LUISS

Dipartimento di Impresa e Management

Corso di laurea triennale in Economia e Management

Cattedra di Statistica Applicata ed Econometria

Modelli ARIMA per la previsione della domanda turistica in Europa

Candidato

Marco Antonio Mirone

Matricola 280131

Relatore

Prof. Mauro Costantini

Anno Accademico 2024/2025

Indice

Introduzione

L'elaborato si propone di mostrare un'analisi previsionale pseudo out of sample condotta dall'autore sulla domanda turistica di tre Paesi europei: Italia, Francia e Spagna. La scelta della statistica con cui rappresentare la domanda è ricaduta su quella della serie storica degli arrivi turistici fornita da Eurostat, in particolare sul numero mensile di check-in nelle strutture di tipo alberghiero e simili.

A seguito di un breve capitolo introduttivo che ambisce ad informare il lettore su pregi e limiti dell'effettuare previsioni sul settore turistico, nel secondo capitolo sono spiegati i dati e i metodi con il quale sono state analizzate le serie storiche in esame. Prima della costruzione dei modelli previsionali, sono state necessarie una correzione logaritmica e una differenziazione di dodicesimo grado.

Il terzo capitolo illustra il processo di scelta dei modelli per un training set esteso dal 1994 all'inizio della pandemia da Covid-19. Le previsioni un passo avanti sono state effettuate con dei modelli fissi fino alla fine del 2020.

Con il modello previsionale ARIMA(p, d, q) come punto di partenza, per ogni Paese sono stati selezionati tre modelli differenti con l'ausilio di tre criteri di decisione diversi: il modello suggerito dall'algoritmo "forecast" del prof. Rob J. Hyndman; un modello selezionato con l'obiettivo di minimizzare il BIC tra degli AR(p) di diverso ordine; un modello selezionato dall'autore a seguito dell'interpretazione delle funzioni di autocorrelazione ed autocorrelazione parziale.

I modelli sono stati confrontati sulla base di due misure di bontà dell'adattamento, eleggendone due migliori per Paese -uno per errore- ed interpretando i risultati. I valori degli errori hanno fornito indicazioni interessanti, seppur dissimili tra loro a causa delle diverse caratteristiche delle tre serie storiche.

Capitolo primo

Serie storiche del turismo

1.1 I flussi turistici e la loro previsione

Passati ormai sei anni dal primo caso registrato di COVID-19, è difficile individuare un aspetto della nostra vita su cui la pandemia non abbia avuto un effetto rilevante. Le conseguenze del lockdown e delle restrizioni hanno "contagiato" la maggioranza dei settori economici, in alcuni casi gonfiandone i numeri ed in altri deprimendoli ai minimi storici.

La fattispecie del settore turistico è sicuramente una delle più interessanti: su di esso l'impatto delle limitazioni alla mobilità è stato inizialmente devastante, per vedere poi i flussi turistici lentamente riprendersi e addirittura toccare vette impronosticabili in precedenza. La crescita avvenuta negli ultimi anni è figlia di diversi fattori, tra cui l'accumulo di maggiore risparmio da parte degli individui; il notevole incremento del turismo *social*; e motivazioni di carattere psicosociale dovute al ritorno della possibilità di viaggiare.

La crescita appena descritta è dimostrata dai dati. L'Organizzazione Mondiale del Turismo (OMT o UNTWO) afferma che il turismo internazionale ha già virtualmente ripreso i livelli pre-pandemici, superando nel 2024 i numeri del 2019. In termini di variazione percentuale, gli 1,4 miliardi di turisti dell'ultimo anno hanno contribuito ad un incremento del 11% rispetto al 2023.

Italia, Spagna e Francia sono tre Paesi che sono stati coinvolti nel fenomeno di ripresa appena citato. Si tratta di Paesi simili tra loro per molti aspetti e li accomuna sicuramente una forza attrattiva molto forte ed efficace nei confronti dei turisti. Sono in grado di attrarre numerosi visitatori grazie alla loro Storia e fascino, e godono di un'ottima reputazione come mete per una vacanza. Tutti e tre hanno temperature e paesaggi variegati, una posizione geograficamente favorevole e al loro interno grandi città così come località più rurali e solitarie.

Questa loro similarità è anche il motivo per il quale sono i tre Paesi presi in esame per il lavoro illustrato in questa tesi.

Entrando più nel dettaglio, la relazione Istat sul turismo italiano nel 2023 ha evidenziato una crescita significativa delle presenze (notti trascorse dai turisti nelle strutture ricettive), oltre la media dei Paesi membri dell'Unione del 1,7%. Ciò ha portato l'Italia ad essere la terza in Europa per presenza turistica, proprio dietro Spagna e Francia, che delle tre è stata quella col più

alto numero di turisti residenti (si veda la figura 1.1). Lo stesso report indica l'area nord-est del Paese come quella più visitata con oltre un terzo delle notti totali, ed il Veneto la Regione con maggiore affluenza, complice probabilmente il turismo di massa verso la città di Venezia. L'occupazione del settore nel medesimo anno ha superato i livelli pre pandemici di quasi due punti percentuali.

I dati più recenti hanno mostrato una flessione delle presenze nel terzo trimestre del 2024: sempre Istat ha rilevato un calo dell'1,4% rispetto a dodici mesi prima. Il fenomeno pare per ora rimanere circoscritto all'attività dei soli turisti residenti in Italia: il flusso estero ha registrato lievi diminuzioni negli arrivi, ma un incremento del numero di notti passate in strutture turistiche.

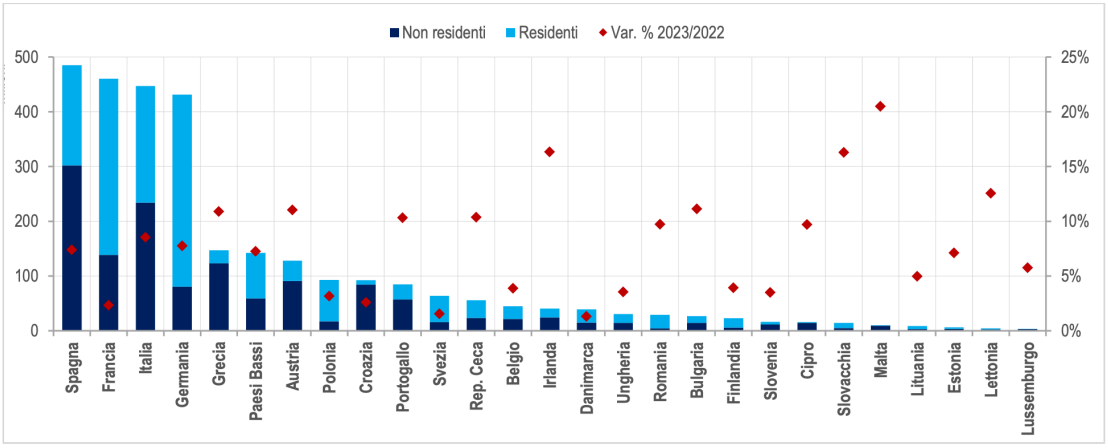

Figura 1.1: Presenze negli esercizi ricettivi per Paese UE di destinazione e residenza clienti (Istat)

In Spagna, il Ministero dell'industria e del turismo si occupa di pubblicare periodicamente gli aggiornamenti sui dati turistici. Al contrario dell'Italia, al terzo trimestre del 2024 è corrisposto un andamento delle presenze del 1,4% superiore a quello dell'anno precedente. Il 2023, a sua volta, era stato un'altro anno da record per il Paese.

L'Istituto nazionale di statistica e degli studi economici francese (Insee) ha reso noto che le presenze turistiche nel proprio Paese sono diminuite nel terzo trimestre del 2024 -precisamente del 1,7%- e che il quarto semestre ha invece sorpassato la vetta dell'anno precedente del 3,2%.

Questa serie di dati si traduce in un andamento crescente ma non lineare, suggerendo come il trend del settore sia mitigato da altre tendenze o fattori, tra cui la stagionalità delle attività. La stagionalità dei dati, approfondita al capitolo secondo, ha riflessi statisticamente rilevanti sui

numeri. È di particolare importanza che, affinché le previsioni in ambito turistico siano precise ed efficaci, siano presi in considerazione questi aspetti.

Per far fronte ai futuri movimenti dei flussi turistici diviene essenziale effettuarne delle previsioni corrette. Il *forecasting* degli arrivi turistici è un punto di partenza fondamentale per prendere decisioni di tipo economico. Ai *policy maker* giovano accurati resoconti sui trend e sul potenziale futuro andamento della domanda turistica. C'è evidentemente una stretta interdipendenza tra politica e settore turistico, soprattutto in un Paese con i connotati dell'Italia, sia a livello nazionale che a livello locale: nel Paese il contributo al PIL del settore turistico è in crescita, posizionandolo al quarto posto nell'Unione Europea con una percentuale superiore al 6% (dati Eurostat, 2024). Gli altri due Paesi presi in esame in quest'elaborato, Spagna e Francia, occupano rispettivamente in questa classifica il terzo (col 6,9%) e il decimo (4%) posto. Senza scomodare altre statistiche, queste singole percentuali rappresentano un peso specifico non da poco sulla produzione di reddito nazionale.
Eventuali interventi indirizzati al settore turistico, riguardino essi l'emanazione di atti legislativi o provvedimenti di altra natura, possono assumere direzioni diverse proprio a seconda dello stato della domanda turistica e delle previsioni sul suo futuro. Legislatore ed amministrazioni pubbliche, ad esempio, potrebbero essere interessati ad arginare una situazione di *overtourism* -vedasi il caso delle città d'arte- qualora questa si faccia portatrice di malfunzionamenti o danni all'ambiente. Viceversa, previsioni troppo negative possono fungere da campanello d'allarme per un'offerta turistica da rinnovare.
Conoscere le aspettative sui potenziali arrivi ha anche utilità pratiche per tutti i lavoratori del settore che devono programmare costi, progetti ed utilizzo di risorse con largo anticipo. Tra di loro non si annovera solamente il management delle strutture d'accoglienza (hotel, b&b ecc.), ma anche aziende di trasporti e compagnie aeree e ferroviarie, oltre che musei ed altre attrazioni. Nel mondo aziendale, le imprese sanno di dover operare in un'ottica sostenibile nel futuro e monitorano costantemente gli obiettivi di lungo termine: per i business che hanno un particolare interesse verso lo stato della domanda turistica di determinati Paesi, lavorare con questo tipo di previsioni rappresenta il pane quotidiano.
Le conseguenze reali dell'andamento dei flussi turistici giocano sicuramente un ruolo chiave nell'economia di una città e della sua vivibilità; così come nei rendimenti delle attività commerciali, che spesso interagiscono con clientele straniere.

La grande disponibilità di dati che contraddistingue la nostra epoca rende forse più difficile individuare quelli adatti a descrivere correttamente un fenomeno. Si fa certamente fatica a scegliere una sola statistica in grado di inquadrare lo stato del turismo. In precedenza sono stati menzionati gli "arrivi", ovvero il numero di check-in effettuati nelle strutture ricettive di una determinata area, che compongono tra l'altro le serie storiche utilizzate ai fini dello studio nei capitoli successivi. Le "presenze", d'altro canto, corrispondono al numero di notti passate nelle medesime strutture, tenendo conto dunque della lunghezza dei soggiorni ma non avendone una contezza del numero.

È chiaro che lavorare con misure quantitative potrebbe portare a tralasciare alcune sfumature. Per esempio, è noto che non tutte le notti trascorse in strutture alberghiere o simili avvengano per scopo turistico, in quanto molti viaggiano per motivi familiari o lavorativi. Si potrebbe dibattere, però, che questa rappresenta comunque una forma di turismo, anche se non "vacanziero", in quanto la presenza di un individuo in una città che non è la sua lo porta comunque ad interagire economicamente con i medesimi soggetti che si interfacciano con i turisti tradizionali.

