In Pursuit of the "Golden Curve"

A Comparison of Functional Data and Partial Least Squares Analyses on Serial Data

Beatrice Blum & Phil Bowtell

Discovery Summit Europe, March 2021

DATA & MODELING SCIENCES
Unlocking Innovation

Introduction

Introduction to the serial / curve sensor data collected and some of the many questions posed.

Understanding K-data curves and linking them to *Yield 1 (smaller is better)*:

- Use of Functional Data Analysis (FDA) to assess this link.
- More traditional modelling via Partial Least Squares (PLS) – analysed using SIMCA
- Common links between the two methods.

Linking P-data curves to *Yield 2 (larger is better)*:

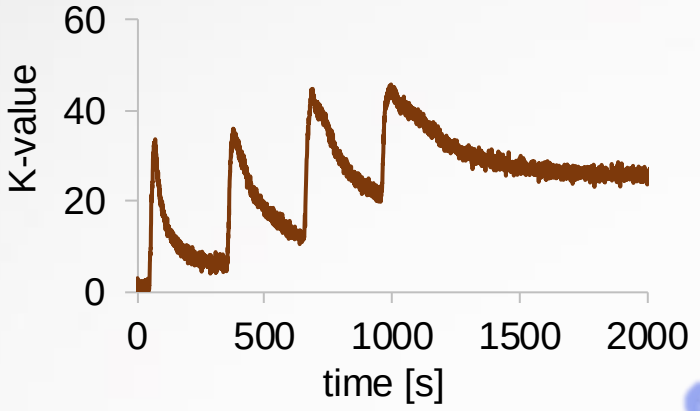
- Making use of some tools in SIMCA for exploratory data analysis and modelling.
- Use of FDA.

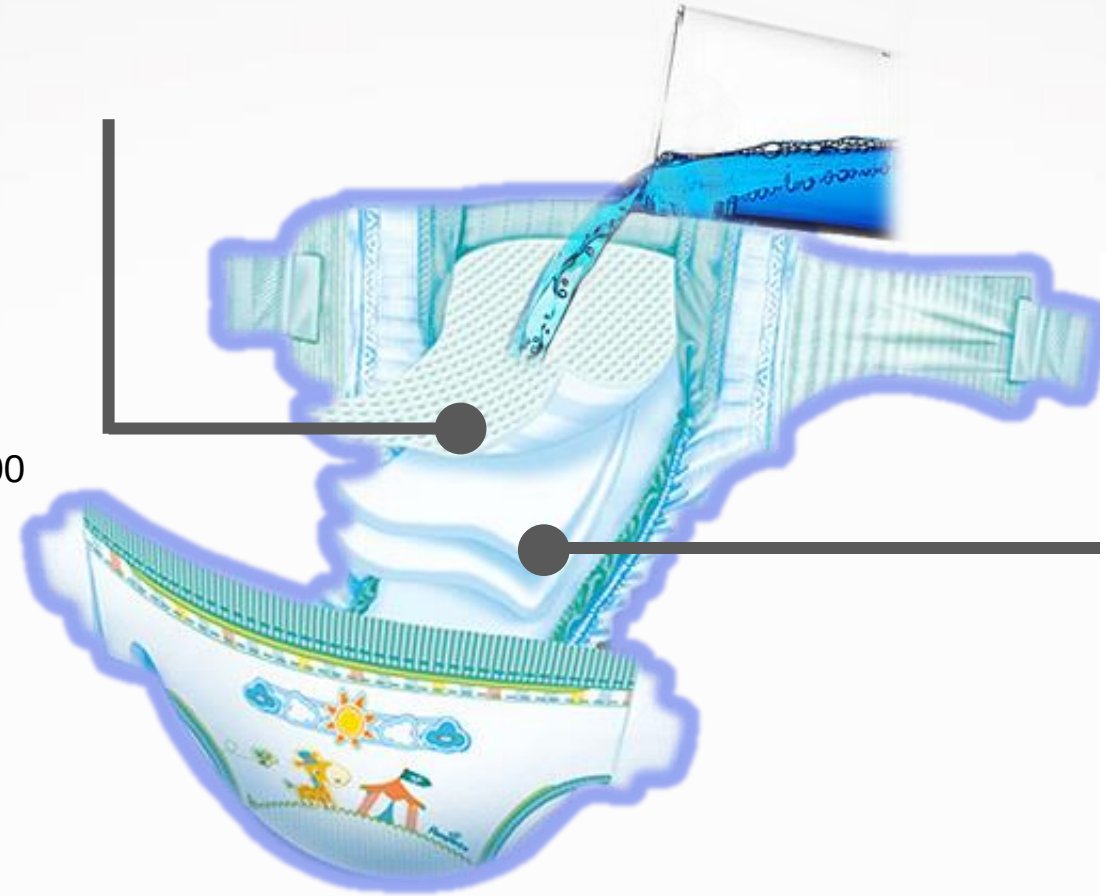
Summary and next steps.

DATA & MODELING SCIENCES
Unlocking Innovation

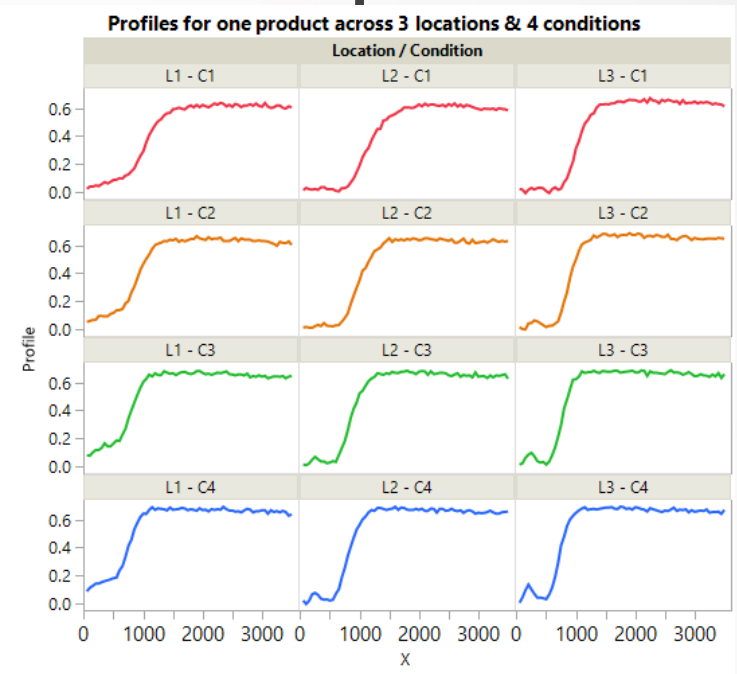
ANALYTICAL

K-data curves





P-data profiles



3 replicates for each endpoint measured

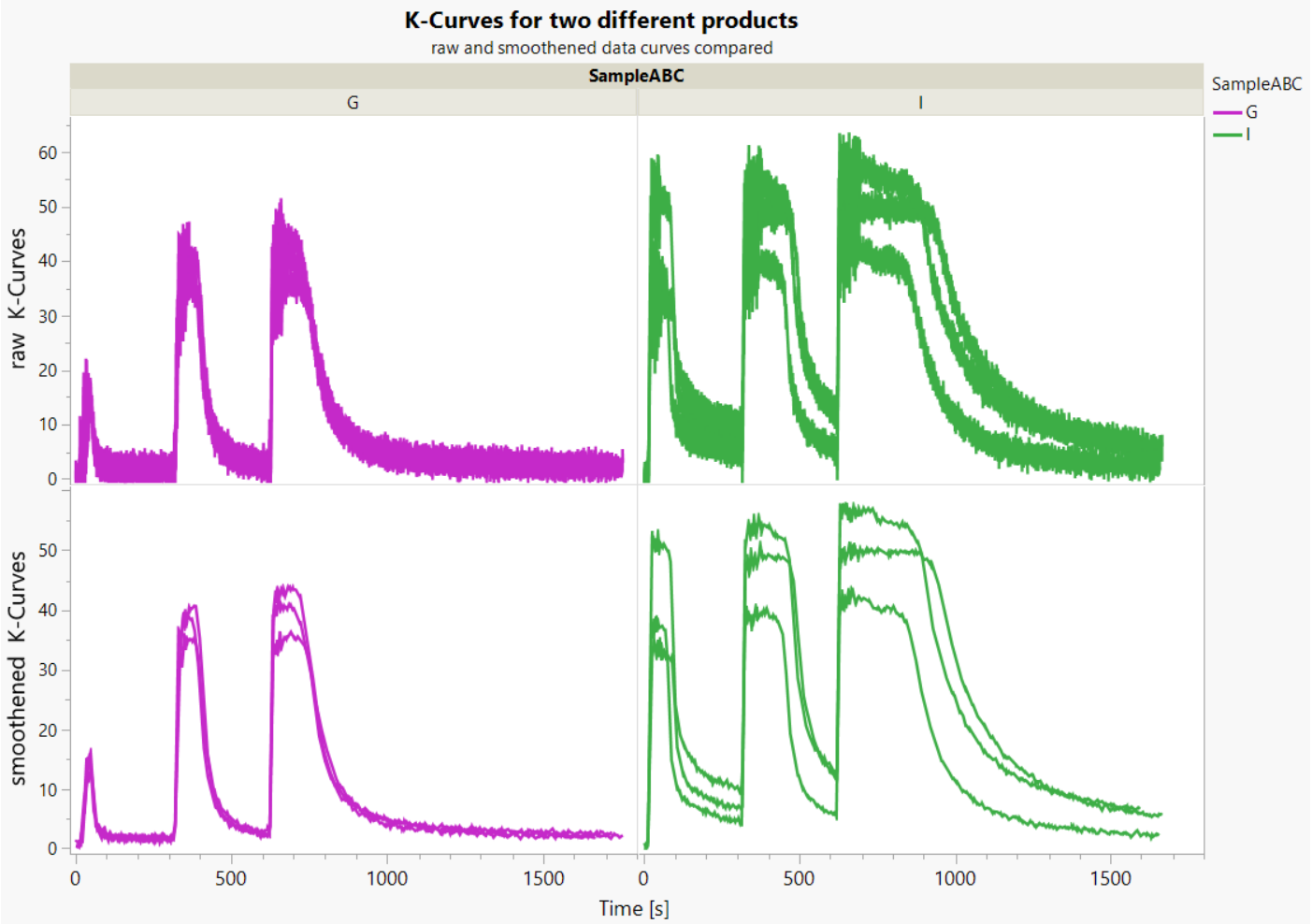
K-data

K-data quite oscillating / noisy due to the very high frequency at which they are measured

Smoothing indicated prior to fitting

Moving Average with “Local width” = 20s used

Curves / derivatives not smooth. FDA analyses smooth curves usually!



FDA Demo

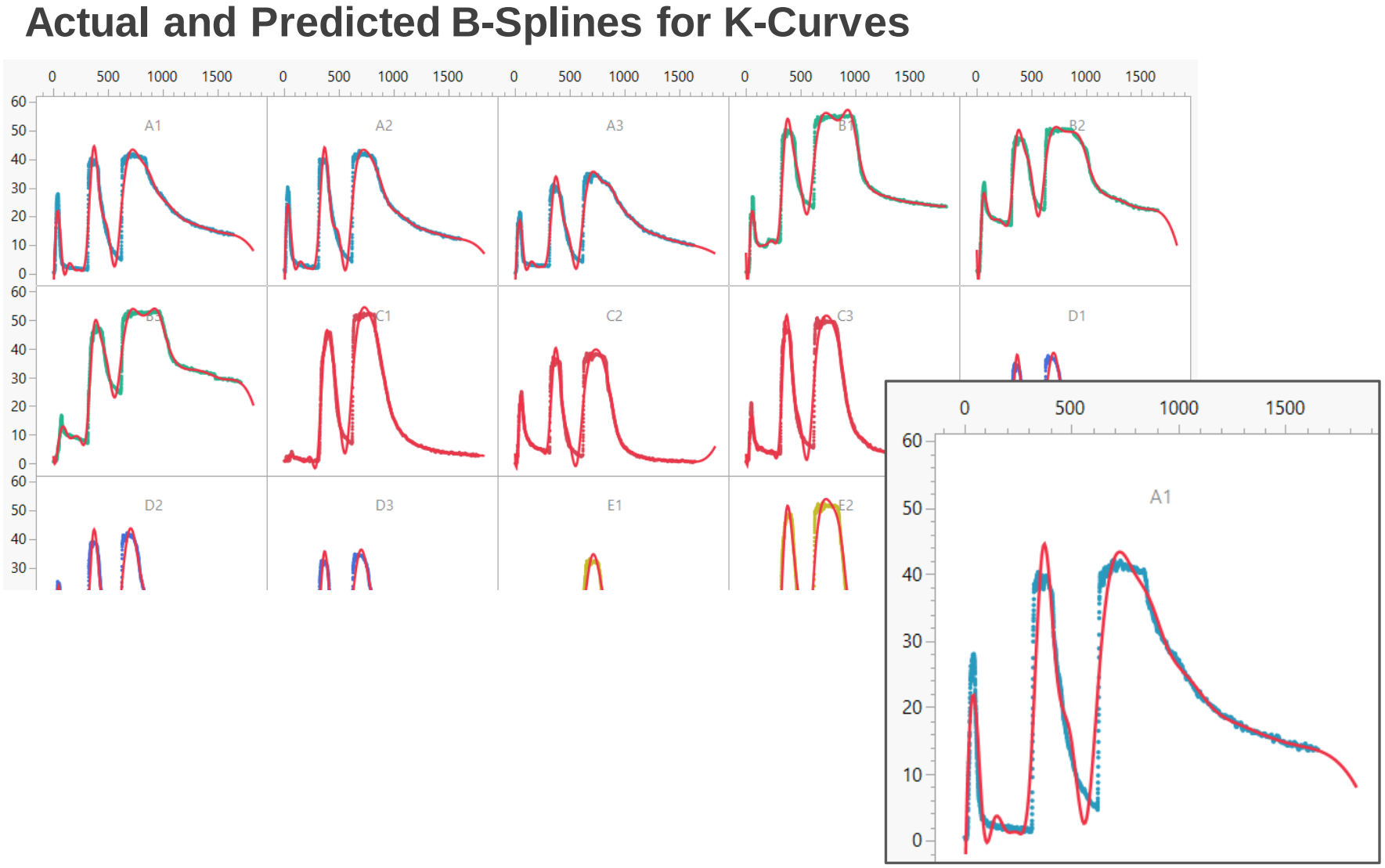
How to analyze K-data curves with JMP Functional Data Explorer (FDE)

