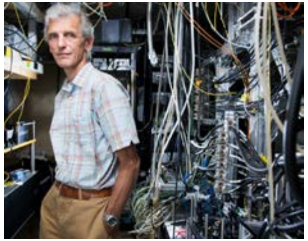
USING FUNCTIONAL DATA ANALYSIS FOR TIME-DEPENDENT OPTIMIZATION OF BATCH PROCESSES

Hadley Myers, JMP

Non Clinical Statistics Conference
October 3-4-5 2018
Maison de la Chimie, Paris

USING FUNCTIONAL DATA ANALYSIS FOR TIME-DEPENDENT OPTIMIZATION OF BATCH PROCESSES

- Wolfgang Ketterle, Nobel Lecture (2001)
  - “Imagine how many aspects of nature we would miss if we lived on the surface of the sun…without refrigerators.”



- If we can create conditions that haven’t been created before, then we’ll make new discoveries.
- If we look at things in a way that hasn’t been done before, then we’ll see new things.
- If we analyze data using new methods, then we’ll gain new insights.

FUNCTIONAL DATA ANALYSIS

- Very often, data will be “telemetric” in nature - many repeated measures of several metrics through time.
  - This is true of data from many sources.
    - Machines output
    - Traditional time series data
    - Sensor data
    - Vibration signals
- A wide variety of specific tools and methods have been created to deal with this type of data:
  - Signal Processing
  - ARIMA Time Series
  - Partial Least Squares
  - Growth Curves via SEM
  - Mixed Models

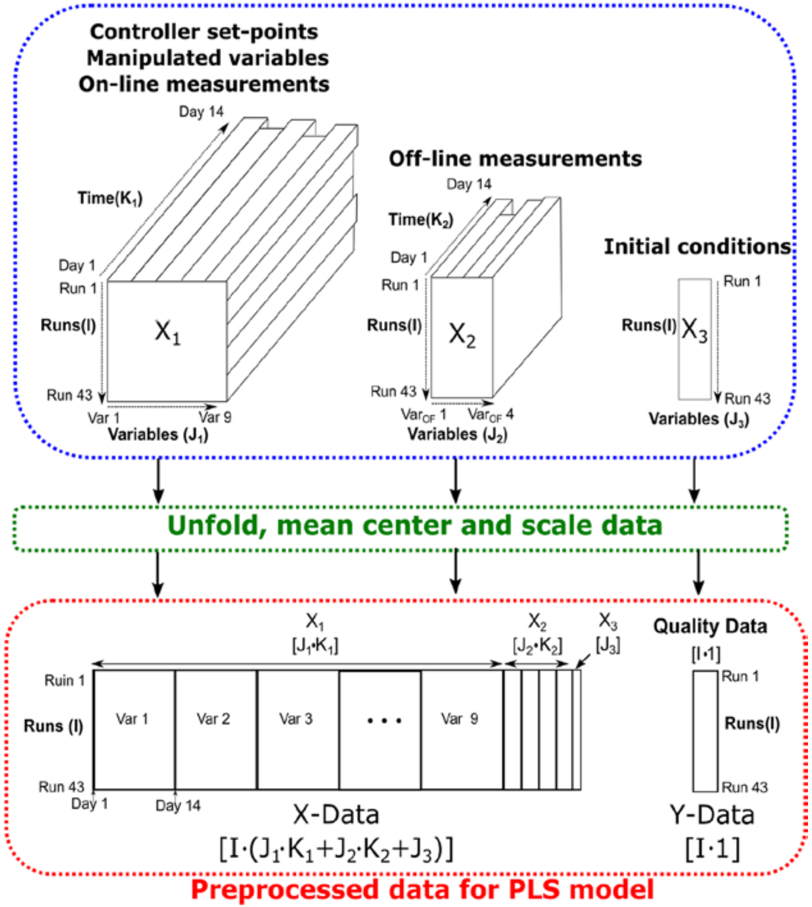
FUNCTIONAL DATA ANALYSIS

- Many products are made in batches by machines that now have many sensors embedded in them
- Sensors record things like temperature, pressure, feed rate, chemical content (ammonia, CO2, ethanol, sugar), vibration, etc.
- Companies care about end results:
  - Yield: the quantity of product created (yield)
  - Quality: Measurable properties of the product (flavor, room temperature viscosity, shear strength, chemical composition)
- They want to understand how the sensor readings relate to the end results
  - To fix ‘bad batches’, or terminate their production early
  - Reduce occurrence of bad batches (process improvement)
- This is not a new problem - Due to the explosion of data access a lot more people want to take advantage of functional data

FUNCTIONAL DATA ANALYSIS

- Traditional approaches are too often inadequate and overcomplicated
  - Converting data to wide format (one input variable per time period) and using PLS.
    - Data cleaning step can be very time-consuming.
    - Sparse table if time-points not aligned, lost data if sizes not equal.
    - Difficult to interpret results for optimization (may be possible for early flagging of batches)
  - Least Squares modeling of summary statistics (mean, min, max, etc.).
    - Too simple, all time-dependent information is lost.
    - Model says nothing about the shape of the functions.
  - Fitting logistic curves, using parameter estimates as features.
    - Very limited in the set of shapes of curves that can be fit.
    - More flexibility needed than simple logistic curves and Gaussian peak models (again, too simple).

DEALING WITH TELEMETRIC DATA
(TRADITIONAL APPROACH)



Goldrick et al.: MVDA of Trisulfide Bond Formation, Biotechnology and Bioengineering, Vol. 114, No. 10, October, 2017

USES OF FUNCTIONAL DATA ANALYSIS

1. Functional factors, constant responses.
2. Constant factors, functional responses.
3. Both functional factors and functional responses.

Example: Maximizing the yield of human insulin produced by modified yeast cells.

