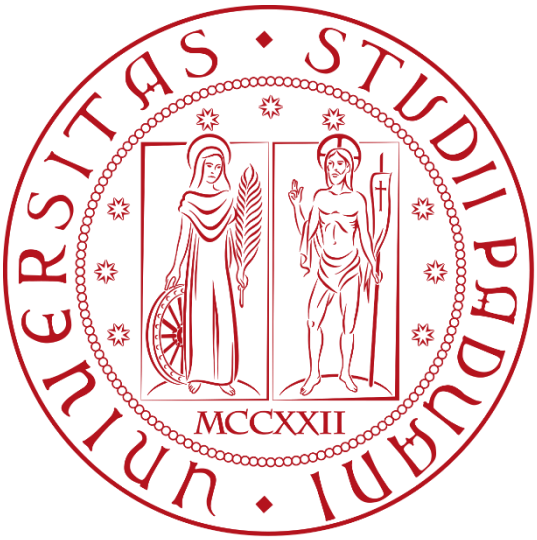
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in Statistica per l'Economia e l'Impresa



Relazione Finale

Modellazione e previsione della produzione di energia fotovoltaica in Italia

Relatore: Prof. Francesco Lisi
Dipartimento di Scienze Statistiche

Laureando: Giacomo Scaglia
Matricola N° 2003669

Anno Accademico 2022/2023

Alla mia famiglia,
che ha sempre creduto in me.

INDICE

INTRODUZIONE

L'energia fotovoltaica è divenuta una delle fonti di energia rinnovabile più diffuse e promettenti a livello mondiale. In Italia, il settore fotovoltaico ha riscontrato una crescita significativa negli ultimi anni, grazie al sostegno di politiche energetiche favorevoli alla produzione di energie rinnovabili e all'aumento di impianti fotovoltaici su scala industriale e domestica. Basti pensare che l'energia fotovoltaica costituisce il 9.9%[1] della produzione di elettricità nel mix energetico italiano nel 2022 che, nonostante non sia una quota particolarmente elevata, risulta essere la terza fonte di produzione di energia elettrica in Italia dopo termoelettrico (70.1%) e idrico (10.7%). Di conseguenza un'analisi accurata della produzione di energia solare consente di comprendere il contributo effettivo di quest'ultima nel soddisfacimento della domanda energetica nazionale e nella riduzione della produzione di energia elettrica proveniente da fonti esauribili che causano emissioni di gas serra nell'atmosfera.

Tuttavia, la produzione di energia fotovoltaica è caratterizzata da forti fluttuazioni stagionali di tipo giornaliero e annuale causate da fattori atmosferici e in particolare dalla dipendenza dalla disponibilità di luce solare, a causa di ciò essa non può garantire una continuità di funzionamento che invece richiede la domanda di energia elettrica nazionale. Ciò rende fondamentale la modellazione e la previsione della produzione di energia fotovoltaica per garantire una gestione efficiente delle risorse energetiche e una programmazione ottimale dell'infrastruttura elettrica, al fine di prevenire situazioni di sovraccarico, mancanza di elettricità o poca tensione, e di ottimizzare l'utilizzo delle risorse disponibili.

Inizialmente verrà fatta una panoramica sul funzionamento del Mercato Elettrico italiano e delle sue suddivisioni in sottomercati, successivamente verranno effettuate delle analisi descrittive preliminari sui dati disponibili, in particolare sulla serie storica della produzione di energia fotovoltaica da impianti distribuiti nelle diverse zone italiane. In seguito, verranno sviluppati alcuni modelli statistici che permetteranno di stimare la produzione di energia fotovoltaica in base a condizioni atmosferiche, utilizzando come variabili esplicative l'irraggiamento solare, l'altezza del Sole, la temperatura dell'aria, la velocità del vento e la capacità di generazione degli impianti installati. Infine, verranno valutate e confrontate le performance predittive dei diversi modelli, al fine di stabilire quali siano i più appropriati da adottare nel caso di studio.

Il seguente lavoro di tesi si propone di contribuire alla comprensione delle dinamiche della produzione dell'energia fotovoltaica in Italia e di fornire strumenti e informazioni, che seppur semplici, possano essere utili per comprendere le caratteristiche e le dinamiche della produzione di energia fotovoltaica nel contesto italiano.

[1] dati di Terna S.p.A.

CAPITOLO PRIMO:
MERCATO ELETTRICO ITALIANO

1.1 Filiera elettrica

La filiera elettrica in Italia comprende tutte le fasi coinvolte nella produzione, trasmissione o dispacciamento, distribuzione e vendita dell'energia elettrica nel Paese. Questa filiera è regolata da diverse entità, tra cui il Governo italiano, l'Autorità di Regolazione per Energia Reti e Ambiente (ARERA), il GSE (Gestore dei Servizi Energetici) e i vari operatori del settore.

L'energia elettrica è uno dei diversi stati che può assumere l'energia, essa è presente in natura solo sottoforma di fulmini e in alcuni animali detti elettrofori, come l'anguilla, che generano elettricità chiamata bioelettricità. Di conseguenza le società di produzione di energia elettrica necessitano di risorse dette fonti di energia primarie (rinnovabili o esauribili) che attraverso processi fisico-chimici generino energia elettrica. La differenza sostanziale tra le due tipologie di fonte è che le fonti esauribili sono risorse naturali disponibili in quantità limitata sulla Terra e che possono esaurirsi nel tempo e il cui processo di produzione di energia genera scarti potenzialmente dannosi per l'ambiente, specialmente in termini di inquinamento atmosferico, un esempio di fonti esauribili sono i combustibili fossili utilizzati nelle centrali termoelettriche o l'uranio e il plutonio impiegati nelle centrali termonucleari a fissione. Le fonti di energia rinnovabili sono risorse naturali che possono essere rigenerate continuamente nel corso del tempo in quanto inesauribili e la cui produzione ha un impatto ambientale estremamente ridotto, alcuni esempi di fonti rinnovabili sono le radiazioni solari sfruttate dagli impianti fotovoltaici e il vento utilizzato nelle centrali eoliche.

Le centrali elettriche possono avere diversi gradi di efficienza che sono legati a molteplici fattori come i tempi di reazione, ovvero la durata temporale per l'attivazione o la disattivazione dell'impianto; i costi economici, come costi di avviamento della centrale, costi di generazione dell'elettricità e costi di manutenzione; fattori intermittenti, come il tempo atmosferico per le fonti rinnovabili. Per la molteplicità degli aspetti da considerare, è difficile stabilire quale sia la fonte di energia migliore, per questo i Paesi diversificano le fonti di produzione elettrica adottando mix energetici diversi a seconda della disponibilità economica e della volontà politica della nazione.

Successivamente alla fase di produzione, avviene la fase di distribuzione e dispacciamento nella quale si verifica il trasferimento dell'energia elettrica dagli impianti di produzione alle reti di distribuzione e misura locali, attraverso una rete ad alta e altissima tensione (220/380 kV), la quale viene gestita in regime di monopolio

naturale dalla società Terna S.p.A., che è il Gestore della Rete di Trasmissione Nazionale in Italia. In seguito, l'energia elettrica viene consegnata e distribuita sul territorio tramite reti regionali e locali da enti di distribuzione che operano in regime di monopolio locale, attraverso atti di concessione temporanei. Lungo la rete di trasmissione, vengono installate sottostazioni elettriche che fungono da punti di connessione e distribuzione dell'energia. Le sottostazioni trasformano l'energia elettrica ad alta tensione in livelli di tensione più bassi (230/400 V) per la fornitura agli utenti finali.

Infine, la filiera elettrica si conclude con la fase di vendita, dove l'energia elettrica prodotta viene commercializzata e venduta ai consumatori finali da società di vendita, dette fornitori, operanti in un mercato concorrenziale, che forniscono il servizio di luce e gas al consumatore finale attraverso la stipula di un contratto di fornitura.

1.2 Mercato Elettrico

Il Mercato Elettrico, noto anche come Borsa Elettrica, è un marketplace telematico che facilita la compravendita all'ingrosso di energia elettrica tra operatori come produttori, importatori, distributori, commercianti e consumatori sul territorio italiano. Il mercato elettrico in Italia è nato col decreto legislativo 79/1999 che ha sancito la liberalizzazione del mercato dell'energia elettrica in Italia dal monopolio dell'industria da parte di grandi compagnie elettriche che controllavano l'intero processo di produzione, dispacciamento e vendita dell'energia elettrica. Ciò è stato fatto allo scopo di: promuovere la concorrenza tra attività di produzione e di vendita nel settore energetico, ridurre i costi dell'energia, incoraggiare l'efficienza del dispacciamento e la trasparenza del mercato, favorire l'innovazione tecnologica e stimolare la produzione da fonti rinnovabili.

La Borsa Elettrica è gestita dal Gestore Mercati Elettrici (GME) ed è un mercato non obbligatorio, ovvero un mercato per cui l'energia elettrica può essere scambiata anche all'esterno del mercato, nel quale, tramite il meccanismo dell'incrocio tra domanda e offerta di energia elettrica, vengono determinati i prezzi e le quantità scambiate di energia per ogni ora del giorno per un totale di 24 fasce orarie. Quindi è grazie al Mercato Elettrico che si definiscono i programmi di immissione e prelievo di energia elettrica, che però devono essere vincolati dalla capacità tecniche della rete nazionale.

Il Mercato Elettrico in Italia si classifica in due tipi principali di mercati:

- Il Mercato a Pronti (o Spot), nel quale la consegna dell'energia elettrica viene effettuata nell'immediato futuro, ovvero per il giorno stesso o per il giorno successivo. Per questo motivo i prezzi dell'energia di questo mercato sono flessibili e determinati in tempo reale dal confronto fra domanda ed offerta;
- Il Mercato a Termine (o Futures), in cui gli accordi per l'acquisto o la vendita di energia elettrica si verificano per una consegna futura. In altre parole, i soggetti

stabiliscono prezzi e quantità dell'energia da scambiarsi per un periodo futuro specificato, consentendo così alle parti in causa di avere una forma di copertura dal rischio di fluttuazioni dei prezzi dell'energia elettrica. Infatti, quest'ultimi sono stabiliti attraverso negoziazioni bilaterali basate su contratti a termine standardizzati.

Per garantire un'efficiente trasmissione dell'energia elettrica su scala nazionale, il sistema elettrico in Italia è suddiviso in zone elettriche, che sono caratterizzate da una serie di reti di trasmissione e sottostazioni che consentono il flusso di energia elettrica tra loro e assicurano l'approvvigionamento elettrico in tutto il Paese. Dal primo gennaio 2021, l'Italia è suddivisa in sette zone di Mercato Energetico raffigurate geograficamente nella figura 1.1 che sono: Nord (NORD), Centro Nord (CNOR), Centro Sud (CSUD), Sud (SUD), Calabria (CALA), Sicilia (SICI) e Sardegna (SARD).

Figura 1.1: Zone del Mercato Elettrico italiano (fonte: Terna S.p.A.)

Queste zone, oltre a esistere per rispettare i limiti fisici di transito dell'energia con le corrispondenti zone confinanti prevenendo sovraccarichi o congestioni nella rete, hanno un ruolo chiave nel Mercato Elettrico in quanto ogni zona ha un suo prezzo dell'energia elettrica, noto come prezzo zonale, che viene determinato attraverso un meccanismo ad aste. In questo modo si tiene conto delle condizioni specifiche di produzione e consumo nell'area, inclusi i costi di trasporto dell'energia elettrica e la disponibilità delle risorse energetiche tra le diverse zone.

Concertandosi sul Mercato a Pronti, esso è articolato in:

- Mercato del Giorno Prima (MGP): luogo in cui i produttori, i fornitori e i consumatori finali autorizzati compravendono energia elettrica per il giorno successivo;
- Mercato Infragiornaliero (MI): permette a produttori, fornitori e clienti finali autorizzati di apportare modifiche ai programmi di immissione/prelievo determinati in precedenza sul MGP;
- Mercato del Servizio di Dispacciamento (MSD): nel quale Terna S.p.A., nel ruolo di gestore della rete, si approvvigiona dei cosiddetti *servizi ancillari* necessari alla gestione e al controllo del sistema elettrico.

1.2.1 Mercato del Giorno Prima (MGP)

Il Mercato del Giorno Prima è un mercato nel quale avvengono la maggior parte delle transazioni per la compravendita di energia elettrica all'ingrosso. In questo mercato, attraverso la negoziazione di blocchi orari di energia elettrica per il giorno successivo, vengono determinati i prezzi e le quantità scambiate di energia, nonché i programmi di immissione e prelievo per il giorno successivo.

Il MGP è basato secondo un meccanismo ad asta dove ogni operatore può fare più offerte di acquisto/vendita in cui viene specificata: la quantità scambiata, il prezzo di acquisto/vendita, la zona del mercato di appartenenza e la fascia oraria del giorno dopo, tutto ciò si svolge senza che gli operatori si preoccupano dei limiti fisici della rete.

La seduta di mercato si apre alle ore 8:00 del nono giorno precedente al giorno di consegna e termina alle ore 9:15 del giorno precedente di consegna. Successivamente, una volta conclusa l'asta, un algoritmo centrale a livello europeo noto come *Euphemia*, risolve il mercato, ovvero combina le offerte di acquisto e vendita in modo tale da massimizzare il surplus sociale del mercato, dato dalla somma del surplus del produttore e dell'acquirente, tutto questo considerando i vincoli fisici delle reti interzonali.

Le offerte di vendita accettate sono valorizzate al prezzo di equilibrio per la zona di appartenenza, mentre le offerte di acquisto accettate sono valorizzate al *Pezzo Unico Nazionale* (PUN), che è la media dei prezzi zonali ponderata per i consumi zonali al netto dei pompaggi e delle importazioni.

Al fine di determinare il prezzo di equilibrio si applica il meccanismo di accettazione chiamato *System Marginal Price*, per il quale si remunerano i produttori pagando loro un prezzo pari al prezzo dell'offerta più costosa tra quelle accettate per soddisfare la domanda di energia elettrica. In sostanza per ogni zona e per ogni ora del giorno si ordinano per prezzo decrescente tutte le offerte di acquisto, formando così la curva di domanda aggregata e si ordinano per prezzo crescente tutte le offerte di vendita, formando così la curva di offerta aggregata. Dall'intersezione della curva di domanda aggregata e di offerta aggregata si determina il punto di equilibrio del mercato, ovvero la quantità totale scambiata sul mercato e il relativo prezzo di equilibrio, per cui vengono accettate tutte le offerte di vendita con prezzo non superiore al prezzo di equilibrio e tutte le offerte di acquisto con prezzo non inferiore al prezzo di equilibrio.

Se i flussi della rete definiti dai programmi non infrangono nessun limite di transito, il prezzo di equilibrio è unico in tutte le zone. Altrimenti se almeno un limite risulta violato, l'algoritmo "divide" il mercato in due zone: una in esportazione che include tutte le zone a monte del vincolo ed una in importazione che include tutte le zone a valle del vincolo. Per ognuna delle due nuove zone si ripete il meccanismo di accettazione delle offerte sopra citato, al fine di definire per ogni nuova zona il relativo prezzo di equilibrio, detto *prezzo zonale di equilibrio*. Il processo di suddivisione zonale appena descritto prende il nome di "*market splitting*" ed esso si reitera fino ad ottenere un esito che soddisfi i vincoli fisici della rete.

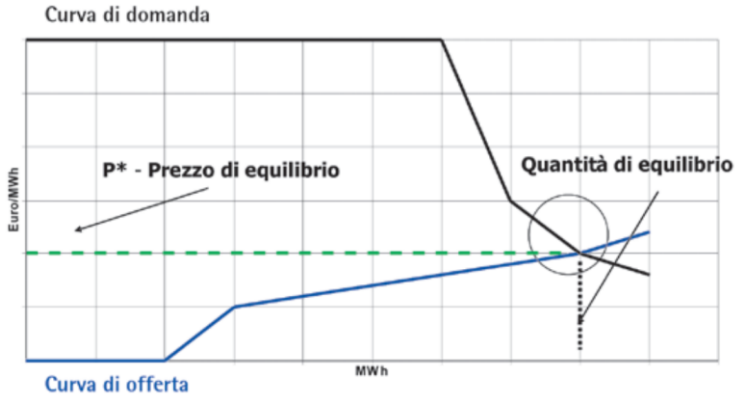


Figura 1.2: Rappresentazione grafica del MGP (fonte: Vademecum della Borsa Elettrica)

Quanto descritto in precedenza è rappresentato dalla figura 1.2, nella quale è possibile osservare che, mentre la curva di domanda aggregata è anelastica in quanto l'elettricità è un bene di prima necessità, la curva di offerta aggregata ha un andamento poco costante, infatti nella parte iniziale delle curva essa è piatta e costantemente nulla, questo perché i produttori di energia elettrica rinnovabile, come il fotovoltaico, per garantirsi un posto sul mercato, fanno offerte di vendita a prezzo pari a zero in quanto hanno costi marginali di generazione decisamente contenuti. Inoltre sanno che il prezzo dell'energia verrà fatto alzare dai produttori di energia proveniente da fonti esauribili, in particolare dalle centrali termoelettriche con impianti a cicli combinati le quali, essendo molto efficienti da un punto di vista tecnico e produttivo, mantengono i costi di produzione relativamente contenuti permettendo ai produttori di proporre un prezzo dell'energia più alto rispetto a quelli di impianti rinnovabili, ma comunque basso abbastanza per rimanere competitivi sul mercato.

In questo modo non stupisce che il prezzo del gas naturale sia uno dei principali fattori che determinano attualmente il prezzo di equilibrio dell'energia elettrica in Italia[2].

Il meccanismo appena descritto è stato creato allo scopo di favorire i produttori di energia elettrica proveniente da fonti rinnovabili poiché garantisce la vendita di energia rinnovabile sul mercato remunerandola allo stesso prezzo dell'energia non rinnovabile, avendo però dei costi di produzione considerevolmente inferiori rispetto a quest'ultima.

1.2.2 Mercato Infragiornaliero (MI)

Il Mercato Infragiornaliero nasce per permettere agli operatori di aggiornare le loro offerte di vendita e di acquisto e le loro posizioni commerciali con una frequenza simile a quella di una negoziazione continua rispetto alle variazioni delle informazioni circa lo stato degli impianti produttivi e le necessità di consumo. La negoziazione continua è una

[2] Behnam Zakeri , Iain Staffell , Paul E. Dodds, Michael Grubb, Paul Ekins, Jaakko Jääskeläinen, Samuel Cross, Kristo Helin, Giorgio Castagneto Gissey (2022). "Role of Natural Gas in Electricity Prices in Europe".

modalità di contrattazione che consiste nell'abbinamento automatico delle proposte di acquisto e di vendita, con l'opportunità di inserimento di nuove offerte in modo continuo durante le sessioni di contrattazione.

Le sessioni di asta nel Mercato Infragiornaliero sono basate su regole di formazione del prezzo simili a quelle del MGP, con la differenza che nel MI non viene calcolato il PUN e tutti gli acquisti e le vendite vengono valorizzate al prezzo zonale. Alla chiusura di ciascuna sessione del MI, il GME comunica a Terna i programmi aggiornati di immissione e prelievo, allo scopo di valutare la fattibilità di transito dell'energia elettrica nella rete ed evitare congestioni.

Il meccanismo che regola questo mercato favorisce la rapidità di esecuzione e modifica dei programmi di fornitura dell'energia elettrica dei produttori fino a un'ora prima dell'inizio della fornitura del servizio; tuttavia, proprio per questo motivo i prezzi dell'energia mutano costantemente, rendendo il mercato meno efficiente.

Inoltre, per come è strutturato il mercato dell'energia elettrica, risulta cruciale la capacità di fare previsioni attraverso modelli statistico-matematici per ottimizzare la produzione delle unità produttive al fine di limitare gli *oneri di sbilanciamento*, ovvero penali di tipo economico imputate ad unità di produzione e/o consumo della rete elettrica che sforano il programma di immissione/prelievo precedentemente dichiarato. Ciò è un punto particolarmente svantaggioso soprattutto per gli impianti di energia rinnovabile, come gli impianti fotovoltaici, perché caratterizzati da una produzione fortemente variabile, in quanto dipendente da fattori atmosferici, e non regolabile a differenza degli impianti termoelettrici.

1.2.3 Mercato del Servizio di Dispacciamento (MSD)

Il Mercato del Servizio di Dispacciamento è il mercato attraverso cui Terna, nel ruolo di gestore della rete, si approvvigiona, dei cosiddetti servizi ancillari, cioè risorse necessarie alla gestione e al controllo del sistema, ovvero al bilanciamento in tempo reale tra produzione e consumi in modo tale che la frequenza di transito della rete sia fissa a 50 Hz, all'istituzione di una riserva di energia e alla risoluzione di congestioni intra-zonali. Terna si approvvigiona di questi servizi stipulando contratti di acquisto e vendita operando come controparte centrale delle negoziazioni.

Le offerte di incremento/decremento della produzione possono essere presentate a Terna solo da soggetti definiti utenti abilitati, ovvero da unità di produzione i cui impianti di generazione soddisfano determinati requisiti tecnici, mentre la partecipazione al MSD è obbligatoria per gli impianti che hanno una capacità di generazione superiore ai 10 Mw/h. Infine, tutte le offerte accettate da Terna sono remunerate col sistema *pay-as-bid*, ovvero allo stesso prezzo che gli utenti abilitati offrono, proprio per questo i prezzi dell'energia in questo mercato sono generalmente più elevati rispetto agli altri mercati della Borsa Elettrica.

