Università
Ca'Foscari
Venezia
Facoltà
di Economia

Corso di Laurea Magistrale
in Statistica per l'impresa

Prova finale di Laurea

Modelli di previsione per il mercato dell'auto in Italia
Tesi di laurea

Relatore
Ch.mo Prof. Claudio Pizzi

Laureando
Gianni Bergamo
Matricola 825828

Anno Accademico
2012-2013

Indice

Indice **i**

Elenco delle tabelle **iii**

Elenco delle tabelle **iv**

Elenco delle figure **v**

Ringraziamenti **vii**

1 Introduzione **1**
1.1 Scopo del lavoro 1
1.2 Le immatricolazioni di auto in Italia dal 1929 al 2012 3

2 Analisi preliminare e stima di un modello stocastico stagionale **10**
2.1 Introduzione 10
2.2 Serie mensile 1985-2011 10
2.2.1 Trend 12
2.2.2 Stagionalità 14
2.3 Un modello stocastico per le immatricolazioni mensili 18
2.3.1 Verifica della stazionarietà 19
2.3.2 Funzioni di autocorrelazione 23
2.3.3 Stima di modelli *SARIMA* 25
2.4 Accuratezza dei modelli 34
2.5 Conclusioni 38

3 Intervention analysis **39**
3.1 Introduzione 39
3.2 Incentivi statali per la rottamazione e l'acquisto di autovetture . . . 39
3.3 L'analisi degli interventi e la ricerca di outlier 40
3.3.1 Outliers additivi e innovativi 44
3.4 Il noise model 49
3.5 Le variabili d'intervento 51
3.6 Stima di modelli d'intervento 55
3.6.1 Un modello iniziale 55
3.6.2 Modifiche al modello iniziale e stima dell'effetto di outliers . 57

3.6.3 Inserimento di ulteriori variabili indicatrici 63
3.7 Accuratezza dei modelli 69
3.8 Conclusioni 70

4 Alla ricerca di nuove variabili predittive: Google Trends come fonte alternativa di dati? 71
4.1 Introduzione 71
4.2 Google Trends 73
4.3 Un possibile predittore per il mercato auto 75
4.4 Stima di modelli di regressione 78
4.5 Accuratezza dei modelli 82
4.6 Conclusioni 83

5 Conclusioni 84

Bibliografia 86

Elenco delle tabelle

2.1 SARIMA$(3,0,0) \times (0,1,0)_{12}$ . . . 26
2.2 SARIMA$(3,0,0) \times (0,1,1)_{12}$ . . . 28
2.3 SARIMA$(3,0,0) \times (1,1,0)_{12}$ . . . 30
2.4 SARIMA(3, 0, 0)$\times$(1, 1, 1)$_{12}$ . . . 30
2.5 AICc, AIC e BIC . . . 33
2.6 Test di normalità sui residui del modello SARIMA(3,0,0)$\times$(0,1,1)$_{12}$ . . . 33
2.7 Test di normalità sui residui dei modelli. . . . 35
2.8 Accuratezza in-sample. . . . 37
2.9 Accuratezza out-of-sample ad un passo. . . . 38

3.1 Contributi statali per la rottamazione e l'acquisto di autovetture nuove 40
3.2 Noise model SARIMA$(1,0,0) \times (1,1,0)_{12}$ . . . 51
3.3 Regressione lineare con gli incentivi come regressori . . . 53
3.4 Regressione lineare con incentivi e variabile "crisi" come regressori . . 54
3.5 Modello d'intervento iniziale . . . 56
3.6 Modelli d'intervento candidati, senza l'effetto di outliers. . . . 58
3.7 Modelli d'intervento candidati con l'inclusione di outliers innovativi. . 59
3.8 Test di normalità sui residui dei modelli Mod1 e Mod2. . . . 62
3.9 Arch LM-Test sui residui dei modelli Mod1 e Mod2. . . . 62
3.10 Modelli d'intervento Mod1.1 e Mod2.1 . . . 64
3.11 Modelli d'intervento Mod1.1 e Mod2.1 escluso l'effetto marzo 2010 . . 65
3.12 Test di normalità sui residui dei modelli Mod1.1 e Mod2.1 . . . 69
3.13 Arch LM-Test sui residui dei modelli Mod1.1 e Mod2.1 . . . 69
3.14 Modelli d'intervento: errori di previsione in-sample. . . . 69
3.15 Modelli d'intervento: errori di previsione ad un passo. . . . 70

4.1 Confronto tra il modello con variabile G_auto8 e il modello con G_brands8 81
4.2 Test di normalità sui residui dei modelli con G_auto8 e G_brands8 . . 83
4.3 Accuratezza out-of-sample ad un passo. . . . 83

Elenco delle tabelle

2.1 SARIMA$(3,0,0) \times (0,1,0)_{12}$ . . . 26
2.2 SARIMA$(3,0,0) \times (0,1,1)_{12}$ . . . 28
2.3 SARIMA$(3,0,0) \times (1,1,0)_{12}$ . . . 30
2.4 SARIMA(3, 0, 0)$\times$(1, 1, 1)$_{12}$ . . . 30
2.5 AICc, AIC e BIC . . . 33
2.6 Test di normalità sui residui del modello SARIMA(3,0,0)$\times$(0,1,1)$_{12}$ . . . 33
2.7 Test di normalità sui residui dei modelli. . . . 35
2.8 Accuratezza in-sample. . . . 37
2.9 Accuratezza out-of-sample ad un passo. . . . 38

3.1 Contributi statali per la rottamazione e l'acquisto di autovetture nuove 40
3.2 Noise model SARIMA$(1,0,0) \times (1,1,0)_{12}$ . . . 51
3.3 Regressione lineare con gli incentivi come regressori . . . 53
3.4 Regressione lineare con incentivi e variabile "crisi" come regressori . . 54
3.5 Modello d'intervento iniziale . . . 56
3.6 Modelli d'intervento candidati, senza l'effetto di outliers. . . . 58
3.7 Modelli d'intervento candidati con l'inclusione di outliers innovativi. . 59
3.8 Test di normalità sui residui dei modelli Mod1 e Mod2. . . . 62
3.9 Arch LM-Test sui residui dei modelli Mod1 e Mod2. . . . 62
3.10 Modelli d'intervento Mod1.1 e Mod2.1 . . . 64
3.11 Modelli d'intervento Mod1.1 e Mod2.1 escluso l'effetto marzo 2010 . . 65
3.12 Test di normalità sui residui dei modelli Mod1.1 e Mod2.1 . . . 69
3.13 Arch LM-Test sui residui dei modelli Mod1.1 e Mod2.1 . . . 69
3.14 Modelli d'intervento: errori di previsione in-sample. . . . 69
3.15 Modelli d'intervento: errori di previsione ad un passo. . . . 70

4.1 Confronto tra il modello con variabile G_auto8 e il modello con G_brands8 81
4.2 Test di normalità sui residui dei modelli con G_auto8 e G_brands8 . . 83
4.3 Accuratezza out-of-sample ad un passo. . . . 83

Elenco delle figure

1.1 Immatricolazioni di auto in Italia 1928-2012 . . . 3
1.2 Immatricolazioni di auto in Italia 1928-1959 . . . 4
1.3 Immatricolazioni di auto in Italia 1960-1990 . . . 6
1.4 Immatricolazioni di auto in Italia 1991-2011 . . . 9

2.1 Immatricolazioni . . . 11
2.2 Scomposizione del logaritmo della serie mensile 1985-2011 . . . 13
2.3 Rette di regressione sulle immatricolazioni mensili 1985-2011 . . . 14
2.4 Boxplot mensili . . . 15
2.5 Autodispersione della serie mensile . . . 16
2.6 Grafico di autocorrelazione globale . . . 17
2.7 Andamento della stagionalità negli anni . . . 17
2.8 Correlogrammi della serie ridotta (logaritmi) . . . 24
2.9 Correlogrammi per la differenza stagionale della serie ridotta (logaritmi) 25
2.10 Residui del modello SARIMA(3,0,0)x$(0,1,0)_{12}$ . . . 27
2.11 p-values stastitica Ljung-Box, residui mod. SARIMA(3,0,0)x$(0,1,0)_{12}$ . . 27
2.12 Residui del modello SARIMA(3,0,0)x$(0,1,1)_{12}$ . . . 29
2.13 p-values stastitica Ljung-Box, residui mod. SARIMA(3,0,0)x$(0,1,1)_{12}$ . . 29
2.14 Residui del modello SARIMA(3, 0, 0) x $(1, 1, 0)_{12}$ . . . 31
2.15 p-values stastitica Ljung-Box, residui mod. SARIMA(3,0,0)x$(1,1,0)_{12}$ . . 31
2.16 Residui del modello SARIMA(3, 0, 0) x $(1, 1, 1)_{12}$ . . . 32
2.17 p-values stastitica Ljung-Box, residui mod. SARIMA(3,0,0)x$(1,1,1)_{12}$ . . 32
2.18 Distribuzione dei residui del modello SARIMA(3,0,0)x$(0,1,1)_{12}$ . . . 34
2.19 Distribuzione dei residui dei modelli . . . 36
2.20 Immatricolazioni reali e stimate in-sample . . . 37

3.1 Step e pulse input . . . 42
3.2 Risposte a diversi tipi di funzioni di trasferimento . . . 43
3.3 Residui, noise model SARIMA(1,0,0)x$(1,1,0)_{12}$ . . . 50
3.4 p-values della statistica Ljung-Box sui residui, noise model . . . 50
3.5 Residui Mod1 . . . 60
3.6 p-values della statistica Ljung-Box sui residui, modello Mod1 . . . 60
3.7 Residui Mod2 . . . 61
3.8 p-values della statistica Ljung-Box sui residui, modello Mod2 . . . 61
3.9 Distribuzione dei residui dei modelli Mod1 e Mod2 . . . 62

3.10 Immatricolazioni reali e stimate, Mod1 e Mod2 63
3.11 Residui Mod1 66
3.12 p-values della statistica Ljung-Box sui residui, modello Mod1 66
3.13 Residui Mod2 67
3.14 p-values della statistica Ljung-Box sui residui, modello Mod2 67
3.15 Distribuzione dei residui dei modelli Mod1 68
3.16 Immatricolazioni reali e stimate, Mod1 68

4.1 Immatricolazioni (logaritmo) e interesse di ricerca per alcune categorie 2004-2012 77
4.2 CCF tra il logaritmo delle immatricolazioni e le serie esogene 80
4.3 Distribuzione dei residui dei modelli com G-auto8 e G-brands8 82

Ringraziamenti

Un ringraziamento doveroso va alla mia famiglia che ha sempre creduto nel raggiungimento di questo obiettivo, che per me rappresenta non un traguardo, bensì una tappa importante per tutto quello che verrà in futuro. Ringrazio inoltre la Cooperativa Sociale N.O.E. Onlus, della quale sono orgoglioso di essere socio da tanti anni, per un'esperienza lavorativa che mi ha permesso di crescere come persona e per avermi supportato dandomi la possibilità di dedicare il tempo necessario per la conclusione di questo impegnativo percorso di studi.

Ringrazio i miei amici, pochi ma veri, per l'incoraggiamento, anche se magari non hanno ancora ben capito cosa abbia studiato durante questi anni e per i quali la parola "statistica" risulta sempre un termine troppo misterioso.

Per la realizzazione di questa tesi di laurea magistrale, ringrazio la società Quintegia per il supporto e che mi ha dedicato e il prof. Claudio Pizzi che ha favorito il contatto con essa per un confronto di idee; in particolar modo ringrazio Elisa Giubilato per l'aiuto nella ricerca di dati e materiale utile, oltre che per i preziosi consigli, e Luca Montagner per aver fornito materiale utile a comprendere il ruolo di internet nel processo decisionale per l'acquisto di un auto, permettendomi di sostenere l'idea alla base dell'ultimo capitolo di questo lavoro.

Capitolo 1

INTRODUZIONE

1.1 Scopo del lavoro

Quello dell'auto in Italia è un mercato di dimensioni considerevoli e, in generale, tutto il settore *automotive* ricopre un peso rilevante per la nostra economia. La filiera *automotive* italiana conta più di 1,2 milioni di addetti e rappresenta oltre il 15% del gettito fiscale nazionale.[1] Negli ultimi anni la crisi economico-finanziaria con l'incertezza dei consumatori, tasso di disoccupazione in crescita e una stretta creditizia da parte delle banche, hanno portato ad una continua decrescita delle immatricolazioni, mettendo in crisi il sistema. A ciò si aggiunge l'alto livello di tassazione sull'auto, l'elevato costo dei carburanti (su cui incide anche la presenza di diverse accise) e delle assicurazioni; l'utilizzo in anni recenti di forme di incentivazione per la rottamazione di vecchi veicoli e l'acquisto di nuovi può a sua volta aver inciso provocando un anticipo di vendite che si è ripercosso negativamente nei periodi successivi al termine dei contributi. La crisi di vendite provoca sia una ripercussione sul sistema produttivo e l'indotto, dato che il costruttore nazionale detiene la maggior quota di mercato, sia sulla rete di vendita: secondo i dati rilevati da Quintegia, nel 2012 ha chiuso il 7% dei concessionari.

In un sistema così complesso sono molte le variabili da considerare per poter

[1] Elaborazione ANFIA su dati ISTAT, EUROSTAT e ACEA e altre organizzazioni, dicembre 2012

effettuare delle stime sul futuro del mercato. La molteplicità delle variabili che possono incidere sul volume di acquisti richiede un certo costo nel reperire i dati utili a sviluppare dei modelli di previsione efficaci o nella richiesta di consulenza a società specializzate in ricerche di mercato, che utilizzano modelli proprietari su dati di settore non liberamente accessibili.

Si trovano disponibili alcuni studi che impiegano diversi approcci che vanno da modelli econometrici all'uso di tecniche di *data mining* [Carlson, 1978, lon, 2012, Hülsmann et al., 2011]. Alcuni utilizzano variabili interne al mercato dell'auto e sfruttano basi di dati dettagliate che permettono varie tipologie di segmentazione, potendo così stimare le diverse dimaniche che si riscontrano nei dati disaggregati. Altri si affidano a variabili esterne principalmente di tipo macroeconomico (PIL, reddito, tasso di disoccupazione, livello dei prezzi e dei consumi o anche indici di borsa). Fonti quali gli istituti di statistica nazionali e delle organizzazioni economiche internazionali permettono l'accesso a banche dati gratuite e danno la possibilità di sviluppare dei modelli più o meno efficaci. Tlalvolta però questi dati sono disponibili in ritardo oppure vengono pubblicati per un intervallo temporale ridotto o in forma di aggregazioni temporali diverse da quelle desiderate.

Lo scopo di questo lavoro è quello di testare la bontà delle previsioni sulle immatricolazioni di auto in Italia ottenute con modelli a basso costo di implementazione, senza l'uso di variabili macroeconomiche esterne che richiedano a loro volta stime dei valori futuri. Nel capitolo 2, dopo un'analisi descrittiva della serie mensile delle immatricolazioni, verranno stimati alcuni modelli di tipo SARIMA e verrà calcolata la loro accuratezza *in-sample* e soprattutto *out-of-sample*. Nel capitolo 3 si affronterà il problema attraverso l'approccio dell'*intervention analysis* sviluppato in particolar modo da G. E. P. Box e G. C. Tiao, dove si cercherà un modello che tenga conto della presenza di particolari disturbi che rendono irregolare la serie, tra i quali ad esempio gli incentivi statali per la rottamazione e

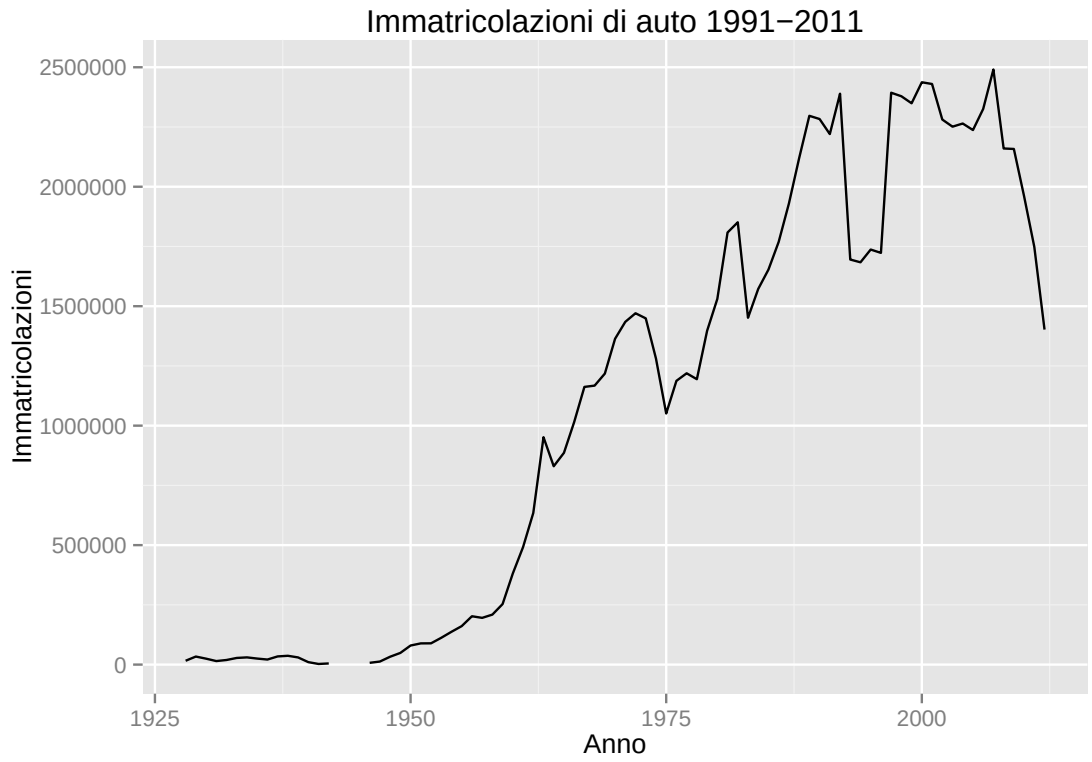


Figura 1.1: Immatricolazioni di auto in Italia 1928-2012

l'acquisto di auto nuove. Nel capitolo 4 si esplorerà la possibilità di utilizzare una fonte alternativa di dati gratuiti rappresentata da indici di interesse delle ricerche effettuate attraverso il motore *Google* verso parole chiave o categorie riguardanti il settore dell'auto.

I dati utilizzati sono forniti dal periodico InterAutoNews nel sito `http://www.interautonews.it`. La serie annuale dal 1928 è reperibile all'indirizzo `http://www.motornet.it/statisticheserver/NEW_storico.html`. I dati sull'interesse di ricerca attraverso *Google* sono stati ricavati da *Google Trends* nella versione italiana all'indirizzo `http://www.google.it/trends/`.

1.2 Le immatricolazioni di auto in Italia dal 1929 al 2012

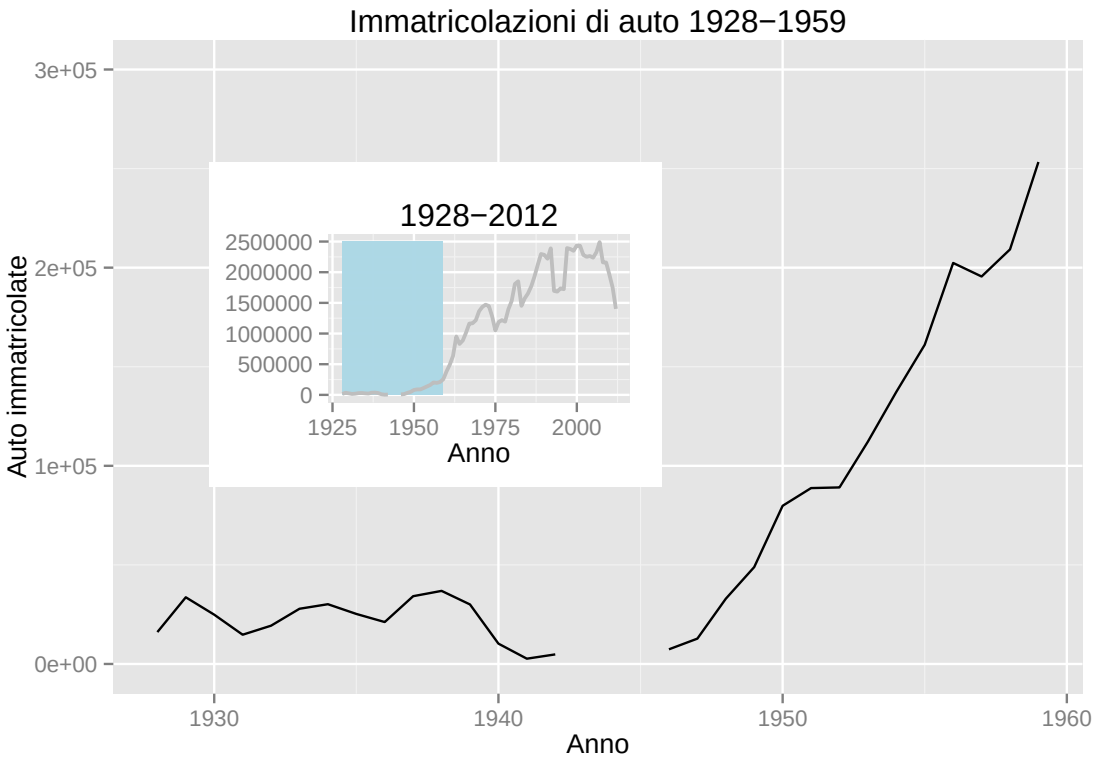


Figura 1.2: Immatricolazioni di auto in Italia 1928-1959

In figura 1.1 è rappresentata la serie storica annuale delle immatricolazioni di auto dal 1928 al 2012, esclusi gli anni dal 1943 al 1945. Si può notare come le auto nuove registrate siano rimaste ad un livello modesto fino alla conclusione del secondo conflitto mondiale. L'automobile era un prodotto non accessibile a gran parte della popolazione. La situazione cambiò con la ripresa economica del secondo dopoguerra, durante il periodo definito "miracolo economico" tra il 1951 e il 1963 (con crescita media del PIL del $5,8\%$ e del 7% tra il 1959 e il 1962) e l'arrivo sul mercato di modelli di autovetture "di massa" (le cosiddette *utilitarie*) che hanno permesso l'acquisto di veicoli a classi sociali con redditi più bassi, la cui condizione economica era in fase di miglioramento nonostante permanessero ancora situazioni di difficoltà, favorendo una forte espansione delle vendite.

Nel 1964, anno di rallentamento dello sviluppo dell'economia italiana, si verificò un arresto della fase di sviluppo del mercato, con le registrazioni che arretra-

rono da 951704 a 830175. Nel 1965, nonostante la ripresa economica, le vendite di auto ebbero un incremento relativamente contenuto (poco più di 56000 unità). Nella *"Relazione sulla situazione economica dello stato"* relativa a quell'anno è riportato l'andamento della produzione nazionale di autovetture, la quale ha registrato una crescita nei primi due trimestri dell'anno e una decrescita nei secondi due (1.4%, 37.6%, -4.9%, -24.4%). Tenendo conto che per l'industria automobilistica italiana è sempre stata importante la quota di domanda interna, queste variazioni si possono considerare in qualche modo indicative della sua tendenza, in mancanza del dettaglio infrannuale sulle immatricolazioni. Nella stessa relazione, per spiegare il calo produttivo nel secondo semestre, viene assunta come ipotesi l'influenza data dall'attesa di nuovi modelli[2] e di ribassi di prezzo, data la ripresa della produzione ad inizio '66.

La fase di forte crescita della motorizzazione in Italia dagli anni '50 fino al 1974 non si spiega soltanto con l'aumento del reddito delle famiglie e con le scelte politiche, pur importanti, volte a favorire l'uso dell'auto. Infatti, tra il 1952 e il 1972 si registravano nell'Italia meridionale (in condizioni economiche decisamente svantaggiate rispetto al resto del Paese) tassi di incremento dei veicoli circolanti in generale superiori rispetto all'Italia settentrionale. Secondo Federico Paolini SISTEMARE CON BIBLIOG(*Un paese a quattro ruote: automobili e società in Italia*), il fenomeno ha tratto forza soprattutto da un contesto sociale e culturale che considerava ormai l'automobile come mezzo indispensabile (tra i cui vantaggi, ad esempio, si sosteneva anche quello di poter rientrare prima dal lavoro e quindi di dedicare più tempo alla famiglia, secondo un'inchiesta di Quattroruote del 1962) ma anche come elemento gratificante, simbolo di benessere e di prestigio, tale per cui anche famiglie che non si trovavano in una situazione economicamente soddisfacente decidevano di acquistarla.

[2]Probabilmente l'imminente uscita nel 1966 del modello *124* della *Fiat*, costruttore che deteneva la quota di mercato più ampia. La 124 fu una vettura di successo, pensata principalmente per la piccola borghesia.

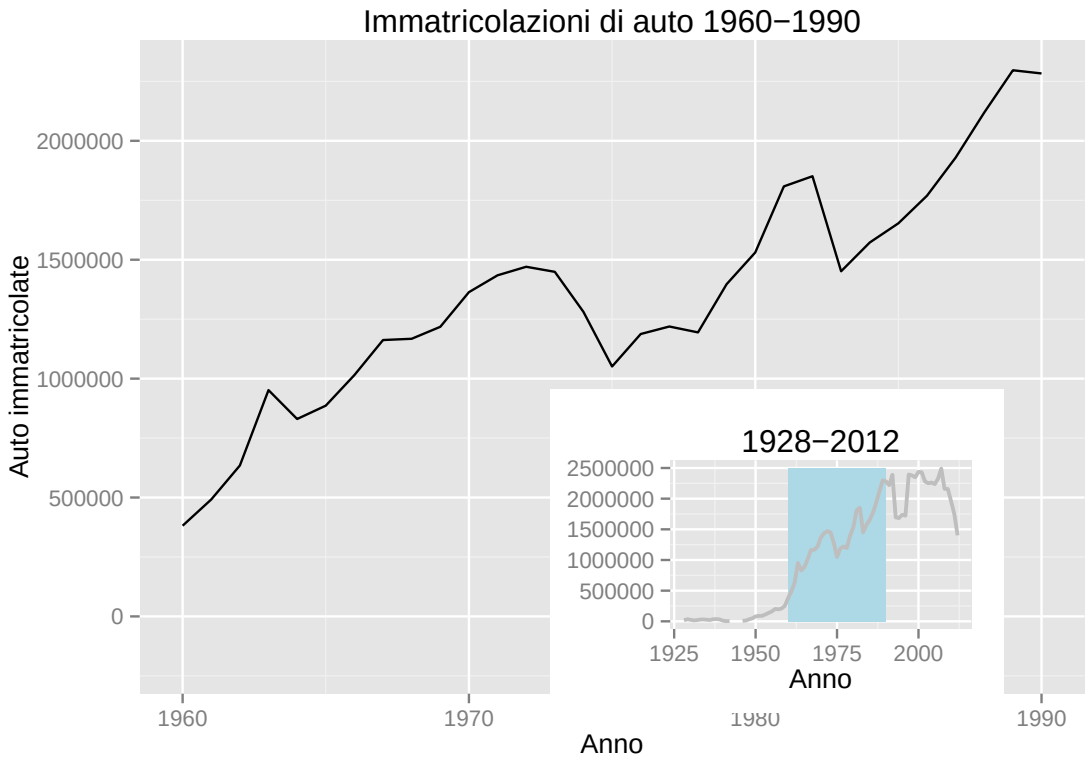


Figura 1.3: Immatricolazioni di auto in Italia 1960-1990

Il primo cedimento di rilievo arrivò conseguentemente alla crisi energetica scoppiata al termine del 1973 che provocò una forte recessione. Il numero di autovetture immatricolate precipitò fino al 1975, recuperò leggermente nel 1976 e rimase in una fase di stallo prima di ripartire con decisione nel 1979. Proprio nel '79 avvenne un secondo *shock* petrolifero che provocò conseguenze nei primissimi anni '80. Questa ulteriore crisi energetica, nell'immediato, non ebbe un'impatto forte come la precedente. Gli effetti tuttavia si distribuirono negli anni successivi, provocando periodi di stagnazione e di recessione, con un tasso di disoccupazione in crescita; dopo una resistenza alla congiuntura sfavorevole all'inizio del decennio, nel 1983 si ebbe un crollo di circa 400000 unità immatricolate. I segnali di ripresa dell'economia al termine dello stesso anno e la crescita nel periodo seguente, permisero una ripartenza del mercato dell'auto già dal 1984.

Il 1989 sembra essere l'ultimo anno di un trend di crescita di lungo termine.

Successivamente si sono riscontrati periodi di crescita e decrescita, oltre a periodi di stagnazione. All'inizio degli anni '90 vi fu un rallentamento dell'economia, in parte anche come effetto del clima di incertezza dovuto alla *Guerra del Golfo*. Fu però nel 1992 che l'Italia entrò in recessione, contemporaneamente ad una fase di profonda crisi politica.[3] Il mercato dell'auto in Italia subì l'anno successivo il più forte crollo mai registrato e rimase per anni ad un livello nettamente inferiore rispetto ai massimi storicamente registrati. Un rimbalzo si ebbe solo nel 1997, favorito dall'introduzione del contributo statale per l'acquisto di autoveicoli nuovi a fronte della rottamazione di veicoli immatricolati precedentemente al 1 gennaio 1987.

Negli anni seguenti, tuttavia, si è verificò una nuova tendenza al ribasso delle vendite, provocando nuovamente il ricorso a forme di incentivazione per ravvivare il mercato, a causa di nuove fasi di rallentamento e stagnazione dell'economia che hanno percorso parte del primo decennio degli anni 2000. Del 2002 sono gli "ecoincentivi" rivolti all'acquisto di auto con potenza non inferiore a 85 *Kw* e conformi alle direttive CE sull'inquinamento, sotto forma di esenzione dal pagamento dell'*Imposta Provinciale di Trascrizione* (IPT) e dalla *Tassa Automobilistica* ("Bollo") per il primo periodo fisso e per i successivi due anni, condizionatamente alla rottamazione di un veicolo non conforme alla direttiva CE n. 91/441 e successive. Tale contributo ha avuto validità per un periodo limitato, ovvero da luglio 2002 a dicembre del medesimo anno con un prolungamento nei primi tre mesi del 2003. Non sembra aver avuto effetti espansivi nel periodo di validità. In aggiunta, i costruttori hanno integrato i benefici previsti dal Governo con promozioni speciali volte ad aumentare i vantaggi economici per l'acquirente, prolungandole per periodi superiori a quanto previsto dall'incentivo statale.

[3]Nel 1992 ci fu un attacco speculativo alla Lira (così come verso la Sterlina britannica) che costrinse l'allora governo Amato ad attuare una pesante svalutazione e che portò l'uscita della valuta italiana dal *Sistema Monetario Europeo* (SME). Oltre a ciò si aggiunsero una forte crisi politica in seguito allo scandalo *"Tangentopoli"*, e un clima di insicurezza in seguito a numerosi attentati di origine mafiosa.

L'effetto che ne è risultato è stato probabilmente quello di bloccare un'ulteriore discesa. Più vantaggiosi sono stati gli ecoincentivi del 2007. Essi prevedevano un contributo per l'acquisto di auto nuove[4] di 800 euro e l'esenzione del pagamento del bollo per le prime due annualità.[5] Durante questo periodo, si è avuto un'incremento delle immatricolazioni arrivando circa a 2.5 milioni di veicoli; nei due anni successivi gli incentivi sono stati riproposti con condizioni differenti, con volumi che si sono mantenuti più bassi.

