



**StatSoft®**

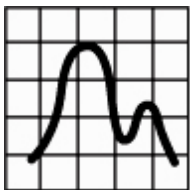
*data analysis • data mining • quality control • web-based analytics*

# **Previsione della Domanda**

**e**

# **STATISTICA**

## **Case Study: Domanda del Volume di Carburante**



### **STATISTICA**

**Soluzioni per Business Intelligence,  
Data Mining, Quality Control, e  
Web-based Analytics**

**StatSoft Italia srl • via Parenzo 3, 35010 • Vigonza (PD) • 049 8934654 • Fax: 049 8932897 • info@statsoft.it • www.statsoft.it**

Australia: StatSoft Pacific Pty Ltd.  
Brazil: StatSoft South America  
Bulgaria: StatSoft Bulgaria Ltd.  
Czech Rep.: StatSoft Czech Rep. s.r.o  
China: StatSoft China

France: StatSoft France  
Germany: StatSoft GmbH  
Hungary: StatSoft Hungary Ltd.  
India: StatSoft India Pvt. Ltd.  
Israel: StatSoft Israel Ltd.

Italy: StatSoft Italia srl  
Japan: StatSoft Japan Inc.  
Korea: StatSoft Korea  
Netherlands: StatSoft Benelux BV  
Norway: StatSoft Norway AS

Poland: StatSoft Polska Sp. z.o.o.  
Portugal: StatSoft Ibérica Lda  
Russia: StatSoft Russia  
Spain: StatSoft Ibérica Lda

S. Africa: StatSoft S. Africa (Pty) Ltd.  
Sweden: StatSoft Scandinavia AB  
Taiwan: StatSoft Taiwan  
UK: StatSoft Ltd.

## Tabella dei Contenuti

**INTRODUZIONE: IN COSA CONSISTE LA PREVISIONE DELLA  
DOMANDA?..... 3**

**CASE STUDY: DOMANDA DEL VOLUME DI CARBURANTE..... 3**

Descrizione..... 3

**ANALISI DEI DATI CON STATISTICA ..... 4**

Selezione delle Caratteristiche ..... 4

Spazio di Lavoro di *STATISTICA Data Miner* ..... 5

Risultati di Modelli Lineari Generali e Spline MAR..... 6

**CONCLUSIONI..... 8**

## Introduzione: In cosa consiste la Previsione della Domanda?

Chiunque operi nel settore delle forniture di beni conosce bene l'importanza rappresentata dalla previsione della domanda. Conoscere quali saranno le fluttuazioni delle richieste di un bene consente ad un fornitore di programmare la produzione e di accumulare la quantità corretta di prodotti necessari a soddisfare la domanda. Se la domanda è sotto-stimata, si avrà come effetto la perdita di vendite derivante da una fornitura insufficiente. Se la domanda invece è sovra-stimata, il fornitore si troverà con un surplus di invenduto dal quale deriverà una certa perdita finanziaria. Prevedere l'andamento della domanda rappresenta quindi un elemento chiave per qualunque azienda desideri ottenere un vantaggio competitivo in qualsiasi mercato.

Per soddisfare i bisogni della clientela, la disponibilità di modelli previsionali appropriati è di vitale importanza. Nonostante nessun modello previsionale possa rivelarsi infallibile, il ricorso a specifici metodi di data mining spesso consente di evitare i costi derivanti da una sotto-stima o da una sovra-stima della domanda. Attraverso l'uso di queste tecniche, un'azienda potrà dirsi ben equipaggiata per soddisfare al meglio possibile la domanda dei propri clienti. *STATISTICA Data Miner* offre un'ampia gamma di strumenti specificamente studiati per la previsione della domanda.

## Case Study: Domanda del Volume di Carburante

### Descrizione

Le grandi catene distributive hanno bisogno di un modello previsionale accurato che consenta loro di determinare la domanda dei propri clienti. Sapere quale sarà il volume della domanda permetterà alle compagnie di disporre della quantità indispensabile di carburante, evitando quanto più possibile qualsiasi accumulo di surplus. L'accuratezza del modello predittivo ridurrà fortemente i costi evitabili associati alla sovra-stima o alla sotto-stima della domanda rispetto alle reali esigenze della clientela.

Le tecniche qui illustrate permetteranno di osservare come costruire un modello predittivo della domanda dei clienti attraverso *STATISTICA Data Miner*, al fine d'identificare gli input o i predittori che influenzano maggiormente l'acquisto di carburante nel particolare distributore rifornito dalla compagnia. Stime accurate consentiranno di essere quanto più certi possibile della domanda di fornitura.

