

Configuration Manual

MSc Research Project
Programme Name

Vikas Khatri
Student ID: x21164894

School of Computing
National College of Ireland

Supervisor: Sasirekha Palaniswamy

National College of Ireland
MSc Project Submission Sheet
School of Computing



Student Name: Vikas Khatri
Student ID: x21164894
Programme: MSc in Data Analytics **Year:** 2023
Module: Research Project
Lecturer: Sasirekha Palaniswamy
Submission Due Date: 14 Dec 2023
Project Title: Machine Learning to forecast Cell growth in Bioreactor using Raman spectroscopy
Word Count: 856 **Page Count:** 11

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

Date: 14 Dec 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Vikas Khatri
Student ID: x21164894

1 Introduction

The configuration manual is the procedure, outlining the key aspects of the research implementation. Hardware/Software requirements, dataset collection and data import, data pre-processing, implementation and evaluation are part of this configuration manual. The configuration manual helps to outline the key steps taken to carry out the research and build the model.

2 Hardware Requirement

Organisation's infrastructure is used to carry out this study. SIMCA (Soft independent modelling by class analogy) workstation in the lab is used to configure the task. The lab workstation with the hardware configuration of 64-bit Windows 10 OS, Intel(R) Core (TM) i5-10500 CPU @ 3.10GHz Processor and 32GB of RAM are used.

3 Software Requirement

SIMCA software is used to build the model and SIMCA version 17.0.2.34594 is used.

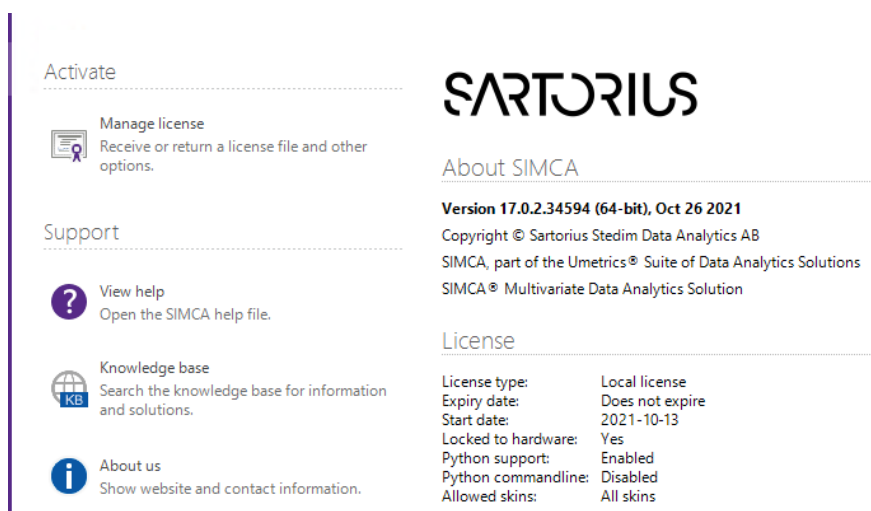


Figure 1: SIMCA software

4 Dataset selection

This study makes use of company data. The bioreactors in the laboratory, which are controlled by a DCU (digital control unit), produce the data. The Modular Fermentation and Culture System (MFCS), which is connected to DCU, analyses and processes bioreactor data, permits trending, and creates recipes that are often used in laboratories. To store and distribute data to various systems, MFCS is linked to a site historian/PI (Process Intelligence) server. Firewalls divide the networks where the MFCS and PI systems are located: the lab network and the company network. Bioreactor data is made available on the enterprise network by the PI system, which is utilised by SIMCA online. The parameters of the bioreactor's batch data are gathered by SIMCA online and made available to the SIMCA client located in the laboratory. Raman spectroscopy is a stand-alone piece of equipment in the lab; to measure the VCD, the Raman probes are directly attached to the bioreactor. VCD is measured on a regular basis for every batch. The Raman system exports the VCD data as a comma-separated values (CSV) file, which is then entered into SIMCA in the laboratory.

Raman spectroscopy is a stand-alone piece of equipment in the lab; to measure the VCD, the Raman probes are directly attached to the bioreactor. VCD is measured on a regular basis for every batch. The Raman system exports the VCD data as a comma-separated values (CSV) file, which is then entered into SIMCA in the laboratory.

