

UNIVERSITÀ DEGLI STUDI DI PISA

SCUOLA DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE CORSO
DI LAUREA TRIENNALE IN INGEGNERIA INFORMATICA



UNIVERSITÀ DI PISA

Isolation forest per la ricerca di
comportamenti anomali nei flussi di mobilità
urbana

Anno Accademico

2019/2020

Relatori

Prof. Mario G.C.A. Cimino
Prof.ssa Gigliola Vaglini
Dott. Antonio Luca Alfeo

Candidato

Salvatore Bonanno

Sommario

1 Introduzione	3
1.1 Motivazioni	3
1.2 Related work.....	4
2 Design	
2.1 Isolation Forest.....	5
2.2 I casi d'uso.....	6
2.2.1 FindAnomalies.....	7
2.2.2 VisualizeTimeSeriesBehaviorLabeling.....	7
2.2.3 VisualizeTimeseriesCluster.....	7
3 Le funzioni.....	8
3 Dati	
3.1 Casi di studio.....	9
4 Risultati sperimentalni	
4.1 Plotting time series e analisi visiva.....	11
4.2 Plotting time series trovate con Isolation Forest e analisi visiva.....	16
4.2.1 Analisi anomalie trovate da isolation forest sul cluster 0.....	19
4.2.2 Analisi anomalie trovate da isolation forest sul cluster 1.....	22
4.2.3 Analisi anomalie trovate da isolation forest sul cluster 2.....	26
5 Risultati, conclusioni e ringraziamenti	
5.1 Risultati.....	29
5.2 Conclusioni.....	29
5.3 Ringraziamenti.....	30
Appendice	
A .Dettagli di implementazione	31
B.Codice	33
Bibliografia e sitografia.....	43

Abstract

I dati di posizionamento offrono una notevole fonte di informazioni per analizzare le dinamiche urbane delle folle. In questa tesi è stato effettuato un lavoro di ricerca all'interno di un dataset contenente i dati dell'affluenza e i giorni anomali del 2015 all'interno di un'area ad alta intensità di Manhattan detta Hotspot. I giorni della settimana sono divisi per gruppi, in base al loro comportamento. Il software sviluppato consente di tracciare analizzare le time series giornaliere suddivise per comportamento e anomalie e di applicare un algoritmo di Machine Learning per rilevare anomalie.

CAPITOLO 1

Introduzione

1.1 MOTIVAZIONI

Al giorno d'oggi, l'uso di tecnologie pervasive genera una grande quantità di tracce digitali associate ad ogni attività umana. Esempi ben noti sono i post sui social media, le tracce GPS dei veicoli, i registri dei cellulari, le transazioni con carta di credito e così via [1]. In tutti questi contesti è utile ed è necessario estrarre conoscenze dai metadati generati dall'attività umana, contenenti informazioni collaterali su eventi correlati: le informazioni estratte possono dare una visione del contesto, altrimenti difficili da ricavare, e permettono inoltre di risolvere i problemi correlati. Uno scenario particolarmente attuale è rappresentato dalle Smart City, nuovo concetto di città moderna "intelligente", nella quale grazie all'impiego di reti di grandi dimensioni, spesso composte da dispositivi ed infrastrutture eterogenee, è possibile monitorare e interagire con fenomeni quali il consumo energetico, lo stato di sicurezza di strutture storiche e edifici pubblici, nonché la sorveglianza dello stato di alcuni ambienti critici. Le smart City hanno come principale obiettivo quello di produrre sia conoscenze teoriche e pratiche per contribuire alla comprensione e alla risoluzione delle problematiche urbane nella società contemporanea [23]. In questo caso, il problema affrontato da questa tesi consiste nell'individuazione di anomalie all'interno di un dataset contenente l'affluenza giornaliera suddivisa per fasce orarie all'interno dell'hotspot D di Manhattan. Tale hotspot include i quartieri Murray Hill, East Village e Gramercy. In questo lavoro, dobbiamo supportare i responsabili dell'urbanistica di New York nel fornire un servizio efficiente di analisi di anomalie del traffico. L'obiettivo è quello di individuare i giorni che presentano dati di affluenza non regolare e, una volta individuate, andare a cercare le motivazioni che hanno causato tali anomalie. Le irregolarità all'interno dell'hotspot possono dipendere da più cause: restrizioni del traffico, manifestazioni, festività o anche fenomeni metereologici.



Figura 1: Taxi in movimento

1.2 RELATED WORKS

La rilevazione di anomalie permette di estrapolare informazioni riguardanti fenomeni inaspettati, che deviino dal comportamento normale, come incidenti o congestioni dovute a qualche particolare evento. In letteratura si possono trovare diverse tecniche ed algoritmi che investigano tali scenari. In passato, il rilevamento manuale delle anomalie era un'opzione praticabile. I set di dati erano abbastanza gestibili per un gruppo di analisi. Oggi giorno, dato che si ha a che fare con dataset enormi si ricorre spesso ad algoritmi di machine learning hanno la capacità di apprendere dai dati e fare previsioni basate su tali dati e inoltre offrono il notevole vantaggio di non richiedere troppa conoscenza del dominio per trovare le anomalie perché creano un modello di normalità e anomalia direttamente dai dati. Esistono tre grandi categorie di tecniche di rilevamento di anomalie basate su machine learning:

-Rilevamento delle anomalie non supervisionato: tecniche di rilevamento delle anomalie nei dati di test un set non etichettati in base al presupposto che la maggior parte delle istanze nel set di dati sono normali, cercando per le istanze che sembrano adattarsi almeno per il resto del set di dati.

-Rilevamento delle anomalie supervisionato: tecniche che richiedono un insieme di dati che è stato etichettato come "normale" e "anormale" e coinvolge la formazione di un classificatore.

-Rilevamento delle anomalie semi-supervisionato tecniche che costruiscono un modello, che rappresenta il comportamento normale, a partire da un normale set di dati di training, e quindi e poi testano la probabilità che un'istanza di test sia generata dal modello appreso. Diverse tecniche di rilevamento delle anomalie sono state proposte in letteratura. Alcune delle tecniche popolari sono:

- Tecniche Density-based (k-nearest neighbour , local outlier factor , isolation forest).
- One class support vector machines.
- Reti neurali artificiali.
- Clustering.

Le prestazioni dei diversi metodi dipendono molto dal set di dati e dai parametri [2].

Vi sono dei problemi da affrontare quando si lavora con le anomalie [3]:

- Definire una regione normale che comprenda ogni possibile comportamento normale è molto difficile. Il confine tra normale e anomalo spesso non è preciso. Così un'osservazione anomala che si trova vicino al confine può sia effettivamente normale, e viceversa.
- l'identificazione di un comportamento che può essere considerato normale è sempre facile, può essere anche impraticabile in quanto non vi è la possibilità di osservare ogni istanza ordinaria appartenente alla popolazione
- La nozione esatta di un'anomalia è diversa per i diversi campi di applicazione. Per esempio, in ambito medico una piccola deviazione dalla norma (ad es. fluttuazioni nella temperatura corporea) potrebbe essere un'anomalia, mentre una simile deviazione nello stock dominio di mercato (ad esempio, le fluttuazioni del valore di un'azione) potrebbero essere considerate come normale. Così l'applicazione di una tecnica sviluppata in un dominio ad un altro, non è semplice.
- Lavorare con dati etichettati (scenario supervisionato) risulta spesso impossibile o troppo oneroso. Perciò risulta necessario in alcune situazioni lavorare con dati non etichettati (scenario non supervisionato) facendo l'implicita ipotesi che le anomalie si verifichino con frequenza molto più bassa rispetto ai comportamenti normali, permettendo in questo modo l'impiego di strumenti statistici per identificarle

Durante questo progetto di tesi il problema della rilevazione di anomalie nel traffico è stato affrontato tramite l'utilizzo di Isolation Forest, un algoritmo di apprendimento non supervisionato per la rilevazione delle anomalie che funziona sul principio dell'isolamento delle anomalie, invece delle più comuni tecniche di profilazione dei punti normali.

CAPITOLO 2

DESIGN

2.1 ISOLATION FOREST

Le tecniche più comuni utilizzate per la rilevazione delle anomalie si basano sulla costruzione di un profilo di ciò che è "normale": le anomalie sono segnalate come quelle istanze del dataset che non sono conformi al profilo normale.

Isolation forest è un algoritmo di machine learning per il rilevamento delle anomalie. Utilizza un approccio diverso: invece di cercare di costruire un modello di istanze normali, isola esplicitamente i punti anomali nel dataset. Il vantaggio principale di questo approccio è la possibilità di sfruttare le tecniche di campionamento, creando un algoritmo molto veloce e con una bassa richiesta di memoria [4]. I metodi avanzati di rilevamento delle anomalie, come Isolation Forest, sono molto efficienti, perché questo metodo rileva le anomalie esclusivamente sulla base del concetto di isolamento senza utilizzare alcuna misura di distanza o densità - fondamentalmente diverso da tutti i metodi esistenti. Isolation Forest si basa sull'algoritmo Decision Tree. Un albero di decisione è un modello predittivo, dove ogni nodo interno rappresenta una variabile, un arco verso un nodo figlio rappresenta un possibile valore per quella proprietà e una foglia rappresenta il valore predetto per la variabile obiettivo a partire dai valori delle altre proprietà, che nell'albero è rappresentato dal cammino dal nodo radice al nodo foglia. Isolation Forest genera N alberi, dove N è un valore che può essere scelto e isola i valori anomali selezionando casualmente una caratteristica da un dato insieme di caratteristiche e poi selezionando casualmente un valore di divisione tra i valori max e min di quella caratteristica. Il punteggio di anomalia di un campione di input è calcolato come il punteggio medio di anomalia degli alberi nella foresta. La soglia di contaminazione può essere determinata in automatico dal dataset oppure settata manualmente e dovrebbe essere compresa tra 0 e 0.5. La misura della normalità di un'osservazione data ad un albero è la profondità della foglia che contiene una determinata osservazione, che è equivalente al numero di frazionamenti necessari per isolare questo punto. Questa suddivisione casuale delle caratteristiche produrrà percorsi più brevi negli alberi per i punti di dati anomali, distinguendoli così dal resto dei dati. La profondità massima di ogni albero è settata alla parte intera superiore del logaritmo in base di N dove N è il numero di alberi scelti.

In genere, il primo passo verso il rilevamento delle anomalie è quello di costruire un profilo di ciò che è "normale", e poi riportare tutto ciò che non può essere considerato normale come anomalo. Tuttavia, l'algoritmo di isolation forest non funziona su questo principio; non definisce prima il comportamento "normale" e non calcola le distanze basate sui punti. Come ci si potrebbe aspettare dal nome, Isolation Forest funziona invece isolando le anomalie.

L'algoritmo di Isolation Forest si basa sul principio che le anomalie sono osservazioni poche e diverse, che dovrebbero renderle più facili da identificare. Isolation Forest utilizza un insieme di alberi di isolamento per i punti dati per isolare le anomalie [5].

Il rilevamento dell'anomalia con Isolation Forest è un processo composto da due fasi principali:

- 1- Nella prima fase, un set di dati di formazione viene utilizzato per costruire gli alberi di isolamento come descritto nelle sezioni precedenti.
- 2- Nella seconda fase, ogni istanza nel set viene passata agli alberi di isolamento costruiti nella fase precedente, e un adeguato "punteggio di anomalia" viene assegnato all'istanza utilizzando l'algoritmo descritto di seguito

Una volta che a tutte le istanze del set di test è stato assegnato un punteggio di anomalia, è possibile contrassegnare come "anomalia" qualsiasi punto il cui punteggio sia superiore ad una soglia predefinita [3].

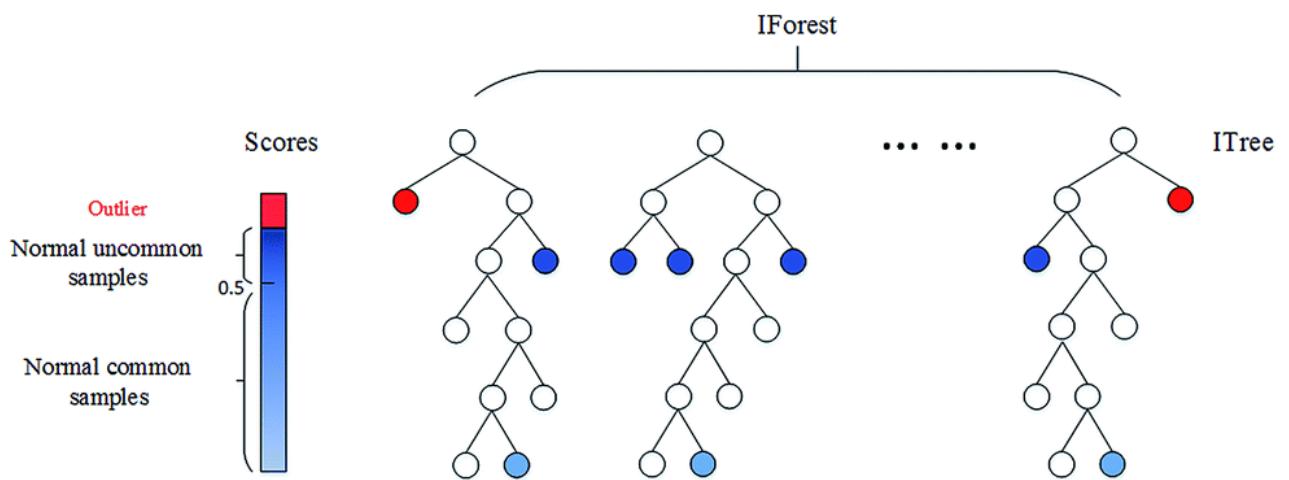


Figura 2: Isolation Forest

2.2 I CASI D'USO

Il software implementato in python, composto da 3 principali casi d'uso.

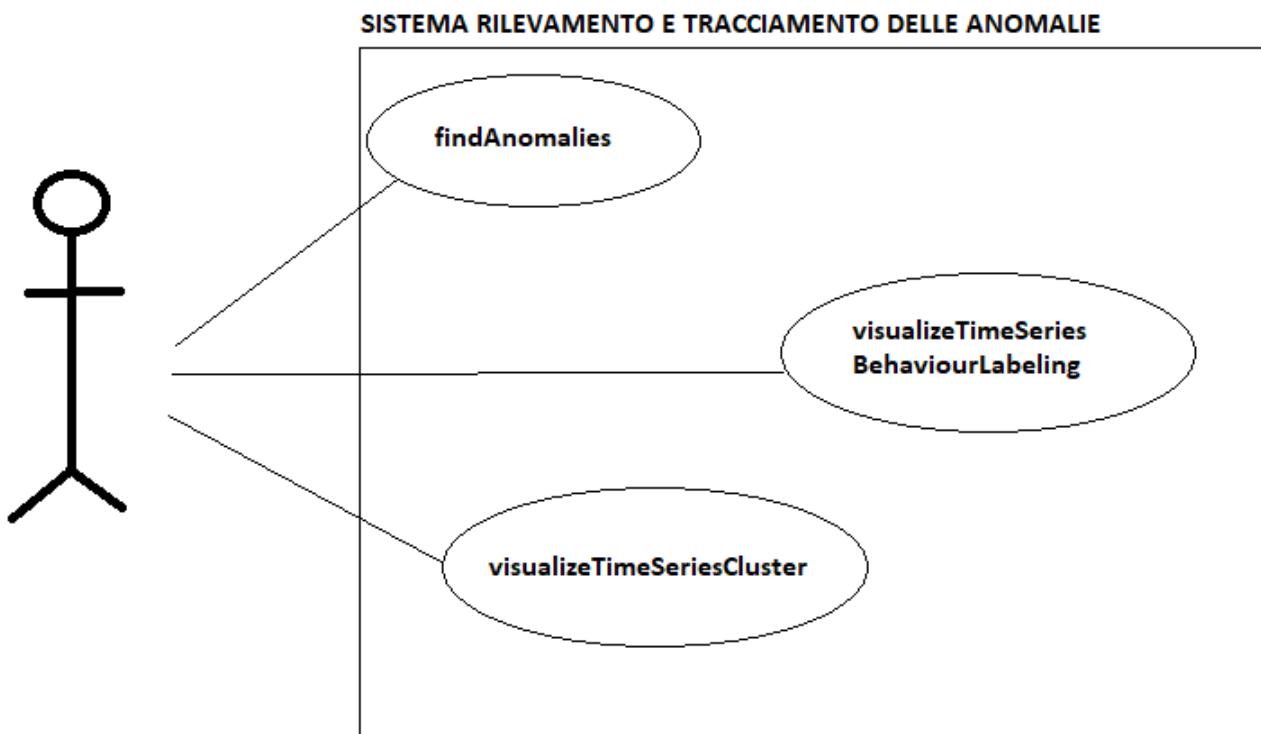


Figura 3 Casi d'uso per il sistema sviluppato

2.2.1 FindAnomalies

Il caso d'uso inizia quando l'utente vuole vedere i risultati che scaturiscono applicando Isolation Forest ad un dataset. L'utente può applicare isolation forest effettuando l'addestramento del modello con parametri di default, oppure può addestrare isolation forest con dei parametri desiderati. Il sistema determina quali sono i valori anomali, li aggiunge ad dataset. Gli outcome sono: creazione di un dataset a cui viene aggiunta una nuova label per identificare le righe del dataset che isolation forest ha riconosciuto come anomale, plot di tali timeseries e della time series media in modo da consentire all'utente di poter effettuare delle analisi. Il file creato da isolation forest con parametri di default ha formato csv e nome "isolation-forest-cluster-x.csv" dove x è uguale a 0, 1 o 2. Il file creato da isolation forest con parametri scelti dall'utente ha formato csv e nome "isolation-forest-cluster(cluster)-estim(n_estimators)-sampl(max_samples)-cont(contamination)-boot(bootstrap)-jobs(n_jobs)-randstate(random_state).csv" dove i valori tra parentesi sono i parametri inseriti. Entrambe le funzioni producono delle immagini con estensione png, cambia il nome delle immagini che è uguale a quello dei file csv creati. Infine viene creato un file con estensione txt in cui vengono inserite le performances di isolation forest. Il file ha nome "prestazion-isolation-forest-cluster-x.txt" dove x è uguale a 0, 1 o 2 nel caso di isolation forest con parametri di default. Nel caso di parametri inseriti dall'utente esso ha nome "prestazioni-isolation-forest-cluster(cluster)-estim(n_estimators)-sampl(max_samples)-cont(contamination)-boot(bootstrap)-jobs(n_jobs)-randstate(random_state).txt". I file vengono salvati nella directory principale.

