Using JMP® to Build Soft Sensors for Efficient Monitoring of Chemical Production Processes

Stephen Pearson, Principal Data Scientist, Syngenta UK Ltd, Jealott's Hill Research Station, Bracknell, RG42 6EY.

Abstract

Production processes are routinely sampled to determine the concentration of key components to meet safety, environmental, or quality criteria. The deployment of temperature-compensated density meters provides an opportunity for live process monitoring to replace offline sampling. Historic process data is not suitable for model building as offline analysis occurs sporadically, with uncertainty about the exact time of sampling and with limited variability in results.

A naïve approach of varying the temperature and concentration before measuring the density (in the laboratory) leads to inflated errors (since concentration is the desired prediction). The method has merits since it only involves solvent addition and temperature control, which can be automated. We show how this model can be used as a first pass, to target evenly spaced temperatures and densities, followed by sampling to determine the concentration, to produce a model with much lower prediction uncertainties.

Exporting the model to Seeq and PowerBI enabled continuous monitoring and decreased costs from daily sampling by €15,000 per year (for a single process). The implementation removed delays waiting on the offline analysis, reduced the risk of operator exposure to process chemicals, and enabled the production team to predict and plan interventions, thus increasing operational time.

Introduction

The widespread deployment of temperature-compensated density meters in production processes makes them an ideal target for the creation of soft sensors. Continued improvements in technology have made it possible to install the same model of sensor in the laboratory and the process.

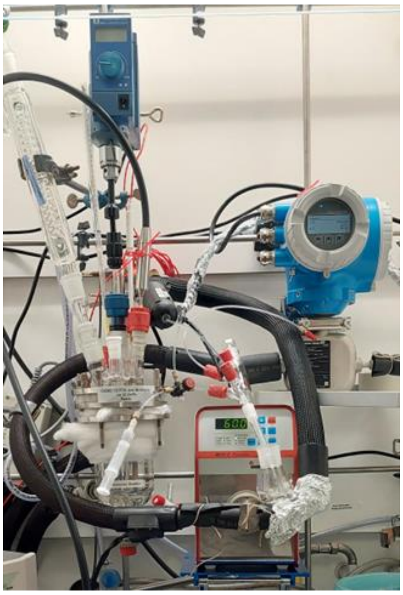


Figure 1: Experimental setup used for model development.

Creation of such soft sensors is a multi-step process that requires:

1. Screening the density meter response over the expected ranges of concentration and temperature.
2. Modifying the density in evenly spaced steps by addition/dilution of the analyte, waiting for temperature to stabilise and then sampling the mixture.
3. Building a model where density and temperature are the input variables, with the analyte concentration as the response.
4. Taking samples from the production process at different analyte concentrations and recording the density and temperature for the time the sample was taken.
5. Validate or recalibrate the model from step 3 to account for differences between the lab and production process.
6. Monitoring of the soft sensor to check for drift between the process sensors and the model.

A key consideration when building a model, is what factors will be inputs and what are the responses. The seemingly quick method of making up solutions of the analyte in the laboratory, varying the temperature and recording the density will not lead to a suitable model. The model inputs are concentration and temperature, and the output is density. A simple equation for this is shown below.

$$density = x_1 \times temperature + x_2 \times concentration + x_3 \times temperature \times concentration$$

Equation 1

In the production process the inputs are temperature and density, the desired output is concentration. A simple equation for this is shown below.

$$concentration = x_4 \times temperature + x_5 \times density + x_6 \times temperature \times density$$

Equation 2

A model built on Equation 1 has the wrong error structure to predict concentration. If there are any interaction effects between temperature and concentration (or temperature and density) it is not possible to rearrange equation 1 to the desired equation 2. Furthermore, any errors in determining the concentration will be correlated with temperature, and a lack of randomisation adds additional correlation between the ability to set the temperature and the recorded density.

