

Report task di tesi

Confronto tra una time-series normale e una anomala

Leonardo Poggiani

October 10, 2020

Contents

1	Motivazione	1
2	Problema	2
3	Risultati	4
4	Conclusioni	10

Abstract

Nel task proposto veniva richiesto di individuare il numero massimo di giorni normali consecutivi, intesi come giorni aventi il valore di anomalous non settato.

Questo viene fatto nell'ottica di predire le giornate anomale, confrontando la time-series di una giornata normale predetta (facilmente predicibile in quanto il numero di giornate normali supera molto il numero di giornate anomale) con quella di una giornata anomala.

Infatti il numero massimo di giornate normali consecutive (chiamato N) sarà la dimensione della finestra che verrà usata per la predizione della serie temporale della giornata $(N+1)$ -esima, che in realtà è una giornata anomala.

A questo punto è possibile confrontare le due time-series per ricavare delle informazioni dalle loro differenze.

Chapter 1

Motivazione

Al fine di riuscire a distinguere e quindi predire le giornate anomale all'interno del dataset fornito, è necessario predire la serie temporale di una giornata "normale" (con valore di anomalous pari a 0).

Questo è agevolato dal fatto nel dataset in questione sono presenti molte più giornate normali rispetto a giornate anomale (12 anomalie su 209 giorni presenti).

Questo consentirà di predire facilmente l'andamento di una giornata normale così da poterlo comparare alla serie temporale di una giornata anomala.

In questa prima fase si farà solamente un'analisi visiva, mediante dei grafici, delle differenze tra le due serie temporali.

Chapter 2

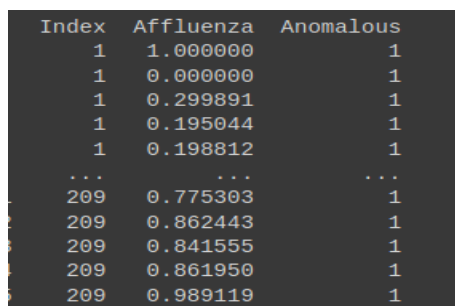
Problema

Il dataset fornito è stato portato in un formato che rende agevole le manipolazioni che si andranno a compiere su di esso.

La prima colonna del dataset ("Index") rappresenta il numero del giorno in questione e il suo valore è ripetuto per tutte e 23 le entrate del giorno corrispondente.

La seconda colonna è quella dell'Affluenza e rappresenta il dato di affluenza oraria.

La terza colonna riporta il valore di anomalous, replicato per ogni entrata.



Index	Affluenza	Anomalous
1	1.000000	1
1	0.000000	1
1	0.299891	1
1	0.195044	1
1	0.198812	1
...
209	0.775303	1
209	0.862443	1
209	0.841555	1
209	0.861950	1
209	0.989119	1

Figure 2.1: Formato del dataset preprocessato

Questo formato ci consente di effettuare le manipolazioni volute semplicemente aggiungendo un indice incrementale, che consente una più semplice realizzazione dei grafici.

A questo punto è necessario individuare la dimensione della finestra ($N - 1$) per la predizione della serie temporale della giornata normale.

Dopo aver individuato il dataframe "target", ovvero quello composto dalle entrate relative alle $N - 1$ giornate normali precedenti alla giornata anomala N si effettua la predizione mediante una rete ricorrente LSTM.

Questo viene fatto ciclicamente per tutte le giornate anomale presenti nel dataset.

	Index	Affluenza	Anomalous
0	1	1.000000	1
1	1	0.000000	1
2	1	0.299891	1
3	1	0.195044	1
4	1	0.198812	1
...
4801	209	0.775303	1
4802	209	0.862443	1
4803	209	0.841555	1
4804	209	0.861950	1
4805	209	0.989119	1

Figure 2.2: Formato del dataset preprocessato con indice incrementale

Chapter 3

Risultati

Per raggiungere l'obiettivo prefissato si è calcolata la dimensione della finestra di predizione iterando sulle tuple del dataset e attraverso un semplice conto si è trovato il numero massimo di giorni consecutivi.