- Note: JMP 16 Early Adopter 8 used for demo

FDA Findings

Curve fitting:

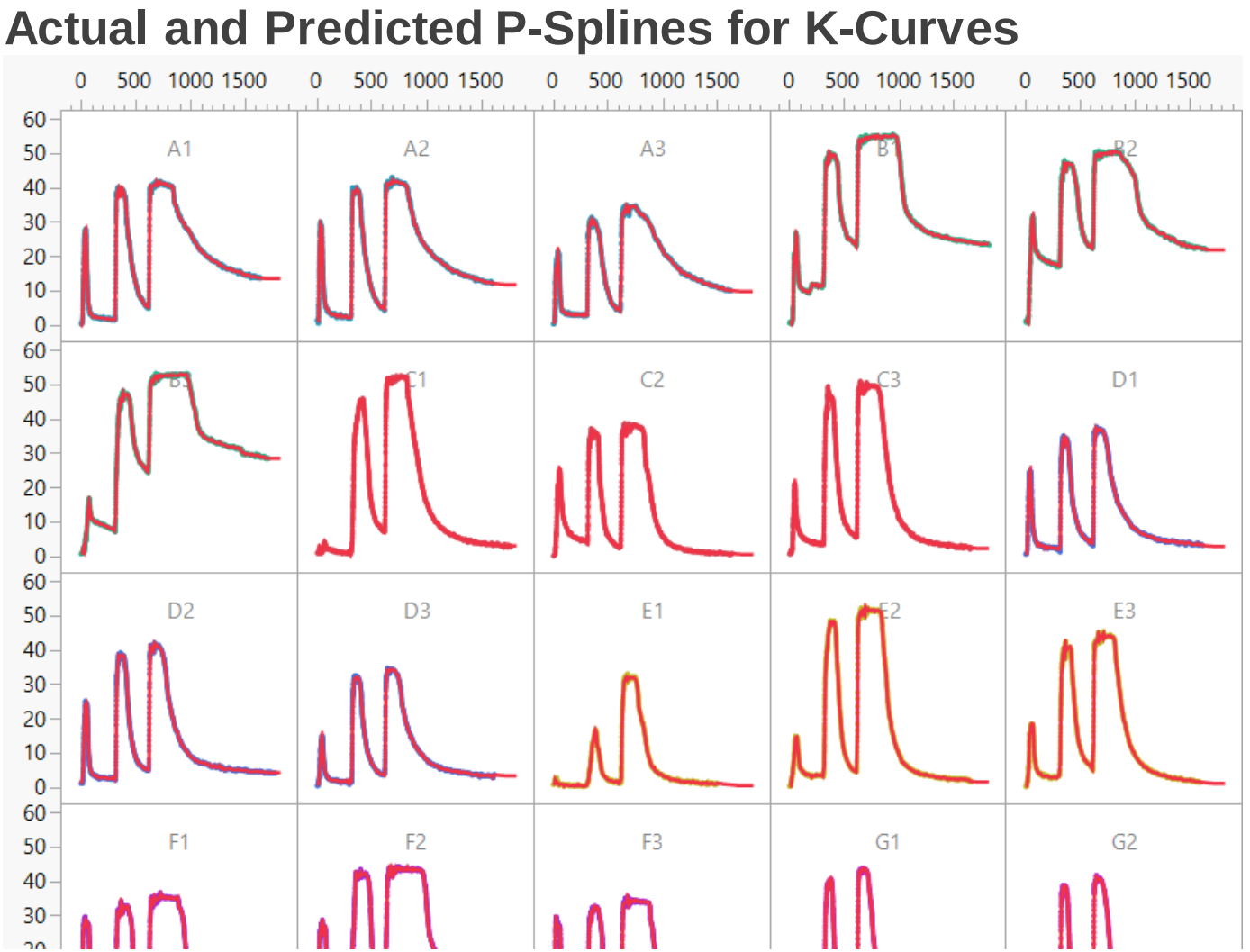
- Red line is fit, any other colour is measurement data
- B-Splines not accurately dealing with step changes in discrete measures.



FDA Findings

Curve fitting:

- P-Splines Step-function much better suited.
- P-Splines fits of smoothened K-curves look really good

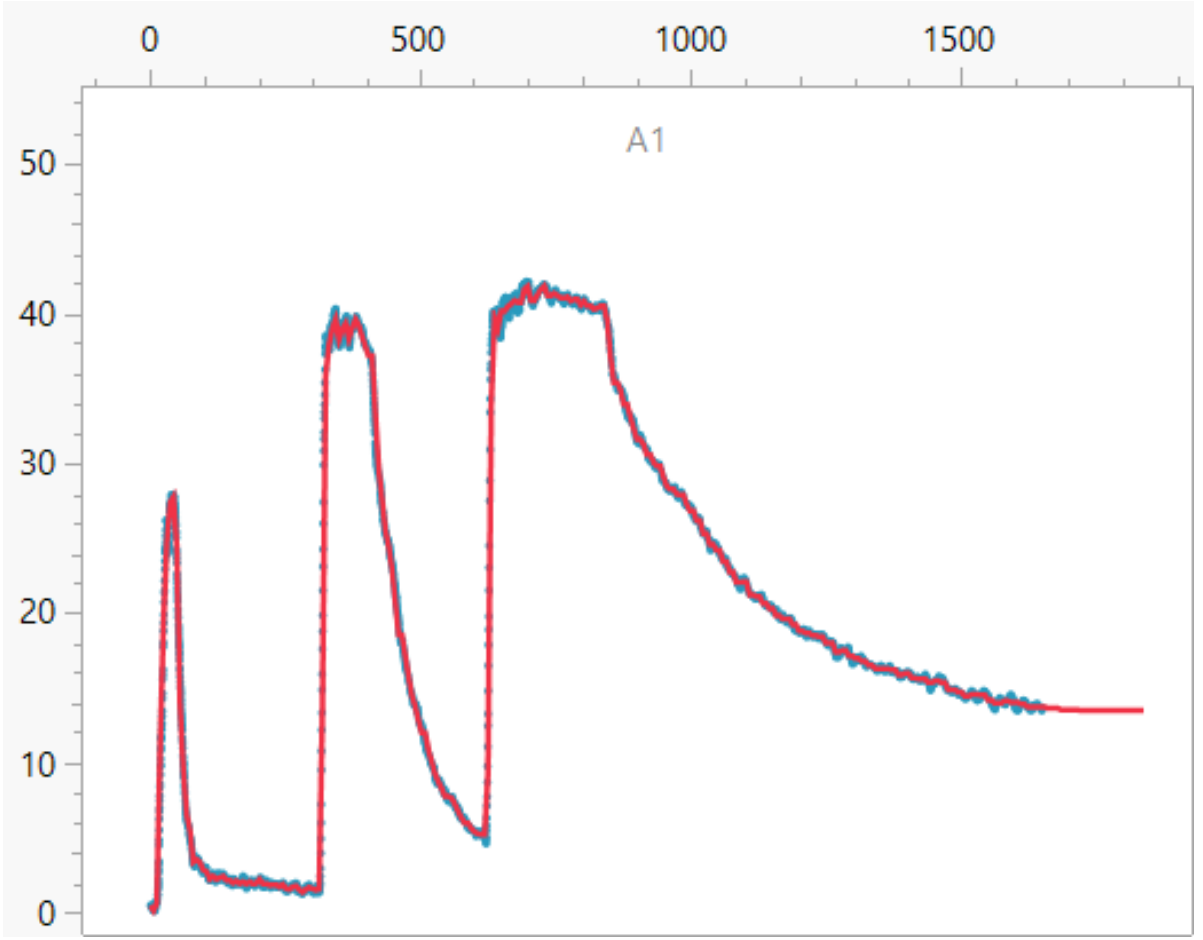


FDA Findings

Curve fitting:

- P-Splines step function still captures some measurement noise.

Actual and Predicted P-Splines for Kinetics-Curves

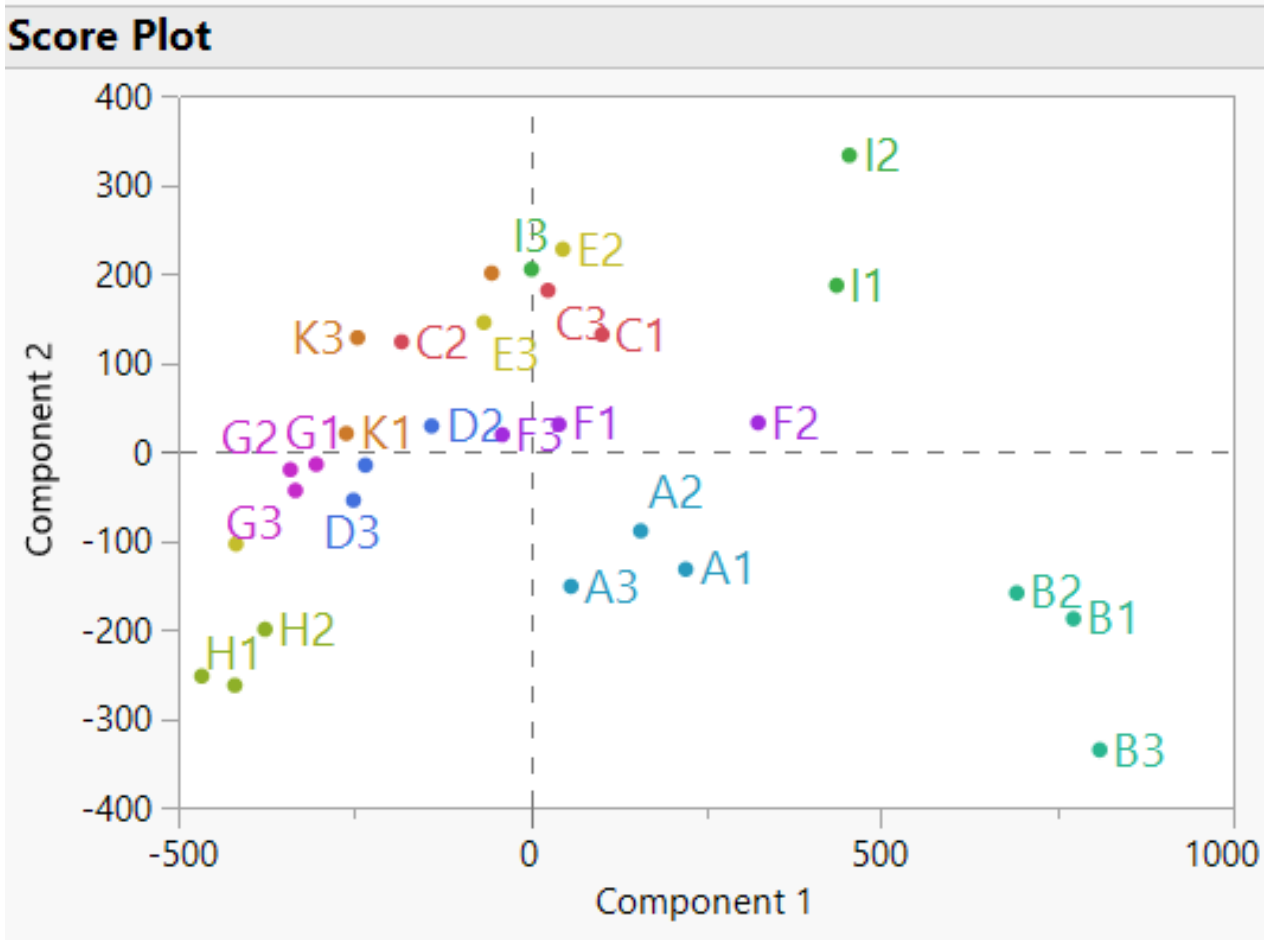


FDA Findings

Calculate Functional Principal Components (FPC) from Curve fit

Plot first 2 FPCs:

