Example Test Sets

The following table is a summary of the example datasets provided in the TESTSET sub-directory.

Name	Columns	Rows	Description
LOGIC	5	4	Contains a truth table for 2 input logical AND, OR and XOR operations.
ENCODE	11	8	Contains a truth table for an 8 input to 3 output binary encoder. If the inputs and outputs are reversed then the table becomes a 3 to 8 decoder.
AIR	13	108	Contains a table of number of airplane tickets sold by month for 9 years. The table is arranged 13 months wide with the first 12 months being the previous 12 months and the NEXT column being the next months number of seats. The final 12 rows are reserved for testing.
VEL	7	609	Contains the distance traveled by a projectile using different angles and initial velocity. Additional columns of angle and velocity are included with random noise added.
COATING	8	128	Contains the results of a coating experiment. Different levels of starch, latex, coating weight, bonding agent and calender pressure are visited and the effects on opacity, brightness and gloss are recorded.
SODIUM	6	220	Contains the results of a designed experiment. Different gases and mixtures were tested to see what combination of gas, time and temperature could be used to convert Na_2SO_4 to Na_2S the most yield in the shortest time.
REDWOOD	15	72	Contains the results of a designed experiment. Different species of wood chips were tested to see if less expensive mixes could be used to make paper board while still guaranteeing a minimum strength and yield.
RING	15	507	Contains a process log of 14 sensors from a paper machine along with one laboratory measurement. The purpose of the log is to see if any process variables could be used to predict the lab variable.
SPECIES	5	2000	Contains a process log of 4 sensors along with 1 field that calculates the wood species exiting a wood digestor.
NOX	23	1340	Contains a process log of 23 sensors from a power boiler. The purpose of the log was to see if the process variables could be used to predict stack gases emitting from the boilers smoke stack.
CLO2	6	30	Contains the results of a designed experiment. Different levels of chemicals were tested to find the ideal setpoints needed to produce ClO_2 most efficiently.
CLOSTAT1	9	15	Contains the results of a designed experiment. Different stream setpoints were simulated to find the most economical setpoints.

PEAK4	3	121	Contains the results of stepping angles X and Y (11 steps) from 0 to and evaluating Z = sin(X) sin(Y)
CURL	9	70	Contains the results of a designed experiment. Paper machine variables were varied to discover any major effects on paper curl.
STR4	24	1178	Contains a process log of a paper machine. The purpose of the log was to see if the process variables could be used to best predict strength properties.

To import any of the aforementioned datasets into the NNMODEL issue the **Import Data From ASCII File** command from the File menu. The files are found in the \nnmodel\testsets sub-directory. Once a raw file has been imported the data matrix can be saved in binary format and reloaded at any time using the Save or Open commands in the File menu.

Example: LOGIC Dataset

<u>LOGIC Detailed Description</u>

File Name - LOGIC.RAW

Description: This dataset contains a truth table of three logical operations (i.e. AND, OR and XOR). The experiment is designed to show the results of the three separate logical operations given the same inputs. The data entered into the table has been translated from the logical language into a numerical representation (i.e. 0 = FALSE and 1 = TRUE).

Column Names	**Column Description**
IN1	First input into the logical operation
IN2	Second input into the logical operation
AND	Logical AND results
OR	Logical OR results
XOR	Logical XOR results

Data Analysis Analysis is not needed due to the small size of the dataset.

Model Building It is suggested to develop 4 models with this dataset. Build a separate model for each of the logical operations and an all inclusive model. The 4 models built are:

```
AND   : IN(IN1, IN2) => OUT(AND)
OR    : IN(IN1, IN2) => OUT(OR)
XOR   : IN(IN1, IN2) => OUT(XOR)
LOGIC : IN(IN1, IN2) => OUT(AND, OR, XOR)
```

The previous notation reads: Model LOGIC has IN1 and IN2 as inputs and generates AND, OR and XOR as outputs.
After creating each model select **Initialize** and **Start Training** commands from the Model menu.

Model Analysis All four models were created and trained using the initial factory default settings for the training parameters. After training the following model statistics were reported.

Analysis of model AND

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
IN1	0.500000	0.577350	0.000000	1.000000	1.000000
IN2	0.500000	0.577350	0.000000	1.000000	1.000000
Measured	0.250000	0.500000	0.000000	1.000000	0.750000
Predicted	0.232930	0.498595	-0.126712	0.970787	0.745791
Residual			-		

	0.017070	0.081808	0.059520	0.126712	0.020078
R Square			0.973230		

Analysis of model OR

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
IN1	0.500000	0.577350	0.000000	1.000000	1.000000
IN2	0.500000	0.577350	0.000000	1.000000	1.000000
Measured	0.750000	0.500000	0.000000	1.000000	0.750000
Predicted	0.754372	0.491654	0.019418	1.056489	0.725170
Residual	-0.004372	0.041884	-0.056489	0.034900	0.005263
R Square			0.992983		

Analysis of model XOR

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
IN1	0.500000	0.577350	0.000000	1.000000	1.000000
IN2	0.500000	0.577350	0.000000	1.000000	1.000000
Measured	0.500000	0.577350	0.000000	1.000000	1.000000
Predicted	0.500708	0.574554	-0.001126	0.999090	0.990337
Residual	-0.000708	0.004522	-0.007405	0.002535	0.000061
R Square			0.999939		

Analysis of model LOGIC

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
IN1	0.500000	0.577350	0.000000	1.000000	1.000000
IN2	0.500000	0.577350	0.000000	1.000000	1.000000
AND					

Measured	0.250000	0.500000	0.000000	1.000000	0.750000
Predicted	0.215229	0.517050	-0.197364	0.969989	0.802023
Residual	0.034771	0.119030	-0.086510	0.197364	0.042504
R Square			0.943328		
OR					
Measured	0.750000	0.500000	0.000000	1.000000	0.750000
Predicted	0.830179	0.532606	0.075665	1.322547	0.851008
Residual	-0.080179	0.175099	-0.322547	0.088393	0.091979
R Square			0.877361		
XOR					
Measured	0.500000	0.577350	0.000000	1.000000	1.000000
Predicted	0.502921	0.339727	0.247580	1.001628	0.346244
Residual	-0.002921	0.471165	-0.418133	0.655659	0.665989
R Square			0.334011		

After reviewing the above model statistics it was noted that the first three separate models predicted the output very well. However, the results of the LOGIC model showed a significant loss of accuracy (as measured by R Square) when combining the three logic functions. The all inclusive model cannot predict as well as the separate models because the default training parameters did not allow the model to build up enough internal complexity. The following table demonstrates that selecting any type of training that will raise the internal complexity will also result in better models. The highlighted model was the initial factory default parameters model shown above.

Training Type	**Count**	**Options**	**AND**	**OR**	**XOR**
AI	1000		0.943328	0.877361	0.334011
Standard 4 Hid	1000		0.894768	0.999530	0.999996
AI	1000	Connect I/O	0.991743	0.996463	0.993545
AI	5000		1.000000	1.000000	1.000000

Standard 4 Hid	1000	CG Train	0.999996	0.999999	0.999997
Equal Spaced	1000		0.999997	0.999999	1.000000

Example: ENCODE Dataset

<u>ENCODE Detailed Description</u>
File Name - ENCODE.RAW

Description: This dataset contains a truth table of three logical operations (i.e. AND, OR and XOR). The experiment is designed to show the results of the three separate logical operations given the same inputs. The data entered into the table has been translated from the logical language into a numerical representation (i.e. 0 = FALSE and 1 = TRUE).

Column Names	**Column Description**
IN1	Input 1 to encoder or output from decoder
IN2	Input 2 to encoder or output from decoder
IN3	Input 3 to encoder or output from decoder
IN4	Input 4 to encoder or output from decoder
IN5	Input 5 to encoder or output from decoder
IN6	Input 6 to encoder or output from decoder
IN7	Input 7 to encoder or output from decoder
IN8	Input 8 to encoder or output from decoder
OUT1	Output 1 from encoder or input to decoder
OUT2	Output 2 from encoder or input to decoder
OUT3	Output 3 from encoder or input to decoder

Data Analysis The following truth table was used as the dataset.