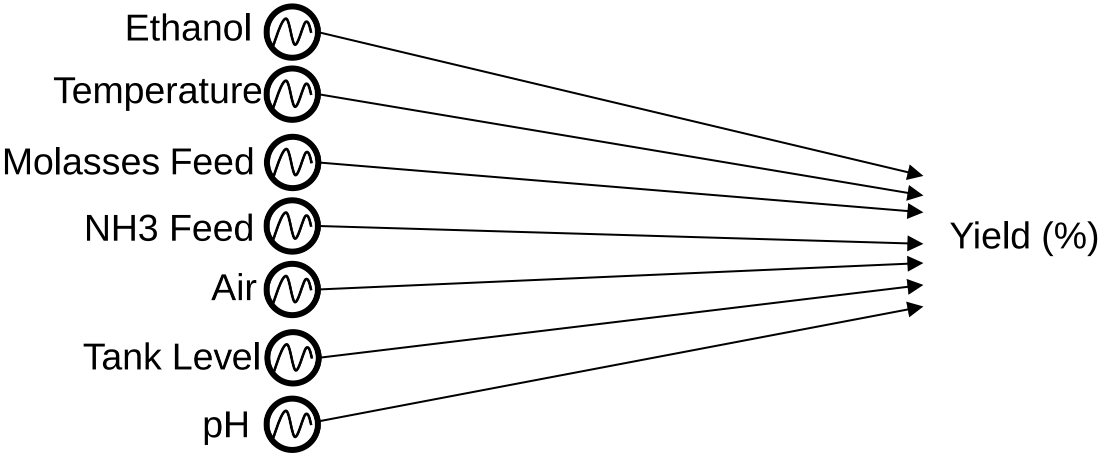


TABLE OF RAW DATA

Fermentation Process Sensor Data - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Six Sigma Tools Tools Add-Ins View Window Help

Fermentat...
Source

Columns (10...
BatchID
Time
Ethanol
Temp
Molas...s Feed
NH3 Feed
Air
Tank Level
pH
Validation