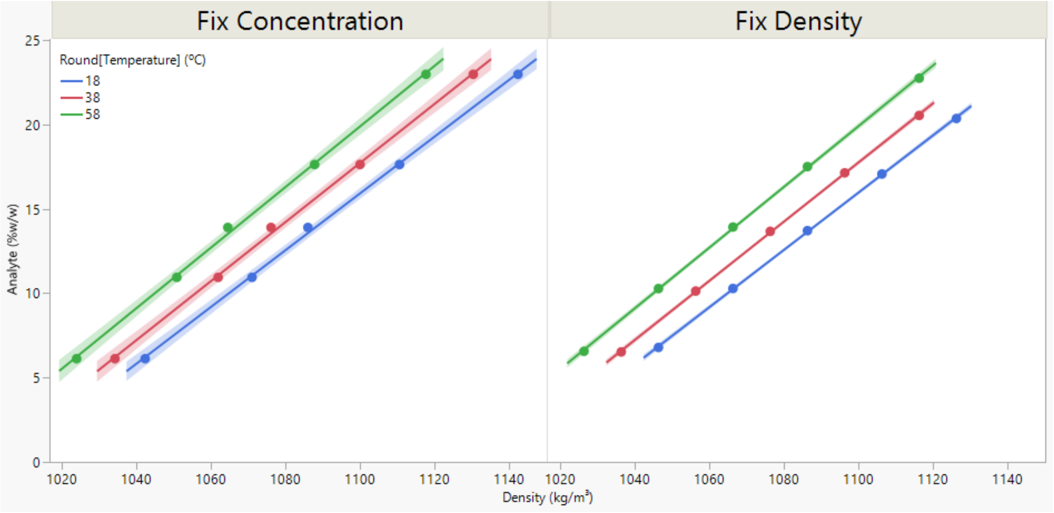


Figure 2: Visualisation of data from the 'quick' approach of using fixed concentrations, versus stepping the density evenly. In the 'quick' approach there are only 5 unique concentrations on which to build the predictions.

Screening the density response

For each new system evaluated, it will not be known how the density varies in response to changes in the temperature and concentration. A preliminary screening is required to estimate these values.

A simple way to do this is systematic addition of either solvent or analyte aliquots, recording the density and calculating the concentration. This data may already be available in the literature for common materials. As a minimum around 9 measurements are needed:

Concentration	Temperature
Low	High
Low	Med
Low	Low
Med	Low
Med	Med
Med	High
High	High
High	Med
High	Low

Table 1 - Basic screening approach

This could be achieved by making up just one solution (or by switching between three). The aim of this step is to check for evidence of non-linear relationships between concentration/temperature with density, and to determine the expected response limits of the model.

This data can also be used to develop a crude model which can help inform the step size in density and starting concentrations. If a model is to be developed it would be better to step the temperature and concentration in 5 steps rather than 3, if time and resource allow.

Stepping the density

A design of experiments approach could be used to build a model using fewer data points than the method below but requires a greater cognitive load than "increase the density in equal steps at each chosen temperature."

Using the preliminary imformation a specific set of densities should be targeted at each temperature. For best results measure the concentration at four or five different densities at each temperature, for five temperatures overall (if there is evidence of curvature, 7 steps in that variable would be better). The steps in density and temperature should be evenly spaced, and it is not a problem if the concentrations end up slightly outside the range expected in the production process.

Be very cautious about including zero concentration in the model. Pure systems can behave wildly different to ones containing only 0.25% analyte, effectively creating a discontinuity.

In an ideal world each of these 20 to 49 solutions would be made up independently and the equipment emptied and returned to a neutral state between each measurement. The order the data is collected in would also be randomised. If during the experiment the solvent or analyte were to become contaminated in a way that would alter the measurements, the randomisation would prevent this noise being incorporated into any models.

Practically this would generate a significant amount of waste and incur a significant time penalty. It is also unclear how much benefit this would provide, given the accuracy is limited by the external technique used to determine the analyte concentration.

A compromise is a blocking design where a solution is fixed at a temperature, and then analyte (or solvent) is added until the target density is achieved. This solution could be reused to step through the densities, provided samples are taken for external analysis of the concentration (and the system contains only two components). Pragmatically this sampling approach can be considered to make the solutions independent, since any error in determining the concentration of analyte is external to the data generation process.

If there is limited material or the measurement system is enclosed rather than in a sampling loop and the analyte is a solid, it can be hard to step through the densities. An alternative blocking approach is shown in appendix 2.

Provided the analyte is stable, all the samples from a day can be analysed in a batch with the order of sample preparation and analysis being randomised (with duplicate analysis for each sample to account for gross errors).

