Predictive Modelling of Silver and Thickness Measurements in Medical Foam Production

Interpretable Machine Learning for Process Understanding

Master's thesis in Engineering Mathematics and Computational Science

VIGGO WESSNER

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Predictive Modelling of Silver and Thickness Measurements in Medical Foam Production

Interpretable Machine Learning for Process Understanding

Viggo Wessner

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2026

Predictive Modelling of Silver and Thickness Measurements in Medical Foam Production
Interpretable Machine Learning for Process Understanding
Viggo Wessner

Supervisor: Holger Rootzén, Department of Mathematical Sciences
External Supervisor: Per Granath, Mölnlycke Health Care
Examiner: Aila Särkkä, Department of Mathematical Sciences

Master's Thesis 2026
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Predictive Modelling of Silver and Thickness Measurements in Medical Foam Production
Interpretable Machine Learning for Process Understanding
Viggo Wessner
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Medical foam used in wound-care products must meet quality requirements for properties such as thickness and silver content. This thesis investigates whether routinely collected production data can be used to predict these properties and to support process understanding. The study uses six months of data from medical foam production, including quality-control and laboratory measurements, roll information, and process sensor data from the production equipment.

The raw data sources were cleaned, linked to finished production rolls or roll positions, and converted into modelling tables for four responses: silver concentration, Ag [mg/g]; silver area-density, Ag [mg/cm^2]; average roll thickness; and thickness log variance, which describes variation in thickness within a roll. Several regression models were compared using time-based cross-validation and a held-out later production period, so that model performance was evaluated in a setting similar to predicting future production.

The results depended strongly on the response. Silver concentration was not usefully predictable from the available data, with test $R^2 = 0.005$. Silver area-density showed some predictive signal during the earlier development period, but the models did not generalise to the later test period because silver levels shifted upward. In contrast, average thickness was predicted very accurately from process sensor summaries alone, with the best model reaching test $R^2 = 0.989$. Thickness log variance was harder to predict but still showed useful signal, with the best model reaching test $R^2 = 0.652$. The most stable interpretation results therefore concern thickness. The fitted models identified production contexts and sensor summaries that can guide expert process review, especially for thickness level and foam uniformity. These results should be read as predictive associations, not as causal evidence or direct control rules.

Keywords: time series data, quality control, predictive modelling, model interpretability, regression, time-based validation.

Acknowledgements

I would like to thank my supervisors, Holger Rootzén and Per Granath, for their guidance, feedback, and support throughout the project. I am also grateful to Jason, Peter, and Magnus at the production plant for their practical insights and for helping make the production and process data understandable for a fool like me. Finally, I would like to thank my partner, Hedda, for her patience, encouragement, and sweet treats during the time of this work.

Viggo Wessner, Göteborg, June 2026

List of Acronyms

ALE	Accumulated Local Effects
CV	Cross-validation
EBM	Explainable Boosting Machine
FL1	Flow Line 1
FL2	Flow Line 2
GAM	Generalised Additive Model
MAE	Mean Absolute Error
PCR	Principal Component Regression
PLS	Partial Least Squares
QC	Quality Control
RMSE	Root Mean Squared Error
SCADA	Supervisory Control and Data Acquisition
SHAP	SHapley Additive exPlanations
XGBoost	Extreme Gradient Boosting

Nomenclature

Indices and dimensions

i	Observation index
j	Predictor index
K	Number of cross-validation time blocks
n	Number of observations
p	Number of predictors

Sets

$\mathcal{D}$	Regression dataset
T	Training index set
E	Evaluation index set for a validation fold or held-out test set

Parameters

β_0	Intercept term
β_j	Regression coefficient for predictor j
$\boldsymbol{\beta}$	Vector of regression coefficients
α	Ridge regularisation strength

Variables

f	Fitted prediction function
f_j	Additive shape function for predictor j
$\mathbf{x}_i$	Predictor vector for observation i

x_{ij}	Value of predictor j for observation i
z_{ij}	Min–max-scaled value of predictor j for observation i
a_j, b_j	Training-fold minimum and maximum of predictor j
y	Response variable for the active prediction task
y_i	Observed response for observation i
$\hat{y}_i$	Predicted response for observation i
$\bar{y}_E$	Mean of the observed response values in evaluation set E
ε_i	Error term for observation i
Ag	Chemical symbol for silver
Ag [mg/g]	Silver concentration response
Ag [mg/cm^2]	Silver area-density response
$T_1, \ldots, T_6$	Six QC thickness profile measurements for a roll
$\bar{T}$	Mean of the six thickness profile measurements
y_{mean}	Thickness-mean response
$y_{\text{log-var}}$	Log-transformed thickness-variance response
V_T	Thickness profile variance before log transformation
$\widehat{V}_T$	Predicted thickness profile variance after back-transformation

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background and motivation

Medical foam is used in wound-care products where properties such as thickness, absorbency, softness, and antimicrobial performance affect product quality. In this thesis, the studied material is medical foam produced at Mölnlycke Health Care. The foam is used in wound-care dressings, including silver-containing antimicrobial foam dressings for wounds that are infected or at risk of infection. Stable production is therefore important both for manufacturing and for ensuring that the final products perform as intended.

Better process understanding can have practical value. If important quality properties can be predicted from routinely collected process data, this may help identify problematic production conditions earlier, reduce waste and rework, and show which parts of the process are linked to variations in the finished material. In a medical-device production setting, this is especially important because stable and well-understood processes support product quality, traceability, and consistency. Ultimately, this helps ensure that patients receive reliable wound-care products.

The practical question studied in this thesis is whether routinely collected production data, in this case data from six months of production, can be used to predict important foam properties before or during quality review. The thesis therefore combines predictive modelling with model interpretation to support both prediction and process understanding.

1.2 Production and data context

The production of the medical foam studied in this thesis can be simplified into the stages shown in Figure 1.1. Raw materials first enter a reactor, where they are combined to form the material that will later become finished foam. The material then continues through one of two parallel flow lines, where it undergoes several processing steps. These steps cannot be described explicitly here due to confidentiality, but they may include operations such as mixing, heating, pressing, or similar. After passing through the flow line, the material has been processed into a continuous foam sheet. This sheet is then cut and rolled up into what is called a *roll*, which is the finished production unit. Each roll is then reviewed before it can be released for use in wound-care products.

Mölnlycke produces several foam products with different material properties. These products follow the same general production stages, but may use different processing settings. This means that both the product type and the process conditions are relevant when studying variation in the finished foam.

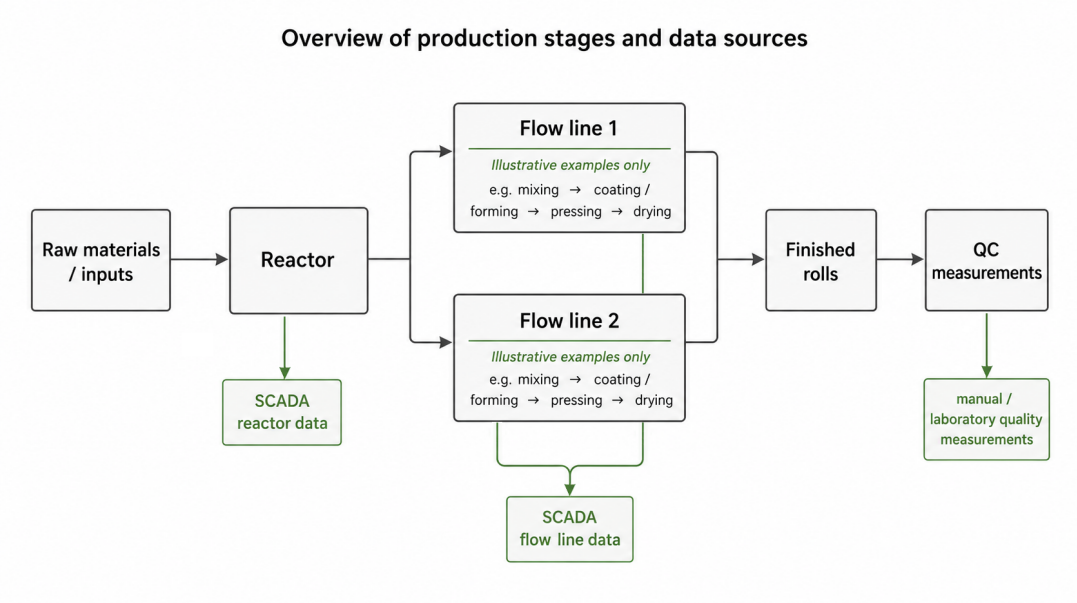


Figure 1.1: Overview of the production stages and data sources used in the thesis. Reactor and flow line SCADA data describe process behaviour before and during roll production, while QC and laboratory measurements describe properties of the finished rolls.

Several types of data were collected during the six months of production studied. This includes metadata about the rolls, such as when they were produced, which of the two flow lines they were processed on, and which product group they belong to. Thus, giving the context of the production for each roll. Quality-control (QC) measurements describe properties of the finished material, such as thickness, density, and related roll characteristics. In the available data, thickness measurements exist for several foam products. Silver measurements, however, are only available for one silver-containing product. In addition, time-stamped process measurements are collected from Supervisory Control and Data Acquisition (SCADA) systems. These measurements describe process behaviour during production and may include variables such as temperature, pressure, flow rates, or similar. The SCADA data are recorded approximately every 10 seconds and are stored separately for the reactor and for the each of the two flow lines.

The data sources therefore differ both in where and when they are collected in the production process, as illustrated in Figure 1.1, and in how they are represented statistically. Reactor SCADA data describe upstream conditions before the material enters the flow lines, whereas flow line SCADA data describe processing conditions during roll production. QC and laboratory measurements instead describe properties of the finished rolls. Table 1.1 summarizes these data sources and shows how they are used in the modelling workflow.

Table 1.1: Summary of the main data sources and their representation in the thesis.

Data source	Raw data	Model input variables	Naming used in thesis
Line (Context) data	Roll metadata, product and line information, time of production	Production line, product number, production start and end	`LineID`, `ProductNumber`, `ProductionStart`, `ProductionEnd`
QC data	Laboratory QC measurements for each roll	Thickness, density, weight, retention, etc	`Left1/2`, `Mid1/2`, `Right1/2`, `Density`, `Weight`, ...
Silver titration data	Laboratory silver measurements at roll positions A, B, and C	Measurements of silver proportion and silver density at specific roll positions	`RollPosition`, Ag [mg/g], Ag [mg/cm^2]
Flow line 1 SCADA	Time series of process variables from flow line 1	Summary statistics of all flow line 1 process variables during the time period when the roll was produced	`FL1__..._A-*__mean`
Flow line 2 SCADA	Time series of process variables from flow line 2	Summary statistics of all flow line 2 process variables during the time period when the roll was produced	`FL2__..._B-*__mean`
Reactor SCADA	Time series of process variables from the reactor	Summary statistics of all reactor process variables during the time period when the material used to create the roll was in the reactor	`Reactor__..._C-*__mean`

The table also shows some naming conventions which will be defined throughout Chapter 3. The SCADA variables are anonymised, meaning that the original process variable names have been replaced by a generic letter combination. These anonymised names are referred to as tags, or process tags throughout the thesis. The tags are always of the form `A-*`, `B-*`, or `C-*`, where the first letter represents which data source they belong to – `A`, `B`, and `C` refer to flow line 1, flow line 2, and reactor SCADA data, respectively – and the asterisk is replaced by some letter which identifies a specific process variable from the respective data source. This protects sensitive information about the production process, but it also makes model interpretation more difficult.

1.3 Response variables

The thesis studies four response variables, meaning four quantities that the models try to predict, which were chosen together with the process experts at Mölnlycke. Two of these responses describe silver properties of the finished foam, while the other two describe thickness properties.

The two silver responses are Ag [mg/g], the silver concentration, and Ag [mg/cm^2], the silver area-density. The first response can be thought of as measuring the proportion of silver in the foam while the second one describes the distribution of silver across the roll. These measurements are based on laboratory samples from the A, B, and C positions of each roll, as shown in Figure 1.2. The figure also shows an important structural detail of the production process: rolls produced in the same batch come from the same continuous foam sheet, which is cut into separate rolls.

