

UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

Laurea in statistica e tecnologie informatiche



Relazione Finale

Google Insight for Search e previsione della
disoccupazione: applicazione ragionata al caso italiano

Relatore: Livio Finos

Correlatore: Dario Solari

Laureando: Marco Rinaldo

Matricola: 585868-STI

Anno accademico 2009 - 2010

Ad Anna

Indice

1. Introduzione	5
2. Disoccupazione	7
2.1. Le misure della disoccupazione	7
2.2. Serie storica della disoccupazione	8
2.3. Modelli Arima e Sarima	10
2.4. Previsioni Sarima	18
3. Variabili esplicative: gli indicatori di Google Insight	21
3.1. Descrizione di Google Insight	21
3.2. Come Google elabora i dati	22
3.3. Scelta della chiave di ricerca	23
3.4. Popolarità delle <i>query</i>	29
4. Modelli di previsione	33
4.1. Adattamento dei dati	33
4.2. Modelli stimati	35

5. Analisi delle <i>performance</i> previsionive	55
5.1. Strumenti di confronto di modelli di forecasting	55
5.2. Confronto delle previsioni utilizzando Diebold Mariano	60
5.3. Analisi previsioni ad un passo avanti utilizzando la Leave-One-Out	63
6. Conclusioni	67
7. Appendice	69
8. Bibliografia	75
9. Ringraziamenti	77

1. Introduzione

I motori di ricerca sono sempre di più lo strumento principe per trovare informazioni su Internet. In Italia si contano circa 28 milioni di persone che ogni giorno utilizzano la rete per i più svariati scopi. Parliamo del 58,3% della popolazione italiana tra gli 11 e i 74 anni. Numeri che fanno riflettere, soprattutto se rapportati alle statistiche sull'uso dei motori di ricerca in Italia.

In particolare il 95% degli italiani che accedono al web, reputano i motori di ricerca lo strumento più efficace per arrivare alle informazioni di cui hanno bisogno.

Dopo essere diventato il motore di ricerca più usato a livello mondiale, Google conferma il suo primato anche in Italia. Le ultime stime disponibili danno un quadro molto netto del peso crescente del motore nelle ricerche *on line* effettuate nel nostro paese, passando dal 45% a marzo 2004 ad oltre il 91% nel 2009.

Google, in particolare, risulta interessante in quanto fornisce delle informazioni costantemente aggiornate sulle indagini effettuate, dando quindi un'indicazione quotidiana e gratuita su ciò che il mondo sta cercando.

I dati forniti da Google possono essere utili in svariati campi applicativi come per esempio nel ramo del marketing in cui le informazioni sono particolarmente importanti per valutare l'impatto di un nuovo prodotto sul mercato, altre applicazioni sono riferibili agli indici macroeconomici: i dati distribuiti possono rivelarsi informazioni utili per capire l'evoluzione di questi.

L'obiettivo di questa relazione è giustappunto valutare l'incidenza degli indici forniti da Google nei modelli di previsione; in particolare è stato scelto di stimare il tasso di disoccupazione a partire dai dati forniti dal motore di ricerca.

La scelta è stata supportata dal fatto che la disoccupazione è uno tra i problemi più sentiti nella nostra società soprattutto in questo periodo, dove il numero delle persone in cerca di lavoro risulta pari al 2 milioni 194 mila unità, in crescita del 2,7% rispetto al mese precedente (Febbraio 2010) e del 12 % rispetto al marzo 2009 (dati Istat).

La prima parte della relazione è dedicata all'analisi della serie del tasso di disoccupazione e alla stima di un modello Sarima, che potrà essere utile come confronto con i modelli che utilizzano l'indice di Google fra le esplicative.

Non tutti i dati che Google fornisce tuttavia contengono informazioni utili, infatti in una seconda parte del documento verranno indicate e confrontate le chiavi di ricerca più interessanti ricavando infine quelle più adatte ad essere inserite all'interno di un modello statistico.

La presentazione dei vari modelli idonei a stimare la disoccupazione rappresenta la parte centrale della relazione: a tal proposito sarà presentato un modello che utilizza esclusivamente l'informazione proveniente dalla serie del tasso di disoccupazione e altri che sfruttano anche i dati provenienti da Google.

Nell'ultima parte verrà valutata l'accuratezza delle previsioni attraverso un confronto tra modelli utilizzando test e procedure di validazione.

Importante è evidenziare che tutti i modelli proposti sono concepiti esclusivamente per effettuare previsioni utilizzando le informazioni più aggiornate fornite da Google e non per cercare di spiegare la causalità tra le variabili in gioco; è evidente infatti che l'indice fornito da Google non influenzi il tasso di disoccupazione ma sia vero il contrario.

2. Disoccupazione

Le misure della disoccupazione

La disoccupazione è la condizione di mancanza di un lavoro per una persona in età lavorativa (da 15 a 74 anni) che lo cerchi attivamente.

Il tasso di disoccupazione è un indicatore statistico del mercato del lavoro ed è compreso tra i principali indicatori di congiuntura economica.

L'obiettivo primario di tale indice è di misurare una tensione sul mercato del lavoro dovuto ad un eccesso di offerta (da parte dei lavoratori) rispetto alla domanda (da parte delle imprese), mentre non è adatto a misurare tensioni dovute a mancanza di manodopera (ricercata dalle imprese).

Il tasso di disoccupazione misura solitamente la percentuale delle forze lavoro che non riescono a trovare occupazione e pertanto viene definito come:

$$\text{Tasso di disoccupazione} = \frac{\text{Persone in cerca di lavoro}}{\text{Forza lavoro}} \times 100 \text{ }^1$$

¹ La forza lavoro è la somma delle "persone in cerca di lavoro" e degli "occupati".

Serie storica della disoccupazione

In questa sezione verrà stimato un modello che utilizza esclusivamente l'informazione della serie osservata, con l'obiettivo di creare un termine di confronto con i modelli che includono la variabile relativa ai volumi di ricerca di lavoro su Google.

Prima di stimare tale modello è necessario analizzare la serie storica mensile della disoccupazione italiana resa disponibile dall' Istat.

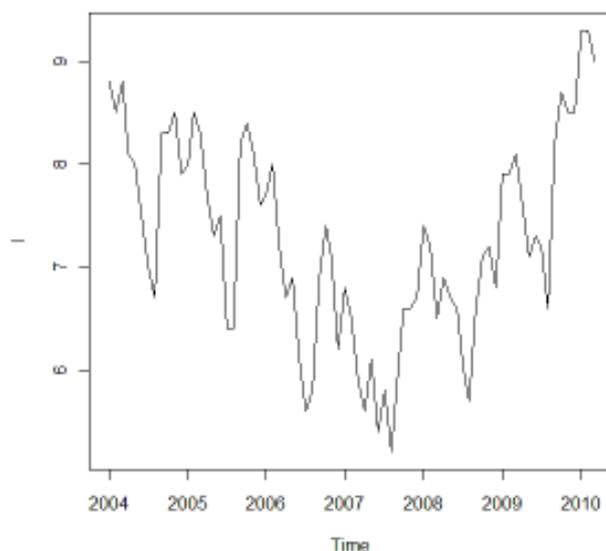


Figura 1: Serie Storica Tasso di disoccupazione.

Da una prima analisi grafica la serie storica non sembra stazionaria in media poiché presenta un trend che nella prima parte è decrescente mentre dalla seconda metà del 2008 l'andamento di questa risulta marcatamente crescente.

La causa dell' inversione di tendenza è attribuibile alla crisi economica che ha colpito il nostro paese.

La serie presenta inoltre una componente stagionale che non sembra influenzata dal trend.

Le considerazioni appena sviluppate risultano ancora più evidenti nel grafico successivo dove vengono scorporati dalla serie la componente stagionale, il trend e l'errore².

² L'estrapolazione del trend, componente stagionale ed errore è resa possibile grazie a funzione presente in R chiamata STL che si basa su una regressione polinomiale locale oppure su medie mobili

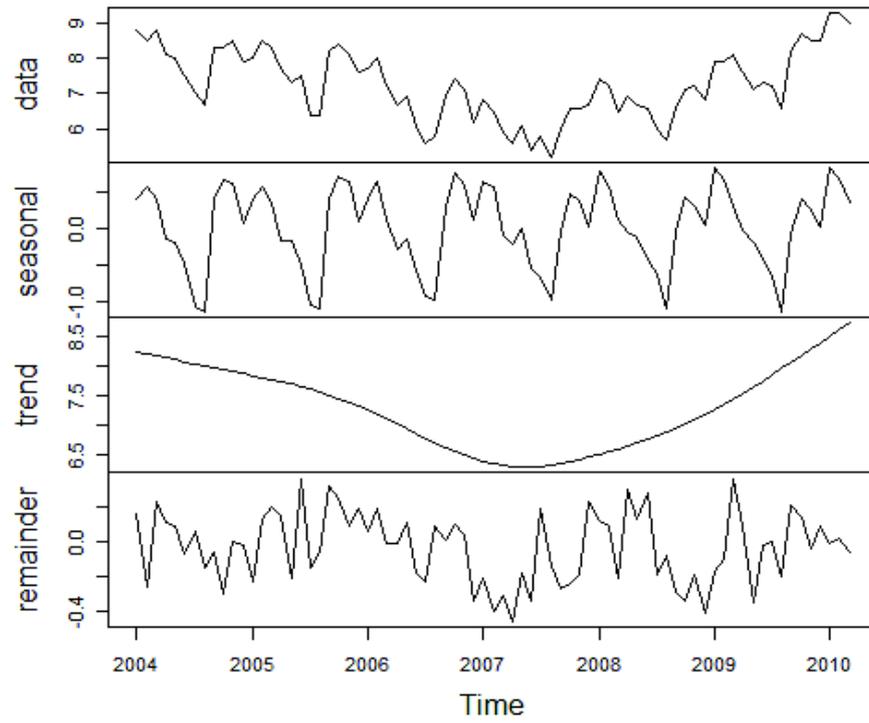


Figura 2: scomposizione delle componenti trend, stagionalità ed errore della serie del tasso di disoccupazione.

Modelli Arima e Sarima³

I modelli ARIMA (autoregressivi integrati a media mobile) di Box e Jenkins partono dal presupposto che fra due osservazioni quello che altera il livello della serie è *white noise*.

Un modello generale di Box-Jenkins viene indicato come: ARIMA (p,d,q) dove AR = AutoRegression (autoregressione) e p è l'ordine della stessa, I = Integration (integrazione) e d è l'ordine della stessa, MA = Moving Average (media mobile) e q è l'ordine della stessa.

Pertanto un modello ARIMA (p,d,q) è analogo ad un modello ARMA(p,q) applicato alle differenze d'ordine "d" della serie dei valori.

Spesso le serie storiche mostrano un andamento periodico, ovvero un comportamento che si ripete in intervalli temporali successivi con periodicità s.

Una serie si definisce stagionale quando mostra un andamento simile di anno in anno, quindi $s = 12$ se le osservazioni sono osservate mensilmente mentre $s = 4$ se la rilevazione è trimestrale.

Una serie storica che presenta un comportamento puramente stagionale sarà sempre la realizzazione finita di un processo non stazionario in media infatti le osservazioni relative a ciascuna stagione fluttuano attorno ad un valore medio diverso da periodo in periodo.

I Modelli SARIMA servono per descrivere oltre alla correlazione tra valori consecutivi (che è quella stimata dagli ARIMA), anche quella presente fra osservazioni che distano di s periodi.

La serie storica mensile del tasso di disoccupazione presenta una media non costante nel tempo, è possibile rendere stazionario il processo utilizzando gli operatori differenza.

³ Jonathan D. Cryer and Kung-Sik Chan (2008). *Time Series Analysis With Applications in R Second Edition*.

Nel caso oggetto di analisi è opportuno differenziare prima stagionalmente la serie: questa operazione permette nella maggior parte dei casi di eliminare oltre alla componente stagionale anche il trend.

La serie differenziata si presenta come nella figura 3.

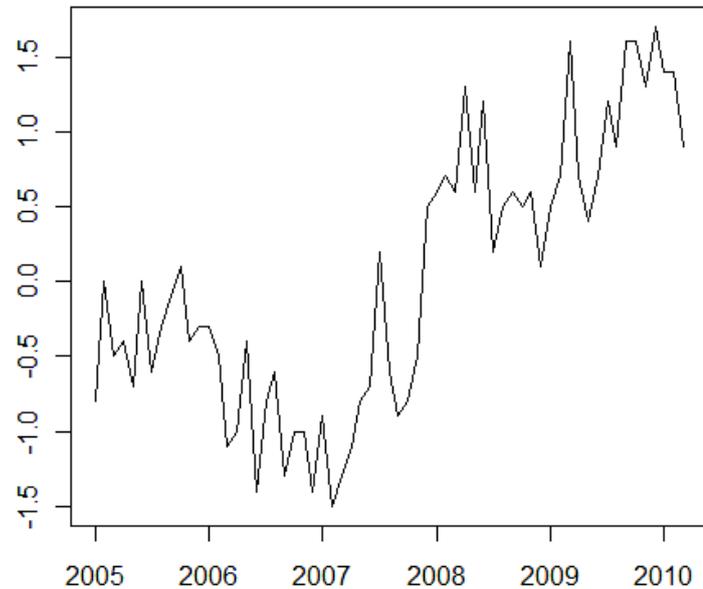


Figura 3: serie storica tasso di disoccupazione differenziata stagionalmente.

Il processo è ancora non stazionario in media, ma la componente stagionale è stata eliminata.

È importante notare che la serie differenziata presenta 12 osservazioni in meno rispetto alla serie di partenza.

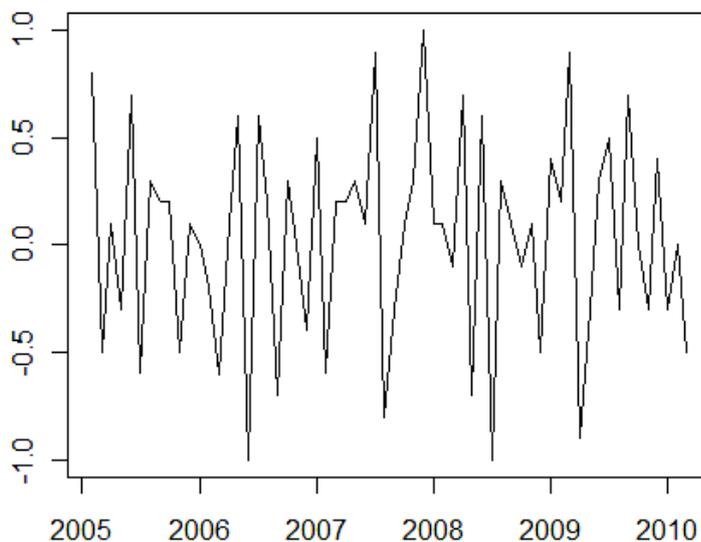


Figura 4: serie storica tasso di disoccupazione differenziata attraverso una differenza stagionale ed una semplice.

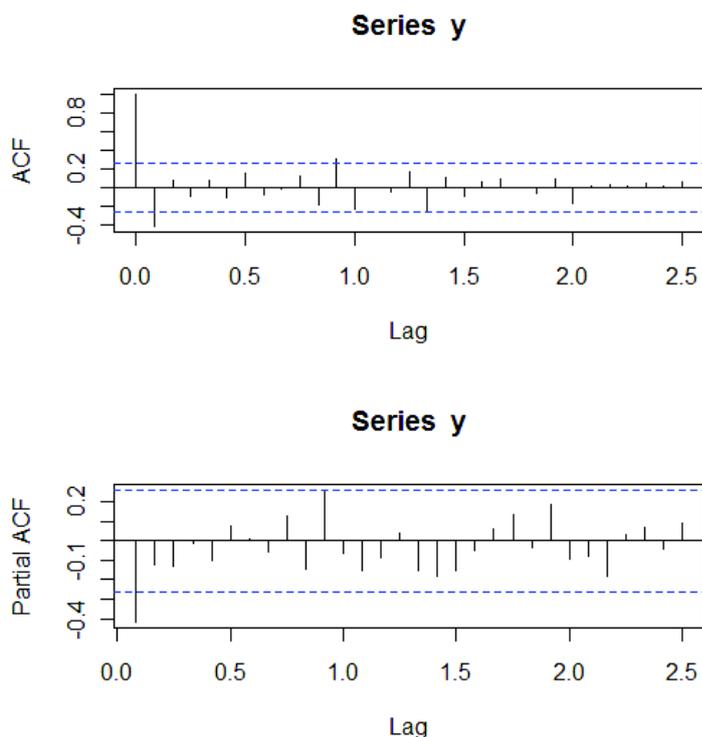
Per eliminare la componente di trend residua è opportuno applicare alla serie oltre alla differenza stagionale, anche una differenza semplice.

Il nuovo processo, il cui grafico è riportato a fianco risulta stazionario e privo di elementi stagionali.

Per stimare il modello è doveroso analizzare i grafici di autocorrelazione

totale (ACF) e parziale (PACF) della serie differenziata.

Al primo e al undicesimo lag l'autocorrelazione del PACF e ACF è significativa: potrebbe



essere opportuno stimare un modello con due parametri: entrambi autoregressivi a ritardi rispettivamente 1 ed 11; un'altra scelta possibile è quella di trascurare la correlazione a lag 11 e stimare modelli che utilizzano esclusivamente quella presente al primo ritardo.

Figura 5: autocorrelogramma totale e parziale della serie del tasso di disoccupazione differenziata.

E' tuttavia difficile riuscire a stimare il modello migliore per cogliere l'andamento del tasso di disoccupazione solo dall'analisi dell'autocorrelogramma totale e parziale, è opportuno quindi stimare i modelli che potrebbero, almeno da un primo esame, adattarsi ai dati, e poi scegliere il modello più idoneo fra questi.

