POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Studio e Costruzione di Modelli di Previsione di Energia Prodotta da un Impianto Eolico

Relatori
prof. Paolo Garza

Candidato
Lorenzo Leoni

Anno Accademico 2021-2022

† *A mia nonna Antonietta*

Sommario

Studio e costruzione di modelli di serie temporali (ARIMA, ARIMAX, SARIMA, SARIMAX) per la previsione dell'energia prodotta da un impianto di pale eoliche. I dati sono stati registrati dalla Supervisory Control And Data Acquisition (SCADA) ed utilizzati per la competizione KDD Cup 2022. Il database è costituito da 134 pale eoliche con registrazioni che si distribuiscono per 245 giorni con un timestamp di 10 minuti tra un'osservazione e l'altra, per un totale di 4,727,520 dati. Oltre ai dati dell'energa prodotta si hanno altre informazioni utili utilizzabili come regressori esterni dei modelli matematici.

Ringraziamenti

Dedico qualche riga alle persone che mi hanno aiutato a completare questo lavoro e mi hanno permesso di essere la persona che sono oggi.

Ringrazio il professore Paolo Garza, per la piena disponibilità nell'aiutarmi a completare il mio progetto.

Ringrazio mia madre Maria Teresa, la donna più importante che esista. L'unica che ha creduto in me anche quando ero io a non crederci.

Ringrazio mio padre Domenico, che mi ha insegnato a riconoscere quali sono le cose importanti della vita.

Ringrazio la mia ragazza Giada, che mi sostiene ogni giorno da diversi anni ormai.

Indice

Elenco delle tabelle

Elenco delle figure

One of the most important things: never, ever quit. Never quit! I've seen people quitting and if they would have held out longer they would been successful. I've seen it so much. I've seen some of the most brilliant people in the world that never made it because they were quitters. You have to also have flexibility though. You can't necessarily say "I'm never giving up"! You have to always be able to change course a little bit, maybe always with that same goal, but don't quit!

[D. J. Trump]

Go after your dreams. I never allowed anybody to tell me what I was capable of accomplishing. I never let anybody tell me that I can't do. If you wanna accomplish something or do something that a lot of people think is difficult to accomplish, you have to put all your eggs in one basket. There's no compromise with that.

[K. Bryant]

It's always the quiet work that you put in with your own two hands when no one else is watching that's the one degree of separation from you and everyone else.

[D. D. Johnson]

Capitolo 1

Introduzione

L'analisi di serie storiche è un ramo molto complesso che permette di modellizzare dati provenienti da qualsiasi settore, come ad esempio economico, climatico, di telecomunicazione e molti altri. Una caratteristica molto importante dei modelli temporali è la possibilità di trattare dati di grande complessità (ad esempio quotazioni in borsa) e nel contempo permettere la personalizzazione dei suoi parametri fondamentali. Ciò consente ad un analista competente di aver a disposizione un modello matematico estremamente potente e duttile, in grado di ottenere un ottimo compromesso tra qualità dei risultati e chiarezza della struttura modellistica.

Con questo lavoro ci si è posti come obbiettivo principale lo studio della costruzione di modelli matematici di serie temporali e l'automatizzazione di esso. Quest'ultima parte permetterebbe di eliminare la necessaria manutenzione a cui vengono sottoposti con il fine di mantenere il modello valido ed efficiente. Riuscire a costruire un unico algoritmo di building & fitting contemporaneamente porterebbe grandi risparmi di tempo e di risorse economiche da parte di privati che hanno necessità di fare previsione su serie temporali.

Le operazioni da svolgere sul codice si ridurrebbero al semplice lancio dello stesso, in modo da consentire l'update di dati più recenti. Chiaramente, il codice sviluppato dipende fortemente dal dataset analizzato. Di conseguenza la selezione delle finestre di training e testing, il timestamp della serie temporale, il numero di modelli da fittare ed altro ancora varierà in base a ciò che si deve studiare. Ciò che però risulta replicabile è l'approccio, più in particolare l'individuazione del modello corretto da costruire, possibile

grazie allo studio automatizzato dei dati.

Il programma è stato creato in ambiente R, in quanto molto potente e flessibile nelle analisi con un elevato livello di matematica e statistica al loro interno. Si hanno a disposizione molte librerie contenenti function per fare test statistici particolari e per estrarre insights tecnici molto utili in modo più diretto. Data la grande quantità di dati da analizzare (4,727,520) e di modelli da fittare (134 pale eoliche moltiplicate per tutti i possibili regressori utili al miglioramento delle metriche), si è fatto girare il codice su HPC, servizio che mette a disposizione cluster di computer per ottenere delle performance di calcolo computazionale ottime, fornito dal Politecnico di Torino.

Il progetto si ferma al training e alla validation dei modelli, non si pone l'obbiettivo di riuscire a migliorare le performance del validation svolto dalla baseline rilasciata per la competizione, in quanto i modelli fittati sono sistemi di reti neurali. Essi hanno spesso performance ottimali ma allo stesso tempo il come si sia riusciti ad ottenere dei buoni risultati è di difficile comprensione. Proprio per questo i neural networks sono dei modelli black-box, dove la struttura dell'approssimazione della funzione che restituirà il risultato non viene esplicitata. La costruzione di modelli di serie temporali è molto piu chiara da questo punto di vista. Tutti i step per la costruzione del modello, dalla scelta degli hyperparameter ai regressori, sono personalizzabili e riconducibili ad una equazione composta da coefficienti e variabili esogene. Oltre a ciò ci sono stati anche limiti tecnici che non hanno permesso di poter generalizzare questo progetto ad un livello di dettaglio superiore. Ad esempio alcuni modelli molto complessi sono dovuti essere stati esclusi per via dell'elevato costo computazionale che l'intero algoritmo portava. In altri casi invece è stato necessario trovare soluzioni alternative meno rigorose ma, come si vedrà, con ottimi risultati in particolare su questo dataset. Ciò però non implica che le casistiche non applicate nella pratica vengano ignorate in questo report. Come si vedrà proseguendo la lettura, prima di applicare delle semplificazioni viene spiegato il metodo migliore che andrebbe utilizzato nel caso in cui si riesca a non avere limiti tecnici.

Nei prossimi capitoli si desciverà l'approccio canonico svolto manualmente da un analista per la costruzione di modelli di serie storiche [2], mostrando come e cosa bisogna osservare per poter determinare le caratteristiche fondamentali della serie, ovvero il trend, la stagionalità e parametri principali.

Ciò è alla base per poter cominciare ad ipotizzare un modello corretto e coerente rispetto ai dati in studio. Verrà mostrato come sono stati trattati i dati e modellizzate le varie turbine in termini temporali [3]. Verranno mostrate parti dell'algoritmo in pseudocodice, commentate nei passaggi chiave e spiegate certe scelte o ipotesi utilizzate lungo lo sviluppo del codice. Successivamente verranno mostrate le metriche ottenute e il controllo delle 134 serie temporali per il check della bontà dello studio automatizzato [4] e osservazioni finali con ipotetiche ed ulteriori generalizzazioni dell'algoritmo e possibili applicazioni [5].

Capitolo 2

Approccio Canonico

Il processo di forecasting è fondamentale in molti ambiti, dal planning di budget aziendale allo studio dei cambiamenti climatici. Con il progresso della tecnologia e la crescente capacità computazionale degli strumenti utilizzabili si può essere in grado di sviluppare algoritmi di machine learning molto complicati e capaci di gestire grandi quantità di dati. Di base il concetto è semplice, si individua l'obiettivo da raggiungere, si studiano i dati a disposizione e si costruisce un modello ottimale per essi che riesca ad ottenere un buon grado di previsione della variabile target, tutto ciò dopo un adeguata fase di training (ovvero il modello studia a sua volta i dati che verrano messi a disposizione dall'analista in modo da comprenderne le logiche e individuare i pattern). In particolare, il forecasting di serie temporali ha degli step precisi da seguire, saltarne uno o non essere rigorosi nel processo potrebbe portare ad ottenere risultati non corretti e talvolta opposti a quelli che si dovrebbero ottenere se si procedesse con attenzione e precisione. Il dato è generalmente strutturato da due componenti: il timestamp (l'instante di tempo quando la registrazione del dato è avvenuta) e il valore dell'osservazione. Data una serie temporale, la frequenza (stagione) di osservazione deve rimanere fissa. In particolare, non si possono fare studi omogenei su serie che hanno osservazioni mensili per n anni e osservazioni bimestrali dall'n+1-esimo anno in poi. Comprendere il tipo di dato che si sta studiando permette di svolgere una data preparation ottimale ed avere già in mente qualche modello o metrica utilizzabile. Una situazione ideale è chiaramente quella dove si ha pieno controllo del dominio da cui questi dati provengono, poichè ciò velocizzerebbe la fase di analisi qualitativa iniziale. Nel caso in cui non si è molto pratici con il dominio, fare della ricerca sul

settore che si sta studiando permette di ridurre il rischio di incomprensione dei dati e dei loro valori.

Da ciò che è stato descritto si può comprendere quanto non sia semplice e veloce la fase di analisi e costruzione di un modello di machine learning. Come detto nell'introduzione, l'idea del progetto è di facilitare il più possibile questi step, in modo da velocizzare ed automatizzare processi spesso ricorrenti.

2.1 Metriche Valutative

Completato tutto il processo di data preparation necessario per lo svolgimento di una qualsiasi data analysis, l'analista si ritroverà dei dati temporali pronti per essere studiati. Il primo passo è sicuramente fare un plot per visualizzare la serie. Ciò è importante poichè comprendere su che range di valori può muoversi la serie è fondamentale in quanto successivamente bisognerà scegliere delle metriche opportune per valutare il nostro modello e la loro scelta non può prescindere da ciò. Ad esempio si potrebbero riscontrare degli errori di codice durante le run. Questi casi possono essere considerati fortunati in quanto la mancata esecuzione del programma è un segnale che qualcosa non è stato fatto in maniera corretta. Molto più grave sarebbe ottenere dei risultati non corretti se non addirittura opposti a ciò che in realtà si dovrebbe avere. Questo potrebbe accadere, appunto, se scegliamo metriche non corrette per il dominio dei dati che stiamo analizzando, restituendo dei score senza alcun vero significato oltre che falsi. Per individuare l'intervallo di valori possibili non bisogna solo basarsi su ciò che attualmente la nostra serie ci sta dicendo, ma bisogna avere sensibilità riguardo il settore da cui questi dati provengono. In caso di vendite automobilistiche per un certo modello di auto, è possibile ottenere in un determinato istante di tempo un valore nullo, anche se magari nei dati sotto osservazione questa casistica non è mai stata riscontrata fino a quel momento.

Sottolineata l'importanza della comprensione del dominio di definizione, è importante cominciare a selezionare quale metriche siano le più adatte per i dati in analisi. Ne esistono diverse, definito Y_t il valore reale al tempo t e $\bar{Y_t}$ il valore predetto al tempo t, le più usate sono le seguenti [Shmueli and Lichtendahl [2016]], [Krispin [2019]]:

- Mean Squared Error (MSE), quantifica la distanza media al quadrato tra i valori effettivi e previsti. L'effetto quadrato dell'errore impedisce ai valori positivi e negativi di annullarsi a vicenda e di penalizzare sempre più all'aumentare del tasso di errore

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y_t})^2$$

- Root Mean Squared Error (RMSE), la radice della distanza media al quadrato dei valori effettivi e previsti. Come l'MSE, l'RMSE ha un alto tasso di errore dovuto all'effetto quadrato ed è quindi sensibile ai valori anomali

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y_t})^2}$$

- Mean Absolute Error (MAE), misura il tasso di errore assoluto della previsione. Analogamente a l'MSE e l'RMSE, evita la cancellazione di valori positivi e negativi. D'altra parte, non vi è alcuna penalizzazione dell'errore, e quindi questo metodo non è sensibile ai valori anomali

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \bar{Y_t}|$$

- Mean Absolute Percentage Error (MAPE), misura la percentuale media dell'errore assoluto. Esso è sensibile alla scala dei valori e non deve essere utilizzato quando si lavora con dati a basso volume. Si noti che siccome il denominatore dell'equazione è composto dal valore reale, il MAPE non è definito quando esso è zero. Inoltre, quando il valore reale non è zero, ma piuttosto piccolo, il MAPE assume spesso valori estremi. Questa sensibilità di scala rende il MAPE quasi privo di valore come misura di errore per dati a basso volume.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \bar{Y_t}|}{|Y_t|}$$

2.2 Componenti di una Serie Temporale

Dopo aver visualizzato la serie temporale, compresi i valori assunti e possibili e pensato alle possibili metriche da utilizzare, si può iniziare ad avere un'idea di come modellare i dati. Osservando il suo grafico è possibile scovare

visivamente le quattro caratteristiche principali di qualsiasi serie temporale [Shmueli and Lichtendahl [2016]]:

- Valor medio L
- Trend T
- Stagionalità S
- Rumore N

Le quali possono essere definite additive o moltiplicative in base al loro singolo comportamento. In particolare si hanno due principali modelli di serie temporali:

- Serie temporale additiva

$$Y = L + T + S + N$$

- Serie temporale moltiplicativa

$$Y = L \cdot T \cdot S \cdot N$$

Si definisce una serie come additiva ogniqualvolta vi è una crescita del trend (rispetto al periodo precedente), o se l'ampiezza della componente stagionale rimane grosso modo la stessa nel tempo. D'altra parte, classifichiamo una serie come moltiplicativa ogni volta che la crescita della tendenza o l'entità della componente stagionale aumenta o diminuisce di una molteplicità da un periodo all'altro nel tempo [Krispin [2019]]. In Figura 2.1 si può notare quanto i plot siano utili per fare delle prime deduzioni riguardo la modellizzazione dei dati tramite l'andamento che essi mostrano.

