

UNIVERSITÀ POLITECNICA DELLE MARCHE

INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

PROGETTO DATA SCIENCE 2021-2022

**Analisi delle serie temporali delle
precipitazioni in Finlandia, Svezia e
Norvegia**



Studenti:

MASSIMO CIAFFONI

SIMONE CAPPANERA

Indice

1	Introduzione	2
1.1	Dataset e Preprocessing	2
2	Finlandia	4
2.1	ARIMA	6
2.2	SARIMAX	9
3	Svezia	11
3.1	ARIMA	13
3.2	SARIMAX	16
4	Norvegia	18
4.1	ARIMA	20
4.2	SARIMAX	23

Capitolo 1

Introduzione

In questo progetto, l'obiettivo è quello di svolgere l'analisi delle serie temporali di un determinato fenomeno tramite l'utilizzo dell'ecosistema Python. In particolare, per questo task, si sono andate ad utilizzare le librerie *Pandas* e *statsmodels* di Python. Quest'ultima è una libreria che fornisce classi e funzioni per la creazione di modelli statistici, per il calcolo dei test statistici e per l'esplorazione dei dati ed è stata costruita sopra le librerie *NumPy*, *SciPy* e *Matplotlib*. *Statsmodels* ha diverse features al suo interno. Tra queste è presente il modulo *statsmodels.tsa* che contiene diverse classi di modelli e funzioni utili per l'analisi delle serie temporali. Parlando dell'analisi effettiva, gli step che sono stati eseguiti per ognuno dei tre casi presi in considerazione sono i seguenti:

- Studio della stazionarietà della serie con relativa stima del parametro D del modello ARIMA;
- Stima dei parametri P e Q del modello ARIMA;
- Fitting del modello ARIMA e relativo forecasting;
- Out-of-time Cross Validation per il modello ARIMA;
- Fitting e relativo forecasting del modello SARIMAX.

1.1 Dataset e Preprocessing

Il dataset utilizzato per l'analisi riguarda le precipitazioni e le temperature in Finlandia, Svezia e Norvegia dall'inizio del 2015 alla fine del 2019. Il dataset è disponibile presso il seguente link:

<https://www.kaggle.com/datasets/adamwurdits/finland-norway-and-sweden-weather-data-20152019>

I dati contenuti al suo interno sono stati raccolti dal Climate Data Online (CDO) del National Centers For Environmental Information (NCEI). Il dataset contiene i dati giornalieri sulle precipitazioni medie e la temperatura dell'aria in ognuna delle tre nazioni (in unità metriche). L'insieme dei dati originale, raccolto dal sito del CDO, consisteva in circa 4.9 milioni di osservazioni individuali, provenienti da 1306 stazioni meteorologiche diverse all'interno dei tre Paesi. Nel dataset presente nel link, invece, i dati mancanti dell'insieme originale sono stati sostituiti con la media giornaliera, mediata su tutte le stazioni meteorologiche della nazione presa in considerazione.

Di seguito sono descritti i vari campi che compongono la tabella:

Campo	Tipo di dato	Descrizione
Country	Country	Nazione in osservazione.
Date	Date	Data dell'osservazione.
Precipitation	Number	Quantità di precipitazioni in centimetri.
Snow_depth	Number	Quantità di neve accumulata al suolo in millimetri.
TAvg (Temperature average)	Number	Media nazionale delle temperature medie giornaliere in gradi Celsius.
TMax (Temperature maximum)	Number	Media nazionale delle temperature massime giornaliere in gradi Celsius.
TMin (Temperature minimum)	Number	Media nazionale delle temperature minime giornaliere in gradi Celsius.

Visto che lo studio delle serie temporali si concentra sulle precipitazioni, sono stati utilizzati unicamente i campi **Country**, **Date** e **Precipitation** del dataset appena descritto. Prima di iniziare l'analisi vera e propria, il dataset è stato estratto dal file *.csv* e salvato in un DataFrame tramite la libreria *Pandas* di Python. Inoltre, si è scelto come indice di questa struttura dati, caratteristica della libreria, il campo **Date**, convertito in un oggetto *datetime* di *Pandas*. Infine, siccome le analisi sulle tre diverse nazioni devono essere eseguite separatamente, utilizzando le operazioni ottimizzate messe a disposizione dalla libreria *Pandas*, si sono ricavati tre DataFrame, uno per ogni nazione, a partire da quello ottenuto al passo precedente.

Capitolo 2

Finlandia

La prima analisi si è concentrata sulla serie temporale delle precipitazioni in Finlandia, che è possibile vedere rappresentata in Figura 2.1. Questa mostra la quantità di precipitazioni, espressa in centimetri, in Finlandia nel periodo considerato.

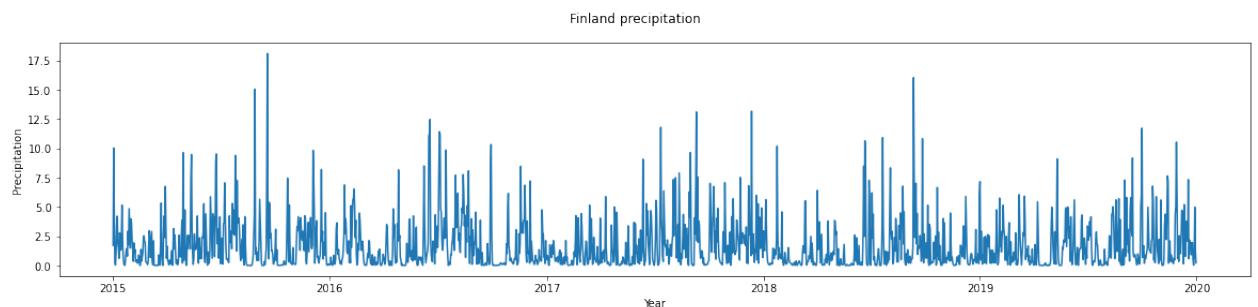


Figura 2.1: Variazione delle precipitazioni in Finlandia nel tempo

Per questa analisi si sono allenati due modelli, così da poterne poi anche confrontare i risultati: un modello ARIMA (acronimo di Auto Regressive Integrated Moving Average, ossia autoregressivo integrato a media mobile) e la sua variante stagionale con regressori esogeni, SARIMAX. Un modello ARIMA, così come le sue varianti, è caratterizzato da 3 termini:

- **P** è l'ordine del termine AR (Auto Regressive);
- **Q** è l'ordine del termine MA (Moving Average);
- **D** è il numero di differenziazioni necessarie per rendere stazionarie le serie temporali (Integrated).

Per ottenere un buon modello bisogna stimare i valori di questi parametri. Il primo passo consiste nello studio della stazionarietà della serie temporale, così da

poter stimare il parametro di differenziazione del modello ARIMA. Nella stima del parametro D, bisogna stare molto attenti a non differenziare eccessivamente la serie, perché una serie troppo differenziata potrebbe perdere di significato. A tal proposito, un modo molto efficace per valutare se la serie è stazionaria o meno è quello di utilizzare l'**Augmented Dickey Fuller test**. Per fare ciò, nel progetto è stata implementata una funzione chiamata *test_stationarity*, che, data una serie temporale in input, ne esegue il test appena citato e stampa in output i risultati, ossia le metriche restituite dall'applicazione del test alla serie temporale data. Tra queste, per valutare la stazionarietà, è fondamentale il *p-value*. Infatti, l'ipotesi nulla del test ADF è che le serie temporali non siano stazionarie. Quindi, se il *p-value* del test è inferiore al livello di significatività, ossia $p < 0.05$, allora si rifiuta l'ipotesi nulla e si deduce che la serie temporale è effettivamente stazionaria. In questo caso, il *p-value* è risultato essere pari a 0.0, valore un po' sospetto essendo così netto, che si è pensato potesse essere un errore. Per questo, nonostante il risultato indichi che la serie sia stazionaria, si è andati poi per sicurezza anche ad analizzare i grafici dell'autocorrelazione della serie e delle sue differenziazioni di primo e secondo ordine. Questi, come è possibile vedere in Figura 2.2, mostrano effettivamente come la prima serie sia già abbastanza stazionaria e come le sue differenziazioni avendo l'autocorrelazione che va nella zona negativa abbastanza velocemente siano state differenziate troppo. Per questo, alla fine, si è scelto di impostare il parametro D a 0.