I metodi che consentono di selezionare modelli alternativi partono dal seguente presupposto: poiché esiste *un trade-off* tra complessità di un modello stimato e adattamento del modello ai dati (tra varianza e distorsione, tra precisione e complessità), è necessario prevedere una penalizzazione crescente al crescere del numero di parametri del modello.

I criteri più frequentemente usati sono⁴:

- Il criterio di informazione di Akaike (AIC) è definito come:

$$AIC = -2 \log L_k + 2k;$$

- Il criterio di informazione di Schwartz (Bayesian Information Criterion, BIC)

$$BIC = -2 \log L_k + k \log(n)^5.$$

Il BIC propone una penalizzazione maggiore per il numero di parametri.

La regola è quella di preferire i modelli con l'AIC e/o il BIC più basso.

Si è operata la scelta di togliere alcune osservazione dalla serie, più precisamente ne saranno tolte 15 che ritorneranno in seguito utili per misurare le capacità predittive dei vari modelli.

I modelli stimati e i relativi criteri sono qui sotto elencati:

Modello	p	d	q	P	D	Q	BIC	AIC
1	1	1	0	0	1	0	59.58296	57.48861
2	0	1	1	0	1	0	59.94307	57.84873
3	1	1	1	0	1	0	63.18767	58.99898

Tabella 1: Modelli Sarima che non utilizzano la correlazione al lag 11 e relativi valori Aic e Bic

I criteri analizzati sono concordi nel classificare il modello migliore che risulta essere il quello autoregressivo di ordine 1 con una differenza stagionale e una semplice.

È opportuno ora confrontare il modello appena menzionato con un altro che sfrutti anche la correlazione presente al lag 11.

⁴ Jonathan D. Cryer and Kung-Sik Chan (2008). *Time Series Analysis With Applications in R Second Edition*.

⁵ k : numero di parametri nel modello;

L_k : è il valore massimo della funzione di verosimiglianza del modello stimato (k parametri).

I due modelli esaminati sono quindi i seguenti:

- Modello 1 = SARIMA((1,1,0),(0,1,0));⁶
- Modello 2 : è un autoregressivo di ordine 11 con parametri $\Phi_{2,\dots,10} = 0$ e con una differenza stagionale ed una semplice

Lo studio ha la finalità di selezionare il modello migliore sulla base dell'adattamento ai dati e del comportamento in fase predittiva.

Utile per verificare l'adattamento del modello ai dati è il test Ljung Box applicato ai residui.

Tale test presenta il seguente sistema di ipotesi:

$$\begin{cases} H0: \text{Residui incorrelati} \\ H1: \text{Residui correlati} \end{cases}$$

L'analisi consiste quindi nel verificare se la funzione di autocorrelazione stimata sui residui è significativamente diversa da quella di un *white noise*.⁷

Il modello sarà buono solo nel caso in cui il test accettasse l'ipotesi nulla.

E' necessario applicare quanto appena esposto nei tre modelli.

⁶ SARIMA((p,d,q),(P,D,Q)) dove p,d, q indicano rispettivamente i numero di : parametri autoregressivi, differenze semplici e parametri a media mobile ; P, D, Q indicano rispettivamente il numero di: parametri autoregressivi stagionali, di differenze stagionali e i parametri a media mobile stagionale

⁷La statistica del test di Ljung Box è:

$$Q(m) = T \cdot (T + 2) \sum_{k=1}^m \frac{\hat{\rho}_k^2(\hat{\epsilon})}{T - k}$$

e sotto l'ipotesi nulla $Q(m) \sim \chi_{m-r}^2$

Modello 1

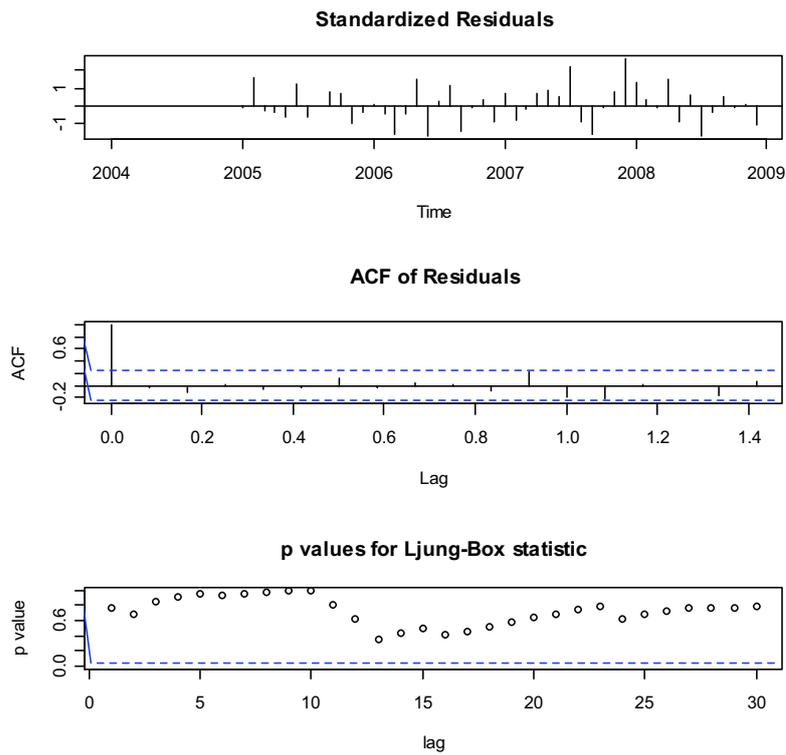


Figura 6: grafico dei residui, ACF dei residui e p-value del test di Ljung-Box dei residui del Modello 1.

I p-value del test Ljung Box relativi al primo modello sono elevati, quindi l'ipotesi nulla viene accettata a tutti i lag.

Il Modello 1 quindi, coglie e spiega molto bene la correlazione utilizzando solamente un parametro autoregressivo.

Modello 2

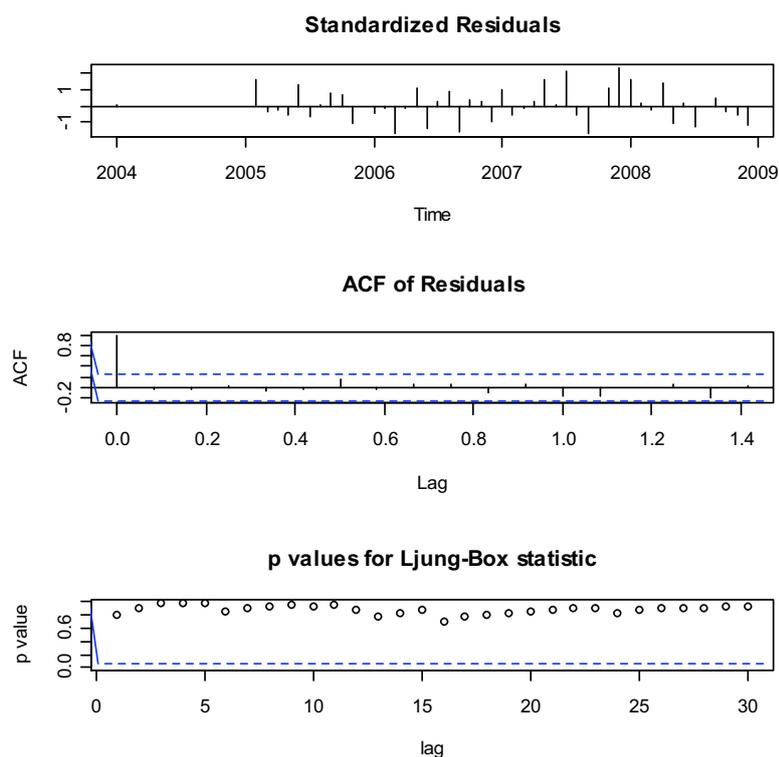
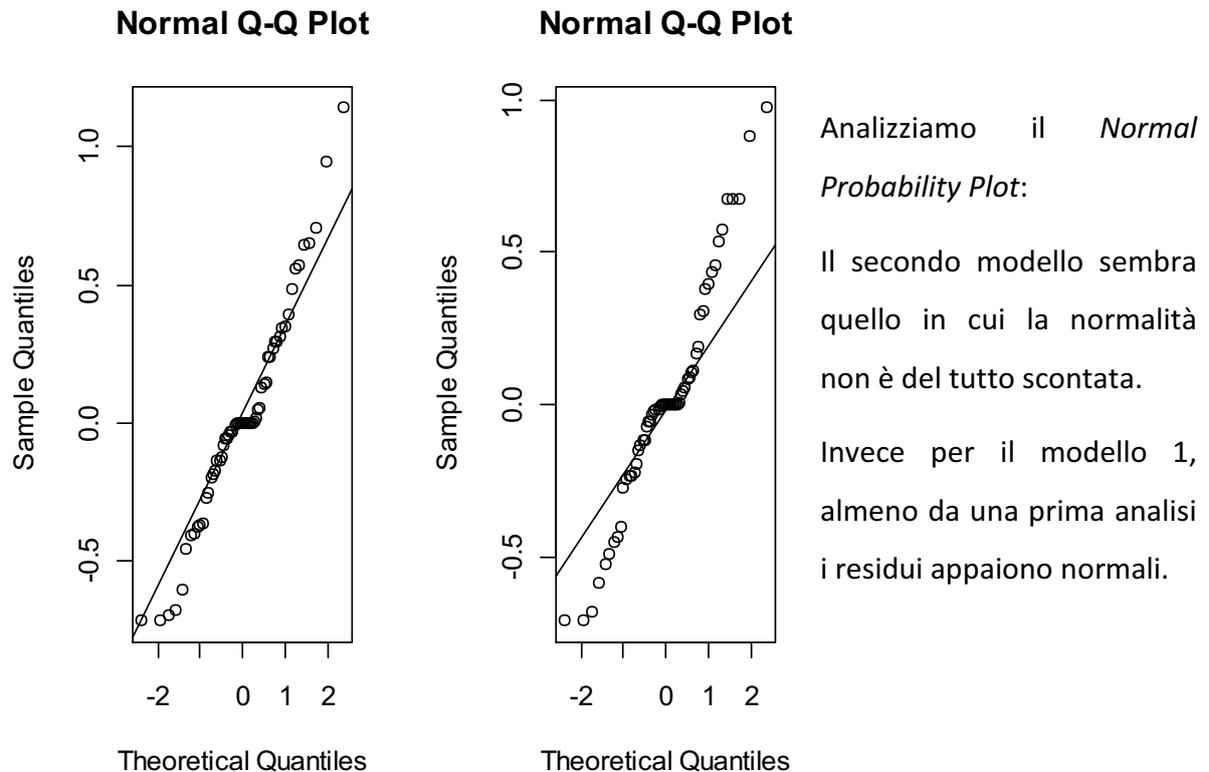


Figura 7: grafico dei residui, ACF dei residui e p-value del test di Ljung-Box dei residui del Modello 2

Anche per il Modello 2 si accetta l'ipotesi nulla, quindi il modello riesce a cogliere in modo soddisfacente la correlazione insita nel processo.

Il test presenta *p-value* più alti se confrontati con quelli del primo modello soprattutto se considerati dopo il decimo ritardo, questo grazie al parametro aggiuntivo che questo modello sfrutta.

Importante è inoltre verificare la normalità dei residui.



Analizziamo il *Normal Probability Plot*:

Il secondo modello sembra quello in cui la normalità non è del tutto scontata.

Invece per il modello 1, almeno da una prima analisi i residui appaiono normali.

Figura 8: Normal Probability Plot del modello 1 e 2.

I dati della serie storica del tasso di disoccupazione sono in tutto 75, ma a causa delle differenze (stagionali e non) abbiamo perso 13 osservazioni, inoltre altre 15 osservazioni sono state tolte per misurare le capacità predittive dei modelli in esame. In conclusione restano 47 osservazioni, sarà quindi opportuno utilizzare un test di normalità efficiente per piccoli campioni per verificare quanto riscontrato dall'analisi del *Normal Probability Plot*.

	Shapiro-Wilk	P-value
Modello 1	0,963	0,07778
Modello 2	0,9539	0,02405

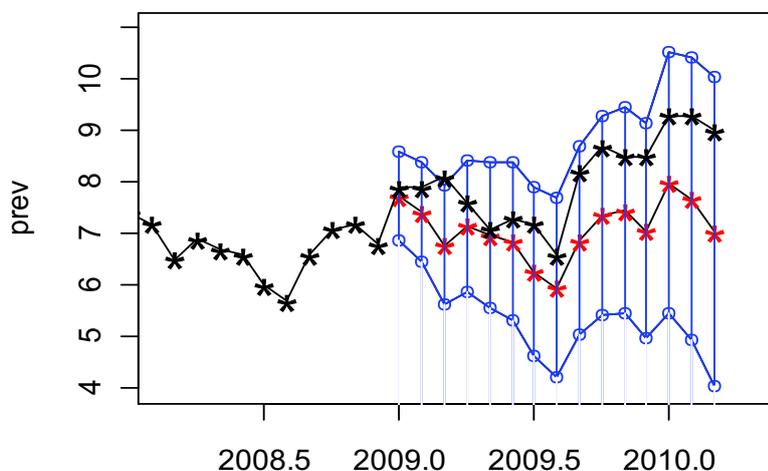
Tabella 2: Test di normalità Shapiro-Wilk.⁸

Solamente i residui del modello 1 si possono considerare normali al livello 5%.

Previsioni SARIMA

E' ora opportuno stimare delle previsioni sulla serie della disoccupazione alla quale sono state tolte le ultime osservazioni in modo da confrontare poi, la serie reale con quella predetta e valutare la bontà del modello.

Previsione Modello 1



All'inizio la previsione segue in modo soddisfacente l'andamento dei valori osservati mentre dal 2009 sembra allontanarsi dalla serie.

Figura 9: Previsione della serie del tasso di disoccupazione attraverso il Modello 1.

⁸ La verifica della normalità avviene confrontando due stimatori alternativi della varianza σ^2 : uno stimatore non parametrico basato sulla combinazione lineare ottimale della statistica d'ordine di una variabile aleatoria normale al numeratore, e la varianza campionaria al denominatore. La statistica test risulta essere:

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Previsione Modello 2

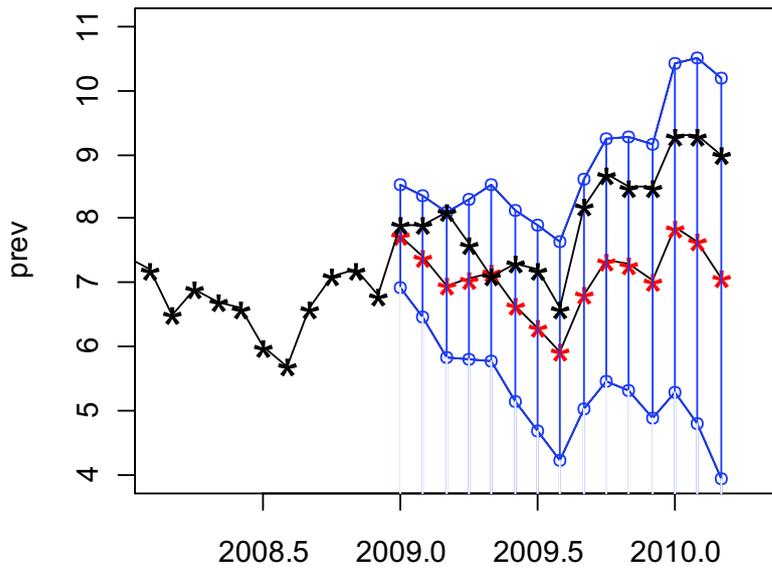


Figura 10: previsione della serie del tasso di disoccupazione attraverso il Modello 2.

La previsione stimata con il Modello 2 presenta un buon adattamento fino a Maggio 2009, per poi discostarsi.

Le previsioni dei due modelli sono simili, il secondo modello però risulta essere meno parsimonioso in quanto

utilizza un parametro in più del modello precedente.

Per selezionare il modello migliore confronto la somma degli errori al quadrato dei due modelli.

Nome Modello	N° Parametri	Somma Errori al Quadrato
Modello 1	1	18,61166
Modello 2	2	19,40187

Figura 11: Somma degli errori al quadrato dei due modelli.

Il primo modello sembra essere quello migliore per stimare la serie del tasso di disoccupazione poiché presenta lo stesso adattamento del Modello 2 utilizzando meno parametri.

Il modello SARIMA(1,1,0)(0,1,0) è quindi il più idoneo fra quelli stimati fin'ora per stimare la serie del tasso di disoccupazione.

Entrambi i modelli tuttavia non considerano la situazione economica nel periodo di analisi, infatti stimerebbero in modo equivalente la disoccupazione in un periodo di recessione economica come in un periodo di crescita: per questo motivo è necessario stimare un modello che tenga in considerazione tali informazioni includendo come variabile esplicativa i volumi di ricerche di lavoro forniti da Google.

3. Variabili esplicative: gli indicatori di Google Insight

Descrizione di Google Insight

E' necessario decidere ora quali variabili includere nel modello idoneo per prevedere la disoccupazione con gli indici di Google.

A tal fine è opportuno in primo luogo capire come Google elabora i dati per poi andare a scegliere le chiavi di ricerca più adatte al nostro scopo e infine per tali *query* sarà opportuno analizzare la popolarità delle stesse.

Google è il motore di ricerca più utilizzato al mondo; in Italia raccoglie oltre il 91% delle ricerche fatte.

Google Index fornisce un'indicazione quotidiana e gratuita su ciò che il mondo sta cercando: dal 2008 è stato giustappunto pubblicato il conteggio numerico delle ricerche effettuate.

