HISTORICAL DATA ANALYSIS WITH AUTOCORRELATION

Stephen Clarke
sclarke@sabic.com
crofut@mac.com

Quality and Productivity Research Conference, June 2017

CHEMISTRY THAT MATTERS™

THE PROBLEM

Manufacturing (continuous) process.

Over 10 years of data (daily averages).

After Data Preparation, over 100 potential independent variables (X's).

Over 4,000 observations, with some missing values.

All variables standardized to have a mean of zero and a standard deviation of 1.

Goal is to increase Y (output)

Software used for this case study was JMP PRO, version 12.0.1

THE CHALLENGE WITH HISTORICAL DATA ANALYSIS

In conducting a Statistical Analysis of Historical Data, two large problems revolve around correlation.

In particular,

Correlation among the potential x's (multicollinearity), and

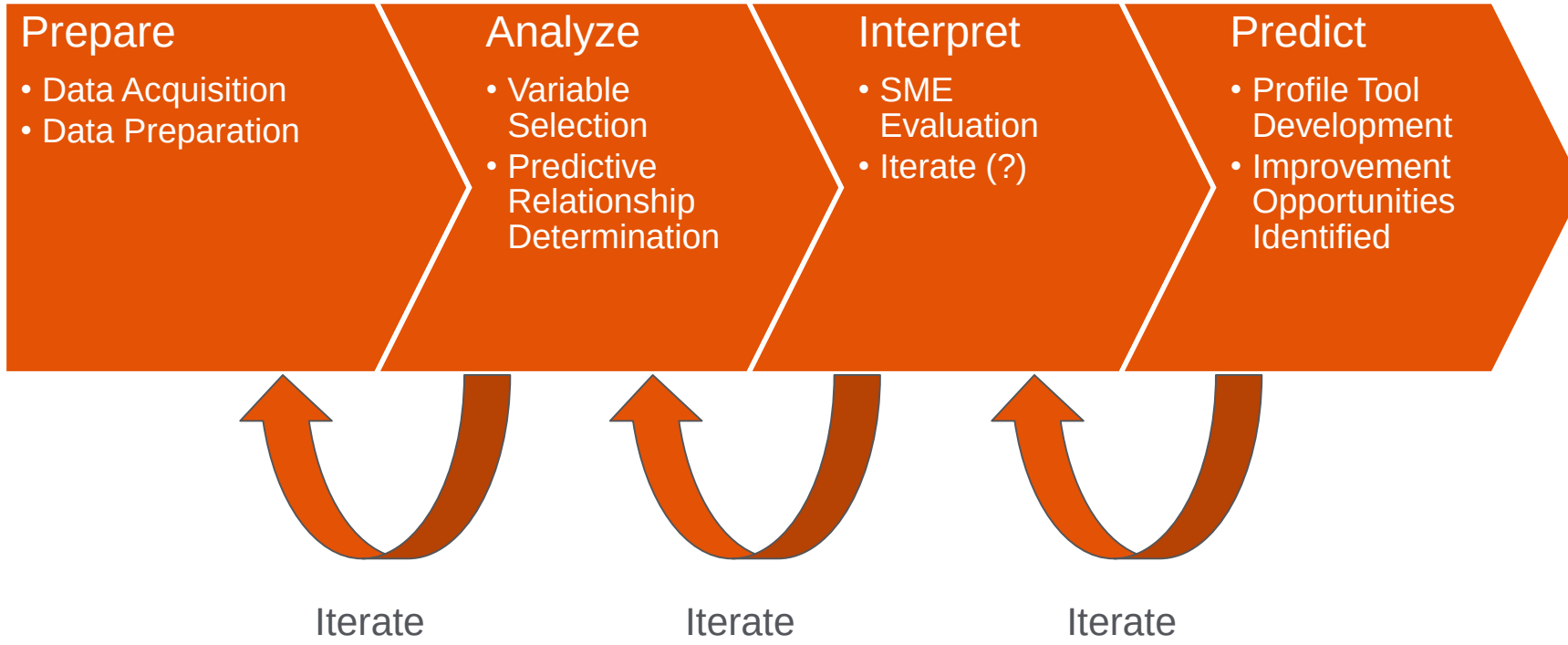
Correlation among the observations over time (autocorrelation).

The autocorrelation violates the assumption of independence in Ordinary Least Squares analysis. The multicollinearity increases the standard error of the estimates, reflecting the instability of the estimates. Both of these are the responsibility of the statistician to address.