CAPITOLO SECONDO:
ANALISI PRELIMINARI DEI DATI

2.1 Presentazione dei dati

In questo capitolo illustrerò in che modo sono stati raccolti i dati utilizzati ai fini dell'analisi e si procederà attraverso analisi descrittive preliminari e alla stima della componente tendenziale per comprendere le principali caratteristiche della serie della produzione di energia fotovoltaica, allo scopo di ottenere quante più informazioni possibili in fase di modellazione per specificare modelli statistici più adatti al contesto di riferimento.

I dati utilizzati in questo studio provengono da due fonti autorevoli, ovvero:

- Terna S.p.A., è il Gestore della Rete di Trasmissione Nazionale in Italia. È l'ente proprietario e responsabile della gestione, del controllo e dello sviluppo della rete di trasmissione dell'energia elettrica a livello nazionale. In particolare, si occupa del servizio di dispacciamento dell'energia elettrica ad alta e altissima tensione dal punto di produzione al centro di distribuzione locale ed occupa un ruolo di monopolio naturale, per questo opera sotto la supervisione dell'Autorità di Regolazione per Energia Reti e Ambiente (ARERA).
- Joint Research Centre (JRC), è il centro di ricerca scientifica della Commissione Europea responsabile di fornire supporto scientifico e tecnico indipendente per le politiche dell'Unione Europea (UE).

I dati di Terna sono le serie storiche della produzione oraria di energia elettrica fotovoltaica espressa in Megawattora (MWh) divise per zone elettriche dal 1° gennaio 2017 alle ore 00:00 fino al 31 dicembre 2020 alle ore 23:00, per un totale di 4 anni.

Inoltre, Terna fornisce i dati sulla capacità annuale di generazione degli impianti fotovoltaici dal 2016 al 2021 per zona misurata in base alla potenza efficiente lorda, ovvero la potenza massima teorica di generazione dell'impianto senza considerare le perdite o gli sprechi energetici. Dato che i dati sulla capacità di generazione degli impianti sono annuali, si è deciso di ipotizzare un andamento lineare costante infra-annuale della capacità di generazione, inserendo nei dati ottenuti nella variabile di capacità installata (Cap).

Il *Photovoltaic Geographical Information System* (PVGIS) è uno strumento sviluppato dalla Commissione Europea, specificatamente dal Joint Research Centre (JRC), per la

valutazione e la pianificazione del potenziale fotovoltaico in Europa. Attraverso il PVGIS si sono ottenute le serie storiche orarie dal 1° gennaio 2017 al 31 dicembre 2020 riguardanti le seguenti variabili: la radiazione solare (G.i.) misurata in watt per metro quadrato (W/m^2), l'altezza del Sole in gradi (H_sun), la temperatura dell'aria a 2 metri di altezza (T2m) in gradi Celsius e la velocità del vento a 10 metri (WS10m) misurata in metri al secondo (m/s).

Dato il numero elevato di zone elettriche, si è deciso di concentrarsi sulle zone nord e sud d'Italia in quanto sono le aree maggiormente rilevanti in termini di produzione di energia fotovoltaica, in quanto hanno costituito nel 2020 rispettivamente il 34.6% e il 22.3% dell'intera produzione di energia solare del Paese (figura 2.1) e sono tra le zone con più capacità efficiente lorda di impianti fotovoltaici (figura 2.2). Inoltre, ciò permette di osservare le possibili differenze e somiglianze produttive tra due aree d'Italia geograficamente e meteorologicamente molto diverse.

Dalle serie storiche della produzione di energia solare delle zone nord e sud (figura 2.3) si colgono delle caratteristiche particolari come una forte componente stagionale, in particolare una stagionalità annuale ed una giornaliera, che rende molto variabile la serie e la presenza di un trend non particolarmente marcata. Com'era prevedibile, l'andamento della serie della produzione di energia solare sembrerebbe simile tra le zone, ma la zona nord produce quasi il doppio di energia fotovoltaica rispetto alla zona sud, questo è soprattutto dovuto alla capacità installata più che doppia tra le due aree geografiche.

Per di più il fatto che le serie di produzione di energia fotovoltaica seguono il ciclo delle stagioni climatiche, delimitate nella figura 1.3 da linee rosse verticali, indica come questo tipo di fonte energetica, come la maggior parte delle fonti rinnovabili, sia fortemente legata alle variabili atmosferiche che utilizzeremo nella nostra analisi.

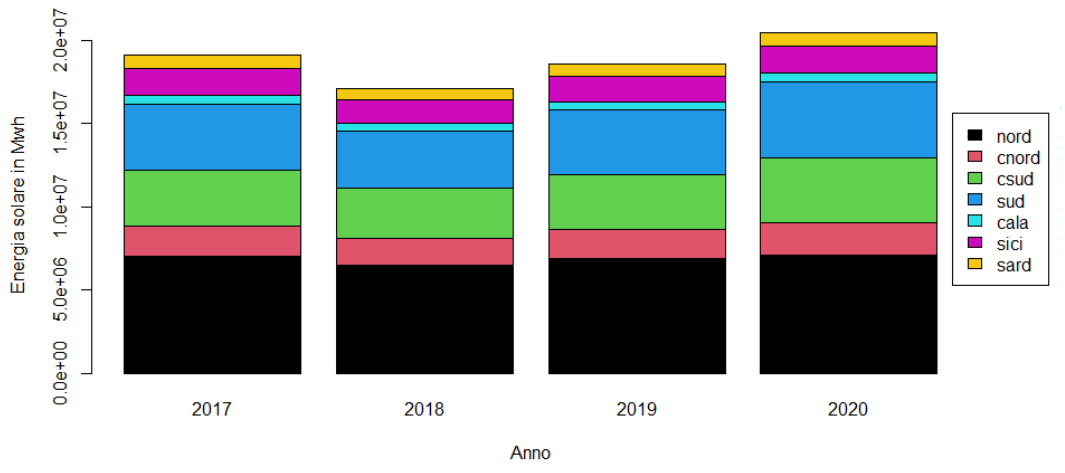


Figura 2.1: Produzione totale di energia solare per anno e zona in Italia (Fonte: Terna S.p.A.)

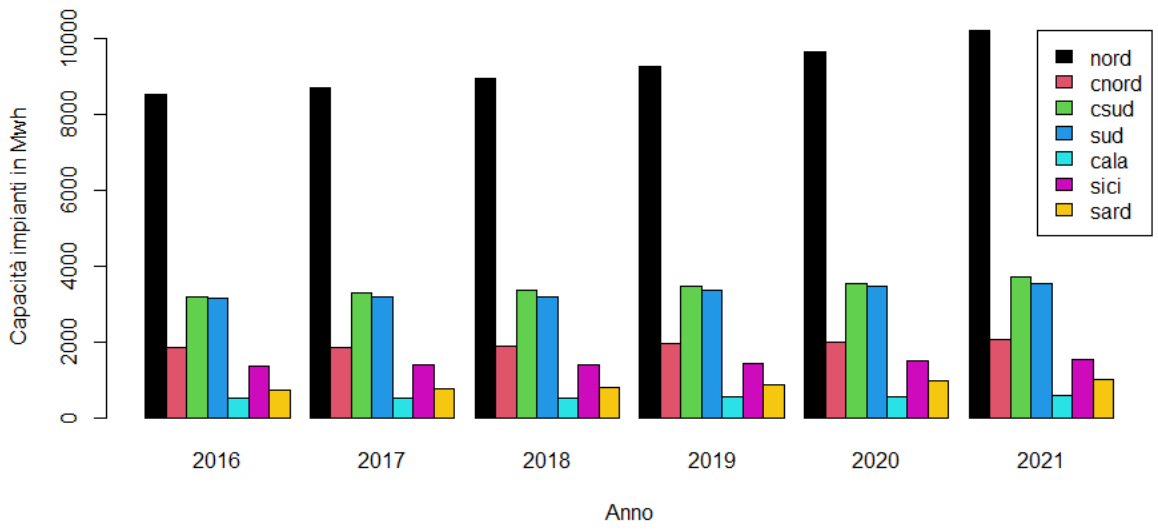


Figura 2.2: Capacità efficiente lorda di generazione degli impianti per anno e zona (Fonte: Terna S.p.A.)

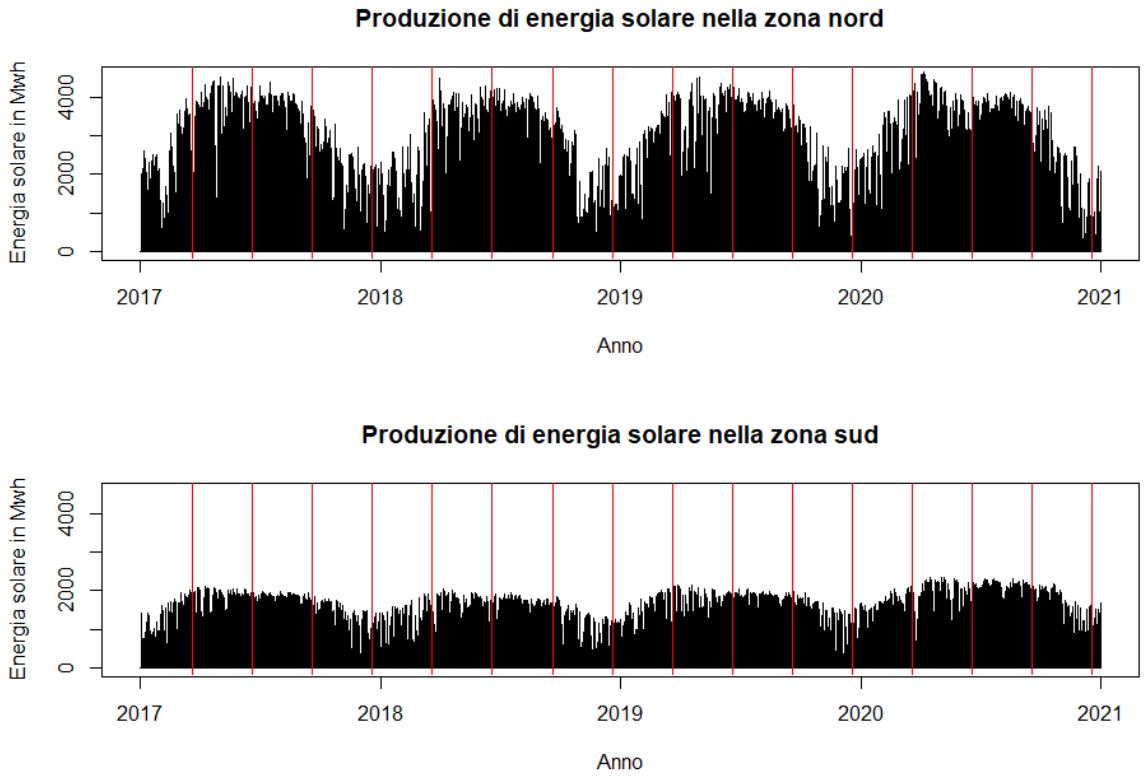


Figura 2.3: Serie storiche della produzione di energia solare nelle zone nord e sud separate per stagioni dell'anno (Fonte: Terna S.p.A.)

2.2 Analisi descrittive

In questa sezione si osserveranno le caratteristiche principali della produzione di energia solare e le differenze zonali tra la zona nord e la zona sud, due aree geograficamente molto diverse. Per comprendere la distribuzione complessiva del fenomeno si è utilizzato il grafico box plot della figura 2.4, da cui si osserva che le distribuzioni delle serie zonali hanno una forte asimmetria positiva[3], in particolare la mediana è prossima allo zero, questo succede perché poco meno della metà delle osservazioni delle distribuzioni sono nulle (42,34% nella zona nord e 45,88% nella zona sud), in quanto rilevate nelle ore in cui è assente luce solare. Inoltre, è possibile vedere che la zona nord produce in media più energia fotovoltaica rispetto al sud, ma presenta una variabilità maggiore, come testimoniato dai numerosi outlier e ciò può essere dovuto alla maggiore capacità installata (più che doppia rispetto al sud) e ad una condizione climatica della zona maggiormente mutevole.

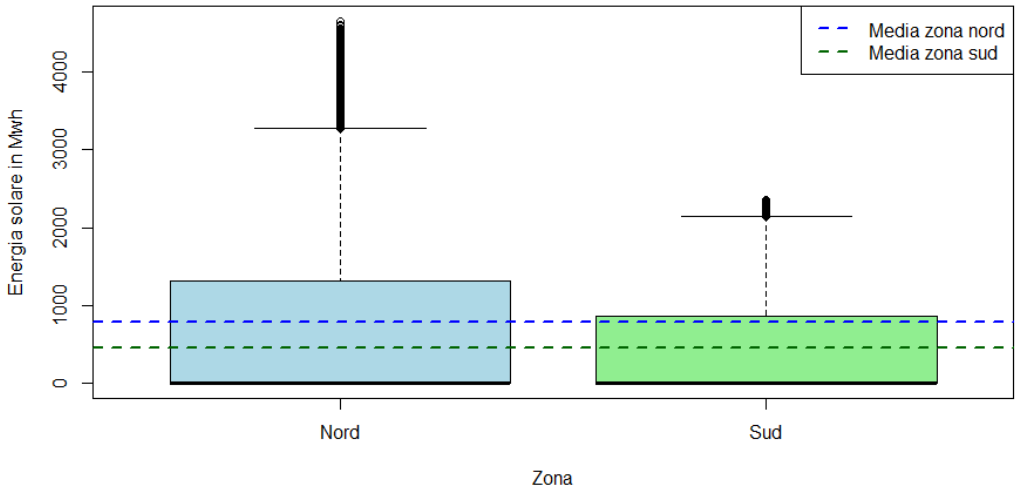


Figura 2.4: Box plot della produzione di energia solare per zona

Un modo efficacie per esaminare la presenza di periodicità nella serie della produzione di energia solare per le due zone prese in esame, è quello di confrontare tra loro alcune variabili temporali, come l'ora del giorno, il giorno della settimana, il mese dell'anno e gli anni, come riportato nelle figure seguenti.

[3] Si verifica quando la media aritmetica è superiore alla mediana della distribuzione

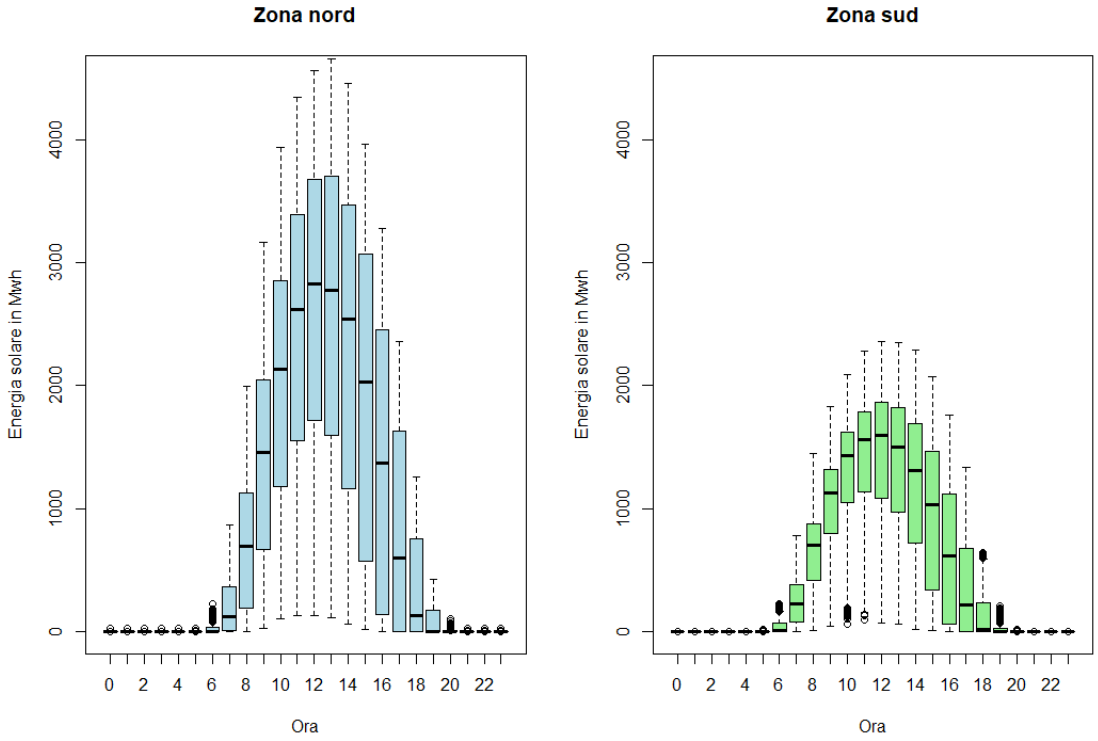


Figura 2.5: Box plot della produzione oraria di energia solare per zona

La figura 2.5 propone il box plot della produzione di energia solare divisa per fasce orarie suggerendo, come prevedibile, una marcata stagionalità giornaliera delle serie storiche zonali, che al netto della grandezza di scala, si comportano in modo simile in termini di distribuzione giornaliera.

Le distribuzioni che vanno dalle 20:00 alle 4:00 possono essere considerate costanti in zero al netto di diversi outlier concentrati sulle fasce orarie meno luminose, la cui presenza è giustificata da eventi atmosferici, e certe volte eventi astronomici anomali, come per esempio eclissi lunari. Infine, si nota una variabilità maggiore nelle ore pomeridiane rispetto a quelle mattutine e serali, data dalla maggiore presenza di radiazione emessa dal Sole in quelle fasce orarie e che può cambiare a seconda del tempo atmosferico.

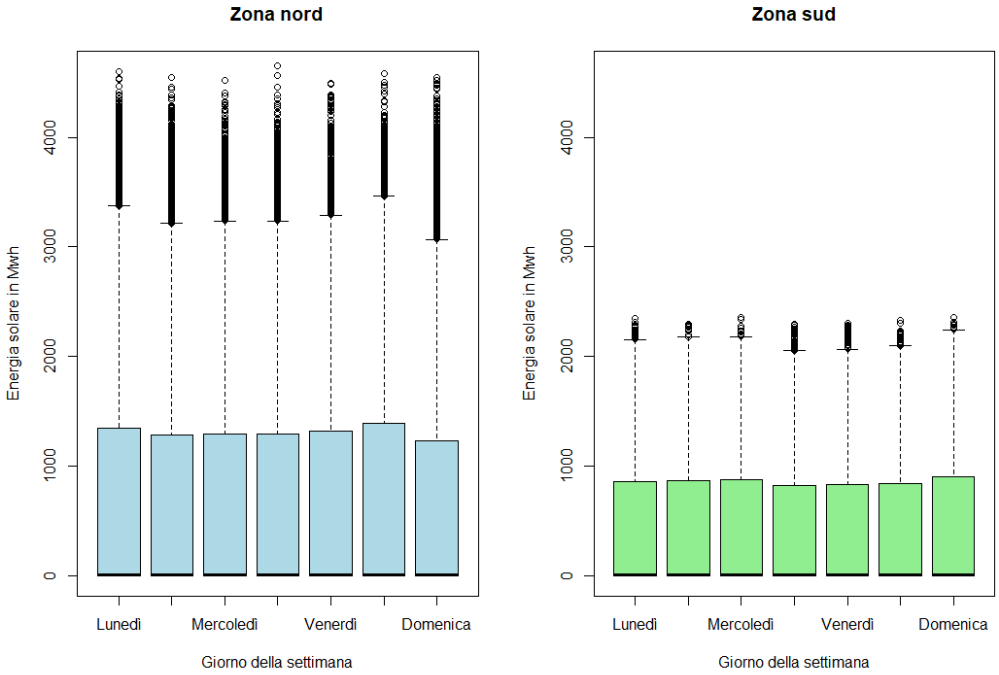


Figura 2.6: Box plot della produzione giornaliera di energia solare per zona

Il grafico della figura 2.6 suggerisce che la produzione di energia solare non è condizionata dal giorno della settimana, ciò è confermato dal test ANOVA che verifica l'ipotesi nulla per cui le medie di tre o più gruppi (in questo caso 7) sono significativamente uguali tra loro, attraverso la seguente statistica:

$$F = \frac{SSB/(k-1)}{SSW/(n-1)} \sim F(k-1, n-1)$$

Dove SSW è la somma dei quadrati all'interno dei gruppi, ovvero la devianza within, SSB è la somma dei quadrati tra i gruppi, ovvero la devianza between e la statistica F converge sotto ipotesi nulla ad una distribuzione F con $k-1$ gradi di libertà per il numeratore e $n-1$ gradi di libertà per il denominatore, dove k e n sono rispettivamente il numero di gruppi e il numero totale di osservazioni.

Applicando il test ai dati si ottiene un p-value pari a 0.6728 per la zona nord e 0.6503 per la zona sud; quindi, è possibile affermare che la produzione di energia fotovoltaica non dipende in modo statisticamente significativo dal giorno della settimana.