2.2.2 VisualizeTimeSeriesBehaviorLabeling

Il caso d'uso inizia quando l'utente vuole vedere il grafico di una time series di un cluster e della timeseries media dello stesso cluster in modo da poter cogliere le differenze tra le due. Il sistema seleziona i valori dal dataset. L'utente, effettuando un'analisi visiva, può stabilire o meno se quella time series ha un comportamento anomalo e aggiungere una label al dataset oppure può decidere di visualizzare le time series in cui l'affluenza è minore o uguale ad un determinato valore in una fascia oraria selezionata dall'utente. Gli outcome sono: plot delle due time series, aggiunta e modifica di labels e valori del dataset. Le immagini prodotte da queste due funzioni hanno estensione png e titoli: "daily-timeseries-(giorno)-(mese)-2015.png" e "time-series-cluster(cluster)-fasciaoraria(fascia)-less-value(value).png", dove i valori tra parentesi sono i parametri inseriti. L'utente se aggiunge delle anomalie visive può salvare il dataset per riusarlo successivamente. I file vengono salvati nella directory principale.

2.2.3 VisualizeTimeseriesCluster

Il caso d'uso inizia quando l'utente vuole vedere il grafico delle time series di un cluster divise per anomale e non anomale. Il sistema seleziona i valori dal dataset. L'utente, effettuando un'analisi visiva, può vedere le principali differenza tra le time series che sono state riconosciute come anomale e quelle non anomale. Gli outcome sono: plot delle time series di un cluster suddivise per anomale e non. Le immagini prodotte da queste due funzioni hanno estensione png e titoli : "plot-timeseries-cluster-(x).png" e "plot-timeseriesvisive-cluster-(x).png", dove x è uguale a 0,1 o 2. I file vengono salvati nella directory principale.

2.3 LE FUNZIONI

Il caso d'uso findAnomalies usa la funzione Isolation Forest importata dalla libreria sklearn.ensemble. Detto X il set dei dati alla funzione IsolationForest è possibile passare i seguenti parametri:

- n_estimators int, default = 10: il numero di stimatori di base nell'insieme.
- max_samples "auto", int o float, default = " auto". Il numero di campioni da prelevare da X per addestrare ciascuno stimatore di base.
- contamination 'auto' o float, default = 'auto'. La quantità di contaminazione del set di dati, ovvero la percentuale di valori anomali nel set di dati. Utilizzato durante la fase di fit per definire la soglia sui punteggi dei campioni. Se 'auto', la soglia è determinata come nel documento originale. Se float, la contaminazione deve essere compresa nell'intervallo [0, 0,5].
- max_features int o float, default = 1.0. Il numero di funzioni da attingere da X per addestrare ciascuno stimatore di base.
- bootstrap bool, default = False. Se True, i singoli alberi si adattano a sottoinsiemi casuali dei dati di addestramento campionati con la sostituzione. Se False, viene eseguito il campionamento senza sostituzione.
- random_state int o RandomState, default = None. Controlla la pseudo-casualità della selezione dell'elemento e dei valori di divisione per ogni passaggio di ramificazione e per ogni albero nella foresta.
- n_jobs int, default = None. Il numero di jobs da eseguire in parallelo.

Il caso d'uso VisualizeTimeSeriesBehaviorLabeling chiama le funzioni plot_daily_timeseries, add_visual_anomalies e plot_timeseries_with_h_value. I parametri della prima e della seconda sono:

- giorno: intero compreso da 1 e 31, nel caso i mesi di interesse siano Gennaio, Marzo, Maggio, Luglio, Agosto, Ottobre e Dicembre oppure un intero compreso tra 1 30 nel caso i mesi di interesse siano Aprile, Giugno, Settembre, Novembre oppure un intero compreso tra 1 e 28 nel caso il mese di interesse sia Febbraio.
- mese: intero compreso tra 1 e 12.

I parametri della terza funzione sono:

- cluster: intero che vale 0,1 o 2. Con 0 si indicano le time series appartenenti ai giorni con comportamenti lavorativi, con 1 quelle appartenenti ai giorni con comportamenti festivi e con 2 quelle appartenenti ai lazy days.
- Fascia: un intero compreso tra 1 e 23 che indica la fascia oraria desiderata.
- Valore: un float compreso tra 0 e 1 che indica l'affluenza.

Il caso d'uso VisualizeTimeSeriesCluster chiama le funzioni plot_timeseries_by_anomalous_and_cluster e plot_timeseries_by_visiveanom_and_cluster. L'unico parametro delle seguenti funzioni è:

- Cluster: che vale 0,1 o 2. Con 0 si indicano le time series appartenenti ai giorni con comportamenti lavorativi, con 1 quelle appartenenti ai giorni con comportamenti festivi e con 2 quelle appartenenti ai lazy days.

Per maggiori dettagli sulle funzioni consultare l'appendice A.

CAPITOLO 3

DATI

3.1 CASI DI STUDIO

Il dataset usato nella tesi contiene informazioni sull'affluenza giornaliera dei taxi all'interno dell'Hotspot D di Manhattan nell'anno 2015. L'hotspot D copre i quartieri di Murray Hill, Gramercy e East Village. Contiene i seguenti attributi:

CLUSTER ANOMALOUS DAY MONTH H1-H2-H3-H4-H5-H6-H7-H8-H9-H10-H11-H12-H13-H14-H15-H16-H17-H18-H19-H20-H21-H22-H23

I giorni della settimana sono stati divisi in gruppi. Vengono indicati con:

- Cluster zero i giorni con comportamenti lavorativi, nonché i giorni che vanno dal lunedì al giovedì dove i movimenti della folla sono per lo più causati dalle routine lavorative.
- Cluster uno i giorni con comportamenti festivi, nonché i venerdì e i sabati dove la gente tende a spendere la serata fuori.
- Cluster due i lazy days, nonché le domeniche dove si ha uno scarso utilizzo dei trasporti.

L'etichetta anomalous con valori binari, zero o uno, indica se nel suddetto giorno è stata riscontrata un'anomalia.

Le etichette day e month indicano il giorno e il mese del 2015 analizzati.

Le etichette che vanno da h1 ad h23 contengono l'affluenza all'interno della specifica fascia oraria. Tale valore è compreso tra 0 e 1, dove zero indica affluenza minima e 1 affluenza massima. Le time series medie dei cluster, mostrate nelle figure sottostanti, sono:

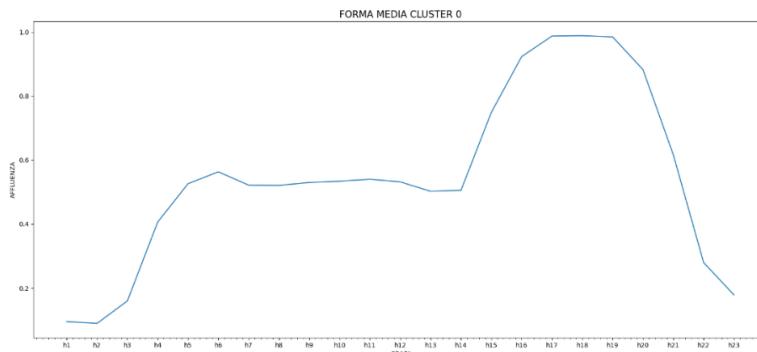


Figura 4 Time Series media del cluster zero

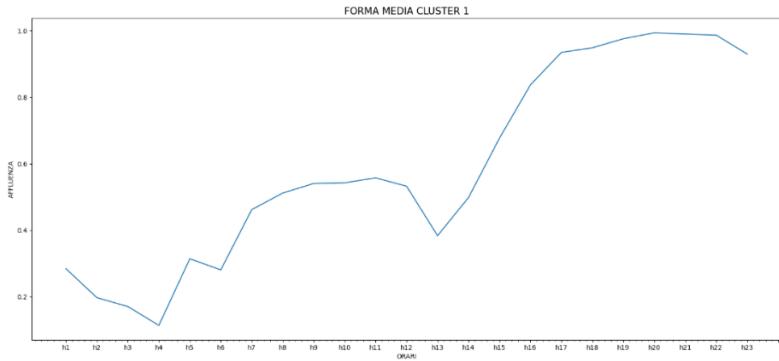


Figura 5: Time Series media del cluster uno

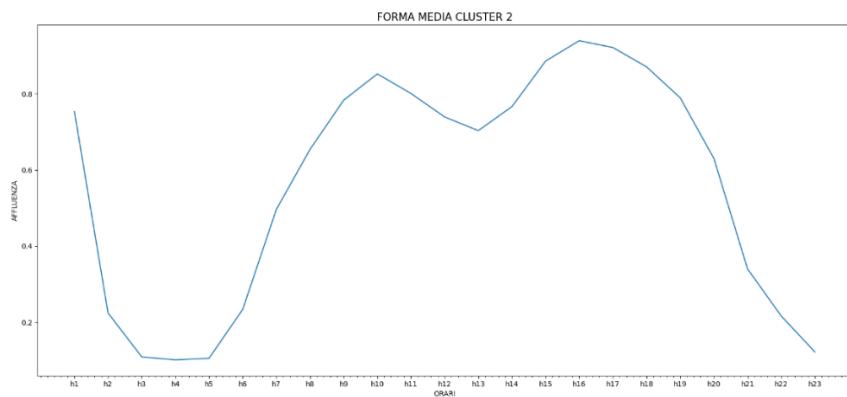


Figura 6: Time Series media del cluster due

Segue una descrizione delle time series medie indicando con valori alti le parti del grafico in cui l'affluenza è maggiore di 0.75 , con valori medi le parti del grafico in cui l'affluenza è compresa tra 0.4 e 0.75 e con valori bassi le parti in cui l'affluenza è compresa tra 0 e 0.4. Indico con parte iniziale la parte della curva che va da h1 a h8, con parte centrale quella che va da h8 a h16 e con finale quella che va da h16 a h23. Indico con il picco in alto il punto in cui viene raggiunta l'affluenza massima e con picco in basso il punto in cui viene raggiunta l'affluenza minima.

Riguardo alla forma media del cluster zero, il picco in alto viene raggiunto nelle fasce h17-h18 e trattandosi di giorni lavorativi corrisponde all'orario di uscita da lavoro. La curva si mantiene su valori medi nella parte centrale e nelle ultime ore della giornata si può notare questo calo di affluenza spiegato dal fatto che l'indomani sarà un giorno lavorativo.

La forma media del cluster uno si mantiene su valori bassi all'inizio della giornata, va su valori medi nella parte centrale e raggiunge il picco in alto nelle fasce h21-h22. Questo è dovuto al fatto che molta gente non lavora il sabato e la domenica e quindi ne approfitta per trascorrere la serata fuori.

La forma media del cluster due inizia decrescendo. Questo andamento può essere motivato dal ritorno a casa delle persone dopo aver trascorso la serata fuori il sabato. Nella parte centrale si può notare un grande innalzamento della curva con raggiungimento del picco in alto nelle fasce h16-h17, poiché trattandosi delle domeniche ed essendo dei giorni liberi le persone si riversano in strada per fare una passeggiata o per fare shopping. Nella parte finale si nota questo decrescere della curva dovuto al fatto che il giorno dopo sarà lunedì e quindi un giorno lavorativo.

CAPITOLO 4

RISULTATI Sperimentali

4.1 PLOTTING TIME SERIES E ANALISI VISIVA

La generazione dei plot è stata effettuata tramite uno script in linguaggio Python e tramite l'uso di librerie di supporto quali PANDAS e MATPLOTLIB.PYPLOT. Gli step per il tracciamento dei grafici sono i seguenti:

- Apertura del dataset (file con estensione csv) in lettura
- Raggruppamento dei dati per coppia CLUSTER- ANOMALOUS con le funzioni LOC and ILOC
- Per ogni riga del dataset, selezione dei valori compresi tra h00-h23
- Tracciamento della curva per tali valori mediante la funzione plot ()

Quello che si ottiene sono le seguenti tre figure:

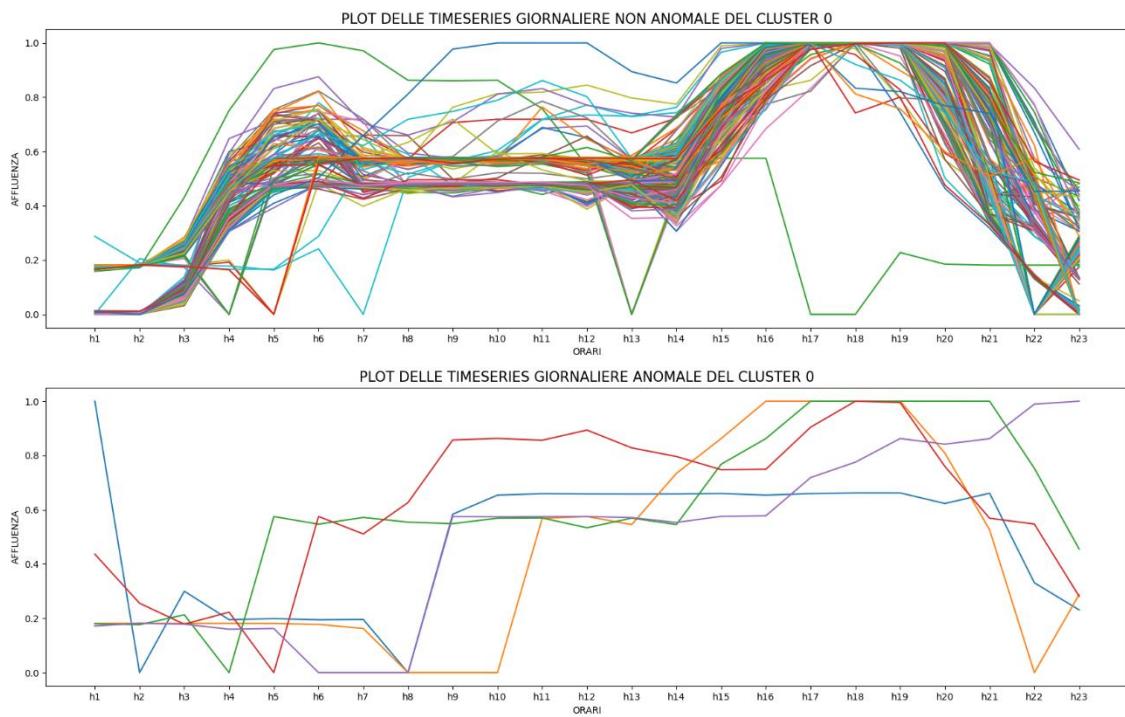


Figura 7: Il primo grafico contiene la rappresentazione grafica delle time series giornaliere del cluster zero che sono state riconosciute regolari. Il secondo grafico contiene la rappresentazione grafica delle time series giornaliere del cluster zero che hanno delle irregolarità. Il cluster zero contiene tutte le time series relative ai giorni con comportamenti lavorativi, che vanno dal lunedì al giovedì. Il valore rappresentato sull'asse delle ascisse rappresenta la fascia oraria in cui è stata effettuata la misurazione. L'asse delle ordinate contiene il valore della misurazione, compreso tra 0 e 1, cioè l'affluenza all'interno dell'hotspot D di manhattan in una specifica fascia oraria.