The results from an example study are shown in Figure 3. In this case, steps of two density units were targeted across each of the five temperatures. In total 6 solutions were used to complete the work (1 screening + 5 different temperatures).

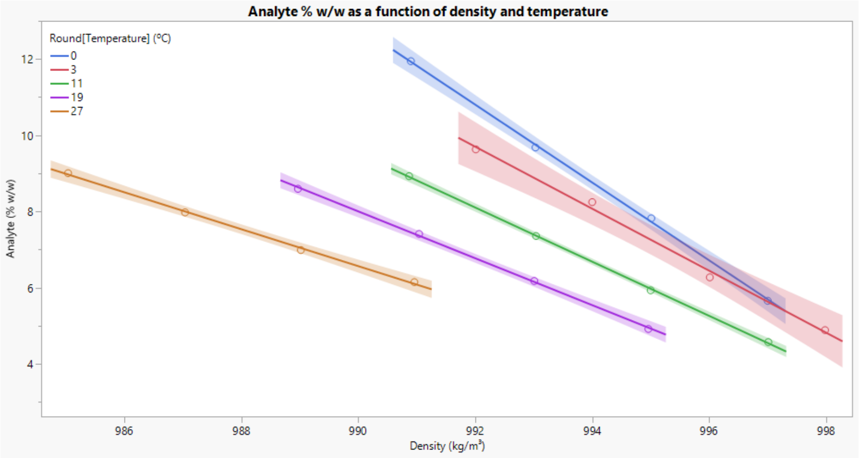


Figure 3: Results from systematic data collection. The data is spaced evenly in temperature and density which leads to uneven spacing in Analyte % w/w. This is the desired outcome as Analyte % w/w is the response to be modelled. The temperature is shown here rounded to the nearest degree to clarify the groupings (data available in appendix 1).

Building the preliminary model

After fitting a model to the results from the lab experiments the prediction uncertainty can be determined. Is it small enough to be practically useful? In the above example the prediction uncertainty was typically ± 0.3 % w/w (95 % individual prediction intervals) and increased slightly at higher temperatures.

It is recommended a penalised technique such as Lasso regression is used to determine if potential quadratic and interaction terms are needed to accurately model the data. If you have access to JMP Pro, a Self-Validating Ensemble Model (SVEM) can help produce a model that has good predictive properties without overfitting to the lab data (JSL code in appendix 3).

Process samples

If the laboratory model has a small enough prediction error to be useful, the next step is to compare it to the production data. Ideally, the number of samples from the process would be at least 50% of the number of observations used to build the model. It is important the samples cover a range of concentrations in the analyte and not just the extremes.

The preliminary model can be applied to the live process data to help target when samples should be taken. *Accurate records of when samples were taken can be critical to obtaining the corresponding temperature and density data from the process historian.*

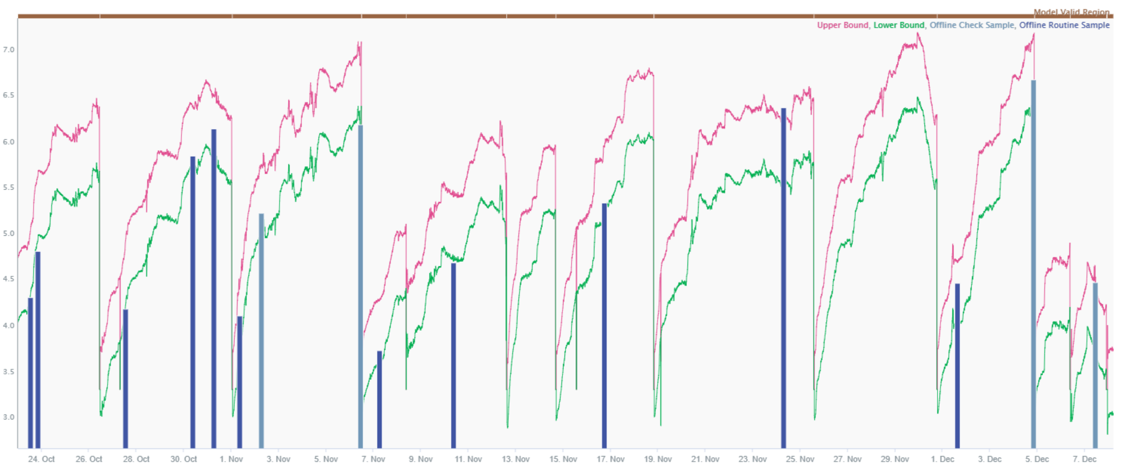


Figure 4: The model was implemented in Seeq workbench to allow for live monitoring of the process. This allowed for targeted collection of validation samples from a range of predicted concentrations. By naming the variable the same in both software, a simple find and replace operation (: to $) can convert a JSL equation to the corresponding Seeq formula.

