Previsioni

Introduzione

09/01/2007 7.30

Indice

- concetti base
- modelli causali
- serie temporali
- errori
- serie stazionarie
- serie con trend
- serie con stagionalità

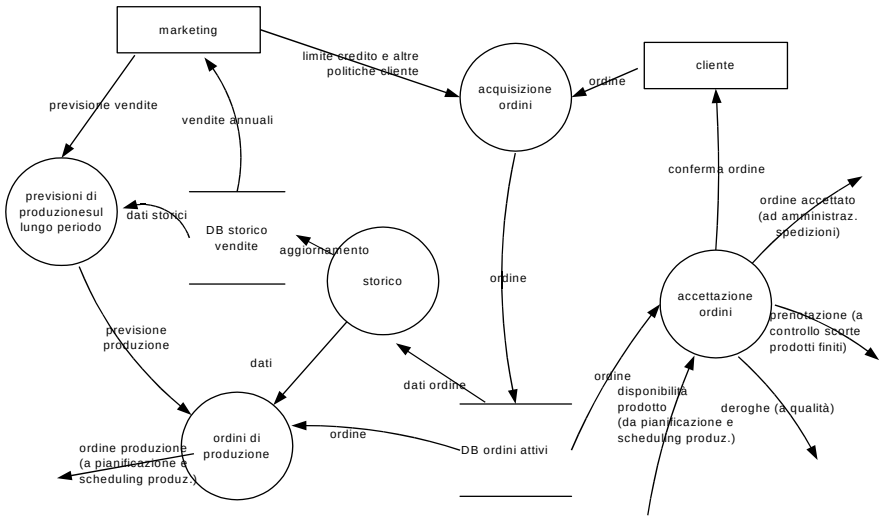
Previsioni

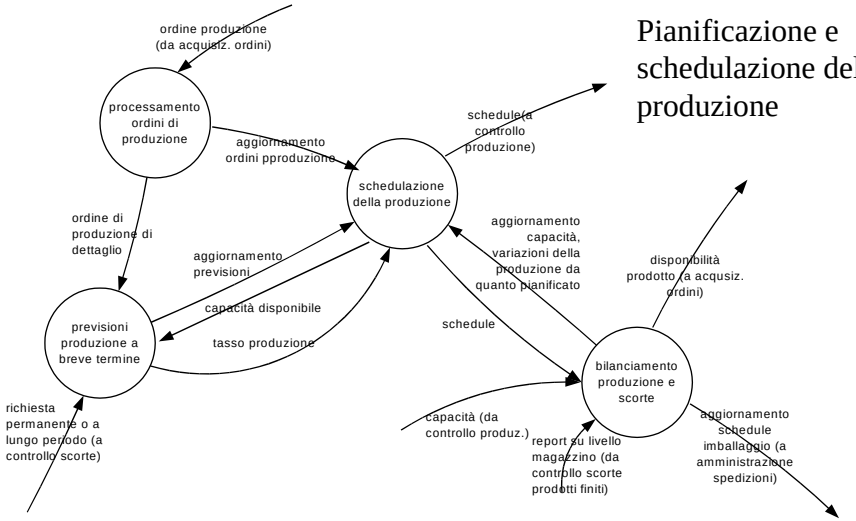
La capacità di prevedere (forecasting) il futuro è fondamentale per un'azienda.

Bisogna prevedere:

- la domanda dei clienti (marketing).
- le esigenze della produzione (logistica).

Acquisizione ed elaborazione ordini





Pianificazione e schedulazione della produzione

Orizzonte

Si fanno previsioni

- a breve termine (giorni)
  - gestione operativa risorse/scorte/personale
  - vendite a breve
- a medio termine (mesi)
  - oscillazioni stagionali
- a lungo termine
  - decisioni strategiche

Proprietà

- Le previsioni sono caratterizzate da:
  - valore atteso
  - varianza/*range*
- sono migliori se aggregate
- sono poco accurate sul lungo periodo
- devono essere supportate da tutte le informazioni disponibili a prezzo ragionevole (anche se non trattabili matematicamente)
- possono essere
  - soggettive
  - oggettive

Previsioni soggettive

fatte da

- riunioni di esperti
- questionari ad esperti
- indagini di mercato

Previsioni oggettive

Modelli

- causali
- serie temporali

Modelli causali / econometrici

Usati quando è noto che la grandezza che si vuole conoscere Y è correlato a grandezze osservabili correntemente $X_1,..., X_n$.