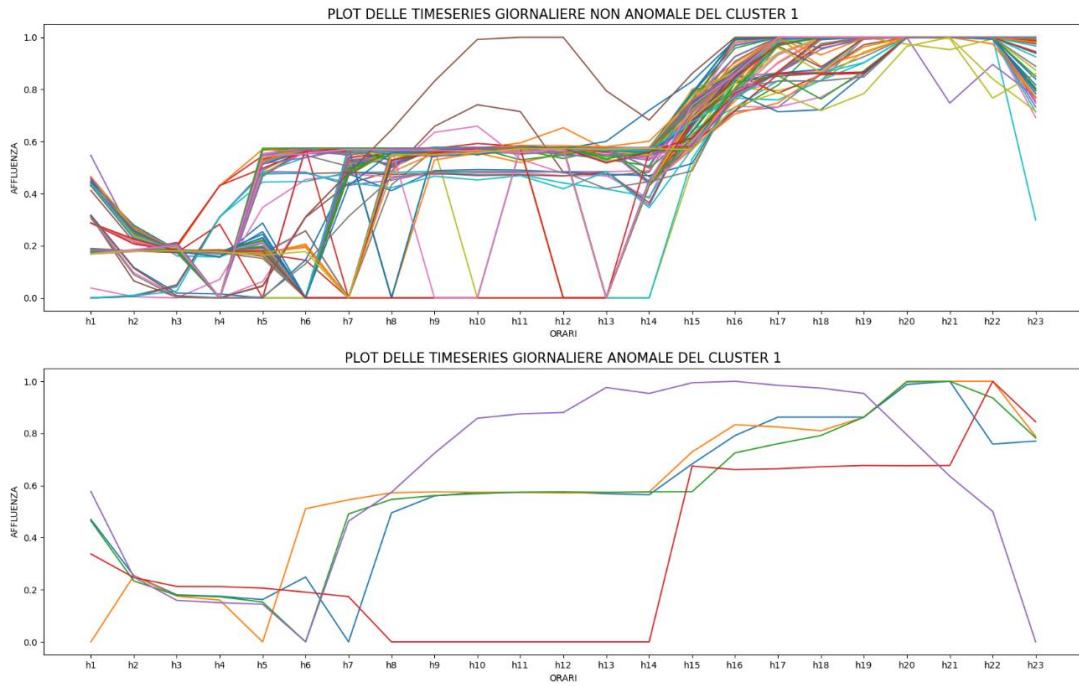


Figura 8: Il primo grafico contiene la rappresentazione grafica delle time series giornaliere del cluster uno che sono state riconosciute regolari. Il secondo grafico contiene la rappresentazione grafica delle time series giornaliere del cluster uno che hanno delle irregolarità. Il cluster uno contiene tutte le time series relative ai giorni con comportamenti festivi, cioè venerdì e sabato. Il valore rappresentato sull'asse delle ascisse rappresenta la fascia oraria in cui è stata effettuata la misurazione. L'asse delle ordinate contiene il valore della misurazione, compreso tra 0 e 1, cioè l'affluenza all'interno dell'hotspot D di manhattan in una specifica fascia oraria.

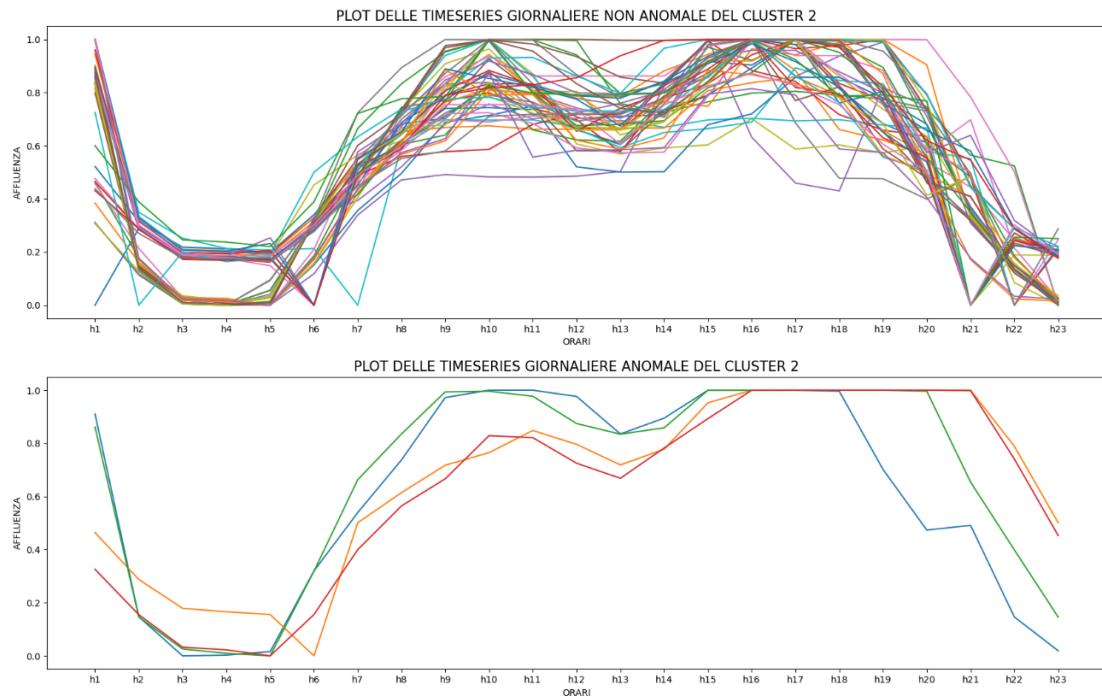


Figura 9 :Il primo grafico contiene la rappresentazione grafica delle time series giornaliere del cluster due che sono state riconosciute regolari. Il secondo grafico contiene la rappresentazione grafica delle time series giornaliere del cluster due che hanno delle irregolarità. Il cluster zero contiene tutte le time series relative ai giorni di svago, corrispondenti alle domeniche. Il valore rappresentato sull'asse delle ascisse rappresenta la fascia oraria in cui è stata effettuata la misurazione. L'asse delle ordinate contiene il valore della misurazione, compreso tra 0 e 1, cioè l'affluenza all'interno dell'hotspot D di manhattan in una specifica fascia oraria.

Dati i seguenti grafici possiamo notare visivamente alcuni comportamenti anomali presenti tra quelli che non lo sono, come nel caso della figura 3. Riguardo al primo grafico contenuto nella Figura1, quello corrispondente ai comportamenti lavorativi non anomali, possiamo scorgere alcune curve che pur appartendendo ai casi non anomali hanno un andamento non regolare:

- La curva verde si differenzia dalle altre dello stesso cluster in quanto nella fascia oraria h5 si trova più in alto dal fascio di curve più denso e ha un picco in basso nelle ore 17-18 invece di averlo in alto.
- La curva blu , a differenza delle altre che si trovano centralmente nel grafico, si discosta in alto.
- La curva azzurra , a differenza delle altre che si trovano centralmente nel grafico, si discosta in basso(in h7).
- Il fascio di curve che in h13, a differenza delle altre che si trovano centralmente nel grafico, si discostano in basso

Relativamente al primo grafico di Figura 2, cluster dei giorni festivi con comportamento non anomalo, a prima vista si possono notare due curve che hanno un andamento diverso:

-La curva marrone che presenta un picco in alto nella fascia oraria compresa tra le ore 10 e le ore 13 invece di trovarsi al centro come le altre.

-La curva celeste che si abbassa vistosamente verso il basso alle ore 23.

Di seguito sono riportati i grafici in cui tali curve vengono isolate dal resto delle altre per una migliore resa visiva (in ordine):

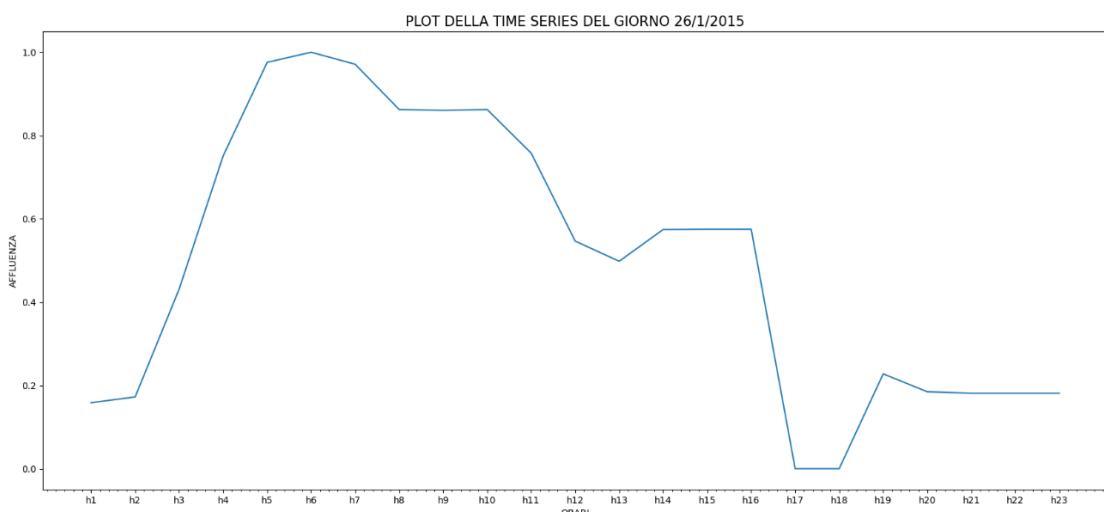


Figura 10: Plot della time series del giorno 26 gennaio 2015. A differenza delle time series dello stesso cluster, questa presenta dei valori di affluenza molto alti nella parte iniziale (dopo h4), raggiunge il picco in alto in h5 mentre la maggior parte delle time series del cluster zero raggiunge il picco in alto nelle ultime fasce orarie della giornata.

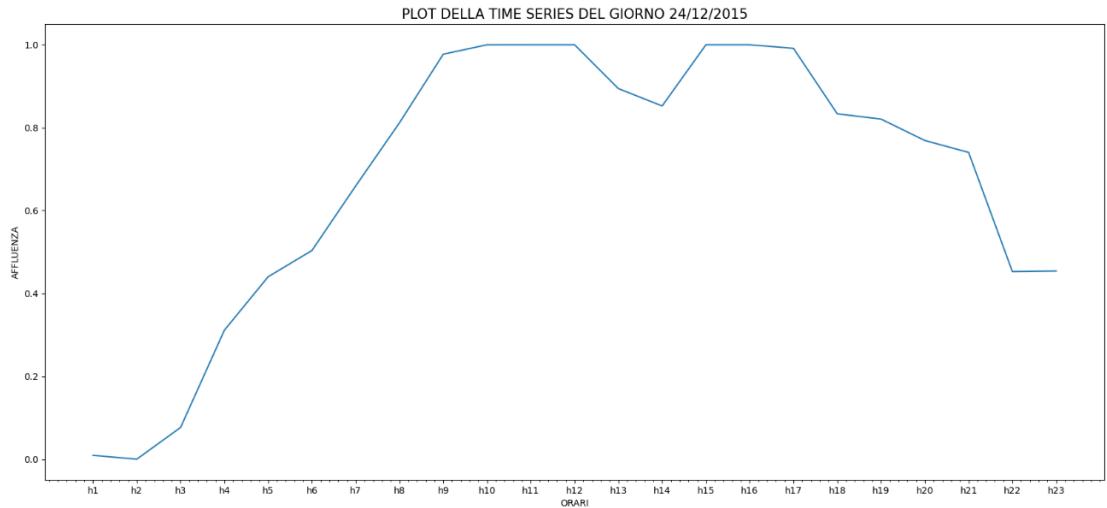


Figura 11 Plot della time series del giorno 24 dicembre 2015. A differenza delle time series dello stesso cluster, questa presenta dei valori di affluenza molto alti nella zona centrale del grafico, raggiungendo il picco in alto in h9 mentre la maggior parte delle time series del cluster zero si mantengono su valori medi nella parte centrale.

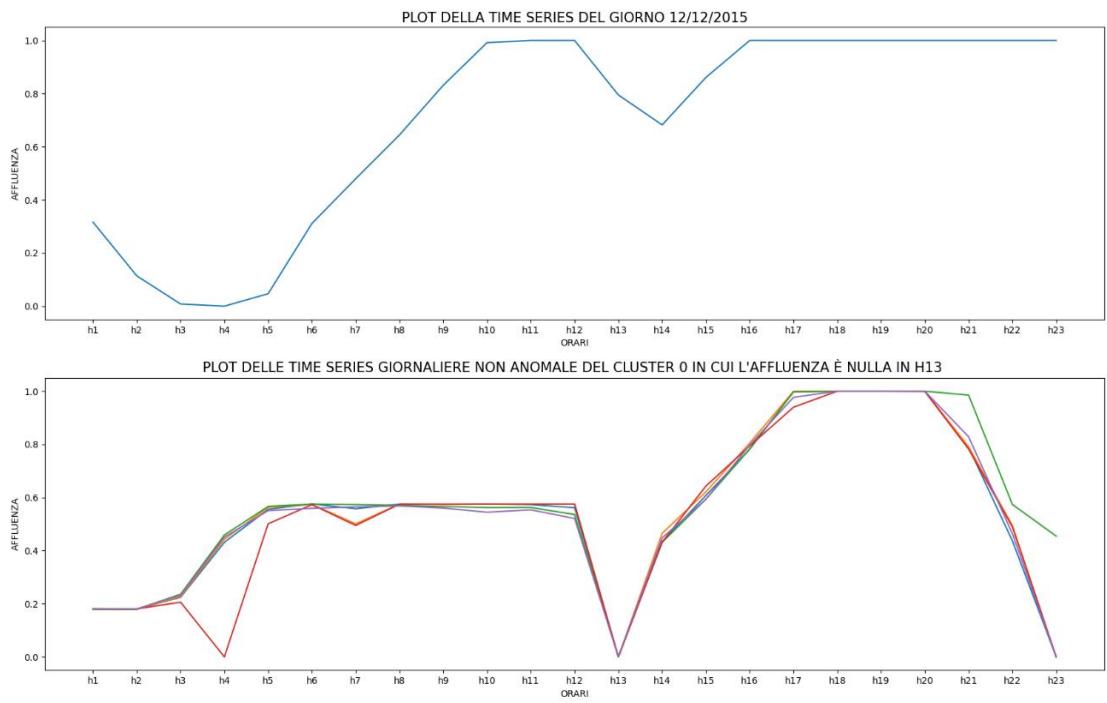


Figura 12: il primo grafico contiene la rappresentazione della time series del giorno 25 maggio 2015. A differenza delle time series dello stesso cluster, ha questa forma a "V" in h7 dove raggiunge il picco in basso e dopo il quale torna ad avere un comportamento simile alla maggior parte delle curve dello stesso cluster. Il secondo grafico rappresenta le time series che si differenziano da quelle dello stesso cluster

in quanto hanno andamento a "V", e quindi un picco in basso, in h13.

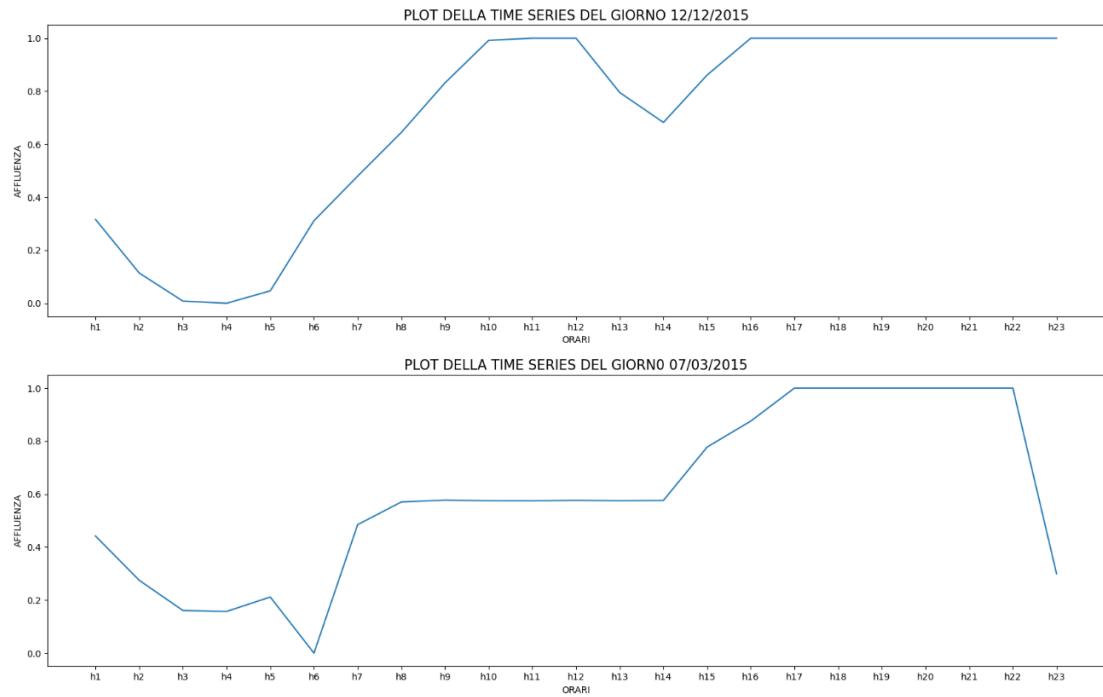


Figura 13 : il primo grafico contiene la rappresentazione della time series del giorno 12 dicembre 2015. A differenza delle time series dello stesso cluster, raggiunge il picco in alto tra h10 e 12 e dopo il quale torna ad avere un comportamento simile alla maggior parte delle curve dello stesso cluster. Il secondo grafico rappresenta la time series del giorno 7 marzo 2015 e a differenze delle curve dello stesso cluster nella parte finale scende molto rapidamente

Andando ad effettuare una ricerca sul file dei giorni e una ricerca sul web degli eventi verificatisi in tali giorni quello che si ottiene è riassunto dalla seguente tabella:

GIORNO	MESE	CLUSTER	FASCIA	FIGURA	GRAFICO	MOTIVAZIONI
26	01	0	H5	4	1	Tempesta di neve Juno [6]
24	12	0	H9-H14	4	2	Vigilia di natale
25	05	0	H7	5	1	Memorial day
19	03	0	H13	5	2	Tempesta di neve [7]
16	04	0	H13	5	2	Tribeca film festival [8]
14	05	0	H13	5	2	Falun dafa Parade [9]
9	07	0	H13	5	2	IGNOTA
10	9	0	H13	5	2	Inizio festa San gennaro [10]
12	12	1	H10-13	6	1	Santacon parade [11]
7	03	1	H23	6	2	IGNOTA

Ciò che possiamo dire ,dopo questa breve analisi visiva, è che effettivamente ci sono degli eventi che hanno portato a delle anomalie ma quest'ultime compaiono nei cluster con comportamenti regolari. Segue l'applicazione dell'algoritmo di Isolation Forest per trovare le anomalie e vedere se quelle trovate visivamente hanno conferma.