Rows
All rows 10.000
Selected 0
Excluded 0
Hidden 0
Labelled 0

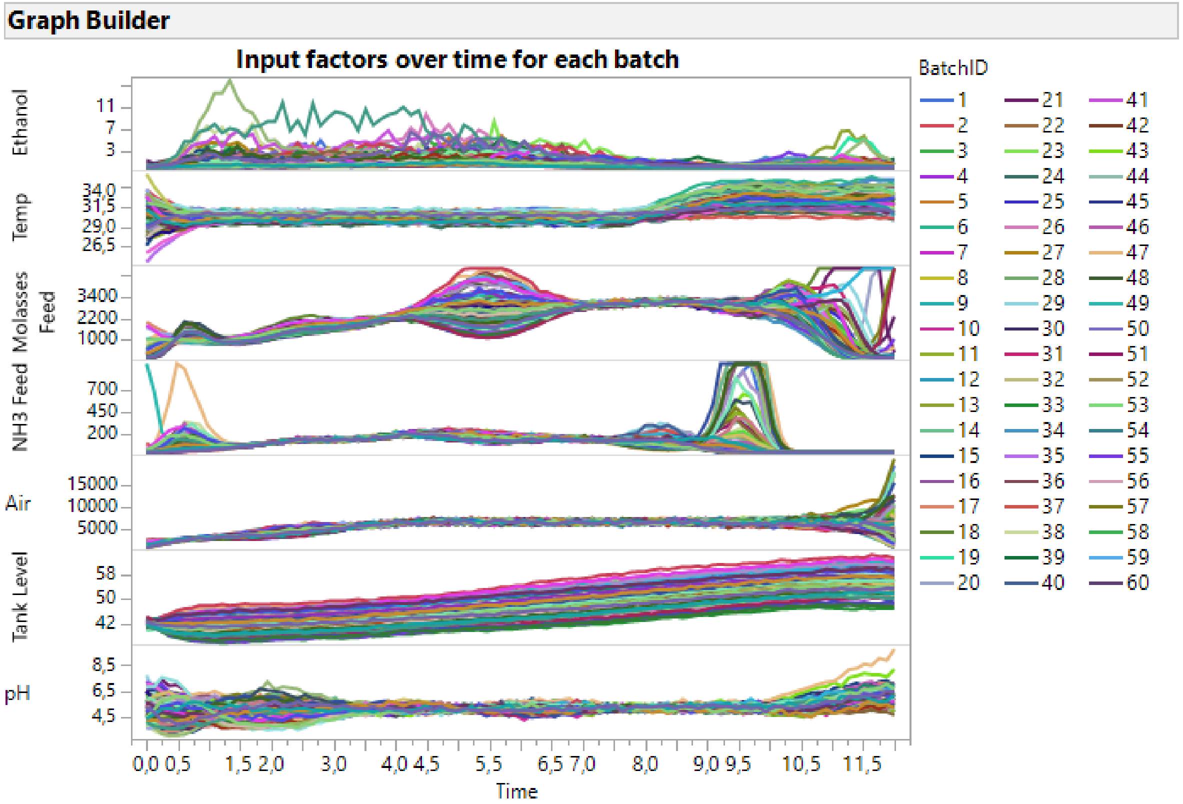
9/0 Cols

	BatchID	Time	Ethanol	Temp	Molasses Feed	NH3 Feed	Air	Tank Level	pH
1	1	0	0,678	30,2	36,6	0,7	1724	43,02	6,13
2	1	0,121	0,797	30,3	129,6	4	1955	42,21	6,22
3	1	0,242	1,06	30,5	361,7	12,1	2134	41,8	6,7
4	1	0,364	1,346	30,6	678,4	28	2247	41,3	6,24
5	1	0,485	0,609	30,4	985,1	53,4	2561	40,75	6,18
6	1	0,606	0,754	30,6	1196,4	71,4	2557	40,93	6,09
7	1	0,727	0,61	30,7	1279,2	77,7	2938	40,48	5,52
8	1	0,848	0,448	30,9	1146,3	82,7	2785	40,61	5,22
9	1	0,97	0,629	30,8	1063,2	73,9	2905	40,76	[illegible]
10	1	1,091	0,417	30,9	869,1	68,4	3110	40,58	4,93
11	1	1,212	0,433	30,9	831,1	67,6	3233	41,16	4,66
12	1	1,333	0,587	31,2	835,9	78,7	3218	40,7	4,57
13	1	1,455	0,82	30,9	826,7	78,9	[illegible]	40,56	4,26
14	1	1,576	0,833	30,9	831,1	70,3	3303	40,94	[illegible]
15	1	1,697	0,996	31,1	849,9	[illegible]	3665	41,01	3,86
16	1	1,818	1,433	31,1	900	98,4	3648	41,15	4,06
17	1	1,939	1,456	31,2	[illegible]	106,1	[illegible]	41,03	3,83
18	1	2,061	1,909	30,9	1128,2	110,6	3857	41,07	3,81
19	1	2,182	1,97	31,2	1181,2	122,2	3940	41,54	4,03
20	1	2,303	2,269	31,1	[illegible]	136,7	4076	41,4	3,91
21	1	2,424	3,546	30,9	1334,1	138,7	4129	41,52	4,12
22	1	2,545	[illegible]	31,2	1378,8	120	4324	41,47	4,17
23	1	2,667	2,963	31	1451,8	158,8	4632	41,59	4,25
24	1	[illegible]	[illegible]	30,9	1448	132,9	4328	41,49	4,36
25	1	[illegible]	3,096	31,1	1544,3	138	5244	41,68	4,7
26	1	3,03	3,515	31	1596,7	144	5042	41,66	4,71
27	1	3,152	3,077	30,9	1666,6	161,3	5358	42,23	4,77
28	1	3,273	4,256	31,1	1714,1	138,1	5378	41,94	4,79
29	1	3,394	2,521	31	1693,4	151,3	6022	42,05	4,81
30	1	3,515	2,992	31,2	1818,4	141,3	6324	42,28	5,05
31	1	3,636	3,24	30,9	1891,3	167	6018	42,34	5,31
32	1	3,758	1,996	31,1	1971,3	162,8	6001	42,31	5,12
33	1	3,879	4,001	31,3	2084,1	176,7	6099	42,84	5,29

Rows
All rows 10.000

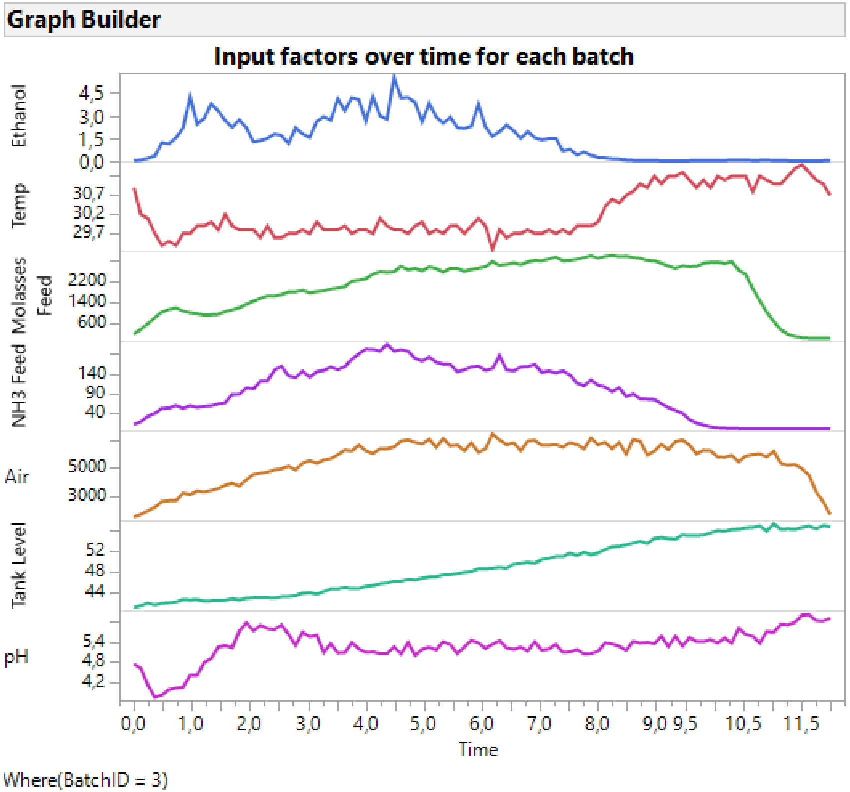
- There are 100 batches in total
- 100 time measurements per batch
- Measurements were taken at fixed time intervals
- This isn't always the case!

INVESTIGATING INPUT FACTORS



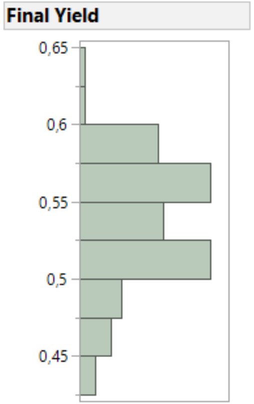
Plotting the factors over time for each batch reveals the complexity of the problem.