- Attraverso dati storici si identifica la funzione f:

$$Y=f(X_1,..., X_n)$$

- si ricercano i valori di $X_1,..., X_n$ e si deduce Y

Modelli causali

$f(X_1,..., X_n)$ è, in generale, una funzione lineare o logaritmica i cui parametri sono determinati attraverso il metodo dei minimi quadrati.

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_n X_n$$

oppure

$$lnY = \alpha_0 + \alpha_1 lnX_1 + \alpha_2 lnX_2 + ... + \alpha_n lnX_n$$

Minimi quadrati

- Nel caso di modelli lineari f è indicata come regressione lineare.
- Metodi specializzati per gestire:
  - l'eteroschedasticità
  - l'instabilità numerica
  - l'acquisizione di nuovi dati.

Correlazione

Si verifica inizialmente che esista una correlazione tra i dati:

- siano date *N* coppie di realizzazioni, $(X_1, Y_1)\ldots, (X_n, Y_n)$, di *X* e *Y*, esiste una correlazione se

$$cov(x,y)=E\{(X_i-\mu_x)(Y_i-\mu_y)\} = E\{X_iY_i\}- \mu_x \mu_y \neq 0$$

- se *X* e *Y* sono correlati, *Y* può essere espresso in funzione di *X* e di una variabile aleatoria indipendente *e*

$$Y=h(X,E)$$

Correlazione

- Nel caso di relazioni lineari tra due variabili casuali distribuite normalmente (*distribuzione normale bivariata*) si definisce il coefficiente di correlazione campionario *r* (stima di ρ) come:

$$r = \frac{\sum (X_i - M_X)(Y_i - M_Y)}{\sqrt{\sum (X_i - M_X)^2 \sum (Y_i - M_Y)^2}}$$

- si verifica che nell'ipotesi nulla di non correlazione H_0: $\rho = 0$

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

- segue la distribuzione di student con *N-2* gradi di libertà.

Correlazione

Misura ordinaria di correlazione è la

correlazione tra ranghi o di Spearman

$$Corr.di\,Spearman = S = \frac{\sum_i (r(X_i)-\bar{r}(X))(r(Y_i)-\bar{r}(Y))}{\sqrt{\sum_i (r(X_i)-\bar{r}(X))^2 \sum_i (r(Y_i)-\bar{r}(Y))^2}}$$

se dati incorrelati $E\{S\} = 0$ e $var\{S\} = 1/(n-1)$

Regressione lineare

- quando si usano modelli di regressione lineare, si suppone:

$$X_i : \text{indipendente}$$

$$Y_i = a + bX_i + e_i$$

- dove $e_i \sim n(0, \sigma_e^2)$

Regressione lineare

- gli stimatori dei parametri sono

$$S_E^2 = \frac{1}{N-2}\left(\sum (Y_i - M_Y)^2 - \hat{b}\sum Y_i (X_i - M_X)\right)$$

$$\hat{a} = M_Y - \hat{b} M_X \qquad \hat{b} = \frac{\sum Y_i (X_i - M_X)}{\sum (X_i - M_X)^2}$$

che si ottengono minimizzando la funzione:

$$g(\hat{a},\hat{b}) = \sum e_i^2 = \sum (Y_i - (\hat{a} + \hat{b} X_i))^2$$

Correlazione: generalizzazioni

- **Correlazione multipla**: una variabile può essere espressa in termini di più di una altra variabile casuale.
- **Regressione non lineare**: una variabile può essere espressa in termini di una relazione non lineare con un'altra variabile casuale. In questo caso r, che esprime quanto è forte la dipendenza, risulta essere t.c.:

$$r^2 = \frac{\text{varianza spiegata}}{\text{varianza totale}} = \frac{\sum (Y_{i,stimato} - M_Y)^2}{\sum (Y_i - M_Y)^2}$$

NB: con sufficienti gradi di libertà si spiega qualunque cosa, ma la stima dei parametri diventa assolutamente inaffidabile. Se non ci sono giustificati motivi conviene sempre usare modelli semplici.

