Case Study Exercise #1

## A simple example of using functional data

This is an example to familiarise you with the Functional Data Explorer platform.

Select Help > Sample Data Library and open Weekly Weather Data.jmp.

This example analyzes weekly weather data collected from 16 weather stations across the United States. Run the Weather Station Locations script in the data table to view a map of the locations. Daily temperatures are summarized as weekly averages. Not every weather station has a weekly temperature measurement for every week of the year. This is an example of sparse functional data.

Plot the data

1. Graph > Graph Builder. Y = TMAX, X = Week of Year, Wrap = STATION.
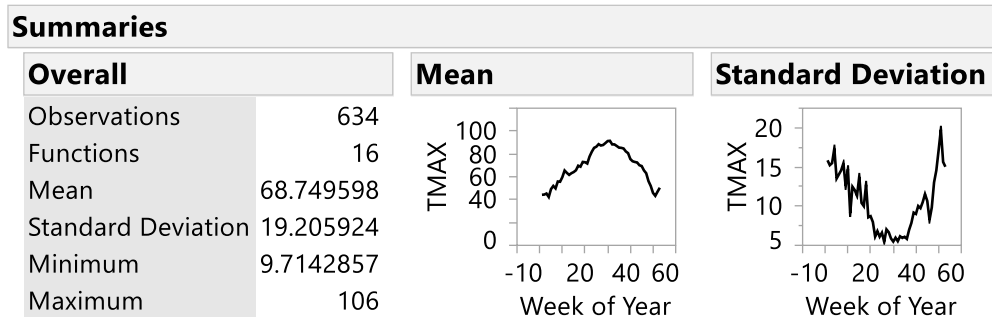2. You will probably want to resize by dragging out one of the corners.
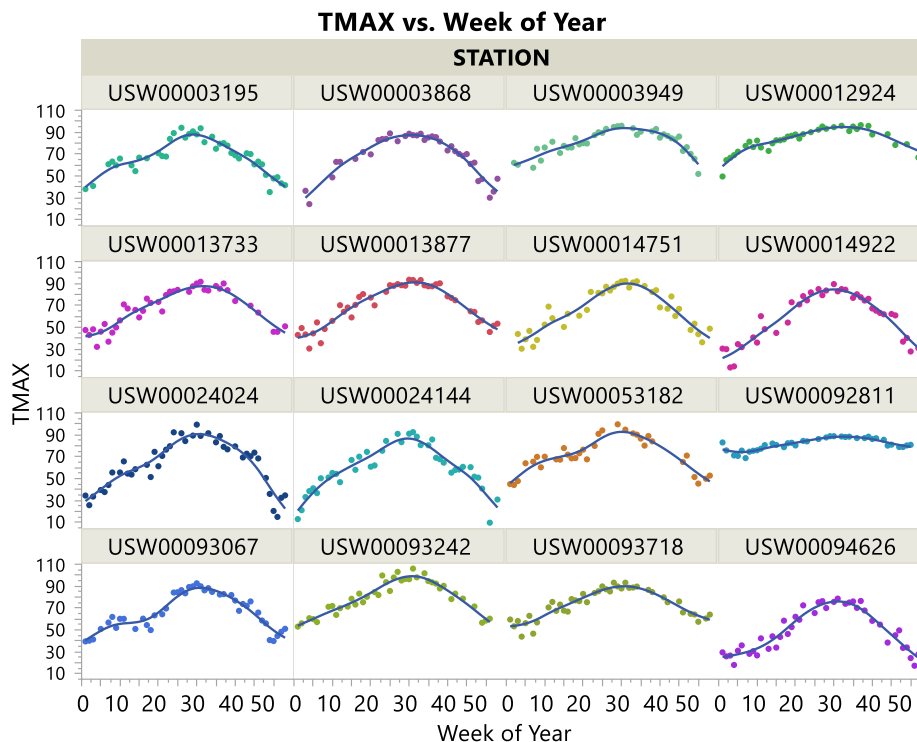
We see that the "curves", "time courses" or "profiles" are similar - they generally peak in the middle of the year, as you would expect. However, there are also differences.

Launching FDE and data processing

1. Select Analyze > Specialized Modeling > Functional Data Explorer.
2. Select TMAX and click Y, Output.
3. Select Week of Year and click X, Input.
4. Select NAME and click ID, Function.
5. Click OK.

Note: The data is in a stacked format. This is probably the easiest format to work in. We don't need to use the other tabs (Rows as Functions, Columns as Functions).