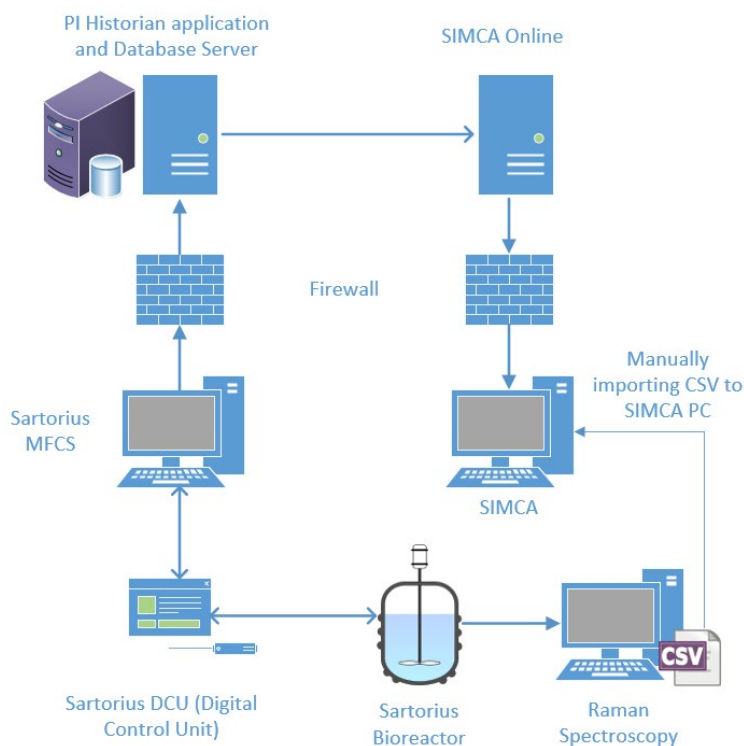


Figure 2: System Architecture

5 Importing required dataset

Dataset from above systems then imported to SIMCA.

KCDM Batch model MS&T Lab.usp - SIMCA - [KCDM Batch model MS&T Lab [M1:M1]]

File Home Data Analyze Predict View Tools Developer

Project Dataset New Edit Delete Statistics Compare models Model type Autofit Add Remove Summary of fit Overview Variable S

Workset [M1:M1]

Overview Variables Observations Transform Lag Expand Scale Spreadsheet

Variables: included: 30 (X=29, Y=1), excluded: 7, selected: 1

Variable ID	Info
X AIRSP Value ccm	
X AIRSP ST PT ccm	
X BASESUB ST PT %	
X BASESUB Value %	
X BASET Value ml	
X CO2SP Value ccm	
X CO2SP ST PT ccm	
X JTEMP Value °C	
X JTEMP ST PT °C	
X O2SP Value ccm	
X O2SP ST PT ccm	
X pH st pt	
X pH Value	
X pO2 ST PT % sat	
X pO2 Value % sat	
X STIRR Value rpm	
X STIRR ST PT rpm	
— SUBS A ST PT %	
X SUBS A Value %	
— SUBS B ST PT %	
X SUBS B Value %	
X TEMP ST PT °C	
X TEMP Value °C	
Y Time days	Shifted
X AIPSP st pt_value difference	
— BASESUB st pt_value difference	
X CO2SP st pt_value difference	
X JTEMP st pt_value difference	
X O2SP st pt_value difference	
X pH st pt_value difference	
X pO2 st pt_value difference	
X STIRR st pt_value difference	
— SUBSA st pt_value difference	
— SUBSB st pt_value difference	
X TEMP st pt_value difference	

Primary ID: TEMP st pt_value difference
Var. Sec. ID:1: TEMP st pt_value difference
Var. Sec. ID:2: TEMP st pt_value difference

X Y Exclude Phases: Set phase

Figure 3: Dataset imported

6 Pre-processing

The batches were produced throughout a range of dates and months, thus SIMCA generated a new parameter called "Time days" that allows all the batches to be compared. Time days reflect the stage of the batch within a day, such as the first, fourth, etc., rather than on a single time or day. To process the input and create the model, SIMCA performed this step. Every batch's "Time days" parameter is displayed in the figure below. Every colour on the trend indicates a batch, while the y axis displays the number of days the batch ran for, and the x axis represents the primary id (there is a unique primary id for each timestamp).

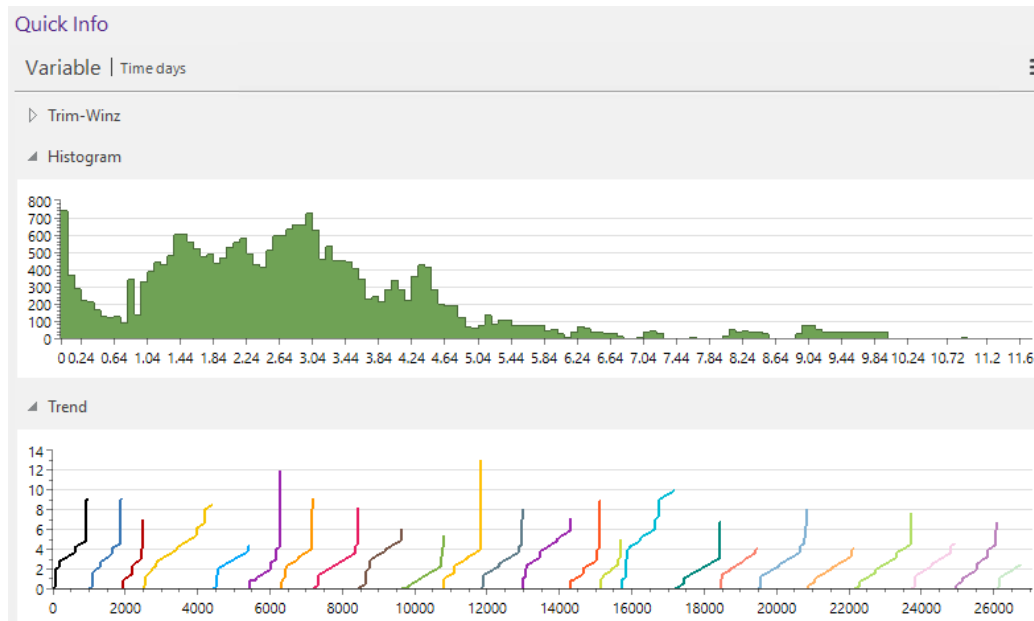


Figure 4: Time days parameter representing each batch's timeline

Outlier value can be selected and removed from the variable window.

