

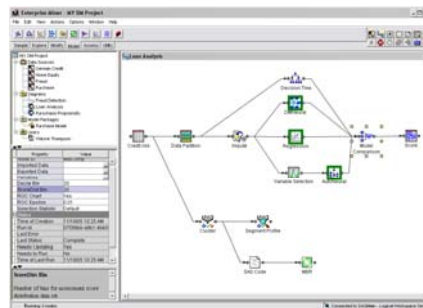
**THE  
POWER  
TO KNOW.**

## Neuerungen in SAS®9.2 zu Analytik und Data Mining

---

Dr. Gerhard Svolba

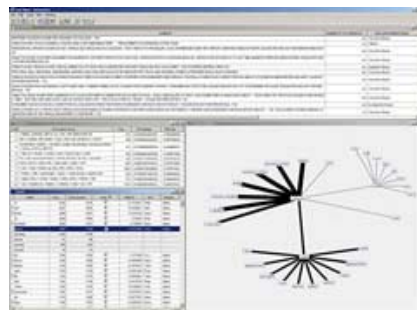
# Neuerungen in SAS®9.2



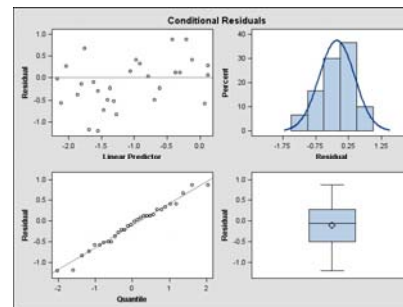
**SAS®Enterprise Miner**



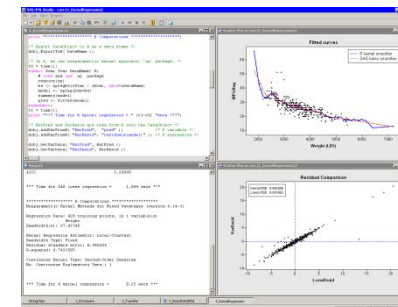
**SAS®Forecast Server**



**SAS®Text Miner**

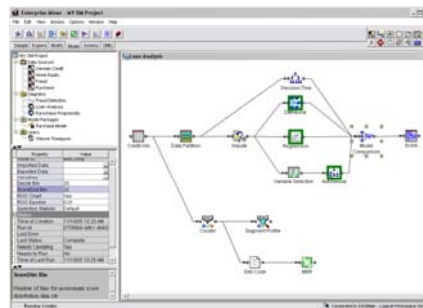


**SAS®STAT (ODS Graphics)**

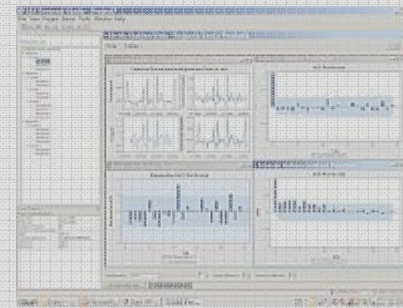


**SAS®IML-Studio (R-Integration)**

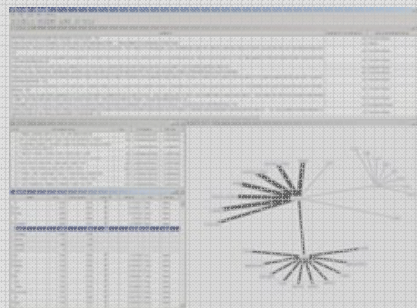
# Neuerungen in SAS®9.2



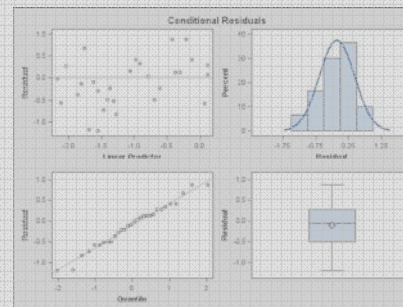
SAS®Enterprise Miner



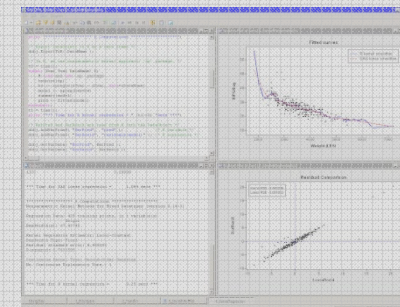
SAS®Forecast Server



SAS®Text Miner



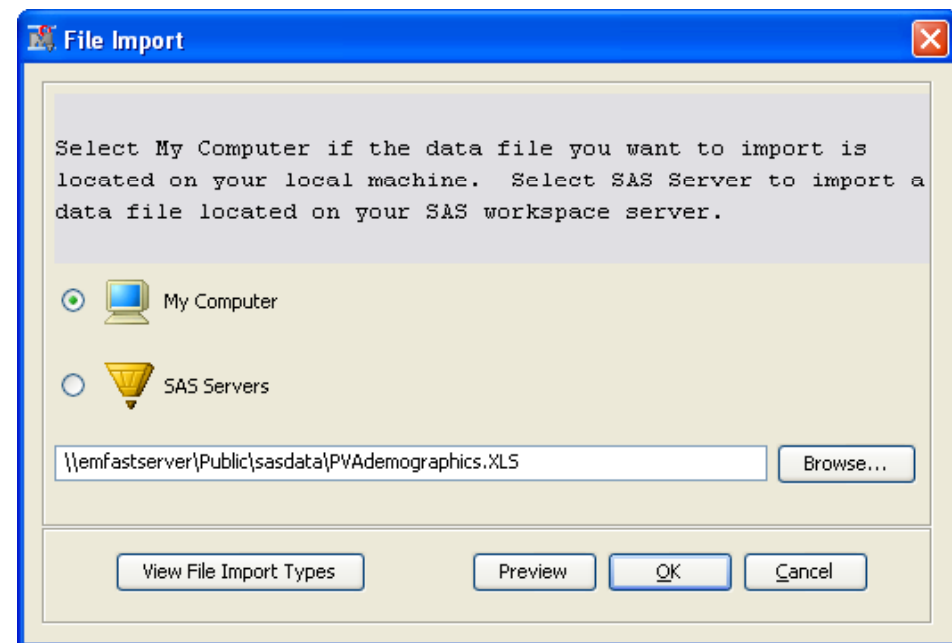
SAS®STAT (ODS Graphics)



SAS®IML-Studio (R-Integration)

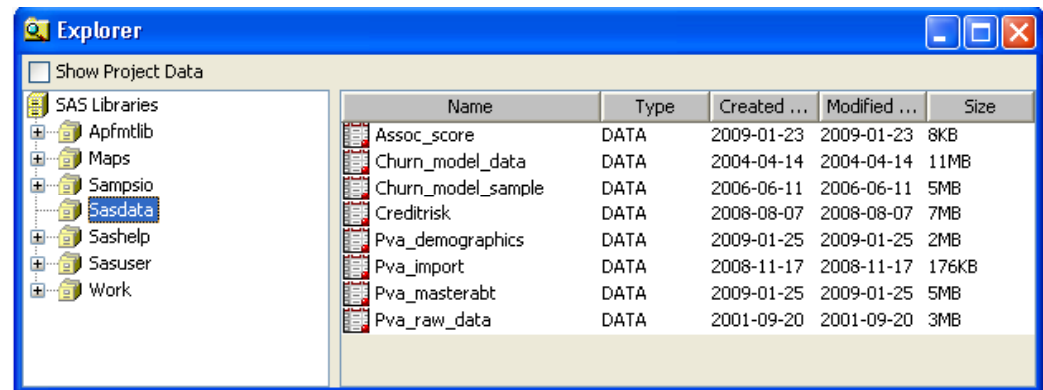
# Datenaufbereitung – File Import

- Dateitypen wie z.B: EXCEL, CSV, Tab delimited, JMP, etc.
- Ideal für Nicht-SAS Programmierer
- Im Table-Preview können Optionen für die Anzahl der Spalten gesetzt werden
- “Advanced Advisor” Integration zur Definition der Eigenschaften der Quelldaten



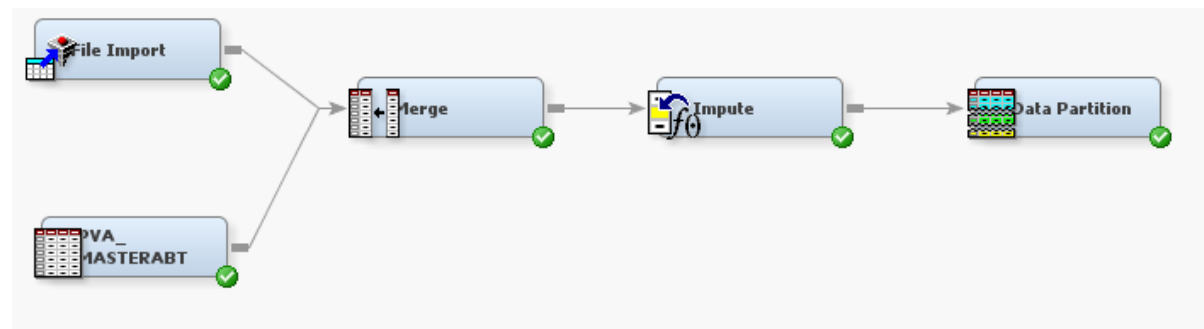
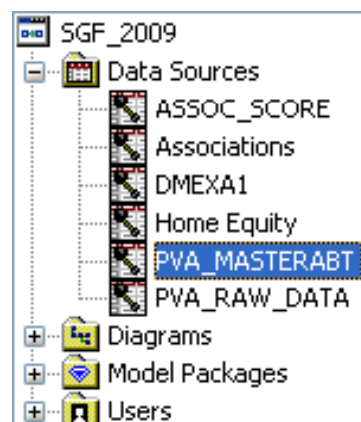
# Datenaufbereitung – Drag and Drop Tables

- SAS Table Explorer
- Startet automatisch den “Data Source Wizard”
- Vordefinierte SAS Tabellen können vom Business Analyst verwendet werden



The screenshot shows the SAS Explorer window. On the left, under 'SAS Libraries', the 'Sasdata' library is selected. On the right, a table lists various data sources with their names, types, creation and modification dates, and sizes.

