Safe Machine Learning Prediction and Optimization via Extrapolation Control

Authors: *Chris Gotwalt* – Chief Data Scientist

Laura Lancaster – Principal Research Statistician Developer

Jeremy Ash – Research Statistician Developer

Presenters: *Tom Donnelly* – Principal Systems Engineer

Elizabeth Claassen – Research Statistician Developer

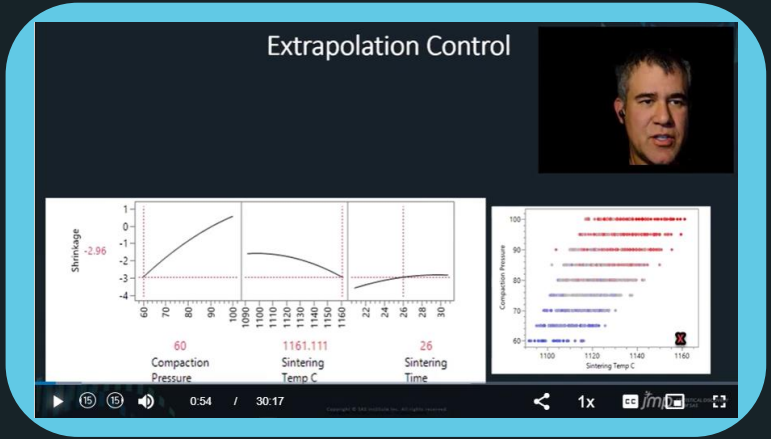
JMP Statistical Discovery LLC (formerly JMP Division of the SAS Institute)

DATAWorks

April 27, 2022

jmp STATISTICAL DISCOVERY FROM SAS

Safe Machine Learning Prediction and Optimization via Extrapolation Control



Link to recording of presentation by the authors:

[Controlling Extrapolation in the Prediction Profiler in JMP Pro 16 (2021-US-45MP... - JMP User Community]

Link to article by authors:

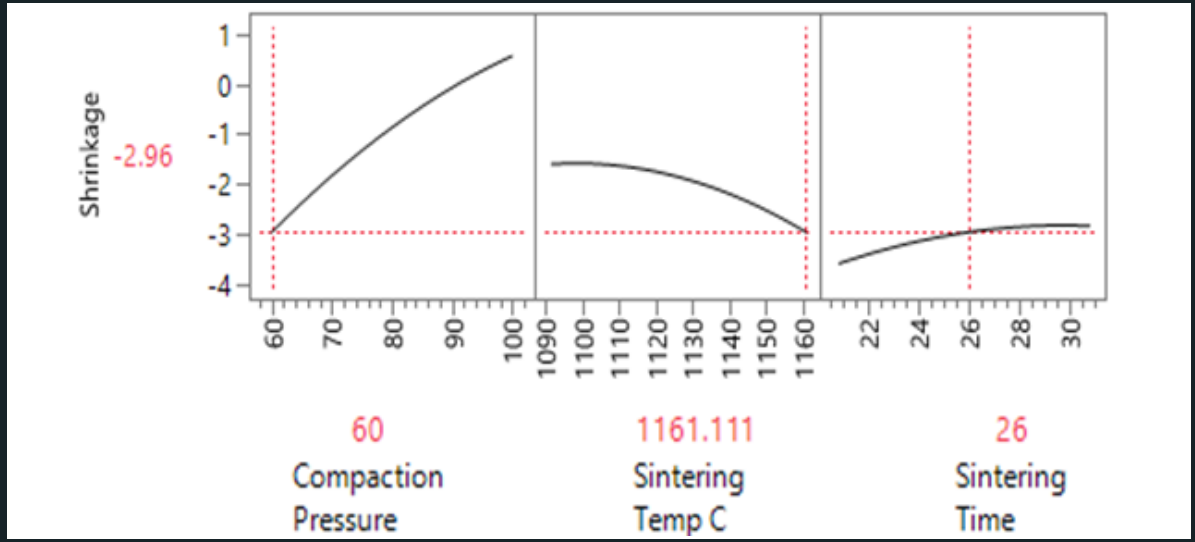
https://arxiv.org/pdf/2201.05236.pdf

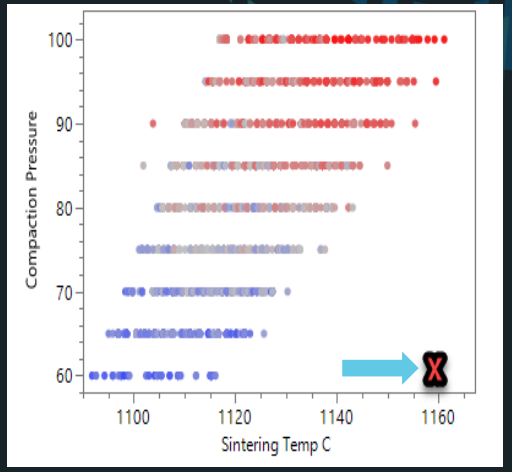
A Method for Controlling Extrapolation when Visualizing and Optimizing the Prediction Profiles of Statistical and Machine Learning Models

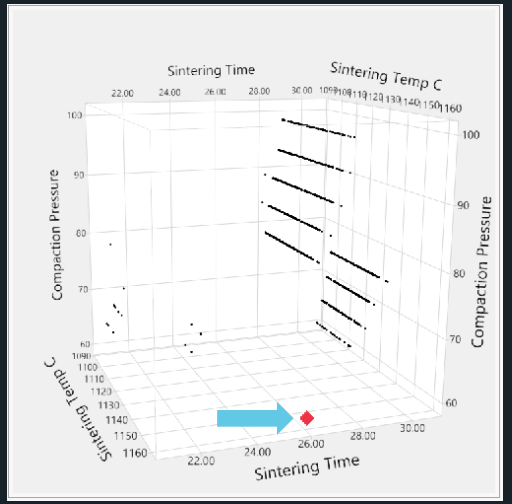
Jeremy Ash,
Laura Lancaster,
and
Chris Gotwalt
JMP Division, SAS Institute

