Università degli studi di
Milano - Bicocca

Streaming Data Management
&
Time Series Analysis

Final Project

Appliance Energy Prediction

Alex Ceccotti
790497
a.ceccotti@campus.unimib.it

January 27, 2019

Abstract

Per le imprese che operano nel mercato dell'energia è importante riuscire a predirre i consumi futuri. In questo elaborato, utilizzando una serie storica relativa ai consumi di energia elettrica per elettrodomestici, si cercherà di fare forecasting di tale serie temporale utilizzando modelli statistici (SARIMA e UCM) e di machine learning (LSTM).

1 Introduzione

Per le aziende produttrici di energia è importante riuscire a prevedere in anticipo il consumo elettrico. Ad oggi vengono quindi spesso implementati modelli predittivi con il fine di evitare sovrapproduzioni di energia (che implicano costi di stoccaggio) o, viceversa, sottoproduzioni di energia (che implicano l'acquisto di essa da altri produttori).

1.1 Obiettivo

L'obiettivo di questo elaborato è l'analisi del consumo elettrico orario riferito agli elettrodomestici di un palazzo. Il fine ultimo è quello di riuscire a predirre i valori futuri della serie storica di tali consumi tramite alcuni modelli predittivi.

2 Dataset

Il dataset utilizzato per l'analisi, disponibile al link `https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction`, è composto da diverse serie storiche con frequenza di campionamento di 10 minuti per un periodo di quattro mesi e mezzo. Nello specifico, ci sono le serie storiche relative ai consumi elettrici per elettrodomestici e per l'illuminazione, oltre ad altre serie temporali relative a dati ambientali e meteorologici. Durante questo elaborato si farà riferimento esclusivamente alla serie dei consumi elettrici per elettrodomestici.

2.1 Data Manipulation

Dal momento che la frequenza di campionamento dei dati è di 10 minuti, cosa che implica una grossa mole di dati, è stato scelto di aggregare i dati

per ora (facendo la somma dei consumi). La serie storica ottenuta a seguito di questa operazione è rappresentata in figura 1. Tuttavia, è stata necessaria un'ultima trasformazione della serie temporale: dal momento che esiste una relazione lineare tra media e varianza della serie (se calcolate ogni 24 ore), è stato applicato il logaritmo. Il risultato così ottenuto è rappresentato in figura 2.

3 Approccio Metodologico

A partire dal logaritmo della serie originale, sono stati sviluppati e validati 3 modelli predittivi diversi: un modello SARIMA, un modello UCM e una rete neurale basata su blocchi LSTM. Per i modelli SARIMA e UCM il dataset è stato diviso in training set (primi 3 mesi) e test set (ultimo mese e mezzo). Per la rete neurale, il training set è stato a sua volta diviso in training set (primi 2 mesi) e validation set (mese restante).

3.1 SARIMA

Il modello SARIMA utilizzato prevede un modello AR(2) più un modello stagionale ARIMA(1,1,1) con periodo 24. Tale modello è stato ottenuto osservando i correlogrammi e modellando la serie in modo iterativo. I correlogrammi dei residui del modello così ottenuto (figura 3) indicano che tutta la correlazione è stata opportunamente modellata. Il grafico in figura 4 mette a confronto le predizioni sul test set (in rosso) e i dati reali (in blu). Come si può notare, il modello coglie abbastanza bene l'andamento della serie, ma non riesce a seguire i picchi di consumo.

3.2 UCM

Il modello UCM utilizzato presenta una componente trend e una componente stagionale. La componente trend è stata imposta senza slope e con varianza di η pari a zero. Nella pratica, questo implica una semplice stima della media della serie. La componente stagionale invece è stata sviluppata tramite 12 sinusoidi (dato che il periodo è 24). Per questo motivo è stato necessario stimare la varianza dei 23 white noise relativi alla componente stagionale. Successivamente, utilizzando il filtro di Kalman, sono state fatte le predizioni di tale modello sui dati di training e di test. Il grafico in figura 5 mette a

confronto le predizioni sul test set (in rosso) e i dati reali (in blu). Anche in questo il caso, nonostante il modello colga abbastanza bene l'andamento della serie, non riesce a seguire i picchi di consumo.

