

# Quality by Design Template

## Identification of Critical Process Parameters

### Table of Contents

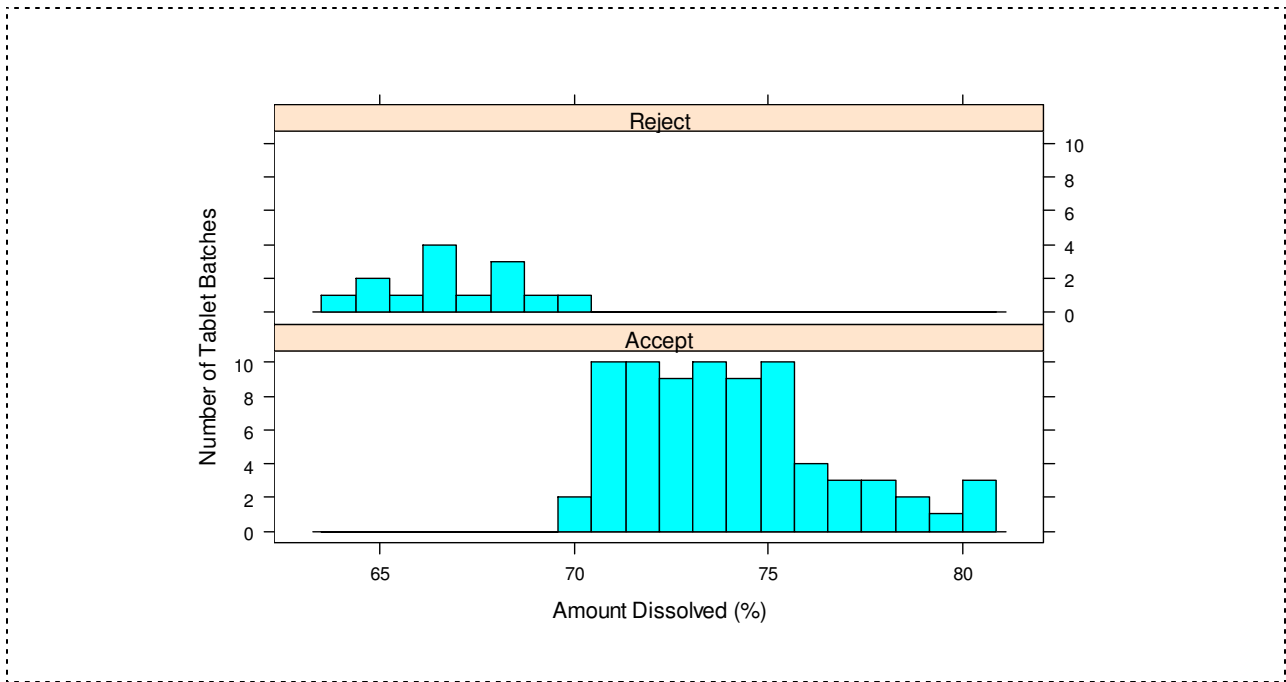
Objectives and Approach.....	2
What is the distribution between accepted and rejected lots? .....	2
Quality Performance Analysis by Random Forests Classification.....	3
Execute the Random Forest Classifier on Manufacturing Data.....	3
Are there major outliers in the data set? .....	4
How good is the Random Forest classifier?.....	4
Which variables are Critical Process Parameters?.....	5
Do Accepted/Rejected Lots cluster into separate groups? .....	6
Trends Exhibited by Six Top Critical Process Parameters .....	7
How do setting values of CPP affect probability of lot rejection?.....	7
Critical Process Parameter Signatures .....	8
Are the CPP signatures for accepted and rejected lots distinguishable?.....	8

## Objectives and Approach

The objective is to link a Critical to Quality Attribute (CQA; e.g., Dissolution at 60 minutes) of the drup product (e.g., formulated tablet) to the collection of processing parameters (e.g., process settings and raw material attributes) that have the largest impact on the Critical to Quality Attribute. This collection of processing parameters is termed Critical Process Parameters (CPPs).

The approach to identification of CPPs in this template is based on Random Forests, which corresponds to using an algorithmic approach to classify tablets (e.g., tablet batch accepted or tablet batch rejected on basis of conformance to CQA) using a multivariate data set of processing parameters. Random Forests employs a nonparametric approach to classification using an ensemble of decision trees. Analysis of the ensemble provides information on the factors that impact the accuracy of classification and are thereby inferred to be critical process parameters.