January 17, 2022

Safe Machine Learning Prediction and Optimization via Extrapolation Control

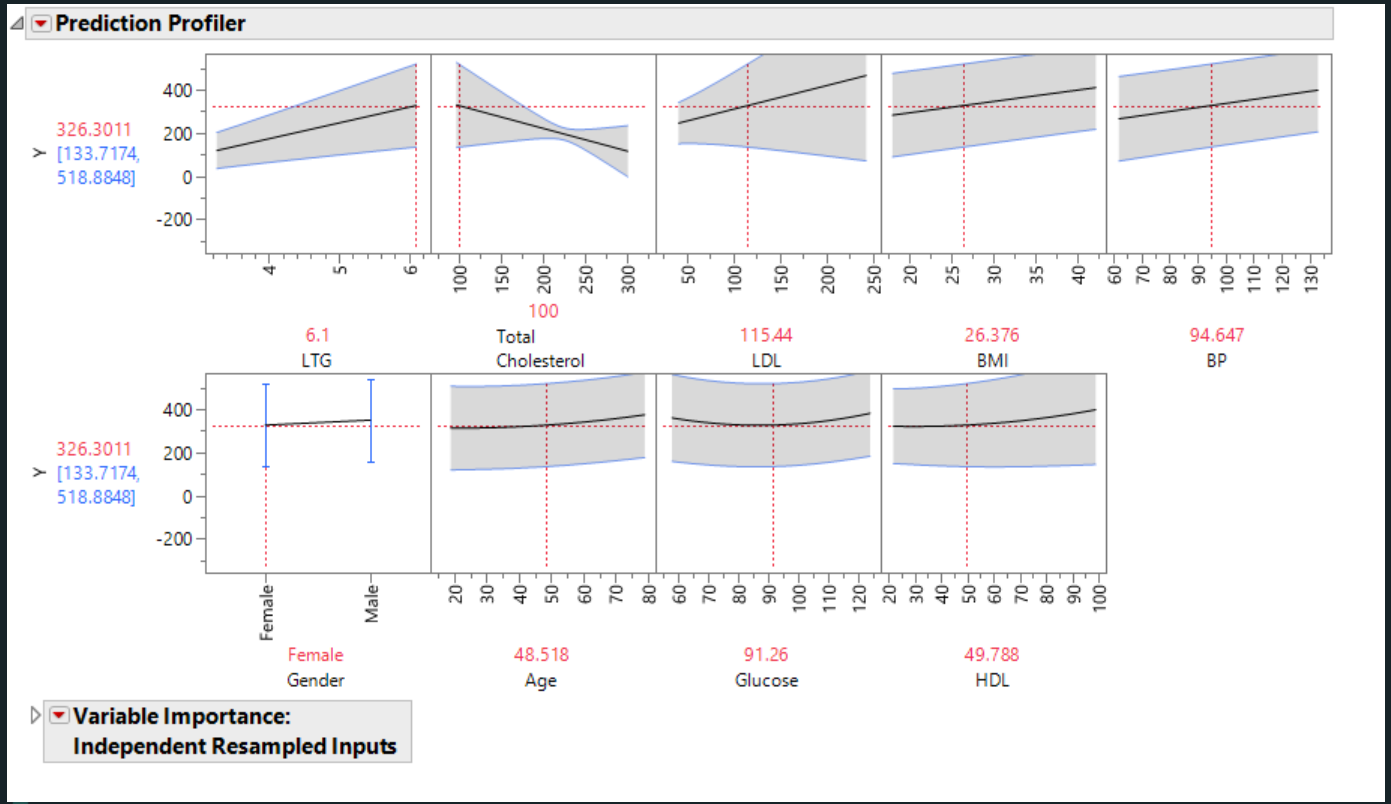




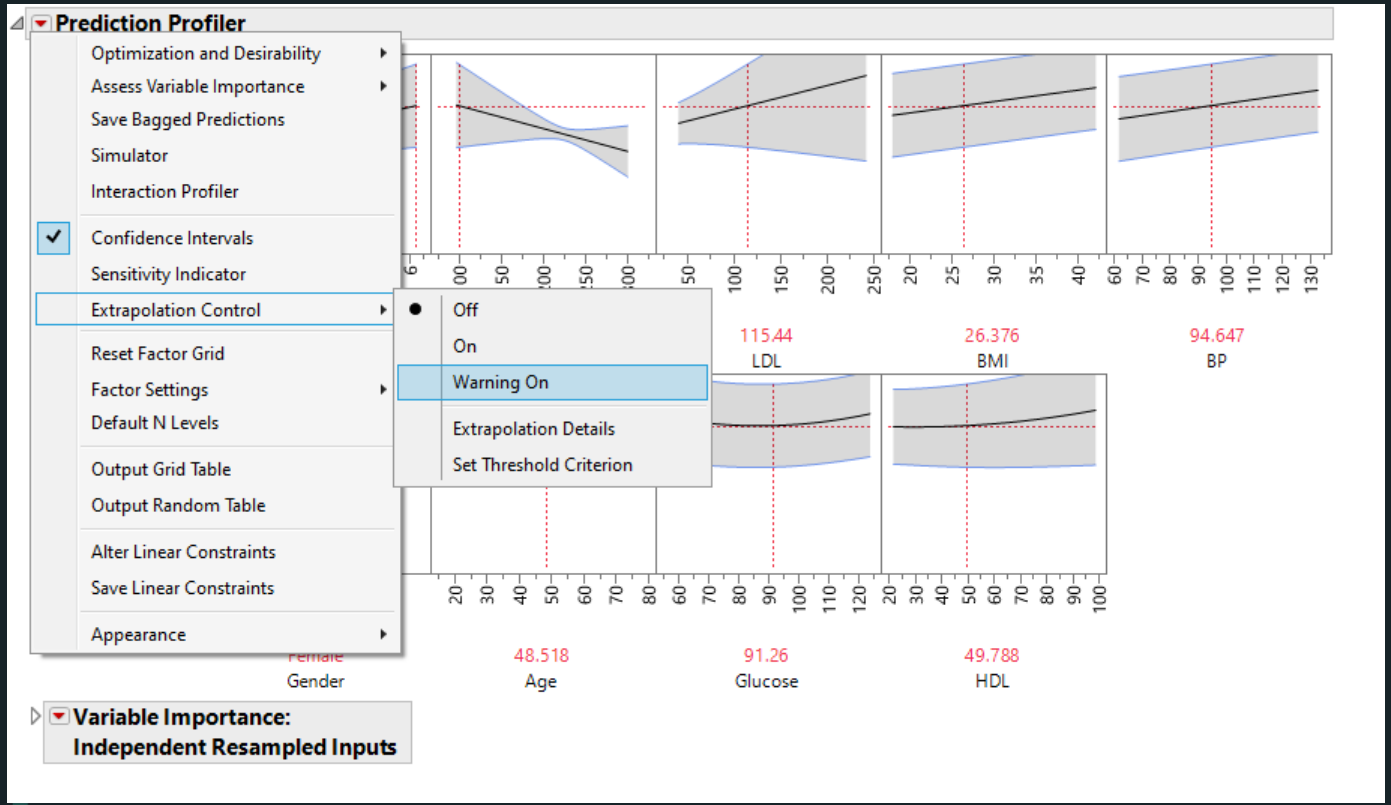


jmp STATISTICAL DISCOVERY FROM SAS

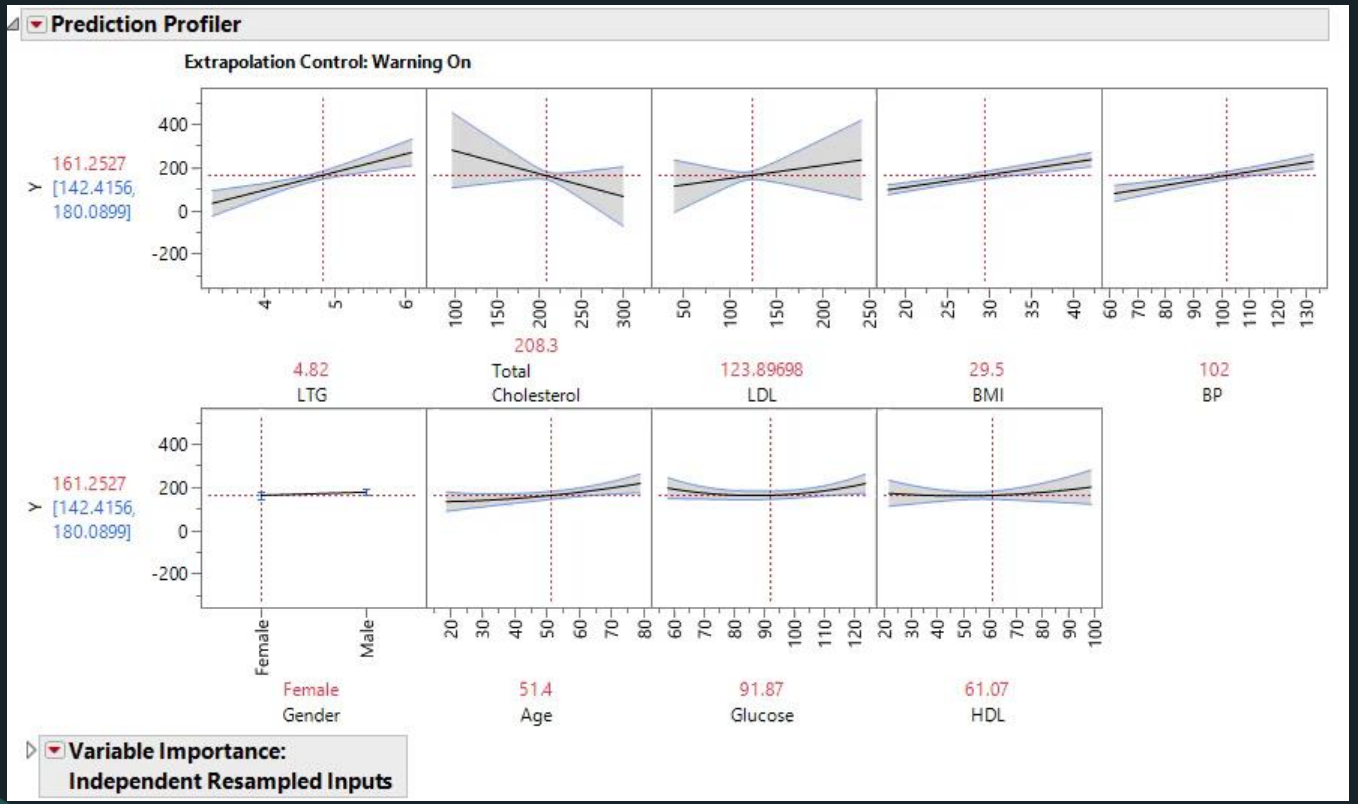
Without Extrapolation Control



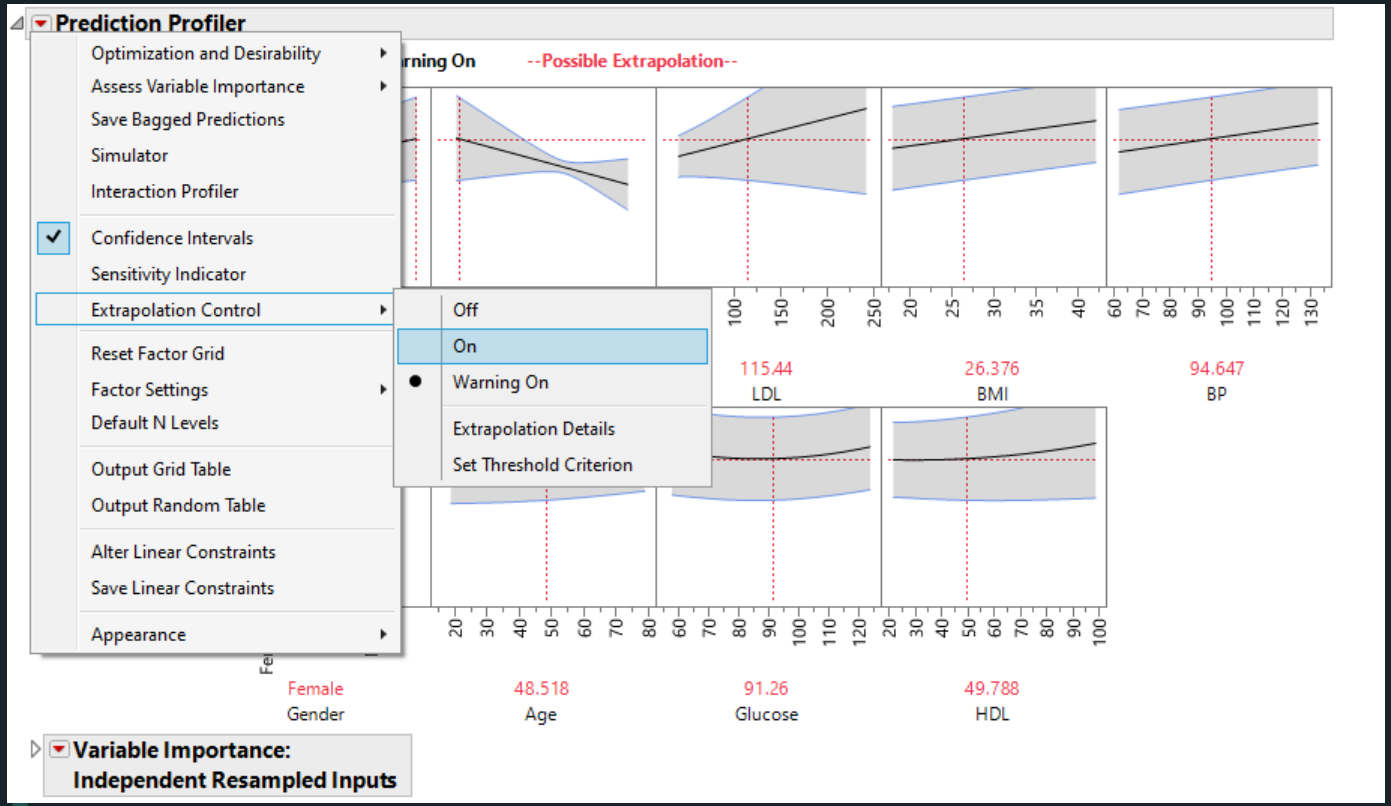
Turn Warning On



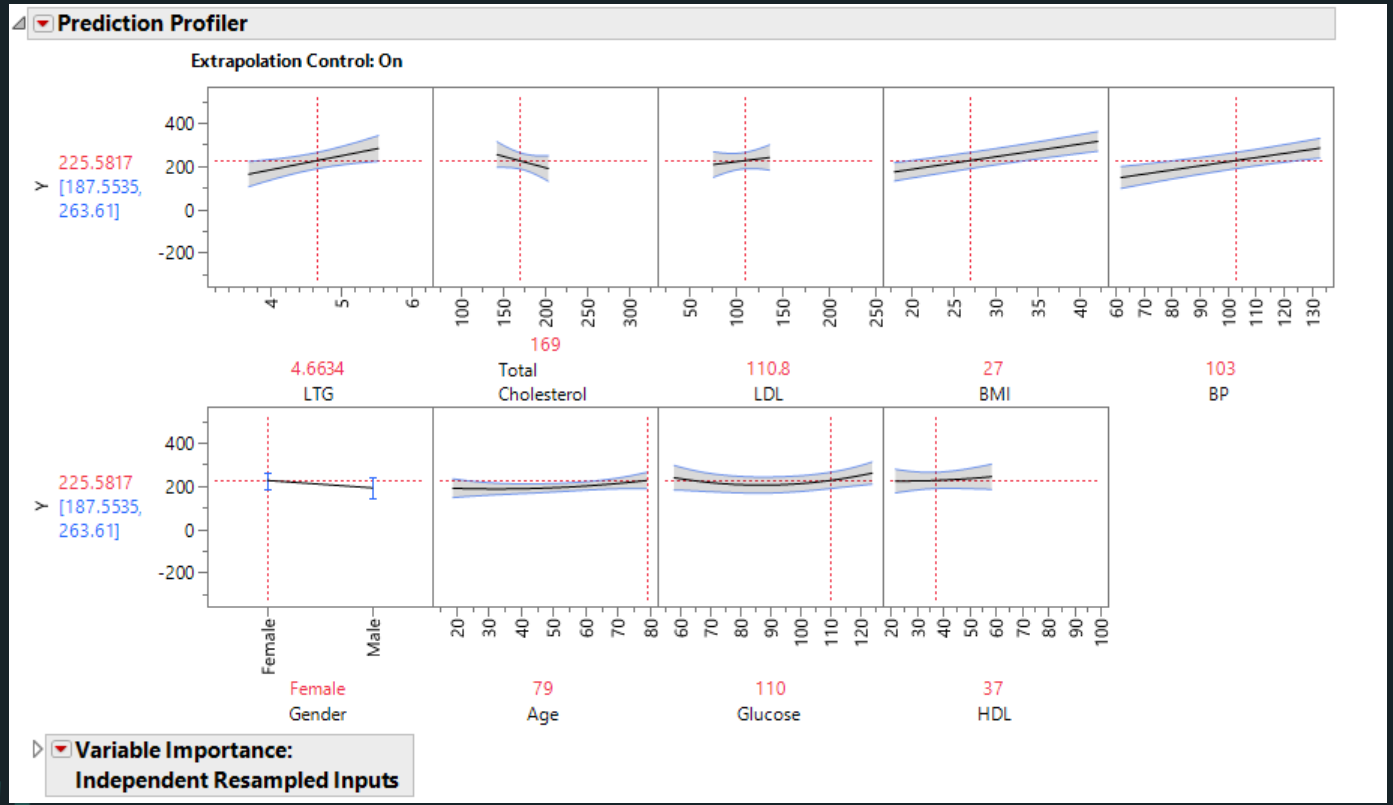
Extrapolation Control Warning in Action



Turn Extrapolation Control On



Turn Extrapolation Control On



Two Types of Extrapolation Control

- Leverage Extrapolation Control
  - Fit Model > Least Squares

- General Extrapolation Control
  - Fit Model > Generalized Regression
  - Graph > Profiler
  - Neural
  - Partial Least Squares
  - Support Vector Machines
  - Naïve Bayes

Least Squares Extrapolation Control

- Fit Model > Least Squares
- Based on Leverage
- $Lev\ (x) = x^T(X^TX)^{-1}x$
- x is the prediction point
- X is the design matrix of training data
- Leverage is equivalent to a scaled prediction variance

General Extrapolation Control Design Goals

- Fast to fit and score
- Continuous and categorical data
- Easy to automate ‘out of the box’
- Unsupervised model of the *Xs*
- Totally separate from prediction model
- Robust to missing cells
- Robust to linear dependencies
- Relatively easy to explain

Generalized Extrapolation Control

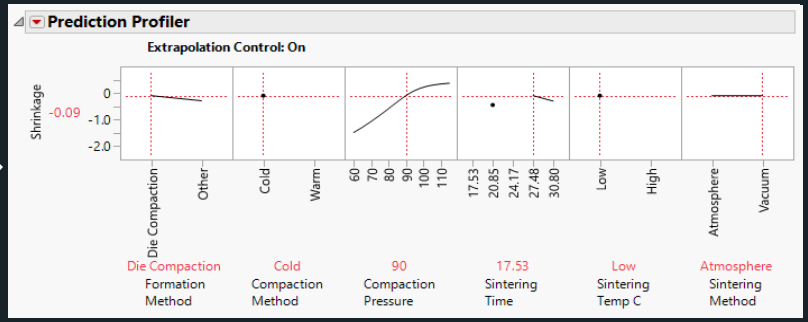
X

	Formation Method	Compaction Method	Compaction Pressure	Sintering Time	Sintering Temp C	Sintering Method
1	Die Compaction	Cold	90	27.48	Low	Vacuum
2	Other	Cold	90	27.48	Low	Atmosphere
3	Other	Cold	100	27.48	Low	Atmosphere
4	Other	Cold	95	27.48	Low	Atmosphere