INVESTIGATING INPUT FACTORS



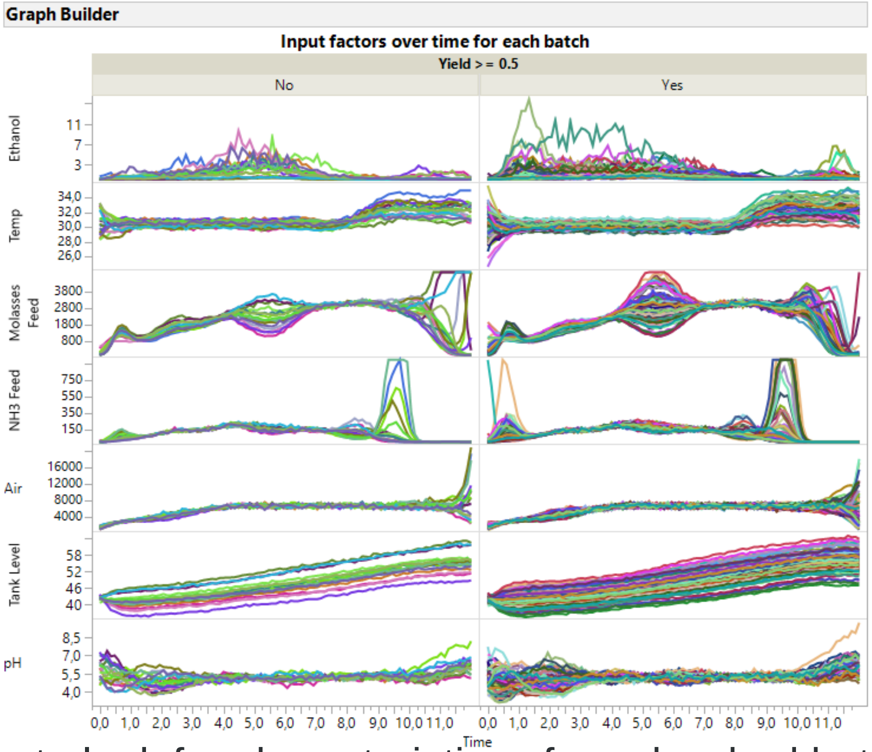
Exploring batches individually further emphasizes the challenge ahead.

GOOD BATCHES VS BAD BATCHES



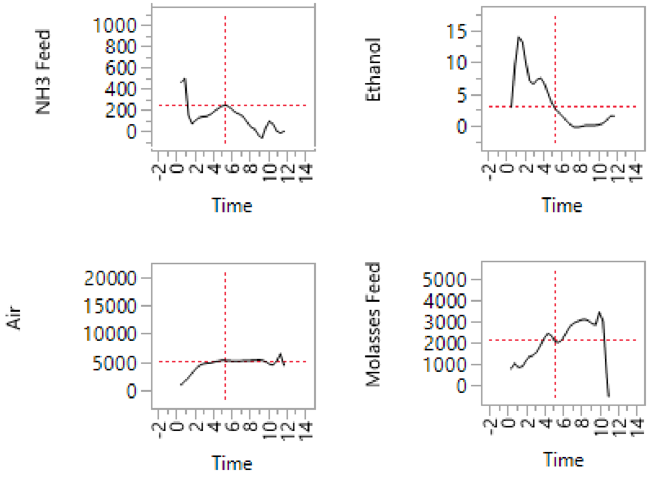
Summary Statistics

Mean	0,53436
Std Dev	0,0392623
Std Err Mean	0,0039262
Upper 95% Mean	0,5421505
Lower 95% Mean	0,5265695
N	100



What we'll find:

1) We can achieve a 74% Yield.
2) Ethanol, Molasses, NH3 and Air are significant.
3) These are their ideal profiles:



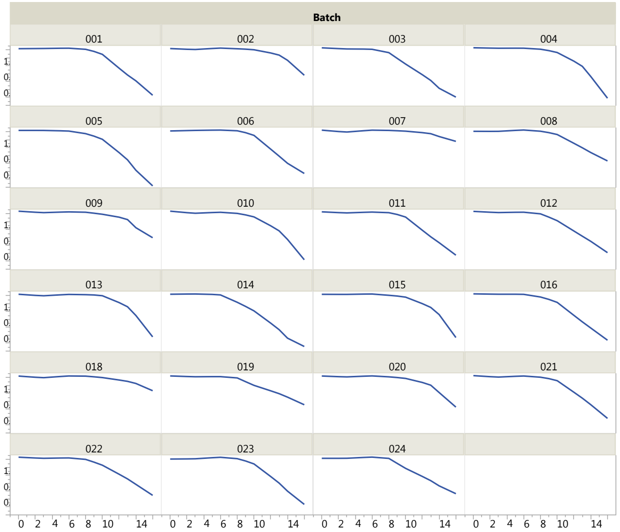
We could try to look for characteristics of good vs bad batches, but what defines "good?"

Can "good" be "better?"

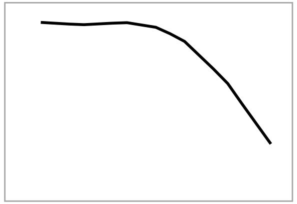
DEALING WITH TELEMETRIC DATA (USING FDA)