The thickness responses are the thickness mean and the thickness log variance, which are based on thickness measurements taken at six prespecified points of each roll. The thickness mean describes the average thickness level of the roll, while

the thickness log variance describes variation in thickness across the roll. Exact definitions of the responses, including the reason for the logarithmic transformation, are given in Section 3.1.

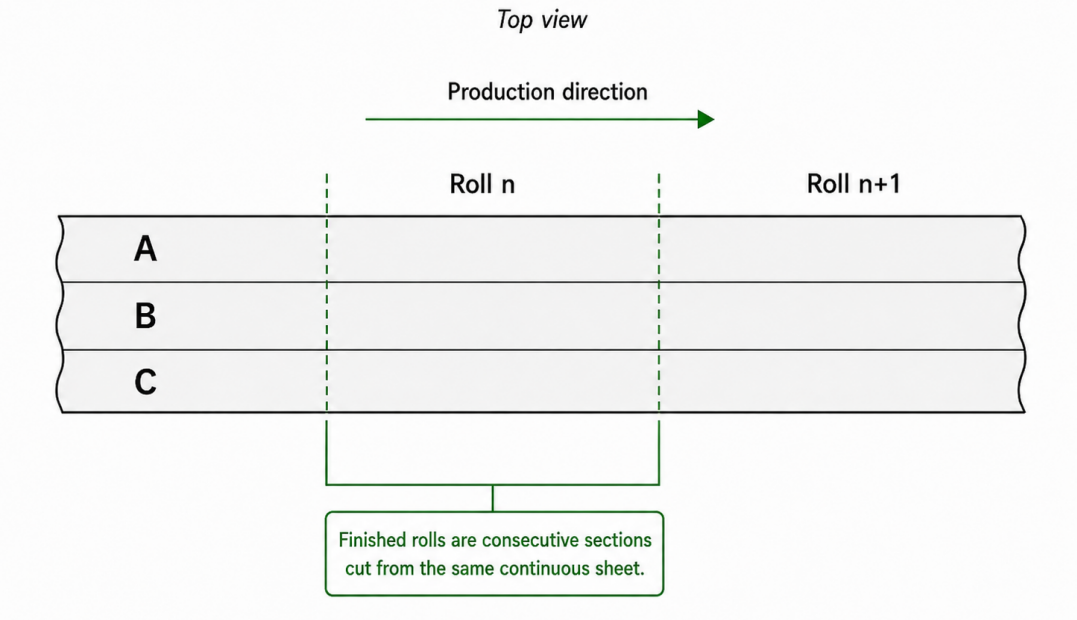


Figure 1.2: Schematic illustration of the A, B, and C roll positions used for the silver measurements.

1.4 Modelling problem and data challenges

The goal of the modelling is to develop predictive models for the four response variables using the available production data. The intended use is to predict the responses for future rolls. The models must therefore only use information that would have been available at the time of prediction.

The modelling also investigates which data sources are most useful for prediction. To study this, models use different groups of data as inputs, including QC data, SCADA summaries, and combinations of these sources. For the thickness responses, additional models use narrower parts of the data, such as observations from a specific product or flow line. These narrower models are used to study whether predictor importance changes between products and flow lines.

Model interpretation is an important part of the thesis. In addition to making accurate predictions, the models should help identify which predictors are most influential and how they are related to the responses. This is difficult for two main reasons. First, many process variables are anonymised, which limits their physical interpretation. Second, many predictors are related to each other and may describe overlapping parts of the production process.

The data also have several practical challenges. A central challenge is that the different data sources must be merged before modelling. This is not straightforward because the sources differ in structure, storage format, and relation to the production

process. For example, the SCADA data are time series recorded continuously over the full study period. To use them as model inputs, they must first be summarised over relevant production time windows and linked to individual rolls. In contrast, QC and laboratory measurements are already linked to specific rolls or to specific positions on a roll. The data also contain missing values, outliers, and other quality issues that must be handled carefully before modelling.

1.5 Aim and research questions

The aim of the thesis is to evaluate whether the data from medical foam production can be used to predict silver and thickness responses, and whether the fitted models can provide useful interpretations which provide insights into the production process.

The thesis addresses four research questions:

1. How predictable are Ag [mg/g], Ag [mg/cm^2], thickness mean, and thickness log variance from the available data?
2. Which data sources contain the most predictive information: QC data, flow-line SCADA data, reactor SCADA data, or a combination of these sources?
3. Which predictors have the strongest and most stable influence on the fitted models, and how can the interpretation results be summarised for operators and process review?
4. Do thickness models using subsets of the data show the same interpretation patterns as thickness models using the full dataset, or do narrower production subsets reveal different predictor rankings?

2

Theory

This chapter introduces the theory needed for the modelling and interpretation used in the thesis. It first describes prediction in industrial process data and the regression setting used for the response variables. It then presents the model families used in the reported results: ridge regression, random forests, XGBoost, and Explainable Boosting Machines. Some additional model types were also tested during model development, but they are not described in detail here because they were not used in the final results. These additional comparisons are shown in Appendix B. The chapter then describes how model performance is evaluated and how the fitted models are interpreted.

2.1 Prediction in industrial process data

Industrial production data often combines data from sensor signals, control system, laboratory analyses, and quality measurements. Such data can in many cases, for instance in this thesis, be used for prediction of product variables using the production data. This will be referred to as the prediction problem. One common method to solve the prediction problem is regression, which is a supervised learning technique where a model is trained on historical data to predict a continuous response variable based on inputs [2, 5, 11].

Although the method is conceptually straightforward, there are practical challenges. Measurements are often noisy, values may be missing, may be on different scales and are often correlated with each other. The data also often has a time structure, observations close in time are more similar than observations far apart, and the data may be grouped by production batches or similar.

These properties affect both modelling and evaluation. For example, a model can look useful if tested in an unrealistic way but fail on later production. The data properties also affect interpretation: an association between a predictor and the response may not be causal if that predictor is correlated with other predictors.

2.2 Supervised learning for regression

We continue with a more formal introduction to the tools needed to solve the prediction problem mentioned above. Let

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$$

denote a data set with n observations, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ is the predictor vector and $y_i \in \mathbb{R}$ is the observed response for observation i. The aim in regression is to learn a function f that predicts the response from the predictors, so that we get a prediction $\hat{y}_i$ for each observation i as

$$\hat{y}_i = f(\mathbf{x}_i).$$

Where the guess $\hat{y}_i$ is an approximation to the true response y_i so that,

$$y_i = \hat{y}_i + \varepsilon_i = f(\mathbf{x}_i) + \varepsilon_i,$$

where the ε_i are the errors from noise and unexplained variation.

The most common way to learn a function, also referred to as model fitting, is to minimise a loss function on training data. Training data refers to a subset of the data $\mathcal{D}$ that is used to fit the model, while the remaining data is used for evaluation. The loss function measures how well the model's predictions match the observed responses in the training data. For regression, a common loss function to minimise is the squared error loss

$$\sum_{i \in T} (y_i - \hat{y}_i)^2$$

where $T \subseteq \{1, \ldots, n\}$ is the set of indices for the training data. However, if no restrictions are placed on the class of functions, there are infinitely many functions that can minimize the error, since the sum is 0 if f passes through all the data points. The question then becomes: which function should be chosen?

Naturally, the function should not only fit the training data well, but also make good predictions on new data. This is referred to as good *generalization performance*. Furthermore, if a model is too flexible, it may learn random noise in the training data instead of real patterns – known as *overfitting* – and if a model is too simple, it may fail to capture important structure in the data – known as *underfitting*. Good regression modelling tries to find a balance between these two problems, often by restricting the class of functions over which the loss is minimized [2, 5].

2.2.1 Linear regression and regularisation

Linear regression restricts f to a linear function,

$$f(\mathbf{x}_i) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta},$$

where $\beta_0 \in \mathbb{R}$ is the intercept and $\boldsymbol{\beta} \in \mathbb{R}^p$ contains the regression coefficients. The intercept and coefficients are often estimated as $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ which are the values that minimise the squared error loss on the training data:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i \in T} \left(y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2.$$

Linear models are simple and interpretable, but they can be unstable when many predictors are correlated, the number of predictors is large relative to the number of observations, or when the true relationship is not linear.

To combat these issues, regularisation can be used. One form of regularisation is ridge regression, which adds an L_2 penalty to shrink the coefficients:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i \in T} \left[\left(y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \right],$$

where $\alpha \geq 0$ controls the strength of the penalty. This stabilises the estimates and can improve prediction when predictors are correlated or when the number of predictors p is large relative to the number of observations n. However, ridge regression still assumes a linear relationship between predictors and response, which may not hold in practice [2, 5].

2.2.2 Tree-based ensemble methods

Restricting f to linear functions can be too limiting, since nonlinear relationships cannot be captured; tree-based ensemble methods provide a more flexible alternative.

Decision trees split the predictor space into regions and assign a constant prediction within each region, as shown in Figure 2.1. Each split is based on one predictor and one threshold, for example separating observations with $x_j \leq c$ from those with $x_j > c$. Repeated splits therefore divide the predictor space into rectangular regions, where the prediction is usually the average response of the training observations in that region. A single tree is often unstable and can overfit, so tree-based ensemble methods combine many trees to improve stability and predictive performance.

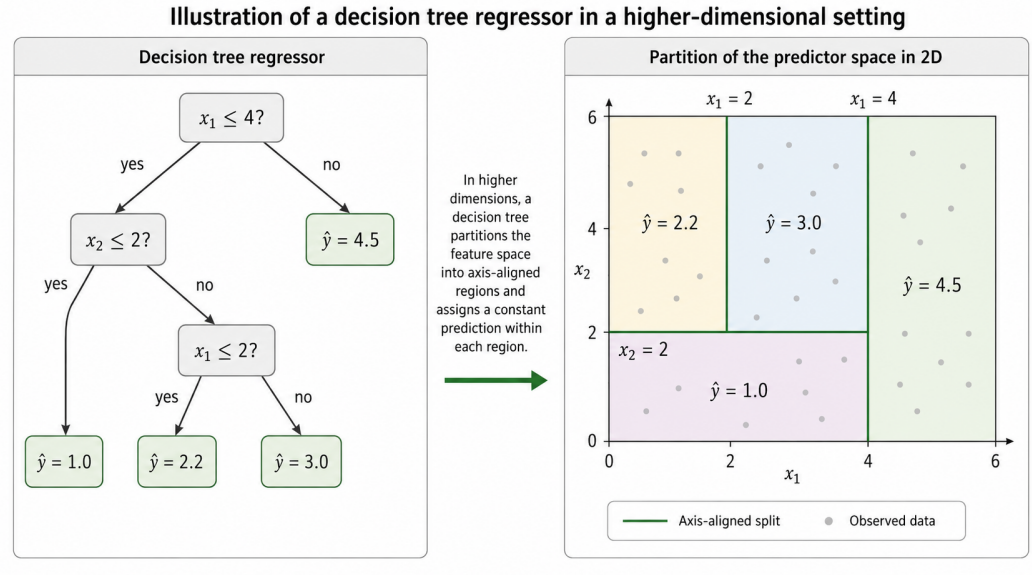


Figure 2.1: Illustration of a decision tree regressor as a partition of the predictor space. Each split creates axis aligned regions, and each region receives a constant prediction based on the training observations that fall within it.

One common ensemble method is the random forest. A random forest fits many trees, where each tree is trained on a slightly different random version of the training

data, and only a random subset of predictors is considered at each split. This makes the trees different from each other. The final prediction is the average of the predictions from all trees.

Another example is gradient boosting, which also fits an ensemble of trees but in a different way. Random forests fits trees independently and then averages them, so the trees are not influenced by each other. Gradient boosting instead fits trees sequentially, where each tree is built to improve the errors made by the earlier trees in order to improve the model in every step. Extreme gradient boosting (XGBoost) [3] is an efficient implementation of gradient boosting that is widely used, for instance, in this thesis.