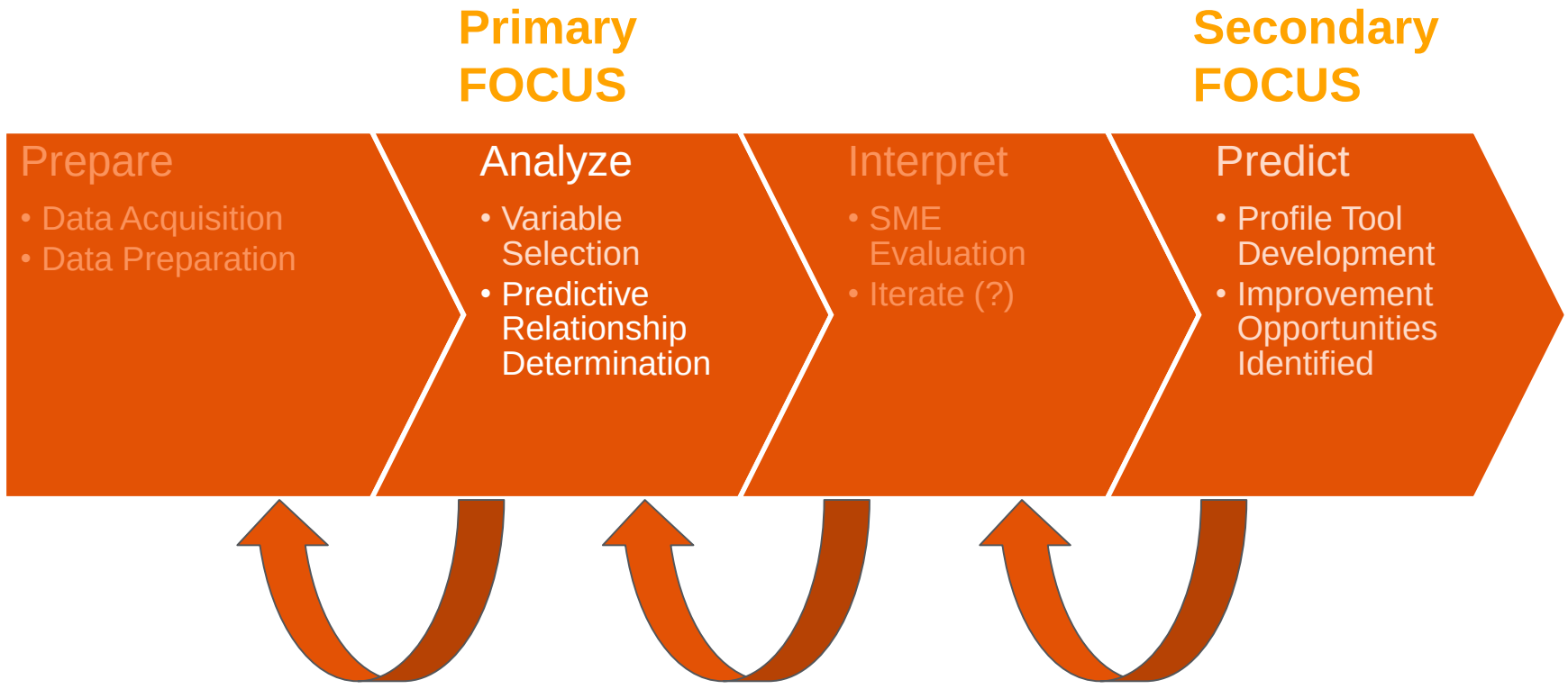
ANALYSIS WITH AUTOCORRELATION

THE ANALYSIS PROCESS

ANALYSIS PROCESS



ANALYSIS PROCESS FOCUS



How to reduce the number of potential variables from hundreds to a more manageable number?

Variable Selection

- Options include
  - Generalized Regression (Lasso)
  - Principle Components Analysis (PCA)
  - Y-Aware PCA
  - Partial Least Squares

Y-Aware PCA has been proposed by Nina Zumel in 2016. Basically, each potential predictor is rescaled to a unit change in y, based on simple linear regression. PCA is then used on these rescaled predictors to identify components. Win-Vector Blog, May 23, 2016.

VARIABLE SELECTION RESULTS

Variable Selection resulted in a subset of 15 candidate predictors

Variable Number	Variable Label
1	X120
2	X118
3	X22
4	X3
5	X71
6	X18
7	X110
8	X111

Variable Number	Variable Label
9	X78
10	X46
11	X20
12	X41
13	X39
14	X101
15	X2

Note: Principle Components Analysis on the 115 Predictors resulted in 18 eigenvalues >1.0
Partial Least Squares suggested 13 latent variables.

THE ANALYSIS PROCESS

The analysis process proceeds in three main steps:

1. Evaluate Main Effects
2. Evaluate Quadratic Effects (and Main Effects)
3. Evaluate Two-Factor Interactions (as well was Quadratic and Main Effects)

Strong model heredity is maintained (in other words, a Main Effect stays in the model as long as the corresponding quadratic term or an interaction containing the Main Effect remains in the model.

Xiang, et al. (2006) reviewed a large number of Full Factorial Designed Experiments and showed that given a Two-Factor Interaction existed, the probability that both of the Main Effects were also significant was over 85%.

CHALLENGE #1: DECISION CRITERION

Prior to the analysis, good statistical practice includes establishing decision criteria.

With very large datasets, what is the appropriate P-value to use to determine a factor should remain in a model (assuming backward elimination)?

As the number of observations (n) increases, the standard error of the estimates decreases in proportion to the square root of n. In the design of an experiment, the sample size is determined so as to calibrate the statistics with the real world. In other words, the conclusion of a statistical difference is calibrated to correspond to a difference that is meaningful to the experimenters. Historically, the P-Value criterion used was P=0.05. Sample sizes were typically less than 100. In analyzing data sets with hundreds or thousands of observations, it is reasonable (if we are to maintain the calibration with real world meaningfulness), to reduce the P-Value criterion.

With hundreds of observations, a recommendation is to use P-value < 0.01. For thousands of observations, use P-value < 0.0001.

Alternatively, one could use BIC, AICc, or max R^2 with a validation subset as the decision criterion.

INITIAL MODEL

Using least squares with the decision criterion P-Value<.0001

Final Model includes:

7 Main Effects:	Quadratic Terms:	6 2FIs:
X120		
X118	**None**	X110*X46
X22		X120*X22
X3		X118*X22
X110		X120*X39
X46		X118*X110
X39		X46*X39

R^2 = .79
RMSE = 0.3273

CHALLENGE #2: MULTICOLLINEARITY

The lack of independence among the predictors can be handled by controlling the Variance Inflation Factors (VIFs) using Ordinary Least Squares (OLS) Analysis. Other options for dealing with multicollinearity include Principle Component Analysis and Partial Least Squares. However, these techniques make interpretation and translation into recommended actions more challenging. Generalized Regression assumes independence, like OLS.

The solution for dealing with multicollinearity is to simplify the model by removing terms. This is usually, but not always, accomplished by removing the term with the highest VIF. In the case of a Main Effect with a high VIF, one must often look to remove a higher order term, such as a Quadratic Effect or a Two-Factor Interaction.

Scaling predictors (mean = 0, standard deviation = 1) reduces the high VIFs attributable to inclusion of 2FIs and Quadratic Terms.

VIF ESTIMATES IN INITIAL MODEL

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.0670182	0.012899	5.20	<.0001*	.
Std X120	0.2260815	0.020274	11.15	<.0001*	10.50105
Std X118	0.3710431	0.020459	18.14	<.0001*	7.2830242
Std X22	0.1372287	0.013561	10.12	<.0001*	3.1199986
Std X3	0.2896224	0.03423	8.46	<.0001*	4.6392884
Std X110	-0.15298	0.056176	-2.72	0.0065*	5.2531762
Std X46	0.0667935	0.011184	5.97	<.0001*	2.6532438
Std X39	0.1416136	0.012738	11.12	<.0001*	1.1490659
(Std X120+0.01481)*(Std X22-0.03614)	-0.140511	0.015133	-9.29	<.0001*	12.556162
(Std X120+0.01481)*(Std X39+0.03068)	0.1156527	0.018826	6.14	<.0001*	3.6367177
(Std X118-0.03273)*(Std X22-0.03614)	0.1477468	0.016658	8.87	<.0001*	14.159911
(Std X118-0.03273)*(Std X110-0.14219)	-0.23953	0.027462	-8.72	<.0001*	13.823503
(Std X110-0.14219)*(Std X46-0.07263)	0.3263422	0.039851	8.19	<.0001*	7.3567628
(Std X46-0.07263)*(Std X39+0.03068)	0.0846604	0.016624	5.09	<.0001*	1.7928013

VARIABLE IDENTIFICATION – VIF CRITERION

Variance Inflation Factors (VIFs) are useful to quantify multicollinearity among predictors.

Recall: $VIF = 1/(1-R^2_k)$

Where R^2_k is from the regression of predictor k on all other terms currently in the model.

