



UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE E AZIENDALI

"M. FANNO"

CORSO DI LAUREA IN ECONOMIA E MANAGEMENT

PROVA FINALE

LA VALUTAZIONE DELL'ACCURATEZZA DELLE PREVISIONI:

INDICI CLASSICI E IL MASE

RELATORE:

CH.MO PROF. TOMMASO DI FONZO

LAUREANDA: VALERIA ZANCAN

MATRICOLA N. 1066205

ANNO ACCADEMICO 2015-2016

Sommario

Introduzione	5
1. Le misure di accuratezza	7
1.1 Le serie storiche	7
1.2 Metodo previsionale	7
1.3 Classificazione misure di accuratezza	8
1.3.1 Misure di accuratezza dipendenti dall'unità di misura	9
1.3.2 Misure di accuratezza basate sull'errore percentuale	9
1.3.3 Misure di accuratezza basate sull'errore relativo	9
1.3.4 Misure di accuratezza relative	10
1.3.5 Misure di accuratezza basate sull'errore scalare	10
1.3.6 Una schematizzazione delle misure di accuratezza presentate.....	11
1.4 Esempio: la produzione quadrimestrale di birra in Australia	12
2 Le competizioni	16
2.1 Accuracy of forecasting: an empirical investigation	16
2.2 M-Competition.....	17
2.2.1 Commenti sulla M-Competition.....	19
2.3 M2-Competition.....	19
2.4 M3-Competition.....	20
2.5 Theta Method	21
3 Un nuovo focus per la ricerca: le misure di accuratezza	23
4 Il RMSE e il MAPE	25
4.1 Il Root Mean Square Error.....	25
4.2 Il Mean Absolute Percentage Error	26
5 Il MASE	29
5.1 Il MASE nel metodo di Diebold e Mariano	30
5.2 Confronto tra MASE, MAPE e RMSE	31
6 Un esempio pratico	33

6.1	Set up esperimento	33
6.1.1	Serie N0472	34
6.1.2	Serie Walt Disney Corporation.....	35
6.1.3	Serie lubrificanti	36
6.2	Risultati dell'esperimento	37
	Conclusioni.....	38
	Indice delle figure	39
	Bibliografia.....	40

Introduzione

La valutazione dell'accuratezza delle previsioni, con l'utilizzo di un determinato indice di accuratezza, è un tema abbastanza recente; infatti, fino a pochi anni fa la scelta di una misura di accuratezza era più una questione di preferenza dello studioso che una scelta ponderata (Armstrong, 2008).

Prima degli studi sulle misure di accuratezza, l'attenzione degli studiosi era focalizzata sul comprendere quale metodo previsionale fosse il più accurato e se a seconda del tipo di dati di una serie storica economica potesse variare l'accuratezza del metodo.

Solo in seguito allo studio di Makridakis e Hibon (1979) e delle successive tre competizioni da loro organizzate, atte a valutare l'accuratezza dei diversi metodi previsionali, l'attenzione degli studiosi si spostò dalla ricerca del miglior metodo previsionale allo studio di quale sia il miglior modo di valutarlo.

Lo studio di Makridakis e Hibon (1979) ha provato che la validità del metodo previsionale applicato a una serie storica varia in base alla misura di accuratezza utilizzare per valutarlo; ciò ha comportato il passaggio dell'interesse di molti studiosi dalla valutazione dei metodi previsionali agli indici di accuratezza utilizzati per valutarli.

In questa tesi, vengono presentate le varie tipologie di indici e il confronto tra loro. Saranno approfondite le misure di accuratezza maggiormente utilizzate dagli studiosi negli anni '80, ovvero il RMSE, e negli anni '90, ovvero il MAPE, e confrontate con una nuova misura di accuratezza, presentata da Hyndman e Koehler (2006), che risulterà superiore alle altre misure per affidabilità e versatilità: il MASE.

Questo elaborato è suddiviso in sei capitoli. Qui di seguito riporto un sunto di ogni capitolo.

- Nel primo capitolo sono descritte le caratteristiche principali delle serie storiche che verranno analizzate tramite le fonti di vari autori nei successivi capitoli, quindi è spiegato brevemente cosa sia un metodo previsionale e come venga calcolato l'errore previsionale, sono quindi descritte le varie tipologie di misure di accuratezza e, infine, è presentato un esempio pratico svolto da Hyndman (2014) nel quale si vede un'applicazione dei concetti illustrati.
- Nel secondo capitolo, è descritta la pubblicazione di Makridakis e Hibon (1979) e le successive M-Competitions nelle quali si cercava, nella maniera più oggettiva possibile, di provare e, se possibile, generalizzare i risultati ottenuti nella pubblicazione del 1979.

Alla fine del capitolo è descritto brevemente il Theta Method metodo risultato “vincitore” della M3-Competition.

- Il fine del terzo capitolo è spiegare come da queste competizioni gli studiosi abbiano compreso che non era più possibile scegliere che misura di accuratezza utilizzare solo in base alle preferenze personali, ma era necessario analizzare le varie misure di accuratezza, comprenderne i vantaggi e i limiti, per poter utilizzare il miglior indice in ogni situazione.
- Nel quarto capitolo sono descritte due delle misure di accuratezza più utilizzate dagli studiosi in passato: ovvero il *Root Mean Square Error*, il RMSE, e il *Mean Absolute Percentage Error*, il MAPE.
- Nel quinto capitolo è presentato il *Mean Absolute Scaled Error*, il MASE, proposto da Hyndman e Koehler (2006), approfondendo il tema lungo le linee proposte da Franses (2016) che mostra la versatilità del MASE e la sua coerenza con l’approccio di Diebold e Mariano (1995) per confrontare l’accuratezza delle previsioni. Infine sono riassunte le differenze fra questi tre indici provando quindi in base a quanto precedentemente affermato la superiorità del MASE.
- Nel sesto capitolo viene analizzata nei dettagli la pubblicazione di Hyndman e Koehler (2006), in cui gli autori presentano il MASE come misura di accuratezza che dovrebbe essere utilizzata nella prassi comune perché, fra i vari vantaggi, dà sempre un risultato finito e facilmente interpretabile. Per provare la superiorità del MASE, Hyndman e Koehler (2006) applicano quattro diverse tecniche previsive a tre serie storiche con diverse caratteristiche e calcolano alcune misure di accuratezza per valutarle.

1. Le misure di accuratezza

1.1 Le serie storiche

Prima di cominciare la trattazione, sarà necessario spiegare cosa sia una **serie storica** e descrivere alcune caratteristiche particolari di quelle che verranno utilizzate nei successivi capitoli. Per serie storica si intende un insieme di dati numerici ordinati in maniera sequenziale nel tempo (Di Fonzo e Lisi, 2013). Le serie storiche considerate negli studi che verranno presentati successivamente sono **serie storiche economiche**. L'ordine delle osservazioni è rilevante e non è totalmente casuale; un'osservazione in un determinato tempo t è dipendente in qualche maniera dalle osservazioni precedenti, ma non dipende dalle osservazioni successive (Wooldridge, 2006).

Le serie storiche che verranno presentate possono essere pensate come realizzazioni di un processo stocastico (processo casuale) e non di un processo deterministico; quando raccogliamo i dati di una serie storica, abbiamo ottenuto una possibile realizzazione di un processo casuale, infatti oggi non possiamo sapere con certezza il valore di un'osservazione futura, per questo motivo le serie storiche determinate da un processo stocastico hanno le caratteristiche necessarie ad essere definite **variabili casuali** (Wooldridge, 2006).

Un altro fattore importante da sottolineare è che, nelle serie storiche che verranno trattate, le osservazioni rilevate sono **equispaziate**, ovvero sono state raccolte sempre con la stessa determinata frequenza; essa può essere settimanale, mensile o durare un qualsiasi periodo di tempo, ma comunque le osservazioni sono state rilevate in modo regolare. Infine, le serie storiche considerate sono **univariate**, ovvero in un determinato tempo t si registra un'unica osservazione del fenomeno.

Dai dati grezzi può essere importate identificare il Trend e la Stagionalità della serie storica per poter poi fare inferenza tenendo conto di queste componenti in modo da non incorrere in risultati distorti. Il **Trend** descrive l'andamento nel tempo della serie storica, esso non si può osservare nel breve periodo ed è la tendenza di fondo del sistema stesso. La **Stagionalità** è un fenomeno che si ripete in maniera più o meno simile nella serie storica nello stesso determinato periodo dell'anno; ad esempio un aumento dei consumi nel periodo natalizio.

1.2 Metodo previsionale

Da una serie storica è possibile, attraverso metodi statistici, calcolare una previsione dei possibili dati futuri basandosi sull'analisi dei dati già osservati. Tramite uno di questi metodi previsionali, sarà quindi possibile studiare, ad esempio, una serie temporale sulle vendite di un

prodotto, per cercare di prevedere le vendite future ed essere in grado di ottimizzarne la produzione.

Poiché non è facile comprendere se un metodo previsionale è adatto a fornire una buona stima dei dati analizzati, è auspicabile compiere dei test sull'accuratezza dei risultati ottenuti da tali metodi. L'accuratezza delle previsioni può essere valutata solo considerando "how well a model performs on new data that were not used when estimating the model." (Hyndman, 2014) e non quanto il modello sembri adatto ai dati storici precedentemente osservati.

Per questo motivo, quando si sceglie che metodo previsionale utilizzare è consuetudine dividere i dati a disposizione in due parti: i "training data" e i "test data" (Hyndman, 2014). I "training data" sono utilizzati per stimare il modello, mentre i "test data", che sono solitamente il 20% delle osservazioni finali della serie storica da analizzare, sono utilizzati per verificare se i dati ottenuti sono sufficientemente accurati per ritenere attendibili le previsioni ottenute. Una valutazione dell'accuratezza del metodo previsionale utilizzando questa prassi è detta "out-of-sample performance".

Un altro modo per valutare l'accuratezza dei dati è considerare l'"in-sample performance" basandosi su "one-step-ahead forecasts". In questo caso, dai "training data" è calcolata un'unica previsione per il primo periodo successivo, poi viene aggiunta un'osservazione ai "training data" precedenti e si calcola un'unica previsione successiva; si continua aggiungendo un'osservazione per volta ai "training data" calcolando solo la previsione successiva (per questo tale metodo è detto "one-step-ahead"). Quindi le misure di accuratezza sono calcolate sulle singole previsioni ottenute da tale metodo (Hyndman, 2014).

1.3 Classificazione misure di accuratezza

Per comprendere se un metodo previsionale di una serie storica univariata sia performante esistono diverse misure per valutarne l'accuratezza. Esse sono basate su una "loss-function" legata all'errore di previsione, ovvero la differenza fra l'osservazione realmente rilevata e il valore previsto da tale metodo.

Definiamo quindi l'errore di previsione: "Let Y_t denote the observation at time t and \hat{Y}_t denote the forecast of Y_t . Then define the forecast error

$$e_t = Y_t - \hat{Y}_t$$

The forecasts may be computed from a common base time, and be of varying forecast horizons." (Hyndman e Koehler, 2006). Le misure di accuratezza saranno quindi legate da una

“loss-function” a questo errore previsionale e possono essere classificate nelle cinque diverse macro-categorie che vengono descritte nel seguito.

1.3.1 Misure di accuratezza dipendenti dall'unità di misura

Questa categoria di misure di accuratezza dipende dalla scala dei dati trattati. Questa tipologia di misure è estremamente **utile per paragonare diversi metodi previsionali applicati sugli stessi dati**, ma inutile se si vuole paragonare metodi previsionali su diverse serie storiche che abbiano una scala differente. Nonostante ciò, questi metodi furono ampiamente applicati in passato; ad esempio, nella M-Competition venne utilizzato il *Mean Square Error* ($MSE = \frac{1}{n} \sum_{t=1}^n e_t^2$) per comparare serie storiche con scala diversa e questa scelta venne immediatamente criticata (Armstrong e Lusk, 1983).