4.2 PLOTTING TIME SERIES TROVATE CON ISOLATION FOREST E ANALISI VISIVA

Il passo successivo consiste nel analizzare i dati presenti nel Dataset attraverso Isolation Forest. Lo scopo è vedere le differenze principali tra le anomalie trovate tramite isolation forest e quelle presenti nel dataset, trovare i giorni di tali anomalie, cercare una motivazione e descrivere le curve in tali giorni confrontandole con quella della forma media.

La generazione dei plot è stata effettuata tramite uno script in linguaggio Python e tramite l'uso di librerie di supporto quali PANDAS, MATPLOTLIB.PYPLOT e SKLEARN.ENSEMBLE. Gli step per il tracciamento dei grafici sono i seguenti:

- Apertura del dataset (file con estensione csv) in lettura
- Raggruppamento dei dati per CLUSTER con la funzione LOC
- Definire il modello model=IsolationForest (n_estimators=100, max_samples='auto', contamination='auto', max_features=1.0, bootstrap=False, n_jobs=-1, random_state =0)
- Effettuare il training del modello mediante i dati raggruppati mediante la funzione fit ()
- Otteniamo i valori delle colonne anomale chiamando la funzione predict () del modello di cui è stato effettuato il training. Quello che si ottiene sono i seguenti grafici:

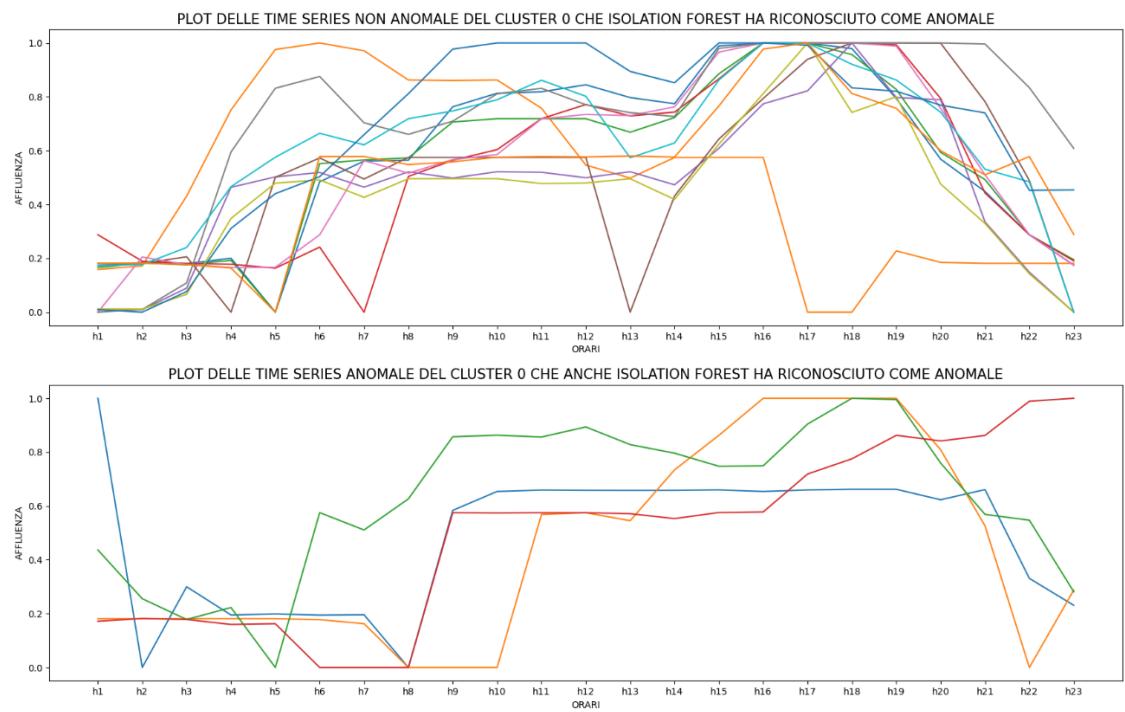


Figura 14: il primo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster zero con anomalous uguale a zero. Il secondo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster zero con anomalous uguale a uno. Si può notare, confrontando questi grafici con quelli di figura 1, che Isolation forest ha riconosciuto come anomale delle time series che erano riconosciute come regolari e viceversa. Per esempio, osservando il secondo grafico di figura 2 e confrontandolo col secondo grafico di questa figura si può notare che in quest'ultimo manca una time series.

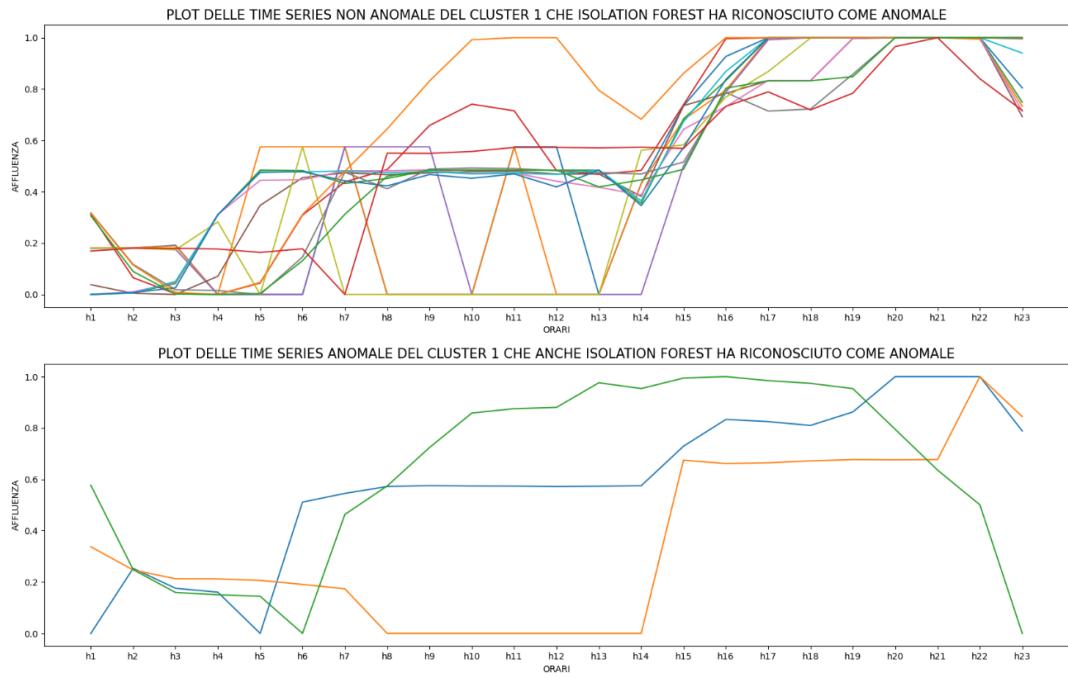


Figura 15: il primo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster uno con anomalous uguale a zero. Il secondo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster uno con anomalous uguale a uno. Si può notare, confrontando questi grafici con quelli di figura 2, che Isolation forest ha riconosciuto come anomale delle time series che erano riconosciute come regolari e viceversa. Per esempio, osservando il secondo grafico di figura 2 e confrontandolo col secondo grafico di questa figura si può notare che in quest'ultimo mancano due time series.

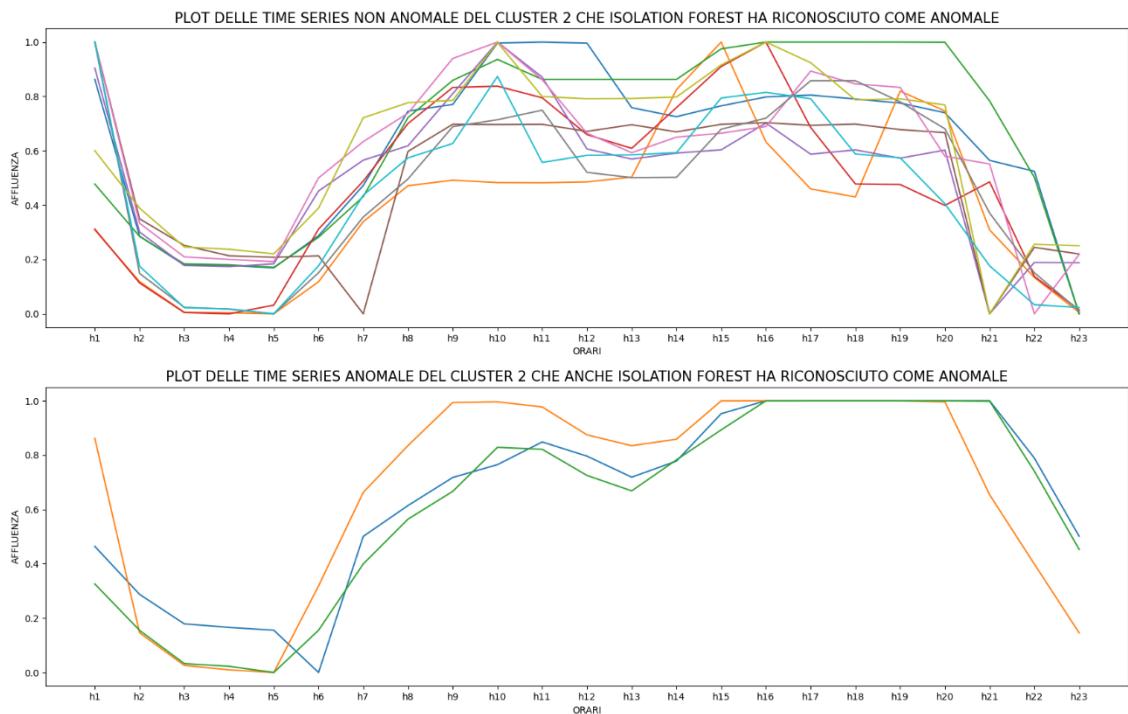


Figura 16 : il primo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster due con anomalous uguale a zero. Il secondo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster due con anomalous uguale a uno. Si può notare, confrontando questi grafici con quelli di figura 3, che Isolation forest ha riconosciuto come anomale delle time series che erano riconosciute come regolari e viceversa. Per esempio, osservando il secondo grafico di figura 2 e confrontandolo col secondo grafico di questa figura si può notare che in quest'ultimo manca una time series.

Indico con il picco in alto il punto in cui viene raggiunta l'affluenza massima e con picco in basso il punto in cui viene raggiunta l'affluenza minima. Indico con valori alti le parti del grafico in cui l'affluenza è maggiore di 0.75, con valori medi le parti del grafico in cui l'affluenza è compresa tra 0.4 e 0.75 e con valori bassi le parti in cui l'affluenza è compresa tra 0 e 0.4.

Effettuando un'analisi visuale si possono formalizzare le differenze tra le anomalie trovate da isolation forest e quelle presenti nel dataset nonché le differenze tra i primi grafici di figura 14, 15 e 16 e i secondi contenuti nelle figure 7,8 e 9:

- Figura 14 e figura 7: nella parte iniziale (prime 4 ore) del primo grafico di figura 14, la maggior parte delle curve partono dal basso e tendono a valori medio bassi sia nella parte iniziale che nella parte finale (ultime 4 ore). Il valore medio viene raggiunto tra h6 e h15. Per un valore di 19h sono presenti valori alti e il valore massimo è raggiunto tra le fasce orarie h15-h20. Nel secondo grafico di figura 7 le curve presentano valori medio-bassi nella parte iniziale e valori medi nella parte finale. Il valore medio viene raggiunto tra h9 e h14. Per un valore di 11h sono presenti valori alti ed il valore massimo è raggiunto mediamente tra le ore 16 e le ore 20
- Figura 15 e figura 8: nella parte iniziale del primo grafico di figura 15, la maggior parte delle curve tendono a valori medio-bassi nella parte iniziale. Nella parte finale sono presenti valori alti. Il valore medio viene raggiunto tra h5 e h14. Per un valore di 10h sono presenti valori alti e il valore massimo è raggiunto mediamente tra le fasce orarie h16-h22. Nel secondo grafico di figura 8 le curve presentano valori medio-bassi nella parte iniziale e valori alti in quella finale. Il valore medio viene raggiunto tra h6 e h14. Per un valore di 14h sono presenti valori alti ed il valore massimo è raggiunto mediamente nelle fasce orarie h13-h22.
- Figura 16 e figura 9: sia nella parte iniziale del primo grafico di figura 16 che in quella finale le curve si mantengono. Il valore medio viene raggiunto tra h5 e h7 e tra h12 e h15. Per un totale di 10h sono presenti valori alti e il valore massimo è raggiunto mediamente tra le fasce orarie h9-h11 e h15-h17. Nel secondo grafico di figura 9 le curve presentano valori medio-bassi nella parte iniziale e valori medio-alti in quella finale. Il valore medio viene raggiunto per la prima volta tra h5 e h7. Per un valore di 14h sono presenti valori alti ed il valore massimo è raggiunto mediamente nelle fasce orarie h9-h11 e h14-h16.

4.2.1 ANALISI ANOMALIE TROVATE DA ISOLATION FOREST SUL CLUSTER 0

Le tabelle che seguono specificano giorno e mese delle anomalie trovate tramite isolation forest, se tali giorni sono etichettati come anomali nel dataset, la possibile motivazione di tale anomalia e se differisce e in cosa differisce dalla forma media del cluster. Ogni tabella è seguita dai grafici contenenti le time series trovate tramite Isolation Forest e la forma media. Indico con FM la forma media delle time series, indico con parte iniziale la parte della curva che va da h1 a h8, con parte centrale quella che va da h8 a h16 e con finale quella che va da h16 a h23. Indico con il picco in alto il punto in cui viene raggiunta l'affluenza massima e con picco in basso il punto in cui viene raggiunta l'affluenza minima.

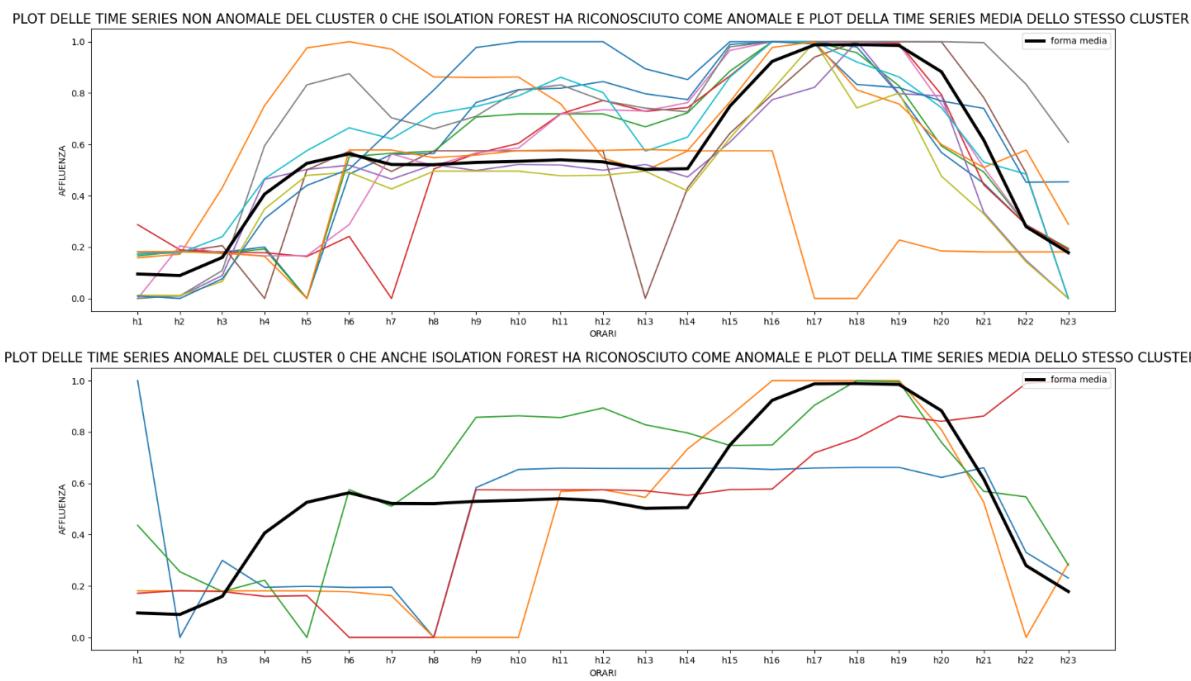


Figura 17 : il primo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster zero con anomalous uguale a zero. Il secondo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster zero con anomalous uguale a uno. Entrambi i grafici contengono la time series di colore nero, che rappresenta la media ora per ora di tutte le time series del cluster zero e serve per effettuare una analisi visiva delle differenze tra quest'ultima e le curve trovate da Isolation forest. Si può vedere come molte time series trovate da Isolation forest si distaccano in alto dalla forma media soprattutto al centro(h8-h16) e come parecchie time series all'inizio(h1-h8) si trovano sotto la forma media. Le differenze tra questa time series e le altre sono elencate nella tabella sottostante.