Encoder/Decoder Truth Table

IN1	IN2	IN3	IN4	IN5	IN6	IN7	IN8	OUT1	OUT2	OUT3
0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0

Model Building It is suggested to develop 2 models from this dataset. Build a encoder model (ENCODE) using IN1 through IN8 as inputs and OUT1 - OUT3 as outputs and build a decoder model (DECODE) using OUT1-OUT3 as inputs and IN1-IN8 as outputs. The 2 models built are:

```
ENCODE   : IN(IN1,...,IN8)   => OUT(OUT1,...,OUT3)
DECODE   : IN(OUT1,...,OUT3) => OUT(IN1,...,IN8)
```

Model Analysis Both models were created and trained using the initial factory default settings plus **Standard BEP** for the training parameters. After training the following model statistics were reported:

Model ENCODE

Predicted	R Square

Outputs	
OUT1	1.000000
OUT2	1.000000
OUT3	1.000000

Model DECODE

Predicted Outputs	R Square
IN1	0.903213
IN2	0.903793
IN3	0.903345
IN4	0.903565
IN5	0.903188
IN6	0.903522
IN7	0.904014
IN8	0.905515

With digital type functions it is hard to get a picture of how well these models are doing. The best way with these particular models is to interactively test them. This can be done using the **Interrogate Model** command in the Model menu.

Example: AIR Dataset

AIR Detailed Description
File Name - AIR.RAW

Description: This dataset was constructed to demonstrate how a neural model can be used to predict a time series. It contains 12 columns of the number of tickets sold during the previous twelve months followed by the number of tickets sold during the next month. The dataset was generated from the following table titled **Airline Ticket Sales 1980-1989** by re-arraigning the first 9 rows for use as a training matrix and the last row as a test matrix.

Airline Ticket Sales 1980-1989

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1980	145	153	171	167	157	179	190	192	178	153	133	151
1981	155	163	183	172	162	194	221	220	204	172	144	182
1982	188	197	231	212	224	231	256	263	242	211	190	218
1983	224	234	253	235	237	286	300	313	275	249	223	253
1984	254	257	309	306	295	315	345	356	308	271	235	261
1985	267	243	303	295	304	344	394	375	338	295	261	299
1986	319	304	342	350	355	410	473	452	402	360	309	365
1987	367	362	413	408	416	487	537	527	460	397	347	399
1988	411	393	464	455	461	546	604	608	523	452	399	438
1989	439	413	468	449	473	565	641	656	527	469	401	439

The following table demonstrates how the previous table was rearranged to be used as a training matrix.

Re-arraigned Ticket Sales

M1	M2	m3	M4	M5	M6	M7	M8	M9	M10	M11	M12	NEXT
145	153	171	167	157	179	190	192	178	153	133	151	155
153	171	167	157	179	190	192	178	153	133	151	155	163
171	167	157	179	190	192	178	153	133	151	155	163	183
167	157	179	190	192	178	153	133	151	155	163	183	172

and so on...

Column Names	**Column Description**
M1	The number of tickets sold twelve months ago
M2	The number of tickets sold eleven months ago
M3	The number of tickets sold ten months ago
M4	The number of tickets sold nine months ago
M5	The number of tickets sold eight months ago
M6	The number of tickets sold seven months ago
M7	The number of tickets sold six months ago
M8	The number of tickets sold five months ago
M9	The number of tickets sold four months ago
M10	The number of tickets sold three months ago
M11	The number of tickets sold two months ago
M12	The number of tickets sold last month
NEXT	The number of tickets that will be sold this month

Data Analysis A **By Row Matrix** graph was printed to see the monthly trend and verify that there were no gross errors in the dataset.

Model One model was constructed from this dataset:

Building

```
AIR   : IN(M1,M2,...,M12) => OUT(NEXT)
```

Model Analysis

The model was created and trained using the initial factory default settings for the training parameters. After training the following model statistics were reported.

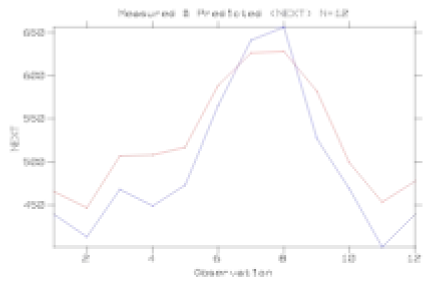
Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
M1	277.88541	93.690969	133.00000	537.00000	833909.78
M2	280.65624	93.652983	133.00000	537.00000	833233.70
M3	283.15624	93.412425	133.00000	537.00000	828958.70
M4	286.20833	94.489759	133.00000	537.00000	848189.88
M5	289.20833	95.234110	133.00000	537.00000	861605.89
M6	292.37499	95.843983	133.00000	537.00000	872676.56
M7	296.19791	98.555103	133.00000	546.00000	922745.29
M8	300.51041	102.82453	133.00000	604.00000	1004423.9
M9	304.84375	106.88885	133.00000	608.00000	1085396.5
M10	308.43750	108.36874	133.00000	608.00000	1115659.5
M11	311.55208	108.15085	133.00000	608.00000	1111177.6
M12	314.32292	106.92875	144.00000	608.00000	1086207.0
NEXT					
Measured	317.31250	106.32475	144.00000	608.00000	1073970.6
Predicted	319.78079	105.50861	171.21496	591.72290	1057546.5
Residual	-2.468285	16.031915	-51.89569	34.427551	24417.117
R Square			0.977265		

To see how the model predicts the next twelve months select **Use Test Matrix** from the Model menu and re-run the model statistics.

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
NEXT					
Measured	495.00001	84.744527	401.00000	656.00000	78997.984
Predicted	524.93071	65.081795	446.55816	628.02099	46592.040
Residual	-29.9307	26.557598	-59.2269	27.979004	7758.3662

0	2
R Square	0.901790

As you can see, the worst case under prediction was around 59 and the worst case over prediction was 28 seats. The following plot graphically demonstrates the result.



The command used was the **Measured and Predicted** command from the
Graph menu.

Example: VEL Dataset

VEL Detailed Description
File Name - VEL.RAW

Description: This dataset was constructed to demonstrate how well a neural model can predict a trajectory. It contains the distance measurement, the angle of launch and the initial velocity. Along with the aforementioned columns the dataset also includes the aforementioned columns with noise added, plus a column of just noise so that you can experiment building neural models with noisy signals and compare them with ideal models.

Column Names	Column Description
ANGLE	Angle measured from horizontal
VEL	Initial velocity
RANGLE	Angle with Gaussian noise added
RVEL	Initial velocity with Gaussian noise added
NOISE	Just Gaussian noise
DIST	Distance traveled by projectile
RDIST	Distance traveled by projectile with Gaussian noise added

Data Analysis A statistics report was generated using the **Basic Statistics** command in the Data menu. This gives us an overall picture of the dataset. If correlations are of interest they can be viewed using the **Correlation Analysis** command also in the Data menu.

By viewing the data matrix it can be observed that the initial velocity was varied from 0 to 100 by 5 and the launch angle was varied from 3 to 87 by 3.

Variable	N	Mean	Std Dev	Minimum	Maximum
ANGLE	609	45.000000	25.120434	3.000000	87.000000
VEL	609	50.000000	30.301392	0.000000	100.00000
RANGLE	609	45.141823	25.438328	-5.960000	92.910004
RVEL	609	49.991724	30.296755	0.000000	102.98999
NOISE	609	-0.023645	3.387809	-10.18000	11.090000
DIST	609	61.804992	65.890137	0.000000	258.10000
RDIST	609	61.801954	65.948653	-0.040000	262.64001

Model Building Two separate models were constructed from this dataset. The first model uses simply the initial velocity and the launch angle:

```
VEL1   : IN(VEL,ANGLE) => OUT(DIST)
```

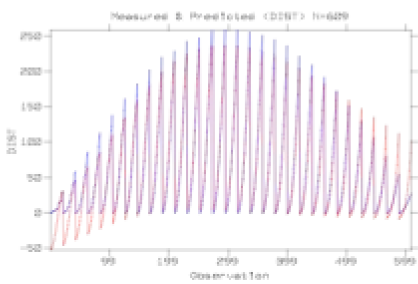
After the previous model was analyzed, and determined to be not good enough, a second model was constructed that used trigonometric functions as inputs rather than the simple angle:

```
VEL2   : IN(VEL,SANG,CANG) => OUT(DIST)
```

Model Analysis The model was created and trained using the initial factory default settings for the training parameters plus CG. CG training was added because a trajectory is known to be trigonometric in nature and harder training is necessary. After training the following model statistics were reported.