Un'altra misura “scale-dependent” estremamente utilizzata in passato è il *Root Mean Square Error* ($RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$). Il RMSE verrà poi approfondito nel quarto capitolo.

1.3.2 Misure di accuratezza basate sull'errore percentuale

Le misure di accuratezza che si basano sull'errore percentuale hanno il vantaggio di essere **indipendenti dalla scala della serie storica trattata**. Hyndman e Koehler (2006) definiscono l'errore percentuale come:

$$p_t = \frac{100e_t}{Y_t}$$

Queste misure hanno lo svantaggio di **divergere a infinito, fino a essere indefinite per $Y_t \rightarrow 0$** . Per questo motivo, nella M3-Competition tutti i dati negativi o nulli vennero sostituiti da valori positivi per poter usare come misura di accuratezza il *Mean Absolute Percentage Error* ($MAPE = \frac{1}{n} \sum_{t=1}^n |p_t|$). Anche il MAPE verrà poi approfondito nel quarto capitolo.

1.3.3 Misure di accuratezza basate sull'errore relativo

Per errore relativo si intende il rapporto fra l'errore di un determinato metodo e quello di un metodo base con il quale si potranno poi confrontare tutti gli altri metodi.

Quindi formalmente l'errore relativo è:

$$r_t = \frac{e_t}{e_t^*}$$

in cui e_t è l'errore del metodo da paragonare, mentre e_t^* è l'errore del metodo di riferimento, ad esempio, il *random walk* anche chiamata previsione *naïve* o *no-change* in cui la previsione è semplicemente $\hat{Y}_t = Y_{t-1}$. Un esempio di misura di accuratezza basata sull'errore relativo è il *Mean Relative Absolute Error*, la cui formula è: $MRAE = \frac{1}{n} \sum_{t=1}^n |r_t|$.

Le misure di accuratezza basate sugli errori relativi hanno il difetto di essere estremamente difficili da calcolare per e_t^* piccoli. Nonostante queste problematiche alcuni autori, tra cui Armstrong e Collopy (1992), raccomandarono l'utilizzo di misure di accuratezza di questo tipo.

1.3.4 Misure di accuratezza relative

Per misura di accuratezza relativa si intende il rapporto fra una misura di accuratezza e la stessa misura calcolata tramite il metodo di riferimento (benchmark method); ovvero, solitamente, il *random walk*. Un esempio di misura relativa proposta da Hyndman e Koehler (2006) è il RelMAE; ovvero $RelMAE = MAE/MAE_b$, in cui $MAE = \frac{1}{n} \sum_{t=1}^n |e_t|$ e il pedice b nel denominatore della formula del RelMAE sta a indicare che il MAE_b è il MAE calcolato per il metodo utilizzato come benchmark.

Utilizzare le misure relative fa in modo che i vari metodi siano facilmente confrontabili tramite una semplicissima interpretazione. Nel caso del RelMAE, se $RelMAE < 1$, allora il metodo proposto è migliore del metodo di riferimento, mentre, se $RelMAE > 1$ il metodo proposto è peggiore di quello di riferimento. Sempre fra le misure relative si trova il *Percent Better* "...the percentage of forecasts for which a given method is more accurate than the random walk." (Hyndman e Koehler, 2006). Il *Percent Better* nel caso del MAE è definito come:

$$PB(MAE) = 100 \frac{1}{n} \sum_{t=1}^n (I\{MAE < MAE_b\})$$

Dove $I\{MAE < MAE_b\}$ è la funzione indicatrice che assume il valore 1 quando $MAE < MAE_b$ e il valore 0 altrimenti. Il limite del *Percent Better* è che non è in grado di evidenziare quale sia il miglioramento possibile e quindi in quale misura un metodo sia migliore o peggiore rispetto al metodo di riferimento.

1.3.5 Misure di accuratezza basate sull'errore scalare

Si è visto che le misure di accuratezza basate sull'errore relativo o su misure relative eliminano il problema della diversa scala fra le serie storiche, ma hanno anche degli svantaggi. Infatti, le misure basate sull'errore relativo hanno il problema che esso ha una distribuzione con media indefinita e varianza infinita. Le misure relative, invece, possono essere calcolate solo quando

vengono calcolati un numero sufficiente di diversi metodi previsionali applicati alla stessa serie storica.

Per ovviare a tutti questi problemi, Hyndman e Koehler (2006) trovano la soluzione "... by scaling the error based on the in-sample MAE from the naïve (random walk) forecast method".

Quindi definiscono l'errore scalare come:

$$q_t = \frac{e_t}{\frac{1}{n-1} \sum_{i=2}^n |Y_i - Y_{i-1}|}$$

questa misura è chiaramente indipendente dalla scala dei dati proposti.

Fra le misure di questa macro-categoria Hyndman e Koehler (2006) propongono come maggiormente performante il *Mean Absolute Scaled Error*, definito come $MASE = \frac{1}{n} \sum_{t=1}^n |q_t|$, che verrà presentato successivamente nel quinto capitolo.

1.3.6 Una schematizzazione delle misure di accuratezza presentate

Le misure di accuratezza presentate nei precedenti paragrafi sono state divise in cinque macro-categorie che ho riassunto, per chiarezza, nella seguente tabella.

TIPO DI MISURA DI ACCURATEZZA	TIPO DI ERRORE CONSIDERATO	ESEMPIO
DIPENDENTE DALL'UNITÀ DI MISURA	$e_t = Y_t - \hat{Y}_t$	$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$
BASATA SULL'ERRORE PERCENTUALE	$p_t = 100(Y_t - \hat{Y}_t)/Y_t$	$MAPE = \frac{1}{n} \sum_{t=1}^n p_t $
BASATA SULL'ERRORE RELATIVO	$r_t = e_t/e_t^*$	$MRAE = \frac{1}{n} \sum_{t=1}^n r_t $
RELATIVA	$e_t = Y_t - \hat{Y}_t$	$RelMAE = MAE/MAE_b$
BASATA SUGLI ERRORI SCALARI	$q_t = \frac{Y_t - \hat{Y}_t}{\frac{1}{n-1} \sum_{i=2}^n Y_i - Y_{i-1} }$	$MASE = \frac{1}{n} \sum_{t=1}^n q_t $

Nella parte successiva dell'elaborato saranno approfondite alcune di queste misure di accuratezza e citate altre facenti parte delle macro-categorie sopra elencate, quindi può risultare utile rifarsi a tale tabella per ricordarsi il tipo di errore su cui si basano.

1.4 Esempio: la produzione quadrimestrale di birra in Australia

È utile adesso considerare un esempio di misure di accuratezza calcolate su una medesima serie storica. L'esempio qui descritto è tratto da Hyndman (2014).

Hyndman (2014) stima un modello ARIMA (AutoRegressive Integrated Moving Average, Box e Jenkins, 1976) ad una serie storica economica della produzione di birra, rilevata quadrimestralmente, in Australia.

L'intervallo temporale della serie storica economica univariata considerata da Hyndman (2014) comincia nel 1992 e si conclude nel secondo quadrimestre del 2010. Egli quindi divide i dati osservati; i "training data" sono i dati dal 1992 al 2006, mentre i dati dal 2007 al 2010 sono i "test data".

Nella figura sotto riportata (Figura 1), si vede l'andamento della serie storica dal 1992 al 2006 della produzione di birra quadrimestrale (parte in nero del grafico), mentre nella seconda parte del grafico si vede la sovrapposizione dei valori predetti col metodo ARIMA da Hyndman (2014) calcolato e i valori realmente osservati nello stesso periodo (sovrapposizione delle spezzate nere e blu).

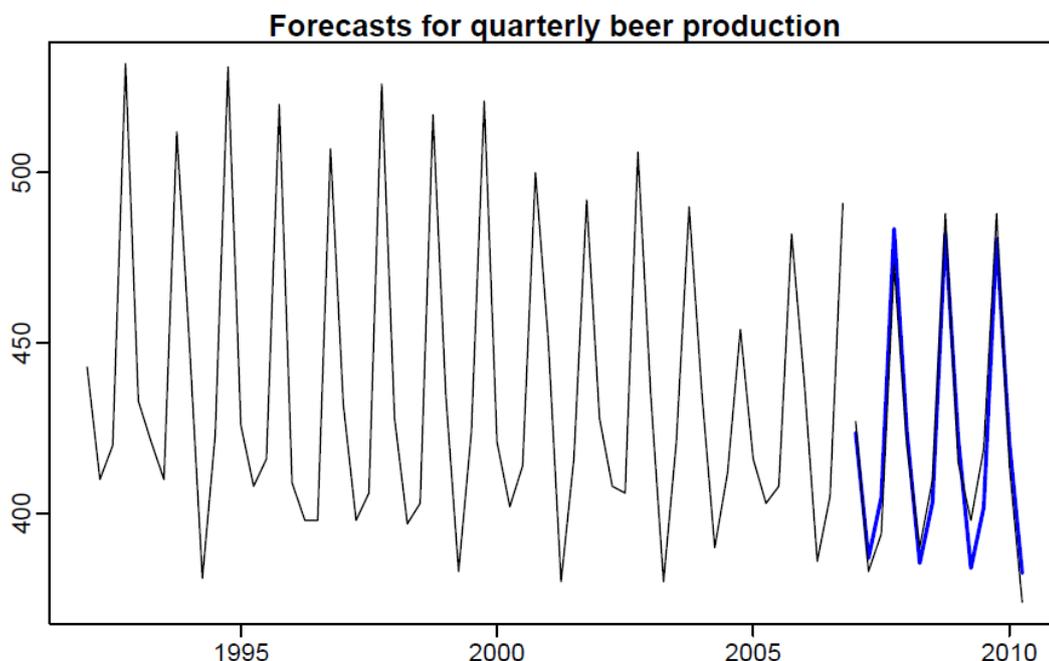


Figura 1- Produzione di birra nel 1992-2010. Fonte : Hyndman (2014)

Da tale grafico (Figura 1) è facile notare che la produzione di birra in Australia ha un andamento stagionale, infatti, la produzione presenta andamenti simili negli stessi periodi dell'anno. Oltretutto, se guardiamo ai valori nella tabella riportata in Figura 2, si ottiene la conferma che la serie ha un andamento stagionale. Nella tabella sono riportate le quattordici osservazioni e le relative previsioni fatte con un modello ARIMA. Nella stessa tabella viene riportato l'errore di previsionale e l'errore percentuale di ogni singola previsione del metodo.