3.3 LSTM

Per poter sviluppare adeguatamente una rete neurale capace di prevedere i consumi orari, è stato necessario manipolare leggermente i dati. Nello specifico, dalla serie di partenza è stata creata una matrice con 26 variabili:

- la serie originale
- la serie ritardata di 1 periodo
- la serie ritardata di 2 periodi
- ...
- la serie ritardata di 24 periodi
- la serie 23 periodi avanti

Sono stati dunque inseriti, per ogni periodo, i lag fino al ventiquattresimo. Inoltre, è stato deciso di predirre anche la serie 23 periodi avanti rispetto quella originale. Potrebbe infatti essere utile per un'impresa produttrice di energia avere una prima stima dei consumi con 24 ore di anticipo, salvo poi avere una stima più accurata l'ora prima. Infine, i dati di tutte le serie sono stati normalizzati nel range (-1,1). Ovviamente, in fase di valutazione, i dati predetti sono stati riportati nel range originale. L'architettura della rete LSTM implementata è illustrata in figura 6. Sono dunque presenti due strati nascosti (con rispettivamente 10 e 5 blocchi LSTM) prima di arrivare allo strato di output con 2 neuroni (con funzione di attivazione lineare).
Il grafico in figura 7 mette a confronto le predizioni sul validation set e sul test set (rispettivamente in rosso tratteggiato e in rosso) e i dati reali (rispettivamente in verde e in blu) della serie originale. Il grafico in figura 8 mette invece a confronto le predizioni sul validation set e sul test set (rispettivamente in rosso tratteggiato e in rosso) e i dati reali (rispettivamente in verde e in blu) della serie originale portata avanti di 23 periodi. Si può notare come le predizioni un'ora in avanti seguono bene sia l'andamento generale, sia i picchi di consumo, mentre le predizioni 24 ore in avanti riescono solo a seguire l'andamento della serie, ignorando i picchi di consumo.

4 Risultati e Valutazioni

La misura di errore considerata per valutare le predizioni è il mean square error (MSE). La scelta del MSE è motivata dal fatto che, rispetto altre misure di errore come ad esempio il mean absolute error (MAE), tramite questa misura i grossi errori commessi sui picchi di consumo assumono un peso maggiore . Inoltre, per ogni modello, il MSE è stato rapportato con il MSE del modello nullo, ovvero quel modello che prevede la media dei dati di training per ogni periodo. Il rapporto tra MSE del modello e MSE del modello nullo sarà da ora in poi chiamato "MSE relativo". Per quanto riguarda il modello SARIMA, il MSE relativo sui dati di test (non è possibile fare predizioni sui dati di training) è risultato essere pari a 0.705. Il modello UCM ha ottenuto risultati simili, con MSE relativo sui dati di training pari a 0.628 e MSE relativo sui dati di test pari a 0.709. La rete LSTM ha invece ottenuto risultati diversi a seconda del periodo da prevedere. Le predizioni 1 ora in avanti hanno riscontrato un MSE relativo pari a 0.418 sul training set, 0.404 sul validation set e 0.463 sul test set. Le predizioni 24 ore in avanti hanno invece riscontrato un MSE relativo pari a 0.613 sul training set, 0.701 sul validation set e 0.683 sul test set. Si può notare dunque come sia il modello UCM sia la rete LSTM overfittano leggermente il training set, ottenendo comunque risultati accettabili in fase di test.

5 Conclusioni

Indubbiamente, il modello che performa meglio in fase di test è la rete neurale LSTM (per predizioni 1 ora in avanti). Tuttavia, per una società che deve decidere l'ammontare di energia da produrre, avere una predizione dei consumi energetici con una sola ora di anticipo potrebbe essere problematico. Dati i modelli e i risultati ottenuti, la strategia vincente potrebbe essere un utilizzo combinato di entrambi gli output della rete. Le predizioni 24 ore in avanti servirebbero per ottenere una prima stima dei consumi del giorno seguente, mentre le predizioni 1 ora in avanti verrebbo utilizzate per avere una stima più affidabile dei consumi dell'ora successiva.