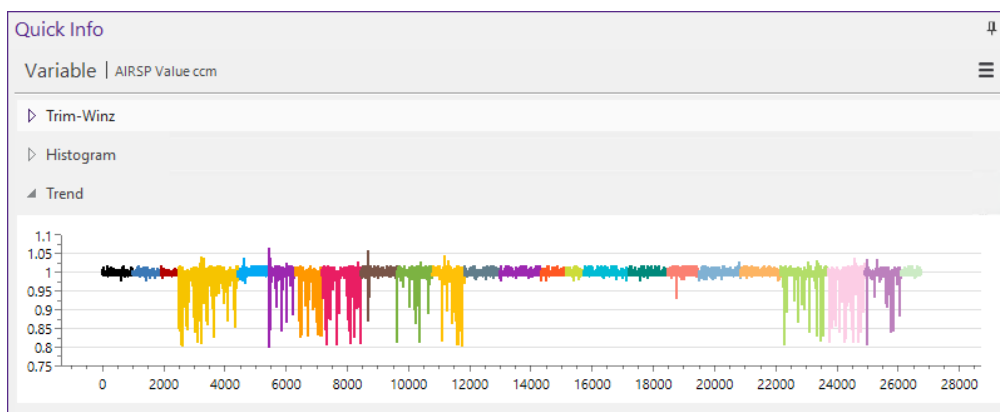


Figure 5: AIRSP Value for each batch

Additionally, looking at each variable's trend, co-related variables which don't add any value to models are dropped from model building.