### File Dati

Il file dati *store data.sta* contiene dati per ognuno dei distributori monitorati. Sono presenti 184 osservazioni e 158 variabili. Questi dati contengono informazioni sul tipo di distributore, sulla loro collocazione geografica, sulle tecniche di marketing impiegate, sul tipo di traffico osservato, sul numero di pompe di benzina, su altri beni e servizi offerti, ecc.

Per ogni distributore viene quindi registrata la domanda di carburante. Questa rappresenterà la variabile dipendente i cui valori saranno previsti attraverso *STATISTICA Data Miner*. Usando funzioni di selezione di caratteristiche, la lista di 158 variabili sarà condensata in un elenco di variabili più maneggevole ed appropriato.

## Analisi dei Dati con **STATISTICA**

Con *STATISTICA Data Miner*, è estremamente facile applicare ai dati potenti strumenti di modellazione e giudicare il valore dei modelli risultanti basati sui rispettivi valori predittivi e descrittivi. Ciò non diminuisce il ruolo assolutamente prezioso rappresentato dalla fase di preparazione preliminare dei dati. Dato infatti che le decisioni maggiormente strategiche saranno determinate da questi risultati, qualsiasi errore commesso in questa fase potrebbe comportare gravi perdite d'informazioni. Quindi, è decisamente importante riuscire a pre-elaborare i dati e migliorare l'accuratezza del modello in modo da poter raggiungere la migliore decisione possibile.

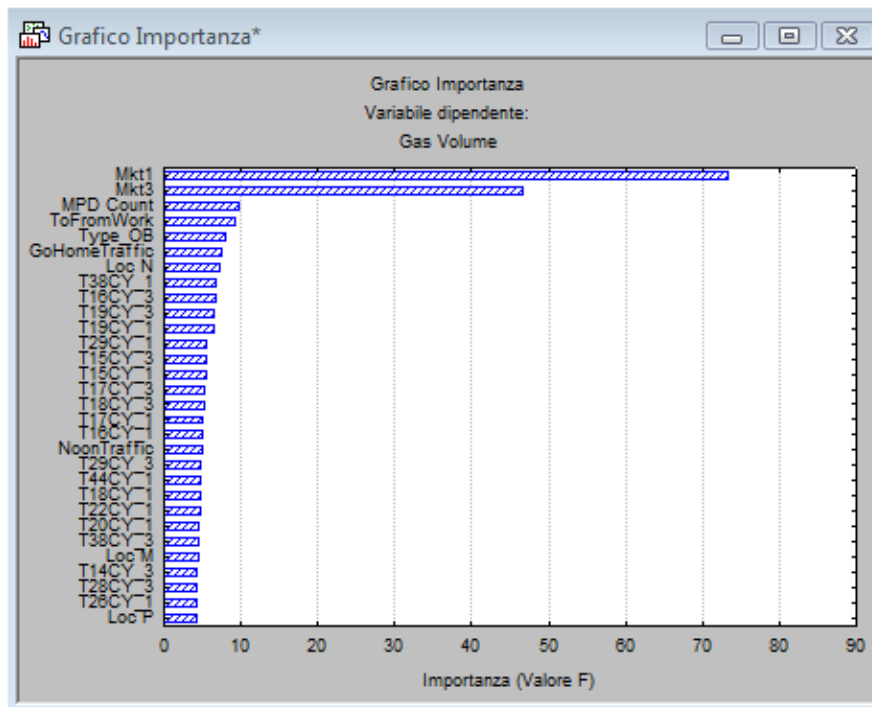
Durante questa fase sono state verificate le seguenti condizioni:

- Suggerimenti forniti dai dati: Statistiche descrittive (osservazione delle distribuzioni, delle medie, dei valori minimo e massimo, dei quartili, ecc.)
- Nei dati non sono presenti outlier
- Nei dati non sono presenti dati mancanti
- Non è richiesta alcuna trasformazione

## Selezione delle Caratteristiche

Per ridurre il livello di complessità del modello, l'insieme di dati può essere trasformato in un insieme di dati di dimensioni inferiori. Lo strumento di *Selezione di Caratteristiche e Screening di Variabili* disponibile in *STATISTICA Data Miner* ha automaticamente individuato i predittori più importanti che influenzano la domanda dei distributori.

Il grafico a barre e lo spreadsheet d'importanza dei predittori forniscono suggerimenti circa quali siano i predittori più efficaci per la previsione della variabile dipendente. Per esempio, di seguito è riportato il grafico a barre relativo all'importanza dei predittori per la previsione della variabile dipendente "Gas Volume".