### What is the distribution between accepted and rejected lots?

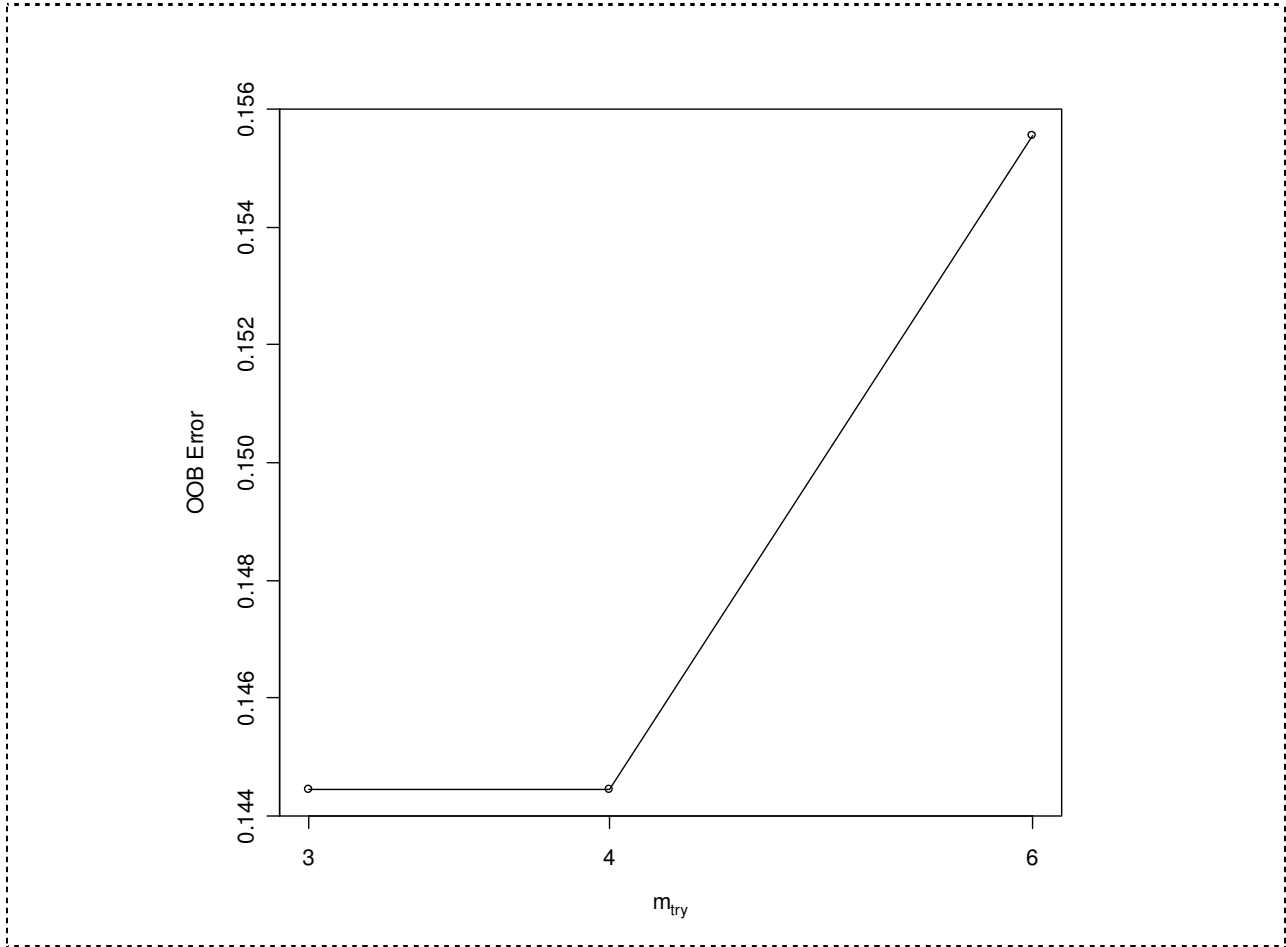


Conclusion is that accepted and rejected lots do not fall into cleanly separated groups. Rather, acceptance and rejection is a continuum. Hence, we anticipate that although we can clearly discern CPP and their settings that discriminate between acceptance and rejection at the extremes, it will be challenging to do so at the acceptance/rejection boundary specification of 70% dissolution.