GIORNO-MESE	MOTIVAZIONE	DIFFERENZE FORMA MEDIA
19/01/2015	Martin Luther King day	Parte iniziale: la curva in tale giorno si trova sopra fino alla fascia oraria h3 dopo va sotto la FM con un picco in basso in h5 ma torna sopra la FM in h7 Parte centrale: la curva si trova molto sopra la curva della FM e raggiunge un picco in alto in h15 prima della FM Parte Finale: la curva mantiene il picco in alto fino alla fascia oraria h18 dopo inizia ad abbassarsi verso lo zero prima rispetto la FM
26/01/2015	TEMPESTA JUNO	Parte iniziale: la curva in tale giorno si trova sopra la FM e raggiunge il picco in h6, dopo inizia a decrescere lentamente Parte centrale: anche qui la curva si trova sopra la FM salvo poi assumere un andamento simile a quest'ultima in h12 Parte finale: la curva si trova sotto la FM, avendo un

		picco in basso nelle fasce orarie h17-h18 invece di averlo in alto, poi risale leggermente e si stabilizza nelle fasce h19-h23
16/02/2015	(President's day)	<p>Parte iniziale: la curva parte sopra la FM ma scende sotto la FM in h3 e presenta un picco in basso in h5, dopo comincia a salire rapidamente</p> <p>Parte centrale: la curva si trova sopra la curva della FM</p> <p>Parte Finale: la curva ha un picco in h16, prima di quello della FM e si abbassa verso lo zero prima rispetto a quest'ultima</p>
25/05/2015	MEMORIAL DAY	<p>Parte iniziale: la curva parte sopra la FM ma scende sotto la FM in h3 e presenta un picco in basso in h7, dopo comincia a salire rapidamente</p> <p>Parte centrale: la curva si alza fino ad arrivare in h9 sopra la curva della FM</p> <p>Parte Finale: la curva ha un picco in alto in h16 prima di quello della FM e si abbassa verso lo zero con lo stesso andamento della FM</p>
8/6/2015	IGNOTA	<p>Parte iniziale: la curva si trova sotto la FM con un picco in basso in h1-h2 ma l'andamento della curva è simile alla FM</p> <p>Parte centrale: la curva ha andamento simile alla curva della FM</p> <p>Parte finale: la curva raggiunge il picco in alto dopo la curva alla FM, tra h18 e h19 de e si abbassa prima della curva FM con un picco in basso in h23</p>
9/7/2015	IGNOTA	<p>Parte iniziale: la curva parte sopra la FM ma scende sotto la FM in h3 e presenta un picco in basso in h4, dopo comincia a salire rapidamente</p> <p>Parte centrale: la curva ha un andamento simili alla FM tranne che si ha un improvviso picco in basso in h13, dopo il quale ritorna a salire</p> <p>Parte finale: la curva risale raggiungendo il picco in alto dopo la FM in h18, mantiene tale picco fino a h21 e si abbassa dopo la FM</p>
7/9/2015	Labour day	<p>Parte iniziale: la curva si trova per dei tratti sopra e per dei tratti sotto la curva della FM ma è maggiore l'intervallo di tempo in cui la curva è sotto la curva della FM, da h3 a h7</p> <p>Parte centrale: la curva sale fino ad arrivare sopra la curva della FM, raggiungendo un picco in h16, prima della FM</p> <p>Parte finale: la curva mantiene il picco fino alla fascia oraria h19 e poi ha un andamento simile alla FM</p>
25/11/2015	The Blackout Wednesday [12]	<p>Parte iniziale: la curva si trova sotto rispetto alla curva della FM per poi ritornare sopra in h3, sale fino a staccarsi molto rispetto alla FM</p> <p>Parte centrale: la curva si trova sopra rispetto alla curva della FM e raggiunge il picco in alto in h16, prima della FM</p> <p>Parte finale: la curva mantiene tale picco fino alla fascia oraria h22, dopo di che si abbassa</p>
14/12/2015	IGNOTA	<p>Parte iniziale: la curva ha un andamento simile pur trovandosi leggermente sotto alla curva della FM</p> <p>Parte centrale: la curva ha un andamento simile pur trovandosi leggermente sotto alla curva della FM, in h14</p>

		inizia a salire rapidamente Parte finale: la curva raggiunge il picco in alto come la FM in h17 e dopo si abbassa molto più rapidamente della FM raggiungendo un picco in basso in h23
23/12/2015	IGNOTA	Parte iniziale: la curva ha un andamento simile pur trovandosi leggermente sopra alla curva della FM Parte centrale: la curva si trova sempre sopra la curva della FM, staccandosi parecchio dalla FM tra h11 e h12 Parte finale: la curva raggiunge il picco in alto prima della curva della FM in h16 e quando quest'ultima raggiunge tale picco, la curva inizia ad abbassarsi andandosi a trovare sotto per la fascia oraria h17-h21 salvo poi ritornare su per h>21. Dopo scende rapidamente raggiungendo un picco in basso in h23
24/12/2015	VIGILIA NATALE	Parte iniziale: la curva ha un andamento simile pur trovandosi leggermente sotto alla curva della FM e presentando un picco in basso in h2, dopo inizia a salire e supera la FM in h6 Parte centrale: la curva si trova molto sopra la curva della FM, raggiungendo un picco in alto in h10-h12, dopo torna a scendere per poi nuovamente salire e raggiunge un nuovo picco in h15 Parte finale: la curva mantiene il picco fino alla fascia oraria h17 dopo inizia ad abbassarsi trovandosi sotto quando la curva della FM raggiunge il picco in alto per poi ritornare sopra la FM nelle fasce orarie h21-h23
28/12/2015	IGNOTA	Parte iniziale: la curva presenta un picco in basso nella fascia oraria h5 trovandosi sotto la curva della FM Parte centrale: la curva ha un andamento simile alla FM pur trovandosi leggermente sopra Parte finale: la curva raggiunge il picco in alto come la FM ed inizia subito ad abbassarsi andando a trovarsi sotto quest'ultima per poi tornare sopra la FM nelle fasce orarie h21-h23
1/1/2015	Capodanno	Parte iniziale: la curva oscilla andandosi a trovare in alcuni tratti sopra la FM e in alcuni sotto la FM, con due picchi in basso nelle fasce orarie h2 e h8 Parte centrale e Parte finale: la curva raggiunge il picco in h10 si stabilizza mantenendosi sopra la FM per le fasce orarie minori di h15 e sotto per le fasce orarie maggiori di h15. Dopo le h21 ha un andamento pressoché uguale alle FM
27/1/2015	TEMPESTA JUNO	Parte iniziale: la curva è costante e si trova sopra fino alla fascia oraria h3 e sotto dopo Parte centrale: la curva si trova sotto la FM e raggiunge un picco in basso nelle fasce orarie h8-h10 per poi risalire e ritornare sopra la FM in h11 e raggiunge il picco in h16 Parte finale: la curva mantiene tale picco ed inizia ad abbassarsi come la FM mantenendosi sotto questa e raggiungendo un picco in basso alle h22
26/11/2015	Thanksgiving day	Parte iniziale: la curva si trova sopra fino alla fascia oraria h3 e sotto dopo e raggiunge un picco in basso alle ore h5 per poi risalire immediatamente Parte centrale: la curva si trova molto sopra rispetto alla FM ma si abbassa e torna sotto la fm in h15

		Parte finale: la curva raggiunge il picco dopo la FM in h18 e si abbassa in modo simile a quest'ultima
31/12/2015	Vigilia di Capodanno	<p>Parte iniziale: la curva si trova sopra fino alla fascia oraria h3 e sotto dopo e raggiunge un picco in basso alle ore h6-h8 per poi risalire immediatamente</p> <p>Parte centrale: la curva si trova sopra rispetto alla FM ma torna sotto la fm in h14</p> <p>Parte finale: la curva cresce gradualmente andando sopra la FM nella fascia oraria h20 e raggiungendo un picco in alto per le h23 a differenza della forma media che scende verso il basso</p>

4.2.2 ANALISI ANOMALIE TROVATE DA ISOLATION FOREST SUL CLUSTER 1

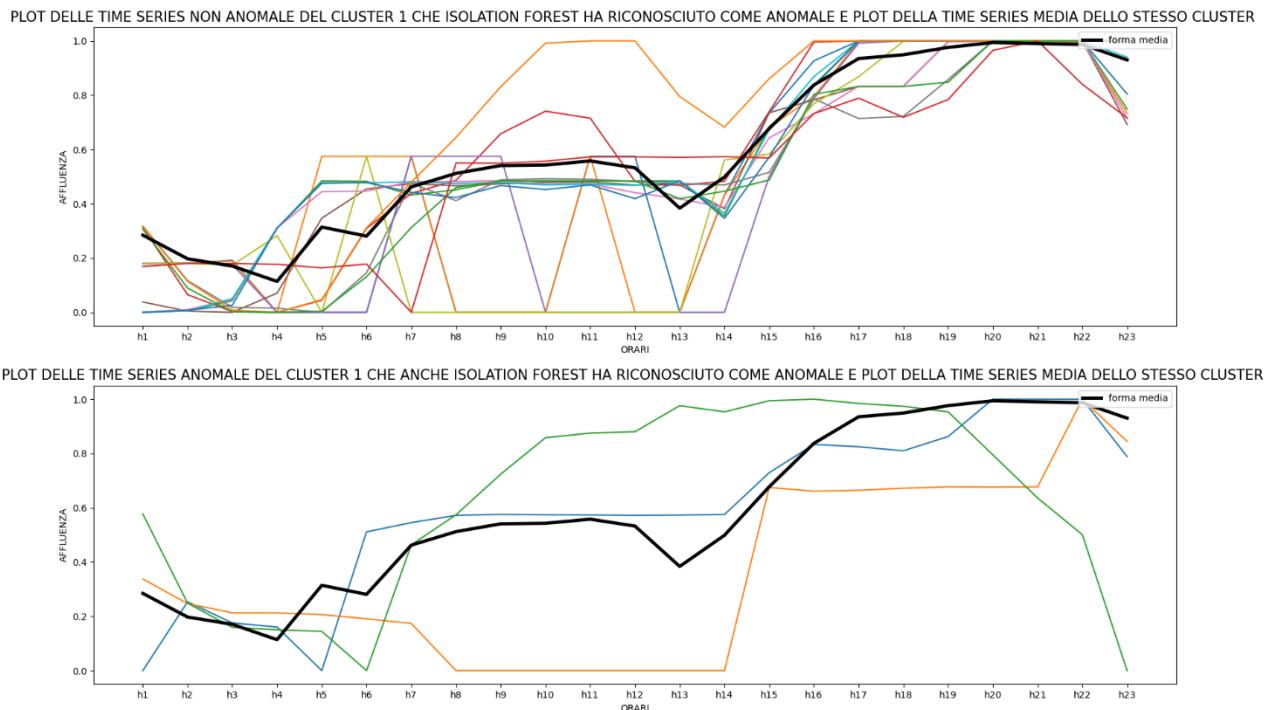


Figura 18: il primo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster uno con anomalous uguale a zero. Il secondo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster uno con anomalous uguale a uno. Entrambi i grafici contengono la time series di colore nero, che rappresenta la media ora per ora di tutte le time series del cluster uno e serve per effettuare una analisi visiva delle differenze tra quest'ultima e le curve trovate da Isolation forest. Si può vedere come all'inizio(h1-h8) molte time series trovate da Isolation forest si distaccano sia in alto che in basso dalla forma media. Le differenze tra questa time series e le altre sono elencate nella tabella sottostante.

GIORNO-MESE	MOTIVAZIONE	DIFFERENZE FORMA MEDIA
9/1/2015	Restrizioni traffico a causa neve [13]	<p>Parte iniziale: la curva si trova sotto la FM, presenta dei picchi in basso nelle fasce h4-h6, sale rapidamente in h7 per poi riscendere</p> <p>Parte centrale: la curva si trova sotto la FM, l'andamento è molto oscillante, ha dei picchi in basso nelle fasce orarie h8-h10, risale andando sopra la FM però poi riscendere e ripresentare un picco in basso nella fascia oraria h13</p> <p>Parte finale: la curva inizia a salire verso l'alto e</p>

		assume un andamento simile alla FM raggiugendo il picco in alto prima di questa, nella fascia oraria h17
16/1/2015	Torneo mondiale di Squash[14]	<p>Parte iniziale: la curva si trova sotto la FM, presenta un picco in basso nella fascia oraria h4, sale rapidamente sopra la FM per poi riscendere e avere un nuovo picco in basso nelle fasce orarie h8-h10</p> <p>Parte centrale: la curva oscilla ma rimane sotto la FM presentando un nuovo picco in basso nelle fasce orarie h12-h13</p> <p>Parte finale: la curva inizia a salire verso l'alto e assume un andamento simile alla FM raggiugendo il picco in alto prima di questa, nella fascia oraria h17 e inizia a decrescere nella fascia h22 in modo molto più rapido della FM</p>
20/2/2015	GHIACCIO E NEVE [15]	<p>Parte iniziale: la curva presenta un picco in basso nella fascia h1, inizialmente si trova sotto la curva della fm ma sale sopra di essa tra la fascia h3 e h4 ma ritorna sotto nella fascia h7</p> <p>Parte centrale: la curva assume un andamento simile alla fm pur rimanendo sotto di essa e inizi a decrescere dopo la FM per poi risalire</p> <p>Parte finale: la curva inizia a salire verso l'alto e assume un andamento simile alla FM raggiugendo il picco in alto prima di questa, nella fascia oraria h17 e inizia a decrescere nella fascia h22 in modo molto più rapido della FM</p>
21/2/2015	GHIACCIO E NEVE [15]	<p>Parte iniziale: la curva si trova sotto la FM, presenta un picco in basso nella fascia oraria h3-h4, sale rapidamente assumendo un andamento simile alla FM</p> <p>Parte centrale: la curva sale sopra la FM presentando una "gobba" tra h8-h12 e si mantiene pressoché costante tra le fasce h12-h14</p> <p>Parte finale: la curva inizia a salire rapidamente raggiugendo il picco in alto nella fascia oraria h16, prima della fm e mantiene tale picco</p>
6/3/2015	TEMPESTA NEVE [16]	<p>Parte iniziale: la curva si trova sotto la FM, presenta un picco in basso nella fascia oraria h4-h6, per poi salire rapidamente in h7</p> <p>Parte centrale: la curva si trova sotto la FM nella fascia oraria h7-h9 ma inizia a scendere in h10 presentando un picco in basso nelle fasce orarie h10-h14</p> <p>Parte finale:la curva sale rimanendo sotto la fm ma raggiunge il picco prima di quest'ultima, in h17 per poi mantenere tale picco</p>
3/4/2015	Pasqua (Venerdì Santo)	<p>Parte iniziale: la curva si trova sotto la FM, presenta un picco in basso nella fascia oraria h3 dopo il quale torna a salire sopra la FM</p> <p>Parte centrale: la curva ha un andamento simile alla FM, trovandosi in alcune fasce sopra essa e in altre sotto</p> <p>Parte finale: la curva inizia a salire con andamento simile alla FM, anche se si trova sotto questa, e raggiunge il picco nella fascia h19 e inizia a decrescere più velocemente della FM in h22</p>
22/05/2015	Inizio saldi memorial day	Parte iniziale: la curva si trova sotto la FM, presenta un picco in basso nella fascia oraria h1, inizia a salire sopra la FM tra h3 e h4 e assumendo un andamento

		<p>simili dopo h7</p> <p>Parte centrale: la curva ha un andamento simile alla FM, trovandosi in alcune fasce sopra essa e in altre sotto</p> <p>Parte finale: la curva inizia a salire con andamento simile alla FM, anche se si trova sotto questa, e raggiunge il picco nella fascia h19 e inizia a decrescere più velocemente della FM in h22</p>
12/09/2015	IGNOTA	<p>Parte iniziale: la curva si trova sotto la FM, presenta un picco in basso nella fascia oraria h5, ma dopo inizia a salire rimanendo sotto la FM</p> <p>Parte centrale: la curva ha un andamento simile alla FM, trovandosi solo sopra essa nelle fasce h13-h14</p> <p>Parte finale: la curva inizia a salire oscillando e rimane sotto la FM, e raggiungono entrambe il picco nella fascia h20. Dopo la curva mantiene andamento costante a differenza della FM che inizia a scendere</p>
25/09/2015	Papa Francesco visita NY	<p>Parte iniziale: la curva oscilla, in tratti si trova sopra la FM e in tratti sotto, presenta un picco in basso in h5, risale velocemente in h6 e riscende in h7</p> <p>Parte centrale: curva si trova sempre sotto la FM, con un costante picco in basso nelle fasce orarie h8-h13, dopo di che comincia a risalire rapidamente</p> <p>Parte finale: la curva ha un andamento simile alla FM rimanendo per un grande tratto sotto di essa ma raggiunge il picco prima di essa, in h18, dopodiché mantiene andamento costante a differenza della FM che inizia a scendere</p>
2/10/2015	Preparazione arrivo uragano Joaquin [17]	<p>Parte iniziale: la curva si trova sotto la FM, presenta un picco in basso nella fascia oraria h1, inizia a salire sopra la FM tra h3 e h4 e assumendo un andamento costante dopo h5</p> <p>Parte centrale: la curva si trova sotto la FM, ritorna sopra in h13 per poi riscendere sotto subito dopo</p> <p>Parte finale: la curva sale con andamento simile alle FM ma raggiunge il picco prima di essa in h17 ma entrambe tornano a scendere in h10 con lo stesso andamento</p>
16/10/2015	IGNOTA	<p>Parte iniziale: la curva si trova sotto la FM, presenta un picco in basso nella fascia oraria h1, inizia a salire sopra la FM dopo h3 e h4 e assumendo un andamento costante dopo h5</p> <p>Parte centrale: la curva si trova sotto la FM, ritorna sopra in h13 per poi riscendere sotto subito dopo</p> <p>Parte finale: la curva sale con andamento simile alla FM ma raggiunge il picco prima di essa in h17 e mantiene tale picco a differenza della FM che dopo le h23 inizia a scendere</p>
12/12/2015	Santacon	<p>Parte iniziale: la curva si trova sotto la FM presentando un picco in basso in h4 e sale lentamente in h5. Dopodiché inizia a salire rapidamente raggiungendo in picco in altro nella fascia oraria h10</p> <p>Parte centrale: la curva mantiene il picco nelle fasce orarie h10-h13, si trova sempre sopra la FM, riscende in h14 ma subito dopo ricomincia a salire</p> <p>Parte finale: la curva raggiunge il picco nella fascia h16, molto prima della FM e rimane costantemente sopra di</p>