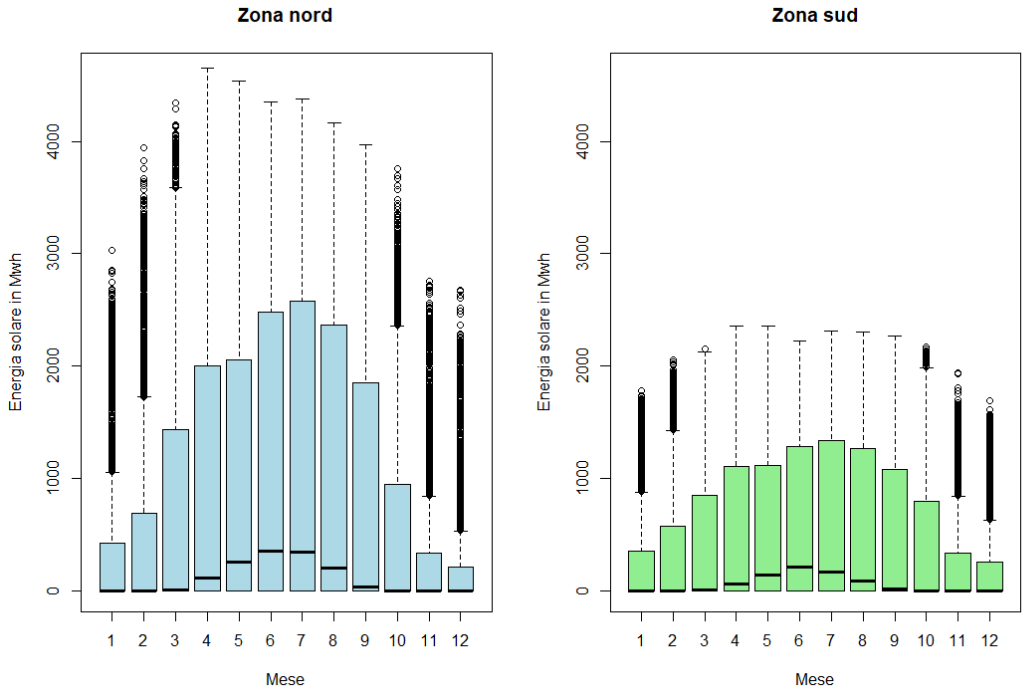


Figura 2.7: Box plot della produzione mensile di energia solare per zona

La figura 2.7 rappresenta le distribuzioni mensili di produzione di fotovoltaico per zona ed è possibile notare la presenzia di una spiccata stagionalità annuale di entrambe le serie, confermata dal test ANOVA applicato sui gruppi formati dai mesi per il quale si ottengono dei p-value minori di 0.0001, ovvero prossimi allo zero; quindi, il mese dell'anno condiziona in modo fortemente significativo la produzione di fotovoltaico.

In particolare, com'era prevedibile, i mesi in cui solitamente è presente maggiore radiazione solare, tipicamente in Italia quelli estivi, sono i mesi più produttivi e, viceversa, i mesi invernali sono mesi poco produttivi a causa delle condizioni climatiche avverse, come copertura nuvolosa e precipitazioni e del minor numero di ore esposizione alla luce solare.

Per di più è possibile osservare che l'andamento mensile, al netto della scala di grandezza, è diverso tra le due zone considerate; infatti, le differenze relative inframensili sono più marcate nella zona nord, mentre sono meno evidenti nella zona sud. Ciò è giustificato dalle differenze di capacità installata tra le due aree e può indicare che l'effetto della capacità installata sia maggiormente rilevante nei mesi più soleggiati.

Infine, si osserva una concentrazione di outlier nei mesi tra ottobre e marzo la cui presenta è legittimata da una forte asimmetria delle distribuzioni mensili, data dal valore quasi nullo della mediana.

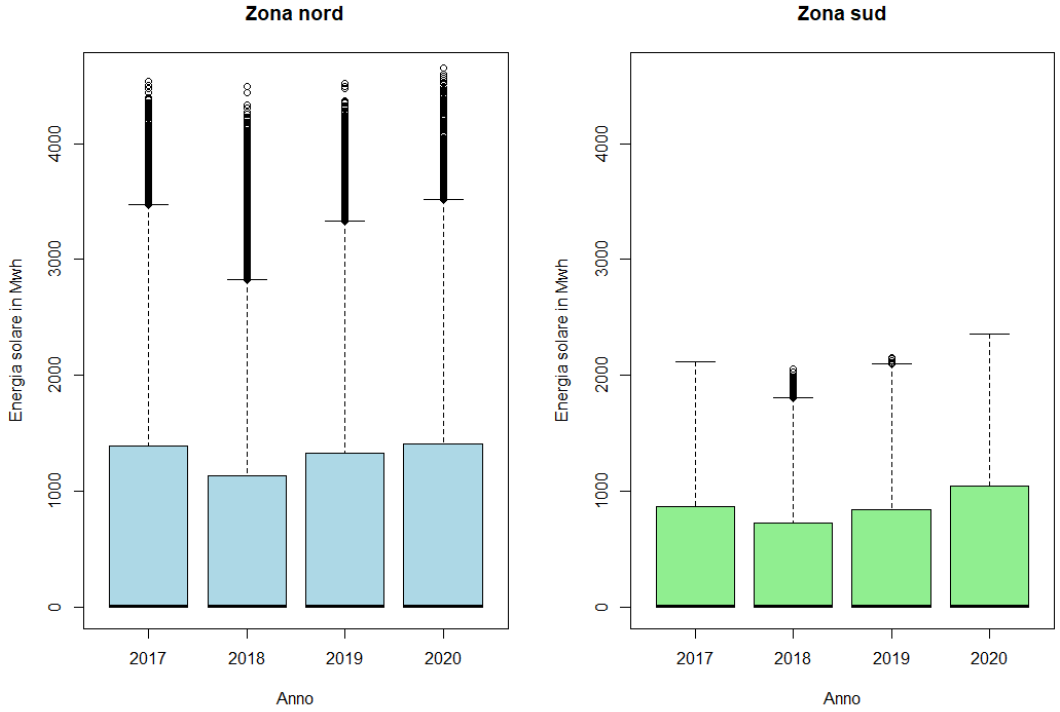


Figura 2.8: Box plot annuale della produzione di energia solare per zona

Analizzando il box plot annuale della produzione di energia fotovoltaica si nota, in entrambe le zone, una crescita della produzione di fotovoltaico interrotta nell'anno 2018 con una riduzione del -7.4% per il nord e -12.3% per il sud rispetto all'anno precedente, ciò è stato causato da un calo della radiazione solare media annuale di -5.62% nella zona nord e -7.48% nella zona sud rispetto al 2017.

Inoltre, con il test ANOVA si è verificata la significatività statistica della variazione tra gli anni per le due zone e mentre per la zona nord questa non è risultata particolarmente significativa con un p-value di 0.2619, per la zona sud la variazione è risultata decisamente significativa con un p-value prossimo a zero (minore di 0.0001). Questi risultati rivelano che la componente di trend sarà più rilevante e definita nella zona sud e più marginale e meno variabile nella zona nord.

L'analisi esplorativa preliminare prosegue concentrandosi sulle serie storiche per fasce orarie della produzione di energia fotovoltaica nelle zone del Mercato Elettrico nord e sud. Nella figura 2.9 sono riportate le serie delle fasce orarie delle 5:00, 8:00, 12:00 e 17:00 in quanto rappresentative delle fasce orarie di tutto l'arco di tempo giornaliero.

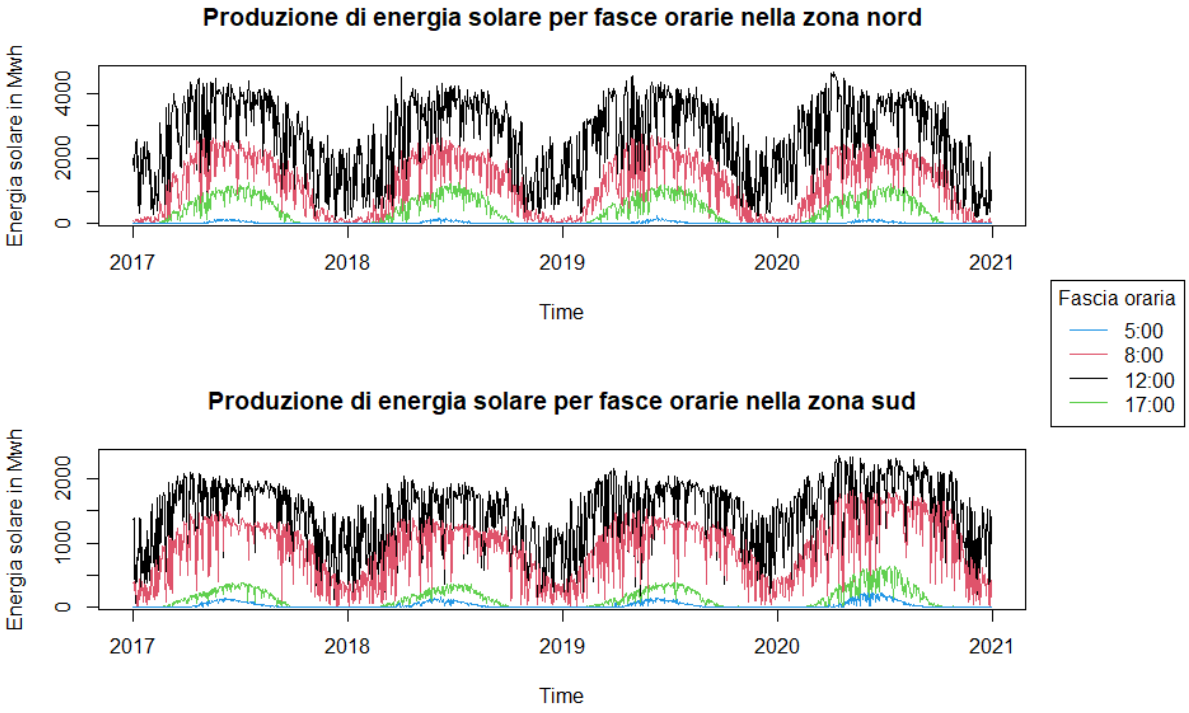


Figura 2.9: Serie storiche della produzione di energia solare per alcune fasce orarie nelle zone nord e sud

Le serie storiche orarie della figura sopracitata mostrano una rilevante componente stagionale annuale e una maggiore variabilità nelle fasce orarie intorno a mezzogiorno, in cui la produzione di energia solare è maggiore. Infine, la differenza più evidente tra le serie delle zone, oltre alla grandezza della scala, va ricercata nella rapidità con la quale la zona sud riesce a raggiungere il livello di massima produzione delle ore 12:00, mentre si osserva un aumento ed una diminuzione della produzione più graduale nella zona nord. Ciò è in gran parte dovuto alle differenze metereologiche delle due aree in esame.

Infine, l'analisi descrittiva si conclude con la matrice di correlazione dei dati, presentata nella figura 2.10, che è uno strumento di analisi statistica utilizzato per studiare la relazione tra le variabili in un insieme di dati. Essa fornisce una rappresentazione tabellare delle correlazioni tra coppie di variabili, consentendo di comprendere meglio come le variabili si condizionano reciprocamente e se esistono tendenze o pattern significativi nei dati.

La matrice di correlazione è generalmente rappresentata come una matrice quadrata, in cui ogni cella contiene il coefficiente di correlazione tra due variabili, che rappresenta la forza e la direzione della relazione lineare tra le variabili e può variare da -1 a 1.

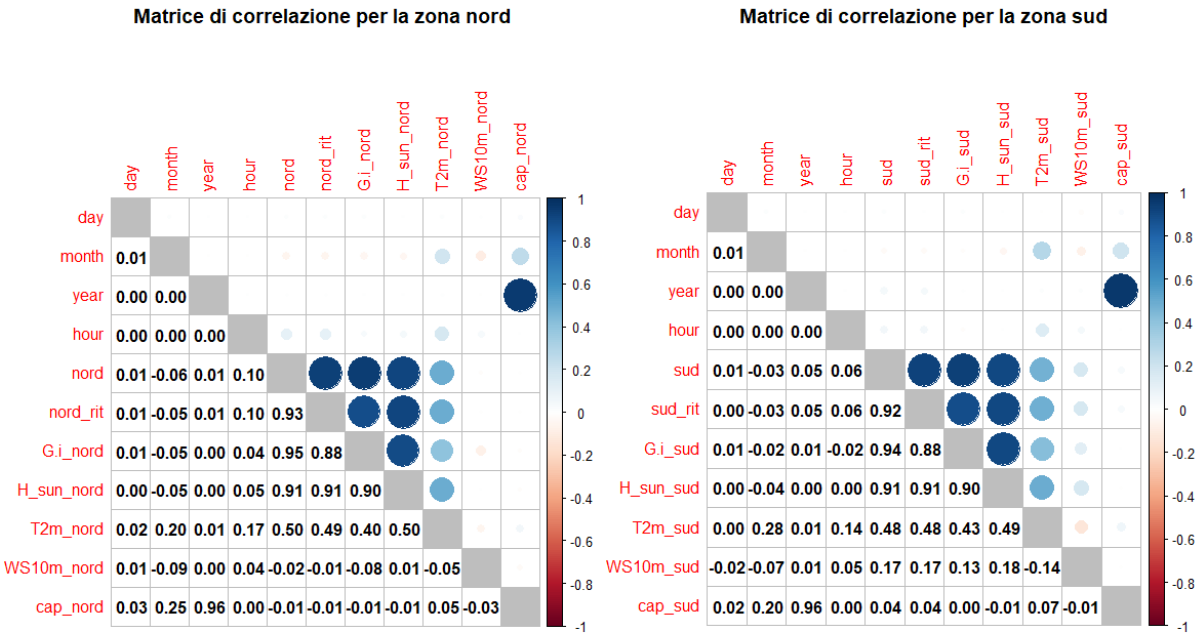


Figura 2.10: Matrici di correlazione dei dati divise per zona

Dai risultati emersi dalle matrici di correlazioni, non si osservano discrepanze significative tra le due zone di studio. Inoltre, gli indici di correlazione più elevati sono da ricercare tra le combinazioni delle variabili della produzione di energia fotovoltaica (nord / sud), delle radiazioni solari (G.i_nord / G.i_sud) e della temperatura dell'aria (T2m_nord / T2m_sud).

Per di più si osserva una forte correlazione tra la produzione di energia fotovoltaica (nord / sud) e la stessa variabile ritardata del giorno precedente (nord_rit / sud_rit), questo suggerisce la possibilità di formare un modello statistico di regressione che includa la variabile ritardata, che verrà specificato nel capitolo successivo.

La grande correlazione tra la variabile dell'anno (year) e quella della capacità di generazione degli impianti (cap_nord / cap_sud) è dovuta per come si è costruita la variabile della capacità, procedura spiegata nel paragrafo 2.1.

Infine, le variabili temporali (day, month, year, hour) non sono correlate con la variabile della produzione di fotovoltaico in quanto la matrice di correlazione tiene conto solo della correlazione lineare tra le variabili; infatti, dalle analisi esplorative precedenti abbiamo osservato che l'effetto del tempo essendo stagionale, si annulla nell'arco temporale del giorno e dell'anno.

2.3 Componente tendenziale

In questa sezione l'obiettivo dell'analisi è quello di osservare la componente tendenziale di fondo delle serie storiche zonali della produzione di energia solare, al netto di comportamenti stagionali; quindi, al fine di confrontare zone diverse si è ritenuto necessario considerare l'andamento della capacità produttiva installata degli impianti.

Si è proceduto a standardizzare i dati, trasformando le serie della produzione di fotovoltaico da capacità installata dell'anno corrente a capacità installata costante, scegliendo come anno di riferimento l'ultimo anno di cui si dispongono i dati, ovvero il 2021; utilizzando la seguente formula:

$$Produzione_{cap\_costante} = \frac{Produzione_{cap\_corrente}}{Cap_{corrente}/Cap_{2021}}$$

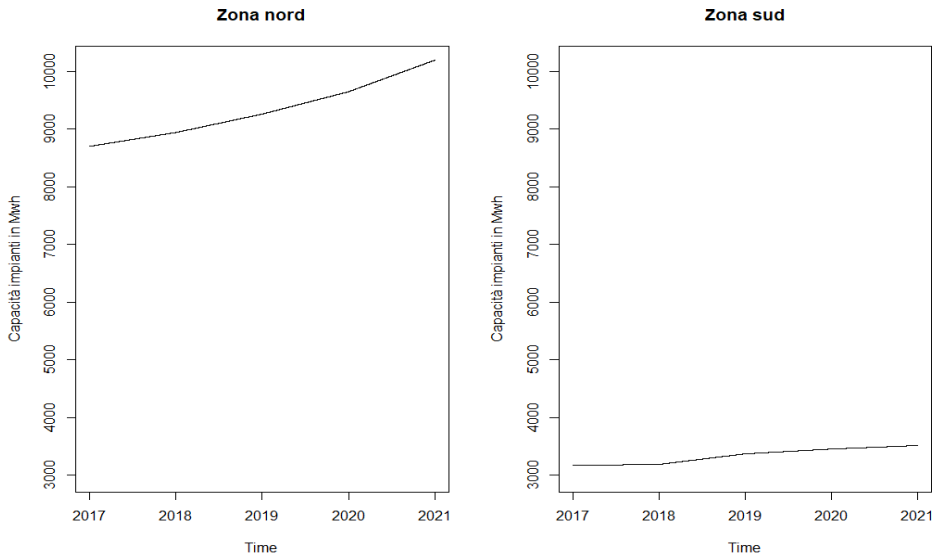


Figura 2.11: Serie della capacità di generazione degli impianti fotovoltaici per zone, secondo l'ipotesi di andamento lineare costante infra-annuale

Dalla figura 2.11 si nota che la capacità installata di impianti fotovoltaici della zona nord non solo è quasi tre volte quella della zona sud, ma ha un andamento crescente di anno in anno superiore alla zona meridionale, ciò spiega la grande differenza di produzione tra le due aree del paese.

Al fine di valutare la variazione del trend della serie occorre destagionalizzare la serie a capacità installata costante eliminando la componente stagionale. Per fare ciò si è deciso di adottare un approccio non parametrico in quanto meno rigido e in grado di cogliere andamenti irregolari del fenomeno rispetto a stime del trend parametriche.

Inoltre, si è trattata un'unica stagionalità che potesse raggruppare la componente periodica annuale (compresa di anno bisestile) e quella giornaliera. Quindi alle serie storiche è stata applicata una media mobile semplice del tipo:

$$M = \left\{[365{,}25 \times 24], \frac{1}{365{,}25 \times 24}\right\}$$

Questa media mobile, oltre ad essere facilmente interpretabile, è simmetrica, ma non gode della proprietà di invarianza in quanto non centrata. Inoltre, applicandola alle serie storiche si evidenzia un'elevata riduzione della variabilità di entrambe le serie, calcolata come il rapporto di riduzione della variabilità residua:

$$\frac{\sigma_{\varepsilon}^{*2}}{\sigma_{\varepsilon}^{2}} = \sum \theta_i^2$$

Dove $\sigma_{\varepsilon}^{*2}$ e σ_{ε}^{2} sono le varianze rispettivamente dell'errore della serie storica con applicata la media mobile e dell'errore della serie storica originale e questo rapporto è uguale alla somma dei quadrati dei pesi della media mobile, che nel caso in esame è di circa 0.0003 per le serie della zona nord e 0.0021 per quella sud, ciò significa che la media mobile riduce la variabilità della serie originale (in quanto minore di 1) ed essendo questo rapporto molto lontano a 1 indica che la media mobile spiega gran parte della variabilità e una parte minore rimane residua. Da ciò si deduce l'importanza della componente stagionale a cui viene imputata la maggior parte della variabilità della serie della produzione di energia fotovoltaica.

Tuttavia, per il fatto che l'applicazione della media mobile causa la perdita di un totale di osservazioni pari alla numerosità dei ritardi di una stagione della serie, nel capitolo successivo si deciderà di adottare un approccio di tipo parametrico per stimare il trend.

Dalle figure di seguito è possibile constatare l'andamento stagionale annuale della serie a capacità costante e come la componente giornaliera sia più variabile nei mesi estivi, rispetto a quelli invernali. Inoltre, si osserva per entrambe le zone una riduzione del trend dalla seconda metà del 2017 alla seconda metà dell'anno successivo ed una successiva ripresa della produzione, che, mentre per la zona nord è stata repentina e successivamente decrescente, nella zona sud la crescita è stata più costante.

Questa differenza di tendenza tra le due aree è dovuta in larga parte alle condizioni climatiche delle due aree, infatti l'Italia settentrionale ha un clima continentale più freddo, caratterizzato da inverni freddi e umidi nel quale non è raro il formarsi di distese di nebbia, mentre l'Italia meridionale ha un clima mediterraneo, ideale per la produzione di energia solare in quanto soleggiato, caratterizzato da inverni più miti, estati calde e secche e correnti di aria marina del Mediterraneo, che tendono a stabilizzare il clima e a portare condizioni di atmosferiche migliori. Tutto ciò rende i pannelli fotovoltaici della zona nord meno produttivi di quelli della zona sud al netto della capacità installata e delle caratteristiche tecniche di efficienza dei pannelli.

