1 Introduzione

Il sito

$$http://it.finance.yahoo.com/q/bc?s=BUL.MI$$

permette di scaricare i dati relativi a tutti i titoli in borsa. E' la pagina di *Bulgari*, che useremo come esempio, ma da questa si può risalire a qualunque altro titolo (*Alitalia*, *Telecom* ecc.). Sotto *Quotazioni storiche* si trovano i dati numerici. Carichiamo su un foglio Excel le quotazioni a partire dal 1 Gennaio 2004. Alla data odierna ci sono i dati fino al 12 Ottobre 2006. Isoliamo la prima colonna, quella con i dati di *Apertura*, salviamola come *Bulgari.txt* e carichiamola da R col nome B. Come sempre costruiamo un file, es. *comandi_ts.txt*, con i comandi che useremo.

Conviene avere un simbolo breve per la lunghezza della serie. Inoltre, conviene rinominare la serie per evitare il formato tabella. I due comandi sono: *lenB<-length(B[,1]), B<-B[1:lenB,1].* Possiamo vedere la serie con *ts.plot(B)* ma suggeriamo fortemente di non farlo, per la ragione che si capirà tra un istante.

Nell'apprendere i metodi di previsione ed anche nel provarli si usa effettuare la seguente ovvia finzione: eliminarne un blocco finale, far finta di avere solo quelli rimasti, e con quelli provare a prevedere il blocco eliminato. Per serie di pochi dati, es. mensili o annuali, in genere si desidera effettuare previsioni di una o poche unità temporali, sia per l'entità temporale di ogni unità, sia per l'importanza di ogni singolo dato. Invece, per serie temporali come *Bulgari.txt*, non serve porre troppa attenzione sulla previsione del prossimo valore giornaliero, mentre serve prevedere l'andamento futuro per un lasso di tempo di vari giorni, es. 15 o 30. Una fluttuazione giornaliera ha un'importanza relativa, quello che conta è l'accumulo relativo ad alcuni giorni. Per questa ragione, eliminiamo ad esempio 30 giorni, costruendo la serie $B0$:

lag <- 30
B0 <- B[1:(lenB-lag)]
lenB0<-length(B0)
ts.plot(B0)

Possiamo anche tracciare in sovrapposizione l'ultimo tratto da prevedere, come linea tratteggiata. Suggeriamo di non farlo! Però se lo si desidera, ecco il metodo. Forse ci sono comandi più brevi, comunque il seguente funziona:

B<-as.ts(B)

B0<-as.ts(B0)
ts.plot(B0,B, lty=c(1,2))

Purtroppo, sapere già com'è fatta la parte da prevedere falsa molto la nostra percezione psicologica del problema. Teniamo comunque presente che durante lo svolgimento del corso diventeranno noti i valori in almeno altri 15 giorni circa, quindi potremo testare la la nostra abilità su dati veramente sconosciuti.

Che fare per prevedere B da $B0$? Apprenderemo vari metodi automatici, però nulla vale quanto un'analisi intuitiva e creativa, con l'immaginazione di alcune metodologie fatte in casa, personalizzate e adattate all'esempio specifico. Che faremmo allora, senza conoscere metodi automatici, sulla serie $B0$?

Analizziamo le nostre impressioni.

- Giornalmente la serie varia con ampia casualità: basta ad es. tracciare trenta valori *ts.plot(B0[(lenB0-30):lenB0]).*
- C'è comunque, nella parte finale di $B0$ (ultime decine di valori) una generale tendenza alla crescita.
- Guardando però il grafico dal 2004, ci sono stati lunghi periodi sia di crescita sia di calo, quindi nulla esclude che possa sopraggiungere uno di questi periodi e che quindi la previsione a 30 giorni di B da $B0$ possa andare in calo.
- Però c'è una certa, se pur vaga, regolarità di oscillazioni tra cali e crescite e sembra che al termine di $B0$ ci troviamo nel mezzo di una fase di crescita.
- Infine, l'ampiezza delle oscillazioni su larga scala aumenta nel tempo, per lo meno se guardiamo le ultime tre grandi oscillazioni. E' un fenomeno simile alla *stagionalità moltiplicativa* di cui parleremo poi, anche se è discutibile parlare qui di stagionalità nell'usuale senso del termine.

In conclusione, l'intuito porta a pensare che nei prossimi 30 giorni ci sia un'ulteriore crescita. Cerchiamo di quantificarla. Un'idea che sembra essere ragionevole è quella di riprodurre il disegno oscillatorio delle ultime tre oscillazioni, opportunamente ampliato in modo moltiplicativo. Inoltre,

se guardiamo il grafico globale dal 2004, vediamo che la crescita generale presenta però un lieve grado di saturazione, cioè la pendenza positiva cala lievemente. Potremmo quindi cercare di costruire un grafico dopo $B0$ che rispecchi questi due elementi:

a) le oscillazioni amplificate moltiplicativamente,

b) un trend crescente medio con un lieve calo.