Recalibrating the model or a paired t-test

Ideally the model was built using chemicals taken from the production process. This ensures it is based on the same analytes as the real system. In this case a paired t-test can be used to validate the model predictions with process samples.

If a ideal system had to be used instead there will be some discrepancy with the process results. In this case the model building process can be re-run, with an additional column to indicate if the temperature, density, concentration triplet originates from the laboratory or process.

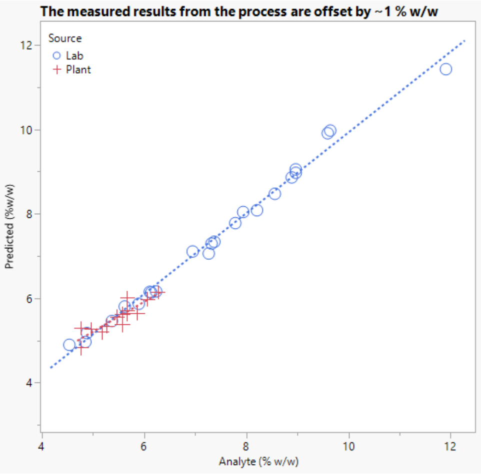


Figure 5: Actual by Predicted Plot. The open circles are laboratory results, whilst the crosses represent process data.

In the example shown in Figure 5, the difference between the laboratory and process could be accounted for with a simple extra term (-1 % w/w for process data). There may be additional interaction terms required. Including the process data in the model caused the individual prediction uncertainty to increase slightly to ± 0.35 % w/w.

Monitoring the model

The model is now ready to be put into production use. *There will be drift in agreement between the model and sampling over time, so it is important that periodic checks are made. Density meters often become coated causing a step change in recorded response.*

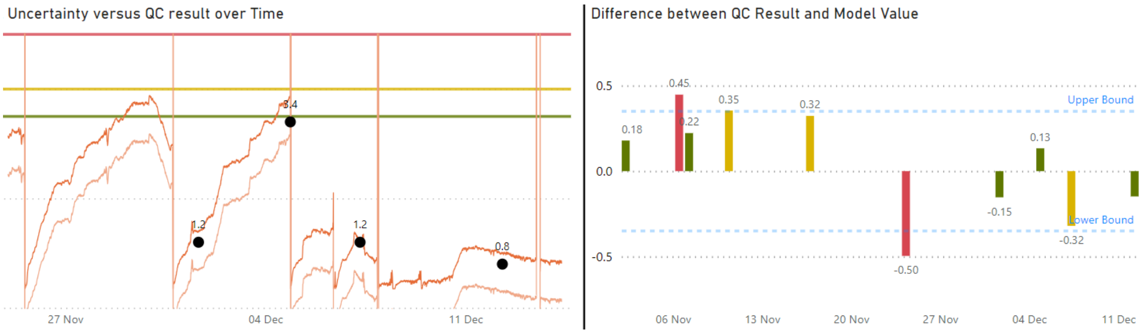
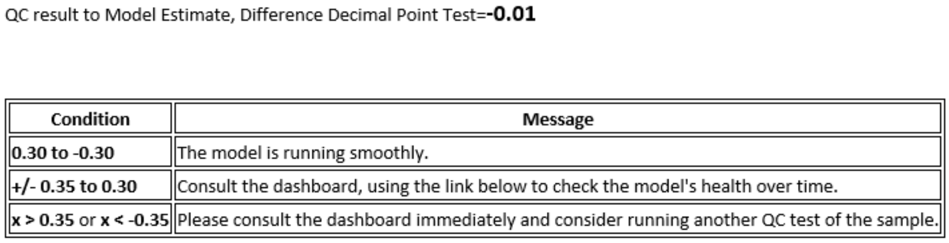


Figure 6: PowerBI dashboard applying the model to live process data and comparing it to the results from routine calibration checks.