Various suggestions as to a maximum acceptable VIF generally vary from 5 to 10. Similar to the more stringent P-Value criterion, a tightening of the VIF criterion is also appropriate. It has been suggested (Klein, 1962) that an acceptable $VIF < 1/(1-R^2)$ of the model under development. This approach, with a model $R^2 = .75$, would result in a maximum allowed VIF = 4.

The proposed decision criterion is to remove any predictor with a VIF > 5.0, when dealing with large historical data sets.

Klein, L. 1962. An Introduction to Econometrics. New York. Prentice Hall

MODEL 2: IMPOSE THE VIF CONSTRAINT

Using least squares with the same decision criteria (P-Value<.0001). Impose the VIF constraint (VIF<5). Ignore the autocorrelation of the data for now

Final Model includes:

5 Main Effects:	2 Quadratic Terms:	1 2FI:
X118	$X22^2$	X110*X46
X22	$X46^2$	
X110		
X46		
X39		

R^2 = .76
RMSE = 0.3492

This model is similar and simpler, but not identical. SME's could explain/understand relationships.

CHALLENGE #3: AUTOCORRELATION

A Durbin-Watson statistic can check for the existence of the lag 1 autocorrelation. Alternatively, more complex autocorrelation structures can be evaluated using time series.

To deal with the autocorrelated structure typical of historical manufacturing data, the Mixed Model Methodology (MMM), utilizing an autoregressive error structure, is a way to model this situation. The downside is that MMM can be quite resource consuming. A single analysis (of many analyses in a backward elimination process) can easily take over an hour on a modern laptop, when dealing with thousands of observations. See the appendix for further details on analysis time.

MIXED MODEL ERROR STRUCTURES

Mixed Model Methodology (MMM), by modeling both Fixed and Random effects, allows the statistician to model (and evaluate) various error structures (Littell, et.al., 2000). Some common error structures that MMM can handle include:

Independent/Simple: This assumes a correlation of zero, the assumption for OLS.

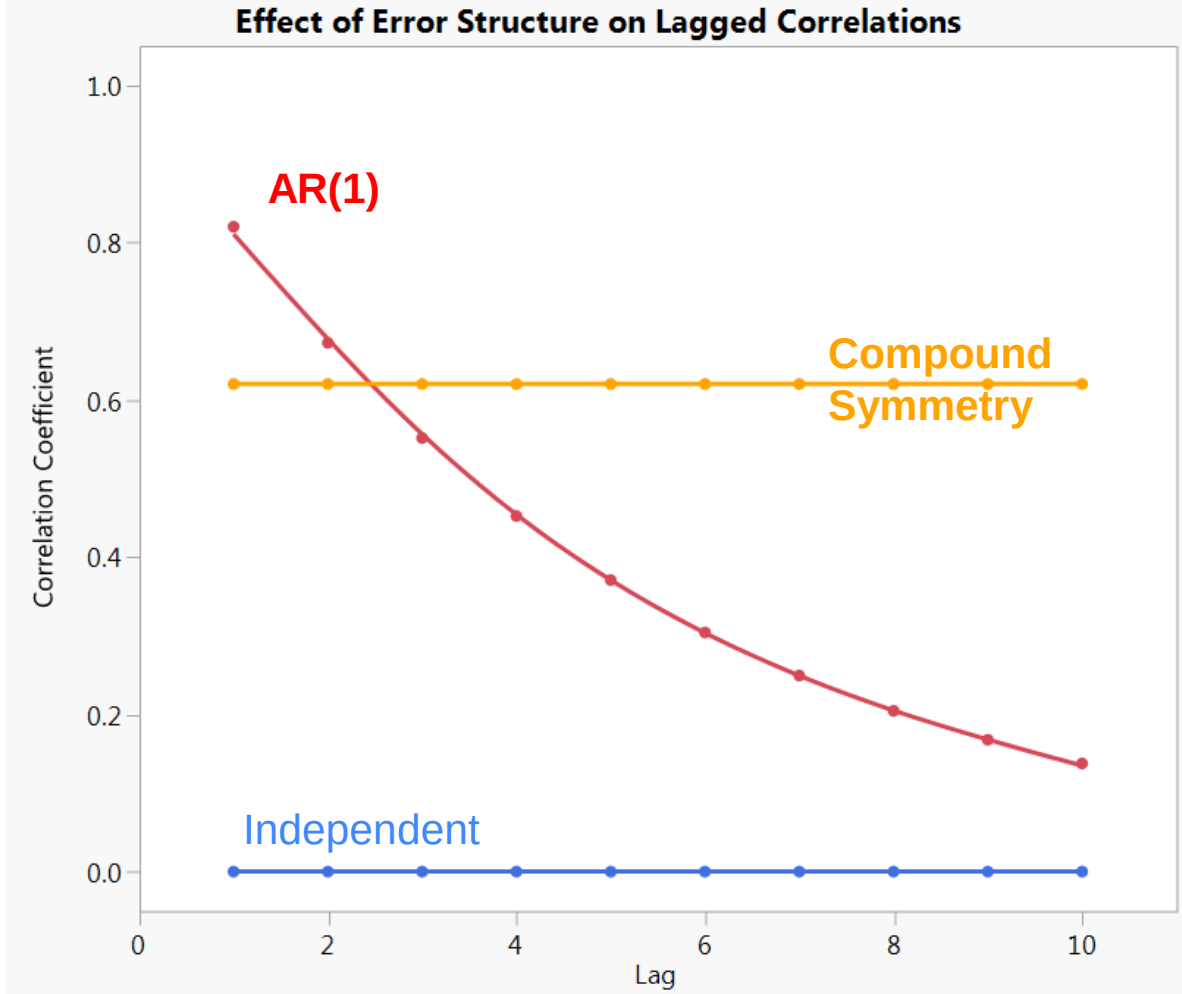
Compound Symmetry: this error structure essentially estimates a single overall correlation among all observations of a group or experimental unit. This is the approach used in split-plots and repeated measures situations.

Unstructured: this error structure estimates the correlation coefficient between each pair of related observations. This uses a lot of degrees of freedom.

AR(1): An autoregressive error structure estimates the relationship between nearest neighbors, and them propagates it to more distant relationships, with a exponential decline. This uses far fewer degrees of freedom than Unstructured. By not indicating a grouping or experimental unit, this approach builds the nearest neighbor relationship across the entire data set.

Littell, R.C., J. Pendergast and R. Natarajan. 2000. Modelling covariance structure in the analysis of repeated measures data. Statist. Med. 19: 1793-1819.