5	Other	Cold	85	27.48	Low	Atmosphere
6	Other	Cold	95	27.48	Low	Atmosphere
7	Other	Cold	100	27.48	Low	Atmosphere
8	Other	Cold	100	27.48	Low	Atmosphere
9	Other	Cold	105	20.85	Low	Atmosphere
10	Other	Cold	85	27.48	Low	Atmosphere
11	Other	Cold	85	27.48	Low	Atmosphere
12	Other	Cold	80	27.48	Low	Atmosphere

Supervised Model

Y

	Shrinkage
1	0.19
2	-0.37
3	-0.16
4	0.41
5	-0.65
6	0.41
7	-0.23
8	0.12
9	0.41
10	-0.72
11	-1.08
12	-0.65



Unsupervised Extrapolation Control

	Formation Method	Compaction Method	Compaction Pressure	Sintering Time	Sintering Temp C	Sintering Method
1	Die Compaction	Cold	90	27.48	Low	Vacuum
2	Other	Cold	90	27.48	Low	Atmosphere
3	Other	Cold	100	27.48	Low	Atmosphere
4	Other	Cold	95	27.48	Low	Atmosphere
5	Other	Cold	85	27.48	Low	Atmosphere
6	Other	Cold	95	27.48	Low	Atmosphere
7	Other	Cold	100	27.48	Low	Atmosphere
8	Other	Cold	100	27.48	Low	Atmosphere
9	Other	Cold	105	20.85	Low	Atmosphere
10	Other	Cold	85	27.48	Low	Atmosphere
11	Other	Cold	85	27.48	Low	Atmosphere
12	Other	Cold	80	27.48	Low	Atmosphere

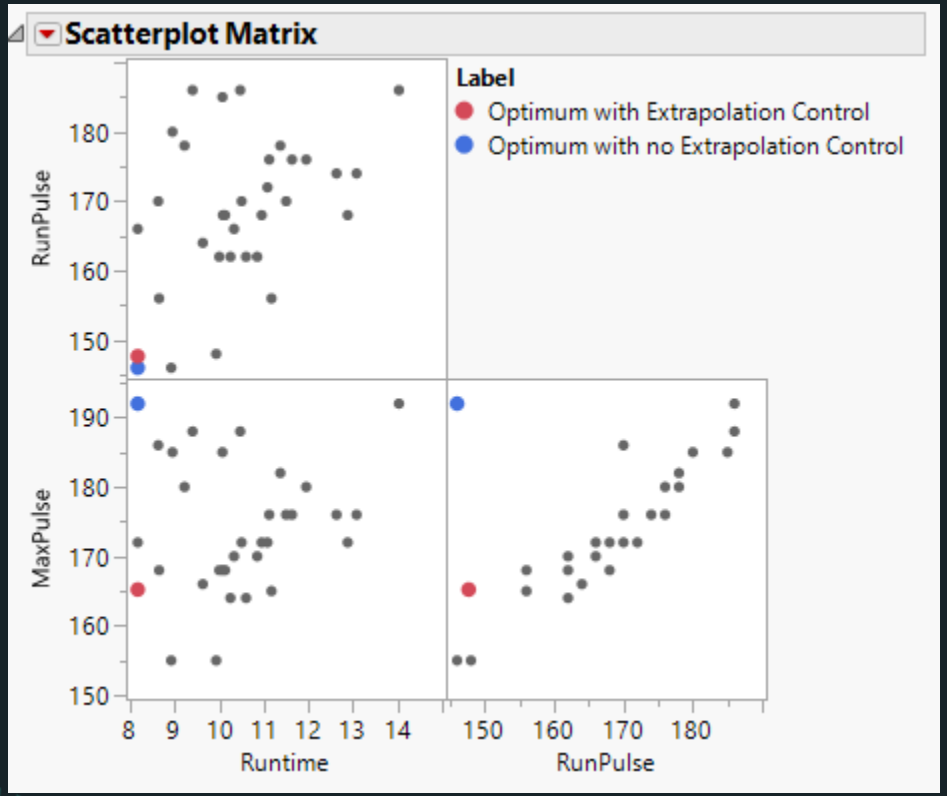
	Formation Method	Compaction Method	Compaction Pressure	Sintering Time	Sintering Temp C	Sintering Method
1	Die Compaction	Cold	90	27.48	Low	Vacuum
2	Other	Cold	90	27.48	Low	Atmosphere
3	Other	Cold	100	27.48	Low	Atmosphere
4	Other	Cold	95	27.48	Low	Atmosphere
5	Other	Cold	85	27.48	Low	Atmosphere
6	Other	Cold	95	27.48	Low	Atmosphere
7	Other	Cold	100	27.48	Low	Atmosphere
8	Other	Cold	100	27.48	Low	Atmosphere
9	Other	Cold	105	20.85	Low	Atmosphere
10	Other	Cold	85	27.48	Low	Atmosphere
11	Other	Cold	85	27.48	Low	Atmosphere
12	Other	Cold	80	27.48	Low	Atmosphere