	Actual	Forecast	Error	Percent. error
2007 Q1	427	423.69	3.31	0.78
2007 Q2	383	386.88	-3.88	-1.01
2007 Q3	394	404.71	-10.71	-2.72
2007 Q4	473	483.59	-10.59	-2.24
2008 Q1	420	423.81	-3.81	-0.91
2008 Q2	390	385.42	4.58	1.17
2008 Q3	410	403.25	6.75	1.65
2008 Q4	488	482.13	5.87	1.20
2009 Q1	415	422.35	-7.35	-1.77
2009 Q2	398	383.96	14.04	3.53
2009 Q3	419	401.79	17.21	4.11
2009 Q4	488	480.67	7.33	1.50
2010 Q1	414	420.89	-6.89	-1.66
2010 Q2	374	382.50	-8.50	-2.27

Figura 2 - « test data », produzione birra. Fonte : Hyndman (2014)

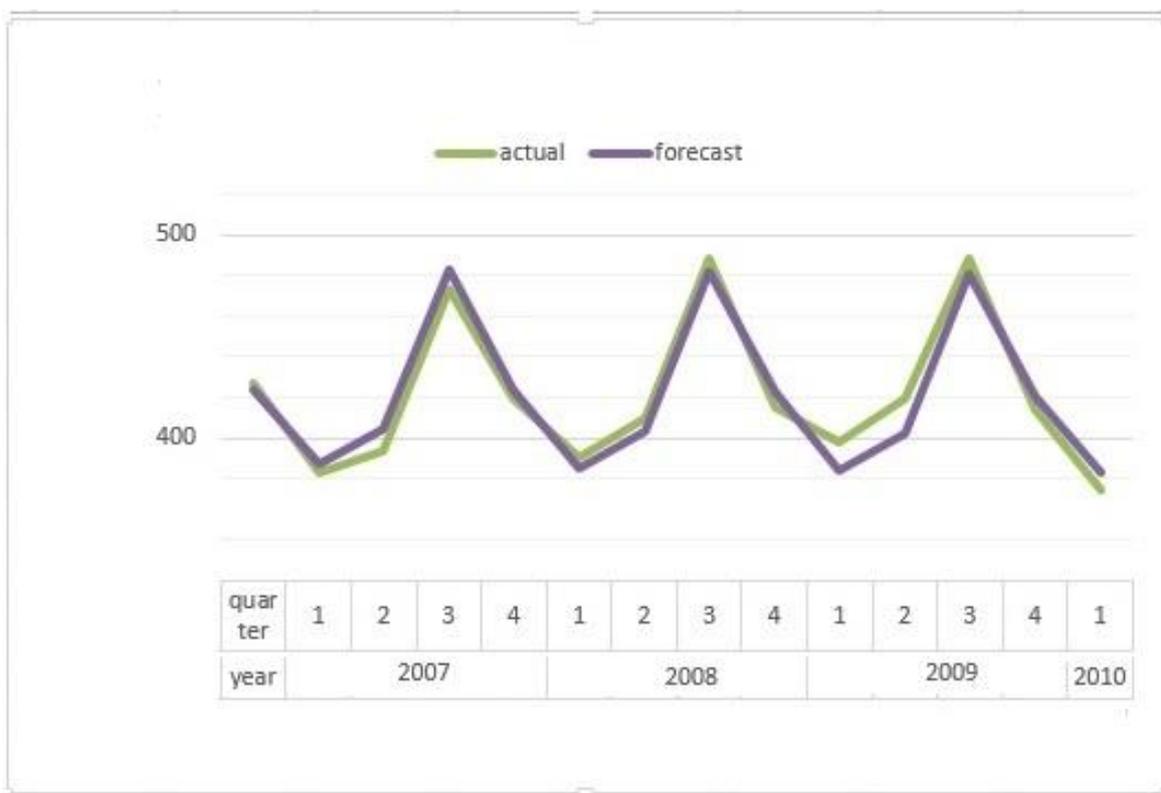


Figura 3- Produzione di birra: 2007-2010. Rielaborazione su dati di Hyndman (2014)

Per aver più chiara la differenza fra i valori veramente verificatisi e quelli previsti dal metodo ARIMA i dati in Figura 2 vengono rappresentati nel grafico in Figura 3, in cui è considerato solamente il periodo dal 2007 al 2010. Vediamo qui che i dati previsti non sembrano discostarsi di molto dai “test data”, ma per avere indicazioni più precise è necessario considerare alcune misure di accuratezza.

Hyndman (2014) calcola quattro misure di accuratezza per valutare la validità del modello previsionale ARIMA. Calcola due indici che dipendono dalla scala dei dati della serie storica, il *Mean Absolute Error*, $MAE = 7,92$, e il *Root Mean Squared Error*, $RMSE = 8,82$. In questo caso questi indici possono essere utilizzati senza avere il dubbio che il loro risultato sia distorto, poiché l’unità di misura delle previsioni ottenute sono le stesse di quelle della serie storica di partenza.

Il valore del *Mean Absolute Percentage Error*, calcolato da Hyndman (2014), è di 1,89%. In questo caso, l’indipendenza del MAPE dall’unità di misura della serie storica non dà maggiori informazioni alla valutazione dell’accuratezza del modello ARIMA applicato per prevedere la produzione futura di birra. Comunque il MAPE può essere utilizzato senza problemi in questo

esempio perché tutti i dati presi in considerazione sono positivi, per questo motivo si è certi che non si ottenga un MAPE infinito o indefinito.

L'ultimo indice che Hyndman (2014) calcola è il *Mean Absolute Scaled Error*, $MASE=0,54$. Tale misura di accuratezza ci dà un'informazione aggiuntiva a quella degli indici precedenti, infatti, senza dover calcolare altri valori con metodi predittivi differenti, col MASE si ha già un confronto con un altro metodo predittivo: per come è calcolato questo indice, si sa che quando il $MASE < 1$ il metodo analizzato produce un risultato più accurato di quello del metodo naïve.

2 Le competizioni

In questo capitolo sono presentate le pubblicazioni di Makridakis e Hibon (1979) e le successive competizioni del 1982, 1993 e del 2000 in cui gli autori cercarono di comprendere quale metodo previsionale fosse maggiormente performante e se alcuni risultati della pubblicazione “*Accuracy of forecasting: an empirical investigation*” potessero essere generalizzati in qualsiasi situazione. È importante parlarne perché in seguito ai risultati di questi studi, molti studiosi compresero che prima di individuare il migliore metodo previsionale, era necessario spostare l’attenzione sulle misure di accuratezza utilizzate per valutarli. Infatti, rifacendosi ai risultati di queste competizioni molti studiosi cominciarono a pubblicare studi sulle misure di accuratezza.

2.1 Accuracy of forecasting: an empirical investigation

Nel 1979 Spyros Makridakis e Michèle Hibon pubblicarono “*Accuracy of forecasting: an empirical investigation*”. Questo studio fu il primo in cui si cercò di comparare nella maniera più oggettiva possibile vari metodi di previsione temporale.

Il focus di tale pubblicazione era la comprensione di quale metodo di previsione temporale fosse in grado di predire gli eventi futuri con maggiore accuratezza. Per far ciò gli autori basarono la loro analisi su un campione di 111 serie storiche, tratte da situazioni reali; esse erano state raccolte da varie fonti, tra cui stati, aziende e industrie. Il numero di osservazioni di queste serie storiche non era lo stesso per tutte le serie, alcune erano serie stagionali e altre no e le osservazioni erano prese ad intervalli di tempo differenti (i.e. mensili, quadrimestrali, annuali...). Come spiegarono gli autori stessi “This selection was made in order to minimize any bias that could arise from the use of a single source of data”. Nonostante i loro sforzi dovettero ammettere che anche questa non poteva comunque dirsi una scelta rappresentativa, perché la maggior parte delle fonti venivano dalla Francia, erano prese ad intervalli mensili, durante gli anni '70 (Makridakis e Hibon 1979).

Per valutare l’accuratezza di questi metodi predittivi utilizzarono $n_j - 12$ osservazioni per ogni serie storica j e predissero i valori per quelle dodici osservazioni successive tramite il modello considerato, ottenendo gli errori di previsione:

$$e_{tj} = X_t - \hat{X}_{tj} \quad t = n_j - 11, n_j - 10, \dots, n_j^1$$

in cui n_j indica la numerosità della j -esima serie storica, $j = 1, \dots, 101$.

¹ L’errore così definito da Makridakis e Hibon (1979) è lo stesso che era stato precedentemente definito nel primo capitolo come errore previsionale ($e_t = Y_t - \hat{Y}_t$).

L'accuratezza del metodo predittivo considerato è dato dal valore dell'errore e_t e può essere calcolato tramite una "loss function" opportuna. Gli autori decisero di utilizzare alcune delle misure di accuratezza più frequentemente utilizzate nella prassi; il *Mean Square Error* (MSE) che è una misura di accuratezza che dipende dalla scala della serie storica e quindi non è adatto per fare paragoni fra serie con scale differenti, il *Theil's U-coefficient* che è una misura relativa ed è estremamente utile per comparare i vari metodi predittivi con un metodo di riferimento - ad esempio il *Naïve Method* o il *Random Walk Method*- e infine il *Mean Absolute Percentage Error* (MAPE) che è una misura di accuratezza che non dipende dalla scala ed è quindi maggiormente utilizzata quando si ha necessità di comparare più serie con una scala differente (Makridakis e Hibon 1979).

Da questa analisi Makridakis e Hibon arrivarono ad alcune conclusioni che si scostavano molto dai risultati precedentemente ottenuti da studi simili.

Primo fra tutti, provarono che **non sempre metodi predittivi più complessi davano risultati più accurati rispetto a metodi più semplici**. Inoltre **l'ordine di accuratezza dei vari metodi predittivi può variare a seconda della misura di accuratezza utilizzata**. Generalmente la combinazione di due o più metodi aumenta l'accuratezza delle previsioni ottenute in confronto a quelle ottenute dal singolo metodo. Infine, per alcuni metodi predittivi l'accuratezza migliora all'aumentare delle osservazioni raccolte.

Questi risultati vennero contestati dalla maggior parte degli studiosi contemporanei, che criticarono il fatto che tale pubblicazione non fosse realmente oggettiva e che si basava solamente sulla correttezza dei calcoli di due singoli studiosi; per questo motivo molti studiosi contestarono la validità dei risultati della pubblicazione.

Per dimostrare l'oggettività dei risultati ottenuti in "*Accuracy of forecasting: an empirical investigation*" Makridakis e Hibon indirono una serie di competizioni, aperte a chiunque, per verificare nella maniera più oggettiva e incontestabile possibile i loro risultati e comprendere quindi quale metodo predittivo fosse maggiormente performante.

2.2 M-Competition

I risultati della prima competizione, da ora chiamata M-Competition, vennero pubblicati nel 1982 (Makridakis, et al. 1982). Lo scopo era di verificare i risultati della pubblicazione di Makridakis e Hibon del 1979 per renderli incontestabili.

Per poter affermare che i risultati ottenuti da questa competizione fossero generalizzabili, **il numero delle serie storiche passò da 111 a 1001**; non era possibile però utilizzare tutte le

1001 serie per tutti i 24 metodi predittivi utilizzati, quindi in alcuni casi i partecipanti alla competizione utilizzarono un sottogruppo di tali serie per un determinato metodo. Makridakis et al. (1982, p. 5) hanno infatti affermato “Running 1001 time series is a formidable and time-consuming task. It was decided, therefore, by the organizer of this competition, to allow some of the participants to run 111 series only. These 111 series were selected through a random systematic sample. ... The Box-Jenkins, Lewandowsky and Parzen methodologies utilized the same systematic sample of 111 series, whereas the rest employed all series.”. Nell’analisi dei risultati distinsero i metodi che utilizzavano tutte le serie storiche da quelli che ne analizzavano solamente un sottogruppo.

La caratteristica principale della M-Competition era che non erano solamente due studiosi ad applicare tutti i metodi predittivi sulle serie storiche analizzate; ma, il fatto che fosse una competizione, fece in modo che chiunque potesse fare i propri studi su questi dati e studiosi maggiormente preparati su un metodo potessero proporlo per la competizione con i propri risultati. Per questa ragione la critica sollevata sulla pubblicazione del 1979, ovvero che i dati riportati si basavano troppo sulle competenze di due sole persone, non poteva essere più utilizzata per sminuire i risultati della M-Competition.

Come affermano Makridakis e Hibon (2000) “Each expert provided his/her forecasts that were compared, in a post-sample fashion, with actual values not used in developing the forecasting model. These forecast errors were then used to compute the various reported accuracy measure.”. Era quindi possibile valutare l’accuratezza previsionale di un determinato metodo tramite misure di accuratezza utilizzate su tutti i metodi proposti in modo da poterli comparare.