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
ANGLE	45.000000	25.120434	3.000000	87.000000	383670.01
VEL	49.999999	30.301391	0.000000	100.00000	558249.97
DIST					
Measured	61.804992	65.890137	0.000000	258.09997	2639638.2
Predicted	58.920517	66.373302	-52.79543	236.05107	2678492.4
Residual	2.884475	15.524349	-73.25933	52.795437	146531.29
R Square			0.944488		

Although the R Square statistic is respectable, a closer examination using the **Measured and Predicted** or **Measured vs. Predicted** graphs reveal significant problems predicting the distance when the angle is near 0 or 90 degrees. The following graph demonstrates the problem.



Therefore, a second model was created using calculated columns to provide more information. Two additional columns were created to include the sine and cosine of the launch angle into the model. To do this, first add the following two equations to the equation string of the data matrix:

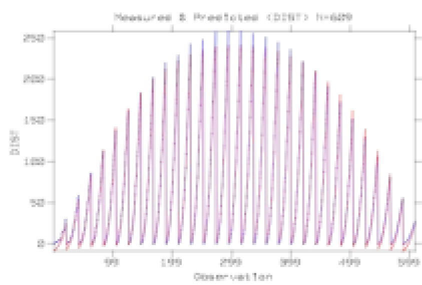
SANG = SIN (ANGLE * 2 * PI / 360)
CANG = COS (ANGLE * 2* PI / 360)

Then create the columns using the **Append Calculated Columns** command in the Edit menu. After training the following model statistics were reported.

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
VEL	49.999999	30.301391	0.000000	100.00000	558249.97
SANG	0.641180	0.298386	0.052336	0.998630	54.132656

CANG	0.641180	0.298386	0.052336	0.998630	54.132656
DIST					
Measured	61.804992	65.890137	0.000000	258.09997	2639638.2
Predicted	61.608327	65.665710	-8.983620	240.97804	2621687.1
Residual	0.196665	4.168274	-10.31471	18.355606	10563.698
R Square			0.995998		

The R Square statistic is better then the previous model and the **Measured and Predicted** or **Measured vs. Predicted** graphs reveal a significant increase in the overall accuracy.



Example: COATING Dataset

COATING Detailed Description
File Name - COATING.RAW

Description:	The coating dataset contains the data from an incomplete designed experiment. This experiment was designed to determine the ideal levels of the five independent variables (STARCH, LATEX, HP91, COATWT and CPSI) necessary to maintain minimum levels of the dependent variables (BRIGHTNESS, OPAC and GLOSS). In this dataset STARCH, LATEX, HP91 and COATWT are varied to five different levels while CPSI is varied to two levels. The independent variables STARCH, LATEX, HP91 and CPSI can set to the desired target and maintained, however, COATWT cannot controlled as accurately. Therefore, the targeted COATWT value is later replaced with the measured value.

Column Names	**Column Description**
STARCH	The percentage of starch added to the coating.
LATEX	The percentage of latex added to the coating. Latex is a rubber used as a binding agent in coatings.
HP91	The percentage of HP91 added to the coating. HP91 is a plastic pigment.
COATWT	The measured amount of coating applied to the paper.
CPSI	The pressure applied by a super-calander to polish the surface of the coated paper.
BRIGHT	The measured brightness of the finished paper/coating. Brightness is the measurement of how white the surface of the piece of paper is.
OPAC	The measured opacity of the finished paper/coating. Opacity is a measurement of how opaque (impenetrable to light) a piece of paper is.
GLOSS	The measured gloss of the finished paper/coating. Gloss is a measurement of how polished the surface of a piece of paper looks.

Data Analysis	A statistics report was generated using the **Basic Statistics** command in the Data menu. This gives us an overall picture of the dataset. If correlations are of interest they can be viewed using the **Correlation Analysis** command also in the Data menu.

Variable	N	Mean	Std Dev	Minimum	Maximum
STARCH	128	17.928281	5.206710	6.500000	26.000000
LATEX	128	12.720312	3.034249	6.500000	19.500000
HP91	128	5.735938	2.644463	0.000000	11.000000
COATWT	128	4.677344	0.904097	2.960000	6.380000
CPSI	128	45.500000	19.576621	26.000000	65.000000
BRIGHT	128	65.760938	1.091934	62.900002	68.400002
OPAC	128	70.977344	1.388775	67.699997	74.400002
GLOSS	128	39.965625	7.816129	25.600000	57.500000

Model Building	Four models were constructed from this dataset. The first model included all dependent variables into one model:

```
COATING   : IN(STARCH, LATEX, HP91, COATWT, CPSI)
```

```
                   => OUT(BRIGHT, OPAC, GLOSS)
```

The next three models were constructed to predict the dependent variables separately:

```
BRIGHT    : IN(STARCH, LATEX, HP91, COATWT, CPSI)
            => OUT(BRIGHT)
OPAC      : IN(STARCH, LATEX, HP91, COATWT, CPSI)
            => OUT(OPAC)
GLOSS     : IN(STARCH, LATEX, HP91, COATWT, CPSI)
            => OUT(GLOSS)
```

Model Analysis

The first model (COATING) was created and trained using the initial factory default settings for the training parameters plus **Standard BEP**. After training the following model statistics were reported.

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
STARCH	17.928281	5.206710	6.500000	26.000000	3442.9479
LATEX	12.720312	3.034249	6.500000	19.500000	1169.2470
HP91	5.735938	2.644464	0.000000	11.000001	888.13480
COATWT	4.677344	0.904097	2.960000	6.380000	103.80870
CPSI	45.500000	19.576621	26.000000	65.000000	48672.000
BRIGHT					
Measured	65.760938	1.091934	62.900002	68.400002	151.42473
Predicted	65.907051	1.093083	63.297192	68.363007	151.74355
Residual	-0.146113	0.368337	-1.586678	0.660271	17.230382
R Square			0.886212		
OPAC					
Measured	70.977344	1.388775	67.699997	74.400002	244.94435
Predicted	70.890938	1.287656	67.926071	74.046333	210.57338
Residual	0.086406	0.491884	-1.036316	2.010513	30.727583
R Square			0.874553		
GLOSS					
Measured	39.965625	7.81612	25.600000	57.500000	7758.6686

		9			
Predicted	39.983750	7.332670	25.655918	58.590076	6828.5421
Residual	-0.018125	2.200280	-6.198959	5.021023	614.83627
R Square			0.920755		

The next three models (BRIGHT, OPAC and GLOSS) were trained using the same training parameters as the first model. This shows that modeling the dependent variables separately can produce higher R Square models under identical conditions.

BRIGHT					
Measured	65.760938	1.091934	62.900002	68.400002	151.42473
Predicted	65.877290	1.118312	63.187656	68.545723	158.82889
Residual	-0.116352	0.321667	-1.404861	0.710152	13.140674
R Square			0.913220		
OPAC					
Measured	70.977344	1.388775	67.699997	74.400002	244.94435
Predicted	71.015069	1.306490	68.109215	74.472168	216.77825
Residual	-0.037725	0.450619	-1.255150	1.767052	25.788266
R Square			0.894718		
GLOSS					
Measured	39.965625	7.816129	25.600000	57.500000	7758.6686
Predicted	39.664502	7.135378	26.043440	55.892056	6466.0295
Residual	0.301123	2.021777	-4.901318	5.476116	519.12310
R Square			0.933091		

The performance of the first model can be increased by tweaking the training parameters. In this case **Connect IO** and **CG Training** was added to the default settings. After training the following model statistics were reported.

BRIGHT					
	65.7609		62.9000	68.4000	151.424

Measured	38	1.091934	02	02	73
Predicted	65.760366	1.048789	63.203236	68.333611	139.69475
Residual	0.000572	0.305982	-1.240593	0.699280	11.890349
R Square			0.921477		
OPAC					
Measured	70.977344	1.388775	67.699997	74.400002	244.94435
Predicted	70.978319	1.314241	67.848198	74.354477	219.35800
Residual	-0.000975	0.438371	-1.064926	1.595451	24.405517
R Square			0.900363		
GLOSS					
Measured	39.965625	7.816129	25.600000	57.500000	7758.6686
Predicted	39.904586	7.683301	24.798870	58.689686	7497.2046
Residual	0.061039	1.899198	-4.346085	4.689075	458.08291
R Square			0.940959		

The final models were exported to a system optimizer to find the answer to: What is the lowest cost coating mixture that can still meet the minimum specifications of BRIGHT, OPAC and GLOSS? In the optimizer the cost of the coating was calculated by the following equation:

```
COST = C1COATWT(C2LATEX + C3STARCH + C4HP91)
```

The solution to the problem would minimize COST while maximizing BRIGHT, OPAC and GLOSS and subject to the following constraints: BRIGHT > 71.5, OPAC > 78 and GLOSS > 48.