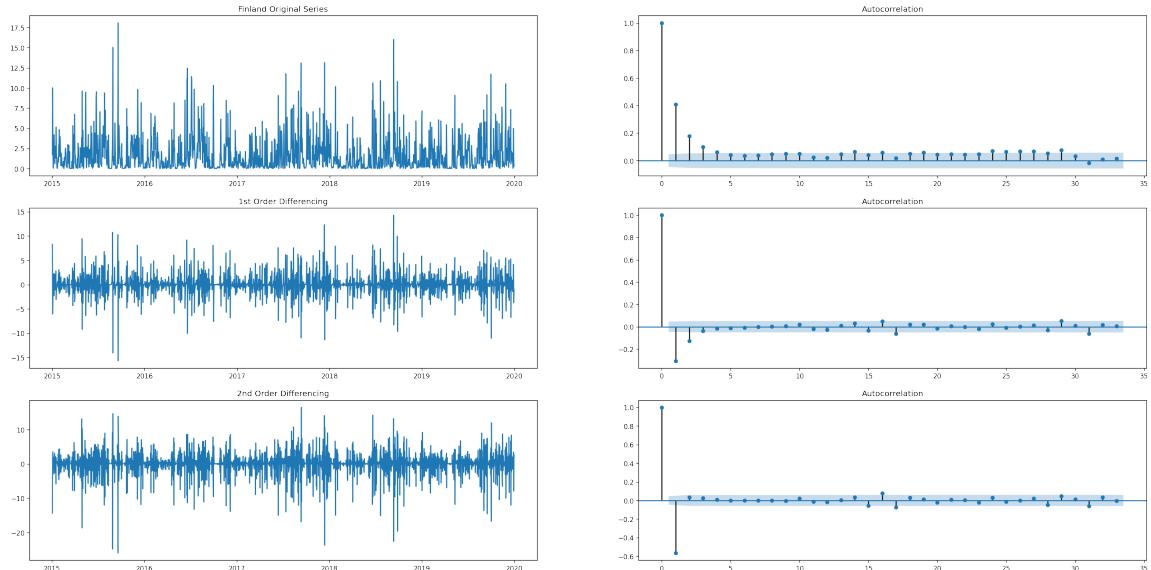


Figura 2.2: Studio Stazionarietà della serie temporale

Il secondo step consiste nella stima dei parametri rimanenti P e Q del modello

ARIMA. A tale scopo sono stati utilizzati i grafici di autocorrelazione e autocorrelazione parziale, presenti in Figura 2.3. Infatti, è possibile scoprire se il modello ha bisogno di un contributo da parte della componente AR (parametro P) ispezionando il grafico dell'autocorrelazione parziale (**PACF**) e scegliendo un numero di termini AR, e quindi il valore del parametro P, pari al numero dei lag che superano il limite in questo grafico. Inoltre, analogamente, si può guardare il grafico dell'autocorrelazione (**ACF**) per stimare il numero di termini MA (parametro Q), visto che l'ACF rappresenta quanti termini MA sono necessari per rimuovere qualsiasi autocorrelazione nella serie stazionaria. In Figura 2.3 si può osservare come il grafico dell'autocorrelazione parziale abbia un lag abbastanza significativo, in quanto si trova ben al di sopra dell'area limite; mentre nell'altro grafico si hanno ben tre lag al di sopra della linea significativa. Per questo si è deciso di assegnare al parametro P il valore 1 e al paramentro Q il valore 3.

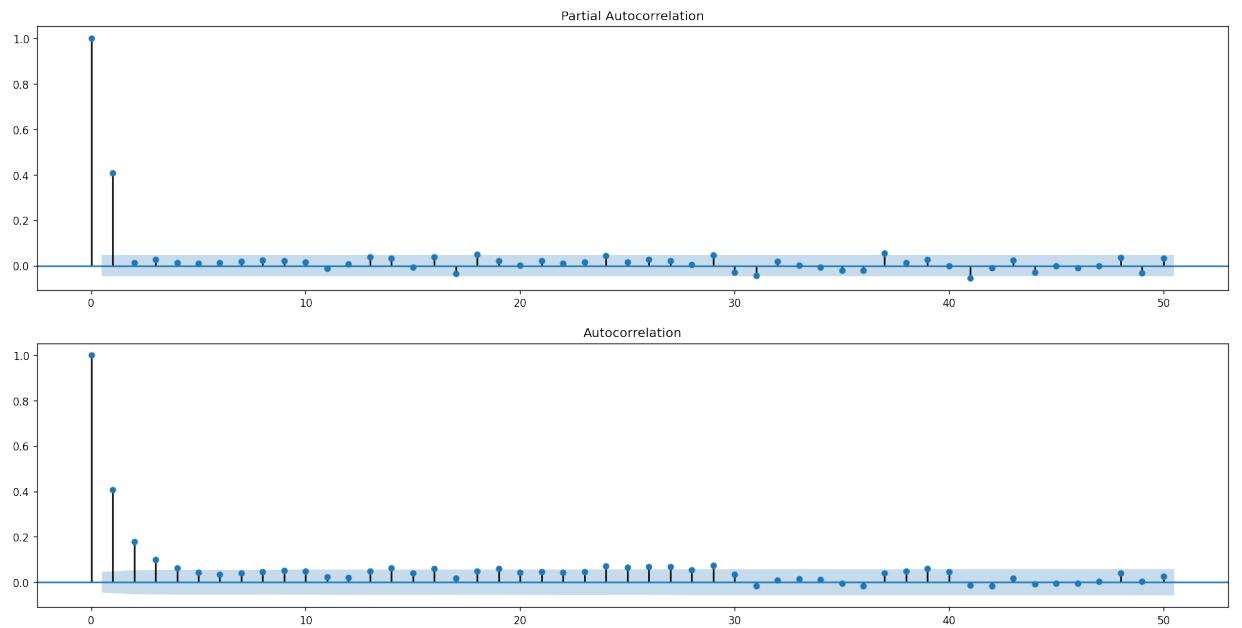


Figura 2.3: Plot dell'autocorrelazione e dell'autocorrelazione parziale della serie differenziata

2.1 ARIMA

Una volta determinati i valori di P, D e Q, si ha tutto il necessario per allenare il modello ARIMA. Visto che i valori rispettivamente di (P, D, Q) pari a (1, 0, 3) corrispondono a un modello un po' complesso, si sono provati inizialmente valori più bassi di Q. Però, alla fine, i risultati dei sommari dei vari modelli addestrati hanno reso evidente come il modello con questi parametri specifici sia il migliore, avendo

un valore di AIC minore rispetto agli altri modelli e tutti i valori della colonna ' $P > |z|$ ' molto significativi ($\ll 0.05$). Infatti, l'AIC, ossia il criterio d'informazione di "Akaike", fornisce una misura della qualità della stima di un modello statistico tenendo conto sia della bontà di adattamento che della complessità del modello ed è utilizzato come metodo per la valutazione e il confronto tra modelli statistici. Per tale scopo, la regola consiste banalmente nel preferire i modelli con l'AIC più basso.

Infine, per assicurarsi che non ci siano pattern particolari, si è andati a effettuare uno studio dei residui, graficandoli. Come è possibile vedere in Figura 2.4, gli errori residui sembrano essere contenuti con una media prossima allo zero, nonostante un paio di picchi, e una varianza costante e relativamente bassa, quindi ci si può ritenere soddisfatti.

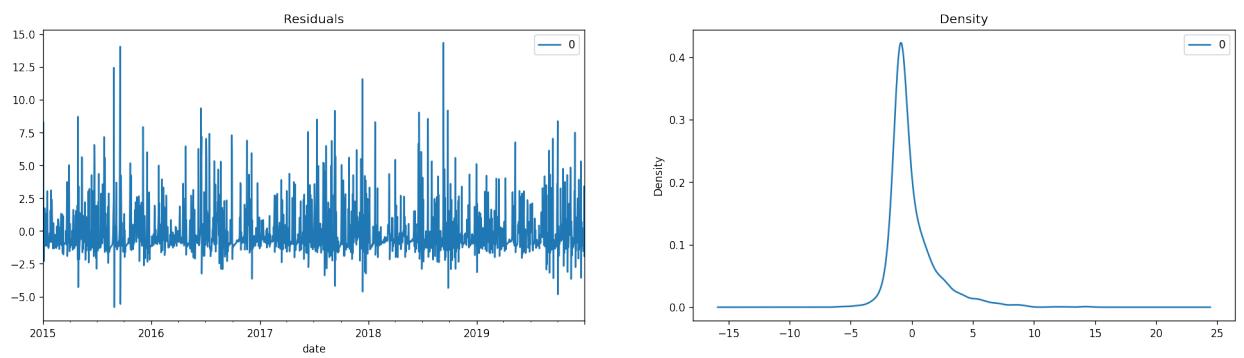


Figura 2.4: Residui

A questo punto, il modello ottenuto è stato impiegato per cercare di prevedere l'andamento delle precipitazioni in Finlandia nei due anni successivi, ossia da inizio 2020 a fine 2021. Come si può notare in Figura 2.5, dopo una risalita ripida, ma estremamente breve, la predizione si assesta su un andamento costante. A un primo sguardo questo andamento, soprattutto se messo a confronto con la serie temporale originale, potrebbe sembrare un risultato non molto impressionante e soddisfacente. Questo però è dovuto alla relativa semplicità del modello di previsione che è stato realizzato. Infatti, bisogna tenere presente che i modelli semplici di previsione spesso risultano essere estremamente competitivi.