You are first given visuals like those we created in Graph Builder. You also see the Mean "trend" and how standard deviation varies with Time.

Summaries

Overall	
Observations	634
Functions	16
Mean	68.749598
Standard Deviation	19.205924
Minimum	9.7142857
Maximum	106

Mean

Standard Deviation

There are also buttons for data processing options. Data processing options are also accessible from the Data Processing red triangle menu. Prior to modelling, there are times when it might be a good idea to standardise your output data.

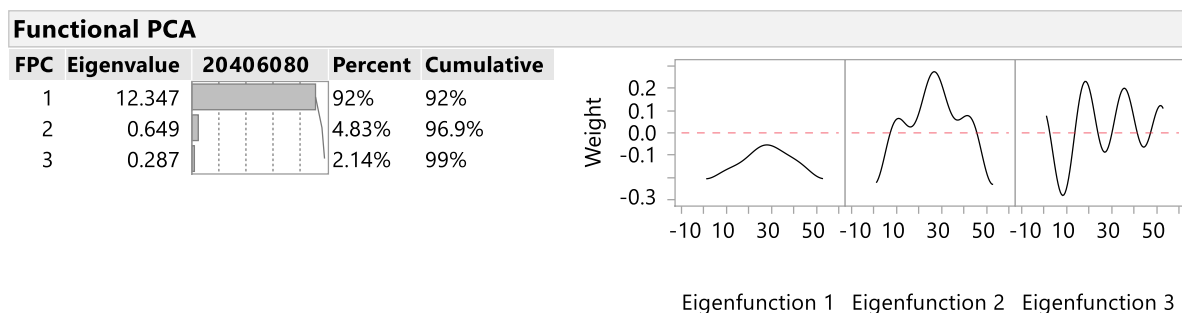6. Click the Standardize button under the Transform menu.

The data plots and summary statistics are updated based on the specified transformation and "Standardized" is added to the Steps list. You can Remove any steps in data processing.

Fitting the model

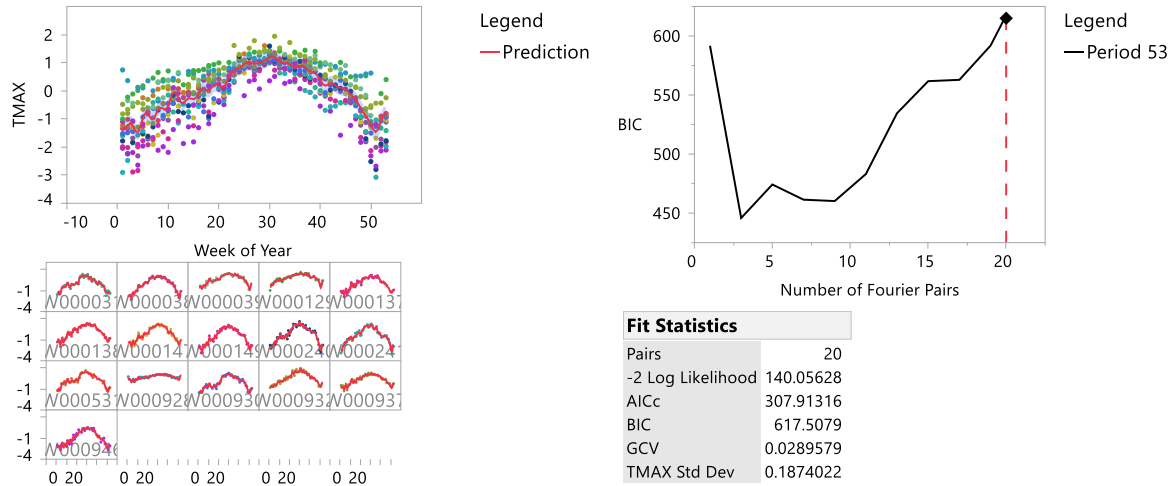1. Click the Functional Data Explorer red triangle and select Models > Fourier Basis.

There are smoothing fits to TMAX vs Week for each NAME.

FDE tells you that a Fourier Basis model with a period of 53 and three basis function pairs has the lowest BIC. FDE has also done a Prinicipal Components Analysis and found that 3 functional principal components account for >99% of the Station to Station variations from the mean trend.