6 Appendice

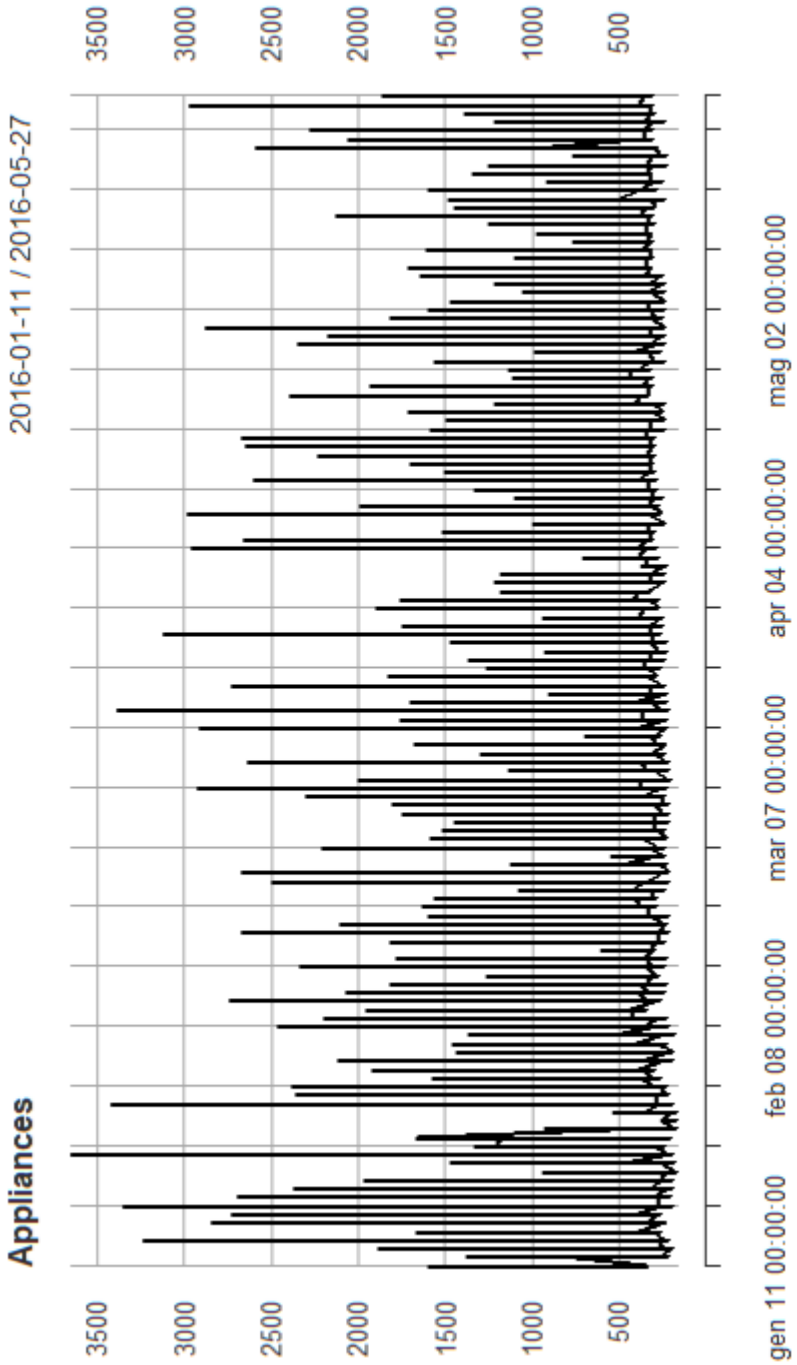

Figure 1: Appliances Time Series