EFFECT OF ERROR STRUCTURE ON TIME-LAGGED CORRELATION COEFFICIENTS



ERROR STRUCTURE OF CASE STUDY

Time Series analysis of the case study data reveal a decreasing lagged autocorrelation. As the time span increases, the correlation decreases. Partial autocorrelations decrease rapidly after lag 1 confirming the autoregressive (lag 1) structure.

Time Series Basic Diagnostics

Lag	AutoCorr	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	0.8187		2632.63	<.0001*	1	0.8187	
2	0.7099		4612.67	<.0001*	2	0.1203	
3	0.6581		6314.81	<.0001*	3	0.1488	
4	0.6310		7879.93	<.0001*	4	0.1185	
5	0.6035		9311.95	<.0001*	5	0.0648	
6	0.5769		10621.0	<.0001*	6	0.0494	
7	0.5562		11838.2	<.0001*	7	0.0487	
8	0.5561		13054.9	<.0001*	8	0.0952	
9	0.5442		14220.7	<.0001*	9	0.0281	
10	0.5349		15347.0	<.0001*	10	0.0488	

MIXED MODEL PREDICTIVE RELATIONSHIP– FINAL MODEL?

After Mixed Model Analysis/Backward Elimination, all terms exhibit a P-Value < .0001. All VIFs < 5.0.

Model includes:

8 Main Effects:	3 Quadratic Terms:	4 2-Factor Interactions:
X118	$X22^2$	X118*X71
X22	$X46^2$	X118*X39
X110	$X39^2$	X71*X20
X46		X71*X39
X39		
X71		
X78		
X20		

Std. Dev. of Residuals = .4738
Approx. R^2 = 0.62 (estimated as the square of the correlation between actual and predicted).

Note: Using AICc or BIC as the model selection criterion resulted in additional 2FIs remaining in the model.

ANALYSIS WITH AUTOCORRELATION

EXAMINATION OF TWO-FACTOR INTERACTIONS

INTERPRETATION OF 2FI

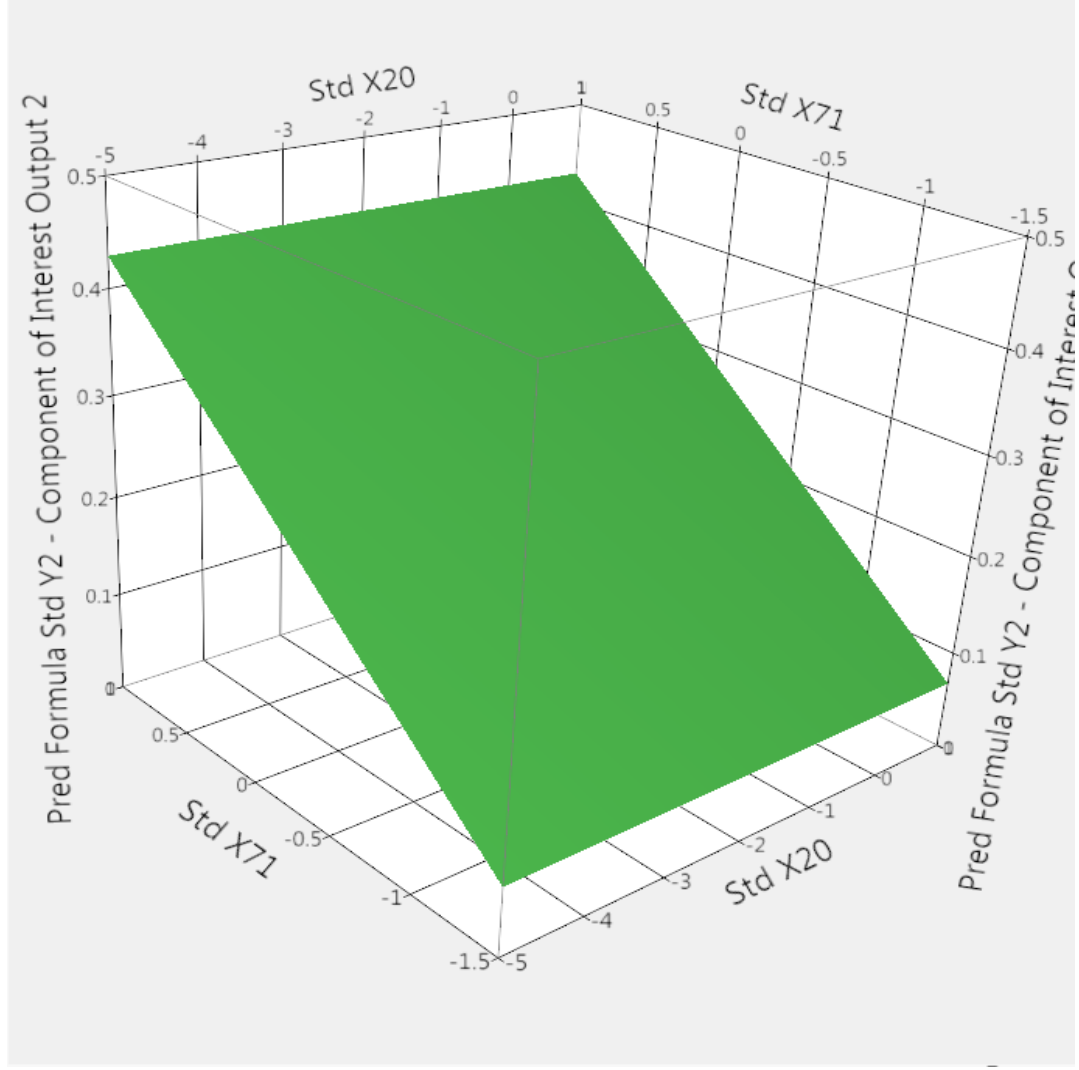
A critical aspect of the process is understanding by the SMEs. Does the model make sense based on an understanding of the chemistry (or whichever relevant subject matter is under study)?

Several possible issues warrant elimination of a statistically significant Two-Factor Interaction (2FI):

- The 2FI may be statistically significant (at whatever level is being used for the study) but a graphical display of the relationship fails to suggest any deviation from the Main Effects in the model. In other words, the effect/shape of this 2FI is not apparent in a three-dimensional plot of the predicted values (over the range of typical predictor values). This is a 2FI that should be dropped from the model. This may also suggest a further decrease in the acceptable P-Value.
- The effect of the 2FI on the predicted output suggests a relationship, but the optimum solution is located in an area devoid of historical observations. This may be of interest to SMEs, as a possible area for future investigation. But should not be relied on simply because the statistics say so.