Vennero quindi utilizzate **cinque misure di accuratezza**, due delle quali erano già state utilizzate precedentemente; ovvero il *Mean Absolute Percentage Error* (MAPE) e il *Mean Square Error* (MSE). Fu quindi introdotto l’*Average Ranking* (AR), il *Medians of absolute percentage error* (Md) e infine il *Percentage Better* (PB).

L’utilizzo di più misure di accuratezza si rese necessario perché, nella pubblicazione del 1979, Makridakis e Hibon avevano osservato che a seconda della misura di accuratezza utilizzata un metodo poteva essere valutato in maniera differente, “... the performance of the various methods differs -sometimes considerably- depending upon the accuracy measure (criterion) being used.” (Makridakis, et al, 1982). Inoltre, dai risultati ottenuti fu nuovamente confermato che non sempre metodi predittivi più complessi danno risultati più accurati di altri più semplici.

Con queste misure di accuratezza vennero quindi analizzati tutti i metodi predittivi utilizzati per le varie serie storiche; attraverso l’analisi dei risultati è stato osservato che alcuni metodi

sono più accurati in alcune situazioni mentre in altre lo sono meno (considerando le cinque misure di accuratezza utilizzate). Infatti se prima di scegliere un metodo predittivo **si è a conoscenza delle caratteristiche dei dati della serie storica** (i.e. annuali o mensili, micro o macro, quantità delle osservazioni in possesso...) **è possibile scegliere un metodo adatto ad ottenere risultati migliori**. Ovviamente, ciò può essere ritenuto valido se si ritiene possibile assumere che i risultati di questa competizione possano essere generalizzati.

2.2.1 Commenti sulla M-Competition

M. D. Geurts (si veda Armstrong e Lusk 1983) afferma che i maggiori contributi della M-Competition sono il fatto che **“there is no “best” forecasting technique, the best technique changes from one forecasting horizon to the next, and the best technique changes when different measures of accuracy are used”**, quindi, in questo modo, riconosce i risultati presentati in Makridakis e al. (1982), ma critica il fatto che gli autori non abbiano utilizzato come misura di accuratezza il *Theil's U-coefficient* come era stato fatto in Makridakis e Hibon (1979). Infatti, l'utilizzo solo del *Mean Square Error*, che funziona perfettamente per valutare l'accuratezza di una singola serie, fa in modo che le varie serie non possano essere comparate in maniera corretta, infatti, comparare serie diverse solo utilizzando il MSE porta infatti a risultati distorti, dato che **“the MSE is not scale invariant”** (Armstrong e Lusk, 1983); mentre il *Theil's U-coefficient*, che è una funzione del MSE, rimuove tale problema e fa in modo che i risultati possano essere facilmente comparabili.

Infatti il *Theil's U-coefficient* è definito da Makridakis e Hibon (1979) come:

$$\textit{Theil's U - coefficient} = \sqrt{\frac{\sum_{t=1}^n e_t^2}{\sum_{t=1}^n (Y_t - Y_{t+1})^2}}$$

quindi una funzione del MSE che non dipende dalla scala dei dati analizzati e che paragona il metodo in analisi con il *Random Walk Method*.

2.3 M2-Competition

Una seconda competizione fu indetta nel 1993 per valutare le performances di metodi previsionali di tipo quantitativo su serie storiche rilevate in tempo reale sfruttando **la possibilità di avere ogni informazione richiesta sui dati stessi** (Makridakis et al. 1993).

In pratica lo scopo di questa competizione era di valutare i metodi quantitativi rispetto ai cosiddetti metodi **“judgementally-based”** una volta che siano disponibili le stesse informazioni

in entrambi i modelli. Per questo motivo, erano disponibili per tutti i concorrenti della competizione tutte le informazioni da loro richieste; ad esempio se un macchinario si era rotto o se era previsto uno sciopero. Infatti, “one of the rules of the competition was that the companies were to provide whatever additional information was requested by the participating forecasters.” (Makridakis et al. 1993, p.4).

Nella M2-Competiton venne usato un numero di serie storiche assai inferiore a quelle usate negli studi precedenti, infatti **vennero usate solo 29 serie storiche**. Tutte queste serie furono rilevate in tempo reale da organizzazioni che collaboravano alla competizione, 23 di queste provenivano da imprese di varia natura, mentre le rimanenti 6 erano ricavate da indicatori macroeconomici. In questa competizione furono messi a confronto 16 diversi metodi predittivi.

Per valutare l'accuratezza dei risultati ottenuti, nella M2-Competition, furono utilizzati solamente **tre misure di accuratezza**: il *Mean Absolute Percentage Error* (MAPE), il *Mean Percentage Error* (MPE) e sono state pure calcolate le *relative medians*, ovvero, come definite da Makridakis et al. (1993) “the ratio of the Absolute Percentage Error of a method divided by that of the Naive 1 or Naive 2”.

Da questa competizione gli autori ottennero dei risultati conformi con quelli ottenuti negli studi precedenti; infatti, i metodi che dalla competizione risultavano migliori erano i più semplici, ovvero il “...Dampen and Single exponential smoothing and the combining of smoothing methods.” (Makridakis et al., 1993).

2.4 M3-Competition

La competizione più recente, ovvero la M3-Competition, vide i propri risultati pubblicati nel 2000. In tale competizione **le serie storiche utilizzate furono 3003**, ovvero un numero estremamente più grande di quello di qualsiasi altra competizione precedente. Anche in questo caso i dati erano di tipo differente (i.e. micro, macro, industriali) e con differenti intervalli temporali fra le osservazioni delle varie serie storiche (Makridakis e Hibon, 2000).

Per essere certi di non aver problemi previsionali con le varie serie storiche, ogni tipo di serie aveva un numero minimo di dati; ad esempio, le serie equispaziate annuali necessitavano di almeno 14 osservazioni e le serie equispaziate mensili almeno 48 osservazioni. Inoltre, “All the time series data are strictly positive; a test has been done on all the forecasted values: in the case of a negative value, it was substituted by zero. This avoids any problem in the various MAPE measures.” (Makridakis e Hibon, 2000).

Come nelle precedenti competizioni, ai partecipanti non vennero fornite tutte le osservazioni delle varie serie, ma solo una parte per poter poi confrontare i risultati ottenuti dai vari metodi predittivi con i relativi “test data”. Per valutare i risultati ottenuti vennero utilizzate cinque misure di accuratezza: “*symmetric MAPE, Average Ranking, Median symmetric APE, Percentage Better, and Median RAE.*” (Makridakis e Hibon, 2000).

La M3-Competition ha confermato i risultati degli studi precedenti; inoltre, durante questa competizione fu proposto un nuovo metodo che ha ottenuto risultati estremamente positivi: il metodo *Theta*, che è stato valutato come metodo più accurato della M3-Competition.

2.5 Theta Method

Il Theta Method venne descritto per la prima volta nel 2000 da Assimakopoulos e Nikolopoulos nella pubblicazione “The theta model: a decomposition approach to forecasting”. Tale metodo si basa sull’alterazione della curvatura della serie storica tramite il coefficiente Theta mantenendone però le altre caratteristiche. Come si è visto nella M3-Competition tale metodo ottiene ottimi risultati, soprattutto se applicato a serie storiche mensili o a dati microeconomici, e, per questo motivo, il Theta Method risultò il metodo “vincitore” della competizione (Assimakopoulos e Nikolopoulos, 2000).

In tale modello si modifica la curvatura locale della serie storica tramite l’applicazione del coefficiente Theta sulle “second differences” dei dati, ovvero:

$$X''_{new}(\theta) = \theta X''_{data} \quad \text{dove} \quad X''_{data} = X_t - 2X_{t-1} + X_{t-2}$$

Con $\theta < 1$ la curvatura locale si riduce fino a $\theta = 0$ ove la serie storica si trasforma in una regressione lineare, in questo modo è facile vedere il trend della serie sul lungo periodo (si veda Figura 3). Invece, con $\theta > 1$ la curvatura locale si amplifica e si enfatizzano le variazioni nel breve periodo (si veda Figura 4).

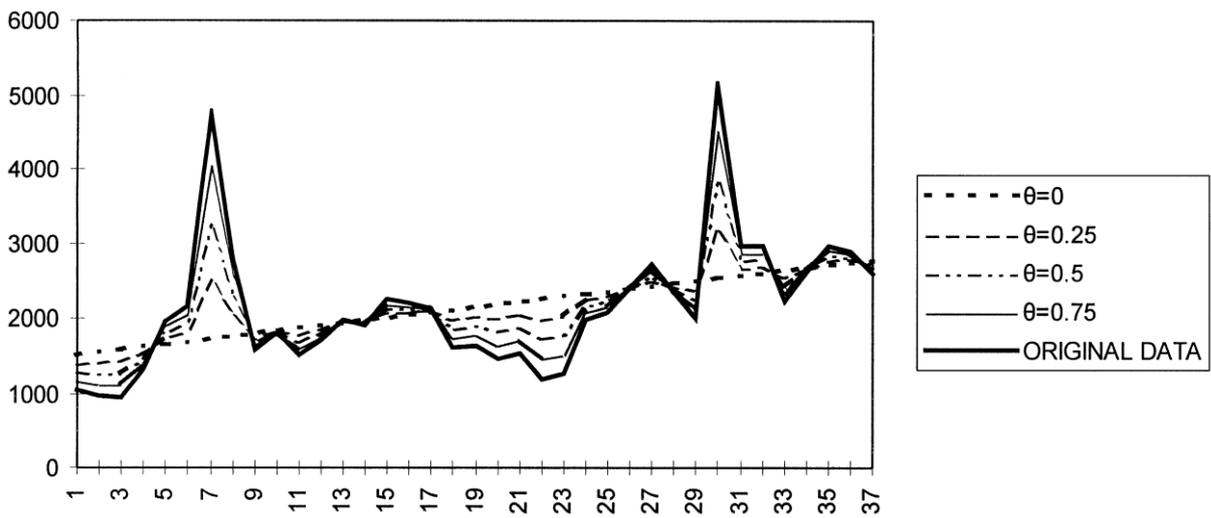


Figura 4- Theta lines con $\theta < 1$. Fonte : Assimakopoulos e Nikolopoulos (2000)

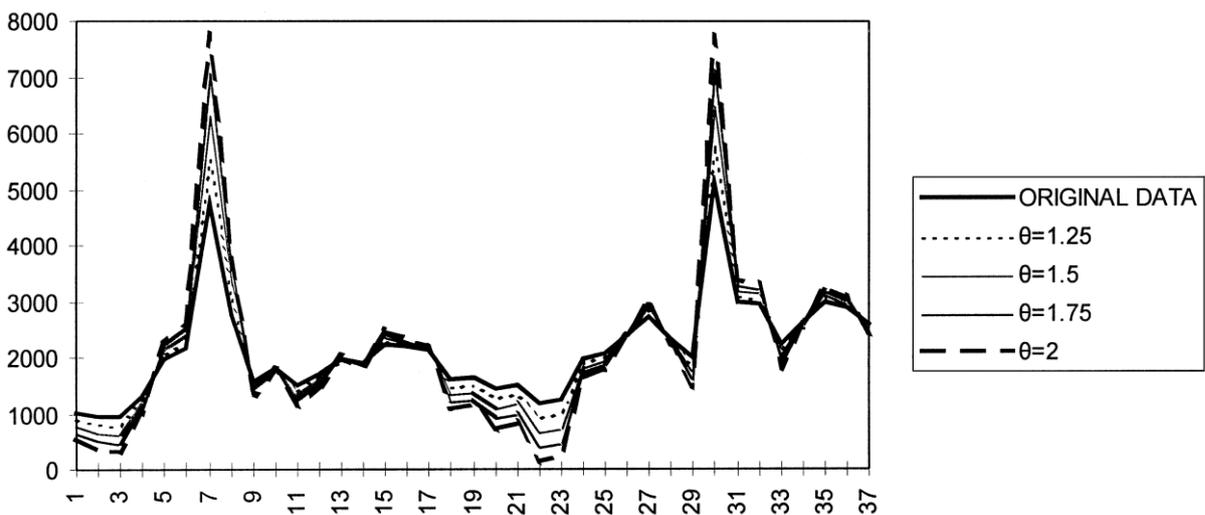


Figura 5- Theta lines con $\theta > 1$. Fonte : Assimakopoulos e Nikolopoulos (2000)

Utilizzando diversi valori di θ “a set of new time series, the so-called Theta-lines, are constructed” (Assimakopoulos e Nikolopoulos, 2000) e si potrà quindi applicare qualsiasi metodo previsionale per calcolare le previsioni di ogni singola “Theta-line”. La combinazione delle previsioni ottenute su ogni singola “Theta-line” darà quindi la previsione finale del modello della specifica serie storica inizialmente considerata.