Terminato l'effetto dei contributi statali con un picco di immatricolazioni nel marzo 2010, il mercato italiano è entrato in una fase di pesante decrescita, in concomitanza con la crisi economica internazionale tuttora in corso. Nonostante la domanda a livello mondiale sia in crescita, il mercato dell'Unione Europea risulta complessivamente in calo [6]. L'Italia chiude il 2012 con un -19.9%.

[4] *Euro 4* oppure *Euro 5* con emissioni non oltre i 140 g/Km di CO2.

[5] Esenzione elevata a tre annualità in caso di vetture con cilindrata inferiore o uguale a 1300 cc oppure per nuclei familiari di almeno 6 persone non intenstatarie di altri veicoli.

[6] I dati provvisori sul 2012, elaborati da UNRAE, indicano un -16.3% rispetto al 2011. Pochi Stati, soprattutto nell'area orientale, presentano un segno positivo nell'area UE, tra i quali spicca il $+17.6\%$ dell'Ungheria. Tra i Paesi dell'Europa occidentale, solo il Regno Unito ($+5.3\%$), Lussemburgo ($+1\%$) e Danimarca ($+0.4\%$) presentano un segno positivo.

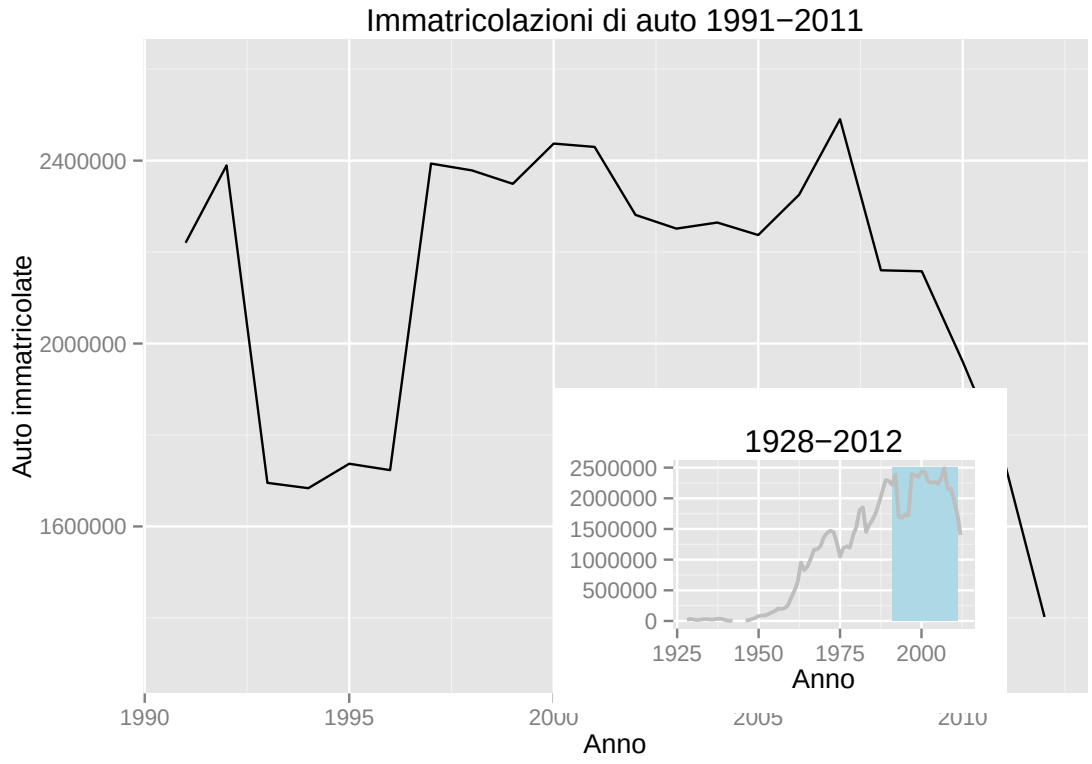


Figura 1.4: Immatricolazioni di auto in Italia 1991-2011

Capitolo 2

ANALISI PRELIMINARE E STIMA DI UN MODELLO STOCASTICO STAGIONALE

2.1 Introduzione

In questo capitolo, dopo aver effettuato un'analisi classica della serie storica delle immatricolazioni attraverso la scomposizione in *trend*, stagionalità ed errore. Successivamente si stimeranno dei modelli stocastici univariati di tipo ARIMA stagionale (SARIMA) sulla trasformazione logaritmica dei dati. Infine si valuterà la bontà di questi modelli in termini di accuratezza delle previsioni.

2.2 Serie mensile 1985-2011

L'arco temporale considerato nella serie delle immatricolazioni va dal gennaio 1985 al dicembre 2011, per un totale di 324 osservazioni mensili.

Prima di affrontare l'analisi dei dati storici attraverso un approccio stocastico, è utile descrivere la serie attraverso una scomposizione classica di quelle che sono le componenti di *trend* (T_t), ciclo (C_t), stagionalità (S_t) ed errore (E_t). Tale

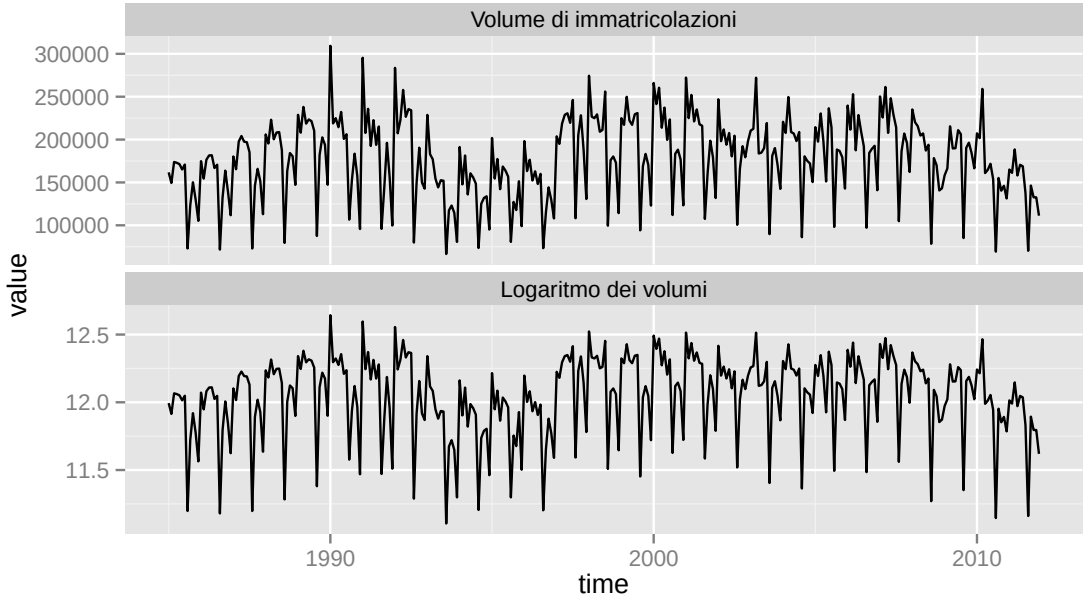


Figura 2.1: Immatricolazioni: serie mensile 1985-2011

scomposizione può essere di tipo additivo (2.1) o moltiplicativo (2.2).

$$Y_t = T_t + C_t + S_t + E_t \tag{2.1}$$

$$Y_t = T_t \cdot C_t \cdot S_t \cdot E_t \tag{2.2}$$

Il modello moltiplicativo è tipico di molti fenomeni economici per i quali ad un cambiamento di livello della serie corrisponde un cambiamento di ampiezza delle oscillazioni stagionali. È lecito pensare che anche la stagionalità del volume di immatricolazioni di auto possa presentare un'andamento sensibile al livello di vendite, con oscillazioni via via crescenti all'aumentare dei volumi. Tuttavia, per osservare in modo evidente questo effetto è necessario disporre di una serie mensile con un arco temporale più lungo, tale da includere la fase di espansione del mercato che ha avuto inizio a partire dal secondo dopoguerra. La scelta di un modello moltiplicativo può essere ricondotta ad un modello adittivo tramite la trasformazione logaritmica dei dati, qualora essi fossero strettamente positivi.

Per *trend* si intende la tendenza di lungo periodo della serie, nell'arco di tempo considerato. Esso può denotare una crescita o una decrescita oppure avere un

andamento stazionario. Le oscillazioni periodiche dovute a fattori di calendario dovuti al trascorrere delle stagioni rappresentano la stagionalità della serie. Essa risulta evidente in fenomeni rilevati con cadenza infrannuale.

La componente di ciclo, a differenza della stagionalità, rappresenta delle oscillazioni che si verificano con una cadenza e con una durata non regolare nel tempo. Essa risente delle fasi di espansione e di contrazione presenti nel sistema economico. La natura, la durata e l'ampiezza di queste fluttuazioni sono determinate da una realtà complessa e numerose teorie economiche cercano di interpretarle sotto punti di vista diversi; tuttavia risulta essere una componente di difficile determinazione a causa della sua irregolarità. In questa analisi descrittiva ciclo e *trend* sono considerati congiuntamente (componente *ciclo-trend*). Pertanto le (2.1) e (2.2) si semplificano nelle (2.3) e (2.4).

$$Y_t = T_t + S_t + E_t \tag{2.3}$$

$$Y_t = T_t \cdot S_t \cdot E_t \tag{2.4}$$

dove T_t è la componente *ciclo-trend*, che in seguito per brevità verrà chiamata semplicemente *trend*.

Nel grafico di figura 2.2 è illustrata la scomposizione del logaritmo della serie in *trend*, stagionalità e componente residuale ottenuta con il comando *stl* presente nel pacchetto *stats*. La componente di stagionalità è stata estratta attraverso una regressione locale polinomiale di tipo *loess* con una finestra stagionale di 27 *lag*[1].

2.2.1 Trend

Osservando la serie appare evidente che, nel periodo considerato, il *trend* non possa essere approssimato efficacemente da una retta attraverso una regressione

[1] I lag sono stati selezionati attraverso il comando *stlId* del pacchetto *ast* (Guido Masarotto, `http://sirio.stat.unipd.it/index.php?id=libast`). Il comando individua i *lag* per la finestra stagionale, se non indicati attraverso il parametro *s.window*, minimizzando la somma delle autocorrelazioni al quadrato della componente residuale, calcolate su diverse finestre.

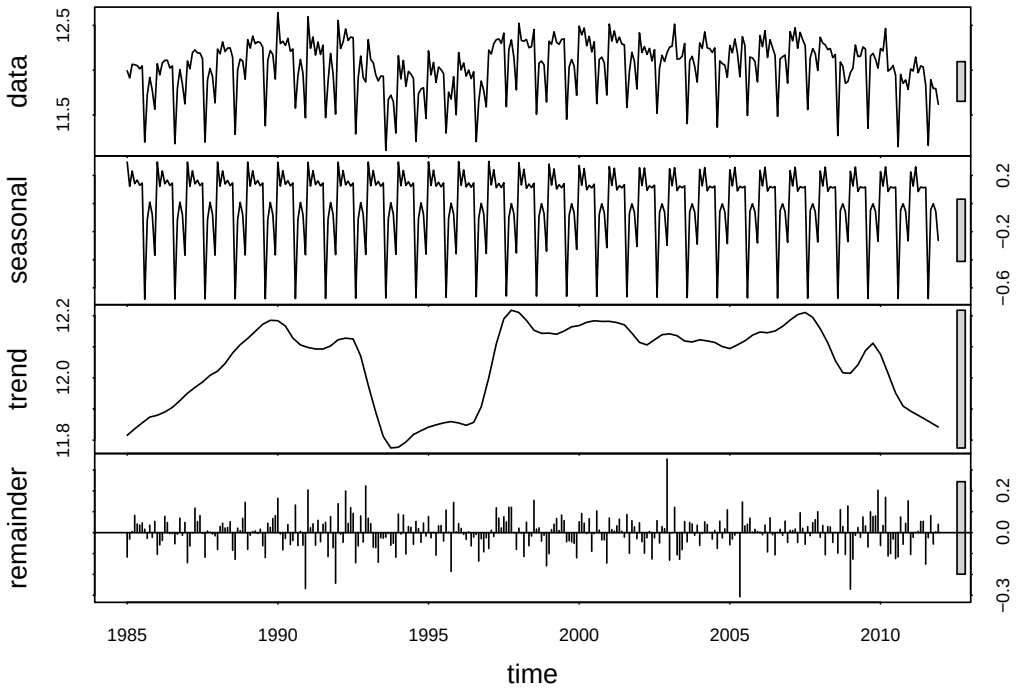


Figura 2.2: Scomposizione logaritmo serie mensile 1985-2011

lineare semplice, poiché essa non è in grado di cogliere periodi in cui sono presenti variazioni di livello temporanee ma di rilievo. Questo approccio descrittivo della tendenza del fenomeno, pur privo dell'effetto stagionale, appare poco rappresentativo di quelle che sono state le dinamiche del mercato, soprattutto negli intervalli di medio periodo. Pur non essendoci grosse indicazioni circa una non stazionarietà complessiva, la serie mostra periodi soggetti a variazioni considerevoli rispetto alla media. Si nota chiaramente come durante la seconda metà degli anni '80 il mercato fosse in crescita, prima di subire un cambio di livello netto durante la crisi degli anni '90 per poi nuovamente tornare a volumi pre-crisi dal 1997; in seguito esso appare più stabile se paragonato al pesante crollo degli anni precedenti, pur mostrando periodi di lenta decrescita e ripresa. Questo comportamento può essere inteso come segnale di un mercato che da anni è giunto (o è prossimo) ai massimi volumi raggiungibili di vendite, ovvero ad una situazione

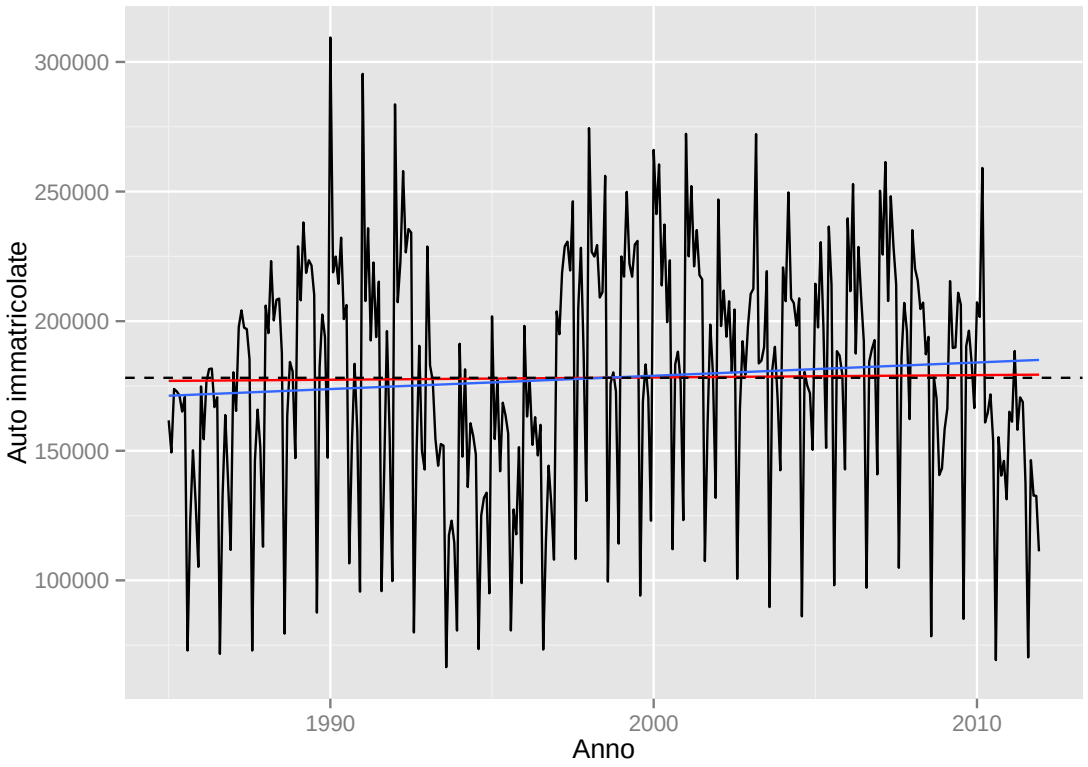


Figura 2.3: Rette di regressione sulle immatricolazioni mensili 1985-2011

di saturazione con una domanda di sostituzione. Considerando che l'economia italiana da anni si trova in una fase di bassa (a volte nulla) crescita rispetto al passato, risulta difficile immaginare il raggiungimento di livelli più alti rispetto a quelli già raggiunti, a meno di temporanei rimbalzi dopo periodi di profondo decremento dell'immatricolato, favoriti magari da forme di contribuzione statale (come avvenne nel 1997); contributi che, tuttavia, non è detto provochino forti incrementi nelle vendite, essendo indubbiamente necessaria una serie di condizioni favorevoli nel sistema economico e sociale del momento.

2.2.2 Stagionalità

Dai grafici di figura 2.1, appare ragionevole pensare che le immatricolazioni di autoveicoli risentano di un effetto stagionale, che si ripete con cadenza annuale. Il minor numero di veicoli immatricolati durante l'intero anno si registra nel mese

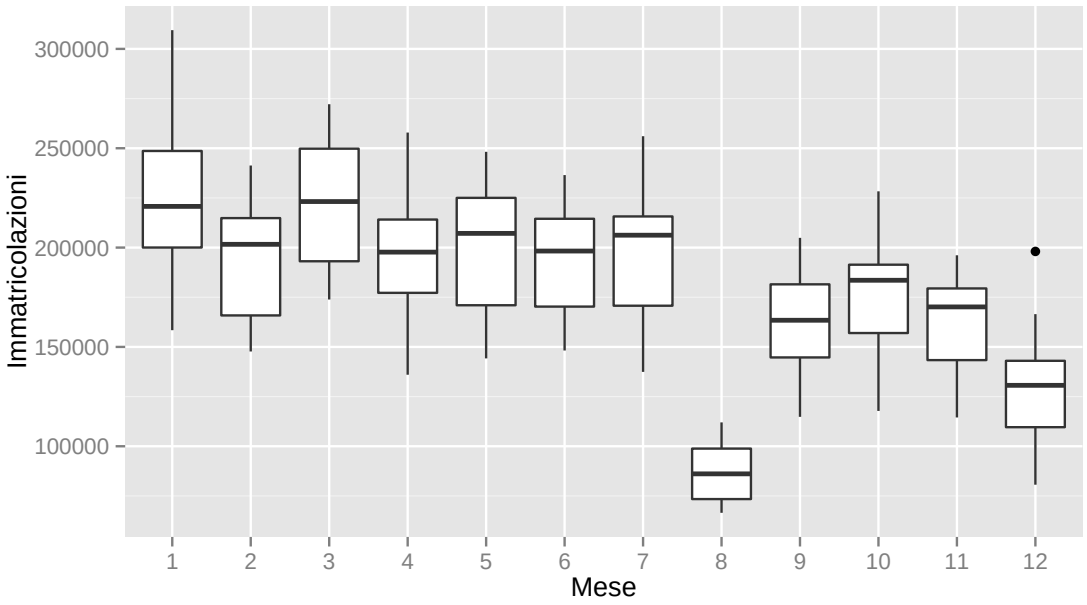


Figura 2.4: Boxplot mensili

di agosto (figura 2.4). I mesi da settembre a dicembre mostrano un numero di immatricolazioni mediamente inferiore rispetto al periodo precedente ad agosto.

Attraverso la rappresentazione della dispersione della serie rispetto ai primi 12 ritardi (figura 2.5), è possibile intuire la presenza di una marcata dipendenza lineare della serie con il dodicesimo *lag*. Ulteriore prova di una correlazione stagionale è visibile osservando le autocorrelazioni globali nel grafico di figura 2.6, dove si hanno picchi positivi ogni 12 mesi. Il grafico di figura 2.7 mostra l'andamento della stagionalità nel corso degli anni.

Ulteriori possibili fonti di variabilità legate al calendario possono derivare dal diverso numero di giorni lavorativi presenti in un mese oppure dalla presenza di festività fisse o mobili. In tal caso diverrebbe opportuno un aggiustamento della serie utilizzando apposite variabili correttive degli effetti di calendario. Esistono tuttavia alcuni motivi che fanno ritenere questa operazione, seppur in modo arbitrario, non strettamente necessaria per la presenza di ulteriori elementi da ritenersi aleatori per questa analisi. Innanzitutto sarebbe necessario verificare,

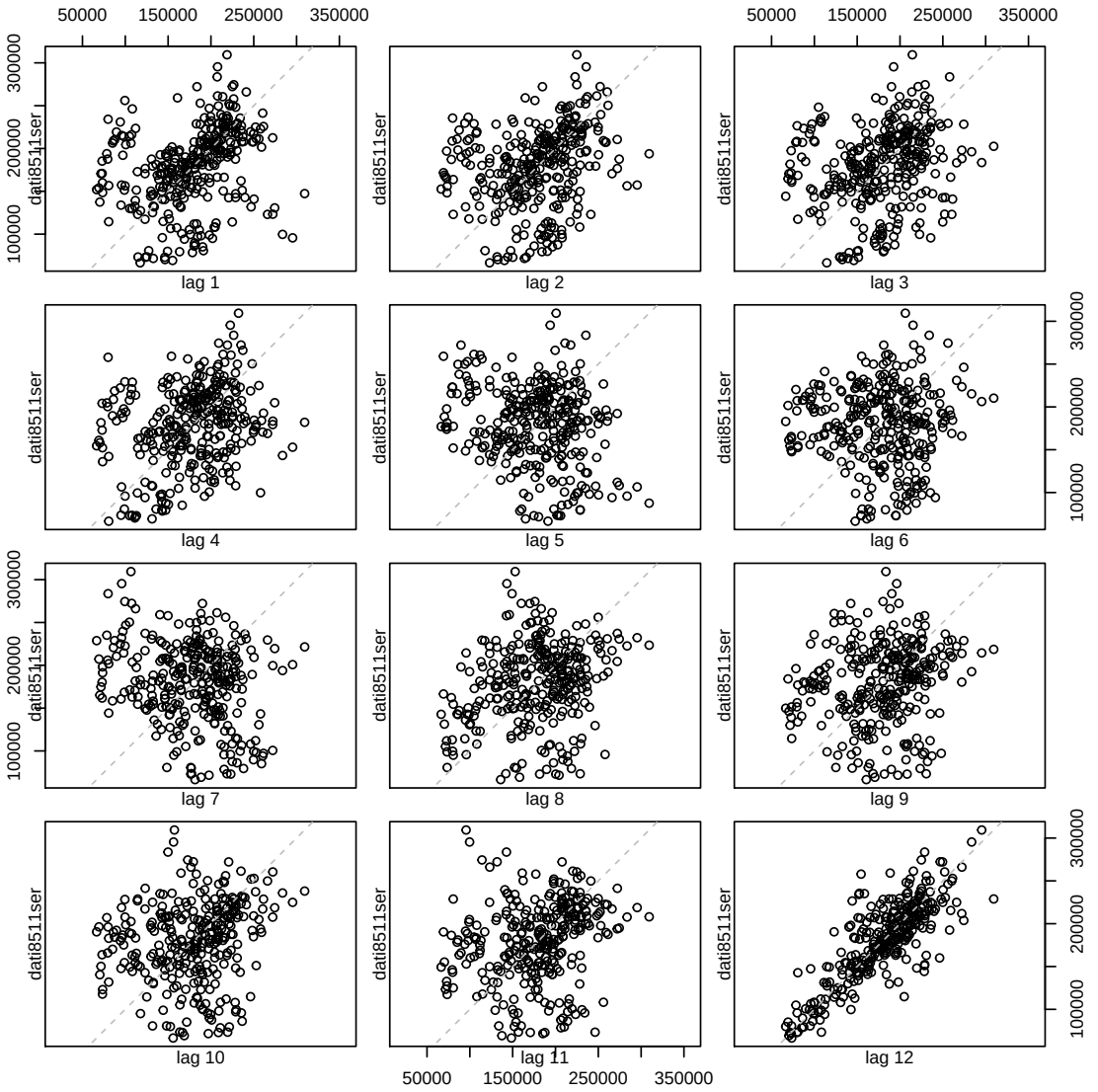


Figura 2.5: Autodispersione della serie mensile

avendo a disposizione dati più dettagliati ma non facilmente accessibili per costo o per riservatezza, quanto sia rilevante l'incidenza del numero di giorni lavorativi in un mese. Il mercato dell'auto non è paragonabile, ad esempio, a quello di generi di largo consumo con flussi di acquisti maggiori, dove un incremento di giorni di apertura degli esercizi può portare globalmente ad un totale mensile di vendite sensibilmente maggiore (a parità di condizioni del sistema economico). In secondo luogo, durante periodi legati a promozioni o all'arrivo di nuovi modelli nel mercato, spesso vengono organizzate aperture straordinarie dei *dealer*

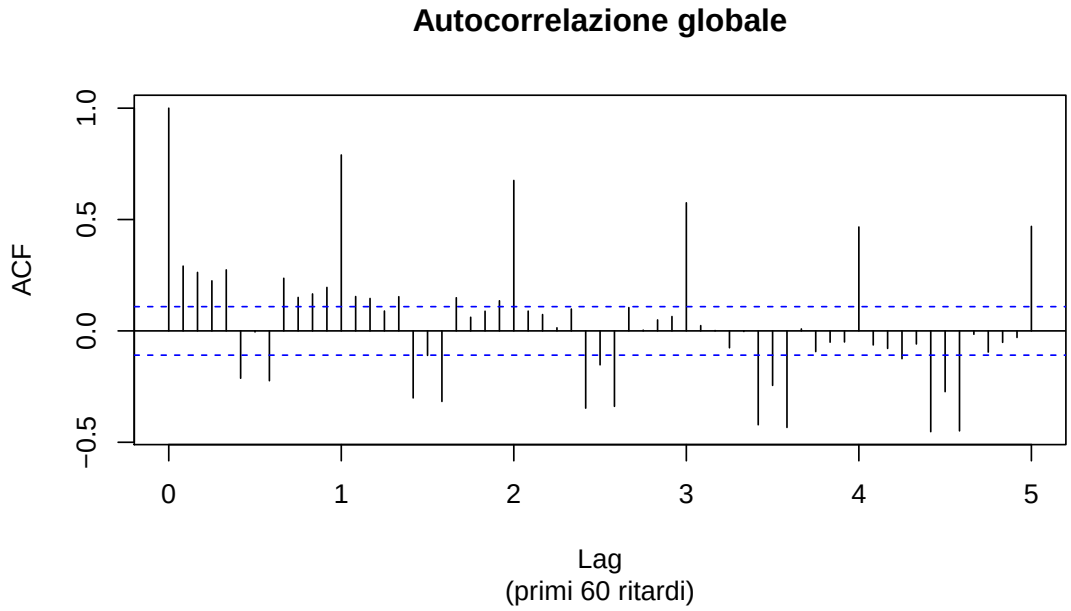


Figura 2.6: Grafico di autocorrelazione globale

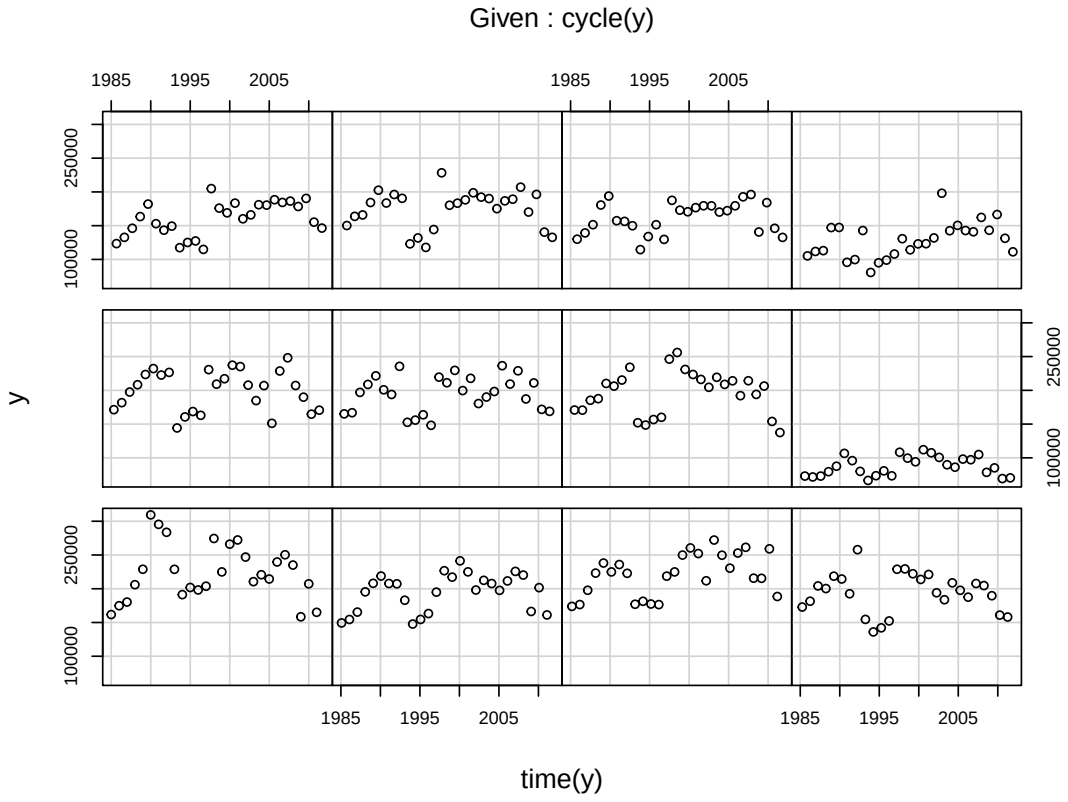


Figura 2.7: Andamento della stagionalità negli anni

(giornate "porte aperte"). Il momento dell'immatricolazione del veicolo, inoltre, non coincide temporalmente con la stipula del contratto di vendita ma può avvenire con tempistiche diverse a seconda del tempo di evasione degli ordini da parte del costruttore[2]. Infine, fatto di rilievo, una percentuale considerevole delle immatricolazioni avviene negli ultimi tre giorni del mese.[3] Questo fenomeno, molto spesso cela in parte al suo interno anche il ricorso ad immatricolazioni di auto che verranno vendute successivamente a *chilometri zero (Km 0)*.[4] Le *Km 0* sono vetture registrate dal *dealer* stesso che successivamente rivende ai clienti a prezzo scontato; questa pratica viene molto spesso utilizzata per raggiungere determinati obiettivi di volumi di vendita stabiliti dalla casa madre per sostenere la quota di mercato [Volpato, 2011]; è possibile quindi che questo fenomeno possa rendere meno rilevante l'effetto del differente numero di giorni lavorativi presenti in un mese, così come essere esso stesso una *proxy* utile per fare previsioni.[5] Pertanto si è scelto di non effettuare correzioni di calendario in questa analisi.

2.3 Un modello stocastico per le immatricolazioni mensili

Come prima analisi, viene calcolato un modello stocastico univariato che tenga conto della stagionalità presente nella serie. Ciò significa che si proverà a modellare le immatricolazioni di auto, seguendo un approccio ti tipo *Box-Jenkins*, con

[2] sui tempi di consegna possono incidere, oltre a fattori strettamente legati al processo produttivo, eventi quali, ad esempio, scioperi interni o nelle aziende dell'indotto, danni agli stabilimenti o eventi naturali straordinari.

[3] All'interno del periodico InterAutoNews vengono riportate le percentuali di immatricolazioni degli ultimi 3 giorni di ciascun mese e la media annuale. Nel 2012 la quota di autovetture immatricolate negli ultimi 3 giorni del mese rispetto al totale è stata in media del 35.48%. Nel dettaglio dei principali marchi, durante il mese di marzo 2012 si va da un minimo del 23.04% di Volkswagen ad un massimo del 57.17% di Fiat.