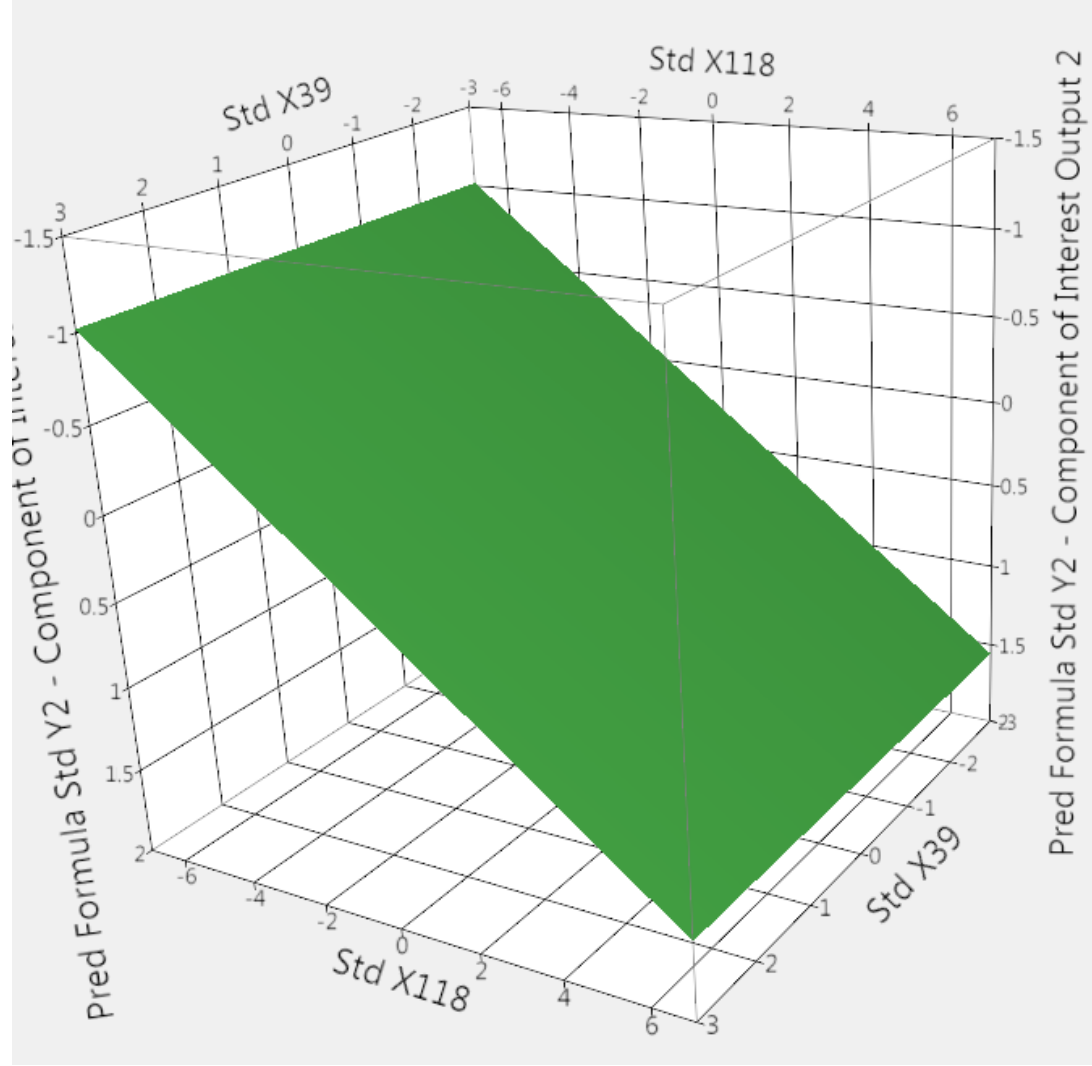
2FI – X71 BY X 20

This interaction, while statistically significant (P<.0001), exhibits no visible effect on predicted values. There is no predictive value to including this interaction. This 2FI should be removed.



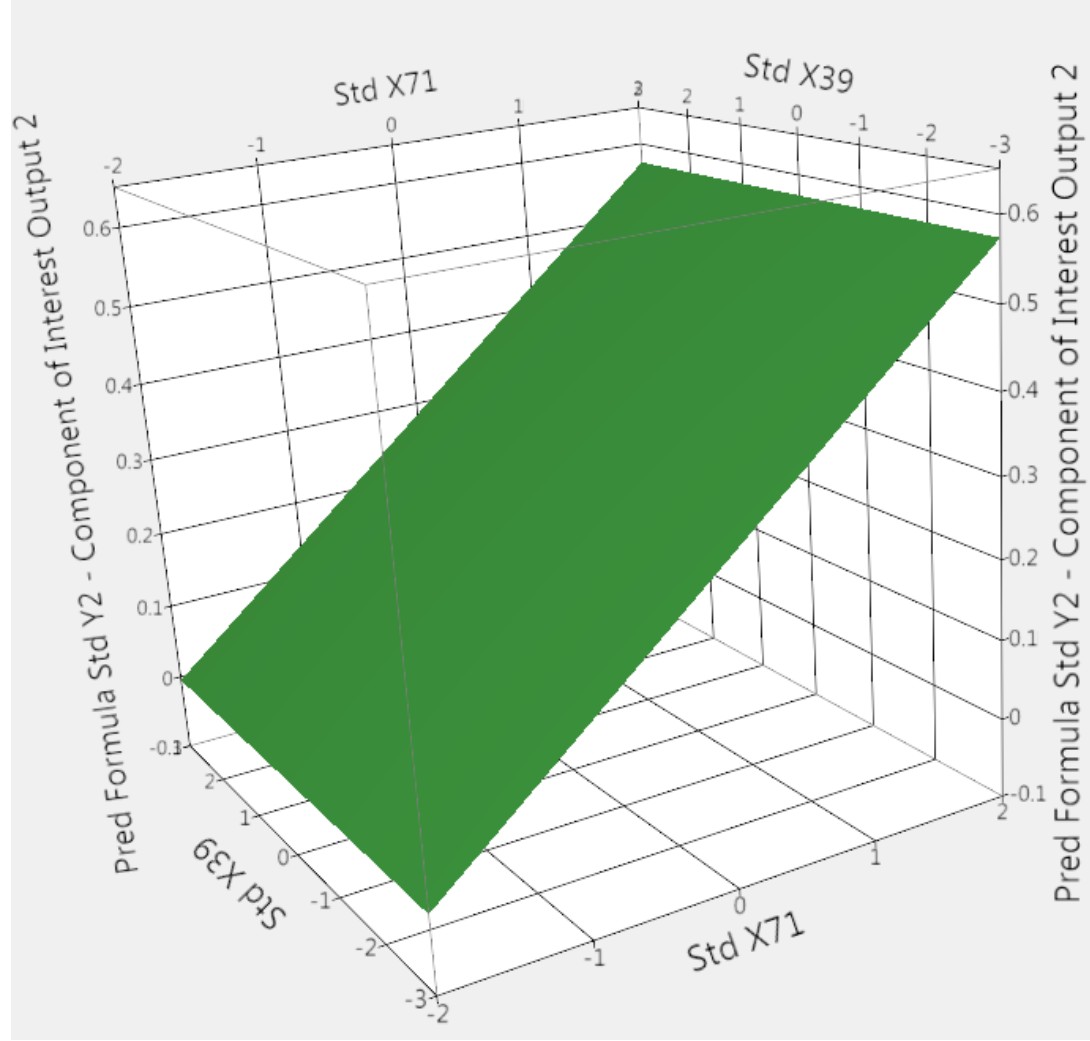
2FI – X39 BY X118

This interaction, while statistically significant (P<.0001), exhibits no visible effect on predicted values. There is no predictive value to including this interaction. This 2FI should be removed.