Nel caso in cui non si fosse sicuri dal semplice plot della serie, si può utilizzare il metodo dei coefficiente di variazione delle differenze stagionali e dei quozienti Dennis et al. [2017]. La differenza stagionale D è stata calcolata prendendo la differenza tra una determinata stagione di un anno e la stessa stagione dell'anno precedente mentre il quoziente stagionale Q è stato calcolato come quoziente di una determinata stagione di un anno e la stessa stagione dell'anno precedente . In particolare le formule sono le seguenti:

$$\begin{cases} D_{i,j} = X_{i,j} - X_{i-1,j} \\ Q_{i,j} = \dfrac{X_{i,j}}{X_{i-1,j}} \end{cases}$$

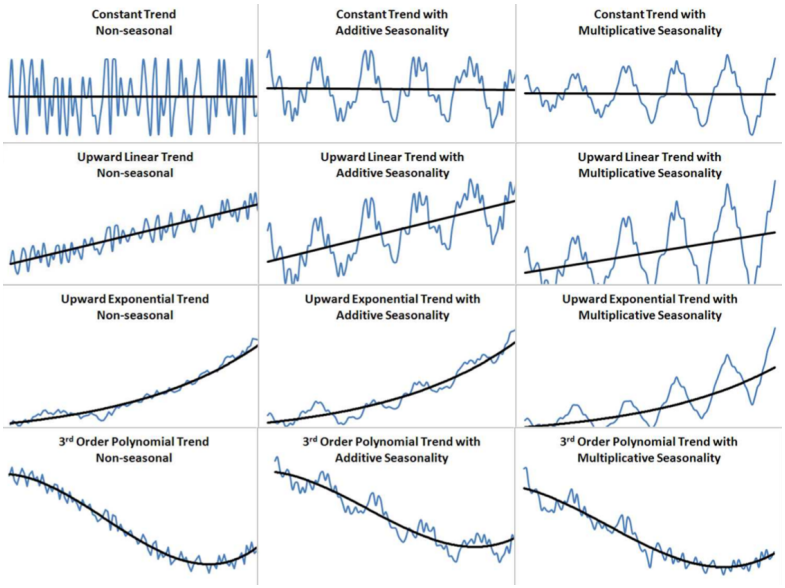

Figura 2.1. Illustrazioni di trend comuni e modelli di stagionalità

dove i è l'anno e j il mese. Successivamente il coefficiente di variazione delle differenze stagionali $CV(D)$ e il coefficiente di variazione dei quozienti stagionali $CV(Q)$ sono calcolati come:

$$\begin{cases} CV(D) = \frac{sd(D)}{media(D)} \\ CV(Q) = \frac{sd(Q)}{media(Q)} \end{cases}$$

dove sd è la deviazione standard. La regola decisionale che aiuta la scelta del modello è stata definita come:

$$\begin{cases} |CV(D)| < |CV(Q)| & \text{additivo} \\ |CV(D)| \geq |CV(Q)| & \text{moltiplicativo} \end{cases}$$

Ora si può passare al determinare la presenza di trend e stagionalità. Per fare ciò bisogna analizzare il grafico più importante per le serie temporali, ovvero l'autocorrelazione (ACF). Esso mostra la correlazione tra la serie e i suoi lags [Krispin [2019]]. In Figura 2.2 viene riportato un plot d'esempio, di una serie avente sia trend che stagionalità. Il decadimento graduale è tipico di una serie temporale contenente un trend mentre il picco a 1 anno indica variazioni stagionali [Cowpertwait and Metcalfe [2009]]. Per visualizzare meglio queste due componenti si può procedere svolgendo la differenza tra

la serie temporale originale e essa stessa con un lag fissato. In particolare, si ricava il trend rimuovendo la componente stagionale, tramite differenza tra la serie originale e il suo N-esimo lag che sarebbe la misurazione nello stesso momento ma avvenuta in un ciclo successivo. Riprendendo l'esempio riportato in Figura 2.2, i dati sono mensili di conseguenza un ciclo di osservazioni avviene un anno dopo, ovvero 12 mesi. Di conseguenza per rimuovere la stagionalità si procede con la differenza tra le osservazioni di ogni mese in un dato anno e quelle registrate l'anno successivo. Per estrapolare la stagionalità va rimossa la componente del trend, ovvero si procede con la differenza tra la serie originale e la stessa un lag successivo. Ciò permette di eliminare un trend crescente/decrescente evidenziato principalmente i picchi dovuti alla stagionalità. Queste procedure di isolamento delle componenti potrebbero essere svolte in modo automatico dalla function *decompose* di R, ma uno degli aspetti negativi è che la stima della componente stagionale si basa sulla media aritmetica, quindi esiste un'unica stima della componente stagionale per ogni unità di ciclo (ad esempio, tutte le osservazioni delle serie verificatesi a gennaio avrà la stessa stima della componente stagionale se la serie è mensile). Ciò non è problematico quando si applica questo metodo a una serie temporale additiva poiché la stagionalità rimane la stessa (o quasi) nel tempo. D'altra parte, questo non è il caso di una serie temporale moltiplicativa, poiché la stagionalità cresce nel tempo. Un eccellente esempio di ciò è il precedente set di dati sulle prenotazioni dei passeggeri aerei, poiché l'entità della componente stagionale aumenta nel tempo [Figura 2.3]. Una media non rappresenterebbe la stagionalità all'inizio e alla fine della serie,

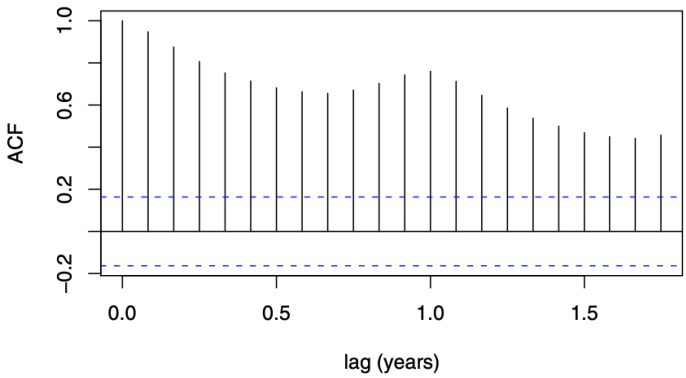

Figura 2.2. ACF delle prenotazioni dei passeggeri aerei nel periodo 1949–1960

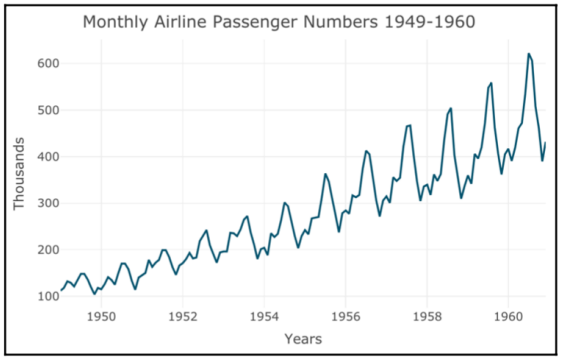

Figura 2.3. Serie temporale delle prenotazioni dei passeggeri aerei nel periodo 1949–1960

poiché le sovrastima e le sottostima rispettivamente. D'altra parte, si adatterebbe bene da qualche parte nel mezzo, dove è vicino alla media [Krispin [2019]].

2.3 Modelli per Serie Temporali

2.3.1 Modelli Principali

Una volta completata l'analisi delle due componenti principali, ed appurata la loro presenza o meno, si dovrà scegliere il modello più adatto alle caratteristiche della serie temporale. Nel caso in cui sia presente trend e/o stagionalità, la serie viene definita non stazionaria, se nessuno dei due fenomeni sussiste invece è detta stazionaria. Stazionarietà significa che i parametri statistici di una serie temporale non cambiano nel tempo. In altre parole, le proprietà di base della distribuzione dei dati delle serie temporali, come la media e la varianza, rimangono costanti nel tempo. Pertanto, i processi di serie temporali stazionari sono più facili da analizzare e modellare perché l'assunto di base è che le loro proprietà non dipendono dal tempo e saranno le stesse in futuro come nel precedente periodo storico. In alternativa, le serie temporali che mostrano variazioni nei valori dei loro dati, come una tendenza o una stagionalità, non sono chiaramente stazionarie e, di conseguenza, sono più difficili da prevedere e modellare Lazzeri [2020]. La scelta dei modelli si basa su questo concetto. Essi devono poter rendere stazionari,

con il tuning dei loro parametri, dati che non lo sono. I modelli più importanti sono i seguenti [Krispin [2019]], [Shmueli and Lichtendahl [2016]], [Nau [2014]]:

- AutoRegression (AR (p)), i modelli AR sono simili ai modelli di regressione lineare, tranne per il fatto che i predittori sono i p valori passati della serie (i lags). Il modello è applicabile a serie temporali stazionarie (senza trend e stagionalità). La sua struttura è la seguente (ϵ_t è il rumore)
$$\bar{Y_t} = \beta_0 + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + \epsilon_t$$

- Moving Average (MA (q)), una media mobile con larghezza della finestra pari a q significa calcolare la media su ciascun insieme di q valori consecutivi, dove q è determinato dall'utente. In generale, ci sono due tipi di medie mobili: una media centered moving average e una trailing moving average. Le centered moving average sono potenti per visualizzare le tendenze perché l'operazione di calcolo della media può sopprimere la stagionalità e il rumore, rendendo così la tendenza più visibile. Le trailing moving average sono utili per il forecasting. La differenza tra i due è il posizionamento della finestra di media sulla serie temporale. Siccome vogliamo fare previsione ci concentreremo su le trailing moving average. Il modello è applicabile a serie temporali stazionarie (senza trend e stagionalità). La sua struttura è la seguente
$$\bar{Y_t} = \frac{Y_{t-q} + ... + Y_{t-1}}{q}$$

- AutoRegression Moving Average (ARMA(p,q)), unione dei due precedentemente elencati. Il modello è applicabile a serie temporali stazionarie (senza trend e stagionalità). La sua struttura è la seguente
$$\bar{Y_t} = \beta_0 + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + \alpha_1 \epsilon_{t-1} + ... + \alpha_q \epsilon_{t-q} + \epsilon_t$$

  È possibile poter includere delle serie temporali esterne che possano aggiungere informazioni utili alla previsione della serie principale. Ciò permette di influenzare i dati stimati tramite l'utilizzo di una o più serie storiche che possano avere una determinata correlazione con l'andamento della serie principale. In questo caso, se venissero usati dei regressori, il modello sarebbe un ARMAX.

- Simple Exponential Smoothing (ES), la sua popolarità deriva dalla sua flessibilità, facilità di automazione, calcolo economico e buone prestazioni. L'Exponential Smooting è simile alla previsione tramite la Moving Average (MA), con la particolarità che invece di prendere svolgere una media semplice sui q valori più recenti, si prende una media ponderata di tutti i valori passati, in modo che i pesi diminuiscano esponenzialmente nel passato. L'idea è di dare più peso alle informazioni recenti, ma di non ignorare completamente le informazioni più vecchie. Come i modelli descritti finora, esso è applicabile a serie temporali stazionarie (senza trend e stagionalità). La susa struttura è la seguente:

  $$\bar{Y_t} = \alpha Y_{t-1} + \alpha(1-\alpha)Y_{t-2} + \alpha(1-\alpha)^2 Y_{t-3} + ...$$

  Dove α è detta smooting constant, e ha valore compreso tra 0 e 1 (in base a quanto peso si vuole dare alle informazioni più recenti). La sua scelta è fondamentale e va fatta con cura in modo da non portare overfitting nel training set e bassa accuracy nel validation/test set.

- AutoRegressive Integrated Moving Average (ARIMA (p,d,q)), generalizzazione del ARMA, che consente di poter lavorare con serie avente trend ma non stagionalità (di conseguenza non stazionarie). Questo grazie al parametro d che indica la necessità di applicare una differenza tra la serie e i suoi lag in modo da poter rimuovere il trend. Un modello senza ordini di differenziazione ($d = 0$) presuppone che la serie originale sia stazionaria. Un modello con un ordine di differenziazione ($d = 1$) presuppone che la serie originale abbia una tendenza media costante. Un modello con due ordini di differenziazione totale ($d = 2$) presuppone che la serie originale abbia una tendenza variabile nel tempo. Per ora ci si limita al caso di ordine due, nel modello successivo si riprenderà il concetto. Nel caso in cui si utilizzino dei regressori esterni il modello sarebbe un ARIMAX.