Name	Type	Created ...	Modified ...	Size
Assoc_score	DATA	2009-01-23	2009-01-23	8KB
Churn_model_data	DATA	2004-04-14	2004-04-14	11MB
Churn_model_sample	DATA	2006-06-11	2006-06-11	5MB
Creditrisk	DATA	2008-08-07	2008-08-07	7MB
Pva_demographics	DATA	2009-01-25	2009-01-25	2MB
Pva_import	DATA	2008-11-17	2008-11-17	176KB
Pva_masterabt	DATA	2009-01-25	2009-01-25	5MB
Pva_raw_data	DATA	2001-09-20	2001-09-20	3MB



# Beschreibende Statistiken während des Datenzugriffs

Variables - FIMPORT

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining ☐ Basic ☒ Statistics

Name	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
CONTROL_NUMBER	31	0	.	.	.	.	.	.
DONOR_AGE	.	24.75222	0	87	58.91905	16.66938	-0.3779	-0.4565
DONOR_GENDER	4	0	.	.	.	.	.	.
HOME_OWNER	2	0	.	.	.	.	.	.
INCOME_GROUP	7	22.6719	.	.	.	.	.	.
MEDIAN_HOME_VALUE	.	0	0	6000	1079.872	960.7534	2.456613	6.99424
MEDIAN_HOUSEHOLD	.	0	0	1500	341.9702	164.2078	1.723597	6.47759
OVERLAY_SOURCE	4	0	.	.	.	.	.	.
PCT_MALE_MILITARY	.	0	0	97	1.029011	4.918297	11.74183	177.909
PCT_MALE_VETERANS	.	0	0	99	30.57392	11.42147	-0.19742	1.22319
PCT_OWNER_OCCUP	.	0	0	99	69.699	21.71102	-1.23559	1.17286
PER_CAPITA_INCOMI	.	0	0	174523	15857.33	8710.63	3.352875	23.1875
PUBLISHED_PHONE	2	0	.	.	.	.	.	.
SES	5	0	.	.	.	.	.	.
URBANICITY	6	0	.	.	.	.	.	.
WEALTH_RATING	10	45.47801	.	.	.	.	.	.
CLUSTER_CODE	31	0	.	.	.	.	.	.

- Beschreibende Statistiken zu den einzelnen Variablen
- Auswahl der gewünschten Statistiken und Metadaten

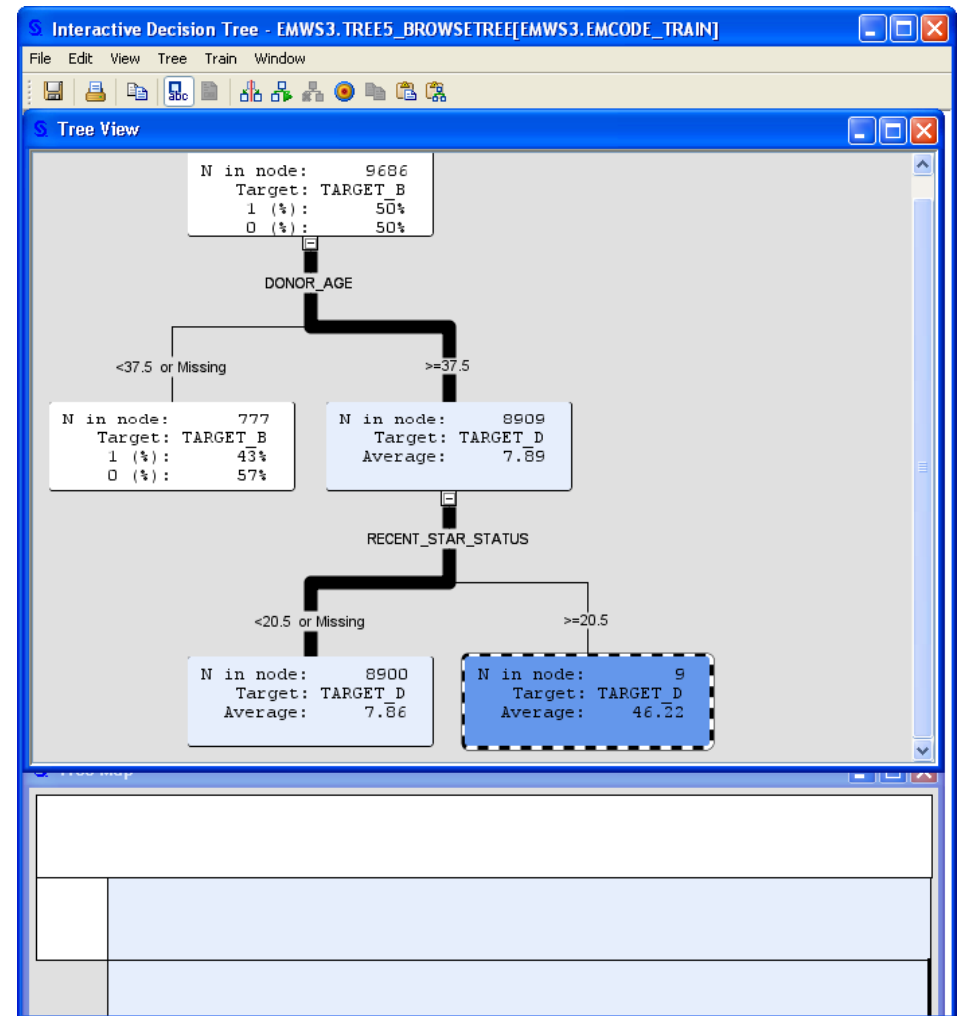
# Zusätzliche Erweiterungen

- Stat Explore node
  - Berechnet die Summary Statistiken auf allen Daten-Partitionen (Train/Validation/Test)
  - Bietet neue Plots, welche den Vergleich von Verteilungen über mehrere Ziel- oder BY-Variablen ermöglichen
- Graph Explore node
  - Stratifiziertes Sampling nach kategoriellen Zielvariablen möglich



# Interactive Decision Tree

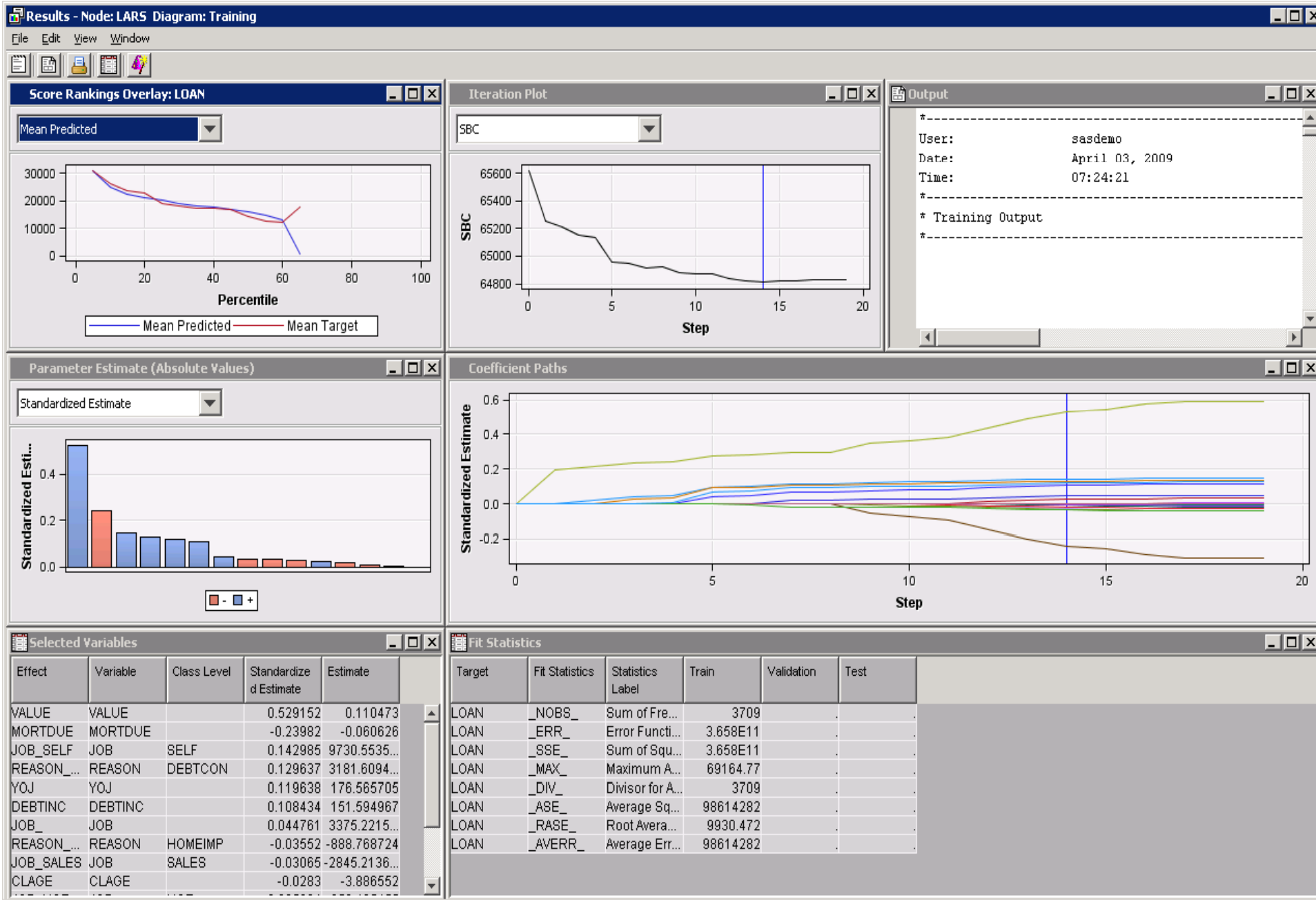
- Native Interactive Tree
  - Ersetzt die "Tree Desktop Application"
  - Kein separater Installationslauf
  - Launch über Java Webstart möglich
- Multiple targets
  - Definition von Splits auf unterschiedlichen Zielvariablen im Baum
  - Entwicklung von Strategien und Segmentierungen





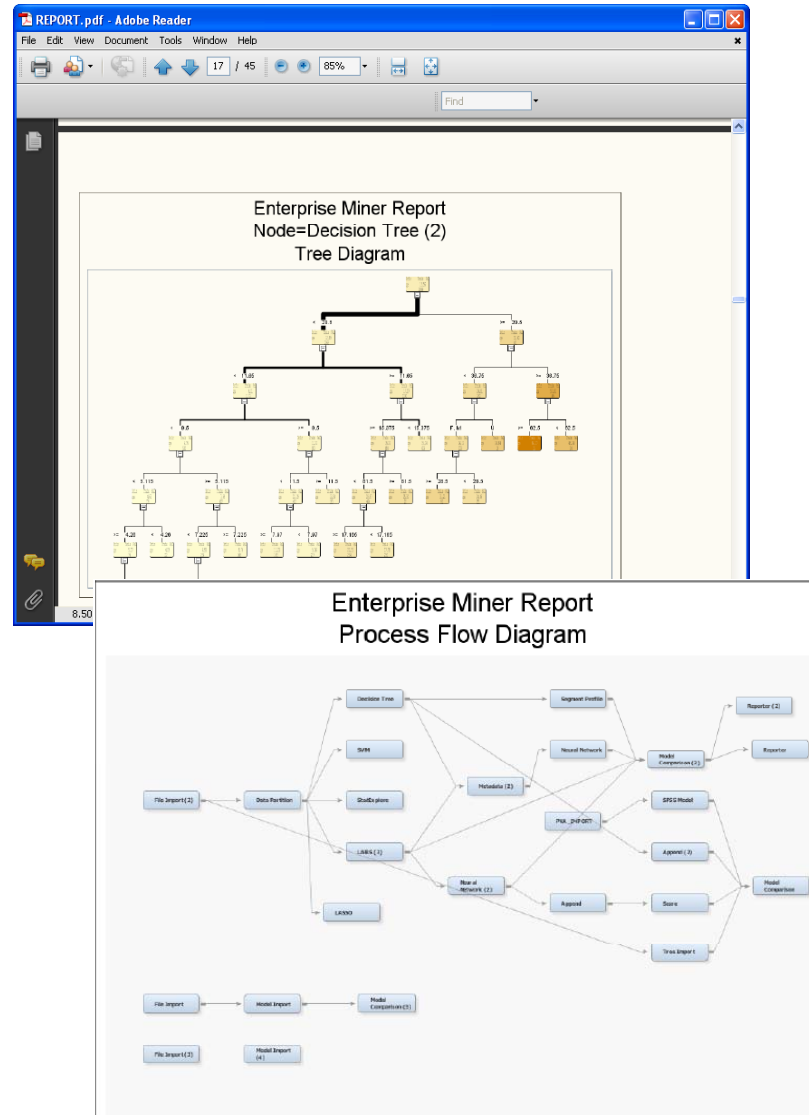
# Least Angle Regression Splines

- LARS und LASSO
- Neues “general linear regression model” für die Variablenselektion
- Effizientere Suche als der Stepwise Algorithmus
- Koeffizienten “grow, decay, enter, exit” im Modell
- Unterstützt Cross validation und hold out sampling



