

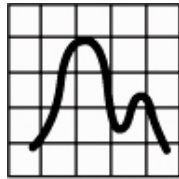


StatSoft®

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

Demand Forecasting and *STATISTICA*

Case Study: Gas Volume Demand



STATISTICA
**Solutions for Business Intelligence,
Data Mining, Quality Control, and
Web-based Analytics**

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

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.

Table of Contents

INTRODUCTION: WHAT IS DEMAND FORECASTING?.....	1
CASE STUDY: GAS VOLUME DEMAND	1
Case Description	1
DATA ANALYSIS WITH <i>STATISTICA</i>.....	2
Feature Selection.....	2
<i>STATISTICA</i> Data Miner Workspace.....	3
<i>General Linear Models</i> and <i>MARSplines</i> Results.....	4
CONCLUSION.....	5

Introduction: What is Demand Forecasting?

Anyone in the business of providing goods knows the importance of forecasting the demand for those goods. Knowledge of how demand will fluctuate enables the supplier to keep the right amount of stock on hand. If demand is underestimated, sales can be lost due to the lack of supply of goods. If demand is overestimated, the supplier is left with a surplus that can also be a financial drain. Understanding demand makes a company more competitive in the marketplace. Understanding demand and the ability to accurately predict it is imperative for efficient manufacturers, suppliers, and retailers.

To be able to meet consumers' needs, appropriate forecasting models are vital. Although no forecasting model is flawless, unnecessary costs stemming from too much or too little supply can often be avoided using data mining methods. Using these techniques, a business is better prepared to meet the actual demands of its customers. *STATISTICA Data Miner* offers a wealth of tools that can aid in forecasting of demand.

Case Study: Gas Volume Demand

Case Description

Gasoline suppliers need an accurate forecast model to determine their customers' demand. Knowing customer demand will enable them to have enough fuel on hand, without too much surplus. The accuracy of the forecasting model will greatly reduce unnecessary costs associated with too much or too little supply to meet customer needs.

The techniques explained in this case study illustrate how to build a customer demand forecasting model using *STATISTICA Data Miner* to identify the inputs or predictors that highly influence the fuel usage of particular stores to which the company supplies. Accurate estimates of these demands will allow for a total assessment of the supplier's demand.

Data File

The data file *store data.sta* consists of store-level data. There are 184 observations and 158 variables. These data contain information about the type of store, location, marketing techniques used, types of traffic seen, number of fuel pumps offered, other goods and services offered, etc.

For each store observation, gasoline demand is recorded. This is the dependent variable that will be forecasted with *STATISTICA Data Miner*. Using feature selection, the 158 variable list will be condensed to a more manageable and appropriate variable list.

Data Analysis with *STATISTICA*

With *STATISTICA Data Miner*, it is straightforward to apply powerful modeling tools to data and judge the value of resulting models based on their predictive or descriptive value. Data preparation efforts are critical. Data is the main resource for data mining – therefore it should be prepared properly before applying any data mining tool. Otherwise, it would be just a case of garbage-in/garbage-out (GIGO). Since major strategic decisions are impacted by these results, any error might give rise to huge losses. Thus, it is important to preprocess the data and improve the accuracy of the model so that the best possible decision can be made.

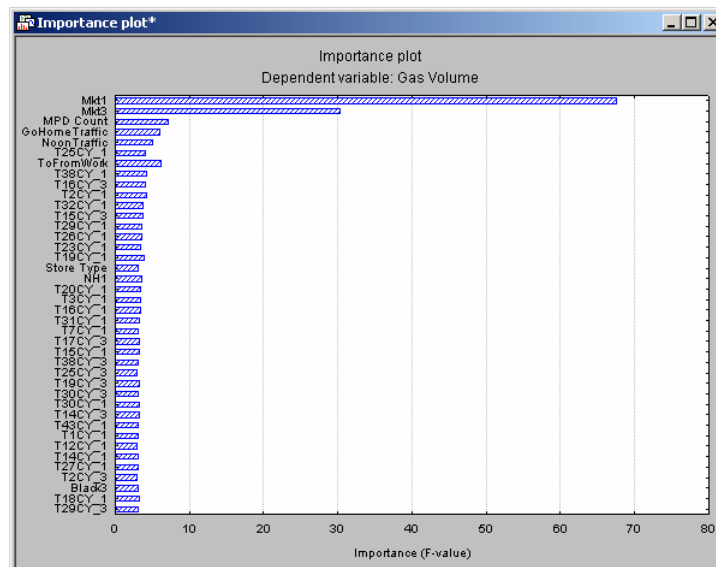
The following aspects of the data were noted during this stage:

- Insight into data: Descriptive statistics (by looking at distributions, means, minimum and maximum values, quartiles, etc.)
- There are no outliers in the data
- There are no missing values in the data
- No transformations are required

Feature Selection

In order to reduce the complexity of the problem, the data set can be transformed into a data set of lower dimension. The *Feature Selection and Variable Screening* tool available in *STATISTICA Data Miner* automatically found the most important predictors that highly influence the store gas demand.

The bar plot and spreadsheet of the predictor importance give insight into the variables that are related to the prediction of the dependent variable of interest. For example, shown below is the bar plot of predictor importance for the dependent variable Gas Volume.



The Feature Selection results show that the marketing campaigns Mkt1 and Mkt3 have the strongest influence on Gas Volume. A number of other variables are significant predictors of Gas Volume as well.

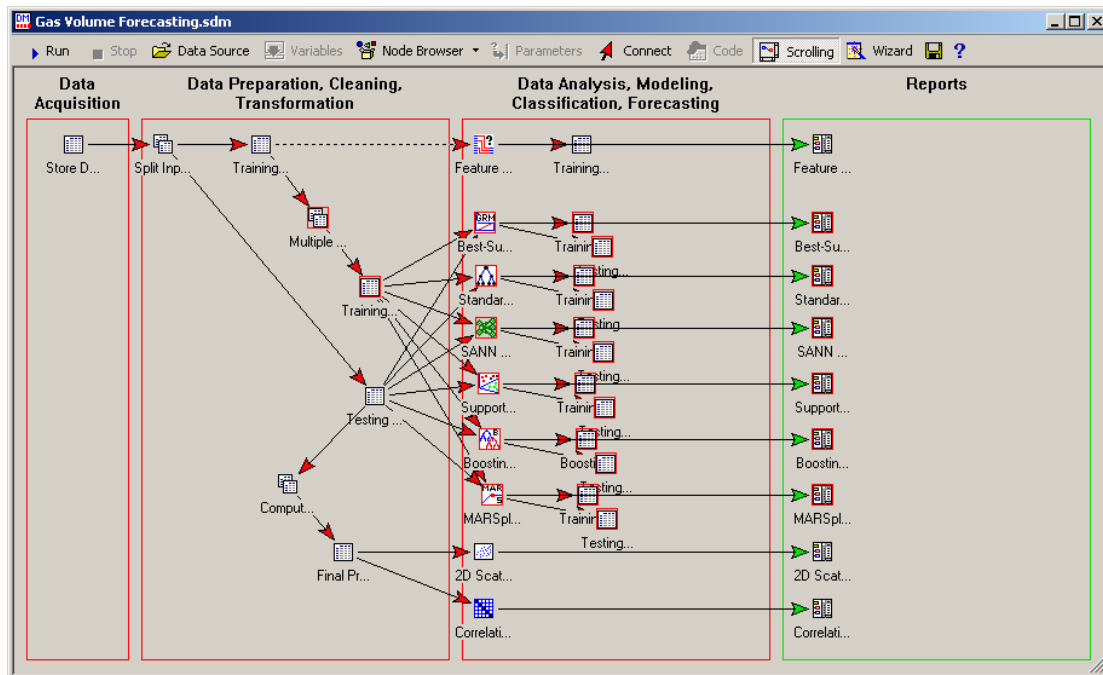
This example examines these predictors using many of the algorithms offered by *STATISTICA Data Miner*. These algorithms include:

- *General Linear Models (GLM)*
- *Classification and Regression Trees (C&RT)*
- *STATISTICA Automated Neural Networks (SANN)*
- *Support Vector Machines*
- *Boosted Regression Trees*
- *Multivariate Adaptive Regression Splines (MARSplines)*

The novelty and abundance of available techniques and algorithms involved in the modeling phase make this the most interesting part of the data mining process. Moreover, it is good practice to experiment with a number of different methods when modeling or mining data. Different techniques may shed new light on a problem or confirm previous conclusions.