- Seasonal AutoRegressive Integrated Moving Average(SARIMA (p,d,q) x (P,D,Q)), ulteriore generalizzazione del ARMA, che consente di lavorare con serie avente sia trend che stagionalità (di conseguenza non stazionarie). Il parametro che indica la necessita di una differenza stagionale è D. Esso non va mai posto maggiore di 1 e, in generale, la somma tra l'ordine di differenza non stagionale (d) e lo stagionale (D) deve essere massimo pari a 2. Ecco il motivo di limitare il caso a $d = 2$

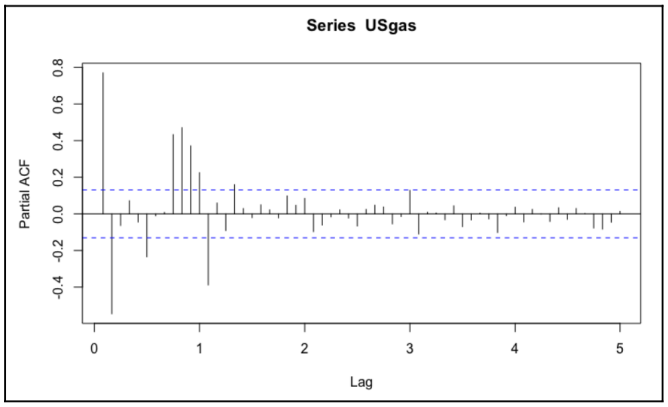

Figura 2.4. PACF di una serie rappresentante il consumo mensile di gas naturale negli Stati Uniti dal gennaio 2000

anche per le serie con solo trend. Nel caso in cui si utilizzino dei regressori esterni il modello sarebbe un SARIMAX. Se invece la serie ha esclusivamente componente stagionale (quindi no trend) il modello sarebbe un SARMAX, dove il parametro d viene posto nullo.

- Holt-Winter's Exponential Smoothing (HWES), generalizzazione dell'ES, che permette di poter lavorare con serie avente sia trend che stagionalità tramite una modellizzazione ad hoc della componente del valor medio (L) del trend (T) e della stagionalità (S).

2.3.2 Scelta degli Hyperparameters

Ogni modello sopraelencato ha degli hyperparameters da dover impostare, la loro scelta ha bisogno di ulteriore studio in modo tale da potersi avvicinare alla costruzione di un modello ottimale, tenendo in cosiderazione quali fattori (trend e stagionalità) influenzano la serie.

Si prenda in considerazione l'AutoRegression (AR (p)), essa è applicabile a serie stazionarie di conseguenza per scegliere il parametro p bisogna visualizzare il correlogramma della autocorrelazione parziale (PACF). Esso mostra la correlazione condizionale della serie con il lag k-esimo data la relazione tra i lag 1, 2, ... e k-1 e la serie. In altre parole, il PACF fornisce una stima della correlazione diretta della serie con il lag k-esimo dopo aver rimosso la correlazione tra essa e i lag precedenti [Krispin [2019]]. In Figura 2.4 viene riportato il grafico della PACF di una serie rappresentante

il consumo mensile di gas naturale negli Stati Uniti dal gennaio 2000. Per individuare il corretto numero di lag da inserire nel modello bisogna contare quanti di essi siano significativi (ovvero si trovano oltre la zona delimitata dalle linee tratteggiate che indicano un livello di significatività inferiore al 5 %). Dal grafico si nota che i lag significativi, prima di essere tagliati fuori dal livello, risultano essere i primi due. Sono presenti altri lag statisticamente importanti ma sono successivi a lag non significativi. Per decidere se integrarli lo stesso bisogna comprendere se il numero di lag precedentemente non significativi non sia relativamente elevato. Se si includessero troppi lag sotto il livello si potrebbe rischiare overfitting solo per cercare di raggiungere lag significativi ma troppo distanti nel tempo. Per scogliere qualsiasi dubbio si può pensare di utilizzare metrice come AIC e BIC, in modo da avere una stima visiva della bontà del nostro modello nel caso o meno di inclusione di alcuni lag non significativi. Di base ciò che possiamo concludere dal grafico è che p sarà almeno pari a 2.

Si prenda in considerazione la Moving Average (MA (q)), essa è applicabile a serie stazionarie. Di conseguenza per scegliere il parametro q bisogna visualizzare il correlogramma della autocorrelazione (ACF). Come per AR, anche qui bisogna controllare quanti lag significativi esistono, facendo attenzione a non overfittare.

Stesso discorso per ARMA, essendo la combinazione dei due modelli precedenti. Ci sono delle variazioni invece per quanto riguarda l'AutoRegressive Integrated Moving Average (ARIMA (p,d,q)). Siccome il modello è in grado di lavorare con serie avente componente trend (e quindi non stazionarie), per potersi ricavare i parametri p e q è importante prima svolgere il processo di differenza tra la serie originaria e i suoi lag. In pratica si cerca di rendere la serie stazionaria rimovendo la componente del trend, e solo successivamente si replicano le analisi su ACF e PACF per individuare i parametri q e p rispettivamente. La scelta del parametro d è stata discussa quando si è definito il modello ARIMA, in generale esso deve essere minore o pari a 2.

Per la sua generalizzazione Seasonal AutoRegressive Integrated Moving Average (SARIMA (p,d,q) x (P,D,Q)) bisogna replicare le analisi svolte ma considerando che si sta trattando serie temporali con componente stagionale oltre che di trend [Krispin [2019]]. La scelta dei parametri non stagionali (p,d,q) risulta identica a come viene effettuata nell'ARIMA. I parametri stagionali possono essere usando due approcci:

- Rimuovendo il trend tramite differenza e analizzando i lag significativi della ACF e PACF
- Analizzando direttamente ACF e PACF ma controllando solo il lag stagionali significativi

Anche qui la scelta del parametro D è stata affrontata nella definizione del modello SARIMA.

Ora che i parametri sono stati fissati, si può procedere con il definire le finestre di training, validation e testing ed iniziare a fittare il modello. Test molto importanti da svolgere sul modello fittato sono sicuramente la distribuzione dei residui (Shapiro) e l'indipendenza degli stessi (Ljung–Box), se uno di questi due test dovesse non ritornare i risultati attesi allora sarebbe un campanello d'allarme nei confronti del modello costruito. Non necessariamente si può essere in grado di costruire un modello che superi questi due test, ma è importante averne conoscenza soprattutto quando si ha necessità di comunicare i risultati delle previsioni ottenute, segnalando che i modelli non sono stati validati al 100%.

Capitolo 3

Approccio Automatizzato

3.1 Problema Risolto

Come detto nell'introduzione, l'obiettivo è automatizzare tutti i processi descritti nel capitolo 2, a partire dall'analisi qualitativa al fitting dei modelli. Tutto ciò che verra descritto successivamente è racchiuso in un unico grande ciclo che permette al codice di studiare e analizzare tutte quante le 134 pale eoliche.

Lo step iniziale di qualsiasi data analysis è la data preparation, svolta seguendo le indicazioni fornite dal report della competizione KDD Cup 2022. Ci sono casistiche di dati per i quali viene descritto l'approccio da usare nel trattarli, mentre per altri viene solo descritto il modo per determinare se siano effettivamente dei valori "particolari". Nella sezione successiva verrà esposto l'approccio utilizzato per quest'ultimi ma in generale, quando il modello verrà validato, i dati che sono stati individuati come "particolari" non avranno influenza sulle metriche di valutazione scelte in quanto tenuti fuori dal calcolo dell'errore per via della loro anormalità.

Fatto ciò si procede con lo step dell'analisi qualitativa, fondamentale per individuare i parametri p, d, q, P, D, Q. Chiaramente il loro calcolo si basa sulle componenti che si trovano nella serie temporale. I plot dell'ACF e PACF per essere chiari hanno bisogno di visualizzare un numero consono di lag in modo da estrapolare le informazioni necessarie senza essere disturbati visivamente da lag non importanti. Questo processo va di pari passo con la stima dei parametri del AR e MA generici (ovvero quelli utilizzati nella costruzione dei modelli se si suppone che la serie sia stazionaria). Questa

supposizione è a solo scopo di visualizzazione qualitativa dei correlogrami, nella pratica i due parametri individuati verranno utilizzati nel building solo se con le successive analisi si potrà afferma che la serie effettivamente sia stazionaria. L'estrapolazione dei parametri di serie non stazionarie, invece, deve seguire il processo descritto nel capitolo precedente. Prettamente lato qualitativo, viene sviluppato un algoritmo in grado di calcolare il numero di lag decrescenti e di picco. Quest'ultimi in particolare saranno di fondamentale importanza anche nel decidere se la serie abbia o meno componente stagionale. Rimane da comprendere come far stabilire al codice in modo automatico se una serie abbia trend o meno. Tramite l'uso di un test statistico (successivamente spiegato) si potrà stabilire ciò e quindi essere pronti a far fittare il modello corretto. Una volta calcolati gli hyperparameters per la turbina in osservazione, essi verranno storati all'interno di una matrice avente nelle colonne l'ID della turbina e i sei parametri individuati. In caso in cui alcuni parametri non siano necessari da calcolare in quanto il modello scelto non li necessita (ad esempio ARIMA non ha i parametri stagionali P,D e Q) essi verranno posti uguali a zero nella matrice.

Estrapolate tutte le informazioni descritte precedentemente, si può procedere con la parte principale dell'algoritmo. Vengono inizializzate tutte le matrici e variabili chiave o di appoggio che permettono al codice di svolgere il suo lavoro. Avendo a disposizione 9 potenziali regressori, la loro selezione si baserà sulle metriche indicate dal report della competizione. Ci si aspetta la presenza del regressore *Wspd*, ovvero la velocità del vento, in quanto fortemente correlato con l'energia prodotta dalle pale eoliche (gli altri regressori verranno presentati nella sezione successiva, quando la loro analisi sarà fondamentale sia per la data preparation e sia come regressori per i modelli). Di conseguenza tutti i modelli saranno per certo degli ARIMAX o SARIMAX. Verranno definite le finestre di training e di validation e tramite il check di trend e stagionalità si procederà al fitting del modello integrando i parametri p, d, q, P, D, Q.

Successivamente si giungerà in un sotto-algoritmo di feature selection dove il regressore che restituisce i risultati migliori in termini di score verra scelto come il nuovo da integrare nel modello. Cosi facendo si selezioneranno un sottinsieme ottimale di regressori e, posti vincoli ai valori che possono essere fittati nel training e predetti nel validation, si potrà fittare il modello definitivo per la turbina in analisi e il calcolo del suo score. Il tutto verrà salvato in due matrici, una avente nelle colonne il nome del modello fittato

(ARIMAX, SARIMAX ecc.), il numero della pala eolica che ha modellizzato e il numero del regressore dove nella cella corrispettiva si avrà il nome del regressore selezionato come n-esimo migliore, ed un'altra matrice avente i punteggi delle metriche.

3.2 Descrizione Soluzione Proposta

In questa sezione verranno presentati tutti i processi descritti sopra, ma nel dettaglio.

3.2.1 Struttura Dataset

Come viene riportato nel report della competizione, il dataset è composto da diversi campi e non solo dal valore dell'energia prodotta. Andando nello specifico si ha:

- TurbID, ID della pala eolica
- Day, Giorno in cui è avvenuta la registrazione
- Tmstamp, Istante della registrazione
- Wspd, Velocità del vento registrata dall'anemometro
- Wdir, L'angolo tra la direzione del vento e la gondola della turbina
- Etmp, Temperatura dell'ambiente circostante
- Itmp, Temperatura all'interno della gondola della turbina
- Ndir, Direzione della gondola ovvero l'angolo dell'azionamento di imbardata della gondola della turbina (l'azionamento di imbardata orienta la gondola sempre in direzione del vento per garantire un'ottimale efficienza della turbina)
- Pab1, Angolo di beccheggio della lama 1
- Pab2, Angolo di beccheggio della lama 2
- Pab3, Angolo di beccheggio della lama 3
- Prtv, Potenza di Reazione
- Patv, Potenza Attiva (variabile target)

Per trattare e individuare tutti i valori "particolari" della variabile target (Patv) presenti nel dataset è necessario fare diversi controlli in base ai valori che essa e gli altri campi possono avere. In particolare si avranno:

- Valori Nulli, Ci sono valori della potenza attiva che possono essere stati registrati come negativi. In questi casi il valore verra settato come nullo.

- Missing Values, Alcuni valori a volte non vengono registrati dal sistema SCADA. Le previsioni di essi non devono essere usate per la valutazione del modello, ovvero $|Patv - \bar{Patv}| = 0$ con $\bar{Patv}$ il valore predetto dal modello.

- Valori Sconosciuti, Le turbine eoliche possono venir interrotte per motivi esterni come il rinnovo di esse e/o la programmazione dell'alimentazione per evitare il sovraccarico della rete. In questi casi la potenza attiva effettiva generata dalla turbina eolica è sconosciuta. Anche in questa circostanza i valori non verranno utilizzati per valutare il modello. Per individuare questa categoria di valori bisogna verificare una delle seguenti condizioni:

  - Se $Patv \leq 0$ e $Wspd > 2.5$ allora il valore della potenza attiva effettiva è sconosciuto
  - Se $Pab1 > 89$ o $Pab2 > 89$ o $Pab3 > 89$ allora il valore della potenza attiva effettiva è sconosciuto

- Valori Anomali, Ci sono alcuni valori anomali registrati dal sistema SCADA e anche in questa circostanza essi non verranno utilizzati per valutare il modello. Per individuare questa categoria di valori bisogna verificare una delle seguenti condizioni:

  - L'intervallo ragionevole per Ndir è $[-720, 720]$, poiché il sistema a turbina consente alla gondola di girare al massimo due giri in una direzione altrimenti la costringerebbe a tornare nella posizione originale. Pertanto, i record oltre l'intervallo possono essere visti come valori anomali causati dal sistema di registrazione. Quindi, se $Ndir > 720$ o $Ndir < -720$ allora la potenza attiva effettiva è anormale.
  - L'intervallo ragionevole per Wdir è $[-180, 180]$. I record al di fuori di questo intervallo possono essere visti come valori anomali causati dal sistema di registrazione. Se $Wdir > 180$ o $Wdir < -180$ allora la potenza attiva effettiva è anormale.