2FI – X71 BY X39

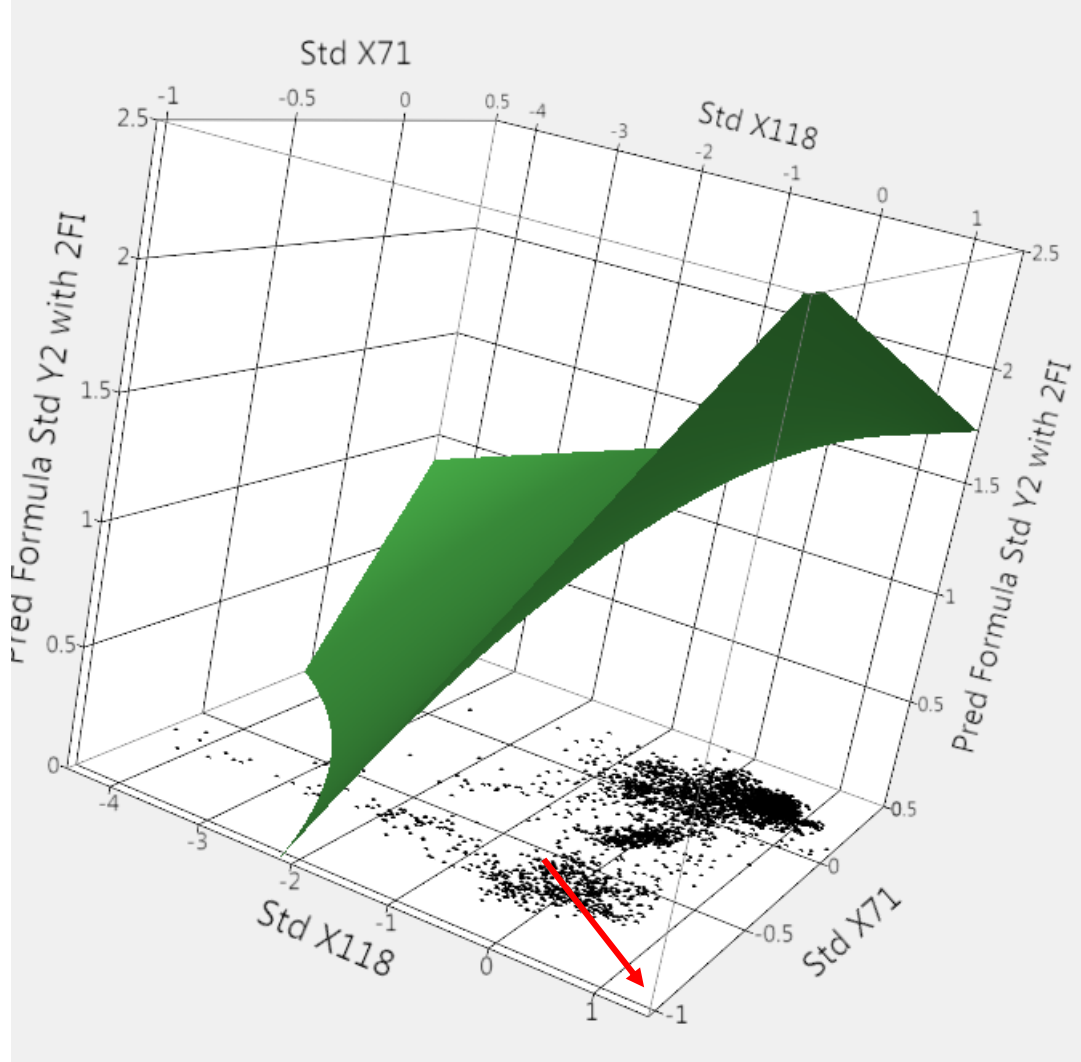
This interaction, while statistically significant (P<.0001), exhibits no visible effect on predicted values. There is no predictive value to including this interaction. This 2FI should be removed.



2FI – X71 BY X118

This interaction suggests moving into areas where there is no historical data. Maximization of y (the goal) is obtained by increasing X118, and decreasing X71 (Red Arrow),. There is no data to support this combination of optimal factor settings. This 2FI should be removed. However, it does warrant a discussion with SMEs.

Note: The surface is the predicted output based on the model. The scatter points on the floor of the diagram represent the observations of X118 by X71 in the data.



REANALYZE WITHOUT 2FI

Model reanalyzed without 2FIs:

At this point, all terms exhibit a P-Value < .0001. All VIFs < 5.0.

Model includes:

8 Main Effects:	3 Quadratic Terms:
X118	$X22^2$
X22	$X46^2$
X110	$X39^2$
X46	
X39	
X71	
X78	
X20	

Std. Dev. of Residuals = 0.4738
Approx. R^2 = 0.62 (estimated as the square of the correlation between actual and predicted.)

No change in either metric from model that included 2FIs!

FINAL MIXED MODEL ANALYSIS

Repeated Effects Covariance Parameter Estimates

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
AR(1)	0.8090003	0.0126178	0.7842699	0.8337307
Residual	0.1924048	0.0122992	0.1704035	0.2189838

Fixed Effects Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	0.1361741	0.0267782	181.2	5.09	<.0001*	0.0833368	0.1890113
Std X118	0.2740321	0.0153922	3395.4	17.80	<.0001*	0.2438532	0.304211
Std X22	0.3876431	0.0253814	614.5	15.27	<.0001*	0.3377982	0.4374879
Std X71	0.2442803	0.0576537	460.7	4.24	<.0001*	0.1309834	0.3575772
Std X110	0.3688901	0.0711366	1327.7	5.19	<.0001*	0.2293378	0.5084424
Std X78	0.0696923	0.0161082	2768.1	4.33	<.0001*	0.038107	0.1012776
Std X46	0.2144208	0.0228244	1064.4	9.39	<.0001*	0.1696348	0.2592068
Std X20	-0.392947	0.0744707	1741.5	-5.28	<.0001*	-0.539008	-0.246886
Std X39	0.5682265	0.0246555	2510.4	23.05	<.0001*	0.5198793	0.6165737
(Std X22-0.03503)*(Std X22-0.03503)	0.0574716	0.0107868	2476.9	5.33	<.0001*	0.0363195	0.0786237
(Std X46-0.07176)*(Std X46-0.07176)	-0.066747	0.0079792	3374.4	-8.37	<.0001*	-0.082391	-0.051102
(Std X39+0.03068)*(Std X39+0.03068)	-0.280449	0.0219998	3030.2	-12.75	<.0001*	-0.323585	-0.237313

ANALYSIS WITH AUTOCORRELATION

PROFILER DEVELOPMENT

PROFILER DEVELOPMENT

After discussions with SMEs to confirm that the model makes sense, a profiler can be developed to visualize relationships and identify opportunities for improvement.