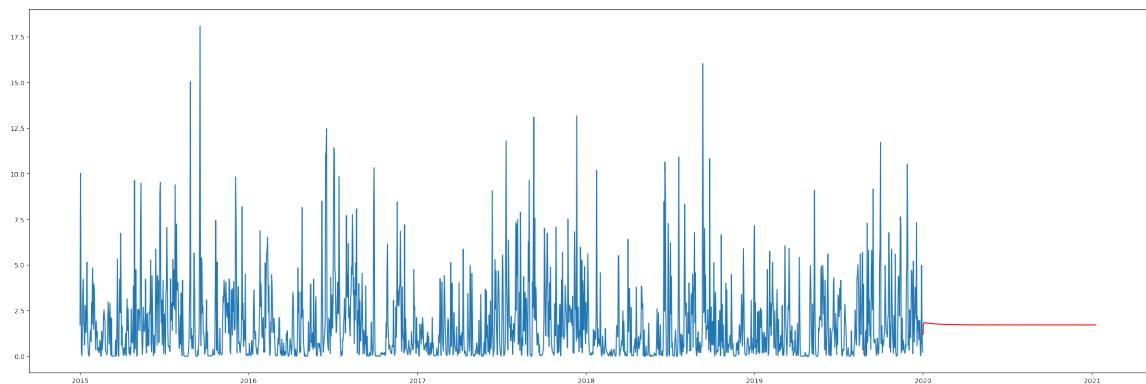


Figura 2.5: Forecasting ARIMA - Finlandia

Per valutare in maniera rigorosa il modello ottenuto, è stato necessario applicare la tecnica della *out-of-time cross validation*, che consiste nel fare alcuni passi indietro nel tempo e ricavare il valore predetto confrontandolo con il valore reale. In particolare, per questa tecnica, bisogna creare due set, uno di training e uno di testing, dividendo la serie temporale presa in considerazione in due parti contigue. Queste verranno poi utilizzate, rispettivamente, per andare ad addestrare il modello e per fare il confronto con il forecast generato a partire da tale modello. In questo caso, si è deciso di utilizzare l'85% del dataset come training set e il rimanente 15% come testing set. Riutilizzando gli stessi parametri P, D e Q utilizzati in precedenza per l'addestramento del modello ARIMA, si è proceduti ad addestrare questo nuovo modello ARIMA con il training set appena descritto, ottenendo così il risultato visibile nella figura sottostante.

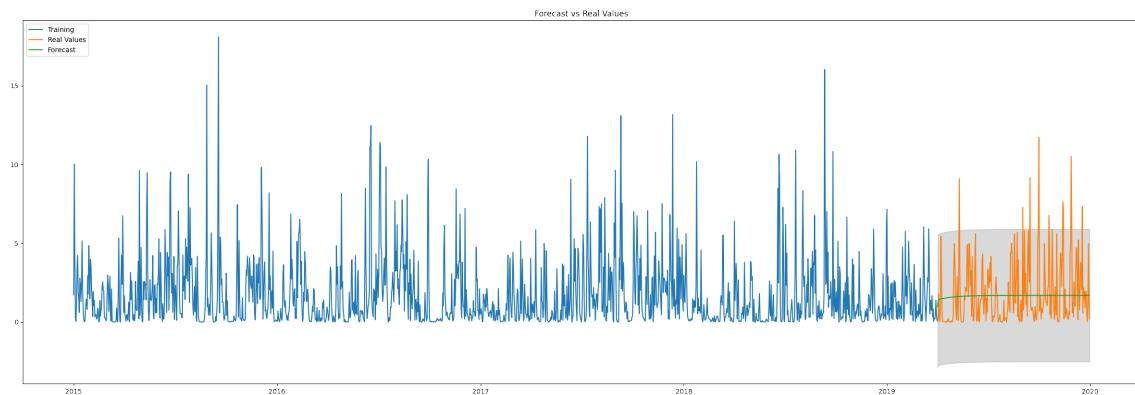


Figura 2.6: Out-of-time Cross Validation - Finlandia

Oltre questo risultato grafico, si è scelto di valutare il modello anche sulla base di alcune metriche di precisione, comunemente usate per analizzare la bontà delle previsioni. I risultati ottenuti sono descritti nella seguente tabella:

Mean Absolute Error	1.53383
Mean Absolute Percentage Error	NA
Mean Error	-0.17758
Mean Percentage Error	NA
Root Mean Squared Error	2.06052
Correlation between the Actual and the Forecast	0.17286
Min-Max Error	0.59620

La prima cosa subito ben visibile è che i valori del **MAPE** e del **MPE** siano non disponibili (*NA*). Questo non è sintomo di qualcosa di anomalo, ma semplicemente significa che tra i dati che compongono la serie temporale ce ne sono alcuni con valore pari a zero. Inoltre, è interessante notare come sia la **RMSE** che la **MAE** siano molto basse, infatti un errore assoluto medio pari a 1.53 è considerabile senza alcun dubbio un ottimo risultato. Purtroppo, però, non si può dire la stessa cosa della correlazione tra i dati predetti e quelli reali, che risulta avere un valore davvero basso, ma che conferma il risultato insoddisfacente che era possibile notare già a livello grafico. Infine, il valore estremamente basso del **ME** indica che gli errori sono presenti in entrambe le direzioni e non ci sono errori solo per eccesso o solo per difetto.

Tirando le somme finali, ci si può ritenere mediamente soddisfatti del risultato, in quanto, nonostante ci sia una bassa correlazione tra i dati predetti e quelli reali, ben visibile anche graficamente, le altre metriche di precisione restituiscono dei risultati davvero buoni.

2.2 SARIMAX

A questo punto, analogamente a quanto fatto per il modello ARIMA, si è andati ad addestrare un modello SARIMAX, per vedere se era possibile ottenere risultati migliori. Avendo in questo caso a che fare con un altro parametro (quello della stagionalità) oltre ai tre visti in precedenza, si è deciso di scegliere un modello più semplice di quello utilizzato sopra. A tale scopo, si sono utilizzati come parametri (P, D, Q, S) i valori (1, 0, 0, 12), dove **S** è il parametro che rappresenta, appunto, la stagionalità. A questo è stato assegnato un valore pari a 12 perché nella serie temporale presa in considerazione si ha una stagionalità annuale. Concluso l'addestramento, i risultati del sommario mostrano come questo modello abbia un valore di AIC leggermente superiore a quello del modello ARIMA analizzato precedentemente, mentre tutti i valori della colonna 'P>|z|' in comune ai due modelli corrispondono. Quindi abbiamo dei risultati leggermente peggiori di quelli

che avevamo nel caso precedente. Per quanto riguarda, invece, lo studio dei residui e il forecast sull'andamento delle precipitazioni in Finlandia negli anni successivi, i risultati tra i due modelli sono analoghi, come è possibile vedere nelle figure sottostanti, che sono estremamente simili a quelle ottenute in precedenza.