A seconda della tipologia di strutture che si considerano, poi, si va incontro all'esclusione di una fetta di turismo. Solitamente in studi di questo tipo ci si limita al numero di check-in in alberghi e simili, contingentando al di fuori della domanda le interazioni con altri luoghi di pernottamento, come i campeggi.

Altri fenomeni turistici impattanti sono purtroppo impossibili da registrare con una statistica di questo tipo. Per esempio, è sempre più popolare effettuare un viaggio in singolo giorno, anche se verso destinazioni lontane. A causa dei crescenti prezzi di alberghi e appartamenti in affitto in alcune città, oggi alcuni preferiscono prendere un volo aereo economico nelle prime ore del giorno, per poi visitare per meno di ventiquattro ore la destinazione e rientrare con l'ultimo aereo disponibile. Questa soluzione viene adottata specialmente da turisti europei verso capitali e grandi città, e ha contribuito alla nascita turismo di massa verso queste mete.

Le città note per la vita notturna sono anch'esse bersaglio di questo genere di viaggio. Sui social è nato un cosiddetto *trend*, quindi un comportamento che viene imitato da più utenti in serie per poi essere condiviso online, che consiste nel trascorrere una notte per le discoteche di alcune città come Barcelona, Palma de Maiorca o Gallipoli, per poi rientrare la mattina dopo; il tutto, ovviamente, senza effettuare alcun check-in in strutture turistiche. Questa tipologia di turismo, fortemente impattante su vita ed economia delle città prese di mira, non può essere registrata dalla statistica delle presenze o da quella degli arrivi.

Infine, va menzionata anche l'ascesa di piattaforme come AirB&B e simili. Non tutte le strutture che si trovano online sono state correttamente inserite nei i registri degli enti territoriali ed è ormai assodato che esista effettivamente un "mercato nero" del turismo, non riflesso nei dati ufficiali ma con effetti reali sulla direzione dei flussi di persone e sulla qualità della vita delle città. Ad oggi, questo è uno dei fenomeni su cui si sta concentrando maggiormente lo sforzo della regolamentazione.

Tenendo in considerazione tuti questi aspetti, è possibile affermare l'impossibilità di inquadrare perfettamente la domanda turistica in termini quantitativi. I dati sono una fonte d'informazione molto utile e inconfutabile, che si scontra però con le difficoltà di conversione in numero di un fenomeno complesso. Come visto prima, è anche l'interpretazione degli esperti del settore a fare la differenza nella previsione dei flussi: si pensi alle grosse divergenze che possono emergere tra uno studio condotto basandosi sugli arrivi e un'altro basato sul numero di presenze.

Nonostante ciò, è oltremodo essenziale ormai sostenere le decisioni strategiche -aziendali o politiche- con delle fondamenta numeriche. Se ben effettuate e in assenza di eventi shockanti come una pandemia, le previsioni dei flussi turistici costituiscono una grande fonte d'informazione per chi vuole operare con un approccio data-driven.

1.2 Letteratura

Introducendo la letteratura sul tema turistico, è necessario un breve inquadramento teorico.

Le serie storiche, o serie temporali, sono dati raccolti per una singola unità a diversi punti nel tempo (Stock e Watson, 2020). Oltre l'analisi dell'andamento storico di queste serie, si è detto come in ambito economico e politico sia fondamentale studiarne la previsione. I metodi di previsione possono essere generalmente divisi in: *judgemental*, univariati e multivariati (Chatfield, 2000). La prima tipologia include i metodi che si avvalgono di intuizioni ed interpretazioni soggettive degli analisti, oltre a tutta una serie di informazioni esterne. La seconda, invece, effettua previsioni sulla semplice base dei valori presenti e passati in relazione allo scorrere del tempo. Col terzo tipo di analisi previsionale, infine, i risultati del *forecasting* dipendono almeno in parte da una diversa variabile protratta nel tempo: è il caso della stima degli effetti causali dinamici.

Un'ulteriore divisione riguarda i dati con i quali si stima il modello. Con la tipologia utilizzata in questo studio, si ricorre alla divisione del dataset in due parti: *training set*, su cui si effettua l'analisi previsionale, e *test set*, confrontata coi risultati. Questa tecnica è chiamata "pseudo

previsione fuori campione" (o *pseudo out of sample*). Al contrario, nel caso della previsione *in sample* la stima del modello avviene sull'intera serie, per poi prendere solo la prima parte dei dati per prevedere le osservazioni successive. Quando si attende che siano effettivamente disponibili i dati previsti, per poi confrontarli coi propri risultati, si parla di previsione "fuori campione" (*out of sample*).
La problematica principale consiste nella verifica della precisione ed efficienza di un modello. Esistono dei parametri come i criteri d'informazione o misure come l'errore di previsione, che sono stati studiati appositamente. Solitamente si verifica l'efficienza del modello con stime in sample o pseudo out of sample, per poi applicarlo alle previsioni out of sample nel caso di risultati positivi.

Tornando al fenomeno turistico, si è visto come esso sia influenzato da numerose variabili qualitative e quantitative, dunque complicato da inquadrare in maniera veritiera. Per ciò, la letteratura è divisa sulla scelta dei modelli di previsione da applicare alla domanda turistica. Inoltre, la grandissima disponibilità di dati ha aperto nuove frontiere rispetto a pochi anni fa e stimolato l'applicazione di tecniche inedite. Gli studiosi hanno la possibilità di scegliere tra diverse misure, o crearne di nuove, per descrivere al meglio il reale andamento dei flussi. L'ampiezza e il focus geografico delle ricerche è anch'essa differente di caso in caso.
Song e Li (2008) hanno raccolto ed esaminato centoventuno studi di tre diversi approcci: il primo contemplante l'uso modelli di serie storiche, il secondo di altri modelli econometrici ed il terzo di modelli emergenti spesso integrati con l'IA. La comparazione tra questi tre metodi non è stata in grado di eleggerne uno sempre migliore degli altri, ma la loro efficienza e la misura degli errori sono necessariamente legati al contesto e alle caratteristiche delle variabili esaminate. È quindi corretto affermare che, per quanto riguarda le previsioni della domanda turistica, vi è molta varietà che contribuisce ad arricchire e stimolare la ricerca.
È presente una letteratura che si avvale di modelli più semplici, basando le previsioni sull'andamento delle serie storiche senza coinvolgere variabili esterne. Una delle tipologie di modello più comune, infatti, è quella degli *ARMA-based*, ovvero autoregressivi a media mobile, che saranno approfonditi tecnicamente al capitolo secondo. Un'applicazione efficace è stata effettuata da Chu (2009), in uno studio previsionale degli arrivi verso i principali Paesi dell'estremo oriente. In particolare, la ricerca utilizza e paragona tre modelli: ARIMA, che è una generalizzazione del modello ARMA; ARAR, che più volte si è dimostrato performante nelle previsioni di lungo termine; e ARIFMA, anch'esso derivato da un ARMA. Quest'ultimo ha otte-

nuto i migliori risultati nella previsione out-of-sample degli arrivi sia mensili che trimestrali, dimostrando maggiore accuratezza.

In altri casi cronologicamente più vicini, gli *ARMA-based* hanno rappresentato il punto di partenza per lo sviluppo di tecniche maggiormente sofisticate.

Ad esempio, appena iniziata la ripresa post-covid, Liu, Vici, Ramos et al. (2021) hanno effettuato un *forecasting* sulle serie storiche trimestrali degli arrivi turistici in venti destinazioni. Trattandosi di una previsione *judgemental scenario-based*, sono stati appositamente creati degli indici statistici con l'aiuto del giudizio di alcuni esperti per riflettere diversi potenziali scenari legati alla pandemia. Nello scenario *mild*, cioè il più fedele a quello effettivamente verificatosi, era stata correttamente pronosticata una piena e veloce ripresa per i Paesi con i flussi in entrata provenienti da mete meno lontane.

L'ARIMA è anche fondamentale nel lavoro di He e Qian (2025), che hanno utilizzato il modello STL-XGBoost sui dati degli arrivi all'aeroporto di Macau. Lo studio ha prodotto delle previsioni out-of-sample molto precise e rappresenta una nuova frontiera della ricerca.

Bufalo e Orlando (2024), invece, hanno condotto uno studio sulle presenze turistiche in Italia, ed hanno effettuato le previsioni mediante il modello CIR#.

Non è da escludere, come anticipato, un approccio differente al tema. Gunter et al. (2024), per esempio, hanno provato a misurare e prevedere la domanda europea con la statistica delle importazioni turistiche reali per Paese. Affiancando al valore delle importazioni quello della crescita economica di ogni Paese, è stato ottenuto un dataset di tipo panel, analizzato poi da un modello FMOLS corretto con una variabile *dummy* per l'anno della pandemia. Anche in questo caso la previsione è stata effettuata per due scenari differenti, fino all'anno 2025. Per quanto riguarda i Paesi presi in esame da questa tesi, era stata prevista una crescita sopra la media europea solamente per la Spagna.

Quest'elaborato si propone di dare il proprio contributo confrontando tre diverse analisi di previsione pseudo out-of-sample sulla domanda turistica pre-covid italiana, francese e spagnola: la prima tramite modelli ARIMA suggeriti dall'algoritmo sviluppato dal prof. Rob J. Hyndman e incluso nel pacchetto "forecasting"; la seconda con dei modelli autoregressivi ottenuti dall'interpretazione dei grafici delle funzioni di autocorrelazione ed autocorrelazione parziale; la terza con dei modelli autoregressivi sulla base della minimizzazione del criterio informativo di Schwartz. Per il confronto tra modelli è stata utilizzata la misura del RMSFE.

Capitolo secondo

Dati e modelli ARIMA

2.1 Dati

Come anticipato, lo studio di quest'elaborato parte dalla statistica mensile degli arrivi turistici in Italia, Francia e Spagna da Gennaio 1993 a Dicembre 2024, misurati come numero di check-in nelle strutture alberghiere e similari. I valori mancanti, corrispondenti al mese di Aprile 2020 per la Spagna e Dicembre 2020 per la Francia, sono stati ottenuti imputando la media delle osservazioni adiacenti. Le serie storiche sono fornite da Eurostat, mentre per l'analisi dei dati è stato utilizzato il software RStudio.