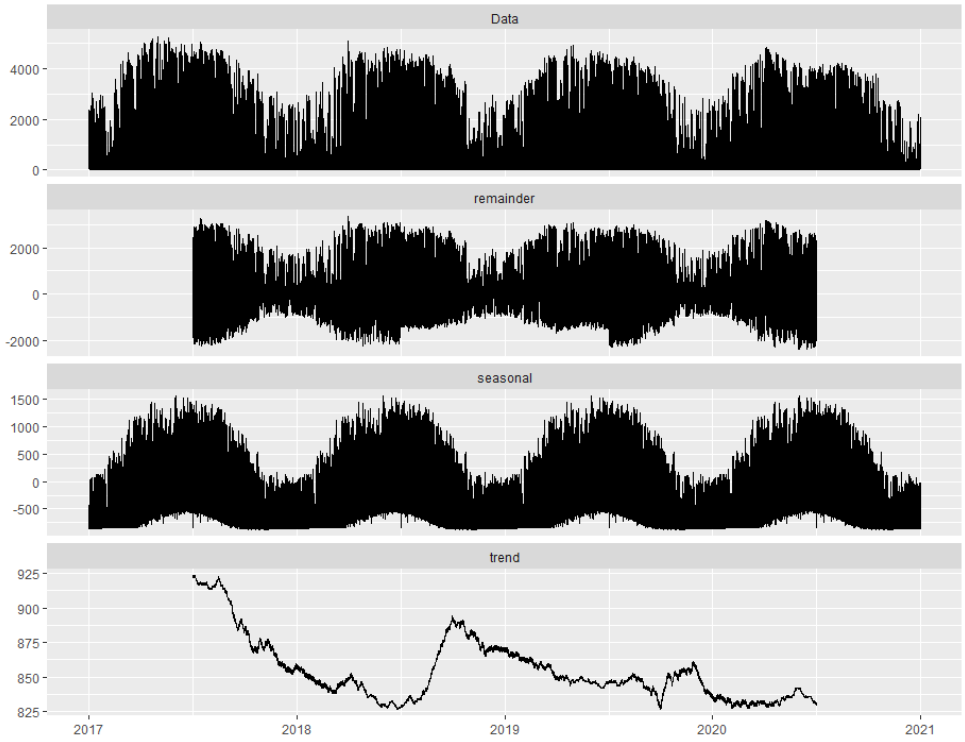


Figura 2.12: Decomposizione della serie della produzione di energia solare a capacità costante (anno base 2021) nella zona nord

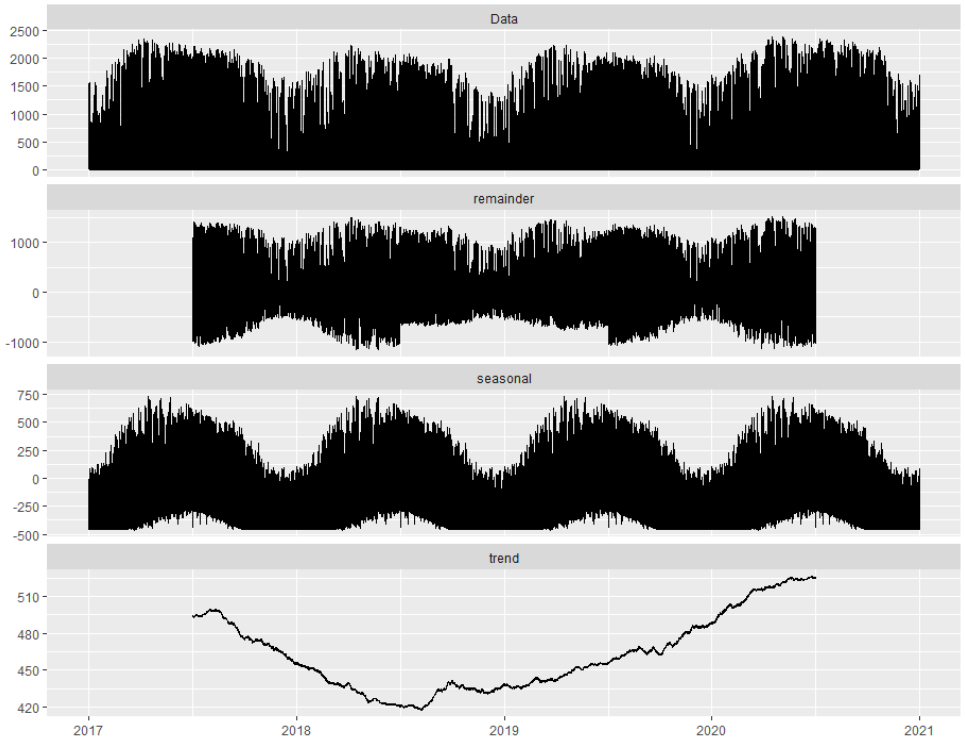


Figura 2.13: Decomposizione della serie della produzione di energia solare a capacità costante (anno base 2021) nella zona sud

Infine, per esaminare l'andamento della componente stagionale annuale si è applicata una media mobile semplice simmetrica a 24×30 termini alle serie zonali della generazione di energia fotovoltaica, in modo tale da escludere dalle serie la componente giornaliera e le variazioni inframensili. I risultati grafici sono riportati nella figura 2.14, in cui è possibile vedere la componente ciclica annuale smussata che è maggiore nella zona nord, ma decrescente nel tempo e nella zona sud viceversa. Inoltre, si osserva che nel periodo invernale la differenza tra le due aree del Paese risulta limitata nonostante la diversa disponibilità di impianti fotovoltaici, questo significa che la capacità degli impianti fotovoltaici è poco significativa nella stagione invernale per determinare il livello della produzione di energia solare, che invece risulta decisiva nella stagione estiva.

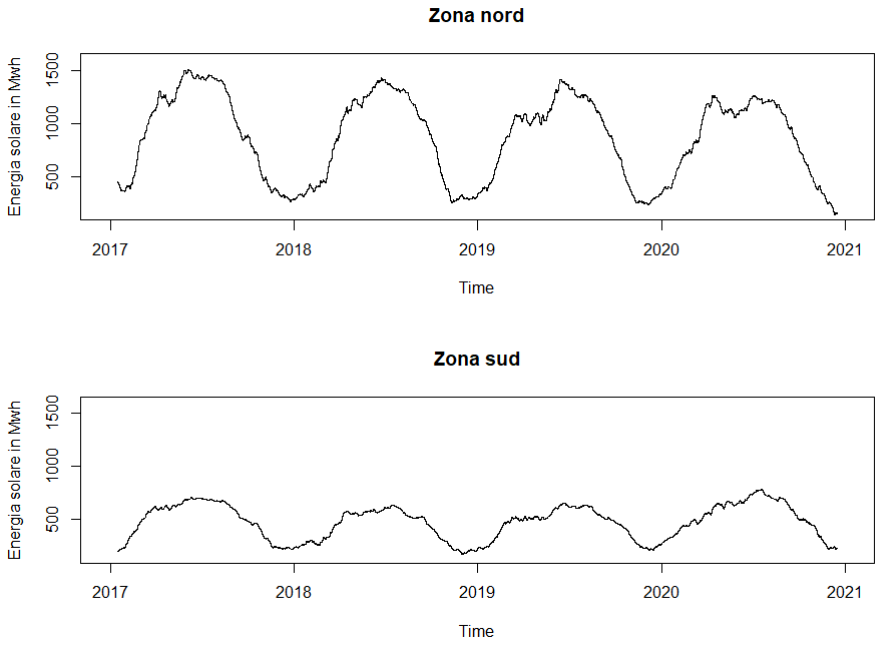


Figura 2.14: Serie zonali a media mobile mensile della produzione di energia solare a capacità costante (anno base 2021)

CAPITOLO TERZO:
MODELLAZIONE DEI DATI

Questo capitolo è dedicato alla presentazione dei risultati relativi all'applicazione di alcuni modelli di regressione utilizzati per stimare le 24 serie storiche della produzione totale di energia proveniente da impianti fotovoltaici nelle diverse fasce orarie del giorno per le zone nord e sud.

In questo modo, adottando l'approccio della modellazione delle serie di ogni ora del giorno si elimina la componente stagionale giornaliera e rimane da stimare solamente la componente annuale.

Dalle analisi esplorative del capitolo precedente si è visto che le serie storiche delle fasce orarie notturne comprese tra le 20:00 e le 4:00 è ragionevole porle costanti nulle a causa della loro esigua variabilità, al netto di sporadici outlier di grandezza sostanzialmente poco significativa. Di conseguenza non saranno oggetto di modellazione o previsione.

3.1 Modello di regressione polinomiale con variabili dummy

Un modello iniziale che riesca a spiegare la variabile della produzione di energia fotovoltaica deve tener conto, oltre della tendenza di fondo della serie, anche della componente stagionale annuale. Per questo, il primo modello che si vuole definire, è un modello di regressione polinomiale con un numero di variabili dummy pari al numero di ritardi stagionali annuali della serie.

Le variabili dummy sono variabili binarie che assumono solo due valori, solitamente 0 e 1, e servono per rappresentare categorie o gruppi distinti, ma nel nostro caso rappresentano i giorni dell'anno solare. Un modello di questo tipo viene specificato nel modo seguente:

$$y_t = \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \beta_1 Cap_t + \beta_2 G.i_t + \beta_3 H\_sun_t + \beta_4 T2m_t + \beta_5 WS10m_t + \gamma_1 d_{1t} + \gamma_2 d_{2t} + \cdots + \gamma_{365} d_{365t} + \varepsilon_t \quad \text{con} \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \ \text{ e } \ t = 1, \dots, n$$

Questo modello tiene conto simultaneamente della componente di trend che è stimata da un polinomio del tempo di grado 3 e dalle variabili esplicative specificate nel paragrafo 2.1, che sono in ordine: la capacità di generazione degli impianti (Cap), l'irraggiamento solare ($G.i$), l'altezza del Sole (H_sun), la temperatura dell'aria a 2 metri

di altezza ($T2m$) e la velocità del vento a 10 metri di altezza ($WS10m$), nonché della componente stagionale annuale calcolata attraverso la regressione di un numero di variabili dummy uguale alla dimensione della stagionalità annuale.

Le ipotesi sottostanti a questo modello sono che la distribuzione dei residui ε_t converga ad una distribuzione White Noise di media zero e varianza costante σ_ε^2. Ciò è possibile osservarlo dalle funzioni di autocorrelazione globale e parziale stimate dai residui del modello e disponibili nell'appendice 1. I correlogrammi dei residui indicano che è presente una componente stagionale annuale che non viene considerata nel modello.

I risultati del modello sono specificati nelle tabelle 3.1 e 3.2, nelle quali sono state modellate le serie storiche della produzione di fotovoltaico nelle zone nord e sud delle fasce orarie delle 5:00, 8:00, 10:00, 12:00 15:00 e 17:00. A scopi descrittivi si è deciso di lasciare all'interno dei modelli tutte le variabili esplicative anche quando gli effetti di quest'ultime non sono statisticamente significativi, questo criterio è stato applicato a tutte le tabelle del capitolo.

Inoltre, per comprendere e selezionare il modello migliore, ovvero quello che si adatta meglio ai dati si sono utilizzate le seguenti misure statistiche:

- Il coefficiente di determinazione, solitamente indicato come R-quadro o R^2, è una misura statistica utilizzata per valutare quanto bene un modello di regressione lineare si adatta ai dati osservati. Esso fornisce una stima della quota di variabilità della variabile risposta viene spiegata dal modello sulla variabilità totale.
  Il coefficiente di determinazione assume valori compresi tra 0 e 1, per cui un valore di 0 indica che il modello non è in grado di spiegare alcuna variazione nella variabile risposta, mentre un valore di 1 indica che il modello spiega completamente la variazione osservata.
  Matematicamente, il coefficiente di determinazione può essere espresso come:
  $$R^2 = 1 - (SSR/SST)$$
  Dove SSR rappresenta la somma dei quadrati dei residui, ovvero le differenze tra i valori osservati della variabile risposta e i valori stimati dal modello e SST è la somma delle differenze tra i valori osservati della variabile risposta e la media dei valori osservati.
  Un valore elevato di R^2 indica che il modello si adatta bene ai dati e spiega gran parte della variazione osservata. Tuttavia, il coefficiente di determinazione da solo non fornisce informazioni complete sulla bontà del modello. È importante considerare anche altri fattori come l'adeguatezza del modello, la significatività statistica delle covariate e la presenza di eventuali violazioni delle assunzioni del modello di regressione lineare.

- Il criterio AIC (Akaike's Information Criterion) è una misura utilizzata per confrontare e selezionare modelli statistici diversi non annidati, in particolare nei contesti di regressione. È stato sviluppato da H. Akaike[4] (1973) e fornisce una stima della qualità del modello, tenendo conto da una parte della sua capacità di adattarsi ai dati osservati, rappresentata dalla log-verosimiglianza del modello ($\log L(\hat{\delta})$) e dall'altra della sua complessità in termini di numerosità dei parametri stimati dal modello (k).
  Il criterio AIC è calcolato utilizzando la seguente formula:
  $$AIC = -2 \log L(\hat{\delta}) + 2k$$
  L'obiettivo è minimizzare il valore del criterio AIC. Quindi tra modelli alternativi, si preferisce il modello con il valore di AIC più basso, poiché indica un migliore equilibrio tra adattamento ai dati e complessità del modello.
  Inoltre, bisogna tenere in considerazione che il criterio AIC non fornisce una misura assoluta della bontà del modello, ma è utile per il confronto tra modelli alternativi.

Qui di seguito sono riportati in formato tabellare i risultati provenienti dalla stima del modello sui dati a disposizione.

Tabella 3.1: Coefficienti stimati del modello di regressione polinomiale con variabili dummy coi dati della zona nord

	Zona NORD					
Fascia oraria:	**5:00**	**8:00**	**10:00**	**12:00**	**15:00**	**17:00**
tempo	0,006	0,523.	-0,381*	1,551**	0,9865*	-0,062
tempo²	-	0,0006**	-	0,001**	0,00096**	-
Cap	-0,006	-1,375*	0,421**	-3,383**	-2,371**	0,06389
G.i.	0,939***	2,581***	2,745***	3,037***	4,402***	8,7366***
H_sun	-5,495	-19,06	-44,78***	-31,92***	-205,7*	-56,11***
T2m	0,3*	8,736***	8,573*	9,729*	8,513**	6,587***
WS10m	-0,609	13,02	28,45*	47,10**	-33,06**	-18,826***
d1	59,077	1187*	-3316*	2968**	19500*	-549,623
…						
d182	132,224.	1326*	134,4	3250**	13020.	462,205
…						
d365	59,582	1184*	-3287*	2974**	19680*	-555,558
R²	0,945	0,9863	0,9876	0,9881	0,9852	0,974
AIC	11594,3	19865,4	21404,9	21693,9	20686,4	17996,5

*Significatività statistica codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

[4] Akaike H. (1973). "Information theory and an extension of the maximum likelihood principle".

Tabella 3.2: Coefficienti stimati del modello di regressione polinomiale con variabili dummy coi dati della zona sud

	Zona SUD					
Fascia oraria:	**5:00**	**8:00**	**10:00**	**12:00**	**15:00**	**17:00**
tempo	0,015.	-0,399***	-0,425***	-0,465***	-0,250**	-0,055.
tempo²	0,000014***	0,00036***	0,00037***	0,00035***	0,00025***	0,000085***
Cap	-0,071**	0,103	0,159	0,377	0,042	-0,079
G.i.	0,698***	1,448***	1,402***	1,365***	2,287***	2,113***
H_sun	-3,83	-26,12**	-18,94***	-18,4***	7,643	-19,35**
T2m	-0,5387*	-6,192**	-9,640***	-3,448	-3,756.	0,8918
WS10m	-0,036	12,71**	1,297	-0,0085	-5,855	4,535**
d1	222,6**	-146,4	465,4	-250,1	-215,2	231,9
…						
d182	335,2***	1489.	1612	887,3	-2435	731**
…						
d365	215,7**	-117,9	277,4	-326,4	-277	213,4
R²	0,928	0,9775	0,9804	0,9764	0,9668	0,931
AIC	12929,53	19553,77	20479,82	20817,72	19707,44	16399,32

*Significatività statistica codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Dai risultati del modello di regressione con stagionalità stimata dalle variabili dummy emergono diversi spunti di riflessione interessati, infatti l'effetto della variabile della capacità installata al nord è statisticamente significativo, tranne nelle ore in cui la produzione di energia fotovoltaica è minore; al sud questa variabile non è particolarmente rilevante, ciò trova giustificazione dal fatto che il trend delle serie storiche orarie del sud sia fortemente parabolico, come il trend della serie della produzione nel suo complesso visto nella figura 2.13. Inoltre, come ci si poteva aspettare, il fattore che maggiormente incide sulla produzione di energia fotovoltaica è la radiazione solare con un effetto positivo maggiore al nord rispetto al sud, mentre l'altezza del sole ha un effetto significativamente negativo in entrambe le zone.

Si nota che la temperatura dell'aria a 2 metri di altezza risulta essere un fattore positivo per la zona nord, ma negativo per le ore del mattino della zona sud, ciò può essere spiegato dal fatto che i pannelli fotovoltaici sono progettati per essere efficienti in maniera ottimale quando le celle solari raggiungono temperature intorno ai 25°C[5]. Infine, si osserva che l'effetto del vento è maggiormente significativo al nord rispetto al sud, con esso che nelle fasce orarie della mattina è positivo e in quelle pomeridiane negativo, ciò fa pensare, dato il clima della regione, che questa variabile possa essere una proxy della copertura nuvolosa.

Il quadro che viene delineato dal modello di regressione è la forte dipendenza tra le variabili atmosferiche e la variabile risposta del modello, ovvero la produzione di energia solare, ciò è molto più evidente nel caso della zona nord.

[5] I produttori di pannelli fotovoltaici utilizzano le Standard Test Conditions (STC) definite dalla Commissione Elettrotecnica Internazionale (IEC) per valutare e ottimizzare il loro prodotto e poter dichiarare la potenza nominale massima di pannello (Pmax).

3.2 Modello di regressione polinomiale con stagionalità trigonometrica

Il modello descritto nel paragrafo precedente avendo 365 dummy è decisamente dispendioso e complesso in termini di numero di parametri da stimare, per questo osservando la stagionalità annuale delle serie storiche delle fasce orarie (figura 3.1), ciò ricorda l'andamento di funzioni trigonometriche. Perciò si è specificato un modello di regressione polinomiale che potesse tener conto della stagionalità annuale delle serie attraverso l'uso di funzioni trigonometriche, il modello specificato è il seguente:

$$y_t = \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \beta_1 Cap_t + \beta_2 G.i_t + \beta_3 H\_sun_t + \beta_4 T2m_t + \beta_5 WS10m_t + \sum_{i=1}^{m} (\beta_{i6} \cos \omega_i t + \beta_{i7} \sin \omega_i t) + \varepsilon_t \quad \text{con} \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \ \text{ e } \ t = 1, \dots, n$$

Dove $m = [\frac{S}{2}]$ è il numero massimo di armoniche che la serie storica può avere, S è il numero di ritardi della stagionalità[6] e $\omega_i = \frac{2\pi i}{S}$ la frequenza angolare.

In questo modello la componente di trend è la stessa del modello di regressione polinomiale con variabili dummy, mentre la componente stagionale è modellata dalle funzioni trigonometriche. Questo tipo di modello è appropriato per cogliere sia le tendenze a lungo termine rappresentate dal polinomio, sia le variazioni stagionali comprese in un ampio arco temporale rappresentate dai termini trigonometrici.

Tabella 3.3: Coefficienti stimati del modello di regressione polinomiale con stagionalità trigonometrica coi dati della zona nord

	Zona NORD					
Fascia oraria:	**5:00**	**8:00**	**10:00**	**12:00**	**15:00**	**17:00**
tempo	0,0011	-0,1165*	-0,169.	-0,082*	-0,189**	-0,0092
tempo²	-	0,00011***	0,0001556**	-	0,00013**	-
tempo³	-	-	-	-	-	-
Cap	-0,0013***	-0,014	-0,021	0,147***	0,035*	0,013***
G.i.	0,952***	2,507***	2,757***	3,054***	4,26***	8,856***
H_sun	2,795***	7,054	6,874.	-27,92***	-7,078	-19,18***
T2m	0,0577	5,983**	17,41***	20,49***	14***	6,217***
WS10m	-0,87706	13,24	69,95***	86,51***	-11,2	-18,48***
cos(2πt/S)	17,652***	-303,6***	-365,4***	-1032***	-358,4**	-144,3***
sin(2πt/S)	1,7183.	88,57***	142,4***	279,5***	114***	92,23***
R²	0,835	0,9806	0,9807	0,9824	0,9775	0,9632
AIC	12472,8	19650,9	21322,2	21541,9	20572,9	17781,8

*Significatività statistica codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

[6] In questo caso è di 365

Tabella 3.4: Coefficienti stimati del modello di regressione polinomiale con stagionalità trigonometrica coi dati della zona sud

	Zona SUD					
Fascia oraria:	**5:00**	**8:00**	**10:00**	**12:00**	**15:00**	**17:00**
tempo	-0,006725	-0,6101***	-0,4997***	-0,4757***	-0,2619***	-0,01657***
tempo²	0,000016***	0,00086***	0,000375***	0,000345***	0,0002623***	0,0002244***
tempo³	-	-0,0000002**	-	-	-	-0,00000006*
Cap	-0,00847***	-0,1638***	0,3809***	0,3921***	0,0578*	0,00907*
G.i.	0,5373***	1,442***	1,405***	1,368***	2,25***	2,011***
H_sun	0,5414***	20,78***	-13,12***	-14,38***	2,861	7,315***
T2m	-0,3532	-3,72*	-8,163***	0,6132	-0,8202	2,105***
WS10m	0,2542	13,36***	1,862	0,8437	-1,681	4,693***
cos(2πt/S)	27,28***	83,99	-0,0535***	-483,7***	-94,55	-60,44***
sin(2πt/S)	3,14*	-25,63*	62,77***	97,04***	12,3	31,24***
R^2	0,8528	0,9697	0,9745	0,9672	0,9562	0,9101
AIC	13248,44	19268,19	20137,49	20573,69	19385,28	16060,11

*Significatività statistica codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

I risultati delle tabelle 3.3 e 3.4, in particolare la significatività statistica della componente trigonometrica e gli alti coefficienti di determinazione, confermano la validità del modello specificato. Per giunta si osserva che gli effetti delle variabili riguardanti le radiazioni solari (G.i.) e la temperatura dell'aria (T2m) sono qualitativamente simili a quelli descritti nel modello del paragrafo precedente.