I dati forniti da Google rappresentano una svolta epocale che permetterà di analizzare in tempo reale le possibilità che il mercato offre.

La comprensione dei trend di ricerca del resto, può essere utile per venditori, economisti, esperti di marketing e per tutte quelle persone che sono interessate a conoscere di più sul loro mondo e su ciò che è attualmente "il pensiero di tutti".

La forza degli indici di Google tuttavia, non sta solo nell'informazione che i volumi possono fornire ma anche nella costanza e nella velocità con cui vengono aggiornati i dati che sono resi disponibili settimanalmente.

Come Google elabora i dati⁹

Il processo con cui Google elabora i dati al fine di inserirli in un grafico rappresentativo è il seguente:

- **Normalizzazione:** il volume di ricerca assoluto della *query* oggetto di analisi in un determinato periodo viene diviso per il numero totale di ricerche effettuate su Google nel medesimo arco temporale.

$$\text{Indice grezzo di Google} = \frac{\text{Volume query}_{(t)}}{\text{Volume ricerche totali}_{(t)}}$$

Nel caso in cui il volume assoluto di una *query* rimanga costante nel tempo ma le ricerche totali aumentino, l'indice grezzo di Google fornirà un valore più basso, quindi di fatto la chiave di ricerca sarà meno popolare;

- **Rappresentazione sul grafico:** dopo aver calcolato l'indice, Google lo riporta in un grafico con una scala che varia da 0 a 100. A tal fine l'indice grezzo viene quindi diviso per il suo valore massimo e successivamente moltiplicato per 100.

I dati che Google fornisce sono inoltre reperibili per regione geografica, per intervalli temporali e per categoria (identificando in modo specifico un settore o un mercato). Google quindi classifica in modo automatico tutte le ricerche fatte e le assegna alla categoria più appropriata.

Nel caso in cui non venga identificata una categoria durante la ricerca, Google analizzerà solo i dati coerenti con la *query* in tutte le categorie.

Qualora venga selezionata una categoria, Google Statistiche di Ricerca fornisce inoltre una rappresentazione grafica della crescita dei volumi della chiave di ricerca in relazione a quelli della categoria. Questa visualizzazione permette quindi di confrontare l'aumento dell'utilizzo del termine di ricerca scelto rispetto a tutti i termini correlati a quella determinata categoria.

Google offre anche la possibilità di raggruppare più termini di ricerca sommando i volumi e presentandoli come un'unica serie storica; per applicare questa funzione è sufficiente utilizzare l'operatore somma tra i termini di ricerca.

⁹ <http://www.google.com/support/insights/bin/topic.py?hl=it&topic=13975>

Scelta della chiave di ricerca

Molte parole possono essere utilizzate per la ricerca di lavoro in Google, è quindi necessario selezionare le parole chiave più opportune per l'analisi.

Per selezionare le chiavi di ricerca ho cercato un compromesso tra l'attinenza della parola con la ricerca del lavoro, ho poi selezionato le *query* che più si adattavano alla disoccupazione supportate però da discreti volumi.

Nel nostro caso non è necessario specificare una determinata categoria in quanto anche la classe più specifica non riesce ad eliminare la componente di rumore creata dalle ricerche non finalizzate al trovare lavoro.

In base al metodo appena menzionato le parole che ho ritenuto più interessanti sono:

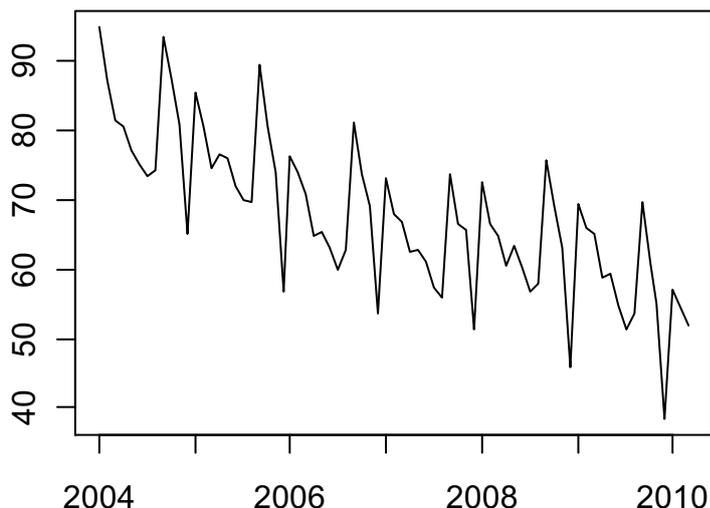
- Lavoro;
- Offerte lavoro;
- *Infojobs*;
- Cerco lavoro.

Query "Lavoro"

La chiave di ricerca "Lavoro" presenta volumi piuttosto elevati.

Tale parola è infatti molto generale, e potrebbe non includere solo i volumi relativi alla ricerca di occupazione ma anche quelli corrispondenti alla "caccia" di informazioni più generali sul lavoro; potrebbero includere, ad esempio, anche le ricerche effettuate al fine di reperire informazioni sulle norme che regolano i contratti di lavoro o semplicemente indagini che riguardano la stessa disoccupazione.

Tenendo presente quanto osservato, è necessario analizzare la serie storica dei volumi di ricerca per la parola chiave oggetto di analisi.



La serie presenta picchi stagionali e sembra avere un trend decrescente.

Figura 12: serie dei volumi per query "Lavoro".

Nella figura 12 viene proiettato il trend stimato per la serie storica delle ricerche corrispondenti alla parola lavoro. Si evince facilmente che l'uso di tale parola come chiave di ricerca diminuisce dal 2004 al 2010; solo nell' anno 2008 i volumi sembrano essersi stabilizzati, per poi ricominciare a scendere.

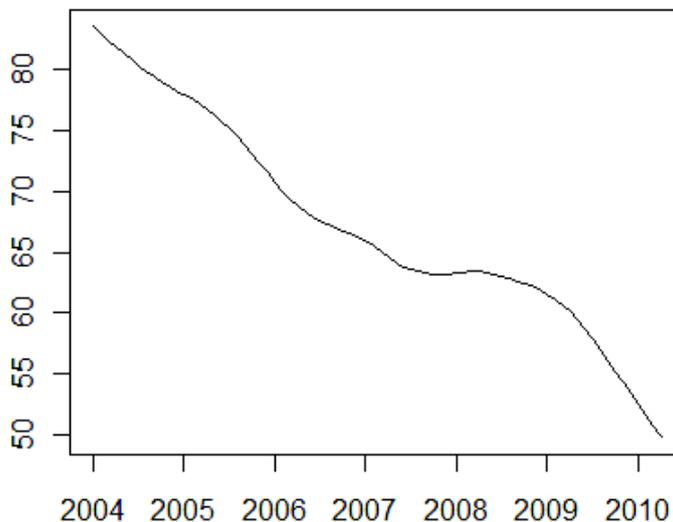


Figura 13: trend dei volumi relativi alla query "Lavoro".

Il trend della serie oggetto di analisi non segue l'andamento della disoccupazione e ciò potrebbe essere dovuto al fatto che tale parola non è usata per la ricerca di lavoro su Google.

Per questo motivo non è opportuno utilizzare tale query fra le esplicative del modello.

Query "Cerco lavoro"

La chiave di ricerca "Cerco lavoro" è abbastanza specifica in quanto si può presumere che i relativi volumi di ricerca facciano riferimento alle sole persone che attraverso Google vogliono ricercare un'occupazione.

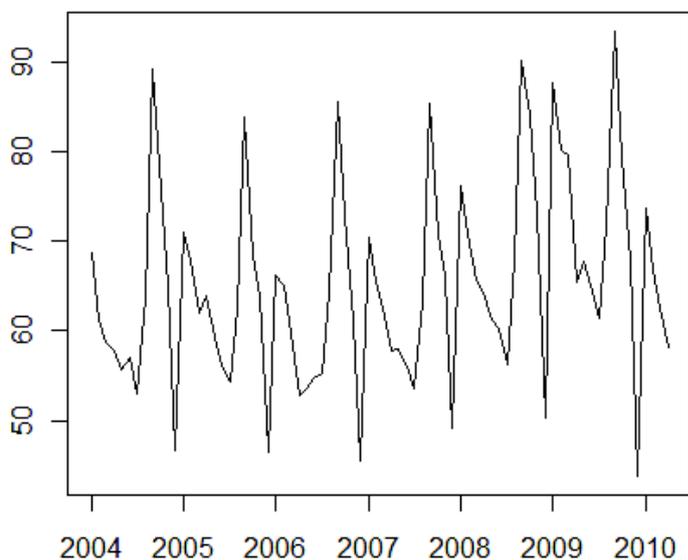


Figura 14: volumi relativi alla query "Cerco lavoro".

Questo grafico rappresenta la serie dei volumi di ricerca per la chiave "Cerco lavoro" e sembra avere una forte componente stagionale che tuttavia risulta più intensa dal 2008.

I picchi dei volumi si riferiscono al mese di Settembre – Ottobre e sono più evidenti nell'ultima parte della serie.

Nell'anno 2007 si sono intensificate anche le ricerche con l'utilizzo della

query in esame, il massimo dei volumi registrati viene raggiunto nell'anno 2009: ciò potrebbe dare una indicazione su un eventuale legame con la disoccupazione che in Italia sembra aumentata proprio in tale anno.

L'analisi del trend della serie storica conferma le osservazioni fatte sulla serie di partenza.

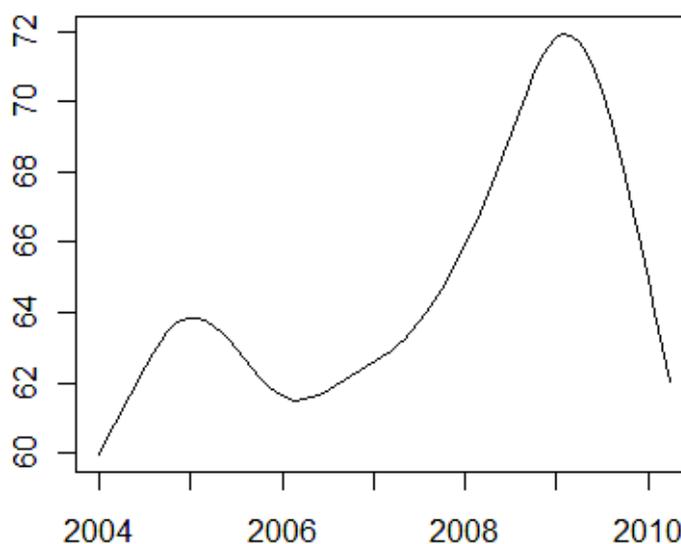


Figura 15: trend relativo alla query "Cerco lavoro".

La figura 15 evidenzia l'aumentare dei volumi nel riferimento temporale 2004-2005 e nel periodo che intercorre tra la metà del 2007 e la seconda parte del 2009.

A differenza dei volumi della parola chiave precedentemente analizzata, quelli relativi alla query "Cerco lavoro" aumentano proprio nei periodi dove ci sarebbe aspettato un incremento della ricerca di lavoro.

Interessante potrebbe quindi essere lo studio della disoccupazione in relazione a tale variabile.

Query "Offerte Lavoro"

Anche questa serie possiamo immaginare includa solo i volumi relativi alle persone che

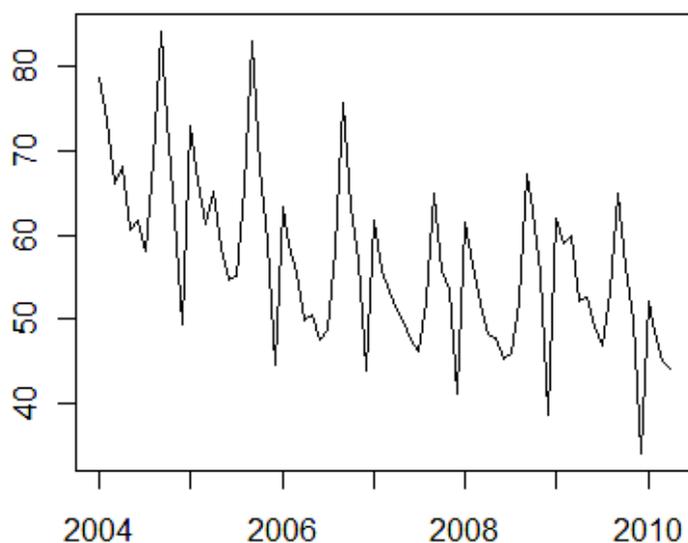


Figura 16: volumi relativi alla query "Offerte lavoro".

cercano impiego.

I volumi delle ricerche riferiti alla parola in esame si comportano in modo simile a quelli relativi alla query "Lavoro", l'unica differenza riscontrabile da questo primo esame è l'aumento della componente stagionale nel periodo compreso tra il 2008 e la seconda metà 2009.

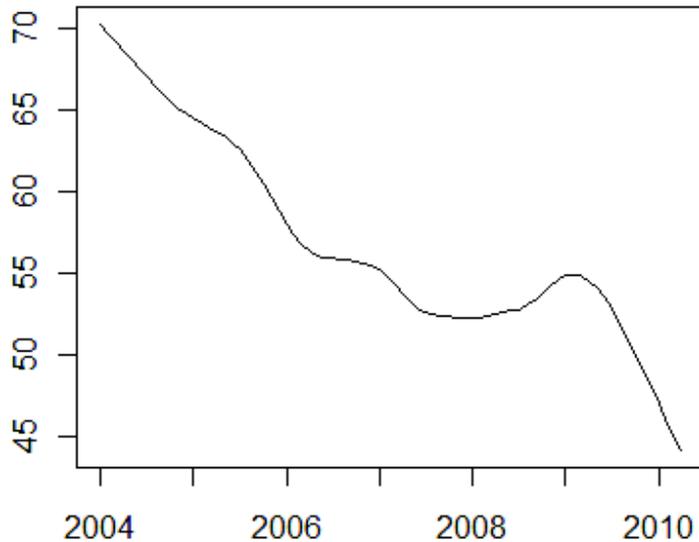


Figura 17: trend per la query "Offerte lavoro".

Effettivamente anche il trend presenta somiglianze con quello della serie dei volumi per la parola chiave "Lavoro".

Queste analisi preliminari ci mostrano che dal 2008 a metà del 2009 c'è stato un effettivo aumento dell'interesse degli italiani verso la ricerca del lavoro utilizzando Google.

Il trend generale tuttavia, risulta non coerente con quella della serie storica della disoccupazione: per questo motivo non è opportuno includere neppure tale variabile nella stima del modello.

Query "Infojobs"

Infojobs è il più importante sito che interfaccia chi offre con chi cerca lavoro.

I volumi relativi a questa chiave di ricerca comprendono sicuramente le persone che

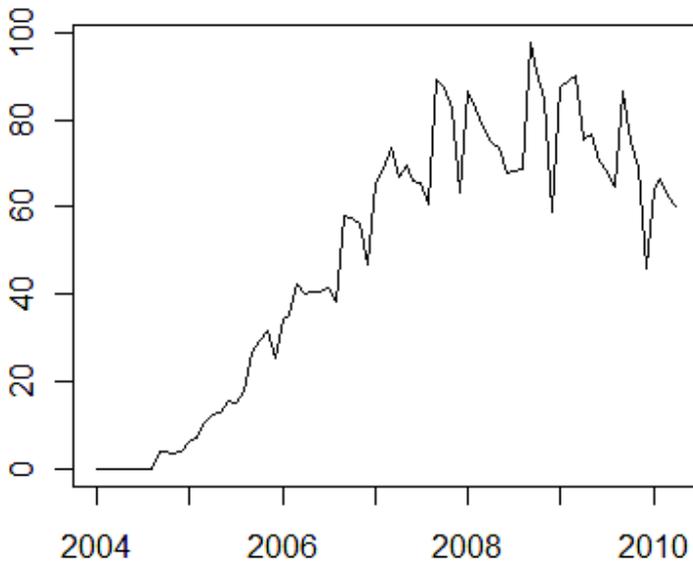


Figura 18: Volumi relativi alla query "Infojobs".

cercano lavoro ma sono ovviamente inferiori a quelli delle precedenti *query* in quanto includono solo le persone che conoscono già il sito trascurando ovviamente quelle che utilizzano direttamente l'url, evitando in questo modo Google.

Dal grafico della serie si può dedurre che nel 2004 la parola "infojobs" non era utilizzata come chiave di ricerca, probabilmente a causa della poca popolarità del sito. Nell'anno 2006 tale *query* realizza un rapido aumento dei volumi di ricerca che corrisponde probabilmente con la diffusione del sito come valido metodo per la ricerca di impiego. Già nell'anno 2007 la serie presenta caratteri stagionali. Il volume di ricerca sembra raggiungere il suo massimo, come le altre *query*, tra il 2008 e fine 2009.

Il grafico della componente di trend sembra confermare le precedenti analisi.

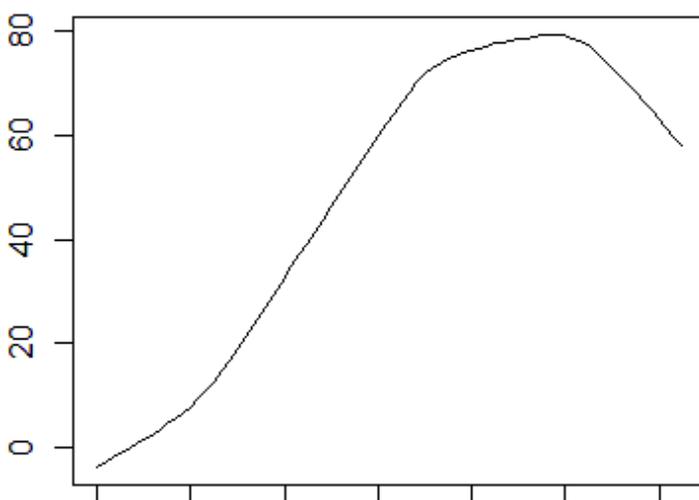


Figura 19: Trend relativo alla query "Infojobs".