Extrapolation Control

Fit Model Least Squares Example

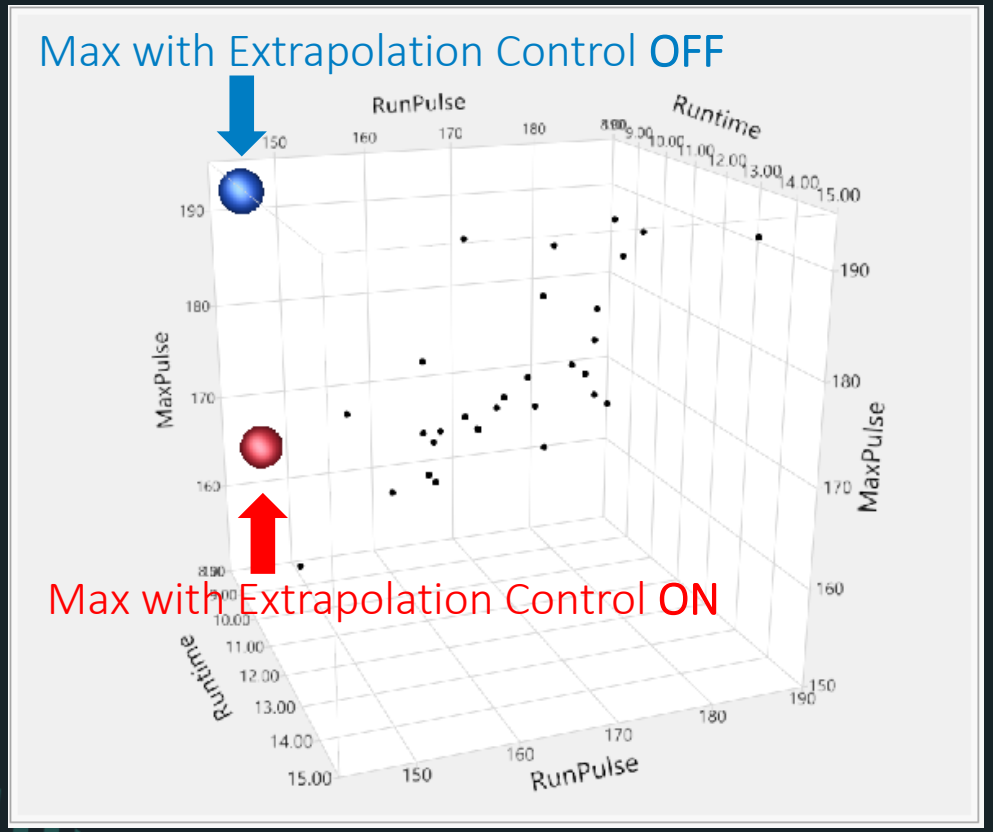
- Fitness Data
  - Response – Oxygen uptake during exercise
  - Predictors – Run Time, Run Pulse, Max Pulse
- Extrapolation Metric – Leverage

Fit Model Least Squares – Fitness Data

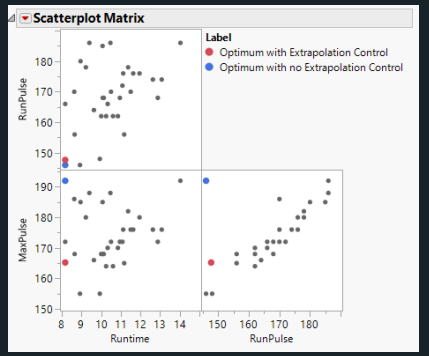


- Extrapolation Metric is leverage.
- Extrapolation Control Threshold is 3*average leverage.

Fit Model Least Squares – Fitness Data



- Extrapolation Metric is leverage.
- Extrapolation Control Threshold is 3*average leverage.

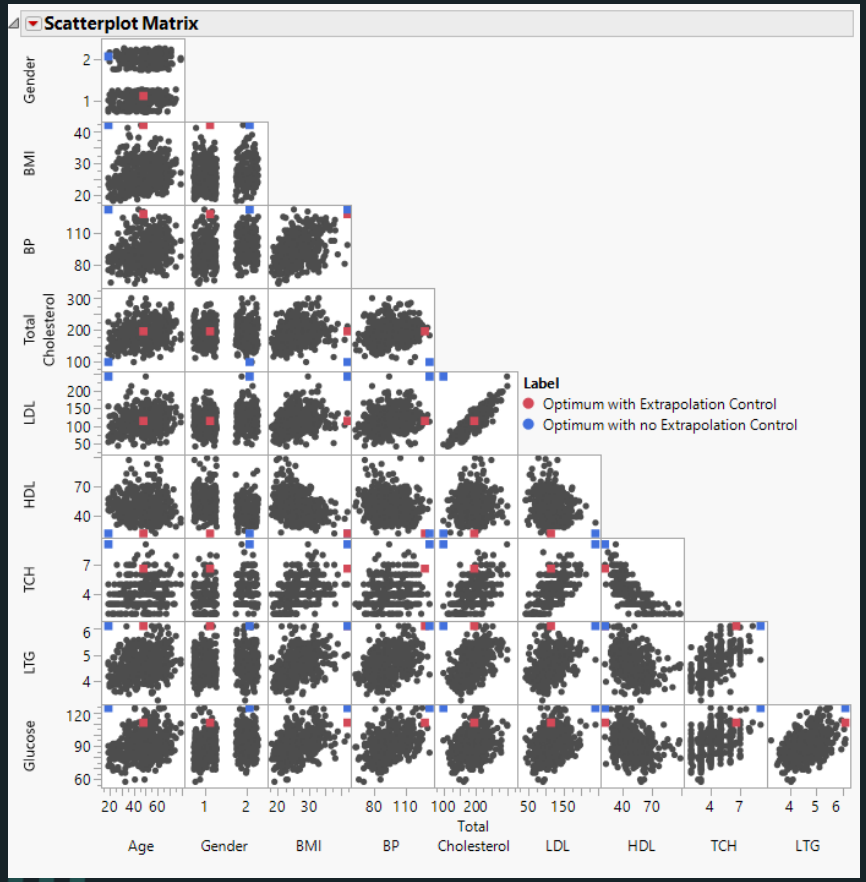


Extrapolation Control

Neural Model Example

- Diabetes Data
  - Response – Measure of disease progression one year after baseline variables taken.
  - Predictors – Baseline variables: age, gender, body mass index, average blood pressure, and six blood serum measurements
- Extrapolation Metric – Regularized T Square

Neural Model - Diabetes Data



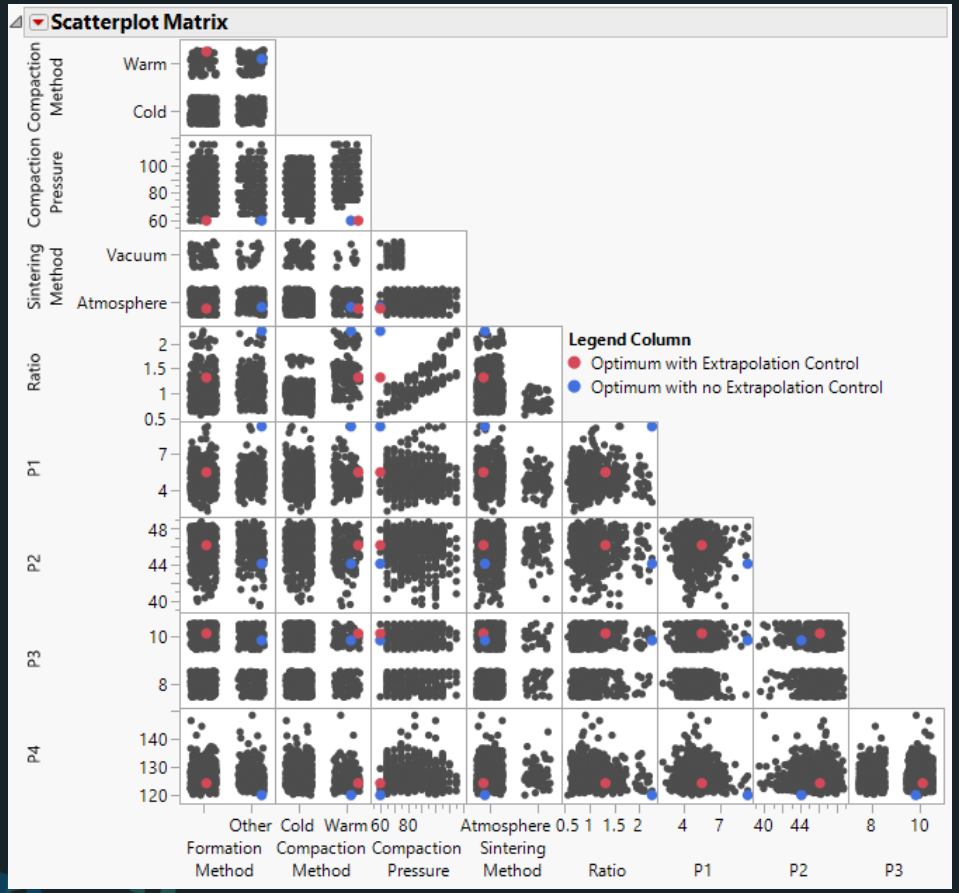
- Extrapolation Metric is regularized T Square.
- Extrapolation Control Threshold is 3*standard deviation of sample regularized T Squares.