		essa
19/12/2015	IGNOTA	<p>Parte iniziale: curva si trova sotto la FM presentando un picco in basso nelle fasce h3-h5 e sale lentamente dopo h5</p> <p>Parte centrale: la curva rimane sempre sotto la FM, pur mantenendo un andamento simile, ad accezione della fascia oraria h13</p> <p>Parte finale: la curva raggiunge il picco insieme alla FM e lo mantiene a differenze della FM che inizia a scendere dopo le h22</p>
26/12/2015	Inizio saldi natalizi (Santo Stefano)	<p>Parte iniziale: la curva presenta un andamento costante nelle fasce h1-h6 trovandosi sopra la fm solo nelle fasce orarie h3-h4, presenta un picco in basso in h7 ma dopo torna a salire</p> <p>Parte centrale: la curva si trova sempre sopra la fm fino alla fascia oraria h14, dopo torna sotto</p> <p>Parte finale: la curva raggiunge il picco dopo la FM, nella fascia oraria h21-h22 e dopo tale picco inizia a scendere molto rapidamente</p>
3/7/2015	Independence day weekend	<p>Parte iniziale: la curva si trova sopra la FM nelle fasce orarie inferiori a h4, sotto dopo e con un picco in basso nelle fasce h7-h8</p> <p>Parte centrale: la curva velocemente andando ad assumere un andamento simile alla FM nelle fasce h9-h12 poi si mantiene sopra la fm per poi iniziare a salire</p> <p>Parte finale: la curva prima un picco in alto nella fascia h17, poi riscende e risale raggiungendo un picco nella fascia h20 insieme la FM che mantiene fino alle h22 dove inizia a scendere verso il basso molto più velocemente della FM</p>
31/10/2015	Halloween	<p>Parte iniziale: la curva si trova sopra la FM ma ha un andamento simile fino alla fascia h4-h5 dove inizia a decrescere fino ad avere un picco in basso nella fascia oraria h8</p> <p>Parte centrale: la curva ha un costante picco in basso che mantiene fino alla fascia oraria h14, dopo la quale la curva inizia a salire rapidamente</p> <p>Parte finale: la curva si mantiene costante sotto la FM, sale rapidamente in h21 per raggiungere il picco in alto in h22 e subito dopo comincia ad abbassarsi</p>
25/12/2015	Natale	<p>Parte iniziale: la curva parte abbastanza in alto, più della fm ed inizia a scendere andando sotto la fm e presentando un picco in basso in h6 ma subito dopo inizia una rapida e continua salita</p> <p>Parte centrale: la curva si trova sempre sopra la FM, con valori molto alti e in h17 raggiunge il picco</p> <p>Parte finale: dopo le h17 la curva inizia a scendere lentamente fino ad h19, dopodiché continua la sua discesa molto rapida verso lo zero in h23</p>

4.2.3 ANALISI ANOMALIE TROVATE DA ISOLATION FOREST SUL CLUSTER 2

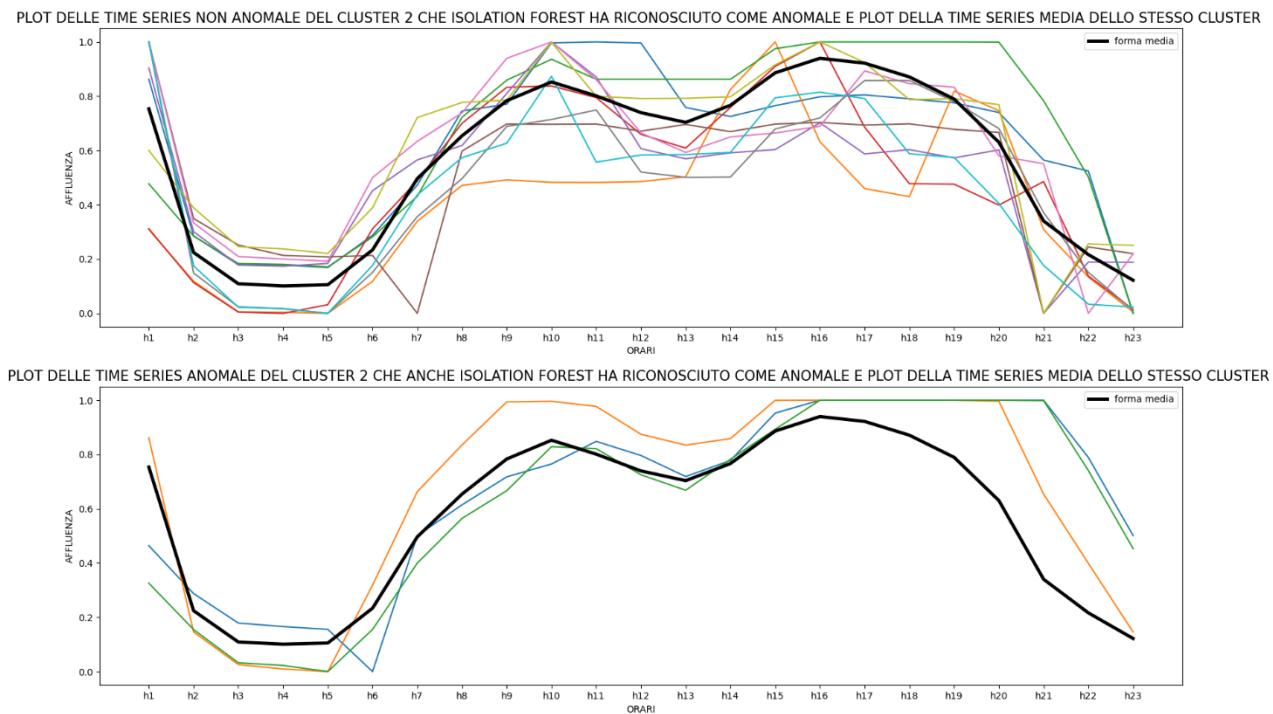


Figura 19: il primo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster due con anomalous uguale a zero. Il secondo grafico contiene le time series che Isolation Forest ha riconosciuto come anomale sul cluster uno con anomalous uguale a uno. Entrambi i grafici contengono la time series di colore nero, che rappresenta la media ora per ora di tutte le time series del cluster due e serve per effettuare una analisi visiva delle differenze tra quest'ultima e le curve trovate da Isolation forest. Si può vedere come la maggior parte delle time series ricavate tramite Isolation forest si distacchino in basso sia al centro(h8-h16) che alla fine (h16-h23) rispetto alla forma media. Le differenze tra questa time series e le altre sono elencate nella tabella sottostante.

GIORNO-MESE	MOTIVAZIONE	DIFFERENZE FORMA MEDIA
18/1/2015	Pioggia incessante [18]	Parte iniziale: l'andamento della curva è molto simile alla FM pur trovandosi leggermente sopra Parte centrale: la curva si alza improvvisamente presentando un picco in alto nella fascia h10-h12 per poi cominciare a scendere Parte finale: la curva si trova sotto la fm nelle fasce comprese tra h14 e h20, dopodiché torna sopra e continua a decrescere
1/2/2015	Superbowl	Parte iniziale: la curva si trova sotto la FM presentando un picco in basso nelle fasce h3-h5, dopo comincia a salire sempre rimanendo sotto la FM Parte centrale: la curva rimane costante per poi salire nella fascia h13, supera la fm in h14 e presenta un picco nella fascia oraria h15 Parte finale: la curva inizia a scendere, trovandosi sotto la fm, torna a salire in h19 e subito dopo ritorna a scendere con andamento simile alla FM
15/2/2015	Tempesta di neve [19]	Parte iniziale: la curva ha un andamento simile alla FM, trovandosi in certe fasce orarie sopra e in certe sotto Parte centrale: la curva ha un andamento simile alla FM, trovandosi sopra essa Parte finale: la curva si trova sopra la FM, presenta un

		picco nelle fasce orarie h16-h21 e inizia a decrescere più velocemente della FM, raggiungendo un picco in basso nella fascia h23
22/2/2015	Lunar New Year Parade [20]	<p>Parte iniziale: la curva si trova sotto la FM presentando un picco in basso nelle fasce h3-h4, dopo comincia a salire assumendo un andamento simile alla FM</p> <p>Parte centrale: la curva ha un andamento simile alla FM e inizia a salire rapidamente nella fascia oraria h13 per raggiungere il picco in h16, poi decresce</p> <p>Parte finale: la curva si trova sotto la FM, decresce più rapidamente di quest'ultima presentando un picco in basso nella fascia oraria h23</p>
1/3/2015	Tempesta di neve [21]	<p>Parte iniziale: la curva si trova sopra la FM, decresce fino ad h5, dopodiché comincia a salire</p> <p>Parte centrale: la curva sale molto più velocemente della FM e raggiunge un picco in h10, dopo il quale inizia una lenta discesa</p> <p>Parte finale: la curva si trova sempre sotto la FM, scende vistosamente verso il basso nella fascia oraria h21, torna leggermente su e si stabilizza nelle fasce orarie h22-h23</p>
8/3/2015	Festa della Donna	<p>Parte iniziale: la curva si trova sopra la FM, decresce fino ad andare sotto quest'ultima in h6 e presentare un picco in basso in h7. Dopo questo picco la curva inizia a salire.</p> <p>Parte centrale: la curva ha un andamento regolare, e si trova costantemente sotto la FM</p> <p>Parte finale: la curva si trova sotto la FM, comincia a scendere nella fascia oraria h20 ed ha un picco in basso nella fascia h21 e dopo di che sale leggermente e si stabilizza nelle fasce h22-h23</p>
31/5/2015	Caduta di equipaggiamento meccanico con 10 feriti [22]	<p>Parte iniziale: la curva si trova sopra la FM pur avendo un andamento simile</p> <p>Parte centrale: la curva sale rapidamente presentando un picco in alto nella fascia oraria h10 e dopo comincia a scendere andandosi a trovare sotto la FM tra h11 e h12</p> <p>Parte finale: la curva sale in h17 e subito dopo scende in basso con un andamento a gradino fino a raggiungere un picco in basso nella fascia oraria h22, dopo il quale torna a salire</p>
7/6/2015	IGNOTA	<p>Parte iniziale: la curva inizialmente si trova sopra la FM, scende rapidamente andando sotto questa e presentando un picco in basso in h5, dopodiché torna a salire</p> <p>Parte centrale: la curva si trova sempre sotto la FM, in certi tratti si stacca da essa in modo vistoso</p> <p>Parte finale: la curva raggiunge il valore massimo nelle fasce orarie h17-18 per poi tornare a scendere con andamento molto simile alla FM e raggiungere un picco in basso nella fascia oraria h23</p>
22/11/2015	IGNOTA	<p>Parte iniziale: la curva inizia sotto la FM ma la si stacca sopra essa in h2, e dopo questa fascia oraria si trova sempre sopra ad essa</p> <p>Parte centrale: la curva si trova sopra la FM, presenta un picco nella fascia oraria h9 dopo la quale inizia a decrescere per poi tornare a salire in h14 per ripresentare un picco in h16</p> <p>Parte finale: la curva decresce molto più velocemente della FM, raggiunge un picco in basso in h21 e ritorna a salire in h22 per poi stabilizzarsi</p>

6/12/2015	IGNOTA	<p>Parte iniziale: la curva inizia sopra la FM ma la si stacca sotto essa in h2 presentando un picco in basso in h5, dopo comincia a salire trovandosi sotto la FM</p> <p>Parte centrale: la curva sale, raggiunge il valore massimo in h10 superando la FM ma torna immediatamente giù</p> <p>Parte finale: la curva ha un andamento simile alla FM pur trovandosi un bel po' sotto essa e raggiunge un picco in basso in h23</p>
24/5/2015	Memorial day weekend	<p>Parte iniziale: la curva inizia sotto la FM ma la si stacca sopra essa in h2, per poi tornare giù tra h4 e h5 e presentare un picco in basso in quest'ultima fascia oraria, dopo di che inizia a salire con andamento simile alla FM</p> <p>Parte centrale: la curva ha un andamento simile alla FM pur trovandosi sopra ma inizia a staccarsi da essa in h15 per poi presentare un picco in h16</p> <p>Parte finale: la curva ha un picco costante tra h16 e h22, dopo inizia a scendere rapidamente verso il centro</p>
28/06/2015	Gay pride	<p>Parte iniziale: la curva inizia sopra la FM ma la si stacca sotto essa in h2 presentando un picco in basso in h5, dopo comincia a salire andando sopra la FM</p> <p>Parte centrale: la curva sale, raggiunge il picco in alto in h9 e riscende leggermente verso il basso, risale in h14 e ripresenta un nuovo picco, si trova sempre sopra alla FM</p> <p>Parte finale: la curva mantiene il picco fino alla fascia oraria h21 dopo inizia a scendere rimanendo sempre sopra la FM</p>
6/9/2015	Labour's day weekend	<p>Parte iniziale: la curva si trova sopra la FM pur avendo un andamento simile, presenta un picco in basso nella fascia oraria h5, dopo inizia a salire</p> <p>Parte centrale: la curva ha un andamento simile alle FM, trovandosi in tratti sopra essa e in tratti sotto, inizia a salire in h13 e raggiunge il picco in alto in h16</p> <p>Parte finale: la curva si trova la FM, mantiene il picco fino alla fascia oraria h21 e subito dopo inizia a decrescere verso il centro</p>

5 RISULTATI, CONCLUSIONI E RINGRAZIAMENTI

5.1 RISULTATI

L'intervallo di confidenza fornisce informazioni riguardo alla precisione dei valori ottenuti attraverso lo studio di un campione. Un intervallo di confidenza 95% comprende un intervallo di valori che tiene conto della variabilità del campione, in modo tale che si può confidare - con un margine di certezza ragionevole (appunto il 95%) - che quell'intervallo contenga il valore vero dell'intera popolazione che si è stata esaminata. Indico le performances di isolation forest come il mean square error tra le anomalie trovate da isolation forest e quelle già presenti nel dataset. Richiamando per un numero pari a 10 volte isolation forest con parametri di default sul dataset e cambiando solo il parametro randState, con valori che vanno da 42 a 51, l'intervallo di confidenza sui valori ottenuti dalla performances è:

INTERVALLO DI CONFIDENZA

CLUSTER 0	(0.05859305077712761, 0.06580886309847049)
CLUSTER 1	(0.14144351348464337, 0.16432571728458742)
CLUSTER 2	(0.20141609700660554, 0.22550697991647137)

5.2 CONCLUSIONI

In questa tesi si è sviluppato un software che consente all'utente di tracciare i grafici delle time series di un dataset, di applicare isolation forest al dataset e di modificare il dataset se l'utente riscontra visivamente delle time series anomale. Dalle osservazioni precedenti possiamo confermare come certe le seguenti anomalie:

GIORNO-MESE	MOTIVAZIONE
19/01/2015	Martin Luther King day
26/01/2015	Tempesta Juno
16/02/2015	(President's day)
25/05/2015	Memorial day
7/9/2015	Labour day
25/11/2015	The Blackout Wednesday
24/12/2015	Vigilia natale
1/1/2015	Capodanno
27/1/2015	Tempesta Juno
26/11/2015	Thanksgiving day
31/12/2015	Vigilia di Capodanno
9/1/2015	Restrizioni traffico a causa neve
16/1/2015	Torneo mondiale squash
20/2/2015	Ghiaccio e neve
21/2/2015	Ghiaccio e neve
6/3/2015	Tempesta neve
3/4/2015	Pasqua (Venerdì Santo)
22/05/2015	Inizio saldi memorial day
25/09/2015	Papa Francesco visita NY
2/10/2015	Preparazione arrivo uragano Joaquin
12/12/2015	Santacon
26/12/2015	Inizio saldi natalizi (Santo Stefano)
3/7/2015	Independence day weekend
31/10/2015	Halloween
25/12/2015	Natale
18/1/2015	Pioggia incessante
1/2/2015	Superbowl
15/2/2015	Tempesta di neve
22/2/2015	Lunar New Year Parade

1/3/2015	Tempesta di neve
8/3/2015	Festa della Donna
31/5/2015	Incidente caduta costruzioni
24/5/2015	Memorial day weekend
28/06/2015	Gay pride
6/9/2015	Labour's day weekend

Isolation forest è stato particolarmente utile per rilevare le anomalie sul cluster 1 e sul cluster due poiché nel primo è molto difficile notare visivamente le anomalie soprattutto nella parte iniziale e centrale (figura 8, grafico 1), nel secondo le anomalie hanno un andamento simile alle time series normali (figura, grafico 2) e quindi scorgerele visivamente risulta difficile. Dall'analisi dei risultati e dalle performances è emerso che isolation forest è stato in grado di riconoscere una maggiore quantità di anomalie rispetto a quelle già presenti del dataset e con molta efficienza. L'algoritmo ha riconosciuto i giorni anomali le cui time series hanno valori alti nella parte iniziale della giornata. Questo è evidente analizzando le time series del cluster 0 trovate da isolation forest (primo grafico figura 11) e confrontandole con quelle già presenti (secondo grafico figura 2).