I risultati della Selezione delle Caratteristiche dimostrano come le campagne di marketing Mkt1 e Mkt2 abbiano l'influenza maggiore sul Volume di Carburante. Un numero di altre variabili risultano essere predittori ugualmente significativi del Volume di Carburante.

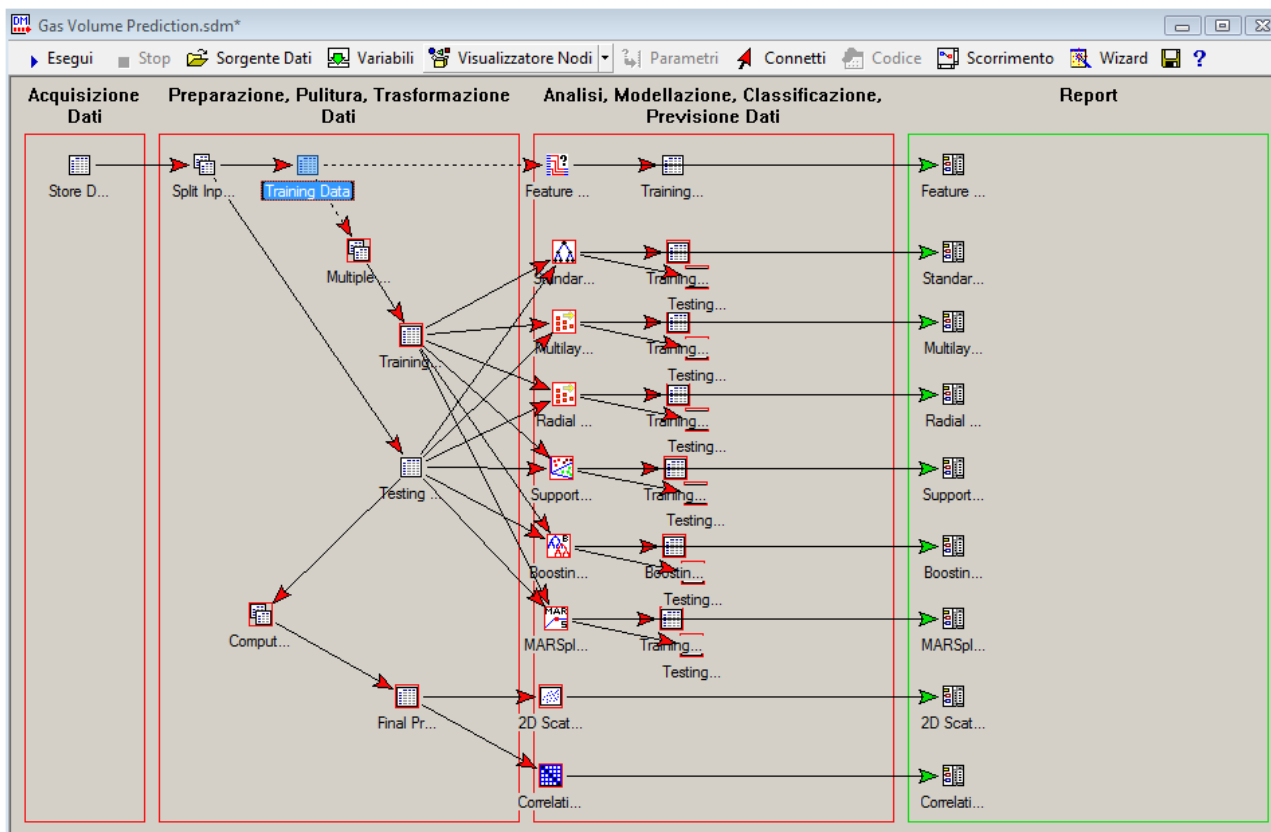
Questo esempio esamina i predittori attraverso i diversi algoritmi offerti in *STATISTICA Data Miner*. Questi algoritmi includono:

- *Modelli Lineari Generali (GLM)*
- *Alberi di Classificazione e di Regressione (C&RT)*
- *STATISTICA Reti Neurali Automatizzate (SANN)*
- *Support Vector Machine*
- *Alberi di Regressione Boosted*
- *Spline Adattabili di Regressione Multivariati*

La novità e l'abbondanza di tecniche e algoritmi disponibili nella fase di modellazione rendono questa fase la parte più interessante del processo di data mining. In più, è buona cosa praticare una serie di esperimenti applicando un numero di metodi differenti durante la fase di modellazione dei dati. Differenti tecniche potrebbero fare luce su un problema o confermare precedenti conclusioni.

## Spazio di Lavoro di *STATISTICA Data Miner*

Lo spazio di lavoro di *Data Miner* funziona sulla base di un flusso di analisi; tutti gli strumenti di *STATISTICA Data Miner* sono disponibili in forma di icona gestibili attraverso semplici operazioni di trascina-e-incolla.