Extrapolation Control

Graph Profiler Example with Two Models

- Powder Metallurgy Data
  - Responses
    - Shrinkage – Least Squares Model
    - Surface Condition (pass/fail) – Nominal Logistic Model
  - Predictors – Formation method, compaction method, compaction pressure, sintering method, ratio, P1-P4.
- Extrapolation Metric – Regularized T Square

Graph Profiler – Powder Metallurgy Data



- Extrapolation Metric is regularized T Square.
- Extrapolation Control Threshold is 3*standard deviation of sample regularized T Squares.

Extrapolation Control Goals

- Fast to fit and score
  - Interactivity of traces and optimization
- Mixed data types
  - Continuous, categorical, ordinal
- Robust to missing cells
- Robust to linear dependencies
- Easy to automate

Extrapolation Control Distance Metrics

Ordinary Least Squares (OLS) – Leverage

- Leverage$(\boldsymbol{x_p}) = \boldsymbol{x_p'(X'X)^{-1}x_p}$
  - where x_p is the prediction point and X is the design matrix.
- Several interpretations
  - Multivariate distance from the center of the training data
  - Scaled prediction variance
- Intuitive thresholds for defining extrapolation.
  - Maximum Leverage – points beyond are outside the convex hull of the training data.
  - 3*Average Leverage = $3\frac{p}{n}$ where *p* is the number of model terms and *n* is the number of observations.
- Non-linear constraint on optimization – Genetic Algorithm

Generalized Extrapolation Control

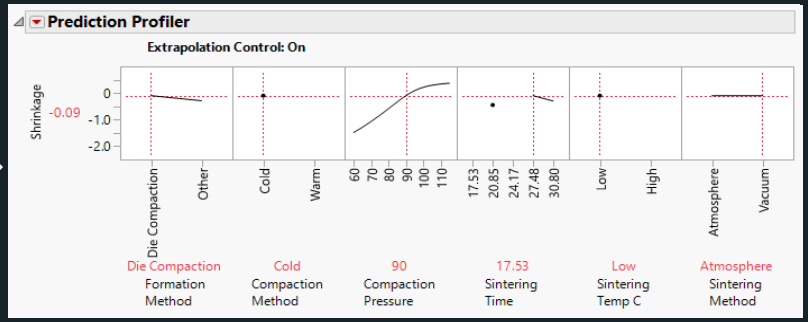
X

	Formation Method	Compaction Method	Compaction Pressure	Sintering Time	Sintering Temp C	Sintering Method
1	Die Compaction	Cold	90	27.48	Low	Vacuum
2	Other	Cold	90	27.48	Low	Atmosphere
3	Other	Cold	100	27.48	Low	Atmosphere
4	Other	Cold	95	27.48	Low	Atmosphere
5	Other	Cold	85	27.48	Low	Atmosphere
6	Other	Cold	95	27.48	Low	Atmosphere
7	Other	Cold	100	27.48	Low	Atmosphere
8	Other	Cold	100	27.48	Low	Atmosphere
9	Other	Cold	105	20.85	Low	Atmosphere
10	Other	Cold	85	27.48	Low	Atmosphere
11	Other	Cold	85	27.48	Low	Atmosphere
12	Other	Cold	80	27.48	Low	Atmosphere

Supervised Model

Y

	Shrinkage
1	0.19
2	-0.37
3	-0.16
4	0.41
5	-0.65
6	0.41
7	-0.23
8	0.12
9	0.41
10	-0.72
11	-1.08
12	-0.65



Unsupervised Extrapolation Control

	Formation Method	Compaction Method	Compaction Pressure	Sintering Time	Sintering Temp C	Sintering Method
1	Die Compaction	Cold	90	27.48	Low	Vacuum
2	Other	Cold	90	27.48	Low	Atmosphere
3	Other	Cold	100	27.48	Low	Atmosphere
4	Other	Cold	95	27.48	Low	Atmosphere
5	Other	Cold	85	27.48	Low	Atmosphere
6	Other	Cold	95	27.48	Low	Atmosphere
7	Other	Cold	100	27.48	Low	Atmosphere
8	Other	Cold	100	27.48	Low	Atmosphere
9	Other	Cold	105	20.85	Low	Atmosphere
10	Other	Cold	85	27.48	Low	Atmosphere
11	Other	Cold	85	27.48	Low	Atmosphere
12	Other	Cold	80	27.48	Low	Atmosphere

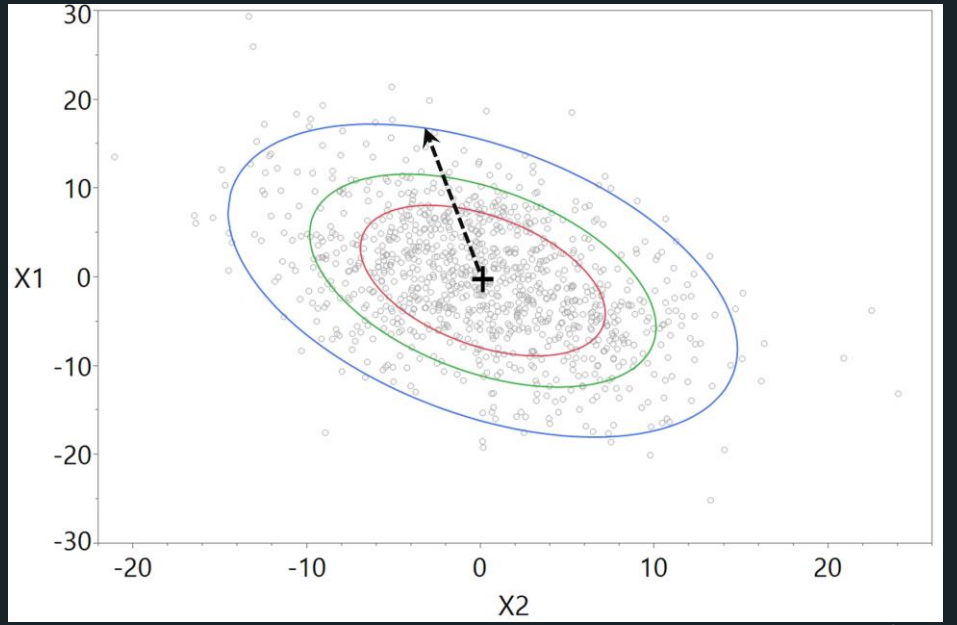
	Formation Method	Compaction Method	Compaction Pressure	Sintering Time	Sintering Temp C	Sintering Method
1	Die Compaction	Cold	90	27.48	Low	Vacuum
2	Other	Cold	90	27.48	Low	Atmosphere
3	Other	Cold	100	27.48	Low	Atmosphere
4	Other	Cold	95	27.48	Low	Atmosphere
5	Other	Cold	85	27.48	Low	Atmosphere
6	Other	Cold	95	27.48	Low	Atmosphere
7	Other	Cold	100	27.48	Low	Atmosphere
8	Other	Cold	100	27.48	Low	Atmosphere
9	Other	Cold	105	20.85	Low	Atmosphere
10	Other	Cold	85	27.48	Low	Atmosphere
11	Other	Cold	85	27.48	Low	Atmosphere
12	Other	Cold	80	27.48	Low	Atmosphere

jmp STATISTICAL DISCOVERY FROM SAS

Extrapolation Control Distance Metrics

General Predictive Modeling - Hotelling's T^2

- **Leverage** can be interpreted as a measure of multivariate distance from the mean of the data.
- This suggests **Hotelling's T^2** as a distance metric for predictive models in general.
  - Distributional assumptions determine an upper control limit.



jmp STATISTICAL DISCOVERY FROM SAS

Hotelling's T^2

- Hotelling's T^2 defined as:

$$T^2 = (x - \bar{x})^T \hat{\mathbf{\Sigma}}^{-1} (x - \bar{x})$$

- If p < n and the predictors are multivariate normal, then:

$$T^2_{pred} \sim \frac{(n+1)(n-1)p}{n(n-p)} F(p, n-p)$$

- We use a more generalized 3 sigma control limit:

$$UCL = \bar{T}^2 + 3\hat{\sigma}_{T^2}$$

Regularized Hotelling's T^2

- It is possible that p > n, so we use Schafer and Strimmer's shrinkage estimator[1]:

$$\hat{\mathbf{\Sigma}} = (1 - \hat{\lambda})\hat{\mathbf{U}} + \hat{\lambda}\hat{\mathbf{D}}$$

- For lambda, there is an analytical expression
- For target matrix (D), use a diagonal matrix with the predictor variances on the diagonal
- Advantages for extrapolation control
  - Prior is that predictors are **uncorrelated**
  - When data is limited, extrapolation control is **conservative**
  - See arXiv paper for simulation details
    - https://arxiv.org/pdf/2201.05236.pdf (Also submitted to Journal of Computational and Graphical Statistics)

J. Schafer, K. Strimmer. Stat. Appl. Genet. Mol. 4.1 (2005).

Regularized Hotelling's T^2: Additional Details

- Categorical variables are dummy encoded
- When there are missing values, a pairwise deletion method is used to estimate the covariance matrix, $\hat{\mathbf{U}} = ((u^{kl}))$, where:

$$u^{kl} = \frac{\sum_{i=1}^{n}(x_i^k - \bar{x}^k)(x_i^l - \bar{x}^l)\mathbb{1}(x_i^k \neq NA, x_i^l \neq NA)}{\mathbb{1}(x_i^k \neq NA, x_i^l \neq NA)}$$

$$\bar{x}^k = \frac{\sum_{i=1}^{n} x_i^k \mathbb{1}(x_i^k \neq NA)}{\mathbb{1}(x_i^k \neq NA)}$$

Additional Advantages of Regularized T^2

- Regularization balances bias-variance tradeoff
- Shown to work well in high dimensional settings
- Closely related to PCA/PLS models with a T^2 and DModX constraint
- Regularized T^2 doesn't require projection
  - Generalizes well to other types of models
  - Robust to non-linear relationships between predictors
  - One metric simplifies usage and interpretation

Summary

- Better visualization of feasible regions for high dimensional models in the profiler
- Genetic algorithm for flexible constrained optimization
- Handles messy observational data
- Available in predictive modeling platforms
- Available in the graph profiler

Future directions

- K-nearest neighbor-based metric for extrapolation control

Email us with any further questions!