Di seguito, la rappresentazione grafica delle tre distribuzioni:

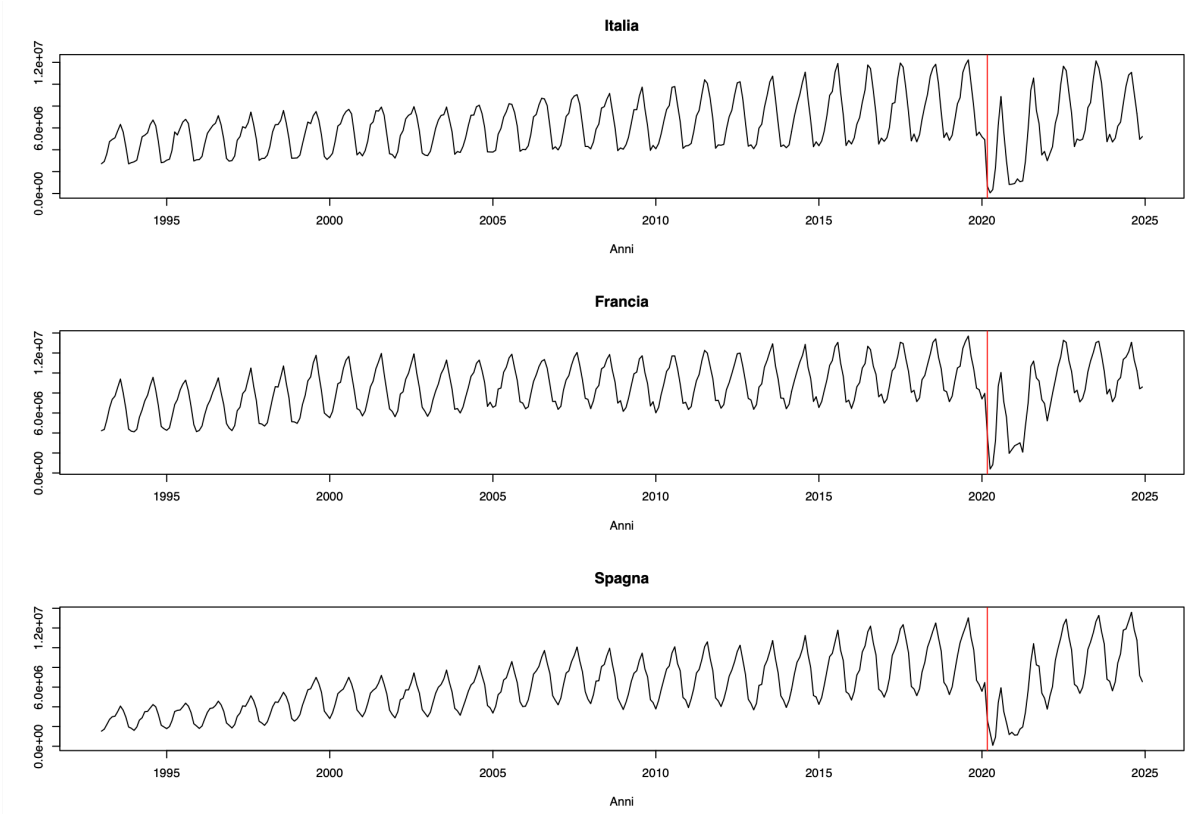

Figura 2.1 Arrivi turistici dal 1993 al 2024 (linea rossa: Marzo 2020, inizio Covid)

In tutti e tre i casi, l'analisi del grafico è utile per giungere a diverse conclusioni.

Innanzitutto, si noti la struttura "spinosa" delle distribuzioni, che raggiungono picchi opposti - uno negativo ed uno positivo- ad intervalli regolari. Questo genere di andamento è causato da una forte stagionalità delle osservazioni, derivante dalla natura del dato raccolto e dal feno-

meno che rappresenta. Il tema della stagionalità è ripreso nei paragrafi successivi del capitolo, che illustrano i metodi con cui è stata eliminata ai fini dell'analisi.
Secondariamente, si riscontra la presenza di un trend. Un trend è il persistente movimento di lungo periodo di una variabile nel corso del tempo (Stock e Watson, 2020), e può assumere una direzione crescente o decrescente. La teoria distingue tra due tipologie. Quando il trend è una funzione non aleatoria del tempo -per esempio, di proporzionalità diretta- prende il nome di "deterministico", altrimenti di trend "stocastico": è il caso delle serie in esame. La presenza di stagionalità rende leggermente meno immediato l'impatto visivo, ma ci si rende conto agevolmente dell'andamento crescente della misura per tutti e tre i Paesi.
La terza caratteristica visibile ad occhio nudo sui grafici è la presenza di un break strutturale, segnalato dalla linea verticale rossa. Con questo termine si intende un cambiamento netto, o graduale evoluzione nel tempo, dei coefficienti autoregressivi (v. sezione 2.2) della serie storica (Stock e Watson, 2020). Come immaginabile, nel caso dei flussi turistici si manifesta un break dopo il mese di Marzo del 2020, quando tutti e tre Paesi hanno imposto restrizioni alla mobilità per contrastare il diffondersi del COVID-19.

Entrando più nel dettaglio rispetto ai grafici precedenti, le serie possono essere analizzate in seconda istanza con l'aiuto di semplici statistiche descrittive.

Tabella 2.1 Statistiche descrittive della domanda turistica dal 1993 al 2024

	Italia	Francia	Spagna
Minimo	61.774	400.752	98.561
Massimo	12.221.964	13.692.822	13.586.095
Media	6.159.840	8.490.944	6.137.381
Scostamento quadratico medio	2.395.619	2.430.992	2.833.465
Asimmetria	0,35	-0,25	0,45
Curtosi	-0,2	-0,15	-0,51

La tabella 2.1 mostra il minimo, il massimo, la media, lo scostamento quadratico medio, l'asimmetria e la curtosi delle serie. Nonostante forniscano una visione meno completa, questi dati sono utili per comprendere al primo impatto almeno gli ordini di grandezza con cui si sta lavorando e alcune caratteristiche della distribuzione.

I dati corrispondenti al valore minimo, come facilmente immaginabile, risalgono ai mesi di Aprile e Maggio 2020. I dati sui valori massimi -registrati tutti ad agosto, ma solo per la Spagna nel periodo post-covid- confermano la recente esplosione del settore turistico, rappresentando numeri impensabili solo pochi anni fa. In questo contesto media e scostamento quadratico hanno un'importanza relativa, dovuta alla presenza della componente stagionale.
L'asimmetria (*skewness* o momento terzo) di una distribuzione indica quanto questa si allontani dalla simmetria. Le distribuzioni delle osservazioni di Italia e Spagna hanno una skewness maggiore di zero, dunque sono asimmetriche verso destra e presentano più osservazioni maggiori della media. La serie francese, al contrario, è asimmetrica verso sinistra.
L'ultima misura, la curtosi (momento quarto), riguarda la massa delle code di una distribuzione, quindi quanto della sua variabilità è determinata dalla presenza di *outlier*. Tutte e tre le serie hanno curtosi minore di 3, che è il valore del momento quarto di una distribuzione normale. Questo ci permette di definirle "platicurtiche", o a code "leggere", ed effettivamente non vi è presenza di valori anomali se non a seguito del break strutturale.

Per studiare queste serie storiche è necessario applicare dei correttivi. Spesso in statistica si preferisce lavorare con il logaritmo naturale delle osservazioni perché in grado di rendere i dati più simmetrici e limitare l'impatto degli outlier smussando la serie. Vista la presenza di un break e le caratteristiche del dato trattato, per questo studio si è ricorsi a tale correzione logaritmica.
Ai fini dell'analisi, il secondo passo è calcolare la funzione di autocorrelazione. Con il termine "autocorrelazione" -o "correlazione seriale"- tecnicamente si intende la correlazione di una serie coi propri valori ritardati (Stock e Watson, 2020). In generale, la correlazione tra due variabili misura la dipendenza tra loro, ed è data dal rapporto tra la loro covarianza e il prodotto delle deviazioni standard. Dovendo calcolare la dipendenza di una variabile con i suoi stessi valori nel passato, nel caso dell'autocorrelazione la formula tramuta.
L'autocorrelazione di una serie storica Y si ottiene dunque dall'equazione:

$$\widehat{\rho_j} = \frac{\widehat{cov(Y_t, Y_{t-j})}}{\widehat{var(Y_t)}} \tag{1}$$

L'equazione (1) mostra come calcolare l'autocorrelazione di j-esimo coefficiente. Il coefficiente della correlazione seriale indica con quale ritardo della variabile Y_t misurare la dipen-

denza. La prima autocorrelazione è la correlazione tra i valori di Y in tempi adiacenti, la seconda autocorrelazione è quella tra Y_t e Y_{t-2} e così via. La funzione di autocorrelazione (AFC) raccoglie l'autocorrelazione di tutti i coefficienti.

L'autocorrelazione rappresenta il primo indizio per verificare la stazionarietà e la stagionalità di una serie storica. A titolo esemplificativo, si riporta di seguito la funzione di autocorrelazione del logaritmo delle serie in esame:

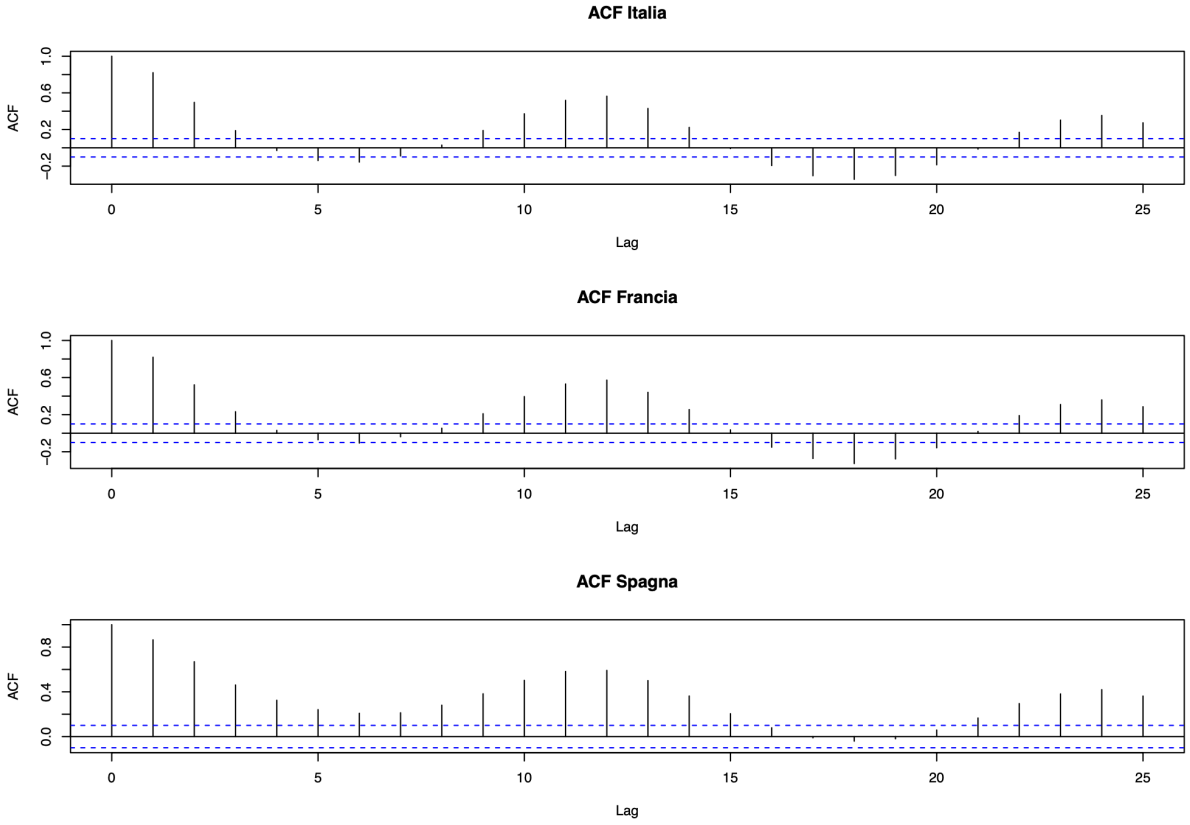

Figura 2.2 ACF fino al venticinquesimo coefficiente del logaritmo degli arrivi turistici

Un'autocorrelazione come nella figura sopra, molto alta a diversi coefficienti, è sintomo di non stazionarietà della serie. Una serie temporale si dice stazionaria se la sua distribuzione congiunta non cambia nel corso del tempo. In altre parole, se una serie è stazionaria la sua media e la sua varianza non dipendono dal tempo e sono costanti. Casi come quello del trend e del break strutturale sono esempi di non stazionarietà, essendo condizioni che implicano un effetto della dimensione temporale sull'andamento della serie.

Osservare la funzione di autocorrelazione è altresì utile in caso di stagionalità. Con questa si intende una componente della serie che si ripete ad intervalli regolari ogni anno, con variazioni di intensità più o meno analoga nello stesso periodo (mese, nel caso in esame) di anni successivi e di intensità diversa nel corso dello stesso anno. La stagionalità è intuibile, oltre che

dal grafico mostrato all'inizio del capitolo, dalla rappresentazione grafica della AFC. Se il valore dell'autocorrelazione tocca le vette più alte in corrispondenza dei coefficienti multipli di dodici, fornisce un chiaro indizio sulla presenza di una componente stagionale. La figura 2.2 riporta chiaramente i picchi stagionali di cui sopra.
Le funzioni di autocorrelazione delle serie temporali di Italia, Francia e Spagna presentano tutte le caratteristiche appena elencate, anche se con intensità diverse e proprie peculiarità.