Tale metodo ottiene buon risultati soprattutto perché, tramite il coefficiente θ , si riesce a dare il giusto peso sia al trend nel lungo periodo della serie ($\theta < 1$), sia a considerare adeguatamente le variazioni nel breve periodo ($\theta > 1$).

3 Un nuovo focus per la ricerca: le misure di accuratezza

Nel capitolo precedente è stata presentata la pubblicazione del 1979 di Makridakis e Hibon e le successive competizioni che furono indette per confermare, nella maniera più oggettiva e aperta possibile, i risultati ottenuti. Il focus di questi studi era l'accuratezza dei metodi predittivi delle serie storiche economiche e la comprensione di quale metodo fosse più adatto a predire risultati migliori data una determinata serie storica.

È importante sottolineare come l'accuratezza previsionale non è solamente un problema teorico, ma ha un riscontro prettamente pratico come si può vedere dall'affermazione di Makridakis e Hibon (2000) "Forecasting accuracy is a critical factor for, among other things, reducing costs and providing better customer service.", quindi, **un miglior metodo previsionale può ridurre i costi sostenuti da un'azienda e migliorare i suoi risultati anche in termini di soddisfazione della clientela**. Si è visto nella M2-Competition che i metodi previsionali quantitativi, una volta che vengono acquisite tutte le informazioni che di solito vengono fornite nei cosiddetti "judgmentally-based methods", riescono ad ottenere risultati estremamente accurati che possono essere utilizzati per implementare i risultati aziendali (Makrdakis et al. 1993).

Si è provato che quasi tutti i metodi previsionali affrontati in queste competizioni hanno utilizzato una tipologia di dati con cui si poteva ottenere un'accuratezza maggiore rispetto ad altri, quindi "If the forecasting user can discriminate in his choice of methods depending upon the type of data (yearly, quarterly, monthly), the type of series (macro, micro, etc.) and the time horizon of forecasting, then he or she could do considerably better than using a single method across all situations..." (Makrdakis et al. 1982). Un altro risultato interessante che si è riscontrato in tutte queste pubblicazioni è che un metodo che, secondo una data misura di accuratezza, risultava migliore rispetto agli altri, utilizzando un'altra misura di accuratezza poteva ottenere una valutazione peggiore e quindi non essere più il metodo più adatto per predire una data serie storica economica.

"Perhaps the greatest benefit of the M-competition has been, not the results as such, but the 'by-products' in making us think more clearly about such issues as error measures and replicability." (C. Chatfield, si veda Armstrong, 2008). Infatti, **non si è raggiunto un risultato univoco su quale metodo di previsione temporale fosse migliore**, ma questa competizione, come quelle successive, ha messo in rilievo un problema che finora era stato considerato marginale, ovvero **se è possibile comprendere quale misura di accuratezza sia più adatta a**

valutare un determinato metodo e se esiste una misura di accuratezza che possa essere utilizzata in tutte le varie situazioni.

A questo proposito è pertinente la domanda posta da Dennis A. Ahlburg: “Is the choice of an error measure to identify the most accurate forecasting method a question of personal taste?” (si veda Armstrong, 2008). Nelle pubblicazioni finora affrontate sembra che la scelta sia stata più casuale che realmente ponderata visto che gli autori non hanno mai motivato la loro scelta. La valutazione dell’accuratezza dipende anche da quale indice viene utilizzato, quindi la scelta della misura di accuratezza dovrebbe essere estremamente rilevante per comprendere se si sta utilizzando il miglior metodo previsionale possibile. Non è quindi secondaria la motivazione degli autori dell’uso di una misura di accuratezza piuttosto che un’altra.

Utilizzare la misura di errore adeguata alla serie storica e al metodo utilizzato è importante, anche perché “Error measures also play an important role in calibrating or refining a model so that it will forecast accurately for a set of time series. That is, given a set of time series, the analyst may wish to examine the effects of using different parameters in an effort to improve a model.” (Armstrong e Collopy, 1992). Infatti la **misura d’accuratezza adeguata può implementare il potere previsionale del modello stesso.**

4 Il RMSE e il MAPE

Come è già stato affermato precedentemente, sarà necessario comprendere quale misura di accuratezza sia più performante. Per far ciò, in questo capitolo, verranno descritte alcune misure di accuratezza che fino a poco tempo fa erano le più utilizzate nella prassi per valutare le serie storiche economiche. Analizzeremo quindi il RMSE e il MAPE descrivendone le caratteristiche principali.

4.1 Il Root Mean Square Error

Il *Root Mean Square Error*, da ora in avanti RMSE, è una misura di accuratezza che si basa sulla scala dei dati analizzati, ed è quindi poco indicata per il confronto fra serie storiche aventi dati con scale diverse. Il RMSE è dato dalla radice quadrata del MSE, ovvero

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

$$RMSE = \sqrt{MSE}$$

Il RMSE è preferito all'MSE perché ha la stessa scala dei dati della serie storica, infatti, una critica alla M-Competition era che era stato usato come misura di accuratezza il MSE. Tale misura di accuratezza era quella maggiormente utilizzata prima dell'introduzione del MAPE, ma il suo utilizzo è anche fortemente criticato; ad esempio Ahlburg (si veda Armstrong, 2008) afferma che “Its use is related to its popularity with statisticians and its interpretability in relation to business decisions and not to its efficiency in choosing the accurate forecasting method. The reliability of RMSE is poor and it is scale dependent.”, quindi, nonostante il suo esteso utilizzo, tale misura di accuratezza ha sia vantaggi che forti svantaggi.

Probabilmente il RMSE fu la misura più utilizzata nei primi anni '80 perché è **una misura ideale per prendere decisioni aziendali**, infatti è semplice interpretarlo: un valore più basso del RMSE solitamente implica un risultato migliore. Inoltre il RMSE è una buona misura per comparare modelli in cui la media dell'errore non sia nulla, perché penalizza le distorsioni da essa derivanti (Nau, 2014).

Nel sondaggio di Carbone e Armstrong (1982) si è visto come il RMSE fosse la misura di accuratezza più utilizzata nella prassi all'inizio degli anni '80, infatti la percentuale di “Practitioners” e soprattutto di “Academicians” che utilizzavano solitamente, nei loro studi, tale misura di accuratezza era la più alta in assoluto. Nonostante il fatto che il RMSE sia una misura dipendente dall'unità di misura della serie storica, la maggior parte di studiosi la preferì a misure di accuratezza, tra cui il MAPE, che non sono dipendenti dalla scala della serie. Comunque il

2% degli studiosi che non preferirono il RMSE, in tale sondaggio, dichiararono di preferire il Theil's U come misura di accuratezza, quindi, in pratica, preferirono una misura che compara il RMSE del metodo analizzato con quello del "random walk method"; quindi, una misura di accuratezza che rimanda comunque al RMSE. Questa preferenza, non è però durata a lungo perché già in un confronto di 17 "papers" condotta da Ahlburg nel 1992 (si veda Armstrong, 2008) solo 4 studi su 17 utilizzarono come misura di accuratezza il RMSE, mentre dieci di questi preferirono utilizzare il MAPE (i restanti preferirono il RMSPE o il Theil's U).

Armstrong e Collopy (1992) studiarono l'affidabilità di alcune misure di accuratezza tra cui il RMSE. Per far ciò, analizzarono se tali misure di accuratezza davano risultati coerenti utilizzando gli stessi metodi previsionali basati su diversi campioni estratti dallo stesso "data set" di serie storiche. Per far ciò calcolarono l'errore previsionale di ogni metodo previsionale utilizzato per ognuno dei cinque sottogruppi di serie storiche utilizzate.

Per comprendere se tale misura stesse dando risultati simili per ogni set di dati calcolarono r_s , ovvero il "Spearman rank-order correlation Coefficients", che misura quanto i risultati fra i diversi "data set" siano coerenti tra loro. Provarono che, in questo caso, il RMSE era "highly unreliable" per comparare diversi metodi previsionali, si è visto infatti che "... the average r_s , for 8-ahead quarterly forecasts was -0.13. Across the four tests, the average pairwise correlation for the RMSE was only 0.20." (Armstrong e Collopy, 1992), un'affidabilità di 0.20 per il RMSE è estremamente bassa.

Calcolarono che per fare in modo che il RMSE fosse affidabile come il "...Percent Better for 18 annual one-ahead forecasts." il "data set" del RMSE doveva comprendere un minimo di 170 serie storiche. Da questa analisi si è visto che **il RMSE non è affidabile per comparare metodi previsionali soprattutto quando ci sono poche serie storiche**, per questo Armstrong e Collopy (1992) affermano che ogni studio precedente, in cui sia stato utilizzato il RMSE su un piccolo "data set", non si deve considerare completamente affidabile senza prima condurre ulteriori analisi.

4.2 Il Mean Absolute Percentage Error

Come già accennato nel primo capitolo, il MAPE è una misura di accuratezza che si basa sull'errore percentuale, che era stato definito come:

$$p_t = \frac{100e_t}{Y_t}$$

Il lato positivo delle misure di accuratezza che si basano sull'errore percentuale è che il rapporto fra e_t e Y_t fa in modo che tale misura sia **indipendente dalla scala utilizzata** e che quindi sia più facile fare confronti fra metodi predittivi utilizzando serie storiche con scale diverse.

Il MAPE è quindi definito come la media dell'errore percentuale in valore assoluto, ovvero:

$$MAPE = \frac{1}{n} \sum_{t=1}^n |p_t|$$

Proprio il fatto che il MAPE è una misura di accuratezza che non è legata alla scala della serie storica lo ha reso una delle **misure maggiormente utilizzate dagli studiosi**, infatti già nei primi anni '90 Ahlburg (si veda Armstrong, 2008) confrontò 17 pubblicazioni di suoi colleghi e notò che in dieci di esse veniva usata come misura di accuratezza il MAPE. Precedentemente la misura di accuratezza più utilizzata nella prassi era il RMSE, ma molti studiosi preferirono il MAPE per la sua caratteristica di poter essere utilizzato per comparare metodi previsionali basati su serie storiche con scala diversa.

Il MAPE, nonostante sia la misura di accuratezza più utilizzata per valutare e comparare i diversi metodi previsionali come afferma Gneiting (2011), ha alcuni svantaggi che lo possono rendere una misura di accuratezza inadatta per queste suddette analisi. Il problema più grande è che la maggior parte degli studiosi che applicano il MAPE non si soffermano su chiedersi se il MAPE possa essere la misura di accuratezza più adatta, infatti come affermano Kolassa e Martin (2011) “using the MAPE for comparisons rewards methods that systematically under-forecast...This problem is poorly understood both among academic forecasters and practitioners in industry” (si veda Tofallis, 2015).