Come catturare questi due elementi, cioè il trend (col suo lieve calo) e la periodicità (moltiplicativa)? Vedremo che il SW offre dei comandi automatici, che però non sempre danno i risultati migliori. Lavoriamo a mano, con un po' di ingegno.

Ci sono vari modi naturali di identificare il trend. Uno ovvio è trovare la retta di regressione:

```
x<-1:lenB0
reg <- lm(B0 ~x)
abline(reg$coefficient)
```

Il difetto è che mescola zone troppo lontane nel tempo (il 2004) e non permette di cogliere la lieve saturazione recente. Si può ad esempio decidere di usare i dati dal 230 (scelto ad occhio) in poi:

```
reg <- lm(B0[230:lenB0] ~x[230:lenB0])
abline(reg$coefficient)
```

La pendenza è cambiata forse troppo. Possiamo cercare delle vie di mezzo. Questo però è un buon esempio dell'indecisione che si incontra ogni volta nella pratica: non c'è nessun criterio (almeno immediato) che dica su quale finestra calcolare la regressione. In fin dei conti, se aggiustiamo ad occhio la finestra fino ad ottenere un risultato che ci piace visivamente, di fatto stiamo sostituendo il metodo numerico di regressione col tracciare *ad occhio* una retta che ci sembri rappresentativa!

Nonostante questi giusti commenti, per ora prendiamo la finestra da 200 in poi, che aggiusta un po' le cose (ma è comunque una scelta soggettiva!):

```
reg <- lm(B0[200:lenB0] ~x[200:lenB0])
abline(reg$coefficient)
```

Eventualmente puliamo il disegno: chiudiamo la finestra grafica e ritracciamo:

```
ts.plot(B0)
abline(reg$coefficient)
```

La retta di regressione non esaurisce le metodologie con cui cercare un trend. Ovviamente potremmo cercare una curva diversa da una retta; Proviamo con un polinomio di secondo grado:

reg2 <- lm(B0 ~x+I(x^2))
summary(reg2)

Dagli indicatori dei test vediamo che tutti i coefficienti sono estremamente importanti (valori p molto piccoli, tre stelle). Quindi il modello quadratico è certamente meglio di quello lineare e di nessun modello. Si noti che i coefficienti trovati sono molto piccoli, ma questo non significa che siano poco rilevanti (se standardizzassimo le variabili, ci accorgeremmo che la loro piccolezza è dovuta solo ad una questione di unità di misura). Tracciamo la parabola in sovrapposizione alla serie storica:

*y<-reg2$coefficient[1]+reg2$coefficient[2]*x+reg2$coefficient[3]*x^2*
ts.plot(B0,y)

Purtroppo vediamo che la parabola già decresce nella parte finale, cosa che contraddice la nostra intuizione (quella che il trend resti positivo, pur calando). E' difficile commentare questo insuccesso, forse inatteso. Scartiamo questo metodo relativamente a questo esempio. Però può essere servito a lasciare il dubbio che davvero il titolo in questione abbia iniziato una fase strutturalmente discendente. In altre parole, mentre la percezione grafica delle oscillazioni ci fa attendere una crescita, il fit globale manifesta una fase di calo.

2 La serie senza trend

Come prima approssimazione della nostra ricerca di un trend, accettiamo la retta ottenuta dal punto 200 in poi. Vogliamo eliminarla, cioè ricondurci ad una situazione senza trend, in modo che sia più semplice analizzare le oscillazioni.

Ci sono due modi di eliminarla. Il primo è *sottrarla.* Se ci aspettiamo che dopo la sottrazione la serie manifesti delle oscillazioni regolari, stiamo presupponendo un modello del tipo

$$d = (at + b) + F(t) + \varepsilon(t)$$

dove $F(t)$ è una funzione periodica ed $\varepsilon(t)$ descrive gli errori o le oscillazioni su piccola scala non incluse nel modello a grande scala. Proviamo con il nostro esempio:

reg <- lm(B0[200:lenB0] ~x[200:lenB0])
*y <- B0-reg$coefficient[1]-reg$coefficient[2]*x*
ts.plot(y)

La parte iniziale in salita non deve ingannare: la retta di regressione è stata scelta per rappresentare il trend nella zona delle tre oscillazioni finali, e lì in effetti il grafico sembra abbastanza orizzontale in media:

ts.plot(y[225:lenB0])

Però è chiaro che il grafico non è quello di una funzione periodica, bensì di una periodica moltiplicata per una funzione crescente (le oscillazioni aumentano di ampiezza). Allora il modello che somiglia ai nostri dati può essere (diciamo da 200 in poi)

$$d = (at + b) + F(t)c\,(t) + \varepsilon\,(t)$$

dove $F(t)$ è una funzione periodica e $c\,(t)$ è una funzione crescente che amplifica le oscillazioni.