Optimization can not be performed in this version of the program.

Example: SODIUM Dataset

SODIUM Detailed Description

File Names - H2.RAW, CO.RAW, COH2.RAW,MIX.RAW,COH2MIX.RAW

Description: This dataset is really made up of 5 separate datasets. It is the result of a chemical experiment to determine the best way to reduce sodium sulfate to sodium sulfide using hydrogen, carbon monoxide or a mixture of both.

The plan was to run each experiment to 160 minutes twice, however, the mixture experiment could not be run longer then 70 minutes due to a problem with the experimental apparatus. The data before sixty minutes is not of any use (all the important stuff happens from 60 to 160 minutes). Due to this problem the MIX experiment yielded only one point per run.

H2	The result of a designed experiment using only hydrogen gas as the agent and varying temperature and gas concentration.
CO	The result of a designed experiment using only carbon monoxide gas as the agent while varying temperature and gas concentration.
COH2	The result of combining both the H2 and CO datasets into one using the **Concatenate Data Matrices** command in the Data menu. The combining of these two datasets is straight forward in that the two experimental designs are similar. It involves creating a new field in both matrices and setting the missing values to zero.
MIX	The result of a designed experiment using a mixture of both hydrogen and carbon monoxide gases as the agent while varying the gas concentrations and temperatures.
COH2MIX	The combined dataset of COH2 and MIX experiments. Combining these two datasets is mechanically easy in that both matrices have the same fields. However, statistically the dataset are very different. COH2 contains experimental runs where time varies from 60 to 160 and MIX only contains the 60 minute values. It is okay to paste these datasets together as long as the consequences are understood. The MIX data will serve as reference points the model must traverse. The MIX data is very important to the model because it contains the only points where both gases are present at the same time. Other reference points could also be entered in this manner (i.e. H2 = 0, CO = 0 and CONV = 0).

Column Names	**Column Description**
TIME	Time elapsed since beginning of the run
H2	Percentage of hydrogen gas used
CO	Percentage of carbon monoxide gas used
TEMP	Temperature during the run
AVTEMP	Average temperature of run

CONV	Percentage of Na2SO4 converted

Data Analysis

H2 and CO contain a central composite design varying concentration of the gas and the reaction temperature. Each run was replicated twice. The design yielded a total of 10 runs. The MIX experiment is a mixture design where the concentrations of H2 and CO are varied and the temperature is held constant at the center point. The following **Basic Statistics** reports were generated for all the datasets.

H2

Variable	N	Mean	Std Dev	Minimum	Maximum
TIME	110	110.00000	31.767504	60.000000	160.00000
H2	110	50.000000	22.463018	25.000000	75.000000
TEMP	110	1203.8877	18.054523	1179.9599	1225.8900
AVTEMP	110	1203.7799	18.949085	1181.9000	1226.1999
CONV	110	0.837782	0.089136	0.629880	0.997350

CO

Variable	N	Mean	Std Dev	Minimum	Maximum
TIME	110	110.00000	31.767504	60.000000	160.00000
CO	110	27.000000	20.241659	5.000000	50.000000
TEMP	110	1200.5162	19.210413	1173.4699	1223.4499
AVTEMP	110	1199.7699	20.422422	1174.6999	1221.9000
CONV	110	0.665540	0.210730	0.163860	0.979830

COH2

Variable	N	Mean	Std Dev	Minimum	Maximum
TIME	220	110.00000	31.694892	60.000000	160.00000
CO	220	13.500000	19.672548	0.000000	50.000000
H2	220	25.000000	29.647857	0.000000	75.000000
TEMP	220	1202.2020	18.675406	1173.4699	1225.8900
AVTEMP	220	1201.7749	19.756972	1174.6999	1226.1999
CONV	220	0.751661	0.183050	0.163860	0.997350

MIX

Variable	N	Mean	Std Dev	Minimum	Maximum
TIME	8	60.000000	0.000000	60.000000	60.000000
CO	8	28.125000	20.863074	0.000000	50.000000
H2	8	29.687500	28.298079	0.000000	75.000000
AVTEMP	8	1202.9000	2.988080	1199.3000	1206.9000

CONV	8	0.599325	0.252241	0.000000	0.758300
		COH2MIX.DM			
TIME	229	108.03493	32.553311	60.000000	160.00000
CO	229	13.951965	19.829147	0.000000	50.000000
H2	229	25.491266	29.901225	0.000000	100.00000
TEMP	229	1202.2224	18.311378	1173.4699	1225.8900
AVTEMP	229	1201.8122	19.371543	1174.6999	1226.1999
CONV	229	0.746724	0.186770	0.000000	0.997350

Model Building

Many models were built during the course of the analysis, but only the last model is reported. The most complete model was built from the COH2MIX dataset.

```
CONV   : IN(TIME, CO, H2, TEMP)
               => OUT(CONV)
```

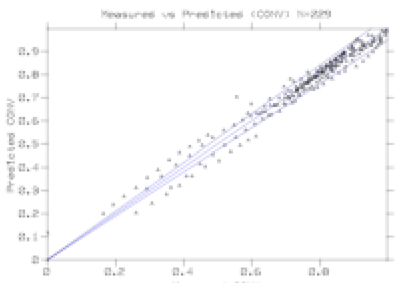
Model Analysis

Model (CONV) was created and trained using the initial factory default settings for the training parameters plus **Standard BEP** and **CG Optimization** .After training the following model statistics were reported.

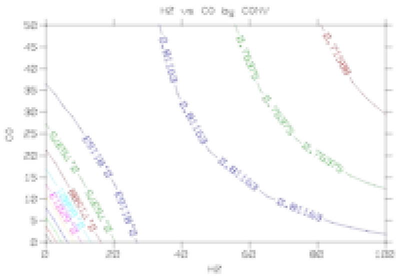
CONV

Measured	0.746724	0.186770	0.000000	0.997350	7.953296
Predicted	0.746847	0.182348	0.116489	0.991212	7.581214
Residual	-0.000123	0.034309	-0.147950	0.078872	0.268383
R Square			0.966255		

A **Measured vs. Predicted** graph was generated to view how the model performed. This graph demonstrates that the model seems to predict CONV fairly well. The blue lines represent the ± 5% tolerance band.



The following contour graph was generated to demonstrate the surface of the CONV variable in relation to the concentrations of H2 and CO, given TEMP=1200 degrees and TIME=110 minutes.



Example: REDWOOD Dataset

REDWOOD Detailed Description
File Name - REDWOOD.RAW

Description: The redwood experiment was done to see if redwood chips could be used to replace the less available Douglas fir chips in making wood pulp for container board. A designed experiment was done to set the various percentages of DFIR, HFIR, PINE, REDW and cooking temperatures. A COOK number was included in the dataset for identification purposes only. After each batch cook the pulp properties TYLD, BPH and KAPN were measured. These pulps were refined to three different levels of (REVS) and the pulp property CSF was measured. Finally paper was made from the pulp batches and the following physical measurements were made on the paper TEAR, BURST, FOLD, SCOT and PORS.

Column Names	Column Description
COOK	The batch number of the cook.
REVS	The number of revolutions the pulp was refined to.
DFIR	The percentage of Douglas fir chips used in the pulp.
HFIR	The percentage of Hemlock fir chips used in the pulp.
PINE	The percentage of Pine chips used in the pulp.
REDW	The percentage of Redwood chips used in the pulp.
TEMP	The temperature the chips were cooked at.
TYLD	The percentage of pulp made as a fraction of total chips (pulp test).
BPH	The pH of the cook (pulp test).
KAPN	The Kappa number (pulp test)
CSF	The freeness number. (pulp test).
BURST	The result of the burst test (paper test).
FOLD	The result of the fold test (paper test).
SCOT	The Scott Bond test (paper test).
PORS	The porosity measurement (paper test).