[4] In Italia la vendita di vettura a *Km 0* supera le 130.000 unità all'anno, oltre il 5% del mercato [Volpato, 2011].

[5] Centri di ricerca specializzati utilizzano insiemi di variabili settoriali tra le quali vengono incluse anche le autoimmatricolazioni e soprattutto le previsioni di vendita delle stesse case automobilistiche (si veda ad esempio il primo numero dell'osservatorio *Previsioni & Mercato* del Centro Studi UNRAE pubblicato ad Aprile 2011), poiché queste ultime possono influenzare gli obiettivi di vendita da raggiungere.

un modello ARIMA stagionale (o SARIMA), che servirà come base di riferimento per il confronto di modelli più complessi. Data la presenza di diverse variazioni di livello significative nell'andamento della serie nei primi 12 anni a partire dal 1985, si è scelto di utilizzare i dati dall'aprile del 1998, mese in cui si esaurisce l'effetto di spinta degli incentivi statali per la rottamazione del 1997 che hanno permesso un rimbalzo dopo la crisi iniziata nel 1992. Come *trainging-set* su cui effettuare le stime dei modelli viene considerato l'insieme delle osservazioni fino dicembre 2010.

2.3.1 Verifica della stazionarietà

Punto di partenza per giungere alla formulazione di un modello stocastico della serie è la verifica della sua stazionarietà. Per fare ciò si possono utilizzare dei test per la verifica della presenza di radici unitarie, come ad esempio l'*Augmented Dikey-Fuller test* (ADF), oppure un test di stazionarietà come il test di *Kwiatkowski-Phillips-Schmidt-Shin* (KPSS).

Il test ADF, variante del test Dikey-Fuller che permette di considerare correlazioni seriali di ordine superiore al primo, verifica l'esistenza di una radice unitaria ipotizzando un processo autoregressivo. I valori critici della statistica test per i principali livelli di significatività sono tabulati rispetto ad alcune numerosità campionarie; sono diversi a seconda del modello ipotizzato per i dati (senza costante, con costante oppure con costante e trend deterministico lineare) e dell'ipotesi nulla considerata. Si è scelto di effettuare il test ipotizzando per il periodo considerato un modello con costante, privo di trend deterministico; tale scelta si è effettuata attraverso una sequenza di test di verifica di ipotesi a partire dal caso più generale di presenza di trend e intercetta.[6] Poiché la serie presenta un andamento che risente di un effetto stagionale, è possibile effettuare il test ADF includendo 12 ritardi. Il risultato porta ad accettare l'ipotesi nulla di non

[6]Procedura descritta nello schema a pag. 194 in [Di Fonzo and Lisi, 2005]

stazionarietà; la statistica test del sistema d'ipotesi unilaterale che verifica che in un modello $Y_t = \phi_0 + \phi_1 Y_{t-1} + u_t$, con $\{u_t\}$ processo a componenti correlate, il coefficiente ϕ_1 sia uguale a 1 contro l'ipotesi che $|\phi_1| < 1$, è infatti superiore al valore critico tabulato τ_2. Pertanto si è portati a ritenere che vi sia la presenza di una non stazionarietà. Anche il secondo test, una statistica F che verifica l'ipotesi $(\phi_0, \phi_1) = (0, 1)$, porta all'accettazione della non stazionarietà poiché si ottiene un risultato superiore ai valori critici Φ_1 (6.52, 4.63, 3.81 rispettivamente ai livelli di significatività 1%, 5% e 10%). Di seguito i risultati con le due statistiche calcolate per la verifica delle ipotesi nulle, nell'ordine $|\phi_1| = 1$ e $(\phi_0, \phi_1) = (0, 1)$.

```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression drift
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3794 -0.0667  0.0169  0.0601  0.3769
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.9682     2.6610    0.74  0.46088
## z.lag.1       -0.1630     0.2193   -0.74  0.45879
## z.diff.lag1   -0.4840     0.2341   -2.07  0.04075 *
## z.diff.lag2   -0.3804     0.2221   -1.71  0.08923 .
## z.diff.lag3   -0.3282     0.2092   -1.57  0.11914
## z.diff.lag4   -0.2813     0.1984   -1.42  0.15875
## z.diff.lag5   -0.4256     0.1818   -2.34  0.02077 *
## z.diff.lag6   -0.4666     0.1628   -2.87  0.00486 **
## z.diff.lag7   -0.5199     0.1464   -3.55  0.00054 ***
## z.diff.lag8   -0.4289     0.1343   -3.19  0.00177 **
## z.diff.lag9   -0.3512     0.1261   -2.79  0.00618 **
## z.diff.lag10  -0.4218     0.1150   -3.67  0.00036 ***
## z.diff.lag11  -0.4215     0.1038   -4.06  8.6e-05 ***
## z.diff.lag12   0.3456     0.0872    3.96  0.00012 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.123 on 126 degrees of freedom
## Multiple R-squared:  0.894,Adjusted R-squared:  0.883
## F-statistic: 81.5 on 13 and 126 DF,  p-value: <2e-16
##
##
## Value of test-statistic is: -0.7431 0.6609
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau2 -3.46 -2.88 -2.57
## phi1  6.52  4.63  3.81
```

Il test di stazionarietà KPSS permette un'ulteriore verifica. Il test ADF, infatti, spesso non è in grado di rifiutare l'ipotesi nulla di non stazionarietà in presenza di situazioni che vi si avvicinano, non essendo sufficientemente potente [Kwiatkowski et al., 1992]; esso può non portare al rifiuto della presenza di una radice unitaria quando in realtà si ha una stazionarietà attorno ad un trend deterministico. Il test KPSS ipotizza la serie come la somma di un trend temporale deterministico, di un processo *random walk* e di residui stazionari e può essere utilizzato per verificare un processo stazionario attorno ad un livello oppure attorno ad un *trend*. Nel primo caso si ottiene un *p-value* <0.1, mentre nel secondo risulta pari a 0.1, quindi porta a rigettare l'ipotesi nulla di stazionarietà.

Gli stessi test si possono applicare sulla differenza prima stagionale della serie. Questa volta i valori del test porterebbero a rigettare la non stazionarietà. Per l'ADF, applicando la stessa sequenza di test a partire dal caso generale (*trend* + intercetta), fin da subito vengono rifiutate le ipotesi di non stazionarietà $\phi_1 = 1$ e $(\phi_1, \beta) = (1, 0)$, con β coefficiente di un *trend* deterministico. Il risultato è il seguente:

```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
```

```
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2735 -0.0614  0.0024  0.0691  0.3899
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.019084   0.022768    0.84  0.40371
## z.lag.1      -1.021856   0.203953   -5.01  2.0e-06 ***
## tt           -0.000432   0.000274   -1.57  0.11807
## z.diff.lag1   0.123349   0.182126    0.68  0.49962
## z.diff.lag2   0.390635   0.169294    2.31  0.02285 *
## z.diff.lag3   0.688033   0.157990    4.35  2.9e-05 ***
## z.diff.lag4   0.615113   0.161885    3.80  0.00024 ***
## z.diff.lag5   0.411394   0.164928    2.49  0.01406 *
## z.diff.lag6   0.396121   0.161019    2.46  0.01540 *
## z.diff.lag7   0.595343   0.147942    4.02  0.00010 ***
## z.diff.lag8   0.598380   0.143380    4.17  5.9e-05 ***
## z.diff.lag9   0.491412   0.150685    3.26  0.00147 **
## z.diff.lag10  0.301882   0.152455    1.98  0.05012 .
## z.diff.lag11  0.324875   0.135579    2.40  0.01821 *
## z.diff.lag12  0.053545   0.099327    0.54  0.59090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.111 on 113 degrees of freedom
## Multiple R-squared:  0.556,Adjusted R-squared:  0.501
## F-statistic: 10.1 on 14 and 113 DF,  p-value: 2.41e-14
##
##
## Value of test-statistic is: -5.01 8.51 12.69
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.99 -3.43 -3.13
## phi2  6.22  4.75  4.07
## phi3  8.43  6.49  5.47
```

Il test KPSS invece riporta un risultato limite che porterebbe (al 5% di significatività) a rifiutare la stazionarietà attorno ad un livello, con un *p-value* =0.1; al contrario, il test induce ad accettare l'ipotesi di stazionarietà attorno ad un *trend*,

con un *p-value*>0.1. Pertanto si è in una situazione in cui uno dei test non concorda con il test ADF e quindi i dati non sono sufficientemente informativi per stabilire con sicurezza la stazionarietà della serie (Kiawt CITARE) con differenziazione dodicesima.

2.3.2 Funzioni di autocorrelazione

Tramite l'osservazione dei correlogrammi generati attraverso le funzioni di autocorrelazione globale (*ACF*) e parziale (*PACF*) è possibile provare ad ipotizzare la struttura di un modello stocastico. In figura 2.8 e 2.9 si osservano i valori delle autocorrelazioni globali e parziali della serie senza e con differenziazione stagionale. La presenza di stagionalità nella serie del logaritmo dei volumi di immatricolazioni trova conferma da autocorrelazioni globali significative che si ripetono ciclicamente ogni 12 ritardi. Una volta differenziata la serie (la quale assume ora il significato di variazioni stagionali delle immatricolazioni, calcolate sempre a partire dalla trasformazione logaritmica), esse scompaiono; la differenziazione porta in questo caso come effetto un'autocorrelazione pari a -0.4. In generale se risulta una correlazione di valore -0.5 o superiore in valore assoluto dopo la differenziazione stagionale, potrebbe significare che non è necessario differenziare la serie; nel caso in esame si ritiene comunque necessaria.

La specificazione di un modello stocastico, identificando le componenti autoregressiva e media mobile attraverso l'analisi delle funzioni di autocorrelazione, non è particolarmente agevole in presenza di modelli misti, nei quali oltretutto è presente anche una componente stagionale. Da segnalare inoltre la presenza nella serie di periodi caratterizzati da interventi esterni (incentivi statali) e da particolari congiunture del sistema economico. Tali disturbi alla serie verranno trattati nel prossimo capitolo; in questa parte ci si è limitati a stimare un modello "grezzo", ovvero senza ulteriori variabili esplicative, pur nella consapevolezza dei limiti che un tale approccio inevitabilmente può comportare.

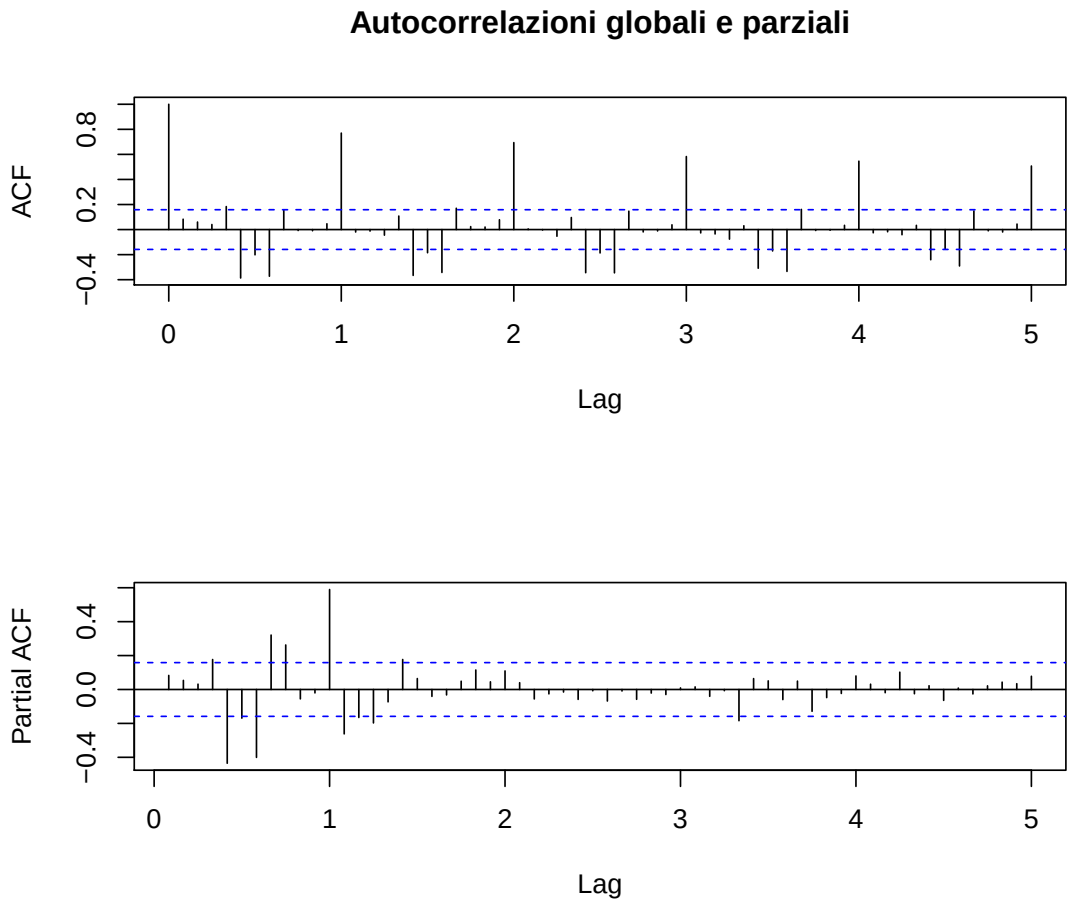


Figura 2.8: Correlogrammi della serie ridotta (logaritmi)

Dai correlogrammi dell'*ACF* e della *PACF* sui dati differenziati, si nota una possibile componente autoregressiva di ordine 3 a livello non stagionale. Sono presenti alcune correlazioni significative a ritardi superiori. Per definire l'ordine più appropriato per il modello ARIMA stagionale si rende necessaria una serie di tentativi, proponendo più modelli candidati. In un'analisi puramente descrittiva dei dati storici generalmente si privilegia il modello che presenta migliori statistiche AIC, AICc o BIC. Se lo scopo è la stima di valori futuri allora la scelta si basa sugli errori di previsione e ciò verrà visto nella sezione 2.4.

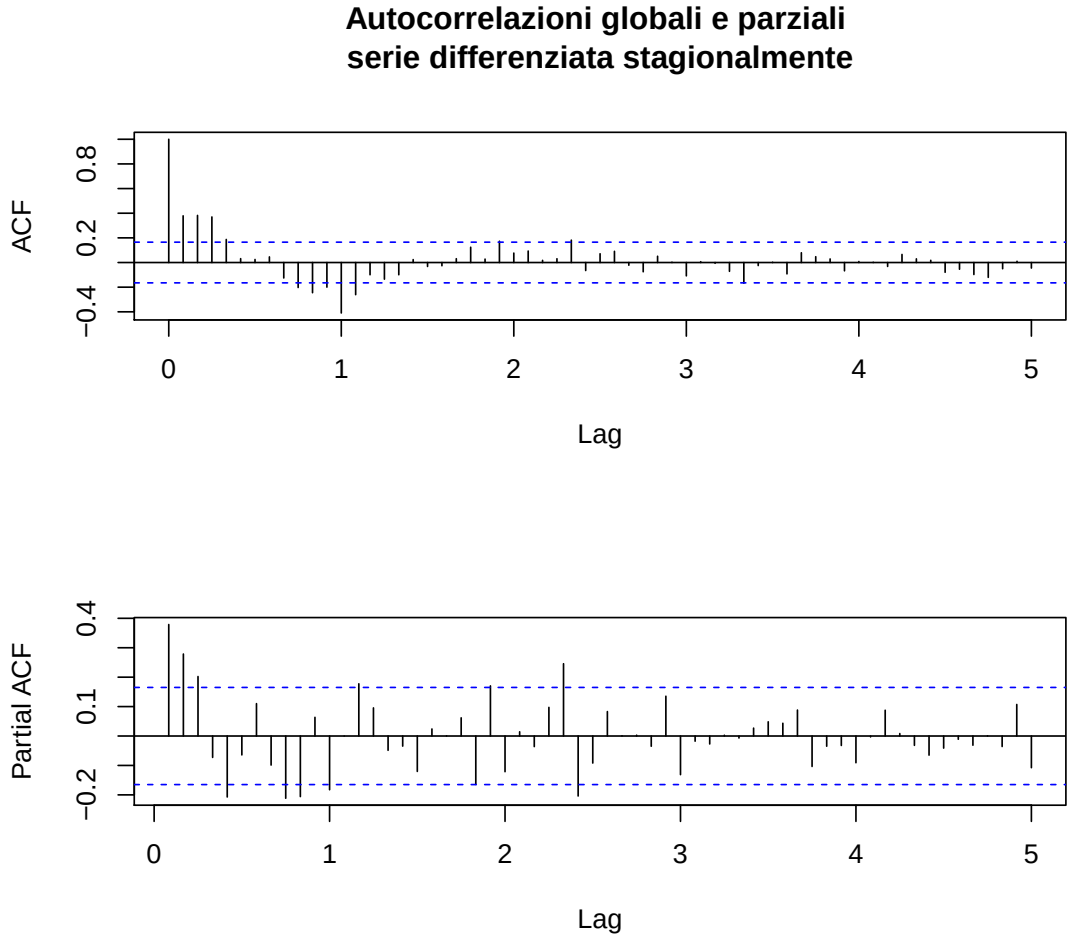


Figura 2.9: Correlogrammi per la differenza stagionale della serie ridotta (logaritmi)

2.3.3 Stima di modelli *SARIMA*

I modelli *SARIMA(p, d, q)×(P, D, Q)$_s$*, detti anche modelli *SARIMA*, sono delle estensioni del modello *ARIMA* che permettono di stimare i coefficienti di un modello stocastico per una serie stagionale. La formulazione generale di un modello *SARIMA* (2.5) non è altro che un modello *ARIMA* con l'aggiunta di fattori $\Phi_P(B^s)$ e $\Theta_Q(B^s)$ che rappresentano rispettivamente le componenti autoregressive e media mobile stagionali.

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D\dot{Z}_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{2.5}$$

con

$$\dot{Z} = \begin{cases} Z_t - \mu, & se\ d = D = 0 \\ Z_t, & altrimenti \end{cases}$$

Nella (2.5) il numero dei parametri della componente autoregressiva sono p non stagionali e P stagionali, mentre per la parte media mobile si considerano q e Q parametri; nel caso in cui si rendessero necessarie delle differenziazioni per rendere stazionario il processo, anch'esse si possono specificare sia per la parte non stagionale, d, sia per la parte stagionale, D.

Come modello iniziale per i dati si è deciso di partire da un *SARIMA(3,0,0)×(0,1,0)*$_{12}$. I risultati in tabella 2.1. In figura 2.10 sono rappresentati i grafici dei residui standardizzati e delle loro funzioni di autocorrelazione. È possibile notare la presenza di un *lag* significativo al ritardo 12. In figura 2.11 sono raffigurati i *p-values* del test *portmanteau* di tipo *Ljung-Box* fino ad un ritardo pari a 2 volte la stagionalità (24 mesi), con i gradi di libertà calcolati sottraendo ai ritardi considerati il numero di parametri del modello (3 in questo caso). Il risultato porterebbe a considerare la presenza di autocorrelazioni tra i residui a partire dal dodicesimo ritardo, il che significa che nella parte stagionale non è sufficiente operare una differenziazione dei dati ma è necessario introdurre dei parametri autoregressivi o media mobile.

	ar1	ar2	ar3
coeff.	0.2164	0.2352	0.2305
s.e.	0.0816	0.0816	0.0833
AIC	-194.33		
AICc	-194.03		
BIC	-182.53		

Tabella 2.1: SARIMA$(3,0,0) \times (0,1,0)_{12}$

Nel secondo modello è stata aggiunta una compenete MA nella parte stagionale, stimando quindi un modello *SARIMA(3,0,0)×(0,1,1)*$_{12}$, i cui grafici relativi ai residui sono rappresentati in figura 2.12. In questo caso, l'autocorrelazione che

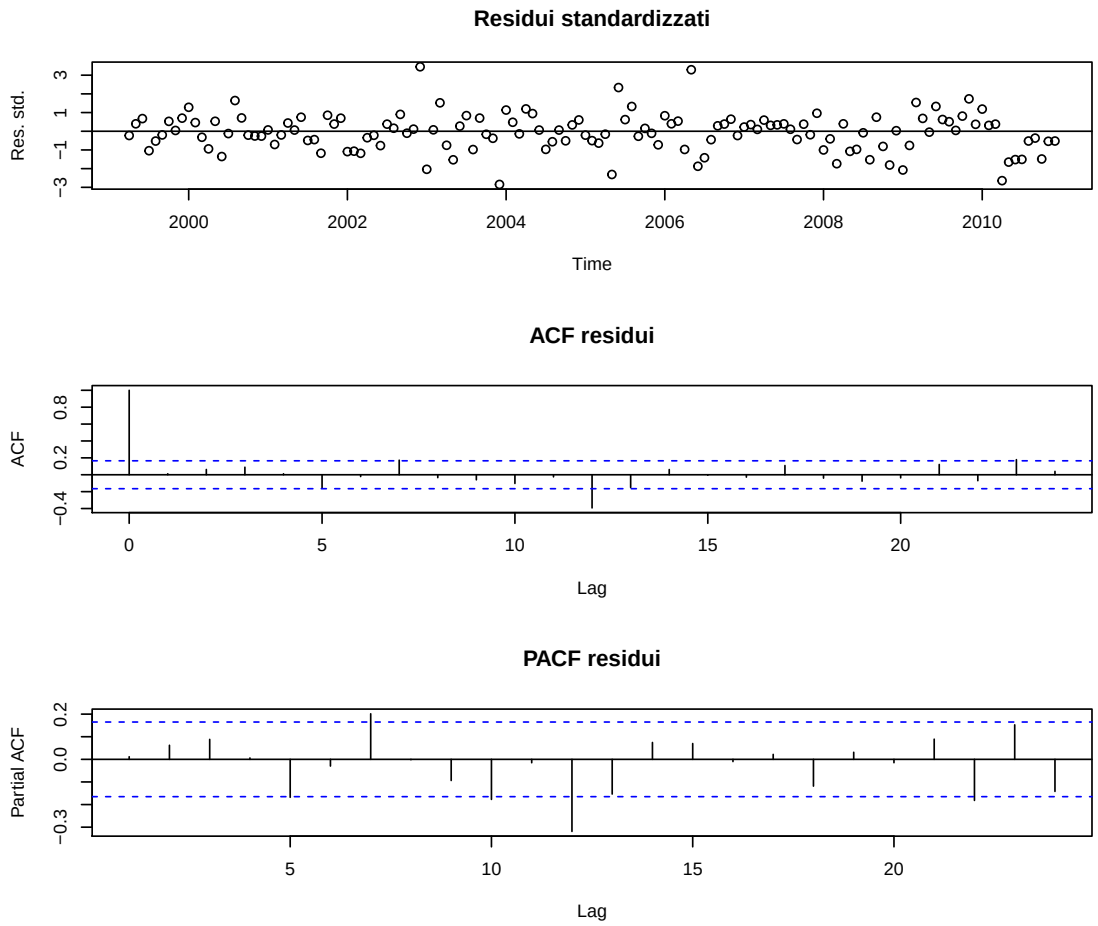


Figura 2.10: Residui del modello SARIMA(3,0,0)x(0,1,0)$_{12}$

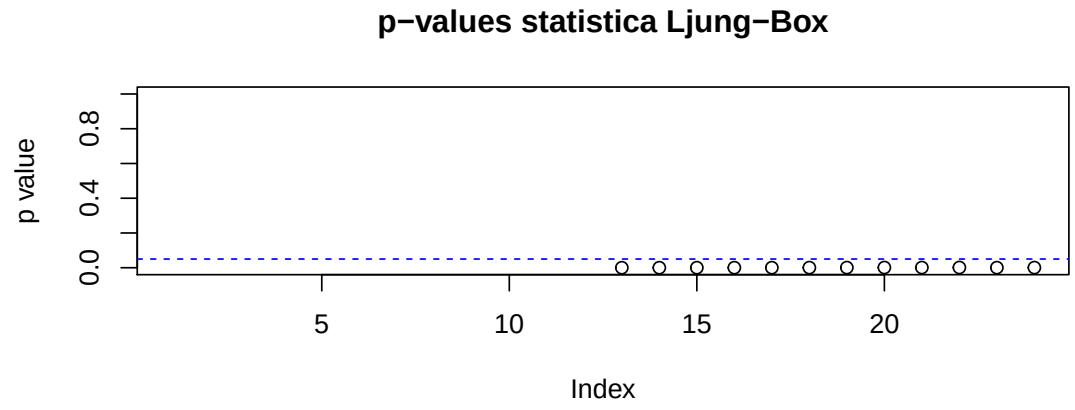


Figura 2.11: p-values della statistica Ljung-Box sui residui del modello SARIMA(3,0,0)x(0,1,0)$_{12}$

compariva al lag 12 non è più significativa e non sembrano esserci ulteriori elementi particolarmente rilevanti da considerare. I *p-values* del test *Ljung-Box* sono tutti superiori ai livelli di significatività del 5% e portano quindi ad escludere la presenza di correlazioni seriali. I coefficienti del modello (tabella 2.2) sono da considerarsi tutti statisticamente significativi.

	ar1	ar2	ar3	sma1
coeff.	0.1614	0.2367	0.3069	-0.6814
s.e.	0.0801	0.0797	0.0827	0.0818
AIC	-232.11			
AICc	-231.66			
BIC	-217.36			

Tabella 2.2: SARIMA$(3,0,0) \times (0,1,1)_{12}$

Non essendo chiaro dalla forma delle funzioni di autocorrelazione se nella parte stagionale sia opportuno considerare la presenza di un parametro MA oppure un AR, si è stimato un ulteriore modello con un parametro AR a livello stagionale in alternativa al parametro MA, ottenendo quindi un SARIMA(3,0,0)$\times$(1,1,0)$_{12}$ (tabella 2.3, figure 2.14 e 2.15); è stato stimato, inoltre, un modello che considera sia un parametro AR che un parametro MA stagionali, del tipo SARIMA(3,0,0)$\times$(1,1,1)$_{12}$, tabella 2.4, figure 2.16 e 2.17.

Nel primo caso il parametro autoregressivo stagionale risulta statisticamente significativo; L'Akaike Information Criterion (sia nella forma classica, AIC, sia in quella corretta per piccoli campioni, AICc[7]) risulta più elevato, mentre in generale sono preferibili modelli con il minor valore di questa misura.

[7]Considerare il valore dell'AICc è preferibile rispetto all'AIC (rispetto al quale vi è l'aggiunta di un termine di correzione della distorsione) fintanto che la numerosità della serie non è sufficientemente maggiore rispetto al numero massimo di parametri considerati, nell'insieme dei modelli alternativi tra cui effettuare la scelta. Una regola, ad esempio, è quella di considerare l'AICc se il rapporto $n/K < 40$, dove n è la numerosità della serie e K è il numero massimo di parametri tra i modelli candidati [Burnham and Anderson, 2002]. Più il rapporto è elevato, più i due criteri portano a valori e decisioni simili. In questa analisi, il rapporto porta a considerare l'AICc.

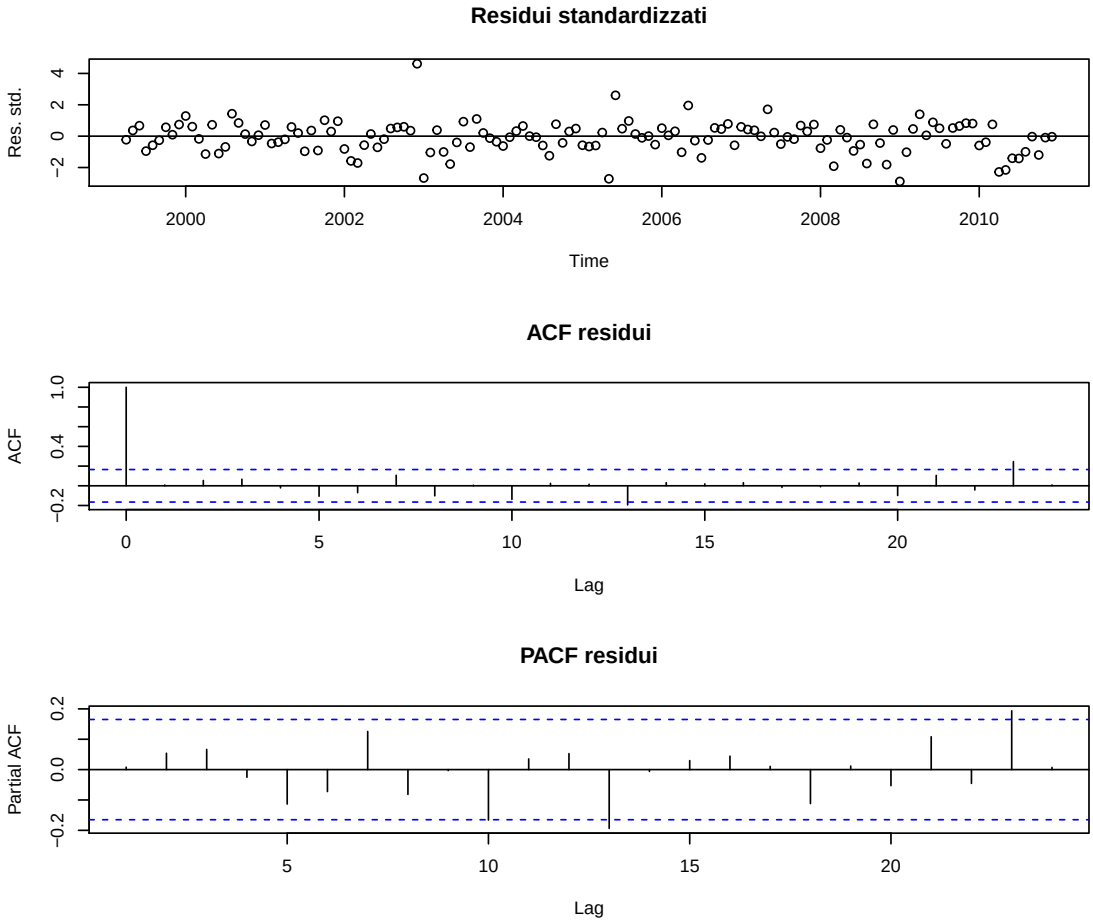


Figura 2.12: Residui del modello SARIMA(3,0,0)x(0,1,1)$_{12}$

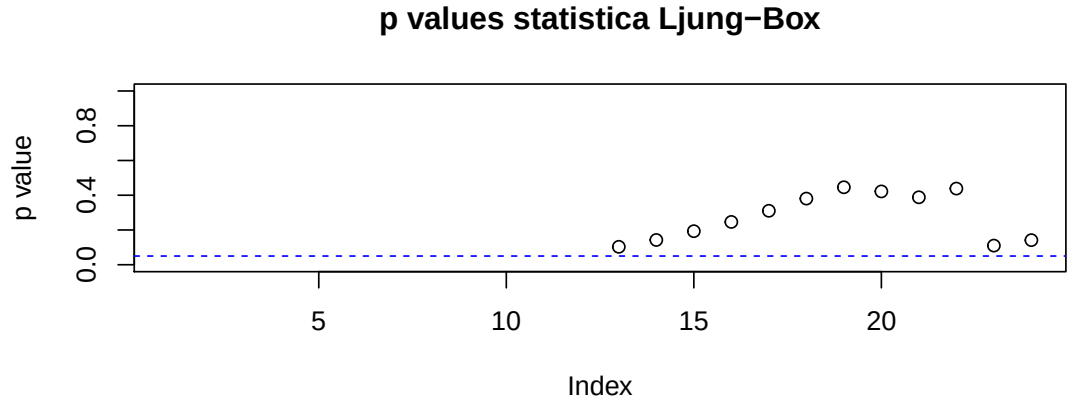


Figura 2.13: p-values della statistica Ljung-Box sui residui del modello SARIMA(3,0,0)x(0,1,1)$_{12}$

Nel secondo caso, il coefficiente autoregressivo stagionale risulta statisticamente non significativo. Si può osservare come i grafici sulle autocorrelazioni dei residui (figura 2.16) risultino molto simili a quelli del modello contentente, nella parte stagionale, un solo parametro MA; ciò rafforza l'ipotesi che il parametro autoregressivo aggiunto sia ininfluente, supportato da un AICc (misura che penalizza le sovraparametrizzazioni) leggermente superiore rispetto al modello SARIMA(3, 0, 0)×(0, 1, 1)$_{12}$.

	ar1	ar2	ar3	sar1
coeff.	0.11	0.2725	0.2901	-0.4589
s.e.	0.0832	0.0781	0.0826	0.0807
AIC	-220.22			
AICc	-219.78			
BIC	-205.48			

Tabella 2.3: SARIMA$(3,0,0) \times (1,1,0)_{12}$

	ar1	ar2	ar3	sar1	sma1
coeff.	0.1742	0.2328	0.2998	0.0886	-0.7334
s.e.	0.0822	0.0803	0.0837	0.1301	0.1049
AIC	-230.57				
AICc	-229.94				
BIC	-212.88				

Tabella 2.4: SARIMA(3, 0, 0)×(1, 1, 1)$_{12}$

Un riepilogo dei criteri di selezione per i modelli analizzati è visibile nella tabella 2.5. Risulta chiaro che, tra questi modelli, la migliore parametrizzazione si ha con un SARIMA(3, 0, 0)×(0, 1, 1)$_{12}$. Questo non significa che sia necessariamente il modello migliore in termini previsionali (si rimanda alla sezione 2.4 per le valutazioni sull'accuratezza).