Remark 1 *Un'osservazione a margine: vedendo questo grafico raddrizzato, non è poi così assurdo pensare che dopo le tre oscillazioni di ampiezza crescente possa essere cominciata una nuova fase, di calo medio, diciamo simmetrica rispetto a quella iniziale (del 2004) di crescita media. Potremmo davvero immaginare che il grafico, alla fine di $B0$, prenda a descrescere, oscillando. Quindi l'ipotesi che la parabola vista sopra sia realistica, non è proprio da buttar via. Per semplicità, rimaniamo nella convinzione che perduri una crescita media - eliminata dal raddrizzamento - alla quale sono sovrapposte delle oscillazioni di ampiezza sempre più grande.*

Come trovare $F(t)$ e $c\,(t)$? Possiamo ingegnarci a mano. Però è spontaneo fermarsi ed introdurre un modello alternativo, in cui le oscillazioni siano *moltiplicative*:

$$d = (at + b)\,F(t) + \varepsilon\,(t)\,.$$

In questo modello, se $a, b > 0$, moltiplichiamo le oscillazioni $F(t)$ per la funzione crescente positiva $(at + b)$, quindi le amplifichiamo all'aumentare del tempo. D'altra parte, se nel modello immaginiamo di sostituire $F(t)$ col suo valor medio, troviamo una retta, appunto la retta media che descrive il trend. Il vantaggio di questo modello è di avere meno parametri incogniti: non c'è $c(t)$. Questo però gli attribuisce meno flessibilità. Un altro vantaggio è l'interpretazione: la grandezza in esame ha una sua oscillazione naturale, ma questa è amplificata dal fatto che certe altre grandezze economiche in gioco aumentano progressivamente. Non che un'interpretazione di questo tipo sia assente per il modello con $c(t)$, solo che in quello è più complessa: ci

sono due meccanismi economici o sociali, per cui c'è una crescita media (il trend) ed una crescita delle oscillazioni.

Proviamo ad usare il modello moltiplicativo. Calcoliamo $F(t)$ dividendo per la retta:

*ym <- B0/(reg$coefficient[1]+reg$coefficient[2]*x)*
ts.plot(ym)
ts.plot(y)

Vediamo che un lieve vantaggio c'è, ma è lieve appunto, nel senso che permane l'amplificazione delle oscillazioni anche dopo la divisione. Il modello moltiplicativo coincide con quelli additivo modificato da $c(t) = (at + b)$. Evidentemente questa funzione $c(t)$ non basta, non è abbastanza crescente per eliminare l'aumento delle oscillazioni.

Torniamo allora al modello additivo corretto da $c(t)$ e (per trovare $F(t)$) dividiamo $y = d - (at + b)$ per una retta $c(t)$ cercata con una regressione lineare nel seguente modo. A mano, troviamo le coordinate dei max e min delle oscillazioni, ribaltiamo le oscillazioni negative, ottenendo quindi una successioni di punti nel primo quadrante, che crescono abbastanza linearmente. Di questi calcoliamo la retta di regressione e la prendiamo come $c(t)$. I punti cercati a mano sono:

281, 0.65
355, -0.81
411, 0.84
494, -0.77
568, 1.12
640, -1.11

Vengono poi ribaltati, messi nel file *estremi.txt*, caricati in R; poi si cerca la retta di regressione:

xx<-est$V1
yy<-est$V2
reg.est <- lm(yy ~xx)
abline(reg.est$coefficient)

il cui risultato grafico è gratificante. Infine si calcola $F(t)$

*F <- y[225:lenB0]/(reg.est$coefficient[1]+reg.est$coefficient[2]*x[225:lenB0])*
ts.plot(F)

trovando finalmente un grafico con una certa periodicità!

Purtroppo la fatica "a mano" non è finita. Questo grafico è solo vagamente periodico. Come estrarne una funzione periodica $F(t)$?

3 La parte periodica

Siamo arrivati ad una funzione, *y[225:lenB0]*, che sembra abbastanza periodica, salvo varie imprecisioni e fluttuazioni. Dobbiamo in qualche modo riassumerla tramite una vera funzione periodica $F(t)$.