- Christopher.Gotwalt@jmp.com
- Laura.Lancaster@jmp.com

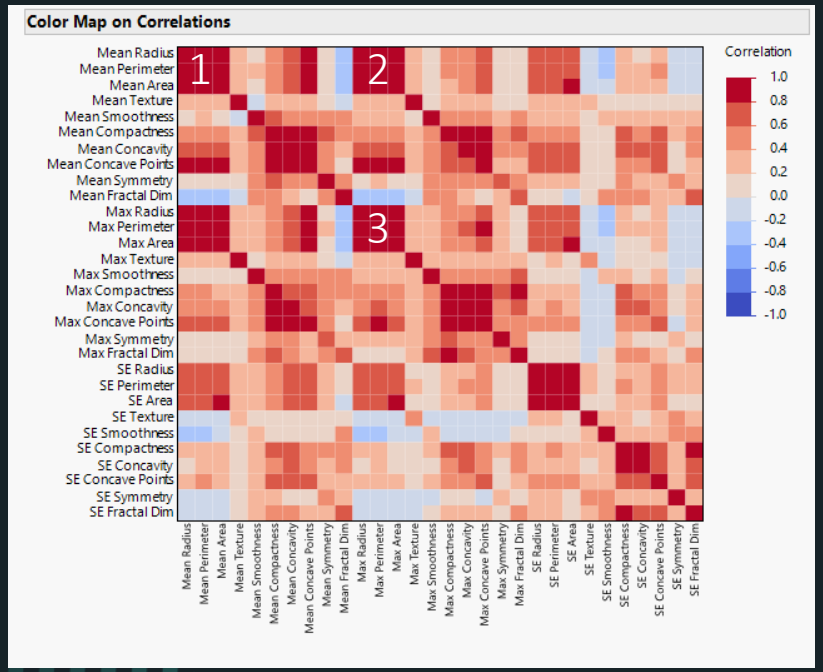
- Tom.Donnelly@jmp.com
- Elizabeth.Claassen@jmp.com

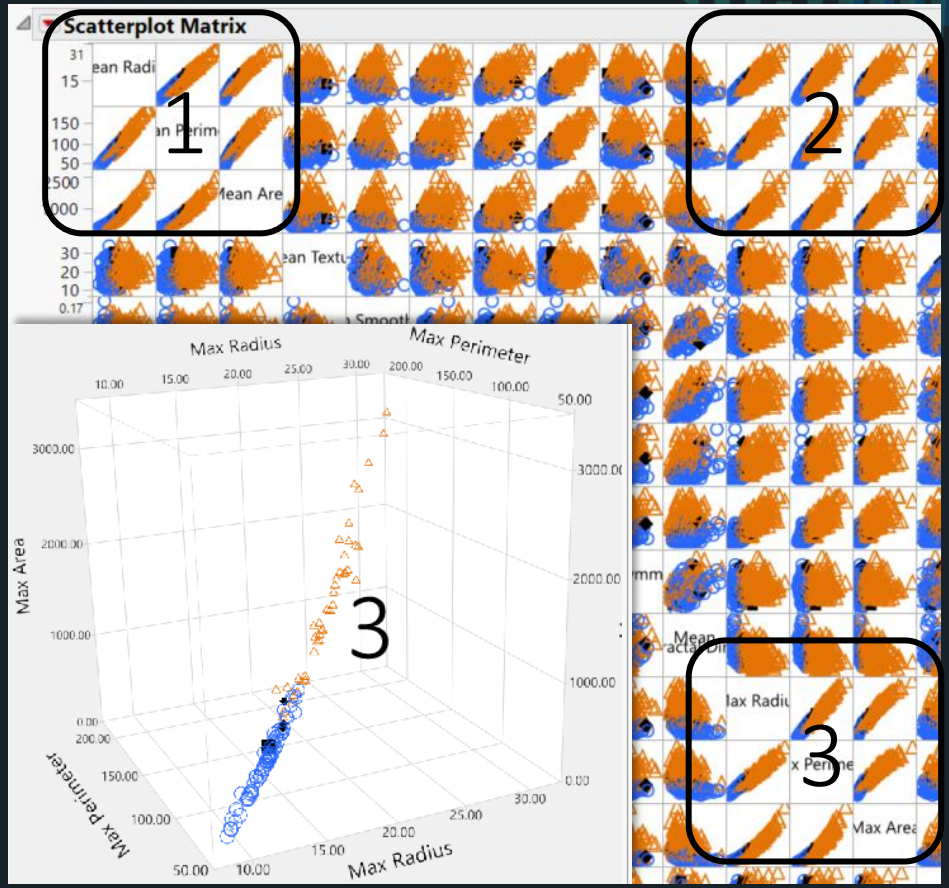
Supplemental Slides Added by Tom Donnelly

- Four slides from a presentation on Generalized Regression with highly correlated data and using Extrapolation Control
- Five slides from an Army presentation at MORSS in 2008 discussing extrapolation with empirical and physics-based models and showing the convex hull

30 Predictors for 569* Breast Cancer Tumors Diagnosed as Malignant (212) or Benign (357)

Some Highly Correlated† Predictors





* 341 Training, 114 Validation, 114 Test

† Tumor: Radius ≈ r, Perimeter ≈ 2πr, & Area ≈ πr²

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Data can be found on UCI Machine Learning Repository.

jmp STATISTICAL DISCOVERY FROM SAS

Interactive Solution Path – Removing Terms

Watching Change in Validation *Generalized Rsquare*

(TVT for Large Data)

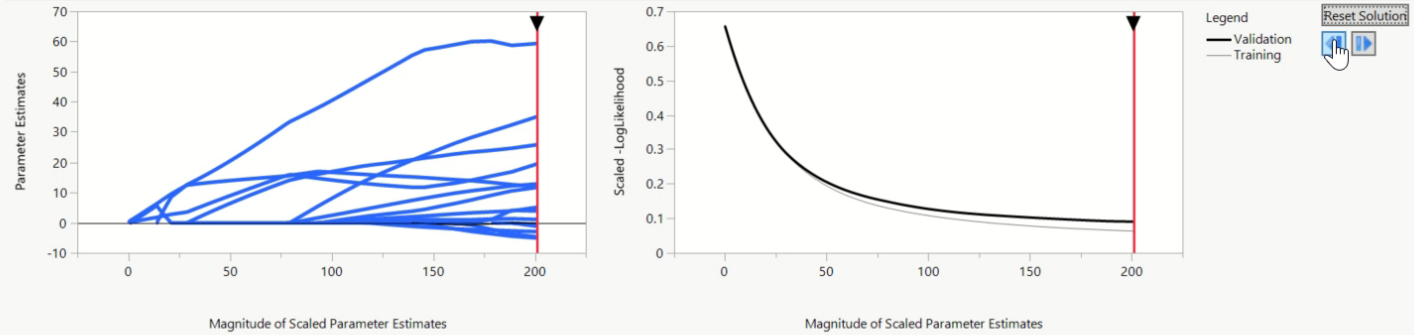
Binomial Lasso with Validation Column

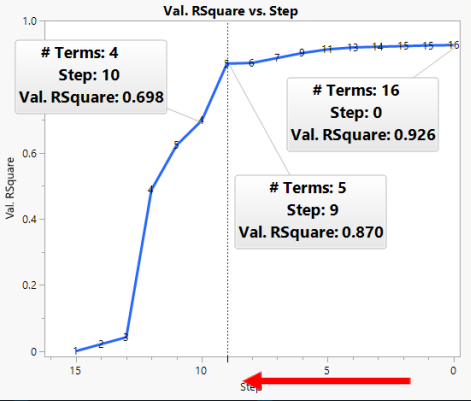
Model Summary

Measure	Training	Validation	Test
-LogLikelihood	21.733031	10.400471	6.7434495
Number of Parameters	16	16	16
BIC	136.77618	96.580118	89.266074
AICc	77.145074	58.40919	51.095146
Generalized RSquare	0.9504785	0.9266583	0.9545543
Lambda Penalty	0.0745664	.	.

Val-R^2 starts at 0.927 for 16-term model (15 effects + int.).
It ***slowly falls 6%*** to 0.870 going to 5-term model (4 effects).
Drop to 4-term model and Val-R^2 ***sharply falls 20%*** to 0.698

Solution Path



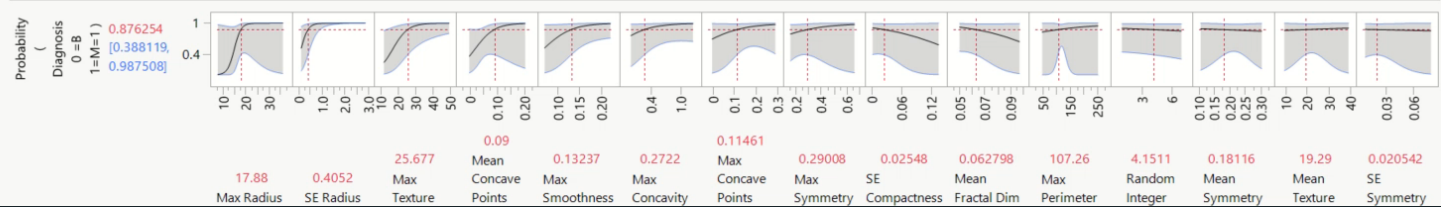


Parameter Estimates for Original Predictors

Effect Tests

Confidence intervals tighten as model shrinks.