***STATISTICA Data Miner* Workspace**

The *Data Miner* workspace depicts the flow of the analyses. All tools of *STATISTICA Data Miner* are available as icons via simple drag-and-drop. The following diagram illustrates how the *Data Miner* workspace looks after all the analyses were performed.



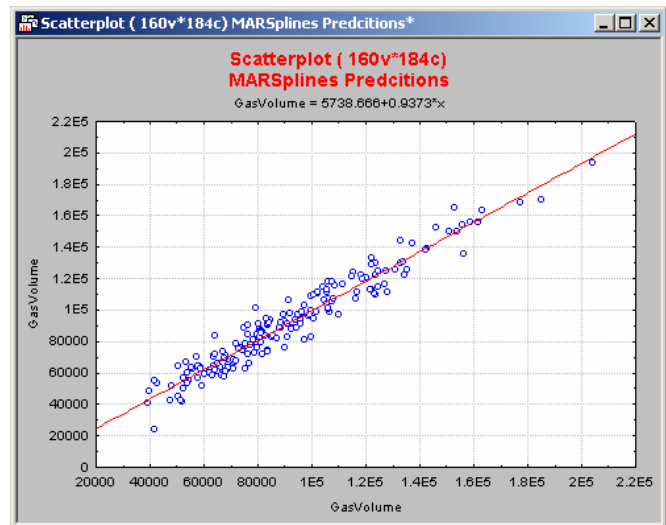
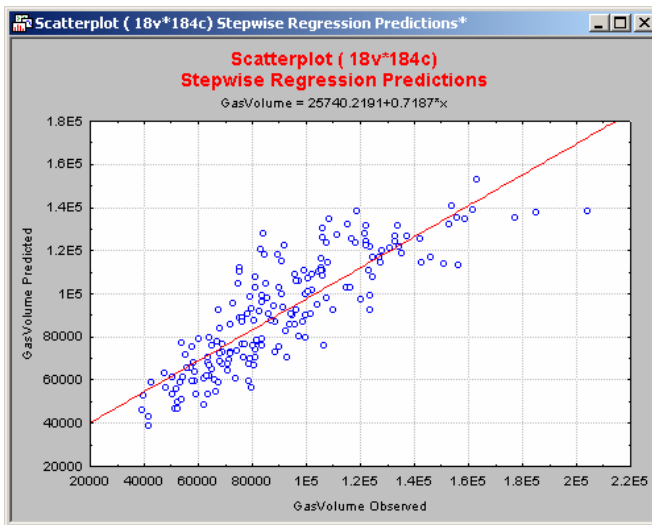
The following steps summarize the data preparation and analysis flow:

- Split the original data set into two subsets; 30% of cases were retained for testing and 70% of cases were used for feature selection and model building.
- Used Feature Selection tool to rank the best predictor variables for predicting gas volume
- Reduced the number of possible predictors from 158 to 40 based on the results of Feature Selection.
- Used different advanced Predictive Models (Machine Learning algorithms) to detect and understand relationships.
- Used scatterplots and correlation tables of the observed vs. predicted values of each model to determine the models with the best predictive accuracy.
- Applied the model to the Test Set (hold-out sample) to validate prediction accuracy.

General Linear Models and MARSplines Results

The output from *General Linear Models (GLM)*, which makes parametric assumptions about the data, and *MARSplines*, a data mining tool, were particularly interesting. The *GLM* analysis uses forward stepwise regression to select a valid prediction model. Variables significant at the 0.05 level are added iteratively until no more variables are found to be significant. The resulting model assumes a linear relationship between Gas Volume and the available predictor variables. In contrast, *MARSplines* is a nonparametric tool that does not assume the linear relationship, but constructs a “data driven” model. This is done by segmenting the input space into regions, each with their own regression equations.

Below are scatterplots showing the observed vs. predicted values from *GLM* and *MARSplines*. The first plot shows observed vs. predicted values from the *GLM* stepwise regression model, and the second shows observed vs. predicted values from the *MARSplines* model. From the two plots, it is easy to see that the *MARSplines* model is the better forecaster of Gas Volume.