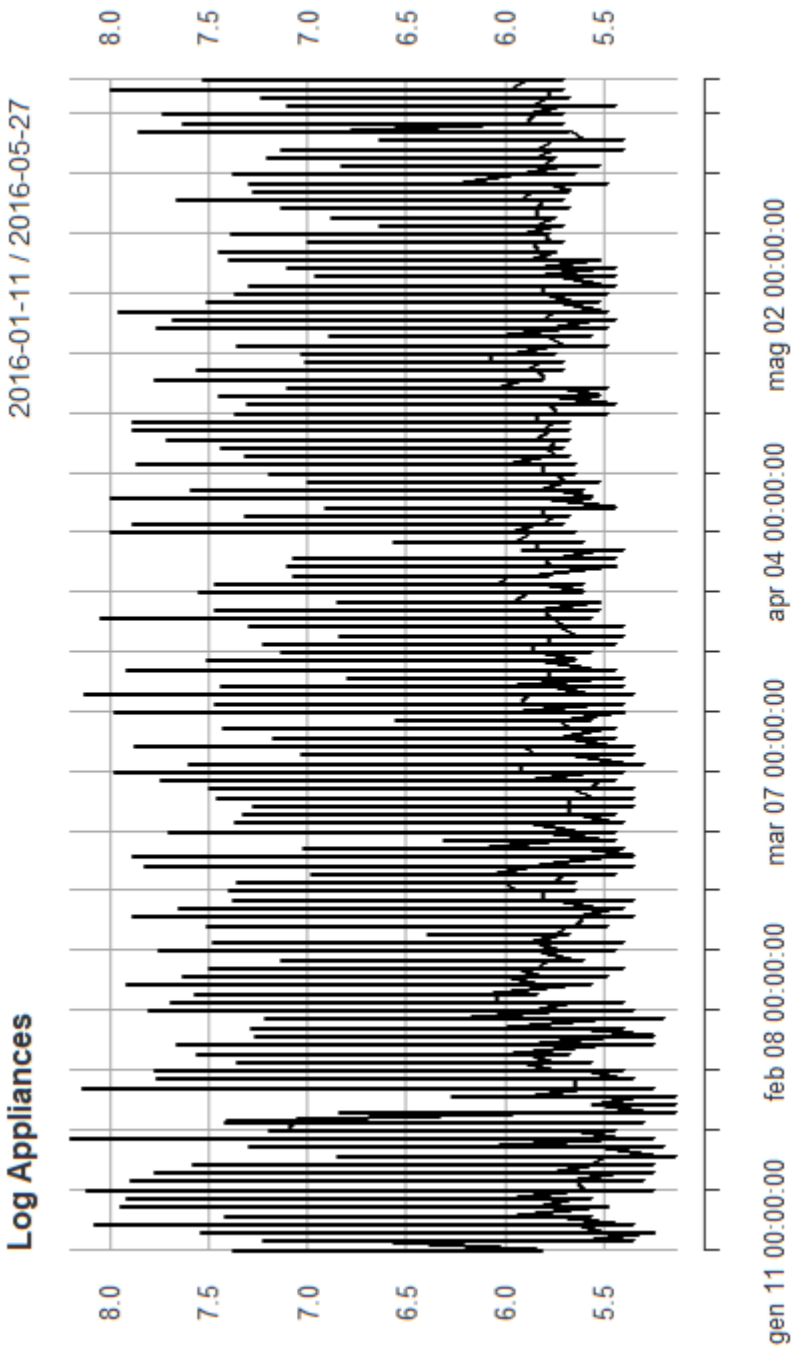

Figure 2: Log Appliances Time Series

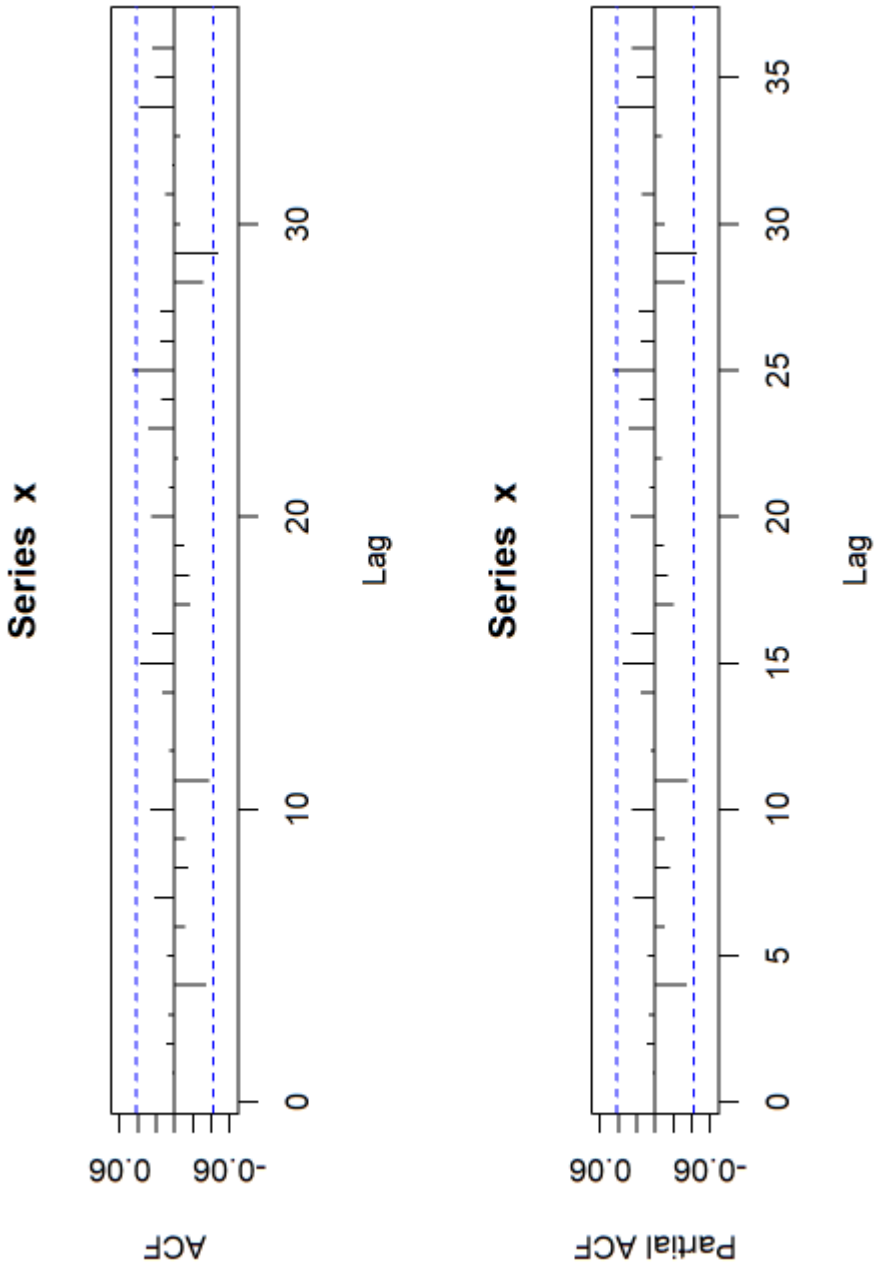