Ai fini dell'applicazione dei modelli ARMA e dell'analisi di previsione, è necessario mitigare la connotazione stagionale della serie storica e renderla stazionaria. In presenza di stagionalità mensile, si può utilizzare il metodo della differenza dodicesima.
Quando si applica la differenza dodicesima per eliminare la stagionalità mensile di una serie storica, nella pratica si svolge l'analisi su un'altra serie -la differenza dodicesima, appunto- ottenuta dall'equazione:

$$\Delta^{12} Y_t = Y_t - Y_{t-12} \tag{2}$$

La nuova serie ottenuta è composta da dodici osservazioni in meno, in quanto le prime dodici della *time serie* originale non hanno un valore ritardato al dodicesimo ordine. A seguito della correzione con differenza dodicesima, la serie da sottoporre ad analisi ha quindi 372 osservazioni.
La rappresentazione grafica dell'andamento delle differenze dodicesime è diversa per tutti e tre i Paesi, non lasciando tracce di non stazionarietà: i valori prima del break oscillano tutti intorno a una media costante. La funzione di autocorrelazione, inoltre, mantiene per tutte e tre le serie dei connotati stagionali ma a dei livelli molto più bassi ed ammissibili ai fini dello studio, come si vedrà più avanti.

La verifica numerica della stazionarietà può essere effettuata con un Augmented Dickey-Fuller (ADF) test. Anche chiamato "test di radice unitaria", è stato effettuato sul logaritmo di tutte e tre le serie con il medesimo risultato: restituendo un p-value minore del 1%, il test ha rifiutato l'ipotesi nulla di non stazionarietà. Per Italia, Francia e Spagna le statistiche test sono rispettivamente: -4,2261 (p-value < 0,01); -46146 (< 0,01); -4,235 (< 0,01). Le statistiche sono inferiori al valore critico del test con trend temporale -3,96 e ricadono quindi nella regione di rifiuto.

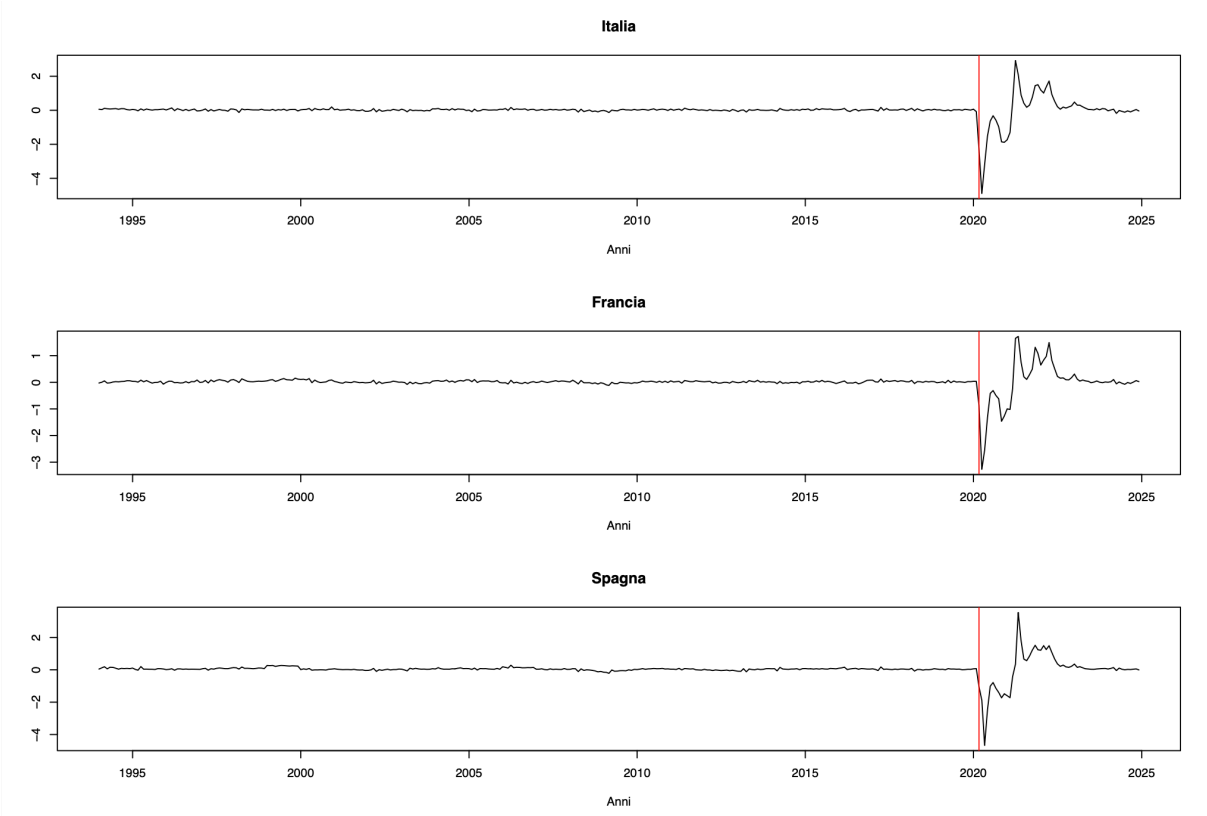

Figura 2.3 Differenza dodicesima del logaritmo degli arrivi turistici dal 1994 al 2024 (linea rossa: Marzo 2020, inizio Covid)

2.2 Modelli ARIMA

Come era stato anticipato al capitolo primo, uno dei modelli utilizzati per lo studio è quello ARIMA (*Autoregressive Integrated Moving Average model*). Il modello ARIMA è il risultato della simbiosi tra un modello autoregressivo AR(p) e un modello a media mobile MA(q), integrato di ordine d. Il modello è anche una generalizzazione del più semplice ARMA, non altro che un ARIMA(p, d, q) con integrazione $d = 0$.

Passando in rassegna le componenti dell'ARIMA, il modello AR(p) è uno dei più semplici nel campo dell'analisi delle serie storiche. Un'autoregressione esprime la media condizionata di una serie storica in funzione dei suoi ritardi (Stock e Watson, 2020). Il modello autoregressivo di ordine p utilizza p ritardi nell'esprimere tale aspettativa condizionata. Pertanto, il valore di Y_t è espresso come funzione lineare dei suoi ritardi (v. Stock e Watson, 2020, p. 452):

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + u_t \tag{3}$$

Nell'equazione (3) i coefficienti $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ sono stimati con il metodo OLS, ovvero il metodo dei minimi quadrati, cioè è lo stesso approccio per il calcolo dei coefficienti di una retta di regressione lineare tra due variabili. Le variabili casuali dei ritardi sono da considerarsi quindi diverse da Y_t e sono a tutti gli effetti dei regressori di quest'ultima. Lo stesso modello può essere riscritto con una notazione diversa (Stock e Watson, 2020, p. 483), che sarà utile più tardi per comprendere meglio il modello ARIMA:

$$a(L)Y_t = \beta_0 + u_t \tag{4}$$

dove $a(L)$ prende il nome di "polinomio nei ritardi" e si configura nella seguente maniera:

$$a(L) = a_0 + a_1 L + a_2 L^2 + \ldots + a_p L^p = \sum_{j=0}^{p} a_j L^j \tag{5}$$

L'equazione (4) ha la stesso significato della (3). L'operatore ritardo (o lag) L esprime il valore ritardato di una variabile casuale quando affiancatogli: $LY_t = Y_{t-1}$. Applicando l'operatore più di una volta è possibile ottenere ritardi oltre il primo ordine, come $L^2 Y_t = Y_{t-2}$; $L^3 Y_t = Y_{t-3}$ e così via. I valori di a_j sono i valori dei coefficienti che nella formula (3) affiancano Y_t e i suoi valori ritardati, quindi: $a_0 = 1$ e $a_j = -\beta_j$.

Così come la regressione lineare può essere utilizzata per supporre il valore della variabile dipendente per un determinato valore dei suoi regressori, il modello AR(p) può essere usato a scopo previsionale "un passo avanti" utilizzando ogni variabile in luogo del loro primo ritardo (Stock e Watson, 2020, p. 452):

$$Y_{T+1|T} = \beta_0 + \beta_1 Y_T + \beta_2 Y_{T-1} + \ldots + \beta_p Y_{T-p+1} \tag{6}$$

Il secondo step per la costruzione di un ARIMA è il passaggio per il modello ARMA(p, q). Questo modello integra un AR(p) di un ulteriore modello a media mobile MA(q), ipotizzando il termine d'errore u_t del modello come una media mobile di un altro termine d'errore inosservato. L'errore del AR(p) si può riscrivere così (Stock e Watson, 2020, p. 483):

$$u_t = b(L)e_t \tag{7}$$

con $b(L)$ polinomio di grado q con $b_0 = 1$, ed e_t una variabile aleatoria *white noise* serialmente incorrelata e non osservata. La versione estesa dell'equazione (7) illustra meglio questo concetto:

$$u_t = b_0 e_t + b_1 e_{t-1} + b_2 e_{t-2} + \ldots + b_q e_{t-q} \tag{8}$$

Un modello ARMA(p, q) si compone allora nel seguente modo:

$$a(L)Y_t = \beta_0 + b(L)e_t \tag{9}$$

Il modello ARMA viene utilizzato per analisi più approfondite di serie storiche rispetto a quelle condotte con un semplice autoregressivo. Questo avviene perché la struttura dell'ARMA riesce a captare due fattori in contemporanea: la componente autoregressiva per mezzo della parte AR, e le tendenze uniformate dei dati tramite la parte MA.

È già stato accennato come l'ARIMA sia in realtà un'estensione generale del modello ARMA. Si utilizza infatti un ARIMA(p, d, q) nel momento in cui è necessario svolgere una differenza di ordine d sulle serie storiche per renderle stazionarie. Non è un caso che la stazionarietà sia una delle condizioni necessarie per la stima dei coefficienti autoregressivi con il metodo OLS. Il modello ARIMA, aggiungendo la componente dell'integrazione, modifica l'equazione (4) (Hyndman, 2014):

$$a(L)(1 - L)^d Y_t = \beta_0 + b(L)e_t \tag{10}$$

Qualora non fosse necessaria alcuna integrazione -ad esempio, quando già prima della scelta del modello viene svolta una differenza della serie- un modello ARIMA(p, d, q) con $d = 0$ non equivale ad altro che ad un ARMA(p, q).

Per l'applicazione del modello a questo studio è stato utilizzato il pacchetto "forecast", sviluppato dal prof. Rob J. Hyndman, che utilizza un algoritmo per il calcolo della migliore combinazione dei parametri p, d e q nella costruzione di un ARIMA.

Ai fini della costruzione di un buon modello, un passaggio obbligatorio è la verifica della struttura della funzione di autocorrelazione parziale.

L'autocorrelazione parziale misura la relazione tra Y_t e Y_{t-j} "controllando" l'effetto dei lag precedenti. I valori della funzione di autocorrelazione parziale (PACF) sono utili a capire quanti ritardi è necessario considerare per inquadrare correttamente una serie in un modello autoregressivo: è utilizzata per trovare l'ordine di un modello AR(p) ottimale, o della componente autoregressiva di un ARIMA(p, d, q).

Nella pratica, si ricerca il lag dal quale il valore della PACF non è più statisticamente significativo. Ad esempio, se la funzione mantiene valori significativi fino al lag 3, questo potrebbe essere un p idoneo al modello autoregressivo.

