimentazione a pendolo vincolato a cura di Eduardo De Francesco e Fabio Leccese.

4) 2019-WEB OF SCIENCE-PUBLONS -TOP PEER REVIEWER 2019 -For placing in the top 1% of reviewers in Cross-Field on Publons global reviewer database.

Elenco di partecipazioni a progetti scientifici

Progetti Internazionali:

1) "PROGETTO DI GRANDE RILEVANZA ITALIA-SERBIA 2016-2018 sul tema di Agriculture and Food Technologies dal titolo **SMART MONITORING OF PESTICIDES IN FARMING AREAS**" Finanziato dal Ministero degli Affari Esteri e della Cooperazione Internazionale. **Ruolo: Responsabile Scientifico**. Durata 3 anni.

Progetti Nazionali:

2) Bando PROGRAMMI DI RICERCA SCIENTIFICA DI RILEVANTE INTERESSE NAZIONALE RICHIESTA DI COFINANZIAMENTO **PRIN 2010-2011** dal titolo: "**Interazione fra minerali e biosfera: conseguenze per l'ambiente e la salute umana**"- sottosezione "**Emissioni antropogeniche di CO2: immobilizzazione per carbonatazione e discriminazione isotopica della componente fossile e non fossile**". PRIN 2010-2011, Area 04, Durata 36 mesi, Protocollo 2010 MKHT9B_007
3) Progetto **Co-Research POR FESR LAZIO 2007-2013-Titolo SIMPLFEX** Progetti di R&S in collaborazione presentati dalle PMI del Lazio con Numero di protocollo assegnato: FILAS-CR-2011-1076 dal 09/01/2012 al 08/01/2014. **Ruolo: Responsabile Scientifico di Sede.** Durata 2 anni.
4) Progetto di ricerca: **"Sviluppo e implementazione di algoritmi per applicazioni di Smart Lighting"** per conto di ENEA-Roma, 2014. **Ruolo: Responsabile Scientifico**. Durata 1 anno.
5) Progetto di ricerca: **"Sviluppo e implementazione di indicatori di prestazione e diagnostica energetica per impianti di illuminazione pubblica"** per conto di ENEA-Roma, 2014. **Ruolo: Responsabile Scientifico**. Durata 5 mesi.

I seguenti progetti sono stati sviluppati all'interno del piano Piano Triennale della Ricerca nell'ambito del Sistema Elettrico Nazionale 2015-2017 finanziato dal Ministero dello Sviluppo Economico (MiSE) e gestito da ENEA all'interno dell'Accordo di Programma MiSE-ENEA 2015-2017.

6) **Progettazione e sviluppo prototipale di strumenti per la gestione del PELL**, per conto di ENEA-Roma, 2015. **Ruolo: Responsabile Scientifico di Sede**. Durata 5 mesi.
7) **Analisi di affidabilità e analisi dei guasti e delle criticità (FMECA) del sistema smart street**, per conto di ENEA-Roma, 2016. **Ruolo: Responsabile Scientifico di Sede**. Durata 5 mesi.
8) **Studio affidabilistico dei componenti di una linea di illuminazione "smart" stradale pubblica operativa in contesto urbano: vantaggi e criticità**, per conto di ENEA-Roma, 2017. **Ruolo: Responsabile Scientifico di Sede**. Durata 5 mesi.
9) **Studio affidabilistico preliminare dei componenti fondamentali del sistema di termoregolazione dell'edificio F-40 ENEA (Casaccia): vantaggi e criticità**, per conto di ENEA-Roma, 2018. **Ruolo di Responsabile Scientifico di Sede**. Durata 3 mesi.
10) **Smart Energy in Sistemi Pubblici: analisi di affidabilità e qualificazione dei dati per ridurre le incertezze di sistema,** per conto di ENEA-Roma, 2019-2021. **Ruolo di Responsabile Scientifico di Sede**. Durata 3 anni.

Progetti Conto Terzi:

11) Progetto di ricerca: **"Studio delle criticità delle PowerLine su Navi da guerra"** per conto della Se.Te.L. group di Roma, 2012. **Ruolo: Responsabile Scientifico**. Durata 1 mese.
12) Progetto di ricerca: **"Valutazione del Processo di Rivitalizzazione degli Accumulatori al Piombo-Acido e del Relativo Liquido Additivo"** per conto della Battery Equalizer Italia s.r.l. di Fiumicino, 2012, **Ruolo: Responsabile Scientifico.** Durata 3 mesi.
13) Progetto di ricerca: **"Evoluzioni del supporto logistico delle power line di unità navali"** per conto della Se.Te.L. group di Roma, 2013. **Ruolo: Responsabile Scientifico**. Durata 1 mese.
14) Progetto di ricerca: **"Sistema di gestione delle telecamere di guida a bordo del Rover SETEL"** per conto della Se.Te.L. group di Roma, 2020-2021. **Ruolo: Responsabile Scientifico**. Durata 7 mesi.