# Verbesserter Reporter node

- PDF und RTF Output
- Verwendet ODS Output
- Verbesserte Abbildung des Prozessflussdiagramms
- Inkludiert auch:
  - Entscheidungsbaum-Darstellung
  - Interaktive Graphiken
- Kennzeichnung der Einstellungen, die vom Default abweichen
- Externer Report Viewer



# Model Export und Import für den Modellvergleich

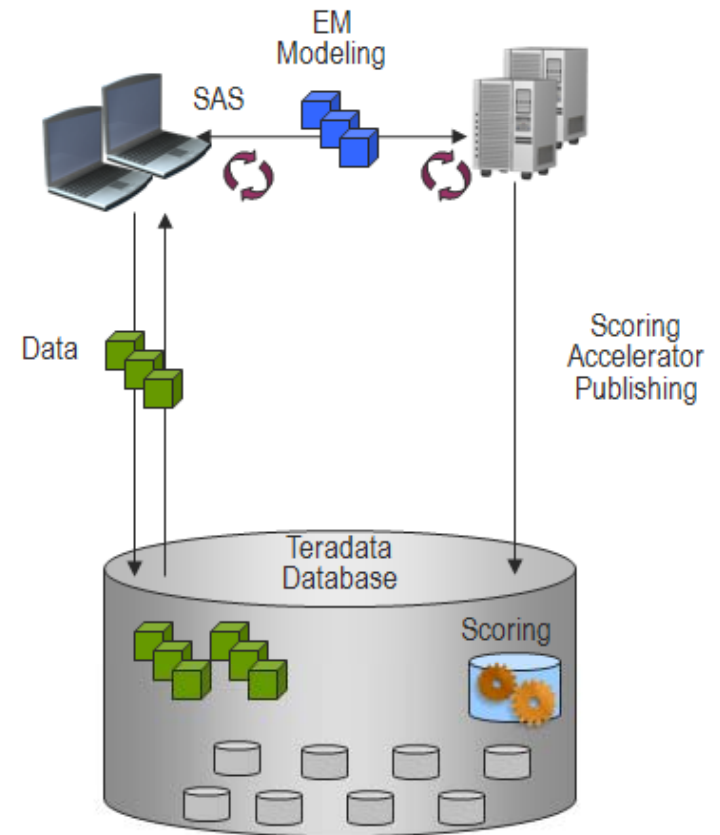
- Speichern des Model Package
- Registrieren am Metadata Server
  - Speichern in SAS Folders
- Einlesen der Modelle im Enterprise Miner
  - Scoring von neuen Daten
  - Berechnung der Modell-Bewertungsstatistiken
- Vergleich mit neuen Modellen
  - Verwenden des Append-Nodes für neue Daten

# Score Node

- Optimierter Score Code
  - Weglassen von temporären Code, der Werte berechnet, die nicht in der finalen Modellgleichung verwendet werden
    - Beispiel: Abgeleitete Variable werden im Prozessfluss erzeugt, aber im finalen Modell nicht verwendet
  - Optimierter Score Code wird per default erzeugt
  - Nicht-optimierter Score Code kann für Vergleichszwecke exportiert werden
  - Optimierter Score Code hat positiven Einfluss auf Scoring und Deployment Prozesse
    - Weniger Variablen müssen im Input Data Set gespeichert sein (weniger Speicherplatz, kürzere Laufzeit, weniger Bereitstellungsaufwand)

# Score Code und Deployment

- Score code optimization
- PMML 3.1 kompatibel
- Scoring Accelerator for Teradata
  - Erzeugt die Scoring-Funktion in der Datenbank
  - Datenbankspezifisches native SQL

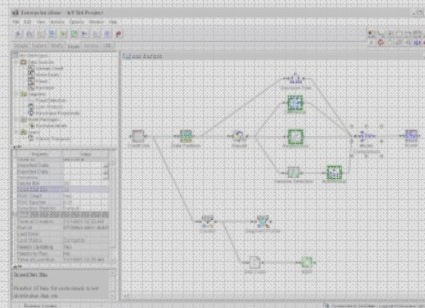


# SAS 9.2 Platform Integration

- Migration von EM 5.2 und EM 5.3 Projekten und Modellen
  - Metadaten werden von SAS 9.1.3 auf SAS 9.2 portiert
  - Projektverzeichnisse am SAS – Server werden nicht verändert
- Zusätzlicher Support
  - Windows Vista 32 bit und 64 bit server support
  - Solaris 64 bit on Intel architecture
  - Linux 64 bit on Intel architecture



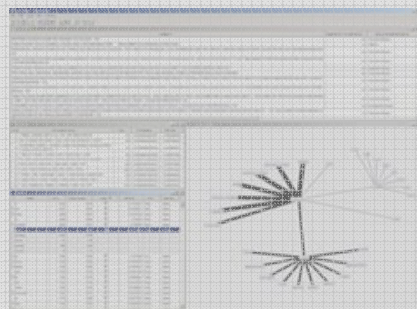
# Neuerungen in SAS®9.2



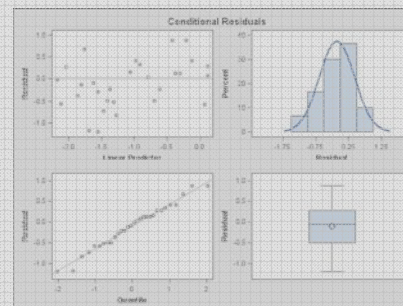
SAS®Enterprise  
Miner



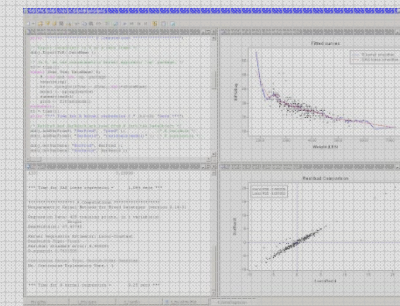
SAS®Forecast  
Server



SAS®Text  
Miner



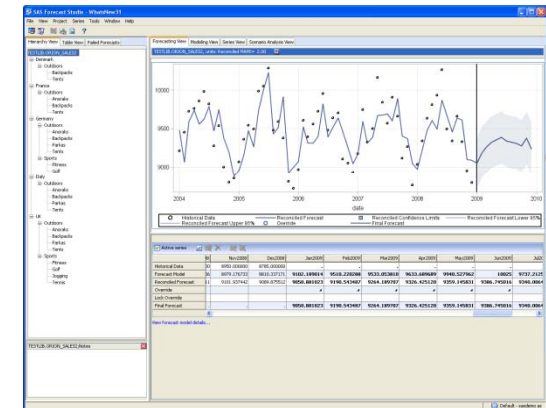
SAS®STAT  
(ODS Graphics)



SAS®IML-Studio  
(R-Integration)

# Was ist der SAS Forecast Server?

- Forecast Server bietet eine unternehmensweite Umgebung für automatisiertes “large-scale” forecasting
- Benutzer können interaktiv mit der graphischen Benutzeroberfläche arbeiten, oder Code in einem Batch-Environment schreiben
- Erzeugt automatisch und rasch eine große Anzahl von Forecasts; Analytiker soll sich auf die “problematischen” Zeitreihen fokussieren
- Bietet Werkzeuge, für das Overwriting von strategischen und “high-value” Zeitreihen



*Ziel ist es, eine große Anzahl von Forecasts zu erzeugen, die so exakt und unverfälscht sind, wie man es rationell erwarten kann; und dies so effizient wie möglich zu tun*

## Spezifikation von „BY“ Variablen und Hierarchiestrukturen gleich zu Beginn der Projektdefinition

**New Project - Step 3 of 8**

**Specify Classification (BY) Variables and Whether to Forecast a Hierarchy**

Available variables:

- regionName
- productLine
- productName
- date
- units
- price
- discount
- cost

Classification (BY) variables selected:

☐ Forecast a hierarchy with levels defined by the above classification (BY) variables:

☐ Reconcile the hierarchy: Top Down

Preview...

< Previous    Next >

**New Project - Step 3 of 8**

**Specify Classification (BY) Variables and Whether to Forecast a Hierarchy**

Available variables:

- date
- units
- price
- discount
- cost

Classification (BY) variables selected:

- regionName
- productLine
- productName

☒ Forecast a hierarchy with levels defined by the above classification (BY) variables:

☒ Reconcile the hierarchy: Top Down

Advanced...

Preview...

Bottom Up  
Top Down  
Middle Out - regionName  
Middle Out - productLine

**Specifying BY variables activates the hierarchy property dialogs**

## Definition der Zeitvariable

**New Project - Step 4 of 8**

**Specify The Properties of the Time Dimension of Your Data**

Time ID variable:

Interval:  Weekend...

Multiplier:

Shift:

Seasonal cycle length:

Format: 9,

**New Project - Step 4 of 8**

**Specify The Properties of the Time Dimension of Your Data**

Time ID variable:

Interval:  Weekend...

Multiplier:

Shift:

Seasonal cycle length:

Format: MONYY7. (e.g. Feb2009)

**Twelve new interval formats**

- Day
- Hour
- ISO 8601 Year
- ISO 8601 week
- Minute
- Month
- None
- Quarter
- Retail 4-4-5 month
- Retail 4-4-5 quarter
- Retail 4-4-5 year
- Retail 4-5-4 month
- Retail 4-5-4 quarter
- Retail 4-5-4 year
- Retail 5-4-4 month
- Retail 5-4-4 quarter
- Retail 5-4-4 year
- Second
- Semimonth
- Semiyear
- Ten-day
- Week
- WeekDay

**Edit Time ID Format**

Select the time ID format to use:

☐ System recommended: MONYY7. (e.g. Feb2009)

☒ Select from list:

- EURDFDT.
- EURDFDWN.
- EURDEMN.

**Select from a list of all supported SAS date formats**

## Zuweisen von abhängigen und unabhängigen Variablen und ihrer Eigenschaften

**New Project - Step 5 of 8**

**Assign Roles to Variables in Your Data**

**i** You must specify a dependent variable.

Variable	Role	Aggregation	Accumulation	Usage in System-Gen...
units	Dependent	Sum of values	Sum of values	
price	None			
discount	Dependent			
cost	Independent			
	Report Only			

☒ Set accumulation to the value used for aggregation

**Accumulation**

- Sum of values
- Average of values
- Corrected sum of squares of values
- First value
- Last value
- Maximum of values
- Median of values
- Minimum of values
- NONE
- Number of missing values
- Number of non-missing values
- Standard deviation of values
- Sum of values
- Total number of values
- Uncorrected sum of squares of values

**New Project - Step 5 of 8**

**Assign Roles to Variables in Your Data**

**i** You must specify a dependent variable.

Variable	Role	Aggregation	Accumulation	Usage in System-Gen...
units	Dependent	Sum of values	Sum of values	
price	Independent	Average of values	Average of values	Try to use
discount	None			Try to use
cost	None			Use if significant
				Force use

☒ Set accumulation to the value used for aggregation

Specify how independent variables are to be used in system-generated forecasts.