5. 3 RINGRAZIAMENTI

Inizialmente, non avendo mai avuto a che fare con l'anomaly detection, ho avuto un po' di difficoltà. Con il tempo, grazie all'aiuto, ai chiarimenti e al materiale che mi ha fornito il Dott. Alfeo, giorno dopo giorno i concetti sono diventati più chiari. Posso affermare di essere molto contento di aver scelto questo percorso perché ho imparato nuovi contenuti molto interessanti. Ringrazio il Dott. Antonio Luca Alfeo per avermi aiutato in queste settimane ed essere sempre stato disponibile per chiarimenti e domande varie. Un ringraziamento al Professor Cimino e alla professoressa Vaglini per la disponibilità mostrata nel farmi da relatori. Ai miei amici, quelli che conoscevo e quelli che ho potuto conoscere durante questo percorso. Infine, il mio ultimo grazie, ma non per importanza, va alla mia famiglia per avermi sostenuto economicamente e moralmente in questo percorso di studi che mi ha permesso di crescere.

APPENDICE

A. DETTAGLI DI IMPLEMENTAZIONE

Function name: add_visual_anomalies_label
ID: 1
BRIEF DESCRIPTION: aggiunge una colonna con label 'visive_anomalous' al dataset
PRECONDITIONS: nessuno
MAIN FLOW: la funzione apre il dataset 'timeSeries2015HotspotD.csv', viene aggiunta la colonna 'Visive_anomalous' i cui campi vengono posti a zero.
POSTCONDITIONS: una nuova colonna viene aggiunta al dataset 'timeSeries2015HotspotD.csv'

Function name: add_visual_anomalies
ID: 2
BRIEF DESCRIPTION: utente setta il campo 'Visive_anomalous' ad 1 se vede che visivamente una time series differisce dalle altre
PRECONDITIONS: add_visual_anomalies_label
MAIN FLOW: la funzione viene richiamata quando l'utente decide di aggiungere un'anomalia visiva, viene aperto il dataset, viene selezionata l'opportuna riga tramite il valore 'Day' e 'Month', e viene modificata la cella la cui etichetta 'Visive_anomaly'
POSTCONDITIONS: il dataset viene modificato

Function name: plot_timeseries_by_anomalous_and_cluster
ID: 3
BRIEF DESCRIPTION: vengono mostrati i grafici delle time series di un determinato cluster, divise per valore dell'etichetta 'Anomalous'
PRECONDITIONS: nessuna
MAIN FLOW: La funzione viene richiamata quando l'utente sceglie di vedere i grafici delle time series di un cluster divise per anomalie 1 viene aperto il dataset, vengono selezionate le righe di un dato cluster. 2 WHILE il numero di righe del dataset con anomalous uguale a zero non è terminato 2.1 Viene selezionata la riga 2.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 2.3 Viene effettuato il plot della time series 3 WHILE il numero di righe del dataset con anomalous uguale a uno non è terminato 3.1 Viene selezionata la riga 3.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 3.3 Viene effettuato il plot della time series
POSTCONDITIONS: il grafico delle timeseries viene mostrato

Function name: plot_daily_timeseries
ID: 5
BRIEF DESCRIPTION: viene mostrato il grafico di una time series di un giorno specifico
PRECONDITIONS: nessuno
MAIN FLOW: La funzione viene richiamata quando l'utente sceglie di vedere il grafico di una time series di uno specifico giorno dell'anno: 1 Viene selezionata la riga 2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 3 Viene effettuato il plot della time series
POSTCONDITIONS: il grafico della timeseries viene mostrato

Function name: plot_medium_form
ID: 6
BRIEF DESCRIPTION: viene mostrato il grafico della time series media di un cluster
PRECONDITIONS: nessuno
MAIN FLOW: La funzione viene richiamata quando l'utente sceglie di vedere il grafico della time series media di un cluster: 1 viene aperto il dataset, vengono selezionate le righe di un dato cluster. 2 WHILE il numero di colonne è minore di 24 2.1 Vengono selezionati tutti i valori di una colonna 2.2 I valori vengono tutti sommati tra di loro, divisi per il numero di righe e infine salvati 2.3 Viene effettuato il plot della time series media
POSTCONDITIONS: il grafico della timeseries viene mostrato

Function name: plot_timeseries_by_visiveanom_and_cluster
ID: 4
BRIEF DESCRIPTION: vengono mostrati i grafici delle time series di un determinato cluster, divise per valore dell'etichetta 'Visive_anomalous'
PRECONDITIONS: add_visual_anomalies_label
MAIN FLOW: La funzione viene richiamata quando l'utente sceglie di vedere i grafici delle time series di un cluster divise per anomalie visive 1 viene aperto il dataset, vengono selezionate le righe di un dato cluster. 2 WHILE il numero di righe del dataset con 'visive_anomalous' uguale a zero non è terminato 2.1 Viene selezionata la riga 2.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 2.3 Viene effettuato il plot della time series 3 WHILE il numero di righe del dataset con 'visive_anomalous' uguale a uno non è terminato 3.1 Viene selezionata la riga 3.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 3.3 Viene effettuato il plot della time series
POSTCONDITIONS: il grafico delle timeseries viene mostrato

Function name: plot_isolation_forest
ID: 8
BRIEF DESCRIPTION: viene mostrato il grafico delle time series trovate da isolation forest con dei parametri specifici
PRECONDITIONS: nessuno
MAIN FLOW: La funzione viene richiamata quando l'utente sceglie di vedere il grafico delle time series che si ottengono applicando isolation forest con dei parametri da lui scelti al dataset: 1 Vengono selezionate tutte le righe del dataset con cluster specificato 2 Viene generato il modello applicando isolation forest con parametri di default 3 Viene applicato Isolation forest al dataset al pt. 1 4 Viene aggiunta alle righe precedentemente selezionate, la colonna 'isolation_forest_anom' contenente le anomalie trovate da isolation forest e viene salvato il dataset come 'isolation-forest-cluster(cluster)-estim(n_estimators)-sampl(max_samples)-cont(contamination)-boot(bootstrap)-jobs(n_jobs)-randstate(random_state).csv' 5 Vengono selezionate le righe con 'Anomalous' uguale zero e 'isolation_forest_anom' uguale meno 1: WHILE il numero di righe del dataset non è terminato 5.1 Viene selezionata la riga 5.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 5.3 Viene effettuato il plot della time series 6 Vengono selezionate le righe con 'Anomalous' uguale uno e 'isolation_forest_anom' uguale meno 1: WHILE il numero di righe del dataset non è terminato 6.1 Viene selezionata la riga 6.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 6.3 Viene effettuato il plot della time series 7 viene aperto il dataset, vengono selezionate le righe di un dato cluster. 8 WHILE il numero di colonne è minore di 24 8.1 Vengono selezionati tutti i valori di una colonna 8.2 I valori vengono tutti sommati tra di loro, divisi per il numero di righe e infine salvati 8.3 Viene effettuato il plot della time series media
POSTCONDITIONS: il grafico delle timeseries viene mostrato e un dataset viene creato

Function name: save_file
ID: 10
BRIEF DESCRIPTION: viene salvato il dataset
PRECONDITIONS: nessuno
MAIN FLOW: La funzione viene richiamata inizia quando l'utente sceglie di salvare il dataset con uno specifico nome 1 L'utente inserisce il nome del file che verrà salvato con estensione .csv e che l'utente potrà riaprire in futuro
POSTCONDITIONS: un nuovo dataset viene salvato

Function name: plot_isolation_forest_default
ID: 7
BRIEF DESCRIPTION: viene mostrato il grafico delle time series trovate da isolation forest con parametri di default ad un determinato cluster
PRECONDITIONS: nessuno
MAIN FLOW: La funzione viene richiamata quando l'utente sceglie di vedere il grafico delle time series che si ottengono applicando isolation forest al dataset: 1 Vengono selezionate tutte le righe del dataset con cluster specificato 2 Viene generato il modello applicando isolation forest con parametri di default 3 Viene applicato Isolation forest al dataset al pt. 1 4 Viene aggiunta alle righe precedentemente selezionate, la colonna 'isolation_forest_anom' contenente le anomalie trovate da isolation forest e viene salvato il dataset come 'isolation-forest-cluster(cluster).csv' 5 Vengono selezionate le righe con 'Anomalous' uguale zero e 'isolation_forest_anom' uguale meno 1: WHILE il numero di righe del dataset non è terminato 5.1 Viene selezionata la riga 5.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 5.3 Viene effettuato il plot della time series 6 Vengono selezionate le righe con 'Anomalous' uguale uno e 'isolation_forest_anom' uguale meno 1: WHILE il numero di righe del dataset non è terminato 6.1 Viene selezionata la riga 6.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 6.3 Viene effettuato il plot della time series 7 viene aperto il dataset, vengono selezionate le righe di un dato cluster. 8 WHILE il numero di colonne è minore di 24 8.1 Vengono selezionati tutti i valori di una colonna 8.2 I valori vengono tutti sommati tra di loro, divisi per il numero di righe e infine salvati 8.3 Viene effettuato il plot della time series media
POSTCONDITIONS: il grafico delle timeseries viene mostrato e un dataset viene creato

Function name: plot_timeseries_with_h_value
ID: 9
BRIEF DESCRIPTION: viene mostrato il grafico delle time series che presentano un valore di affluenza minore uguale a quello specificato nella fascia oraria desiderata
PRECONDITIONS: nessuno
MAIN FLOW: La funzione viene richiamata quando l'utente sceglie di vedere il grafico delle time series che presentano un valore di affluenza minore uguale a quello specificato nella fascia oraria desiderata 1 Vengono selezionate le righe in cui l'affluenza è minore uguale a quella scelta nella fascia oraria inserita 2 WHILE il numero di righe del dataset non è terminato 2.1 Viene selezionata la riga 2.2 Vengono selezionati i valori da h1 a h23 contenuti nella riga 2.3 Viene effettuato il plot della time series 3 Vengono stampati i giorni che presentano l'affluenza minore o uguale a quella inserita
POSTCONDITIONS: L'utente può vedere i grafici delle time series, può notare quelle che visivamente possono sembrare anomale. Tramite la funzione plot_daily_timeseries può vedere i singoli grafici e aggiungere un'anomalia visiva tramite add_visual_anomalies

B. CODICE

CODICE MAIN.PY

```
import class_def as model
|
while True:
    try:
        filename = input("Inserisci nome del file: ")
        csv=model.Model(filename)
        break
    except FileNotFoundError:
        print("il file non esiste, inserisci nome valido")
        continue
csv.add_visual_anomalies_label(filename)
columns = ['h1', 'h2', 'h3', 'h4', 'h5', 'h6', 'h7', 'h8', 'h9', 'h10', 'h11', 'h12', 'h13', 'h14', 'h15',
           'h16', 'h17', 'h18', 'h19', 'h20', 'h21', 'h22', 'h23']
print("Benvenuto digita:")
while True:
    choise=input("1 - Trova anomalie\n"
                 "2 - Visualizza il comportamento di una specifica time series\n"
                 "3 - Visualizza time series di un cluster divise per anomalie\n"
                 "4 - Termina l'esecuzione\n"
                 "5 - Salva il dataset\n")
    if choise == str(1):
        option=input("Vuoi usare parametri di deafault? S\\N: ")
        if (option == 'S' or option == 's'):
            cluster =int(input("Inserisci il cluster: "))
            if (cluster != 2 and cluster != 0 and cluster != 1):
                print("Cluster deve essere uguale a zero, uno o due")
                continue
            csv.plot_isolation_forest_default(int(cluster))

        elif(option=='N' or option=='n'):
            cluster = int(input("Inserisci il cluster: "))
            if (cluster != 2 and cluster != 0 and cluster != 1):
                print("Cluster deve essere uguale a zero, uno o due")
                continue
            n_estimators=input("Inserisci n_estimators: ")
            max_samples=input("Inserisci max_samples: ")
            contamination=input("Inserisci contamination: ")
            max_features=input("Inserisci max_features: ")
            if (float(max_features)>1.0):
                print("Max features non può essere maggiore di 1")
                continue
            bootstrap=input("Inserisci bootstrap: Si\\No: ")
```

```

if(bootstrap=='Si' or bootstrap=='si'):
    bootstrap=True
elif (bootstrap == 'No' or bootstrap == 'No'):
    bootstrap = False
else:
    print("bootstrap deve valere true o false")
    continue
random_state=input("Inserisci randomstate: ")
n_jobs=input("Inserisci il njobs: ")
csv.plot_isolation_forest(int(cluster),int(n_estimators),float(max_samples),float(contamination)
                           ,float(max_features),bool(bootstrap),int(random_state),int(n_jobs))
else:
    print("Hai inserito un valore errato")
    continue
elif choise == str(2):
    giorno = int(input("Inserisci il giorno: "))
    if (giorno < 1 or giorno > 31):
        print("Giorno deve essere compreso tra 1 e 31")
        continue
    mese = int(input("Inserisci il mese: "))
    if (mese < 1 or mese > 12):
        print("Mese deve essere compreso tra 1 e 12")
        continue
    if (giorno == 31 and (
            mese != 1 or mese != 3 or mese != 5 or mese != 7 or mese != 8 or mese != 10 or mese != 12)):
        print("Il mese inserito non ha il giorno 31")
        continue
    if (giorno > 28 and mese == 2):
        print("Febbraio ha 28 giorni")
        continue
    csv.plot_daily_timeseries(int(giorno), int(mese))
    print("Digita:")
    choise2 = input("1 - Aggiungere anomalia visiva\n"
                   "2 - Per ottenere il grafico delle time series con un valore di affluenza minore uguale a quello "
                   "scelto in una specifica fascia oraria\n\n"
                   "3 - Tornare al menu principale\n")
    print(choise2)
    if(choise2==str(1)):
        csv.add_visual_anomalies(int(giorno),int(mese))
    elif(choise2==str(2)):
        cluster = int(input("Inserisci il cluster: "))
        if (cluster != 2 and cluster != 0 and cluster != 1):
            print("Cluster deve essere uguale a zero, uno o due")
            continue
        fascia = input("Inserisci la fascia oraria: h seguita da un intero compreso tra 1-23: ")
        valore = float(input("Inserisci valore affluenza compreso tra 0 e 1: "))
        if valore not in columns:
            print("Fascia oraria non valida")
        if (valore < 0 or valore > 1):
            print("Hai inserito un valore non valido")
            continue
        csv.plot_timeseries_with_h_value(cluster, fascia, valore)
    elif (choise2 == str(3)):
        continue
if choise == str(3):
    cluster = int(input("Inserisci il cluster: "))
    if (cluster != 2 and cluster != 0 and cluster != 1):
        print("Cluster deve essere uguale a zero, uno o due")
        continue
    print("Digita:")
    choise2 = input("1 - Per ottenere il grafico delle time series giornaliere raggruppate per cluster-anomalia\n"
                   "2 - Per ottenere il grafico delle time series giornaliere raggruppate per cluster-anomalia visiva\n")
    if(choise2==str(1)):
        csv.plot_timeseries_by_anomalous_and_cluster(int(cluster))
    else:
        csv.plot_timeseries_by_visiveanom_and_cluster(int(cluster))

if choise == str(4):
    break
if choise == str(5):
    csv.save_file(csv.save_file(input("Inserisci nome del file: ")))

```

Codice classe

```
from matplotlib import ticker
import pandas as pd
import matplotlib.pyplot as mat
import numpy as np
from sklearn.ensemble import IsolationForest
from sklearn.metrics import mean_squared_error
pd.options.mode.chained_assignment = None # default='warn'
class Model:
    def __init__(self, filename):
        self.csv_file= pd.read_csv(filename)
        self.columns = ['h1', 'h2', 'h3', 'h4', 'h5', 'h6', 'h7', 'h8', 'h9', 'h10', 'h11', 'h12', 'h13', 'h14', 'h15',
                        'h16','h17', 'h18', 'h19', 'h20', 'h21', 'h22', 'h23']
    def add_visual_anomalies_label(self,filename):
        if(filename=='timeSeries2015HotspotD.csv'):
            self.csv_file['Visive_anomalous']=0
    def add_visual_anomalies(self,day,month):
        self.csv_file.loc[(self.csv_file['Day'] == int(day)) & (self.csv_file['Month'] == int(month)), 'Visive_anomalous']=1

    def plot_timeseries_by_anomalous_and_cluster(self,cluster):
        mat.rcParams["figure.figsize"] = [20.8, 12.7]
        fig, ax = mat.subplots(2)
        ax[0].set_xlabel('ORARI', fontsize='medium')
        ax[0].set_ylabel('AFFLUENZA', fontsize='medium')
        ax[0].set_title('PLOT DELLE TIMESERIES GIORNALIERE DEL CLUSTER ' + str(cluster) + ' ANOMALE',
                         fontsize=15)
        ax[1].set_title('PLOT DELLE TIMESERIES GIORNALIERE DEL CLUSTER ' + str(cluster) + ' NON ANOMALE',
                         fontsize=15)
        ax[1].set_ylabel('AFFLUENZA', fontsize='medium')
        ax[1].set_xlabel('ORARI', fontsize='medium')
        data=self.csv_file
        data = data.loc[(self.csv_file['Anomalous'] == 0) & (data['Cluster'] == int(cluster))]
        num_it = len(data)
        i = 0
        while i < num_it:
            data = self.csv_file
            data = data.loc[(data['Anomalous'] == 0) & (data['Cluster'] == int(cluster))]
            data = data.iloc[i]
            # print(hotspot.loc[['Day','Month']])
            data = data.loc[self.columns]
            ax[0].plot(data)
            i += 1
        data=self.csv_file
```

```

data = data.loc[(data['Anomalous'] == 1) & (data['Cluster'] == int(cluster))]
num_it = len(data)
i = 0
while i < num_it:
    data = self.csv_file
    data = data.loc[(data['Anomalous'] == 1) & (data['Cluster'] == int(cluster))]
    data = data.iloc[i]
    # print(hotspot.loc[['Day', 'Month']])
    data = data.loc[self.columns]
    ax[1].plot(data)
    i += 1
self.plot_medium_form(int(cluster), ax[0])
self.plot_medium_form(int(cluster), ax[1])
mat.savefig("plot-timeseries-cluster-"+str(cluster)+".png")
mat.show()

def plot_timeseries_by_visiveanom_and_cluster(self, cluster):
    mat.rcParams["figure.figsize"] = [20.8, 12.7]
    fig, ax = mat.subplots(2)
    ax[0].set_xlabel('ORARI', fontsize='medium')
    ax[0].set_ylabel('AFFLUENZA', fontsize='medium')
    ax[0].set_title('PLOT DELLE TIMESERIES GIORNALIERE DEL CLUSTER ' + str(cluster) + ' VISIVAMENTE NON ANOMALE',
                     fontsize=15)
    ax[1].set_title('PLOT DELLE TIMESERIES GIORNALIERE DEL CLUSTER ' + str(cluster) + ' VISIVAMENTE ANOMALE',
                     fontsize=15)
    ax[1].set_ylabel('AFFLUENZA', fontsize='medium')
    ax[1].set_xlabel('ORARI', fontsize='medium')
    data=self.csv_file
    data = data.loc[(self.csv_file['Visive_anomalous'] == 0) & (data['Cluster'] == int(cluster))]
    num_it = len(data)
    i = 0
    while i < num_it:
        data = self.csv_file
        data = data.loc[(data['Visive_anomalous'] == 0) & (data['Cluster'] == int(cluster))]
        data = data.iloc[i]
        # print(hotspot.loc[['Day', 'Month']])
        data = data.loc[self.columns]
        ax[0].plot(data)
        i += 1