3.2.2 Data Preparation

Nella sottosezione precendente sono stati elencati i modi per individuare i valori particolari della potenza attiva, ma non come trattarli o sostituirli (solo nel caso di valori negativi viene esplicitato di porli uguali a 0) nello pseudocodice sottostante viene mostrato l'approccio di data preparation utilizzato:

1. INIZIO
2. Per ogni osservazione della pala eolica i
   - 2.1. Se si sta analizzando la prima osservazione in assoluto
     - 2.1.1. Se essa è un Valore Nullo, Missing, Sconosciuto o Anomalo
       - 2.1.1.1. Ponila uguale alla prima osservazione successiva che non è un valore "particolare"
     - 2.1.2. Altrimenti mantienila invariata
   - 2.2. Altrimenti:
     - 2.2.1. Se essa è un Valore Nullo, Missing, Sconosciuto o Anomalo
       - 2.2.1.1. Ponila uguale alla osservazione che la precede
       - 2.2.1.2. Salva in lista d'appoggio l'indice della riga del valore particolare individuato
     - 2.2.2. Altrimenti mantienila invariata
   - 2.3. Rimuovi tutti gli indici duplicati presenti nella lista d'appoggio
3. FINE

Si è scelto di trattare questi valori tramite sostituzione all'indietro tranne nel caso della prima ossservazione dove, chiaramente, ciò non è possibile. La sostitutzione in avanti di tutti i valori è da escludere per i seguenti motivi:

- Se l'osservazione i-esima è un valore particolare, sostituirla con la sua successiva non garantisce che essa sia effettivamente un valore normale. Se la i+1 è a sua volta un valore particolare non si risolverebbe il problema (ciò motiva il perchè nella riga 2.1.1.1 viene fatta una ricerca in avanti fino a quando viene individuato un valore normale)

- Siccome si vuole costruire un modello di previsione, sostituire tutti i valori particolari presenti/passati con valori futuri andrebbe a falsare il modello successivo, in quanto avrebbe informazioni future che non dovrebbe avere in certi istanti di tempo

La creazione della lista d'appoggio è fondamentale per tenere traccia degli indici per i quali la valutazione del modello non dovrà essere influenzata. Inoltre è importante il passo 2.3 in quanto potrebbero essere presenti casi speciali dove il valore della potenza effettiva è sia, ad esempio, un Missing Value che Valore Anomalo. In questo caso l'algoritmo andrebbe a salvare più volte l'indice in considerazione e di conseguenza se si andassero a rimuovere le previsioni del modello negli indici della lista si avrebbe un errore di codice in quanto si vorrebbe eliminare l'indice n-esimo (quello del caso speciale) tante volte quante sono le sue ripetizioni in lista.

3.2.3 Analisi Qualitativa e Hyperparameters

Puliti i dati inizia la fase dell'analisi qualitativa. Tramite un algoritmo si riesce a contare quanti lag significativi sono decrescenti (insight utile per poter pensare alla presenza di un trend) e quanti picchi significativi sono individuati nel correlogramma. Come descritto nella sezione 3.1, il plot dell'ACF per essere chiaro ha bisogno di visualizzare un numero consono di lag in modo da estrapolare le informazioni necessarie senza essere disturbati visivamente da lag non importanti. Per fare ciò è necessaria la stima del parametro MA. Di conseguenza si calcola l'ACF e si contano quanti valori sono significativi (ovvero il loro valore assoluto è maggiore o uguale a 0.025). Questo valore sarà il parametro q della Moving Average. È solito visualizzare un numero di lag pari ad almeno un ciclo, in modo da notare possibili picchi stagionali. Nel caso del dataset in analisi abbiamo 245 giorni per 134 pale eoliche ed osservazioni ogni 10 minuti. Un ciclo sarebbe quindi giornaliero, in particolare avremo 144 osservazioni ogni giorno ($245 \cdot 144 \cdot 134 = 4{,}727{,}520$ ovvero il numero totale di dati). Ora possiamo calcolarci nuovamente l'ACF ma ci fermeremo ad un numero di lag pari ad un ciclo (144) più il parametro q trovato precedentemente. In questo modo siamo sicuri di aver colto tutti i lag significativi fino ad una coda più o meno lunga di lag non significativi, ma sempre utili per determinare le caratteristiche della serie.

Salvati i valori della ACF inizia l'algoritmo di individuazione della stagionalità e del numero di lag decrescenti. Il pseudocodice è il seguente:

1. INIZIO
2. Creazione lista d'appoggio per salvare lag corrispondenti a picchi
3. Per ogni lag dell'ACF i
   - 3.1. Se si sta analizzando il secondo lag in assoluto (viene escluso il lag 1 in quanto il suo valore sarebbe la correlazione tra la serie e se stessa senza lag, quindi il suo valore è sempre pari ad 1)
     - 3.1.1. Se il lag precedente i-1 non è nella lista d'appoggio (quindi non è un picco) e il valore del lag presente i è minore rispetto all'ACF del lag precedente i-1
       - 3.1.1.1. Il lag presente i è decrescente
   - 3.2. Altrimenti:
     - 3.2.1. Se il lag precedente i-1 è nella lista d'appoggio (quindi è un picco) e il valore del lag presente i è minore rispetto all'ACF di due lag precedenti i-2
       - 3.2.1.1. Il lag presente i è decrescente
   - 3.3. Se il lag presente i non è l'ultimo in assoluto e il suo valore è superiore all'ACF del lag precedente i-1 e successivo i+1
     - 3.3.1. Il lag presente i è un picco (quindi viene salvato nella lista d'appoggio)
     - 3.3.2. Se la lista d'appoggio ha almeno due picchi al suo interno
       - 3.3.2.1. Se il valore del picco precedente è maggiore del valore del picco presente
         - 3.3.2.1.1. Il lag presente i è decrescente
4. FINE

Andando nel dettaglio per quanto riguarda il conteggio dei lag decrescenti, in riga 3.1.1 semplicemente si escludono i lag picchi e si controlla se il lag è effettivamente decrescente rispetto a quello precedente. In riga 3.2.1 invece si considera la presenza di un picco precedentemente al lag in analisi, di conseguenza la possibile decrescenza la si valuta rispetto al lag che precede il picco. Infine in riga 3.3.2.1 si controlla che anche se i picchi possano essere decrescenti tra di loro. Con questi check si ottiene una variabile contatore che indica il numero dei lag decrescenti nel correlogramma, e con il

quale si può già cominciare a presupporre l'esistenza di un qualsivoglia trend decrescente. Principalmente, però, questa variabile va interpretata come descrittiva, in quanto non è ragionevole appesantire l'output del codice con 134 plot di ACF oltre al già dispendioso calcolo computazionale dei modelli fittati. La lista d'appoggio per i picchi invece è essenziale per determinare la presenza di stangionalità e non è puramente descrittiva.

Quindi per determinare il modello corretto da utilizzare, questa lista dei picchi verrà accompagnata (in linea teorica) da un test statistico sulla presenza o meno di trend. In particolare il WAVK test, un test non parametrico per individuare trend non-monotoni. Tramite l'utilizzo della function *notrend_test()*, che testa l'ipotesi nulla di non avere trend contro l'ipotesi alternativa del WAVK test [funtimes R Package]. Tra i diversi parametri da poter settare verrà utilizzato come metodo di calcolo dei coefficienti autoregressivi quello di Burg, in quanto in alcuni casi speciali il metodo standard di Yule-Walker potrebbe portare ad una stima errata di essi [de Hoon et al. [1996]]. Considerando il modello AR riportato nella sezione 2.3, si definisce l'operatore di backward shift z^{-1} tale per cui $z^{-1}Y_t = Y_{t-1}$. Allora un processo AR può essere espresso come:

$$\bar{Y_t} = z^p \Big(\sum_{i=0}^{p} a_i z^{p-i} \Big)^{-1} \epsilon_t$$

Se si studiano i poli dell'equazione (ovvero le radici in cui si annulla il denominatore) e se le radici giacciono su un cerchio unitario, il processo autoregressivo sarà stazionario solo nel caso in cui ϵ_t sia identico a zero. In tal caso risulterà un processo armonico, costituito da una somma di funzioni coseno. Poiché i poli sul cerchio unitario rappresentano un processo armonico, ci si può aspettare che un processo autoregressivo con i poli vicini al cerchio unitario dimostri una sorta di comportamento pseudo-periodico. In questo caso la funzione di autocovarianza può essere descritta come una somma di funzioni periodiche debolmente smorzate. Inoltre, poiché i termini di rumore ϵ_t sono ancora presenti, il processo autoregressivo può mostrare una sorta di comportamento quasi non stazionario. Siccome il metodo di Yule-Walker si basa sull'uso della funzione di autocovarianza, questa casistica comporta una stima non sempre corretta. Quindi per determinare l'hyperparameter d basta controllare l'esito del test. Se il p-value è minore di 0.05 si rifiuterà l'ipotesi nulla di "no trend" e di conseguenza $d = 1$. Nel caso in cui non si possa rifiutare l'ipotesi nulla, $d = 0$.

Quello che si è descritto è un processo di individuazione del trend molto rigoroso che comporta un costo computazionale molto elevato (solo l'individuazione del trend di una serie temporale di una pala eolica tramite testing impiegava un'ora di calcolo sul cluster di computer dell'HPC). Di conseguenza si è scelto un metodo più empirico, ovvero controllare che se il 50% dei lag presenti nel correlogramma dell'ACF sia discendente allora $d = 1$. Questa soglia è puramente qualitativa, basata su esperienza visiva delle serie temmporali aventi trend. Nonostante non sia rigorosa quanto un testing essa riesce a cogliere la maggior parte dei casi possibili e verrà validata (in relazione a questo studio) nel capitolo 4. Per quanto riguarda la stagionalità invece si utilizzerà la lista d'appoggio definita nel pseudocodice. In particolare bisogna controllare le seguenti condizioni:

- Se la lista è vuota, serie non stagionale
- Se la differenza tra gli indici dei lag, a due a due, non è uguale ad un valore fissato, serie non stagionale. Altrimenti è stagionale

La prima condizione è ovvia, ovvero non sono stati individuati potenziali picchi e di conseguenza la stagionalità è da escludere. Perciò viene posto l'hyperparameter $D = 0$. La seconda condizione è ciò che visivamente un analista controllerebbe per poter dedurre la presenza della componente stagionale, ovvero se i picchi avvengono in modo regolare (ogni n lag, con n fisso). Svolgendo la differenza a due a due degli indici dei lag nella lista, si deve ottenere un valore unico pari ad n in modo da concludere ciò. Avere diversi potenziali picchi ma in intervalli di lag non fissi potrebbe essere semplice rumore, o comunque non si può stabilire con certezza la presenza o meno della stagionalità. Quindi in caso di intervallo dei picchi fisso, l'hyperparameter viene posto $D = 1$ altrimenti $D = 0$.

Dopo la scelta degli hyperparameters q, d e D, si procede con il determinare i restanti. Partendo dal parametro dell'AR, ovvero p, la procedura svolta per q (Moving Average) viene replicata anche per esso ma con l'utilizzo dei lag provenienti dalla PACF. L'aver determinato prima degli altri i parametri d e D serve per prevenire un superfluo calcolo dei p e q in casistiche non possibili. Ovvero, se $d = 1$ e $D = 0$ allora verranno calcolati solamente i parametri p e q della serie detrendalized mentre P e Q verranno ignorati in quanto il modello che verrà scelto non utilizzerà parametri stagionali. Viceversa, se $d = 0 e D = 1$ allora verranno calcolati solamente i parametri P e

Q della serie deseasonalized mentre p e q verranno ignorati. Nel caso in cui $d = 1$ e $D = 1$ allora verranno calcolati sia i parametri non stagionali (p e q) dalla serie detrendalized e sia i parametri stagionali (P e Q) dalla serie deseasonalized. Questo approccio permette di velocizzare notevolmente il processo successivo di fitting in quanto si escludono i modelli non stagionali (quindi $D = 0$) ma che hanno parametri di P e Q diversi da zero, ponendoli pari zero a prescindere. Questa semplificazione è dovuta solo a problemi di tempi di calcolo, con dataset di queste dimensioni è difficile poter costruire modelli che tengano conto di molti parametri e i modelli con hyperparameters stagionali diversi da zero sono molto impegnativi dal punto di vista computazionale.