Prediction Profiler

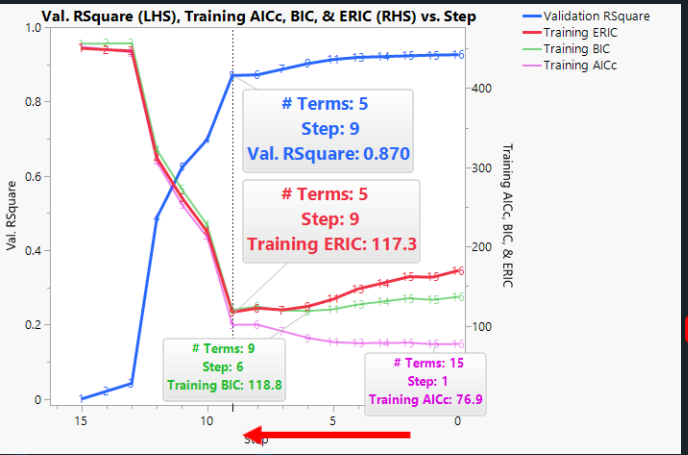


Interactive Solution Path –

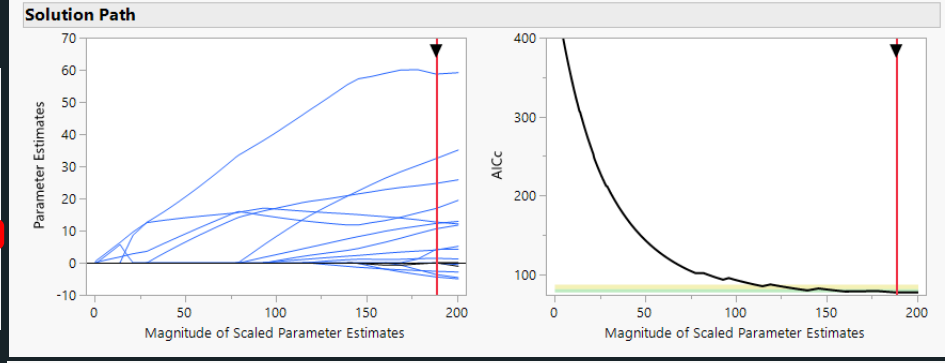
Removing Terms with Information Criteria

(Typically for Small Data)

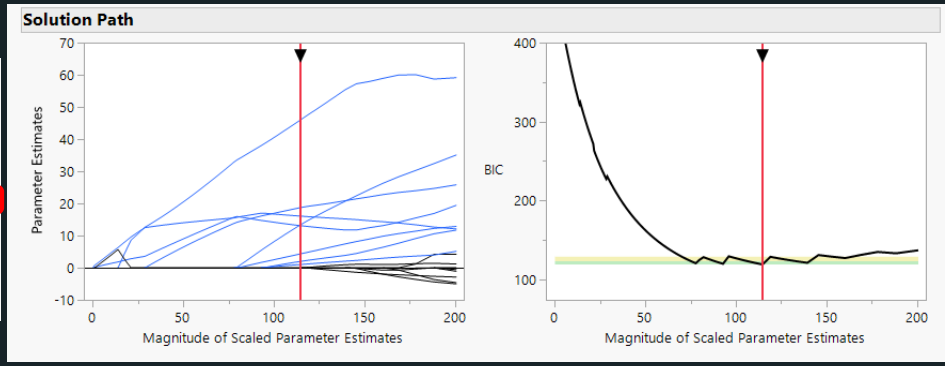
Sharp Increase in Training ***AICc***, ***BIC***, & ***ERIC*** occurs between Step 9 (5 terms) and Step 10 (4 terms)



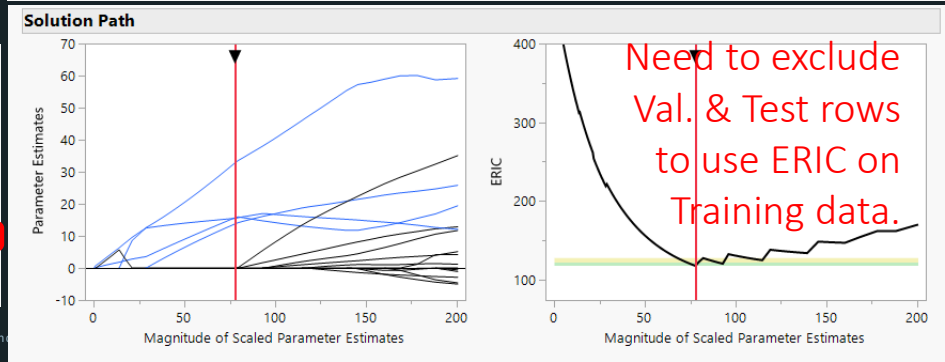
Measure	
Number of rows	569
Sum of Frequencies	341
-LogLikelihood	22.707893
Number of Parameters	15
BIC	132.89402
AICc	76.892709
ERIC	161.65703
Generalized RSquare	0.9481058
Lambda Penalty	0.0844949



Measure	
Number of rows	569
Sum of Frequencies	341
-LogLikelihood	33.144815
Number of Parameters	9
BIC	118.77657
AICc	84.833436
ERIC	124.54195
Generalized RSquare	0.9218349
Lambda Penalty	0.2346507



Measure	
Number of rows	569
Sum of Frequencies	341
-LogLikelihood	45.566649
Number of Parameters	5
BIC	120.29271
AICc	101.3124
ERIC	117.36299
Generalized RSquare	0.8883978
Lambda Penalty	0.4840634



Correlation between *Mean* & *Max Concave Points* is 91%

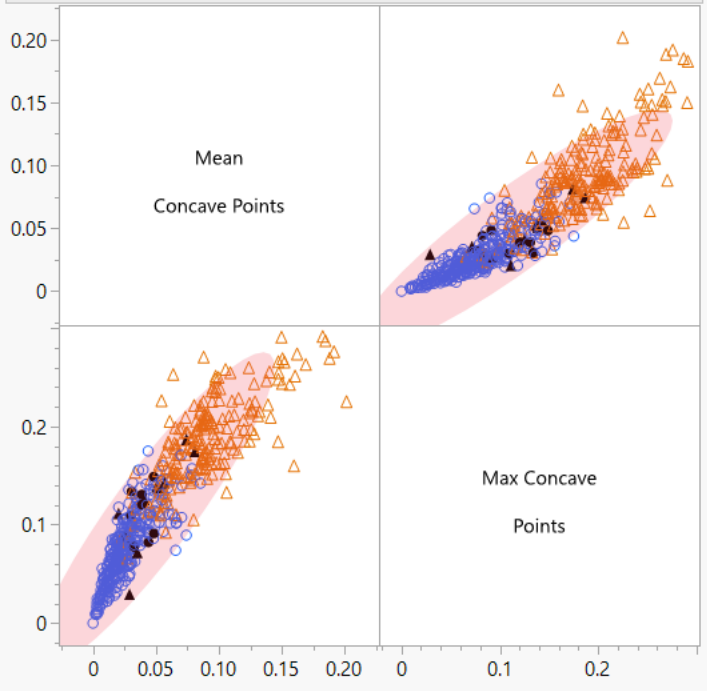
Multivariate

Correlations

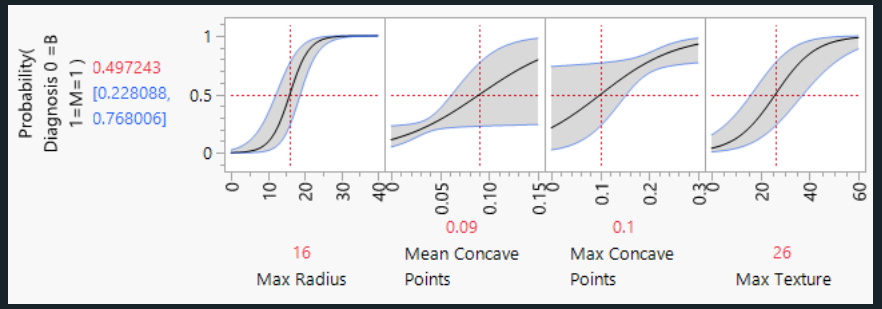
	Mean Concave Points	Max Concave Points
Mean Concave Points	1.0000	0.9102
Max Concave Points	0.9102	1.0000

The correlations are estimated by Row-wise method.

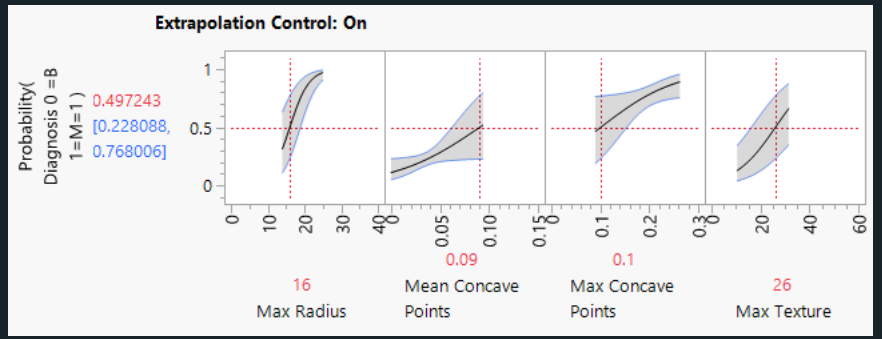
Scatterplot Matrix



Factors contain much the same information



Expanded confidence intervals indicate where little/no data



Extrapolation Control: On Prevents Profiler from making predictions where no data

jmp STATISTICAL DISCOVERY FROM SAS

Comparison of 10-term Quadratic and 4-term Linear Models

$$\log_{10}(y) = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2$$

constant + linear

+ 2-way interactions

+ curvature terms

The quadratic model can support many shapes – including; mountain, valley, ridge, saddle and plane.

$$\log_{10}(y) = A_0 + A_1X_1 + A_2X_2 + A_3X_3$$

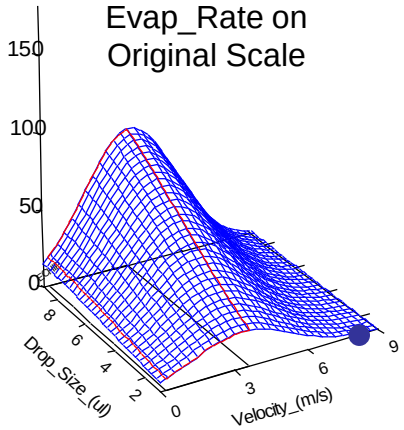
$$\text{and } X_1 = (x_1)^{-1},\ X_2 = (x_2)^{1/2},\ X_3 = (x_3)^{1/3}$$

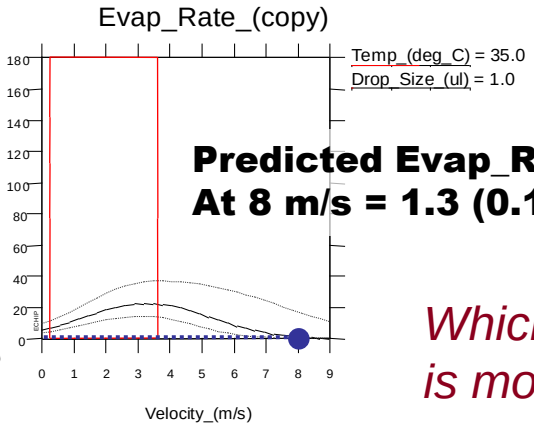
constant + linear terms

exponents used to “linearize” model

The linear model can only support a plane.