La **distorsione dei risultati** del MAPE è data dal fatto che **tale misura non è simmetrica**, ovvero scambiare Y_t e \hat{Y}_t non porta allo stesso risultato, nonostante il fatto che l'errore assoluto non cambia dopo questo scambio; l'asimmetria è dovuta al denominatore della formula ove scambiare Y_t e \hat{Y}_t porta a un differente risultato (Tofallis, 2015). Proprio per questa asimmetria il MAPE “puts a **heavier penalty on forecasts that exceed the actual than on those that are less than the actual**. For example, the MAPE is bounded on the low side by an error of 100%, but there is no bound on the high side.” (Armstrong e Collopy, 1992), ciò vuol dire che alcuni metodi potrebbero essere scartati per il fatto che ottengono risultati di poco superiore alla reale osservazione.

Per fare un **esempio pratico**: il Metodo 1 predice \hat{Y}_{t1} tale che $\hat{Y}_{t1} = Y_t - \varepsilon$, mentre il Metodo 2 predice \hat{Y}_{t2} tale che $\hat{Y}_{t2} = Y_t + \varepsilon$, in questo caso i due metodi dovrebbero essere valutati in

maniera equivalente dato che si discostano dalla osservazione reale della stessa misura, ma, data l'asimmetria del MAPE, il Metodo 1 sarà valutato migliore del Metodo 2.

Per comprendere meglio questo concetto, Tofallis (2015) spiegò la distorsione data dall'utilizzo del MAPE in maniera grafica: possiamo immaginare la serie storica trattata come un insieme di punti in un piano, è possibile tracciare una linea che rappresenti l'andamento di tali dati, parte di questi saranno sopra la linea e altri sotto, i "... points below the line or in the lower part of the data cloud will have a smaller denominator in the percentage error calculation and so will be more influential in pulling the line towards them.", infatti, il MAPE "... is a weighted form of median regression, where the weights are $1/Y_i$.".

Capiamo quindi che i punti che sono disposti sotto tale linea hanno un peso maggiore perché il denominatore della formula del MAPE sarà più piccolo. Questo maggior peso dei punti disposti sotto la linea farà in modo di spostarla verso il basso così che alla fine ci saranno più punti al di sopra della linea che al di sotto. Il risultato così ottenuto sarà distorto verso il basso, infatti, Foss et al. (2003) argomentano che "MMRE [qui per MMRE si intende il MAPE] may be lower (i.e., "better") for a bad model than for a good model even when the good model happens to be the true model" (si veda Tofallis, 2015), quindi, paradossalmente, da questi studi, si vede come il MAPE tenda a valutare migliore un metodo con previsioni distorte verso il basso piuttosto che i valori che si sono realizzati nella realtà.

Come già spiegato nel paragrafo relativo alla M3-Competition, il MAPE non è in grado di computare misure negative o nulle e per questo motivo, nelle competizioni, tutti i valori negativi o nulli nella serie storica erano stati sostituiti da valori positivi. Inoltre, Armstrong e Collopy (1992) affermano che "A disadvantage of the MAPE is that it is relevant only for ratio-scaled data (i.e., data with a meaningful zero)." ciò vuol dire che misurare l'accuratezza tramite il MAPE di una serie storica di temperature atmosferiche non ha senso. Il fatto che il MAPE non dia risultati soddisfacenti con valori nulli è dato dal fatto che, per $Y_t = 0$, il MAPE è divergente a infinito o indefinito nel caso di una forma $0/0$.

Un altro svantaggio del MAPE espresso da Armstrong e Collopy (1992) è che "The MAPE and MdAPE are less appealing because percentages do not have obvious implications for decision making. For example, in inventory control, losses would be related to dollars of additional inventory or to opportunity costs from lost sales, not to percentage errors. Relative measures have the least relationship to decision making." per questo motivo **il MAPE non è una delle misure di accuratezza migliori se si ha l'intenzione di fare previsioni per decisioni aziendali.**

5 II MASE

Da quanto visto nel capitolo precedente, il MAPE e il RMSE, nonostante siano le misure di accuratezza maggiormente utilizzate per valutare i metodi previsionali di serie storiche economiche, hanno alcuni difetti che possono produrre risultati distorti o non confrontabili tra loro. Per questo motivo Hyndman e Koehler (2006) cercarono una misura di accuratezza che riuscisse a dare ottimi risultati in qualsiasi situazione e che fosse facilmente interpretabile; a seguito della loro analisi proposero come misura di accuratezza standard per valutare i metodi previsionali il *Mean Absolute Scaled Error*, da ora indicato come MASE.

Il MASE è una misura di accuratezza che si basa sull'errore scalare, che avevamo precedentemente definito come:

$$q_t = \frac{e_t}{\frac{1}{n-1} \sum_{i=2}^n |Y_i - Y_{i-1}|}$$

che ha il vantaggio di essere **indipendente dalla scala dei dati utilizzati**, quindi adatto a comparare metodi previsionali con “data set” differenti. Già dall'errore scalare è possibile dare una prima valutazione dei metodi previsionali, infatti se $q_t < 1$ il metodo da cui è stato ricavato è migliore “... than the average one-step naïve forecast computed in-sample” (Hyndman e Koehler, 2006). Invece se $q_t > 1$ il metodo considerato è peggiore del valore predetto con il metodo naïve per il periodo successivo calcolato all'interno del campione.

Una volta calcolato l'errore scalare è estremamente semplice ricavare il MASE, infatti esso è semplicemente la media del valore assoluto dell'errore scalare:

$$MASE = \frac{1}{n} \sum_{t=1}^n |q_t|$$

Il MASE, come l'errore scalare, è **facilmente interpretabile**; ovvero, quando il $MASE < 1$ il metodo analizzato ha, in media, un errore minore a quello del metodo naïve ed è quindi preferibile (Hyndman e Koehler, 2006).

Il vantaggio più grande del MASE è che, a meno del caso limite in cui tutte le osservazioni della serie storica hanno lo stesso valore, esso non diverge mai ad infinito o ha un valore indefinito come capitava con le misure di accuratezza precedentemente analizzate. Per il MASE, a differenza del MAPE, è possibile analizzare serie storiche in cui ci siano valori negativi o nulli senza avere problemi di degenerazione dei valori ottenuti. Per questo motivo con il MASE si ottiene sempre un valore reale che può essere utilizzato per confrontare la validità dei metodi previsionali.

5.1 Il MASE nel metodo di Diebold e Mariano

Per valutare la versatilità del MASE, P. H. Franses (2016) prova che esso può essere utilizzato col metodo di Diebold e Mariano (1995), nel quale si verifica se ci sono differenze statisticamente significative fra i risultati ottenuti utilizzando due diversi metodi predittivi calcolati sulla stessa serie storica. Ovvero, utilizzando le parole degli autori stessi “... we propose widely applicable tests of the null hypothesis of no difference in the accuracy of two competing forecasts.” Diebold e Mariano, vogliono quindi verificare l’ipotesi nulla per la quale non ci sia differenza di accuratezza fra due diversi metodi predittivi.

Il DM test (Diebold e Mariano test) paragona due diversi metodi predittivi basandosi su una “loss function” tale che:

$$l_{i,t} = f(y_t, \hat{y}_{i,t}) \quad i = 1,2$$

Ovvero la “loss function” è una funzione dell’osservazione realmente rilevata al tempo t , y_t , e il valore predetto per il modello 1 o 2, $\hat{y}_{i,t}$. Quindi si calcola la cosiddetta “sample mean loss differential” (Franses, 2016) come:

$$\bar{d}_{12} = \frac{1}{T} \sum_1^T (l_{1,t} - l_{2,t})$$

Si calcola quindi la deviazione standard di \bar{d}_{12} e il DM test per i metodi 1 e 2 è data da:

$$DM = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}}$$

Questa quantità, sotto l’ipotesi nulla, si distribuisce asintoticamente come una Normale standardizzata, usata per verificare l’ipotesi nulla che i due metodi non hanno differenze di accuratezza statisticamente significative.

Franses (2016), dopo aver verificato che “...one can only expect the familiar DM test to have an asymptotic $N(0,1)$ distribution for squared errors, absolute errors and absolute scaled errors.”, provò che il MASE può essere utilizzato come “loss function” per compiere il DM test sui risultati dei diversi metodi applicati e essere in grado di testare “... whether any observed differences in forecast performances are statistically significant.”. Provò, tramite una simulazione, che il MASE può essere utilizzato nel DM test mentre altre misure di accuratezza non possono essere utilizzate anche se Diebold e Mariano nella pubblicazione del 1995 avevano affermato che qualsiasi “loss function” potesse essere utilizzata.

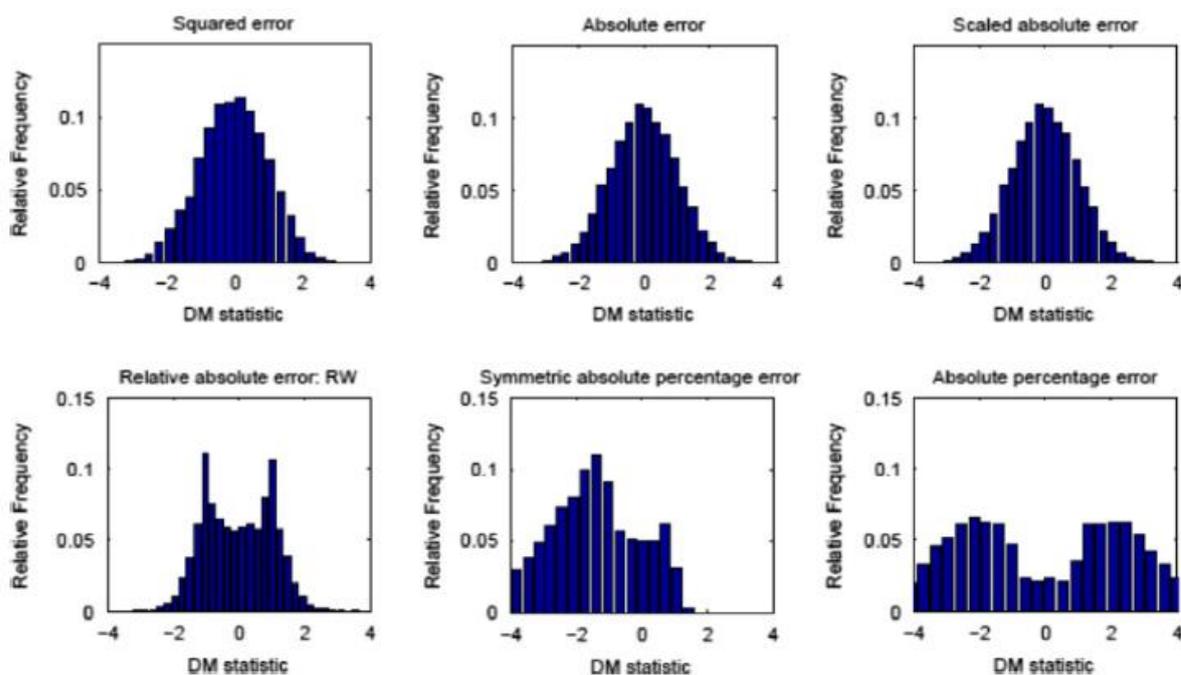


Figura 6- Distribuzione del DM test per vari tipi di errore previsionale. Fonte : Franses (2016)

In Figura 6 sono riassunti i risultati della simulazione compiuta da Franses (2016), ovvero il DM test può essere utilizzato per “... squared errors, absolute error and absolute squared errors.”, quindi per misure come il MSE, il RMSE, il MAE o il MASE, infatti la distribuzione del DM test in figura è una Normale; mentre per misure come il MRAE, il sMAPE o il MAPE non può essere utilizzato, e la distribuzione ottenuta non è una Normale.