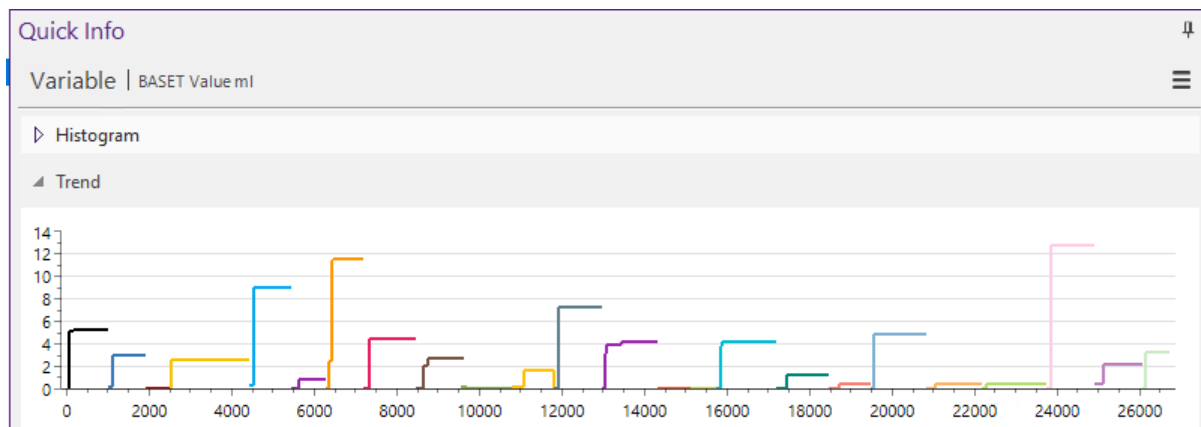


Figure 6: BASET Value for each batch

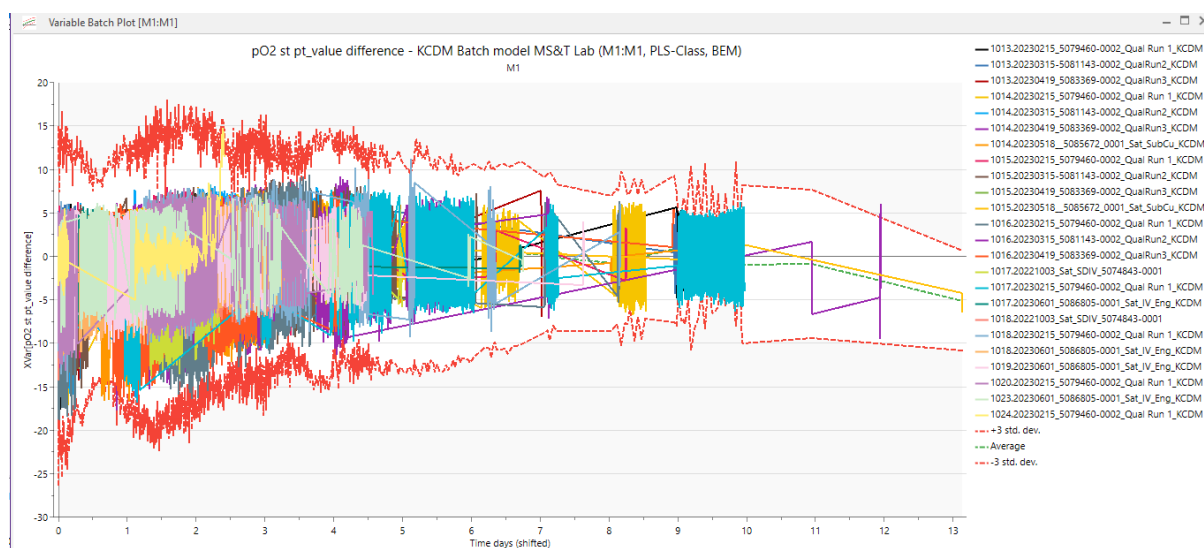


Figure 7: pO2 set point value difference for each batch

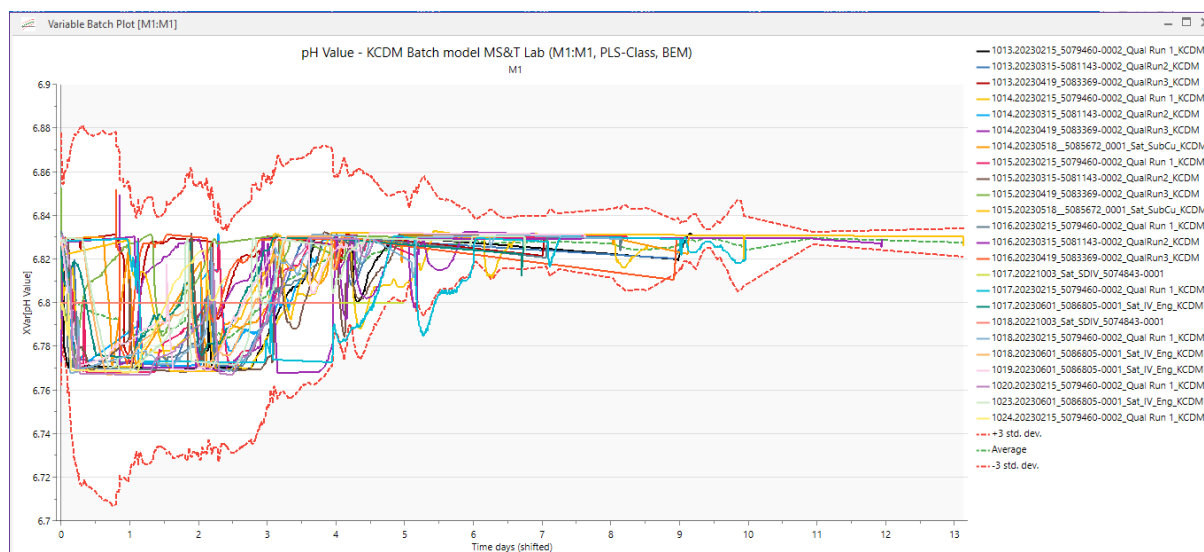


Figure 8: pH value for each batch

7 Model building

Offline data from Raman spectroscopy was added to batch evolution model to make the batch level model.

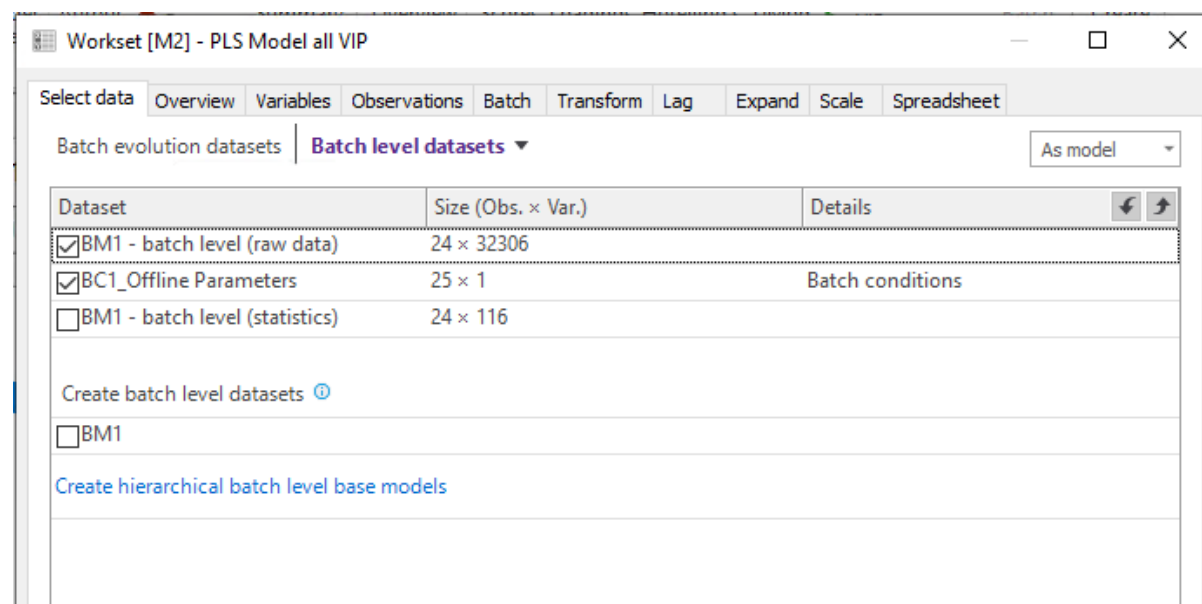


Figure 9: Batch level models

Summary of fit can be shown by selecting the model and clicking at “Summary of fit” icon on the menu bar highlighted in yellow.