If the offline cross-check data is published to a database, PowerBI can be used to email a relevant distribution list each time a sample is analysed, advising if the model prediction and sample result agree.



QC result to Model Estimate, Difference Decimal Point Test=**-0.01**

Condition	Message
0.30 to -0.30	The model is running smoothly.
+/- 0.35 to 0.30	Consult the dashboard, using the link below to check the model's health over time.
x > 0.35 or x < -0.35	Please consult the dashboard immediately and consider running another QC test of the sample.

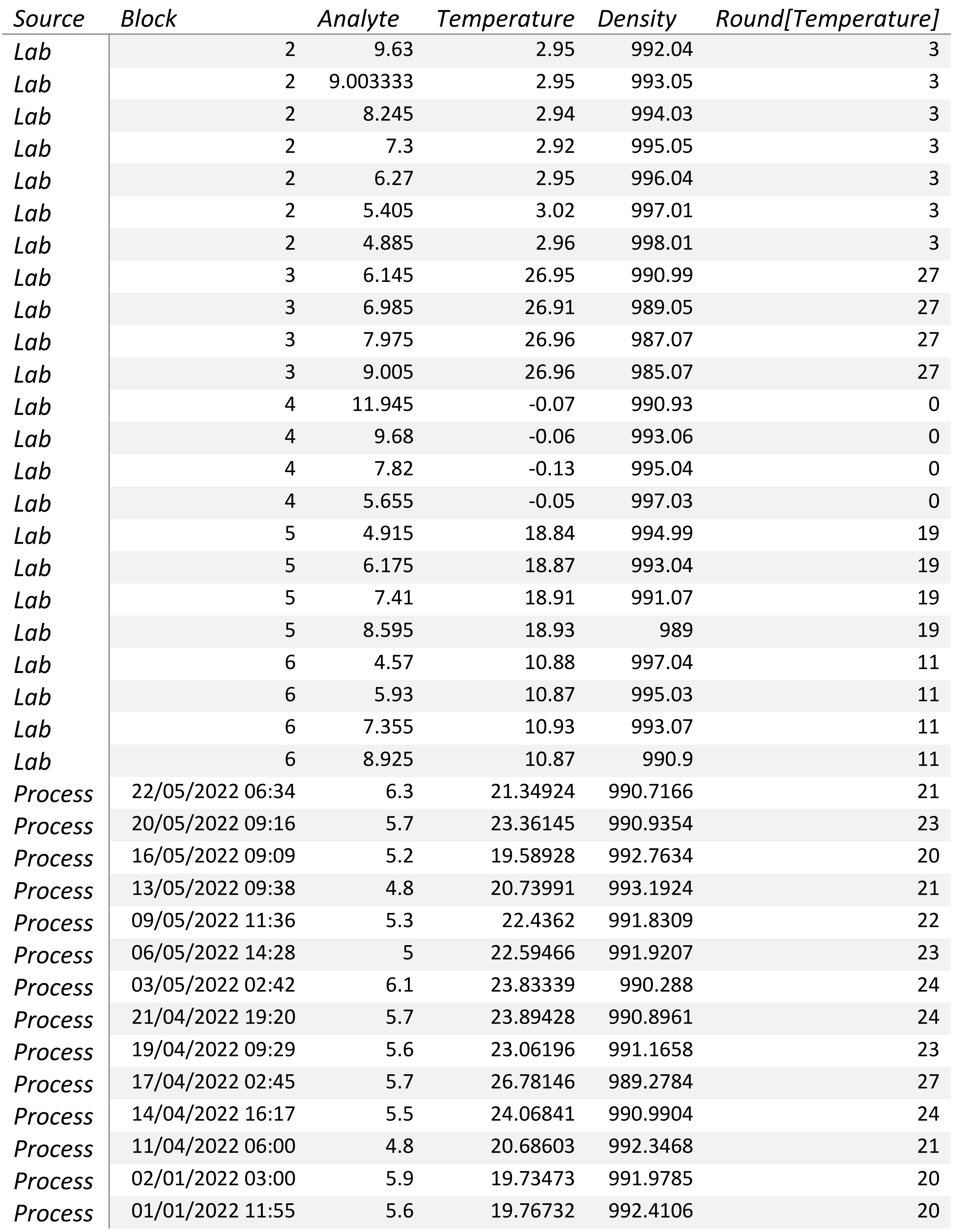
Figure 7: Example of automated report email.