Data Analysis A statistics report was generated using the **Basic Statistics** command in the Data menu. This gives us an overall picture of the dataset. If correlations are of interest they can be viewed using the **Correlation Analysis** command also in the Data menu.

Variable	N	Mean	Std Dev	Minimum	Maximum
COOK	72	252.66666	7.253945	241.00000	266.00000
REVS	72	2520.0000	2072.0106	0.000000	5040.0000
DFIR	72	0.267500	0.134675	0.080000	0.430000
HFIR	72	0.245000	0.088795	0.130000	0.340000
PINE	72	0.280000	0.068669	0.170000	0.340000
REDW	72	0.062500	0.057132	0.000000	0.130000
TEMP	72	447.00000	8.056141	439.00000	455.00000
TYLD	72	0.681931	0.034129	0.603000	0.796000
TEAR	72	27.023889	4.171943	21.440001	36.639999
BPH	69	15.636232	0.596798	14.300000	16.500000
KAPN	72	79.548611	6.466925	69.099998	96.699997
CSF	72	621.61111	42.098997	536.00000	672.00000
BURST	72	5.463750	1.255601	3.230000	6.900000

FOLD	72	2464.0694	715.55764	984.00000	4070.0000
SCOT	72	0.169958	0.070813	0.039000	0.299000
PORS	72	4.729708	2.154128	1.442000	7.824000

Model Building

4 models of unrefined pulp properties were constructed from this dataset. The pulp properties modeled are TYLD, BPH and KAPN and the only numbers to be included into the model(s) are when the REVS is equal to zero (definition of unrefined). To exclude all other rows of data except the REVS=0 add to the exclusions string the following formula:

XIF (REVS != 0)

The first model included all independent variables (except REVS) of the pulp cook into one model predicting the pulp properties:

```
PULP    : IN(DFIR, HFIR, PINE, REDW, TEMP)
            => OUT(TYLD, BPH, KAPN)
```

The next three models were constructed to predict the dependent variables separately:

```
TYLD    : IN(DFIR, HFIR, PINE, REDW, TEMP)
            => OUT(TYLD)
BPH     : IN(DFIR, HFIR, PINE, REDW, TEMP)
            => OUT(BPH)
KAPN    : IN(DFIR, HFIR, PINE, REDW, TEMP)
            => OUT(KAPN)
```

One model of refined pulp properties was created to predict CSF. This is the only pulp property (in this experiment) that varies with REVS so it is treated separately:

```
CSF     : IN(DFIR, HFIR, PINE, REDW, TEMP, REVS)
            => OUT(CSF)
```

Finally a model is constructed to predict all paper properties:

```
ALL     : IN(DFIR, HFIR, PINE, REDW, TEMP, REVS)
            => OUT(TEAR, BURST, FOLD, SCOT, PORS)
```

Model Analysis

The first model (PULP) was created and trained using the initial factory default settings for the training parameters plus **Standard BEP**, **CG Training** and **Connect IO**. After training the following model statistics were reported.

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
DFIR	0.262174	0.137112	0.080000	0.430000	0.413591
HFIR	0.250000	0.088626	0.130000	0.340000	0.172800
PINE	0.277391	0.070014	0.170000	0.340000	0.107843
REDW	0.065217	0.057672	0.000000	0.130000	0.073174
TEMP	446.65217	8.17206	439.00000	455.00000	1469.2173

		3			
TYLD					
Measured	0.681913	0.037294	0.617000	0.796000	0.030598
Predicted	0.681315	0.029041	0.634339	0.765630	0.018554
Residual	0.000598	0.021773	-0.041405	0.049940	0.010430
R Square			0.659141		
BPH					
Measured	15.634783	0.608731	14.300000	16.500000	8.152171
Predicted	15.629808	0.554882	14.611290	16.605103	6.773670
Residual	0.004975	0.267684	-0.486740	0.735995	1.576404
R Square			0.806628		
KAPN					
Measured	79.917391	6.442589	69.199997	96.699997	913.15291
Predicted	79.998469	6.075874	71.612045	92.992897	812.15743
Residual	-0.081079	2.005435	-3.490204	3.707100	88.478890
R Square			0.903106		

After viewing the rather low R Square statistic it was decided to create separate models to increase the performance. The following three models were trained using the same parameters as the previous model.

TYLD					
Measured	0.682625	0.036640	0.617000	0.796000	0.030878
Predicted	0.682765	0.035731	0.625390	0.797267	0.029364
Residual	-0.000140	0.008722	-0.019358	0.021583	0.001750
R Square			0.943335		
BPH					
Measured	15.634783	0.608731	14.300000	16.500000	8.152171

Predicted	15.633953	0.548757	14.371428	16.485111	6.624947
Residual	0.000830	0.236424	-0.522560	0.495054	1.229720
R Square			0.849154		
KAPN					
Measured	79.550000	6.552994	69.199997	96.699997	987.65988
Predicted	79.535829	6.383595	69.700935	92.437851	937.25656
Residual	0.014170	1.421623	-2.365013	4.262146	46.483297
R Square			0.952936		

A single model was constructed to predict CSF. The following model was trained using the same parameters as the first model.

CSF					
Measured	621.61111	42.098997	536.00000	672.00000	125835.11
Predicted	621.70089	41.537126	535.97607	670.44061	122498.63
Residual	-0.089779	6.975161	-13.03338	21.181396	3454.3534
R Square			0.972549		

A single model was constructed to predict all paper properties. The following model was trained using the same parameters as the first model.

TEAR					
Measured	27.023889	4.171943	21.440001	36.639999	1235.7629
Predicted	27.003455	3.946271	21.924404	35.265640	1105.6870
Residual	0.020434	1.333415	-2.842812	3.702446	126.23768
R Square			0.897846		
BURST					
Measured	5.463750	1.255601	3.230000	6.900001	111.93390
Predicted	5.464549	1.239189	3.265778	6.739976	109.02691
Residual	-		-		

	0.000799	0.195209	0.414065	0.485160	2.705563
R Square			0.975829		
FOLD					
Measured	2464.0694	715.55765	984.00006	4070.0000	36353615.
Predicted	2464.0329	639.97692	1189.0734	3518.6354	29079503.
Residual	0.036491	323.19581	-498.2634	998.62963	7416343.1
R Square			0.795994		
SCOT					
Measured	0.169958	0.070813	0.039000	0.299000	0.356031
Predicted	0.170150	0.066934	0.050595	0.260065	0.318087
Residual	-0.000192	0.023071	-0.059773	0.053572	0.037792
R Square			0.893851		
PORS					
Measured	4.729708	2.154128	1.442000	7.824000	329.45902
Predicted	4.724768	2.139457	1.539140	7.431902	324.98666
Residual	0.004940	0.242487	-0.495115	0.500623	4.174795
R Square			0.987328		

The final question. What mixture of wood chips, cooking temperature and REVS would allow us the meet the minimum paper properties while minimizing DFIR and maximizing TYLD?

subject to the following constraints:
FOLD > 2500
SCOT > 0.14
REDW > 0.10
DFIR+HFIR+PINE+REDW < 1.0

Optimization can not be performed in this version of the program.

Example: RING Dataset

<u>RING Detailed Description</u>
File Name - RING.RAW

Description: The RING dataset was captured during the normal operation of a paper machine. The intent of the data capture was to see if any of the standard logged process variables could be used to predict a physical property (MDRING) of the manufactured paper board. This experiment is really a fishing expedition in that no designed experiment was performed on the process variables. However, there may be enough information in the log to point to variables that have a major effect.

Column Names	**Column Description**
MDRING	Ring crush measured in machine direction
CONDWT	Basis weight measurement
AVEMO	Average moisture of the paper board measurement
SPEED	Machine speed measurement
FL1	Flow rate measurement
CS1	Consistancy measurement
FL2	Flow rate measurement
FL3	Flow rate measurement
FL4	Flow rate measurement
HP1	Horse power measurement
FL5	Flow rate measurement
FL6	Flow rate measurement
CS2	Consistancy measurement
AN1	Freeness measurement
CS3	Consistancy measurement

Data Analysis

A statistics report was generated using the **Basic Statistics** command in the Data menu. This gives us an overall picture of the dataset. With this much data it is highly recommended that the data be viewed using the **By Row Matrix** command in the graph menu.