Similmente, anche la funzione di autocorrelazione semplice (ACF) fornisce delle indicazioni sui modelli da costruire su una serie. In particolare, la stessa struttura descritta sopra -una funzione significativa fino al lag 3- indicherebbe in questo caso l'ordine q migliore da usare per la componente a media mobile MA(q) di un ARIMA.

Nel caso studio, la rappresentazione grafica della PACF è stata utilizzata per individuare quali modelli AR(p) potessero meglio predire la domanda turistica.

Un metodo per valutare l'efficienza dei modelli dopo la loro costruzione si sostanzia nell'utilizzo dei criteri informativi. Sono misure ideate dagli statistici per stimare il numero di ritardi ottimale di una regressione temporale. Solitamente se ne sceglie uno che viene calcolato per più modelli e si seleziona quello in grado di minimizzarlo. Maggiormente noti sono il criterio d'informazione di Akaike (AIC); il criterio d'informazione Hannan-Quinn (HQIC); e il criterio d'informazione bayesiano o di Schwartz (BIC o SIC).

Gli statistici preferiscono spesso quest'ultimo rispetto agli altri perché più "parsimonioso", cioè tende a suggerire un numero di ritardi inferiore e dunque meno complesso da applicare. Tuttavia, la parsimonia ha talvolta il difetto di consigliare modelli troppo semplici che non si adattano al meglio alla serie.

Il BIC si calcola con la formula (Stock e Watson, 2020, p. 460):

$$\text{BIC}(p) = ln(\frac{SSR(p)}{T}) + (p+1)\frac{lnT}{T} \tag{11}$$

dove p è il numero di lag e T il numero di osservazioni. Lo stimatore $\hat{p}$ è il numero di ritardi che minimizza il BIC per un certo numero di osservazioni. La formula (11) è il risultato di un bilanciamento dei due addendi: il secondo cresce all'aumentare del numero dei ritardi; il primo, al contrario, diminuisce o non cresce per valori più grandi di p perché contiene al suo interno la SSR (somma dei quadrati dei residui). Per i diversi ordini dell'autoregressione, i due addendi variano in senso opposto e minimizzano il criterio proprio all'ordine $\hat{p}$. Anche il criterio bayesiano è stato utilizzato nel caso studio, come criterio di scelta di alcuni modelli.

Nel capitolo successivo, i nove forecast sono confrontati tra loro con la radice dell'errore quadratico medio di previsione (RMSFE) come misura di riferimento, con l'obiettivo di minimizzarla. Vi sono diversi approcci per calcolare l'RMSFE. Lo studio prende come parametro di confronto l'errore calcolato con la tecnica del SER e l'errore pseudo out of sample (Stock e Watson, 2020, p. 456).

Il SER (*Standard Error of Regression*) è un indice di bontà dell'adattamento di un modello autoregressivo a una serie di dati. Si ottiene nella seguente maniera:

$$SER = \frac{RSS}{T - p - 1} \tag{12}$$

dove RSS è la *Residual Sum of Squares*, quindi la somma dei residui $(y_i - \widehat{y_i})$ della regressione elevati al quadrato, e $T - p - 1$ il numero di osservazioni sottratto dal numero di coefficienti autoregressivi stimati più l'intercetta. Quando non si hanno a disposizione dei valori reali con cui confrontare le previsioni si ricorre all'uguaglianza $MSFE = SER$, e calcolando la radice quadrata si ottiene una stima del RMSFE.

L'errore pseudo out of sample, invece, si calcola confrontando le previsioni con le osservazioni del test set . Se i residui out of sample vengono calcolati solo sui periodi del test set, quindi $\widetilde{u_i} = (p_i - \widehat{p_i})$; allora l'errore out of sample si ottiene con:

$$RMSFE_{POOS} = \sqrt{\frac{\sum_{i=T-P+1}^{T} \widetilde{u_i}^2}{P}} \tag{13}$$

dove P è il numero di previsioni . Nel caso in analisi, si è scelto $P = 10$.

Quando si stimano dei valori futuri out of sample, il metodo migliore è ovviamente quello di attendere che i dati reali sui periodi disponibili diventino disponibili.

Capitolo terzo

Analisi empirica

3.1 Selezione dei modelli di previsione

Riassumendo brevemente quanto detto in precedenza, per questo studio sono state fatte delle previsioni un passo avanti out of sample sugli arrivi turistici in Italia, Francia e Spagna per i periodi successivi al mese di marzo 2020, quindi subito prima del break strutturale dovuto alla pandemia. I modelli sono stati calcolati sulla base della differenza dodicesima del logaritmo delle serie per il periodo da Gennaio 1994 e Febbraio 2020 (314 osservazioni).
Si è scelto di utilizzare un modello *fixed* (fisso), tralasciando metodi eccessivamente complessi come l'uso di modelli *rolling*.

Al capitolo secondo sono stati menzionati i criteri di scelta dei modelli usati. Per ognuna della serie, si è dapprima calcolato il modello ARIMA(p,d,q) suggerito dal pacchetto statistico "forecast", sviluppato dal prof. Rob J. Hyndman. Una funzione del pacchetto è in grado di individuare quello che dovrebbe essere il modello migliore per prevedere la serie sulla base dei dati passati e della loro autocorrelazione. A questo, va aggiunto un modello ARMA(p,q) di uno tra i primi quattro ordini, selezionato con l'obiettivo di minimizzare il criterio d'informazione bayesiano. Infine, partendo dall'interpretazione dei grafici di AFC e PACF, si è utilizzato un modello aggiuntivo per serie di scelta dell'autore.

La prima serie storica su cui sono stati adattati i tre modelli è quella relativa all'Italia. Le sue funzioni di autocorrelazione ed autocorrelazione parziale sono rappresentate nell'immagine seguente.

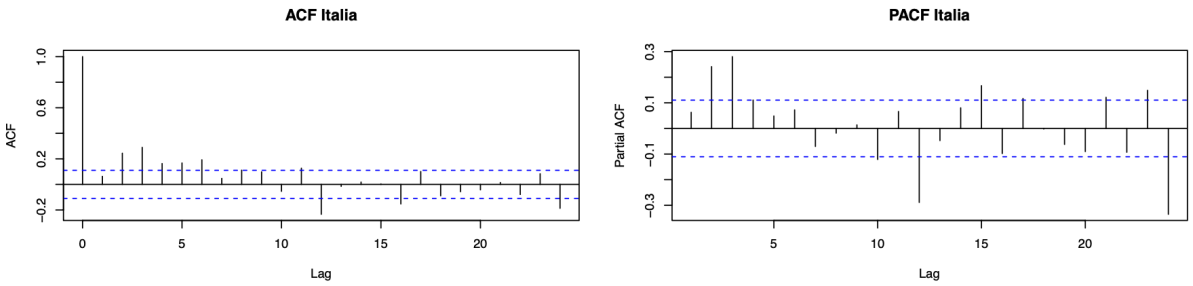

Figura 3.1 Funzioni di autocorrelazione e di autocorrelazione parziale

Osservando la figura 3.1 si può notare come la differenza dodicesima e la correzione logaritmica non siano riuscite del tutto ad eliminare la stagionalità. Durante lo studio le serie si sono rivelate molto particolari e fortemente autocorrelate, anche effettuando ulteriori differenziazioni. Per ciò, almeno con il modello selezionato in maniera soggettiva, non si è voluto andare oltre la differenza di lag dodici effettuata in precedenza, che rappresenta una procedura standard per la rimozione della stagionalità. Basandosi quindi sui grafici di ACF e PACF, sono stati scelti una p e una q che racchiudessero i suoi valori più alti tra i primi sei coefficienti. Questa restrizione è stata fatta con la volontà di non rendere eccessivamente complesso il modello. Il risultato è un ARMA(3, 3).
Sottoponendo i dati all'analisi dell'algoritmo di Hyndman, il modello suggerito è invece un ARIMA(1, 0, 2).

La serie storica degli arrivi turistici in Francia ha delle differenze rispetto a quella analizzata in precedenza. La serie è significativamente autocorrelata, ma a livelli maggiori e per un numero superiore di lag iniziali, come si può vedere in figura 3.2.

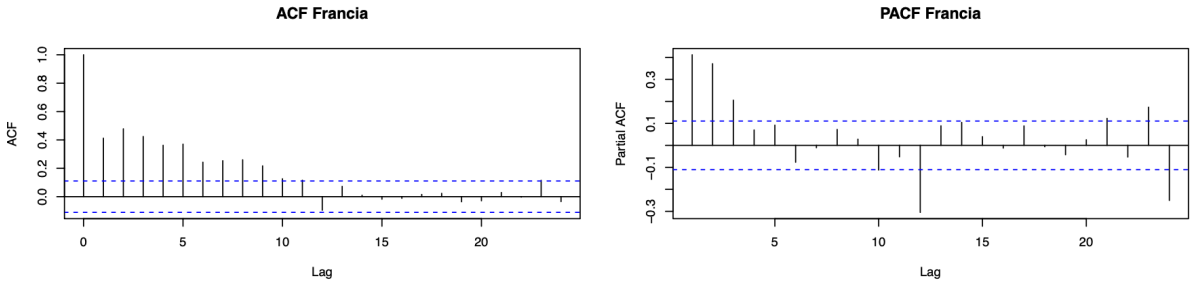

Figura 3.2 Funzioni di autocorrelazione e di autocorrelazione parziale

L'ACF, che dovrebbe farci intuire la componente a media mobile, è molto difficile da decifrare a causa dei suoi valori molto alti e significativi fino all'undicesimo lag. Applicando in maniera rigorosa il criterio di scelta utilizzato per gli arrivi in Italia, il modello ARIMA dovrebbe avere una componente MA di ordine $q = 2$. il valore dell'autocorrelazione rimane alto per quasi un anno, ma al secondo ritardo tocca la vetta più alta e inizia una discesa, anche se tutt'altro che ripida.
La componente autoregressiva è presente: lo indica l'autocorrelazione parziale, che rimane significativa anche al terzo lag. In questo caso, decidere l'ordine p del modello con la stessa

logica precedente porterebbe ad ignorare il valore del secondo ritardo, prossimo a quello del primo e dalla significatività difficilmente ignorabile. Per evitare quella che sarebbe una grave omissione, uno strappo alla regola consente di costruire in definitiva un ARMA(2,2).
Il modello suggerito dall'algoritmo forecast per la previsione di questa serie è un ARIMA(1,1,2).

L'ultima serie è, per certi versi, più simile a quella francese che a quella italiana. Di seguito, la rappresentazione delle sue ACF e PACF.

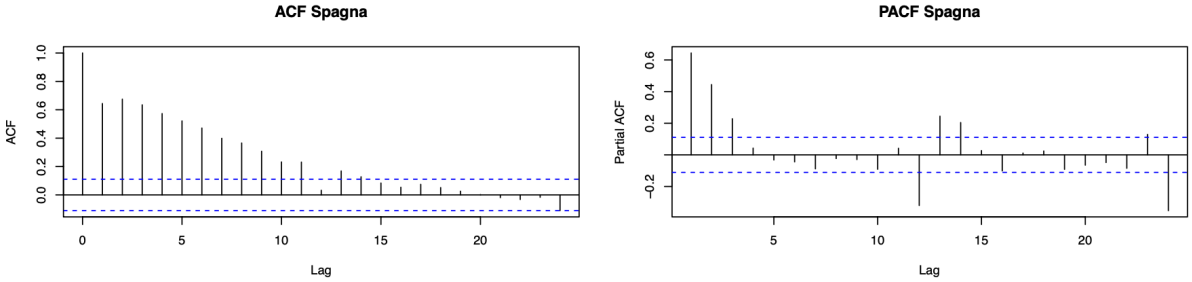

Figura 3.3 Funzioni di autocorrelazione e di autocorrelazione parziale

L'autocorrelazione, come nel caso della Francia, è significativa fino all'undicesimo ritardo, con un andamento decrescente a partire dal lag 2. L'autocorrelazione parziale ha una struttura più frastagliata, significantiva per i primi tre ritardi e decrescente già dal primo. Analizzando la serie precedente si era già optato per una $p = 2$ con l'obiettivo di non tralasciare un valore della PACF vicino a 0,4 al secondo coefficiente. In questo caso un ARMA(2,2) garantirebbe maggiore coerenza nella scelta dei modelli: il valore al secondo lag, infatti, è addirittura superiore allo 0,4 considerato per la domanda Francese. Il modello suggerito dell'algoritmo è in questo caso un ARIMA(3,1,0).