**Quality Performance Analysis by Random Forests Classification**

**Execute the Random Forest Classifier on Manufacturing Data**

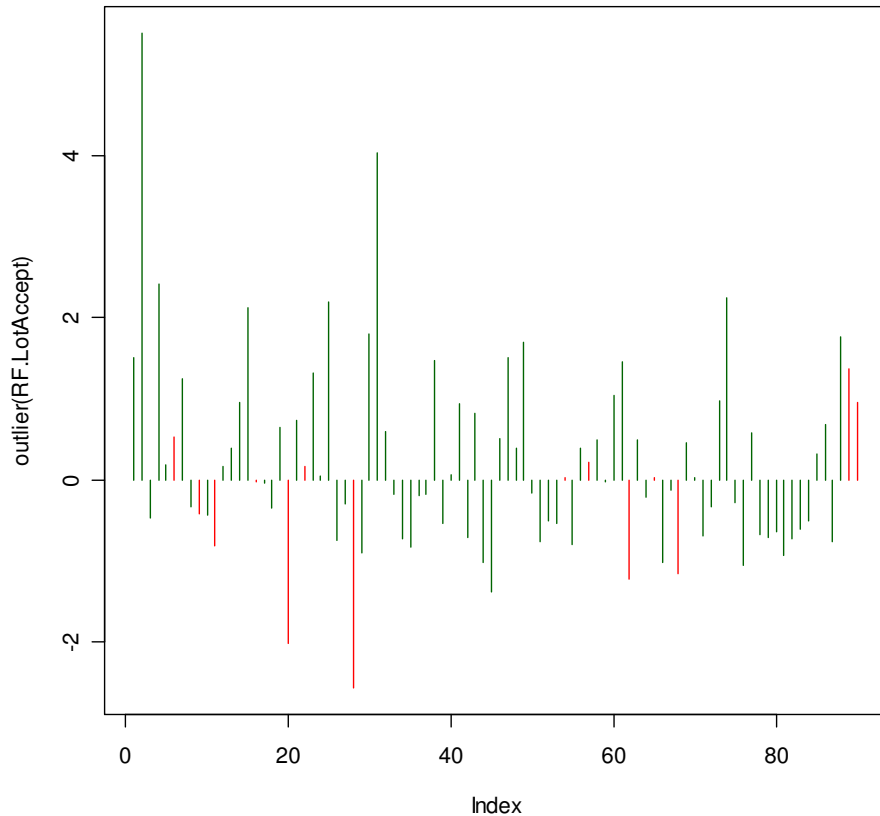
Here we will execute a Random Forest analysis on the data. The only tunable parameter is  $m_{try}$ , which corresponds the number of columns randomly sampled for each decision tree in the ensemble. We will use the `tuneRF` function to automate the tuning and select the best value.



Three actions taken on the data:

- Remove the API lot number from variable importance considerations since each tablet lot used a unique API lot.
- Also remove the two responses in columns 19 and 20 corresponding to the %dissolved and reject/accept status from the input set.
- Adjust the sample size to reflect the fact that the data set is imbalanced—that is, there are only 14 rejected lot cases relative to 76 accepted cases.

## Are there major outliers in the data set?



No remarkable outliers (values  $< -10$  or  $> +10$ ) are indicated by the analysis.