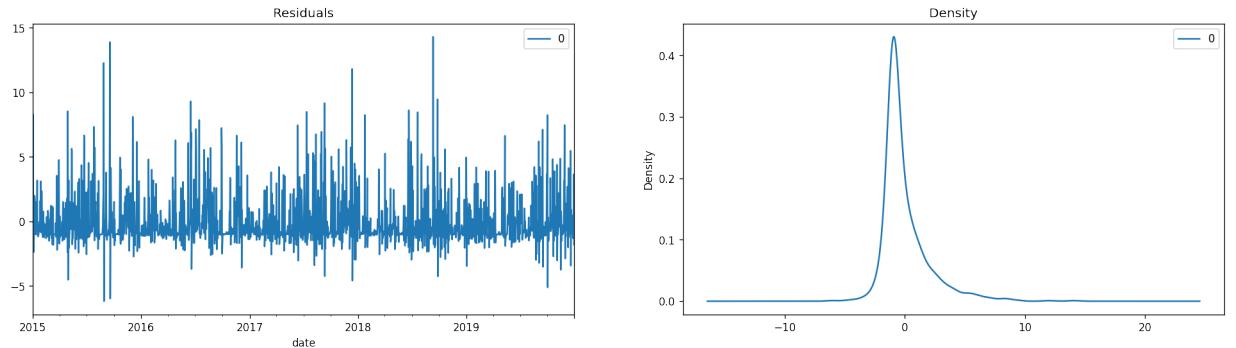


Figura 2.7: Residui

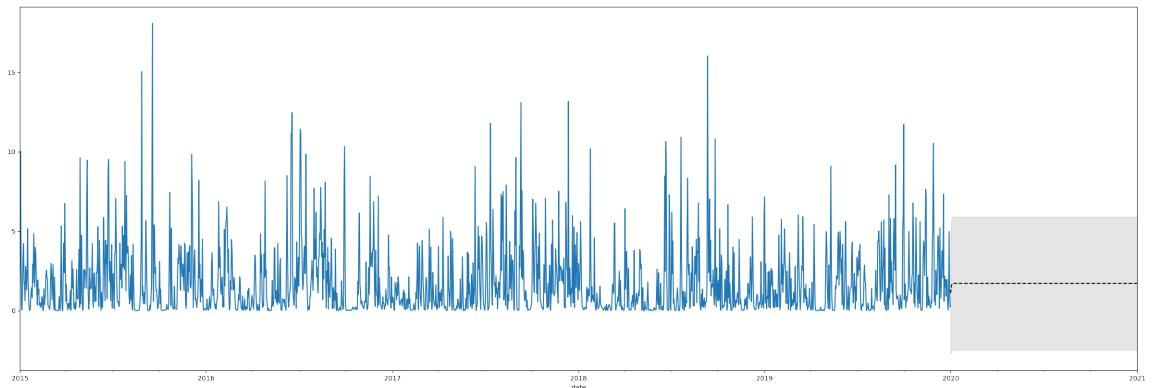


Figura 2.8: Forecasting SARIMAX - Finlandia

Capitolo 3

Svezia

La seconda analisi, invece, si concentra sulla serie temporale delle precipitazioni in Svezia. In Figura 3.1, dove è raffigurata la serie, è visibile quindi la quantità di precipitazioni, espressa in centimetri, in Svezia nel periodo che va dall'inizio del 2015 alla fine del 2019.

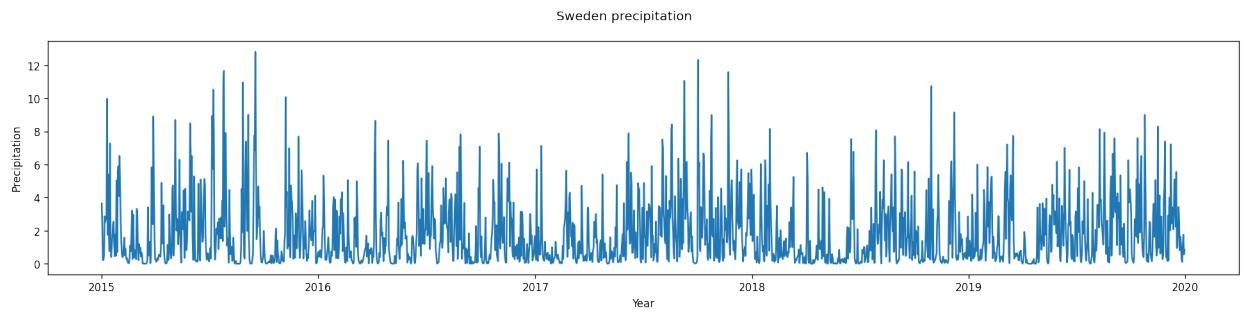


Figura 3.1: Variazione delle precipitazioni in Svezia nel tempo

Anche in questo caso, l'obiettivo è di addestrare due modelli, quello ARIMA e quello SARIMAX, iniziando dal primo, per svolgere poi delle previsioni future sui dati della serie e confrontare i risultati dei due. Quindi, si procede anche questa volta con i vari step illustrati in precedenza, partendo dallo studio della stazionarietà della serie temporale, così che sia possibile stimare il parametro di differenziazione D del modello ARIMA. Per valutare se la serie è stazionaria o meno, si è usato nuovamente l'**Augmented Dickey Fuller test** tramite la funzione `test_stationarity` implementata nel progetto. Il *p-value*, in questo caso, non è risultato essere nullo, ma pari a una quantità davvero molto piccola ($3.146373\text{e-}29$), ben al di sotto della soglia di 0.05, e quindi la serie è stazionaria. Un'ulteriore conferma di ciò viene dai grafici dell'autocorrelazione della serie e delle sue differenziazioni di primo e secondo ordine, rappresentati in Figura 3.2. Per tutti questi motivi, alla fine, la scelta del parametro D è ricaduta sul valore 0.

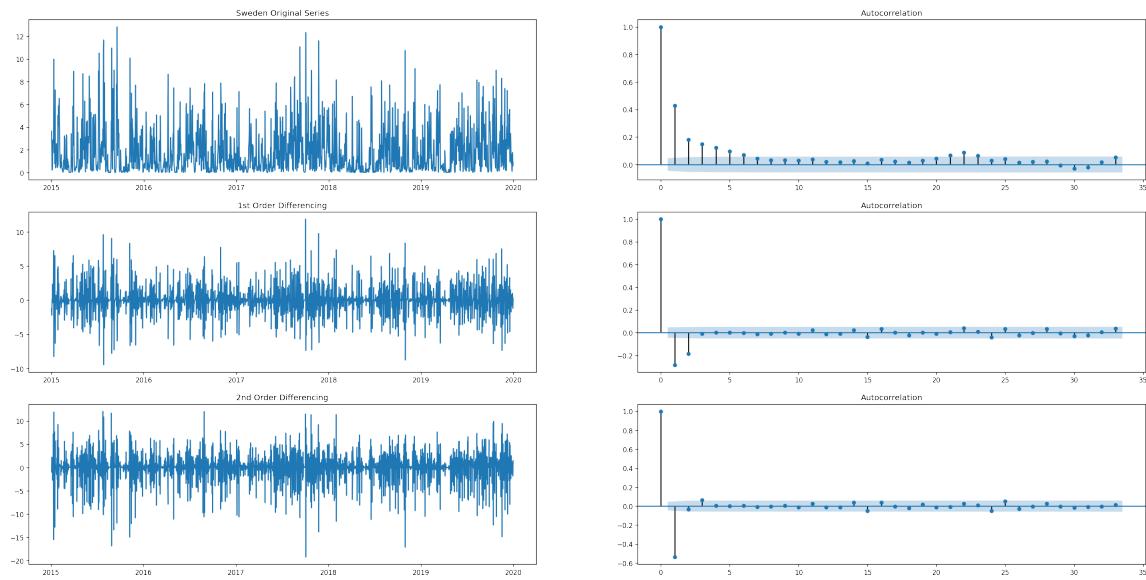


Figura 3.2: Studio Stazionarietà della serie temporale

Il secondo passo è la stima dei parametri rimanenti P e Q del modello ARIMA. Per fare ciò, bisogna utilizzare i grafici di autocorrelazione e autocorrelazione parziale della serie temporale, visibili in Figura 3.3. Analizzandoli, è possibile osservare come il grafico dell'autocorrelazione parziale abbia solo un lag al di sopra della linea significativa e per questo è possibile fissare il valore 1 al parametro P con abbastanza sicurezza. Invece, per quanto riguarda il paramentro Q , il grafico dell'autocorrelazione ha addirittura quattro lag al di sopra dell'area limite, di cui però principalmente solo uno è abbastanza significativo. Per questo motivo, al fine di essere conservativi, si è scelto di fissare inizialmente il parametro Q a 1.

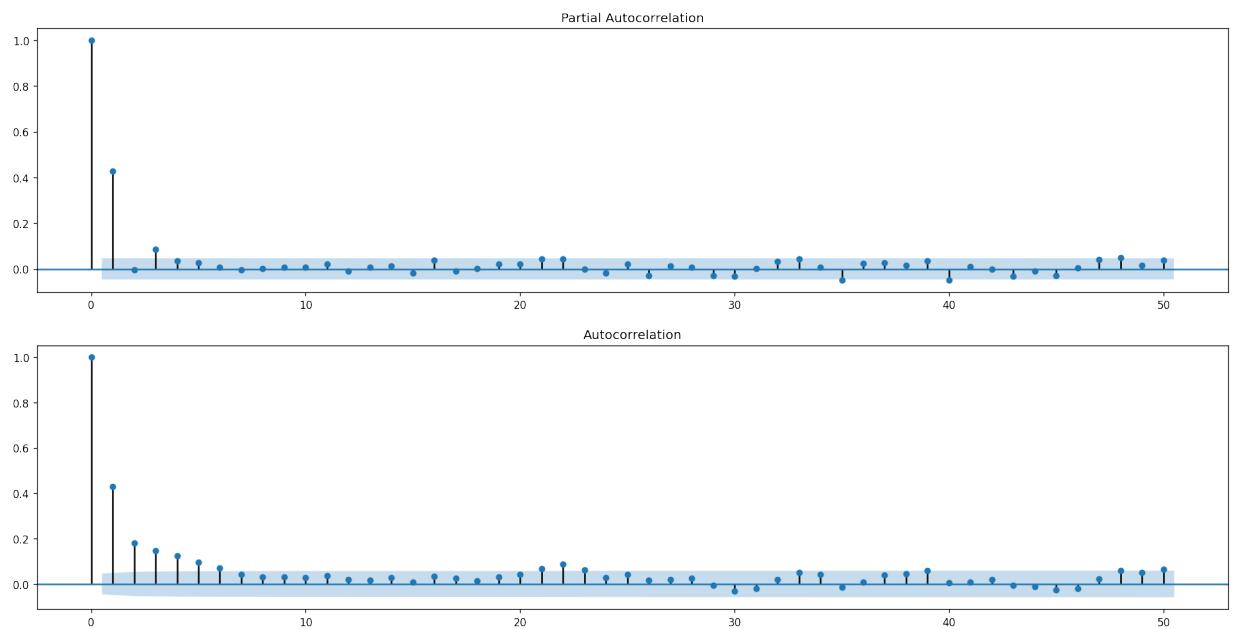


Figura 3.3: Plot dell'autocorrelazione e dell'autocorrelazione parziale della serie differenziata