Variable	N	Mean	Std Dev	Minimum	Maximum
MDRING	507	120.27810	15.541973	75.000000	150.00000
CONDWT	507	40.058619	3.296797	32.849998	46.259998
AVEMO	507	6.249132	0.443878	4.250000	7.990000
SPEED	507	2132.8500	125.89145	1606.0000	2305.0000
FL1	507	21.466075	2.696087	13.000000	26.100000
CS1	507	3.231894	0.412474	2.500000	5.410000
FL2	507	67.242604	11.245570	35.799999	103.00000
FL3	507	8704.5956	575.66911	6849.0000	9805.0000
FL4	507	51733.443	3700.4051	10000.000	61023.000
HP1	507	1.094359	0.303432	0.500000	2.150000
FL5	507	0.064083	0.084342	0.000000	0.470000
FL6	507	42.958383	6.451481	27.700001	58.900002
CS2	507	3.379487	0.317734	3.050000	4.100000
AN1	507	684.21696	63.229086	500.00000	800.00000
CS3	507	5.599053	0.678285	2.850000	6.730000

Model Building

A model was built that included all independent variables to predict the MDRING property:

```
MDRING    : IN(CONDWT, AVEMO, SPEED,FL1, CS1, FL2,
            FL3, FL4, HP1, FL5, FL6, CS2, AN1,
             CS3 ) => OUT(MDRING)
```

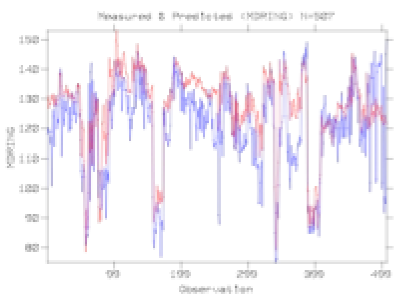
Model Analysis

The model was created and trained using the initial factory default settings for the training parameters plus **Standard BEP**. After training the following model statistics were reported.

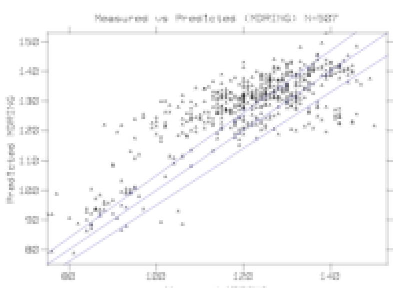
MDRING

Variable	Mean	Std Dev	Minimum	Maximum	Sum Sq
Measured	120.27810	15.541973	75.000000	150.00000	122225.78
Predicted	126.97214	13.398455	78.760902	152.96443	90836.403
Residual	-6.694042	9.308496	-33.88612	28.481262	43843.936
R Square			0.641287		

A **Measured and Predicted** graph was generated to view how the model performed as a time series. This graph demonstrates that the model seems to capture much of the variability, but there are major gaps.



A **Measured vs. Predicted** graph was also generated to demonstrate the lack of fit.



A sensitivity analysis was run to see which variables account for most of the variability of MDRING. The results are presented below.

Sensitivity Analysis of MDRING

Variable Name	Initial Setting	Percent Total
FL1	19.6	+0.13543
FL4	47204.5	+0.12555
HP1	1.33	+0.12218
CS2	3.58	-0.11213

SPEED	1955.5	-0.09160
AVEMO	6.12	-0.08170
FL5	0.24	-0.07192
CS1	3.96	-0.05456
AN1	650.0	+0.05361
CONDWT	39.56	+0.04894
CS3	4.79	-0.03340
FL6	43.3	+0.03116
FL2	69.4	-0.02752
FL3	8327.0	+0.01030

Example: SPECIES Dataset

SPECIES Detailed Description
File Name - SPECIES.RAW

Description: The species dataset was downloaded from a process control system in a paper mill. It was the result of an experiment to see if an algorithm could be developed that could predict when the wood species changed in the output of a continuous wood digestor. A continuous digestor converts wood chips into paper pulp. It is like a long pipe that you dump chips in a the top and pulp falls out at the bottom. The digestor is a hydraulic system that operates under high pressure and temperature. The inside of a digestor is a very corrosive and hence cant be well instrumented. The wood chips usually spend 3-5 hours making the trip from the top to the bottom.

Paper is made of a mixture of two species of wood (hardwood and softwood). Because the two species cook (digest) so differently they must be processed and stored separately. The ideal process would have two digestors (one for softwood and one for hardwood), however due to the expense, many mills have only one. In these mills the digestor is swung between the two species. Temperatures, chemicals, flows and cooking time vary between the two species. Pulp manufactured during this swing is called twilight pulp because it is nether hardwood or softwood. The twilight pulp must be treated as if it was hardwood thus reducing the profitability of the process. If s detector could be developed that could more exactly determine when the crossover was between the species the process would be more efficient.

The species dataset represents a 33 hour period. Each row is a one minute scan. Signal A3 was captured by an automatic sampling device that bottled the pulp. The A3 sample was then measured in a laboratory at a later time. The two questions to be answered by this experiment are 1) can the species change be detected and 2) what signals are the most important?

Column Names	**Column Description**
A1	Blow line gamma process measurement
A2	Refractivity index process measurement
A3	Softwood present calculation (laboratory test)
A4	Triple D calculation (from process measurements)
A5	Consistency process measurement

Data Analysis A statistics report was generated using the **Basic Statistics** command in the Data menu. This gives us an overall picture of the dataset. With this much data it is highly recommended that the data be viewed using the **By Row Matrix** command in the graph menu.

Variable	N	Mean	Std Dev	Minimum	Maximum
A1	2000	0.345829	0.125509	0.176045	0.715970
A2	2000	0.493294	0.166214	0.176530	0.715647
A3	2000	0.543000	0.498272	0.000000	1.000000

A4	2000	0.300089	0.150496	-0.074310	0.882878
A5	2000	0.366992	0.191050	0.136625	0.742250

Model Building

Three models were constructed to predict A3 from the input variables:

```
A3a    : IN(A1, A2, A4, A5) => OUT(A3)
A3b    : IN(A1, A4, A5)     => OUT(A3)
A3c    : IN( A4, A5)        => OUT(A3)
```

Signal A2 was eliminated from model A3b because it didnt appear to be significant. Likewise signals A1 and A2 were eliminated from model A3c.

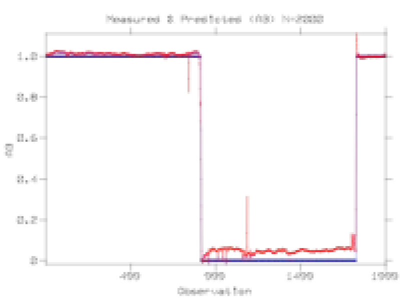
Model Analysis

The model was created and trained using the initial factory default settings for the training parameters. After training the following model statistics were reported.

A3

Measured	0.543000	0.498272	0.000000	1.000000	496.30200
Predicted	0.569940	0.478081	-0.018100	1.110421	456.89343
Residual	-0.026940	0.033566	-0.400754	0.499685	2.252166
R Square			0.995462		

A **Measured and Predicted** graph was generated to view how the model performed as a time series. This graph demonstrates that the model seems to predict A3 very well.



A sensitivity analysis was run to see if any of the variables could be eliminated from the model. The signal A2 is a candidate for elimination.

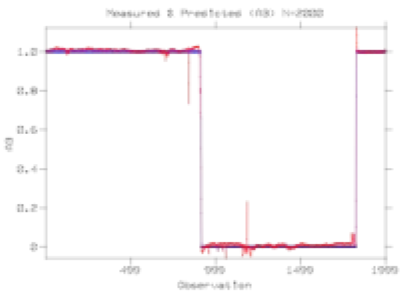
Sensitivity Analysis of A3

Variable Name	Initial Setting	Percent Total
A4	0.404284	+0.52183
A1	0.446008	-0.23331
A5	0.439438	-0.19154
A2	0.446089	+0.0533

Another model (without A2) was created to see if the performance is severely effected. As you can see from the statistics and the **Measured and Predicted** plot the performance actually increased.