While tree-based ensemble methods often have better predictive performance than linear models, they have some drawbacks. They are less transparent than linear models, so the interpretation is more difficult and requires more complex tools to understand how predictors contribute to predictions. They also require more careful tuning of hyperparameters to achieve good performance and avoid overfitting [2, 5].

2.2.3 Explainable Boosting Machines

Explainable Boosting Machines (EBM) are fitted in the form of a generalised additive model (GAM) but use boosting to estimate the shape functions [7, 12]. In a standard GAM, the learnt function f has the form

$$f(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}),$$

where each f_j is a one-dimensional shape function for predictor j. Thus, the additive structure is more flexible than a linear model, because each predictor can have a nonlinear effect on the response.

The difference between a GAM and an EBM is therefore mainly the fitting procedure. Classical GAMs are often fitted with smooth basis functions or splines, whereas EBMs estimate the shape functions with gradient boosting of shallow trees, see [6, p. 152] for details.

Thus, EBMs combine the interpretability of GAMs with the flexibility of tree-based methods [12]. For interpretation, an EBM can be summarized using term importance and plots of the estimated shape functions. These plots show how the fitted prediction changes over the observed range of a predictor, while the other additive terms remain part of the model. In this sense, EBMs sit between simple linear models and unrestricted tree ensembles: they can model nonlinear predictor effects while still keeping much of the additive structure visible.

2.3 Evaluation and interpretation

So far, we have described how different types of models can be fitted on a training index set T. However, we have not yet discussed how the training and test data should be chosen, how appropriate hyperparameters can be selected, or how model

performance should be estimated on unseen data. Model interpretation has also been mentioned in earlier sections, but has not yet been defined or discussed in detail. The following sections therefore describe both how the models are evaluated and how their results are interpreted.

2.3.1 Validation and data leakage

The full data set is often split into a development set and a held-out test set, for example using an 80/20 split. The development set is used during model development, mostly for comparison of models and hyperparameter tuning. The held-out test set is kept separate and is used only for the final evaluation after *all* the modelling choices have been made.

To avoid overfitting to the development set, cross-validation (CV) can be used within the development set. In standard K-fold CV, the development set is split into K folds. Each fold is used once as a validation set, while the model is trained on the remaining $K - 1$ folds. The validation performance is then averaged across the K folds to estimate how well the model is expected to perform in multiple scenarios and on unseen data.

When splitting data for CV or test evaluation, it is important to avoid data leakage. Data leakage occurs when information that would not be available at prediction time is used during training, preprocessing, tuning, or evaluation. Data leakage can lead to overly optimistic performance estimates and poor generalization to new data. For example, if observations from the same production batch, roll, or time period are present in both the training and validation sets, then the validation may be too similar to the training data and not reflect the true performance on new, unseen batches or time periods.

Leakage can also occur during preprocessing. Steps such as imputation and scaling should be fitted only on the training part of each split and then applied to the corresponding validation or test data. This ensures that the validation and test data remain unseen during model development.

To reduce leakage, the splitting strategy should respect both the grouping and time structure of the data. For example, observations from the same physical roll should remain in the same split, and time-based splits should be used when the goal is to predict future production. The specific splitting strategy used in this thesis is described in Section 3.3.

2.3.2 Performance metrics for regression

Recall that $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ denotes the full data set and that $T \subseteq \{1, \ldots, n\}$ contains the indices of the observations used to fit a model. For a particular validation or test evaluation, let $E \subseteq \{1, \ldots, n\}$ denote the indices of the observations on which that fitted model is evaluated, with $T \cap E = \varnothing$. In a cross-validation fold, T is the training part of the fold and E is its validation part. For the final evaluation, T contains the development observations used to fit the selected model and E is the held-out test set. Thus, for each $i \in E$, $\hat{y}_i$ is a prediction produced by a model fitted without using observation i.

The metrics in this thesis compare observed values y_i with predictions $\hat{y}_i$ for $i \in E$ and are standard for regression [2, 5]. The mean absolute error (MAE) is

$$\text{MAE} = \frac{1}{|E|} \sum_{i \in E} |y_i - \hat{y}_i|.$$

The root mean squared error (RMSE) is

$$\text{RMSE} = \sqrt{\frac{1}{|E|} \sum_{i \in E} (y_i - \hat{y}_i)^2}.$$

Both MAE and RMSE are measured in the same unit as the response. MAE describes the average absolute prediction error, while RMSE gives more weight to large errors.

The coefficient of determination is

$$R^2 = 1 - \frac{\sum_{i \in E} (y_i - \hat{y}_i)^2}{\sum_{i \in E} (y_i - \bar{y}_E)^2}, \qquad \bar{y}_E = \frac{1}{|E|} \sum_{i \in E} y_i.$$

Here, $\bar{y}_E$ is the mean response in the same validation fold or test set. The R^2 value compares the model with a baseline that always predicts this mean value. An R^2 close to one means that the model explains a large part of the variation in the response on the evaluation set. An R^2 close to zero means that the model performs similarly to the mean-value baseline, while a negative R^2 means that the model performs worse than this baseline.

Because the response variables have different units and scales, MAE and RMSE are mainly interpreted within each response, while R^2 is used as a relative measure of predictive performance.

2.3.3 Model interpretability

Model interpretability is a broad and somewhat ambiguous concept, but in this thesis it refers to the ability to understand which variables contribute to the model's predictions and how they contribute. This is important for several reasons. First, it can provide insight into the production process by identifying which predictors are most influential for the response. Second, it can help build trust in the model by providing explanations for its predictions. Third, it can guide future data collection and process improvements by highlighting predictors that may be worth studying further. The specific way interpretability is used in this thesis is described in Section 3.5.

Previously, we mentioned that linear models and EBMs are more interpretable than tree-based ensembles. This is because the linear models have a simple structure where each predictor has a coefficient that directly indicates its contribution to the prediction, while EBMs have shape functions that show how the prediction changes with each predictor. In contrast, tree-based ensembles combine many trees in a complex way, making it harder to understand how individual predictors contribute to the predictions.

To compensate for the lower interpretability of tree-based ensembles, various tools have been developed to explain their predictions. For example, permutation importance measures how much the model's performance decreases when a predictor's values are randomly permuted. Another example is SHapley Additive exPlanations (SHAP), which assigns an importance value to each predictor for a given prediction based on cooperative game theory. These tools can help identify which predictors are most important for the model's predictions. Furthermore, accumulated local effects (ALE) plots summarise average fitted effects of predictors in tree-based ensembles, similar to the shape functions in EBMs, and are designed to avoid some of the limitations of partial dependence plots when the predictors are correlated [1, 4, 8–10].

However, these interpretation tools show fitted predictive associations, not causal effects. A predictor may appear important because it is related to the response, or because it is correlated with other important predictors. This is especially relevant when many predictors describe overlapping parts of the production process. Therefore, predictor importance should be interpreted with caution and in relation to the broader structure of the data, rather than as direct evidence that one variable causes changes in the response.

3

Methods

This chapter describes the modelling workflow used to construct, evaluate, and interpret the predictive models in the thesis. The workflow is summarised in Figure 3.1. We first define the four response variables and the response-specific modelling tables. We then describe how the raw data sources were cleaned, merged, and converted into predictors, how the time-based development and held-out test splits were created, and how preprocessing was applied within each split. Finally, we describe model tuning, model selection, and the interpretation methods used for the final models.

All the modelling steps were implemented in Python, and the code is available in a Bitbucket repository to which Mölnlycke has access. The software implementation is described in more detail in Appendix A.

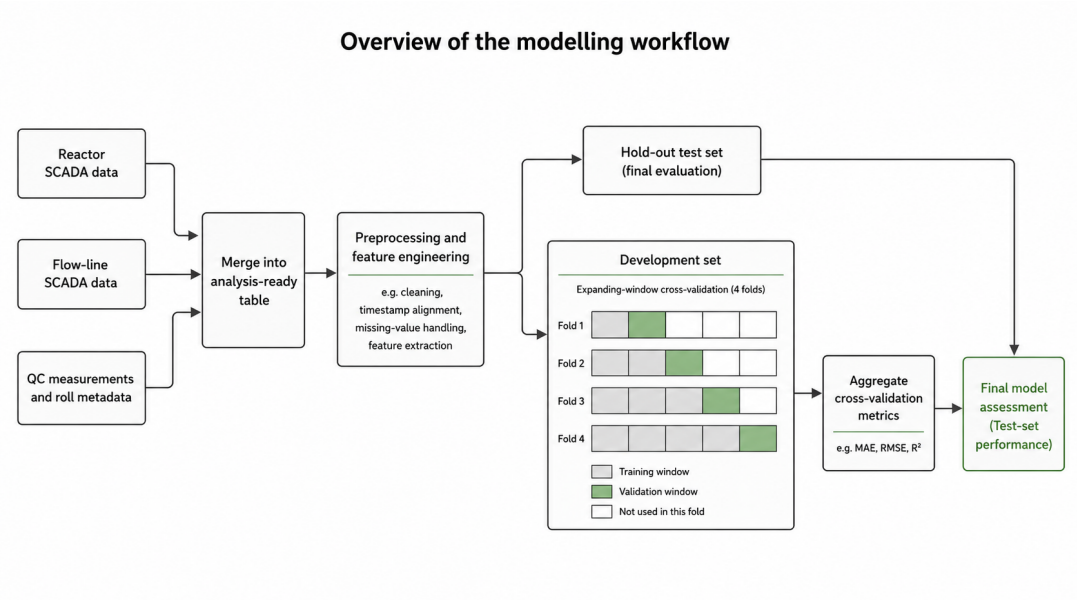


Figure 3.1: Overview of the modelling workflow. Data sources are merged into modelling tables for each of the four response variables and each response is evaluated using expanding window cross-validation on the development set, and finally assessed on a held-out test set.

3.1 Response-specific modelling tables

The thesis considered four prediction tasks: two with silver titration measurements as the response and two with thickness profile measurements as the response. The previous chapter described a regression dataset as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. In the modelling workflow, this dataset is stored as a modelling table: each row is one observation i, one column contains the response y_i, and the remaining columns form the predictor vector $\mathbf{x}_i$. Thus, before a model can be fitted, the raw data must be cleaned, linked, and arranged into such a table. Because the four responses differ in observation unit and predictor columns, the responses are defined first, before the construction of the modelling tables is described.

3.1.1 Silver responses

In the silver workflow, the response was either the silver concentration response,

$$y = \text{Ag } [\text{mg/g}],$$

or the silver area-density response,

$$y = \text{Ag } [\text{mg/cm}^2].$$

Both these measurements were in the silver-titration data, which contained laboratory measurements of silver content at roll positions A, B, and C, see Figure 1.2. The silver concentration response is the silver content per gram of sample, while the silver area-density response is the silver content per square centimetre of roll surface. The area-density response is calculated from the concentration response together with the sample weight.

The observation unit for the silver responses was one roll position, defined by the combination of job number, roll number, and roll position. This means that each row in the silver modelling tables corresponds to one of the A, B, or C positions on a particular roll.

3.1.2 Thickness responses

The thickness responses are derived from the QC data. Let

$$T_1, \ldots, T_6$$

denote the six thickness measurements taken for each roll at the predefined profile positions. The thickness mean response is then defined as

$$y_{\text{mean}} = \frac{1}{6} \sum_{j=1}^{6} T_j.$$

With $\bar{T} = y_{\text{mean}}$, the thickness profile variance was computed as

$$V_T = \frac{1}{6} \sum_{j=1}^{6} (T_j - \bar{T})^2.$$

The thickness log variance response was then defined as

$$y_{\log\text{-var}} = \log\Big(V_T + 10^{-6}\Big).$$

The logarithmic transformation was used because variance is non-negative, while ordinary regression models can produce negative predictions. By modelling the log variance instead of the variance directly, predictions can be transformed back to the variance scale while preserving the non-negative interpretation.