Improved model description readability

#### Subset ARIMA Model

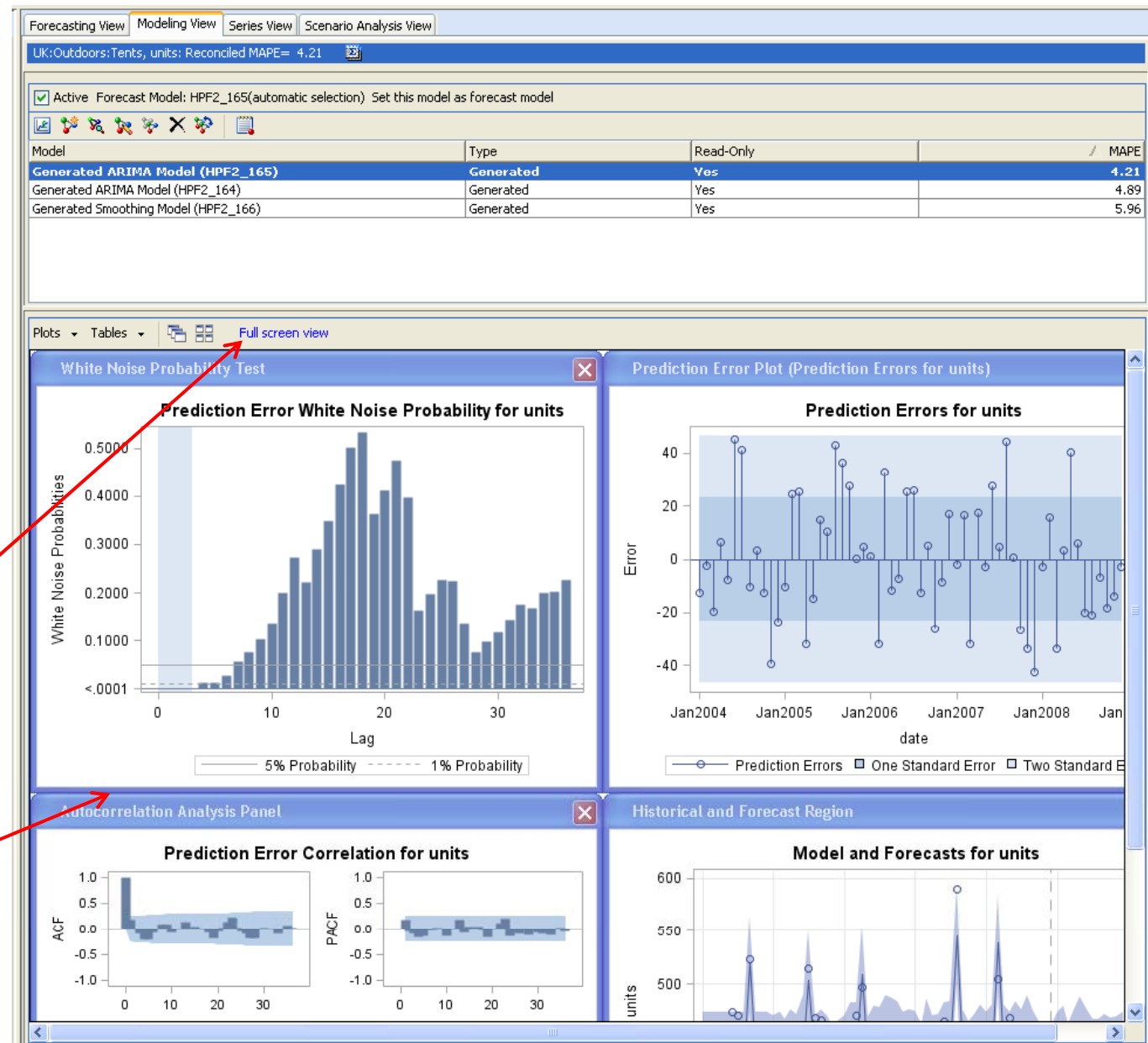
**Name:** HPF2\_165  
**Description:** "ARIMA: units ~ Q = (12) + INPUT: price"  
**Details:** "ARIMA: units ~ Q = (12) + INPUT: price"  
**Model family:** ARIMA  
**Model type:** GENERALARIMA  
**Source:** Generated by HPFDIAGNOSE

**Intercept:** Yes  
**Forecast variable:** units  
**Delay:** 0  
**Differencing:** ( 0 ) Q: ( 12 )

Toggle to "full screen view" for plot area

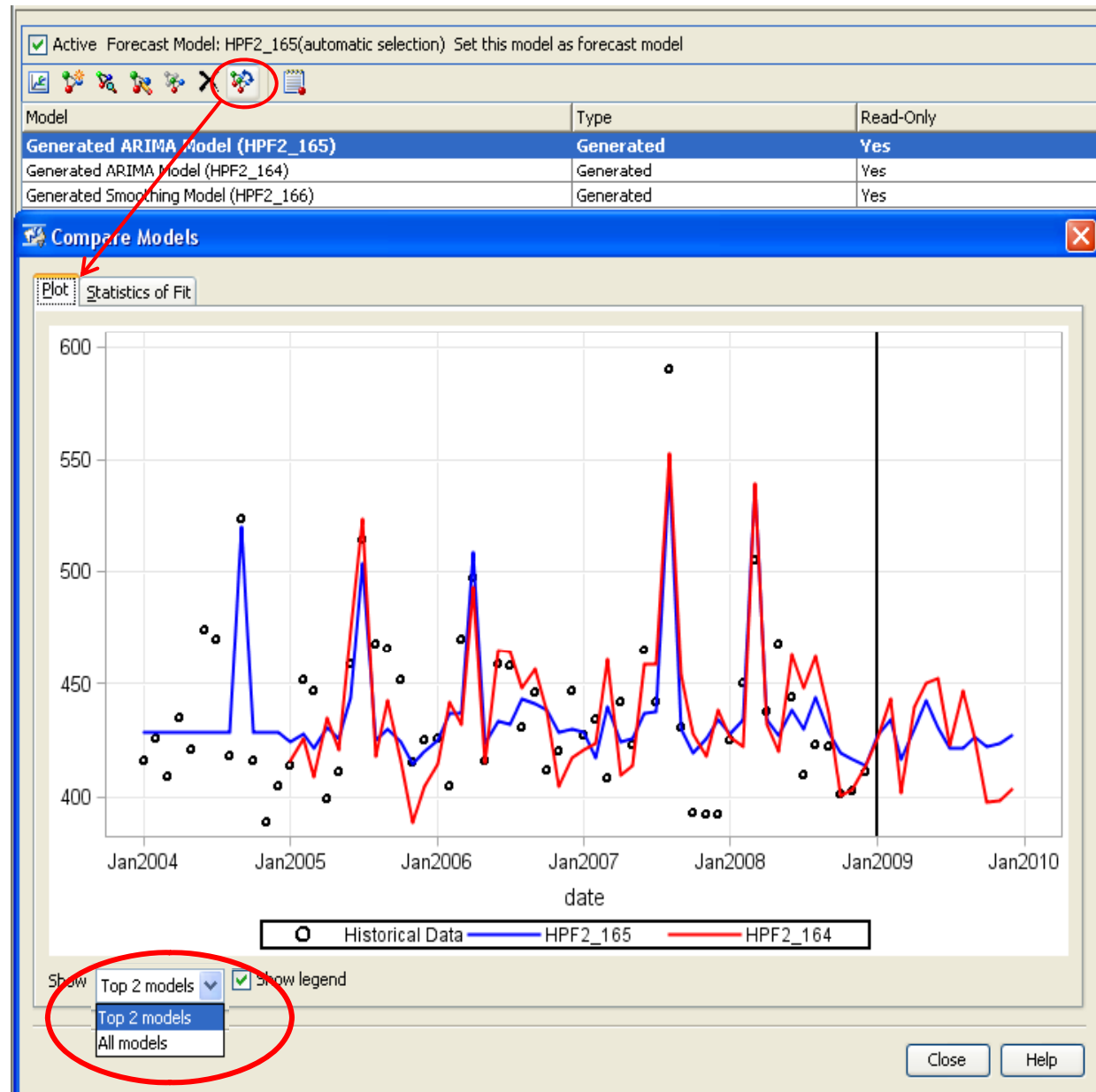
Four plots automatically shown:

- ✓ while noise probability plot
- ✓ prediction error plot
- ✓ autocorrelation plot
- ✓ historical / forecast plot

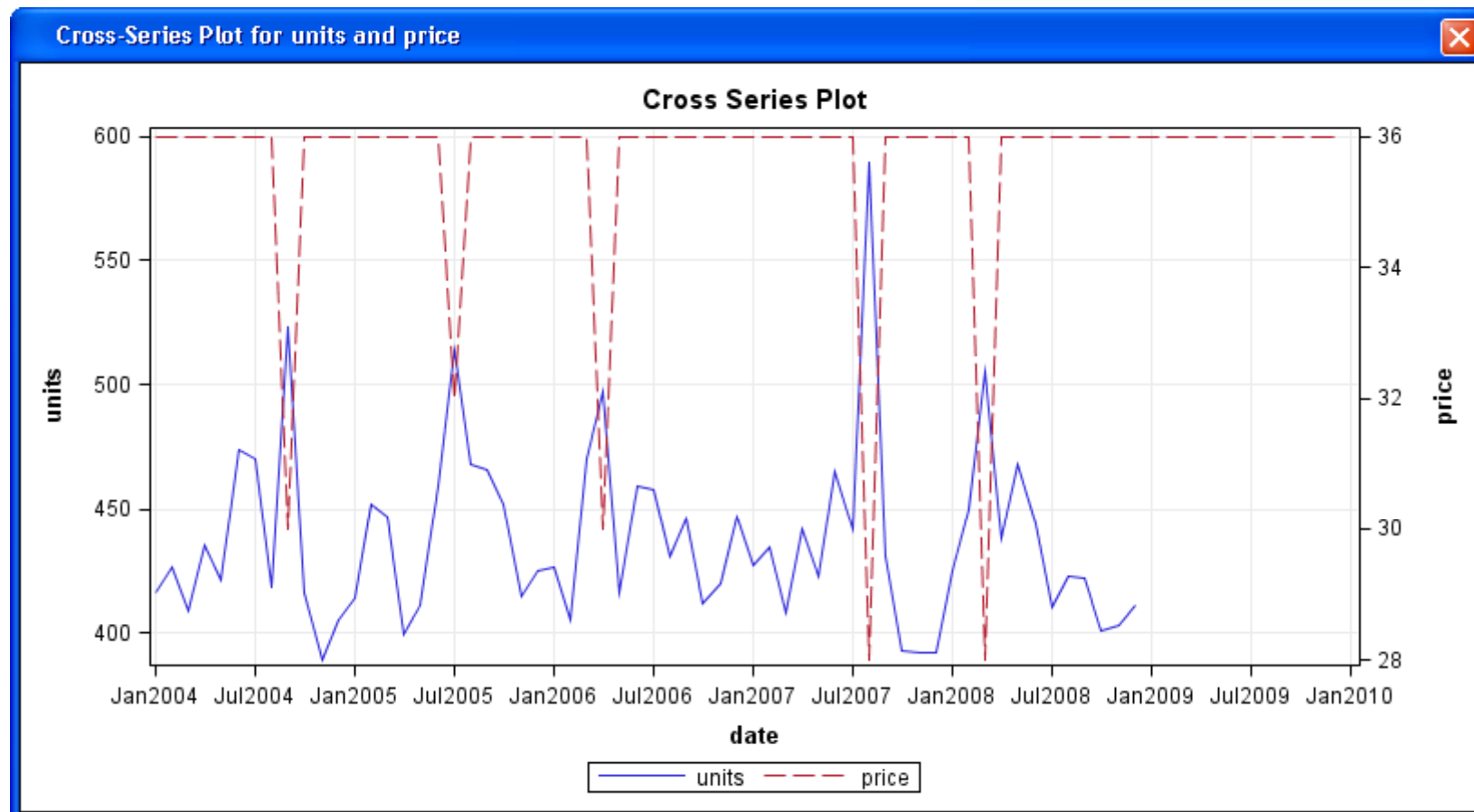




# Vergleichs- graphiken für Modelle



Szenario-Analysen behandeln die Frage „Wie wirkt sich die Veränderung der Variable „PREIS“ auf die Zielvariable „VERKAUFTE STÜCK“ aus?“