Per la scelta del terzo modello si è fatto affidamento alla minimizzazione del criterio d'informazione bayesiano. In tutti e tre i casi, tale indice ha riscontrato i valori più bassi in corrispondenza di un modello autoregressivo di terzo ordine AR(3).
La tabella 3.1 illustra il confronto e affianca i valori per i tre Paesi al fine di individuare il modello più efficiente di tutti. Confrontando i valori tra di loro, si percepisce un'importante distanza tra i valori raggiunti dalle tre diverse serie. Essendo l'obiettivo una minimizzazione

del criterio, i modelli per la Francia sono i più efficienti tra quelli esaminati; al contrario gli spagnoli sono i peggiori.

Tabella 3.1 Criterio informativo di Schwartz

	AR(1)	AR(2)	AR(3)	AR(4)
Italia	-1004,168	-1017,443	-1037,847	-1036,301
Francia	-1131,595	-1172,457	-1180,048	-1175,782
Spagna	-919,348	-982,6176	-994,6976	-989,4809

Questa discrepanza è un indizio di quali serie temporali siano meglio descrivibili dall'autoregressione, o abbiano in generale una componente autoregressiva più forte rispetto a quella a media mobile.

Sintetizzando quanto fatto fino ad ora, i tre modelli ARIMA scelti per ogni Paese sono rappresentati nella tabella 3.2. L'algoritmo di Hyndman ha ritenuto necessaria un'ulteriore differenziazione nel caso di Francia e Spagna. In precedenza si era detto di come le serie avessero mantenuto una connotazione stagionale nonostante la differenza dodicesima: un modello autoregressivo probabilmente più efficiente sarebbe stato infatti un AR(12), che richiede però un grado di complessità maggiore ed è stato scartato al momento della scelta dei modelli sulla base di ACF e PACF. La stagionalità delle serie temporali è molto forte, intrinseca nella natura stessa dei dati raccolti, e rappresenta uno degli ostacoli principali nel trovare dei modelli ben performanti.
È interessante osservare, inoltre, come non sia stato possibile ignorare la componente *moving average* al momento dell'analisi delle autocorrelazioni: l'intenzione iniziale dello studio era, escludendo il caso dell'algoritmo, di costruire dei semplici modelli autoregressivi, ma i notevoli valori della ACF non lo hanno permesso.

Tabella 3.2 Modelli di previsione

	Algoritmo	BIC	ACF e PACF
Italia	ARMA(1,2)	AR(3)	ARMA(3,3)
Francia	ARIMA(1,1,2)	AR(3)	ARMA(2,2)
Spagna	ARIMA(3,1,0)	AR(3)	ARMA(2,2)

Nella sezione successiva sono analizzati i risultati empirici e confrontate tra loro le varie previsioni, cercando di trarre le migliori conclusioni possibili dall'overview che questo studio è stato in grado di fornire.

3.2 Analisi dell'accuratezza delle previsioni

I valori della radice degli errori quadratici medi di previsione sono riportati nelle tabelle di seguito.

Tabella 3.3 RMSFE Italia

	Algoritmo	BIC	ACF e PACF
Modello	ARMA(1,2)	AR(3)	ARMA(3,3)
$RMSFE_{SER}$	0,002502	0,002521	0,002514
$RMSFE_{POOS}$	2,261385	2,260036	2,261462

Tabella 3.4 RMSFE Francia

	Algoritmo	BIC	ACF e PACF
Modello	ARIMA(1,1,2)	AR(3)	ARMA(2,2)
$RMSFE_{SER}$	0,002047	0,002010	0,002011
$RMSFE_{POOS}$	1,583657	1,581086	1,581197

Tabella 3.5 RMSFE Spagna

	Algoritmo	BIC	ACF e PACF
Modello	ARIMA(3,1,0)	AR(3)	ARMA(2,2)
$RMSFE_{SER}$	0,002749	0,002697	0,002681
$RMSFE_{POOS}$	2,118764	2,117015	2,120629

Il mero numero non è in grado di fornirci un'indicazione univoca, ma serve interpretare i risultati tenendo a mente tutti i punti di partenza dell'analisi e gli obiettivi posti.

Come presupposto, è necessario ricordare che non esiste una statistica precisa in grado di inquadrare la domanda turistica, ma la scelta degli arrivi nelle strutture alberghiere è una delle più comuni in letteratura ed è una delle più ragionevoli da usare.

La difficoltà a limitare il fenomeno ad un contesto quantitativo è una delle ragioni per cui in realtà un'analisi multivariata sembrerebbe più adatta a prevedere i flussi turistici in maniera ottimale. Nonostante ciò, è presente una vasta letteratura che lavora con le tecniche e i modelli che sono stati utilizzati in questo elaborato. L'obiettivo, si ribadisce, è il loro confronto sulla base degli indici di bontà dell'adattamento mostrati sopra.

Osservando le tabelle precedenti in ordine, la 3.3 riporta gli errori per la serie storica italiana. L'errore calcolato col metodo del SER è stato minimizzato dal modello ottenuto con l'algoritmo, l'errore pseudo out of sample dal modello autoregressivo di terzo ordine. Questo, selezionato in base al criterio del BIC, ha raggiunto i risultati migliori in entrambi gli errori per la serie francese, come mostra la tabella 3.4. L'unico caso in cui il modello scelto soggettivamente ha minimizzato l'errore è quello del $RMSFE_{SER}$ della Spagna (tabella 3.5). Anche per quest'ultima serie l'AR(3) ha prodotto l'errore pseudo out of sample più basso.
I modelli più efficienti e il loro criterio di selezione sono riassunti nella tabella sotto, al fine di garantire una comparabilità facile ed immediata.

Tabella 3.6 Modelli più efficienti

Paese	$RMSFE_{SER}$	$RMSFE_{POOS}$
Italia	ARMA(1,2) - Algoritmo	AR(3) - BIC
Francia	AR(3) - BIC	AR(3) - BIC
Spagna	ARMA(2,2)	AR(3) - BIC

Ciò che salta subito all'occhio è che l'errore pseudo out of sample è sempre stato minimizzato dal AR(3) trovato col criterio di Schwartz. Come prima cosa, dunque, si potrebbe affermare che il modello che è stato maggiormente in grado di prevedere l'effettiva domanda turistica in tutti e tre i Paesi, da Marzo 2020 fino al dicembre dello stesso anno, è l'AR(3). Tecnicamente ciò è vero, ma non bisogna dimenticarsi che il taglio tra training set e test set è stato effettuato all'inizio di un break strutturale i cui effetti si sono protratti sicuramente almeno fino a dicembre. Con lo scoppio della pandemia, i valori reali hanno subito un calo notevole e le previsioni -tutte errate *de facto*, in quanto la possibilità di un break non era tenuta in considerazione- hanno dato vita a residui molto alti: richiamando le tabelle dalla 3.3 alla 3.5, vedasi come l'errore pseudo out of sample ha sempre raggiungo un valore superiore ad 1,5.

Proprio per questo, se si rappresentassero sullo stesso grafico il valore reale della serie e delle tre previsioni per il periodo considerato, questo risulterebbe con una sovrapposizione delle previsioni, ben al di sopra del livello reale della variabile. In realtà i tre forecast differiscono tra di loro.

I modelli che sono stati premiati dall'errore out of sample, quindi, sono quelli che hanno restituito le previsioni "peggiori", ovvero le più negative. Sul perché siano sempre gli AR(3), si può fornire un'interpretazione. Si era già parlato dei residui della componente stagionale individuati dalle funzioni di autocorrelazione: a diversi coefficienti risulta ancora significativa, anche se con un ordine di grandezza inferiore rispetto alla sua versione con valori reali.

In questo contesto, un modello AR(12) -lo si accennava prima- fornisce probabilmente risultati migliori, anche se a scapito della semplicità di calcolo. Questo accade perché nella previsione di un valore futuro si considera anche il valore del suo dodicesimo ritardo, corrispondente allo stesso mese dell'anno precedente, e influenzato dalla stagionalità nello stesso modo in cui dovrebbe esserlo il valore da predire. Tutto ciò in un AR(3) non avviene, in quanto si utilizzano per la previsione solo tre ritardi. La motivazione potrebbe risiedere nel fatto che i valori vicini all'osservazione da prevedere di solito hanno una connotazione stagionale diversa, se non opposta: se si prevede il mese di giugno i ritardi sono maggio, aprile e marzo, non certo simili dal punto di vista turistico.

Conseguentemente a quanto è appena stato detto, per la verifica dell'efficienza dei modelli probabilmente è opportuno concedere un maggiore peso all'errore di previsione calcolato col metodo del SER. Questo errore è implicito all'autoregressione e viene "proiettato" alla previsione: facendo ciò, il break strutturale non causa scostamenti eclatanti. Le tabelle precedenti mostravano che per l'Italia il $RMSFE_{SER}$ viene minimizzato dal modello ARMA(1,2) scelto dall'algoritmo. In generale, l'errore calcolato col SER ha registrato dei valori abbastanza bassi da poter affermare una buona efficienza del modello.

Le previsioni Francia e Spagna si configurano in una forma maggiormente distintiva rispetto alle proprie alternative. L'operazione dell'algoritmo forecast di integrare i propri modelli, effettuando una differenziazione di un grado, ha uniformato tra loro le proprie previsioni.

Tra i tre modelli francesi si è visto che è l'AR(3) a restituire l'errore di previsione minore, anche calcolandolo come SER. In effetti, nella sezione precedente, la tabella riassuntiva del BIC mostrava i valori più piccoli in corrispondenza della serie francese; ciò era stato interpretato

come un potenziale indizio della predominanza della componente autoregressiva per questo Paese in particolare, e la misura degli errori ha dato ragione a questa supposizione.
Il modello AR(3) sembra essere quindi maggiormente efficiente anche dell'ARIMA(1,1,2) stimato dall'algoritmo, oltre che del modello nato dall'interpretazione delle autocorrelazioni.

La serie storica spagnola è quella che ha dato i risultati più diversi tra loro, in parte per come sono stati costruiti i modelli. Anche in questo caso la previsione dell'ARIMA è più stabile. Le previsioni del ARMA(2,2) si conformano in modo diverso dalla fattispecie della Francia, eppure finisce con l'avere l'errore più piccolo. Effettivamente, in termini di struttura di ACF e PACF, Francia e Spagna hanno caratteristiche simili, quest'ultima però arriva ad ordini di grandezza maggiori, di circa 0,2 (vedasi figure 3.2 e 3.3). I grafici delle funzioni di autocorrelazione palesavano una forte presenza della componente a media mobile rispetto all'autoregressiva; inoltre, la Spagna aveva ottenuto i risultati peggiori per quanto riguarda il BIC. Per questo, un modello con una componente moving average forte, al pari all'autoregressiva, ha restituito migliori valori dell'indice di bontà di adattamento.