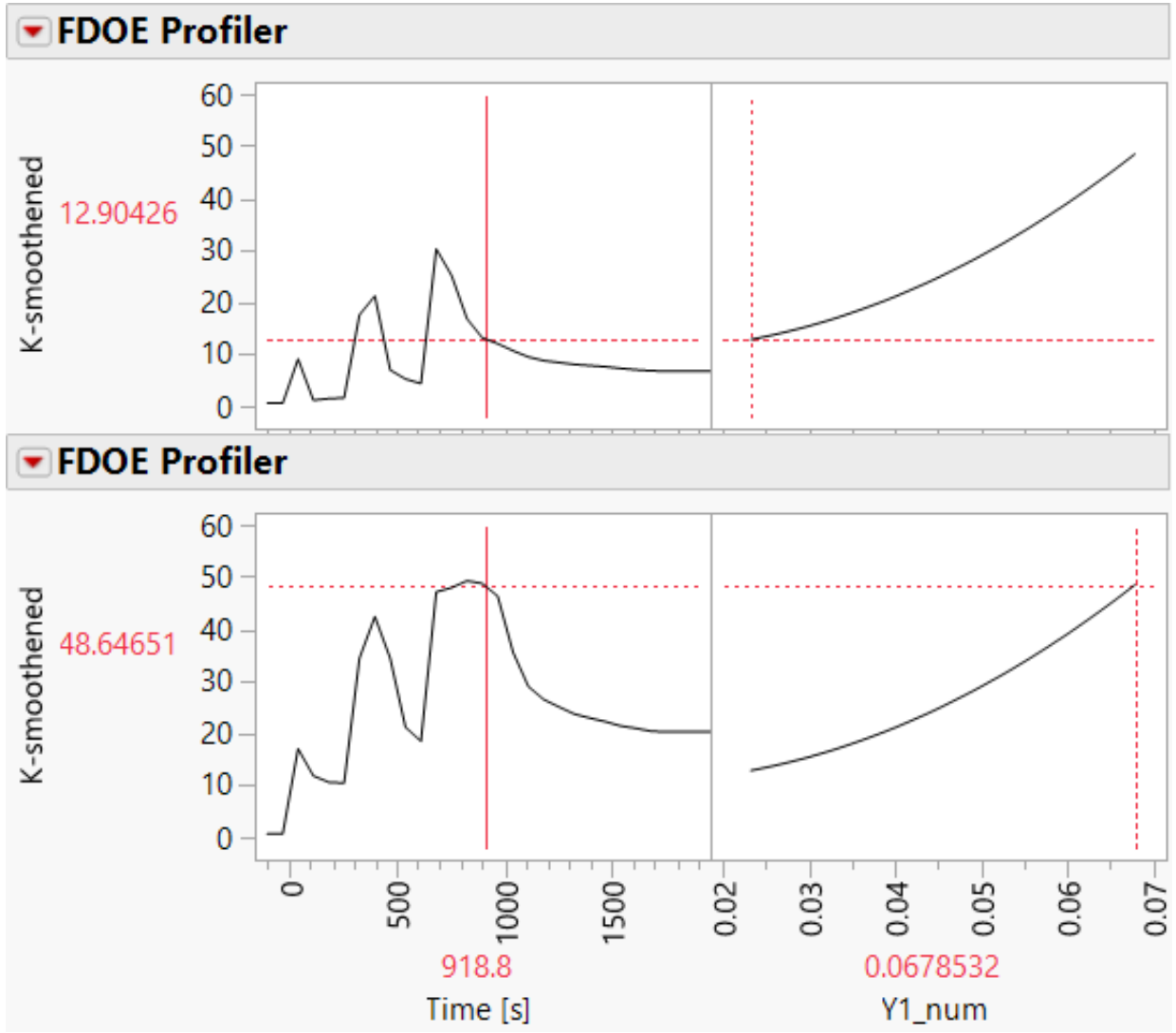
- Can see some product differences. Product replicates often group together; some not
- Most of the variability seen is coming from products



FDA Findings

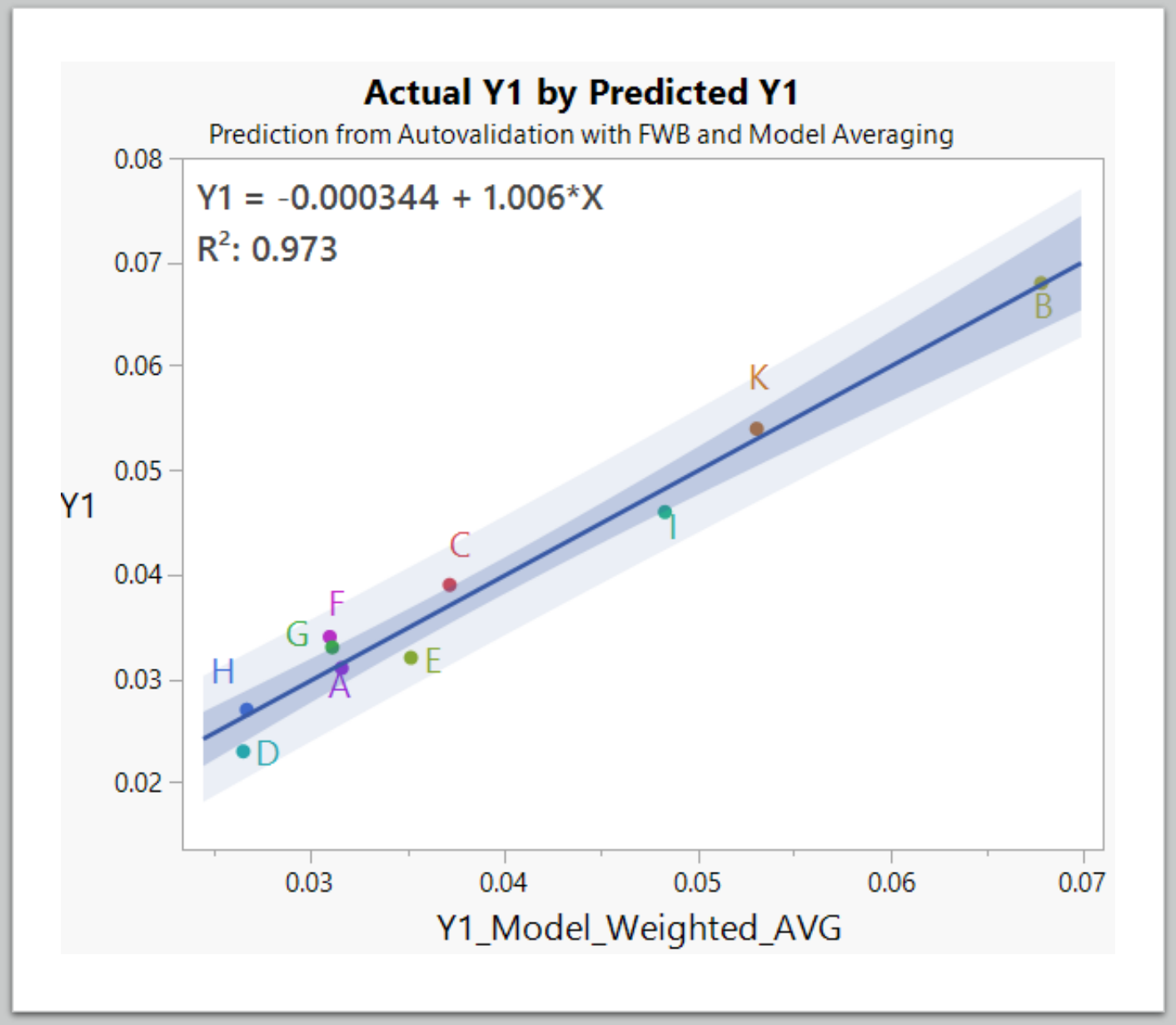
Functional DOE Profiler:

- The real purpose of JMP's FDE: Use the DOE capability of JMP
- Plotting K-curve predictions against Yield 1 performance
- Identify how good (top) Y1 curve compares to bad (bottom) Y1 curve
- Can do that for each Response of interest to visually understand what curves drive which consumer perception



JMP Y1 Predictions

- Use FDA *Function Summaries* with Auto-Validation, Weighing, and Model Averaging (*)
- R-Square = 0.97
- Press R-Square: 0.90
- Too good to be true?!!



(*) reference to auto-validation and model averaging (JMP Community)

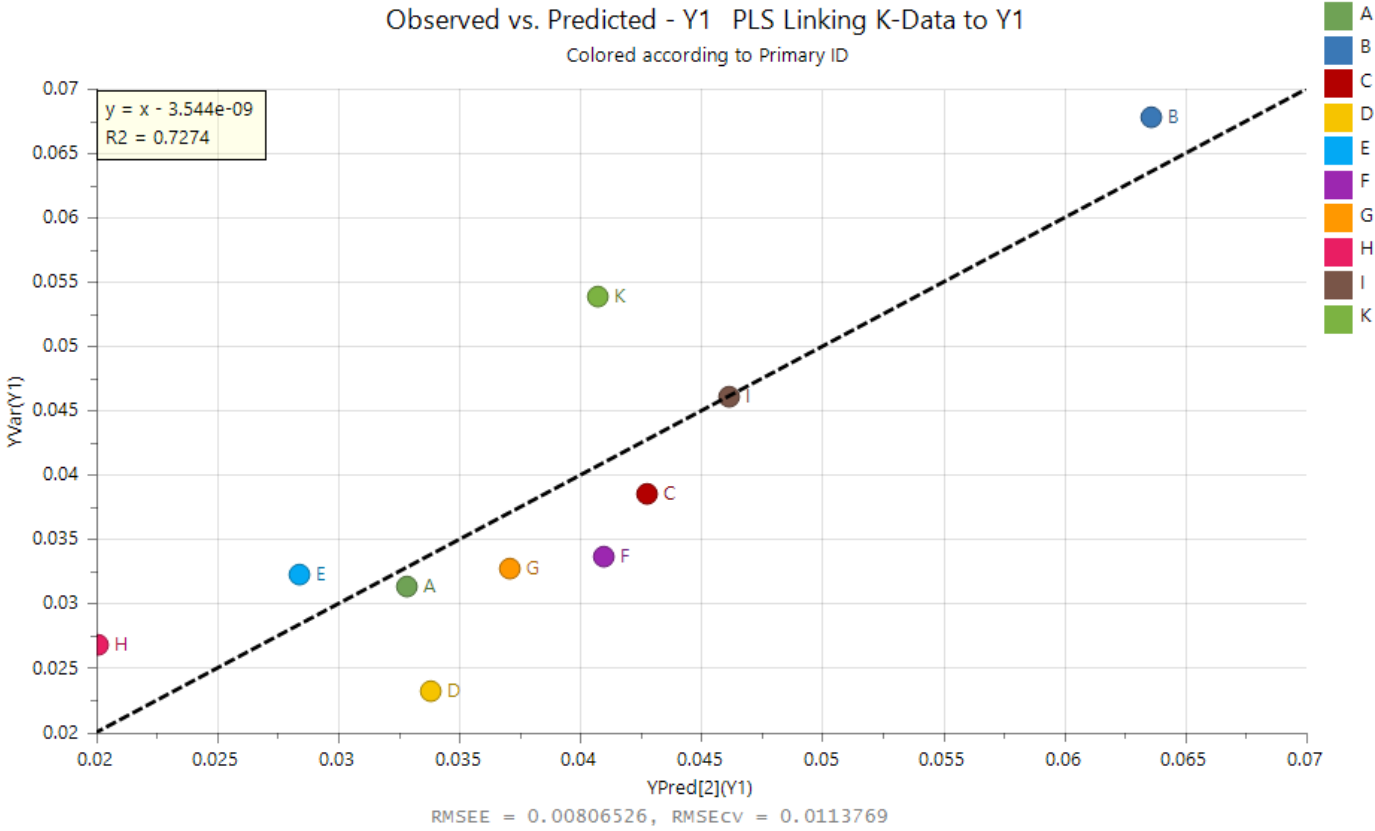
PLS - Predicting outcomes

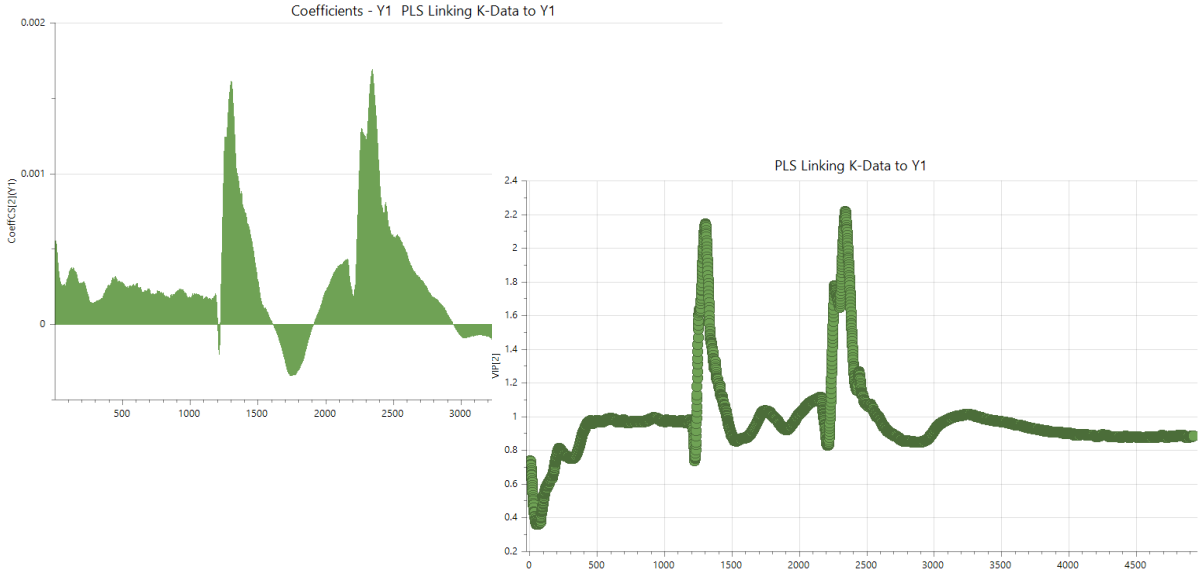
Looks like an OK(ish) model:

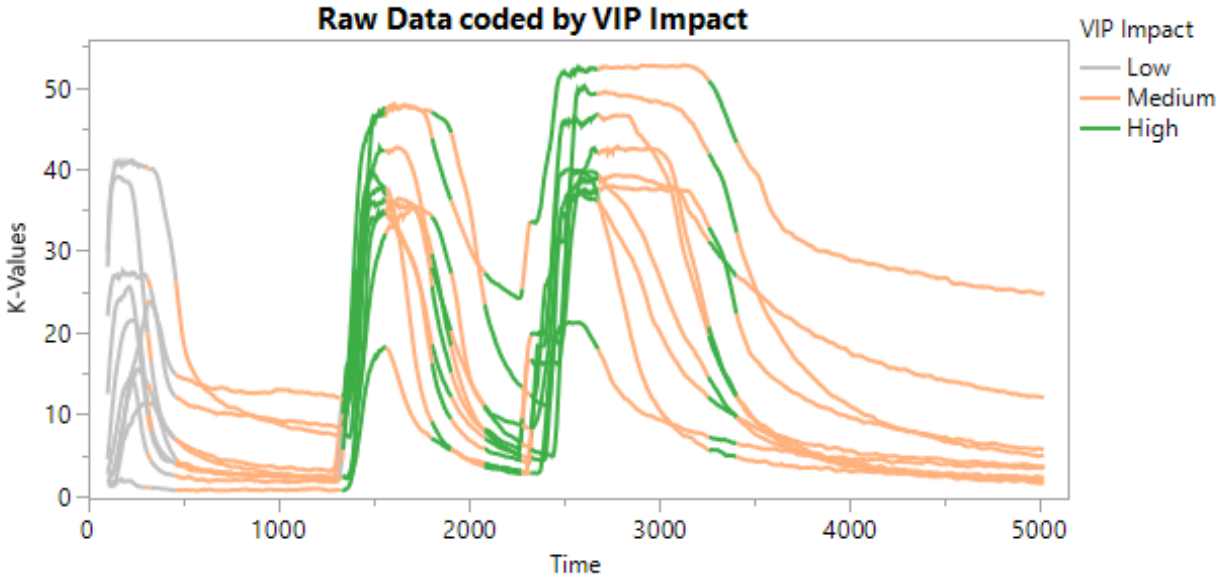
- R^2 is 73% and fit looks suitable
- Maybe an influence from B?

Cross-Validation measures (Q^2) are quite low at only 33%.

However, can get some idea of the regions that impacts predictions.







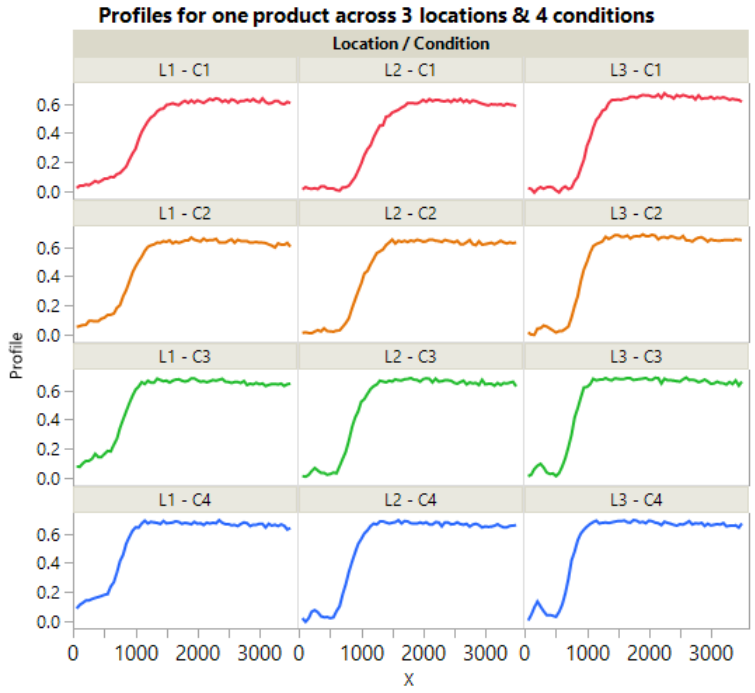
P-Data Curves – Multiple Traces

When assessing curves, we look at 4 conditions at three locations

- Conditions: C1 – C4
- Locations: 1 - 3.

We also have measures taken at C5 at location 1 only

- Different curve obtained as on a different device.



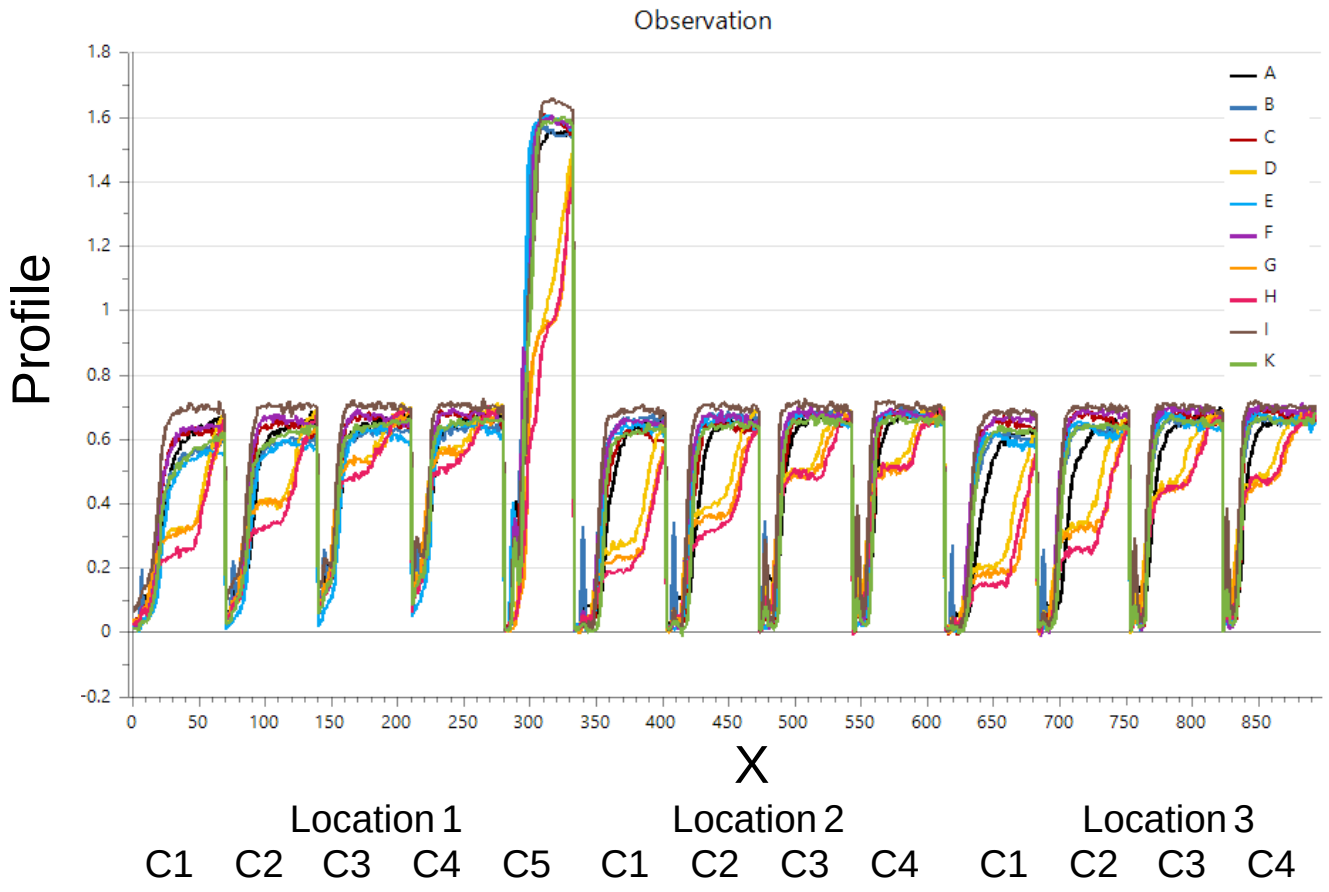
P-Data Curves – Multiple Traces

When assessing curves, we look at 4 conditions at three locations

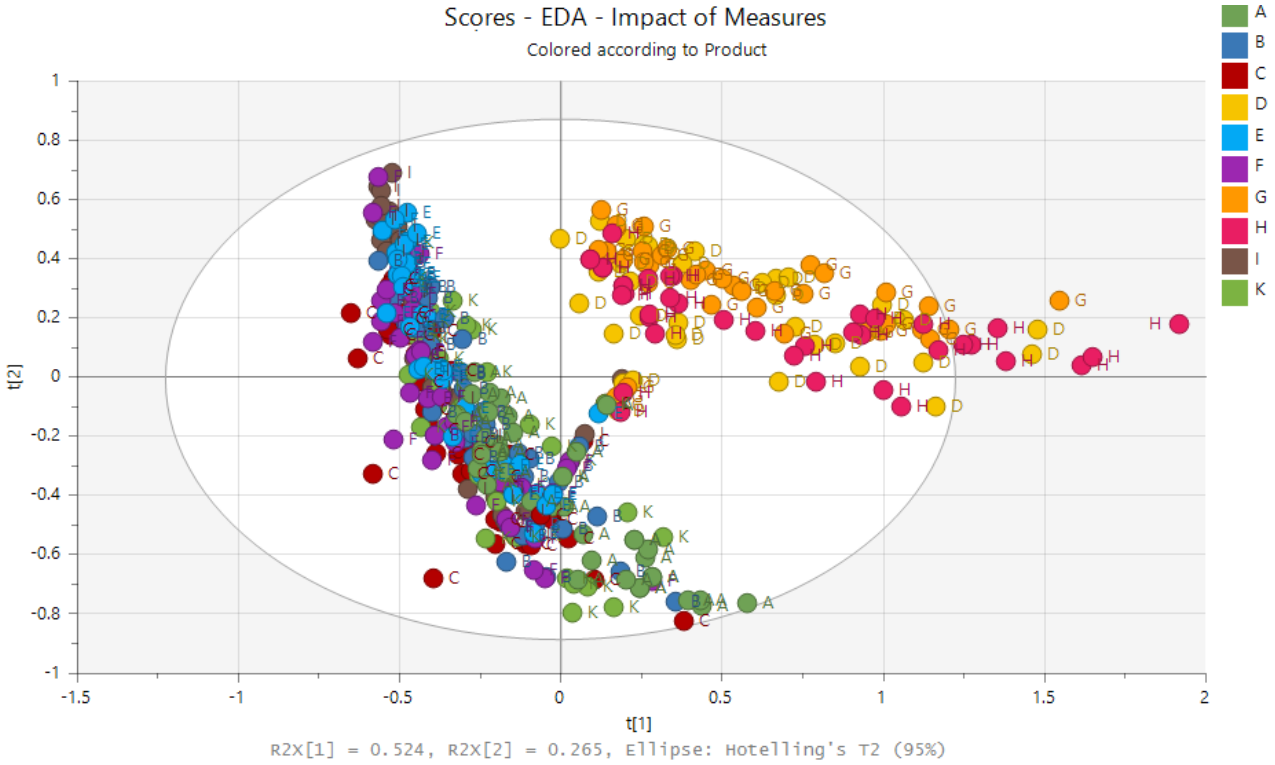
- Conditions: C1 – C4
- Locations: 1 - 3.