Erzeugen eines Szenarios, das auf einem der Forecast Modelle und einer unabhängigen Variable (PREIS) basiert

Create New Scenario

Name: PriceEffect2009

Description: test unit impact of dropping sales price |

Choose a model to create a scenario:

Model	Type	Rank	/ MAPE	All Input...	No Input...	price Used
<b>Generated ARIMA Model (HPF2_165)</b>	<b>Generated</b>	<b>1</b>	<b>4.21</b>	<b>X</b>		<b>X</b>
Generated ARIMA Model (HPF2_164)	Generated	2	4.89	X		X
Generated Smoothing Model (HPF2_166)	Generated	3	5.96		X	

Quick View...

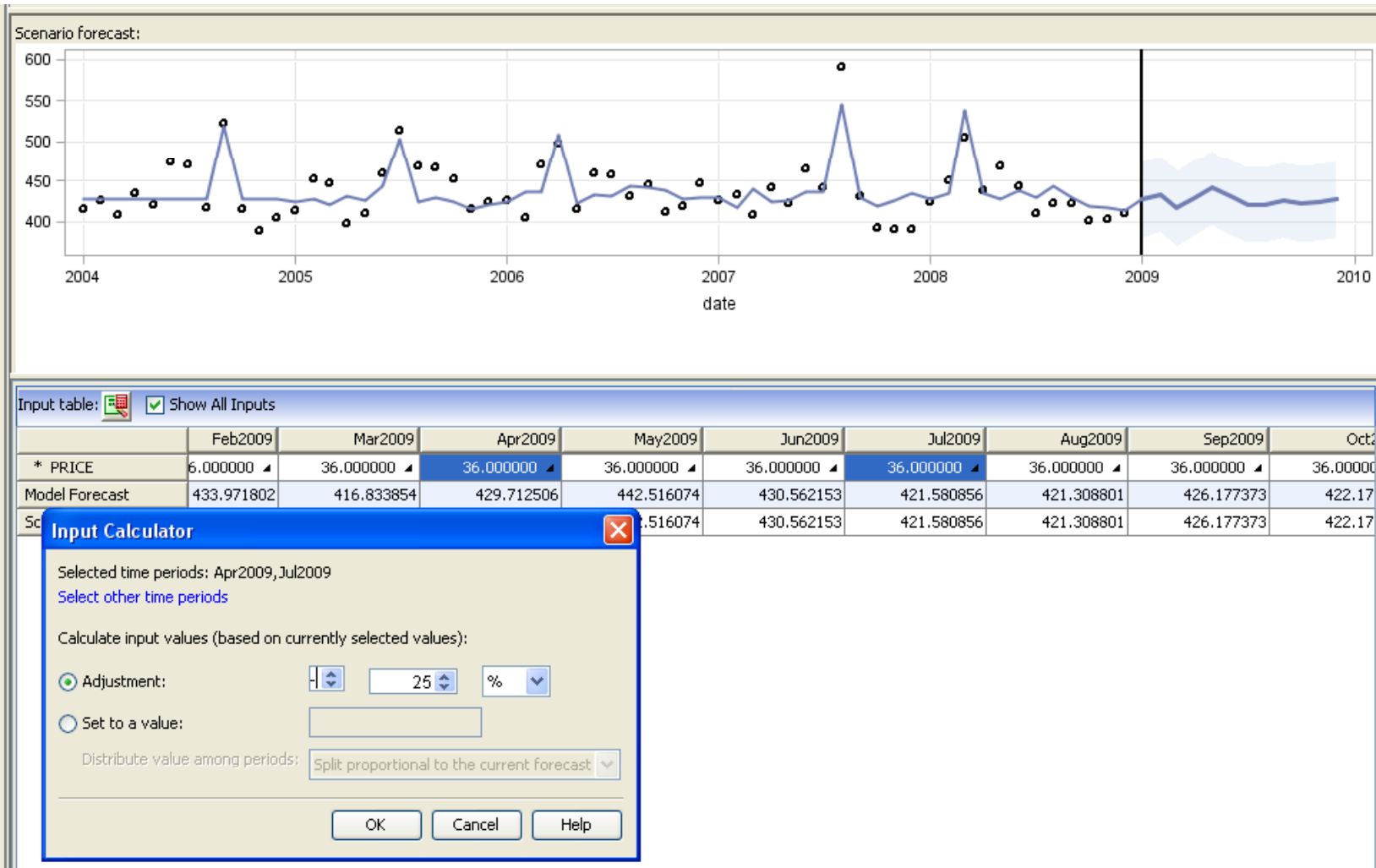
OK

Cancel

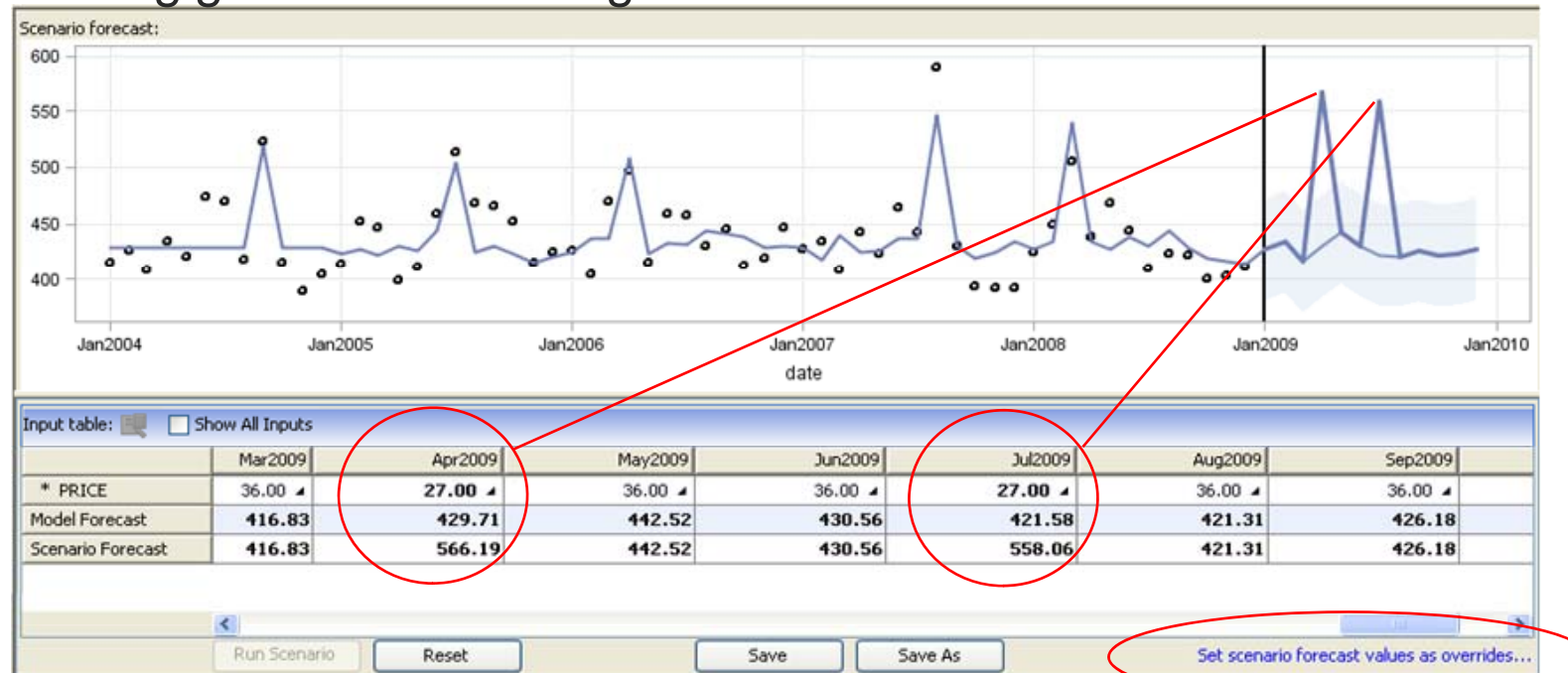
Help

Eingeben neuer Werte für die unabhängige Variable (PREIS)

Verwendung des Forecast-Modells um Vorhersagen für „VERKAUFTE STÜCK“ zu prognostizieren



Ein neuer SZENARIEN FORECAST wird für die abhängige Variable erzeugt

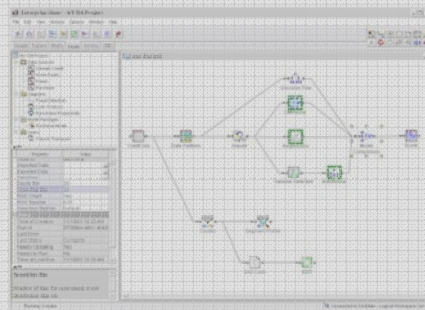


Scenario values can be saved as overrides in the forecasting view, and the project can be reconciled.