Il fatto che per i tre Paesi il responso sul modello migliore indichi tre modelli diversi ha un significato. Così come non è fattibile stabilire una o più variabili fisse per la definizione della domanda turistica, è ugualmente sbagliato pensare che possa esistere un modello unico per la costruzione dei modelli. I tre Paesi, seppur molto simili, presentano delle serie storiche molto diverse per caratteristiche, dal livello delle autocorrelazioni fino ad arrivare alle statistiche descrittive. Ai fini di un'analisi previsionale la natura del dato può dare delle indicazioni sullo svolgimento e sulle tecniche -ad esempio, sulla correzione della stagionalità- ma lo specifico ordine autoregressivo dei modelli applicati, così come la definizione della sua componente a media mobile, è solo e soltanto campo della specifica serie temporale e deriva dalle sue caratteristiche intrinseche. Possono essere svariati i motivi per cui la serie spagnola ha un'autocorrelazione così forte rispetto alle altre, o per cui gli arrivi in Francia sembrano avere una componente autoregressiva predominante, e così via. Per spiegare queste caratteristiche, un'analisi multivariata potrebbe accostare altre variabili legate alle condizioni economiche dei Paesi o ad altri fattori sociali, o anche climatici.
Queste sono tutte implicazioni che rientrano nell'interesse degli *stakeholder* dei tre Paesi, tra cui ai policy maker vanno aggiunti tutti i business i cui profitti sono correlati alla domanda turistica. A grandi linee, per i portatori d'interesse della domanda turistica francese è impor-

tante conoscerne la forte componente autoregressiva, che con maggiori esplorazioni ed indagini più dettagliate potrebbe aiutare a trovare degli specifici periodi di riferimento col quale prevedere i flussi turistici futuri. Similmente, chi è interessato all'andamento del turismo in Italia è consapevole dei valori minori delle autocorrelazioni e potrebbe spostare l'asse della ricerca su variabili esogene in grado di spiegare maggiormente il fenomeno.

Oltre alla tipologia di modello, è anche interessante osservare come sia stata anche la metodologia di selezione a non vedere nascere una sorta di standard univoco. Si osservino nuovamente le tabelle a inizio sezione: i SER ottenuti dai tre modelli si scostano tra loro quasi sempre alla quinta -talvolta alla quarta- cifra decimale, e hanno in generale un valore abbastanza basso. I risultati ottenuti legittimano in due casi su tre le tecniche alternative all'algoritmo, che potenzialmente potrebbe essere la prima scelta in fase di progettazione di un'analisi previsionale. I soggetti interessati dovrebbero allora non solo conoscere e tenere a mente le caratteristiche proprie della domanda, ma superare un'impostazione dove viene scelto un singolo criterio di selezione e tentare l'applicazione di più metodologie in una singola analisi.

Conclusioni

Il lavoro mostrato in questa tesi ha in primo luogo affrontato il tema del turismo e della sua crescita avvenuta negli ultimi anni. I dati riportati hanno restituito il ritratto di un settore in espansione, soprattuto dopo il periodo di lockdown dovuto alla pandemia del 2020.
In questo contesto, divengono fondamentali un'ampia conoscenza e miglior previsione possibile del suo andamento. La domanda del settore turistico, si è detto, è quindi un punto di riferimento sia per grandi business che per piccoli imprenditori, così come per i decisori dello scenario politico. La regolamentazione del turismo, infatti, deve far i conti con i suoi flussi futuri e tentare di orientarli nella direzione voluta.
Proprio quanto alla previsione della domanda turistica, ne si sono evidenziate le peculiarità e caratteristiche che la rendono un interessante oggetto di studio e di analisi. Si è spiegato come una grande quantità di dati e di fattori aumentino l'incertezza nel definire quantitativamente un fenomeno così complesso. Tra di essi, lo sviluppo tecnologico e i trend nati dal turismo social portati avanti dalle nuove generazioni, ma anche motivazioni riguardanti il comportamento degli individui e l'accumulo di maggiore risparmio.

In seguito è stata illustrata l'analisi di previsione effettuata dall'autore sulla domanda turistica di tre Paesi europei, misurata per mezzo della statistica degli arrivi turistici. Per la lavorazione delle time series sono stati necessari correttivi come la differenziazione di dodicesimo grado e il calcolo del logaritmo naturale, al fine di mitigare gli effetti degli outlier e quelli della componente stagionale.
Sono stati costruiti tre ARIMA(p, d, q) per le domande di Italia, Francia e Spagna, per un totale di nove modelli. Per la selezione dei modelli si è fatto ricorso a tre criteri: calcolo di un modello ottimale da parte dell'algoritmo "forecast" del prof. Rob J. Hyndman, minimizzazione del criterio d'informazione bayesiano con un AR(p), infine una scelta soggettiva a seguito dell'interpretazione dei grafici di ACF e PACF delle serie storiche. Il risultato è stato un insieme di modelli diversi, individuati tramite approcci anch'essi distanti tra loro.

I modelli sono stati costruiti nell'ottica di una previsione pseudo out of sample con modello *fixed* a partire dal mese di Marzo 2020, su un training set di 314 osservazioni. Successivamente, è seguito un confronto tra modelli e metodi di selezione finalizzato alla ricerca di un even-

tuale maggior efficienza di uno di essi. I nove modelli sono stati comparati con la radice dell'errore quadratico medio di previsione pseudo out of sample, e con la radice dell'errore calcolato col metodo del SER.

Avendo sezionato la serie all'inizio di un break strutturale, tra le due radici degli errori quadratici medi di previsione si è riservato maggiore peso a quello calcolato col metodo del SER; questo perché lo pseudo out of sample ha favorito i modelli che avevano prodotto le previsioni maggiormente negative, ovvero gli autoregressivi di terzo ordine. Quanto al SER, le tre serie storiche dei tre differenti Paesi si sono adattate a modelli diversi tra loro, sia per ordine che per criterio di selezione: le proprietà "intrinseche" delle serie si sono rivelate cruciali. Gli arrivi in Francia, nonostante una forte componente a media mobile, sembrano essere ben rappresentati dallo stesso AR(3) selezionato minimizzando il BIC. Per gli arrivi spagnoli ha avuto la meglio il modello ARMA(2,2), scelto soggettivamente, anche perché i due modelli "concorrenti" non assecondavano l'autocorrelazione molto alta della serie anche per grandi coefficienti. La time serie italiana, dalle caratteristiche più moderate, è stata rappresentata in modo ottimale dall'algoritmo forecast con un ARMA(1,2).

L'analisi previsionale sulle tre serie ha prodotto risultati soddisfacenti, sia in termini di valore assoluto degli errori che in termini di interpretazione.
Seppur con i limiti di un'analisi univariata, forse esigua nel definire e prevedere un fenomeno così complesso, i risvolti possono considerarsi sicuramente di valore nella misura in cui confermano la necessità di attenzionare ogni serie con le proprie caratteristiche, non fermandosi a dei modelli generali, pur descrivendo lo stesso fenomeno. Inoltre, questo lavoro avvalora tutti e tre i metodi utilizzati per la selezione dei modelli, avendo prodotto ognuno un risultato migliore degli altri pur partendo da una base comune.