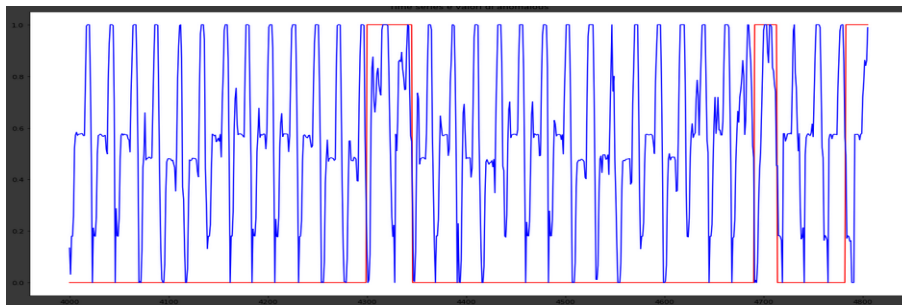


Figure 3.1: Serie temporali con i giorni anomali evidenziati

Data l'evidente differenza tra una serie temporale anomala e una normale, si procede quindi a predire l'andamento della giornata "normale" per poterla confrontare con quella (non predetta) della giornata anomala.

A questo punto per ogni giorno anomalo viene effettuata la predizione, con le N giornate normali precedenti e viene poi plottato il confronto.

Nella figura 3.11 vengono illustrate le azioni intraprese per predire ciclicamente l'andamento della finestra di N giorni normali.

Si fa un'iterazione su tutte le righe del dataset, conservando le righe già controllate in modo da non doverle più predire.

Vengono calcolati i giorni normali disponibili per un dato giorno anomalo andando a contare le righe relative nel dataset. Se questi sono sufficienti per la formazione della finestra (almeno maggiori a $N*23$) e la riga è anomala, si individua il dataframe target.

Il dataframe target è composto unicamente da giorni normali e su di questo

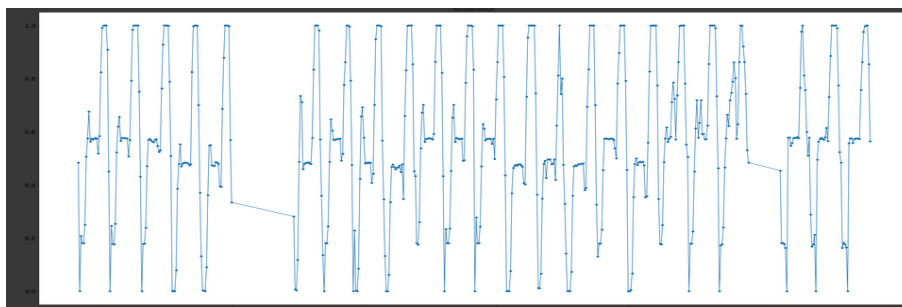


Figure 3.2: Solo le serie temporali normali, gli spazi vuoti sono le giornate anomale

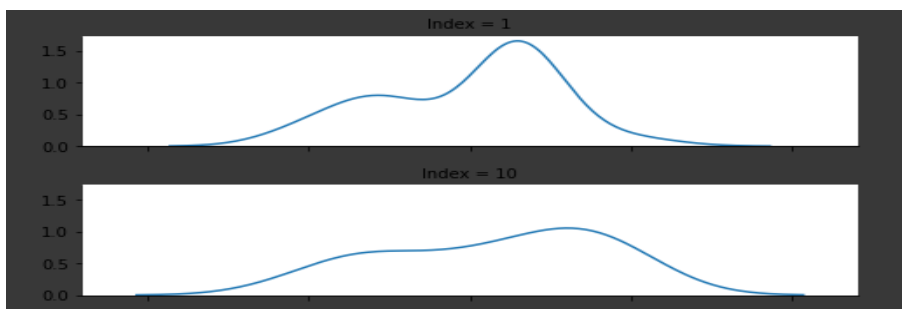


Figure 3.3: Solo le serie temporali anomale, giorno per giorno

viene fatta la predizione.

Infine vengono plottati i risultati comparando le serie temporali.