3.2.4 Fitting Modelli e Feature Selection

Prima di entrare fittare i modelli e procedere con la feature selection bisogna fissare le finestre di training e di validation. Verranno usati 205 giorni per il train set e 20 giorni per il validation set. Vengono tenuti fuori 20 giorni per il test set, ma per via della mancanza di risultati che possano fare da benchmark di test (la competizione dopo la sua chiusura non ha permesso più la visualizzazione degli score ottenuti nei test) esso non verrà utilizzato. Di seguito viene riportato lo pseudocodice del fitting e della feature selection:

1. INIZIO
2. Fissato valore iniziale dello score a 999999
3. Creazione subset degli indici dei regressori da testare (dalla colonna 3 alla 11, vengono esclusi TurbID, Day, Tmstamp e chiaramente la variabile target Patv)
4. Creazione lista d'appoggio per salvataggio delle features selezionate
5. Per 9 volte (numero massimo di regressori inseribili)
   - 5.1. Per ogni regressore presente nel subset i
     - 5.1.1. Se la lista d'appoggio delle features selezionate è vuota
       - 5.1.1.1. Estrazione serie temporale ad una dimensione corrispondente al regressore i
     - 5.1.2. Altrimenti:

5.1.2.1. Estrazione serie temporale ad n+1 dimensioni corrispondenti agli n regressori presenti nella lista d'appoggio delle features selezionate più il regressore i

5.1.3. Definizione train e validation set

5.1.4. Se la serie è con solo trend ($d = 1$ e $D = 0$)

5.1.4.1. Fitting modello ARIMAX sul train set con parametri p e q calcolati dalla serie detrendalized e come regressori esterni la serie temporale ad 1 o n+1 dimensioni estratta precedentemente

5.1.5. Altrimenti se la serie è con solo stagionalità ($d = 0$ e $D = 1$)

5.1.5.1. Fitting modello SARMAX sul train set con parametri P e Q calcolati dalla serie deseasonalized e come regressori esterni la serie temporale ad 1 o n+1 dimensioni estratta precedentemente

5.1.6. Altrimenti se la serie è con trend e stagionalità ($d = 1$ e $D = 1$)

5.1.6.1. Fitting modello SARMIAX sul train set con parametri p e q calcolati dalla serie detrendalized e i parametri P e Q calcolati dalla serie deseasonalized. Come regressori esterni la serie temporale ad 1 o n+1 dimensioni estratta precedentemente

5.1.7. Altrimenti se la serie è stazionaria ($d = 0$ e $D = 0$)

5.1.7.1. Fitting modello ARMAX sul train set con parametri p e q calcolati dalla serie originale. Come regressori esterni la serie temporale ad 1 o n+1 dimensioni estratta precedentemente

5.1.8. Forecasting del modello fittato sul validation set

5.1.9. Tutti i valori predetti negativi vengono posti nulli

5.1.10. I residui dei valori predetti negativi vengono posti nulli

5.1.11. Esclusione dei valori predetti negli indici in cui si era marcata la presenza di un valore "particolare" (viene utilizzata la lista d'appoggio definita nella sezione 3.2.2

5.1.12. Calcolo della media tra le metriche RMSE e MAE

5.1.13. Se lo score calcolato è minore di quello fissato inizialmente/precedente

5.1.13.1. Nuovo score pari a quello appena calcolato

5.1.13.2. Salvataggio dell'indice del regressore attualmente ottimale

5.1.14. Se l'indice del regressore ottimale è nel subset dei papabili regressori

5.1.14.1. Rimuovere indice dalla lista dei papabili regressori rimanenti
5.1.14.2. Aggiungerlo alla lista d'appoggio delle features selezionate
5.1.15. Altrimenti forza uscita dall'algoritmo

6. FINE

La feature selection si basa semplicemente sull'individuare la combinazione di regressori che permette di ottenere lo score migliore possibile, di conseguenza ogni volta che una feature migliora il modello essa viene salvata, vengono valutate tutte le altre e se continua a risultare la migliore viene salvata nella lista d'appoggio ed esclusa dao regressori pababili (il subset creato inizialmente)

3.2.5 Storage dei Regressori e delle Metriche

Il salvataggio dei regressori e delle metriche avviene una volta individuato il modello ottimale. I primi verranno salvati all'interno di una matrice composta come segue:

- Turbina, il suo ID
- Model, il nome del modello fittato (ARMAX, SARMAX, ARIMAX, SARIMAX)
- Regressor 1,...,9, la lista delle features selezionate in ordine di selezione (nella colonna regressor 1 sarà presente il primo campo che ha migliorato gli score rispetto agli altri, e cosi per le restanti colonne)

Per il calcolo e salvataggio delle metriche si procede con il fittare nuovamente il modello con gli hyperparameters opportuni, passando la lista d'appoggio delle features selezionate come regressore esterno. Vengono posti nulli tutti i valori predetti negativi e tutti i corrispettivi residui. Verranno escluse tutte le osservazioni marcate come "particolari" dalla lista d'appoggio creata nella fase di 3.2.2. Le metriche calcolare e salvate saranno la RMSE, MAE e la loro media denominata "Score". Vengono storati in una matrice di output che racchiuderà tutte le informazioni modellistiche per le 134 pale eoliche presenti nel sistema.

La sua struttura è la seguente:

- Turbina, il suo ID
- RMSE, Root Mean Squared Error
- MAE, Mean Absolute Error
- Score, la media dei due sopracitati

Capitolo 4

Risultati Ottenuti

Come anticipato nell'introduzione, in questo capitolo verrà validato l'algoritmo di studio automatizzato, in particolare la soglia euristica imposta sui lag decrescenti a sostituzione del WAVK test e l'approccio utilizzato per lo studio della lista dei picchi. Successivamente verranno riportati i parametri usati per fittare il modello, i regressori selezionati tramite features selection e gli score, il tutto per ogni pala eolica nel database.

In Tabella 4.1 vengono riportati il numero di lag e di picchi individuati per ciascuna turbina. Si può subito notare il numero elevato di lag decrescenti, ciò è un segnale di presenza della componente trend. Infatti vedendo i grafici in Figura 4.1, Figura 4.2, Figura 4.3, Figura 4.4, Figura 4.5, Figura 4.6 e Figura 4.7, si nota in modo chiaro un andamento decrescente dell'ACF. Di conseguenza la validazione dell'approccio per individuare il trend può essere confermata (in relazione a questo dataset) e di conseguenza $d = 1$. Con il WAVK test si potrebbe portare la validazione ad un livello più generale in quanto la soglia è, appunto, euristica e non rigorosa come potrebbe essere invece un processo di testing statistico. Ma per ovviare ai problemi di costo computazione era necessario porsi in una situazione più agevole. Per quanto riguarda i picchi, invece, sembrerebbero essere relativamente pochi (sempre in riferimento alla Tabella 4.1). L'analisi importante da validare è quella descritta nella sottosezione 3.2.3. Tutte le turbine mostrano almeno un numero di picchi pari a 8, di conseguenza la lista d'appoggio non sarà vuota e non si può escludere in questa maniera la stagionalità. Bisogna quindi valutare il secondo punto, ovvero come questi picchi avvengono, se in un intervallo di lag fisso o in modo irregolare. In tutti i casi il risultato restituito è di intervallo non fisso, di conseguenza l'algoritmo non è riuscito

ad individuare un pattern ricorrente di lag picchi. Questa conclusione può essere validata osservando l'andamento della correlazione per ogni turbina in Figura 4.1, Figura 4.2, Figura 4.3, Figura 4.4, Figura 4.5, Figura 4.6, Figura 4.7. Si noti infatti che principalmente lungo le code esistono dei picchi, ma non ricorrenti. Spesso essi si individuano prima dei 100 lag, ma nei successivi 100 non viene riscontrato un pattern simile. In generale il comportamento è molto diverso rispetto ad una serie con trend e stagionalità, come quella riportata precedentemente in Figura 2.2. Di conseguenza, sia con l'approccio canonico sia con quello automatizzato, non si può supporre la presenza di stagionalità per cui si pone $D = 0$.

TurbID	Lag Discendenti	Picchi	TurbID	Lag Discendenti	Picchi	TurbID	Lag Discendenti	Picchi	TurbID	Lag Discendenti	Picchi
1	393	20	41	306	17	81	299	15	121	303	14
2	348	14	42	327	11	82	311	16	122	277	12
3	313	16	43	316	18	83	326	11	123	292	18
4	356	16	44	308	17	84	303	17	124	271	16
5	373	29	45	312	20	85	327	16	125	311	17
6	322	26	46	310	16	86	294	14	126	377	41
7	328	25	47	321	15	87	285	18	127	287	18
8	325	12	48	305	20	88	304	18	128	308	16
9	323	18	49	314	18	89	299	19	129	279	18
10	320	11	50	319	22	90	298	22	130	297	19
11	334	14	51	315	21	91	306	19	131	297	18
12	325	17	52	328	28	92	295	31	132	306	21
13	310	14	53	338	9	93	312	14	133	285	16
14	331	17	54	312	13	94	311	14	134	269	20
15	313	15	55	316	18	95	313	18			
16	301	8	56	317	17	96	320	13			
17	325	14	57	317	13	97	318	11			
18	305	13	58	337	12	98	301	11			
19	322	17	59	301	19	99	400	23			
20	315	11	60	317	22	100	312	12			
21	333	15	61	308	14	101	306	12			
22	331	15	62	328	14	102	290	10			
23	324	14	63	297	12	103	309	12			
24	322	18	64	306	15	104	311	14			
25	314	18	65	301	22	105	321	18			
26	330	21	66	307	15	106	307	17			
27	310	20	67	316	18	107	302	17			
28	335	17	68	289	18	108	297	15			
29	328	15	69	296	25	109	307	22			
30	362	17	70	327	20	110	301	15			
31	287	18	71	305	16	111	287	15			
32	319	14	72	313	16	112	313	16			
33	317	23	73	311	17	113	314	18			
34	319	17	74	320	13	114	303	15			
35	335	14	75	304	18	115	297	14			
36	325	14	76	322	20	116	315	20			
37	327	18	77	307	14	117	293	18			
38	327	10	78	299	14	118	321	28			
39	321	14	79	307	16	119	302	14			
40	320	11	80	313	19	120	297	23			

Tabella 4.1. Tabella del numero di lag discendenti e dei picchi individuati tramite approccio automatizzato per ogni turbina dalla 1-134