```

```

data=self.csv_file
data = data.loc[(data['Visive_anomalous'] == 1) & (data['Cluster'] == int(cluster))]
num_it = len(data)
i = 0
while i < num_it:
    data = self.csv_file
    data = data.loc[(data['Visive_anomalous'] == 1) & (data['Cluster'] == int(cluster))]
    data = data.iloc[i]
    # print(hotspot.loc[['Day', 'Month']])
    data = data.loc[self.columns]
    ax[1].plot(data)
    i += 1
self.plot_medium_form(int(cluster), ax[0])
self.plot_medium_form(int(cluster), ax[1])
mat.savefig("plot-timeseriesvisive-cluster-"+ str(cluster) +".png")
mat.show()

def plot_daily_timeseries(self, day, month):
    mat.rcParams["figure.figsize"] = [20.8, 9]
    mat.rcParams["xtick.minor.visible"] = True

    fig, ax = mat.subplots()
    ax.set_xlabel('ORARI', fontsize='medium')
    ax.set_ylabel('AFFLUENZA', fontsize='medium')

    data=self.csv_file
    data = data.loc[(data['Day'] == int(day)) & (data['Month'] == int(month))]
    data = data.iloc[0]
    cluster=data['Cluster']
    data = data.loc[self.columns]
    ax.set_title('PLOT DELLA TIME SERIES DEL GIORNO ' + str(day) + '/' + str(month) + '/2015 CLUSTER '
    + str(cluster) ,| fontsize=15)
    ax.plot(data)
    self.plot_medium_form(int(cluster),ax)
    mat.savefig("daily-timeseries-"+str(day)+"-"+str(month)+"-2015.png")
    mat.show()

def plot_medium_form(self,cluster,ax):

    data = self.csv_file
    data = data.loc[(data['Cluster'] == int(cluster))]
    num_it = len(data)

```

```

array = np.empty(23)
for columns_index in range(1, 24):
    column = 'h' + str(columns_index)
    x = np.array(data[[column]])
    media = 0;
    for i in x:
        media += i
    media = media / num_it
    array[columns_index - 1] = media

ax.plot(array,color='black', label="forma media", linewidth=3.5)
ax.legend(loc='upper right')

def plot_isolation_forest_default(self,cluster):
    mat.rcParams["figure.figsize"] = [20.8, 12.7]
    fig, ax = mat.subplots(2)
    ax[0].set_xlabel('ORARI', fontsize='medium')
    ax[0].set_ylabel('AFFLUENZA', fontsize='medium')
    ax[0].set_title('PLOT DELLE TIME SERIES NON ANOMALE DEL CLUSTER '+str(cluster)+ ' '
    'CHE ISOLATION FOREST CON PARAMETRI DI DEFAULT HA RICONOSCIUTO COME ANOMALE E '
    'PLOT DELLA TIME SERIES MEDIA DELLO STESSO CLUSTER ', fontsize=15)
    ax[1].set_title('PLOT DELLE TIME SERIES ANOMALE DEL CLUSTER '+str(cluster)+ ''
    ' CHE ANCHE ISOLATION FOREST CON PARAMETRI DI DEFAULT HA RICONOSCIUTO COME ANOMALE '
    'E PLOT DELLA TIME SERIES MEDIA DELLO STESSO CLUSTER ', fontsize=15)
    ax[1].set_ylabel('AFFLUENZA', fontsize='medium')
    ax[1].set_xlabel('ORARI', fontsize='medium')
    data=self.csv_file

    data = data.loc[(data['Cluster'] == int(cluster))]
    model = IsolationForest(n_estimators=100, max_samples='auto', contamination='auto',
                            bootstrap=False, n_jobs=-1,
                            random_state=0)
    model.fit(data[self.columns])
    data['isolation_forest_anom'] = model.predict(data[self.columns])

    data.to_csv('isolation-forest-cluster'+str(cluster)+'.csv')
    data_copy = data.loc[(data['isolation_forest_anom'] == -1) & (data['Anomalous'] == 0)
                        & (data['Cluster'] == int(cluster))]
    data_copy2 = data.loc[(data['isolation_forest_anom'] == -1) & (data['Anomalous'] == 1)
                        & (data['Cluster'] == int(cluster))]
    num_it = len(data_copy)

```

```

i = 0
while i < num_it:
    x = data_copy
    x = x.iloc[i]
    x = x.loc[self.columns]
    ax[0].plot(x)
    i += 1
num_it = len(data_copy2)

i = 0
while i < num_it:
    x = data_copy2
    x = x.iloc[i]
    x = x.loc[self.columns]
    ax[1].plot(x)
    i += 1

self.plot_medium_form(cluster, ax[0])
self.plot_medium_form(cluster, ax[1])
x=data['isolation_forest_anom'].values
c=len(x)
for i in range(0,c):
    if(x[i]==-1):
        x[i]=(1)
    else:
        x[i]=0

error = mean_squared_error(x, data['Anomalous'].values)
print("Prestazioni:" + str(error))
f = open('prestazioni-isolation-forest-cluster'+str(cluster)+'.txt', "a+")
f.write("Prestazioni:" + str(error)+"\n")
f.close()
mat.savefig('isolation-forest-cluster'+str(cluster)+'.png')
mat.show()

def plot_isolation_forest(self,cluster, n_estimators, max_samples, contamination,
                           max_features, bootstrap, random_state, n_jobs):
    mat.rcParams["figure.figsize"] = [20.8, 12.7]
    fig, ax = mat.subplots(2)
    ax[0].set_xlabel('ORARI', fontsize='medium')
    ax[0].set_ylabel('AFFLUENZA', fontsize='medium')
    ax[0].set_title('PLOT DELLE TIME SERIES NON ANOMALE DEL CLUSTER '+str(cluster)+ ' ')

```

```

'CHE ISOLATION FOREST CON PARAMETRI SCELTI HA RICONOSCIUTO COME ANOMALE '
'E PLOT DELLA TIME SERIES MEDIA DELLO STESSO CLUSTER ', fontsize=15)
ax[1].set_title('PLOT DELLE TIME SERIES ANOMALE DEL CLUSTER '+str(cluster)+'
'CHE ANCHE ISOLATION FOREST CON PARAMETRI SCELTI HA RICONOSCIUTO COME ANOMALE'
' E PLOT DELLA TIME SERIES MEDIA DELLO STESSO CLUSTER ', fontsize=15)
ax[1].set_ylabel('AFFLUENZA', fontsize='medium')
ax[1].set_xlabel('ORARI', fontsize='medium')
data=self.csv_file

data = data.loc[(data['Cluster'] == int(cluster))]
model = IsolationForest(n_estimators=n_estimators, max_samples=max_samples,
                         contamination=contamination , bootstrap=bootstrap, n_jobs=n_jobs,
                         random_state=random_state)
model.fit(data[self.columns])
data['isolation_forest_anom'] = model.predict(data[self.columns])

data.to_csv('isolation-forest-cluster'+str(cluster)+'-estim'+str(n_estimators)+
            '-sampl'+str(max_samples)+'-cont'+str(contamination)+
            '-boots'+str(bootstrap)+'-jobs'+str(n_jobs)+'-randstate'+
            str(random_state)+'.csv')
data_copy = data.loc[(data['isolation_forest_anom'] == -1) & (data['Anomalous'] == 0)
                     & (data['Cluster'] == int(cluster))]
data_copy2 = data.loc[(data['isolation_forest_anom'] == -1) & (data['Anomalous'] == 1)
                     & (data['Cluster'] == int(cluster))]
num_it = len(data_copy)

i = 0
while i < num_it:
    x = data_copy
    x = x.iloc[i]
    x = x.loc[self.columns]
    ax[0].plot(x)
    i += 1
num_it = len(data_copy2)

i = 0
while i < num_it:
    x = data_copy2
    x = x.iloc[i]
    x = x.loc[self.columns]
    ax[1].plot(x)
    i += 1

```

```

    self.plot_medium_form(cluster, ax[0])
    self.plot_medium_form(cluster, ax[1])
    c = len(x)
    for i in range(0, c):
        if (x[i] == -1):
            x[i] = (1)
        else:
            x[i] = 0

    error = mean_squared_error(x, data['Anomalous'].values)
    print("Prestazioni:"+str(error))
    f = open('prestazioni-isolation-forest-cluster'+str(cluster)+'-estim'+
              str(n_estimators)+'-sampl'+str(max_samples)+'-cont'+str(contamination)+
              '-boots'+str(bootstrap)+'-jobs'+str(n_jobs)+'-randstate'+
              +str(random_state)+'.txt', "a+")
    f.write("Prestazioni:"+str(error)+"\n")
    f.close()
    mat.savefig('isolation-forest-cluster'+str(cluster)+'-estim'+str(n_estimators)+
                +'-sampl'+str(max_samples)+'-cont'+str(contamination)+
                '-boots'+str(bootstrap)+'-jobs'+str(n_jobs)+'-randstate'+
                +str(random_state)+'.png')
    mat.show()

def save_file(self, filename):
    self.csv_file.to_csv(filename+'.csv')
def plot_timeseries_with_h_value(self,cluster,fascia,valore):
    mat.rcParams["figure.figsize"] = [20.8, 9]
    mat.rcParams["xtick.minor.visible"] = True

    fig, ax = mat.subplots()
    ax.set_xlabel('ORARI', fontsize='medium')
    ax.set_ylabel('AFFLUENZA', fontsize='medium')

    data = self.csv_file
    data = data.loc[(data[fascia] <= float(valore)) & (data['Cluster'] == int(cluster))]
    num_it=len(data)
    i = 0
    x=[]
    while i < num_it:
        data = self.csv_file
        data = data.loc[(data[fascia] <= float(valore))& (data['Cluster'] == int(cluster))]

```

```

data = data.loc[(data[fascia] <= float(valore))& (data['Cluster'] == int(cluster))]
data = data.iloc[i]
x.append(data.loc[['Day', 'Month']].values)
data = data.loc[self.columns]
ax.plot(data)
i+=1
ax.set_title('PLOT DELLE TIME SERIES CON VALORE AFFLUENZA MINORE DI ' +
             str(valore) + ' NELLA FASCIA ORARIA ' + str(fascia) + ' CLUSTER ' +
             str(cluster) , fontsize=15)
print("Giorni: \n")
for a in x:
    print(a)
ax.plot(data)
mat.savefig("time-series-cluster"+str(cluster)+"-fasciaoraria"+str(fascia)+
            "-less-value"+str(valore)+".png")
mat.show()

```

BIBLIOGRAFIA E SITOGRADIA

- [1] A. L. Alfeo, M. G. C. A. Cimino, S. Egidi, B. Lepri, A. Pentland, and G. Vaglini, "Stigmergybased modeling to discover urban activity patterns from positioning data", in proceedings of "Social, Cultural, and Behavioral Modeling: 10th International Conference", (SBP-BRIMS 2017), vol. 10354, p.p. 292-302. Springer. Washington, DC, USA, July 5-8, 2017.
- [2] https://en.wikipedia.org/wiki/Anomaly_detection
- [3] Chandola, V., Banerjee, A., and Kumar, V. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. DOI = 10.1145/1541880.1541882 <http://doi.acm.org/10.1145/1541880.1541882>.
- [4] https://en.wikipedia.org/wiki/Isolation_forest
- [5] <https://blog.paperspace.com/anomaly-detection-isolation-forest/>
- [6] https://en.wikipedia.org/wiki/January_2015_North_American_blizzard
- [7] https://www.123rf.com/photo_114717719_new-york--march-19-2015-cars-taxi-cabs-and-people-rushing-on-busy-streets-of-downtown-manhattan-duri.html, <https://www.thrillist.com/news/new-york/snow-in-nyc-on-the-first-day-of-spring-2015>
- [8] <https://www.tribecafilm.com/festival/faq> - Il Tribeca Film Festival è una rassegna internazionale di film indipendenti ideata e diretta da Robert De Niro, all'indomani dell'11 settembre (Italiafilm.org). La manifestazione ha luogo in vari cinema e in vari teatri tra i quali il Village East Cinema (181-189 2nd Ave, New York, NY 10003) che si trova all'interno dell'hotspot D
- [9] <http://it.clearharmony.net/articles/a114367-New-York-Oltre-8000-praticanti-della-Falun-Dafa-hanno-partecipatoalla-grande-parata-di-Manhattan.html#.XskMTGgzaUk> - Venerdì 14 Maggio 2015, oltre 8.000 praticanti della Falun Dafa provenienti da più di 50 paesi hanno preso parte alla grande manifestazione con sfilata di Manhattan, per celebrare il 23° anniversario dell'introduzione al pubblico della Falun Dafa. La sfilata è iniziata a mezzogiorno, partendo dalla piazza delle Nazioni Unite in Dag Hammarskjold Plaza che si trova nell'hotspot D (Murray H ill)
- [10] <http://murphguide.com/outdoor-dining/feast-of-san-gennaro-2015/> - La **festa di san Gennaro** è una fiera che si tiene verso la metà di settembre a New York. La festa, originariamente, era una commemorazione religiosa, mentre, ora, è la celebrazione del rapporto tra italiani e statunitensi. L'edizione del 2015 è iniziata giorno 10 e come si può vedere dal sito include aree e posti che si trovano all'interno dell'hotspot D, come ristoranti e bar che danno su Houston st. (Estela, Milano's ecc.)
- [11] <https://www.google.com/maps/d/viewer?mid=1rMMhvsdXEeqBgx0dRZOdN7foGll8&ll=40.729955774341875%2C-73.97560538305734&z=15> - Il santaCon è un pub crawl annuale in cui le persone vestite da babbo Natale o da altri personaggi natalizi sfilano in varie città del mondo, tra cui New York. Il santacon del 12 dicembre di New York è stato parecchio movimentato in quanto la polizia ha dovuto gestire più di 100 sommosse (Wikipedia). Nel link possiamo trovare la serie di sedi del santaCon 2015 e si può notare che molte di esse risiedono all'interno dell'Hotspot D.
- [12] <https://www.startribune.com/how-the-night-before-thanksgiving-became-the-biggest-drinking-day-of-the-year/459149303/> - La sera prima del Ringraziamento, noto anche come "Drinksgiving", è diventata una delle più grandi vacanze per bere dell'anno. I festeggiamenti coincidono con quello che è un periodo di traffico molto intenso.
- [13] <https://nypost.com/2015/01/09/motorists-angry-over-alternative-side-parking-rules-during-snowfall/>
- [14] <https://www.dnainfo.com/new-york/20150116/midtown-east/worlds-top-squash-players-battle-grand-central-tournament/>
- [15] <https://www.ilpost.it/2015/02/21/foto-ghiaccio-manhattan/>
- [16] https://en.wikipedia.org/wiki/Early_March_2015_North_American_winter_storm
- [17] <https://www.ny1.com/nyc/all-boroughs/news/2015/10/2/city-prepares-for-hurricane-joaquin-s-worst>
- [18] <https://www.currentresults.com/Yearly-Weather/USA/NY/New-York-City/extreme-annual-new-york-city-precipitation.php>
- [19] https://en.wikipedia.org/wiki/Mid-February_2015_North_American_blizzard
<https://earthobservatory.nasa.gov/images/85317/another-blizzard-piles-up-the-snow-in-new-england>
- [20] <http://www.turismoitalianews.it/i-luoghi-piu-divertenti/10228-capodanno-cinese-tre-quartieri-per-festeggiare-a-new-york-city-chinatown-a-manhattan-flushing-nel-queens-e-sunset-park-a-brooklyn>
- Una sfilata di carri allegorici, bande musicali, leoni e draghi danzanti, maghi e acrobati, per dare il benvenuto all'anno della pecora che si svolge a chinatown e little italy, che si trovano in basso nell'hotspot D
- [21] https://en.wikipedia.org/wiki/Early_March_2015_North_American_winter_storm
- [22] <https://www.dailymail.co.uk/news/article-3104812/NYPD-Material-lifted-crane-building-falls.html>
- [23] https://en.wikipedia.org/wiki/Urban_science