Functional PCA

FPC	Eigenvalue	20 40 60 80	Percent	Cumulative
1	12.347		92%	92%
2	0.649		4.83%	96.9%
3	0.287		2.14%	99%

Eigenfunction 1   Eigenfunction 2   Eigenfunction 3

2. See what happens as you increase the number of pairs by dragging the slider from left to right.

Model Selection

Fit Statistics	
Pairs	20
-2 Log Likelihood	140.05628
AICc	307.91316
BIC	617.5079
GCV	0.0289579
TMAX Std Dev	0.1874022

3. As you change the number of pairs look at the effect on the diagnostic plots.

Now try B-spline models.

4. Click the Functional Data Explorer red triangle and select Models > B-splines.

The Functional PCA is less parsimonious - 5 FPCs are required to account for 99% of deviation from the mean trend.

5. Click on Quadratic in the legend to change the order of the fit from Cubic - the BIC is almost the same.
6. Try different orders of fit (Step, Linear, ...) and numbers of knots to get a feel for model selection.

In this case the Fourier Basis model looks like the most useful. Fourier Basis models are often good for periodic data. In other cases B-Spline or P-Spline models will be preferred. FDE makes it simple and quick to explore the different possibilities.

7. Remove the B-spline model by clicking the B-Spline on Standardize red triangle and select Remove Fit.

Looking at the Functional Principal Components

Now return to the Fourier Basis model with 3 basis pairs. You can see the 3 Eigenfunctions that describe the variation in the Station temperature trends.

Functional PCA

| FPC	Eigenvalue	20406080	Percent	Cumulative
1	12.347		92%	92%
2	0.649		4.83%	96.9%
3	0.287		2.14%	99%
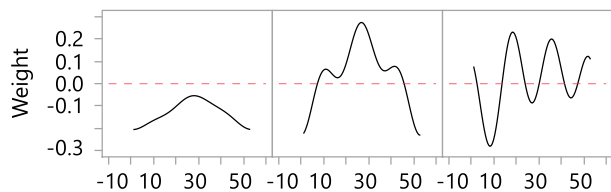
Eigenfunction 1   Eigenfunction 2   Eigenfunction 3

You can re-create any of the original temperature trends from the mean trend plus different amounts of these Eigenfunctions. The score plot tells you how much of each Eigenfunction is required for each Batch.

1. In the Score Plot deselect Label variables. Now as you hover over the data points you will see the names.
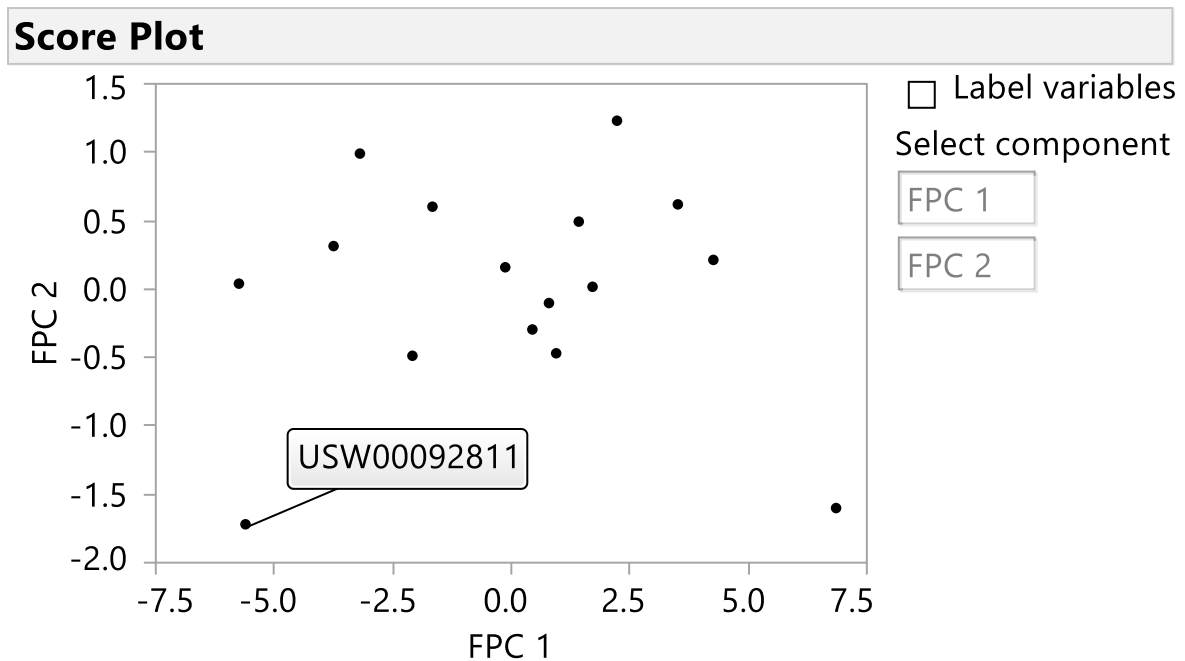
Score Plot

Explore how the FPC scores relate to the shape of the profiles. You should find that profiles that are near to each other on this score plot are more similar. Outliers in the score plot correspond to stations with very different temperature trends , e.g. USW00092811 in the lower left corner of the Score Plot.