Correlazione: generalizzazioni

- **Multivarianza**: un vettore $\boldsymbol{X}$ di variabili casuali correlate espresso in funzione di variabili indipendenti $\boldsymbol{e}$.

  Se è dato il vettore $\boldsymbol{X}$ di variabili aleatorie normali correlate con media $\boldsymbol{\mu}$ e matrice di covarianza Σ. $\boldsymbol{X}$ può essere espresso come

  $$\boldsymbol{X} = \boldsymbol{\mu} + C\boldsymbol{E}$$

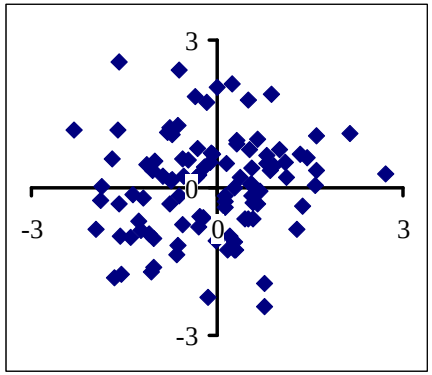
  dove
  - $\boldsymbol{e} \sim \boldsymbol{n}(0,1)$ e
  - C è una matrice triangolare inferiore di Cholesky, t.c.

    $$CC^T = \Sigma .$$

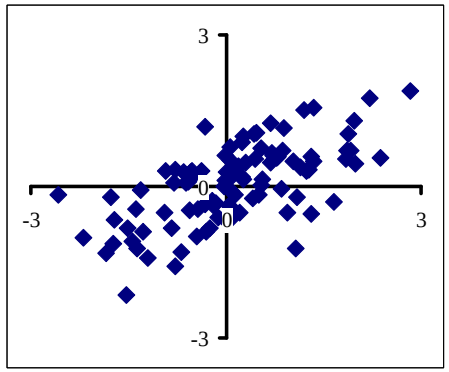
Commenti

- Esistono altre regressioni:
  - regressione quadratica
  - regressione esponenziale
  - reti neurali
  - ....
- usare la regressione più semplice possibile, tra quelle ragionevoli (sempre graficare prima i dati per avere un'idea): più parametri si devono identificare, più il risultato si adatta ai dati passati, ma richiede un maggior numero di informazioni per avere lo stesso errore sui dati futuri.

Diagrammi a scattering (100 dati)



non correlati



correlati

Commenti

- se non si riesce ad avere una stima statistica dell'errore, suddividere i dati in insieme di "addestramento", con cui stimare i parametri, ed insieme di verifica, con cui controllare la correttezza delle previsioni.

  Non correggere mai i parametri alla luce dei risultati dell'insieme di verifica, altrimenti li si adatta ai dati di tale insieme, ma non si ha nessuna garanzia di previsioni migliori.

  Generare altri due insiemi di addestramento e di verifica più grandi ed eseguire nuove stime.

Serie temporali

Usate quando si ritiene che l'andamento dei valori del passato si mantenga nel futuro.

I valori passati sono, in genere, campionati ad intervalli regolari.

Le serie vengono espresse come composizione di andamenti "regolari":

- trend
- stagionalità
- ciclicità
- pura casualità

Serie temporali

L'ipotesi base nei modelli di previsione è che i dati siano autocorrelati e che quindi l'andamento (in media, trend, ecc..) dei dati temporalmente vicini fornisca una stima migliore che le stesse statistiche calcolate rispetto a tutti i dati della serie.

Questa ipotesi è, per motivi di semplicità analitica, dimenticata nelle dimostrazioni.

Simbologia

- D_t : dato osservato nel periodo t;
- F_{t+1} : previsione per il periodo $t+1$, in generale:

$$F_t = \sum_i a_i D_{t-i}$$

- e_t : errore tra previsione e osservazione del il periodo t:

$$e_t = F_t - D_t$$

Accuratezza della previsione

misure di accuratezza della previsione:

- deviazione assoluta media: $MAD = E\{|e_t|\}$
- errore quadratico medio: $MSE = E\{e_t^2\}$
- errore percentuale assoluto: $MAPE = E\{|e_t/D_t|\}$ %

Proprietà

- *MAD*: se errore distribuito normalmente, la deviazione standard errore σ è t.c.