La *query* in esame comunque non presenta i caratteri necessari per la stima del modello come unica variabile esplicativa, l'indice non appare

infatti ancora del tutto maturo per svolgere le analisi opportune: troppo pochi i dati che possono essere utilizzati e i volumi nonostante il rapido aumento figurano ancora

modesti. Il processo non presenta una stagionalità ben definita e quindi l'informazione ricavabile risulta limitata.

La serie per di più è influenzata dall'efficienza del sito infatti qualora risultasse un ottimo strumento nella ricerca d'impiego naturale sarebbe il rapido aumento dei volumi, mentre in caso contrario il trend si aggiornerebbe al ribasso, e tutto ciò a prescindere dal tasso di disoccupazione presente in Italia.

Ritengo quindi che questa *query* non possa essere utilizzata singolarmente per la stima del tasso di disoccupazione, potrebbe invece assumere significato se utilizzata congiuntamente a volumi di altre *query*.

Popolarità delle *query*

E' opportuno valutare infine come varia l'interesse di una specifica *query* rispetto alla propria categoria, tralasciando i volumi di ricerca.

Le uniche chiavi tra quelle esaminate, che presentano risultati interessanti sono quelli relative a "Infojobs" e "Cerco lavoro".

Per tali *query* infatti si è registrato un aumento della popolarità nello stesso periodo in cui la disoccupazione è aumentata.

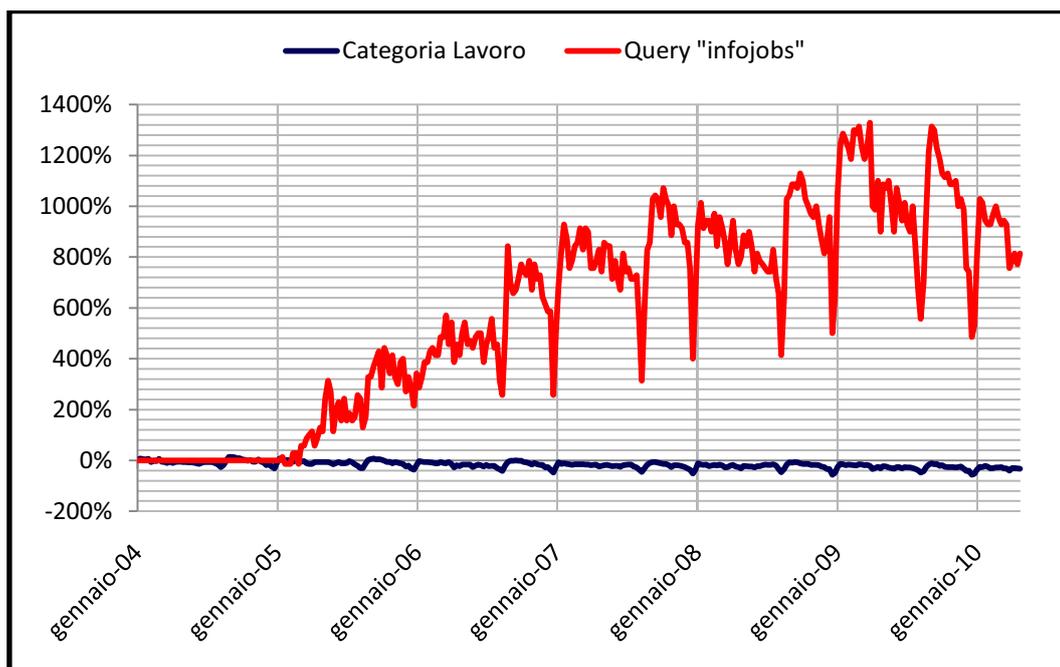


Figura 20: crescita di "Infojobs" in relazione alla categoria "Lavoro".

La gran parte dell'aumento dell'interesse nella *query* "infojobs" è da imputare alla popolarità del sito, forse dalla seconda metà della serie la crescita potrebbe essere supportata dal reale interesse verso la ricerca di lavoro con Google.

Comunque la chiave in esame si distingue dall'intera categoria presentando un aumento considerevole della popolarità, quindi anche se tuttora non è adatta ad essere inserita in un modello si può pensare che questa lo diventi in futuro.

La *query* "cerco lavoro" sembra invece, come osservato nelle analisi precedenti, seguire in modo soddisfacente la disoccupazione.

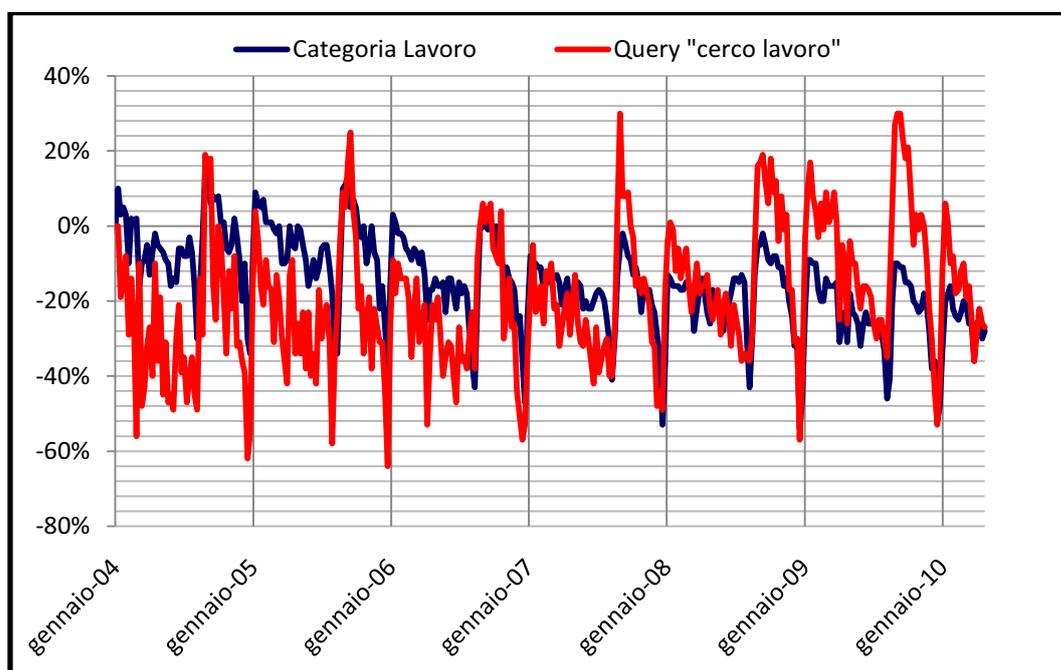


Figura 21: crescita di "Cerco Lavoro" in relazione alla categoria Lavoro.

Si può notare infatti come i picchi di popolarità si vadano ad affievolire tra il 2006 e 2007 per poi aumentare.

E' ora opportuno confrontare l'aumento della popolarità dell'insieme delle due chiavi di ricerca in relazione alla categoria.

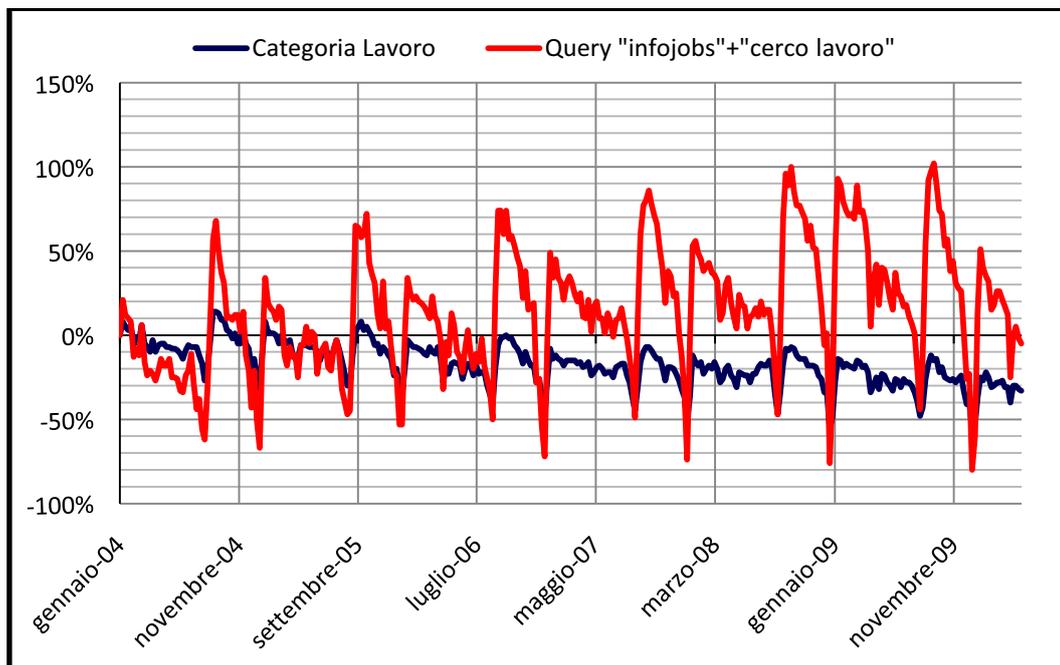


Figura 22: Crescita di "Cerca Lavoro" in relazione alla categoria Lavoro.

L'accorpamento delle *query* presentano un trend nettamente diverso da quello dell'intera categoria, infatti nell'arco temporale che va dal 2008 alla fine del 2009 si sono registrati aumenti di oltre il 100%, mentre l'interesse verso la categoria Lavoro è decrescente.

La bontà della scelta delle *query* viene confermata da queste ultime analisi: le chiavi di ricerca selezionate sono le uniche con discreti volumi a presentare un trend coerente con quello del tasso di disoccupazione.

Variabili esplicative: gli indicatori di Google Insight

4. Modelli di previsione

Adattamento dei dati

In questa sezione verranno comparate la serie della disoccupazione con l'indice di Google "cerco lavoro".

È opportuno riportare in un grafico le due serie storiche sovrapposte:

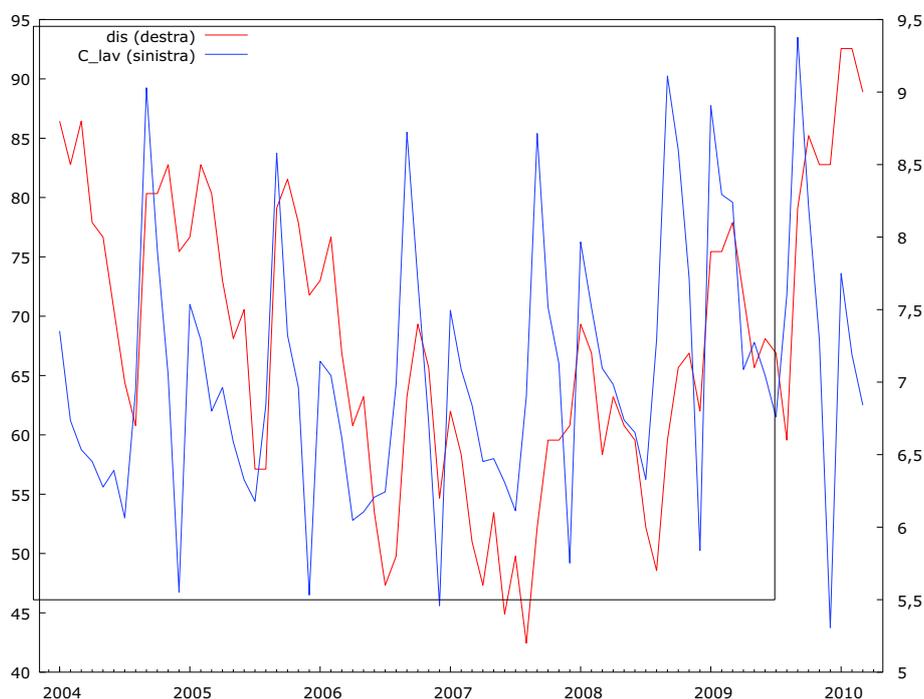


Figura 23: serie del tasso di disoccupazione e dei volumi della query "cerco lavoro".

Le serie "cerco lavoro" e quella del tasso di disoccupazione presentano caratteri simili, tuttavia i picchi stagionali sembrano essere più accentuati nella serie fornita da Google statistiche di ricerca.

Il trend della serie della disoccupazione (curva rossa) invece risulta maggiormente marcato rispetto a quello della query (curva blu), malgrado ciò anche il corso della chiave "Cerca lavoro" sembra abbassarsi leggermente nell'arco temporale 2005-2006, per poi cambiare tendenza nell'anno 2007.

Una cosa da notare è il valore molto basso di ricerche che caratterizza la serie del Google Index (GI) durante il mese di dicembre, soprattutto nel 2009.

Questo può essere causato da molti fattori ma ritengo si possa attribuire alle vacanze natalizie: in questo periodo, infatti, è probabile che molte persone siano in vacanza o che non si dedichino alla ricerca di lavoro facendo così decrescere i volumi della *query*.

Per limitare questo fenomeno ho pensato di sostituire il valore assunto dall'indice nel mese di dicembre (calcolato come media delle quattro settimane) con quello della prima settimana dello stesso mese. Ho ritenuto adeguata questa soluzione perché l'effetto "vacanze natalizie" è molto più marcato nella seconda metà del mese cioè in concomitanza con le festività.

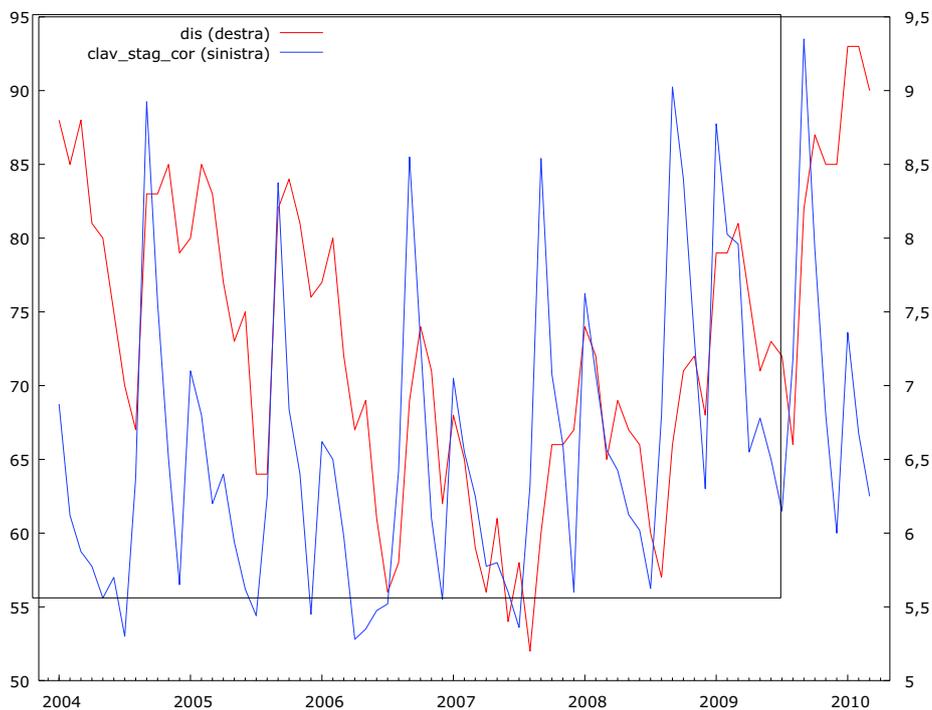


Figura 24: serie del tasso di disoccupazione e dei volumi della query "cerco lavoro" corretta per l'effetto "vacanze natalizie".

La serie del GI pur mantenendo la componente stagionale non presenta più i forti picchi negativi che caratterizzavano il mese di dicembre. Oltretutto come si può notare dal grafico, il GI sembra adattarsi meglio alla disoccupazione: in particolare si nota che l'inversione di tendenza dell'indice di Google è in concomitanza con il cambiamento del trend della disoccupazione.

Modelli stimati

In questa sezione verranno presentati diversi modelli: solo uno non sfrutta l'informazione proveniente da Google, mentre tutti gli altri includono anche se in diverse forme l'indicatore riferito alle *query* selezionate.

Tutti i modelli successivi saranno stimati utilizzando il campione ridotto di osservazioni adoperato per il modello SARIMA.

Come visto dai grafici proposti in precedenza la serie della disoccupazione è un processo non stazionario, presenta infatti trend e stagionalità; ho deciso tuttavia di non utilizzare filtri per rendere stazionario il processo ma di utilizzare dei modelli lineari che sostanzialmente usufruiscono implicitamente del trend e della stagionalità della serie osservata. Questo è reso possibile dall'introduzione nel modello di una o più variabili riferite alla disoccupazione ritardata di uno o più lag.

Il concetto su cui si basano i modelli che verranno presentati è il fatto che per stimare la disoccupazione al tempo t è intuitivo usare il suo livello al tempo $t-1$ ed un parametro riferito al valore della serie undici mesi prima per catturare la componente stagionale¹⁰ e la correlazione presente a quel ritardo.

Inoltre è presente la variabile relativa al GI che come abbiamo visto in precedenza segue abbastanza bene la serie della disoccupazione sia per quanto riguarda il trend sia per la componente stagionale. Si suppone quindi che questa variabile fornisca un'informazione aggiuntiva del livello del processo. Questo salto dipenderà ovviamente dal livello del GI.

Vista la natura del Google Index ossia una variabile facile da reperire e costantemente aggiornata è ragionevole utilizzarla al tempo t in modo tale da usufruire delle informazioni aggiornate.