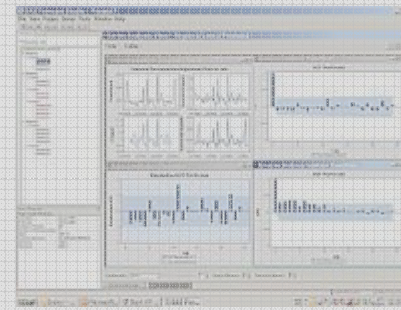
Active series

	Mar2009	Apr2009	May2009	Jun2009	Jul2009
Historical Data	.	.	.	.	.
Forecast Model	416.83	429.71	442.52	430.56	421.58
Reconciled Forecast	416.83	429.71	442.52	430.56	421.58
Override	416.83 ▲	566.19 ▲	442.52 ▲	430.56 ▲	558.06 ▲

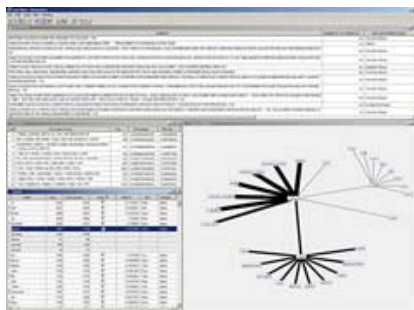
# Neuerungen in SAS®9.2



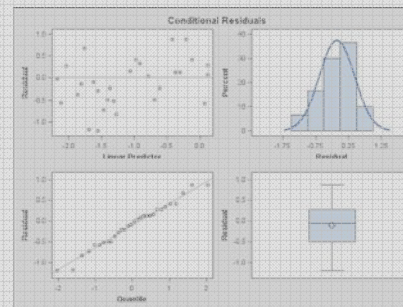
SAS®Enterprise Miner



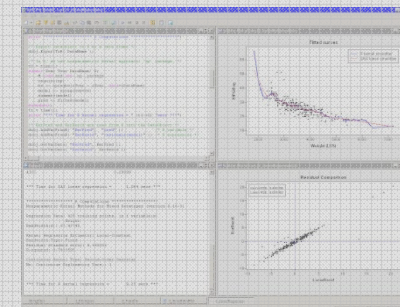
SAS®Forecast Server



SAS®Text Miner



SAS®STAT (ODS Graphics)



SAS®IML-Studio (R-Integration)

# Text Mining Definition

Der Prozess der **ENTDECKUNG** und **EXTRAKTION** bedeutender Muster und Zusammenhänge aus großen Textsammlungen



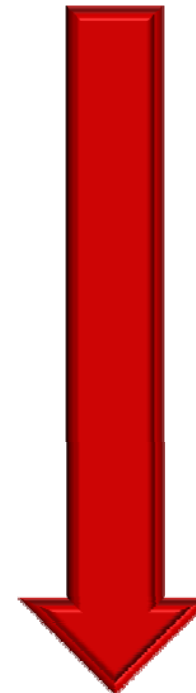


# SAS® Content Categorization

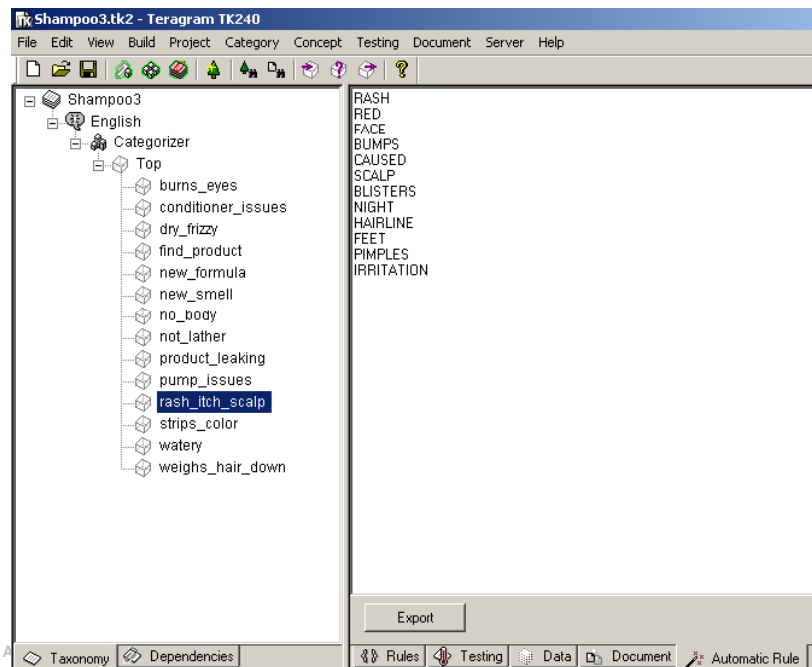
## ■ DEFINE:

- Kategorien
- Konzepte (entities/facts/events)

TOP



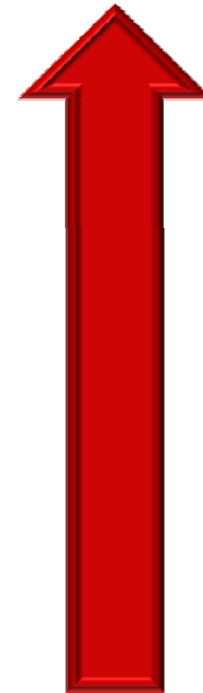
DOWN



# SAS® Text Miner: Discovery

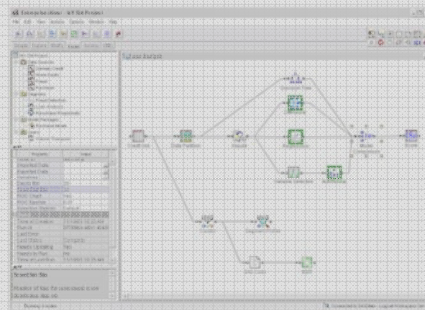
- Unstrukturierte und strukturierte Daten
- Explorative Analyse und Visualisierung
- Automatische Klassifikation
- Predictive Modeling

UP

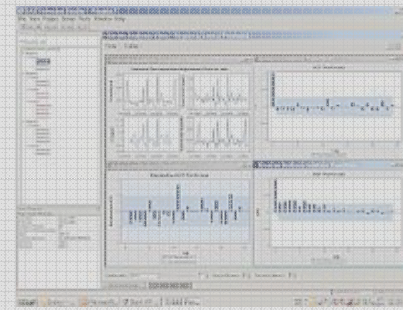


BOTTOM

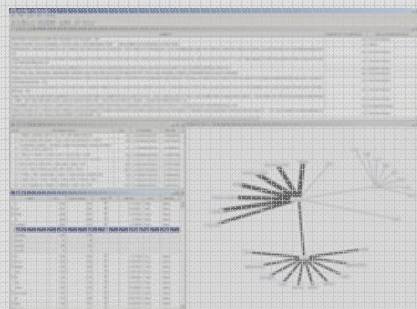
# Neuerungen in SAS®9.2



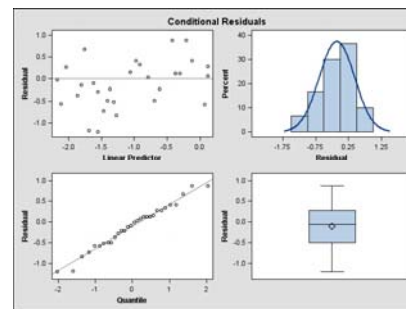
SAS®Enterprise Miner



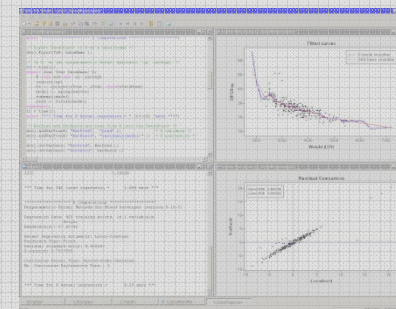
SAS®Forecast Server



SAS®Text Miner



SAS®STAT (ODS Graphics)



SAS®IML-Studio (R-Integration)

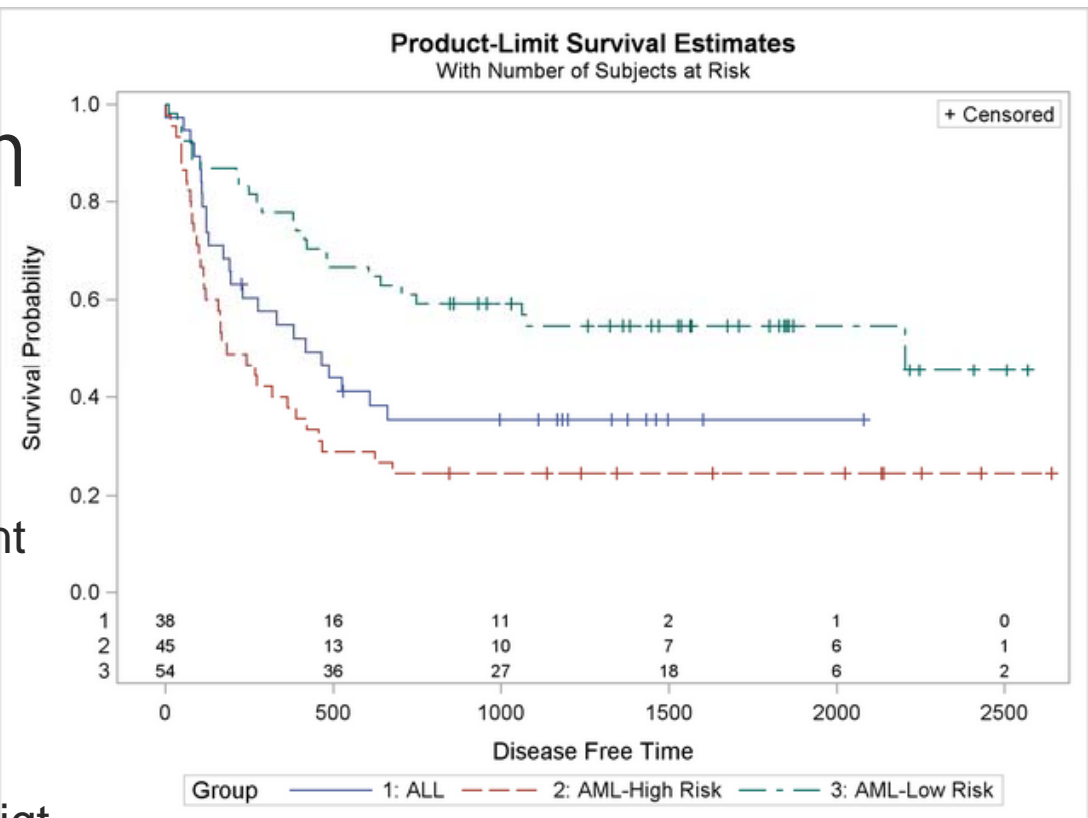
# Neue Optionen in den Survival Analysis Procedures

## ■ Proc LIFETEST

- SURVIVAL statement ermöglicht die Berechnung von Konfidenz-Bändern für die Survivor Function  $S(t)$
- Number of subjects at risk kann in Kaplan Meier Kurven angezeigt werden
- Smoother hazard function basierend auf der Kernel Method kann spezifiziert werden

## ■ Proc PHREG

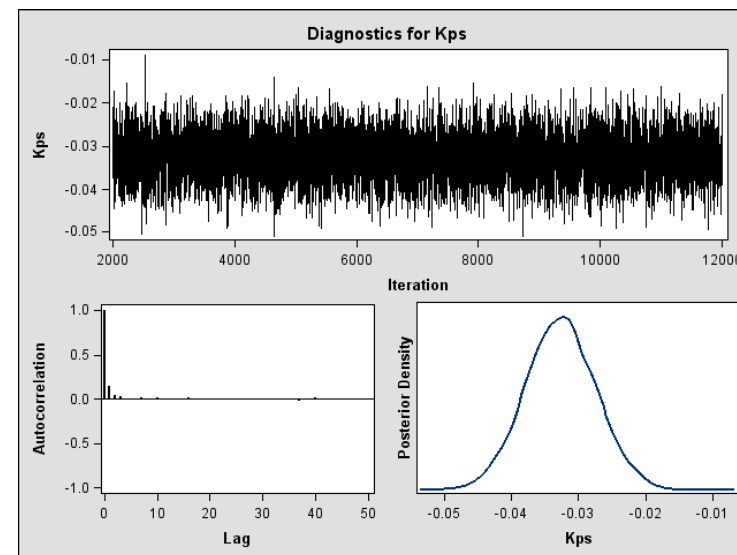
- CLASS statement ist verfügbar
- HAZARDRATIO statement erlaubt die Berechnung von hazard ratios bei Interaktionen
- Firth's penalized likelihood method is verfügbar