3.1 ARIMA

Fissati i parametri P , D e Q si può passare all'addestramento del modello ARIMA. Inizialmente, quindi, si è andati ad allenare il modello con i valori di (P, D, Q) rispettivamente pari a $(1, 0, 1)$, però, purtroppo, con questi parametri, nei risultati del sommario emerge che il coefficiente del termine MA è vicino allo zero e il suo valore P nella colonna ' $P>|z|$ ' è molto alto (0.839) e quindi ben al di sopra del livello di significatività pari a 0.05. Questo risultato è sicuramente legato al fatto che nel grafico dell'autocorrelazione erano quattro i lag al di sopra dell'area limite e quindi bisogna aumentare il valore del parametro Q . Dopo alcuni esperimenti, si è trovato che il valore ottimale del parametro Q che è un buon compromesso tra complessità del modello e buoni risultati è 2. Quindi, si è riaddestrato il modello ARIMA con i valori di (P, D, Q) rispettivamente pari a $(1, 0, 2)$ e così facendo si è ottenuto un modello con un AIC nettamente minore di quello del precedente e che avesse tutti i valori della colonna ' $P>|z|$ ' molto significativi ($\ll 0.05$).

Inoltre, anche in questo caso, per essere sicuri che non ci siano pattern particolari, si è andati a rappresentare graficamente i residui, così da poterli studiare. In Figura 3.4, possiamo quindi vedere gli errori residui. Si può notare che la media è circa pari a zero, a parte qualche picco sporadico, e la varianza è uniforme, quindi ci si può ritenere soddisfatti.

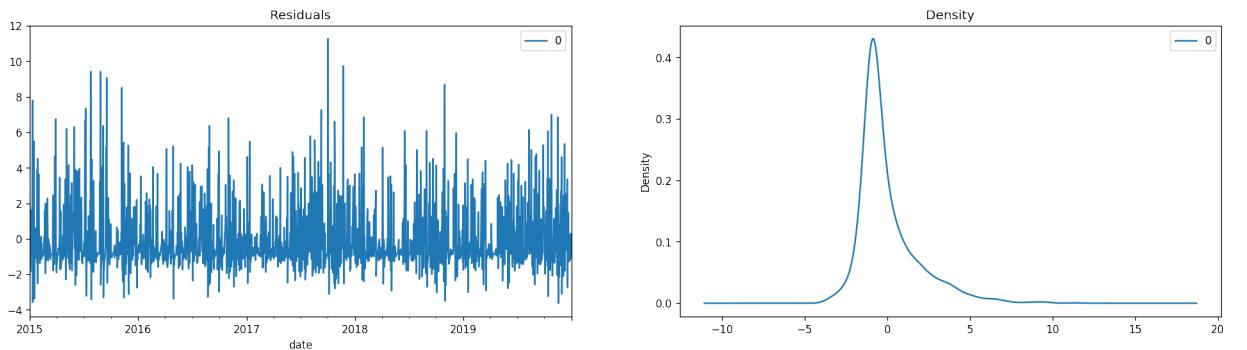


Figura 3.4: Residui

Il modello così ottenuto è stato poi utilizzato per effettuare delle previsioni sull'andamento delle precipitazioni in Svezia negli anni a venire, in particolare per tutto il 2020 e il 2021. Come è possibile vedere in Figura 3.5, dopo una breve risalita della predizione, questa si mantiene su un livello stabile. Anche in questo caso la previsione potrebbe non sembrare particolarmente impressionante, in quanto a parte la breve risalita iniziale è praticamente una linea retta. Tutto ciò è dovuto nuovamente alla semplicità del modello ARIMA addestrato e quindi si può considerare un risultato accettabile.

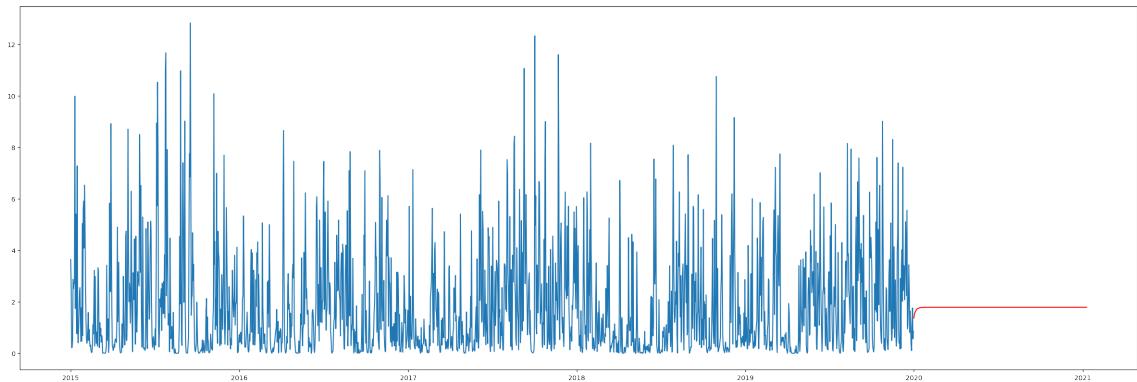


Figura 3.5: Forecasting ARIMA - Svezia

A questo punto, come nell'analisi per la Finlandia, si è andati ad applicare la tecnica della *out-of-time cross validation* per valutare in maniera rigorosa il modello usato. Anche in questo caso, l'85% del dataset è stato utilizzato come training set, mentre il rimanente 15% come testing set. Dopo questa suddivisione dei dati, si è addestrato un nuovo modello ARIMA con gli stessi valori (1, 0, 2) dei parametri (P, D, Q), ma che ha come input il training set appena definito, ottenendo così la previsione in Figura 3.6.

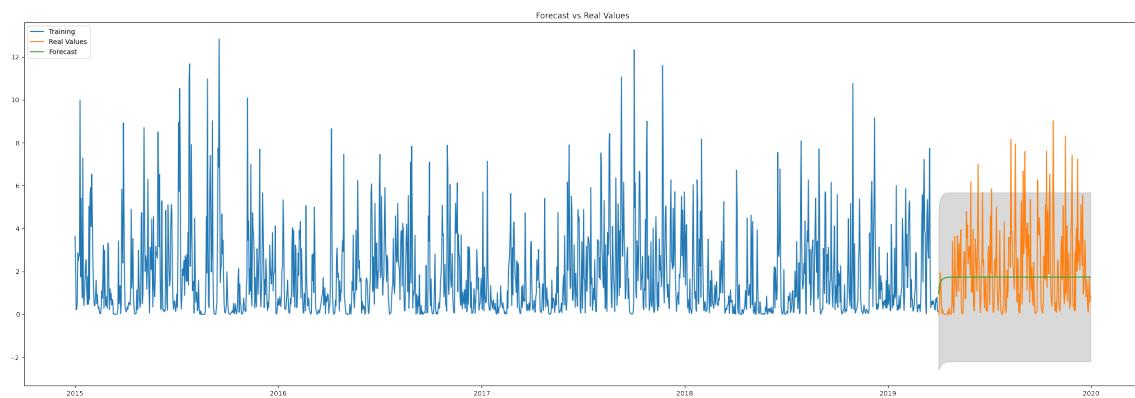


Figura 3.6: Out of time cross validation - Svezia

Come già fatto in precedenza, oltre al risultato grafico, si è scelto di valutare il modello anche sulla base di alcune metriche di precisione, di cui possiamo vedere i valori nella tabella sottostante.

Mean Absolute Error	1.50886
Mean Absolute Percentage Error	NA
Mean Error	-0.32537
Mean Percentage Error	NA
Root Mean Squared Error	1.94581
Correlation between the Actual and the Forecast	0.13812
Min-Max Error	0.55748

Anche in questo caso, i valori del **MAPE** e del **MPE** sono non disponibili (*NA*), indice della presenza di alcuni valori pari a zero tra i dati che compongono la serie temporale. Inoltre, si hanno nuovamente valori molto bassi della **RMSE** e della **MAE**, che sono considerabili degli ottimi risultati. Purtroppo, però, ancora una volta, si ha un valore estremamente basso della correlazione tra i dati predetti e i dati reali, che non è affatto un buon risultato e che era già visibile in realtà dalla Figura 3.6.