Figure 3: Correlogramma dei residui

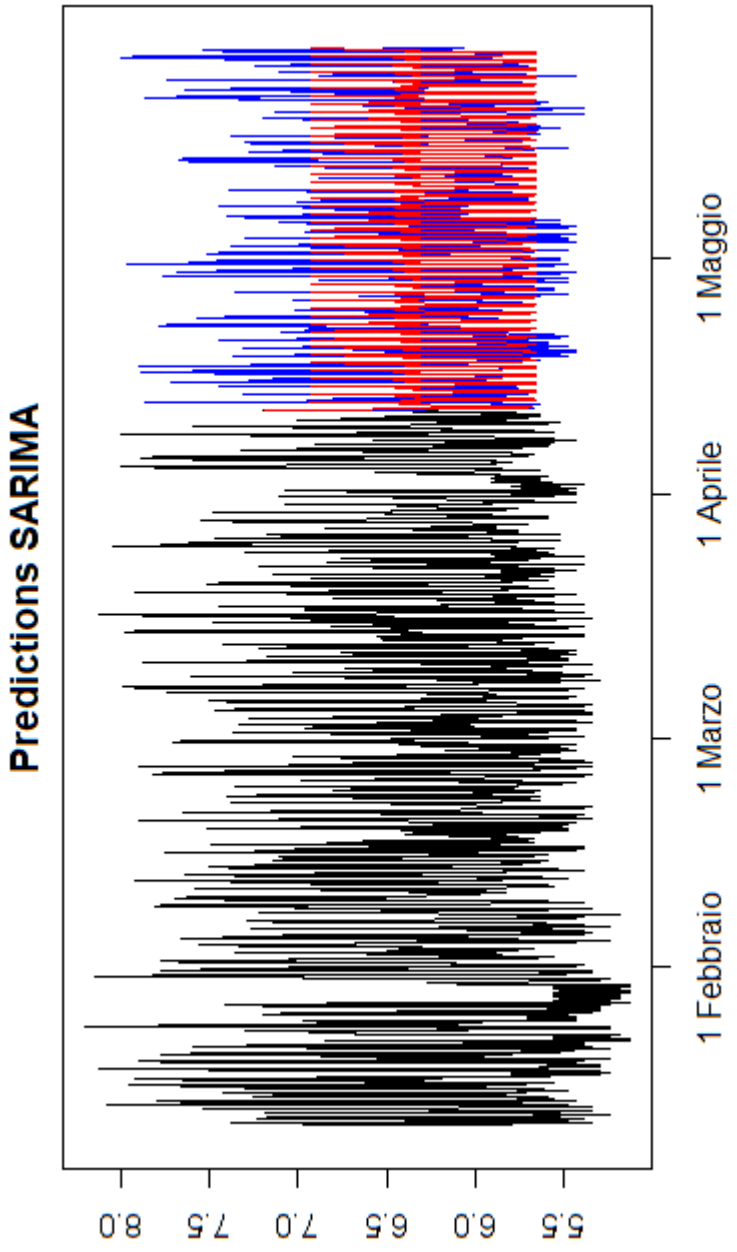