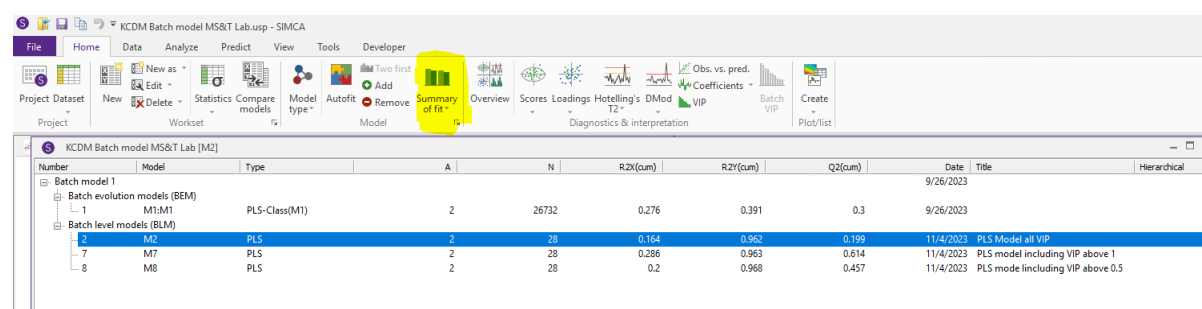


Figure 10: Summary of fit

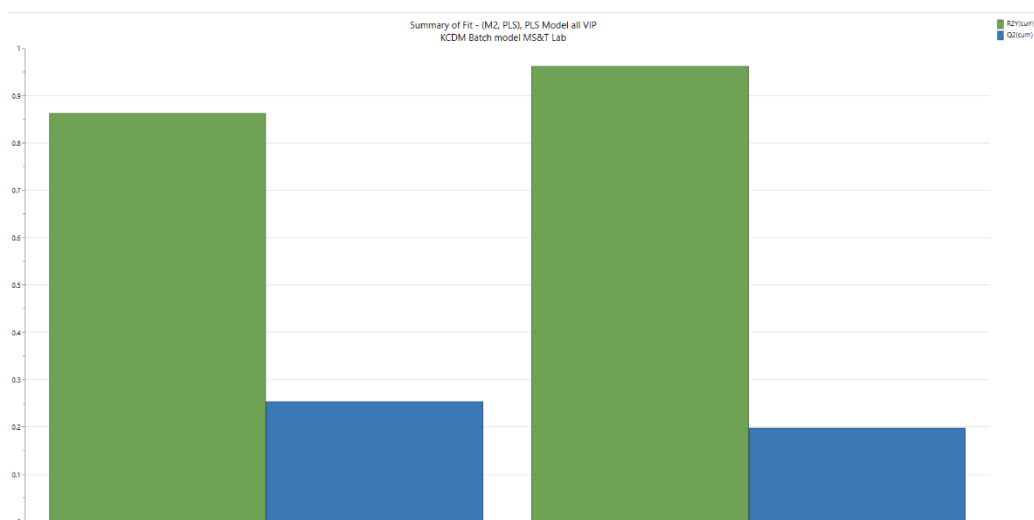


Figure 11: Summary of fit plot – all VIP value

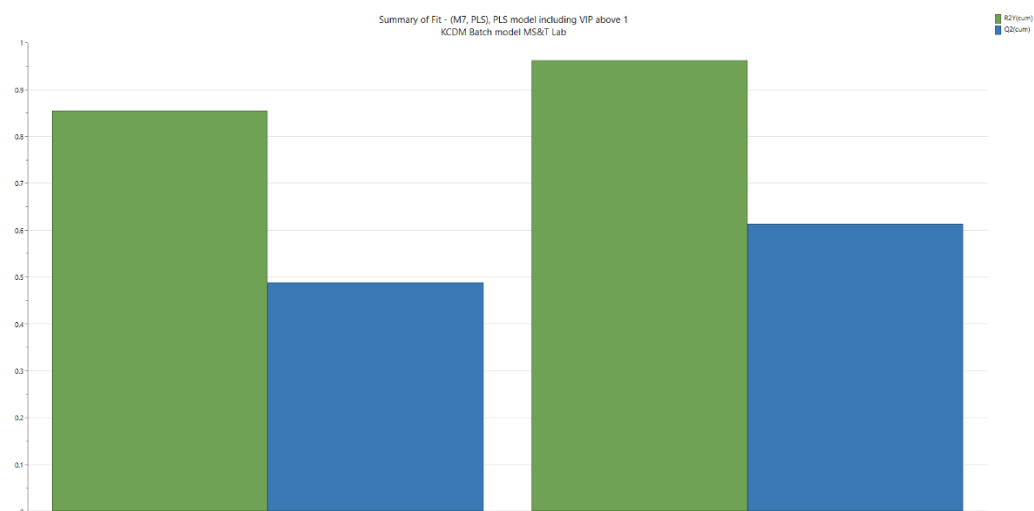


Figure 12: Summary of fit plot – VIP above 1

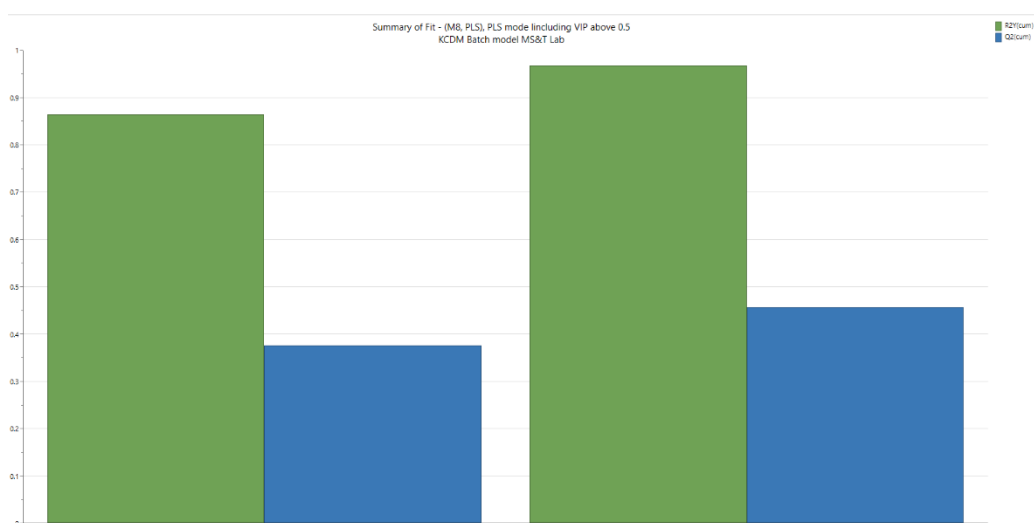
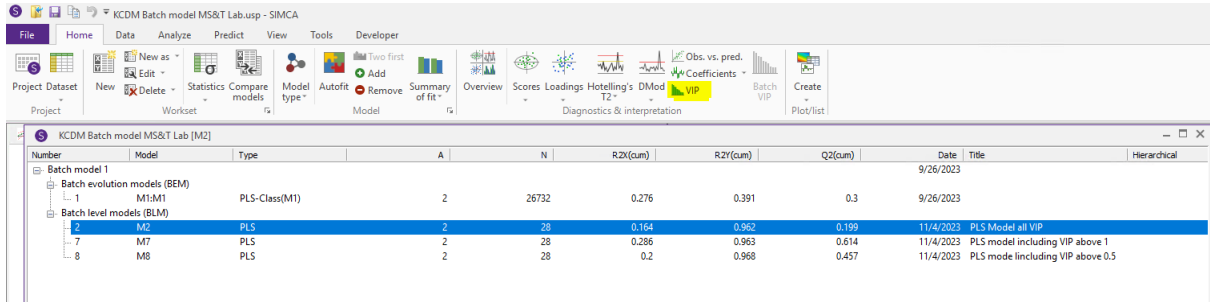


Figure 13: Summary of fit plot – VIP above 0.5

VIP plot can be shown by selecting the model and clicking at “VIP” icon on the menu bar highlighted in yellow.



The screenshot shows the SIMCA software interface with the 'VIP' icon highlighted in yellow on the menu bar. Below the menu bar is a table titled 'KCDM Batch model MS&T Lab [M2]'.