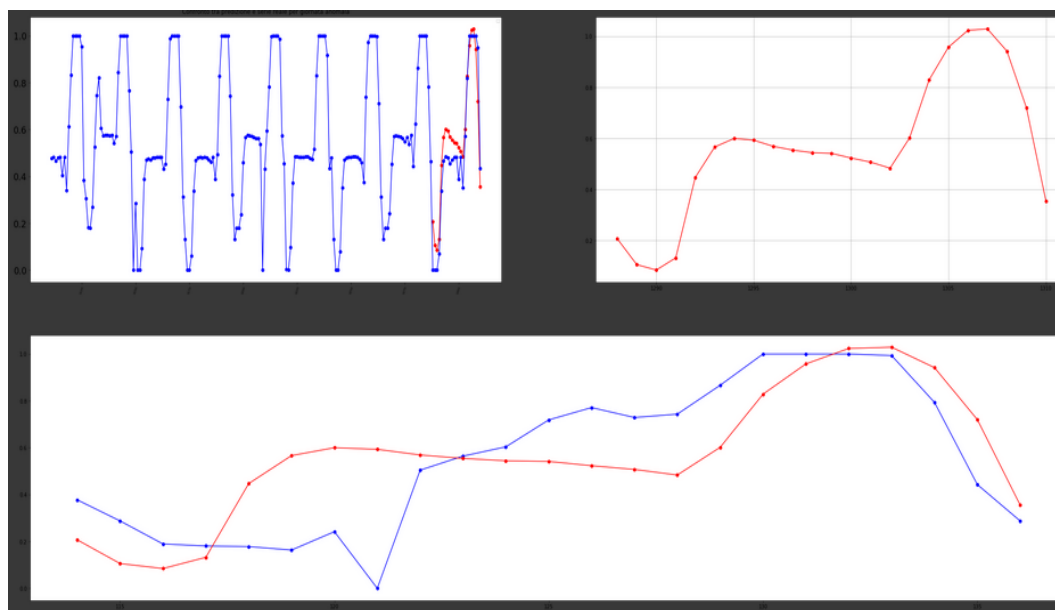


Figure 3.4: In ordine: in alto a sinistra il confronto tra serie temporali normali del df target (blu) e quella predetta (in rosso); in alto a destra la serie predetta; in basso il confronto tra la serie predetta (rosso) e la serie anomala di riferimento (blu)