The offset 10^{-6} was added before taking the logarithm because the variance can in principle be zero. A prediction $\hat{y}_{\log\text{-var}}$ can be returned to the variance scale as

$$\widehat{V}_T = \exp(\hat{y}_{\log\text{-var}}) - 10^{-6},$$

with any small negative numerical value interpreted as zero.

For the thickness tasks, the observation unit was one roll rather than one roll position. This means that each row in the thickness modelling tables corresponds to one roll, with the response derived from the six profile measurements for that roll.

3.2 Construction of the modelling tables

The modelling tables were constructed from the raw data sources summarised in Table 1.1. Each table consists of rows representing observations of the response variable being studied and columns representing predictors. The construction process was carried out separately for the silver and thickness workflows, since they use different responses, observation units, and predictor rules. In both workflows, the main steps were to clean the source data, apply response-specific inclusion and exclusion rules, summarise time series SCADA data over the relevant production windows, and join the resulting predictors to the active response table. The following sections describe these steps in more detail.

3.2.1 Base tables

Each modelling table started from a base table containing the response variable and the information needed to link that response to a roll or roll position. For the thickness responses the QC data, with the small modification of adding the thickness mean and log variance as columns, formed the base table. This is because the responses were derived from the thickness measurements in the QC data, and the data included the information needed to link each measurement to a specific roll: the production batch and the roll's sequence number within that batch.

Similarly for the silver responses, the base table was the silver-titration data, which contained the silver measurements. This table contained the same roll identifiers as the QC data, but it also included the roll position (A, B, or C). Thus allowing us to link the silver measurements to a specific roll and position on that roll.

3.2.2 Cleaning and merging

Starting from the base tables, the data was cleaned and then merged with the other relevant data sources. The cleaning steps included removing duplicate rows, removing rows with missing key identifiers or impossible values such as negative, zero, or excessively large values. Furthermore, for rows with identical identifiers, one row was kept using a fixed rule based on source ordering or metadata.

After the cleaning, the base tables were merged with the line data. The line data contains context information, such as production start and end times, which flow line each roll was produced on, and which product was produced. The silver base table was additionally merged with the QC data. Since the silver is measured at roll positions the QC and line data is identical for all positions on the same roll. This means that the same line and QC information was attached to the A, B, and C observations for each roll.

After the merging, the resulting tables were cleaned again by removing rows with missing or unusable timestamps or impossible production windows. An additional check for the silver workflow was to remove rolls without a complete A/B/C set of silver measurements, so that the observations were balanced. This gave cleaned datasets of valid rolls with complete production windows for later feature creation using the SCADA data.

3.2.3 Merging SCADA data and feature engineering

After the previous cleaning and merging steps, the only remaining data sources to include were the flow line and reactor SCADA data. For the flow line SCADA data this was fairly straightforward, since the production start and end times were available for each roll as well as information regarding which of the two flow lines each roll was produced on. The reactor SCADA data was more challenging to merge, since the time between the reactor and the flow line is not fixed and can vary greatly between rolls. Therefore, an auxiliary table which linked each roll to a batch of material used to produce the foam and also linked the batches to the relevant reactor windows. This made it possible to merge the reactor SCADA data to the production data by linking each roll to the relevant reactor windows through the material batches.

Once the time windows for the flow line and reactor SCADA data had been defined, the next step was to summarise the time series data over these windows so that the summaries could be used as predictors. This was done in three main ways. First, the full window was summarised for each sensor using statistics such as the mean, standard deviation, minimum, maximum, range, and selected quantiles. Second, additional summaries were computed to describe how each signal changed over time, such as overall drift, mean absolute successive difference, number of missing values, and related change-based summaries. Third, each window was divided into three equal-duration subwindows, and the same summary statistics were calculated separately for the early, middle, and late parts of the window.

These summaries could then be added to the modelling tables as predictors for each roll or roll position. In order to make the predictors identifiable, a consistent naming

convention was used to indicate the data source, time window, process variable, and summary type for each predictor. For example, a predictor named `FL1__early_ _A-S__mean` would denote the mean of the anonymised flow line 1 variable `A-S` over the early subwindow of the production period. Other predictors would follow the same convention, with the data source prefix (e.g., `FL1`, `FL2`, `Reactor`), the time window label (e.g., `full`, `early`, `mid`, or `late`), the anonymised tag, and the type of summary statistic.

Additionally, from the QC data, several new predictors were created that summarised the thickness profile measurements in different ways. For example, the mean, standard deviation, minimum, maximum, and range. Simple contrasts, such as left–right or edge–centre differences, were also included. These predictors were of course only included for the silver modelling tables and were mainly exploratory to see if they could capture useful information for the silver responses.

3.2.4 Feature screening and feature set definition

After the modelling tables had been defined for each response, several screening steps were applied before modelling. Predictors with only missing values were removed, as were constant or nearly constant predictors. Predictors with very high missingness were also dropped. Perfectly correlated predictors within the same anonymised SCADA tag were reduced using a priority rule, where simpler statistics were prioritised, so that only one representative predictor was kept.

Columns that directly described the response, or were used to calculate it, were excluded from the predictor set. For example, Ag [mg/cm^2] was not used as a predictor for Ag [mg/g], or vice versa. For the Ag [mg/cm^2] modelling table, the sample-weight information used to calculate the area-density was also excluded.

For the thickness workflow, this meant that the six raw thickness measurements were not used as predictors, since the thickness mean and thickness log variance were derived from them.

The remaining predictors after screening were then grouped into feature sets for modelling, that is sets of predictors. The feature sets were defined to allow comparisons between different types of predictors, such as comparing models that only use QC predictors to models that only use SCADA predictors, or to models that use all available predictors after screening. The four feature set labels used throughout the thesis are summarised in Table 3.1.

Table 3.1: Feature set labels used in the modelling experiments. All feature sets include the same available contextual predictors, such as roll position, product identity, and line identity, but differ in which QC, flow line SCADA, and reactor SCADA predictors are included.

Feature set	Included predictors
QC-only	QC predictors for the active response
SCADA-only	Flow line and reactor SCADA summaries, no QC predictors
SCADA-no-reactor	Only flow line SCADA summaries
combined	All remaining predictors after screening

Additionally, for the predictors derived from the SCADA data, the models could be fitted using any combination of the summary statistics and time windows. For example, a model could be fitted using only the mean summaries, only the early-window summaries, or only the full-window mean and standard deviation summaries. This allowed us to check whether certain types of summaries or time windows were more predictive than others and to identify the most effective combinations.

3.3 Data splitting and preprocessing

After the modelling tables had been constructed, the next step was to define the data splits for model development and evaluation, and to specify the preprocessing steps to be applied within each split. The data splitting and preprocessing were designed to give a realistic estimate of how well the fitted models would generalise to future production data and were designed to avoid data leakage and overfitting. The following sections describe the data splitting and preprocessing in more detail.

3.3.1 Training and held-out test splits

The data is split at the roll level for all the responses. This means that silver measurements at the A, B, and C positions from the same roll are always placed in the same split. This prevented closely related observations from the same roll from appearing in both training and evaluation data during the silver modelling.

Due to the time-dependent nature of the data, the splits were defined based on production time. This means that the training and test data were separated in time, with the test data representing later production than the training data. This design was chosen to give a more realistic estimate of how well the fitted models would generalise to future production data. By keeping the training and test data separate in time, the risk of data leakage and overfitting is reduced, and the evaluation results are more likely to reflect the true predictive performance of the models on unseen data.

Hence, for the training and held-out test split, the latest 20% of rolls were assigned to the test set, while the remaining 80% formed the development set. The common choice of a 80/20 split happened to result in the training set of silver observations approximately covering the first five production batches, while the test set covered the sixth and final production batch, as seen in Figure 3.2 which shows the observations on a time axis. This is because the silver measurements were only available for one of the foam products which was produced in batches. This means that the held-out test set represented a future production batch that was not seen during model development, giving a strong test of how well the fitted models would generalise to new production data.

For the thickness responses, the same 80/20 split was used, but the thickness measurements were available continuously over the production period of six months, so the split resulted in a training set covering the first five months and a test set covering the last month instead of seperate batches. This also gave a strong test of generalisation to future production data, since the test set represented a later time

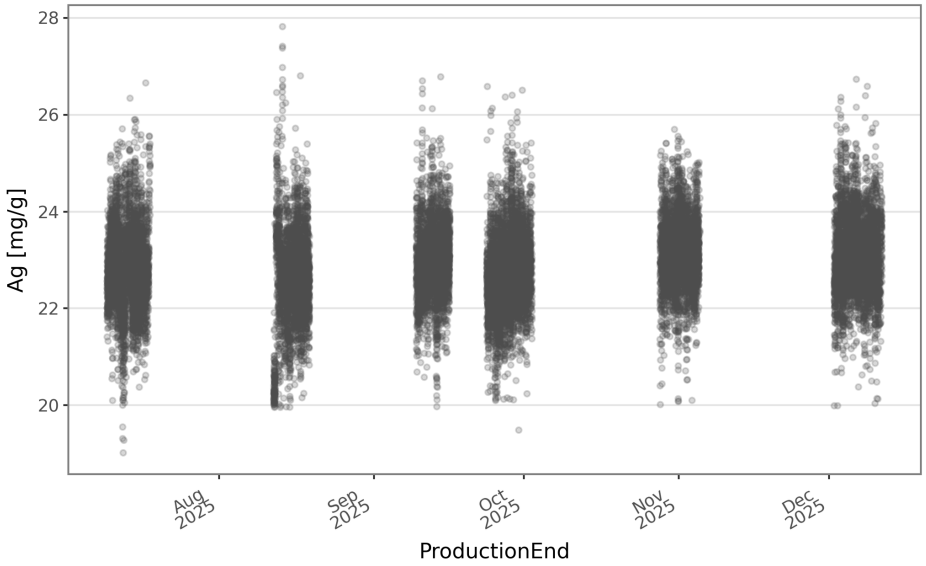


Figure 3.2: Cleaned Ag [mg/g] observations by production end time. Observe that the silver observations occur in separated production batches.

period than the training set.

3.3.2 Expanding window cross-validation

Within the development set, candidate models and hyperparameters were evaluated using expanding window cross-validation. In this procedure, the development set was divided into K consecutive blocks according to production time. In fold k, the model was trained on blocks $1, \ldots, k$ and validated on the following block, $k+1$, for $k = 1, \ldots, K-1$. Thus, the amount of training data increased in each fold, and validation was always carried out on later production data. For each candidate model, MAE, RMSE, and R^2 were computed on each validation block. The mean value of each metric across the $K-1$ folds was then used to compare models and select hyperparameters.

For the silver responses, five equal time blocks were defined within the development set, so that each block approximately represented one production batch. Thus, there were four folds in the expanding window validation, with each fold training on the earlier batches and validating on the next batch. The thickness responses had more observations, so the development set was split into six equal time blocks, giving five folds.

This design was chosen to match the intended use case as closely as possible. In practice, a model would be trained on past production and then used to predict later production. The cross-validation therefore respected time order, giving a more realistic estimate of predictive performance than a random split.

3.3.3 Preprocessing within each split

After a training, validation, or test split had been defined, the final predictor table was passed through a fixed preprocessing pipeline. Numeric predictors were imputed using the median and then scaled with min–max scaling to the interval $[0, 1]$. Categorical predictors were imputed using the most frequent category and then one-hot encoded, meaning that each category was converted into a binary indicator variable that equals one when the observation belongs to that category and zero otherwise.

For a numeric predictor x_j, min–max scaling was fitted on the training part of the active fold. If a_j and b_j are the training-fold minimum and maximum, the transformed value was

$$z_{ij} = \frac{x_{ij} - a_j}{b_j - a_j}.$$

Values below a_j or above b_j in validation or test data were set to 0 or 1 after transformation. This means that the predictors were on a bounded scale, with 0 corresponding to the smallest value observed in the training data and 1 corresponding to the largest value observed in the training data.

Median imputation was used because many process variables were skewed or contained unusual values. One-hot encoding was used for categorical variables such as roll position and product identity, avoiding an artificial numerical ordering. The min–max scaling also made ridge coefficients and fitted effects easier to compare as opposed to standard scaling. This is because a change from 0 to 1 corresponds to moving across the training range of a predictor.