Nonostante tutto, nel complesso, ci si può ritenere mediamente soddisfatti di questi risultati, perché anche se si ha una bassa correlazione tra i dati predetti e quelli reali, i risultati davvero buoni delle altre metriche vanno a compensare il bilancio finale.

3.2 SARIMAX

Come fatto anche per l'analisi della Finlandia, dopo il modello ARIMA si è andati ad addestrare un modello SARIMAX, per confrontare poi i risultati dei due. Dopo una serie di test, si sono scelti come parametri P, D e Q gli stessi del modello ARIMA addestrato in precedenza, aggiungendo però il parametro S della stagionalità e ponendolo uguale a 12, visto che la serie temporale considerata ha una stagionalità annuale. Una volta concluso l'addestramento, i risultati del sommario hanno mostrato una situazione davvero interessante e degna di nota. Infatti, sia il valore di AIC che tutti i valori della colonna ' $P>|z|$ ', così come la quasi totalità dei valori in comune tra i due modelli, sono uguali. Questo ci potrebbe portare ad ipotizzare che se un modello ARIMA e un modello SARIMAX vengono addestrati con gli stessi dati e gli stessi parametri P, D e Q, i risultati del sommario per quanto riguarda i campi in comune saranno in gran parte (se non addirittura totalmente) uguali. Questa ipotesi è ulteriormente avvalorata andando ad analizzare i grafici sottostanti, sullo studio dei residui e sul forecast dell'andamento delle precipitazioni in Svezia nei due anni a venire, che risultano estremamente simili, se non identici, a quelli ottenuti in precedenza per il modello ARIMA.

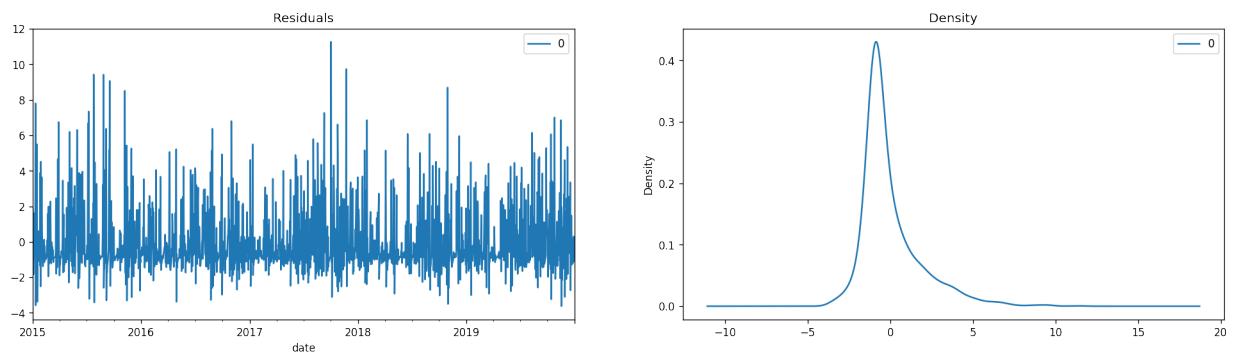


Figura 3.7: Residui

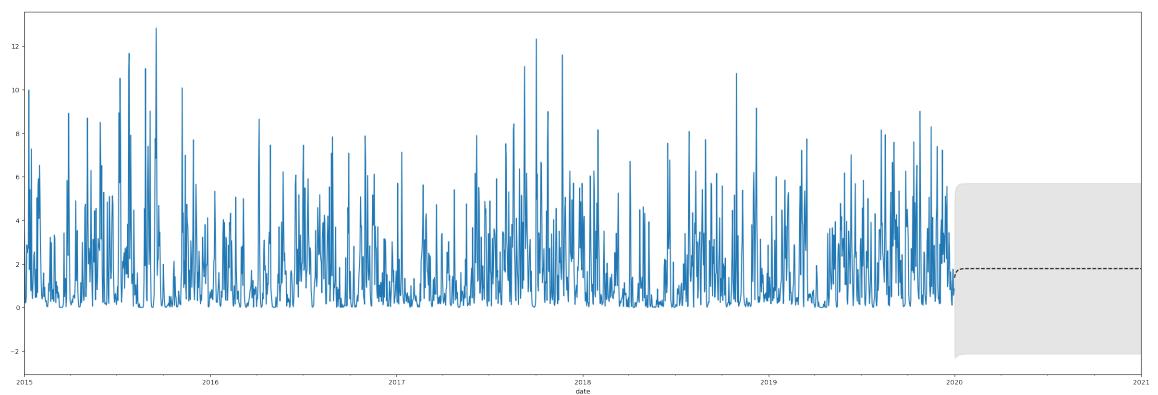


Figura 3.8: Forecasting SARIMAX - Svezia

Capitolo 4

Norvegia

Infine, la terza ed ultima analisi si è concentrata sulla serie temporale delle precipitazioni in Norvegia, che è visibile in Figura 4.1 e che rappresenta la quantità di precipitazioni, espressa in centimetri, in questa nazione nel periodo preso in considerazione.

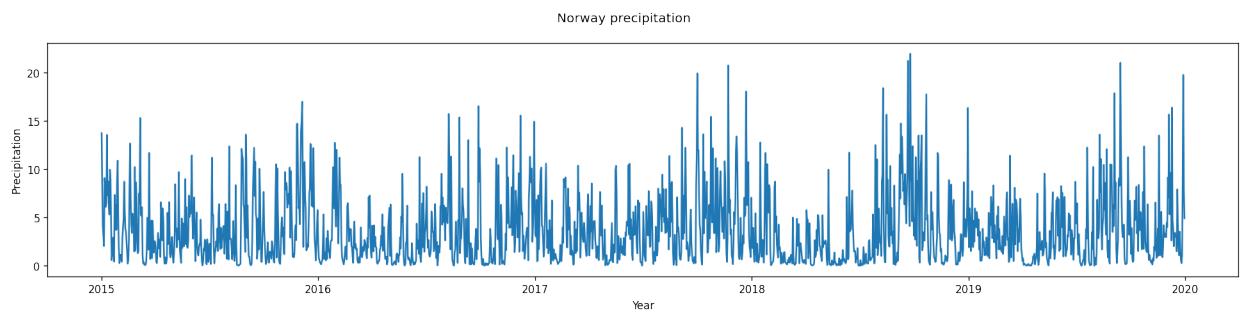


Figura 4.1: Variazione delle precipitazioni in Norvegia nel tempo

Anche in questo caso, si sono addestrati due modelli, uno ARIMA e uno SARIMAX, iniziando dal primo, per svolgere poi delle previsioni future sui dati della serie e confrontarne i risultati. Quindi, si procede nuovamente con i vari passi già utilizzati nelle due analisi precedenti. Il primo step è lo studio della stazionarietà della serie temporale per poter stimare il parametro di differenziazione D del modello ARIMA. Per valutare se la serie è stazionaria o meno, si è usato ancora una volta l'**Augmented Dickey Fuller test** tramite la funzione `test_stationarity` implementata nel progetto. Il *p-value*, anche in questo caso, è risultato essere pari a una quantità davvero molto piccola (4.159407e-28), ben al di sotto della soglia di 0.05, e quindi si può dedurre che la serie sia stazionaria. Si può avere un'ulteriore conferma di ciò analizzando i grafici dell'autocorrelazione della serie e delle sue differenziazioni di primo e secondo ordine, rappresentati in Figura 4.2. Quindi, anche questa volta, al parametro D è stato assegnato il valore 0.