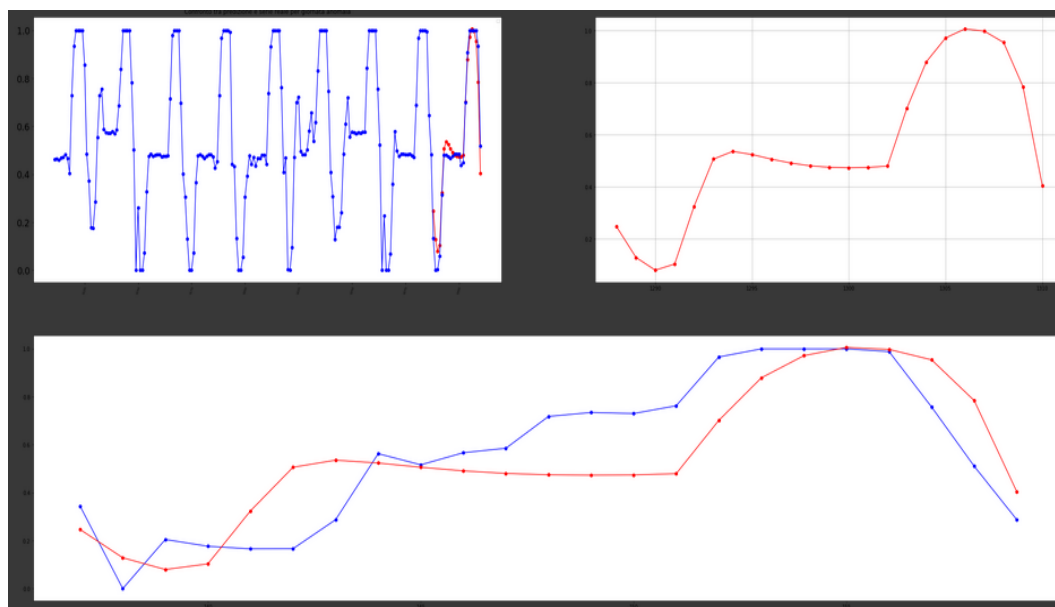


Figure 3.5: Come sopra ma per una giornata anomala diversa

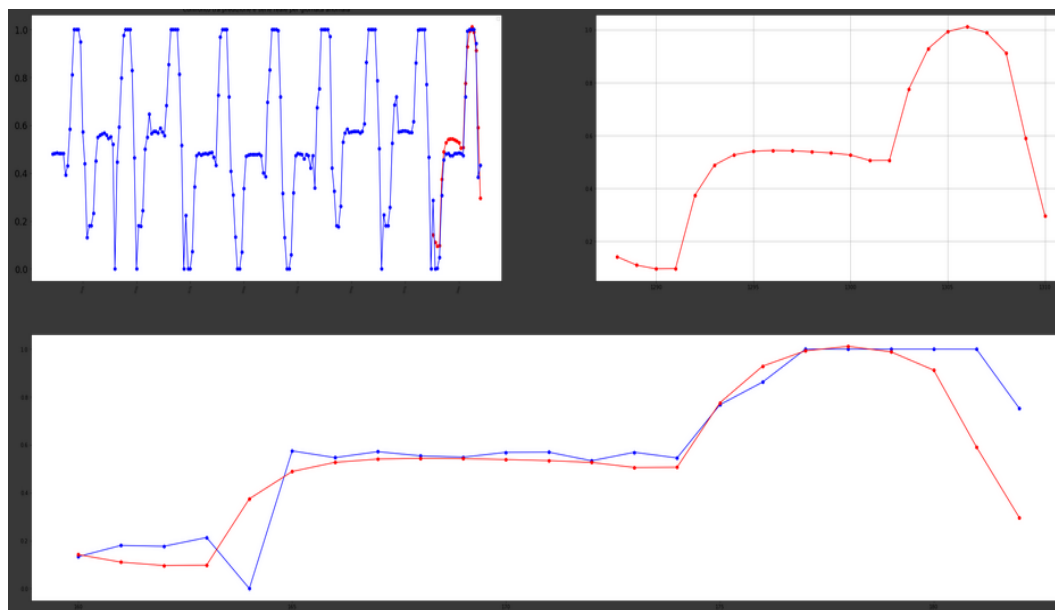


Figure 3.6: Come sopra ma per una giornata anomala diversa

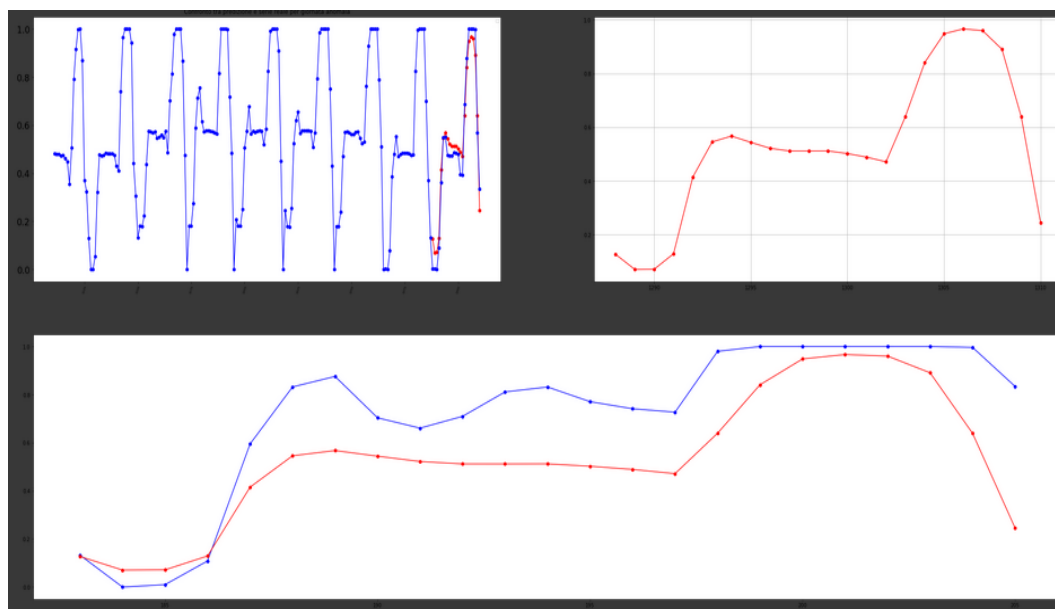


Figure 3.7: Come sopra ma per una giornata anomala diversa

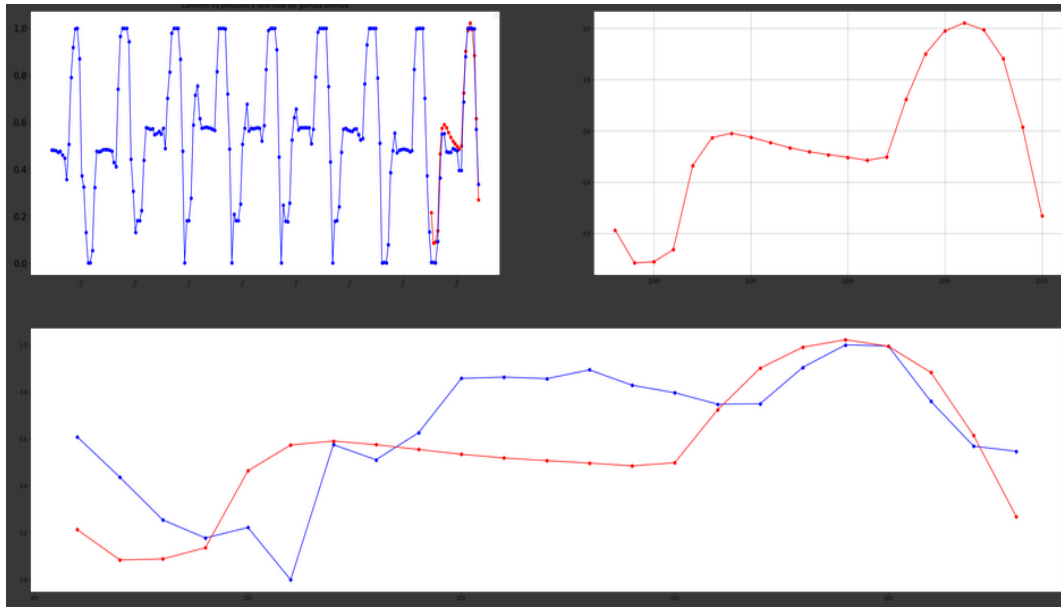


Figure 3.8: Come sopra ma per una giornata anomala diversa

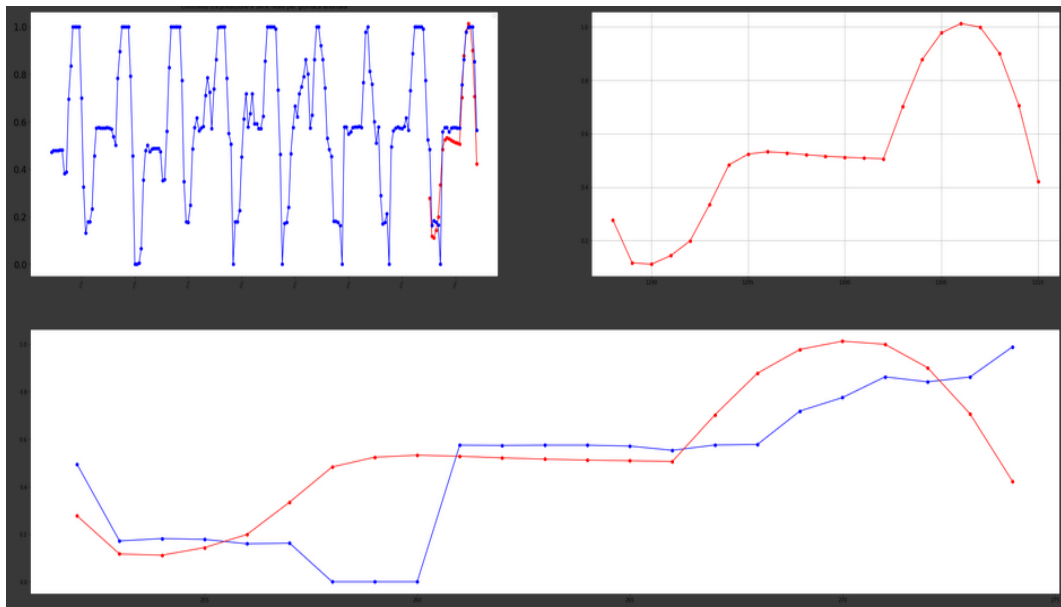


Figure 3.9: Come sopra ma per una giornata anomala diversa

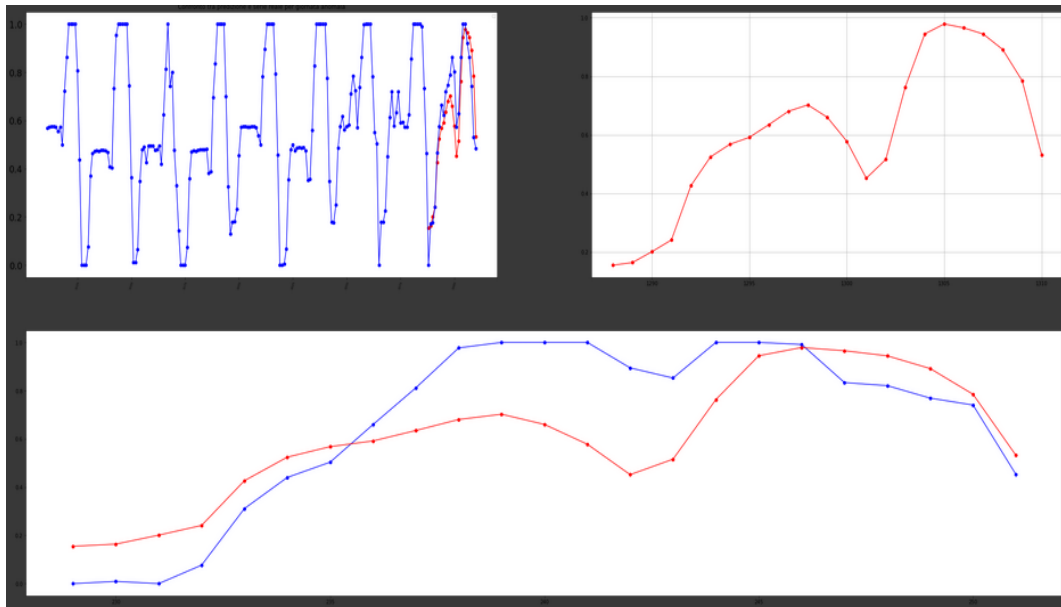