## How good is the Random Forest classifier?

The objective here is to get a sense of how accurate the Random Forest classifier in predicting acceptance or rejection of tablet batches given the manufacturing conditions.

```
Call:
  randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1],      sampsize = ..2,
  proximity = TRUE)
  Type of random forest: classification
  Number of trees: 500
  No. of variables tried at each split: 3

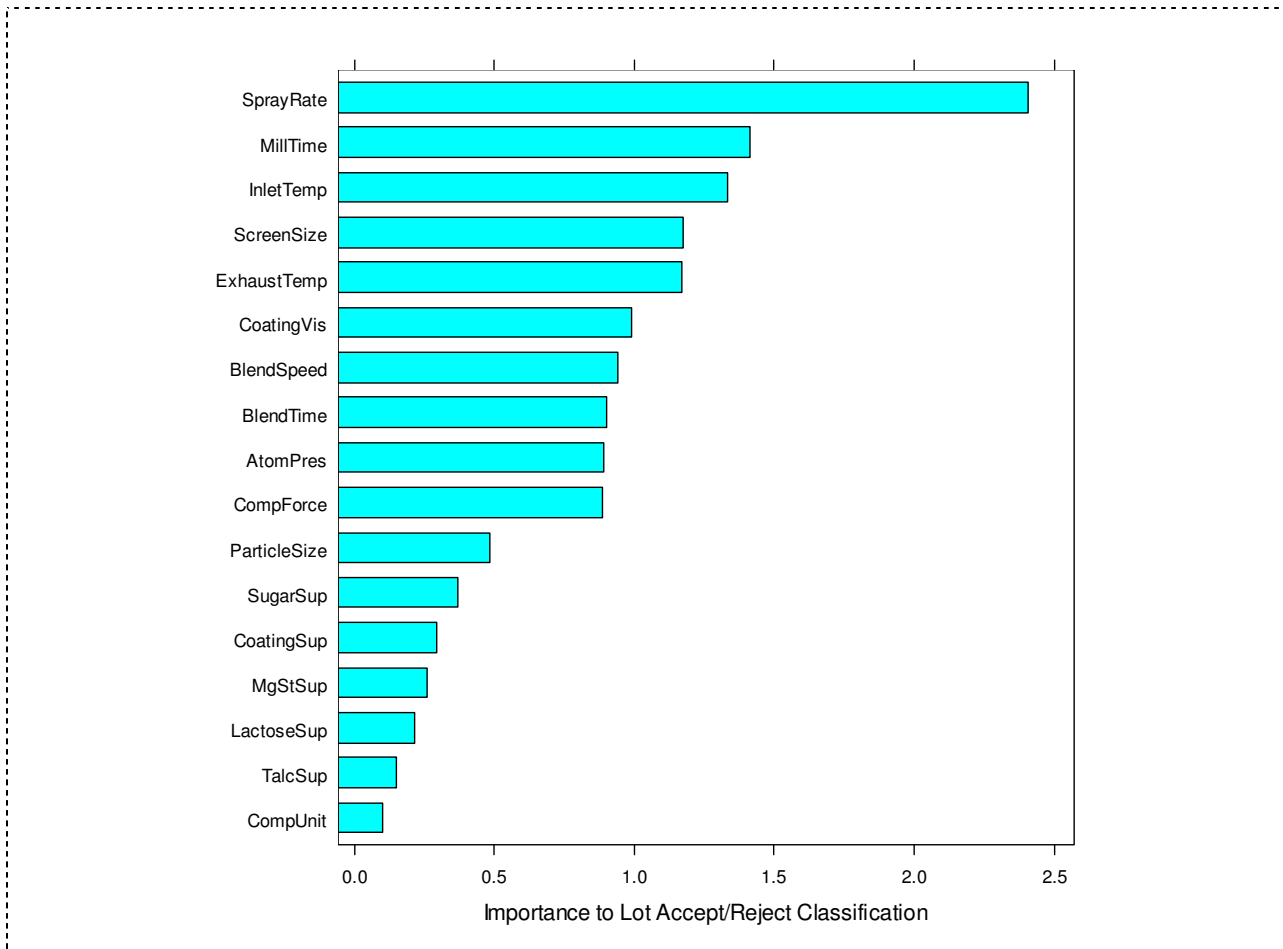
  OOB estimate of error rate: 13.33%
  Confusion matrix:
    Accept Reject class.error
Accept    71     5 0.06578947
Reject     7     7 0.50000000
```

Results of the confusion matrix are summarized and explained in the table below:

LotAccept Class	Cases Actual	Cases Predicted Accept	Cases Predicted Reject	Prediction Error
Accept	[1] 76	TruePos:[1] 71	FalsePos:[1] 5	[1] 0.06578947
Reject	[1] 14	FalseNeg:[1] 7	TrueNeg:[1] 7	[1] 0.5