The preprocessing steps were fitted inside the model pipeline within each training fold. This ensured that imputation and scaling were learned from the training data only, with no information from the validation or test data being used.

3.3.4 Thickness subset-data runs

As has been briefly mentioned, the thickness responses contained observations from both of the flow lines as well as from multiple products. To further investigate the predictive structure within these subsets of the data, additional runs were carried out where the modelling tables were restricted to certain subsets of the data. For example, only using observations from flow line 1, or only using observations from flow line 2, or only using observations from a specific product. These subset-data runs were only carried out for the thickness responses, since the silver responses were only available for one product and had already been analysed in the full dataset. The subset-data runs used the same modelling workflow as the main thickness models, but with a narrower modelling table that only included the relevant observations. This allowed us to check whether the predictive structure changed after restricting the data to a more homogeneous line or line-product setting.

3.4 Model tuning and selection

Within the modelling workflow, several candidate models were fitted for each response, feature set, and combination of SCADA predictors. See Section 3.2.4 for a

description of the feature sets and how the SCADA summary predictors were used. The candidate models were compared using the expanding window cross-validation results on the development set, and the best-performing models were selected for final evaluation on the held-out test set.

The candidate models were chosen to represent a range of model families with different levels of flexibility and interpretability. The main comparisons were between ridge regression, random forests, XGBoost, and EBMs. These model families were selected because they provide complementary views of the same prediction problem: ridge regression gives a linear benchmark, EBMs give an interpretable nonlinear additive model, and random forests and XGBoost give more flexible tree-based models.

Additional model families, including elastic net, principal component regression (PCR), partial least squares (PLS), and RuleFit, were tested during development and are documented in Appendix B. They were not included in the final reporting because they did not add a clear advantage over the retained model families.

3.4.1 Hyperparameter tuning

Model tuning was carried out within the development set by comparing candidate hyperparameter settings across the cross-validation folds. The held-out test set was therefore not used for hyperparameter selection. Each training run was defined by an explicit experiment specification, including the modelling dataset, response column, predictor set, split design, preprocessing setup, model family, and model settings. This made the experiments repeatable and ensured that model comparisons differed only in planned ways.

For simpler models, such as ridge regression, tuning used a one-dimensional search over the ridge penalty parameter. For XGBoost, random forests, and EBMs, several hyperparameters were tuned using grids. The grid searches were carried out in stages, starting with broad candidate ranges based on the default values and then moving to narrower grids around settings that performed well in the development folds. This approach made tuning feasible across the many models considered for each response, while avoiding an excessively large search space.

3.4.2 Model selection

Model selection was based mainly on the average performance across the cross-validation folds in the development set. The main selection metric was R^2, since it gives a relative measure of how well the model explains variation in the response compared with a baseline mean prediction. MAE and RMSE were used as supporting metrics to check that the selected models had sufficiently small prediction errors.

For models using SCADA predictors, selection also considered interpretability. If two models had similar validation performance, the simpler model was preferred. For example, a model using only mean summaries was preferred over a model using mean, standard deviation, and quantile summaries if the additional summaries did not clearly improve performance. This was done because one aim of the thesis was to

identify predictors that could support process understanding, not only to maximize predictive accuracy.

Fold-level predictions were also inspected before final model selection. For each validation fold, predictions were saved together with roll identifiers, production time, observed and predicted responses, and roll position when relevant. Residuals were defined as $e_i = y_i - \hat{y}_i$ and standardised as

$$r_i = \frac{e_i}{s_e},$$

where s_e is the sample standard deviation of the residuals in the corresponding validation fold. Standardisation expresses each error relative to the typical residual variation, which makes residual patterns easier to compare across models and response variables with different scales. Predicted-versus-observed plots, standardised residual histograms, and grouped standardised residual plots were used to identify systematic bias, restricted prediction ranges, time-dependent errors, and errors concentrated in particular groups of rolls.

After selection, each final model was refitted on the full development set and evaluated once on the held-out test set. The same diagnostics were then applied to the held-out predictions, but only to support interpretation of the reported test performance, not to retune the models. Diagnostics therefore described how well the models predicted and where their errors occurred, while the model interpretation described below in Section 3.5 examined which predictors or predictor groups were most influential in the fitted predictions.

3.5 Model interpretation

Model interpretation was applied to the final selected models. The aim was to identify which predictors contributed most to the fitted predictions and to inspect how these predictors were associated with the response in the fitted model. The results were interpreted as predictive associations in the available data, not as causal effects.

3.5.1 Interpretation scope

Interpretation was performed separately for each combination of response variable, feature set, model family, and, where relevant, data subset. Interpretation was treated as reliable only when the corresponding model showed useful predictive performance on the held-out test set. Results from models with weak or negative held-out R^2 were included only as exploratory findings.

The interpretation also had to account for correlation between predictors. Many of the predictors were correlated with each other, especially within the same data source. Under such correlation, importance can be shared between predictors or move when the feature set or model family changes. For this reason, the reported rankings should be read as evidence for predictor groups.

For example, if a SCADA predictor, which could summarise a time series of the variable "Fan Speed", was ranked highly that would suggest that the underlying process variable was important for prediction. However, it might be the case that other fan settings are also recorded and these change together with the fan speed, so the conclusion would be that the fan settings are important for prediction, rather than the specific fan speed. This is not something this study can resolve, due to the anonymised nature of the SCADA data, but the domain experts can use the predictor rankings as a starting point for further investigation of which underlying process variables are important for prediction and how they are related to the response.

3.5.2 Evidence tiers

Predictors of a model were first ranked, in each cross-validation fold and in the held-out period, using permutation importance. This measures how much model performance decreases when the values of a predictor are randomly permuted. The permutation was repeated five times for each predictor, and the resulting importance values were averaged.

Robustness checks were then used to avoid basing interpretation on one split, one model, or one importance ranking. Three checks were used: stability across development folds, confirmation on the held-out period, and comparison with a shuffled-response baseline. The idea was that a predictor that appears repeatedly among the top 12 predictors across development folds, also appears among the top 12 predictors in the held-out period, and has much larger importance than a shuffled-response baseline is more likely to be stable and meaningful than a predictor that only appears in one fold or has similar importance to the shuffled baseline.

The shuffled-response baseline was used as a null comparison. The response values were randomly shuffled and the importance analysis was repeated using three shuffle iterations. Any remaining importance was interpreted as the level of apparent importance that could arise without a meaningful predictor-response relationship. By comparing the original importance with the shuffled baseline, we could check whether the observed importance was larger than what would be expected by chance. This helped to filter out predictors that had high importance due to random correlations or noise rather than due to a meaningful relationship with the response.

A predictor was classified as *high evidence* if it satisfied all three criteria:

1. it appeared among the top 12 predictors in at least 80% of the development folds,
2. it appeared among the top 12 predictors for the held-out diagnostics, and
3. its held-out importance was at least 1.5 times larger than the corresponding shuffled-response baseline.

Predictors satisfying two of the three criteria were classified as *medium evidence* in a model. Predictors with weak or inconsistent support across models were not used for the main interpretation claims. The predictors that passed these filters were discussed in detail.

3.5.3 Predictor ranking

A response-level predictor ranking was created to give a compact overview of recurring non-context predictor groups across interpreted runs. This ranking used the same interpretation summaries as the evidence-supported predictor selection, but served a different purpose. Evidence-supported predictors were selected for detailed interpretation claims, whereas the response-level ranking was used to compare broader predictor groups across responses.

Context variables such as `LineID`, `ProductNumber`, and `RollPosition` were excluded before scoring, so that the ranking focused on process, reactor, and QC predictor groups rather than contextual identifiers.

Predictor names were simplified before aggregation. For SCADA and reactor predictors, time window labels and summary statistics were removed so that summaries of the same underlying anonymised tag were grouped together. For example, `FL1__full__A-S__mean` and `FL1__early__A-S__mean` were grouped as `A-S`. QC roll variables kept their descriptive names. This grouping was used because adjacent windows and summary statistics for the same sensor are often strongly correlated and are better interpreted as one predictor family than as isolated variables.

For each response, simplified predictor, and model family, a score on a 0–100 scale was computed from the development-period importance and fold-stability summaries. Higher scores corresponded to predictors that appeared repeatedly among the top 12 predictors, had a median rank close to the top of the ranking, and had stable ranks across folds rather than widely spread ranks. A score of 100 would require a predictor to be ranked first consistently across the available fold summaries with no rank instability. When several retained model families were available, the final score was averaged across model families. The score was used only as an overview measure requested by the stakeholders.

3.5.4 Model-specific interpretation

For the predictors that passed the evidence-tier filters, model-specific interpretation methods were used to inspect how these predictors were associated with the response in the fitted models. The interpretation methods differed between model families, since they have different structures.

For ridge regression, the fitted coefficients were inspected. The coefficients were read on the min–max-scaled predictor scale, so a coefficient describes the fitted change in the response when the predictor moves from the smallest to the largest value observed in the training data, with the other predictors held fixed in the fitted linear model. The main interpretation focused on the direction and relative size of the coefficients.

XGBoost and random forest models were interpreted using SHAP values and ALE curves. The ALE curves show how the fitted contribution of a predictor changes over the observed range of that predictor, while the SHAP values give a local explanation of how each predictor contributes to the prediction for each observation.

The EBM models were interpreted using the shape functions. The shape functions

show how the fitted additive contribution of a predictor changes over the observed range of that predictor. This allows us to see whether the fitted relationship is linear, monotonic, or more complex, and to identify any thresholds, plateaus, or other patterns. These plots are also useful for comparing against the interpretation of the ridge regression coefficients and the XGBoost ALE plots.

If multiple model families were available for a response and feature set, the interpretation was used to check whether the same predictors had similar associations across models. For example, if a predictor had a positive coefficient in ridge regression and also had a increasing ALE curve in XGBoost, that would give more confidence that the predictor has a consistent positive association with the response. If the associations differed between models, that would suggest that the relationship between the predictor and the response is more complex and that the fitted associations are more model-dependent. This would be taken into account when interpreting the results and making conclusions about which predictors are most important for understanding the response.

4

Results

This chapter presents the results from the modelling workflow described in Chapter 3. It first summarises the four response distributions and then compares predictive performance across model families, feature sets, and the held-out test period. The chapter then reports the stable predictor interpretation results and ends with the additional thickness subset-data runs used to examine line- and product-specific structure.

4.1 Response distributions

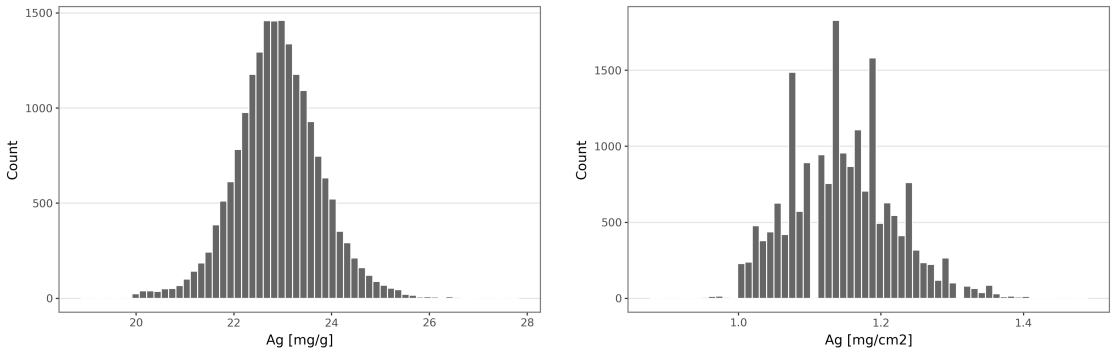
Figure 4.1 compares the four response distributions and reports the number of observations in the final modelling table for each. The thickness responses show more mixture-like structure than the silver responses.