Dal coefficiente di determinazione è evidente che il modello di regressione polinomiale con stagionalità trigonometrica spiega meglio la variabilità della produzione di energia fotovoltaica nelle fasce orarie in cui è più variabile e ben definita la componente stagionale, ovvero nelle ore con maggiore luce solare.

La figura seguente propone un confronto tra i valori stimati del modello con variabili dummy e il modello trigonometrico, nella quale è possibile osservare una migliore performance del secondo modello nelle fasce orarie pomeridiane in cui c'è maggiore volatilità della produzione.

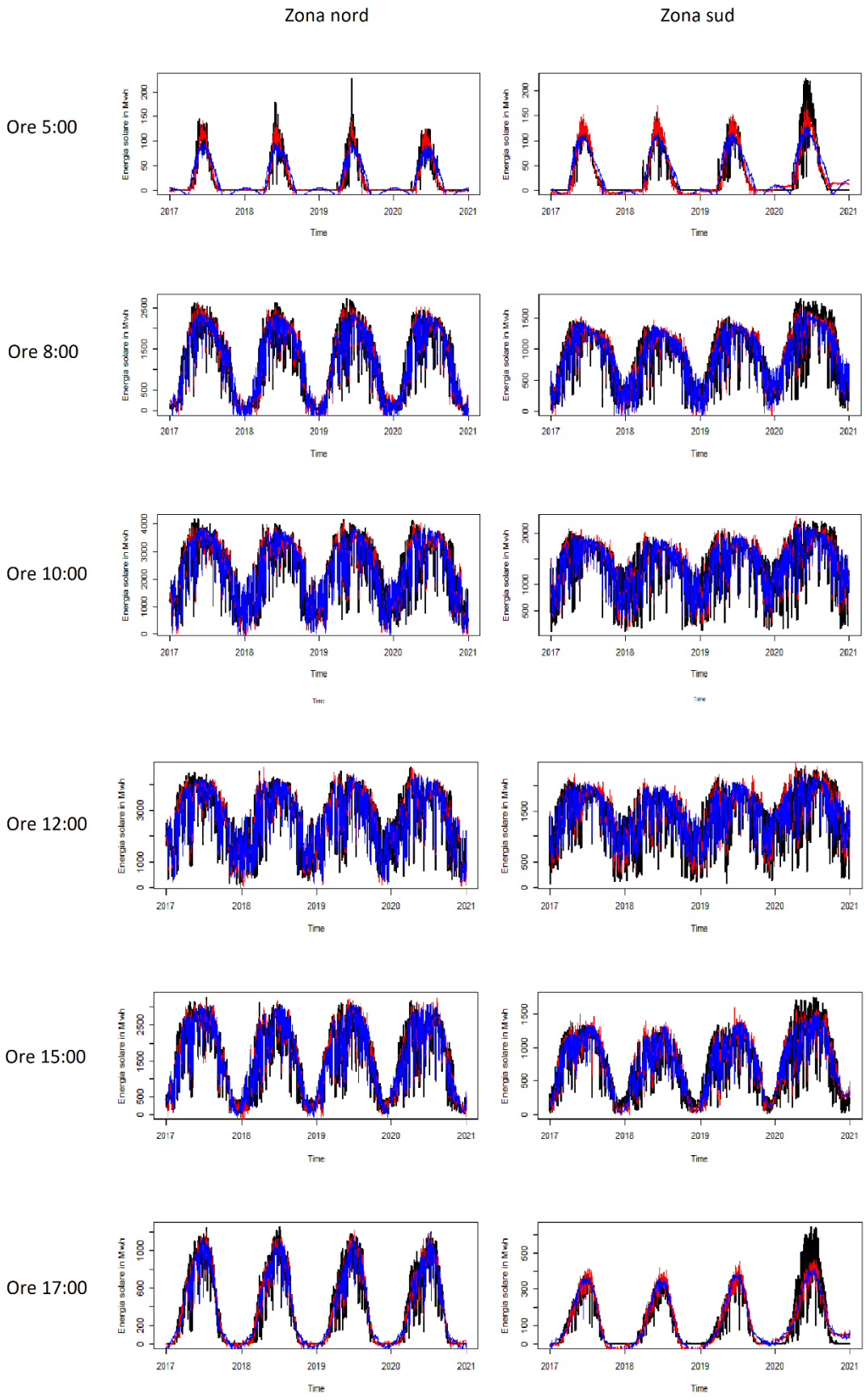


Figura 3.1: Grafici della produzione di energia solare stimata dal modello di regressione con dummy (rosso) e dal modello di regressione trigonometrico (blu) confrontati coi valori reali (nero) per fascia oraria e zona

3.3 Modello di regressione polinomiale con variabile risposta ritardata

I modelli che fino ad ora si sono tenuti in considerazione sono stati vincolati dall'ipotesi di indipendenza tra le osservazioni, ma trattandosi di serie storiche quest'ipotesi è a dir poco inverosimile. Una dimostrazione di ciò è data dalla correlazione tra la variabile di risposta, ovvero la generazione di energia fotovoltaica e la stessa variabile con ritardo di un giorno, osservabile nella figura 2.10. Per questo motivo si è proceduto a specificare un modello di regressione nel quale, oltre alle consuete variabili regressori, è presente la variabile ritardata della produzione di energia fotovoltaica, che funge da variabile proxy utile per controllare il resto delle variabili esplicative non osservate dal modello nell'istante temporale precedente all'osservazione.

Il modello appena descritto si specifica nel modo seguente:

$$y_t = \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \beta_1 Cap_t + \beta_2 G.i_t + \beta_3 H\_sun_t + \beta_4 T2m_t + \beta_5 WS10m_t + \varphi_1 y_{t-1} + \varepsilon_t \quad \text{con} \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \text{ e } t = 1, \dots, n$$

Questo tipo di modello è anche noto come "modello di regressione con componente auto-regressiva" in questo caso di termine 1, AR(1). Questo metodo viene spesso utilizzato per analizzare serie storiche o dati temporali in cui esiste una relazione tra il valore presente della variabile e i suoi valori passati.

Tabella 3.5: Coefficienti stimati del modello di regressione polinomiale con variabile risposta ritardata coi dati della zona nord

	Zona NORD					
Fascia oraria:	**5:00**	**8:00**	**10:00**	**12:00**	**15:00**	**17:00**
tempo	0,0002	-0,09.	0,1193***	-0,71**	-0,143*	-0,253***
tempo²	-	0,0001***	-	0,001**	0,0001**	0,0004***
tempo³	-	-	-	-0,00005**	-	0,0000002***
Cap	0,0003.	-0,065***	-0,107***	-0,07***	-0,014***	0,008***
G.i.	0,742***	2,422***	2,657***	2,891***	4,072***	7,164***
H_sun	-1,080***	23,33***	23,258***	16,32***	12,22***	-8,243***
T2m	-0,462***	-1,457	6,267**	13,33***	8,514***	3,11***
WS10m	-1,002	14,55.	69,563***	95,79***	-10,97	-11,64**
Y_{t-1}	0,622***	0,102***	0,115***	0,114***	0,098***	0,316***
R^2	0,902	0,981	0,9805	0,982	0,9779	0,959
AIC	11706,8	19627,7	21316,35	21602,5	20528,2	17945,7

*Significatività statistica codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Tabella 3.6: Coefficienti stimati del modello di regressione polinomiale con variabile risposta ritardata coi dati della zona sud

	Zona SUD					
Fascia oraria:	**5:00**	**8:00**	**10:00**	**12:00**	**15:00**	**17:00**
tempo	0,006***	-0,471***	-0,345***	-0,713***	-0,192***	-0,127***
tempo²	-	0,0006***	0,0003***	0,0009**	0,021***	0,0002***
tempo³	-	-0,0000002*	-	-0,0000003*	-	-0,00000007**
Cap	-0,0005	-0,112***	0,009	0,076***	0,013.	0,001
G.i.	0,538***	1,396***	1,329***	1,28***	2,035***	1,54***
H_sun	-0,815**	12,73***	8,516***	6,591***	5,778***	3,734*
T2m	-0,822***	-3,633***	-5,056***	1,573	-0,565	0,8**
WS10m	0,146	14,94***	2,758	1,451	0,329	4,533***
Y_{t-1}	0,635***	0,131***	0,084***	0,078***	0,157***	0,561***
R^2	0,912	0,971	0,9741	0,966	0,9581	0,926
AIC	12485,1	19198,3	20143,2	20604,4	19307,2	15767,5

*Significatività statistica codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Le due tabelle sopra indicano che la componente auto-regressiva è positiva e particolarmente significativa, ciò significa che il valore presente della variabile dipendente, ovvero la produzione di energia fotovoltaica, è influenzato in modo significativo dal suo valore del giorno prima. Inoltre, anche le variabili regressori riguardanti il tempo atmosferico sono statisticamente rilavanti, anche se a differenza dei modelli precedenti, l'altezza del sole nel modello specificato ha un effetto positivo sulla variabile risposta.

Osservando i correlogrammi dei residui stimati dal modello nell'appendice 1, si nota la capacità del modello sopra specificato di cogliere l'autocorrelazione presente nelle serie, rendendo quindi più credibile l'ipotesi che i residui si distribuiscano secondo una distribuzione White Noise di media zero e varianza costante.

3.4 Modello REG-ARIMA

La classe di modelli ARIMA è una delle classi di modelli lineari fra le più utilizzate per modellare serie storiche e farne previsioni; infatti, essa considera la natura casuale e la struttura di autocorrelazione globale e parziale tra istanti temporali diversi. Il modello ARIMA(p,d,q) sta per AutoRegressive Integrated Moving Average, ciò fa riferimento a tre processi stocastici[7] del modello:

- Processo auto-regressivo con numero di parametri pari a p, AR(p), definito come segue:
$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$$
$$\text{con} \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \quad \text{e} \quad t = 1, \dots, n.$$
Con l'operatore ritardo $\varphi(\beta) = (1 - \varphi_1\beta - \varphi_2\beta^2 - \cdots - \varphi_p\beta^p)$ è possibile abbreviare il processo AR(p) in $\varphi(\beta)y_t = \varphi_0 + \varepsilon_t$;

- Processo a media mobile con numero di parametri pari a q, MA(q), definito come segue:
$$y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} + \cdots + -\theta_q\varepsilon_{t-q} \quad \text{con} \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \quad \text{e} \quad t = 1, \dots, n.$$
Con l'operatore ritardo $\theta(\beta) = (1 - \theta_1\beta - \theta_2\beta^2 - \cdots - \theta_q\beta^q)$ è possibile abbreviare il processo MA(q) in $y_t = \theta(\beta)\varepsilon_t$;

- Processo di differenziazione di grado d con drift μ, definito come segue:
$$y_t = \mu + y_{t-1} + y_{t-2} + \cdots + y_{t-d} + \varepsilon_t \quad \text{con} \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \quad \text{e} \quad t = 1, \dots, n.$$
Con l'operatore ritardo è possibile abbreviare il processo di differenziazione di grado d in $(1 - \beta)^d y_t = \mu + \varepsilon_t$.
Nel caso in cui d = 1 questo processo è un Random Walk: $y_t = y_{t-1} + \varepsilon_t$.

Quindi, in generale, un processo ARIMA(p,d,q) si definisce come:

$$\varphi(\beta)(1 - \beta)^d y_t = \varphi_0 + \theta(\beta)\varepsilon_t \quad \text{con} \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \quad \text{e} \quad t = 1, \dots, n$$

dove il parametro φ_0 svolge un ruolo diverso a seconda del valore assunto da d, infatti:

- Se d = 0: la serie è stazionaria e φ_0 è collegato alla media del processo;
- Se d > 0: la serie non è stazionaria in quanto contiene d radici unitarie e φ_0 è il drift, quindi è collegato al trend deterministico del processo.

Il modello REG-ARIMA è un modello che combina l'approccio di regressione con il modello ARIMA per incorporare variabili indipendenti nel processo di previsione di una serie storica. Le variabili indipendenti vengono introdotte nel modello come regressori, che possono influenzare la variabile dipendente e contribuire a spiegare la variazione delle serie storiche. In questo modo la parte deterministica della serie viene stimata dai

[7] Un processo stocastico è una successione di variabili casuali indicizzate da un certo istante temporale.

regressori, mentre la parte stocastica di modellazione dei residui viene catturata dal modello ARIMA.

Ciò si traduce nella specificazione del seguente modello:

$$y_t = \beta_1 Cap_t + \beta_2 G.i_t + \beta_3 H\_sun_t + \beta_4 T2m_t + \beta_5 WS10m_t + \varepsilon_t \quad \text{con}$$
$$\varepsilon_t \sim ARIMA(p, d, q) \text{ e } t = 1, \dots, n.$$

La scelta di adattare le serie orarie a questo modello è stata presa in quanto ARIMA è un modello flessibile che può gestire efficacemente le tendenze e i pattern temporali presenti nelle serie storiche. Grazie alle componenti auto-regressiva (AR) e a media mobile (MA), il modello ARIMA può catturare le dipendenze dai valori passati della serie e gli effetti degli errori residui. Inoltre, la presenza di variabili esplicative stagionali come l'irradiazione solare (G.i.) e l'altezza del Sole (H_sun) permette di controllare in modo deterministico la componente stagionale annuale delle serie.

Le matrici di correlazione della figura 2.10 ci suggeriscono una forte correlazione tra la produzione di energia solare oraria con quella del giorno precedente, ciò ci induce a mettere in seria discussione l'ipotesi di stazionarietà in media delle serie orarie, perciò dopo aver verificato la non stazionarietà tramite le funzioni di autocorrelazione globale e parziale presente nell'appendice 2, si è giunti alla decisione di differenziare in media la serie, quindi d = 1 e l'errore del modello assume la seguente forma: $\varepsilon_t \sim ARIMA(p, 1, q)$.

Per determinare il numero ideale di parametri delle componenti auto-regressiva e a medie mobile, rispettivamente p *e* q, si è proceduto ad utilizzare un approccio di tipo stepwise, utilizzando come criterio di selezione del numero di parametri appropriato l'AIC e il test t di significatività statistica.

Per questo modello si è utilizzato il test di Ljung-Box[8] che è un test statistico utilizzato nell'analisi delle serie storiche per verificare la presenza di autocorrelazione nei residui di un modello. L'autocorrelazione si verifica quando i residui del modello non sono indipendenti tra loro nel tempo, ciò indica che il modello non riesce a catturare completamente la struttura dei dati e quindi non risulta adeguato.

L'ipotesi nulla del test di Ljung-Box (H_0) è che i residui del modello siano indipendenti, mentre l'ipotesi alternativa (H_1) è che vi sia autocorrelazione nei residui. Il test calcola una statistica Q basata sui residui del modello e confronta questa statistica con una distribuzione chi-quadrato con un numero di gradi di libertà pari a $m - p - q$. La statistica Q è la seguente:

$$Q(m) = n(n+2)\sum_{k=1}^{m} \frac{1}{n-k}\hat{p}_k^2 \; \dot{\sim} \; \chi^2(m-p-q)$$

Dove m è il numero di autocorrelazioni prese in esame, n il numero di osservazioni e $\hat{p}_k$ è l'autocorrelazione campionaria dei residui al ritardo k-esimo.

[8] G. M. Ljung and G. E. P. Box (1978). "On a Measure of Lack of Fit in Time Series Models".

Se il p-value del test è inferiore a un livello di significatività predefinito (tipicamente 0,05 o 0,01), allora si rifiuta l'ipotesi nulla e si conclude che vi è autocorrelazione nei residui, questo indica la necessità di migliorare il modello o di prendere in considerazione altre variabili che potrebbero spiegare la struttura dei dati. Altrimenti se si accetta l'ipotesi nulla si conclude che i residui sono indipendenti e il modello è appropriato.

Di seguito sono riportati i coefficienti del modello REG-ARIMA stimati per ogni serie storica delle fasce orarie considerate.

Tabella 3.7: Coefficienti stimati del modello REG-ARIMA coi dati della zona nord

	Zona NORD					
Fascia oraria:	**5:00**	**8:00**	**10:00**	**12:00**	**15:00**	**17:00**
AR(1)	-0,244***	0,153***	-	-	0,88***	-1,124***
AR(2)	0,157**	-	-	-	-	-0,985***
AR(3)	0,829***	-	-	-	-	-
AR(4)	-	-	-	-	-	-
MA(1)	-0,412***	-0,935***	-0,743***	-0,753***	-1,728***	0,409***
MA(2)	-0,502***	-	-0,094***	-0,081**	0,728***	0,062**
MA(3)	-0,763***	-	-	-	-	-0,87***
MA(4)	0,566***	-	-	-	-	-0,127***
MA(5)	0,238***	-	-	-	-	-0,024
Cap	0,0004	0,069	0,253	0,075	0,036	-0,013
G.i.	0,759***	2,516***	2,71***	2,948***	4,225***	7,793***
H_sun	-0,806	21,42***	-5,92	-15,555***	15,053***	-13,882***
T2m	0,210	7,848**	16,333***	22,373***	12,261***	10,654***
WS10m	-0,677	6,308	19,619	39,8555**	-26,445**	-19,175***
AIC	11389,2	19572,8	21183,6	21462,5	20451,3	17717,3
p-value del test Ljung-Box	0,274	0,958	0,004	0,079	0,526	0,399

*Significatività statistica codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Tabella 3.8: Coefficienti stimati del modello REG-ARIMA coi dati della zona sud

	Zona SUD					
Fascia oraria:	**5:00**	**8:00**	**10:00**	**12:00**	**15:00**	**17:00**
AR(1)	-0,981***	0,077**	-	-	0,592***	0,677***
AR(2)	-0,215**	-	-	-	-	0,869***
AR(3)	-0,232***	-	-	-	-	-0,674***
AR(4)	-0,156***	-	-	-	-	-
MA(1)	0,245***	-0,932***	-0,819***	-0,871***	-1,401***	-1,358***
MA(2)	-0,601***	-	-0,138***	-0,046.	0,416**	-0,452***
MA(3)	-	-	-	-	-	1,172***
MA(4)	-	-	-	-	-	-0,339***
MA(5)	-	-	-	-	-	-
Cap	0,002	0,114	0,34	0,287	0,255	0,031
G.i.	0,604***	1,395***	1,328***	1,272***	2,104***	2,383***
H_sun	1,17.	13,664***	6,29***	-3,288	8,032***	1,27
T2m	0,119	-1,607	-0,786	6,355*	2,501	3,718***
WS10m	-1,156**	7,755.	0,212	-1,948	-3,689	3,73**
AIC	12038,8	19123,2	20122,9	20560,7	19269,7	15488,2
p-value del test Ljung-Box	0,036	0,803	0,927	0,908	0,564	0,082

*Significatività statistica codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Le tabelle sopra mostrano i risultati del modello REG-ARIMA per zona ed è possibile notare dal p-value del test di Ljung-Box come i residui del modello siano incorrelati fra loro, quindi il modello è adeguato. Inoltre, si osserva per entrambe le zone lo stesso processo stocastico, quando vengono modellate le serie storiche delle fasce orarie più centrali del giorno. Mentre le fasce orarie delle 5:00 e delle 17:00 hanno una componente stocastica più complessa e meno definita. I regressori poco statisticamente significativi per determinare la produzione di energia fotovoltaica nella zona nord sono la capacità installata e il vento nelle ore del mattino, mentre gli unici regressori significativi per la zona sud sono la radiazione solare e l'altezza del sole in gradi.

I grafici seguenti mettono a confronto le stime del modello di regressione con variabile ritardata e il modello REG-ARIMA, con quest'ultimo che sembra adattarsi meglio ai valori osservati delle serie. Ciò è possibile osservarlo anche nei correlogrammi dell'appendice 2 per cui il modello REG-ARIMA riesce a modellare in modo accurato le correlazioni tra i ritardi della stessa serie.