Conclusion

Soft sensor models permit continuous process monitoring and decrease costs from daily sampling by €15,000 per year (for a single process). The implementation removes delays waiting on the offline analysis, reduces the risk of operator exposure to process chemicals, and enables the production team to predict and plan interventions, thus increasing operational time.

This paper has discussed pragmatic compromises for collecting data in a timely and simple manner whilst still obtaining a model that is useful. This approach has been successfully applied for varied chemical processes in multiple countries.

Appendix 1 – Sample loop and process data



Source	*Block*	*Analyte*	*Temperature*	*Density*	*Round[Temperature]*
Lab	2	9.63	2.95	992.04	3
Lab	2	9.003333	2.95	993.05	3
Lab	2	8.245	2.94	994.03	3
Lab	2	7.3	2.92	995.05	3
Lab	2	6.27	2.95	996.04	3
Lab	2	5.405	3.02	997.01	3
Lab	2	4.885	2.96	998.01	3
Lab	3	6.145	26.95	990.99	27
Lab	3	6.985	26.91	989.05	27
Lab	3	7.975	26.96	987.07	27
Lab	3	9.005	26.96	985.07	27
Lab	4	11.945	-0.07	990.93	0
Lab	4	9.68	-0.06	993.06	0
Lab	4	7.82	-0.13	995.04	0
Lab	4	5.655	-0.05	997.03	0
Lab	5	4.915	18.84	994.99	19
Lab	5	6.175	18.87	993.04	19
Lab	5	7.41	18.91	991.07	19
Lab	5	8.595	18.93	989	19
Lab	6	4.57	10.88	997.04	11
Lab	6	5.93	10.87	995.03	11
Lab	6	7.355	10.93	993.07	11
Lab	6	8.925	10.87	990.9	11
Process	22/05/2022 06:34	6.3	21.34924	990.7166	21
Process	20/05/2022 09:16	5.7	23.36145	990.9354	23
Process	16/05/2022 09:09	5.2	19.58928	992.7634	20
Process	13/05/2022 09:38	4.8	20.73991	993.1924	21
Process	09/05/2022 11:36	5.3	22.4362	991.8309	22
Process	06/05/2022 14:28	5	22.59466	991.9207	23
Process	03/05/2022 02:42	6.1	23.83339	990.288	24
Process	21/04/2022 19:20	5.7	23.89428	990.8961	24
Process	19/04/2022 09:29	5.6	23.06196	991.1658	23
Process	17/04/2022 02:45	5.7	26.78146	989.2784	27
Process	14/04/2022 16:17	5.5	24.06841	990.9904	24
Process	11/04/2022 06:00	4.8	20.68603	992.3468	21
Process	02/01/2022 03:00	5.9	19.73473	991.9785	20
Process	01/01/2022 11:55	5.6	19.76732	992.4106	20

Table 2: Example data from the laboratory used to build a model and the corresponding process data