Un'altra strada percorribile visto il buon adattamento della componente stagionale del GI con quella della disoccupazione, è quella di utilizzare al posto della stagionalità ricavata dalla serie della disoccupazione, quella ricavata dal GI.

Non verranno presentati ARMAX, ovvero modelli che sfruttano oltre alla correlazione presente nel processo anche l'informazione fornita da una variabile esogena. Questi

¹⁰ Questa soluzione è equivalente a stimare un modello autoregressivo con lag 1 e 11.

modelli per essere adatti alla previsione (modello proprio) devono considerare l'influenza dell'esogena almeno con un istante di ritardo (Modelli empirici lineari, appendice A4, pag 3-4).

Come ribadito in precedenza il fine di questa relazione non è studiare la causalità tra le variabili in gioco, ma quello di utilizzare l'informazione, più aggiornata possibile riferita alle ricerche di lavoro su Google; inoltre la scelta di non utilizzare ARMAX è rafforzata dal fatto che il dato della disoccupazione mensile è fornito dall'Istat costantemente in ritardo mentre le stime di Google Index sono disponibili settimanalmente.

Le stime mensili fornite dall'Istat sono diffuse a distanza di circa 30 giorni dalla fine di ciascun mese di riferimento causando disagi a coloro che invece hanno bisogno di una stima tempestiva.

Ne è un esempio il fatto che il tasso di disoccupazione fornito dall'Istat per il mese di Marzo è stato pubblicato nei primi giorni di Maggio, con Google Index invece sarebbe stato possibile avere tale dato già alla data 1 Aprile anticipando di ben un mese il dato Istat.

I modelli plausibili saranno di seguito commentati, ne verrà discusso l'adattamento e la bontà in fase di previsione.

Per ogni modello verranno messi in evidenza le seguenti caratteristiche

- stime e significatività dei parametri;
- dati riassuntivi del modello utili per valutarne l'adattamento;
- analisi dei residui che consistono nello studio della normalità, omoschedasticità e indipendenza;

Modello 1

Il primo modello è caratterizzato dall'equazione:

$$Y_t = Costante + Disoccupazione_{t-1} + Disoccupazione_{t-11} + CercoLavoro_t + \varepsilon_t$$

Il modello così definito stima la disoccupazione basandosi sul livello di questa ad uno e undici mesi prima .

La disoccupazione con ritardo 11 oltre che approssimare abbastanza bene la componente stagionale riesce a cogliere anche la correlazione presente a questo ritardo.

L'ultima variabile inserita nel modello fa riferimento alla *query* "cerco lavoro" che rappresenta l'indice di Google. L' influenza della variabile esplicativa "cerco lavoro" sulla risposta è immediata cioè la disoccupazione viene stimata utilizzando il valore del GI allo stesso istante t.

Stime e significatività dei parametri

	Coefficiente	Errore std.	Rapporto t	P-value
Costante	-1.86973	0.649869	-2.877	0.0061 ***
Cerco_lavoro	0,0358885	0,00630206	5,695	8,85e-07 ***
Dis.t-1	0,750801	0,0733450	10,24	2,49e-013 ***
Dis.t-11	0,171732	0,0637980	2,692	0,0099 ***

Tabella 3: stima delle variabili riferite al Modello 1.

La variabile relativa al GI (versione corretta per la stagionalità), come anche le altre, sono significative significativa ad ogni livello di alpha usuale .

Dati riassuntivi

R-quadro	0,820730	R-quadro corretto	0,808779
Test F(2,41)	68,67281	P-value (F)	7,92e-17
Log-verosimiglianza	-18,77854	Criterio di Akaike	45,55707
Criterio di Schwarz	53,12435	Hannan-Quinn	48,42809

Tabella 4: dati riassuntivi Modello 1.

L'R quadro e l'R quadro corretto manifestano un buon adattamento del modello ai dati, infatti questo interpreta correttamente circa il 82% della variabilità della risposta.

Il test F, che confronta il modello con la sola intercetta e il modello completo, conferma la significatività dei parametri.

Analisi residui

Normalità dei residui

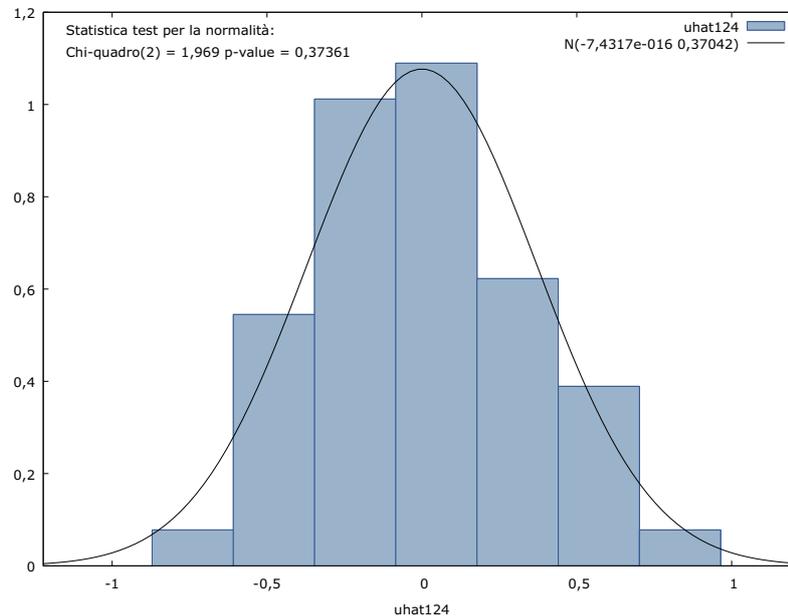


Figura 25: istogramma e test di Doornik-Hansen sui residui del Modello 1.

La normalità sembra essere confermata sia a livello grafico che dal p-value fornito dal test di Doornik-Hansen.

Test Shapiro-Wilk	P-value
0,978491	0,503726

Anche il test Shapiro-Wilk, efficiente per piccoli campioni, accetta la normalità:

Omoschedasticità dei residui

Per verificare l'omoschedasticità dei residui è possibile effettuare il test di White che presenta il seguente sistema d'ipotesi:

$$\begin{cases} H0: \text{Omoschedaticità} \\ H1: \text{Eteroschedaticità} \end{cases}$$

Test White	P-value
11,107488	0,268413

Viene accettata l'omoschedasticità dei residui.

Indipendenza dei residui

In questa sezione verrà valutata l'indipendenza dei residui utilizzando l'ACF e PACF.

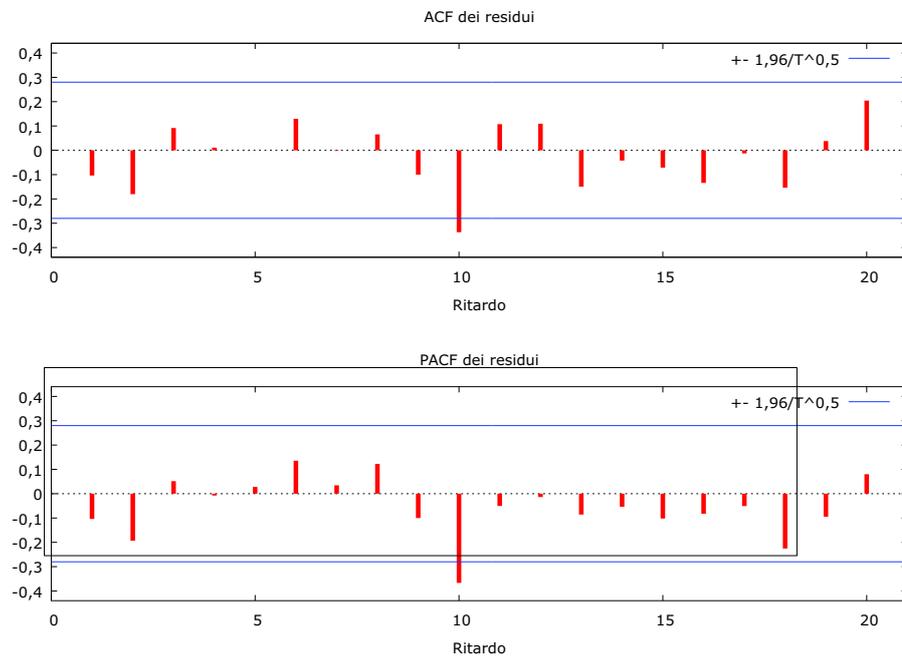


Figura 26: ACF e PACF dei residui del Modello 1.

Non si rilevano correlazioni particolarmente evidenti tranne a lag 10 dove sia nell' ACF che nel PACF viene segnalata la presenza di una correlazione anomala.

In generale i residui si possono ritenere incorrelati.

Modello 2

Il secondo modello è stato ottenuto attraverso la procedura *backward* ossia eliminando una variabile alla volta partendo da un modello completo che include le seguenti esplicative: una costante, una variabile relativa all'indice di Google e dodici variabili relative ai diversi lag della serie della disoccupazione.

Il modello ottenuto è il seguente:

$$Y_t = Costante + Dis_{t-1} + Dis_{t-10} + Dis_{t-11} + CercoLavoro_t + \varepsilon_t$$

Stime e significatività dei parametri

	Coefficiente	Errore std.	Rapporto t	P-value
Costante	-1.44836	0.645793	-2.243	0.0300 **
Cerco_lavoro	0,0317603	0,00567412	5,597	1,31e-06 ***
Dis._{t-1}	0,750517	0,0699748	10,73	7,37e-014 ***
Dis._{t-10}	-0,210543	0,0902755	-2,332	0,0243 **
Dis._{t-11}	0,340105	0,0944279	3,602	0,0008 ***

Tabella 5: stima delle variabili riferite al Modello 2.

Dati riassuntivi

R-quadro	0,840453	R-quadro corretto	0,825949
Test F(2,41)	57,94542	P-value (F)	5,67e-17
Log-verosimiglianza	-15,92291	Criterio di Akaike	41,84583
Criterio di Schwarz	51,30493	Hannan-Quinn	45,43460

Tabella 6: dati riassuntivi Modello 2.

Tutte le variabili sono statisticamente significative inoltre, i dati riassuntivi del modello sono migliorati rispetto al precedente modello; in particolare l' R^2 corretto è aumentato di oltre due punti.

Analisi residui

Normalità dei residui

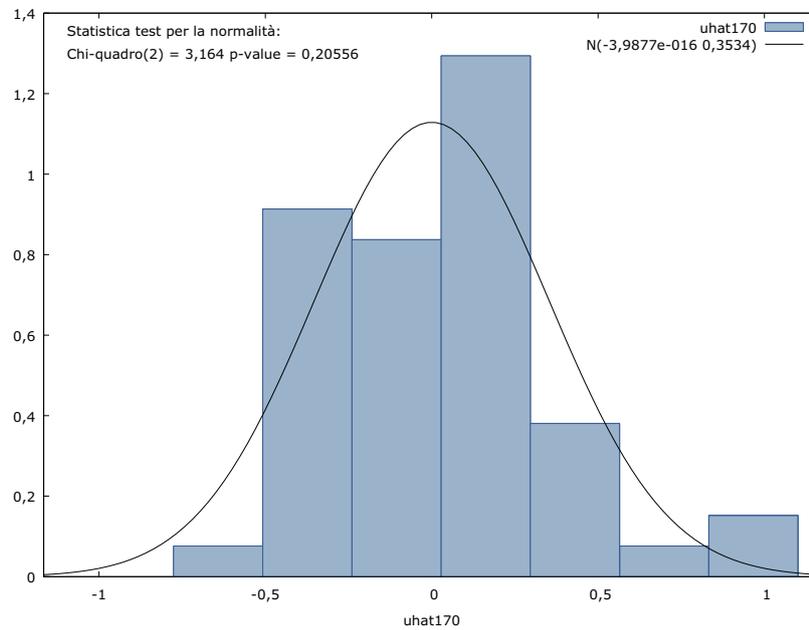


Figura 27: istogramma e test di Doornik-Hansen sui residui del Modello 2.

Test Shapiro-Wilk	P-value
0,968037	0,202035

Viene accettata l'ipotesi nulla di normalità.

Omoschedasticità dei residui

Test White	P-value
21,004941	0,101506

Viene accettata omoschedasticità per ogni livello di alpha usuale.

Indipendenza dei residui

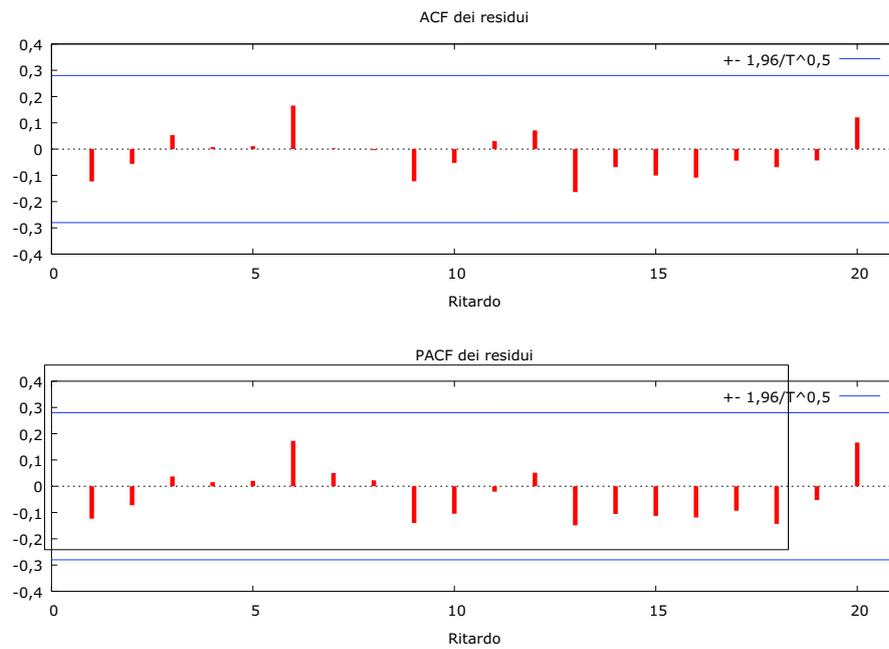


Figura 28: ACF e PACF dei residui del Modello 2

L'indipendenza è confermata; rispetto al modello precedente è stata anche eliminata la correlazione significativa del decimo ritardo.

Il secondo modello presenta un buon adattamento ai dati. I residui sono incorellati, omoschedastici e indipendenti.

Modello 3 (senza Google Index)

Il terzo modello proposto non include la variabile GI, può quindi essere utilizzato per un confronto con l'obiettivo di verificare l'importanza dei volumi di ricerca forniti da Google per la stima del tasso di disoccupazione.

L'equazione del modello è la seguente:

$$Y_t = Costante + Disoccupazione_{t-1} + Disoccupazione_{t-11} + \varepsilon_t$$

Stime e significatività dei parametri

	Coefficiente	Errore std.	Rapporto t	P-value
Costante	0,981915	0,607389	1,617	0,1130
Dis.t-1	0,509302	0,106559	4,780	1,91e-05 ***
Dis.t-11	0,329476	0,0906795	3,633	0,0007 ***

Tabella 7: stima delle variabili riferite al Modello 3.

Dati riassuntivi

R-quadro	0,67697	R-quadro corretto	0,662620
Test F(2,41)	47,15445	P-value (F)	9,07e-12
Log-verosimiglianza	-32,22951	Criterio di Akaike	70,45903
Criterio di Schwarz	76,07263	Hannan-Quinn	72,58041

Tabella 8: dati riassuntivi Modello 3.

Le variabili relative alla disoccupazione sono significative.

Il modello riesce a spiegare il circa il 67% della varianza totale.

Analisi dei residui

Normalità dei residui

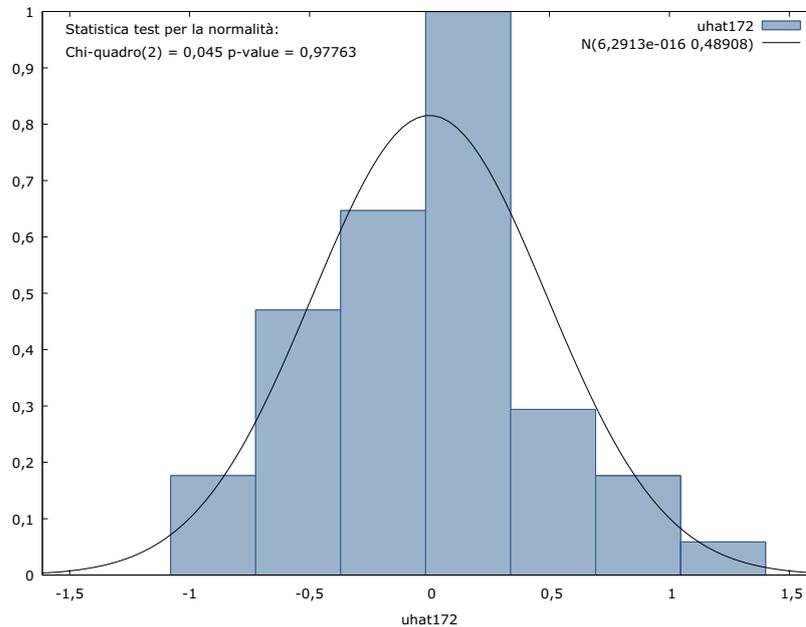


Figura 29: Istogramma e test di Doornik-Hansen sui residui del Modello 3

Test Shapiro-Wilk	P-value
0,976548	0,444246

Viene accettata anche per questo modello l'ipotesi di normalità dei residui.

Omoschedasticità dei residui

Test di White	P-value
5,835142	0,32259

Non sembrano esserci problemi per quanto riguarda l'eteroschedasticità.