$$\sigma = 1.25 \text{ MAD}$$

- *MSE*: più facilmente trattabile in modo analiticamente perché derivabile.
- l'errore deve essere non deviato (a media nulla), per cui

$$\sum_t e_t$$

deve avere valore atteso nullo e in valore assoluto non crescere più velocemente della radice quadrata del numero degli addendi (*random walk*)

Serie stazionarie

Ipotesi:

$$D_t = \mu + \varepsilon_t$$

μ : costante incognita

ε_t : disturbo di media zero e deviazione standard σ

Media mobile

F_t : media ultimi N valori osservati

$$F_t = (D_{t-1} + D_{t-2} + \ldots + D_{t-N})/N = MA(N)$$

potrebbe essere anche pesata

Il valore di N permette di regolare la sensibilità della media rispetto a fluttuazioni dei valori della serie osservata: N deve essere sufficientemente grande per non risentire di disturbi casuali, ma sufficientemente piccolo per accorgersi di fluttuazioni stagionali.

Proprietà statistiche

Previsioni non deviate

$$E\{F_t - D_t\} = \frac{1}{N}\sum_{i=1}^{N} E\{D_{t-i}\} - E\{D_t\} = \frac{N\mu}{N} - \mu = 0$$

$$Var\{F_t - D_t\} = Var\{F_t\} + Var\{D_t\} = \frac{1}{N^2}\sum_{i=1}^{N} Var\{D_{t-i}\} + Var\{D_t\} = \sigma^2\frac{N+1}{N} = \sigma_e^2$$

La varianza osservata in realtà è minore quando, come è auspicabile, D_t e F_t sono correlati positivamente:

$$Var\{D_t - F_t\} = Var\{D_t\} + Var\{F_t\} - 2Covar\{D_t F_t\}$$

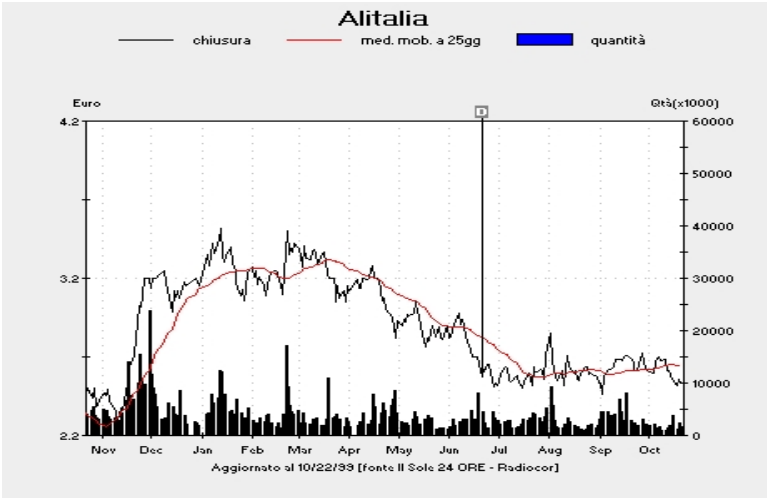
Vantaggi / Svantaggi

- **vantaggi**
  - semplice comprensione, facile da usare, molto diffusa,
  - di facile aggiornamento

$$F_{t+1} = F_t + (D_t - D_{t-N})/N$$

- **svantaggi**
  - sottostima trend positivi, sovrastima trend negativi (presenta un ritardo di fase),
  - si devono ricordare i dati di N intervalli precedenti.

Esempio



Livellamento esponenziale

F_t : correzione della previsione precedente alla luce della nuova osservazione

$$F_t = (1-\alpha)F_{t-1} + \alpha D_{t-1} \qquad 0 < \alpha \leq 1$$

Il valore di α permette di regolare la sensibilità del valore previsto rispetto a fluttuazioni dei valori della serie osservata: α deve essere sufficientemente piccolo per non risentire di disturbi casuali, ma sufficientemente grande per accorgersi di fluttuazioni stagionali (valori generalmente usati $0.1 \leq \alpha \leq 0.2$)

Proprietà statistiche

- Espressioni alternative

$$F_t = (1-\alpha)F_{t-1} + \alpha D_{t-1} = F_{t-1} - \alpha(F_{t-1} - D_{t-1}) = F_{t-1} - \alpha e_{t-1}$$

$$F_t = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i D_{t-i-1}$$

Il valore di α permette di regolare il peso dei valori lontani nel tempo della serie osservata: influenza di questi è tanto minore tanto più α è grande.