Se il modello è stato identificato correttamente, i residui devono distribuirsi gaussianamente. In particolare essi devono essere realizzazioni di un processo *white noise*, ovvero con media pari a 0 e varianza costante e devono risultare

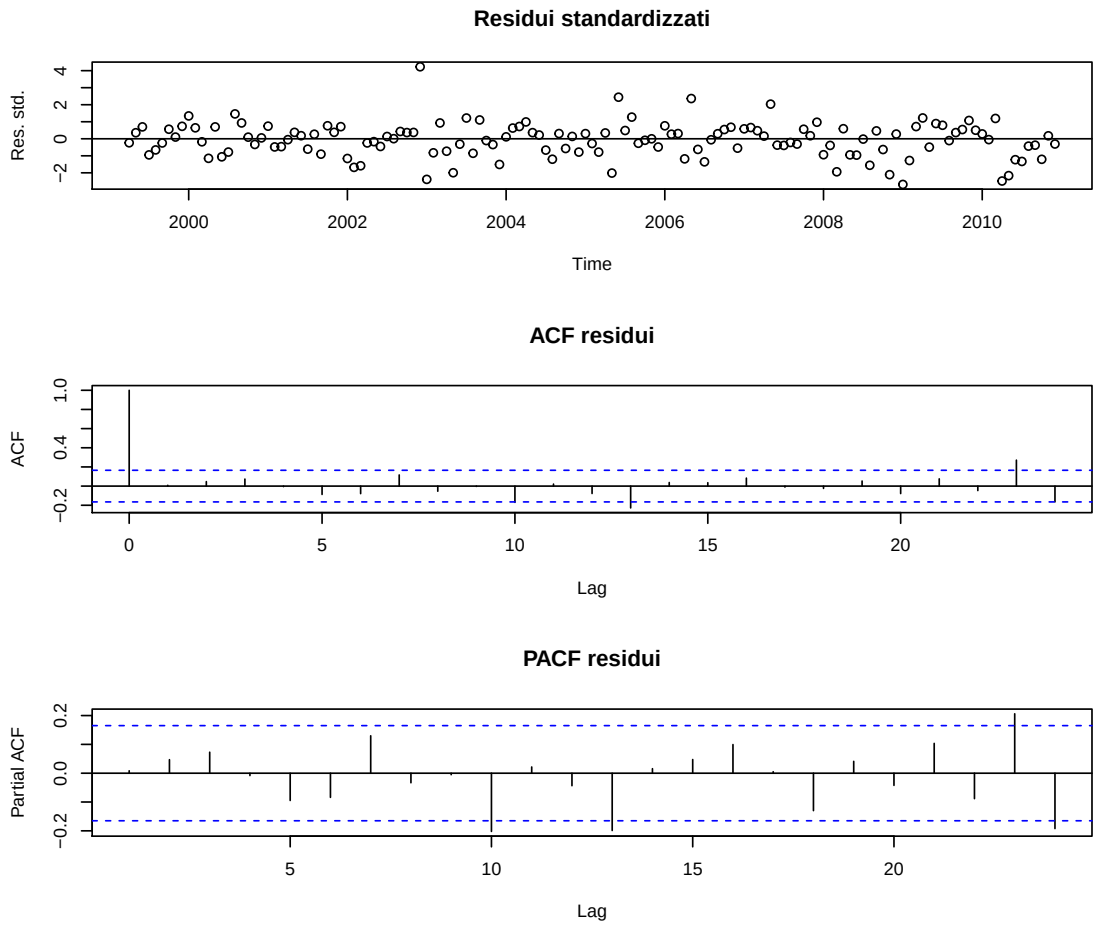


Figura 2.14: Residui del modello SARIMA(3, 0, 0) x $(1, 1, 0)_{12}$

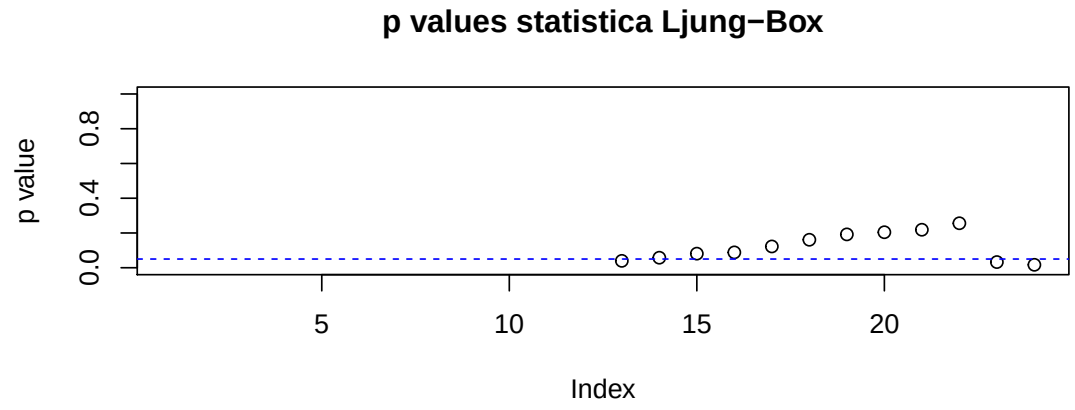


Figura 2.15: p-values della statistica Ljung-Box sui residui del modello SARIMA(3, 0, 0) x $(1, 1, 0)_{12}$

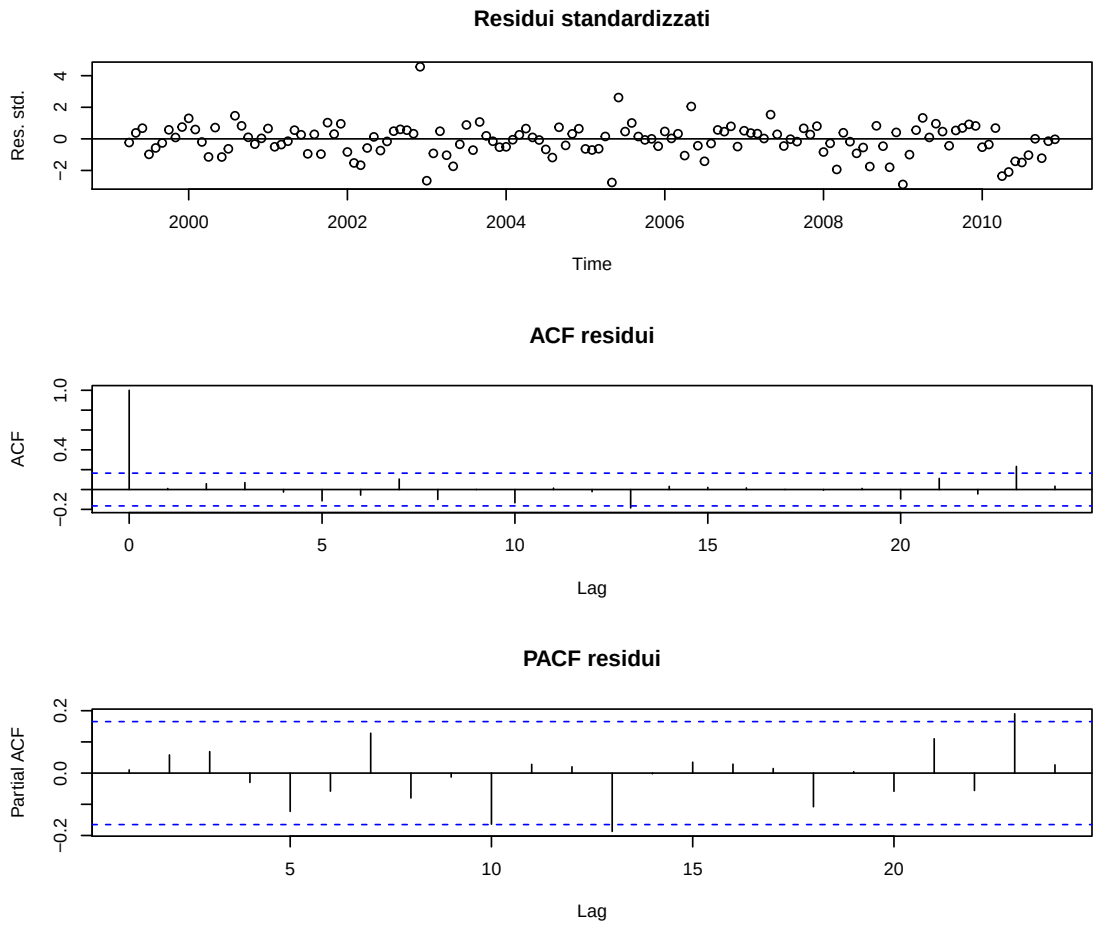


Figura 2.16: Residui del modello SARIMA(3, 0, 0) x $(1, 1, 1)_{12}$

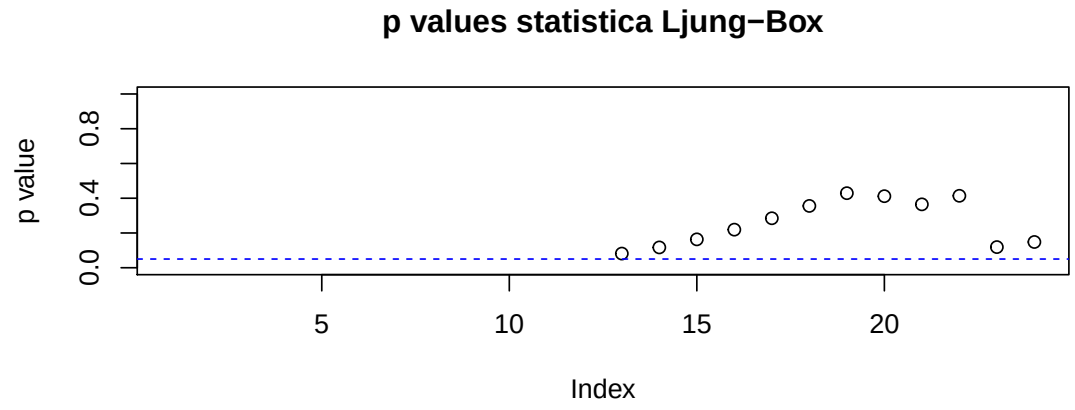


Figura 2.17: p-values della statistica Ljung-Box sui residui del modello SARIMA(3,0,0) x $(1,1,1)_{12}$

SARIMA(p, d, q)×(P, D, Q)$_s$	AICc	AIC	BIC
(3, 0, 0)×(0, 1, 0)$_{12}$	-194.03	-194.33	-182.53
(3, 0, 0)×(0, 1, 1)$_{12}$	-231.66	-232.11	-217.36
(3, 0, 0)×(1, 1, 0)$_{12}$	-219.78	-220.22	-205.48
(3, 0, 0)×(1, 1, 1)$_{12}$	-229.94	-230.57	-212.88

Tabella 2.5: AICc, AIC e BIC

tra loro incorrelati. Si è già visto come i correlogrammi dei residui del modello scelto (figura 2.12) e i *pvalues* del test Ljung-Box (figura 2.13) portano a ritenere che tra di essi vi sia incorrelazione seriale. Per quanto riguarda la normalità, un istogramma e un grafico quantile-quantile rispetto ad una distribuzione Normale (figura 2.18) sono strumenti utili per valutare visivamente la distribuzione dei dati. Si può notare come sia presente una leggera asimmetria, dovuta alla presenza di alcuni valori elevati nella coda di destra. I risultati di alcuni test di normalità sono riportati nella tabella 2.6. Ad un livello di significatività del 5%, i test di *Shapiro* e *Jarque-Bera* portano a rifiutare l'ipotesi che i residui seguano una distribuzione Normale; il test di *Kolmogorov-Smirnov* presenta un *pvalue* circa pari al 5%.

Test	H_0	valore	p-value
Kolmogorov-Smirnov	Normale	0.1077	0.0761
Shapiro	Normale	0.9489	4.666×10^{-5}
Jarque-Bera	Normale	63.6751	1.4877×10^{-14}

Tabella 2.6: Test di normalità sui residui del modello SARIMA(3,0,0)×(0,1,1)$_{12}$

È lecito pensare che la serie contenga alcune "anomalie", le quali rendono difficile una migliore stima attraverso un modello ARIMA stagionale e che pertanto i residui difficilmente possano distribuirsi gaussianamente. È possibile ipotizzare che il processo di fondo risenta dell'effetto dell'erogazione di incentivi statali in diversi periodi e, specie a partire dal 2010, dell'effetto delle crisi economica;

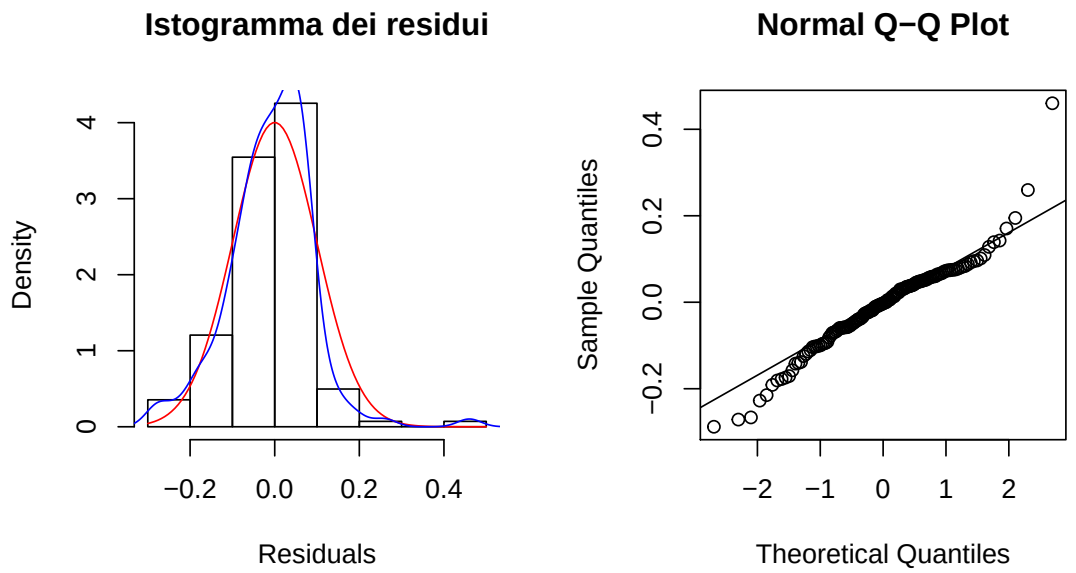


Figura 2.18: Distribuzione dei residui del modello SARIMA(3,0,0)x$(0,1,1)_{12}$

risulta opportuno, pertanto, tener conto della presenza di queste variabili e oltre a questo ricercare potenziali singoli *outliers* di rilievo.

Un aspetto che può influire la bontà della stima della deviazione standard del modello è la presenza di eteroschedasticità nei residui. Il test ARCH-LM in questo caso porta ad accettare l'ipotesi nulla di assenza di eteroschedasticità con un *p-value* pari a 0.55.

Per completezza, nella tabella 2.7 sono riportati i risultati dei test di normalità e di eteroschedasticità calcolati sui residui degli altri modelli stimati.

Nel grafico di figura 2.20 sono raffigurati i valori stimati dai modelli rispetto al training-set: in rosso il SARIMA$(3,0,0)\times(0,1,0)_{12}$, in verde il SARIMA$(3,0,0)\times(0,1,1)_{12}$, in blu il SARIMA$(3,0,0)\times(1,0,0)_{12}$ e in viola il SARIMA$(3,0,0)\times(1,1,1)_{12}$.

2.4 Accuratezza dei modelli

Il modello SARIMA$(3,\ 0,\ 0)\times(0,\ 1,\ 1)_{12}$ si è già visto come essere il migliore tra quelli proposti se si stabilisce l'indice AICc come criterio di scelta tra modelli al-

SARIMA(3,0,0)×(0,1,0)$_{12}$	Test	p-value
Kolmogorov-Smirnov	0.0845	0.2665
Shapiro	0.9758	0.0133
Jarque-Bera	11.3122	0.0035
ARCH-LM	17.354	0.1368
SARIMA(3,0,0)×(1,1,0)$_{12}$	Test	p-value
Kolmogorov-Smirnov	0.0592	0.706
Shapiro	0.9669	0.0017
Jarque-Bera	25.8285	2.4627×10^{-6}
ARCH-LM	10.6629	0.558
SARIMA(3,0,0)×(1,1,1)$_{12}$	Test	p-value
Kolmogorov-Smirnov	0.0665	0.5614
Shapiro	0.9508	6.5322×10^{-5}
Jarque-Bera	58.8064	1.6998×10^{-13}
ARCH-LM	10.5219	0.5703

Tabella 2.7: Test di normalità sui residui dei modelli.

ternativi. Tuttavia si deve considerare come non necessariamente il modello migliore in AICc, così come in generale il modello migliore in accuratezza *in-sample*, sia sinonimo di una migliore accuratezza *out-of-sample*. L'*Akaike Information Criterion*, infatti, permette una stima relativa dell'informazione persa descrivendo i dati reali attraverso un particolare modello, tenendo conto della sua complessità; è quindi una misura strettamente legata alla stima *in-sample*. Ciò che più interessa per un modello che possa essere utilizzato per scopi di previsione è la sua accuratezza rispetto alla stima dei dati futuri. Nella pratica è difficile da verificare a causa dei tempi necessari ad ottenere una serie abbastanza lunga di rilevazioni future. Per questo motivo spesso si usa suddividere la serie disponibile in due campioni, il primo utilizzato per la stima dei parametri (*training-set*) e il secondo per la verifica della bontà delle stime (*test-set*).

Tra gli indici comunemente usati per la misura dell'accuratezza vi sono l'errore assoluto medio (*Mean Absolute Error* o MAE) e la radice dell'errore quadratico

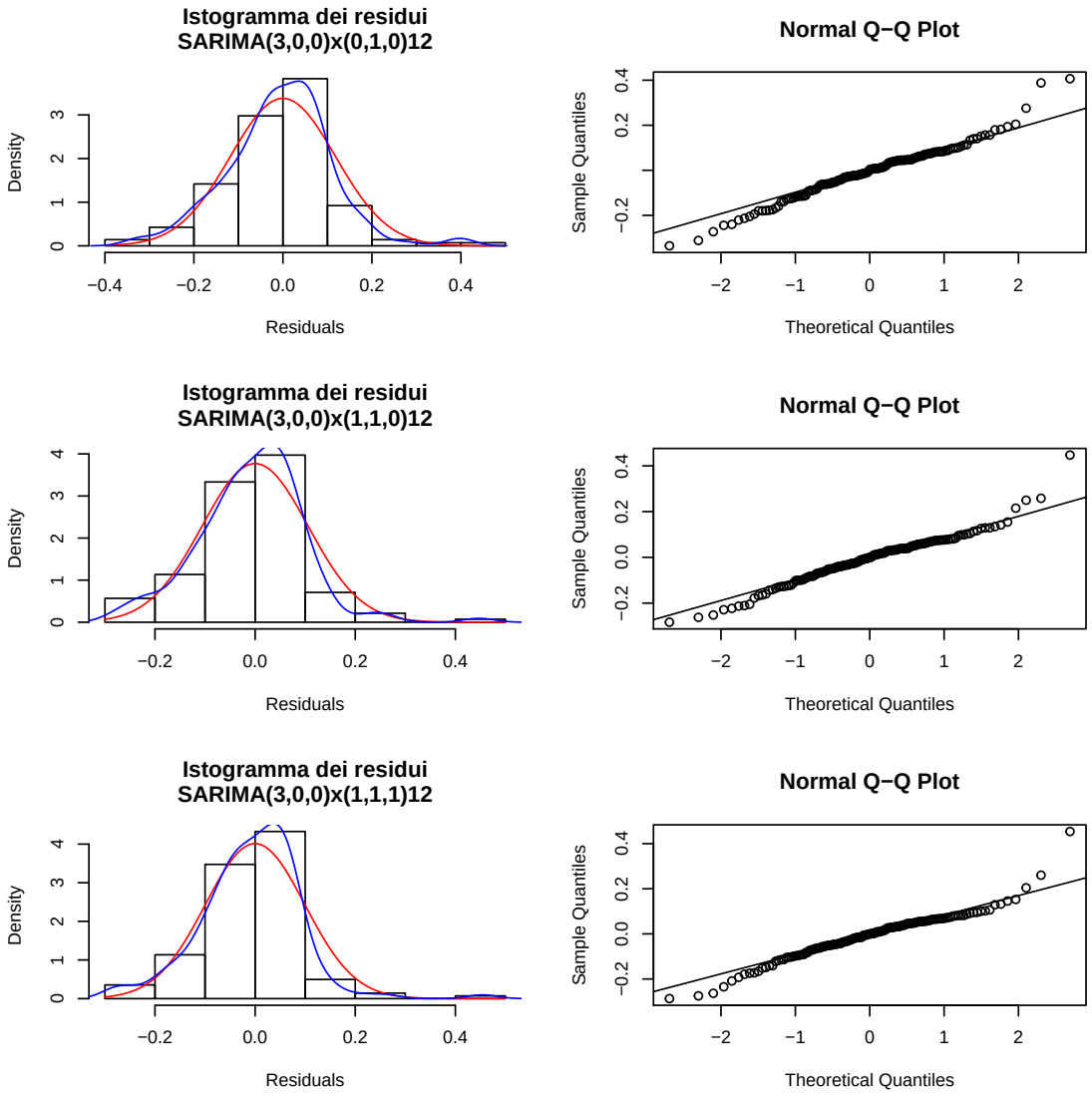


Figura 2.19: Distribuzione dei residui dei modelli.

medio (*Root Mean Squared Error* o RMSE). Il MAPE è l'errore percentuale medio assoluto (*Mean Absolute Percentage Error*), mentre l'errore medio (*Mean Error* o ME) può dare un'indicazione circa la distorsione delle previsioni. Per semplicità di interpretazione il calcolo è stato effettuato in relazione ai volumi di immatricolazioni, pertanto si è reso necessario trasformare esponenzialmente i dati stimati dai modelli.

L'errore *in-sample* si calcola sui residui del modello stimato sul *training-set* ed è un'indicazione di quanto il modello riesca a descrivere i dati storici. Dalla tabel-

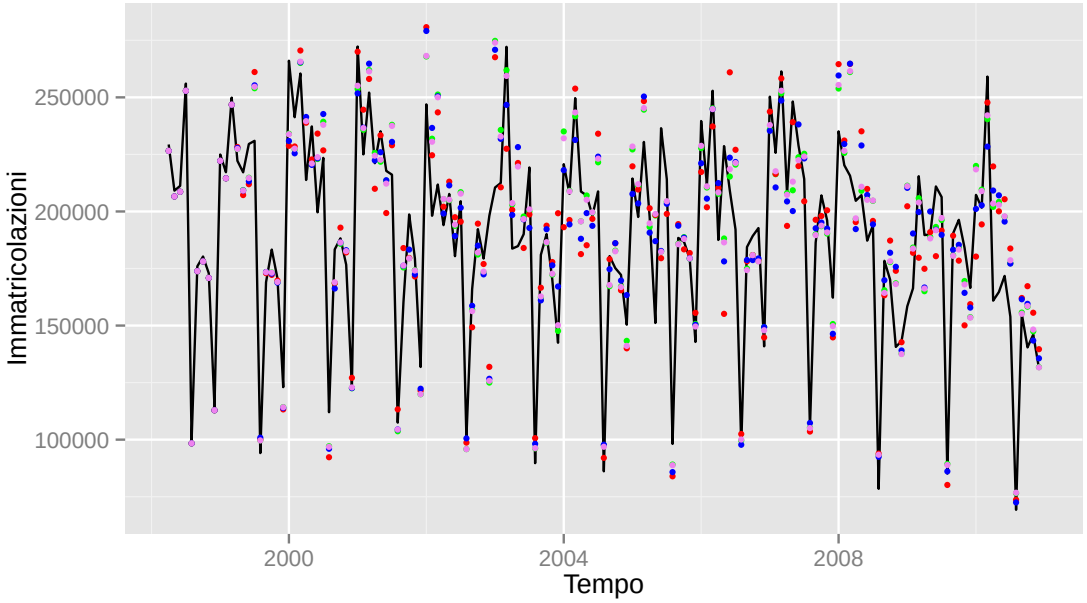


Figura 2.20: Immatricolazioni reali e stimate in-sample.

la 2.4 si osserva come il modello con una migliore performance sia il SARIMA(3, 0, 0)×(1, 1, 0)$_{12}$, che tuttavia non si discosta molto dal modello SARIMA(3, 0, 0)×(0, 1, 1)$_{12}$, migliore in termini AICc.

SARIMA(p, d, q)×(P, D, Q)$_s$	ME	RMSE	MAE	MAPE
(3, 0, 0)×(0, 1, 0)$_{12}$	-1206	2.1489×10^4	1.558×10^4	8.41
(3, 0, 0)×(0, 1, 1)$_{12}$	-1358	1.8165×10^4	1.2958×10^4	6.97
(3, 0, 0)×(1, 1, 0)$_{12}$	-1076	1.953×10^4	1.42×10^4	7.54
(3, 0, 0)×(1, 1, 1)$_{12}$	-1371	1.8084×10^4	1.2933×10^4	6.97

Tabella 2.8: Accuratezza in-sample.

La misura dell'accuratezza delle previsioni *out-of-sample* si effettua rispetto ai dati contenuti nel *test-set*, di cui fanno parte le rilevazioni più recenti. Nella tabella 2.4 sono indicati ME, RMSE, MAE e MAPE delle previsoni ad un passo, aggiornando di volta in volta il campione con un nuovo dato reale e mantenendo fissi i coefficienti del modello.

Dai risultati è evidente come il modello SARIMA(3, 0, 0)×(0, 1, 1)$_{12}$ non sia il migliore in termini di accuratezza *out-of-sample*, nonostante sia il migliore in

SARIMA(p, d, q)×(P, D, Q)$_s$	ME	RMSE	MAE	MAPE
(3, 0, 0)×(0, 1, 0)$_{12}$	-6707.4167	1.4815×10^4	1.2305×10^4	9.18
(3, 0, 0)×(0, 1, 1)$_{12}$	-1.299×10^4	1.7285×10^4	1.4062×10^4	11.2
(3, 0, 0)×(1, 1, 0)$_{12}$	-1.0321×10^4	1.6788×10^4	1.3343×10^4	10.35
(3, 0, 0)×(1, 1, 1)$_{12}$	-1.2771×10^4	1.7208×10^4	1.4064×10^4	11.15

Tabella 2.9: Accuratezza out-of-sample ad un passo.

termini di AICc, così come il SARIMA(3, 0, 0)×(1, 1, 1)$_{12}$ che è risultato il più accurato *in-sample*. L'errore minore si ottengono con il modello SARIMA(3, 0, 0)×(0, 1, 0)$_{12}$.

2.5 Conclusioni

In questo capitolo, dopo aver analizzato la serie in termini descrittivi evidenziando in particolar modo le componenti di *trend* e stagionalità, sono stati stimati alcuni modelli di tipo ARIMA stagionale (SARIMA) sui logaritmi dei dati mensili sulle immatricolazioni di auto. È stata utilizzata la serie che va dall'aprile 2008 a dicembre 2010 come *training-set*, mentre i dati degli anni 2011 e 2012 sono stati impiegati come *test-set* per valutare l'accuratezza delle previsioni *out-of-sample* in termini di MAE, RMSE e MAPE.

La serie presenta un andamento con diverse irregolarità dovute sia alla presenza di periodi di crisi del mercato, sia ad effetti dovuti all'introduzine di contributi statali per l'acquisto di nuove autovetture. Ciò si evidenzia nella difficoltà di ottenere modelli che rispettino le condizioni di ottimalità sui residui per modelli di tipo ARIMA. Il modello che risulta essere migliore in termini di AICc è un SARIMA(3,0,0)×(0,1,1)$_{12}$ mentre il modello che registra un errore minore nelle previsioni ad un passo *out-of-sample* è un SARIMA(3,0,0)×(0,1,0)$_{12}$ (i cui residui sembrano però presentare un certo grado di dipendenza seriale). Tutti i modelli sono stati calcolati sui dati logaritmici mentre gli errori sui volumi di immatricolazioni.

Capitolo 3

Intervention analysis

3.1 Introduzione

Nel precedente capitolo si è giunti alla stima di un modello stocastico ARIMA stagionale, o SARIMA, considerando i soli dati sulle immatricolazioni e operando su di essi una trasformazione logaritmica. I residui del modello, pur presentando una forma campanulare simile a quella di una distribuzione gaussiana, non hanno superato i test di normalità e mostrano una leggera asimmetria, dovuta alla presenza di alcuni valori che "allungano" la coda di destra. È necessario quindi analizzare l'effetto della presenza di eventi esterni, quali sono gli incentivi alla rottamazione di vecchi veicoli e all'acquisto di nuovi mezzi che rispettino normative antinquinamento vigenti (chiamati anche ecoincentivi), volti sopratutto a sostenere il mercato; è opportuno inoltre verificare la presenza di eventuali outlier di tipo additivo e innovativo che possono avere effetti di diverso tipo.

3.2 Incentivi statali per la rottamazione e l'acquisto di autovetture

Nel periodo considerato in questa analisi si possono considerare come eventi esterni 4 periodi di contributi alla rottamazione di autovetture (con vita superiore ad un determinato numero di anni) e all'acquisto di nuovi veicoli. Essi si differenziano per regole di accesso agli incentivi, importo sostenuto dallo Stato e

durata del periodo in cui sono stati vigenti. Il riassunto delle caratteristiche dei diversi incentivi[1] si trova nella tabella 3.2.