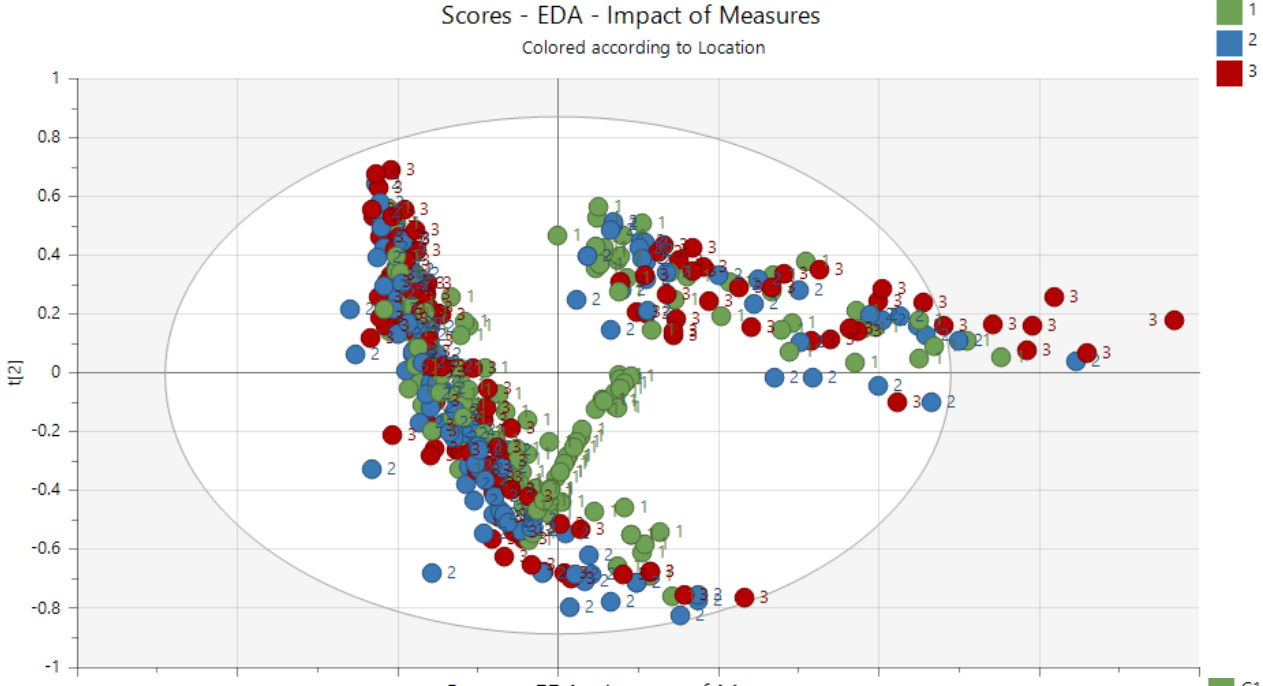
We also have measures taken at C5 at location 1 only

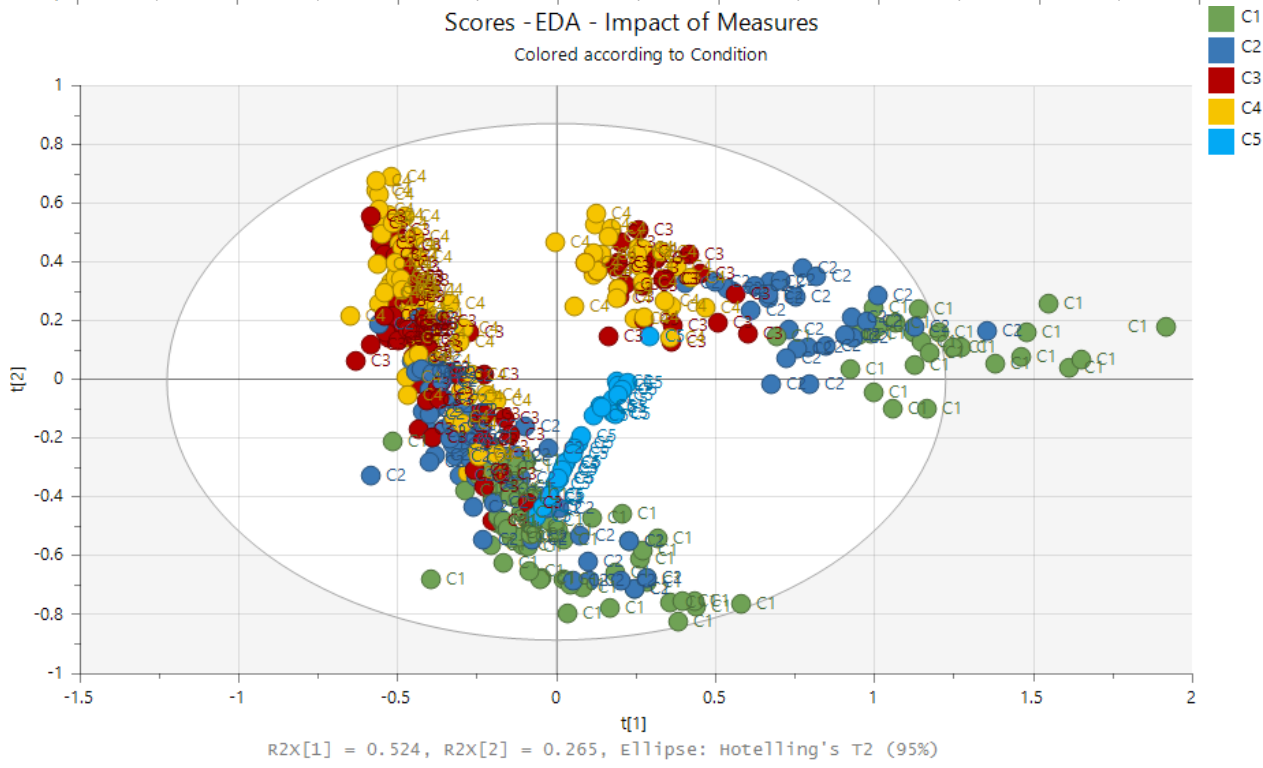
- Different curve obtained as on a different device.



P-Data Curves – PCA





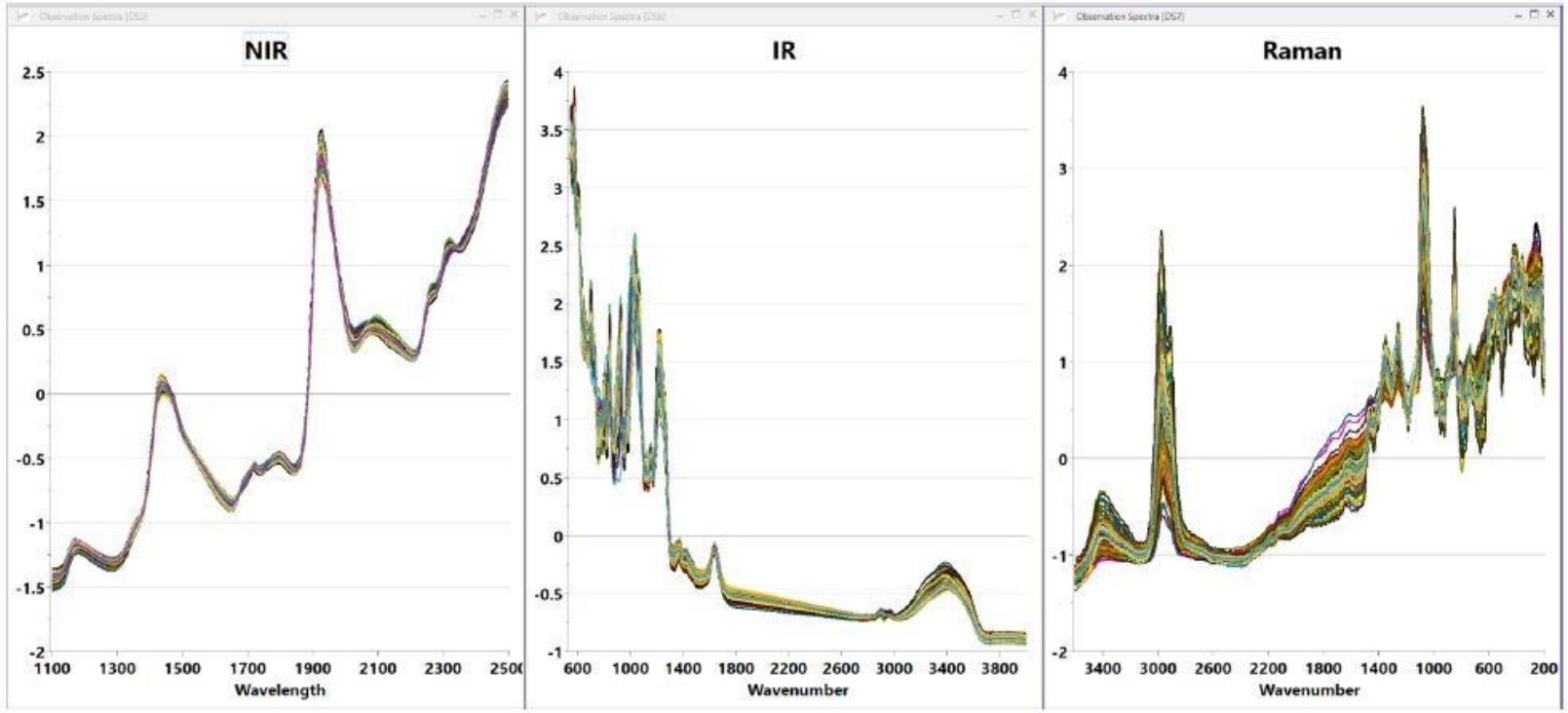


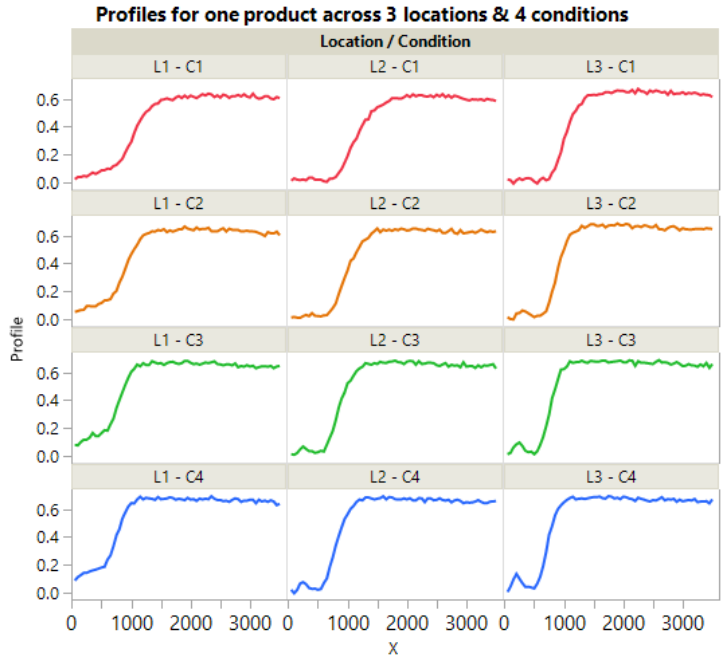
- Clear to differentiate the products into two groups/clusters – they make sense;-)
- No real patterns when colour by location
- Can see patterns due to the condition applied during measurement.

We shall focus on the location = 3 for now.

Multiblock Orthogonal Component Analysis (MOCA) & Hierarchical Modelling

- Can look at ‘blocks’ of data - normally different spectra.
  - Assess links between blocks – if they are not unique, potential redundancy.
  - Assess impact of each block on a response.





For our data – do we see overlap between the different locations and conditions?