A3

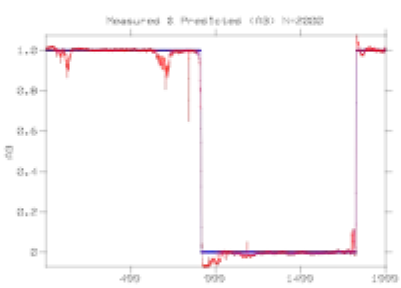
Measured	0.543000	0.498272	0.000000	1.000000	496.30200
Predicted	0.549675	0.496866	-0.059234	1.122736	493.50542
Residual	-0.006675	0.025613	-0.405061	0.493989	1.311433
R Square			0.997358		



So another sensitivity analysis was done and A1 was eliminated. As you can see the model is starting to fall apart but it is still very significant. Further attempts at reducing the number of inputs to one failed.

A3

Measured	0.543000	0.498272	0.000000	1.000000	496.30200
Predicted	0.535189	0.499161	-0.072939	1.074120	498.07506
Residual	0.007811	0.034226	-0.539843	0.389317	2.341613
R Square			0.995282		



Example: NOX Dataset

NOX Detailed Description
File Name - NOX.RAW

Description: The NOX dataset was captured during the normal operation of a power boiler. The intent of the data capture was to see if any of the standard logged process variables could be used to predict the four stack gas variables. This experiment is really a fishing expedition in that no designed experiment was performed on the process variables. However, there may be enough information in the log to point to variables that have a major effect.

Column Names	Column Description
AMBAIR	Ambient air temperature.
BARKFEED	Amount of bark fed into the boiler.
BARKOFP	Air pressure over bark bed.
BARKUAIR	Air pressure under bark bed.
COALUPL	Amount of coal fed to upper level.
COALMILV	Amount of coal fed to middle level.
COALLOLV	Amount of coal fed to lower level.
1LVATEMP	Level 1 flame/gas temperature.
2LVATEMP	Level 2 flame/gas temperature.
3LVATEMP	Level 3 flame/gas temperature.
GASBURN	Amount of natural gas fed into boiler (main burners).
GASIGN	Amount of natural gas fed into boiler (ignitor).
OILUPLV	Amount of fuel oil upper level.
OILMILV	Amount of fuel oil lower level.
PAIRUPLV	Primary air feed upper level.
PAIRMILV	Primary air feed middle level.
PAIRLOLV	Primary air feed lower level.
SECUPLV	Secondary air feed upper level.
SECMILV	Secondary air feed middle level.
SECLOLV	Secondary air feed lower level.
STEAMPR	Output steam pressure.
STEAMFLO	Output steam flow.
STEAMTMP	Output steam temperature.
NOX	Nitrogen oxides exhaust from stack.
O2	Free oxygen exhaust from stack.
SO2	Sulfur dioxide exhaust from stack.
OPAC	Opacity of exhaust gases from stack.

Data Analysis A statistics report was generated using the **Basic Statistics** command in the Data menu. This gives us an overall picture of the dataset. With this much data it is highly recommended that the data be viewed using the **By Row Matrix** command in the graph menu. If you look closely at the NOX, SO2 and OPAC you will see a re-occurring blip. This was traced to a particular maintenance item done once a day.

Variable	N	Mean	Std Dev	Minimum	Maximum
AMBAIR	1340	104.44056	10.743963	85.480003	128.58999
BARKFEED	1340	84.214254	10.266984	51.849998	103.26000
BARKOFP	1340	25.602993	2.868492	-0.030000	26.530001

BARKUAIR	1340	410.85285	54.698823	27.520000	477.77999
COALUPL	1340	0.840000	2.806041	0.000000	15.620000
COALMILV	1340	0.294045	1.433840	-0.070000	10.690000
COALLOLV	1340	0.278000	1.336544	0.010000	14.230000
1LVATEMP	1340	169.54458	27.795764	100.05000	233.13000
2LVATEMP	1340	137.48353	39.120769	85.930000	227.55999
3LVATEMP	1340	113.24388	31.235535	75.570000	238.50999
GASBURN	1340	16.029052	32.871236	-0.020000	155.42999
GASIGN	1340	7.511463	2.763247	-0.040000	14.110000
OILUPLV	1340	0.602015	3.001349	-0.040000	21.070000
OILMILV	1340	0.979067	4.984428	-0.060000	32.340000
PAIRUPLV	1340	26.027418	6.590400	8.530000	35.990002
PAIRMILV	1340	13.760216	10.276768	3.860000	33.669998
PAIRLOLV	1340	9.025209	6.088600	6.140000	31.410000
SECUPLV	1340	81.394933	18.186028	54.389999	160.39999
SECMILV	1340	94.725522	20.819705	70.080002	172.58999
SECLOLV	1340	100.07086	15.736587	74.120003	137.91000
STEAMPR	1340	1638.9450	24.963765	1519.9499	1715.2800
STEAMFLO	1340	489.19210	95.430005	115.61000	694.27002
STEAMTMP	1340	1201.7549	8.199598	1130.4699	1226.1199
NOX	1340	91.407224	27.824783	32.880001	240.52000
O2	1340	9.999037	2.547359	2.930000	23.280001
SO2	1340	32.512619	39.967415	14.300000	370.76001
OPAC	1340	4.133246	1.442123	2.370000	19.900000

Model Building

Due to the large amount of data in this dataset, 80% of it was reserved for testing. The first model was constructed to predict NOX from all input variables:

```
NOX1    : IN(AMBAIR, BARKFEED, BARKOFP, BARKUAIR,
             COALUPL, COALMILV, COALLOLV, 1LVATEMP,
             2LVATEMP, 3LVATEMP, GASBURN, GASIGN,
             OILUPLV, OILMILV, PAIRUPLV, PAIRMILV,
             PAIRLOLV, SECUPLV, SECMILV, SECLOLV,
             STEAMPR, STEAMFLO, STEAMTMP)
          => OUT(NOX)
```

Model Analysis

The first model (NOX1) was created and trained using the initial factory default settings for the training parameters. After training the following model statistics were reported.

NOX1 - Training matrix statistics based on 146 observations.

Measured	93.838493	27.536246	38.169998	181.63000	109945.50
Predicted	92.675019	26.058615	46.374672	171.66412	98462.453
Residual	1.163474	9.156508	-20.98502	45.171471	12157.038
R Square			0.889427		

The model was tested using the test matrix and the following model statistics were reported.

NOX1 - Test matrix statistics based on 1194 observations.

Measured	91.109933	27.856727	32.879997	240.52002	925764.67
Predicted	90.027941	25.187365	50.670448	173.30337	756843.17
Residual	1.081992	10.132322	-54.08673	70.059113	122478.08
R Square			0.867701		

The model performance did not collapse on the test matrix indicating that the model is probably OK. The next step is to run a sensitivity analysis on the model to see if any input variables could be removed. The following table was generated using the **Sensitivity Report** command in the Model menu.

Sensitivity Analysis of NOX

Variable Name	Initial Setting	Percent Total
COALUPL	7.81	+0.15803
COALLOLV	7.12	+0.15003
BARKOFP	13.25	-0.12880
1LVATEMP	166.59	+0.08929
AMBAIR	107.04	-0.07536
3LVATEMP	157.04	+0.05892
COALMILV	5.31	+0.05778
STEAMFLO	404.94	+0.04864
OILMILV	16.14	-0.04636
OILUPLV	10.52	+0.03677
2LVATEMP	156.75	-0.03403
PAIRLOLV	18.78	+0.02443
STEAMTMP	1178.30	+0.02078
SECMILV	121.34	-0.01325
GASIGN	7.04	-0.01165
PAIRMILV	18.77	-0.01096
SECLOLV	106.02	+0.00982
SECUPLV	107.40	-0.00617
BARKUAIR	252.65	+0.00617
PAIRUPLV	22.26	-0.00480
BARKFEED	77.56	+0.00343
GASBURN	77.71	-0.00297
STEAMPR	1617.61	-0.00160

After reviewing the previous report, it was decided that variables that had less then a 2% effect on NOX should be eliminated. The following variables were eliminated: SECMILV, GASIGN ,PAIRMILV, SECLOLV, SECUPLV, BARKUAIR, PAIRUPLV, BARKFEED, GASBURN and STEAMPR. A new model (NOX2) was then created and trained using the paired down input list.

NOX2 - Training matrix statistics based on 146 observations.

Measured	93.838493	27.536246	38.169998	181.63000	109945.50
Predicted	92.261639	26.254764	41.677437	174.91064	99950.333
Residual	1.576855	9.132643	-22.12422	43.526802	12093.750
R Square			0.890002		

The model was tested using the test matrix and the following model statistics were reported.