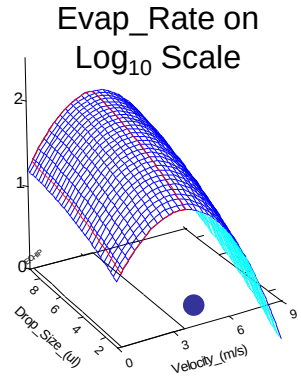
Compare Extrapolations for Empirical (Quadratic) & Physics-Based (Linear) Models (Response shown on Original Scale)

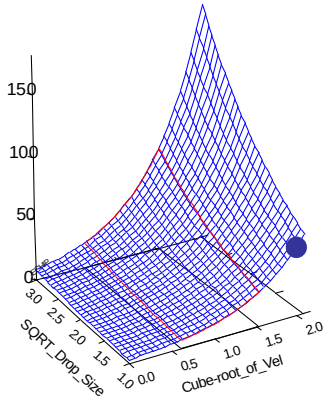


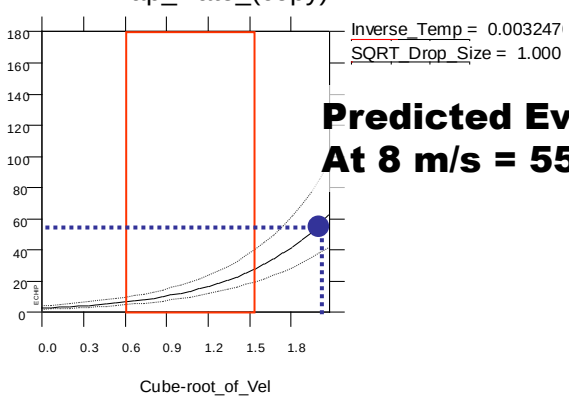


Predicted Evap_Rate
At 8 m/s = 1.3 (0.1, 17.1)

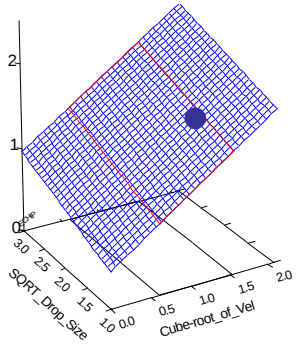
Which prediction is more plausible?



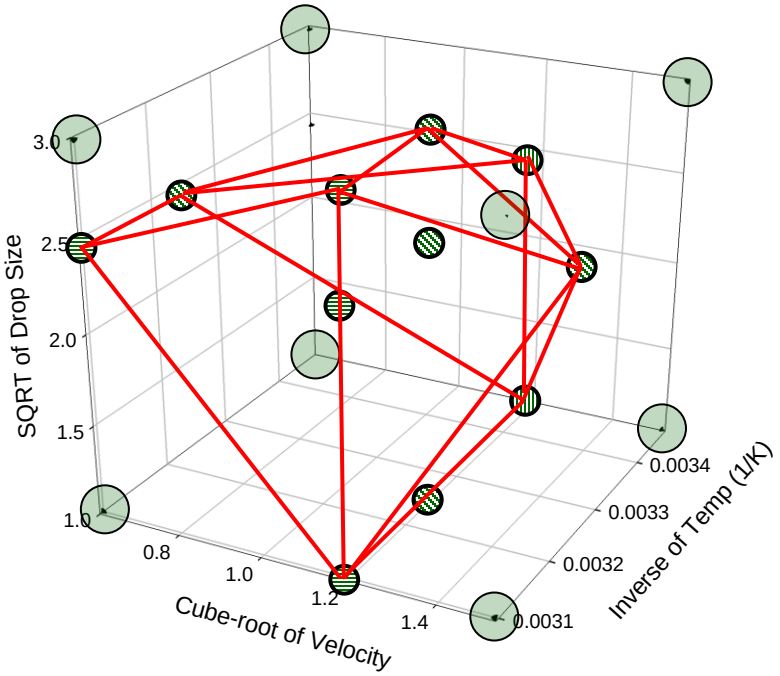




Predicted Evap_Rate
At 8 m/s = 55.1 (36.7, 82.9)



Volume Enclosed by the 11 Unique Interior Trials for the 5-cm Tunnel



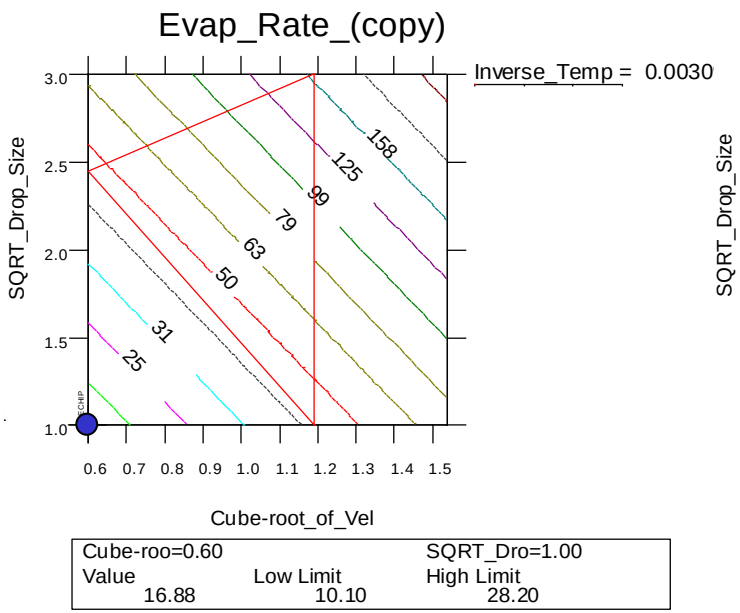
The red polyhedral shape results from “shrink wrapping” the 11 non-corner design trials for the 5-cm tunnel.

Predictions at the 8 corners of the design region made using a model fit to these 11 points are *extrapolated* predictions.

Comments on 5-cm Tunnel Analyses

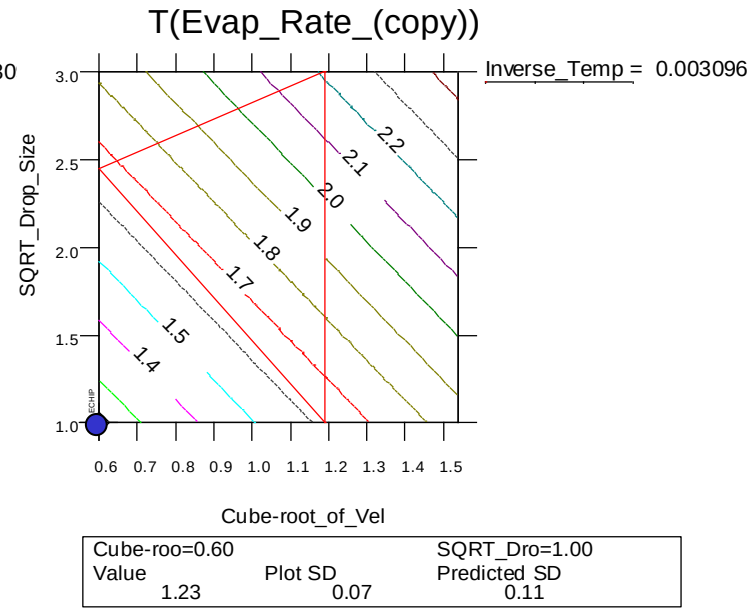
- Good news: Physics-based linearized model fits well - has slightly smaller model error (residual std. dev.) and higher Adjusted-R^2 than empirical model
- Better news: *Interpolated* model predictions based on fitting data at 8 corner design points are validated by data at locations of 11 interior design trials - which were not used in fitting model
- Even better news: Reversing the situation, the *extrapolated* model predictions based on fitting data at 11 interior points are validated by data at 8 corners - which were not used in fitting model
- Maybe best news: As few as 4 corner points + 1 center point are needed for the 80% solution…

Predictions on Raw and Transformed Scales at Location of the Poorest Performing Extrapolated Checkpoint



Cube-roo=0.60		SQRT_Dro=1.00
Value	Low Limit	High Limit
16.88	10.10	28.20

Predicted value is 16.88 on raw scale
with 95% Prediction Limits of 10.10 to 28.20
Observed value was **21.6** on raw scale



Cube-roo=0.60		SQRT_Dro=1.00
Value	Plot SD	Predicted SD
1.23	0.07	0.11

Predicted value is 1.23 on log10 scale
Within one Predicted SD (0.11) of the
Observed value of **1.33** on log10 scale

Observed $Log_{10}(21.6) = 1.33$

Predicted $10^{1.23} = 16.98$