Figure 4: Predizioni SARIMA

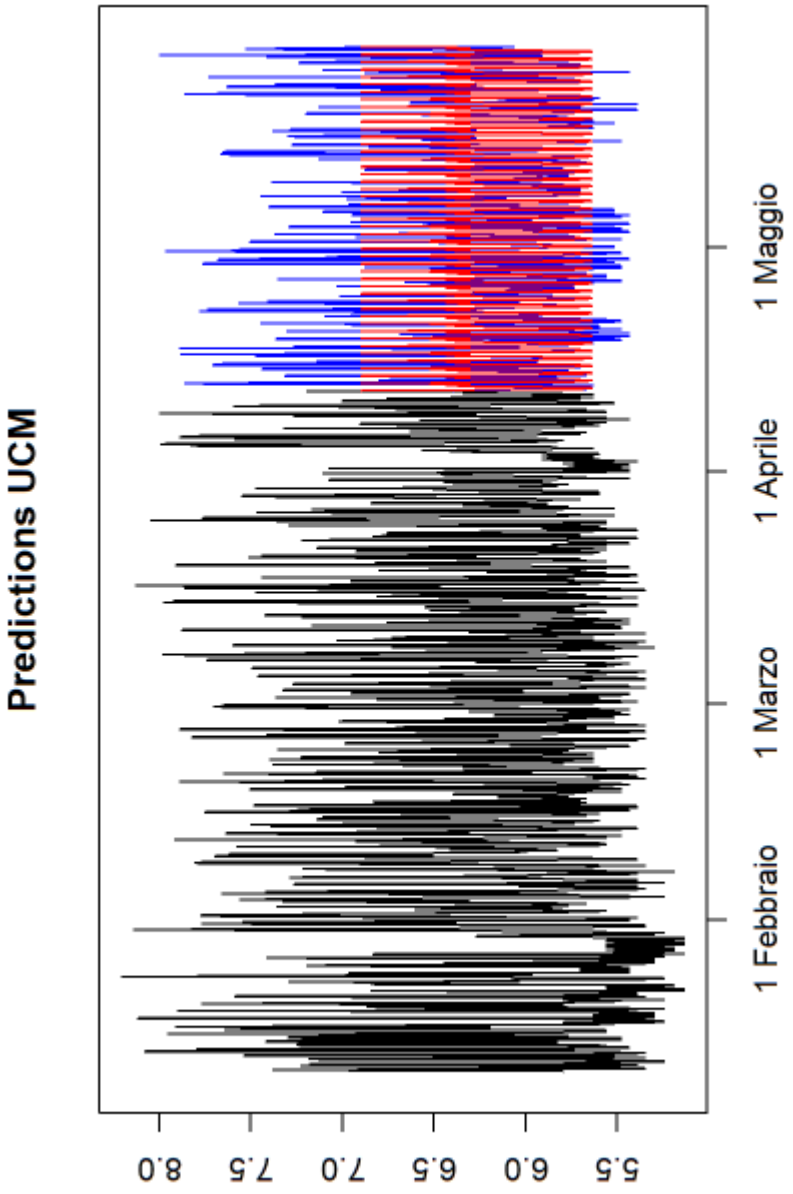

Figure 5: Predizioni UCM