Number	Model	Type	A	N	R2(cum)	R2Y(cum)	Q2(cum)	Date	Title	Hierarchical
Batch model 1										
Batch evolution models (BEM)										
1	M1:M1	PLS-Class(M1)	2	26732	0.276	0.391	0.3	9/26/2023		
Batch level models (BLM)										
2	M2	PLS	2	28	0.164	0.962	0.199	11/4/2023	PLS Model all VIP	
7	M7	PLS	2	28	0.286	0.963	0.614	11/4/2023	PLS model including VIP above 1	
8	M8	PLS	2	28	0.2	0.968	0.457	11/4/2023	PLS model including VIP above 0.5	

Figure 14: VIP plot for models

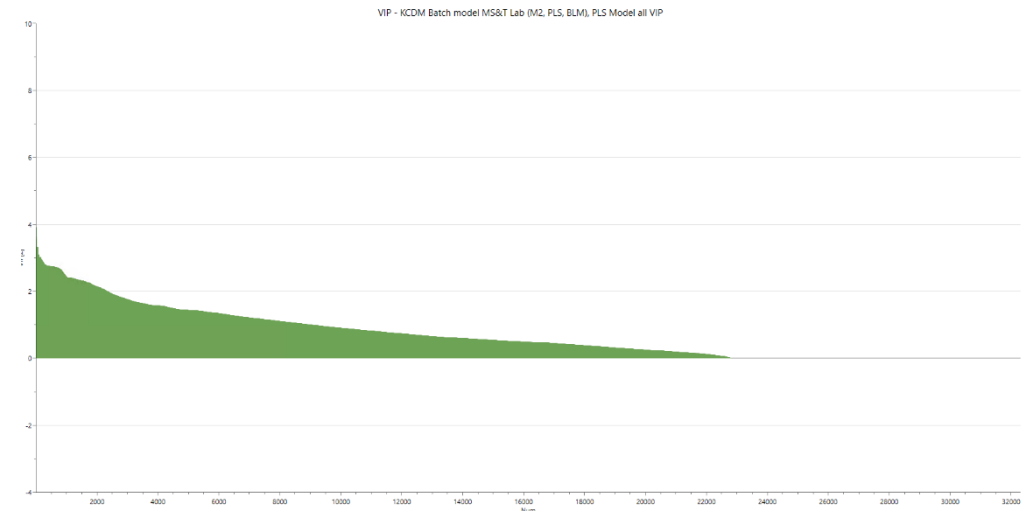


Figure 15: VIP plot (all values)