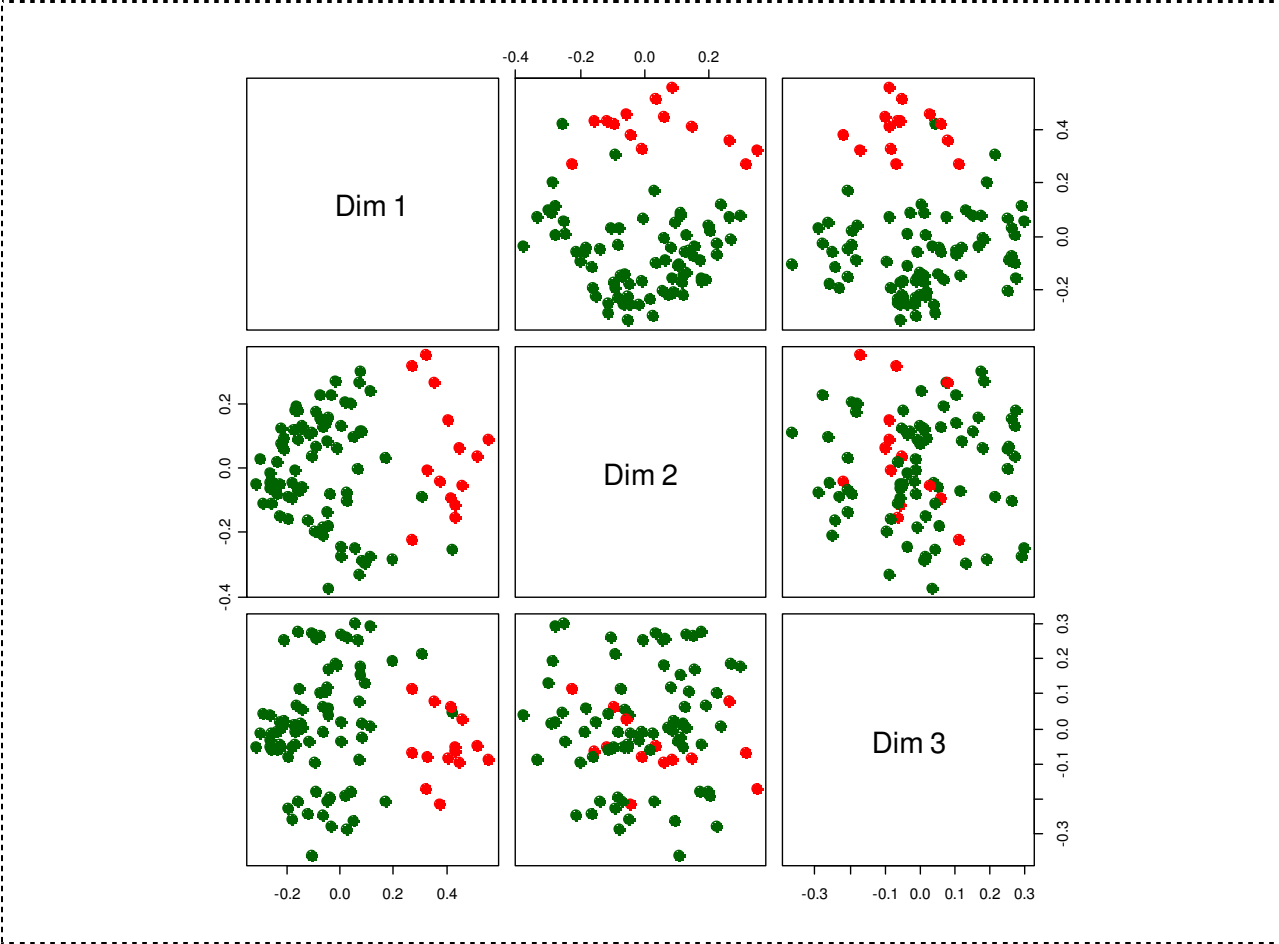
### Which variables are Critical Process Parameters?

Variables that impact the accuracy of lot acceptance/rejection classification are identified as critical process parameters (CPP). If they do not impact the accuracy, they are deemed not to be important and thus can be eliminated for further study.