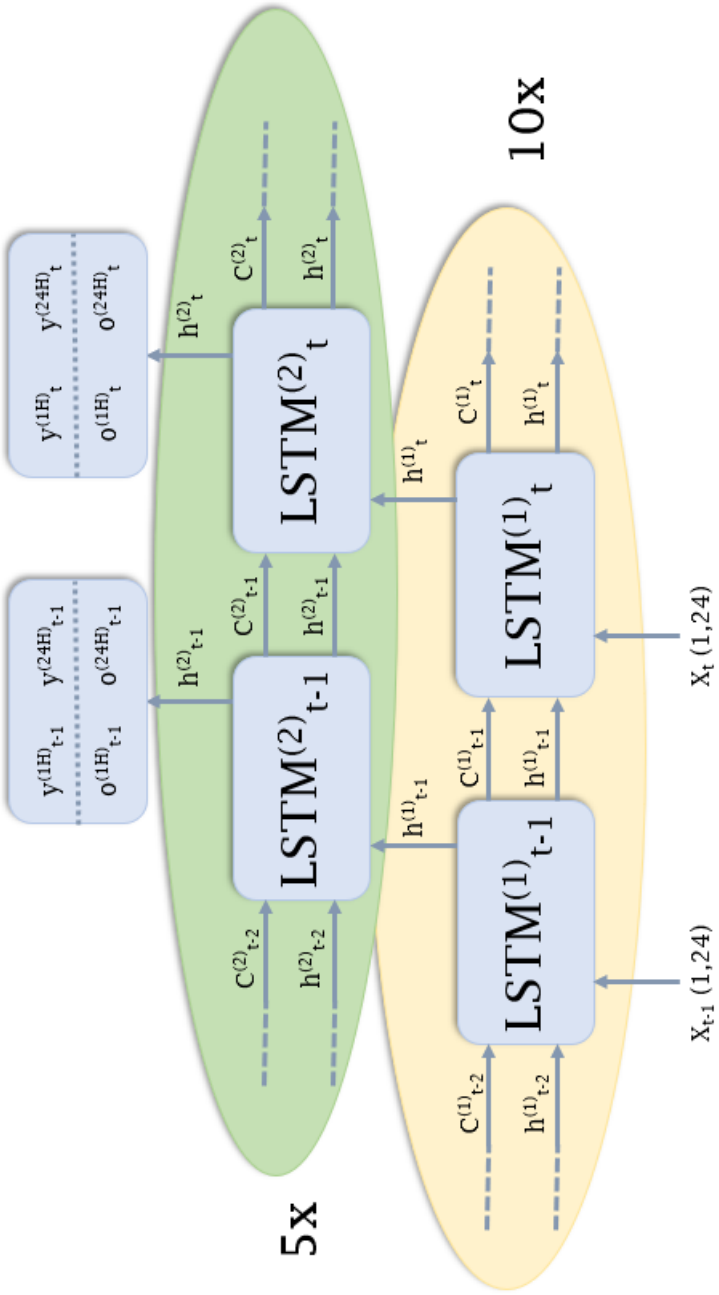

Figure 6: Architettura Rete LSTM

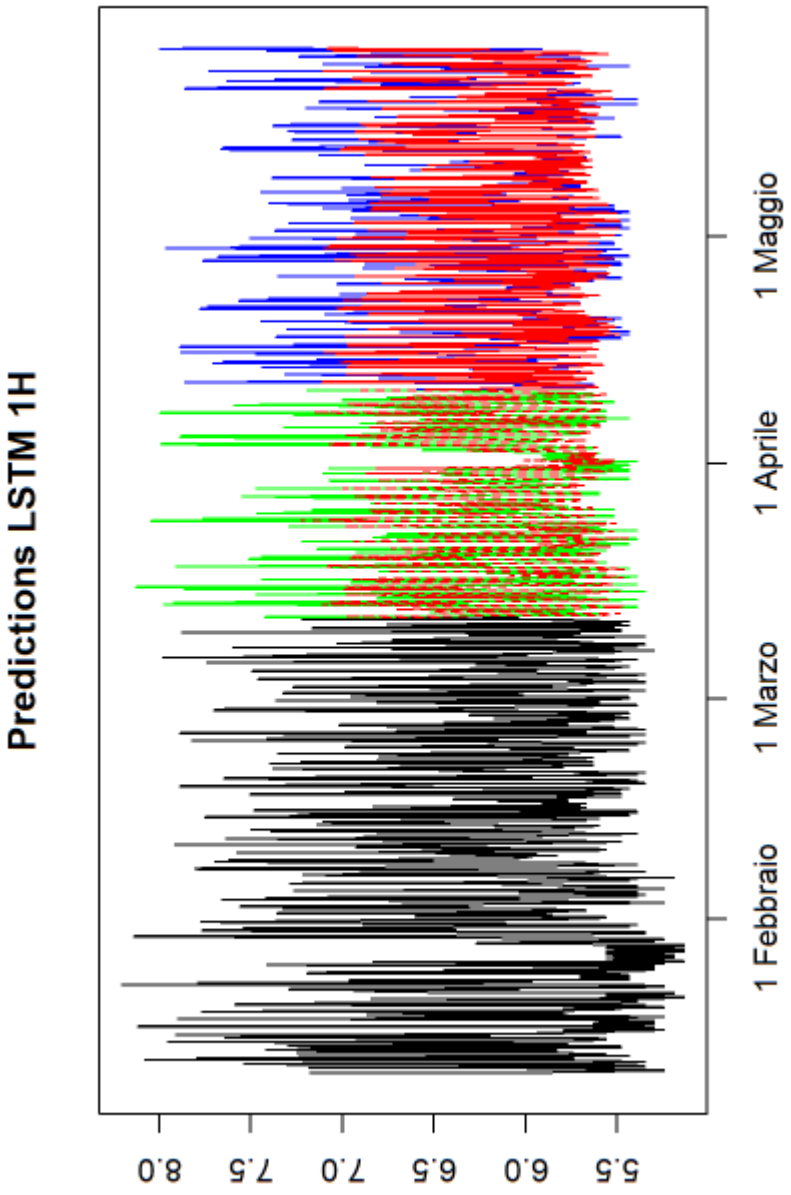