Periodo	Condizioni	Contributo
07/2002 - 12/2002	DL 08 luglio 2002 n. 138 - Acquisto autoveicolo entro il 31/12/2002 e consegna di un veicolo non conforme alla direttiva CE 91/441 e successive; potenza non superiore a 85 Kw e conforme alle direttive CE sull'inquinamento.	Esenzione dall'imposta provinciale di trascrizione, dalla tassa automobilistica per il primo periodo fisso e per le due annualità successive, dall'imposta di bollo e dagli emolumenti dovuti agli uffici del Pubblico Registro Automobilistico (P.R.A.).
01/2003 - 03/2003	DL 13 gennaio 203 n.2 - Valgono le condizioni del precedente incentivo.	Stessi benefici del precedente incentivo.
10/2006 - 12/2007 (immatr. entro 03/2008)	DL 262/2006 e legge finanziaria 2007 - Acquisto autoveicolo *euro 4* o *euro 5* (con emissioni di CO_2 <140 g/km) con contratto tra il 03/10/2006 e il 31/12/2007 e immatricolato entro il 31/03/2008, con rottamazione vettura *euro 0* o *euro 1*. Acquisto di autovetture con alimentazione, esclusiva o doppia, a GPL, metano, elettrica o ad idrogeno.	Esenzione dal pagamento della tassa automobilistica e contributo di 800 Euro; esenzione per 2 anni dal "bollo auto", che aumenta a 3 per veicoli di cilindrata inferiore a 1300 cc o superiore o uguale a 1300 cc per nuclei familiari di almeno 6 persone. Contributo (cumulabile con il precedente) di 1500 Euro per autovetture con alimentazione, esclusiva o doppia, a GPL, metano, elettrica o ad idrogeno, più ulteriori 500 Euro per emissioni di C02 <120 g/km; il contributo è pari a 2000 Euro per vetture con alimentazione esclusivamente elettrica o a idrogeno.
01/2008 - 12/2008 (immatr. entro 03/2009)	Acquisto veicolo *euro 4* o *euro 5* (con emissioni di CO_2 <140 g/km) con contratto entro il 31/12/2008 e immatricolato entro il 31/03/2009, con rottamazione vettura *euro 0* o *euro 1* (immatricolata entro il 31/12/1998. Acquisto di autovetture con alimentazione, esclusiva o doppia, a GPL, metano, elettrica o ad idrogeno.	Contributo 1500 Euro per veicoli con massa complessiva <3 t e 2500 tra 3 e 3,5 t (veicolo rottamato e veicolo nuovo devono essere dello stesso tipo). Per l'acquisto di autovetture con alimentazione, esclusiva o doppia, a GPL, metano, elettrica o ad idrogeno, l'importo del contributo è identico al precedente incentivo e viene esteso ai veicoli immatricolati entro il 03/2010 (acquistati entro 12/2009).
02/2009 - 12/2009	DL 10 febbraio 2009 n. 5 - Condizioni analoghe ai precedenti incentivi. I veicoli da rottamare devono essere stati immatricolati entro il 31/12/1999.	Benefici analoghi ai precedenti.

Tabella 3.1: Contributi statali per la rottamazione e l'acquisto di autovetture nuove

3.3 L'analisi degli interventi e la ricerca di outlier

La presenza di fenomeni esterni può influenzare l'andamento di una serie storica. Tra di essi si possono includere vacanze, scioperi, eventi naturali, periodi promozionali, novità legislative e interventi di varia natura sul mercato. Gli ecoincentivi statali rientrano tra la tipologia di eventi volti a modificare l'andamento delle vendite, cercando di far fronte a segnali di arretramento del mercato.

[1]Si riporta solo il caso di acquisto di autoveicoli di prima immatricolazione.

Per valutare l'effetto di un intervento esterno dopo la sua manifestazione, un *test* t per due campioni non è uno strumento idoneo perché particolarmente sensibile alla violazione del requisito di indipendenza [Box and Tiao, 1965].

I modelli d'intervento [Box and Tiao, 1975] permettono di stimare contemporaneamente gli effetti di eventi esterni, codificati attraverso delle variabili di tipo *dummy*, e i parametri di un modello stocastico.

In generale vengono distinti due tipi di interventi: quelli rappresentabili da una variabile indicatrice "scalino" (*step*) e quelli identificabili con una variabile indicatrice "impulso" (*pulse*).

Una variabile *step* (3.1) è rappresentata da una sequenza di 0 fino al tempo $t-1$ e assume valore 1 a partire dal tempo t.

$$S_t^{(T)} = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases} \tag{3.1}$$

Una variabile *pulse* (3.2), invece, assume valore 1 solamente al tempo t.

$$P_t^{(T)} = \begin{cases} 0, & t \neq T \\ 1, & t = T \end{cases} \tag{3.2}$$

Si tenga presente che è possibile rappresentare un input $S^{(T)}$ con uno di tipo $P^{(T)}$ (e viceversa) nel seguente modo:

$$(1-B)S_t^{(T)} = P_t^{(T)} \tag{3.3}$$

L'effetto provocato da variabili *step* e *pulse* può essere di vario tipo. Esse possono provocare una risposta immediatamente al tempo t, oppure ad un tempo successivo ($t+1, t+2, \ldots$). In generale possiamo scrivere

$$\omega B^b S_t^{(T)} \tag{3.4}$$

oppure

$$\omega B^b P_t^{(T)}, \tag{3.5}$$

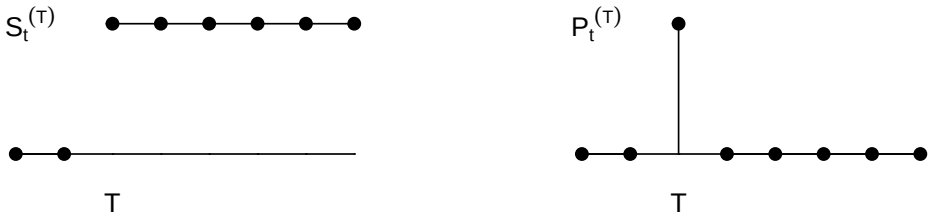


Figura 3.1: Step e pulse input

dove ω rappresenta l'entità del salto di livello nella serie e b il numero di periodi di ritardo dopo i quali l'intervento ha impatto sul processo. Nel caso in cui $b = 0$, l'effetto avviene al tempo t.

In altri casi l'intervento può avere effetto sul fenomeno dopo b periodi ma con una risposta graduale, avendo così nei due casi

$$\frac{\omega B^b}{1 - \delta B} S_t^{(T)} \tag{3.6}$$

e

$$\frac{\omega B^b}{1 - \delta B} P_t^{(T)}, \tag{3.7}$$

con $0 < \delta < 1$ a determinarne la gradualità. Nel caso in cui $\delta = 0$ si ottengono la 3.4 e la 3.5. Qualora δ fosse uguale a 1, con un input di tipo *step* si produrrebbe un effetto "rampa" [Box and Tiao, 1975]. È possibile inoltre creare varie combinazioni delle funzioni per ottenere forme diverse nella risposta.

In figura 3.1 sono raffigurati i due diversi tipi di input mentre in figura 3.2 sono riportati alcuni esempi di risposta alle funzioni di trasferimento $[\omega(B)/\delta(B)]S_t^{(T)}$ e $[\omega(B)/\delta(B)]P_t^{(T)}$, come illustrati da Box e Tiao, con $b = 1$ [Box and Tiao, 1975].

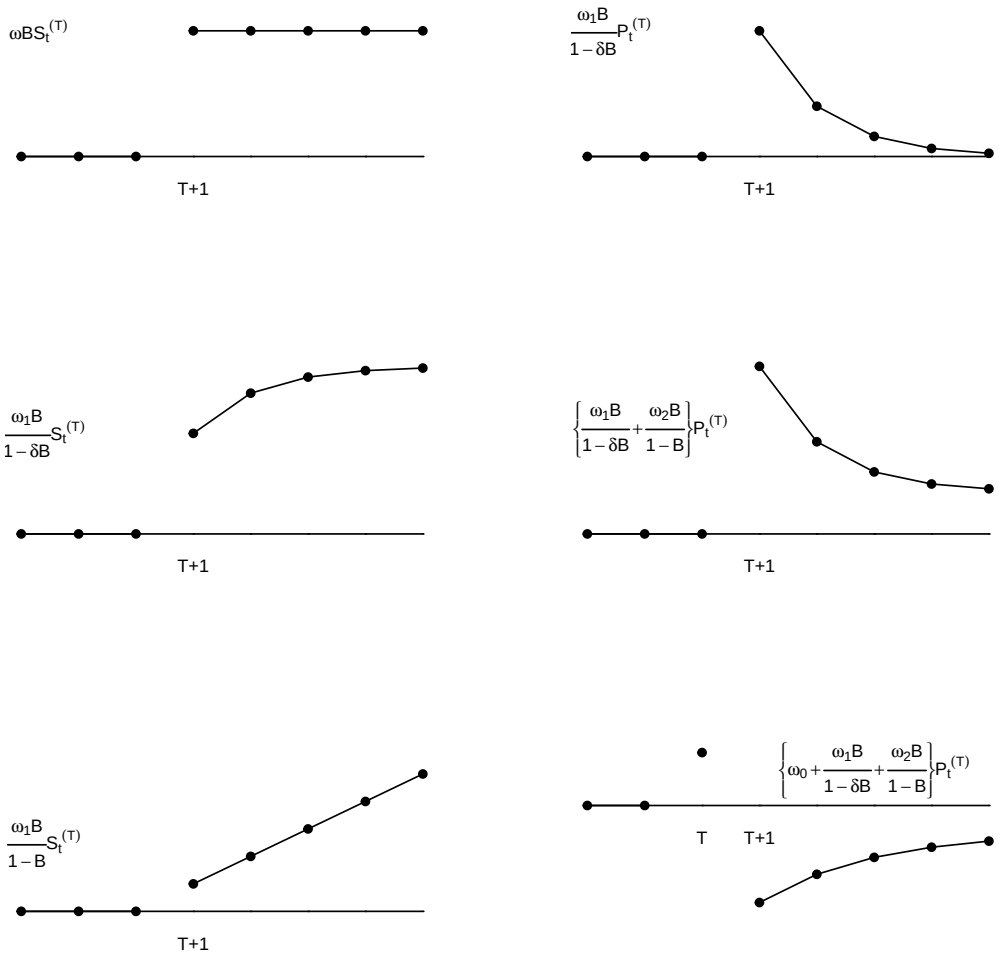


Figura 3.2: Risposte a diversi tipi di funzioni di trasferimento

In generale, sia X_t un processo *ARIMA(p,d,q)* (estendibile eventualmente al caso stagionale), privo dell'effetto di *outliers* del tipo

$$\phi(B)(1-B)^d x_t = \theta(B) a_t \ , \tag{3.8}$$

dove B è l'operatore ritardo, $\phi(B)$ e $\theta(B)$ sono rispettivamente la componente autoregressiva e la componente media mobile, con $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ e $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ con radici al di fuori del cerchio unitario,

il modello d'intervento è

$$Z_t = \frac{\omega(B)}{\delta(B)} I_t^{(T)} + X_t \tag{3.9}$$

nel caso di un solo intervento e

$$Z_t = \sum_{j=1}^{k} \frac{\omega_j(B)}{\delta_j(B)} I_t^{(T_j)} + X_t \tag{3.10}$$

in presenza di interventi multipli.

La procedura di stima di un modello prevede come prima fase l'identificazione del modello sulla *noise series*, ovvero sui dati precedenti al verificarsi di un intervento. Il modello $[\theta(B)/\psi(B)]a_t$ viene definito *noise model*. Successivamente, se il modello risulta adeguato, si procede all'inclusione delle variabili che rappresentano gli interventi e alla stima del modello d'intervento.

3.3.1 Outliers additivi e innovativi

L'introduzione degli interventi nell'analisi di una serie storica si può ricondurre al caso più generale dell'inclusione degli effetti di *outliers* in un modello stocastico. La differenza di fondo riguarda la conoscenza del manifestarsi del fenomeno causa dell'anomalia (ad esempio l'emanazione di una legge) e il tempo t in cui esso si verifica, informazioni solitamente note quando si affronta un'analisi degli interventi. La presenza di un valore *outlier*, invece, non sempre è facilmente individuabile e collocabile temporalmente. Nei casi in cui esso è dovuto ad un'errata registrazione del dato, può risultare evidente se l'errore è particolarmente rilevante. Tuttavia un *outlier* può essere determinato da un evento esterno non noto e databile a priori; non necessariamente è da considerarsi un evento ignoto o imprevisto in assoluto, ma lo si considera tale in relazione alle informazioni a disposizione dell'analista; pertanto è possibile che uno o più valori risultati anomali ad una prima analisi, possano rappresentare, dopo l'individuazione delle cause,

l'effetto di interventi (ad esempio l'emanazione di una norma di cui il ricercatore non era a conoscenza).

È possibile distinguere principalmente due tipi di valori anomali: *outliers* additivi (AO) e *outliers* innovativi (IO) [Chang et al., 1988]. Sia X_t una serie stazionaria e priva dell'influenza di *outliers* e che segua un modello *ARMA(p,q)*

$$\phi(B)x_t = \theta(B)a_t \ , \tag{3.11}$$

con a_t sequenza di white noise, ovvero di valori indipendenti e identicamente distribuiti come $N(0,\ \sigma_a^2)$; sia inoltre $I_t^{(T)}$ una variabile indicatrice del tipo

$$I_t^{(T)} = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases} .$$

Il modello per un AO è del tipo

$$z_t = \begin{cases} x_t, & t \neq T \\ x_t + \omega, & t = T \end{cases} \tag{3.12}$$

$$z_t = x_t + \omega I_t^{(T)} \tag{3.13}$$

$$= \frac{\theta(B)}{\phi(B)} a_t + \omega I_t^{(T)} \ . \tag{3.14}$$

L'effetto di un AO è limitato al solo istante in cui esso si verifica. Esso può manifestarsi in seguito ad un evento che una perturbazione che cessa nell'istante temporale immediatamente successivo. In molti casi questo tipo di *outlier* può essere dovuto ad errori di registrazione di un dato. Si è invece in presenza di un IO quando le conseguenze sulla serie si propagano nei periodi successivi al tempo T in cui esso si verifica. Il modello quindi risulta essere

$$z_t = x_t + \frac{\theta(B)}{\phi(B)} \omega I_t^{(T)} \tag{3.15}$$

$$= \frac{\theta(B)}{\phi(B)} (a_t + \omega I_t^{(T)}) \ . \tag{3.16}$$

Oltre generalmente a non conoscere l'esatta collocazione temporale, non è sempre immediato a priori identificare la tipologia di *outlier* che un dato rappresenta. È necessario quindi effettuare dei test di verifica d'ipotesi successivamente alla stima ai minimi quadrati del parametro ω, in modo da verificare che z_t sia un *outlier* ed eventualmente se di tipo additivo o innovativo. Si ipotizzi quindi il caso in cui i parametri del modello *ARMA(p,q)* della (3.11) siano noti. Sia

$$\pi(B) = \frac{\phi(B)}{\theta(B)} = (1 - \pi_1 B - \pi_2 B^2 - \dots) \tag{3.17}$$

e inoltre

$$e_t = \pi(B) z_t \tag{3.18}$$

con $t = 1, \dots, n$; si ottiene quindi

$$IO: \; e_t = \omega I_t^{(T)} + a_t \tag{3.19}$$

$$AO: \; e_t = \omega \pi(B) I_t^{(T)} + a_t \; . \tag{3.20}$$

Tenendo conto che

$$\pi(B) I_t^{(T)} = \begin{cases} 0 & \text{, per } t < T \\ 1 & \text{, per } t = T \\ -\pi_{t-T} & \text{, per } t > T \end{cases} ,$$

gli stimatori ai minimi quadrati di ω risultano

$$IO: \; \tilde{\omega}_I = e_T \tag{3.21}$$

$$\begin{aligned} AO: \; \tilde{\omega}_A &= \rho^2 \pi(F) e_T \; , \\ &= \rho^2 (1 - \pi_1 F - \pi_2 F^2 - \dots - \pi_{n-T} F^{n-T}) e_t \; , \end{aligned} \tag{3.22}$$

con $\rho^2 = (1 + \pi_1^2 + \pi_2^2 + \cdots + \pi_{n-T}^2)^{-1}$ e F operatore di anticipo, tale per cui $Fe_t = e_{t+1}$. Le varianze degli stimatori così ricavati sono:

$$IO: \ Var(\tilde{\omega}_I) = \sigma_a^2 \tag{3.23}$$

$$AO: \ Var(\tilde{\omega}_A) = \rho^2 \sigma_a^2 \ . \tag{3.24}$$

Si ponga come ipotesi nulla H_0 che ω sia pari a 0 e vi siano due ipotesi alternative H_1 e H_2, rispettivamente $\omega \neq 0$ in un modello per IO (3.15) e $\omega \neq 0$ in un modello per AO (3.12). Attraverso il criterio del rapporto di verosimiglianza si possono ottenere tre statistiche test (3.25), a seconda del sistema d'ipotesi considerato.

$$\begin{aligned} &H_0 \ vs. \ H_1: \ \lambda_{1,T} = \tilde{\omega}_I/\sigma_a \ , \\ &H_0 \ vs. \ H_2: \ \lambda_{2,T} = \rho\tilde{\omega}_A/\sigma_a \ , \\ &H_1 \ vs. \ H_2: \ \lambda_{3,T} = [\rho^{-2}\tilde{\omega}_A^2 - \tilde{\omega}_I^2]/[2\sigma_a^2(1-\rho^2)^{1/2}] \ . \end{aligned} \tag{3.25}$$

Le statistiche $\lambda_{1,T}$ e $\lambda_{2,T}$ hanno distribuzione $N(0,1)$ sotto H_0. La particolare distribuzione di $\lambda_{3,T}$ è tabulata, tuttavia per distinguere la natura del possibile *outlier* si può seguire la regola per cui se $|\lambda_{1,T}| > |\lambda_{2,T}|$, allora esso è di tipo IO, altrimenti è un AO.

Nel caso in cui i parametri del modello (3.11) sono ignoti, è necessario stimare $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$ e σ_a attraverso i loro stimatori di massima verosimiglianza, ottenendo quindi $\hat{\phi}_1, \ldots, \hat{\phi}_p$, $\hat{\theta}_1, \ldots, \hat{\theta}_q$ e $\hat{\sigma}_a$; le stime vanno effettuate sulla serie come se non fossero presenti *outliers*. Le statistiche λ di conseguenza saranno:

$$\begin{aligned} &\hat{\lambda}_{1,T} = \hat{\omega}_I/\hat{\sigma}_a \ , \\ &\hat{\lambda}_{2,T} = \hat{\rho}\hat{\omega}_A/\hat{\sigma}_a \ , \\ &\hat{\lambda}_{3,T} = [\hat{\rho}^{-2}\hat{\omega}_A^2 - \hat{\omega}_I^2]/[2\hat{\sigma}_a^2(1-\hat{\rho}^2)^{1/2}] \ , \end{aligned} \tag{3.26}$$

con $\hat{\omega}_I = \hat{e}_T$, $\hat{\omega}_A = \hat{\rho}^2\hat{\pi}(F)\hat{e}_T$ e $\hat{\rho}^2 = (1 + \hat{\pi}_1^2 + \hat{\pi}_2^2 + \cdots + \hat{\pi}_{n-T}^2)^{-1}$. Le statistiche $\hat{\lambda}_{1,T}$, $\hat{\lambda}_{2,T}$ e $\hat{\lambda}_{3,T}$ sono asintoticamente equivalenti a $\lambda_{1,T}$, $\lambda_{2,T}$ e $\lambda_{3,T}$.

Un metodo per l'individuazione di *outliers* (con T ignoto) e la stima del loro effetto è quello proposto da Chang e Tiao, attraverso una procedura iterativa in 3 passi [Chang et al., 1988]:

1. stima del modello iniziale sulla serie Z_t ignorando l'eventuale presenza di outlier;

2. individuazione degli *outliers*;

3. stima simultanea di tutti parametri includendo gli *outliers* identificati per la serie Z_t, secondo un intervention model.

Riguardo al secondo punto, per ogni istante t si calcolano $\hat{\lambda}_{1,T}$ e $\hat{\lambda}_{2,T}$, scegliendo tra le due statistiche quella con valore massimo. Il dato al tempo T è un possibile IO se risulta essere massima $\hat{\lambda}_{1,T}$ e se essa è in valore assoluto maggiore di una costante positiva C; viceversa, se al tempo T risulta massima $\hat{\lambda}_{2,T}$ ed essa è in valore assoluto maggiore di C, il dato è un possibile AO. Individuato il possibile *outlier*, si calcola una nuova stima $\check{\sigma}_A^2$ dopo aver modificato i residui al tempo T con $\check{e}_T = \hat{e}_T - \hat{\omega}_I = 0$, nel caso di un IO, oppure agli istanti $t \geq T$ con $\check{e}_t = \hat{e}_t - \hat{\omega}_A\hat{\pi}(B)I_t^{(T)}$. Successivamente si procede nuovamente alla stima di $\hat{\lambda}_{1,T}$ e $\hat{\lambda}_{2,T}$ a partire dallo stesso modello ma con i residui $\check{e}_t$ e la varianza $\check{\sigma}_a^2$, ripetendo la procedura fino a che non vengono trovati ulteriori possibili *outliers*.

La costante positiva C, che permette di discriminare le stime $\hat{\lambda}_{1,T}$ e $\hat{\lambda}_{2,T}$ scegliendo quelle che possono indicare un possibile IO o AO, solitamente è compresa tra 3 e 4. È possibile eventualmente ricorrere alla regola di Bonferroni ponendo $C = \mathrm{z}_{1-\frac{\alpha}{2}n^{-1}}$, ovvero il quantile di una distribuzione Normale standard di ordine $1 - \frac{\alpha}{2}n^{-1}$, dove α è il livello di significatività e con n la numerosità della serie [Chang et al., 1988].

3.4 Il noise model

Come è stato anticipato nella sezione 3.3, la stima di un modello d'intervento ha inizio con l'individuazione di un modello stocastico che rappresenti la *noise series*. In via teorica, la *noise series* N_t è la serie di dati precedente all'introduzione di un intervento esterno. Quindi, nel caso in cui sia presente un solo intervento al tempo T, essa è difinita come $N_t = \{Z_t : t < T\}$. Nei dati in esame sono presenti più interventi in diversi periodi e non vi è una finestra temporale ampia su cui effettuare l'identificazione del modello. Il primo incentivo è stato introdotto a a partire da luglio del 2002, pertanto

$$N_t = \{Z_t : t < \text{Luglio 2002}\} \, . \tag{3.27}$$

La serie $\{N_t\}$ ha una numerosità n_{N_t} pari a 51; considerando che è necessario operare, come già visto, la differenza stagionale con il dodicesimo ritardo $\nabla^{12} N_t$, la numerosità dei dati con cui si vanno a calcolare le funzioni ACF e PACF si riduce a 39, il che rende debole la loro interpretazione, con bande di confidenza per la non significatività molto ampie. Alcuni tentativi empirici hanno portato a definire un modello del tipo SARIMA(1, 0, 0)$\times$(1, 1, 0)$_{12}$ risultando migliore rispetto ad altri in termini di AICc. I grafici nelle figure 3.3 e 3.4 mostrano i residui, le loro autocorrelazioni e i *p values* delle statistiche *Ljung-Box*. Il test Jarque-Bera sui residui porta a non rigettare l'ipotesi nulla di normalità, con un *p value* pari a 0.7529. Il modello stimato quindi è:

$$(1 + 0.3B^{12})(1 - 0.4275B)(1 - B^{12})N_t = a_t \tag{3.28}$$

Si può notare come il *noise model* nella (3.28) risulti ben diverso dai modelli stimati nel precedente capitolo sulla serie completa, dove sono stati individuati tre parametri AR non stagionali. Questa differenza può essere imputata o alla

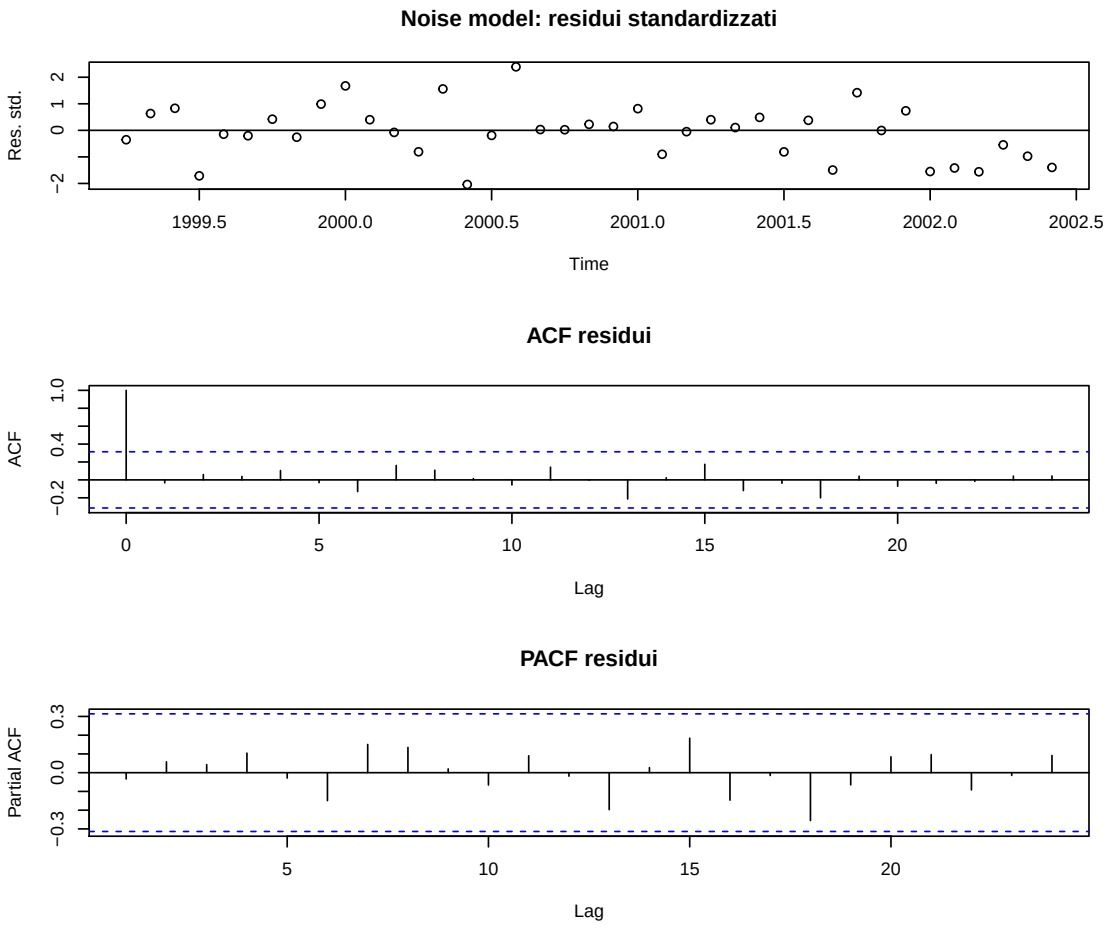


Figura 3.3: Residui, noise model SARIMA(1,0,0)x(1,1,0)$_{12}$

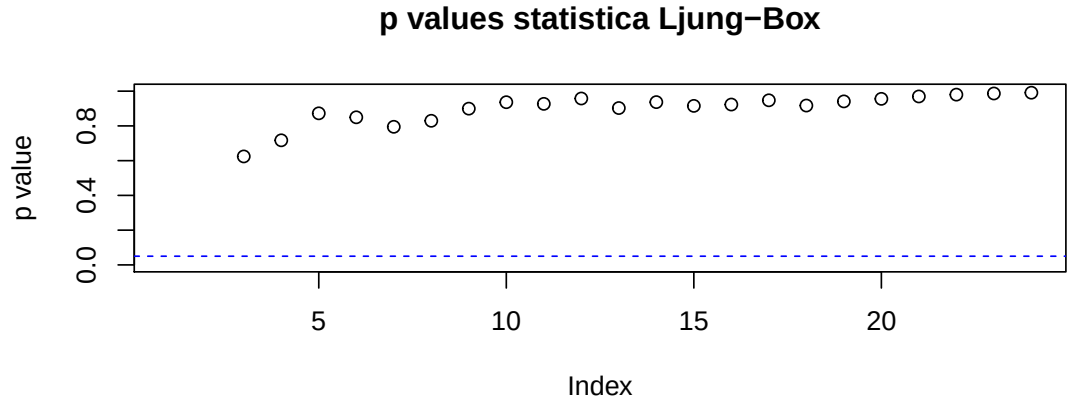


Figura 3.4: p-values della statistica Ljung-Box sui residui, noise model

	ar1	sar1
coeff.	0.4275	-0.3
s.e.	0.1505	0.1783
AIC	-81.75	
AICc	-81.06	
BIC	-75.09	

Tabella 3.2: Noise model SARIMA$(1,0,0) \times (1,1,0)_{12}$

scarsa numerosità della *noise series* che non permette di cogliere al meglio la struttura del modello, oppure all'influenza di fattori esterni che portano a modificarla temporaneamente o permanentemente; sotto quest'ultima ipotesi, tali influenze non considerate potrebbero rendere i modelli SARIMA, stimati in precedenza, distorti in termini di previsione.

3.5 Le variabili d'intervento

I diversi incentivi per l'acquisto di autovetture illustrati nella sezione 3.2, che sono stati introdotti negli anni, sono rappresentabili attraverso delle variabili indicatrici; esse, nei casi più semplici, in genere assumono valore 1 per $t = T$ oppure per $t \geq T$ e 0 altrove. Nel caso in esame, invece, esse assumono valore 1 in un intervallo tra due istanti temporali corrispondenti all'inizio e al termine del periodo di incentivo $T_{start} \leq t \leq T_{end}$. Le codifiche codifiche delle variabili sono indicate nelle 3.29, 3.30, 3.31 e 3.32. Per quanto riguarda la variabile descritta nella 3.29, è stato unito il contributo del 2002 con il successivo del 2003, avendo questo pari condizioni e benefici ed essendo entranto in vigore al termine del precedente. Come effettiva durata di ogni incentivo è stata considerato il periodo che va dal primo mese a partire quale devono essere stati stipulati i contratti, fino all'ultimo mese possibile per l'immatricolazione (quasi sempre successivo al termine ultimo per i contratti d'acquisto); questo fa sì che si possa cogliere l'effetto dell'incentivo anche nei 3 mesi successivi al termine ultimo di acquisto, dove si

manifestano le vendite effettuate entro la fine dell'anno.