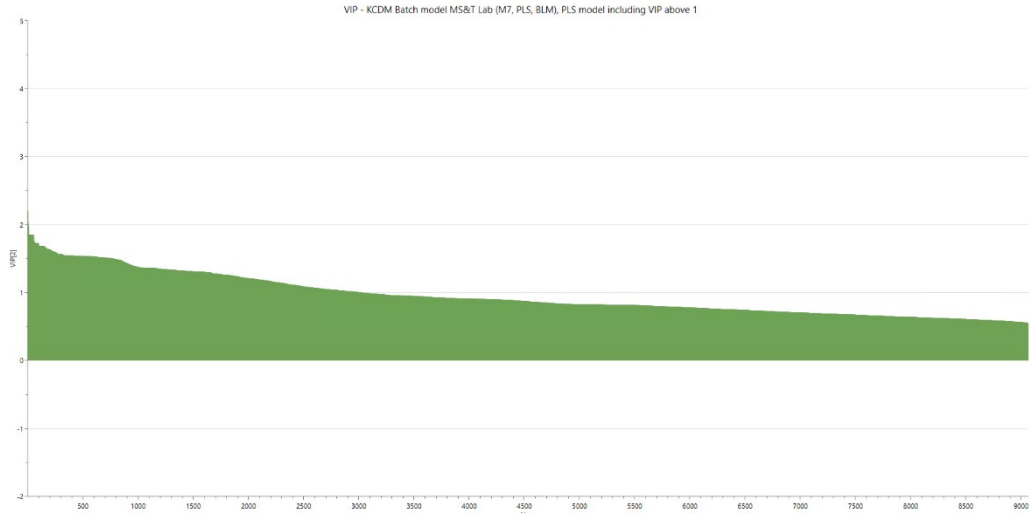


Figure 16: VIP plot – VIP above 1

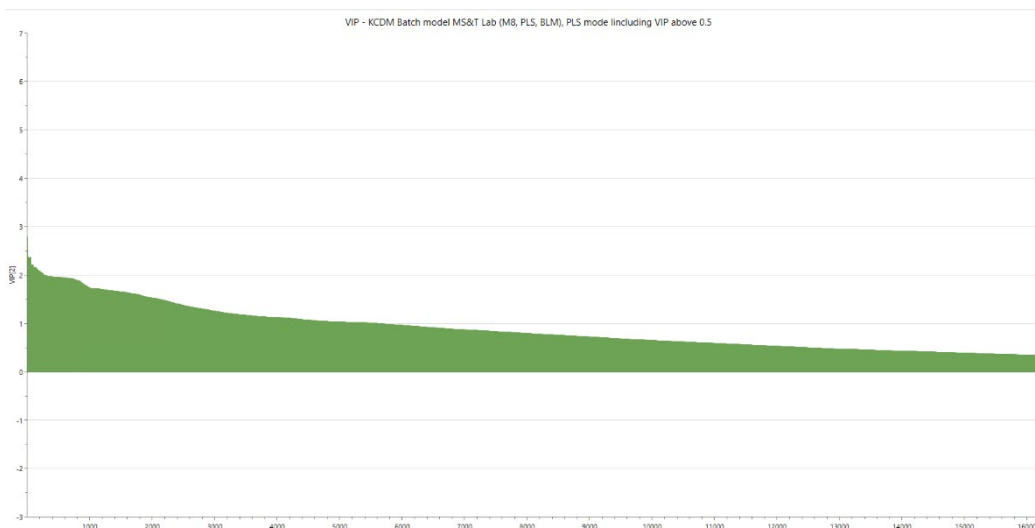


Figure 17: VIP plot – VIP above 0.5

8 Evaluation of models

For each of the three studies, the PLS algorithm was employed to assess the model. The observed and projected values of VCD for each batch are displayed below. The model predicted (X axis) the VCD, while the observed value (Y axis) is the one imported from Raman spectroscopy. With a good model all the points will fall close to the 45-degree line. The fit of the observations to the model is indicated by the RMSEE in the footer. The similar metric, the RMSECV, is calculated through the cross-validation process and indicates how predictable the model is.

Observed vs predicted plot can be shown by selecting the model and clicking at “Obs. Vs. pred.” icon on the menu bar highlighted in yellow.