4.2 Predictive results by response

For each response, predictive performance is summarised by the mean cross-validation (CV) R^2, with error bars showing approximate 95% confidence intervals based on the sample standard deviation of the fold-level R^2 values. The best performing models on the development set were then selected for evaluation on the held-out test set, and the test metrics are reported in Table 4.1. After the predictive results, the diagnostic plots for one of the selected models are shown and discussed.

The mean CV R^2 figures in this section compare model families across the different feature sets. The x-axis labels in the figures show the feature set as defined in Section 3.2.4 above, *QC* means the model used only QC predictors. If the feature set is *SCADA no reactor* then the predictors are SCADA summaries from the flow lines, while if it is *SCADA* then the predictors come from both flow lines and reactor SCADA summaries. *Combined* contains all of the previously mentioned predictors.

Furthermore, if the feature set is described as *all mean*, it refers to SCADA summaries using the mean value of each sensor over the full, early, middle, and late roll production time windows. One major result is that the all mean SCADA summaries give the best balance between predictive performance and interpretation. Using the mean and standard deviation over the full time window gave similar performance, but it was decided that the standard deviation summaries were harder to interpret for the production experts. Consequently, all of the models reported in this chapter use the all mean SCADA summaries when SCADA predictors are included.



(a) Ag [mg/g].

(b) Ag [mg/cm^2].

(c) Thickness mean.

(d) Thickness log variance.

Figure 4.1: Marginal distributions of the four response variables used. The silver modelling tables contain 19,035 observations from 6,345 rolls with one observation per A/B/C roll position. The thickness modelling tables contains 16,738 roll observations, with 8,648 observations from Line 1 and 8,090 observations from Line 2.

Table 4.1 summarises the final best performing models selected for each response, the CV metrics come from the development set while the test metrics come from the held-out test set. From the table it is clear that thickness mean is predicted very accurately from SCADA-only predictors, the thickness log variance models show moderate performance, Ag [mg/cm^2] models show moderate development set performance that does not transfer to the held-out period, and Ag [mg/g] is not predictable.

Table 4.1: Performance of the final models selected. The selected models are those with the highest mean CV R^2 on the development set, with a preference for SCADA-only models when the performance was close. The held-out test metrics are reported for the same model configurations as in the development set. Hyperparameters for each model are reported in Appendix B.

Response	Model	Feature set	CV RMSE	CV MAE	CV R^2	Test RMSE	Test MAE	Test R^2
Ag [mg/g]	XGBoost	combined, all mean	0.8471	0.651	0.004	0.8001	0.6223	0.005
Ag [mg/cm^2]	XGBoost	combined, all mean	0.0550	0.0432	0.298	0.0815	0.0665	-0.234
Ag [mg/cm^2]	Ridge	combined, all mean	0.0606	0.0475	0.150	0.0807	0.0656	-0.208
Thickness mean	XGBoost	SCADA-only, all mean	0.0162	0.00816	0.875	0.0071	0.00499	0.989
Thickness mean	Ridge	SCADA-only, all mean	0.0128	0.00786	0.936	0.0170	0.00865	0.938
Thickness mean	EBM	SCADA-only, all mean	0.0260	0.0164	0.812	0.0225	0.0137	0.891
Thickness log variance	XGBoost	SCADA-only, all mean	0.8097	0.638	0.384	0.8654	0.693	0.449
Thickness log variance	Ridge	SCADA-only, all mean	0.7784	0.626	0.435	0.6880	0.546	0.652
Thickness log variance	EBM	SCADA-only, all mean	0.9768	0.777	0.118	1.2176	0.989	-0.090

4.2.1 Ag [mg/g]

Ag [mg/g] was the weakest response in the final comparison. Figure 4.2 shows that the mean CV R^2 values are close to zero for the best configurations and negative for many others. The best development set result came from the combined XGBoost model using all mean SCADA summaries, but its mean CV R^2 was only 0.004. Thus, the available models do not provide a useful prediction of Ag [mg/g].

The held-out test period confirms the same conclusion. The selected combined XGBoost model reached test $R^2 = 0.005$, with test RMSE 0.800, as reported in Table 4.1. Due to the lack of predictive signal, the residual diagnostics are not included.

4.2.2 Ag [mg/cm^2]

Reformulating silver as the Ag [mg/cm^2] silver area-density response gave a clearer development set signal. In Figure 4.3, the best CV results are obtained by XGBoost models, with the combined feature set reaching mean CV $R^2 = 0.298$. QC-only models only perform slightly worse, which suggests that much of the development set signal is present in the quality measurements and not in the SCADA data.

The XGBoost combined model is used for the diagnostic plots, while the ridge combined model is retained for comparison in Table 4.1. Despite the moderate development set signal, the held-out test R^2 values are negative for both models, which indicates that the fitted associations do not transfer to the later production period. The later interpretation of predictors for this response is therefore more tentative, and it should be read as exploratory evidence of silver structure during the development period.

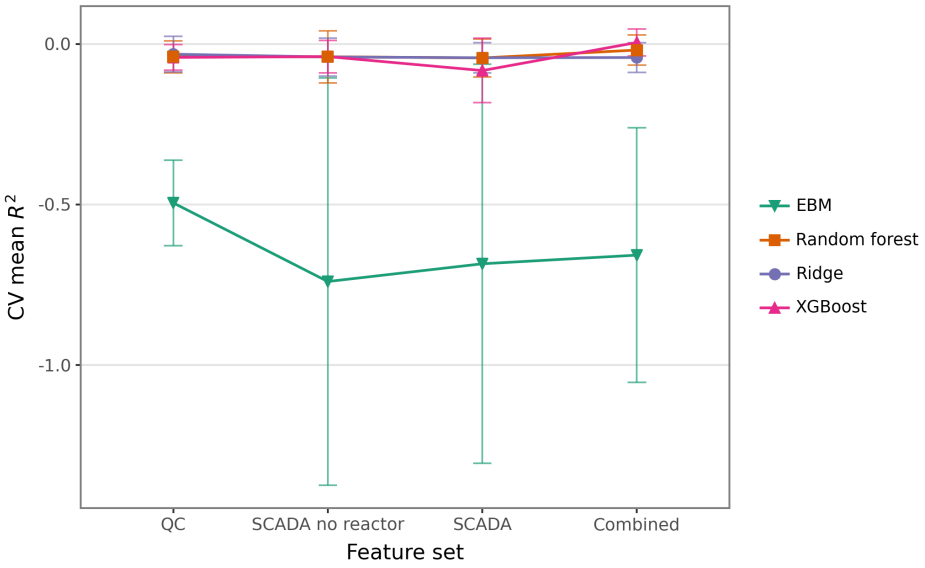


Figure 4.2: Mean CV R^2 for Ag [mg/g] by model family and feature set family on the development set. Error bars show approximate 95% confidence intervals for the fold mean.

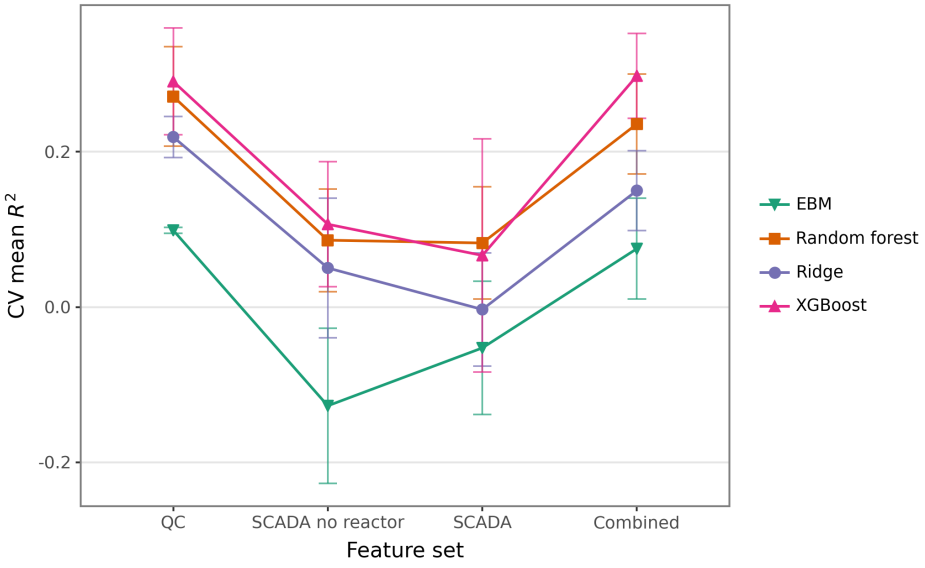


Figure 4.3: Mean CV R^2 for Ag [mg/cm^2] by model family and feature set family on the development set. Error bars show approximate 95% confidence intervals.

One important reason for the weak held-out transfer is visible in Figure 4.4. The Ag [mg/cm^2] boxplots show that the silver measurements suddenly increase during the last production period, which is the held-out test period. The upward shift is visible both in the mean across A, B, and C roll positions and in the separate A, B, and C measurements. The later production period therefore has a different distribution of silver levels than the earlier development period, and the fitted models do not capture this shift. The models therefore fail to predict the later silver levels, which gives negative held-out R^2 values even though the models capture predictive associations during development.

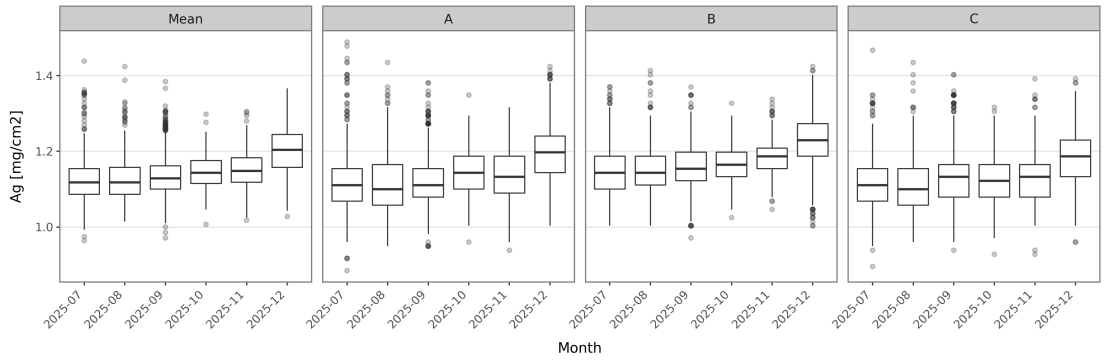


Figure 4.4: Monthly boxplots of Ag [mg/cm^2]. The left most window shows the roll level mean across A, B, and C, while the remaining windows show the A, B, and C roll position measurements separately. The upward monthly shift indicates response drift over the production period.

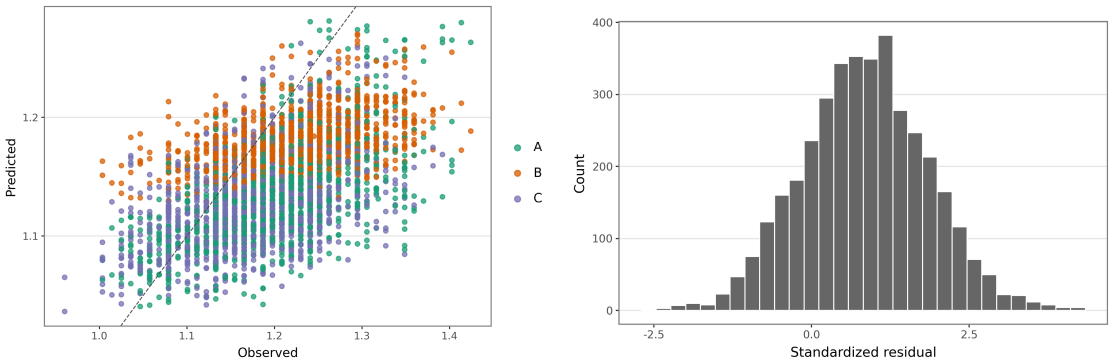
Figure 4.5 shows that the residuals are skewed toward underprediction, which is consistent with the upward shift described above. Other than this, the predictions show some structure by roll position and consistent variation around the mean.