With large data sets, even after data preparation, the large number of observations can result in predictor values many standard deviations away from the mean. Using the prediction equation across the full range of the predictor variables will likely result in unreasonable combinations of predictors.

One solution is to bound the allowable ranges of the predictors. A reasonable approach, when there are thousands of observations, is to bound at the

0.5 % quantile
99.5% quantile

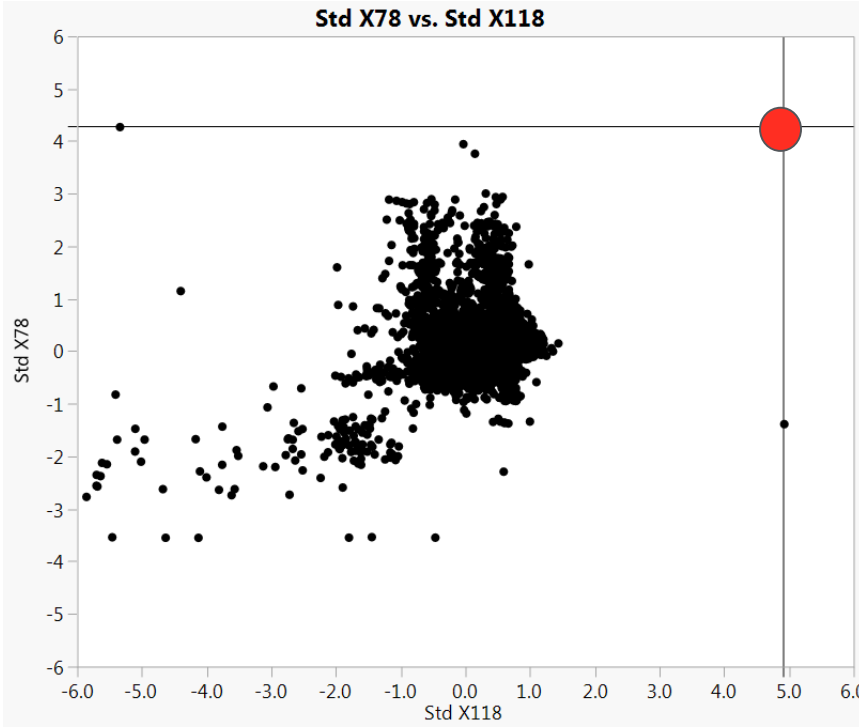
This eliminates the extreme values that may exist in the data.

RESULTING PREDICTOR BOUNDS

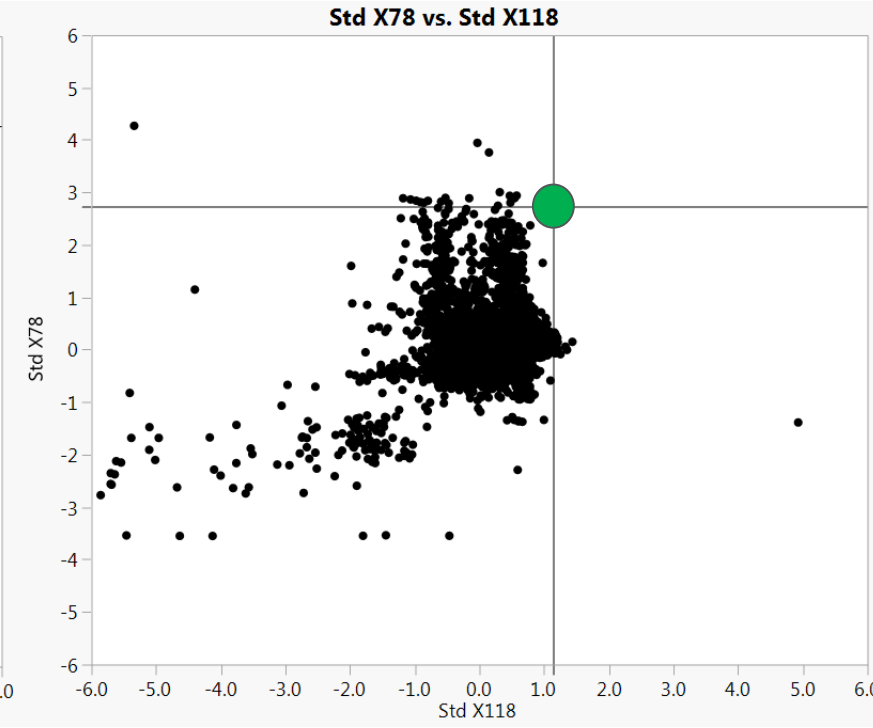
Predictor	Lower Bound 0.05 %	Upper Bound 99.5 %
X118	-4.14	1.14
X22	-2.50	1.47
X110	-0.97	0.34
X46	-2.28	1.78
X39	-1.53	1.18
X71	-0.79	0.35
X78	-2.38	2.72
X20	-0.58	0.41