Di seguito sono riportate le descrizioni delle diverse fasi di preparazione dei dati e di flusso delle analisi:

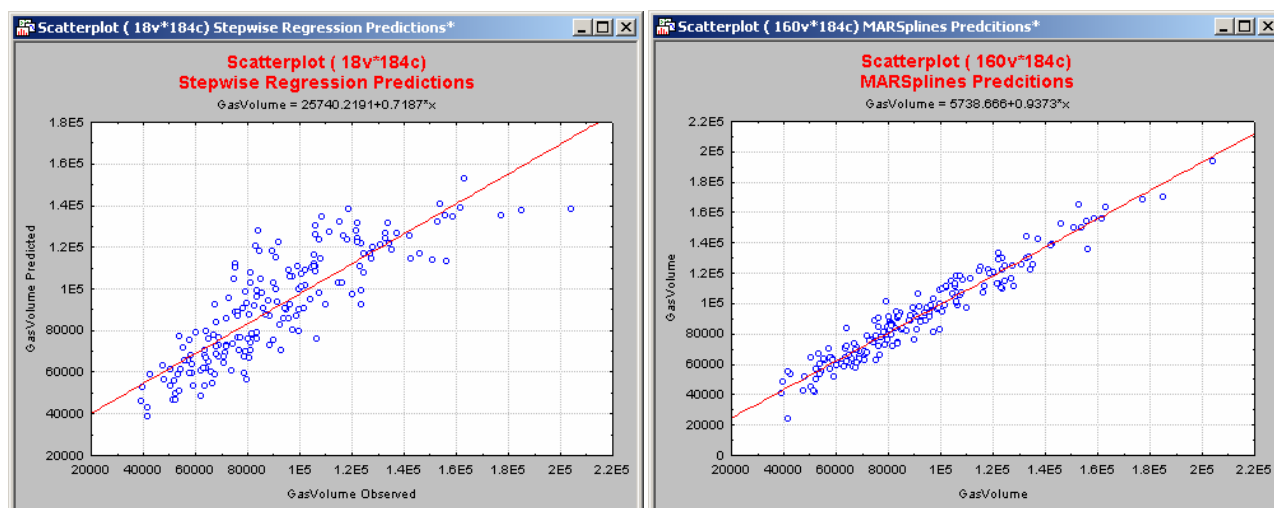
1. Suddivisione del file dati originale in due sottoinsiemi: il 30% dei casi viene trattenuto per scopi di verifica mentre il 70% dei casi viene utilizzato per la costruzione dei modelli.
2. Applicazione dello strumento di Selezione delle Caratteristiche per ordinare per importanza le migliori variabili predittrici da utilizzare per prevedere il volume di carburante.
3. Riduzione del numero di possibili predittori da 158 a 40 sulla base dei risultati della Selezione delle Caratteristiche.
4. Utilizzo di differenti Modelli Predittivi (algoritmi di Apprendimento Automatico) da utilizzare per individuare e comprendere eventuali relazioni esistenti.
5. Uso di scatterplot e di tabelle di correlazione per confrontare i valori osservati con quelli previsti per ogni modello, allo scopo di stabilire quale siano i modelli con la migliore accuratezza predittiva.
6. Applicazione del modello all'Insieme di Test (campione conservato a parte) per la validazione della capacità/accuratezza predittiva.

## Risultati di Modelli Lineari Generali e Spline MAR

L'output di *Modelli Lineari Generali (GLM)*, che si basa su assunti parametrici sui dati, e di *Spline MAR*, strumento tradizionale di data mining, è particolarmente interessante. L'analisi *GLM* usa la regressione stepwise forward per selezionare un modello predittivo valido. Le variabili significative al livello 0,05 vengono aggiunte iterativamente finché non vengono individuate ulteriori variabili significative. Il modello risultante assume che vi sia una relazione lineare tra il Volume di

Carburante e le variabili predittrici disponibili. Contrariamente, *Spline MAR* è uno strumento non parametrico che non assume l'esistenza di una relazione lineare, bensì costruisce un modello "guidato dai dati". Questo processo viene compiuto segmentando lo spazio di input in diverse regioni, ognuna delle quali caratterizzata dalla propria equazione di regressione.

Di seguito sono riportati gli scatterplot in cui vengono riportati i valori previsti rispetto ai valori attesi, rispettivamente per *GLM* e per *Spline MAR*. Dall'osservazione dei due grafici è evidente come il modello *Spline MAR* esegue previsioni migliori sul Volume di Carburante.