USW00092811
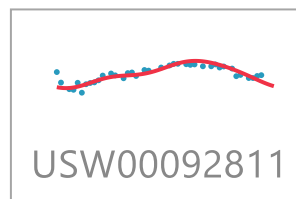
2. Save the FPC scores: from the red triangle next to Function Summaries > Save Summaries.

This gives you a table with one row per weather station. The characteristic shape of the temperature trend is described in that one row for each station. You have seen how you can comprehensively describe semi-continuous profiles, curves or time-courses with functional summaries. In later exercises you will see how you can use these functional summaries
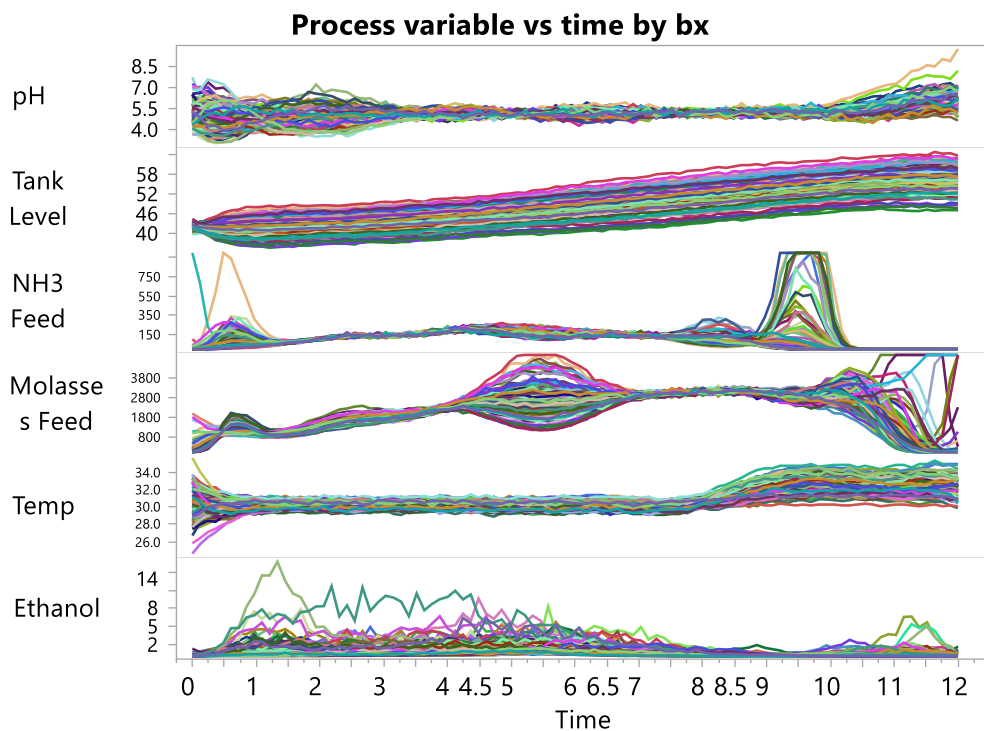
Case Study Exercise #2

Using sensor data in a model to predict quality in a manufacturing process

You oversee a fermentation process for the manufacture of an enzyme. Yield is variable and low. You need to get it under control and improve it.
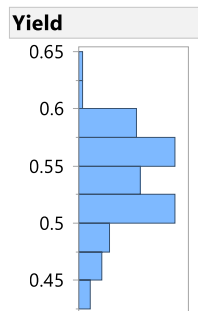
You have data on how process variables have varied throughout each batch.

Which variables do we need to control and what should their ideal profiles look like?

Fit Functional Models

1. Get the data: Help > Sample Data Library and Open "Fermentation Process.jmp".
2. Select Analyze > Special Modeling > Functional Data Explorer.
3. In the Stacked Data Format tab, select Ethanol through to pH and click Y, Output.
4. Select Time and click X, Input.
5. Select Batch ID and click ID, Function.
6. Click OK.
7. Click the Functional Data Explorer Group red triangle and select Data Processing > Align > Align Range 0 to 1. This aligns the input variable to be between 0 and 1 in each Functional Data Explorer report.
8. Click the Functional Data Explorer Group red triangle and select Models > B-Splines. This fits a B-spline model to each of the functional processes.