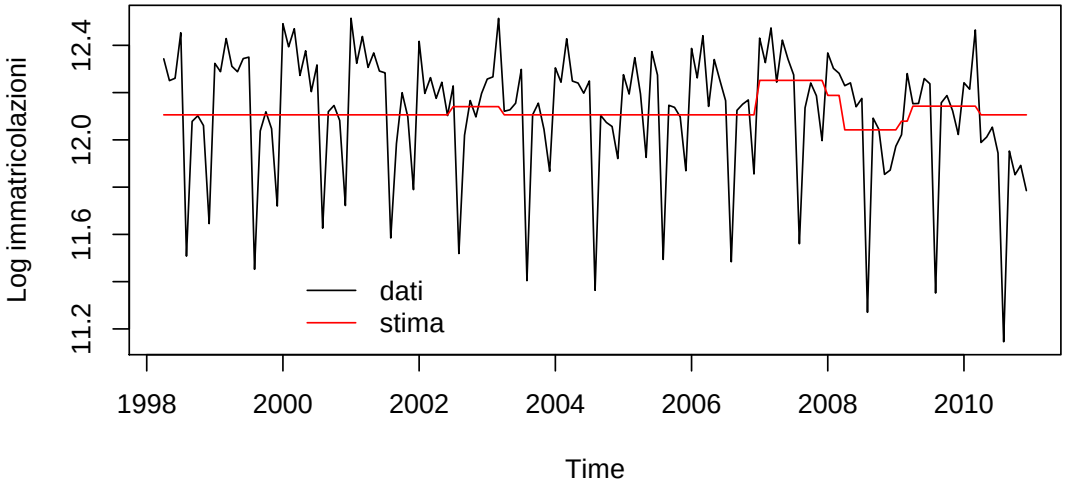
$$I02 = \begin{cases} 1, & \text{luglio 2002} \leq t \leq \text{marzo 2003} \\ 0, & \text{altrove} \end{cases} \tag{3.29}$$

$$I07 = \begin{cases} 1, & \text{ottobre 2006} \leq t \leq \text{marzo 2008} \\ 0, & \text{altrove} \end{cases} \tag{3.30}$$

$$I08 = \begin{cases} 1, & \text{gennaio 2008} \leq t \leq \text{marzo 2009} \\ 0, & \text{altrove} \end{cases} \tag{3.31}$$

$$I09 = \begin{cases} 1, & \text{febbraio 2009} \leq t \leq \text{marzo 2010} \\ 0, & \text{altrove} \end{cases} \tag{3.32}$$

Nella figura **??** è rappresentato il risultato di una regressione lineare della serie logaritmica sulle variabili indicatrici poste come regressori. In questo modo è possibile rappresentare graficamente i salti di livello che essi provocano nella serie. Questo modello lo si può ritenere "naive" e puramente dimostrativo, in quanto i coefficienti così stimati non risultano statisticamente significativi (tranne per la variabile I07, significativa al 5%) (tabella 3.3) e colgono solamente una variazione media rispetto alla media della serie.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.1059	0.0258	468.43	0.0000
incentivi$I02	0.0346	0.0931	0.37	0.7105
incentivi$I07	0.1458	0.0742	1.97	0.0512
incentivi$I08	-0.0636	0.0738	-0.86	0.3898
incentivi$I09	0.0367	0.0761	0.48	0.6304

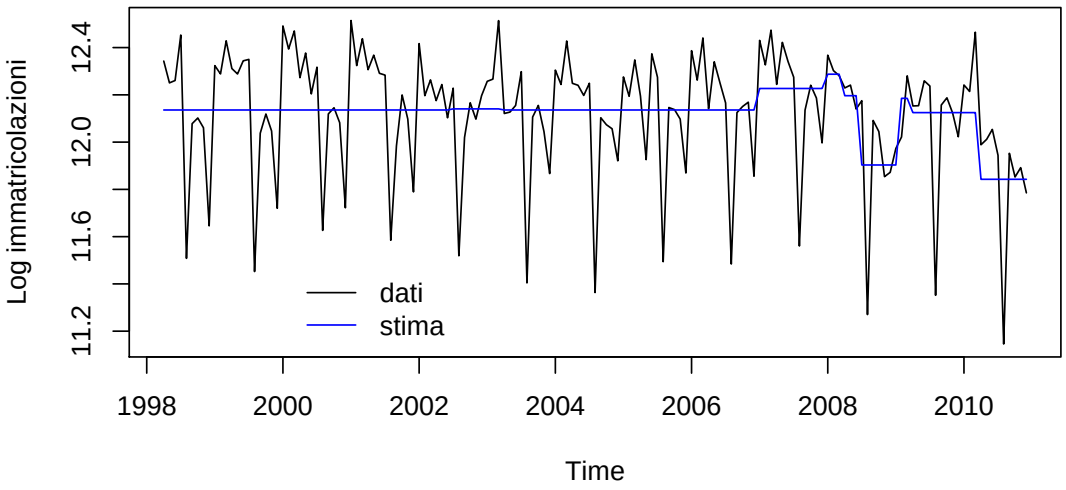
Tabella 3.3: Regressione lineare con gli incentivi come regressori

La linea rossa stimata nel grafico di figura **??** segue la media della serie, salvo deviare nei periodi in cui è presente l'effetto di un incentivo. Tuttavia, nel tratto finale della serie quest'ultima inizia una decrescita apparentemente lineare, che la porta nel tempo a discostarsi significativamente dalla popria media complessiva, facendo ipotizzare una rottura della struttura di origine del modello. La ricerca di uno o più *break* strutturali attraverso un algoritmo in grado di rilevare l'eventuale presenza multipla, come quello proposto da Baj e Perron[2] [Bai and Perron, 2003], individua un punto di rottura corrispondente a luglio 2008. Esso viene segnalato sia applicando la ricerca sui dati originali, sia sulla serie depurata dall'effetto "naive" calcolato degli incentivi (ovvero sui residui della regressione). Questa informazione proveniente dai dati può essere trattata alla pari di un intervento

[2]Si è utilizzata l'implementazione per R nel pacchetto *strucchange*, comando *breakpoints*

esterno e può essere identificata come l'effetto della crisi prima finanziaria, emersa negli Stati Uniti tra la fine del 2006 e il 2007 e propagatasi successivamente nel resto del mondo, e poi di quella economica. La nuova variabile è rappresentata nella (3.33) e l'effetto della sua inclusione nel modello di regressione porta al risultato nella figura **??** con i coefficienti stimati nella tabella 3.4. Da osservare come il coefficiente della variabile $Icrisi$ risulti significativo e, com'è ovvio pensarlo, di segno negativo.

$$Icrisi = \begin{cases} 0, & t < \text{giugno 2008} \\ 1, & t \geq \text{luglio 2008} \end{cases} \tag{3.33}$$



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.1359	0.0259	468.44	0.0000
incentivi$I02	0.0046	0.0894	0.05	0.9586
incentivi$I07	0.0910	0.0723	1.26	0.2105
incentivi$I08	0.0605	0.0775	0.78	0.4362
incentivi$I09	0.2822	0.0967	2.92	0.0041
Icrisi	-0.2932	0.0760	-3.86	0.0002

Tabella 3.4: Regressione lineare con incentivi e variabile "crisi" come regressori

3.6 Stima di modelli d'intervento

Definita la struttura del *noise model* e individuate la variabili esterne che rappresentano la presenza di incentivi statali e l'effetto della crisi, i parametri del modello d'intervento vengono stimati contemporaneamente attraverso il metodo della massima verosimiglianza. Come primo passo si individua un possibile modello inserendo tutti gli interventi a partire dal *noise model*, valutando successivamente possibili aggiustamenti a seconda della significatività dei parametri e scegliendo dei possibili candidati per il modello finale. Successivamente sui candidati si effettua la ricerca di valori che il modello non è stato in grado di spiegare, i cui residui possono denotare la possibile presenza di *outliers* di tipo innovativo o additivo. L'effetto di eventuali IO o AO viene poi stimato nei modelli contemporaneamente agli interventi per giungere poi alla scelta definitiva del migliore modello tenendo conto dell'AICc.

3.6.1 Un modello iniziale

La (3.34) rappresenta il modello iniziale con $[\theta(B)/\psi(B)]a_t$ il *noise model* e $I02$, $I07$, $I08$ e $I09$ le variabili d'intervento degli incentivi individuate nella sezione 3.5 con le relative funzioni di trasferimento; La variabile $Icrisi$ viene considerata anch'essa come un intervento con la propria funzione di trasferimento.

$$Z_t = \frac{\theta(B)}{\psi(B)}a_t + \frac{\omega_{I02}(B)}{\delta_{I02}(B)}I02 + \frac{\omega_{I07}(B)}{\delta_{I07}(B)}I07 + \frac{\omega_{I08}(B)}{\delta_{I08}(B)}I08 + \frac{\omega_{I09}(B)}{\delta_{I09}(B)}I09 + \frac{\omega_{Icrisi}(B)}{\delta_{Icrisi}(B)}Icrisi \tag{3.34}$$

Per quanto riguarda le funzioni di trasferimento bisogna stabilire il numero di parametri della componente AR di $\delta(B)$ e il numero di parametri MA di $\omega(B)$ per le variabili che rappresentano gli incentivi. Dato che dall'acquisto di un auto nuova alla consegna al cliente e quindi all'immatricolazione generalmente trascorre del tempo più o meno lungo, è lecito aspettarsi che l'effetto venga rilevato

in modo ritardato rispetto all'inizio del periodo di validità del contributo per un periodo più o meno lungo a seconda dei tempi di consegna del costruttore per un determinato modello. Si può inizialmente fissare semplicemente a 1 la componente AR della funzione di trasferimento, evitando di aggiungere eccessiva complessità.

Si è deciso di non inserire elementi MA ma di mantenere esclusivamente un coefficiente ω (in seguito chiamato MA0[3]). Essendo ω il salto di livello provocato dall'intervento, qualora dovesse risultare non significativo porterebbe necessariamente a ritenere nullo l'effetto della variabile, per la quale pertanto si può ipotizzare la rimozione. Il risultato del modello iniziale è illustrato nella tabella 3.5.

	coef	st dev
ar1	-0.120	0.084
sar1	-0.341	0.079
I02-AR1	0.057	0.862
I02-MA0	0.038	0.039
I07-AR1	0.255	0.490
I07-MA0	0.056	0.038
I08-AR1	0.353	0.960
I08-MA0	-0.034	0.045
I09-AR1	0.652	0.131
I09-MA0	0.091	0.027
Icrisi-AR1	0.871	0.060
Icrisi-MA0	-0.039	0.015

Tabella 3.5: Modello d'intervento iniziale

Dalla tabella 3.5 si possono trarre alcune indicazioni. Come prima osservazione, il coefficiente di $ar1$ (parametro AR non stagionale) risulta non significativo a differenza di quanto accade nella stima del *noise model* (tabella 3.2). Anche nella stima di varianti del modello d'intervento iniziale esso risulta sempre non significativo, pertanto risulta opportuno escluderlo. Risultano non significativi i parametri I02-MAO e I02-AR1; anche il coefficiente $I08 - MA0$ risulta non si-

[3]Denominazione utilizzata nell'output del comando *arimax*, libreria TSA per R.

gnificativamente diverso da zero, senza contare la violazione della condizione $0 < \delta < 1$ vista nella sezione 3.3. È ipotizzabile quindi l'esclusione dell'effetto degli incentivi del 2002 e del 2008 dal modello, rimuovendo le relative variabili $I02$ e $I08$. Anche l'intervento $I07$ ha il coefficiente ω non significativo, tuttavia anche dall'osservazione del grafico della serie, appare esserci un effetto di crescita delle immatricolazioni. L'intervento che rappresenta l'incentivo del 2009 è significativo, così come l'effetto crisi.

3.6.2 Modifiche al modello iniziale e stima dell'effetto di outliers

Come si è visto, il modello iniziale in cui è stato considerato un ritardo per ogni funzione di trasferimento ha fatto emergere come non tutti gli incentivi abbiano avuto la stessa tipologia di effetto oppure come l'effetto non sia stato tale da essere rilevato significativamente. Fa specie l'incentivo del 2008 che, pur essendo di tipologia simile al precedente e temporalmente continuo, probabilmente ha solamente in parte attutito l'effetto negativo della crisi, senza riuscire però a sostenere sufficientemente il mercato.

Per formulare delle nuove e diverse ipotesi sugli interventi può essere utile osservare i grafici delle figure **??** e **??**, dove erano stati stimati dei semplici effetti a gradini attraverso una semplice regressione della serie sulle variabili indicatrici $I02$, $I07$, $I08$, $I09$ (tabella **??**) con l'aggiunta della variabile $Icrisi$ nel secondo caso (tabella 3.4). Tralasciando la non significatività dei coefficienti in tale modello, alcune indicazioni si possono trarre dall'andamento delle linee stimate tracciate sui grafici. È evidente come l'incentivo del 2002 non produca rilevanti cambi di livello nel breve periodo della sua durata, confermando la rimozione della variabilie $I02$. Per quanto riguarda il contributo del 2007, sembra plausibile poterlo considerare come un semplice salto duraturo della serie e non un incremento graduale. La variabile $I08$, invece, non porta ad incrementi delle immatricolazioni

nel corso del 2008, le quali invece diminuiscono. Solo l'aggiunta della variabile $Icrisi$ che rivela una tendenza di più lungo periodo, porta a mettere in evidenza la decrescita in atto. Infine, i dati della serie corrispondenti al periodo in cui interviene la variabile $I09$ appaiono crescere gradualmente e pertanto è ipotizzabile il mantenimento di un ritardo nella funzione di trasferimento.

Si propongono due possibili varianti. La prima (*Mod1*) prevede nessun parametro AR o MA non stagionale, eliminando di fatto il coefficiente AR rilevato nel *noise model* e che risulta non significativo dopo l'aggiunta degli interventi. Nella seconda variante (*Mod2*) sono stati aggiunti 3 parametri AR in modo analogo al modello **??** (tabella 2.2); tuttavia solo il terzo è risultato significativo e quindi i primi due sono stati fissati pari a zero. Nella tabella 3.6 sono riportati i risultati ottenuti.

	coef Mod1	st dev	coef Mod2	st dev
ar1	0.000	0.000	0.000	0.000
ar2	0.000	0.000	0.000	0.000
ar3	0.000	0.000	0.207	0.083
sma1	-0.593	0.085	-0.602	0.083
I07-MA0	0.076	0.024	0.077	0.027
I09-AR1	0.602	0.129	0.597	0.132
I09-MA0	0.107	0.027	0.109	0.029
Icrisi-AR1	0.783	0.078	0.793	0.077
Icrisi-MA0	-0.058	0.018	-0.056	0.018

Tabella 3.6: Modelli d'intervento candidati, senza l'effetto di outliers.

I coefficienti degli interventi rispettano la condizione $0 < \delta < 1$ e sono significativi in entrambi i modelli. Essi sono simili in entrambi i modelli e la differenza sostanziale è nella presenza di un coefficiente AR sul *noise model* che è cambiato rispetto alla formulazione iniziale.

La ricerca di valori anomali sui residui, svolta attraverso la procedura descritta nella sezione 3.3.1, porta ad individuare in entrambi i casi due punti come possibili *outliers* innovativi. Il primo possibile IO corrisponde a dicembre 2002. Si può ipotizzare come esso sia il momento in cui si è manifestato con maggiore im-

patto l'effetto dell'incentivo del 2002, il quale non si riesce a cogliere attraverso un input di tipo *step*. La breve durata di tale incentivo e il suo inizio nel mese di luglio che precede il naturale calo delle vendite di agosto, ha prodotto conseguenze avvertibili in particolar modo nell'ultima parte dell'anno. Tuttavia l'effetto non si è esaurito a dicembre ma con un picco a marzo 2003 in seguito al prolungamento del contributo nei primi tre mesi dell'anno. Il secondo *outlier* viene rilevato nel mese di maggio 2005, periodo in cui uno sciopero delle bisarche[4] ha aggravato la tendenza negativa del mercato già presente nei primi mesi dell'anno.

	coef Mod 1	st dev	coef Mod 2	st dev
ar1	0.000	0.000	0.000	0.000
ar2	0.000	0.000	0.000	0.000
ar3	0.000	0.000	0.243	0.084
sma1	-0.472	0.100	-0.500	0.090
IO-57	0.378	0.066	0.359	0.063
IO-Dic02	-0.343	0.066	-0.342	0.063
IO-Mag05	0.074	0.020	0.076	0.023
I09-AR1	0.610	0.103	0.599	0.109
I09-MA0	0.106	0.022	0.110	0.024
Icrisi-AR1	0.789	0.063	0.800	0.062
Icrisi-MA0	-0.058	0.015	-0.055	0.015

Tabella 3.7: Modelli d'intervento candidati con l'inclusione di outliers innovativi.

Dalla diagnostica dei residui, si può osservare come nel *Mod2* il test Ljung-Box dia una più forte indicazione di assenza di una loro autocorrelazione rispetto ai residui del *Mod1* (figure **??** e **??**). Sebbene le ACF non presentino particolari strutture che indichino in maniera netta eventuali grossi errori di specificazione del modello, nel *Mod1* sembrano essere presenti in maggior numero delle autocorrelazioni prossime al limite della significatività o leggermente significative, mentre nell'autocorrelazione parziale risulta significativo il lag 3 a differenza di quanto accade nel *Mod2*, nel quale questo ritardo è stato incluso. Infine, i residui in entrambi i casi superano i principali test di normalità (tabella **??**) e il risultato del

[4]Lo sciopero dei trasportatori di autoveicoli viene comunemente chiamato sciopero delle bisarche prendendo il nome dal tipico autocarro utilizzato per il trasporto delle autovetture.

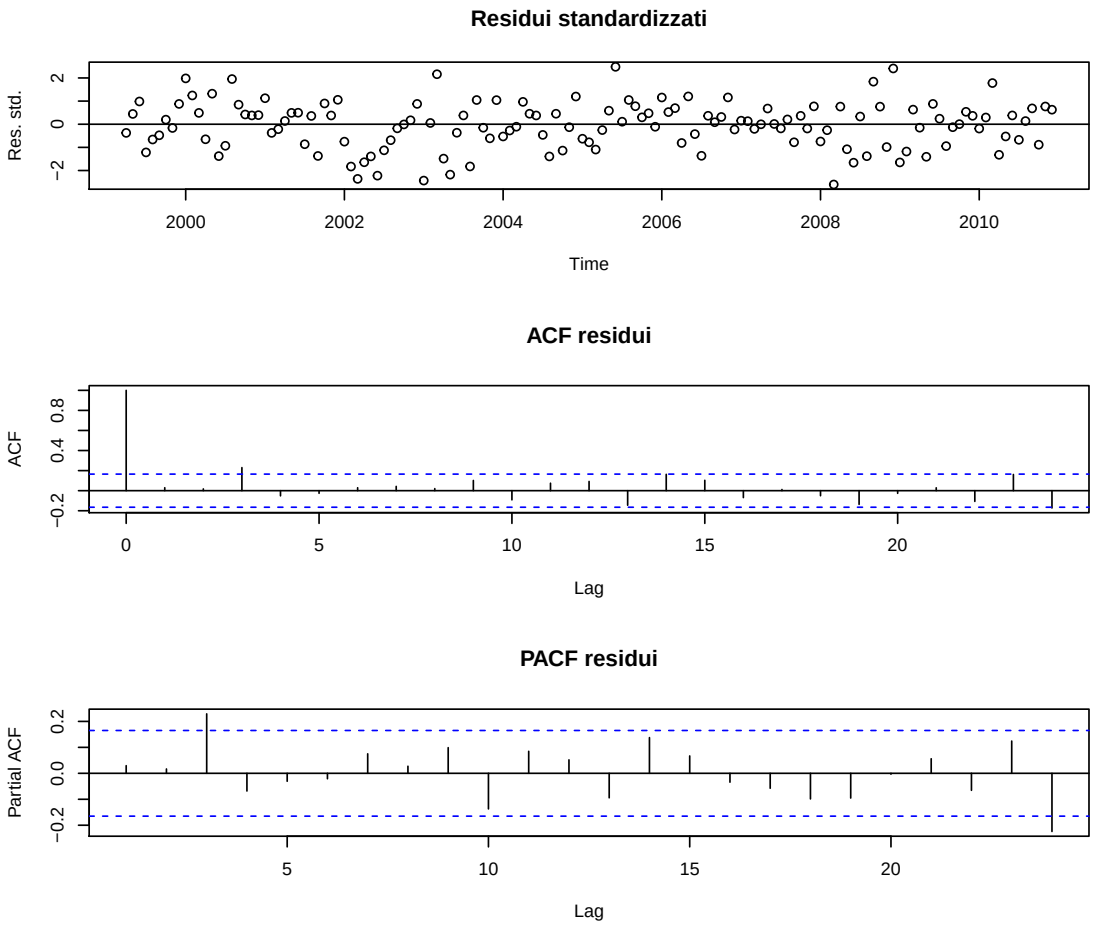


Figura 3.5: Residui Mod1

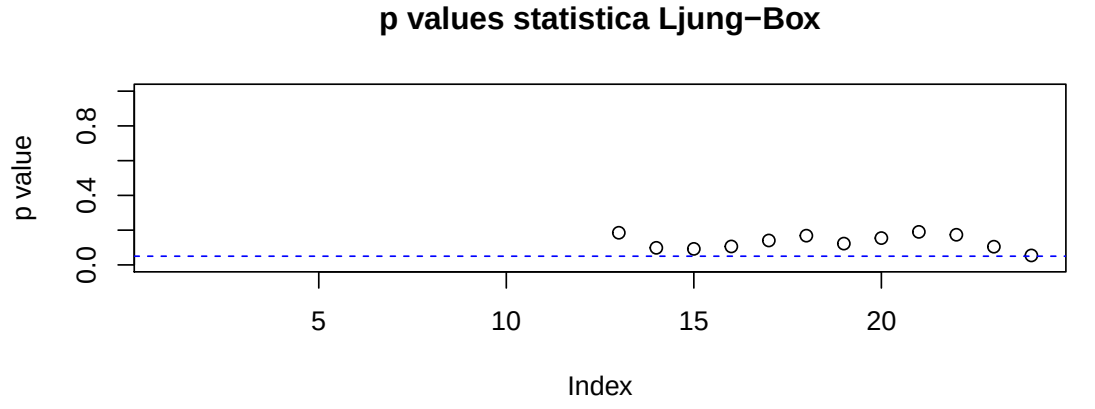


Figura 3.6: p-values della statistica Ljung-Box sui residui, modello Mod1

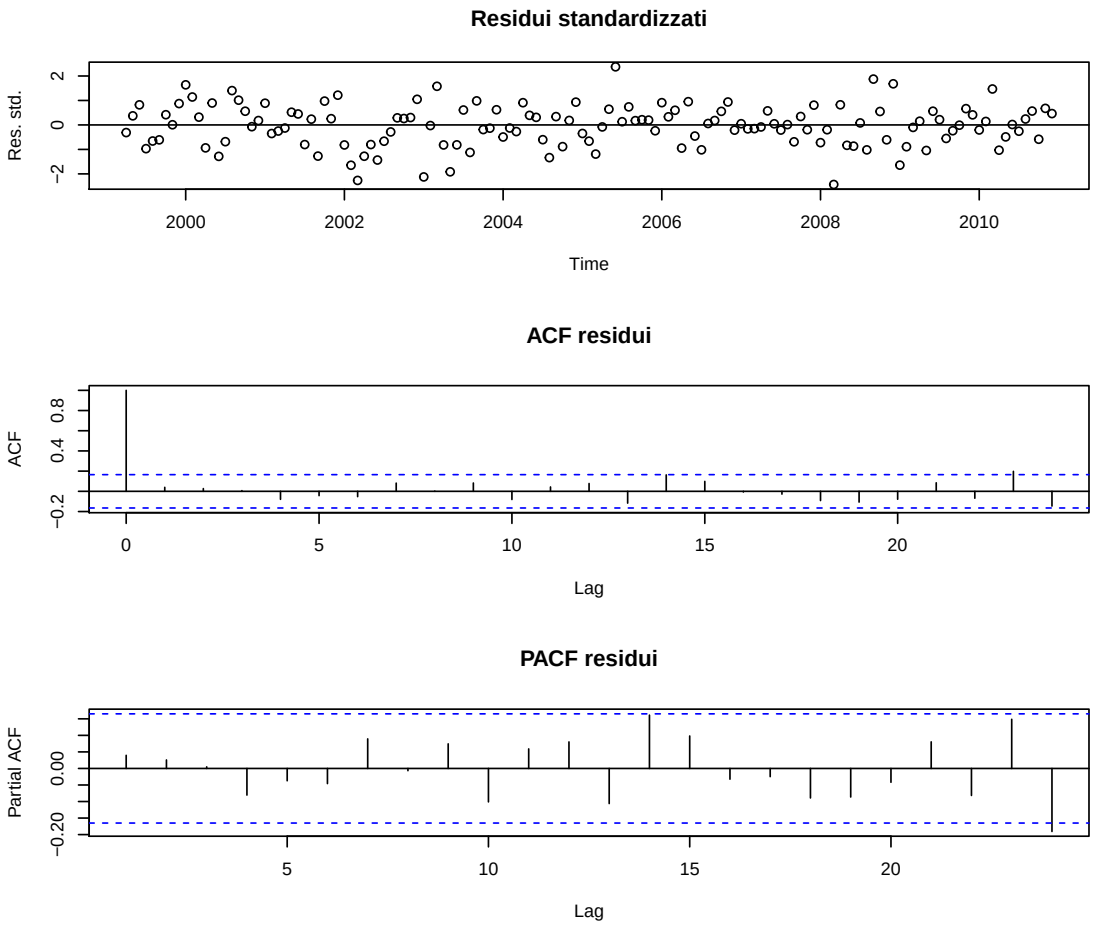


Figura 3.7: Residui Mod2

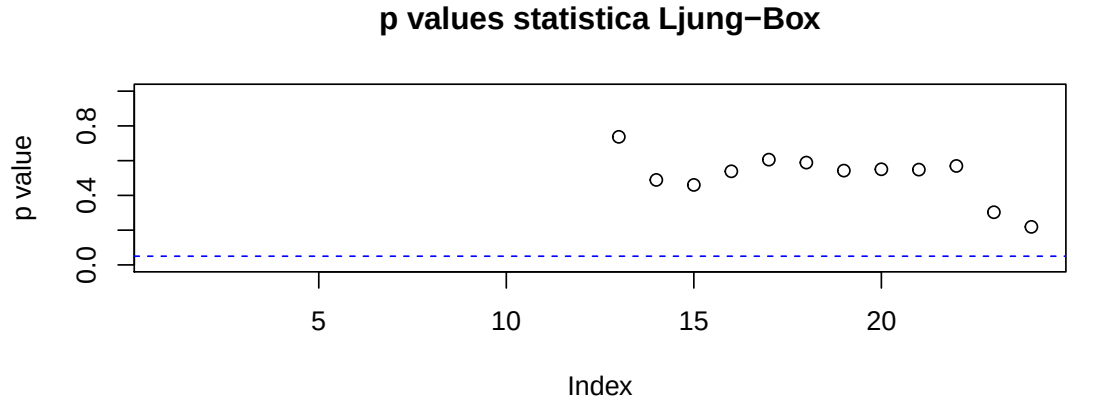


Figura 3.8: p-values della statistica Ljung-Box sui residui, modello Mod2

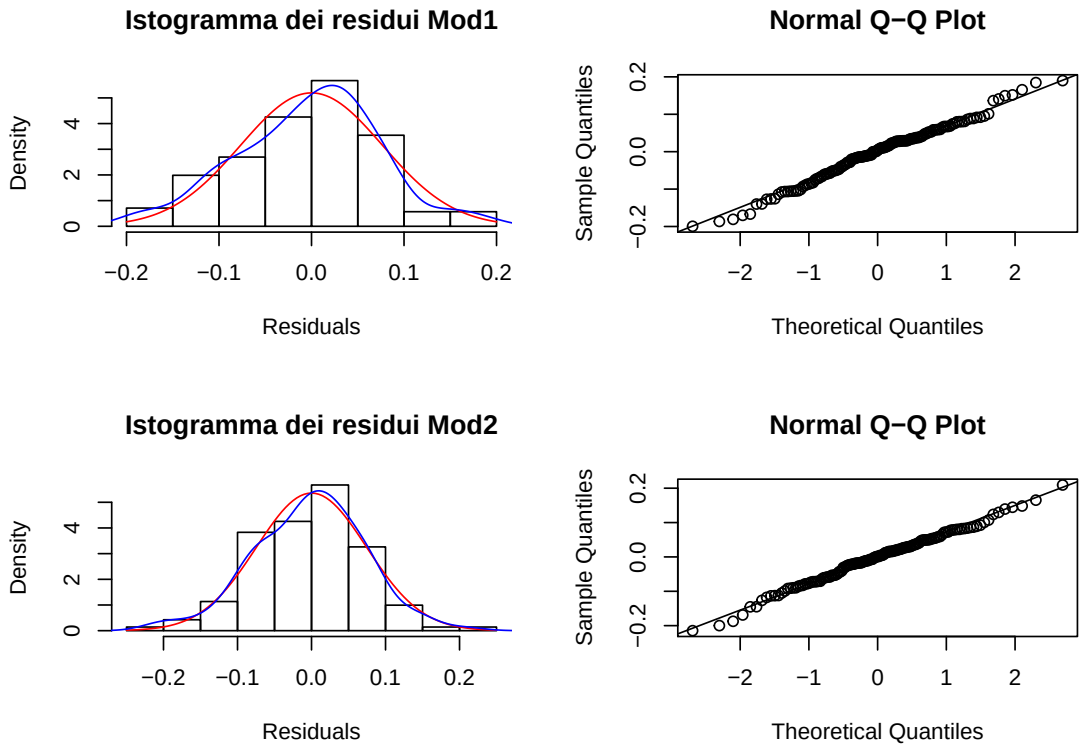


Figura 3.9: Distribuzione dei residui dei modelli Mod1 e Mod2

test ARCH porta a non rifiutare l'ipotesi nulla di assenza di eteroschedasticità in entrambi i modelli (tabella 3.9).

		Mod1		Mod2	
Test	H_0	valore	pvalue	valore	pvalue
Kolmogorov-Smirnov	Normale	0.0514	0.8507	0.0445	0.9431
Shapiro	Normale	0.9887	0.3061	0.993	0.7225
Jarque-Bera	Normale	0.5589	0.7562	1.1715	0.5567

Tabella 3.8: Test di normalità sui residui dei modelli Mod1 e Mod2.

Modello	ARCH LM-Test	p value
Mod 1	11.2388	0.5086
Mod 2	9.3699	0.6711

Tabella 3.9: Arch LM-Test sui residui dei modelli Mod1 e Mod2.

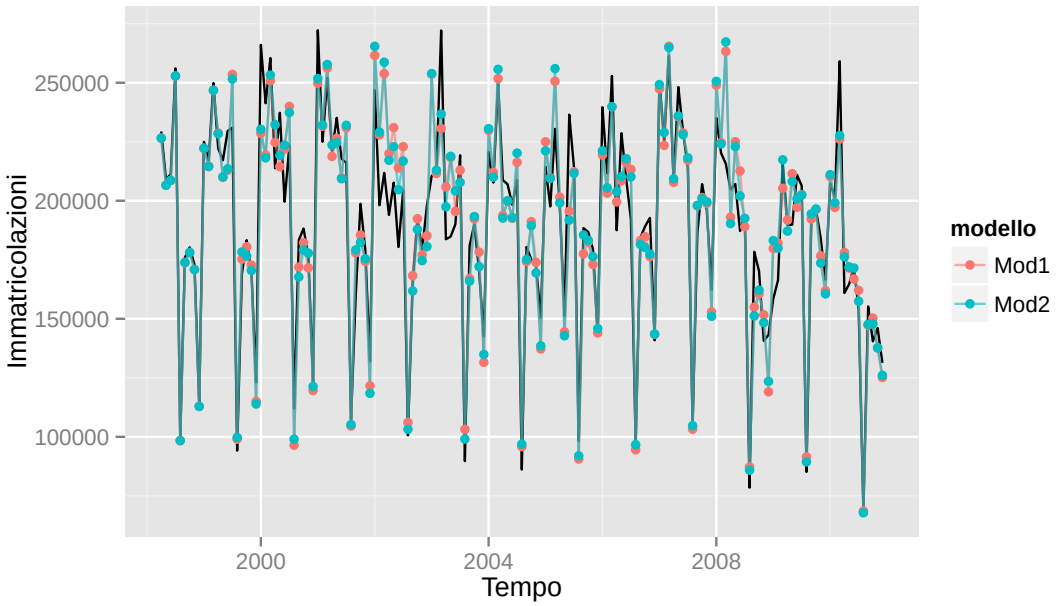


Figura 3.10: Immatricolazioni reali e stimate, Mod1 e Mod2

Nel grafico di figura 3.10 sono rappresentati i dati reali (linea nera) e le stime dei modelli Mod1 e Mod2 espresse in volumi di auto immatricolate.