Functional Data Analysis

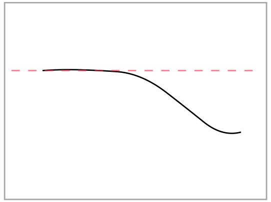
X3 vs. Time



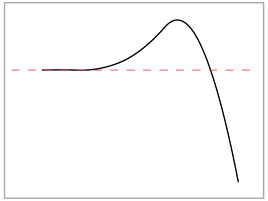
= Mean



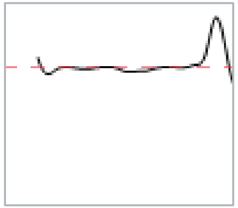
+ *FPC1* Eigenfunction 1



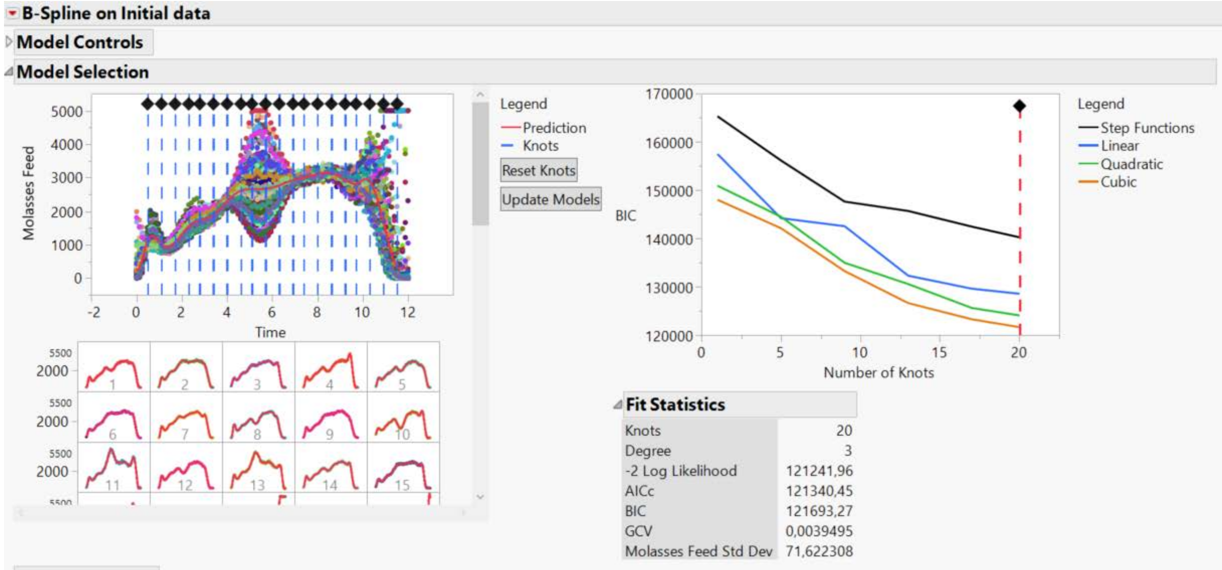
+ *FPC2* Eigenfunction 2



+ … + *FPCn* Eigenfunction n



FUNCTIONAL DATA ANALYSIS

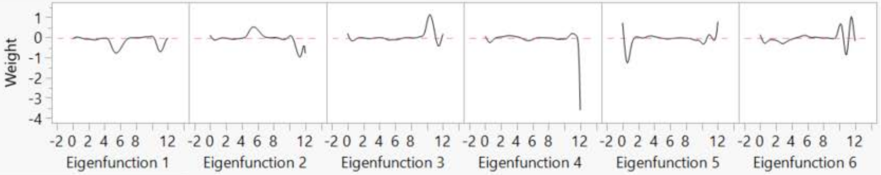


Function Summaries

BatchID	FPC 1	FPC 2	FPC 3	FPC 4	FPC 5	FPC 6
1	580,57767	-267,7805	28,573012	147,18336	55,508213	-40,88853
2	-286,5936	543,56012	-411,2534	-18,71583	28,023103	-16,11789
3	310,61672	440,80251	-75,12021	34,650718	68,248024	95,491374
4	610,62427	-295,9634	944,9418	178,77983	-154,218	18,311561
5	613,31246	-357,7327	-124,0855	201,97364	147,03689	-145,2244
6	-23,01349	459,66476	-167,9752	7,8868092	117,11506	110,54716
7	453,57879	214,81655	-161,5337	37,1434	-84,7756	-1,366664
8	1330,6241	-42,51764	-584,4787	26,178107	-153,5792	11,219651
9	259,84521	471,10044	-617,4147	-38,18937	69,281452	92,57684
10	1588,9261	-453,3788	395,10516	255,93291	135,0555	-1,14896
11	-1829,969	986,67051	558,25093	-80,39355	-187,2943	-67,82698