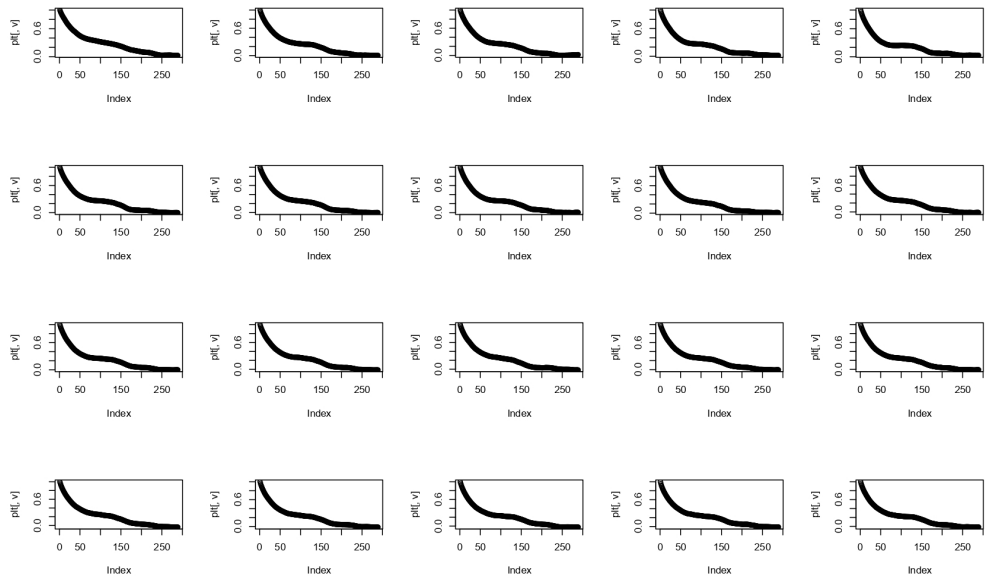

Figura 4.1. ACF delle serie temporali delle pale eoliche dalla 1-20

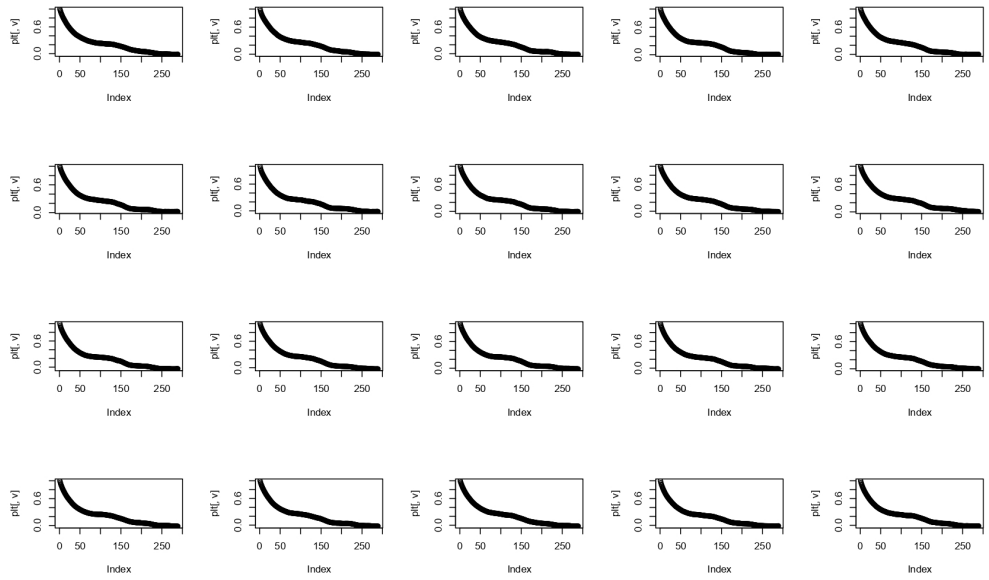

Figura 4.2. ACF delle serie temporali delle pale eoliche dalla 21-40

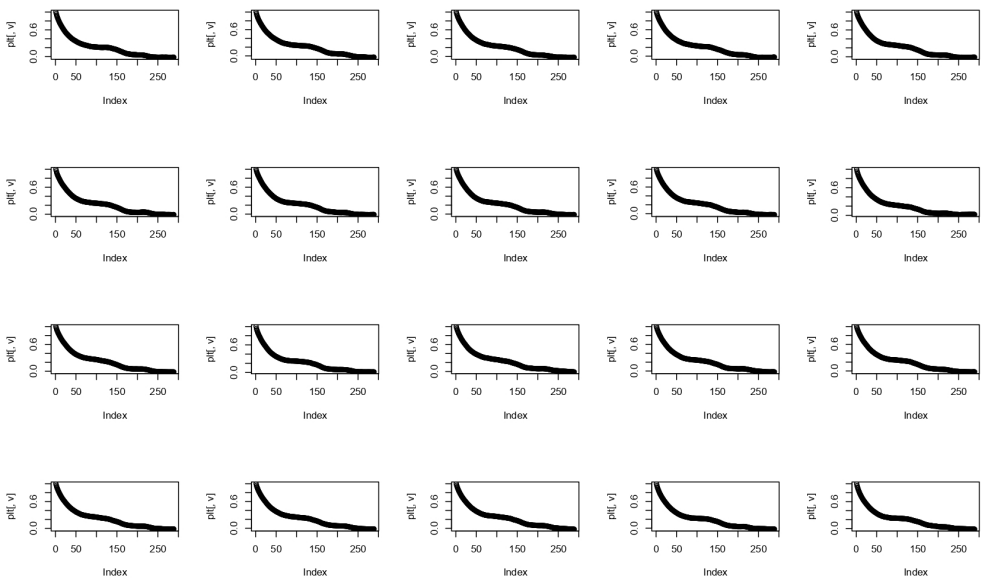

Figura 4.3. ACF delle serie temporali delle pale eoliche dalla 41-60

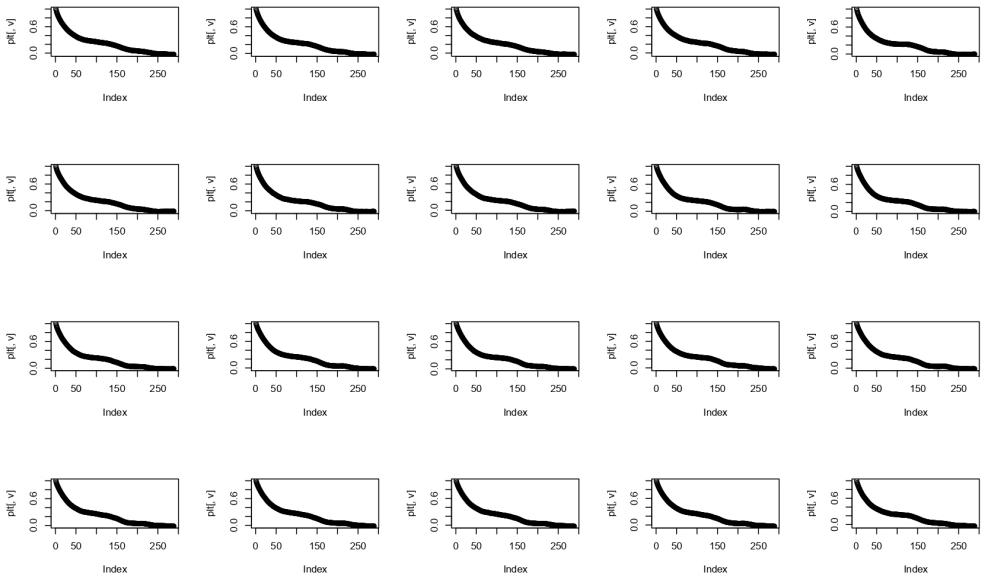

Figura 4.4. ACF delle serie temporali delle pale eoliche dalla 61-80

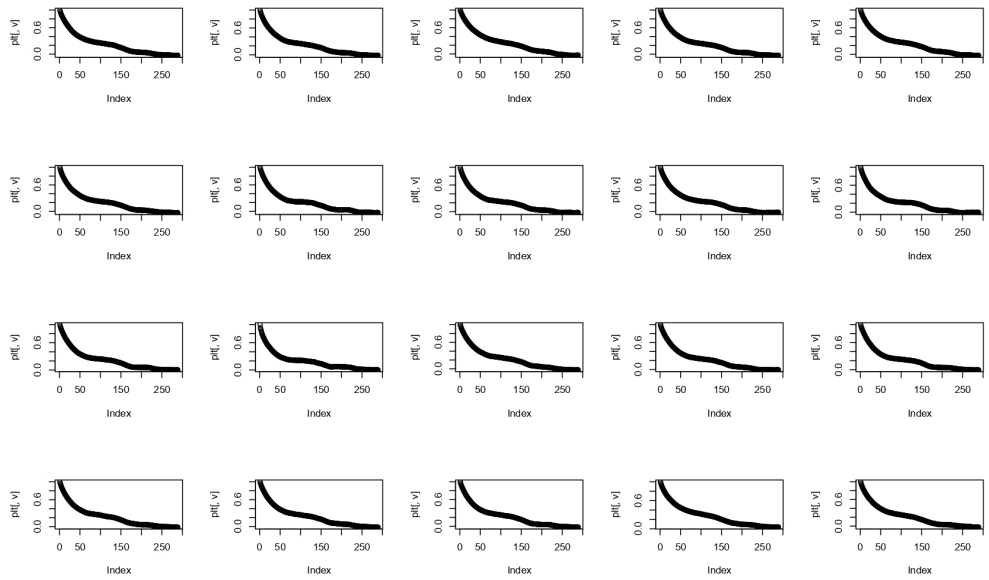

Figura 4.5. ACF delle serie temporali delle pale eoliche dalla 81-100

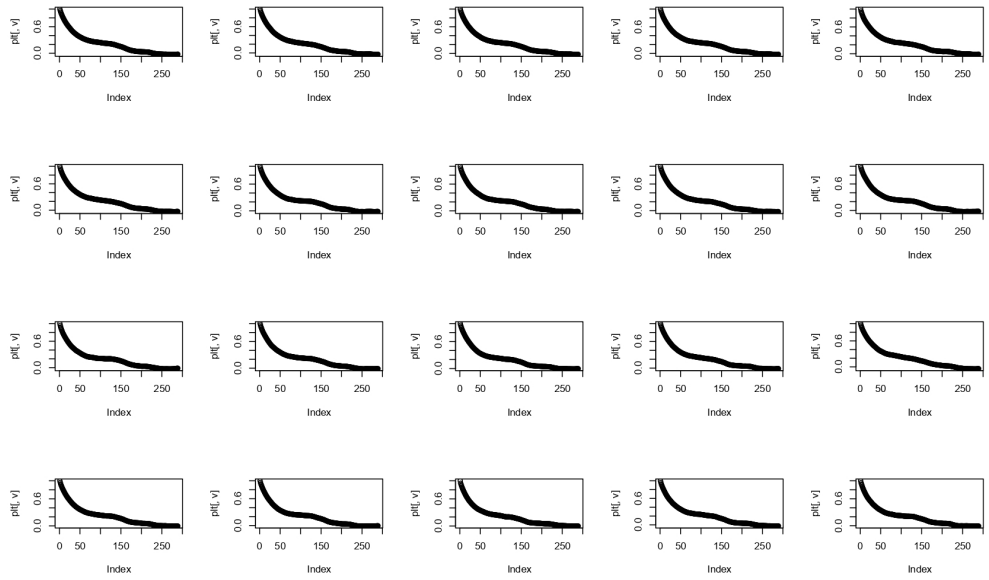

Figura 4.6. ACF delle serie temporali delle pale eoliche dalla 101-120

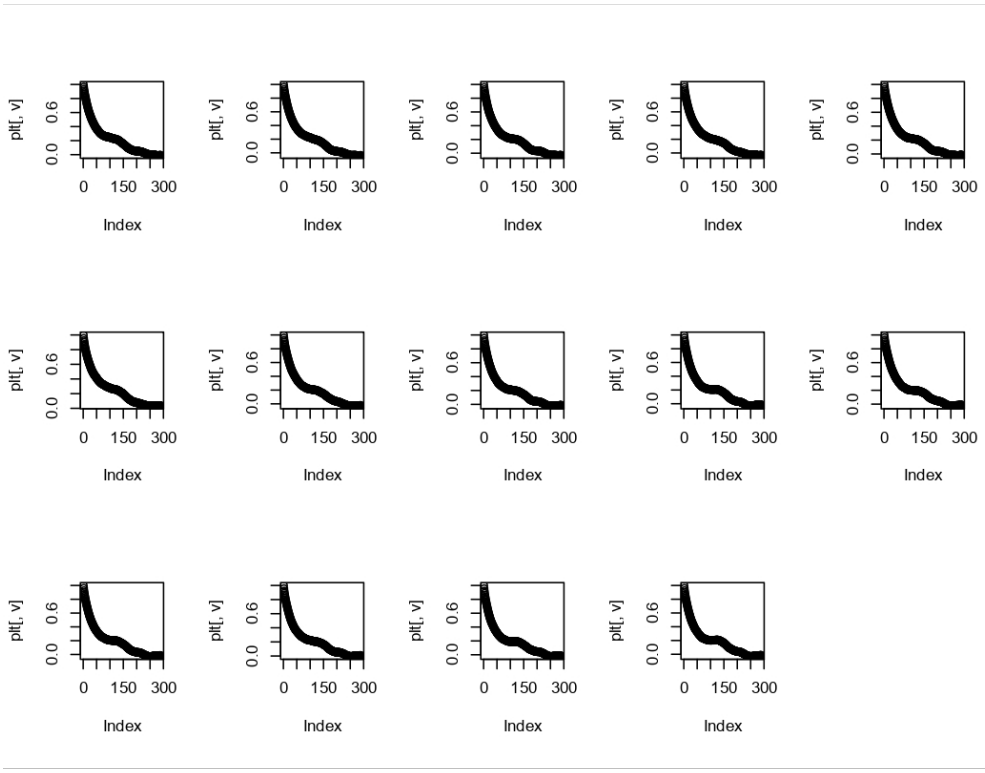

Figura 4.7. ACF delle serie temporali delle pale eoliche dalla 121-134

I risultati ottenuti sono interessanti, in particolare si può notare la complessità dei modelli fittati dagli hyperparameters calcolati. In Tabella 4.2 vengono riportati i valori dei parametri, il Moving Average viene fatto in media su circa 226 lag mentre i coefficienti dell'AutoRegression sono in media 3.5. Per quanto riguarda i regressori selezionati si nota che tutte le turbine hanno come regressore la velocità del vento *Wspd*, risultato atteso in quanto è strettamente correlato con la variabile target. In Tabella 4.3 e Tabella 4.4 vengono riportati ogni subset ottimale di regressori per tutti i modelli fittati. I punteggi invece sono riportati in Tabella 4.5, lo Score medio è pari a 139.4125 con MAE medio uguale a 110.96 e RMSE pari a

167.865. L'obiettivo, come anticipato nell'introduzione, non era migliorare le metriche ottenute dalla baseline, poichè:

- Impossibile poter parallelizzare il codice e sfruttare più nodi del cluster di HPC, le function utilizzate non consentivano ciò e di conseguenza alcune casistiche sono state semplificate (ad esempio l'uso della soglia euristica invece che del test statistico sul trend e l'esclusione dei parametri stagionali P e Q se la serie non fosse risultata stagionale)
- Caso particolare del punto precedente è il metodo di validazione, dove nella baseline è stato svolto calcolandosi la media delle previsioni ottenute su una finestra di validation mobile. Procedura non replicabile con gli strumenti a disposizione in quanto avrebbe portato ulteriore sforzo computazionale e causato uno stallo nello studio.
- L'utilizzo di modellizzazioni diverse, la baseline si basa su una neural network (quindi modello black-box) mentre in questo studio si sono utilizzati modelli che permettono di rendere esplicitabile la function approssimata.

In Tabella 4.6 viene mostrato come il processo di feature selection ha distribuito i diversi regressori. Come prevedibile, il campo *Wspd* è sempre risultato essere il primo regressore in termini di miglioramento degli score del modello. Per quanto riguarda il numero di modelli per numero di regressori, si hanno le seguenti numeriche:

- 1 modello avente 1 regressore esterno
- 22 modelli aventi 2 regressori esterni
- 24 modelli aventi 3 regressori esterni
- 35 modelli aventi 4 regressori esterni
- 30 modelli aventi 5 regressori esterni
- 15 modelli aventi 6 regressori esterni
- 7 modelli aventi 7 regressori esterni
- 0 modelli aventi 8 o 9 regressori esterni

Sommando il numero di modelli si ottengono esattamente 134, ovvero il numero delle turbine nel dataset.

Come si può intuire dai risultati mostrati in questo capitolo, il costo computazionale è stato molto elevato nonostante alcune semplificazioni citate precedentemente. Si parla circa di un giorno di run, riducibile a circa 16 ore lanciando più job in contemporanea studiando un sottoinsieme di turbine alla volta. Per avere un idea del costo computazionale, possiamo calcolarci il numero di modelli fittati dal codice, ricavabile dalla lista riportata sopra. Di conseguenza avremo la seguente formula:

$$\sum_{i=1}^{8} N_i \cdot \frac{9! \cdot (9-i)}{(9-i)!} + N_9 \cdot 9!$$

Dove N_i è il numero di modelli aventi i regressori. il valore $9 - i$ al numeratore indica il numero di modelli fittati successivamente alla scelta dell'ultimo regressore ma che non hanno portato a migliorare lo score e di conseguenza a nessuna selezione di un nuovo regressore. Ad esempio, ad inizio algoritmo vengono fittati almeno 9 modelli. Viene individuato un regressore in grado di migliorare il più possibile lo score, una volta selezionato si fitteranno 8 modelli (ora con al loro interno il regressore scelto precedentemente) per individuarne uno nuovo. Nel caso in cui ciò non avvenga, il numero di modelli fittati è 9 che moltiplica 8, dove i primi 9 hanno portato a selezionare un regressore mentre gli altri 8 no. Di conseguenza si osservi che nel caso di modelli con 8 o 9 regressori, il numero totale di modelli fittati è il medesimo in quanto l'algoritmo proverà a vedere se l'integrazione della variabile esterna rimasta possa portare un aumento della qualità del fitting, a prescindere se successivamente essa verrà integrata. Ecco spiegato il motivo della somma con l'ultima componente. Nel caso in analisi il numero totale di modelli fittati è pari a 7,689,096. Un numero considerevolmente elevato, che aggiunto alla complessità dei modelli fittati rende i tempi computazionali molto lunghi.