Proprietà statistiche

Si osservi che

$$\sum_{i=0}^{\infty} \alpha(1-\alpha)^i = 1 \qquad \sum_{i=0}^{\infty} \alpha(1-\alpha)^{2i} = \frac{\alpha}{1-(1-\alpha)^2}$$

quindi, si hanno previsioni non deviate

$$E\{F_t\} = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i E\{D_{t-i-1}\} = \mu \sum_{i=0}^{\infty} \alpha(1-\alpha)^i = \mu$$

$$Var\{F_t - D_t\} = Var\{F_t\} + Var\{D_t\} = \sigma^2 \frac{\alpha}{2-\alpha} + \sigma^2 = \sigma^2 \frac{2}{2-\alpha} = \sigma_e^2$$

Vantaggi / Svantaggi

- **vantaggi**
  - semplice comprensione, facile da usare, molto diffusa,
  - si devono ricordare solo i dati del periodo precedente;
- **svantaggi**
  - sottostima trend positivi, sovrastima trend negativi,
  - "dimentica" lentamente possibili *outlier**.

*dato non rappresentativo della serie temporale (tipicamente dovuto ad eventi non ripetibili)

Commento

Se si confrontano le “età” dei dati utilizzati nella media mobile e nello smorzamento esponenziale si ottiene

- media mobile

$$(1 + 2 + \ldots + N)/N = (N + 1)/2$$

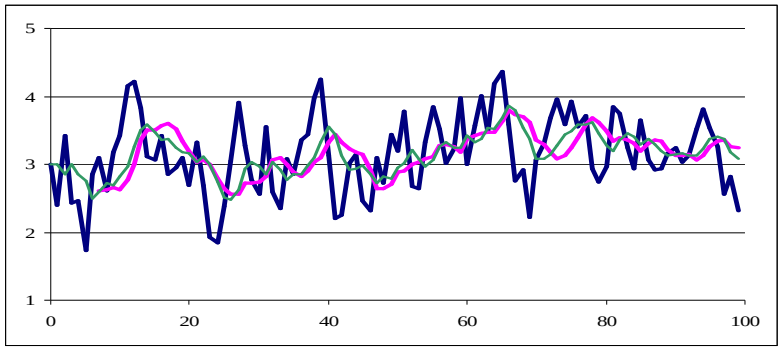
- livellamento esponenziale

$$\sum_{i=0}^{\infty} i\alpha(1-\alpha)^i = \frac{1}{\alpha}$$

Ponendo $\alpha = 2/(N+1)$ si ottengono serie con la stessa distribuzione dell’errore di previsione (se questo è normalmente distribuito). Le previsioni saranno però in generale diverse.

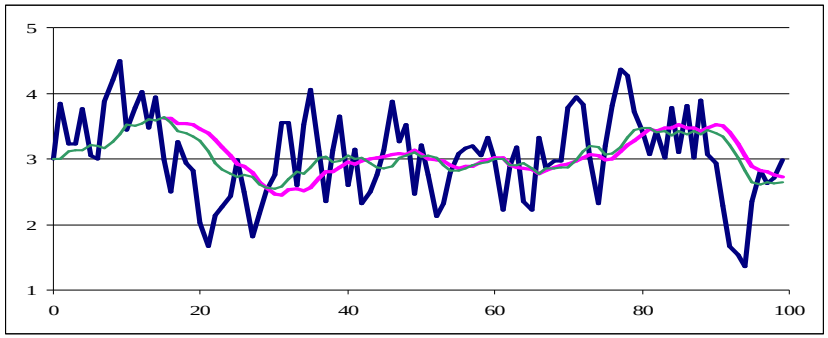
Esempio

media mobile in rosso $N = 7$
livellamento esponenziale in verde $\alpha = 0.25$



Esempio

media mobile in rosso $N = 15$
livellamento esponenziale in verde $\alpha = 0.125$



Esempio

media mobile in rosso $N = 15$
livellamento esponenziale in verde $\alpha = 0.125$