Indipendenza dei residui

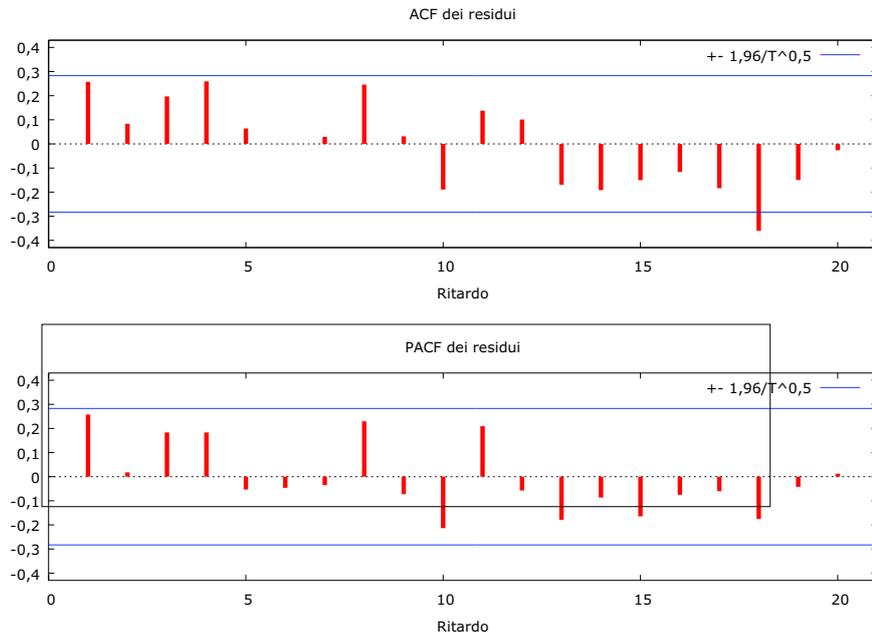


Figura 30: ACF e PACF dei residui del Modello 3.

I residui sembrano indipendenti.

Modello 4

L'equazione del quarto modello è la seguente:

$$Y_t = Costante + Dis_{t-1} + Dis_{t-11} + CLav_Infojobs_t + \varepsilon_t$$

In questo caso la variabile del GI contiene l'informazione fornita da Google relativamente alla ricerca lavoro con le chiavi "Infojobs" e "Cerco lavoro" assimilando in un'unica esplicativa i volumi congiunti delle due *query*.

Ci si aspetta che il modello presenti un migliore adattamento alla parte finale della serie del tasso di disoccupazione: la variabile relativa a GI fino al 2007 presenta un trend non conforme alla serie oggetto di analisi, l'esplicativa è infatti condizionata dai volumi di "infojobs" che assumono rilevanza solamente negli ultimi due anni di osservazione.

Il problema esposto non dovrebbe comunque compromettere la capacità previsiva del modello.

Le altre variabili incluse servono per modellare trend e componente stagionale sulla base della disoccupazione ai lag 1 e 11.

Stime e significatività dei parametri

	Coefficiente	Errore std.	Rapporto t	P-value
Costante	-2.45223	0.899779	-2.725	0.0091 ***
Dis_{t-1}	0,786021	0,0878101	8,951	1,48e-011 ***
Dis_{t-11}	0,350020	0,0677014	5,170	5,21e-06 ***
Clav_Infojobs	0,0254519	0,00598407	4,253	0,0001 ***

Tabella 9: stima delle variabili riferite al Modello 4.

Tutti i parametri del modello sono altamente significativi.

Dati riassuntivi

R-quadro	0,779985	R-quadro corretto	0,765317
Test F(2,41)	53,17716	P-value (F)	7,75e-15
Log-verosimiglianza	-23,79625	Criterio di Akaike	55,59249
Criterio di Schwarz	63,15977	Hannan-Quinn	58,46351

Tabella 10: dati riassuntivi Modello 4.

Il modello presenta un discreto adattamento ai dati cogliendo il 77% della varianza totale.

Analisi dei residui

Normalità dei residui

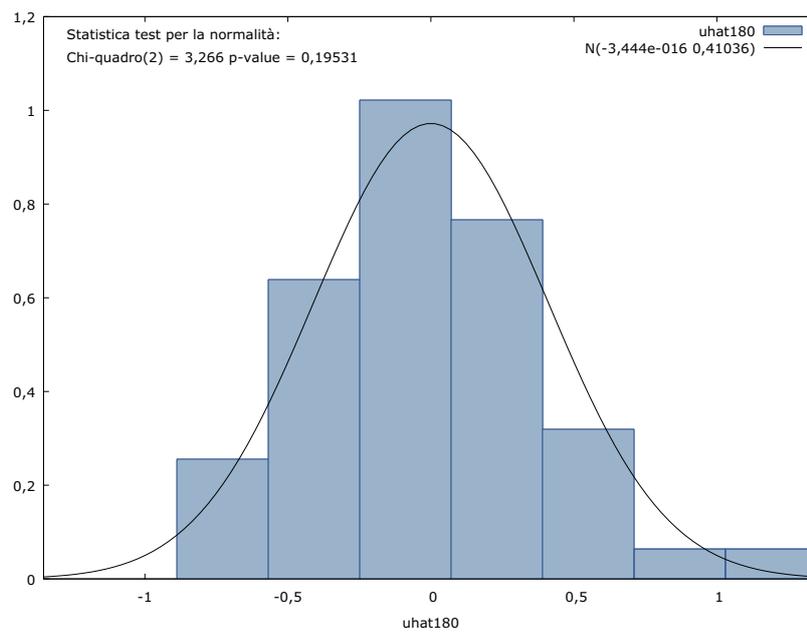


Figura 31: Istogramma e test di Doornik-Hansen sui residui del Modello 4

Test Shapiro-Wilk	P-value
0,971695	0,282489

Viene accettata l'ipotesi di normalità dei residui.

Omoschedasticità dei residui

Test di White	P-value
5,74781	0,764869

Anche l'omoschedasticità è confermata.

Indipendenza dei residui

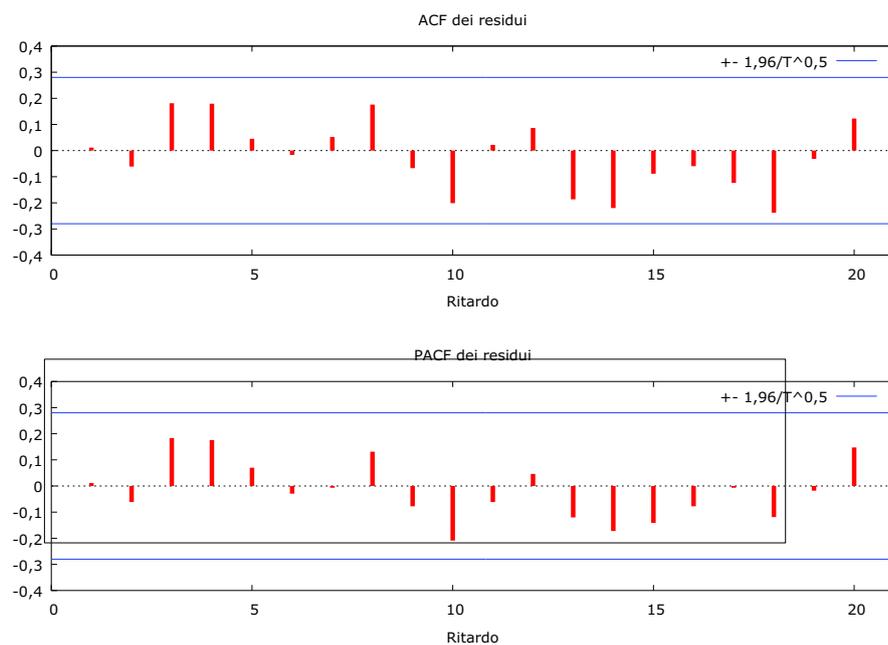


Figura 32: ACF e PACF dei residui del Modello 4.

I residui sembrano indipendenti

In conclusione anche questo modello presenta un buon adattamento ai dati.

I parametri sono infatti significativi e i residui sono incorrelati.

Modello 5

L'ultimo modello è identificato dalla seguente equazione:

$$Y_t = Costante + Disoccupazione_{t-1} + CercoLavoro_{t-12} + CercoLavoro_t + \varepsilon_t$$

Il modello stima la disoccupazione al tempo t utilizzando fra le esplicative la disoccupazione al tempo t-1 e la variabile Google Index ritardata di dodici mesi (che risulta appropriata per modellare la componente stagionale), e utilizza inoltre l'informazione aggiornata fornita dalla variabile riferita al GI al tempo t.

La motivazione alla base della scelta è il fatto che Google Index, come detto in precedenza, presenta una componente stagionale che si adatta molto bene alla serie della disoccupazione. Di fatto con questo modello si sta utilizzando l'informazione proveniente dal GI ritardato di dodici mesi che, combinata con l'informazione attuale fornita dall'indice di Google, permette di cogliere e stimare la componente stagionale aggiornandone i livelli con l'informazione attuale.

E' possibile supporre tuttavia che Google Index abbia la proprietà di anticipare i cambiamenti repentini nella serie della disoccupazione, ciò è deducibile dallo studio della psicologia delle persone, infatti, talvolta queste possono essere portate a ricercare lavoro ancor prima di entrare nel reale stato di disoccupazione facendosi influenzare dal solo timore che ciò si realizzi.

Il modello stimato con la Google Index riferito alla variabile "cerco lavoro" è il seguente:

Stime e significatività dei parametri

	Coefficiente	Errore std.	Rapporto t	P-value
Costante	-1.76771	0.615430	-2.872	0.0174 **
Cerco_lavoro_12	0,0276585	0,0119894	2,307	0,0253 **
Dis._{t-1}	0,823193	0,0613932	13,41	5,11e-018 ***
Cerco_lavoro	0,0184157	0,0107107	1,719	0,0919 *

Tabella 11: Stime delle variabili riferite Modello 5.

Dati riassuntivi

R-quadro	0,814685	R-quadro corretto	0,803340
Test F(2,41)	71,80506	P-value (F)	5,95e-18
Log-verosimiglianza	-21,37058	Criterio di Akaike	50,74117
Criterio di Schwarz	58,62234	Hannan-Quinn	53,77189

Tabella 12: Dati riassuntivi Modello 5.

La variabile di Google Index e la disoccupazione al tempo t-1 sono variabili significative nel modello. Si nota inoltre che il valore del parametro riferito al Google Index è il più elevato se confrontato con gli altri modelli che lo includono.

L' R^2 e l' R^2 corretto manifestano un buon adattamento del modello ai dati, infatti questo interpreta correttamente circa l'80% della variabilità della risposta.

Il test F, che confronta il modello con la sola intercetta e il modello completo, conferma la validità dei parametri.

La componente stagionale di Google Index sembra effettivamente rendere flessibile il modello facendo in modo che segua il più possibile l'evoluzione del processo in esame senza farlo rimanere ancorato ai vecchi livelli di stagionalità.

Analisi dei Residui

Test Normalità

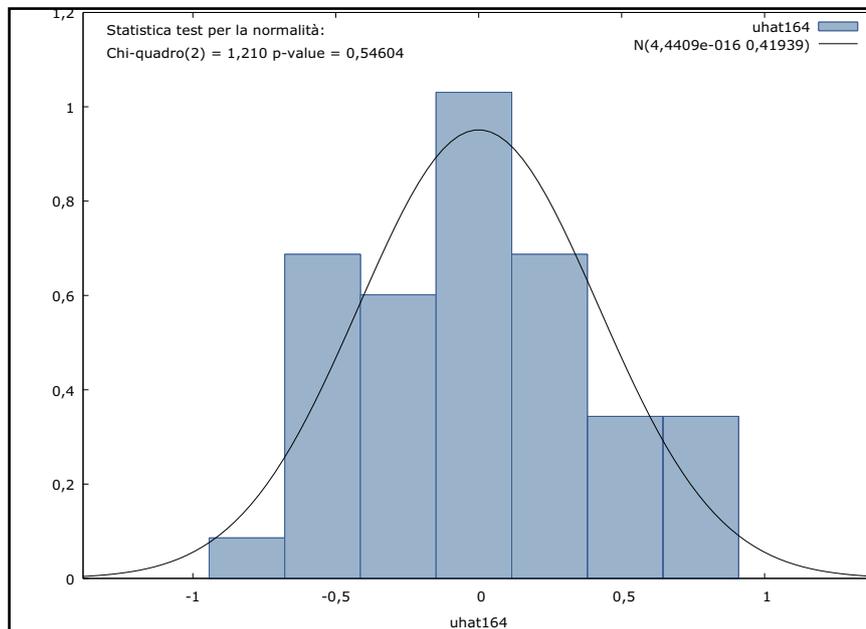


Figura 33: Istogramma e test di Doornik-Hansen sui residui del Modello 5

La normalità dei residui sembra confermata :il test di normalità Doornik-Hansen accetta l'ipotesi nulla con un p-value di 0,5464 e l'istogramma dei residui sembra avere una distribuzione coerente con quella normale.

Test Shapiro-Wilk	P-value
0,973317	0,279515

Anche il test Shapiro-Wilk accetta la normalità.

Omoschedasticità dei residui

Per verificare l'omoschedasticità dei residui è opportuno, come in precedenza, effettuare il test di White:

Test White	P-value
9,828261	0,323404

Il test accetta l'ipotesi nulla si omoschedasticità

Indipendenza dei residui

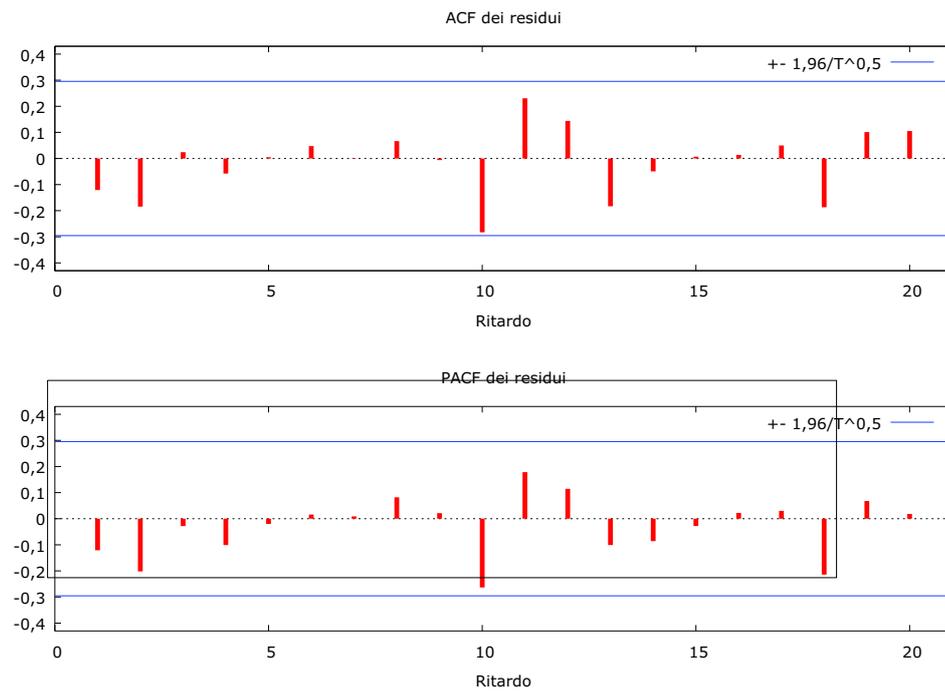


Figura 34: Correlogrammi totale e parziale dei residui del Modello 5.

Dalla figura 33 non si rilevano correlazioni significative per i residui.

5. Analisi delle *performance* previsive

Strumenti per il confronto di modelli di forecasting

In questa sezione verranno proposti degli strumenti usati per l'analisi e il confronto dei vari modelli: il coefficiente di determinazione, il test di Daibold e Mariano e la *Leave-One-Out*.

Coefficiente di Determinazione R^2

R^2 o coefficiente di determinazione misura la bontà dell'adattamento del modello di regressione lineare ai dati osservati.

L' R^2 è definito come:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Dove:

$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ è la devianza spiegata dal modello (Explained Sum of Squares)

$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ è la devianza totale (Total Sum of Squares)

$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ è la devianza residua (Residual Sum of Squares)

y_i sono i dati osservati

\bar{y} è la loro media

\hat{y}_i sono le stime ottenute utilizzando il modello di regressione

L' R^2 varia tra 0 e 1 e misura la parte della variabilità delle y_i che il modello lineare riesce a spiegare. L'indice R^2 non misura però, se effettivamente sussista una relazione tra la risposta e i regressori, ma solamente quanto il modello lineare riesce ad approssimare la realtà dei dati osservati.

Un aumento dei regressori al modello porta ad un incremento dell' R^2 , cioè ad un migliore adattamento ai dati; tuttavia questo non significa che il modello sia più buono, non è detto infatti che migliori la sua capacità predittiva su dati futuri.

L'indice R^2 non è uno strumento di confronto sempre valido, in quanto non è possibile concludere che un modello che presenta l'indice di determinazione più alto sia il migliore.

Per ovviare questo problema è stata presentata una forma corretta dell' indice R^2 che inserisce una penalizzazione crescente man mano che il numero dei regressori nel modello aumenta: il valore dell' R^2 corretto è legato quindi alla significatività delle variabili aggiuntive.

L' R^2 corretto si presenta come:

$$\bar{R}^2 = 1 - \frac{N - 1}{N - k - 1} (1 - R^2)$$

Test Diebold Mariano

Il test di Diebold Mariano ci permette di confrontare le performance di previsione di due modelli.