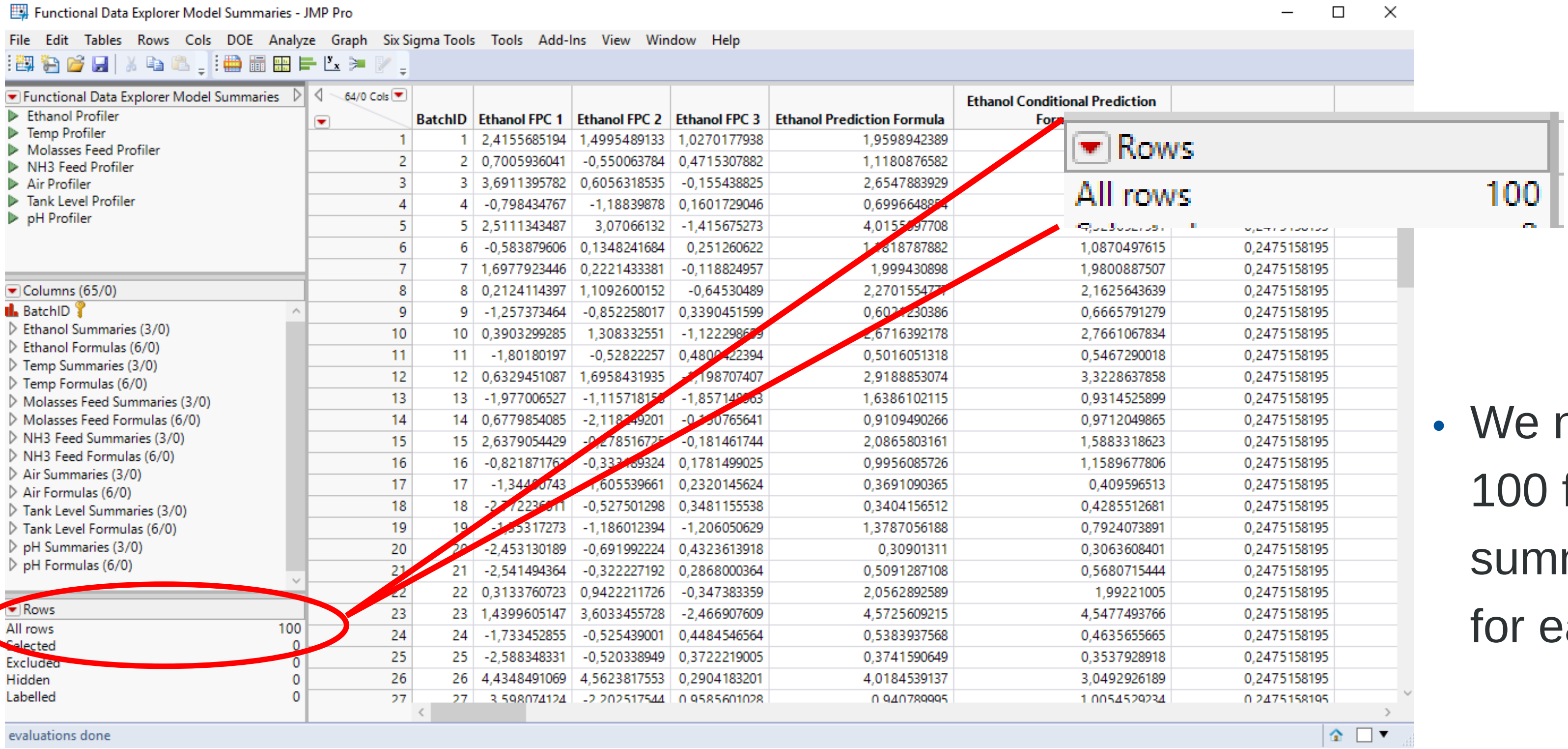
Functional PCA

FPC	Eigenvalue	20 40 60 80	Percent	Cumulative
1	1088713,1		56,1%	56,1%
2	477300,89		24,6%	80,7%
3	189338,88		9,75%	90,4%
4	89394,683		4,6%	95%
5	35637,809		1,84%	96,9%
6	28947,867		1,49%	98,3%



Functional Data Analysis modeling for determining Functional Principal Components to be used in prediction.

SAVING FPC AND EIGENFUNCTIONS FOR EACH BATCH



	BatchID	Ethanol FPC 1	Ethanol FPC 2	Ethanol FPC 3	Ethanol Prediction Formula	Ethanol Conditional Prediction For[illegible]	
1	1	2,4155685194	1,4995489133	1,0270177938	1,9598942389		
2	2	0,7005936041	-0,550063784	0,4715307882	1,1180876582		
3	3	3,6911395782	0,6056318535	-0,155438825	2,6547883929		
4	4	-0,798434767	-1,18839878	0,1601729046	0,69966488[illegible]		
5	5	2,5111343487	3,07066132	-1,415675273	4,015[illegible]97708		
6	6	-0,583879606	0,1348241684	0,251260622	[illegible]818787882	1,0870497615	0,2475158195
7	7	1,6977923446	0,2221433381	-0,118824957	1,999430898	1,9800887507	0,2475158195
8	8	0,2124114397	1,1092600152	-0,64530489	2,270155477[illegible]	2,1625643639	0,2475158195
9	9	-1,257373464	-0,852258017	0,3390451599	0,60[illegible]230386	0,6665791279	0,2475158195
10	10	0,3903299285	1,308332551	-1,122298[illegible]	[illegible],6716392178	2,7661067834	0,2475158195
11	11	-1,80180197	-0,52822257	0,480[illegible]422394	0,5016051318	0,5467290018	0,2475158195
12	12	0,6329451087	1,6958431935	[illegible],198707407	2,9188853074	3,3228637858	0,2475158195
13	13	-1,977006527	-1,115718[illegible]	-1,857149[illegible]63	1,6386102115	0,9314525899	0,2475158195
14	14	0,6779854085	-2,118[illegible]49201	-0,[illegible]30765641	0,9109490266	0,9712049865	0,2475158195
15	15	2,6379054429	-0,[illegible]78516725	-0,181461744	2,0865803161	1,5883318623	0,2475158195
16	16	-0,82187176[illegible]	-0,33[illegible]89324	0,1781499025	0,9956085726	1,1589677806	0,2475158195
17	17	-1,344[illegible]0743	[illegible]1,605539661	0,2320145624	0,3691090365	0,409596513	0,2475158195
18	18	-2,[illegible]7223[illegible]11	-0,527501298	0,3481155538	0,3404156512	0,4285512681	0,2475158195
19	19	-1,[illegible]55317273	-1,186012394	-1,206050629	1,3787056188	0,7924073891	0,2475158195
20	20	-2,453130189	-0,691992224	0,4323613918	0,30901311	0,3063608401	0,2475158195
21	21	-2,541494364	-0,322227192	0,2868000364	0,5091287108	0,5680715444	0,2475158195
22	22	0,3133760723	0,9422211726	-0,347383359	2,0562892589	1,99221005	0,2475158195
23	23	1,4399605147	3,6033455728	-2,466907609	4,5725609215	4,5477493766	0,2475158195
24	24	-1,733452855	-0,525439001	0,4484546564	0,5383937568	0,4635655665	0,2475158195
25	25	-2,588348331	-0,520338949	0,3722219005	0,3741590649	0,3537928918	0,2475158195
26	26	4,4348491069	4,5623817553	0,2904183201	4,0184539137	3,0492926189	0,2475158195
27	27	3,598074124	-2,202517544	0,9585601028	0,940789995	1,0054529234	0,2475158195

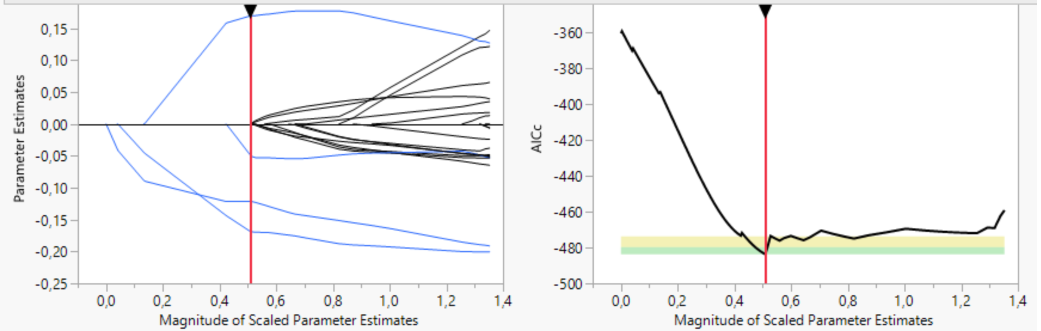
- We now have 100 functional summaries (one for each batch)