4.2.3 Thickness mean

Thickness mean was the most predictable response by a large margin. In Figure 4.6, several models and feature sets reach high mean CV R^2. Although the combined feature set gives the best development set performance, the SCADA-only models are very close behind, and they provide a more insightful interpretation because they isolate summaries of process variables from downstream quality variables. Hence, the final models selected all use the SCADA-only all mean feature set, and the interpretation will focus on these models.

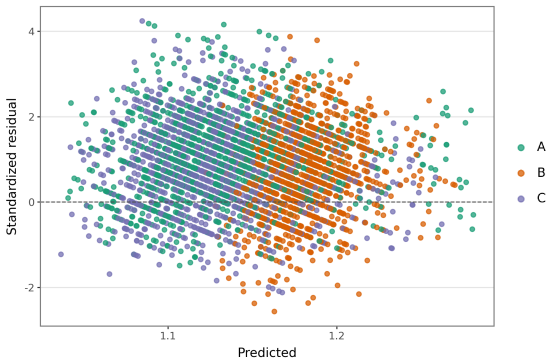
The SCADA-only models perform very well, as summarised in Table 4.1. Ridge regression had the highest development set mean CV R^2 among the SCADA-only reporting models (0.936), while the SCADA-only XGBoost model had the strongest held-out fit, with test RMSE 0.0071 and test $R^2 = 0.989$. The SCADA-only EBM was weaker but still useful, with test $R^2 = 0.891$ and test RMSE 0.0225. This shows that the SCADA data contains substantial information about the thickness response.

The held-out diagnostics in Figure 4.7 show predictions close to the diagonal and



(a) Predicted-versus-observed values by roll position.

(b) Standardised residual distribution.



(c) Standardised residuals by roll position.

Figure 4.5: Diagnostics for the Ag [mg/cm^2] XGBoost combined model calculated on the held-out set. The model fitted on the development set does not predict the later silver level well.

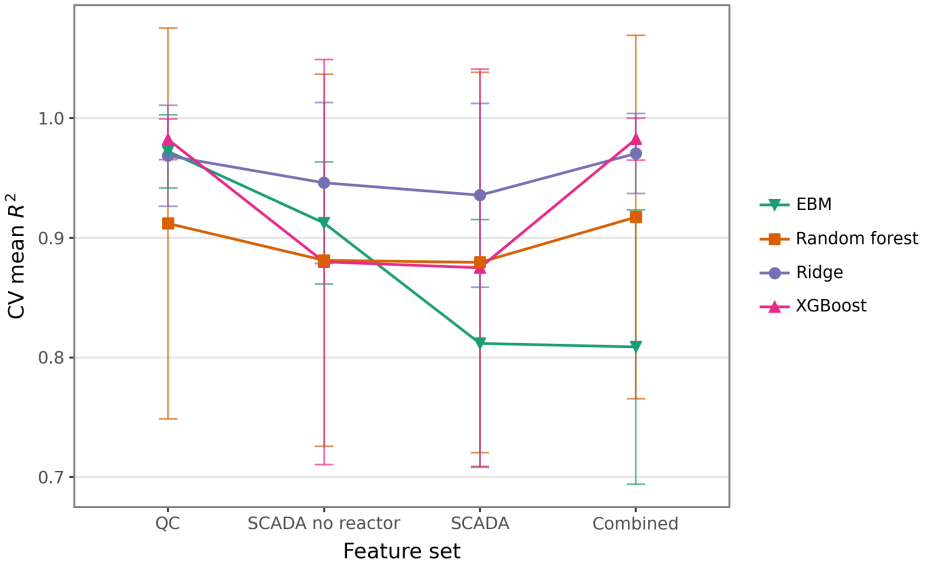


Figure 4.6: Mean CV R^2 for thickness mean by model family and feature set family on the development set. The error bars represent 95% confidence intervals. The SCADA-only feature set retains strong performance using only process variable summaries.

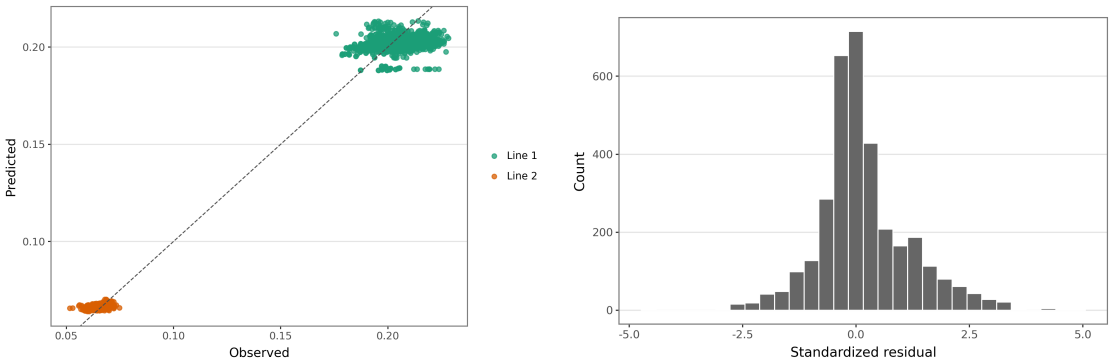
small residuals centred near zero. The residuals show structure by line and product. The residuals on line 1 are larger than on line 2. The product grouping is also visible in the residuals, with some product groups showing worse estimation than others. Thus, the model performance is not uniform across the production space, but the overall fit is very good. This motivated the later follow-up analyses of thickness mean models using subsets of the data, which are reported in Section 4.4.

4.2.4 Thickness log variance

Thickness log variance was more difficult for the models to predict than thickness mean, but it still was moderately predictable. The response represents within-roll thickness variation rather than the average thickness itself, so a lower predictive ceiling is expected. Figure 4.8 shows that ridge regression and XGBoost SCADA-only models both capture most of the development set signal. These models are therefore selected as the final models for this response, for the same reason as for thickness mean: they give good predictive performance while isolating process variable summaries from downstream quality variables, which gives a more insightful interpretation below.

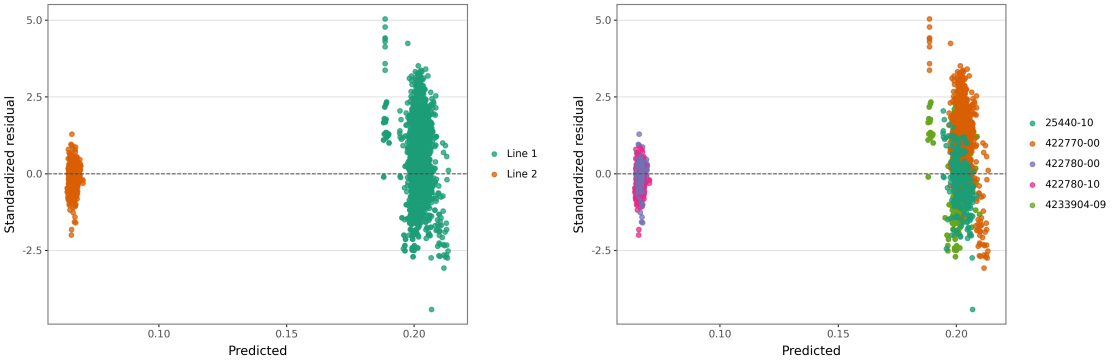
As seen in Table 4.1, the ridge and XGBoost SCADA-only models perform well on both the development and held-out test sets. Ridge regression achieved test RMSE 0.6880 and test $R^2 = 0.652$. XGBoost was slightly weaker, with test $R^2 = 0.449$ and test RMSE 0.8654. The SCADA-only EBM model performed poorly, with negative held-out $R^2 = -0.090$, so it is not used for interpretation.

Figure 4.9 shows the diagnostic plots for the XGBoost model. The plots show that the residuals almost have consistent variation around the mean, but the model



(a) Predicted-versus-observed values by line.

(b) Standardised residual distribution.



(c) Standardised residuals by line.

(d) Standardised residuals by product.

Figure 4.7: Diagnostics calculated on the held-out set for the SCADA-only XGBoost thickness mean model, which achieved test $R^2 = 0.989$. Predicted-versus-observed values are labelled by line, while standardised residuals are labelled both by line and by product number. The residuals show structure by line and product, but the overall fit is very good.

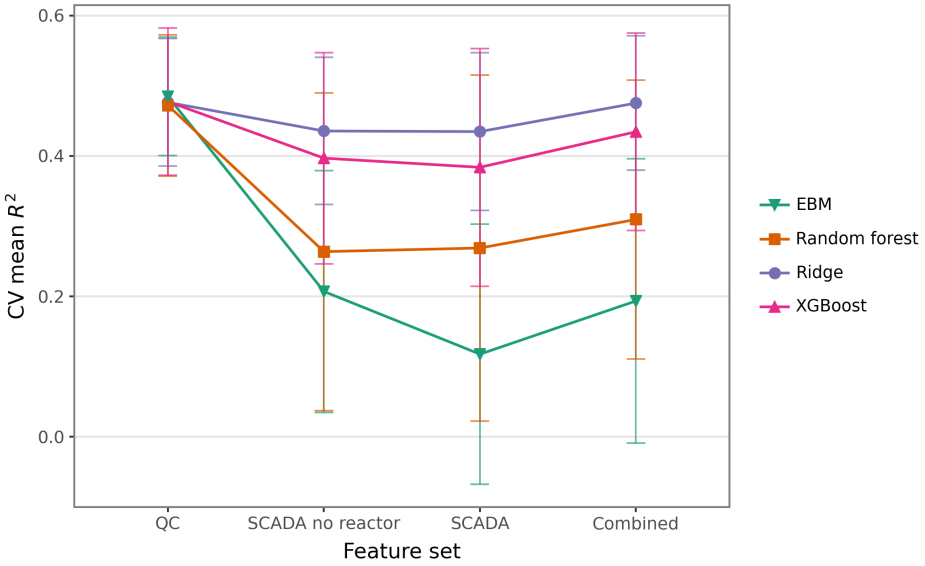


Figure 4.8: Mean CV R^2 for thickness log variance by model family and feature set family on the development folds. Error bars show approximate 95% confidence intervals. The SCADA-only feature set captures most of the development set signal, which motivates the selection of SCADA-only models for this response.

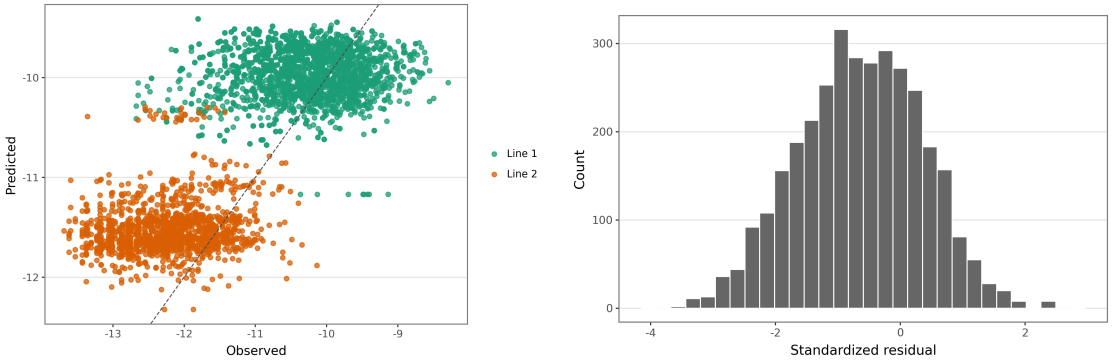
consistently overpredicts the log variance. As before, the residuals show structure by line and product, but the line structure is less clear than for thickness mean. The broader error distribution, as compared with thickness mean, is consistent with the lower held-out R^2 value. Again, these diagnostics motivated the later follow-up analyses of thickness log variance models using subsets of the data, which are reported in Section 4.4.

4.3 Interpretation of stable predictors

The interpretation results consist of two main parts: a broad response-level predictor ranking and a narrower set of evidence-supported predictors selected for detailed interpretation. How these are derived is described in Section 3.5, and the results are summarised in Figure 4.10 and Table 4.2 below.