NOX2 - Test matrix statistics based on 1194 observations.

Measured	91.109933	27.856727	32.879997	240.52002	925764.67
Predicted	89.746692	25.100364	46.305305	178.31256	751623.71
Residual	1.363241	10.558678	-51.73450	65.622284	133002.42
R Square			0.856332		

Example: CLO2 Dataset

CLO2 Detailed Description
File Name - CLO2.RAW

Description: The CLO2 dataset was the result of an chemical experiment to find the best operating points for ACID, TEMP, H2O2 and NaClO3 to product ClO2.

Column Names	**Column Description**
ACID	Amount of acid used in the reaction
TEMP	The temperature of the reaction
H2O2	Amount of hydrogen peroxide used in the reaction
NACLO3	Amount of sodium chlorate used in the reaction
CLO2	Amount of chlorine dioxide produced
PROD	Amount of chlorine converted relative to total available

Data Analysis The basic statistics report follows:

Variable	N	Mean	Std Dev	Minimum	Maximum
ACID	30	12.000000	2.729153	6.000000	18.000000
TEMP	30	60.000000	9.097177	40.000000	80.000000
H2O2	30	1.980333	0.991966	0.140000	4.050000
NACLO3	30	56.000000	20.943273	20.000000	100.00000
CLO2	30	3.206000	2.199261	0.160000	8.370000
PROD	30	71.833333	28.118080	3.000000	100.00000

Model Building Build a model of CLO2 and PROD using the other variables as inputs. Export the models to an optimizer to find the maximum production.

Example: CLOSTAT1 Dataset

CLOSTAT1 Detailed Description
File Name - CLOSTAT1.RAW

Description: The CLOSTAT1 dataset was the result of an chemical simulation to find the best operating points for DIL, CONS and RECY. WAT, D0CS, COSW, SOL D1CW and D1CS are process streams resulting from the simulation.

Column Names	**Column Description**
DIL	Dilution
CONS	Consistancy
RECY	Recycle water
WAT	
D0CS	
COSW	
SOL	
D1CW	
D1CS	

Data Analysis The basic statistics report follows:

Variable	N	Mean	Std Dev	Minimum	Maximum
DIL	15	5.000000	0.590399	4.000000	6.000000
CONS	15	8.794607	3.232301	3.291297	14.284348
RECY	15	0.941179	0.100738	0.778182	1.105573
WAT	15	7019.7693	1613.4971	5327.4257	11238.688
D0CS	15	767.34628	311.13754	270.94604	1285.0131
COSW	15	313.60096	56.263328	211.49833	405.66909
SOL	15	0.080857	0.019873	0.047394	0.118405
D1CW	15	204.52833	4.186309	196.89950	211.28370
D1CS	15	808.47947	18.348460	775.10199	838.51586

Model Building Build a models of WAT, D0CS, COSW, SOL, D1CW and D1CS using the other variables as inputs. Export the models to an optimizer to find ???

Example: PEAK4 Dataset

PEAK4 Detailed Description
File Name - PEAK4.RAW

Description:	Contains the results of stepping angles X and Y (11 steps) from 0 to and evaluating Z = sin(X) sin(Y)
Column Names	**Column Description**
X	The X variable
Y	The Y variable
Z	The result of the equation
Data Analysis	The basic statistics report follows:

Variable	N	Mean	Std Dev	Minimum	Maximum
X	121	0.500000	0.317543	0.000000	1.000000
Y	121	0.500000	0.317543	0.000000	1.000000
Z	121	0.680000	0.200671	0.200000	1.000000

Model Building	Build a model of Z using X and Y as inputs.

Example: CURL Dataset

CURL Detailed Description
File Name - CURL.RAW

Description: The was the result of a designed experiment to find which independent variables have the most effect on paper curl.

Column Names	**Column Description**
JET	Jet to wire ratio measurement
MOIST	Moisture measured on the paper machine
DD	Dryer differential measurement (between top and bottom of the sheet)
CDPOS	Position across the paper machine (physical)
FOT	Fiber orientation angle (lab)
SCURL	Simplex curl (lab)
DCURL	Duplex curl (lab)
RCURL	Reel curl (lab)
RMOIST	Reel moisture (lab)

Data Analysis The basic statistics report follows:

Variable	N	Mean	Std Dev	Minimum	Maximum
JET	70	26.001604	17.569578	6.442120	45.548595
MOIST	70	5.591592	0.587862	4.907064	6.282612
DD	70	16.901340	8.196261	7.749969	26.027664
CDPOS	70	16.428571	10.965785	1.000000	32.000000
FOT	70	4.145139	9.372166	-18.19136	20.694777
SCURL	70	-1.858674	21.378735	-55.25967	39.023815
DCURL	70	0.183420	17.469644	-39.00546	26.024122
RCURL	70	-4.502250	11.028441	-25.98776	19.509588
RMOIST	70	6.236317	0.506186	5.269414	6.997371

Model Building Build models of FOT, SCURL, DCURL and RCURL using JET, MOIST, DD and CDPOS as inputs. Try using RMOIST (lab moisture) in place of MOIST (on-line measurement).

Example: STR4 Dataset

STR4 Detailed Description
File Name - STR4.RAW

Description: The STR4 dataset was captured during the normal operation of a paper machine. The intent of the data capture was to see if any of the standard logged process variables could be used to predict paper strength properties. This experiment is really a fishing expedition in that no designed experiment was performed on the process variables. However, there may be enough information in the log to point to variables that have a major effect.

Column Names	Column Description
KSOFT	Percent softwood pulp used in furnish
KHARD	Percent hardwood pulp used in furnish
KBROKE	Percent broke pulp used in furnish
KDEINK	Percent deinked pulp used in furnish
KGRDW	Percent groundwood pulp used in furnish
STARSLD	Starch solids
SPEED	Paper machine speed
HDBXPH	Head box pH
HDBXFREE	Head box freeness
HDBXCONS	Head box consistancy
SOFTCONS	Softwood consistancy
SOFTFREE	Softwood freeness
HARDCONS	Hardwood consistancy
HARDFREE	Hardwood freeness
SBSWGT	Supered basis weight
STAF	Supered TAF (strength test)
STEARMD	Supered MD tear (strength test)
STEARCD	Supered CD tear (strength test)
RAWSTOCK	Raw stock basis weight
REELMO	Reel moisture
UBSWGT	Un-supered basis weight
COUCH	Couch vacuum
REELASH	Reel ash
LABMO	Lab moisture

Data Analysis The basic statistics report follows:

Variable	N	Mean	Std Dev	Minimum	Maximum
KSOFT	1178	35.300509	5.127401	0.000000	41.000000
KHARD	1178	11.530560	14.534154	0.000000	47.000000
KBROKE	1178	30.334465	3.890627	10.000000	40.000000
KDEINK	1178	6.057725	4.218549	0.000000	15.000000
KGRDW	1178	16.782683	14.455638	0.000000	34.000000
STARSLD	265	1.216679	0.075545	0.900000	1.600000
SPEED	1178	2254.4295	100.00711	1845.0000	2313.0000
HDBXPH	1178	7.184550	0.133124	6.900000	7.400000
HDBXFREE	1178	144.56536	73.165909	54.00000	330.00000

				0	
HDBXCONS	1178	0.584888	0.044584	0.500000	0.740000
SOFTCONS	1176	3.716556	0.206896	2.980000	4.300000
SOFTFREE	1176	501.39881	34.988070	398.00000	635.00000
HARDCONS	491	3.878411	0.268816	3.360000	4.560000
HARDFREE	491	417.72301	33.024367	351.00000	483.00000
SBSWGT	642	44.566963	7.246427	36.830002	71.330002
STAF	157	35.529618	6.042086	20.400000	54.430000
STEARMD	340	22.358529	4.784218	14.600000	45.099998
STEARCD	340	26.862941	6.009229	19.400000	56.099998
RAWSTOCK	718	30.471086	4.376988	25.950001	56.759998
REELMO	741	3.851309	0.464810	2.280000	5.420000
UBSWGT	739	45.397253	6.907419	37.099998	70.580002
COUCH	240	6.915833	1.587740	4.000000	13.900000
REELASH	197	27.450254	2.671402	22.500000	34.500000
LABMO	228	4.524561	0.654777	2.400000	6.200000

Model Building Build a model of STAF and find the variables that most effect it.