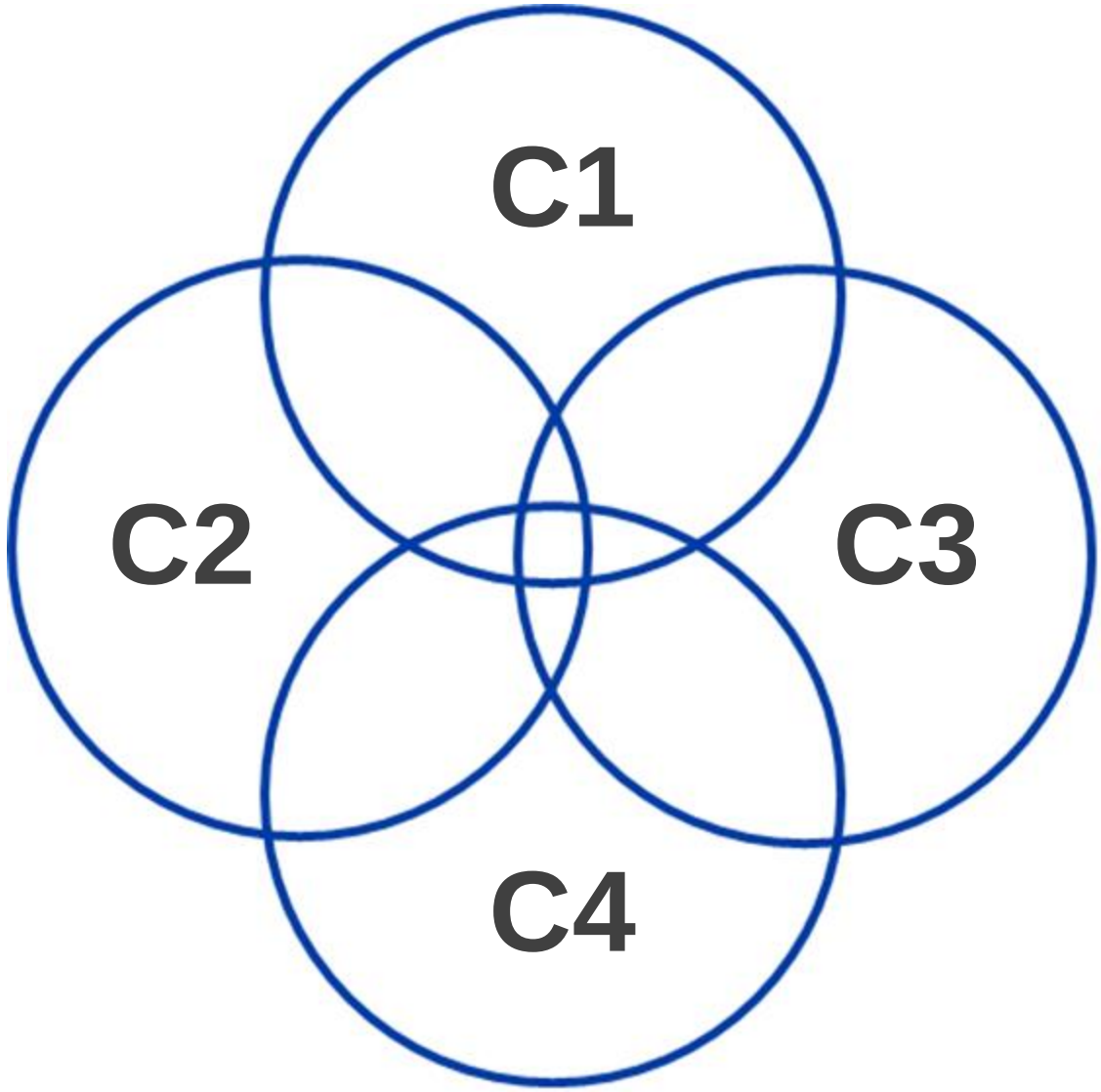
Multiblock Orthogonal Component Analysis
MOCA

Consider the four conditions at location **3**, and how the data may overlap.

The figure shows the case where we see some overlap between conditions.

We can see three forms of ‘overlap’

- **Globally Joint information**
- **Locally Joint information**
- **Unique information**



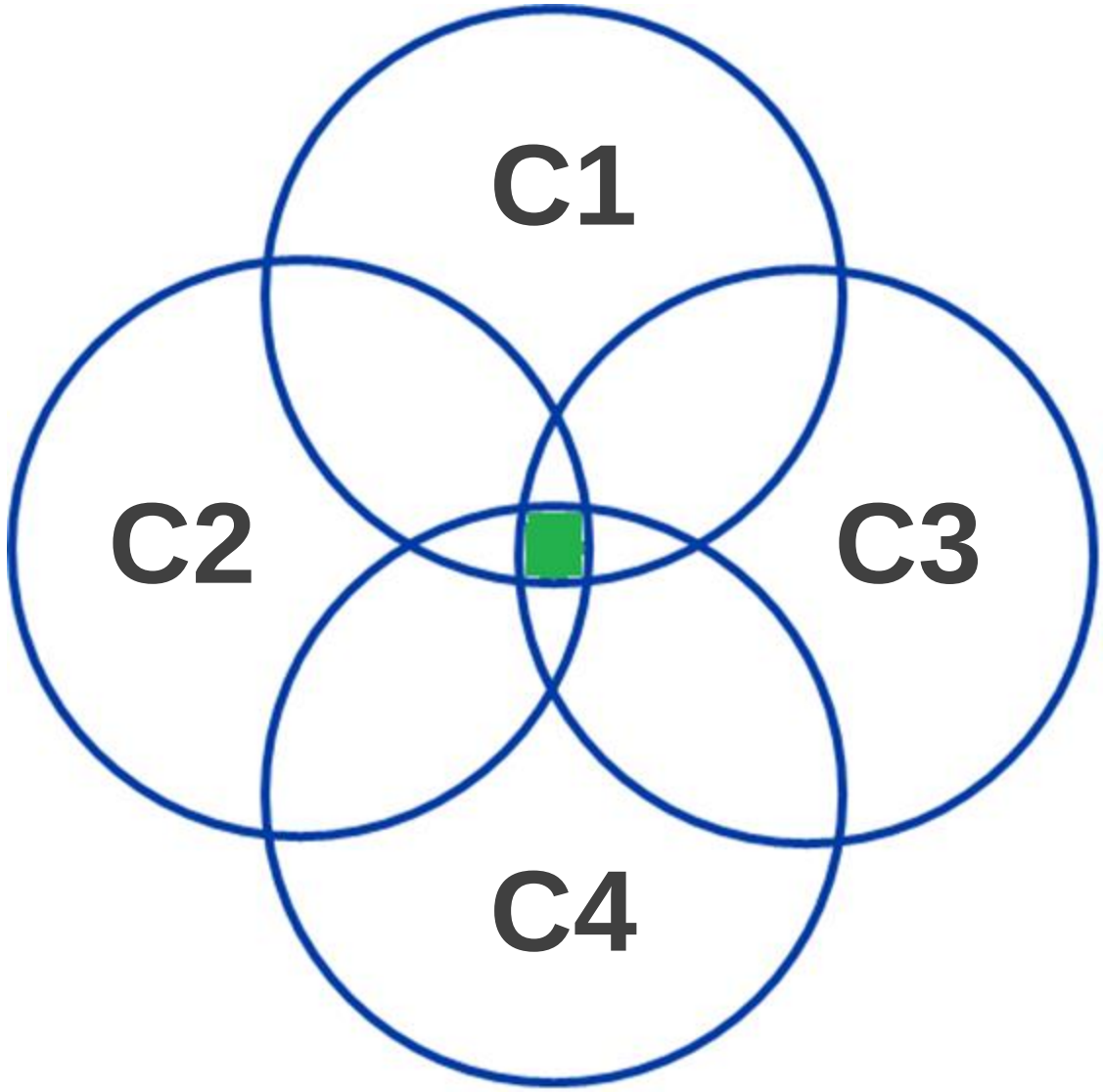
Multiblock Orthogonal Component Analysis
MOCA

Consider the four conditions at location **3**, and how the data may overlap.

The figure shows the case where we see some overlap between conditions.

We can see three forms of 'overlap'

- **Globally Joint information**
- **Locally Joint information**
- **Unique information**



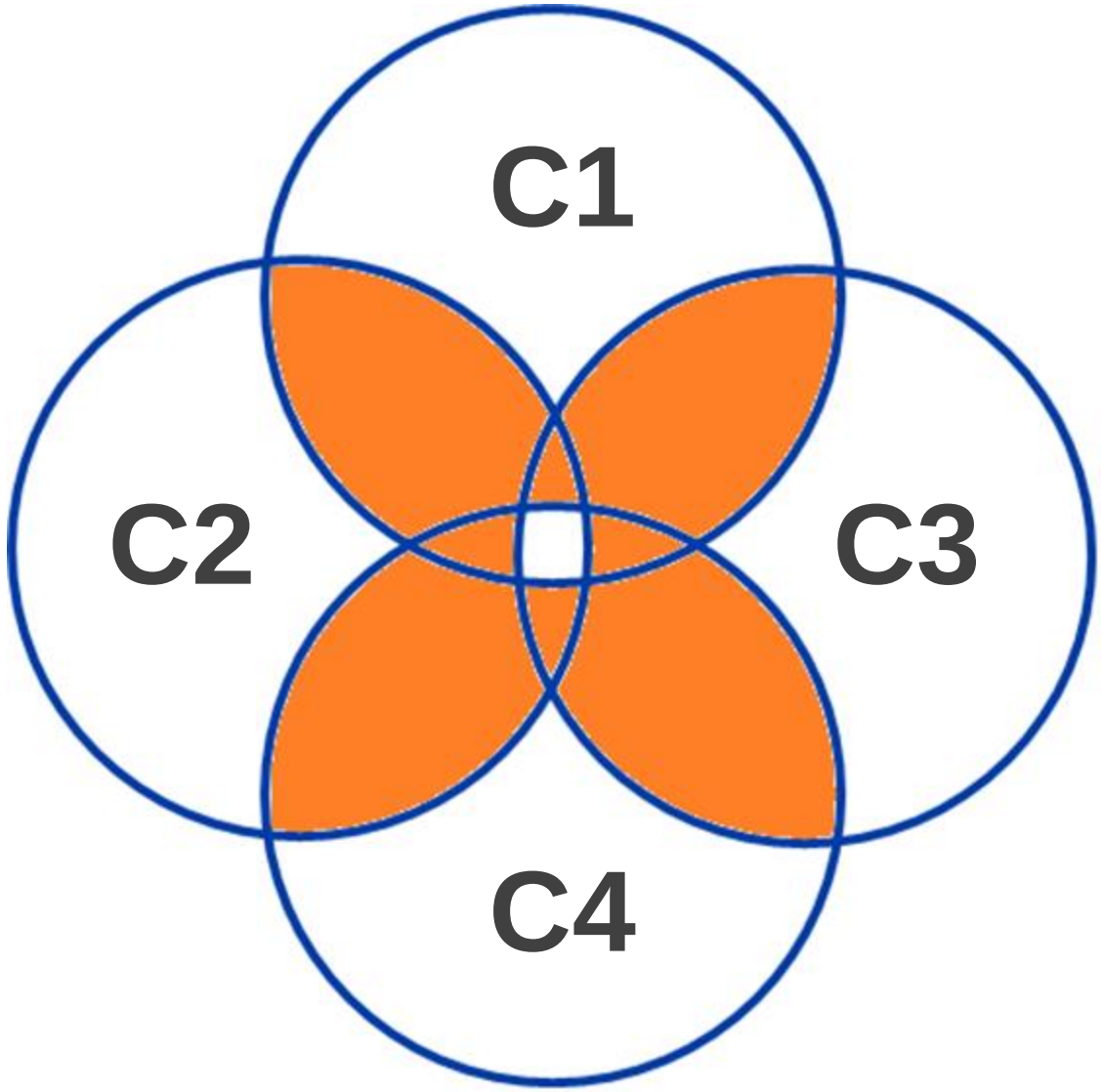
Multiblock Orthogonal Component Analysis
MOCA

Consider the four conditions at location **3**, and how the data may overlap.

The figure shows the case where we see some overlap between conditions.

We can see three forms of ‘overlap’

- **Globally Joint information**
- **Locally Joint information**
- **Unique information**



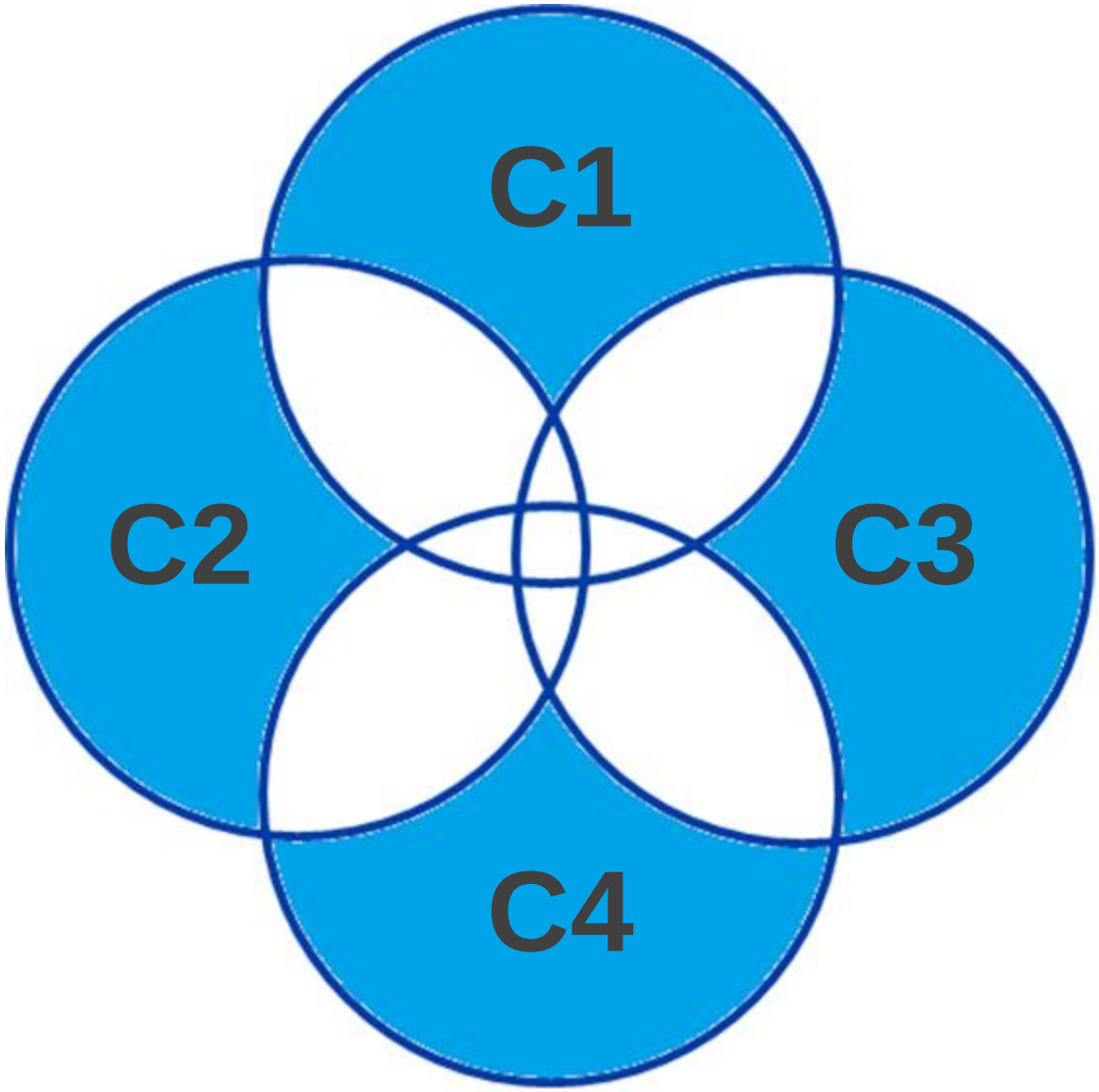
Multiblock Orthogonal Component Analysis
MOCA

Consider the four conditions at location **3**, and how the data may overlap.