3.6.3 Inserimento di ulteriori variabili indicatrici

Calcolati i due modelli candidati visti nel paragrafo 3.6.2 e rappresentati i dati stimati (figura 3.10), è possibile aggiungere ulteriori variabili indicatrici qualora si ritenga possano essere rilevanti nel modello.

La prima modifica ai due modelli è l'inserimento di un effetto additivo per i mesi di marzo 2003, 2008 e 2010, al pari di *outliers* additivi, sebbene non individuati come tali dalla procedura vista in precedenza. Nel caso di marzo 2003 e marzo 2010 i modelli non riescono a cogliere il picco dovute al termine ultimo per l'immatricolazione dei veicoli per i contributi statali all'epoca vigenti. Nel mese di marzo 2008, al contrario, i modelli sovrastimano ampiamente il dato reale. Una seconda modifica è l'aggiunta di una variabile indicatrice dei primi 6 mesi del 2002, periodo in cui il mercato italiano ha registrato un'importante flessione e precedente all'introduzione degli eco-incentivi; l'effetto di questa crisi può essere

considerato attraverso un semplice salto di livello.

Le nuove variabili sono illustrate nelle (3.35), (3.36), (3.37) e (3.38). Nella tabella 3.11 le nuove stime dei modelli precedenti, denominati ora Mod1.1 e Mod2.2.

$$Cr02 = \begin{cases} 1, & \text{gennaio 2002} \leq t \leq \text{giugno 2002} \\ 0, & \text{altrove} \end{cases} \tag{3.35}$$

$$Mar03 = \begin{cases} 1, & t = \text{marzo 2003} \\ 0, & \text{altrove} \end{cases} \tag{3.36}$$

$$Mar08 = \begin{cases} 1, & t = \text{marzo 2008} \\ 0, & \text{altrove} \end{cases} \tag{3.37}$$

$$Mar10 = \begin{cases} 1, & t = \text{marzo 2010} \\ 0, & \text{altrove} \end{cases} \tag{3.38}$$

	coef Mod1.1	st dev	coef Mod 2.1	st dev
ar1	0.000	0.000	0.000	0.000
ar2	0.000	0.000	0.000	0.000
ar3	0.000	0.000	0.211	0.088
sma1	-0.397	0.110	-0.440	0.098
Cr02	-0.071	0.024	-0.071	0.026
Mar03	0.121	0.060	0.139	0.059
Mar08	-0.182	0.062	-0.147	0.062
Mar10	0.108	0.076	0.102	0.073
IO-Dic02	0.376	0.059	0.391	0.059
IO-Mag05	-0.348	0.059	-0.347	0.057
I07-MA0	0.083	0.019	0.086	0.021
I09-AR1	0.608	0.097	0.602	0.102
I09-MA0	0.105	0.020	0.106	0.021
Icrisi-AR1	0.798	0.056	0.807	0.056
Icrisi-MA0	-0.057	0.013	-0.053	0.013

Tabella 3.10: Modelli d'intervento Mod1.1 e Mod2.1

Il coefficiente relativo alla variabile $Mar10$ (3.38) non è significativo. Nonostante visivamente sembri necessaria l'inclusione di un'indicatore dell'ultimo meso di immatricolazione per l'incentivo 2009 rispetto ai modelli *Mod1* e *Mod2*, la "cattiva" rappresentazione di marzo 2010 che essi hanno prodotto è frutto degli altri valori anomali il cui effetto è considerato nei modelli *Mod1.1* e *Mod2.1*. Pertanto è possibile rimuovere la variabile (3.38), ottenendo le stime definitive (tabella **??**).

	coef Mod1.1	st dev	coef Mod 2.1	st dev
ar1	0.000	0.000	0.000	0.000
ar2	0.000	0.000	0.000	0.000
ar3	0.000	0.000	0.212	0.089
sma1	-0.403	0.109	-0.455	0.096
Cr02	-0.072	0.025	-0.071	0.027
Mar03	0.121	0.060	0.138	0.060
Mar08	-0.200	0.061	-0.166	0.061
IO-Dic02	0.376	0.059	0.392	0.060
IO-Mag05	-0.348	0.060	-0.346	0.058
I07-MA0	0.082	0.019	0.085	0.021
I09-AR1	0.640	0.085	0.630	0.093
I09-MA0	0.100	0.018	0.103	0.020
Icrisi-AR1	0.800	0.056	0.808	0.056
Icrisi-MA0	-0.057	0.013	-0.054	0.013

Tabella 3.11: Modelli d'intervento Mod1.1 e Mod2.1 escluso l'effetto marzo 2010

L'analisi dei residui mostra come nel caso del modello Mod1 una possibile presenza di autocorrelazione tra i residui nel modello *Mod1.1*, di cui è una derivazione, stando ai bassi *p values* della statistica Ljung-Box (figura 3.12). Risulta migliore la situazione relativa al modello *Mod2.1* (figura 3.14). L'ipotesi di normalità dei residui non viene rigettata dai test *Jarque-Bera*, *Shapiro* e *Kolmogorov-Smirnov*, così come al 5% non viene rigettata l'ipotesi di assenza di eteroschedasticità dal test ARCH (tabella 3.13).

Nel grafico di figura 3.16 sono rappresentati i dati reali (linea nera) e le stime dei modelli *Mod1.1* e *Mod2.1* espresse in volumi di auto immatricolate.

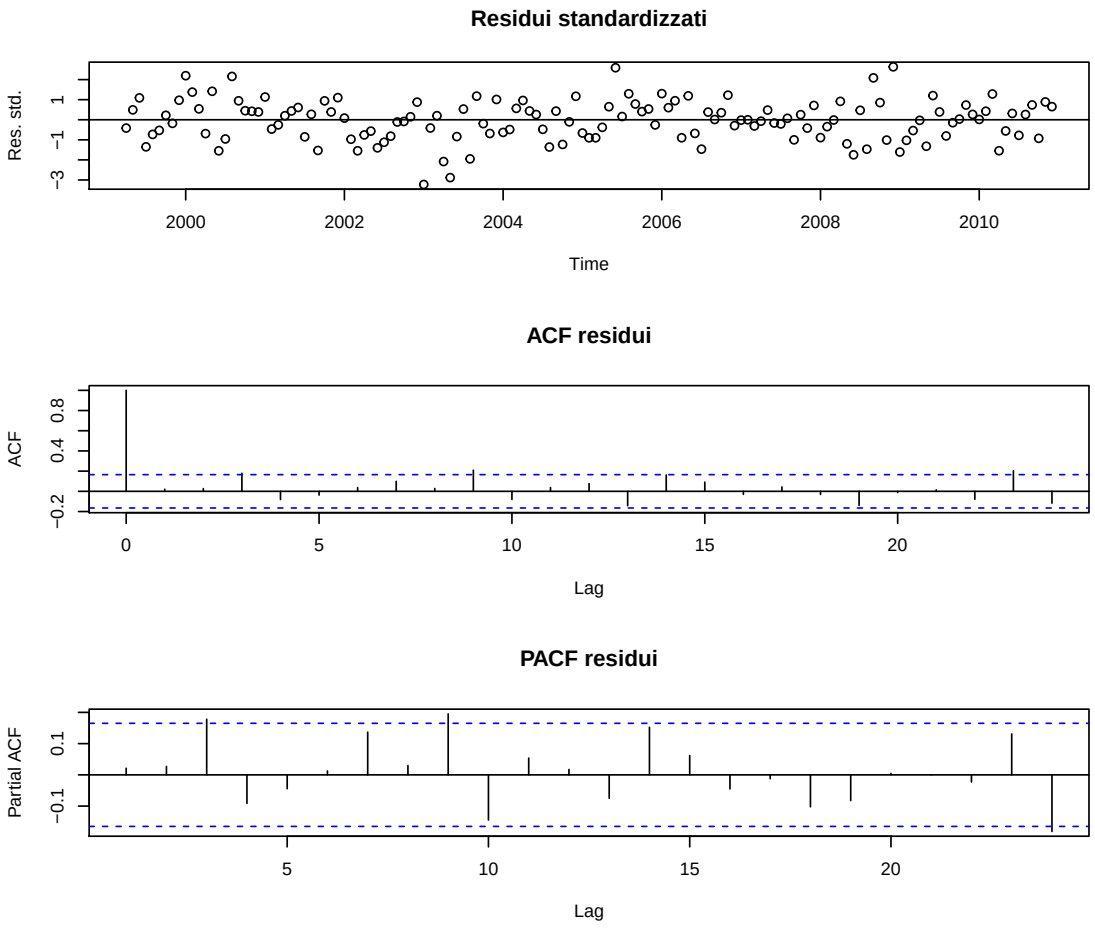


Figura 3.11: Residui Mod1.1

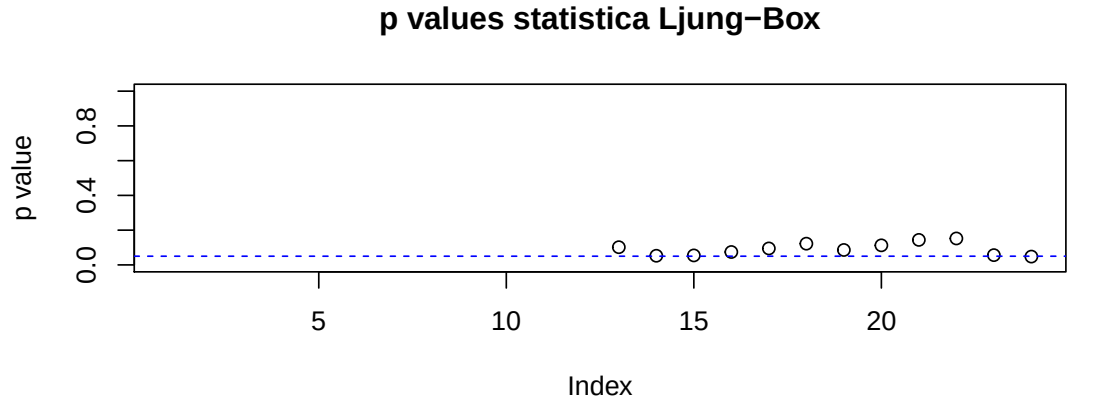


Figura 3.12: p-values della statistica Ljung-Box sui residui, modello Mod1.1

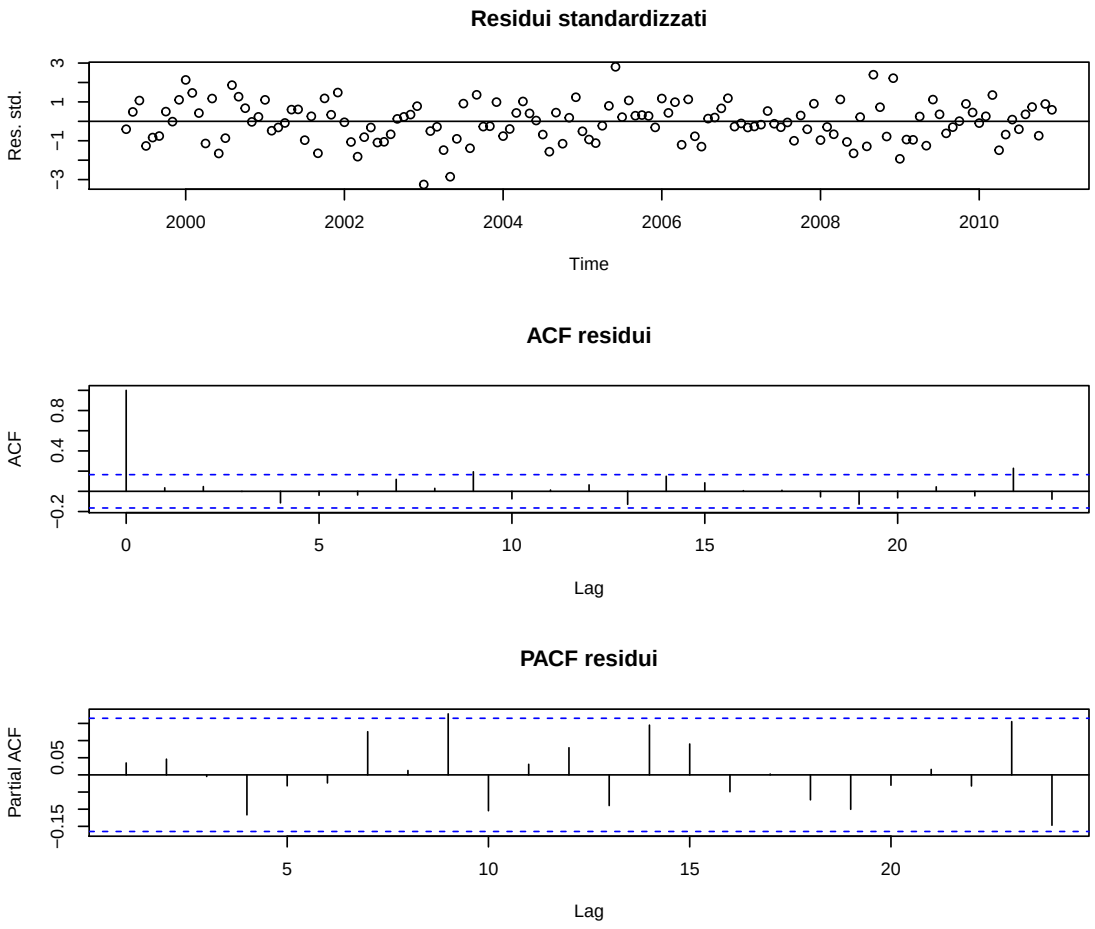


Figura 3.13: Residui Mod2.2

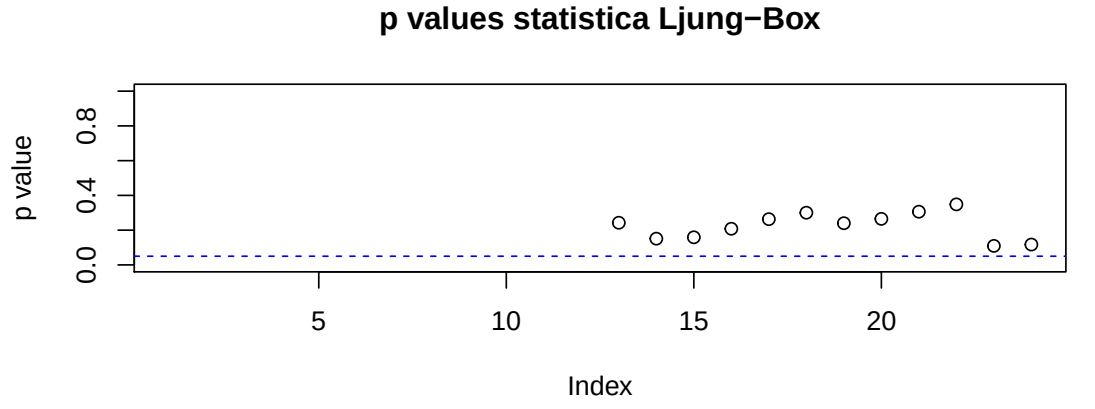


Figura 3.14: p-values della statistica Ljung-Box sui residui, modello Mod2.1

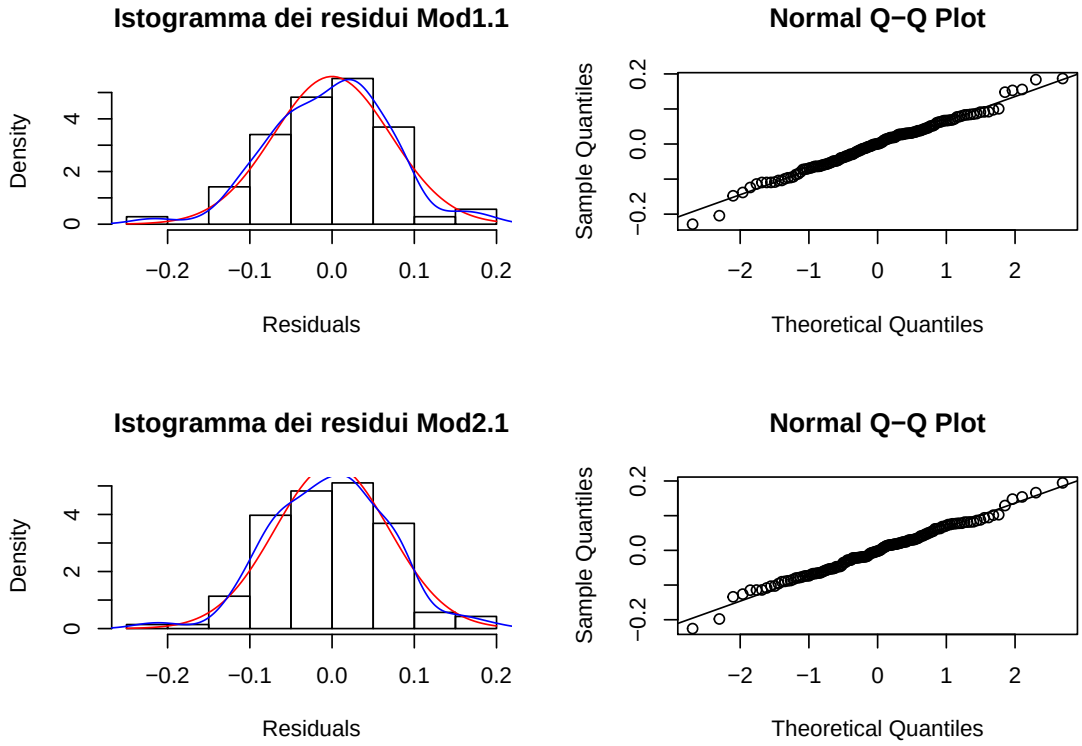


Figura 3.15: Distribuzione dei residui dei modelli Mod1.1 e Mod2.1

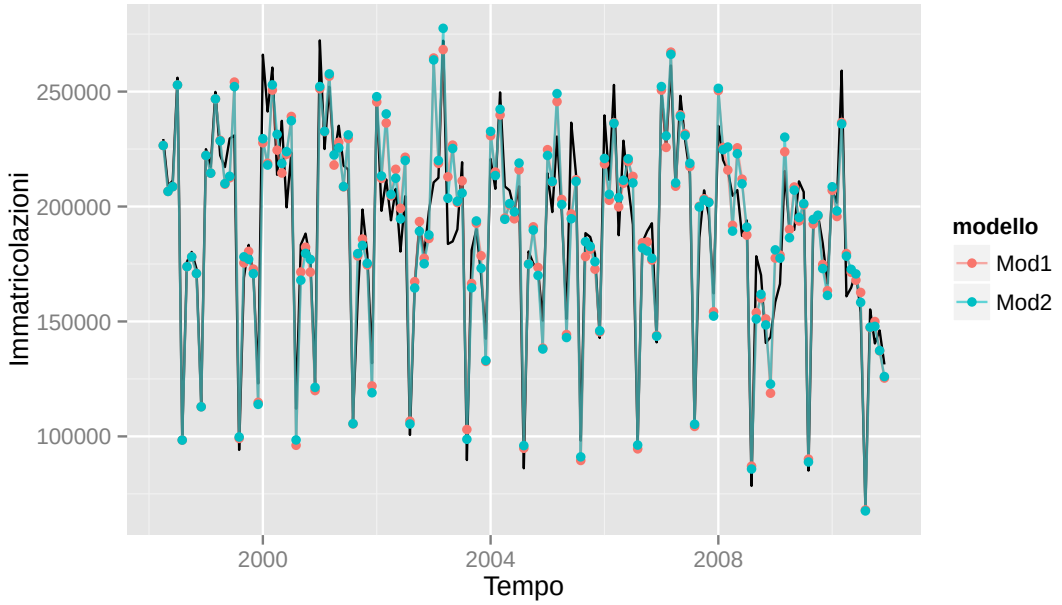


Figura 3.16: Immatricolazioni reali e stimate, Mod1.1 e Mod2.2

		Mod1.1		Mod2.1	
Test	H_0	valore	pvalue	valore	pvalue
Kolmogorov-Smirnov	Normale	0.0477	0.9059	0.0449	0.9392
Shapiro	Normale	0.9896	0.3737	0.9915	0.5592
Jarque-Bera	Normale	1.6297	0.4427	0.9917	0.6091

Tabella 3.12: Test di normalità sui residui dei modelli Mod1.1 e Mod2.1

Modello	ARCH LM-Test	p value
Mod 1	19.1943	0.0839
Mod 2	16.8506	0.1553

Tabella 3.13: Arch LM-Test sui residui dei modelli Mod1.1 e Mod2.1

3.7 Accuratezza dei modelli

Come già visto nella sezione 2.4, è opportuno valutare la performance di un modello attraverso l'accuratezza *out-of-sample*. Nella tabella 3.7 sono indicati gli errori ME, RMSE, MAE, e MAPE relativi alle previsioni ad un passo sugli anni 2011 e 2012 che formano il *test set*. I risultati mostrano performance peggiori rispetto a quelle ottenute con i semplici modelli SARIMA, mentre migliori sono gli errori *in-sample* 3.7. Il modello con minore errore di previsione a breve termine è il Mod2 che include un parametro AR sul terzo ritardo non stagionale e un parametro MA stagionale, con le variabili d'intervento $IO7$, $I09$ e $Icrisi$ che denotano un effetto statisticamente significativo, oltre all'effetto di due *outliers innovativi* (3.7, colonna 2). Tutti i modelli d'intervento proposti dimostrano un netto miglioramento della stima *in-sample*

Modello	ME	RMSE	MAE	MAPE
Mod1	-520	1.4607×10^4	1.0772×10^4	5.72
Mod2	-510	1.4367×10^4	1.0478×10^4	5.5
Mod1.1	-290	1.3127×10^4	9851	5.3
Mod2.1	-403	1.3001×10^4	9807	5.22

Tabella 3.14: Modelli d'intervento: errori di previsione in-sample.

Modello	ME	RMSE	MAE	MAPE
Mod1	-1.8152×10^4	2.4811×10^4	2.0382×10^4	16.93
Mod2	-1.4331×10^4	2.0577×10^4	1.7006×10^4	13.9
Mod1.1	-1.8078×10^4	2.4925×10^4	2.0465×10^4	16.9
Mod2.1	-1.4749×10^4	2.1279×10^4	1.7552×10^4	14.3

Tabella 3.15: Modelli d'intervento: errori di previsione ad un passo.

3.8 Conclusioni

I modelli d'intervento stimati in questo capitolo, offrono un valido strumento di rappresentazione della serie storica dei dati sulle immatricolazioni, riusciendo a cogliere le anomalie della serie e l'effetto degli ecoincentivi statali, fornendo buone stime *in-sample*. Essi si rivelano inoltre migliori dal punto di vista della distribuzione dei residui. Tuttavia accade che questa loro flessibilità nel modellare i dati diventi un limite, nel caso in esame, circa la capacità di prevedere con minore errore i dati futuri. Come spesso accade, un over-fitting rispetto al *training-set* su cui si stima il modello fa perdere in parte la capacità di previsione, soprattuto quando si ha una serie che non mostra una struttura completamente regolare nel tempo. Va considerato inoltre il fatto che negli anni recenti l'effetto della crisi economica, inizialmente in parte mascherato dall'effetto degli incentivi, ha iniziato a produrre una discesa dei volumi di immatricolazioni che nel training-set si presenta solo in parte, rendendo così difficile una stima efficace della funzione di trasferimento della variabile $Icrisi$.

Capitolo 4

Alla ricerca di nuove variabili predittive: Google Trends come fonte alternativa di dati?

4.1 Introduzione

Il web sta diventando una fonte informativa di primaria e crescente importanza nel processo di acquisto di un'automobile [Research and Google, 2011]. Attraverso internet i futuri acquirenti possono visitare i siti dei vari costruttori o dei loro *dealer*, oppure siti terzi specializzati (riviste di auto, forum di discussione di appassionati) dove poter confrontare modelli, conoscere il parere di altri utenti o di esperti o ancora vedere video di prove su strada dei veicoli. Ricerche svolte dall'ICDP[1] mettono in luce come il ruolo della rete nel processo di acquisto di un'auto abbia ridotto il numero di visite presso i concessionari: gli acquirenti continuano a ritenere importante la visita presso il salone di un dealer, tuttavia svolgono la maggior parte del processo decisionale *online* [Tongue and Guillaneuf, 2013].

Uno degli strumenti principali per il recupero di informazioni è l'uso dei motori di ricerca. L'indagine "Gli italiani e i motori di ricerca 2011"[2] ha fatto emer-

[1] International Car Distribution Programme.

[2] Indagine condotta dall'agenzia di *search engine marketing* Sems S.r.l. su 2000 casi rappresentativi della popolazione internet italiana over 16 anni, tramite questionario con struttura dinamica e metodolgia C.A.W.I., reperibile all'indirizzo `http://www.sems.it/about/survey/`.

gere come per il 90% degli italiani che accedono a internet essi rappresentino lo *"strumento più efficace per trovare informazioni"* su un'argomento d'interesse, inclusi prodotti e servizi; il 65% ritiene i motori di ricerca come *"strumento principe per reperire informazioni essenziali in vista dell'acquisto di un prodotto o di un servizio"*. Per il 92%, il motore di ricerca preferito è *Google* [sem, 2011].

Negli ultimi anni sta aumentando sempre più l'attenzione verso i dati provenienti dal web, non più soltanto limitatamente alle statistiche riguardanti i visitatori di un sito, utili a capire la "presenza online" dell'organizzazione (o della persona) a cui appartiene il sito e l'efficacia dei contenuti pubblicati nell'attrarre utenza. Diventano interessanti statistiche che riguardano il comportamento degli utenti nel *web*, come ad esempio poter conoscere l'interesse generale verso particolari termini di ricerca oppure verso un *brand*, un prodotto, così come la misura del "sentimento" degli utenti verso specifici argomenti, anche di ordine sociale e politico[3].

Il quesito che ci si pone è se la vasta mole di dati continuamente raccolti in rete, disponibili in tempi brevi e spesso gratuitamente, possa rivelarsi come un'alternativa alle fonti di dati più tradizionali per analizzare fenomeni sociali ed economici. In particolar modo, nel tema di questo lavoro, può il *web* contribuire allo sviluppo di modelli di previsione multivariati per le immatricolazioni di auto fornendo dati a basso costo o nullo?

[3]Su questo aspetto sta evolvendo negli ultimi tempi la *sentiment analysis* volta a misurare "ciò che si dice in rete" su determinati temi, analizzando i contenuti di blog e strumenti di social network quali Twitter. Risultati interessanti sono quelli ottenuti dal progetto di ricerca *Voices from the Blogs* (VfB) dell'Università degli studi di Milano (*spin-off* da dicembre 2012), ideato da Luigi Curini, Stefano M. Iacus e Giuseppe Porro. Rispetto ad altri progetti di *sentiment analysis* si distingue per l'affiancare alle procedure automatiche una supervisione umana di precodifica, in modo da poter cogliere il contenuto semantico dei testi analizzati.

4.2 Google Trends

Trends è un servizio gratuito messo a disposizione da *Google Inc.*.[4] Esso permette di conoscere l'interesse di ricerca verso parole chiave utilizzate per trovare informazioni attraverso l'omonimo motore di ricerca *Google*. [5] I dati, forniti sotto forma di serie storica settimanale in scala 0-100 a partire dal 2004, riguardano il volume di ricerca di un particolare termine normalizzato rispetto al volume totale di ricerche effettuate nello stesso periodo in un'area geografica; le fluttuazioni dell'indice nel tempo quindi non indicano necessariamente una corrispondente variazione nei volumi di ricerca ma una variazione della popolarità della parola chiave o della categoria sul totale delle ricerche. Oltre a ciò è possibile estrarre il volume generato da una parola chiave in relazione all'interesse di ricerca della categoria di appartenenza. L'estrazione della sola categoria permette di ottenere una serie con le variazioni percentuali dell'interesse rispetto al primo dato del 2004.

Esistono diversi studi che mirano a comprendere le potenzialità dei dati ricavabili da *Google Trends*: essi sono soprattutto *working papers* e *technical reports*, sottolineando come il tema sia ancora nuovo e richieda approfondimenti e revisioni da parte della comunità scientifica e studi indipendenti[6]. In tema di previsione, ad esempio, i ricercatori di *Google* hanno studiato un modello per prevedere "istantaneamente" (o a brevissimo termine) il tasso di diffusione delle epidemie stagionali di influenza in una determinata area geografica[7], anticipando i dati raccolti dagli *U.S. Centers for Disease Control and Prevention* (CDC) [Ginsberg et al., 2009]. Tale modello (che beneficia tuttavia dell'accesso ad una quantità maggiore di dati rispetto a quelli pubblicamente disponibili) è risultato essere

[4] `http://www.google.com/trends/`

[5] Le categorie non sono altro che raccolte di parole chiave inerenti allo stesso tema.

[6] I *technical reports* sono realizzati da *Google* e pubblicati in rete con lo scopo di mostrare le potenzialità dei propri dati, quindi sono necessari maggiori studi a livello accademico, con gli opportuni processi di revisione da parte della comunità scientifica.

[7] `http://www.google.org/flutrends/intl/it/`

sufficientemente accurato in condizioni non distorte da altri fattori.[8]

In campo economico si è visto come alcune chiavi di ricerca mostrino una forte correlazione con il tasso mensile di disoccupazione in Germania [Askitas and Zimmerman, 2009]. Choi e Valrian, senza voler sostenere di poter fare previsioni future, hanno voluto dimostrare come i dati di *Google Trends* possono essere di grande aiuto per "prevedere il presente"[9] anticipando e dati che altrimenti sarebbero disponibili con settimane di ritardo [Choi and Hal, 2011]. Essi hanno inoltre notato come le interrogazioni del motore di ricerca circa la vendita di auto durante la seconda settimana di giugno possa essere utile nel prevedere i risultati rapporto sulle vendite dello stesso mese, solitamente pubblicato nelle successive settimane in luglio. Circa l'utilizzo per previsioni future, ritengono possa essere possibile prevedere le vendite di auto del mese successivo ma molto dipende dalla serie in esame. Nel loro lavoro preliminare [Choi and Hal, 2009] hanno provato a stimare le vendite per alcuni singoli marchi e la presenza o meno di periodi di vendite promozionali poteva in alcuni casi rendere meno accurate le stime.

Uno studio recente invece ha dimostrato come i volumi di ricerca su determinati termini di natura finanziaria (in particolar modo la parola *"debt"*) sono utili nel prevedere le variazioni dei prezzi del *Dow Jones Industrial Average* [Preis et al., 2013]. Tuttavia in questo caso esiste il rischio che una volta scoperta questa correlazione, essa provochi un crescente interesse degli operatori finanziari possa portare ad una distorsione dei volumi di ricerca e quindi rendere meno efficace il modello di previsione.

[8] A febbraio 2013 il modello è stato fortemente criticato dai CDC, a difesa del loro sistema di monitoraggio e diffusione di dati sulla salute nazionale, a causa di una forte sovrastima del picco di ammalati durante l'ultima epidemia di influenza stagionale negli Stati Uniti. Alcuni ipotizzano che la causa si possa ricondurre al maggior interesse dato dai media rispetto agli anni precedenti verso il virus influenzale dimostratosi particolarmente aggressivo, amplificando la preoccupazione della popolazione e quindi provocando una distorsione rispetto al normale volume di ricerca. *Google* dal canto suo conta di poter perfezionare ulteriormente il proprio algoritmo per ridurre questi effetti. Si veda `http://www.nature.com/news/when-google-got-flu-wrong-1.12413`

[9] *"Predicting the present"* che gli autori definiscono anche come *"contemporaneous forecasting"* o *"nowcasting"*.