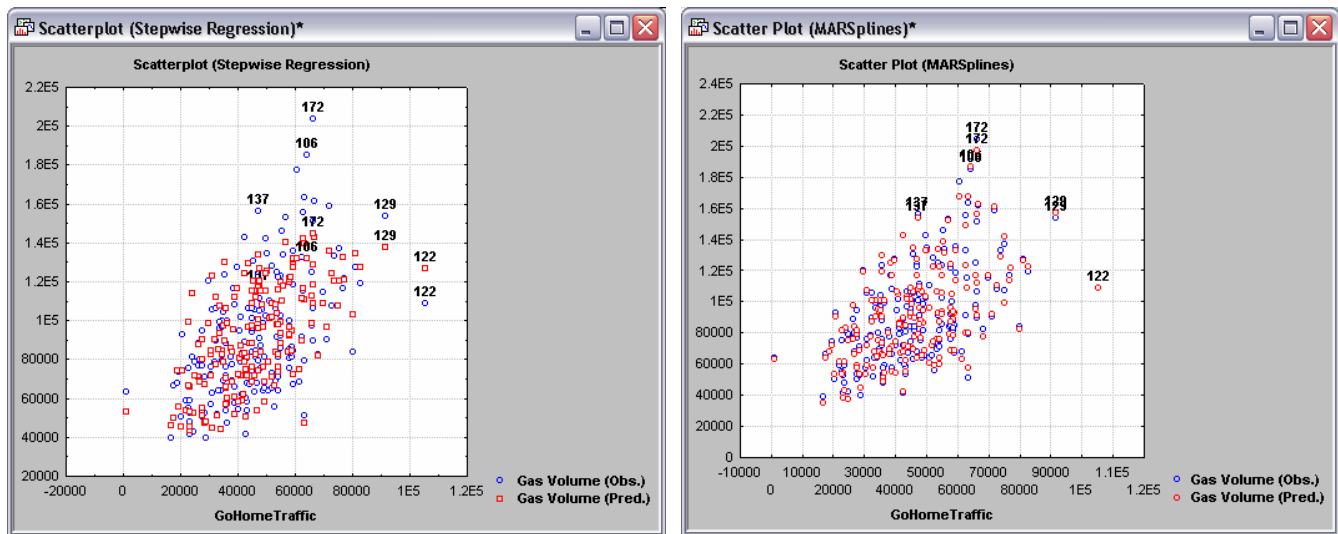


Specifically, the plot of *GLM* stepwise regression results shows a cluster of values not well predicted by the model. This cluster contains the cases where Gas Volume demand was high. The high-demand customers are of particular interest and are important to be accurately predicted. Stores with Gas Volume above 150,000 were consistently underestimated by the regression model.

The results of *MARSplines* show a much stronger correlation between observed and predicted values. This is especially important for the high-demand customers. The predictive accuracy is far better with *MARSplines* than with stepwise regression.

Another comparison of these models is R^2 and adjusted R^2 . These are measures of model performance in which values close to 1 indicate the best fit. Adjusted R^2 is adjusted for the number of variables in the model, penalizing the measure for many variables. This measure will always be less than R^2 and favors a more simplified model. For these two models, we see a much better R^2 for the *MARSplines* model. For the regression results, R^2 is 0.72 and adjusted R^2 is 0.68. For *MARSplines* results, R^2 is 0.97 and adjusted R^2 is 0.96. These statistics highlight the improvement we find with the *MARSplines* algorithm over the conventional regression approach.

The following scatterplots show the discrepancy of observed and predicted cases using the *GLM* stepwise regression model and the *MARSplines* model. Five cases of interest are labeled with their case numbers. Both observed Gas Volume and predicted Gas Volume are plotted in each scatterplot. In the stepwise regression spreadsheet, the discrepancy in observed Gas Volume and predicted Gas Volume is quite large. In the *MARSplines* spreadsheet, observed Gas Volume and predicted Gas Volume are very close, as expected from an accurate model.



Conclusion

This example highlights the shortcoming of traditional regression analysis and the power that is gained by utilizing complex data mining algorithms for demand forecasting. Although the data mining algorithms are complex and high-powered, they are just as easy to use (if not easier) in the *STATISTICA* environment.