The figure shows the case where we see some overlap between conditions.

We can see three forms of ‘overlap’

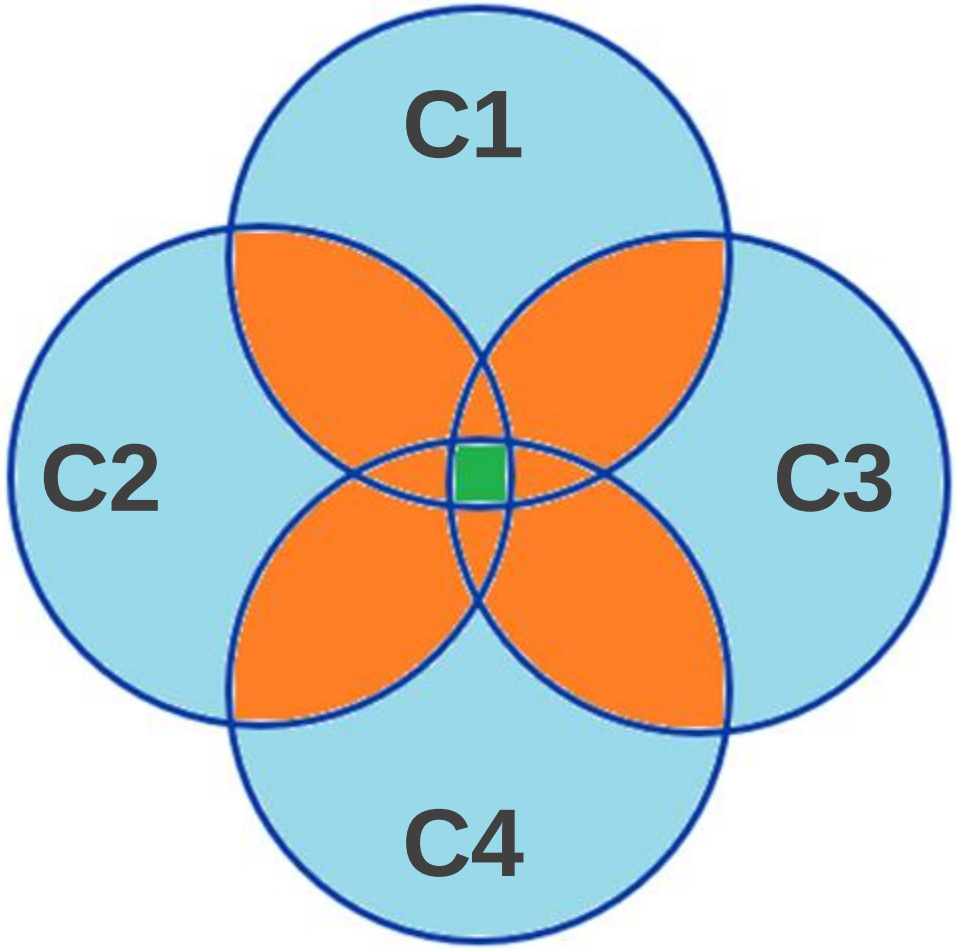
- **Globally Joint information**
- **Locally Joint information**
- **Unique information**



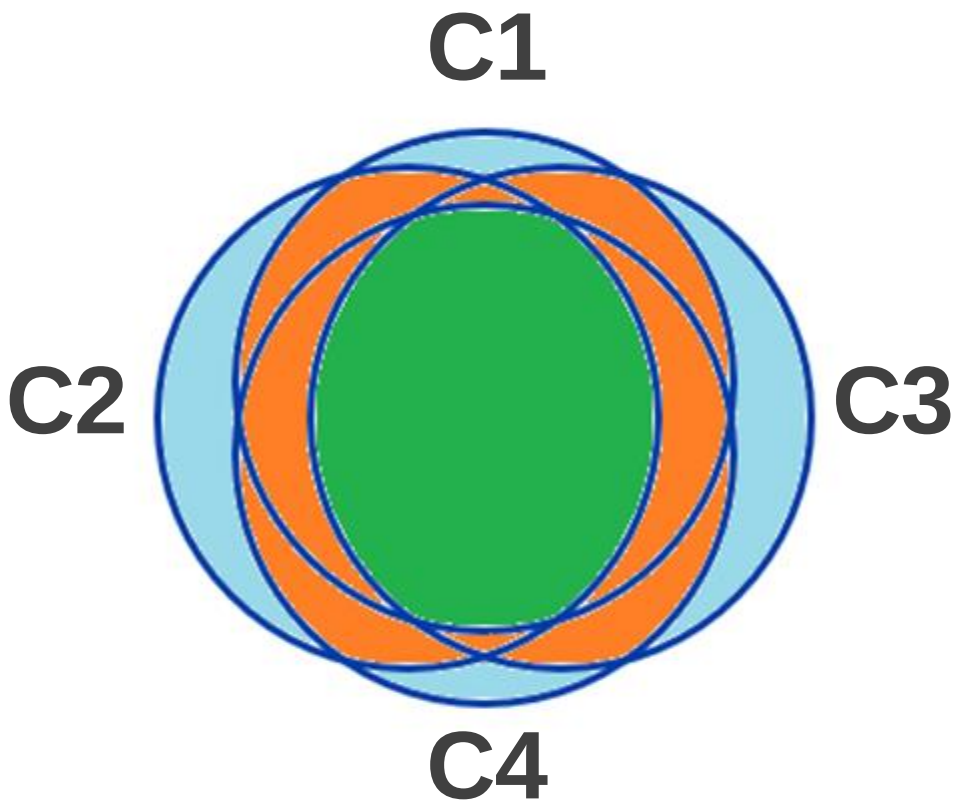
Multiblock Orthogonal Component Analysis

MOCA

Globally Joint information
Locally Joint information
Unique information



Relatively Unique Information
from each condition



Lots of Joint Information
across conditions

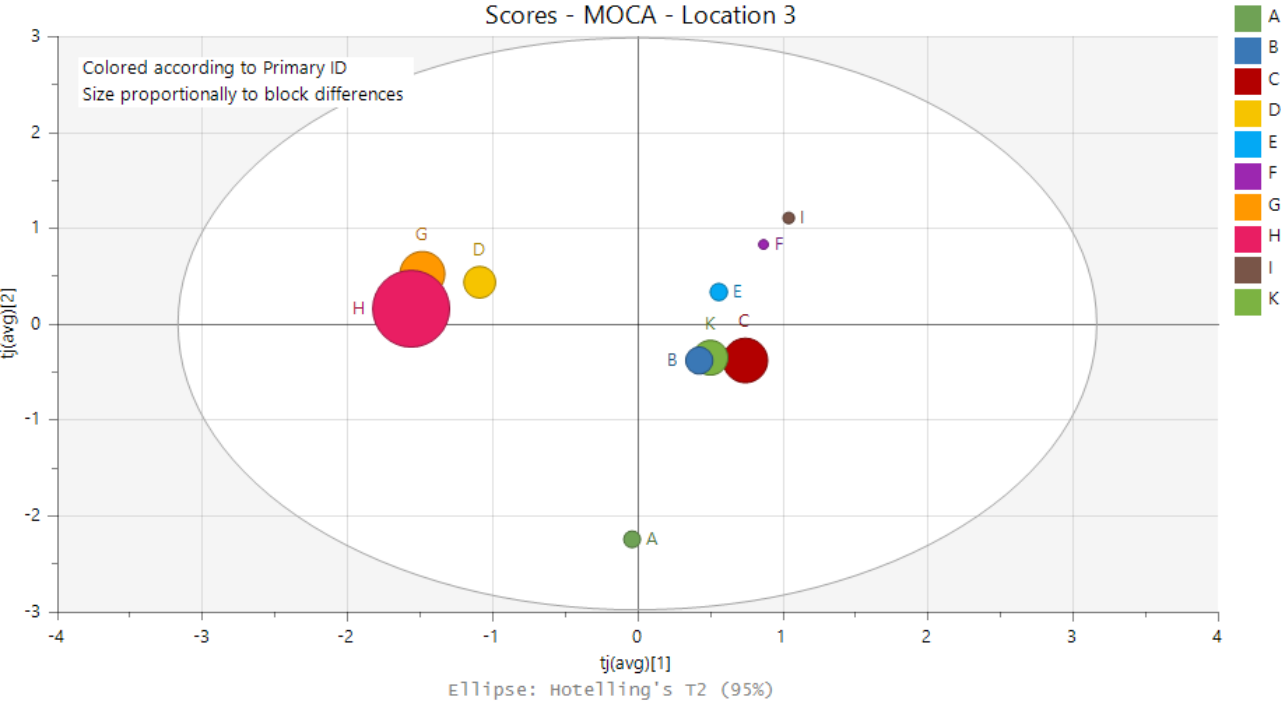
SIMCA MOCA

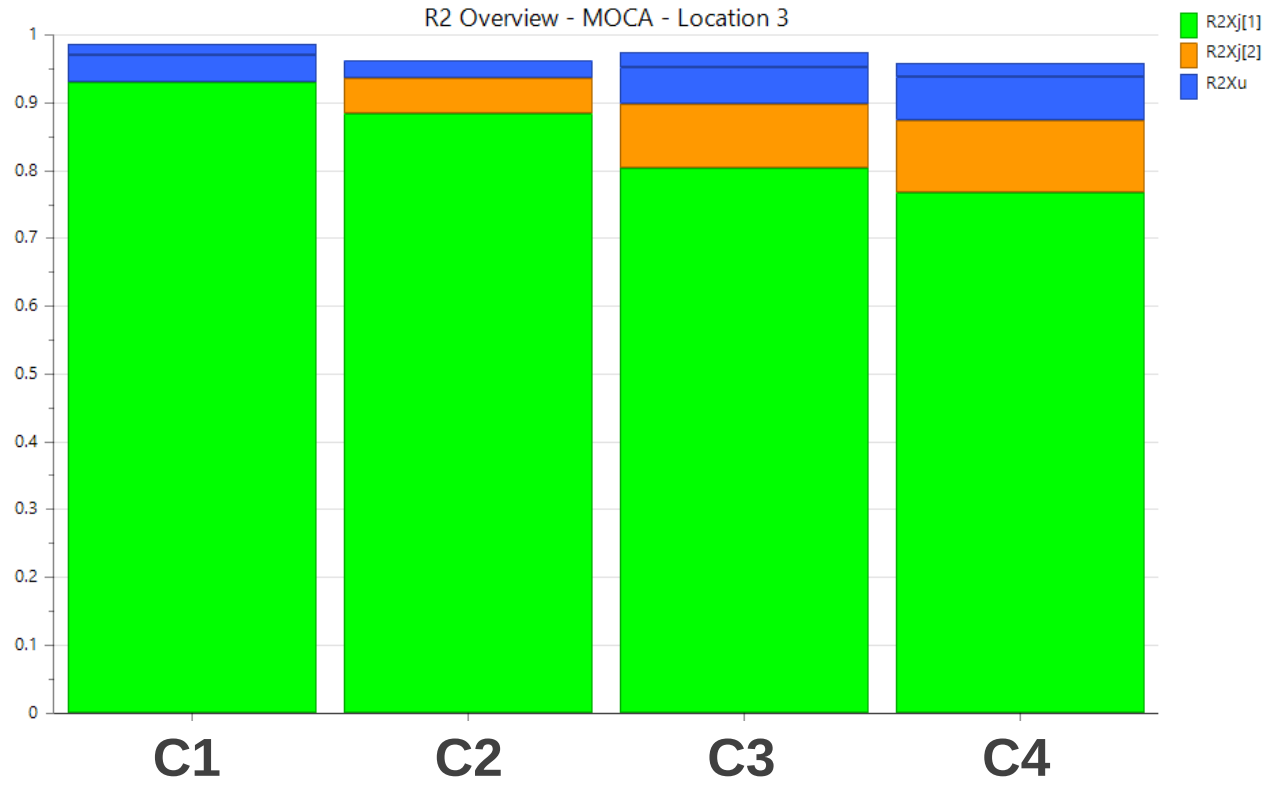
Analysis indicates much overlap between conditions (green & orange) and little uniqueness (blue). Also, product dependent.