Note: There is also a validation role in the Functional Data Explorer launch dialog (not used here). You can holdback functional observations to ensure robustness of predictive models.

Save FPCs and Link Yield Results

1. Press Ctrl, click any Function Summaries red triangle, and select Customize Function Summaries.
2. In the box next to Enter number of FPCs to show, type 3.
3. Click the Deselect All Summaries box.
4. Click OK.
5. Click the Functional Data Explorer Group red triangle and select Save Summaries.
6. Select Help > Sample Data Library and open Fermentation Process Batch Yield Results.jmp.
7. In the Functional Data Explorer Model Summaries.jmp data table, right-click BatchID and deselect Link ID.
8. In the Functional Data Explorer Model Summaries.jmp data table, right-click BatchID and select Link Reference > Fermentation Process Batch Yield Results.jmp. This virtually joins the yield data table and the summaries data table.

Fit a Generalized Regression Model

Here we will use Generalized Regression to fit a model to determine how Yield is affected by the functional process variables.

1. In the Functional Data Explorer Model Summaries.jmp data table, select Analyze > Fit Model.
2. Click the triangle next to referenced by BatchID to Fermentation Process Batch Yield Results.
3. Select Yield[BatchID] and click Y.
4. Select the remaining columns, except Time and BatchID, and click Add.
5. Change the Personality to Generalized Regression.
6. Click Run.
7. Click Go.
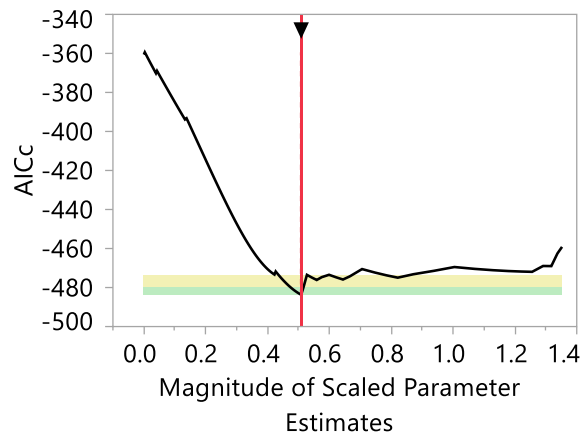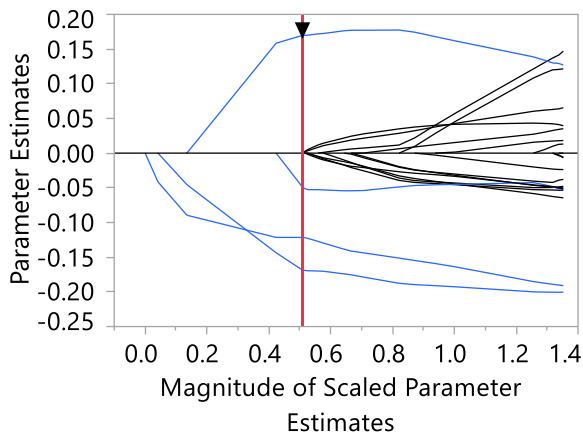
Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	0.53436	0.0020203	69956.999	<.0001 *	0.5304003	0.5383197
Ethanol FPC 1	0	0	0	1.0000	0	0
Ethanol FPC 2	-0.012168	0.0018192	44.736472	<.0001 *	-0.015733	-0.008602
Ethanol FPC 3	0	0	0	1.0000	0	0
Temp FPC 1	0	0	0	1.0000	0	0
Temp FPC 2	0	0	0	1.0000	0	0
Temp FPC 3	0	0	0	1.0000	0	0
Molasses Feed FPC 1	0	0	0	1.0000	0	0
Molasses Feed FPC 2	0.0000244	4.0897e-6	35.568946	<.0001 *	1.6375e-5	0.0000324
Molasses Feed FPC 3	0	0	0	1.0000	0	0

You have seen how you can identify the key "in-process" variables that most affect the outcome. More generally, you can use functional Xs to build predictive models of Ys.

Bonus Exercise (Advanced) – What do the best and worst process profiles look like?