Figure 7: Predizioni LSTM H1

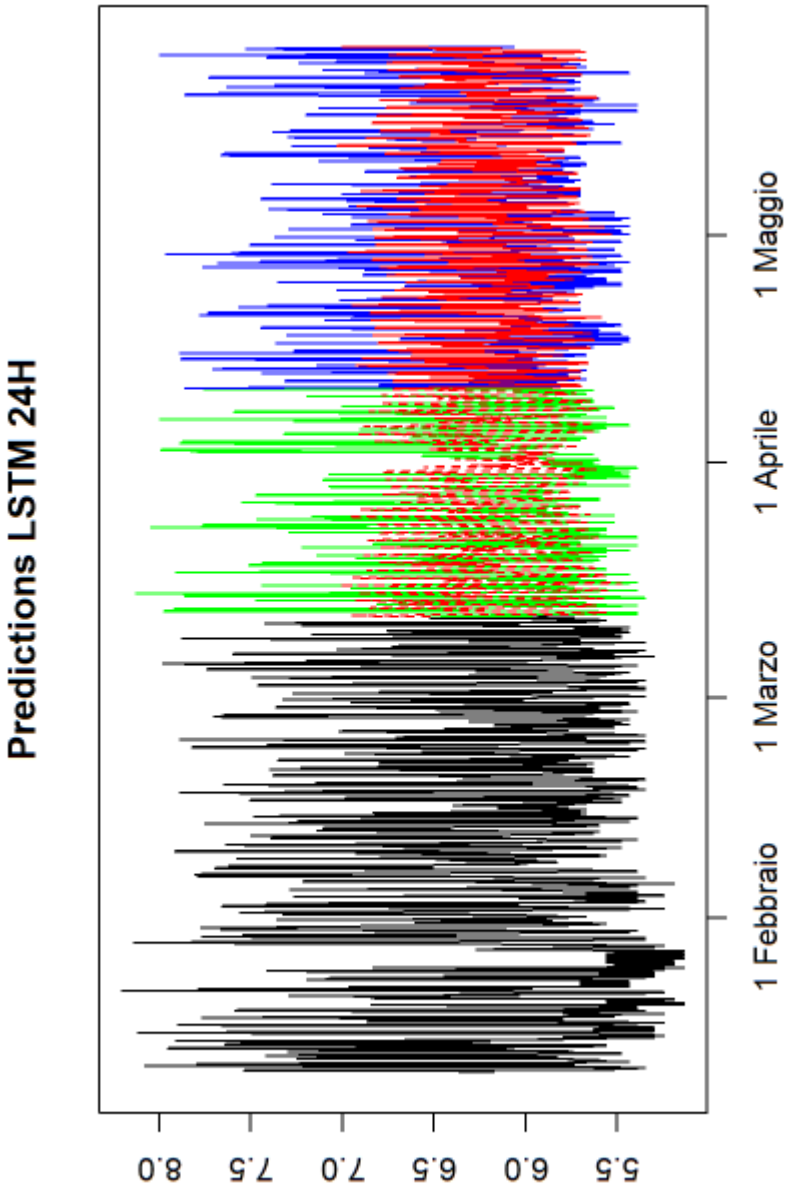

Figure 8: Predizioni LSTM H24