MODELING WITH FPCS AS INPUTS

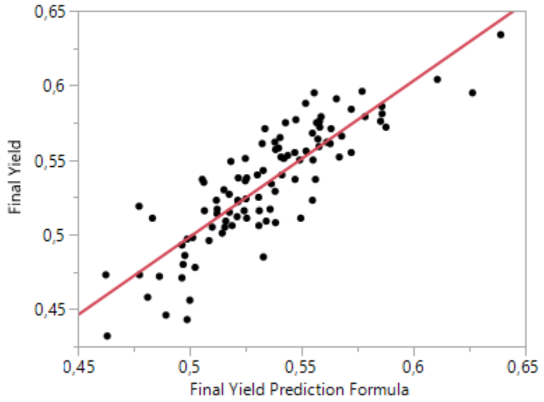
Generalized Regression for Final Yield

Adaptive Lasso with AICc Validation

Solution Path



Actual by Predicted plot



RSquare	0,733979
RSquare Adj	0,731265
Root Mean Square Error	0,020353
Mean of Response	0,53436
Observations (or Sum Wgts)	100

Parameter Estimates for Original Predictors

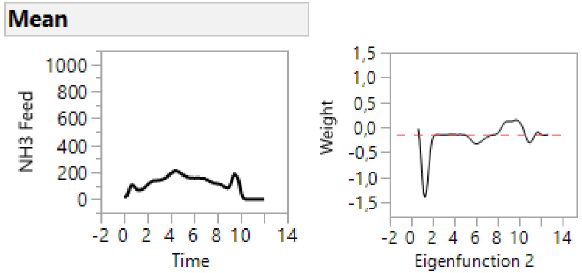
Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	0,53436	0,0020203	69956,999	<,0001*	0,5304003	0,5383197
NH3 Feed FPC 2	-0,000173	2,4377e-5	50,628148	<,0001*	-0,000221	-0,000126
Ethanol FPC 2	-0,012168	0,0018192	44,736472	<,0001*	-0,015733	-0,008602
Molasses Feed FPC 2	0,0000244	4,0897e-6	35,568946	<,0001*	1,6375e-5	0,0000324
Air FPC 1	-2,63e-6	1,1457e-6	5,2684672	0,0217*	-4,875e-6	-3,842e-7
Ethanol FPC 1	0	0	0	1,0000	0	0
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 3	0	0	0	1,0000	0	0
NH3 Feed FPC 1	0	0	0	1,0000	0	0
NH3 Feed FPC 3	0	0	0	1,0000	0	0

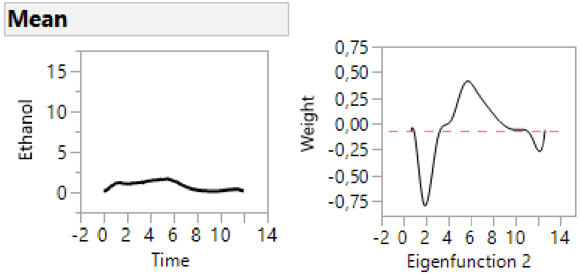
- Significant factors

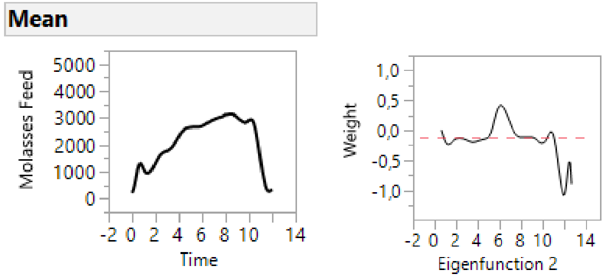
Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare
Intercept	0,53436	0,0020203	69956,999	<,0001*
NH3 Feed FPC 2	-0,000173	2,4377e-5	50,628148	<,0001*
Ethanol FPC 2	-0,012168	0,0018192	44,736472	<,0001*
Molasses Feed FPC 2	0,0000244	4,0897e-6	35,568946	<,0001*
Air FPC 1	-2,63e-6	1,1457e-6	5,2684672	0,0217*

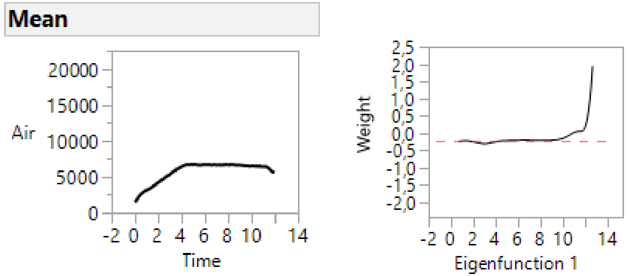
INTERPRETATION OF RESULTS

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare
Intercept	0,53436	0,0020203	69956,999	<,0001*
NH3 Feed FPC 2	-0,000173	2,4377e-5	50,628148	<,0001*
Ethanol FPC 2	-0,012168	0,0018192	44,736472	<,0001*
Molasses Feed FPC 2	0,0000244	4,0897e-6	35,568946	<,0001*
Air FPC 1	-2,63e-6	1,1457e-6	5,2684672	0,0217*