**Do Accepted/Rejected Lots cluster into separate groups?**

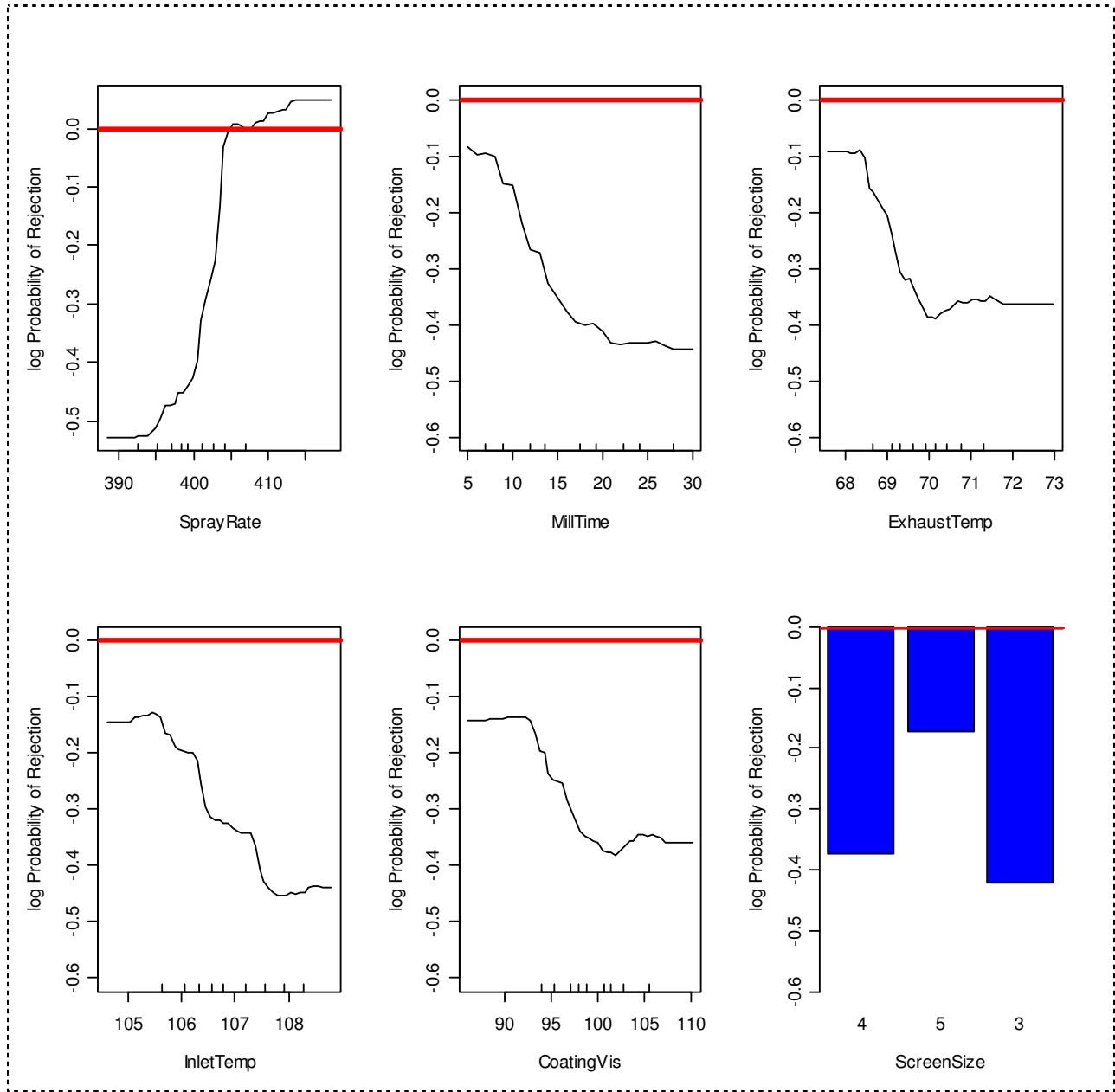
One of the tests of critical process parameters is that their settings cause the tablet batches to cluster into separate accept and reject groups. Using Random Forest proximities tests this hypothesis.



**Trends Exhibited by Six Top Critical Process Parameters**

**How do setting values of CPP affect probability of lot rejection?**

**Note:** The analysis focuses on the LotAccept class "Reject"—that is, what causes the batch to be rejected. The vertical axis corresponds to the log of the probability that the batch will be rejected given the setting of the CPP, all other factors being equal, with the red line corresponding to 100% probability ( $\log(1) = 0$ )



**Critical Process Parameter Signatures**

**Are the CPP signatures for accepted and rejected lots distinguishable?**

Here we use a parallel coordinates plot, where each line comprises the settings of the ordered CPP (most important at the bottom; least important at the top) for each production batch, making each line a CPP signature for a batch. We create separate plots for accepted and rejected batches. By design, there is no scale on the x-axis; instead, a plotting range for CPP is selected to show the span of parameter values for that CPP.