4.3 Un possibile predittore per il mercato auto

Ci sono molti motivi per cui le persone possono effettuare una ricerca inerente al settore *automotive* attraverso *Google*. Chi naviga in rete può voler cercare informazioni circa i modelli di auto nei siti dei costruttori ma anche siti per l'acquisto di vetture usate, così come siti di concessionari o indirizzi di officine per operazioni di riparazione e manutenzione, sostituzione di pneumatici o ancora l'acquisto di accessori e siti specializzati sul mondo dell'auto.

Per quanto riguarda l'acquisto di auto nuove, l'uso di un motore di ricerca si rivela utile per poter confrontare i modelli presenti sul mercato, leggere recensioni e vedere filmati di prove su strada. È ragionevole pensare che chi prevede di acquistare un'automobile nuova metta in atto un processo di ricerca di informazioni attraverso internet con un margine di anticipo più o meno lungo rispetto alla decisione di acquisto finale (sempre che si manifesti). Secondo un'indagine del 2011 svolta su 13 Paesi (di cui 8 europei)[10], il 65% degli intervistati ha indicato i motori di ricerca come una delle principali fonti per l'ottenimento di informazioni sulle auto, appena dietro alle concessionarie (69%), mentre i siti dei costruttori lo sono per il 42%. Secondo il Google Gearshift 2012, il tempo medio trascorso da quando l'acquirente pensa all'acquisto di un'auto (e inizia a cercare informazioni) al momento dell'acquisto finale è di circa 4.2 mesi.[11]

Le indicazioni provenienti da queste indagini portano in sintesi a considerare che:

- indici sull'interesse di ricerca per parole chiave o per categorie inerenti all'*automotive* potrebbero rappresentare dei predittori per per le immatri-

[10] Netpop Research - Google, *The Role of the Internet in New Car Purchases*.

[11] Nel 2011 risultava di 3.6 mesi e questo porta a pensare che l'uso sempre maggiore dei nuovi media tenda ad allungare i tempi medi del processo anziché accorciarli, probabilmente perché permette una serie di valutazioni più attento. A parere di chi scrive questa tendenza però va valutata in un periodo più lungo perché potrebbe semplicemente essere un riflesso dell'attuale clima di incertezza economica nel mercato europeo che porta a rallentare una decisione d'acquisto importante e magari attendere eventuali periodi promozionali più vantaggiosi. Tuttavia non si hanno dati a supporto di questa ipotesi.

colazioni di auto;

- tali indici potrebbero anticipare la dinamica della serie mensile sulle immatricolazioni.

Per costruire un modello per la previsione delle immatricolazioni di auto si può pensare di estrarre la categoria generale "auto e veicoli" oppure una delle sottocategorie che vengono proposte da *Google Trends*. Per lo scopo di questa analisi sono state scelte le categorie di ricerca con le seguenti denominazioni:

- auto e veicoli;
- acquisto veicoli;
- specifiche, recensioni e confronto di veicoli;
- marche di automobili.

Di queste, tre sono sottocategorie di "auto e veicoli" mentre "specifiche, recensioni e confronto di veicoli" è a sua volta sottocategoria di "acquisto veicoli". Per tutte quante, i dati rappresentano la variazione percentuale dell'interesse di ricerca verso la categoria rispetto al primo dato del 2004. Essendo rilevazioni settimanali, si è resa necessaria un'aggregazione a livello mensile, optando per una media.[12] I grafici delle serie sono visibili in figura 4.1.

Come prima osservazione la categoria "auto e veicoli" sembra essere quella che segue un andamento più regolare e per certi tratti simile a quello delle immatricolazioni. Essendo una categoria di livello più generale rispetto alle altre, si può intendere come una misura dell'interesse complessivo verso il tema auto. Inoltre è evidente come nel periodo in cui erano in vigore gli incentivi statali per l'acquisto di autovetture nuove, la serie delle immatricolazioni mantenga un

[12]Data la presenza di settimana a cavallo tra due mesi per uno o più giorni, l'aggregazione risulta non essere perfettamente fedele al reale dato mensile.

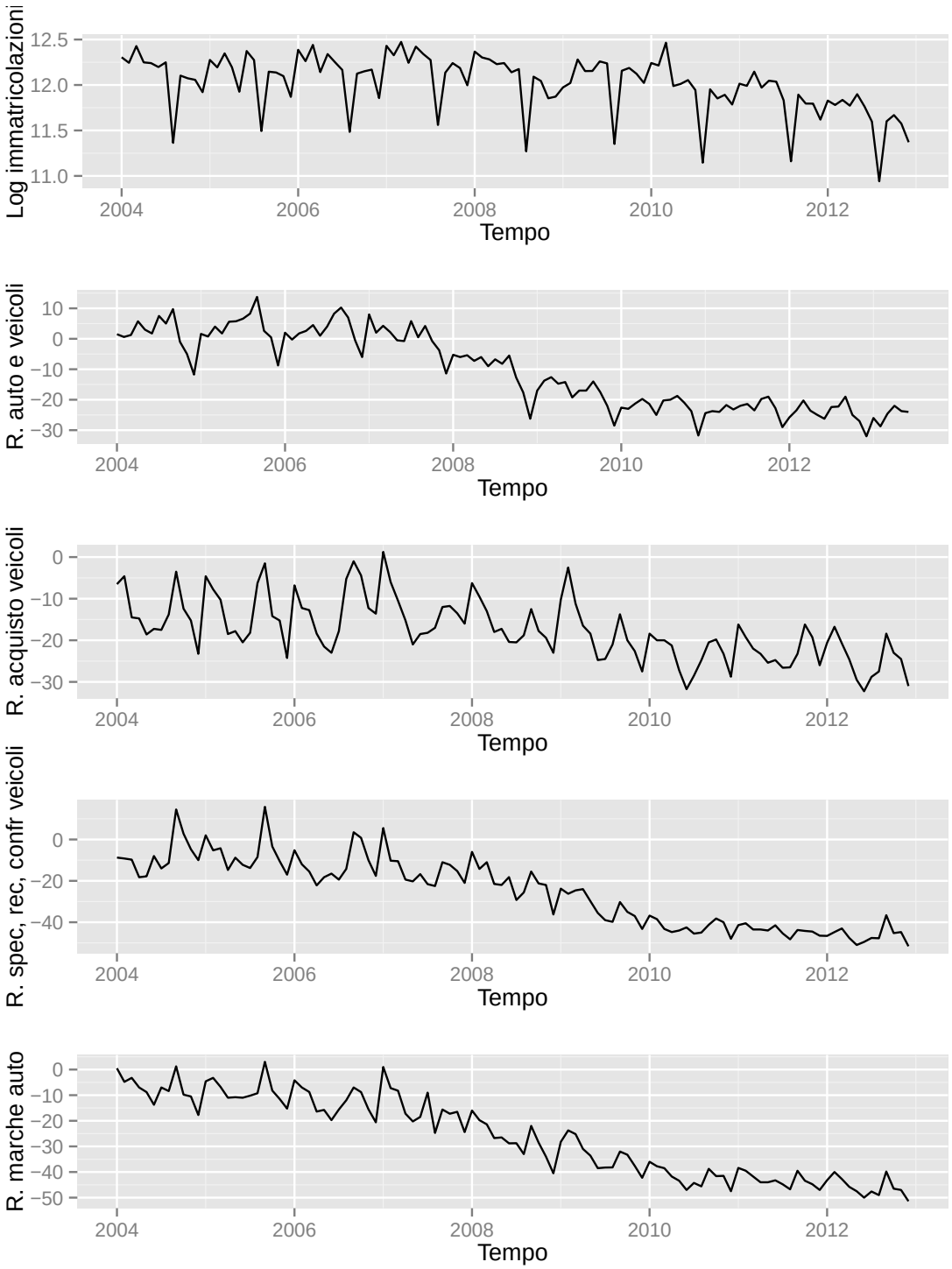


Figura 4.1: Immatricolazioni (logaritmo) e interesse di ricerca per alcune categorie 2004-2012

livello più alto rispetto all'andamento delle altre serie, facendo presupporre che sia necessaria l'inclusione di opportune variabili *dummies*[13].

Alcune prove effettuate attraverso dei semplici modelli autoregressivi stagionali con variabili esogene portano a considerare come predittori relativamene migliori le categorie "auto e veicoli" e "marche di automobili".

4.4 Stima di modelli di regressione

Si affronta ora la costruzione di un semplice modello di regressione lineare per il logaritmo delle immatricolazioni con l'aggiunta di regressori derivanti dai dati sulle categorie di ricerca "auto e veicoli" e "marche di automobili", oltre a considerare come regressore un ritardo di ordine 12 per la serie stessa oggetto di stima. Data anche la scarsa numerosità dei dati a disposizione, questa vuole essere una verifica della possibilità di sfruttare questo nuovo tipo di dati per poter creare dei modelli di facile e veloce implentazione. Pertanto non ci si soffermerà in modo restrittivo sul rispetto delle condizioni sui residui dei modelli per serie storiche, con la consapevolezza che i risultati ottenuti possano essere da spunto per l'individuazione di modelli più sofisticati.

Immaginando il processo d'acquisto di un auto, il futuro acquirente tende ad avviare la ricerca di informazioni attraverso il motore di ricerca Google k periodi in anticipo, ovvero al tempo $t-k$. Ovviamente possono esserci anche persone che iniziano questo processo più di k periodi precedenti (ad esempio al tempo $t-k-1$ o $t-k-2$) e altre meno di k periodi ($t-k+1, t-k+2, \dots$). Per poter inviduare quanto può valere il ritardo k si può ricorrere al calcolo della correlazione incrociata (*cross-correlation*) dopo aver effettuato un pre-sbiancamento delle serie in modo da rimuovere effetti di autodipendenza. Questo pre-sbiancamento avviene ipotizzando che una serie sia indotta dai ritardi dell'altra dopo che entrambe siano

[13] Probabilmente il periodo di incentivi ha portato alla decisione di acquisto anche molta popolazione non abituata ad usare internet i motori di ricerca ma attratta dalla pubblicità nei media tradizionali, elevata in quel periodo

state ripulite dalla struttura di autodipendenza della serie indotta. Questo genera delle correlazioni incrociate per i periodi $\ldots, t-2, t-1, t, t+1, t+2, \ldots$. Se un'ipotetica serie Y_t è indotta dalla serie X_t, saranno presenti una o più correlazioni signficative nei periodi precedenti a t. Nel caso in cui invece si individuassero correlazioni significative per i tempi successivi a t, significa in realtà che è la serie Y_t a indurre la serie X_t, invertendo la relazione ipotizzata. Non è necessario quindi formulare la corretta ipotesi su quale variabile sia dipendente dall'altra. Nel caso in esame ad esempio è stata calcolata la funzione di correlazione incrociata (CCF) invertendo l'ipotesi di dipendenza lineare delle immatricolazioni da un indice di interesse di ricerca poiché la loro struttura ARMA è di più facile individuazione e ha portato a risultati più chiari.[14] Ci si aspetta quindi di trovare una correlazione significativa al tempo $t+k$ (o addirittura più correlazioni significative). Nel grafico di figura 4.2 sono rappresentate le funzioni CCF ottenute tra il logaritmo della serie e l'indice di interesse verso la categoria di ricerca "auto e veicoli" nel primo caso e l'interesse verso la categoria "marche di automobili" nel secondo.

Nella prima CCF risulta plausibile che la serie delle immatricolazioni segua l'indice di interesse per la categoria "auto e veicoli" in particolar modo al tempo $t+8$; gli altri ritardi significativi non sono risultati tali in fase di modellazione, così come appare poco sensato ritenere che l'interesse di ricerca segua l'andamento delle immatricolazioni, come possono far pensare le correlazioni significative ai tempi $t-2$, $t-3$ e $t-4$. [15]. È probabile si tratti di un effetto pi un pre-sbiancamento non ottimale, sufficiente tuttavia a dare le informazioni necessarie. Nella secon-

[14]Non è necessario individuare l'esatto modello ARMA: l'importante è che ci si avvicini ad avere dei residui *white noise*.

[15]Altro non è che un effetto relativo al punto di osservazione. Se si osserva il grafico delle le due serie nella figura 4.1 e si prendono come riferimento i picchi, può sembrare ad esempio che le immatricolazioni anticipino di 4 mesi il successivo picco negativo dell'indice di interesse ma quest'ultimo è in realtà quello che influenza il picco negativo delle immatricolazioni 8 mesi dopo (esattamente 12 mesi dopo il picco negativo delle immatricolazioni da cui si è partiti). Ovviamente viene in aiuto l'idea che sia la ricerca di informazioni tramite Google ad anticipare l'acquisto di un'auto e non il contrario.

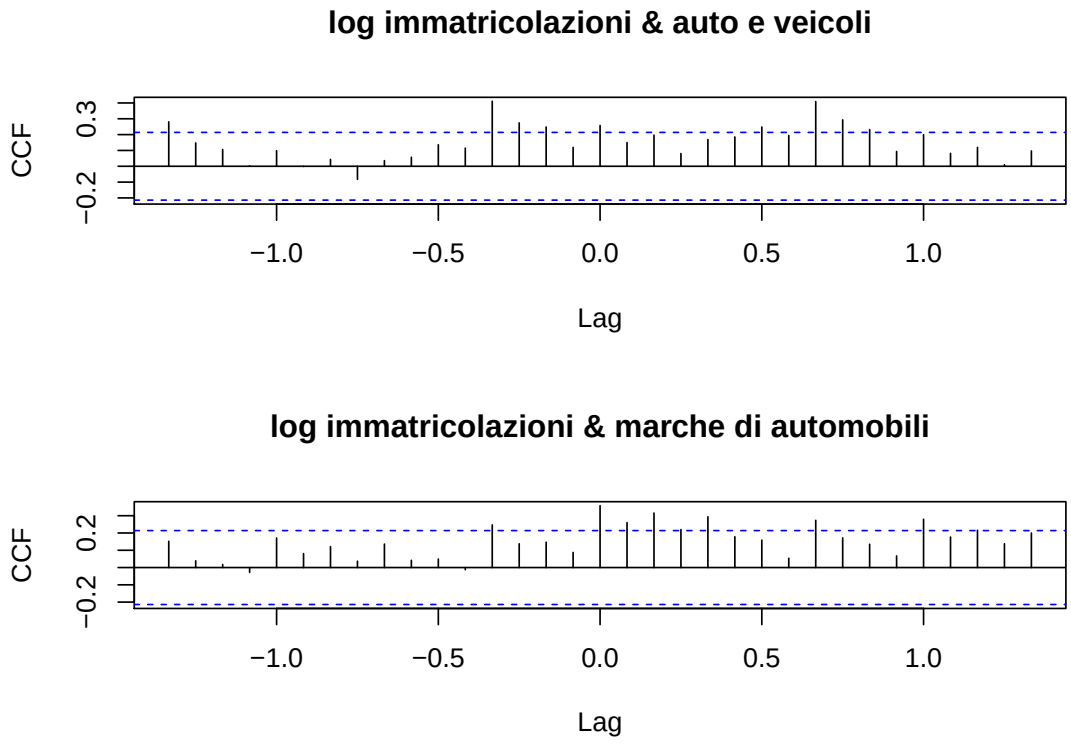


Figura 4.2: CCF tra il logaritmo delle immatricolazioni e le serie esogene

da CCF appare più chiaro come l'interesse di ricerca verso i marchi delle case automobilistiche possa indurre la serie sulle immatricolazioni indicativamente rispetto a diversi ritardi.

Si presentano ora due modelli, tenendo conto dei risultati delle CCF. Varie combinazioni sono state provate utilizzando i diversi *lag* che presentano una correlazione più evidente nelle CCF e le variabili indicatrici per gli incentivi del 2007 (3.30), 2008 (3.31) e 2009 (3.32) già utilizzate nel capitolo precedente. L'inserimento contemporaneo di tutti i ritardi individuati o di gruppi di essi non ha portato a migliorare i modelli, risultando molti di essi non significativi se considerati contemporaneamente agli altri. Vengono proposti dei modelli che considerano in entrambi i casi la variabile esogena con ritardo $k = 8$: nonostante non compaia nella CCF relativa alla variabile "marchi di automobili", l'ottavo ritardo ha dato migliori risultati in termini di bontà di adattamento . La struttura dei modello è

rappresentata nella 4.1

$$Y_t = \alpha + \phi Y_{t-12} + \beta X_{t-k} + \omega I09 + \epsilon_t \ , \tag{4.1}$$

dove X_t rappresenta la variabile esogena (G_auto8 o G_marchi8) la variabile indicatrice per l'effetto dei contributi statali del 2009. In ciascun modello ϵ_t rappresenta la componente d'errore. Le stime sono effettuate su un *training-set* nell'intervallo che va da gennaio 2004 a dicembre 2010. I risultati nella tabella 4.1 denotano per entrambi i modelli un indice R^2 (aggiustato e non) identico. In figura 4.3 e tabella 4.1 le verifiche sulla normalità dei residui. I test di Shapiro e Jarque-Bera propendono per il rifiuto dell'ipotesi di normalità mentre ciò non avviene con il test di Kolmogorov-Smirnov[16]. Dal punto di vista dei grafici di diagnostica non sembrano comunque esserci preoccupnati anomalie nella loro distribuzione.

	Model 1	Model 2
(Intercept)	3.15^{***}	2.27^{**}
	(0.77)	(0.73)
ytm12	0.74^{***}	0.82^{***}
	(0.06)	(0.06)
G_auto8	0.01^{***}	
	(0.00)	
I09	0.23^{***}	0.24^{***}
	(0.04)	(0.04)
G_brands8		0.01^{***}
		(0.00)
N	72	72
R^2	0.80	0.80
adj. R^2	0.79	0.79
Resid. sd	0.12	0.13

Standard errors in parentheses

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Tabella 4.1: Confronto tra il modello con variabile G_auto8 e il modello con G_brands8

[16]Il confronto è svolto tra la distribuzione dei residui e una distribuzione Normale con media 0 e deviazione standard pari a quella dei residui stessi.

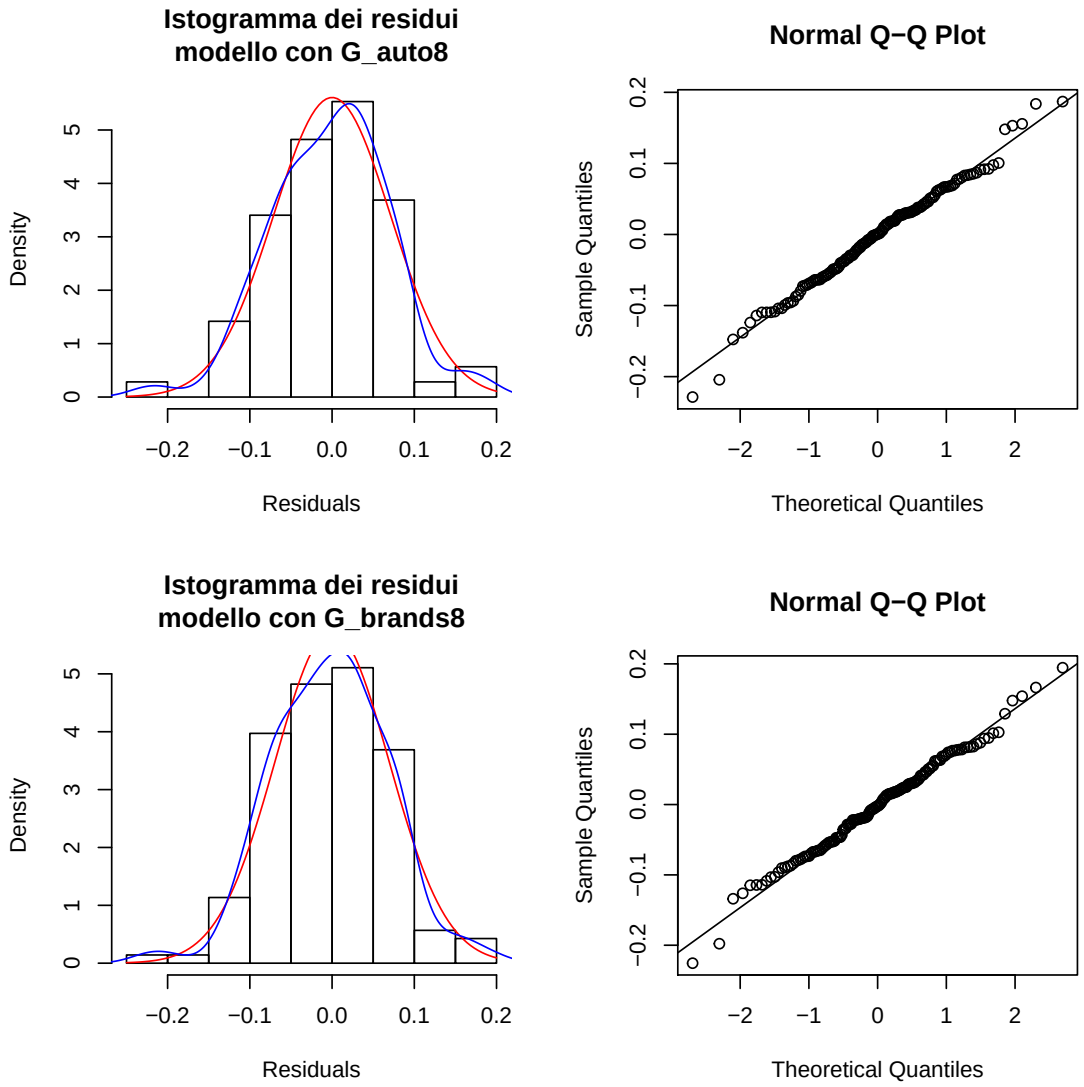


Figura 4.3: Distribuzione dei residui dei modelli com G-auto8 e G-brands8

4.5 Accuratezza dei modelli

Ribandendo il concetto che i modelli stimati in questo capitolo, considerata soprattutto la scarsa numerosità del campione a disposizione, non hanno la pretesa di modellare nel modo migliore possibile i dati, sopratutto dal punto di vista della rimozione di ogni eventuale segnale ancora presente nei residui dovuto alla non inclusione di altri parametri autoregressivi o di tipo media mobile che potrebbero essere presenti, ma hanno lo scopo di esplorare la possibilità che *Google*

		Mod G_auto8		Mod G_brands8	
Test	H_0	valore	pvalue	valore	pvalue
Kolmogorov-Smirnov	Normale	0.1106	0.3184	0.0968	0.4804
Shapiro	Normale	0.9481	0.0049	0.9661	0.0491
Jarque-Bera	Normale	11.354	0.0034	7.6198	0.0222

Tabella 4.2: Test di normalità sui residui dei modelli con G_auto8 e G_brands8

Trends possa essere una fonte di aiuto per effettuare delle previsioni in mancanza di altri dati, è interessante verificare la loro performance *out-of-sample*.

Modello	ME	RMSE	MAE	MAPE
con G_auto8	-7491	1.4978×10^4	1.1855×10^4	9.84
con G_brands8	-9782	1.5608×10^4	1.1738×10^4	9.76

Tabella 4.3: Accuratezza out-of-sample ad un passo.

Il risultato è incoraggiante. A confronto con i modelli dei precedenti capitoli (tenendo conto il campione di numerosità ridotta) i MAPE ottenuti sono il secondo e terzo migliori.

4.6 Conclusioni

L'uso dei motori di ricerca ormai sempre più diffuso nella popolazione può dare indicazione su quelli che sono gli interessi "offline" degli utenti di internet e potrebbero prestarsi ad essere utilizzate per analisi di fenomeni economico e sociali. I risultati ottenuti con dei semplici modelli lineari in cui si fa dipendere la serie delle immatricolazioni con un ritardo di 8 mesi dall'interesse di ricerca verso categorie "auto e veicoli" e "marche di automobili", danno risultati soddisfacenti in termini di errore a breve termine rispetto agli altri modelli considerati in questo lavoro, con un MAPE pari a 9.84 e 9.76. Ulteriori approfondimenti, con l'uso di eventuali altri regressori, potrebbe portare a risultati ancora migliori.

Capitolo 5

Conclusioni

Lungo tutto questo lavoro si è cercato di individuare dei modelli rappresentativi delle immatricolazioni in Italia senza far ricorso a variabili costose ed è stata la loro performance a breve termine. Si è visto come il ricorso a modelli d'intervento che tendono ad essere eccessivamente adattati alle anomalie della serie ha portato ha risultati migliori *in-sample* ma decisamente peggiori in termine di previsione rispetto a dei normali modelli SARIMA. È chiaro come per ottenere risultati migliori, il fenomeno vada analizzato tenendo conto di altri fattori, considerando modelli multivariati più complessi oppure segmentando opportunamente il mercato, suddividendo ad esempio i volumi dei vari marchi o le diverse tipologie di auto, oppure ancora eventuali fattori territoriali. Tutto ciò porta necessariamente all'aumento dei costi di realizzazione dei modelli e competenze specifiche elevate da parte degli analisti. Nel caso in cui si faccia ricorso a dati macroeconomici liberamente disponibili, si aggiunge un problema di scarsa tempestività della loro pubblicazione.

Una possibile nuova fonte di dati, per la quale si potrà conocere meglio nei prossimi anni se si dimostrerà efficace in modo stabile e soprattutto rappresentativa, se non anticipatrice di fenomeni economici e sociali, può essere ricercata nei dati sulle attività di ricerca degli utenti di internet attraverso il motore di ricerca *Google*, disponibili gratuitamente in tempo reale grazie al servizio *Google*

Trends. I risultati su un semplice modello lineare a partire dal 2004 (primo periodo disponibile), hanno permesso di ottenere dei MAPE non molto diversi dal modello SARIMA che ha dato i migliori risultati nelle previsioni di breve termine a un passo. La possibilità di estrarre dati filtrati per diverse categorie o per singole parole chiave e localizzati per aree geografiche permette di sondare ulteriori possibili modelli allo scopo di migliorare le previsioni.

Bibliografia

[sem, 2011] (2011). Gli italiani e i motori di ricerca 2011.

[lon, 2012] (2012). *Long-Term Demand Prediction using Long-Run Equilibrium Relationship of Intrinsic Time-Scale Decomposition Components*, Proceedings of the 2012 Industrial and Systems Engineering Research Conference.

[ANFIA, 2012] ANFIA (2012). Presentazione presidente assemblea dic. 2012 v10 def.

[Askitas and Zimmerman, 2009] Askitas, N. and Zimmerman, K. F. (2009). Google econometrics and unemployment forecasting. *Discussion Paper*, (899).

[Bai and Perron, 2003] Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

[Bardi et al., 2006] Bardi, A., Garibaldo, F., and Telljohann, V. (2006). *A passo d'auto: impresa e lavoro nel settore automobilistico*. Maggioli Editore.

[Box and Tiao, 1965] Box, G. E. P. and Tiao, G. C. (1965). A change in level of a non-stationary time series. *Biometrika*, 52(1-2):181–192.

[Box and Tiao, 1975] Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70(349):pp. 70–79.

[Burnham and Anderson, 2002] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer.

[Carlson, 1978] Carlson, R. L. (1978). Seemingly unrelated regression and the demand for automobiles of different sizes, 1965-75: A disaggregate approach. *The Journal of Business*, 51(2):pp. 243–262.

[Chan and Ripley, 2012] Chan, K.-S. and Ripley, B. (2012). *TSA: Time Series Analysis*. R package version 1.01.

[Chang et al., 1988] Chang, I., Tiao, G. C., and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30(2):193–204.

[Choi and Hal, 2009] Choi, H. and Hal, V. (2009). Predicting the present with google trends - australian conference of economists keynote address.

[Choi and Hal, 2011] Choi, H. and Hal, V. (2011). Predicting the present with google trends - australian conference of economists keynote address. http://people.ischool.berkeley.edu/ hal/Papers/2011/ptp.pdf.

[Cryer and Chan, 2008] Cryer, J. and Chan, K. (2008). *Time Series Analysis: With Applications in R*. Springer Texts in Statistics. Springer.

[dell'Economia e delle Finanze, vari] dell'Economia e delle Finanze, M. (vari). Relazione generale sull'economia del paese.

[Di Fonzo and Lisi, 2005] Di Fonzo, T. and Lisi, F. (2005). *Serie storiche economiche: analisi statistiche e applicazioni*. Università / [Carocci]. Carocci.

[from the Blogs,] from the Blogs, V. Faq.

[Gaynor and Kirkpatrick, 1994] Gaynor, P. and Kirkpatrick, R. (1994). *Introduction to Time Series Modelling and Forecasting in Business and Economcis*. Economic series. McGraw-Hill Ryerson, Limited.

[Gearshift, 2012] Gearshift, G. (2012). Indagine.

[Ginsberg et al., 2009] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Science*.

[Graves, 2012] Graves, S. (2012). *FinTS: Companion to Tsay (2005) Analysis of Financial Time Series*. R package version 0.4-4.

[Hülsmann et al., 2011] Hülsmann, M., Borscheid, D., Friedrich, C. M., and Reith, D. (2011). General sales forecast models for automobile markets based on time series analysis and data mining techniques. In *Proceedings of the 11th international conference on Advances in data mining: applications and theoretical aspects*, ICDM'11, pages 255–269, Berlin, Heidelberg. Springer-Verlag.

[Hurvich and Tsai, 1989] Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

[Hyndman and Koehler, 2006] Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688.

[Italiano and dei Ministri, vari] Italiano, P. and dei Ministri, C. (vari). Leggi italiane e decreti legge.

[Kwiatkowski et al., 1992] Kwiatkowski, D., Phillips, P. C., and Schmidt, P. e Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1–3):159 – 178.

[la Repubblica.it, 2005] la Repubblica.it (2005). Lo sciopero delle bisarche affonda l'auto.

[Li, 2003] Li, W. (2003). *Diagnostic Checks in Time Series*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

[Masarotto, 2009] Masarotto, G. (2009). *ast: ast*. R package version 0.61.

[Paolini, 2005] Paolini, F. (2005). *Un paese a quattro ruote: automobili e società in Italia*. Saggi Marsilio. Marsilio.

[Petri, 2002] Petri, R. (2002). *Storia economica d'Italia: dalla Grande guerra al miracolo economico (1918-1963)*. Vie della civiltà. Il mulino.

[Pfaff, 2008] Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*. Springer, New York, second edition. ISBN 0-387-27960-1.

[Preis et al., 2013] Preis, T., Moat, H. S., and Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Sci. Rep.*

[R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Research and Google, 2011] Research, N. and Google (2011). The role of the internet in new car purchases.

[Ryan and Ulrich, 2013] Ryan, J. A. and Ulrich, J. M. (2013). *xts: eXtensible Time Series*. R package version 0.9-3.

[Shumway and Stoffer, 2006] Shumway, R. and Stoffer, D. (2006). *Time Series Analysis And Its Applications: With R Examples*. Springer Texts in Statistics. Springer-Verlag GmbH.

[Tongue and Guillaneuf, 2013] Tongue, A. and Guillaneuf, C. (2013). Distribution in a digital age. how is costumer behavior evolving and how should the sector respond?

[Trapletti and Hornik, 2013] Trapletti, A. and Hornik, K. (2013). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-32.

[Tsay, 1988] Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, 7(1):1–20.

[Volpato, 2011] Volpato, G. (2011). *Fiat group automobiles: le nuove sfide*. Economia e Management. Il Mulino.

[Wei, 1990] Wei, W. (1990). *Time series analysis: univariate and multivariate methods*. Advanced book program. Addison-Wesley Pub.

[Wickham, 2009] Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

[with contributions from George Athanasopoulos et al., 2013] with contributions from George Athanasopoulos, R. J. H., Razbash, S., Schmidt, D., Zhou, Z., and Khan, Y. (2013). *forecast: Forecasting functions for time series and linear models*. R package version 4.04.

[Zeileis et al., 2002] Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: An r package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38.