# Bayes Analysen in SAS - Überblick

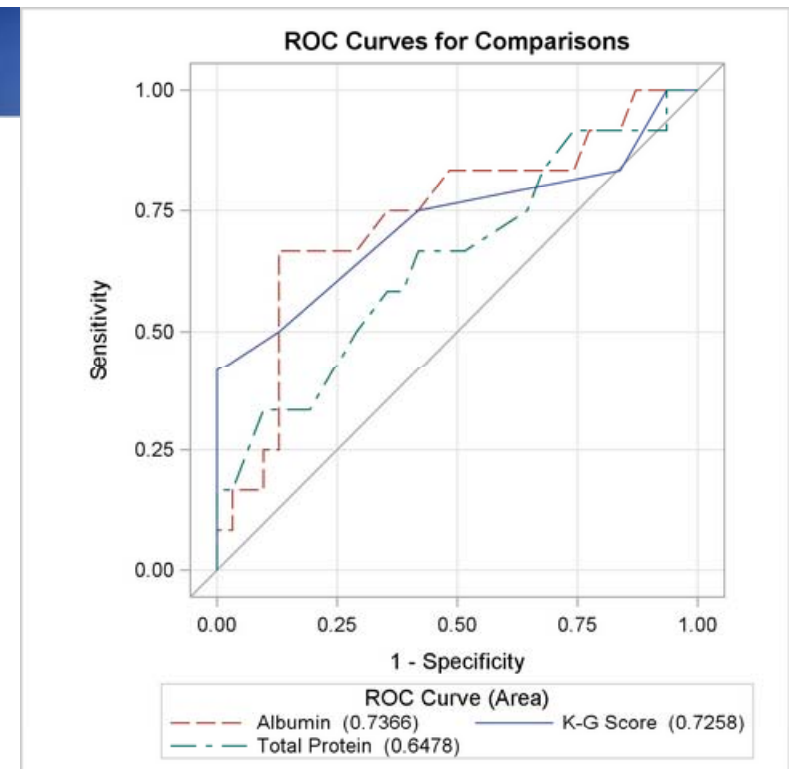
- Bayes-Analysen wurde zu bestehenden Prozeduren hinzugefügt
  - BAYES statement in Proc GENMOD, Proc LIFEREG, Proc PHREG
  - Gibbs sampling
- Proc MCMC
  - Markov Chain Monte Carlo simulations
  - Flexible simulation-basierte procedure die für eine große Bandbreite von Bayes-Modellen geeignet ist

Experimental

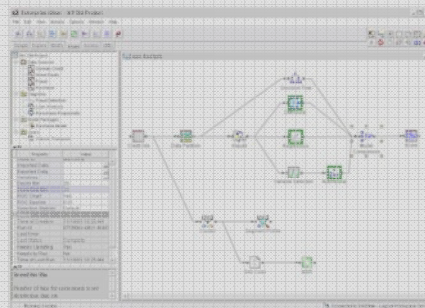


# Proc LOGISTIC

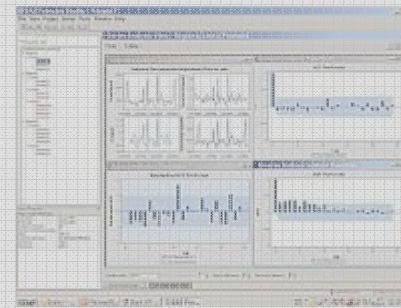
- Modellen können mit der OUTMODEL und INMODEL Option eingelesen und weiterverwendet werden
- SCORE statement erlaubt das Scoring neuer Beobachtungen
  - ROC Statistiken werden auch für die neuen Beobachtungen berechnet
- Odds ratios für Interaktionen werden berechnet
- ROCCONTRAST vergleicht unterschiedliche ROC Modelle
- Performs Firth's penalized maximum likelihood



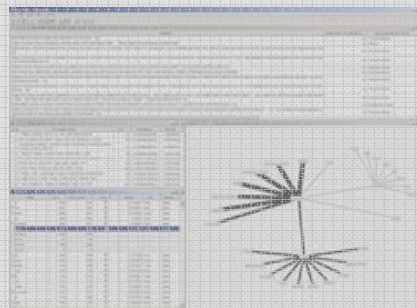
# Neuerungen in SAS®9.2



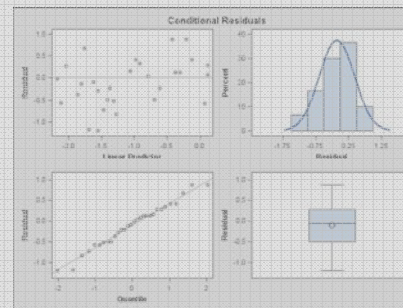
SAS®Enterprise Miner



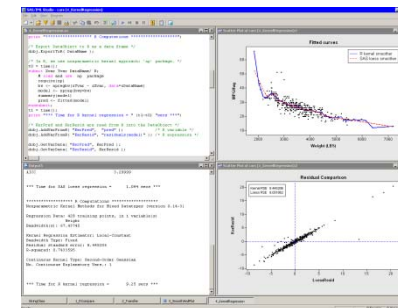
SAS®Forecast Server



SAS®Text Miner

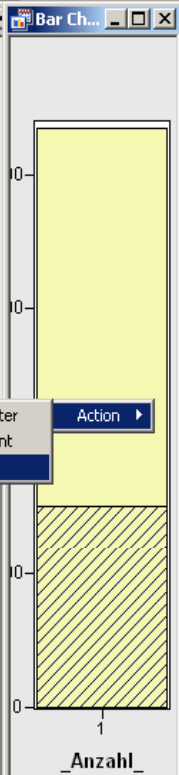
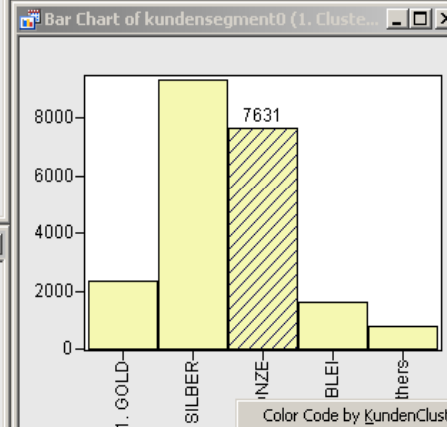
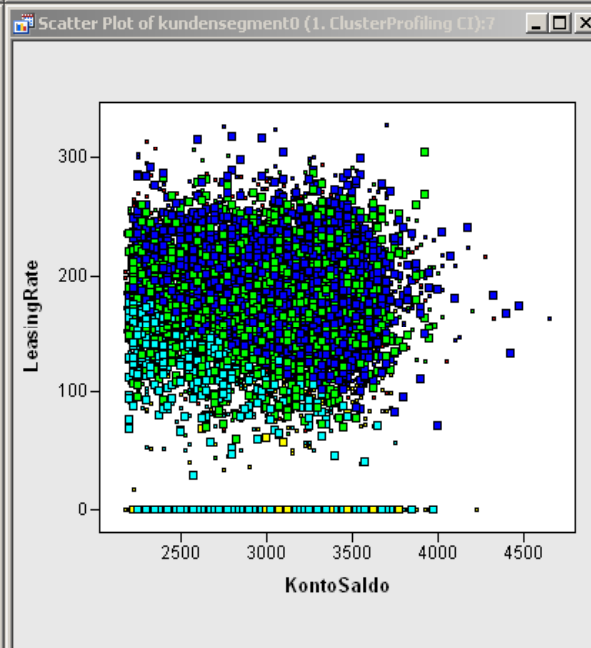
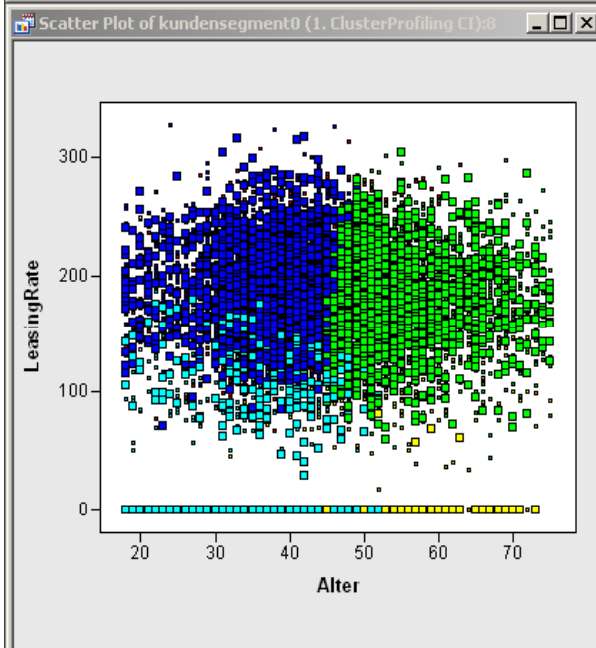
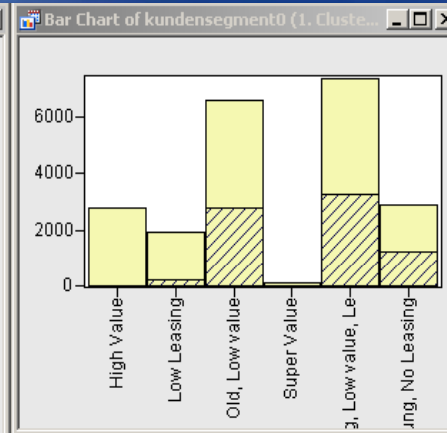
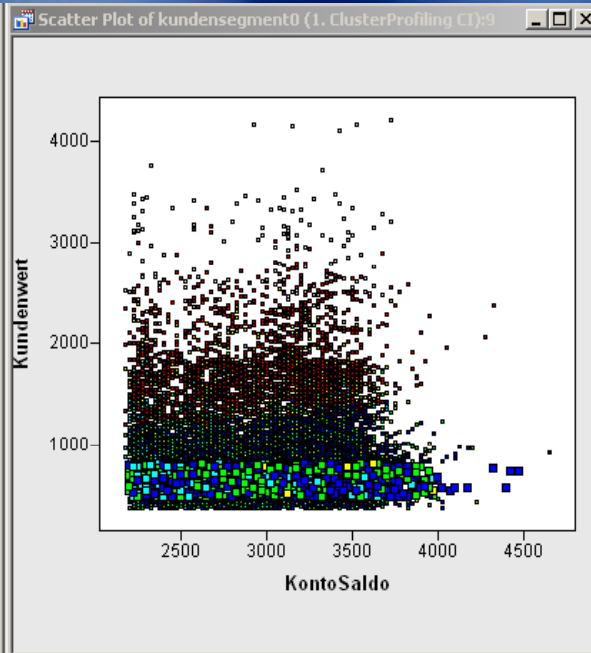
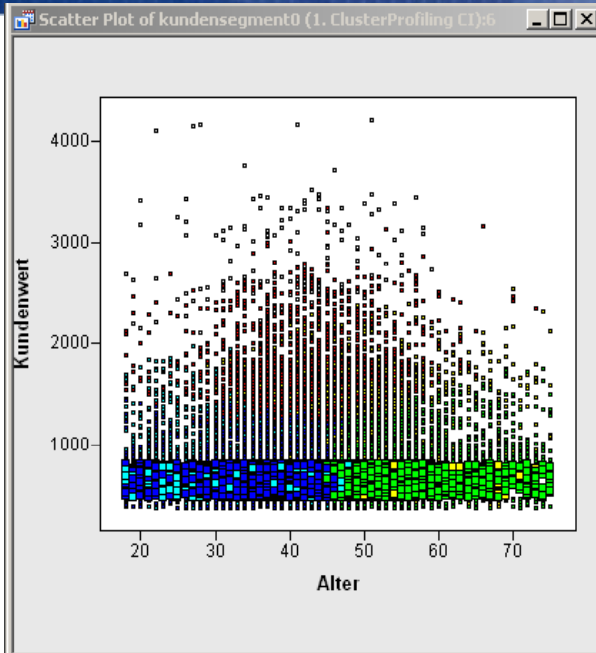