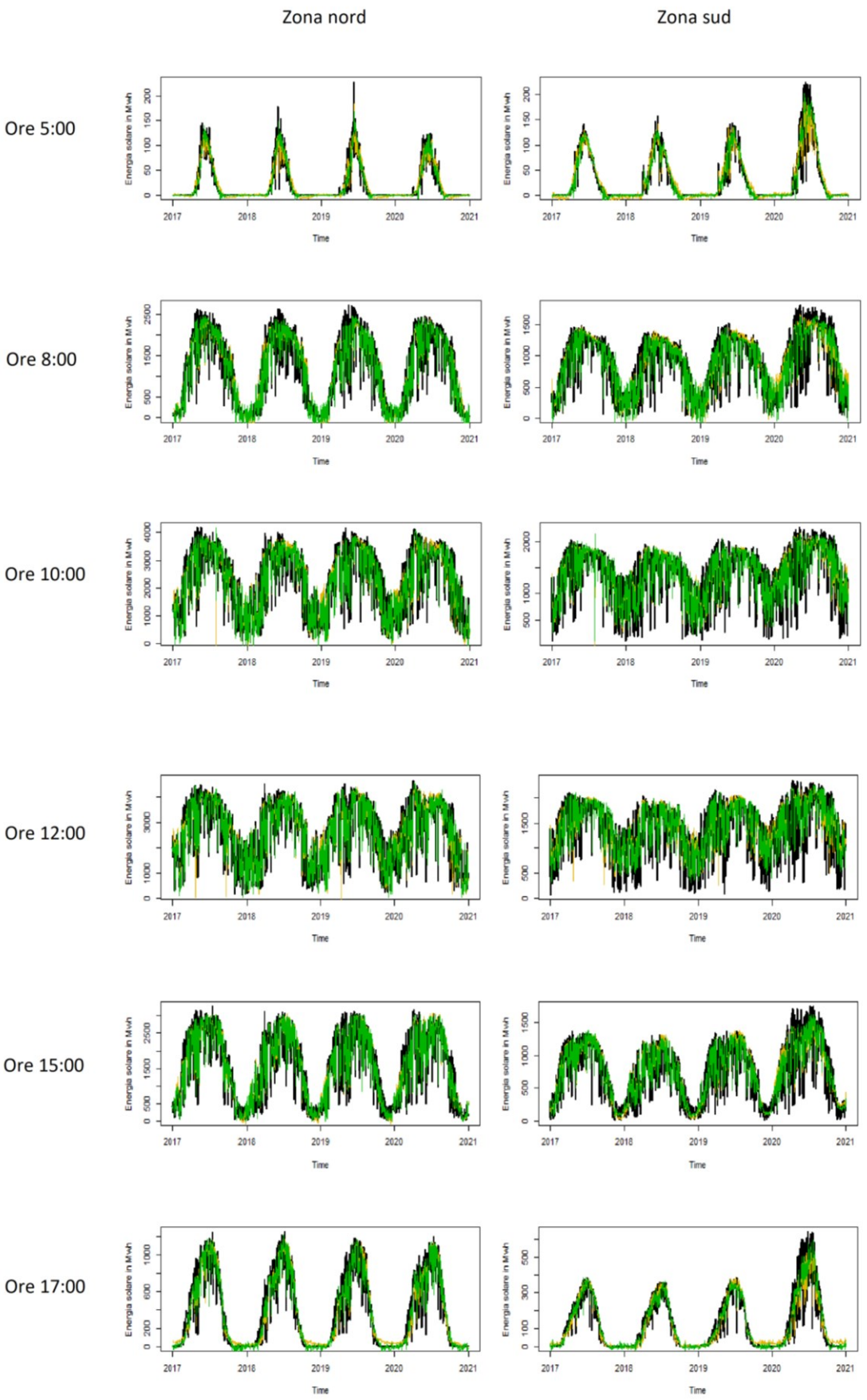


Figura 3.2: Grafici della produzione di energia solare stimata dal modello di regressione ritardato (giallo) e dal modello REG-ARIMA (verde) confrontati coi valori reali (nero) per fascia oraria e zona

3.5 Confronto dell'accuratezza dei modelli

In questa sezione si esporranno degli indicatori fondamentali per misurare l'accuratezza delle capacità predittive dei modelli statistici specificati nei paragrafi precedenti. A tale scopo si confrontano i valori stimati dai modelli (p_t) con i valori reali osservati (r_t) attraverso la misura chiamata errore di previsione (e_t), definita in questo modo:

$$e_t = p_t - r_t \quad \text{con} \quad t = 1, \dots, n$$

Da questa misura si ricavano i principali indicatori per l'accuratezza delle stime che sono:

- L'Errore Medio di previsione (EM): rappresenta la differenza media tra le previsioni del modello e i valori osservati. È la somma degli errori di previsione divisa per il numero di osservazioni.

$$EM = \frac{1}{n}\sum_{t=1}^{n} e_t$$

- La Media Quadratica degli Errori di previsione (MQE): è la radice quadrata della media dei quadrati degli errori tra le previsioni del modello e i valori osservati. Misura quanto gli errori di previsione si discostano in media dai valori osservati, dando un peso maggiore agli errori più elevati.

$$MQE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

- Errore Assoluto Medio di previsione (EAM): è la media delle differenze assolute tra le previsioni del modello e i valori osservati. È un'altra misura dello scostamento medio delle previsioni dai valori effettivi.

$$EAM = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

- Mean Absolute Scaled Error (MASE)**:** è una misura di accuratezza delle previsioni che normalizza l'errore medio assoluto del modello ($EAM_{modello}$) rispetto all'errore medio assoluto di un modello naive di riferimento (EAM_{naive}), in questo caso un modello Random Walk, il quale prevede che il valore futuro sia uguale a quello osservato più recente, ciò si traduce nella seguente specificazione:

$$y_t = y_{t-1} + \varepsilon_t \quad \text{con} \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \text{ e } t = 1, \dots, n$$

Un valore MASE pari a 1 indica che il modello ha prestazioni simili al modello naive, valori inferiori a 1 indicano che il modello è migliore rispetto al modello naive, mentre valori superiori a 1 indicano che il modello è meno accurato del modello naive.

La formula per calcolare il MASE è la seguente:

$$MASE = \frac{EAM_{modello}}{EAM_{naive}}$$

- Statistica U di Theil (U Theil): La statistica U di Theil[9] è un indice di accuratezza delle previsioni che tiene conto sia della differenza tra previsioni e valori reali, sia della variazione temporale dei dati. È spesso utilizzata per analizzare simultaneamente previsioni relative a più serie storiche, in quanto è un indice che normalizza la misura MQE.

$$U = \frac{MQE}{\sqrt{\frac{1}{n}\sum_{t=1}^{n} r_t^2}}$$

- Errore Sistematico (ES): si riferisce a una discrepanza costante tra le previsioni e i valori realizzati. Indica se le previsioni del modello tendono a essere costantemente sovrastimate o sottostimate.

$$ES = \frac{(\bar{p} - \bar{r})^2}{MQE^2}$$

- Errore nelle Variabilità (EV): misura quanto le varianze delle previsioni si discostano dalle varianze dei valori realizzati. Indica come le previsioni del modello riescono a catturare la variazione dei dati reali.

$$EV = \frac{(\sigma_p - \sigma_r)^2}{MQE^2}$$

- Errore nelle Covarianze (EC): si riferisce a una discrepanza tra le covarianze tra le previsioni del modello e le variabili di input e le covarianze tra i valori realizzati e le variabili di input. Valuta come il modello cattura le relazioni tra le variabili. Tale misura assume meno importanza rispetto alle altre in quanto può essere attribuita a fattori accidentali.

$$EC = \frac{2\sigma_p\sigma_r(1 - \rho_{pr})}{MQE^2}$$

Inoltre, dato che le ultime tre misure sono normalizzate si ha che: $ES + EV + EC = 1$

Per calcolare gli indici della seguente tabella si è calcolata la media aritmetica degli indici riguardanti le serie storiche delle fasce orarie considerate in precedenza per ogni modello specificato nel capitolo.

[9] Theil, H. (1950) "A Rank Invariant Method of Linear and Polynomial Regression Analysis".

Tabella 3.9: Indici per valutazione dell'accuratezza dei modelli proposti

	ZONA NORD				ZONA SUD			
TIPO DI MODELLO:	**con Dummy**	**Trigonometrico**	**Ritardato**	**REG-ARIMA**	**con Dummy**	**Trigonometrico**	**Ritardato**	**REG-ARIMA**
EM	5,747E-15	-0,00621	-0,03034	-0,455	-3,694E-15	0,00878	0,02629	-0,169
MQE	182,238	223,569	225,355	215,987	136,853	159,111	158,117	153,809
EAM	139,119	169,965	167,634	159,891	103,836	119,179	116,201	112,103
MASE	0,193	0,27	0,244	0,453	0,323	0,384	0,344	0,494
U THEIL	0,142	0,193	0,18	0,169	0,193	0,235	0,231	0,202
ES	2,26E-33	1,068E-08	4,706E-08	0,00002742	1,75E-33	8,181E-08	0,00001017	2,115E-06
EV	0,015	0,028	0,024	0,008	0,037	0,054	0,054	0,034
EC	0,984	0,971	0,975	0,991	0,962	0,945	0,945	0,965

Gli indici della tabella 3.9 presentano un quadro nel quale il modello di regressione con variabili dummy è il modello più accurato nello stimare le serie orarie della produzione di energia fotovoltaica in entrambe le zone d'Italia, lo si evince dagli indicatori EM, MQE, EAM e la statistica U di Theil. Il risultato è giustificabile dal fatto che questo modello per stimare la stagionalità annuale utilizza un numero elevato di parametri, il che lo rende più complesso, ma allo stesso tempo più preciso nell'accuratezza delle stime.

I modelli di regressione trigonometrico e con variabile risposta ritardata hanno caratteristiche simili per quanto riguarda la precisione delle stime, mentre il modello REG-ARIMA nonostante abbia un EM abbastanza elevato e un EAM di poco meno della metà del modello naive, ha una discreta precisione delle stime, in particolare coglie molto bene la variabilità dei valori effettivi.

CAPITOLO QUARTO:
PREVISIONI

In questo capitolo si riportano i risultati ottenuti in termini di previsioni sulle serie orarie della produzione di energia solare per le zone elettriche nord e sud applicando i modelli di regressione sviluppati nel capitolo precedente introducendo inoltre un modello naive di riferimento, che in questo caso sarà un Random Walk, già specificato nel paragrafo 3.5.

Le previsioni e i relativi indici saranno di tipo ex-post, ovvero per stimare il modello si considererà a disposizione un determinato numero di osservazioni della serie in esame e sui quali si baseranno i valori previsti, che successivamente verranno confrontati coi valori reali osservati.

Inoltre, si procederà con l'esaminare i risultati in termini predittivi dei modelli proposti con una combinazione di previsioni di questi ultimi, calcolata attraverso una media pesata utilizzando la MQE. Infatti, in letteratura[10], la combinazione di previsioni di serie storiche è un approccio molto diffuso che mira a migliorare l'accuratezza delle previsioni combinando più modelli predittivi per una determinata serie storica. Questo metodo è basato sull'idea che diversi modelli possono catturare diverse sfaccettature del comportamento della serie storica, e combinandoli assieme è possibile ottenere una previsione più affidabile e statisticamente robusta del fenomeno oggetto di studio.

4.1 Combinazione di previsioni

La combinazione di previsioni è un approccio che consiste nella sintesi di diverse previsioni provenienti da modelli statistici differenti al fine di ottenere previsioni più accurate e affidabili. Questo approccio è stato sviluppato per contenere i limiti dei singoli modelli di previsione e cogliere i diversi aspetti predittivi della serie, cercando di migliorare le prestazioni complessive della previsione.

L'idea di combinare previsioni è stata introdotta oltre 50 anni fa[11], ed è stata ulteriormente sviluppata nel corso degli anni con risultati empiricamente rilevanti

[10] Xiaoqian Wang, Rob J. Hyndman, Feng Li, Yanfei Kang (2022). "Forecast combinations: An over 50-year review".
[11] John Bates, Clive Granger (1969). "The Combination of Forecasts".

anche nel campo di interesse di questa tesi[12]. L'obiettivo era affrontare il problema di incertezza nei modelli di previsione e sfruttare le differenze tra i modelli per ottenere stime più accurate.

I principali vantaggi di questa tecnica sono: il miglioramento dell'accuratezza della previsione rispetto a un singolo modello, in quanto sfrutta le diverse prospettive dei vari modelli e cerca di compensare i loro errori; la riduzione del rischio, perché combinando diverse previsioni si può ridurre l'azzardo di fare scelte errate basate su un singolo modello e l'adattamento ai cambiamenti del fenomeno di studio, poiché modelli diversi possono reagire in modo diverso a tali cambiamenti.

I limiti maggiori riguardano la complessità dell'implementazione della procedura; la dipendenza delle previsioni dai modelli iniziali e la necessità di aggiornamento dei pesi attribuiti alle previsioni dei modelli con il passare del tempo.

Il metodo della combinazione di previsioni si articola nelle seguenti fasi:

1. Selezione dei modelli di base: scegliere una gamma di modelli di previsione che abbiano dimostrato buone performance e che siano adatti al contesto specifico.
2. Generazione di previsioni: creare singole previsioni utilizzando ciascuno dei modelli selezionati.
3. Definizione dei pesi: assegnare dei pesi ai modelli in base alla loro performance predittiva come la MQE o ad altri indicatori rilevanti come il criterio di Akaike.
4. Combinazione delle previsioni: utilizzare i pesi assegnati per combinare le previsioni dei modelli in moda da ottenere una previsione combinata.
5. Valutazione e aggiornamento: valutare l'accuratezza della previsione combinata, monitorare le performance dei modelli di base e, se necessario, aggiornare i pesi.

In questo lavoro si è applicata la combinazione delle previsioni dei quattro modelli specificati in precedenza svolgendo una media pesata calcolando i pesi in base alla MQE, poiché, come è possibile vedere dalle tabelle 4.1 e 4.2, le differenze di previsione dei modelli sono tali da giustificare questo tipo di combinazione di previsioni.

Per definire i pesi si è adottato l'inverso della misura di MQE normalizzata in modo che i pesi sommino ad 1, in questo modo:

$$\theta_m = \frac{\sum_{m=1}^{M} MQE_m}{MQE_m}$$

Dove θ_m è il peso associato al modello m e M è il numero di modelli utilizzati per la combinazione di previsioni, in questo caso sono i 4 modelli specificati nel capitolo precedente.

[12] Chaman Lal Dewangan, S.N. Singh, S. Chakrabarti (2020). "Combining forecasts of day-ahead solar power".

Infine, basta applicare i pesi di ciascun modello alle previsioni del modello di riferimento per ottenere la combinazione di previsioni specificata:

$$\hat{y}_{comb,t} = \sum_{m=1}^{M} \theta_m \hat{y}_{m,t} \qquad \text{con} \qquad t = T+1, \dots, n$$

In cui $\hat{y}_{m,t}$ è la previsione di stima del modello m al passo t che comincia da $T+1$, cioè il ritardo per il quale inizia la previsione ex-post.

4.2 Confronto dell'accuratezza delle previsioni

In quest'ultima parte del lavoro si è proceduto al confronto dei risultati predittivi stimati dai modelli calcolati coi valori realmente osservati. Per fare ciò si è adottato un approccio definito ex-post nel quale si divide l'insieme dei dati a disposizione in due sottoperiodi: nel primo sottoperiodo in cui $t = 1, \dots, T$, si costruiscono i modelli su cui si faranno le previsioni del secondo sottoperiodo in cui $t = T+1, \dots, n$, le quali verranno confrontate con i valori realmente osservati della serie e su cui si baseranno le misure di accuratezza delle previsioni già definite nel paragrafo 3.5.

In questo caso il primo sottoperiodo in cui si stima il modello va dal 1° gennaio 2017 al 31 dicembre 2019, mentre il secondo sottoperiodo in cui si fa previsione va dal 1° gennaio 2020 al 31 dicembre 2020, in questo modo si otterranno dei risultati che coprono l'intero arco temporale della componente stagionale annuale della serie.

Tabella 4.1: Indici per valutazione dell'accuratezza delle previsioni ex-post dei modelli proposti e della combinazione di previsioni per la zona nord

	ZONA NORD					
TIPO DI MODELLO:	**Random Walk**	**con Dummy**	**Trigonometrico**	**Ritardato**	**REG-ARIMA**	**Combinazione di previsioni**
EM	-636,271	73,804	69,187	14,76	-366,152	-6,946
MQE	1019,917	281,431	262,86	262,576	544,888	252,7
EAM	841,998	217,134	206,195	267,015	457,986	249,862
U THEIL	0,78	0,212	0,214	0,22	0,353	0,2
ES	0,395	0,088	0,063	0,059	0,284	0,027
EV	0,607	0,022	0,029	0,041	0,188	0,084
EC	-0,002	0,89	0,908	0,9	0,528	0,889

Tabella 4.2: Indici per valutazione dell'accuratezza delle previsioni ex-post dei modelli proposti e della combinazione di previsioni per la zona sud

	ZONA SUD					
TIPO DI MODELLO:	**Random Walk**	**con Dummy**	**Trigonometrico**	**Ritardato**	**REG-ARIMA**	**Combinazione di previsioni**
EM	-304,266	-133,393	-29,578	-28,099	-165,903	-80,916
MQE	515,993	254,991	203,21	206,531	280,889	219,293
EAM	419,107	204,494	160,543	161,595	229,092	175,629
U THEIL	0,69	0,321	0,287	0,303	0,4	0,312
ES	0,351	0,261	0,059	0,084	0,297	0,149
EV	0,651	0,271	0,401	0,4	0,391	0,404
EC	-0,002	0,468	0,54	0,516	0,312	0,447

I risultati sull'accuratezza delle previsioni ex-post indicano che i modelli di regressione trigonometrico e con variabile risposta ritardata hanno prestazioni predittive simili in entrambe le zone, mentre il modello REG-ARIMA risulta inadeguato per la previsione di un fenomeno fortemente condizionato dalla componente stagionale, e che per questo non riesce a prevedere correttamente.

Come pronosticato dalla letteratura empirica, nella zona nord la combinazione di previsioni ponderata per la MQE ottiene dei risultati predittivi migliori dei singoli modelli di base; invece, nella zona sud il modello di regressione trigonometrico è migliore in termini predittivi della combinazione di previsioni, da ciò si deduce che la componente stagionale annuale di questa regione sia più regolare e quindi meglio stimabile da questo tipo di modello rispetto alla zona nord. Questo risultato sottolinea le grandi differenze delle caratteristiche legate alla produzione di energia fotovoltaica che sono presenti nelle due aree del Paese.

Come è possibile osservare dalla figura 4.1 le previsioni del modello trigonometrico sono simili alle previsioni formate dalla combinazione di previsioni, inoltre si nota come le previsioni siano più accurate nelle ore del giorno con la maggior produzione di energia fotovoltaica, ciò è dovuto all'incapacità del modello trigonometrico di catturare l'autocorrelazione di serie per le quali la produzione è nulla in certi periodi dell'anno, come testimoniato nell'appendice 1.

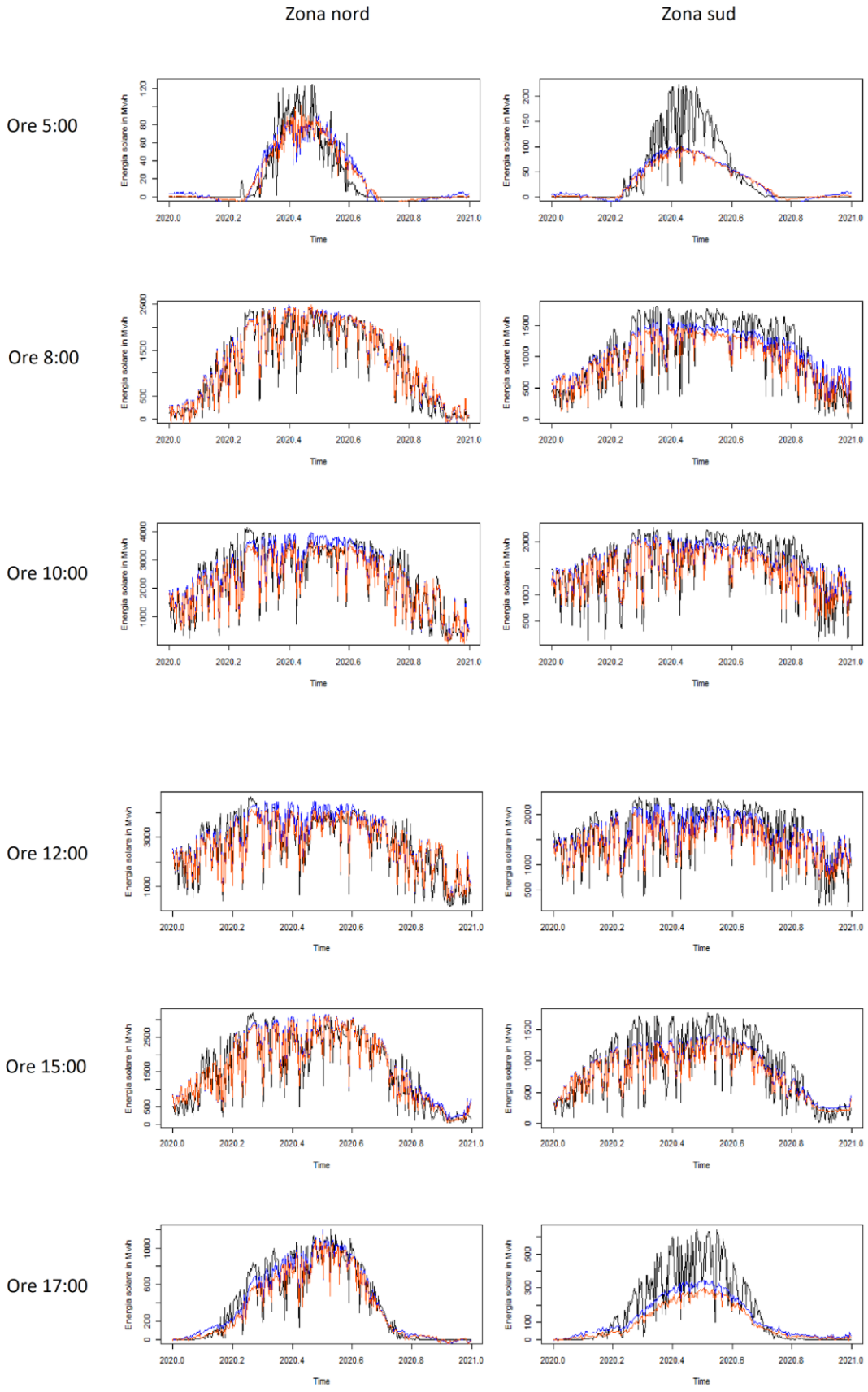


Figura 4.1: Grafici delle previsioni della produzione di energia solare nel 2020 del modello trigonometrico (blu) e della combinazione di previsioni (arancio) confrontati coi valori reali (nero) per fascia oraria e zona

CONCLUSIONI

Nella prima parte di questa tesi si è descritto il funzionamento del Mercato Elettrico allo scopo di comprendere l'importanza di fare previsioni sulla produzione di energia fotovoltaica, e si sono svolte delle analisi preliminari per conoscere e descrivere le caratteristiche peculiari della variabile d'interesse.