In questo modo, Franses (2016) ha provato che tramite l’uso del MASE nel metodo di Diebold e Mariano è possibile comprendere se le differenze fra i risultati di due diversi metodi previsionali siano significative.

5.2 Confronto tra MASE, MAPE e RMSE

La tabella sottostante è stata creata per riassumere le caratteristiche più significative degli indici appena descritti.

Caratteristiche	MASE	MAPE	RMSE
Dipende dall'unità di misura?	no	no	sì
Funziona il metodo D&B?	sì	no	sì
Degenera con osservazioni nulle?	no	sì	no
È facilmente interpretabile per prendere decisioni aziendali?	sì	no	sì
Ottiene risultati per serie con zero solo informativo?	sì	no	sì

Da tale tabella riassuntiva, si nota subito che il MASE, per le caratteristiche considerate, appare senz'ombra di dubbio una misura di accuratezza migliore del MAPE e del RMSE. Sono state confrontate proprio queste caratteristiche, perché nelle fonti considerate sono quelle a cui diversi autori hanno dato più rilevanza.

Il MASE, a differenza del RMSE, non dipendendo dall'unità di misura, può essere utilizzato per confrontare metodi previsionali basati su serie storiche con unità di misura differenti. Il MASE e il RMSE possono essere utilizzati col metodo di Diebold e Mariano per comprendere se la differenza dei risultati di due metodi predittivi sia statisticamente significativa, mentre non è possibile utilizzare il MAPE. A differenza del MAPE, il MASE non diverge a infinito o dà risultati indefiniti con valori di $Y_t = 0$. Il MAPE, a differenza del MASE e del RMSE non può essere utilizzato con serie storiche in cui lo zero è un valore informativo. Infatti, gli errori percentuali assumono che la scala della serie storica sia basata su quantità, e per serie storiche che non misurano una quantità, come può essere una serie storica di temperature atmosferiche (Fahrenheit o Celsius), non ha senso una valutazione con questo tipo di misure. Per prendere decisioni aziendali è meglio utilizzare il MASE o il RMSE perché facilmente interpretabili, mentre il MAPE è sconsigliato per questo fine.

6 Un esempio pratico

Nei capitoli precedenti si è discusso tramite l'utilizzo di fonti solo di tipo teorico della superiorità del MASE sulle altre misure di accuratezza. Per concludere tale discussione è presentato in questo capitolo la dimostrazione pratica di Hyndman e Koehler (2006) in cui quattro metodi previsionali vengono utilizzati su tre diverse serie storiche e sono valutati mediante alcune tra le misure di accuratezza fin qui viste, più altri indici classici precedentemente usati nella letteratura sull'argomento.

6.1 Set up esperimento

Hyndman e Koehler (2006) intendevano trovare la misura di accuratezza migliore per la valutazione e il confronto di diversi metodi previsionali. I due autori affermarono, prima ancora di descrivere l'esperimento, che molte misure proposte in passato non possono applicarsi a qualsiasi situazione e che spesso possono portare a risultati infiniti o indefiniti (quindi non confrontabili) o comunque a risultati distorti.

Gli autori dimostrarono tramite questa applicazione che la miglior misura di accuratezza è il MASE e che, a meno di particolari serie storiche, la maggior parte degli indici raccomandati in passato possono dare risultati non soddisfacenti.

Le tre serie storiche economiche univariate che Hyndman e Koehler (2006) decisero di utilizzare nella loro dimostrazione sono: la serie N0472 della M3-Competition che comprendeva solo osservazioni strettamente positive, il "monthly log stock returns" della Walt Disney Corporation che comprendeva osservazioni anche negative o nulle, le vendite mensili di un prodotto lubrificante venduto in grandi containers con osservazioni non negative.

Per ognuna di queste serie storiche Hyndman e Koehler (2006) calcolarono quattro diversi metodi predittivi: "(1) the historical mean using data up to the most recent observation; (2) the "naïve" or random-walk method based on the most recent observation; (3) simple exponential smoothing and (4) Holt's method.". Per ognuno di questi metodi, confrontano la "in-sample performance" dei metodi basandosi sulla previsione "one-step-ahead" e l'"out-of-sample performance" paragonando i dati predetti con i "test data" non utilizzati per calcolare le previsioni.

Hyndman e Koehler (2016) selezionarono alcune misure di accuratezza per valutare quattro metodi previsionali da loro calcolati, la loro scelta fu di considerare alcuni indici che sono stati spesso raccomandati dalla letteratura precedente al loro studio. Le misure di accuratezza che

sono state scelte per valutare questi quattro metodi previsionali sono riportate nella tabella seguente:

Indice		Tipo di misura di accuratezza
MAPE	Mean Absolute Percentage Error	Basata sull'errore percentuale
MdAPE	Median Absolute Percentage Error	Basata sull'errore percentuale
sMAPE	Symmetric Mean Absolute Percentage Error	Basata sull'errore percentuale
sMdAPE	Symmetric Median Absolute Percentage Error	Basata sull'errore percentuale
MdRAE	Median Relative Absolute Error	Misura relativa
GMRAE	Geometric Mean Relative Absolute Error	Misura relativa
MASE	Mean Absolute Scaled Error	Basata sull'errore scalare

Gli autori sottolinearono che “To our knowledge, the MASE has not been proposed before” (Hyndman e Koehler, 2006) e, quindi, tale indice non era stato utilizzato nemmeno durante la M3-Competition.

6.1.1 Serie N0472

La serie storica N0472 della M3-Competition, come si può vedere dal grafico sotto riportato (Figura 7) prodotto da Hyndman e Koehler (2006), presenta sempre valori strettamente positivi.

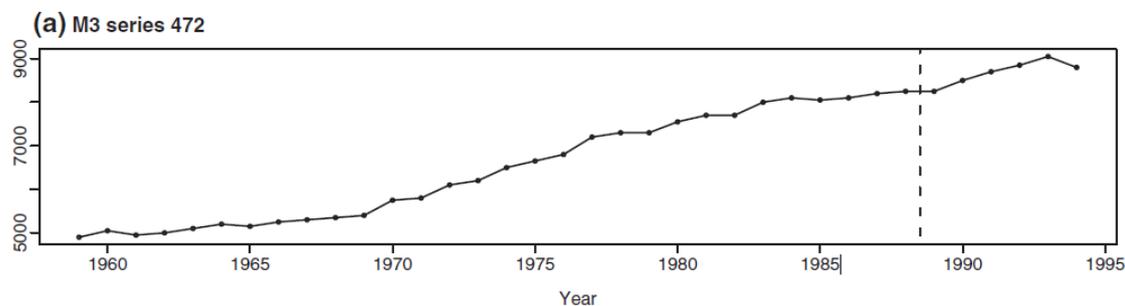


Figura 7- Serie N0472. Fonte : Hyndman e Koehler (2006)

Dal fatto che tutti i valori delle osservazioni sono strettamente positivi, possiamo dedurre che il MAPE darà come risultato un valore finito. Nella tabella (Figura 8) calcolata da Hyndman e Koehler (2006), sono riassunti i valori degli indici per ognuno dei quattro metodi applicati alla serie col valore dell'indice sia per valutare l'”in-sample performance”, ovvero la valutazione della performance basata su una previsione “one-step-ahead”, sia per l'”out-of-sample performance”, ovvero la previsione in cui i dati sono divisi fra “training data” e “test data”.

Table 1
Forecast error measures for M3 series N0472

Example A	Mean		Random walk		SES		Holt	
	In	Out	In	Out	In	Out	In	Out
MAPE	14.09	25.57	2.01	5.00	2.08	5.03	1.31	3.94
MdAPE	17.44	26.13	1.61	5.71	1.86	5.74	1.04	4.20
sMAPE	0.16	0.29	0.02	0.05	0.02	0.05	0.01	0.04
sMdAPE	0.19	0.30	0.02	0.06	0.02	0.06	0.01	0.04
MdRAE	6.50	4.61	Undefined	Undefined	1.02	1.01	0.50	0.80
GMRAE	∞	∞	Undefined	Undefined	∞	∞	∞	∞
MASE	7.88	17.23	1.00	3.42	1.04	3.44	0.66	2.69

Figura 8- Misure di accuratezza, serie N0472. Fonte : Hyndman e Koehler (2006)

In questa serie la maggior parte degli indici ha un valore finito, ma il GMRAE e il MdRAE per la “random walk” risultano divergenti a infinito o indefiniti. Questi valori non confrontabili sono il risultato del fatto che, nella serie N0472, ci sono osservazioni successive che presentano lo stesso valore. Hyndman e Koehler (2006) stessi affermarono che osservazioni successive con lo stesso valore è “... something that very often occurs with real data”, quindi misure di accuratezza che con dati di questo tipo non danno un risultato finito e quindi non confrontabile risultano inadatte per valutare questa tipologia di dati.

6.1.2 Serie Walt Disney Corporation

La serie storica mensile della Walt Disney Corporation è relativa ai rendimenti azionari dal 1990 al 2000 dell’azienda. Tali dati sono divisi in “training data”, dal 1990 al 1998, e in “test data”, dal 1998 al 2000. Questa serie storica, come si può vedere chiaramente dal grafico (Figura 9), ha sia valori negativi che nulli, ma non presenta due osservazioni successive con lo stesso identico valore.

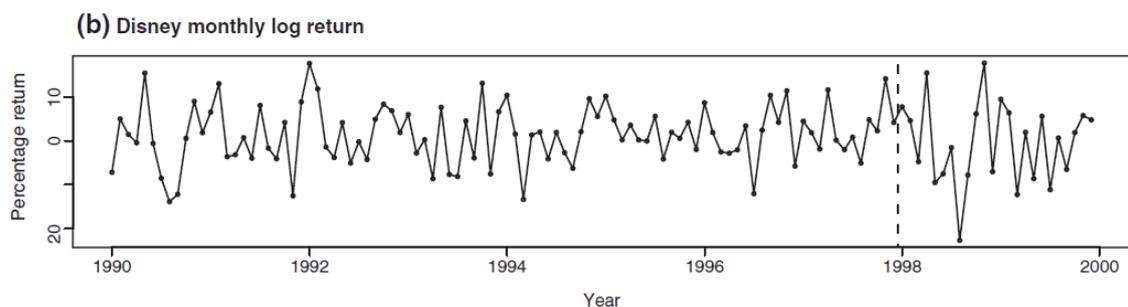


Figura 9- Serie Walt Disney Corporation. Fonte : Hyndman e Koehler (2006)

Table 2
Forecast error measures for Disney stocks

Example B	Mean		Random walk		SES		Holt	
	In	Out	In	Out	In	Out	In	Out
MAPE	∞	96.56	∞	125.90	∞	102.32	∞	113.90
MdAPE	101.61	96.92	159.25	119.19	97.06	98.63	98.70	97.91
sMAPE	-0.80	-2.04	∞	-3.03	-1.10	-0.57	-0.80	-1.12
sMdAPE	0.91	0.35	0.66	0.22	1.01	0.43	0.91	0.48
MdRAE	0.71	0.82	1.00	1.00	0.63	0.88	0.64	0.94
GMRAE	0.66	1.06	1.00	1.00	0.62	1.01	0.60	1.10
MASE	0.72	1.01	1.00	1.06	0.70	1.02	0.70	1.04

Figura 10- Misure di accuratezza, serie Walt Disney Corporation. Fonte : Hyndman e Koehler (2006)

Nella tabella riassuntiva (Figura 10), calcolata da Hyndman e Koehler (2006), notiamo che il MAPE, soprattutto per quanto riguarda la comparazione fra metodi di “one-step-ahead forecasts”, dà risultati che divergono a infinito, quindi non confrontabili tra loro. Anche il sMAPE calcolato per il metodo “random walk” per il valore “in-sample” diverge ad infinito e quindi non può essere utilizzato per questa serie storica.