Using MOCA & Hierarchal Modelling of Y2:

- C1 @ Locations 1 – 3, and C5 (Location 1)
- $R^2 = 85.5$ & $Q^2 = 69.3$

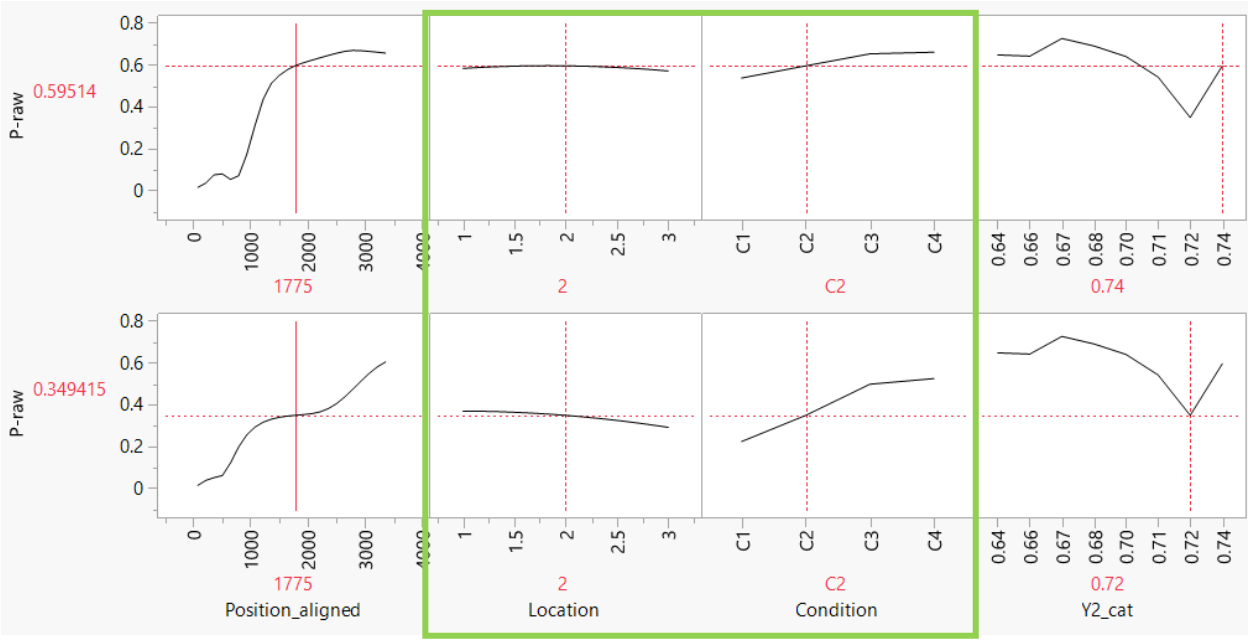
Component	R2X	R2X(cum)	R2X	R2X(cum)	R2X	R2X(cum)	R2X	R2X(cum)
Model		0.987		0.962		0.976		0.959
B	**Location 4**		**C1**		**C2**		**C3**	**C4**
Joint components		0.93		0.937		0.9		0.875
1	0.93	0.93	0.884	0.884	0.805	0.805	0.768	0.768
2	--	--	0.0528	0.937	0.0951	0.9	0.108	0.875
Unique components		0.0565		0.0255		0.076		0.0842
1	0.041	0.041	0.0255	0.0255	0.0524	0.0524	0.0629	0.0629
2	0.0156	0.0565	--	--	0.0236	0.076	0.0213	0.0842





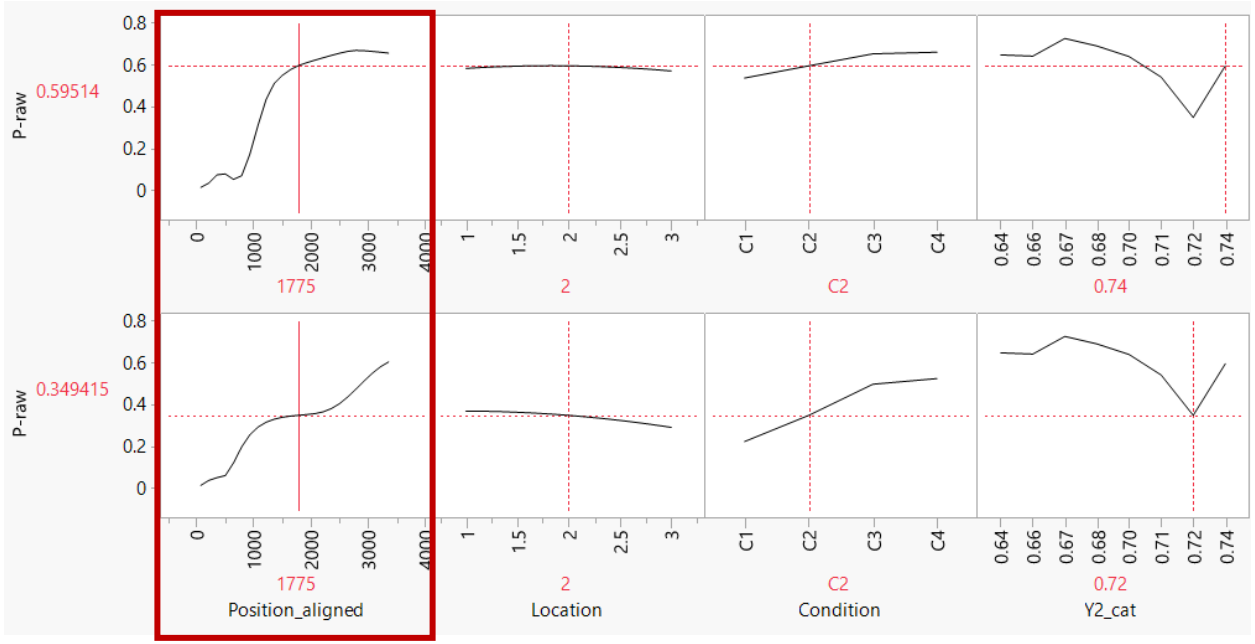
P-data Profiles and Y2: FDA Findings

- Location has less impact on curve shape than condition



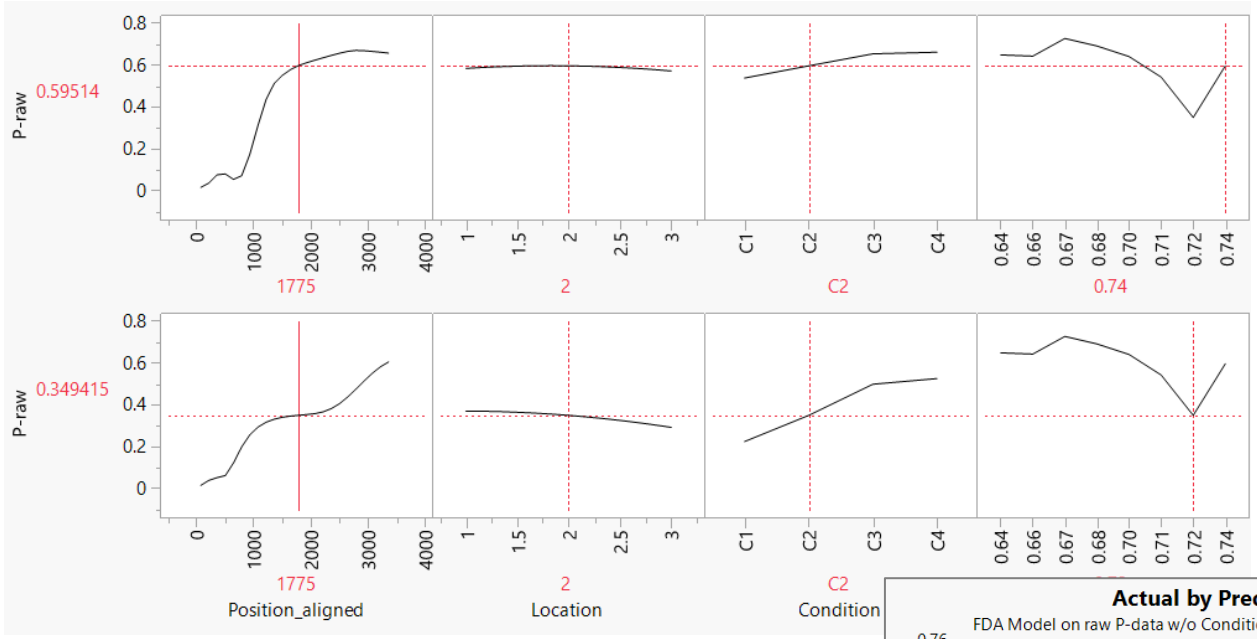
P-data Profiles and Y2: FDA Findings

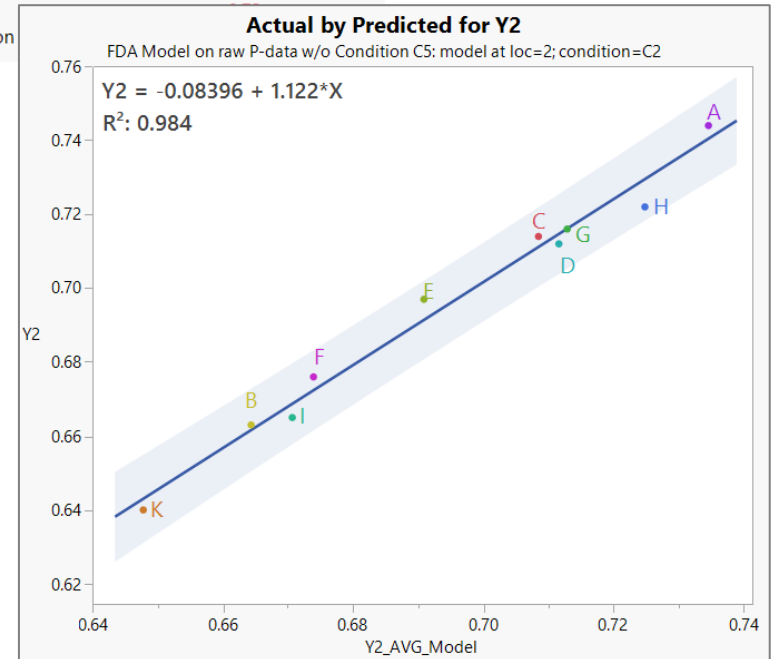
- Location has less impact on curve shape than condition
- Curve shapes not “commonly” related to product performance . Is “average curve” = “Golden Curve”? Are there several golden curves?
- Can’t answer yet what location-condition combinations meaningful



P-data Profiles and Y2: FDA Findings

- Location has less impact on curve shape than condition
- Curve shapes not “commonly” related to product performance . Is “average curve” = “Golden Curve”? Are there several golden curves?
- Can’t answer yet what location-condition combinations meaningful
- Auto-Validation + Model averaging:
  $R^2 = 0.98$
  Press $R^2 = 0.97$





Summary of models

K-Data

FDA:

- Very strong model from FDA
- We can predict curve shape from Yield 1. Do not understand what drives deviations.
- Simple data preparation, but an element of ‚black box‘ modelling.

PLS:

- A relatively strong model (not as good as FDA!).
- Can assess regions of curve that drive Yield 1 – potential for variable selection?
- Not so simple to analyse – element of data prep.

P-Data

PLS:

- MOCA and Hierarchical PLS yield a good model.
- Understanding of regions of curve as well as conditions and locations that drive predictions.

FDA:

- Understanding location and condition impact on curve
- Very good, despite questionable model on Yield 2
- Cannot identify which condition-location combinations needed for Golden Curve understanding.

Summary and Next Steps

Summary

- Simple EDA using FDA and PLS/PCA shows clear patterns, and we can differentiate products.
- With the (limited) data we have, we have a proof of principle to model our Yield responses better than we currently can do.
- The modelling tools have shown which aspects of the data collected drive these predictions and product differentiation.
- Perfect example that ‘too many cooks spoil the broth’ is not always correct – the more tools, the greater the understanding in this case – even if we don't agree.
- Work is ongoing – bugs, new data, feedback, new understandings drive what we are doing.

Next Steps

- External validation of models!!
- More understanding of how different technical measures drive each other – can we simplify what we collect?
- Make use of FDA DoE tools to assess product making and material composition impact on curve shapes.
- Follow-up with JMP on explaining which part/aspect of the curve most impacts Yield predictions. Combine B- and P-Splines?!
- We have nearly caught the ‘golden curve’. However, answers to some of the above will hopefully mean we will eventually capture the curve entirely.

DATA &
MODELING
SCIENCES

Unlocking Innovation

Thank You

Any Questions?