Definiamo:

- $\{y_t\}$ la serie su cui effettuare le previsioni
- $\hat{y}_{t+h|t}^1$ previsioni utilizzando il Modello 1
- $\hat{y}_{t+h|t}^2$ previsioni utilizzando il Modello 2

Gli errori di previsione dei due modelli:

$$\varepsilon_{t+h|t}^1 = y_{t+h} - \hat{y}_{t+h|t}^1$$

$$\varepsilon_{t+h|t}^2 = y_{t+h} - \hat{y}_{t+h|t}^2$$

Si ipotizza quindi di conoscere il processo fino al tempo T con $t = t_0, \dots, T$ e poi calcolare le previsioni per h passi successivi. Per misurare l'accuratezza delle previsioni si possono utilizzare diverse funzioni di perdita che confrontano le previsioni con i dati realmente osservati:

$$L(y_{t+h}, y_{t+h|t}^i) = L(\varepsilon_{t+h|t}^i) \text{ con } i = 1, 2$$

Le più popolari *Loss Function* sono:

- *Squared error loss*: $L(\varepsilon_{t+h|t}^i) = (\varepsilon_{t+h|t}^i)^2$ ossia i quadrati degli errori;

- *Absolute error loss*: $L(\varepsilon_{t+h|t}^i) = |\varepsilon_{t+h|t}^i|$ ossia il valore assoluto dell'errore commesso.

Nei test che verranno effettuati per valutare i modelli che meglio prevedono la disoccupazione, verrà utilizzata la prima funzione di perdita ossia il quadrato degli errori.

Per determinare quale dei due modelli sia il migliore si può costruire un test con il seguente sistema d'ipotesi:

$$\begin{cases} H0: E[L(\varepsilon_{t+h|t}^1)] - E[L(\varepsilon_{t+h|t}^2)] = 0 \\ H1: E[L(\varepsilon_{t+h|t}^1)] - E[L(\varepsilon_{t+h|t}^2)] \neq 0 \end{cases}$$

Il test confronta le medie della due *loss function* e presuppone che qualora le funzioni di perdita dei due modelli siano non statisticamente diversi tra loro (ipotesi nulla accettata) questi forniscano la stessa precisione nelle previsioni.

La statistica test è la seguente:

$$S = \frac{\bar{d}}{\sqrt{avar(\bar{d})}} = \frac{\bar{d}}{\sqrt{\frac{LVR_{\bar{d}}}{T}}} = \frac{\bar{d} \sqrt{T}}{\widehat{LVR}_{\bar{d}}}$$

Il numeratore della statistica test si basa sulle differenze tra gli errori di ogni singola previsione previa applicazione della *loss function*.

$$d_t = L(\varepsilon_{t+h|t}^1) - L(\varepsilon_{t+h|t}^2)$$

Dove:

$$\bar{d} = \frac{1}{T} \sum_{t=1}^T d_t$$

$$LVR_{\bar{d}} = \gamma_0 + \sum_{j=1}^{\infty} \gamma_j \quad \text{con } \gamma_j = cov(d_t, d_{t-j})$$

$\widehat{LVR}_{\bar{d}}$ è una stima consistente della varianza di lungo periodo di $\bar{d} \sqrt{T}$

Qualora l'ipotesi nulla venga rifiutata il test fornisce anche un'indicazione sul modello che presenta previsioni maggiormente accurate:

- Se \bar{d} è positivo significa che mediamente gli errori di previsione del primo modello sono maggiori rispetto a quelli del secondo

- Se \bar{d} è negativo, viceversa, gli errori del primo modello saranno mediamente inferiori a quelli del secondo.

Sotto l'ipotesi nulla il test assume una distribuzione normale:

$$S \sim N(0,1)$$

Viene rifiutata l'ipotesi nulla di equità delle previsioni al 5% se:

$$|S| > 1,96$$

Leave One Out ed R^2 reale

Per conoscere le capacità predittive del modello si può utilizzare il metodo della validazione incrociata che consiste nel calcolare il modello togliendo un oggetto alla volta (metodo *leave-one-out*).

La *leave one out* permette di stimare una sorta di R^2 reale: quello fornito dall'*output* dei modelli è infatti ottimistico in quanto calcolato sulla base dell'adattamento del modello sui dati utilizzati per la stima.

Il concetto è quindi quello di valutare se il processo si adatta al modello e non il contrario.

Sia la variabile risposta che le esplicative sono serie storiche e quindi non si possono permutare o togliere dati, in caso contrario le informazioni contenute verrebbero alterate o perse. La funzione che ho sviluppato per implementare la *leave-one-out* non è quindi quella usuale, ovvero dalla totalità dei dati ne verrà destinata una parte per la stima ed un'altra per la validazione; queste due parti verranno aggiornate di volta in volta.

Il modo con cui ho calcolato questo indice è il seguente:

- per prima cosa ho definito la parte della serie destinata alla stima del modello e quella alla validazione dello stesso (eliminando, nei casi in cui vengano utilizzate variabili ritardate, i dati mancanti);
- successivamente ho calcolato la devianza totale TSS (*Total Sum of Squares*) della parte destinata alla validazione:

$$TSS = Var (Validazione) * (N - 1)$$

Ho cioè moltiplicato la varianza per i gradi di libertà;

- utilizzando il modello desiderato ne ho studiato i parametri e calcolato gli errori di previsione ad un passo elevati al quadrato:

$$e(t + 1) = (\widehat{y}_{t+1} - y_{t+1})^2$$

dove y_{t+1} rappresenta il valore realmente osservato contenuto all'interno della validazione;

- infine ho integrato la parte destinata alla stima con il valore appena utilizzato per la validazione, non compromettendo in questo modo l'informazione utilizzata dal modello. Per evitare che gli ultimi errori di previsione vengano stimati con maggiore informazione rispetto a primi e che quindi siano caratterizzati da una varianza più bassa, ho deciso, ricordando le considerazioni precedenti sulla tipologia dei dati, di escludere dalla stima il dato più lontano ovvero il primo della suddetta parte. In questo modo il numero di dati utilizzati e quindi l'informazione sfruttata dal modello rimane costante per tutte le previsioni ottenendo così una stima realistica dell'errore commesso.

Dopo aver calcolato tutti gli errori di previsione ad un passo al quadrato è possibile calcolare l' R^2 reale nel modo usuale:

$$RSS_{CV} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2 \quad ^{11}$$

$$R_{Reale}^2 = \frac{RSS_{Cross\ validation}}{TSS_{validazione}}$$

In questo modo l'indice non sarà più ottimistico in quanto calcolato a partire da previsioni fuori campione, e fornirà una misura di adattamento del processo studiato al modello.

¹¹ Somma degli errori di previsione fuori campione

Confronto delle previsioni utilizzando Diebold Mariano

In questa sezione andremo a confrontare le varie previsioni stimate attraverso i modelli proposti in precedenza.

L'interesse maggiore sarà rivolto al confronto tra i modelli che contengono il GI e il modelli che non lo utilizzano.

In una prima parte verranno confrontate le previsioni utilizzando il test di Diebold Mariano.

È utile ricordare che:

Modello 1: Usa la *query* "cerco lavoro"

Modello 2: Usa la *query* "cerco lavoro" più ulteriori ritardi per la disoccupazione

Modello 4: Usa i volumi congiunti di "cerco lavoro" e "infojobs"

Modello 5: Usa la *query* "cerco lavoro" al tempo t e ritardata di dodici mesi

Mentre il Modello 3 e il Sarima non utilizzano il supporto del GI.

Per ogni confronto è riportato il valore della statistica e il p-value associato.

I dati verranno riportati all'interno della seguente tabella.

	Sarima	Modello 1	Modello 2	Modello 3 (senza GI)	Modello 4	
Modello 1	6,554659 5,58E-11					Stat. Test P-Value
Modello 2	6,544274 5,98E-11	-1,560456 0,1186522				Stat. Test P-Value
Modello 3 (senza GI)	1,963861 0,04954624	-15,76635 5,30E-56	-13,16947 1,32E-39			Stat. Test P-Value
Modello 4	6,673474 2,50E-11	5,246588 1,55E-07	2,9522 0,003155188	17,17279 4,25E-66		Stat. Test P-Value
Modello 5	6,673021 2,51E-11	2,575833 1,00E-02	3,076698 0,002093071	14,48502 1,51E-47	4,257411 2,07E-05	Stat. Test P-Value

Legenda
HO accettata all'1%
HO è rifiutata
HO è accettata ad ogni livello usuale

Tabella 13: test di Diebold Mariano fra i vari modelli

Il numeratore della statistica test è stato costruito effettuando una sottrazione tra il quadrato gli errori dei modelli presenti nelle colonne e quelli delle righe:

$$d_t = (\text{Errori modello colonna})^2 - (\text{Errori modello riga})^2$$

L'interpretazione della tabella è la seguente:

- Qualora l'ipotesi nulla venga accettata i due modelli presentano la stessa precisione nelle previsioni;
- nel caso in cui l'ipotesi nulla venga rifiutata con una statistica test positiva il modello con migliori performance previsive sarà quello citato nella riga corrispondente;
- se la statistica test assume un valore negativo e viene rifiutata H_0 , il modello migliore sarà questa volta quello menzionato sulla colonna.

Per prima cosa valutiamo i confronti tra il modello Sarima stimato all'inizio con gli altri modelli lineari. Si nota come il test di Deibold Marianno rifiuti l'ipotesi nulla con p-value molto bassi e statistiche test positive, tranne nel caso in cui viene confrontato il Sarima con il Modello 3 cioè quello senza Google Index (all'1% infatti l'accuratezza nelle previsioni dei due modelli si può assumere simile).

È opportuno ora analizzare i risultati del test tra il Modello senza il GI e gli altri modelli che invece lo utilizzano. L'ipotesi nulla viene sempre rifiutata con p-value significativi e con statistiche test che volgono a premiare, in modo molto marcato, sempre i modelli che includono l'indice di Google.

Il modello 1 e 2 (cioè rispettivamente quello che utilizza la *query* "cerco lavoro" e quello contenente la stessa chiave con l'aggiunta della variabile riferita alla disoccupazione con ritardo 10) non presentano discrepanze statisticamente significative per quel che riguarda l'accuratezza delle previsioni, tuttavia esiste una piccola differenza tra i due modelli: il primo ha performance previsive leggermente migliori del secondo pur avendo un parametro in meno; da questa analisi viene dunque preferito il modello più semplice.

Per quanto riguarda il quinto e quarto modello (il modello 5 usa la variabile "cerco lavoro" al tempo t e ritardata di 12 lag, mentre il quarto utilizza tra le altre variabili i volumi congiunti di Infojobs e Cercolavoro) il test di Deibold e Marianno chiarisce che il

modello 5 presenta previsioni migliori rispetto all'altro modello e tale differenza risulta essere significativa.

In conclusione i modelli che includono l'indice di Google fra le esplicative presentano previsioni più accurate rispetto agli altri.

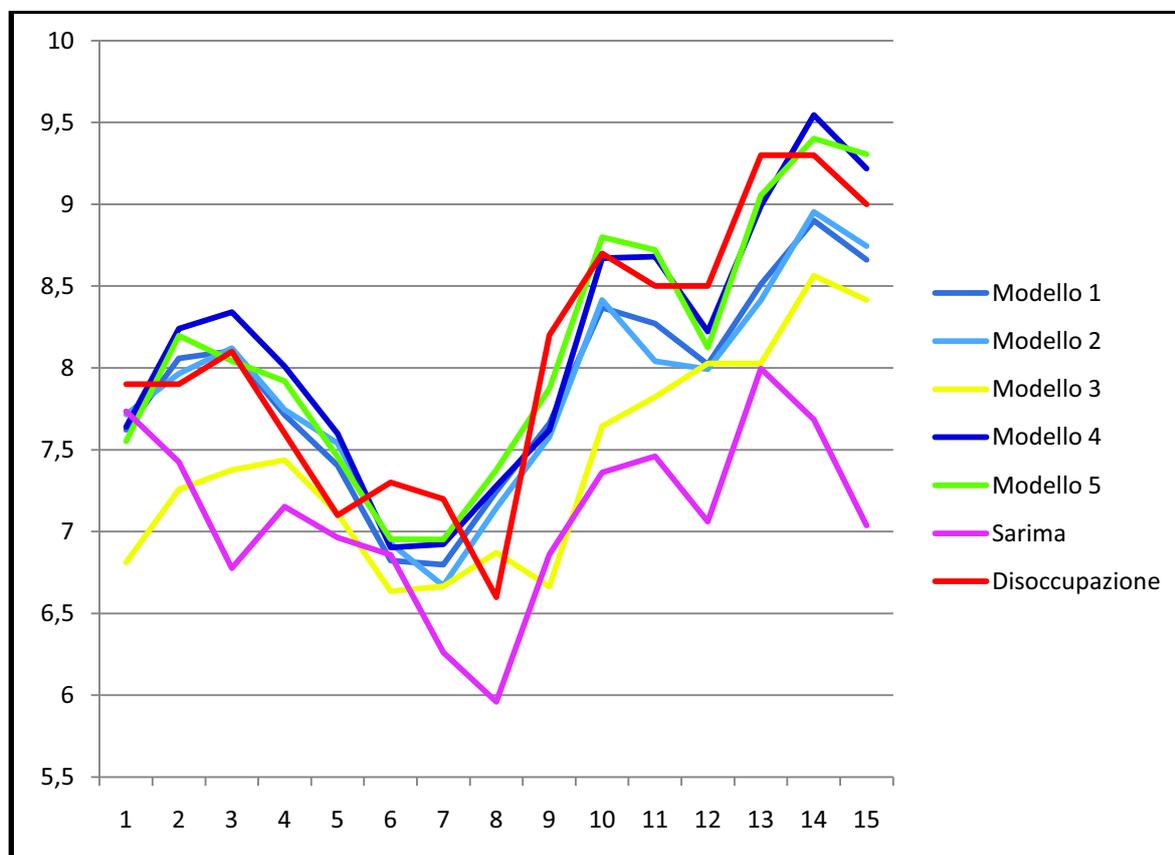


Figura 35: Serie del tasso di disoccupazione e le previsioni out of sample di ogni modello.

Anche dal punto di vista grafico le conclusioni del test Deibold Mariano sono confermate.

I modelli 1 e 2 e quelli 4 e 5, ovvero quelli che contengono le variabili riferite alle query selezionate, presentano previsioni con un andamento molto simile che riescono a seguire in modo soddisfacente la serie del tasso di disoccupazione nonostante il forte incremento.

I modelli Sarima ed il modello 3 che non utilizzano il supporto di Google sono invece quelli con previsioni più lontane dalla serie del tasso di disoccupazione.

Analisi previsioni ad un passo avanti utilizzando la Leave-One-Out

	R ²	R ² Corretto	R ² Reale (cv)	AIC
Modello 1	0,820730	0,808779	0,7440656	45,55707
Modello 2	0,840453	0,825949	0,7139218	41,84583
Modello 3 (senza GI)	0,691537	0,678126	0,2865477	70,14984
Modello 4	0,779985	0,765317	0,763829	55,59249
Modello 5	0,795400	0,786307	0.7902296	48,53901
Sarima	\	\	0,6579915	57.48861

Tabella 14: principali criteri per la valutazione dei modelli proposti.

I modelli con Google Index fra le esplicative presentano un buon adattamento dei dati, e nel caso peggiore i modelli riescono a cogliere il 76% della variabilità .

Il modello che non tiene in considerazione i dati forniti da Google presenta un R² corretto circa dieci punti inferiore agli altri modelli.

Anche con il criterio AIC è chiaro il distacco per quanto riguarda l'adattamento fra i modelli che utilizzano l'indice di Google e gli altri.

Un'osservazione particolare va fatta per il modello Sarima, questo infatti, sempre in base al criterio AIC, sembra presentare un adattamento migliore rispetto al modello senza GI; questo potrebbe essere dovuto al fatto che il modello Sarima sfrutta direttamente la correlazione presente nel processo.

L'R² reale calcolato utilizzando la *leave one out* è uno strumento ottimo per valutare l'adattamento del modello ai dati futuri, tale criterio risulta essere più efficiente rispetto al test Diebold Mariano.

R² reale, essendo calcolato a partire da errori di previsione ad un solo passo avanti, permette di valutare la vera capacità previsiva del modello senza compromettere l'informazione che questo può sfruttare. L'R² reale, come dice il nome, è in questo caso il criterio più attendibile nella valutazione dei modelli.

In particolare l' R^2 reale mette in luce differenze tra accuratezza delle previsioni prima nascoste come quella tra il Sarima e modello 3 (senza GI); il primo infatti risulta più preciso (il Sarima presenta un R^2 reale di 0,65 contro lo 0,28 del modello 3).

È ora opportuno approfondire le differenze tra R^2 corretto e quello reale.

È usuale rilevare un R^2 reale inferiore al R^2 corretto, tale fatto è riscontrabile anche nel nostro caso, ma questa differenza si presenta molto lieve nei modelli che includono Google Index.

Solo nel modello 2 tra R^2 corretto e reale è presente un gap di circa 10 punti: questo conferma quanto ottenuto con il test DB (il modello 1 si presenta migliore del modello 2), il parametro aggiunto nel modello 2 ossia quello riferito al lag 10 della disoccupazione identifica una correlazione spuria non realmente significativa.