Appendix 2 – Enclosed device/hard to change factors

ID	Group	Analyte	Temperature	Density	Round[Temperature]	Pred Analyte
1	Fix Concentration	6.13	17.9	1043	18	
2	Fix Concentration	6.13	37.9	1034.9	38	
3	Fix Concentration	6.13	57.9	1024.6	58	
4	Fix Concentration	10.95	17.9	1071.7	18	
5	Fix Concentration	10.95	37.9	1062.6	38	
6	Fix Concentration	10.95	57.9	1051.5	58	
7	Fix Concentration	13.91	17.9	1086.7	18	
8	Fix Concentration	13.91	37.9	1076.8	38	
9	Fix Concentration	13.91	57.9	1065.2	58	
10	Fix Concentration	17.64	17.9	1111.3	18	
11	Fix Concentration	17.64	37.9	1100.7	38	
12	Fix Concentration	17.64	57.9	1088.5	58	
13	Fix Concentration	22.99	17.9	1143.1	18	
14	Fix Concentration	22.99	37.9	1131.1	38	
15	Fix Concentration	22.99	57.9	1118.4	58	
16	Fix Density		17.9	1047	18	6.8
17	Fix Density		17.9	1067	18	10.3
18	Fix Density		17.9	1087	18	13.7
19	Fix Density		17.9	1107	18	17.1
20	Fix Density		17.9	1127	18	20.4
21	Fix Density		27.9	1037	28	5.7
22	Fix Density		27.9	1057	28	9.3
23	Fix Density		27.9	1077	28	12.8
24	Fix Density		27.9	1097	28	16.2
25	Fix Density		27.9	1117	28	19.6
26	Fix Density		37.9	1037	38	6.5
27	Fix Density		37.9	1057	38	10.1
28	Fix Density		37.9	1077	38	13.7
29	Fix Density		37.9	1097	38	17.1
30	Fix Density		37.9	1117	38	20.6
31	Fix Density		47.9	1027	48	5.6
32	Fix Density		47.9	1047	48	9.3
33	Fix Density		47.9	1067	48	12.9
34	Fix Density		47.9	1087	48	16.4
35	Fix Density		47.9	1107	48	19.9
36	Fix Density		57.9	1027	58	6.6
37	Fix Density		57.9	1047	58	10.3
38	Fix Density		57.9	1067	58	13.9
39	Fix Density		57.9	1087	58	17.5
40	Fix Density		57.9	1117	58	22.8

Table 3:Example Data From the screening process, with corresponding prediction model used to identify steps in temperature and density.

In situations with limited raw materials or difficulty adjusting the density within the experimental setup, the number of solutions to be made up could be reduced by grouping concentrations that differ by only 0.1 % w/w. In the crude prediction example from Table 3 this would reduce the total number of solutions required from 25 to 13. For example, the density for a predicted concentration of 6.6 % w/w would be measured at 20, 40 and 60 °C (IDs 16, 26 & 36).

Concentrations to study
5.7
6.6
9.3
10.2
12.9
13.7
16.4
17
17.3
19.8
20.3
20.7
22.6

Table 4: Reduced number of solutions from table above by grouping similar values to accommodate hard to change concentration.

Appendix 3 – JSL script for SVEM approach

The Self Validating Ensemble Model approach was utilised to fit 100 models using the JMP Scripting Language [JSL]. The final model is then the average of these.

```
// Part I - Set up data table
dt = current data table();
nrow = n rows(dt);  // number of rows in dataset
// Duplicate data
dt_copy = dt << Subset( All rows, Selected columns only( 0 ) );
dt << Concatenate(
        dt_copy,
        "Append to first table",
        "Keep Formulas"
);
Close(dt_copy, No Save);
// Validation column
dt << New Column( "Validation",
        "Numeric",
        "Nominal",
        <<Set Each Value( 0 )
);
wait (1);
For Each Row(dt, If(row()>nrow, :Validation = 1));
// null and FW columns
dt << New Column( "Null Factor", "Numeric","Continuous");
dt << New Column( "Fractional Weight", "Numeric", Formula(If( Row() == 1,
        _u = J( 1, N Row() / 2, Random Uniform() )
);
If( Row() <= N Row() / 2,
        :Null Factor[Row()] = Random Normal();
        Gamma Quantile( _u[Row()], 1, 1 );
,
        :Null Factor[Row()] = :Null Factor[Row() - N Row() / 2];
```

```
    Gamma Quantile( 1 - _u[Row() - N Row() / 2], 1, 1 );
)));
// Part 2 - Launch dialogue
wait (1);
for(i=1, i<=100, i++,
dt = Current Data Table();
fm = dt << Fit Model(
    Freq( :Fractional Weight ),
    Validation( :Validation ),
    Y( :Analyte ),
    Effects(
        :Temperature, :Density, :Temperature * :Density, :Temperature * :Temperature,
        :Density * :Density, :Temperature * :Temperature * :Temperature,
        :Density * :Density * :Density
    ),
    Personality( "Generalized Regression" ),
    Generalized Distribution( "Gamma" ),
    Run( Fit( Estimation Method( Lasso ), Validation Method( Validation Column ) ) )
);
fm << (fit[1] << Save Prediction Formula);
fm << Close Window();
dt << Rerun Formulas;);
```