1. Repeat the steps in "Save FPCs and Link Yield Results". BUT - at step 3 after Deselect All Summaries, click the Save Formulas box.
2. After you have fit the generalized regression model, click the red triangle next to Adaptive Lasso with AICc Validation and Save Columns > Save Prediction Formula.
3. Now Graph > Profiler.
4. Select the Yield[BatchID] Prediction Formula column and cast into the Y, Prediction Formula role. Also add the "Prediction Formula" columns for Ethanol, Molasses Feed, Air and NH3 Feed. The dialog should look like this:
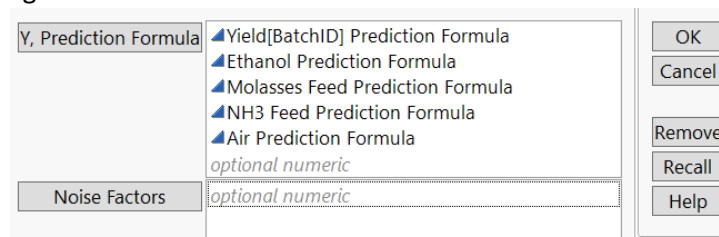

5. Make sure to click the box for Expand Intermediate Formulas. Click OK.

6. You might want to rearrange the X-axes: click the red triangle next to Prediction Profiler > Appearance > Reorder X Variables. Drag Time to the top of the list. Click OK.
7. You might also want to increase the size of the profiler plots by dragging out the corner of one.
8. Look at Ethanol FPC 2 and how changing the setting affects Yield.
9. Look at how the Ethanol vs Time profile changes as you change Ethanol FPC 2.
10. Now you can see what ethanol profiles resulting in a high (left) and low (right) yield look like…
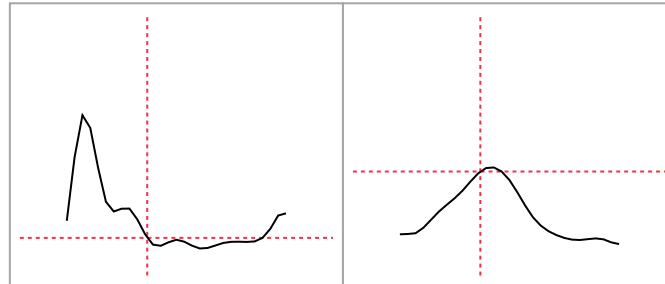
Looking at the raw data in Graph Builder, this fits with what we have seen with the functional data analysis. (Ask if you want to know how to create this plot)
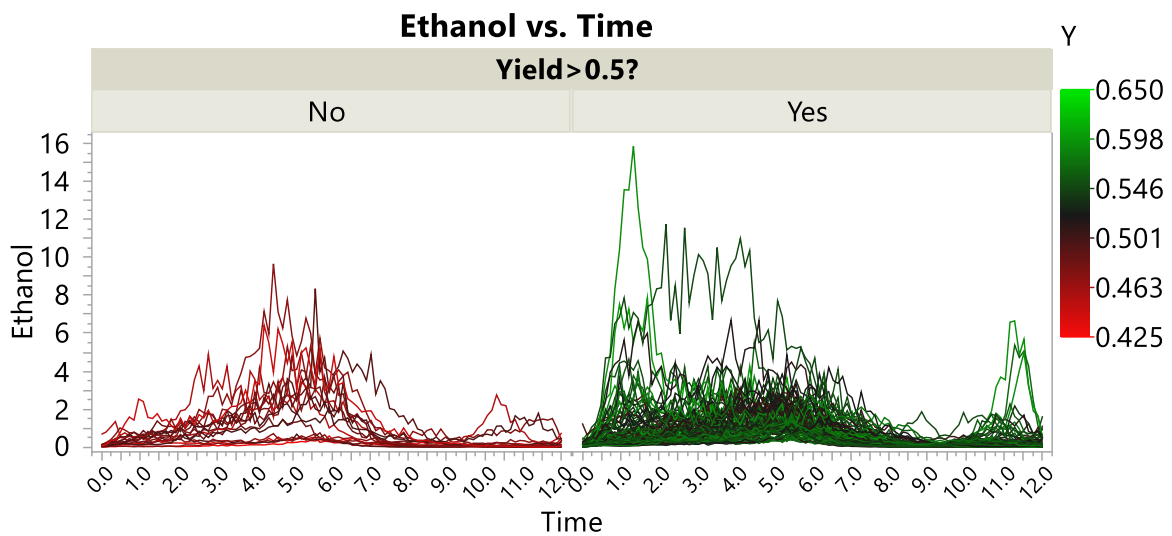
Look at the other significant Xs in the model of Yield and their corresponding time profiles.