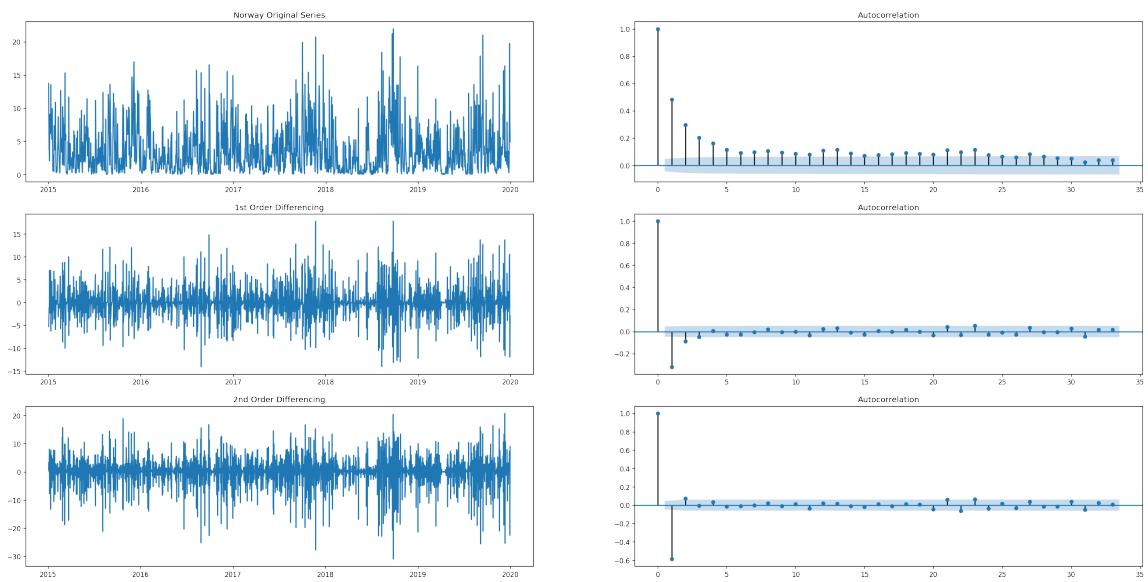


Figura 4.2: Studio Stazionarietà della serie temporale

A questo punto, nel secondo step, si procede a stimare i due parametri rimanenti P e Q del modello ARIMA, utilizzando i grafici di autocorrelazione e autocorrelazione parziale della serie temporale, presenti in Figura 4.3. A partire da questi, infatti, è possibile notare come il grafico dell'autocorrelazione parziale abbia solo un lag al di sopra dell'area limite, mentre quello dell'autocorrelazione ne ha diversi, di cui tre in particolare sono abbastanza elevati. Da queste osservazioni, si è deciso di assegnare al parametro P il valore 1 con abbastanza sicurezza; mentre, per quanto riguarda il parametro Q , la scelta è stata di fissarlo inizialmente a 3 e svolgere successivamente delle prove anche con altri valori, per vedere quale portasse ad avere risultati migliori.

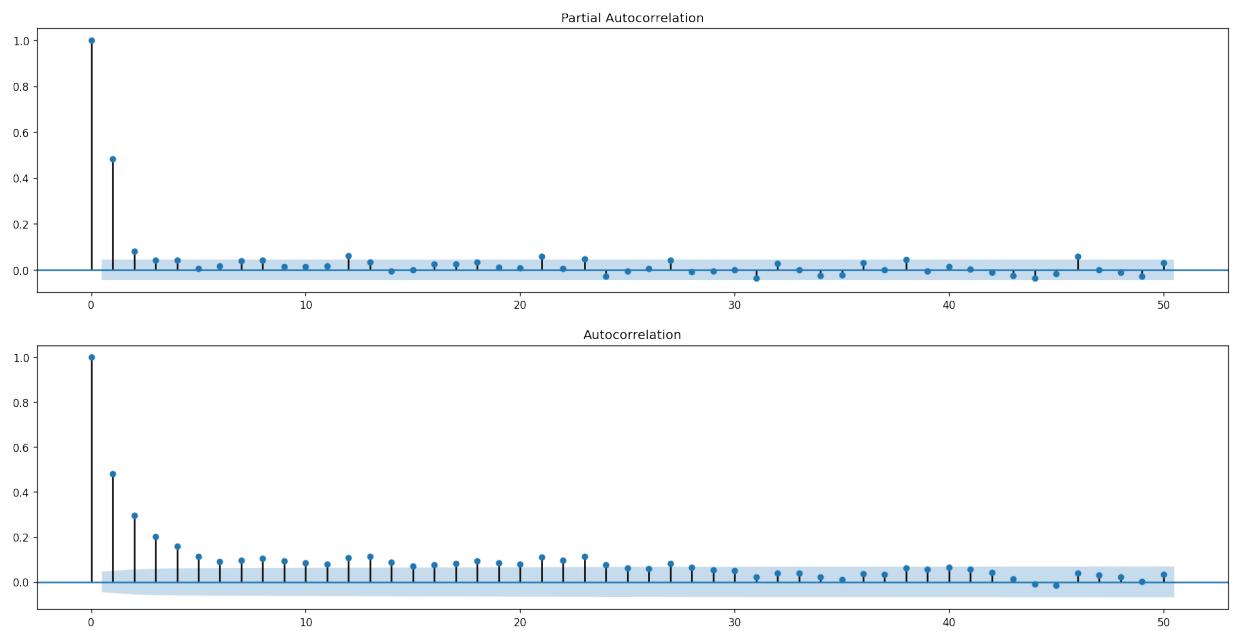


Figura 4.3: Plot dell'autocorrelazione e dell'autocorrelazione parziale della serie differenziata

4.1 ARIMA

Dopo aver fissato i valori dei parametri P, D e Q si può passare all’addestramento del modello ARIMA. Per prima cosa si è addestrato il modello con i parametri scelti di P, D e Q pari rispettivamente a 1, 0 e 3 ottenendo dei buoni valori sulla colonna ‘P>|z|’; infatti tutti i valori sono risultati molto significativi ($\ll 0.05$). Come anticipato, però, avendo visto nel grafico dell’autocorrelazione che il parametro Q aveva diversi lag al di sopra della linea significativa, si sono effettuate anche altre prove con altri valori di Q per vedere se ci fossero modelli più semplici che portassero a risultati migliori o comunque accettabili. Ciò non si è verificato avendo tali modelli valori di AIC peggiori di quello del modello con il parametro Q pari a 3, così come, in alcuni casi, anche valori più alti del livello di significatività (0.05) nella colonna ‘P>|z|’.

Successivamente, come nelle due analisi precedenti, si è andati a rappresentare graficamente i residui per garantire che non ci fossero pattern particolari. Come si può notare in Figura 4.4, gli errori residui sembrano essere contenuti con una media all’incirca pari a zero, nonostante alcuni picchi e le oscillazioni più marcate rispetto ai casi precedenti. Inoltre, la varianza è abbastanza uniforme e relativamente bassa, quindi ci si può ritenere soddisfatti.

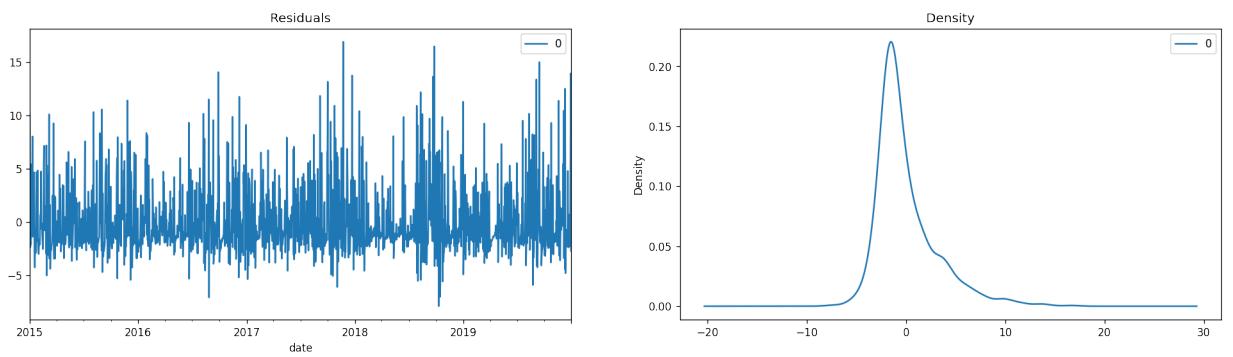


Figura 4.4: Residui

Una volta addestrato il modello, questo è stato impiegato per andare a prevedere l’andamento delle precipitazioni in Norvegia nei due anni successivi al periodo considerato nel dataset, ossia per gli anni 2020 e 2021. In Figura 4.5 si può notare come, dopo una piccola e ripida risalita che culmina in un picco, la predizione tende a decrescere per un breve tratto, fino ad assestarsi su un andamento costante. Ovviamente, anche in questo caso, valgono le considerazioni fatte nelle due analisi precedenti sui risultati del forecast.