SAS®STAT (ODS Graphics)



SAS®IML-Studio (R-Integration)





Color Code by KundenCluster  
Color Code by WertSegment  
Export Selection

Action

## Aufruf von "R"-Routinen

```
y = {60 70 54 56 ... 48 52 49 53}`;  
x = t( 1:nrow(y) );  
run ExportMatrixToR( y, "Ry" );
```

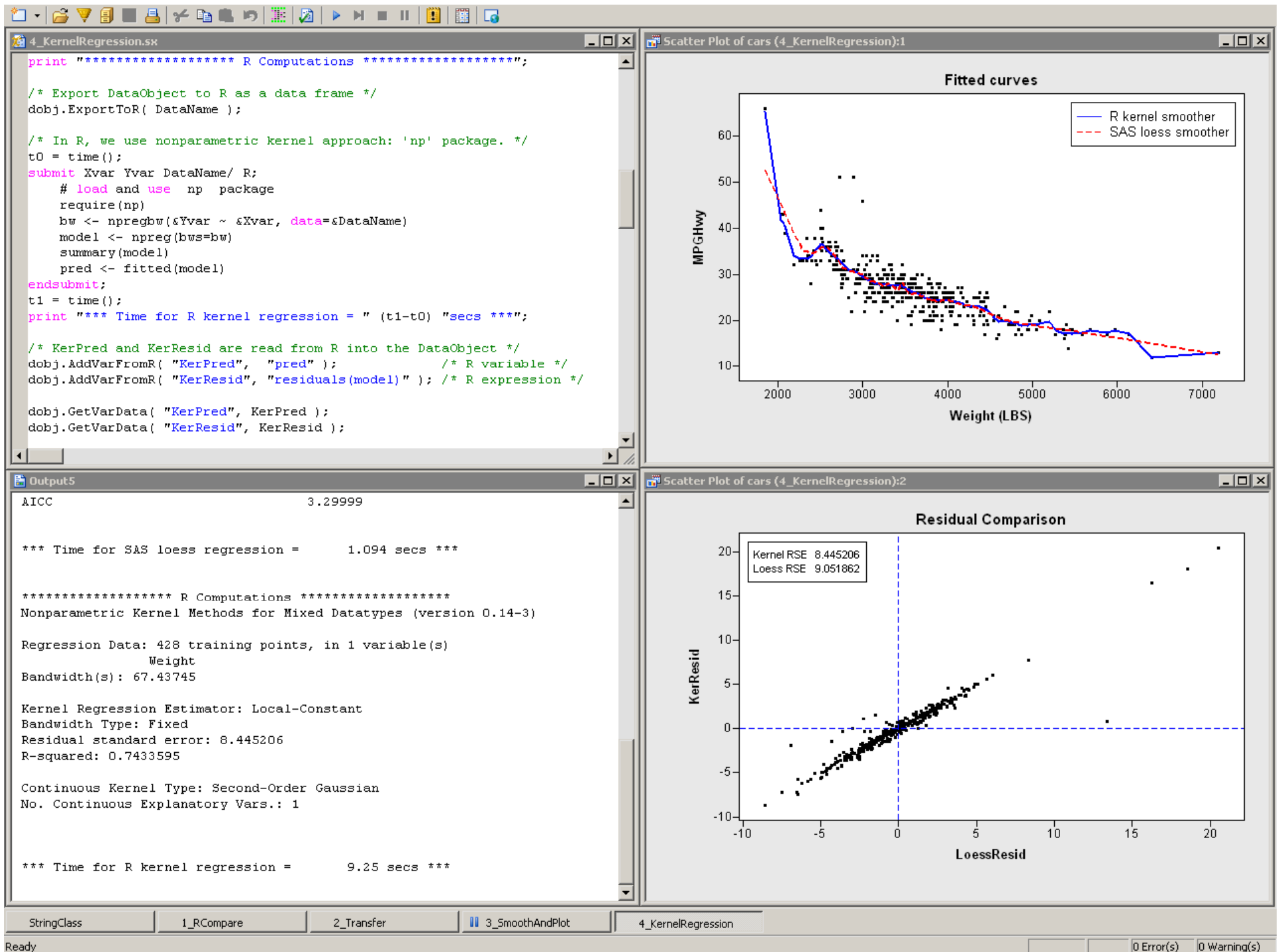
```
submit / R;  
    Rs <- smooth( Ry)    # time series smoother  
    Rs <- as.matrix(Rs)  # Rs is an object, not matrix  
endsubmit;
```

```
run ImportMatrixFromR( s, "Rs" );
```

```
declare ScatterPlot plot;  
plot = ScatterPlot.Create("Tukey", x, y);  
plot.DrawUseDataCoordinates();  
plot.DrawLine( x, s );
```

[RSmooth](#)

[RKernelReg](#)



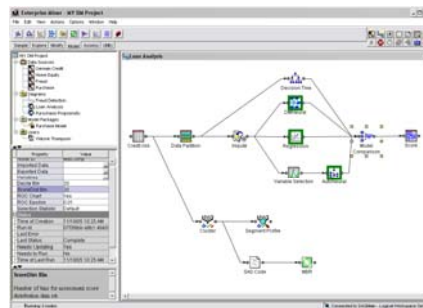
# Aufruf von SAS Procedures

```
declare DataObject dobj;
dobj = DataObject.CreateFromFile( "Hurricanes" );
/* write only a subset of variables */
vars = { "wind_kts" "min_pressure" "latitude"};
dobj.WriteVarsToServerDataSet( vars, "Work", "In", true );

submit;
proc glm data=In;
    model wind_kts = min_pressure min_pressure*min_pressure;
    output out=ProcOut P=Pred R=Residual;
    ods output ParameterEstimates = ParamEst;
run;
endsubmit;
```

[ProcGLM](#)

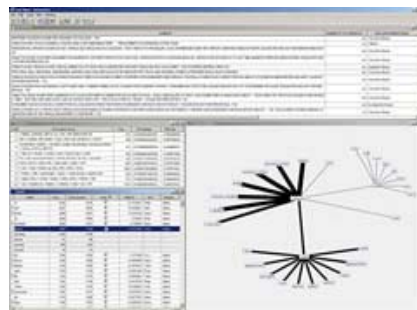
# Neuerungen in SAS®9.2



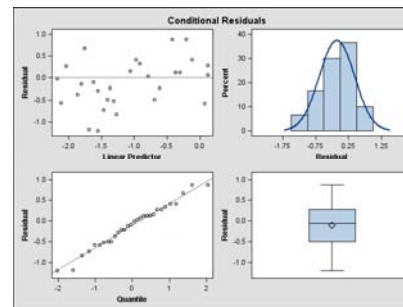
**SAS®Enterprise Miner**



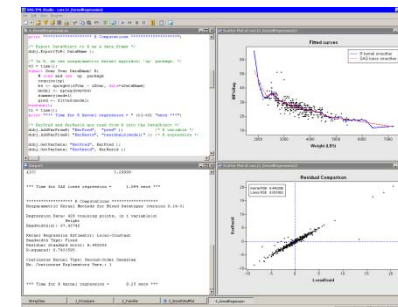
**SAS®Forecast Server**



**SAS®Text Miner**



**SAS®STAT (ODS Graphics)**



**SAS®IML-Studio (R-Integration)**

# Links and other resources

- What's new in SAS Enterprise Miner 6.1  
<http://support.sas.com/resources/papers/proceedings09/14-2009.pdf>
- What's new in SAS®STAT 9.2  
<http://support.sas.com/documentation/cdl/en/whatsnew/61982/HTML/default/statugwhatsnew.htm>
- What's new in SAS®STAT 9.0, 9.1  
<http://support.sas.com/documentation/whatsnew/91x/statugwhatsnew900.htm>
- An Introduction to Quantile Regression and the QUANTREG Procedure  
<http://www2.sas.com/proceedings/sugi30/213-30.pdf>
- Introducing the GLMSELECT PROCEDURE for Model Selection  
<http://www2.sas.com/proceedings/sugi31/207-31.pdf>
- Robust Regression and Outlier Detection with the ROBUSTREG Procedure  
<http://www2.sas.com/proceedings/sugi27/p265-27.pdf>
- An introduction to partial least squares regression  
<http://support.sas.com/techsup/technote/ts509.pdf>
- Sample-Size Analysis in Study Planning: Concepts and Issues, with Examples Using PROC POWER and PROC GLMPOWER  
<http://www2.sas.com/proceedings/sugi29/211-29.pdf>
- Updates to SAS® Power and Sample Size Software in SAS/STAT® 9.2  
<http://www2.sas.com/proceedings/forum2008/368-2008.pdf>
- A Comparison of the Mixed Procedure and the Glimmix Procedure  
<http://www2.sas.com/proceedings/sugi31/189-31.pdf>
- Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models  
<http://www2.sas.com/proceedings/sugi30/196-30.pdf>
- Growing Up Fast: SAS 9.2 Enhancements to the GLIMMIX Procedure  
<http://www2.sas.com/proceedings/forum2007/177-2007.pdf>
- Old versus New: A Comparison of PROC LOGISTIC and PROC GLIMMIX  
<http://www2.sas.com/proceedings/forum2008/226-2008.pdf>
- Advanced Statistical and Graphical features of SAS® PHREG  
<http://www2.sas.com/proceedings/forum2008/375-2008.pdf>
- SAS Online Help for SAS®STAT, Chapter 7: Introduction to Bayesian Analysis Procedures
- Data Preparation for Analytics  
[http://www.sascommunity.org/wiki/Data\\_Preparation\\_for\\_Analytics](http://www.sascommunity.org/wiki/Data_Preparation_for_Analytics)
- Makewide and Makelong Macro  
[http://www.sascommunity.org/wiki/Gerhard%27s\\_Samples](http://www.sascommunity.org/wiki/Gerhard%27s_Samples)