Il primo problema è trovare il periodo. Questo non è identificato con precisione: abbiamo solo tre periodi e le fluttuazioni fanno sì che anche il periodo sia vago. Dopo un po' di prove si capisce che 145 può essere un discreto compromesso. In realtà, conviene eliminare un ulteriore pezzetto di F, riconducendoci ad una F lunga 150:

```
length(F)
F<-F[19:468]
ts.plot(F)
```

Questa sembra un po' meglio come periodicità. Ci sono dei comandi interessanti che aiutano a vedere 145:

```
acf(F,300)
ts.plot(diff(F,1))
ts.plot(diff(F,40))
ts.plot(diff(F,75))
ts.plot(diff(F,100))
ts.plot(diff(F,150))
ts.plot(diff(F,145))
```

però non li commentiamo ora in dettaglio, per non interrompere il discorso. Supponiamo quindi che abbiamo scelto 145 in qualche modo, ad occhio oppure con l'aiuto di qualche calcolo.

Scelto 145, sovrapponiamo le tre oscillazioni:

```
per<-145
P1<-F[1:per]
P2<-F[(per+1):(2*per)]
P3<-F[(2*per+1):(3*per)]
P1<-as.ts(P1)
P2<-as.ts(P2)
P3<-as.ts(P3)
ts.plot(P1,P2,P3,lty=c(1,2,3))
```

Il risultato non è ottimo ma nemmeno pessimo. Ora si tratta di trovare una funzione che passi vicino a tutti questi punti. Ecco tre idee facili.

- Cercare una spezzata a triangolo che interpoli P1, P2, P3 (ottenuta con 2 regressioni). Non portiamo avanti questa idea.

- calcolare la curva che media P1, P2, P3:

  Pmean <- (P1+P2+P3)/3

  Pmean <-as.ts(Pmean)

  ts.plot(Pmean,P1,P2,P3,lty=c(1,2,3,4))

  ts.plot(Pmean)

  Il risultato è una curva rappresentativa dei dati, ma quasi altrettanto irregolare. La potremmo usare come $F(t)$, però è naturale pensare che la $F(t)$ del modello debba essere una curva regolare, e che le fluttuazioni che vediamo in *Pmean* siano dovute al termine $\varepsilon(t)$ del modello.

- Possiamo cercare una curva localmente polinomiale che interpoli *P1, P2, P3* con la funzione "Loess"; possiamo usare *Pmean* se lo abbiamo già calcolato:

  xxx<-1:per

  loe<-loess(Pmean~xxx)

  F.loe<-predict(loe, data.frame(time = seq(1, per, 1)), se = TRUE)

  ts.plot(F.loe$fit)

Il risultato è ottimo. Tra parentesi, sui dati grezzi y si poteva notare una certa asimmetria rispetto all'asse delle ascisse, plottando $-y$; questa asimmetria è perfettamente colta dall'interpolata loess.

A questo punto bisogna *rimettere tutto insieme.* Non scordiamoci che abbiamo preso come periodo 145, quindi 3 volte loess non arriva in fondo ai nostri dati, ma arriva a 15 dati prima. Suggeriamo di provare a mettere tutto insieme, predire B da $B0$, ma trattenersi dal visualizzare B per capire se il risultato è buono. Sarebbe molto gratificante, ma annullerebbe l'aspetto psicologico delle domande che ora ci porremo:

- Una volta visualizzata la previsione ed i dati veri, come giudicheremo se la previsione era buona? Infatti la previsione non coinciderà mai coi dati, quindi bisogna valutare se lo scarto è piccolo o grande.

- Più semplice è il seguente problema: se sviluppiamo un secondo metodo di previsione, altrettanto ragionevole ma diverso, come giudicheremo quale sia il migliore una volta noti i dati veri?

- C'è invece la versione più difficile di questo problema (alla cieca dei dati veri): una volta sviluppati altri metodi di previsione, prima di conoscere i dati veri, possiamo espriemere un giudizio a priori su quale ci aspettiamo sia il migliore?

- La versione alla cieca del primo problema è: ponendo l'attenzione su un singolo metodo, prima di conoscere i dati veri, possiamo dichiararne le prestazioni?

- Infine osserviamo che sarebbe utile automatizzare alcune delle idee sviluppate nel metodo sopra descritto. Si possono costruire metodi universali, automatici, che rispecchino queste idee?

In una frase, le prime quattro domande si chiedono: prima di conoscere i dati, le nostre previsioni saranno state fatte bene? E dopo, le giudichiamo buone o no? La prima domanda, alla cieca dei dati, è soprattutto impellente se ci si cala nella realtà, in cui sbagliare le previsioni abbia un costo importante. Per questo suggeriamo di non vedere i dati veri e subire l'effetto dell'incertezza, per sforzarsi di arrivare a metodi e dichiarazioni che ci convincano in una situazione "al buio".