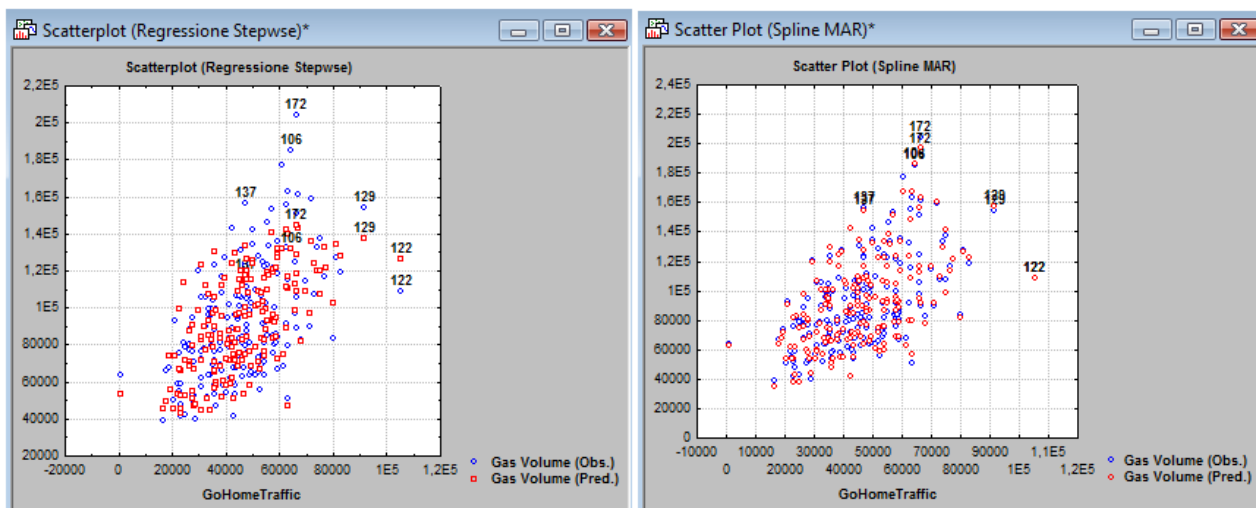


In particolare, il grafico relativo ai risultati della regressione stepwise *GLM* evidenzia la presenza di un gruppo di valori non ben previsti dal modello. Questo gruppo contiene i casi in corrispondenza dei quali il Volume di Carburante è più alto. I clienti con la domanda più alta sono di particolare interesse e per costoro è importante che la previsione sia il più accurata possibile. I distributori caratterizzati da Volumi di Carburante al di sopra di 150.000 sono considerevolmente sottostimati dal modello di regressione.

I risultati del modello *Spline MAR* mostrano una correlazione molto più forte tra valori osservati e valori previsti. Questo è specialmente importante per i clienti con domanda più elevata. L'accuratezza predittiva è molto migliore, quindi, per il modello *Spline MAR* che per la regressione stepwise.

Un ulteriore confronto tra questi modelli può esser fatto attraverso l'osservazione dei valori di  $R^2$  e di  $R^2$  corretto. Queste sono le principali misure di prestazione dei modelli, per i quali un valore prossimo ad 1 indica un adattamento pressoché perfetto. L' $R^2$  corretto è un valore aggiustato in base al numero di variabili presenti nel modello, consistente nella penalizzazione di modelli contenenti un grande numero di variabili. Questa misura sarà sempre inferiore rispetto a  $R^2$  e favorisce modelli più semplici. Per questi due modelli, possiamo osservare valori di  $R^2$  migliori per il modello *Spline MAR*. Per i risultati di regressione,  $R^2$  è pari a 0,72 mentre il suo corrispettivo corretto è pari a 0,68. Per i risultati *Spline MAR*,  $R^2$  è pari a 0,97 mentre l' $R^2$  corretto è pari a 0,96. Queste statistiche evidenziano il miglioramento ottenuto con l'algorithmo *Spline MAR* rispetto al risultato raggiunto attraverso il ricorso ad una tecnica convenzionale quale la regressione stepwise.

I seguenti scatterplot illustrano la discrepanza tra casi osservati e casi previsti usando il modello di regressione stepwise ed il modello *Spline MAR*. Cinque casi d'interesse sono etichettati con i rispettivi numeri di caso.



## Conclusioni

Questo esempio evidenzia i difetti della tradizionale analisi di regressione così come i vantaggi conseguibili con l'utilizzo di algoritmi più complessi di data mining. Nonostante gli algoritmi di data mining siano più complessi e molto più potenti, nell'ambiente di *STATISTICA* la loro integrazione è molto semplice.