TurbID	p	d	q	P	D	Q
1	3	1	301	0	0	0
2	1	1	227	0	0	0
3	3	1	226	0	0	0
4	4	1	263	0	0	0
5	1	1	313	0	0	0
6	1	1	231	0	0	0
7	1	1	230	0	0	0
8	3	1	225	0	0	0
9	1	1	230	0	0	0
10	1	1	220	0	0	0
11	3	1	229	0	0	0
12	3	1	224	0	0	0
13	3	1	226	0	0	0
14	1	1	227	0	0	0
15	1	1	221	0	0	0
16	4	1	207	0	0	0
17	4	1	227	0	0	0
18	5	1	217	0	0	0
19	5	1	228	0	0	0
20	3	1	216	0	0	0
21	1	1	225	0	0	0
22	1	1	231	0	0	0
23	4	1	228	0	0	0
24	4	1	226	0	0	0
25	4	1	228	0	0	0
26	1	1	243	0	0	0
27	6	1	227	0	0	0
28	4	1	229	0	0	0
29	5	1	230	0	0	0
30	6	1	252	0	0	0
31	5	1	199	0	0	0
32	5	1	214	0	0	0
33	4	1	228	0	0	0
34	3	1	221	0	0	0
35	3	1	233	0	0	0
36	3	1	225	0	0	0
37	5	1	226	0	0	0
38	3	1	220	0	0	0
39	3	1	220	0	0	0
40	5	1	222	0	0	0
41	5	1	218	0	0	0
42	3	1	222	0	0	0
43	5	1	216	0	0	0
44	3	1	216	0	0	0
45	4	1	226	0	0	0
46	1	1	230	0	0	0
47	4	1	225	0	0	0
48	3	1	227	0	0	0
49	4	1	221	0	0	0
50	4	1	235	0	0	0
51	4	1	227	0	0	0
52	5	1	238	0	0	0
53	4	1	238	0	0	0
54	4	1	229	0	0	0
55	3	1	221	0	0	0
56	3	1	228	0	0	0
57	3	1	230	0	0	0
58	5	1	233	0	0	0
59	4	1	222	0	0	0
60	3	1	224	0	0	0
61	5	1	220	0	0	0
62	3	1	220	0	0	0
63	3	1	211	0	0	0
64	3	1	215	0	0	0
65	5	1	222	0	0	0
66	5	1	214	0	0	0
67	3	1	221	0	0	0
68	3	1	197	0	0	0
69	4	1	226	0	0	0
70	3	1	227	0	0	0
71	3	1	220	0	0	0
72	4	1	225	0	0	0
73	4	1	229	0	0	0
74	3	1	227	0	0	0
75	4	1	224	0	0	0
76	4	1	226	0	0	0
77	3	1	230	0	0	0
78	4	1	223	0	0	0
79	3	1	222	0	0	0
80	3	1	212	0	0	0
81	4	1	217	0	0	0
82	4	1	230	0	0	0
83	4	1	224	0	0	0
84	4	1	220	0	0	0
85	3	1	224	0	0	0
86	3	1	208	0	0	0
87	3	1	220	0	0	0
88	5	1	217	0	0	0
89	5	1	215	0	0	0
90	4	1	220	0	0	0
91	3	1	231	0	0	0
92	4	1	235	0	0	0
93	4	1	222	0	0	0
94	3	1	223	0	0	0
95	3	1	231	0	0	0
96	4	1	228	0	0	0
97	4	1	227	0	0	0
98	3	1	224	0	0	0
99	3	1	384	0	0	0
100	3	1	221	0	0	0
101	3	1	215	0	0	0
102	3	1	212	0	0	0
103	3	1	222	0	0	0
104	3	1	218	0	0	0
105	4	1	229	0	0	0
106	3	1	213	0	0	0
107	3	1	218	0	0	0
108	6	1	214	0	0	0
109	5	1	214	0	0	0
110	4	1	216	0	0	0
111	4	1	213	0	0	0
112	3	1	221	0	0	0
113	3	1	225	0	0	0
114	3	1	224	0	0	0
115	3	1	207	0	0	0
116	4	1	229	0	0	0
117	3	1	220	0	0	0
118	4	1	233	0	0	0
119	3	1	220	0	0	0
120	3	1	224	0	0	0
121	4	1	227	0	0	0
122	4	1	196	0	0	0
123	3	1	214	0	0	0
124	3	1	195	0	0	0
125	3	1	219	0	0	0
126	3	1	348	0	0	0
127	4	1	208	0	0	0
128	3	1	216	0	0	0
129	3	1	211	0	0	0
130	3	1	217	0	0	0
131	4	1	212	0	0	0
132	5	1	227	0	0	0
133	4	1	219	0	0	0
134	4	1	214	0	0	0

Tabella 4.2. Tabella degli hyperparameters calcolati tramite approccio automatizzato per ogni turbina dalla 1-134

Turbine	Model	Regressor 1	Regressor 2	Regressor 3	Regressor 4	Regressor 5	Regressor 6	Regressor 7	Regressor 8	Regressor 9
1	ARIMAX	Wspd	Itmp							
2	ARIMAX	Wspd	Itmp	Pab1	Pab2	Wdir				
3	ARIMAX	Wspd	Pab2	Etmp						
4	ARIMAX	Wspd	Pab3	Itmp	Etmp	Wdir				
5	ARIMAX	Wspd	Pab3	Ndir	Prtv	Pab2				
6	ARIMAX	Wspd	Pab3	Ndir	Pab2					
7	ARIMAX	Wspd	Itmp	Pab1	Ndir	Pab3				
8	ARIMAX	Wspd	Pab2	Ndir	Itmp					
9	ARIMAX	Wspd	Pab2	Itmp	Ndir					
10	ARIMAX	Wspd	Pab1	Ndir	Etmp	Itmp	Prtv			
11	ARIMAX	Wspd	Pab1	Ndir	Pab3	Etmp	Itmp	Pab2		
12	ARIMAX	Wspd	Pab3	Ndir	Itmp	Pab1				
13	ARIMAX	Wspd	Pab1	Prtv	Etmp	Itmp	Wdir			
14	ARIMAX	Wspd	Pab2	Etmp						
15	ARIMAX	Wspd	Pab2	Ndir	Etmp	Itmp	Pab1	Pab3		
16	ARIMAX	Wspd	Pab2	Pab1	Etmp					
17	ARIMAX	Wspd	Pab1	Ndir						
18	ARIMAX	Wspd	Pab2	Ndir						
19	ARIMAX	Wspd	Pab1	Prtv	Etmp	Pab1				
20	ARIMAX	Wspd	Pab1	Pab3						
21	ARIMAX	Wspd	Pab3	Wdir						
22	ARIMAX	Wspd	Pab1	Itmp	Prtv	Pab2				
23	ARIMAX	Wspd	Pab1	Pab2	Pab3					
24	ARIMAX	Wspd	Pab1	Pab2	Pab3					
25	ARIMAX	Wspd	Pab2	Ndir	Etmp					
26	ARIMAX	Wspd	Pab1	Wdir	Etmp	Itmp				
27	ARIMAX	Wspd	Pab1	Prtv	Itmp					
28	ARIMAX	Wspd								
29	ARIMAX	Wspd	Pab2	Ndir	Pab1	Etmp	Itmp			
30	ARIMAX	Wspd	Ndir	Etmp	Pab2	Prtv	Pab1			
31	ARIMAX	Wspd	Pab3	Itmp						
32	ARIMAX	Wspd	Pab3	Pab1	Prtv	Ndir	Pab2			
33	ARIMAX	Wspd	Pab3	Prtv	Wdir					
34	ARIMAX	Wspd	Pab2	Wdir	Pab3					
35	ARIMAX	Wspd	Pab1	Etmp	Prtv	Ndir				
36	ARIMAX	Wspd	Pab3	Ndir	Pab1					
37	ARIMAX	Wspd	Pab2	Ndir	Etmp					
38	ARIMAX	Wspd	Pab1							
39	ARIMAX	Wspd	Pab2	Ndir	Pab1					
40	ARIMAX	Wspd	Pab2							
41	ARIMAX	Wspd	Pab3							
42	ARIMAX	Wspd	Pab1							
43	ARIMAX	Wspd	Wdir							
44	ARIMAX	Wspd	Pab1	Ndir	Pab2					
45	ARIMAX	Wspd	Pab2	Etmp	Prtv	Itmp				
46	ARIMAX	Wspd	Pab3	Etmp	Prtv	Ndir	Pab2	Pab1		
47	ARIMAX	Wspd	Pab3	Etmp	Itmp	Ndir				
48	ARIMAX	Wspd	Pab1	Ndir	Itmp	Wdir				
49	ARIMAX	Wspd	Pab3	Etmp	Ndir	Pab2				
50	ARIMAX	Wspd	Pab2							
51	ARIMAX	Wspd	Pab3	Wdir						
52	ARIMAX	Wspd	Pab3	Ndir						
53	ARIMAX	Wspd	Pab1	Prtv	Wdir	Pab3				
54	ARIMAX	Wspd	Pab3							
55	ARIMAX	Wspd	Pab1	Ndir	Etmp	Itmp	Pab1			
56	ARIMAX	Wspd	Pab1	Itmp						
57	ARIMAX	Wspd	Pab3	Itmp	Etmp	Ndir	Wdir			
58	ARIMAX	Wspd	Pab1							
59	ARIMAX	Wspd	Pab2	Pab1	Wdir	Prtv	Pab3	Itmp		
60	ARIMAX	Wspd	Pab1	Etmp	Prtv	Ndir	Wdir			
61	ARIMAX	Wspd	Pab3							
62	ARIMAX	Wspd	Pab3	Ndir	Etmp	Itmp	Pab2	Wdir		
63	ARIMAX	Wspd	Ndir	Pab1	Etmp	Itmp				
64	ARIMAX	Wspd	Pab2	Etmp	Itmp	Ndir				
65	ARIMAX	Wspd	Pab3	Etmp	Pab2					
66	ARIMAX	Wspd	Pab1	Ndir						
67	ARIMAX	Wspd	Ndir	Etmp	Itmp	Pab2				

Tabella 4.3. Tabella dei regressori selezionati tramite l'algoritmo di features selection per ogni turbina dalla 1-67