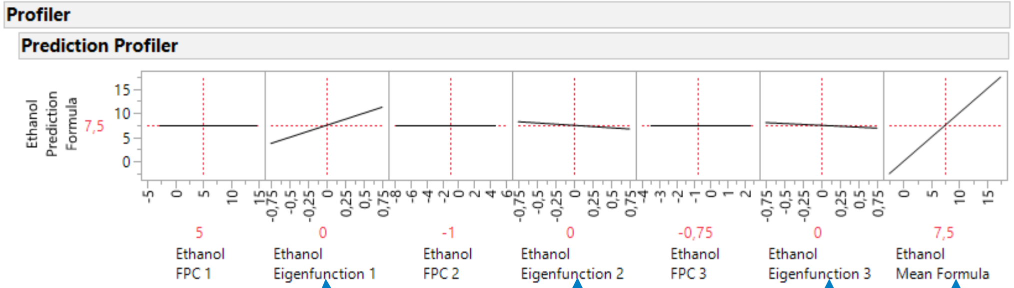




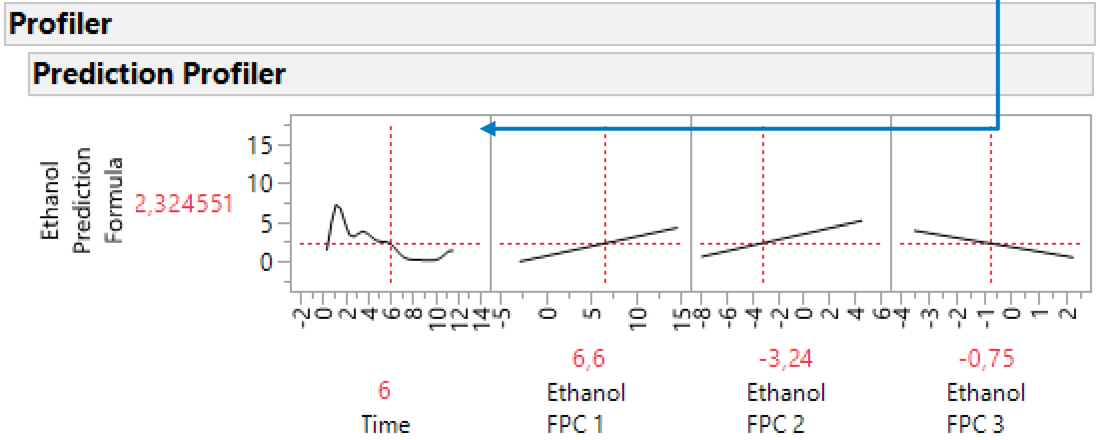
INTERPRETATION OF ONE FACTOR (ETHANOL)

Predictive formula for Ethanol, from FDA B-spline fitting

Ethanol FPC 1 • Ethanol Eigenfunction 1
+ Ethanol FPC 2 • Ethanol Eigenfunction 2
+ Ethanol FPC 3 • Ethanol Eigenfunction 3
+ Ethanol Mean Formula



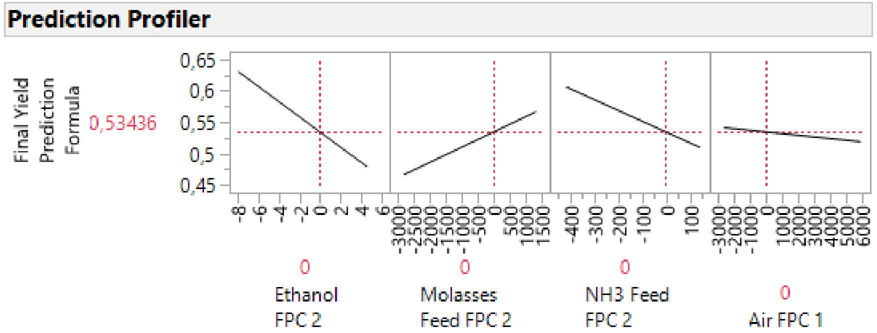
These are functions of time, which can be consolidated.

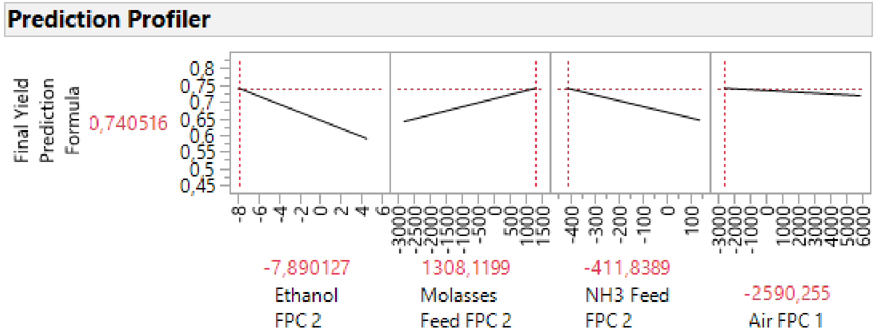


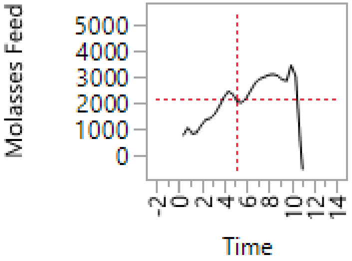
The Prediction Profiler is used to visualize and explore predictive models.

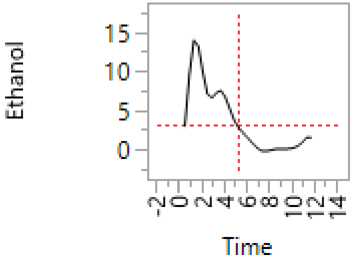
OPTIMAL YIELD RECIPE

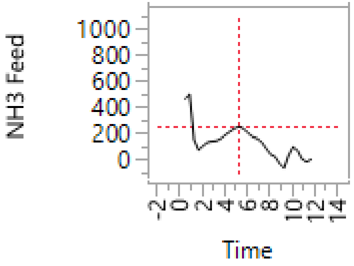
Optimizing the values of the FPCs to maximize Yield, we have everything needed to find the optimal profiles.

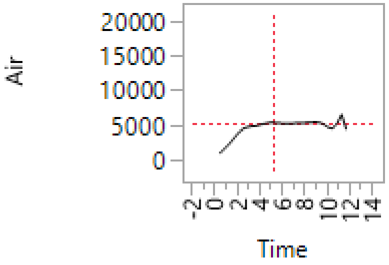












These profiles are predicted to result in a Yield of 74%.

USING FUNCTIONAL DATA ANALYSIS FOR TIME-DEPENDENT OPTIMIZATION OF BATCH PROCESSES

- Using Functional Data Analysis over traditional methods, we can:
  - Dramatically reduce total time to meaningful results,
  - Utilize all data while preserving time-dependent info,
  - Generate more interpretable knowledge output,
  - Better engage with subject-matter experts.

- Question: Can you think of opportunities within your organization where you could apply this method?

Thanks for the support of my colleagues Phil Kay, Chris Gotwalt, and Malcolm Moore