6.1.3 Serie lubrificanti

L’ultima serie storica analizzata da Hyndman e Koehler (2006) è relativa alla vendita in grandi containers di lubrificante, le osservazioni sono state rilevate mensilmente per tre anni. Dal successivo grafico (Figura 11), vediamo che la serie presenta solo osservazioni positive o nulle e che ci sono alcune osservazioni successive uguali e nulle.

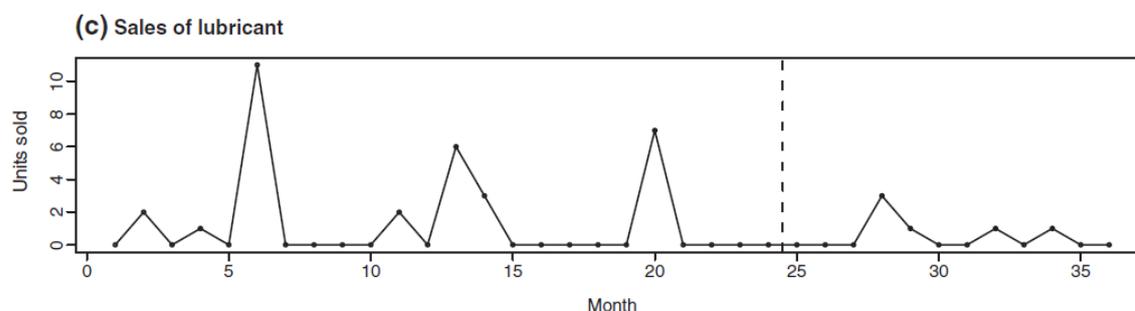


Figura 11- Serie lubrificanti. Fonte : Hyndman e Koehler (2006)

Come visto nelle due precedenti serie, osservazioni nulle e osservazioni successive uguali fanno in modo che molte misure di accuratezza diano risultati che divergono a infinito o della forma 0/0, e quindi siano indefinite.

Table 3
Forecast error measures for lubricant sales

Example C	Mean		Random walk		SES		Holt	
	In	Out	In	Out	In	Out	In	Out
MAPE	∞	∞	Undefined	Undefined	∞	∞	∞	∞
MdAPE	∞	∞	Undefined	Undefined	∞	∞	∞	∞
SMAPE	1.73	1.47	Undefined	Undefined	1.64	1.46	1.66	1.65
SMdAPE	2.00	2.00	Undefined	Undefined	2.00	2.00	2.00	2.00
MdRAE	0.95	∞	Undefined	Undefined	0.89	∞	0.93	∞
GMRAE	∞	∞	Undefined	Undefined	∞	∞	∞	∞
MASE	0.86	0.44	1.00	0.20	0.81	0.43	0.78	0.27

Figura 12- Misure di accuratezza, serie lubrificanti. Fonte : Hyndman e Koehler (2006)

Come si può vedere dalla tabella (Figura 12) il MAPE, il MdAPE e il GMRAE, divergono ad infinito su tre metodi su quattro sia “in-sample” che “out-of-sample” (mentre in MdRAE diverge ad infinito solo “out-of-sample”) e, nel caso della “random walk” sono indefiniti a causa di una forma 0/0. Tutte le misure di accuratezza, a parte il MASE, sono indefinite quando calcolate col metodo del “random walk”.

Il valore del MASE per il metodo “random walk” calcolato “in-sample” è uguale a 1 per definizione, mentre per tutti gli altri metodi dà un risultato finito, compreso fra 0 e 1, che significa che il metodo valutato ha un errore inferiore a quello del metodo utilizzato come riferimento.

6.2 Risultati dell’esperimento

L’analisi compiuta da Hyndman e Koehler (2006) su tre serie storiche aventi dati con caratteristiche diverse, ha illustrato come il MASE è l’unica misura di accuratezza, fra quelle utilizzate nell’esperimento, che dà sempre un valore finito sia “in-sample” che “out-of-sample”. Per questa motivazione, Hyndman e Koehler (2006), hanno proposto che le misure di accuratezza basate sull’errore scalare, nello specifico il MASE, debbano essere utilizzate per valutare e comparare i metodi predittivi; anche perché l’unica occasione in cui il MASE può essere infinito o indefinito è nel caso limite in cui tutte le osservazioni sono uguali tra loro, quindi un caso non rilevante per le applicazioni pratiche.

Conclusioni

In questo elaborato è stato esposto il processo attraverso il quale si è compreso che, prima di poter porsi il problema di quale sia il miglior metodo previsionale, è necessario valutare se le misure di accuratezza utilizzate per valutare tale metodo siano adatte allo scopo. Molti studiosi hanno dato peso a questa problematica solo in seguito alle M-Competition e, per questo motivo, è stato utile illustrarne gli obiettivi e i risultati che hanno dato adito alle successive pubblicazioni sugli indici di accuratezza.

Si è quindi passato ad un confronto delle misure di accuratezza maggiormente utilizzate dalla letteratura: il RMSE, il MAPE e il MASE. Dal semplice confronto delle caratteristiche più salienti di questi indici si è constatato che il MASE può essere utilizzato in un maggior numero di situazioni, dando sempre risultati soddisfacenti e facilmente interpretabili.

Descrivendo poi la pubblicazione di Hyndman e Koehler (2006) si è provato, tramite un esempio reale, che ciò che era stato affermato precedentemente solo in via teorica è vero anche nella pratica. Infatti, gli autori hanno dimostrato come il MASE sia superiore agli altri indici utilizzati e come tale indice possa essere utilizzato in qualsiasi situazione senza degenerare (a meno del caso limite di dati tutti uguali).

La dimostrazione che il MASE funzioni bene in qualsiasi situazione non esclude la possibilità di utilizzare un'altra misura di accuratezza per valutare un determinato metodo previsionale nel caso in cui la scelta di tale misura sia giustificata (ad esempio è più facilmente calcolabile) e se si è certi che, nella situazione concreta presa in considerazione, tale misura di accuratezza non fornisca risultati distorti.

Indice delle figure

Figura 1- Produzione di birra nel 1992-2010. Fonte : Hyndman (2014).....	12
Figura 2 - « test data », produzione birra. Fonte : Hyndman (2014).....	13
Figura 3- Produzione di birra: 2007-2010. Rielaborazione su dati di Hyndman (2014).....	14
Figura 4- Theta lines con $\theta < 1$. Fonte : Assimakopoulos e Nikolopoulos (2000).....	22
Figura 5- Theta lines con $\theta > 1$. Fonte : Assimakopoulos e Nikolopoulos (2000).....	22
Figura 6- Distribuzione del DM test per vari tipi di errore previsionale. Fonte : Franses (2016)	31
Figura 7- Serie N0472. Fonte : Hyndman e Koehler (2006).....	34
Figura 8- Misure di accuratezza, serie N0472. Fonte : Hyndman e Koehler (2006).....	35
Figura 9- Serie Walt Disney Corporation. Fonte : Hyndman e Koehler (2006)	35
Figura 10- Misure di accuratezza, serie Walt Disney Corporation. Fonte : Hyndman e Koehler (2006)	36
Figura 11- Serie lubrificanti. Fonte : Hyndman e Koehler (2006)	36
Figura 12- Misure di accuratezza, serie lubrificanti. Fonte : Hyndman e Koehler (2006).....	37

Bibliografia

- ARMSTRONG, J. S., (2008), *A Commentary on Error Measures*. International Journal of Forecasting Vol. 8. pp. 99-111. Disponibile su:<http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1164642> [Data di accesso: 01/09/2016]
- ARMSTRONG, J. S., CARBONE, R., (1982), *Evaluation of Extrapolative Forecasting Methods: Results of a Survey of Academicians and Practitioners*. Journal of Forecasting, Vol. 1, 215-217
- ARMSTRONG, J.S., COLLOPY, F., (1992), *Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons*. International Journal of Forecasting, Vol. 8, 69-80
- ARMSTRONG, J.S., LUSK, E.J., (1983), *Commentary on the Makridakis Time Series Competition (M-Competition)*. Journal of Forecasting, Vol. 2, 259-311.
- ASSIMAKOPOULOS, V., NIKOLOPOULOS, K., (2000), *The theta model: a decomposition approach to forecasting*. International Journal of Forecasting, Vol. 16, 521-530
- BOX, G., JENKINS, G., (1976), *Time series analysis, forecasting and control*, Holden-Day, San Francisco.
- DIEBOLD, F., MARIANO, R., (1995), *Comparing Predictive Accuracy*. Journal of Business and Economic Statistics, Vol. 13, 253-263
- DI FONZO, T., LISI, F., (2013), *Serie storiche economiche: analisi statistiche e applicazioni*. Roma: Carrocci editore.
- FRANCES, P. H., (2016), *A note on the Mean Absolute Scaled Error*. International Journal of Forecasting, Vol. 32, 20-22
- GNEITING, T., (2011), *Making and Evaluating point forecasts*. Disponibile su: <<https://arxiv.org/pdf/0912.0902.pdf>> [Data di accesso: 08/11/2016]
- HYNDMAN, R. J., (2014), *Measuring forecast accuracy*. Disponibile su: <<http://www.robjhyndman.com/papers/forecast-accuracy.pdf>> [Data di accesso: 20/10/2016]
- HYNDMAN, R. J., BILLAH, B., (2003), *Unmasking the Theta method*. International Journal of Forecasting, Vol. 19, 287-290
- HYNDMAN, R. J., KOEHLER, A. B., (2006), *Another look at measures of forecast accuracy*, International Journal of Forecasting 22 (2006) 679-688

- MAKRIDAKIS, S., et al. (1982), *The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition*. Journal of Forecasting, Vol. 1, 111-153
- MAKRIDAKIS, S., et al. (1993), *The M2-Competition: a real-time judgmentally based forecasting study*. International Journal of Forecasting, Vol. 9, 5-22. Disponibile su: <https://www.researchgate.net/publication/222322115_The_M2-competition_A_real-time_judgmentally_based_forecasting_study>[Data di accesso: 29/09/2016]
- MAKRIDAKIS, S., HIBON, M., (1979), *Accuracy of forecasting: an empirical investigation*. Journal of the Royal Statistical Society. Wiley. Vol. 142, pp. 97-145
- MAKRIDAKIS, S., HIBON, M., (2000), *M3-Competition*. Disponibile su:<<http://sites.insead.edu/facultyresearch/research/doc.cfm?did=1094>> [Data di accesso: 30/08/2016]
- MAKRIDAKIS, S., HIBON, M., (2000), *The M3-Competition: results, conclusions and implications*. Francia: International Journal of Forecasting 16 (2000) 451-476
- MAKRIDAKIS, S., HIBON, M., (2000), *The M3-Competition: results, conclusions and implications*. Journal of Forecasting, Vol. 16, 451-476
- NAU, R., (2014), *Forecasting with moving averages*. Disponibile su <http://people.duke.edu/~rnau/Notes_on_forecasting_with_moving_averages--Robert_Nau.pdf> [Data di accesso: 29/09/2016]
- TOFALLIS, C., (2015), *A better measure of relative prediction accuracy for model selection and model estimation*. Journal of the Operational Research Society, Vol. 66, 1352-1362
- WOOLDRIDGE, J. M., (2006), *Introductory Econometrics: A Modern Approach*. 4° ed. Mason: South-western Cengage learning. pp. 339-376

Numero parole: 10.387