Successivamente si è cercato di proporre dei modelli statistici di regressione sempre più elaborati per stimare la produzione di energia fotovoltaica per fasce orarie, eliminando in questo modo la componente giornaliera. Grazie alla stima dei modelli di regressione si è potuto capire gli effetti dei principali fattori atmosferici, come l'irraggiamento solare, l'altezza del sole, la temperatura dell'aria e la velocità del vento, sulla produzione di fotovoltaico per due zone elettriche nord e sud d'Italia.

Infine, utilizzando un approccio ex-post si sono fatte previsioni di un intero anno, al fine di verificare la capacità predittiva dei modelli specificati, inoltre si è adottata una combinazione di previsioni per ottenere delle previsioni più accurate, con risultati rilevanti e diversi in base all'area geografica di riferimento.

Un fattore da tenere in considerazione è che, è stato possibile implementare le variabili atmosferiche nei modelli di regressione in quanto i valori di quest'ultime sono quelli realmente osservati. Infatti, ciò non sarebbe possibile quando si tratta di previsioni future poiché occorrerebbe fare affidamento alle previsioni delle variabili atmosferiche utilizzante nel modello come regressori, complicando ulteriormente la previsione della produzione di energia solare.

I risultati ottenuti indicano che la struttura di generazione di energia fotovoltaica cambia radicalmente a seconda della zona elettrica di attinenza, in quanto cambia il clima atmosferico a cui i pannelli fotovoltaici sono sottoposti, mentre la capacità installata di generazione dell'impianto è un fattore rilevante, ma meno di quanto ci si aspetterebbe, in particolare nelle fasce orarie caratterizzate da limitate radiazione solare. Inoltre, si conferma la tesi secondo cui la combinazione di previsioni di diversi modelli produce performance migliori rispetto alla previsione di un singolo modello.

Si è comunque consapevoli del fatto che esistono modelli più avanzati adatti a stimare in modo più accurato serie storiche con stagionalità complesse come quelle in esame. Nonostante ciò, si ritiene che il lavoro svolto in questa tesi sia utile a comprendere la dinamica della produzione dell'energia fotovoltaica in diverse zone d'Italia e risulta essere un primo passo per la modellazione e previsione dell'attività oggetto di studio, in quanto è un mezzo importante per affrontare la minaccia del cambiamento climatico.

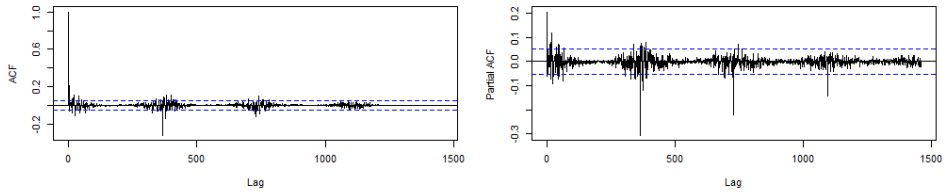
APPENDICE 1

Correlogrammi dei residui dei modelli di regressione stimati

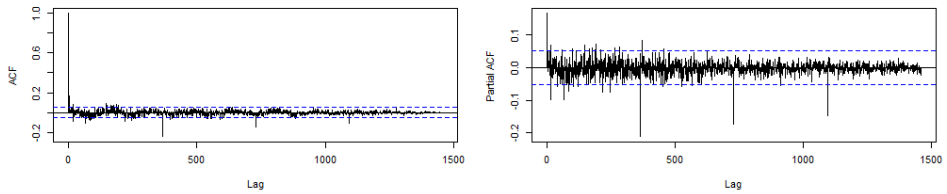
a) Correlogrammi dei residui del modello di regressione con variabili dummy:

Zona nord

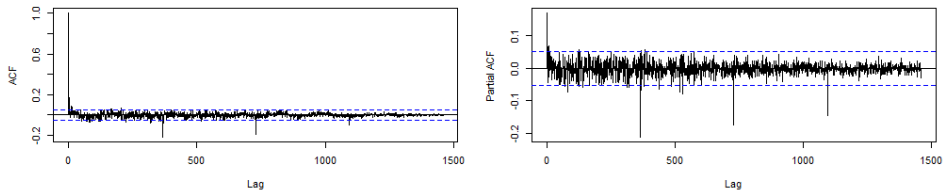
Ore 5:00



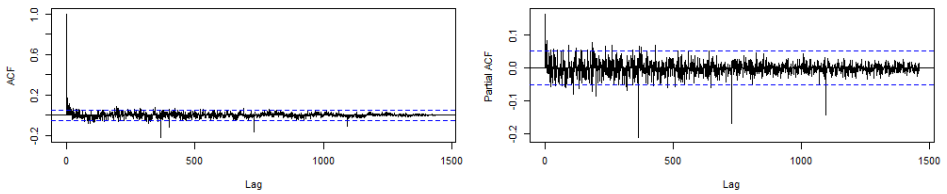
Ore 8:00



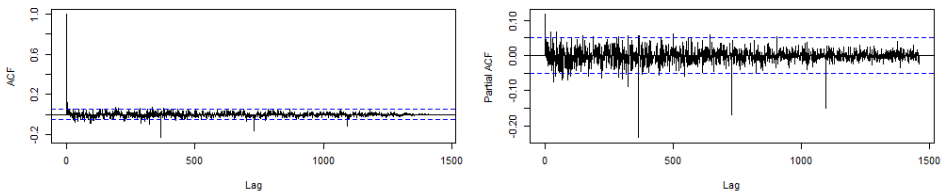
Ore 10:00



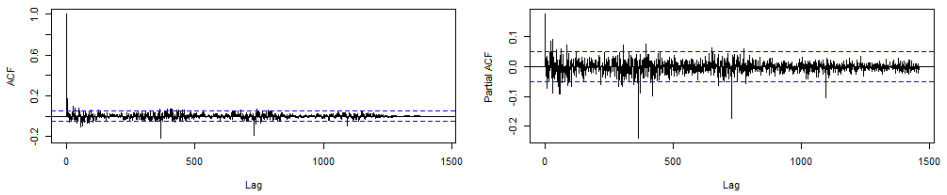
Ore 12:00



Ore 15:00

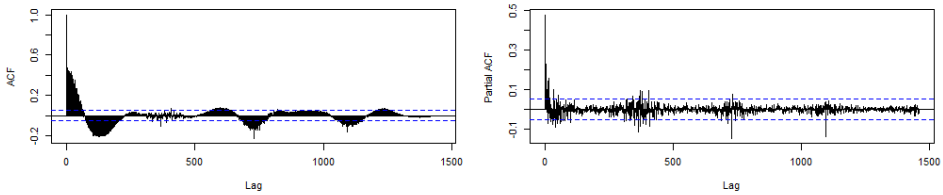


Ore 17:00

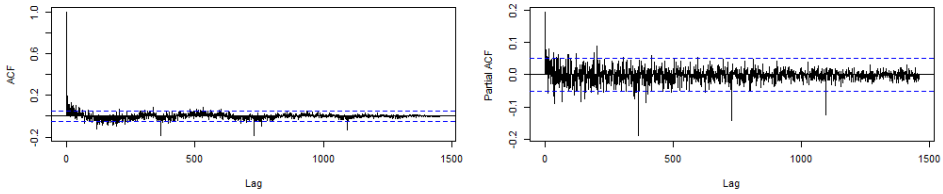


Zona sud

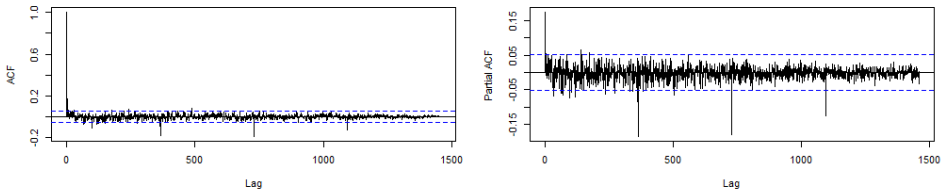
Ore 5:00



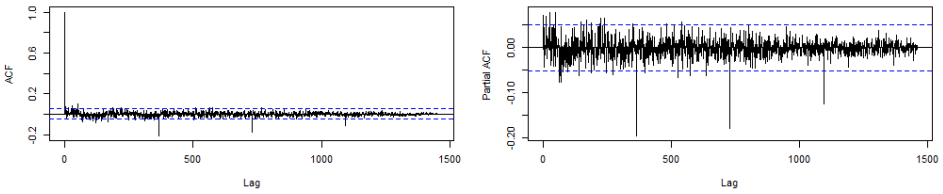
Ore 8:00



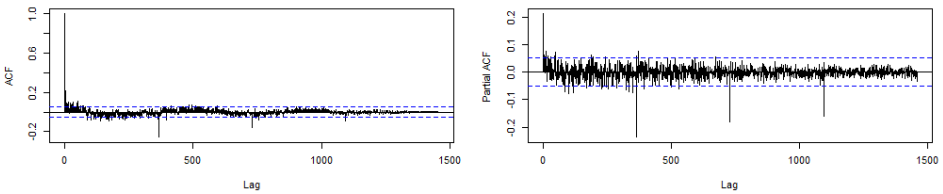
Ore 10:00



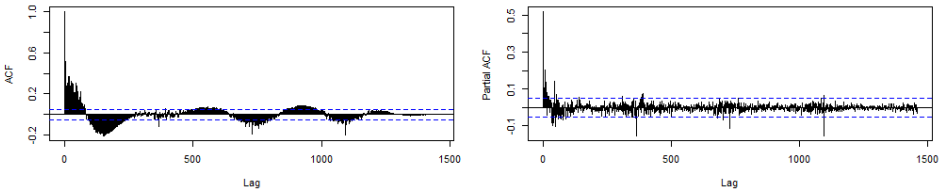
Ore 12:00



Ore 15:00



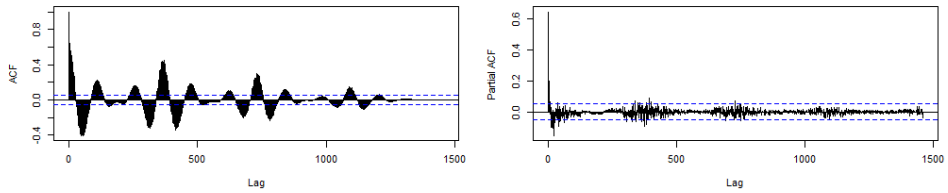
Ore 17:00



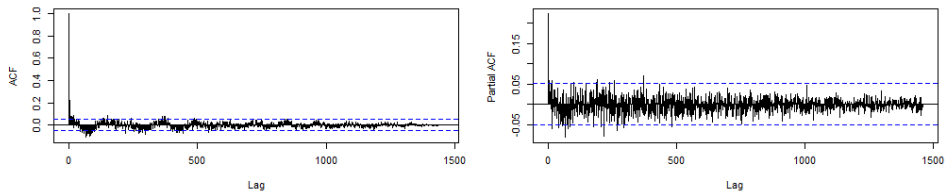
b) Correlogrammi dei residui del modello di regressione trigonometrico:

Zona nord

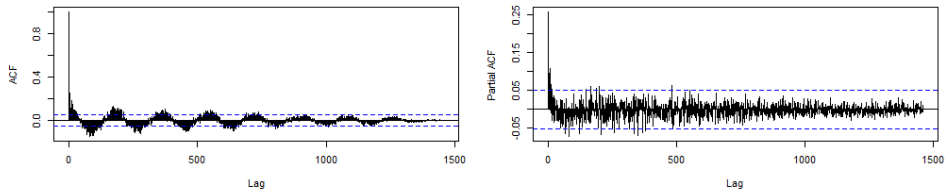
Ore 5:00



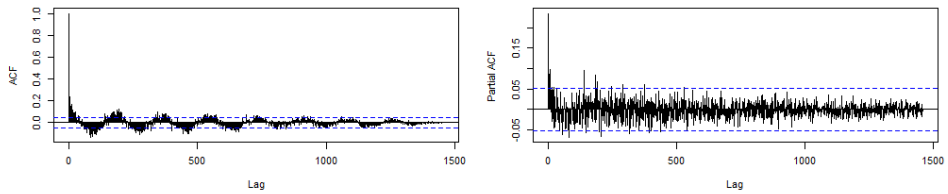
Ore 8:00



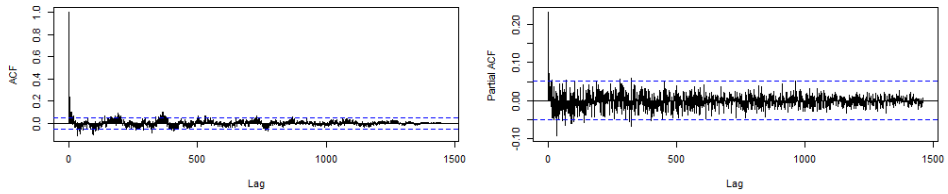
Ore 10:00



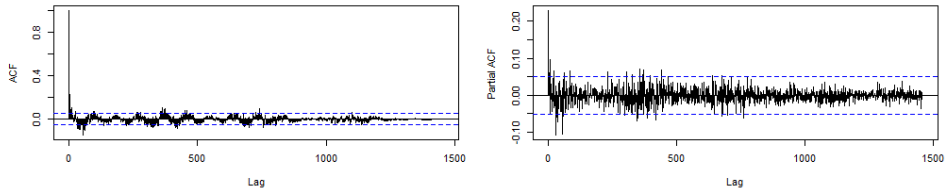
Ore 12:00



Ore 15:00

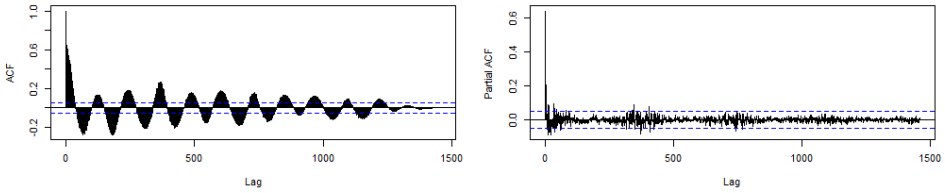


Ore 17:00

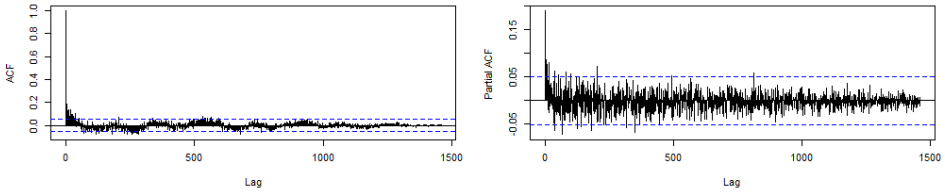


Zona sud

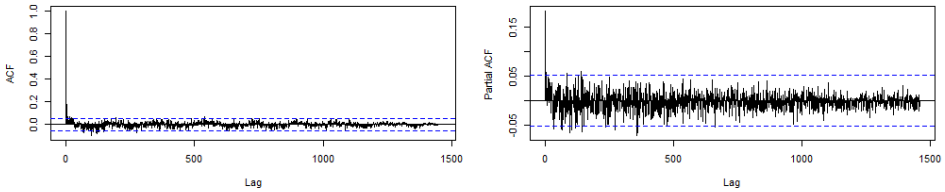
Ore 5:00



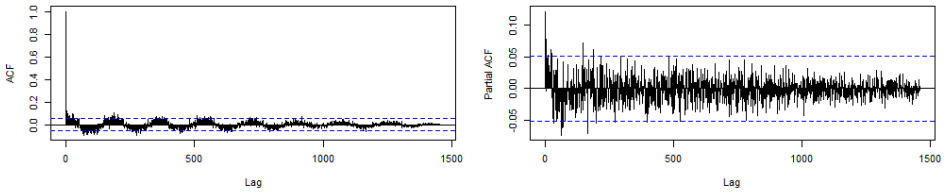
Ore 8:00



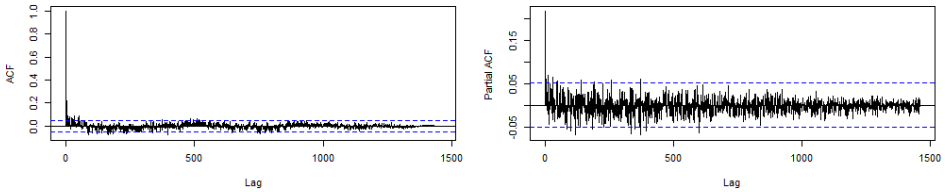
Ore 10:00



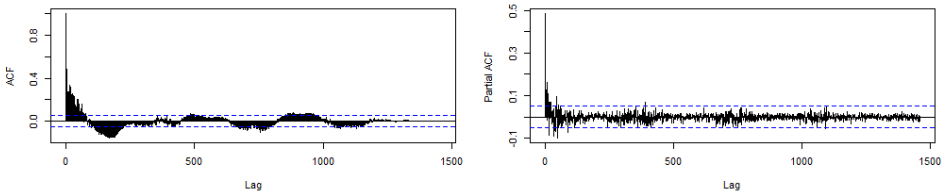
Ore 12:00



Ore 15:00



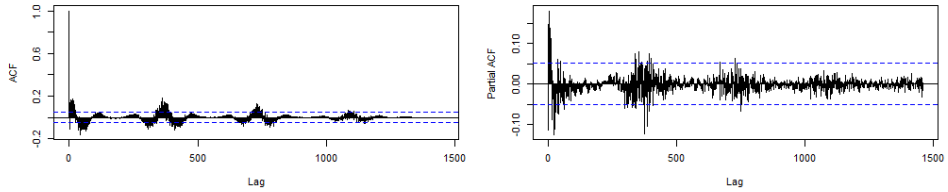
Ore 17:00



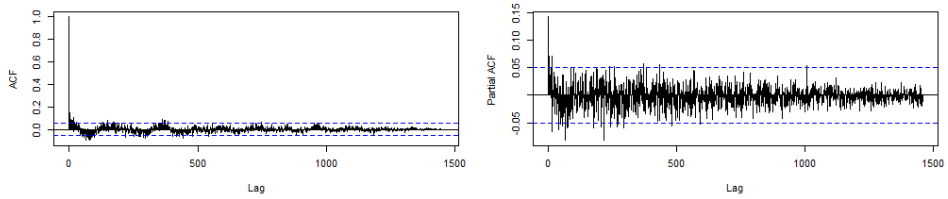
c) Correlogrammi dei residui del modello di regressione con variabile risposta ritardata:

Zona nord

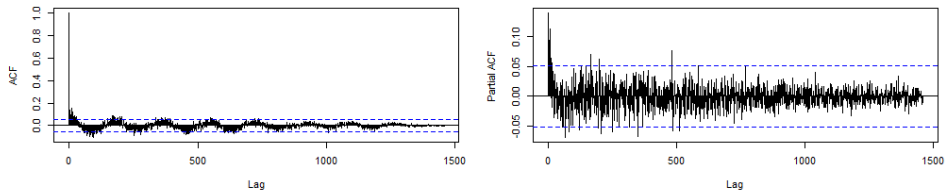
Ore 5:00



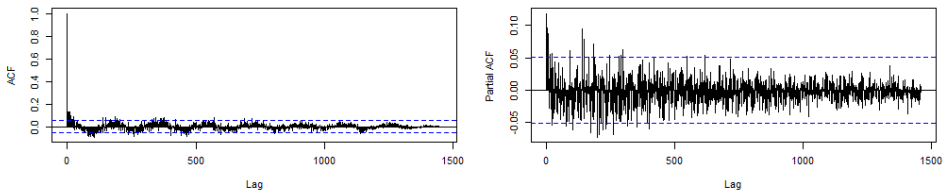
Ore 8:00



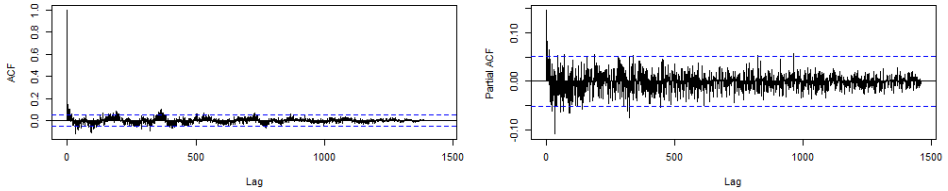
Ore 10:00



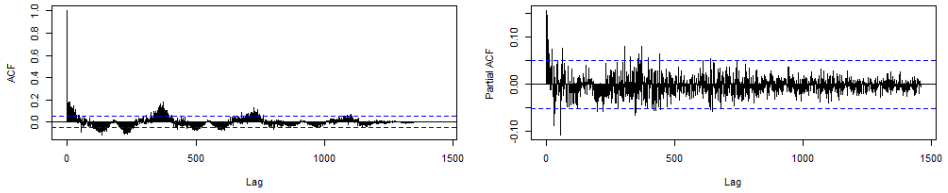
Ore 12:00



Ore 15:00

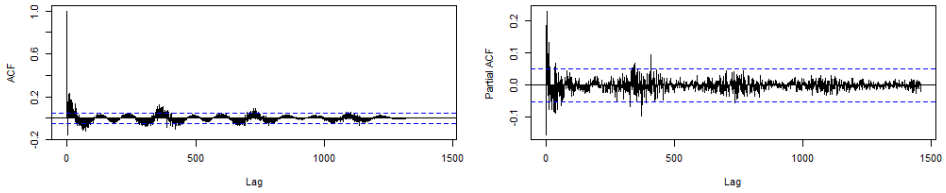


Ore 17:00

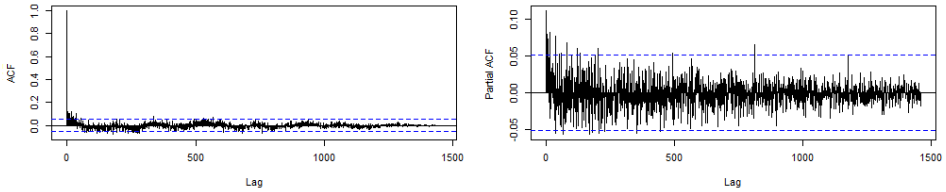


Zona sud

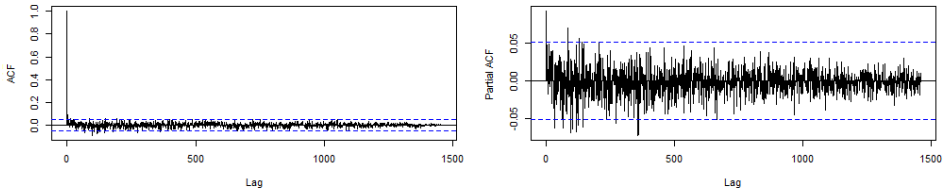
Ore 5:00



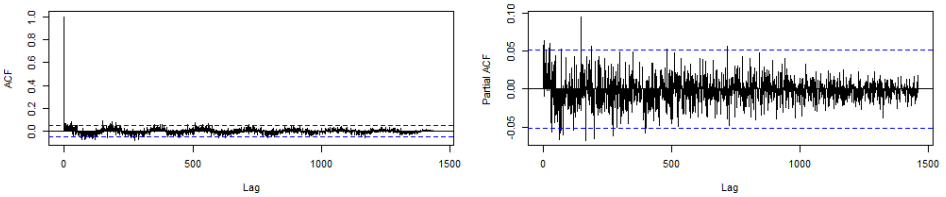
Ore 8:00



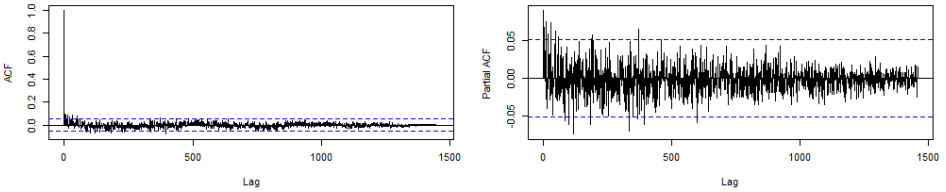
Ore 10:00



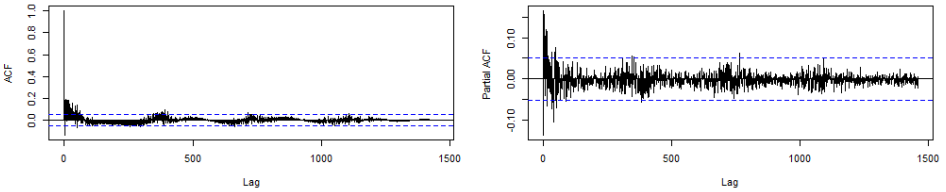
Ore 12:00



Ore 15:00



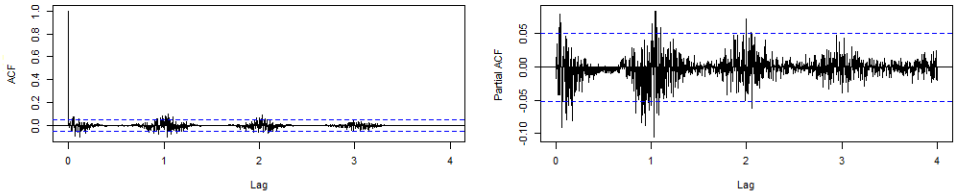
Ore 17:00



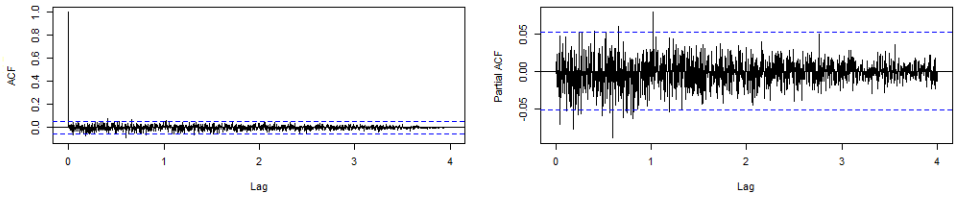
d) Correlogrammi dei residui del modello di REG-ARIMA:

Zona nord

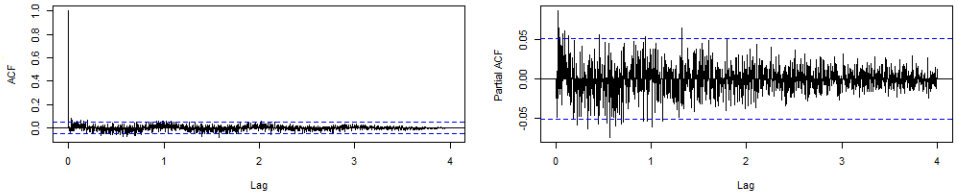
Ore 5:00



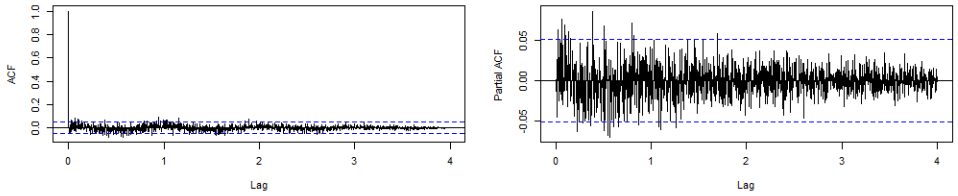
Ore 8:00



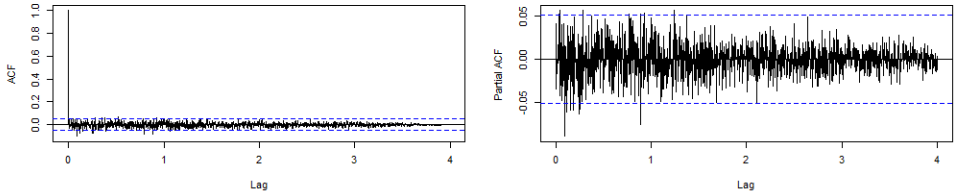
Ore 10:00



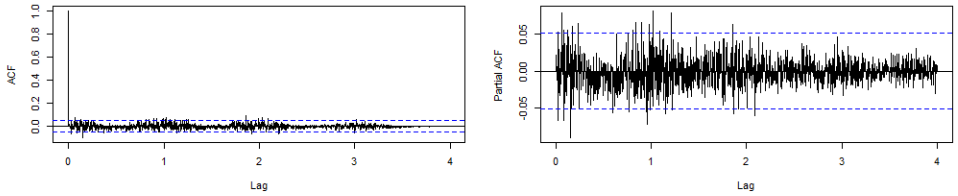
Ore 12:00



Ore 15:00

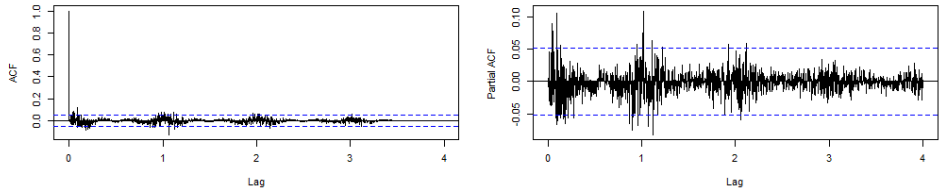


Ore 17:00

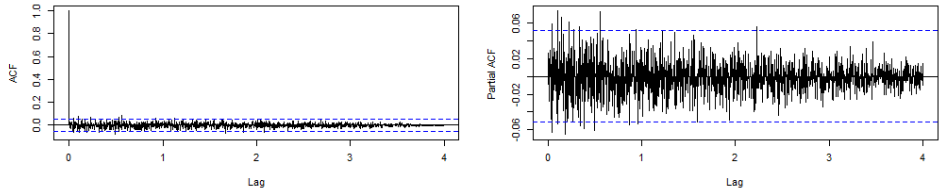


Zona sud

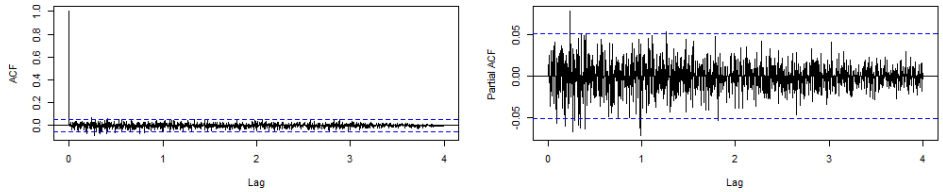
Ore 5:00



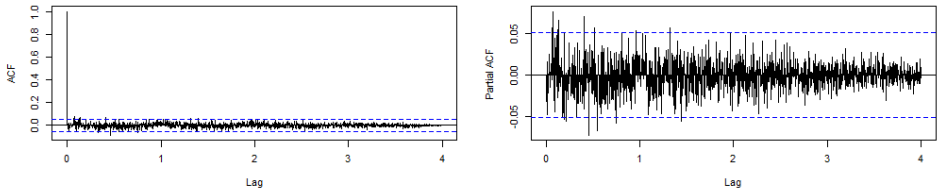
Ore 8:00



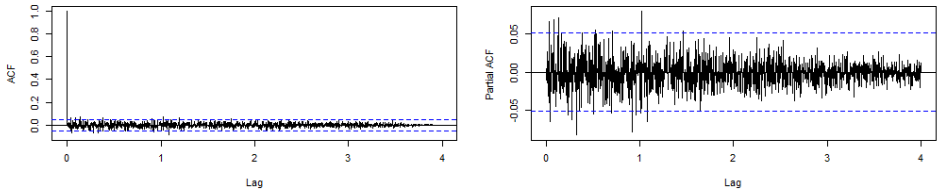
Ore 10:00



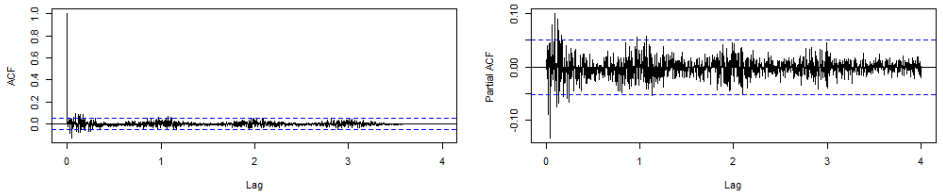
Ore 12:00



Ore 15:00



Ore 17:00

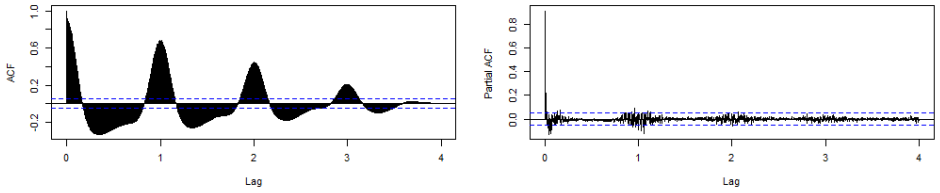


APPENDICE 2

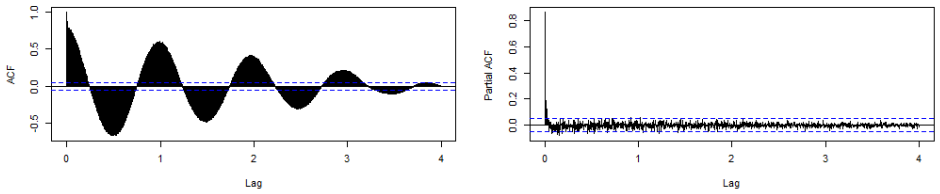
Correlogrammi delle serie orarie della produzione di energia fotovoltaica

Zona nord

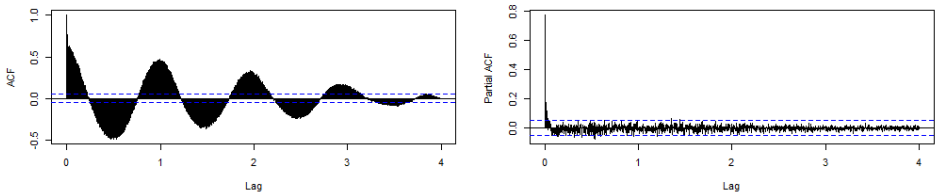
Ore 5:00



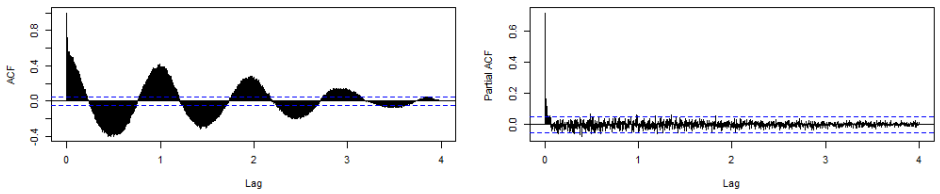
Ore 8:00



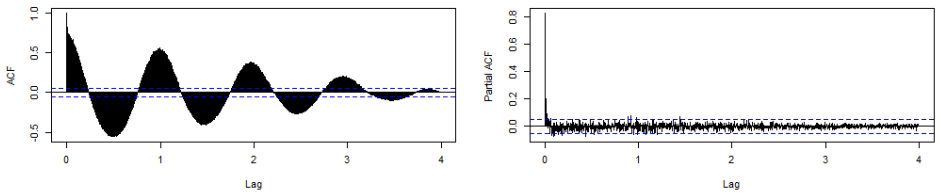
Ore 10:00



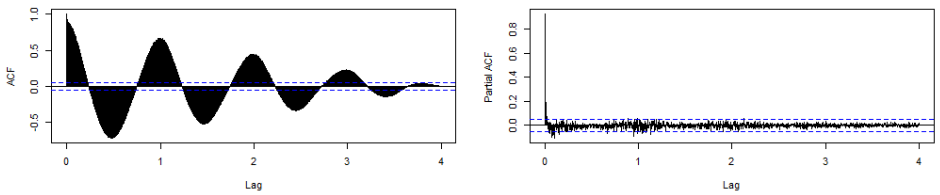
Ore 12:00



Ore 15:00

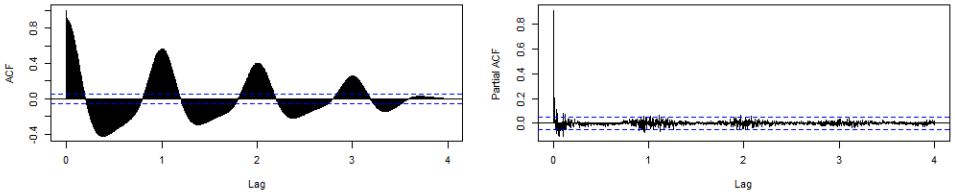


Ore 17:00

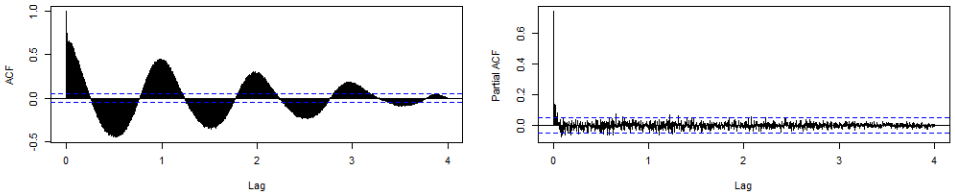


Zona sud

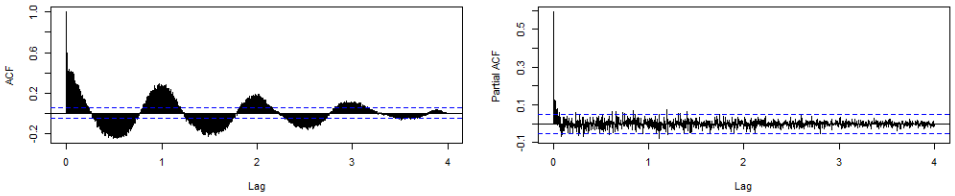
Ore 5:00



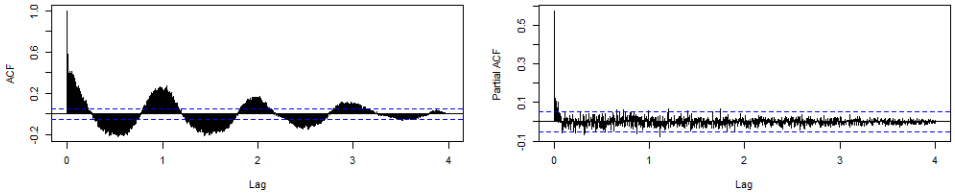
Ore 8:00



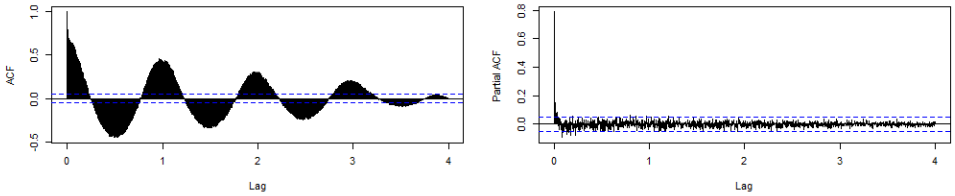
Ore 10:00



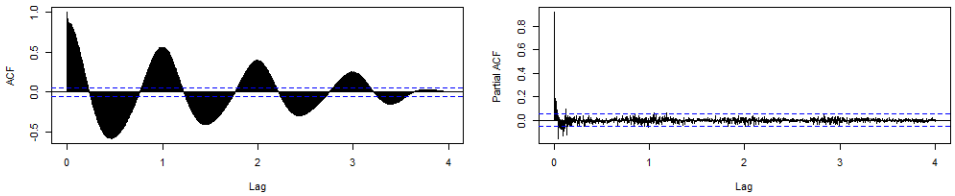
Ore 12:00



Ore 15:00



Ore 17:00



BIBLIOGRAFIA

[1] Akaike H. (1973). “Information theory and an extension of the maximum likelihood principle”.

[2] Behnam Zakeri, Iain Staffell, Paul E. Dodds, Michael Grubb, Paul Ekins, Jaakko Jääskeläinen, Samuel Cross, Kristo Helin, Giorgio Castagneto Gissey (2022). “Role of Natural Gas in Electricity Prices in Europe”.

[3] Chaman Lal Dewangan, S.N. Singh, S. Chakrabarti (2020). “Combining forecasts of day-ahead solar power”.

[4] Gestore dei Mercati Elettrici (2009). “Vademecum della borsa elettrica”.

[5] Gestore dei Mercati Elettrici – Glossario: https://www.mercatoelettrico.org/it/Tools/Glossario.aspx

[6] G. M. Ljung and G. E. P. Box (1978). “On a Measure of Lack of Fit in Time Series Models”.

[7] John Bates, Clive Granger (1969). “The Combination of Forecasts”.

[8] Mathieu David, Mazorra Aguiar Luis, Philippe Lauret (2018). “Comparison of intraday probabilistic forecasting of solar irradiance using only endogenous data”.

[9] Muhammad Naveed Akhter, Saad Mekhilef, Hazlie Mokhlis, Noraisyah Mohamed Shah (2019). “Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques”.

[10] Rob J. Hyndman, George Athanasopoulos (2021). “Forecasting: Principles and Practice” OTexts.

[11] Sito web del Photovoltaic Geographical Information System (PVGIS): https://re.jrc.ec.europa.eu/pvg_tools/it/tools.html

[12] Sito web di Terna S.p.A.: https://www.terna.it/it

[13] Tommaso Di Fonzo, Francesco Lisi (2013). “Serie storiche economiche” Carocci.

[14] Xiaoqian Wang, Rob J. Hyndman, Feng Li, Yanfei Kang (2022). “Forecast combinations: An over 50-year review”.

[15] Theil, H. (1950). “A Rank Invariant Method of Linear and Polynomial Regression Analysis”.