Figure 3.10: Come sopra ma per una giornata anomala diversa

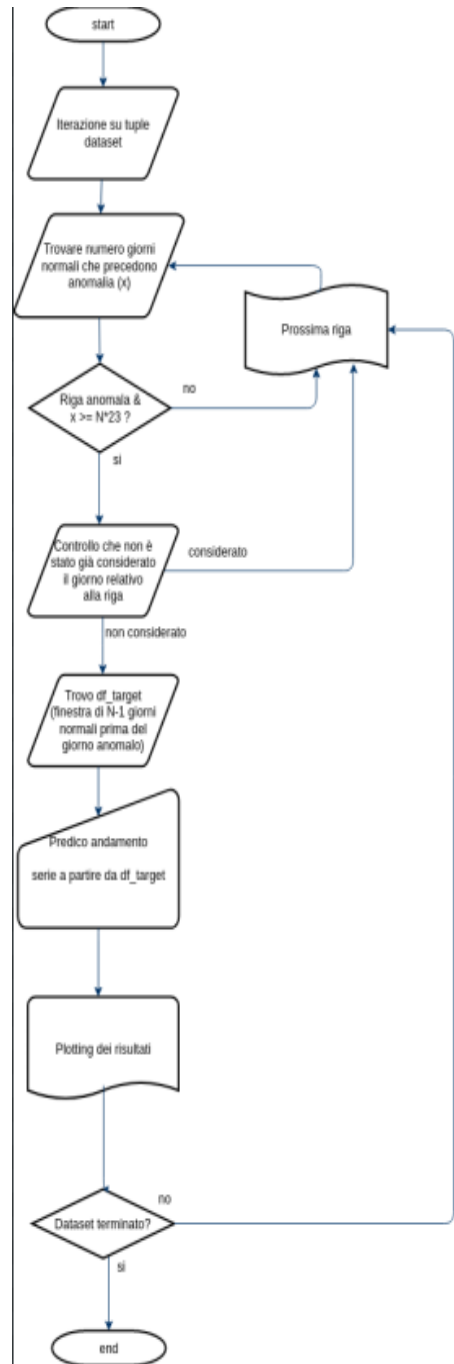


Figure 3.11: Diagramma di flusso che illustra le azioni svolte per la predizione ciclica

Chapter 4

Conclusioni

Osservando le immagini riportate sopra (Figura 3.5 e 3.6) si possono trarre delle conclusioni, in particolare osserviamo che le predizioni fatte sulla base della finestra di giorni normali che precedono il giorno anomalo è abbastanza precisa e si discosta poco dalle serie temporali normali.

Va notato che le immagini che sono riportate sono relative ai giorni anomali per cui si potesse creare la finestra di $(N-1)$ giorni.

Inoltre possiamo notare che le serie temporali anomale si discostano evidentemente dalle serie normali.

Questo può rivelarsi molto utile nell'ottica di predire le giornate anomale in base alla differenza tra queste serie.

Si può osservare che le predizioni risultano sempre piuttosto simili tra di loro e questo è dovuto alla prevalenza di giorni normali rispetto alle anomalie, come già detto in precedenza e soprattutto le predizioni sono molto somiglianti alle serie temporali reali.