Turbine	Model	Regressor 1	Regressor 2	Regressor 3	Regressor 4	Regressor 5	Regressor 6	Regressor 7	Regressor 8	Regressor 9
68	ARIMAX	Wspd	Etmp							
69	ARIMAX	Wspd	Pab2	Pab1						
70	ARIMAX	Wspd	Pab2	Ndir	Prtv	Itmp	Pab3	Wdir		
71	ARIMAX	Wspd	Pab2	Etmp	Prtv	Wdir	Pab1			
72	ARIMAX	Wspd	Pab1	Pab3	Prtv					
73	ARIMAX	Wspd	Pab2	Etmp						
74	ARIMAX	Wspd	Pab3	Wdir						
75	ARIMAX	Wspd	Pab1	Ndir	Prtv	Itmp				
76	ARIMAX	Wspd	Pab3	Prtv	Ndir	Etmp	Wdir			
77	ARIMAX	Wspd	Pab1	Ndir	Etmp	Itmp	Pab2			
78	ARIMAX	Wspd	Pab1	Pab2	Prtv					
79	ARIMAX	Wspd	Pab1	Ndir	Prtv					
80	ARIMAX	Wspd	Pab1	Etmp	Prtv	Wdir				
81	ARIMAX	Wspd	Wdir	Prtv						
82	ARIMAX	Wspd	Wdir	Pab1	Prtv					
83	ARIMAX	Wspd	Wdir	Pab2						
84	ARIMAX	Wspd	Wdir	Pab3	Prtv	Ndir				
85	ARIMAX	Wspd	Wdir							
86	ARIMAX	Wspd	Pab2	Pab1	Itmp	Pab3				
87	ARIMAX	Wspd	Pab3	Ndir	Prtv	Pab1				
88	ARIMAX	Wspd	Pab2	Wdir						
89	ARIMAX	Wspd	Pab2							
90	ARIMAX	Wspd	Pab3							
91	ARIMAX	Wspd	Pab1							
92	ARIMAX	Wspd	Pab3	Prtv	Pab1	Wdir				
93	ARIMAX	Wspd	Pab3	Prtv						
94	ARIMAX	Wspd	Pab3	Ndir	Etmp	Itmp	Prtv			
95	ARIMAX	Wspd	Pab2	Prtv	Etmp	Pab1	Wdir			
96	ARIMAX	Wspd	Pab1	Ndir	Prtv	Itmp				
97	ARIMAX	Wspd	Pab3	Prtv	Etmp					
98	ARIMAX	Wspd	Pab2	Prtv	Pab3					
99	ARIMAX	Wspd	Pab3	Itmp	Prtv	Etmp				
100	ARIMAX	Wspd	Pab3	Prtv	Etmp					
101	ARIMAX	Wspd	Pab3							
102	ARIMAX	Wspd	Pab3	Prtv	Wdir					
103	ARIMAX	Wspd	Prtv							
104	ARIMAX	Wspd	Pab2	Prtv	Itmp	Etmp	Ndir			
105	ARIMAX	Wspd	Pab3							
106	ARIMAX	Wspd	Pab2	Ndir	Wdir					
107	ARIMAX	Wspd	Pab2	Pab3	Ndir					
108	ARIMAX	Wspd	Pab2							
109	ARIMAX	Wspd	Pab3	Prtv	Wdir	Ndir	Pab2			
110	ARIMAX	Wspd	Pab1	Etmp	Ndir	Pab2				
111	ARIMAX	Wspd	Ndir	Prtv	Pab1	Itmp				
112	ARIMAX	Wspd	Pab2	Etmp	Itmp					
113	ARIMAX	Wspd	Pab1							
114	ARIMAX	Wspd	Pab1	Etmp	Pab2					
115	ARIMAX	Wspd	Pab2	Prtv	Pab3	Ndir				
116	ARIMAX	Wspd	Pab3	Prtv						
117	ARIMAX	Wspd	Ndir	Pab3	Pab2					
118	ARIMAX	Wspd	Pab1	Prtv	Etmp					
119	ARIMAX	Wspd	Pab1	Pab3						
120	ARIMAX	Wspd	Prtv	Pab2	Etmp	Itmp	Pab1	Wdir		
121	ARIMAX	Wspd	Pab2	Ndir	Prtv	Etmp				
122	ARIMAX	Wspd	Ndir	Wdir						
123	ARIMAX	Wspd	Pab2	Ndir	Prtv	Pab1				
124	ARIMAX	Wspd	Pab2	Etmp	Prtv					
125	ARIMAX	Wspd	Pab3	Etmp	Ndir					
126	ARIMAX	Wspd	Prtv	Etmp	Itmp					
127	ARIMAX	Wspd	Pab1	Prtv	Pab3					
128	ARIMAX	Wspd	Pab3	Prtv						
129	ARIMAX	Wspd	Ndir	Pab1	Prtv					
130	ARIMAX	Wspd	Pab1	Ndir	Pab3					
131	ARIMAX	Wspd	Ndir							
132	ARIMAX	Wspd	Pab2	Prtv						
133	ARIMAX	Wspd	Pab2	Pab1						
134	ARIMAX	Wspd	Pab2							

Tabella 4.4. Tabella dei regressori selezionati tramite l'algoritmo di features selection per ogni turbina dalla 68-134

Turbina	RMSE	MAE	Score	Turbina	RMSE	MAE	Score	Turbina	RMSE	MAE	Score	Turbina	RMSE	MAE	Score
1	166.125	108.069	137.097	41	161.717	114.444	138.081	81	159.545	100.089	129.817	121	172.762	116.813	144.788
2	247.49	137.736	192.613	42	144.511	102.520	123.516	82	151.209	77.403	114.306	122	130.134	79.044	104.589
3	220.184	168.114	194.149	43	132.291	87.308	109.799	83	153.515	107.295	130.405	123	142.462	105.396	123.929
4	201.947	134.413	168.18	44	146.121	95.833	120.977	84	168.008	101.973	134.990	124	150.615	95.945	123.280
5	231.771	127.596	179.684	45	184.208	144.423	164.315	85	138.564	84.521	111.543	125	170.768	123.451	147.109
6	244.442	197.481	220.961	46	185.474	107.355	146.414	86	140.220	96.694	118.457	126	116.684	69.578	93.131
7	206.481	128.256	167.368	47	163.988	104.036	134.012	87	130.336	78.866	104.601	127	141.184	89.206	115.195
8	211.506	156.292	183.899	48	181.377	116.634	149.006	88	140.415	86.227	113.321	128	135.381	92.840	114.111
9	201.603	141.41	171.507	49	178.719	123.103	150.911	89	134.540	80.892	107.716	129	143.034	92.554	117.794
10	175.838	129.503	152.670	50	182.681	100.328	141.505	90	128.347	81.679	105.013	130	153.881	111.321	132.601
11	191.606	135.117	163.362	51	173.312	109.547	141.429	91	187.191	150.333	168.762	131	140.992	81.247	111.119
12	206.486	152.725	179.606	52	154.066	100.207	127.137	92	134.413	72.341	103.377	132	143.701	103.465	123.583
13	195.026	117.722	156.374	53	156.883	83.291	120.087	93	193.779	118.619	156.199	133	136.843	80.790	108.816
14	187.357	128.090	157.723	54	194.849	126.484	160.667	94	144.151	97.184	120.668	134	141.047	87.718	114.383
15	193.946	133.840	163.893	55	152.748	96.534	124.641	95	277.266	160.218	218.742				
16	192.889	119.846	156.368	56	167.241	115.893	141.567	96	157.066	114.472	135.769				
17	165.310	105.325	135.317	57	164.752	108.042	136.397	97	122.151	74.418	98.284				
18	177.249	138.466	157.857	58	185.439	125.909	155.674	98	191.548	128.477	160.012				
19	154.758	110.250	132.504	59	166.588	111.862	139.225	99	248.719	147.036	197.877				
20	174.197	122.778	148.488	60	162.812	97.099	129.955	100	159.908	119.230	139.569				
21	169.099	103.884	136.491	61	156.551	101.562	129.056	101	179.608	133.229	156.419				
22	206.700	135.885	171.293	62	143.890	97.010	120.450	102	151.134	96.459	123.796				
23	176.109	117.637	146.873	63	161.536	113.217	137.376	103	142.230	93.256	117.743				
24	215.150	159.663	187.407	64	165.922	86.189	126.056	104	117.542	74.572	96.057				
25	200.088	127.623	163.856	65	135.140	89.320	112.230	105	155.199	95.947	125.573				
26	198.267	139.649	168.958	66	150.726	90.277	120.502	106	128.517	73.651	101.084				
27	189.289	124.431	156.860	67	154.626	94.121	124.374	107	129.785	90.069	109.927				
28	203.987	152.931	178.459	68	139.657	96.737	118.197	108	132.966	82.617	107.792				
29	168.764	122.078	145.421	69	183.232	125.291	154.262	109	142.090	97.167	119.629				
30	162.524	97.207	129.865	70	177.358	115.075	146.216	110	147.864	101.905	124.884				
31	188.523	127.946	158.235	71	180.560	116.160	148.360	111	123.779	79.903	101.841				
32	191.415	128.146	159.780	72	189.446	127.900	158.673	112	170.545	119.766	145.156				
33	152.947	109.966	131.456	73	193.311	136.902	165.106	113	195.568	116.487	156.027				
34	178.625	113.761	146.193	74	168.066	106.745	137.406	114	156.158	86.938	121.548				
35	189.153	124.394	156.774	75	185.210	121.483	153.346	115	178.452	126.201	152.326				
36	183.089	117.899	150.494	76	195.819	136.266	166.043	116	174.891	122.668	148.780				
37	181.795	110.736	146.265	77	179.173	105.242	142.207	117	134.239	94.532	114.385				
38	199.024	131.260	165.142	78	139.925	95.707	117.816	118	166.506	116.398	141.452				
39	183.035	120.793	151.914	79	155.276	100.494	127.885	119	171.543	124.795	148.169				
40	160.604	115.539	138.072	80	153.780	95.440	124.610	120	101.110	68.739	84.924				

Tabella 4.5. Tabella dei valori delle metriche selezionate per ogni turbina dalla 1-134

Var	Regressor 1	Regressor 2	Regressor 3	Regressor 4	Regressor 5	Regressor 6	Regressor 7	Regressor 8	Regressor 9	Total
Wspd	134	0	0	0	0	0	0	0	0	134
Wdir	0	6	7	6	6	5	3	0	0	33
Etmp	0	1	21	20	6	0	0	0	0	48
Itmp	0	3	7	11	15	2	1	0	0	39
Ndir	0	8	31	7	10	1	0	0	0	57
Pab1	0	37	11	5	5	5	1	0	0	64
Pab2	0	38	5	7	5	5	1	0	0	61
Pab3	0	37	6	8	3	2	1	0	0	57
Prtv	0	3	23	23	2	2	0	0	0	53

Tabella 4.6. Tabella della distribuzione dei regressori

Capitolo 5

Conclusioni e Possibili Sviluppi

Il progetto può ritenersi concluso. L'obiettivo di automatizzare lo studio qualitativo, la scelta degli hyperparameters, la features selection e il fitting dei modelli è stato raggiunto. Soprattutto lo studio qualitativo, il quale serve per mettere delle basi solide alla modellizzazione della serie temporale sotto osservazione. Lo step più difficile era cercare di strutturare un algoritmo in grado di cogliere tutti i casi particolari di andamento di una qualsiasi serie ed essere in grado di restituire dei plot dai quali estrapolare insights chiave come le componenti del trend e stagionali. Le logiche dietro la scelta di una certa modellizzazione devono essere le più generali possibili, in quanto l'idea di questo approccio era quello di poter ridurre l'effort del lavoro di modellizzazione al solo run del codice. Ciò permetterebbe all'algoritmo di avere in pasto nuovi dati aggiornati (ove disponibili) e poter a sua volta aggiornare lo studio della serie precedentemente svolto. Inoltre è importante anche ridurre il più possibile la complessità di calcolo, in modo da poter avere il giusto equilibrio tra qualità di risultato e tempi di attesa. A fronte di ciò, la selezione di un numero ottimale di lag ha permesso di non prolungare calcoli computazioni su valori non significativi, stesso discorso per la soglia euristica (posta per l'eccessiva durata delle run che causava il fare inferenza) e l'esclusione di modelli molto complessi (come quelli stagionali). Avendo fisso in mente di sviluppare un algoritmo che possa replicare con ottime performance le analisi che un analista farebbe, viene da se che anche il tempo computazionale gioca un ruolo importante. Se l'algoritmo impiega molto a restituire un output, a prescindere dalla sua qualità, l'analista potrebbe far

prima ad analizzare nuovamente in modo canonico i nuovi dati e modificare dove necessario il modello.

Ponendosi in un ipotetico scenario dove non si abbiano limiti tecnici, si potrebbe portare questo approccio ad un livello ancora più generale e rigoroso allo stesso tempo. Generale poichè si potrebbe fare in modo di offrire un servizio al cliente di previsione riguardo qualsiasi dataset, di qualsiasi dimensione e caratteristica. Si potrebbe implementare la possibilità di far individuare in modo autonomo la frequenza della serie che bisogna studiare (ovvero fare in modo che l'algoritmo riconosca in autonomia se la serie è mensile, annuale ecc.). Poter far inserire al cliente parametri esterni se ritenuti utili, il tutto tramite un'interfaccia digitale e interattiva sul web. Un servizio simile viene già offerto da Facebook, si tratta di Prophet. La sua descrizione è la seguente *"Esso Implementa una procedura per la previsione dei dati delle serie temporali basata su un modello additivo in cui le tendenze non lineari si adattano alla stagionalità annuale, settimanale e giornaliera, oltre agli effetti delle vacanze. Funziona meglio con serie temporali che hanno forti effetti stagionali e diverse stagioni di dati storici"* prophet R Package. Da come si legge ci sono diversi vincoli con questo approccio, a partire dall'ipotesi che la serie sia additiva e che riesca ad analizzare effetti stagionali fino ad un livello settimanale. Già per il dataset in questione questo approccio era da escludere a priori.

L'automatizzazione dei modelli di previsione è un ramo ancora acerbo ma con grandi potenzialità, che può portare molti benefici a chi ne ususfruisce e in generale al progresso tecnologico e scientifico.

Bibliografia

Paul S. P. Cowpertwait and Andrew V. Metcalfe. *Introductory Time Series with R*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 0387886974.

M.J.L. de Hoon, T.H.J.J. van der Hagen, Hijlke Schoonewelle, and Hugo van Dam. Why yule-walker should not be used for autoregressive modelling. *Annals of Nuclear Energy*, 23:1219–1228, 1996.

Enegesele Dennis, Iheanyi Iwueze, Maxwell Ijomah, and Taiwo Owolabi. Methods for choice of model in descriptive time se-ries : A review with example. *International Journal of Advanced Statistics and Probability*, 6: 10, 12 2017. doi: 10.14419/ijasp.v6i1.8606.

funtimes R Package. https://cran.r-project.org/web/packages/funtimes/funtimes.pdf.

R. Krispin. *Hands-On Time Series Analysis with R: Perform Time Series Analysis and Forecasting Using R*. Packt Publishing, 2019. ISBN 9781788629157. URL https://books.google.it/books?id=F9KytQEACAAJ.

F. Lazzeri. *Machine Learning for Time Series Forecasting with Python*. Wiley, 2020. ISBN 9781119682363. URL https://books.google.it/books?id=fKN_zQEACAAJ.

R. Nau. *Notes on nonseasonal ARIMA models*. Fuqua School of Business, Duke University, 2014.

prophet R Package. https://cran.r-project.org/web/packages/prophet/index.html.

G. Shmueli and K.C. Lichtendahl. *Practical Time Series Forecasting with R: A Hands-On Guide [2nd Edition]*. Practical Analytics. Axelrod Schnall Publishers, 2016. ISBN 9780997847918. URL https://books.google.it/books?id=SOtgvgAACAAJ.