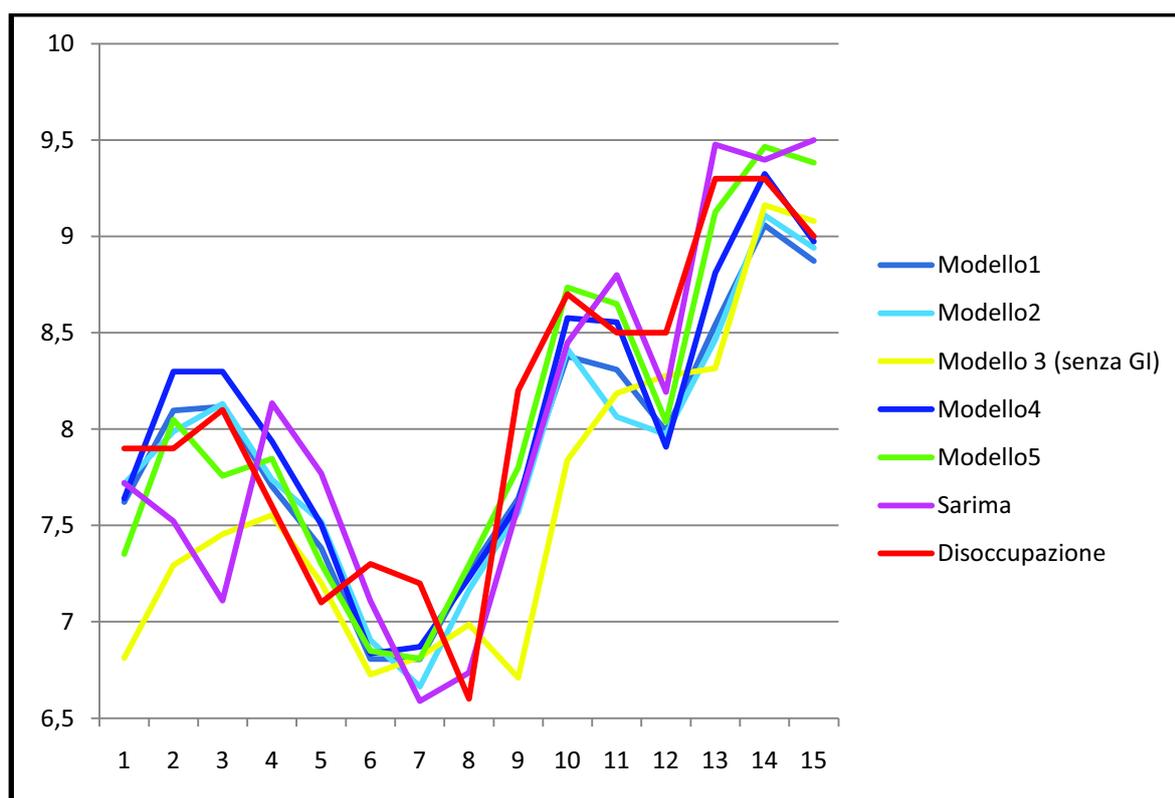


Figura 36: Previsioni ad un passo dei modelli proposti.

Anche da un punto di vista grafico il modello 3 risulta il meno adatto per la previsione ad un passo della serie della disoccupazione.

Pure il SARIMA non rientra fra i modelli più appropriati in questa fase: soprattutto da dicembre 2008 a luglio 2009 non riesce a seguire in modo soddisfacente la serie della disoccupazione.

I modelli che contengono Google Index fra le esplicative invece riescono ad approssimare il processo in tutto il periodo considerato.

Un altro grafico che manifesta la bontà di adattamento dei modelli con Google Index è il grafico degli errori di previsione ad un passo elevati al quadrato.

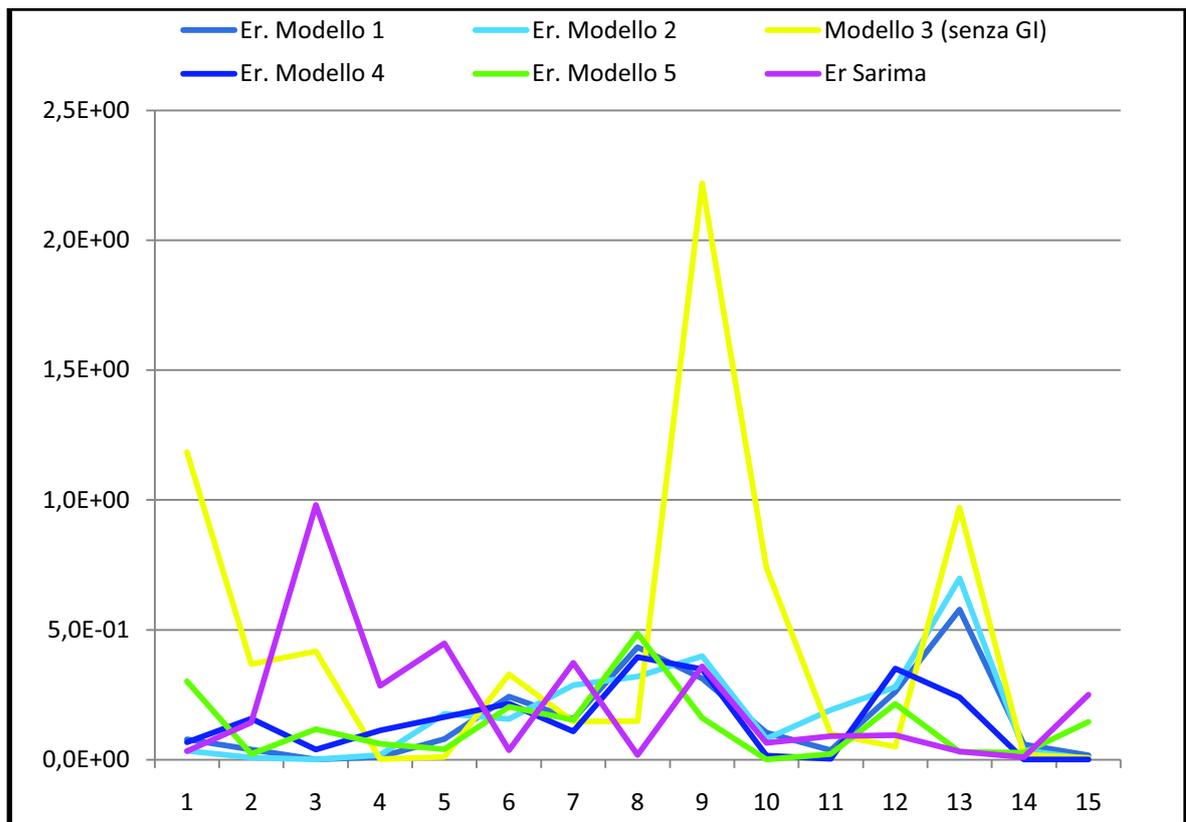


Figura 37: Errori ad un passo al quadrato per i sei modelli.

Il grafico non necessita di ulteriori delucidazioni: come già affermato gli errori commessi dai Modelli Sarima e dal Modello 3 risultano molto più marcati.

6. Conclusioni

L'indice di Google fornisce una preziosa informazione; in particolare nel caso in analisi ne è stata riscontrata l'utilità per la previsione del tasso di disoccupazione.

Google Index presenta anche taluni aspetti negativi.

Il primo di questi risiede nel fatto che i volumi delle chiavi di ricerca vengono normalizzati e riscaldati per proteggere la privacy degli utenti, e come argomentato nel lavoro "A Poisson Regression Examination of the Relationship between Website Traffic and Search Engine Queries", ciò potrebbe causare stime e significatività diverse dei parametri relativi a Google in base alla frequenza dei dati utilizzati.

Un altro aspetto negativo consiste nel fatto che i disoccupati che cercano lavoro utilizzando Google potrebbero non rappresentare un campione casuale, è possibile infatti che questi possano presentare caratteristiche omogenee in base alla classe di età e dal tipo di lavoro ricercato.

Ad oggi non è possibile inoltre dare con tutta certezza valutazioni sui modelli stimati a causa del ridotto numero di dati a disposizione (l'indice di Google fornisce serie storiche che partono solamente dall'anno 2004 e il tasso di disoccupazione fornito dall'Istat è a cadenza mensile).

Nonostante tali aspetti, i modelli stimati mostrano una buona approssimazione della serie della disoccupazione, per di più le informazioni sfruttate sono gratuite e aggiornate settimanalmente.

In conclusione ritengo importante aggiungere che l'indice di Google diventerà sempre più accurato e preciso per la veloce espansione di Internet sulla popolazione in generale, e dell'uso di motori di ricerca come fonte principale per reperire informazioni.

7. Appendice

In questa sezione sarà presentato il codice R relativo alla:

- preparazione delle variabili da inserire nel modello;
- stima di un modello che utilizza il Google Index;
- adattamento del modello.

Preparazione delle variabili:

Per creare ritardi della disoccupazione:

```
dis_1 = lag(dis, -1)
dis_1
```

Il comando ha effettivamente ritardato la serie di un passo ma in questo modo è andato a modificare l' arco temporale originale della serie.

Per ovviare a questo problema basta, utilizzando il comando `window`, estrarre una sottoserie definita nel periodo originario.

```
dis_1 = window(dis_1, start=c(2004, 1), end=c(2010, 3))
dis_1
```

Se si vuole utilizzare la variabile `dis_1` ed inserirla all'interno di un dataset questa deve avere la medesima lunghezza della serie risposta; è necessario inserire quindi degli NA in corrispondenza degli spazi creati dalla funzione `lag`.

```
window(dis_1, start=c(2004, 1), end=c(2004, 1)) = NA
```

In questo modo si può identificare l'arco temporale dove inserire gli NA.

Cosa molto importante è il fatto che se non si inserisce nessuna variabile a cui assegnare il risultato di `window` questa andrà a modificare la serie `dis_1`.

Dopo aver applicato la funzione sarà comunicato un messaggio di attenzione.

Questo procedimento si può applicare per tutte le altre variabili ritardate.

Stima del modello:

```
cerco_lavoro = lm(dis ~ dis_1 + dis_11 + clav_cor,  
data=dati)  
summary(cerco_lavoro)
```

Adattamento:

```
res = residuals(cerco_lavoro)  
qqnorm(res)  
qqline(res)  
shapiro.test(res)
```

Funzioni Usate:

In ordine verranno presentate le funzioni per effettuare:

- Leave one out e calcolo R^2 reale per i modelli lineari;
- Leave one out e calcolo R^2 reale Sarima: la funzione utilizza al suo interno un'altra funzione chiamata "inc_data"; i dati utilizzati all'interno della funzione esposta sono in formato serie storica;
- Test DM: la suddetta funzione usa al suo interno altre funzioni chiamate "quadrato" e "modulo" che identificano le funzioni di perdita;

Funzione validazione e calcolo R^2 reale modelli lineari

```

cv_lm = function
(form,dati,col_risp=1,serie,E_esclusi=c(2004,12),S_stima=c(2
005,1),E_stima=c(2008,12)){
  S = start(serie)
  E = end(serie)

  esclusi = window(serie,start=S,end=E_esclusi)
  n_esclusi = length(esclusi)
  dati = dati[-c(1:n_esclusi),]

  n = dim(dati)[1]

  stima = window(serie,start=S_stima,end=E_stima)
  fine_stima = length(stima)
  stima = dati[c(1:fine_stima),]

  inizio_validazione = fine_stima+1
  validazione = dati[c(inizio_validazione:n),]
  B = dim(validazione)[1]
  k = inizio_validazione

  TSS = var(validazione[,1])* (B-1)

  er = rep(NA,B)
  prev = rep(NA,B)
  l = rep(NA,B)

  for (i in 1:B)
  {
    validazione = dati[(k:k),]
    stima = dati[-c(k:n),]

    if (i>1) {stima = stima[-c(1:(i-1)),]}
    l[i] = dim(stima)[1]
    fit = lm(as.formula(form),data=stima)
    prev[i] = predict(fit,newdata=validazione)
    er[i] = (prev[i]-validazione[,col_risp])^2
    if (i!=B) { k = k + 1}
    else {break()}
  }

  # calcolo R2 reale
  RSS = sum(er)
  R_quad = 1 - (RSS/TSS)
  lista=list()
  lista$error = mean(er)
  lista$er = er

```

```
lista$B = B
lista$lung_s = 1
lista$R_quadro = R_quad
lista$prev = prev
lista
}
```

Funzione validazione e calcolo R^2 reale Sarima

```
inc_data = function (data_attuale) {
  anno = data_attuale[1]
  mese = data_attuale[2]
  if (mese == 12)
    { mese = 1
      anno = anno + 1}
  else { mese = mese + 1}
  data_attuale[1] = anno
  data_attuale[2] = mese
  data_attuale
}
```

```
cv_arima = function
(serie,E_esclusi=c(2005,1),S_stima=c(2005,2),E_stima=c(2008,
12)){
  S = start(serie)
  E = end(serie)
  x = serie
  esclusi = window(serie,start=S,end=E_esclusi)
  n_esclusi = length(esclusi)
  x = x[-c(1:n_esclusi)]

  n = length(x)

  stima = window(serie,start=S_stima,end=E_stima)
  fine_stima = length(stima)

  inizio_validazione = fine_stima+1
  validazione = x[c(inizio_validazione:n)]
  B = length(validazione)
  k = inizio_validazione

  TSS = var(validazione) * (B-1)
```

```

er = rep(NA,B)
prev = rep(NA,B)
l = rep(NA,B)
data_attuale = E_stima
data_iniziale = S_stima

for (i in 1:B)
{
  validazione = x[(k:k)]
  stima = x[-c(k:n)]

  if (i>1) {stima = stima[-c(1:(i-1))]}
  l[i] = length(stima)
  # i dati utilizzati per la stima sono in forma di serie
storica
  stima
  =
ts(stima,start=data_iniziale,end=data_attuale,freq=12)
  fit = arima(stima,c(1,1,0),c(0,1,0))
  prev[i] = predict(fit,1)$pred
  er[i] = (prev[i]-validazione)^2
  if (i!=B) { k = k + 1
              data_attuale = inc_data(data_attuale)
              data_iniziale = inc_data(data_iniziale) }
  else {break()}
}
# calcolo R2 reale
RSS = sum(er)
R_quad = 1 - (RSS/TSS)
lista=list()
lista$errorre = mean(er)
lista$er = er
lista$B = B
lista$lung_s = l
lista$R_quadro = R_quad
lista$prev = prev
lista$stima = stima
lista
}

```

Funzione per il test Diebold Mariano

```

modulo = function(x)
{
  abs(x)
}

quadrato = function(x)
{
  (x)^2
}

diebold.mariano.test <- function(x, alternative =
c("two.sided", "less", "greater"), k) {
  if (NCOL(x) > 1) { stop("x is not a vector or univariate
time series")}
  if (any(is.na(x))) {stop("NAs in x")}
  alternative <- match.arg(alternative)
  DNAME <- deparse(substitute(x))
  n <- NROW(x)
  cv <- acf(x, lag.max=k, type="covariance",
plot=FALSE)$acf[,,1]
  eps <- 1.0e-8
  vr <- max(eps, sum(c(cv[1], 2*cv[-1]))) / n)
  STATISTIC <- mean(x) / sqrt(vr)
  names(STATISTIC) <- "Standard Normal"
  METHOD <- "Diebold-Mariano Test"
  if (alternative == "two.sided")
    PVAL <- 2 * pnorm(-abs(STATISTIC))
  else if (alternative == "less")
    PVAL <- pnorm(STATISTIC)
  else if (alternative == "greater") {
    PVAL <- pnorm(STATISTIC, lower.tail = FALSE)
    PARAMETER <- k
    names(PARAMETER) <- "Truncation lag"
    structure(list(statistic = STATISTIC, parameter =
PARAMETER, alternative = alternative, p.value = PVAL, method =
METHOD, data.name = DNAME), class = "htest")
  }
  lista = list()
  lista$z = STATISTIC
  lista$pvalue = PVAL
  lista
}

```

8. Bibliografia

Jonathan D. Cryer and Kung-Sik Chan (2008). *Time Series Analysis With Applications in R Second Edition*.

Hyunyoung Choi and Hal Varian (April 10, 2009). *Predicting the Present with Google Trends*.

Nikolaos Askatas and Klaus F. Zimmermann (June 2009). *“Google Econometrics and Unemployment Forecasting”*. IZA Discussion Paper No. 4201

Francesco D’Amuri and Juri Marcucci (November 2009). *“Google it!” Forecasting the US unemployment rate with a Google job search index*.

Yair Shimshoni and Niv Efron and Yossi Matias (August 17, 2009). *“On the Predictability of Search Trends”*. Google, Israel Labs.

Tierney, Heather L.R. and Pan, Bing College of Charleston (05. November 2009). *“A Poisson Regression Examination of the Relationship between Website Traffic and Search Engine Queries”*. “MPRA Paper No. 19895”.

Diebold and Mariano, R.S. (1995), *“Comparing predictive accuracy”*, *“Journal of Business and Economic Statistics”*.

Web:

<http://www.google.com/support/insights/bin/topic.py?hl=it&topic=13975>

Ringraziamenti

Mi sembra dovuto iniziare i ringraziamenti con la persona a cui la tesi è dedicata. Voglio ringraziare immensamente Anna che mi è sempre stata vicino aiutandomi nei momenti difficili portando allegria e felicità con il suo sorriso. Non ho parole per rendere onore all'aiuto e al sostegno che mi ha fornito durante la stesura di questo lavoro, senza di te non ce l'avrei mai fatta.

Voglio ringraziare il Prof. Livio Finos, relatore di questa tesi, per la disponibilità e l'aiuto che mi ha concesso, ringrazio inoltre il correlatore Dario Solari per avermi fatto scoprire gli interessanti strumenti forniti da Google.

Ringrazio i miei genitori, che, con il loro sostegno sia morale che economico mi hanno permesso di raggiungere questo importante traguardo.

Ringrazio gli amici, quelli di sempre: Riccardo, Marco, Antonio, Federico e quelli trovati durante il mio periodo di studi: Erica, Massimo, Davide e Luisa; tutti hanno contribuito a rendere indimenticabili questi tre anni.

Ringrazio Enry e la Vale per la loro bravura con i papiri.

Ringrazio infine tutte le persone che non ho citato ma che mi sono state vicine e mi hanno accompagnato durante questo percorso.