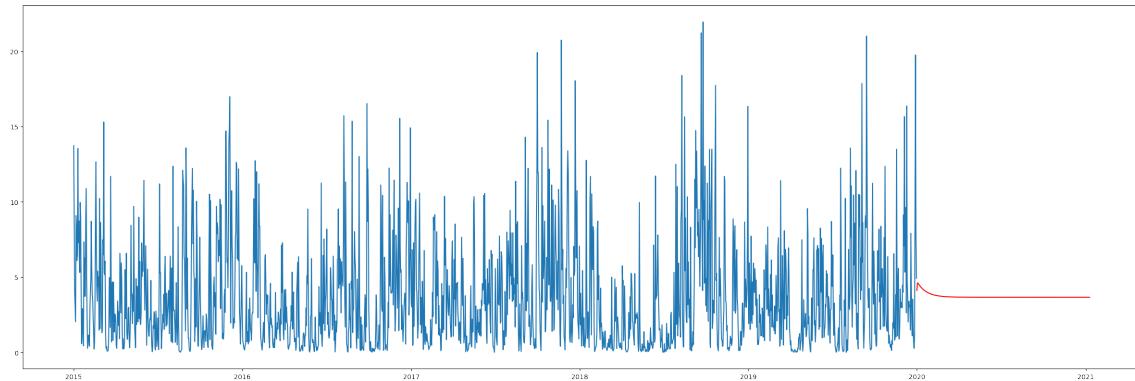


Figura 4.5: Forecasting ARIMA - Norvegia

Anche in questo caso, come per le analisi della Finlandia e della Svezia, si è scelto di utilizzare la tecnica della *out-of-time cross validation* per valutare in maniera rigorosa il modello ottenuto. Ancora una volta si è deciso di suddividere i dati utilizzando l’85% del dataset di partenza come training set e il restante 15% come testing set. Fatto ciò, si è proceduto a riutilizzare gli stessi valori (1, 0, 3) dei parametri (P, D, Q) per addestrare un nuovo modello ARIMA, con in input, però, il training set appena descritto, ottenendo così la previsione visibile nella figura sottostante.

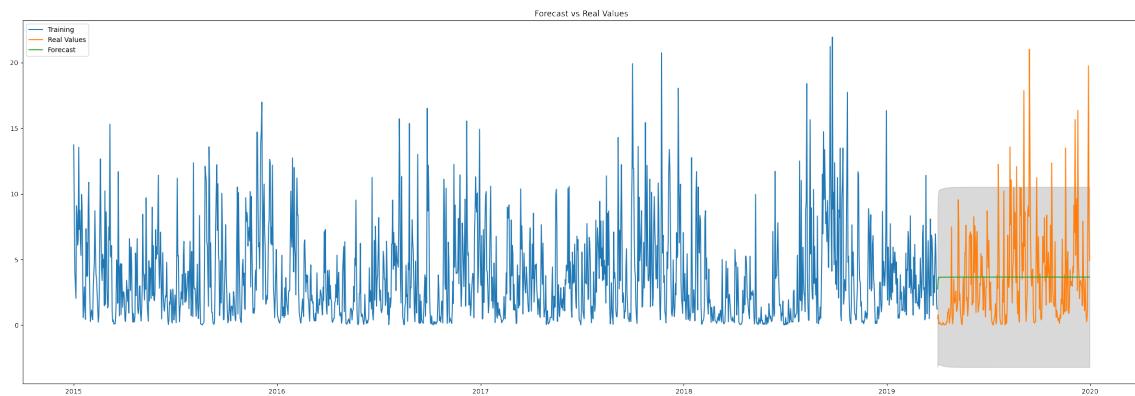


Figura 4.6: Out of time cross validation - Norvegia

Infine, per concludere l'analisi di questo modello ARIMA, oltre al risultato grafico della *out-of-time cross validation*, si è deciso di valutare il modello anche sulla base di alcune metriche di precisione, elencate con i rispetti valori nella tabella seguente.

Mean Absolute Error	2.78425
Mean Absolute Percentage Error	9.25786
Mean Error	0.04694
Mean Percentage Error	8.95342
Root Mean Squared Error	3.67153
Correlation between the Actual and the Forecast	0.07048
Min-Max Error	0.54009

In questo caso, a differenza delle analisi precedenti, i valori del **MAPE** e del **MPE** sono disponibili. Questo significa che nessuno dei valori che compongono la serie temporale è pari o prossimo allo zero. Però, analizzando le varie metriche, purtroppo è abbastanza palese come queste siano peggiori di quelle che erano state ottenute nei modelli ARIMA sviluppati per i casi della Finlandia e della Svezia. Infatti, oltre ad avere un valore estremamente basso della correlazione tra i dati predetti e quelli reali, anche le altre metriche, per quanto accettabili, non sono molto buone.

Quindi, tenendo in considerazione queste metriche e il risultato grafico ottenuto in Figura 4.6, si può concludere che il modello è comunque accettabile, anche se non ci si può ritenere particolarmente soddisfatti. Per tale motivo, si potrebbe eventualmente procedere in futuro ad addestrare ulteriori modelli, non necessaria-

mente ARIMA, per cercare di ottenere dei risultati ancora migliori nelle previsioni di questa serie temporale.

4.2 SARIMAX

Anche in questo caso, dopo il modello ARIMA si è passati ad allenare un modello SARIMAX, per confrontare i risultati dei due e soprattutto per vedere se era possibile ottenere dei risultati migliori di quelli poco soddisfacenti appena visti. Dopo alcune prove, si è deciso nuovamente di scegliere come parametri P, D e Q gli stessi del modello ARIMA considerato in questa analisi, aggiungendo però, ovviamente, il parametro S della stagionalità con valore pari a 12, essendo questa una serie temporale con stagionalità annuale. Una volta concluso l'addestramento, i risultati del sommario hanno mostrato la stessa situazione interessante e degna di nota che era già emersa nell'analisi della Svezia. Infatti, anche questa volta, allenando il modello SARIMAX con gli stessi parametri P, D e Q del modello ARIMA abbiamo ottenuto gli stessi valori in tutti quei campi in comune tra i due. Questa cosa avvalorava quindi ulteriormente l'ipotesi fatta in precedenza, durante l'analisi della serie temporale della Svezia. Inoltre, come ennesima conferma di ciò, si ha che i grafici sullo studio dei residui e sul forecast dell'andamento delle precipitazioni in Norvegia negli anni successivi per il modello SARIMAX risultano praticamente identici a quelli ottenuti per il modello ARIMA.

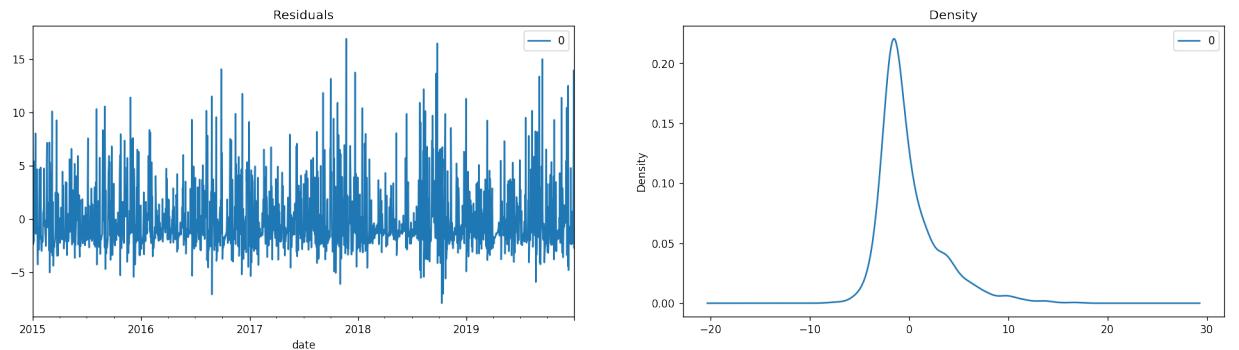


Figura 4.7: Residui

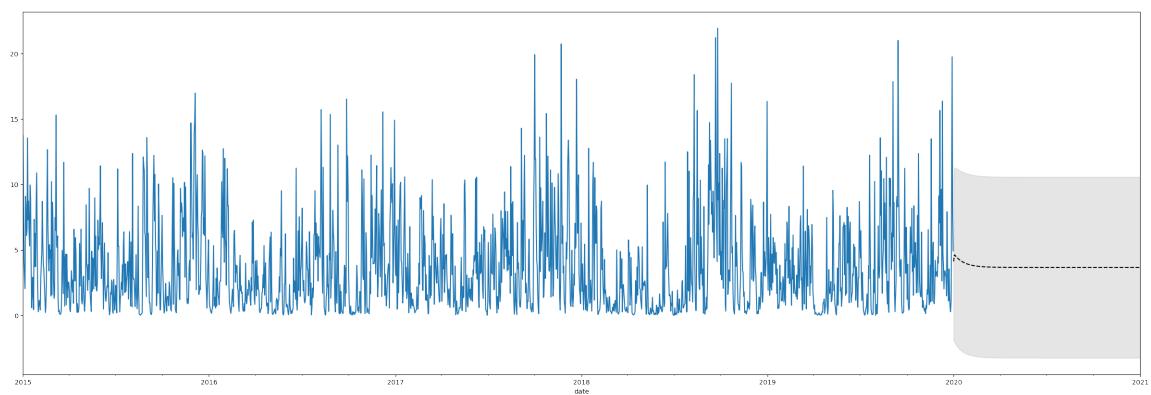


Figura 4.8: Forecasting SARIMAX - Norvegia