Number	Model	Type	A	N	R2X(cum)	R2Y(cum)	Q2(cum)	Date	Title	Hierarchical
Batch model 1										
Batch evolution models (BEM)										
1	M1:M1	PLS-Class(M1)	2	26732	0.276	0.391	0.3	9/26/2023		
Batch level models (BLM)										
2	M2	PLS	2	28	0.164	0.962	0.199	11/4/2023	PLS Model all VIP	
7	M7	PLS	2	28	0.286	0.963	0.614	11/4/2023	PLS model including VIP above 1	
8	M8	PLS	2	28	0.2	0.968	0.457	11/4/2023	PLS mode including VIP above 0.5	

Figure 18: Obs. vs pred. plot

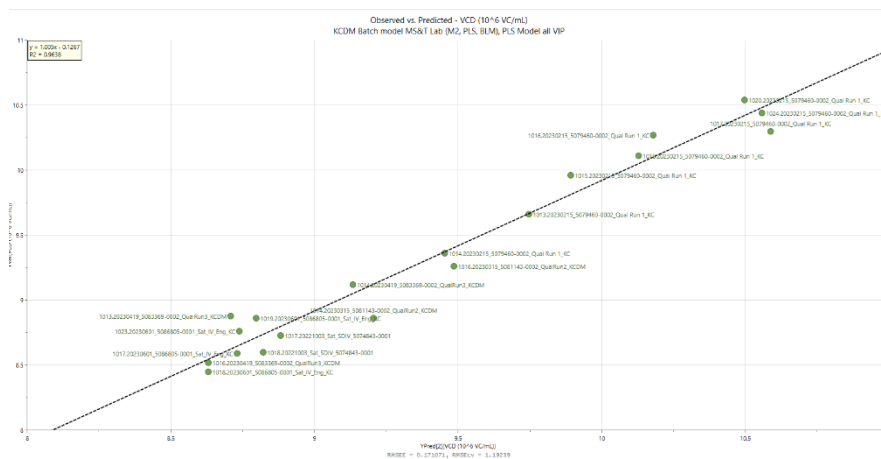


Figure 19: Observed vs predicted plot – all VIP model

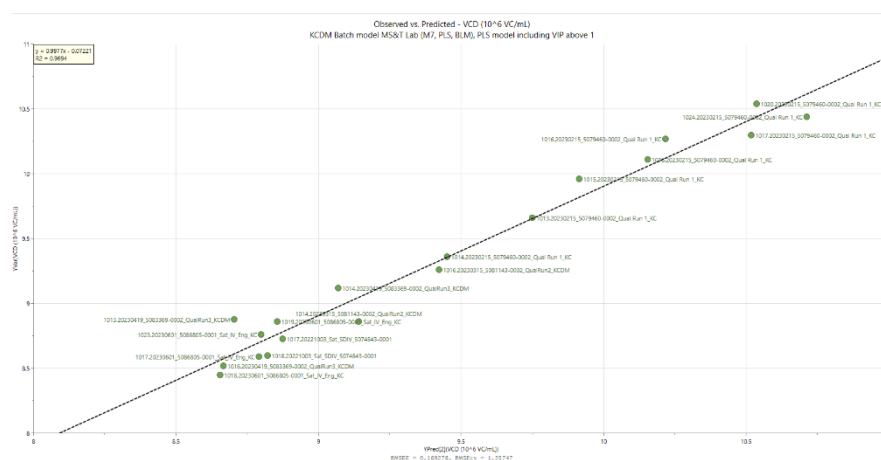


Figure 20: Observed vs predicted plot –VIP above 1 model

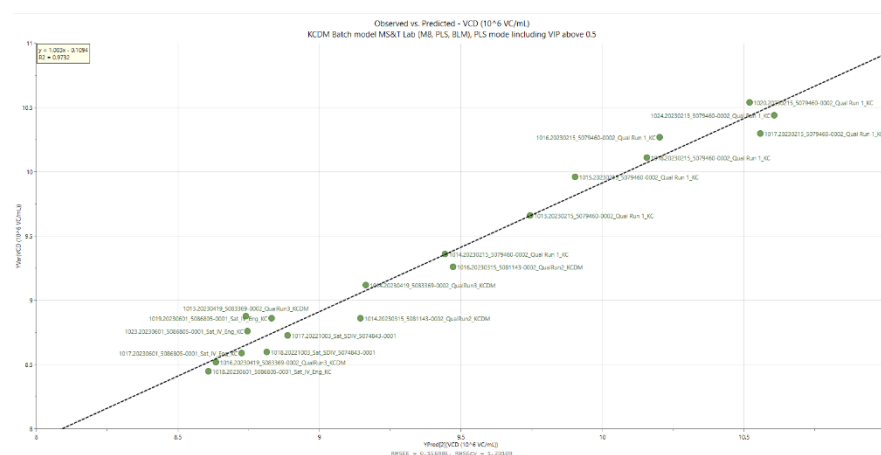


Figure 21: Observed vs predicted plot –VIP above 0.5 model

One batch is used as a “test” batch and predicted using predict functionality within SIMCA. Batch is specified as the training batch and result can be seen in below figure.

1	2	3	4	5	6	7	8	9
Primary ID	SBatchID	Observation	Number	VCD (10 ⁶ VC/mL)	YPredPS2[VCD (10 ⁶ VC/mL)]	Set	PModXPS+[2]	DModXPS+[2](Norm), Weighted
1015.20230419_5083369-0002_QualRun3_KCDM	1015.20230419_5083369-0002_QualRun3_KCDM	4/20/2023 10:30:00 AM	10	9.09	12.4724	TS	0	

Figure 22: Test model prediction