Appendice

Pacchetti e comandi utilizzati su RStudio

```
#IMPORTAZIONE E LAVORAZIONE DEL DATASET
---------------------------------------
library(readxl)
library(tseries)
library(forecast)
library(xts)
library(psych)
library(tidyr)
library(ggplot2)

#Dati riportati su un file excel
dataset=read_excel(“arrivi.data.xlsx", #I dati erano riportati su un file excel
                   sheet="Foglio4")
dataset=data.frame(dataset[37:420,1:4])
colnames(dataset)=c("date","esp", "fra", "ita")

#dati singoli Paesi
data.esp=data.frame(dataset[,c(1:2)])
data.fra=data.frame(dataset[,c(1,3)])
data.ita=data.frame(dataset[,c(1,4)])

#ANALISI DELLA SERIE STORICA
-----------------------------------
date=seq(as.Date("1993-01-01"), length=384, by="months")
ITA= xts(dataset[,4], order.by=date)
FRA= xts(dataset[,3], order.by=date)
ESP= xts(dataset[,2], order.by=date)

#Rappresentazione grafica
v = as.Date("2020-03-01", format = "%Y-%m-%d", tz = "UTC")
par(mfcol=c(3,1))
plot(date, ITA,type = "l",main="Italia",ylab="",xlab="Anni")
abline(v=v, col="red")
plot(date, FRA,type = "l",main="Francia",ylab="",xlab="Anni")
abline(v=v, col="red")
plot(date, ESP,type = "l",main="Spagna",ylab="",xlab="Anni")
abline(v=v, col="red")

#STATISTICHE DESCRITTIVE
-------------------------------------------------------
summary(ITA)
describe(ITA)
summary(FRA)
describe(FRA)
summary(ESP)
describe(ESP)

#Correzione logaritmica
ITAln=log(ITA)
FRAln=log(FRA)
ESPln=log(ESP)
par(mfcol=c(3,1))
```

```
plot(date, ITAln,type = "l",main="Logaritmo naturale degli arrivi turistici in
Italia",ylab="",xlab="Anni")
abline(v=v, col="red")
plot(date, FRAln,type = "l",main="Logaritmo naturale degli arrivi turistici in
Francia",ylab="",xlab="Anni")
abline(v=v, col="red")
plot(date, ESPln,type = "l",main="Logaritmo naturale degli arrivi turistici in
Spagna",ylab="",xlab="Anni")
abline(v=v, col="red")

par(mfcol=c(3,2))
acf(ITAln, main="ACF Italia") #Autocorrelazione
acf(FRAln, main="ACF Francia")
acf(ESPln, main="ACF Spagna")
pacf(ITAln, main="PACF Italia") #Autocorrelazione parziale
pacf(FRAln, main="PACF Francia")
pacf(ESPln, main="PACF Spagna")

#DESTAGIONALIZZAZIONE DELLE SERIE STORICHE
-------------------------------------
ITA.ns= diff(ITAln, lag=12)
ITA.ns=subset(ITA.ns[13:384])
FRA.ns= diff(FRAln, lag=12)
FRA.ns=subset(FRA.ns[13:384])
ESP.ns= diff(ESPln, lag=12)
ESP.ns=subset(ESP.ns[13:384])
date.ns=seq(as.Date("1994-01-01"), length=372, by="months")
par(mfcol=c(3,1))
plot(date.ns, ITA.ns,type = "l",main="Italia",ylab="",xlab="Anni")
abline(v=v, col="red")
plot(date.ns, FRA.ns,type = "l",main="Francia",ylab="",xlab="Anni")
abline(v=v, col="red")
plot(date.ns, ESP.ns,type = "l",main="Spagna",ylab="",xlab="Anni")
abline(v=v, col="red")

par(mfcol=c(3,2))
acf(ITA.ns, main="ACF Italia") #Autocorrelazione
acf(FRA.ns, main="ACF Francia")
acf(ESP.ns, main="ACF Spagna")
pacf(ITA.ns, main="PACF Italia") #Autocorrelazione parziale
pacf(FRA.ns, main="PACF Francia")
pacf(ESP.ns, main="PACF Spagna")

adf.test(ITA.ns) #Augmented Dickey-Fuller test per la stazionarietà
adf.test(FRA.ns)
adf.test(ESP.ns)

#ANALISI PRE-COVID
-------------------------------------------------------------
ITA.pre = subset(ITAln[1:326])
ITA.pre = diff(ITA.pre, lag=12)
ITA.pre = subset(ITA.pre[13:326])
FRA.pre = subset(FRAln[1:326])
FRA.pre = diff(FRA.pre, lag=12)
FRA.pre = subset(FRA.pre[13:326])
ESP.pre = subset(ESPln[1:326])
ESP.pre = diff(ESP.pre, lag=12)
ESP.pre = subset(ESP.pre[13:326])

date.pre=seq(as.Date("1994-01-01"), length=314, by="months")
par(mfcol=c(3,1))
```

```
plot(date.pre, ITA.pre,type = "l",main="Italia pre-covid",ylab="",xlab="Anni")
plot(date.pre, FRA.pre,type = "l",main="Francia pre-covid",ylab="",xlab="Anni")
plot(date.pre, ESP.pre,type = "l",main="Spagna pre-covid",ylab="",xlab="Anni")


par(mfcol=c(3,2))
acf(ITA.pre, main="ACF Italia") #Autocorrelazione
acf(FRA.pre, main="ACF Francia")
acf(ESP.pre, main="ACF Spagna")
pacf(ITA.pre, main="PACF Italia") #Autocorrelazione parziale
pacf(FRA.pre, main="PACF Francia")
pacf(ESP.pre, main="PACF Spagna”)

adf.test(ITA.pre) #Augmented Dickey-Fuller test per la stazionarietà
adf.test(FRA.pre)
adf.test(ESP.pre)

#MODELLI ARIMA ITALIA
-----------------------------------------------------------
#Italia con auto.arima()
ITApre.auto = auto.arima(ITA.pre)
summary(ITApre.auto)
ITAf.auto = predict(ITApre.auto, n.ahead=10)
ITAf.auto$pred
forecastdate=seq(as.Date("2020-03-01"), length=10, by="months")
forecastITA.auto= xts(ITAf.auto$pred, order.by=forecastdate)
ITA.auto = rbind(ITA.pre, forecastITA.auto)
v1= seq(as.Date("1994-01-01"), length=324, by="months")
par(mfcol=c(1,1))
plot(v1, ITA.auto,type = "l",main="Italia",ylab="",xlab="Anni")

#Italia in base a PACF e ACF
par(mfcol=c(2,1))
acf(ITA.pre, main="ACF Italia")
pacf(ITA.pre, main="PACF Italia")
ITApre.mdl = Arima(ITA.pre, order=c(3,0,3))
summary(ITApre.mdl)
ITAf.mdl=predict(ITApre.mdl, n.ahead=10)
ITAf.mdl$pred
forecastITA.mdl= xts(ITAf.mdl$pred, order.by=forecastdate)
ITA.mdl = rbind(ITA.pre, forecastITA.mdl)
par(mfcol=c(1,1))
plot(v1, ITA.mdl,type = "l",main="Italia",ylab="",xlab="Anni")

#Italia in base al BIC
ITApre.bic1 = Arima(ITA.pre, order=c(1,0,0))
ITApre.bic2 = Arima(ITA.pre, order=c(2,0,0))
ITApre.bic3 = Arima(ITA.pre, order=c(3,0,0))
ITApre.bic4 = Arima(ITA.pre, order=c(4,0,0))
BIC(ITApre.bic1)
BIC(ITApre.bic2)
BIC(ITApre.bic3)
BIC(ITApre.bic4)
ITApre.bic = ITApre.bic3
summary(ITApre.bic)
ITAf.bic=predict(ITApre.bic, n.ahead=10)
ITAf.bic$pred
forecastITA.bic= xts(ITAf.bic$pred, order.by=forecastdate)
ITA.bic = rbind(ITA.pre, forecastITA.bic)
par(mfcol=c(1,1))
plot(v1, ITA.bic,type = "l",main="Italia",ylab="",xlab="Anni")
```

```
#MODELLI ARIMA FRANCIA
-------------------------------------------------------
#Francia con auto.arima()
FRApre.auto = auto.arima(FRA.pre)
summary(FRApre.auto)
FRAf.auto = predict(FRApre.auto, n.ahead=10)
FRAf.auto$pred
forecastFRA.auto= xts(FRAf.auto$pred, order.by=forecastdate)
FRA.auto = rbind(FRA.pre, forecastFRA.auto)
par(mfcol=c(1,1))
plot(v1, FRA.auto,type = "l",main="Francia",ylab="",xlab="Anni")

#Francia in base a PACF e ACF
par(mfcol=c(2,1))
acf(FRA.pre, main="ACF Francia")
pacf(FRA.pre, main="PACF Francia")
FRApre.mdl = Arima(FRA.pre, order=c(2,0,2))
summary(FRApre.mdl)
FRAf.mdl = predict(FRApre.mdl, n.ahead=10)
FRAf.mdl$pred
forecastFRA.mdl= xts(FRAf.mdl$pred, order.by=forecastdate)
FRA.mdl = rbind(FRA.pre, forecastFRA.mdl)
par(mfcol=c(1,1))
plot(v1, FRA.mdl,type = "l",main="Francia",ylab="",xlab="Anni")

#Francia in base al BIC
FRApre.bic1 = Arima(FRA.pre, order=c(1,0,0))
FRApre.bic2 = Arima(FRA.pre, order=c(2,0,0))
FRApre.bic3 = Arima(FRA.pre, order=c(3,0,0))
FRApre.bic4 = Arima(FRA.pre, order=c(4,0,0))
BIC(FRApre.bic1)
BIC(FRApre.bic2)
BIC(FRApre.bic3)
BIC(FRApre.bic4)
FRApre.bic = FRApre.bic3
summary(FRApre.bic)
FRAf.bic=predict(FRApre.bic, n.ahead=10)
FRAf.bic$pred
forecastFRA.bic= xts(FRAf.bic$pred, order.by=forecastdate)
FRA.bic = rbind(FRA.pre, forecastFRA.bic)
par(mfcol=c(1,1))
plot(v1, FRA.bic,type = "l",main="Francia",ylab="",xlab="Anni")

#MODELLI ARIMA SPAGNA
-------------------------------------------------------
#Spagna con auto.arima()
ESPpre.auto = auto.arima(ESP.pre)
summary(ESPpre.auto)
ESPf.auto = predict(ESPpre.auto, n.ahead=10)
ESPf.auto$pred
forecastESP.auto= xts(ESPf.auto$pred, order.by=forecastdate)
ESP.auto = rbind(ESP.pre, forecastESP.auto)
par(mfcol=c(1,1))
plot(v1, ESP.auto,type = "l",main="Spagna",ylab="",xlab="Anni")

#Spagna in base a PACF e ACF
par(mfcol=c(2,1))
acf(ESP.pre, main="ACF Spagna")
pacf(ESP.pre, main="PACF Spagna")
ESPpre.mdl = Arima(ESP.pre, order=c(2,0,2))
summary(ESPpre.mdl)
```

```
ESPf.mdl = predict(ESPpre.mdl, n.ahead=10)
ESPf.mdl$pred
forecastESP.mdl= xts(ESPf.mdl$pred, order.by=forecastdate)
ESP.mdl = rbind(ESP.pre, forecastESP.mdl)
par(mfcol=c(1,1))
plot(v1, ESP.mdl,type = "l",main="Spagna",ylab="",xlab="Anni")

#Spagna in base al BIC
ESPpre.bic1 = Arima(ESP.pre, order=c(1,0,0))
ESPpre.bic2 = Arima(ESP.pre, order=c(2,0,0))
ESPpre.bic3 = Arima(ESP.pre, order=c(3,0,0))
ESPpre.bic4 = Arima(ESP.pre, order=c(4,0,0))
BIC(ESPpre.bic1)
BIC(ESPpre.bic2)
BIC(ESPpre.bic3)
BIC(ESPpre.bic4)
ESPpre.bic = ESPpre.bic3
summary(ESPpre.bic)
ESPf.bic=predict(ESPpre.bic, n.ahead=10)
ESPf.bic$pred
forecastESP.bic= xts(ESPf.bic$pred, order.by=forecastdate)
ESP.bic = rbind(ESP.pre, forecastESP.bic)
par(mfcol=c(1,1))
plot(v1, ESP.bic,type = "l",main="Spagna",ylab="",xlab="Anni")

#CALCOLO DEGLI RMSFE
------------------------------------------------------------
#Italia col metodo del SER
sqrt(ITApre.auto$sigma2/(314-2)) #Migliore
sqrt(ITApre.mdl$sigma2/(314-3))
sqrt(ITApre.bic$sigma2/(314-2))

#Italia pseudo out-of-sample
pi = subset(ITA.ns[315:324])
sqrt(sum((pi-forecastITA.auto)^2)/10)
sqrt(sum((pi-forecastITA.mdl)^2)/10)
sqrt(sum((pi-forecastITA.bic)^2)/10) #Migliore

#Francia col metodo del SER
sqrt(FRApre.auto$sigma2/(314-2))
sqrt(FRApre.mdl$sigma2/(314-4))
sqrt(FRApre.bic$sigma2/(314-2)) #Migliore

#Francia pseudo out-of-sample
pf = subset(FRA.ns[315:324])
sqrt(sum((pf-forecastFRA.auto)^2)/10)
sqrt(sum((pf-forecastFRA.mdl)^2)/10)
sqrt(sum((pf-forecastFRA.bic)^2)/10) #Migliore

#Spagna col metodo del SER
sqrt(ESPpre.auto$sigma2/(314-4))
sqrt(ESPpre.mdl$sigma2/(314-4)) #Migliore
sqrt(ESPpre.bic$sigma2/(314-2))

#Spagna pseudo out-of-sample
pe = subset(ESP.ns[315:324])
sqrt(sum((pe-forecastESP.auto)^2)/10)
sqrt(sum((pe-forecastESP.mdl)^2)/10)
sqrt(sum((pe-forecastESP.bic)^2)/10) #Migliore
```

Bibliografia

Bufalo, M. & Orlando G. (2024). Improved tourism demand forecasting with CIR# model: a case study of disrupted data patterns in Italy. *Tourism review*, 79(2), 445-467.

Chatfield, C. (2000). *Time-series forecasting.* Chapman & Hall/CRC.

Chu, F. L. (2009). Forecasting tourism demand with ARMA-based methods. *Tourism management* 2009, 30(5), 740-741.

Gunter, U., Smeral, E., & Zekan, B. (2024). Forecasting Tourism in the EU after the COVID-19 Crisis. *Journal of Hospitality & Tourism Research*, 48(5), 909–919.

He, M., & Qian, X. (2025). Forecasting tourist arrivals using STL-XGBoost method. *Tourism Economics,* 13548166241313411.

Hyndman (2014). *Forecasting: Principles and Practice.* University of Western Australia

Lim, C., & McAleer, M. (2002). Time series forecast for international travel in Australia. *Tourism Management*, 23(4), 389-396.

Liu, A., Vici, L., Ramos, V., Giannoni, S., &Blake, A. (2021). Visitor arrivals forecasts amid COVID-19: A perspective from the Europe team. *Annals of Tourism Research*, 88, 103182

Ramsey, F. L. (1974). Characterization of the Partial Autocorrelation Function. *The Annals of Statistics*, 1296–1301.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting. *Tourism Research*, 29(2), 203-220.

Song, H., Li, G., Witt, S. F., & Athanasopoulos, G. (2011). Forecasting tourist arrivals using time-varying parameter structural time series model. *International Journal of Forecasting*, 27(3), 855-869

Stock, J. H., &Watson, M. W. (2020). *Introduction to econometrics.* Pearson

Sitografia

I flussi turistici – Anno 2023. *Istat*, 27/11/2024

Destagionalizzazione di serie storiche con metodologia Arima model based (AMB) implementata nel software JDemetra+. *Istat*, 2015

Flussi turistici terzo trimestre 2024. *Istat*, 27/11/2024

In Q4 2024, collective tourist attendance rose by 3,2% over a year. *Insee*, 20/02/2025

In Q3 2024, collective tourist attendance decreased by 1,7% over a year. *Insee*, 19/11/2025

Encuesta de ocupación hotelera e indicadores de rentabilidad. Boletín estadístico, *Ministerio de Industria y Turismo*, 27/11/2025 (ultimo accesso 28/02/2025)

UN Tourism Barometer (ultimo accesso 28/02/2025)

Fonte dei dati:

https://ec.europa.eu/eurostat/web/tourism/database (ultimo accesso 04/04/2025)

Ringraziamenti

Un sentito ringraziamento va al relatore, Prof. Mauro Costantini, per il tempo dedicatomi e per il costante supporto dimostrato in questi mesi.

Grazie di cuore alla mia famiglia e ai miei nonni per non aver mai fatto mancare presenza e affetto.

Grazie a Matilde, Claudio, Gigi e tutte le altre persone che mi hanno accompagnato al meglio lungo questo percorso.