SUGGESTED PROFILE SOLUTION WITHOUT AND WITH BOUNDS PLACED ON PREDICTORS

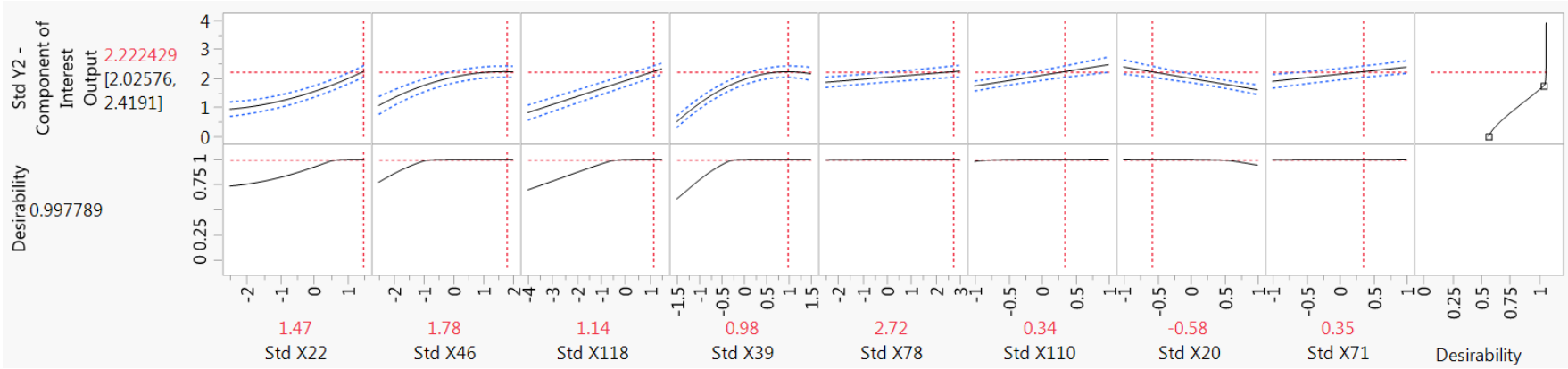
Optimum Solution without Bounds



Optimum Solution with Bounds



PROFILE WITH RECOMMENDED SOLUTION (BOUNDED)



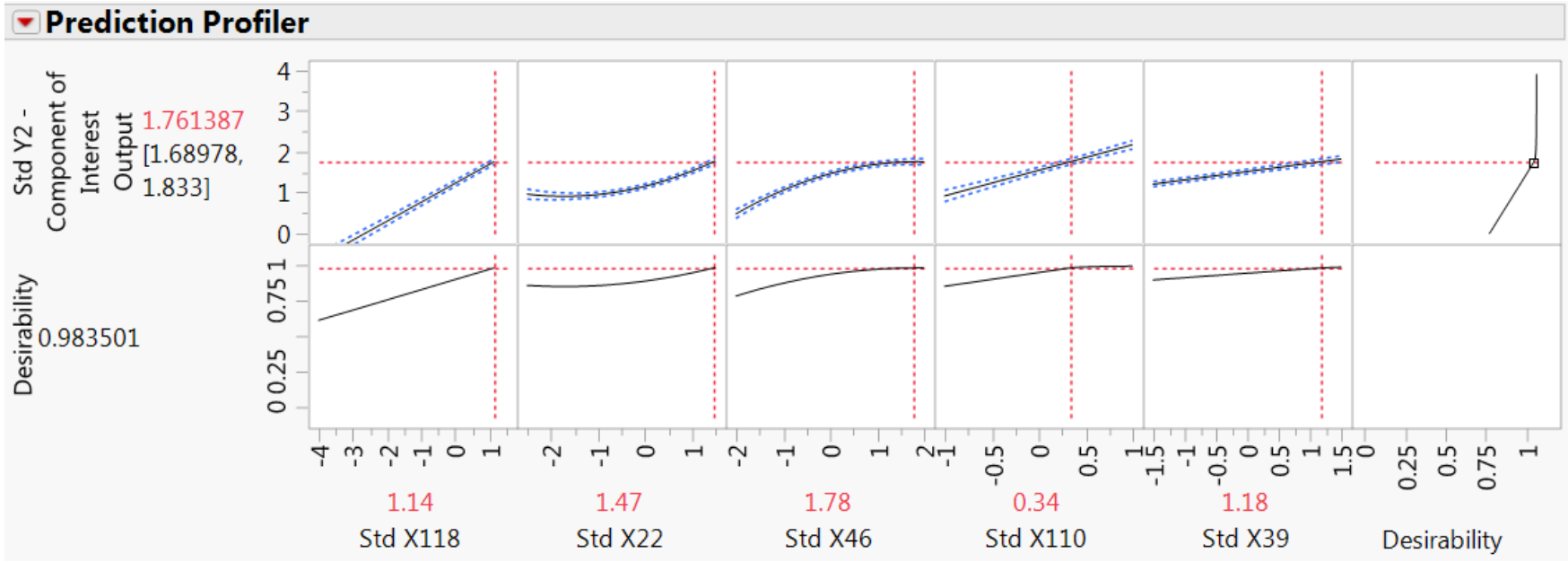
Recommended Solution (2.22) is an increase from the previous observed maximum of 1.63. In other words, a predicted increase in output of 36 %. All predictors except X39 are at a boundary.

Note: If the X118*X71 interaction remained in the model, the final solution is the same, however, the predicted output is slightly less, 2.19.

ANALYSIS WITH AUTOCORRELATION

COMPARISON WITH LEAST SQUARES ANALYSIS

LEAST SQUARES PROFILER AND OPTIMAL SOLUTION



Expected maximum output is 1.76, for an 8 % increase.

OPTIMAL SOLUTION COMPARISON

Variable	Least Square Solution	Mixed Model Methodology
Predicted Y - Output	1.76	2.22
X118	1.14	1.14
X22	1.47	1.47
X110	0.34	0.34
X46	1.78	1.78
X39	1.18	0.98
X71	-	0.35
X78	-	2.72
X20	-	-0.58

Rectangle highlights differences in solution. X39 is linear in the LS model, and quadratic in the MM.

CONCLUSIONS

The Mixed Model Methodology offers a solution to the autocorrelation that exists in large historical data sets.

When dealing with large historical data sets, more stringent decision criteria are recommended.

Correcting for the autocorrelation results in a different model and optimal solution to the problem under study.

Careful examination of Two-Factor Interactions for meaning and reasonableness is critical to determination of a useful model.

Thank You!

ANALYSIS WITH AUTOCORRELATION

APPENDIX
ANALYSIS TIME FOR MIXED MODELS

MIXED MODEL ANALYSIS TIMES

A study was conducted to determine the impact of Model Complexity and Number of Observations on the time to perform a Mixed Model Methodology Analysis.

An artificial dataset was created exhibiting an autoregressive error structure. The dependent variable was a function of five predictors (linear), two quadratic terms, and three two-factor interactions.

Elapsed analysis (Mixed Model, AR(1)) time was measured to the nearest second with a stopwatch.

Model Complexity included 3 levels: Linear Terms Only (5 modeling terms plus an intercept, 6), Linear and Quadratic (10 modeling terms plus an intercept, 11), and Full Response Surface (20 terms plus intercept, 21).

Number of observations: 100, 500, 1000, 1500, 2000, 2500, 3000.

As can be seen on the graph on the next page, both factors impact analysis time in a non-linear manner. While the specific times are a function of the particular computer, the general shape of the relationship should hold for different configurations (processor speed, RAM, etc.).

EFFECT OF NUMBER OF OBSERVATIONS AND MODEL COMPLEXITY ON MIXED MODEL ANALYSIS TIME