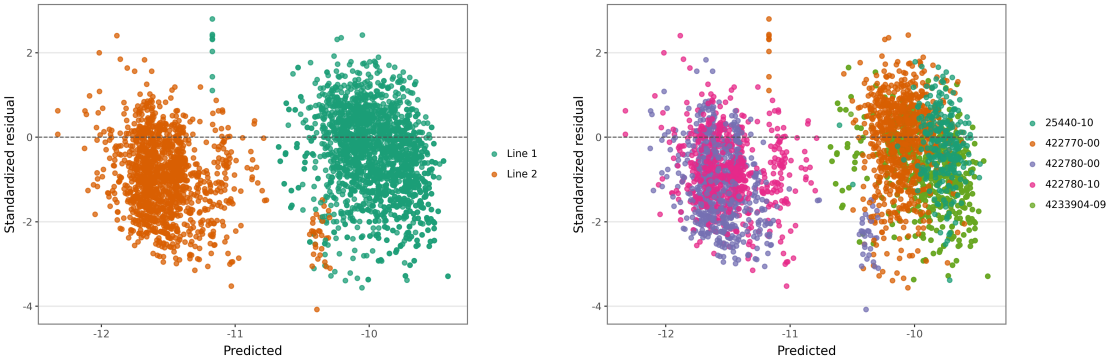
The predictor rankings in Figure 4.10 is based on the full set of predictors, but it groups repeated windows and summary statistics into simplified labels. The ranking scores are on the 0–100 scale described in Section 3.5.3. The ranking is useful for a quick overview of which process or QC variables recur near the top across multiple folds for the interpreted responses.

The evidence-supported predictors in Table 4.2 are a narrower set of specific predictors that consistently show high or medium evidence across the model families. These predictors are selected for detailed discussion, including their effect curves and potential process implications. The table can also include contextual operating-regime markers, such as `LineID`, `ProductNumber`, and `RollPosition`, whereas the



(a) Predicted-versus-observed values by line.

(b) Standardised residual distribution.



(c) Standardised residuals by line.

(d) Standardised residuals by product.

Figure 4.9: held-out diagnostics for the SCADA-only XGBoost thickness log variance model, which achieved test $R^2 = 0.449$ in Table 4.1. Predicted-versus-observed values are labelled by line, while the standardised residuals are labelled both by line and by product number. The model consistently overpredicts the log variance, and the residuals show structure by line and product.

ranking figure removes these to focus on process and QC variables.

It is important to remember that many predictors are correlated and form groups of predictors describing similar information in the production process. Due to the anonymised data and the complexity of the production process, it is left to the production experts at Mölnlycke to interpret the predictors using their domain knowledge.

In Figure 4.10 we see that Ag [mg/cm^2] is dominated by QC thickness and density related predictors; thickness mean is dominated by flow line 1 and reactor process variables. For thickness log variance the top ranking predictors are more distributed but `A-S` and `B-C` are the most consistent top ranking variables. Notably, `A-S` is consistently important for both thickness mean and thickness log variance, which suggests that it is a key process variable for both the mean thickness and the thickness variation within rolls.

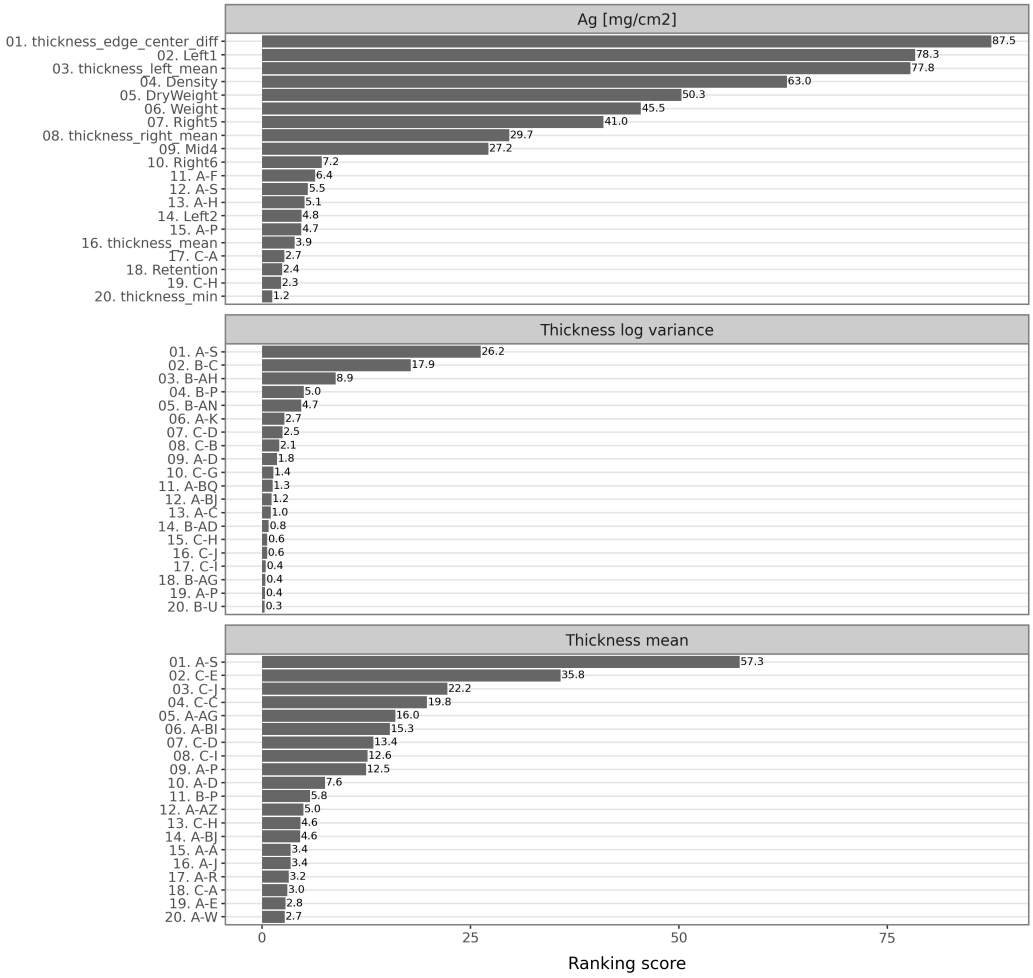


Figure 4.10: Top 20 grouped predictor ranking for the interpreted responses. Repeated SCADA windows and summary statistics are collapsed into simplified predictor labels before scoring, so the figure highlights recurring predictor families rather than individual predictors. Context variables are excluded.

In Table 4.2 we see that context variables such as `LineID`, `ProductNumber`, and

RollPosition were consistently important, but they are interpreted as different operating regimes rather than as process variables with standalone effect curves. Thus, no further interpretation is given for these contextual predictors. For the remaining predictors we discuss them below.

Table 4.2: Evidence-supported predictors selected for interpretation in the full-dataset models. Evidence is summarised by model family using the medium and high evidence tiers.

Response	Predictor	Source	Evidence level
Ag [mg/cm^2]	RollPosition	Context	Ridge high; XGBoost high
Ag [mg/cm^2]	thickness_edge_center_diff	QC	XGBoost high
Ag [mg/cm^2]	thickness_left_mean	QC	XGBoost high; Ridge medium
Ag [mg/cm^2]	Density	QC	XGBoost high; Ridge medium
Ag [mg/cm^2]	Mid4	QC	XGBoost high
Ag [mg/cm^2]	Left1	QC	Ridge medium; XGBoost medium
Thickness mean	LineID	Context	EBM high; Ridge high; XGBoost high
Thickness mean	ProductNumber	Context	EBM high; Ridge high; XGBoost high
Thickness mean	FL1__full__A-S__mean	Flow line 1	XGBoost high; EBM medium; Ridge medium
Thickness mean	Reactor__full__C-D__mean	Reactor	EBM medium; Ridge medium
Thickness log variance	ProductNumber	Context	Ridge high; XGBoost high
Thickness log variance	LineID	Context	XGBoost high; Ridge medium
Thickness log variance	FL2__early__B-C__mean	Flow line 2	XGBoost high
Thickness log variance	FL2__B-AN__mean family	Flow line 2	Ridge medium across several windows

4.3.1 Ag [mg/g] and Ag [mg/cm^2]

For Ag [mg/g], no predictor interpretation is reported. This is because the predictive performance was very weak, with mean CV R^2 values close to zero and negative for many model configurations. The lack of predictive signal means that any interpretation of predictors would not be supported by the model's ability to predict the response.

On the other hand, Ag [mg/cm^2] had moderate predictive performance in the development set, but the held-out test performance was poor. The predictor rankings are therefore only exploratory. The strongest predictors are all QC variables related to thickness or density. The interpretation of these predictors is that they capture the thickness and density structure that is associated with the silver area-density during the development period, but this structure does not transfer to the later production period due to the upward shift in silver levels.

The SCADA predictors in Figure 4.10 are much weaker than the QC predictors for this response. The highest ranked SCADA groups are A-F, A-S, A-H, and A-P, with ranking scores under 10, meaning they are consistently in the bottom 10% of the top 12 predictor rankings.

Thus, the ranking and evidence-supported predictors agree that the predictions of Ag [mg/cm^2] are driven by QC variables and roll position in the fitted models, while the SCADA signal is too weak and unstable to emphasize.

Due to the poor held-out performance of the Ag [mg/cm^2] models, their interpretation is treated with caution. ALE plots and ridge regression coefficients for the QC predictors, computed on the held-out data, are therefore provided in Appendix B as exploratory results rather than as final process findings.

4.3.2 Thickness mean

Among the SCADA predictors for the thickness mean response, `FL1__full__A-S__mean` and `Reactor__full__C-D__mean` were the predictors with the highest evidence of being associated with the response. The FL1 A-S mean predictor was high evidence in XGBoost and medium evidence in EBM and ridge regression models, while the Reactor C-D mean predictor was medium evidence in EBM and ridge regression models.

The ridge coefficients for `FL1__full__A-S__mean` and `Reactor__full__C-D__mean` are -0.0091 and -0.0102, respectively (Table B.2), so higher values of these predictors are associated with lower predicted thickness mean in the linear model. Figure 4.11 shows that the ALE and EBM shape plots for FL1 A-S mean both indicate a negative association with thickness mean. For Reactor C-D mean, the ALE plot is not included since there was no evidence for this predictor in the XGBoost model, but the shape plot shows a negative association for values near the bottom third of the observed range. The size of the ridge coefficients and the contribution difference between the extreme values of the shape plots have similar magnitudes, which suggests that the different model families are capturing similar associations.

The ranking supports the same conclusion but also shows a broader pattern. The highest ranked process variables come from `A-S`, `C-E`, `C-J`, `C-C`, and `A-AG`, with `A-S` having a score of 57.3, which means it was consistently slightly above the middle of the top 12 rankings.

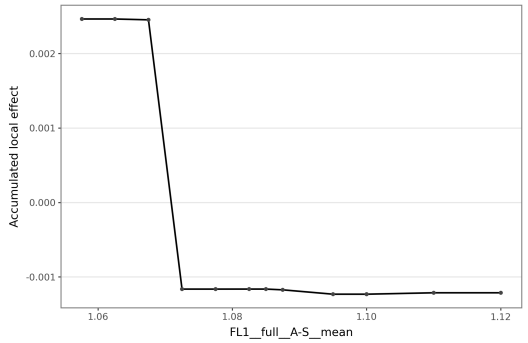
4.3.3 Thickness log variance

For thickness log variance, `FL2__early__B-C__mean` and the `FL2__B-AN__mean` family, that is the `B-AN` mean predictors across all four time windows, were the predictors with the strongest evidence of being associated with the response. The FL2 B-C early mean predictor was high evidence in XGBoost and medium evidence in ridge regression, while the FL2 B-AN mean predictors were medium evidence across several windows in ridge regression but not in XGBoost. Since the EBM models had negative held-out R^2 for this response, they are not used for interpretation, so there is no EBM shape plot for these predictors.

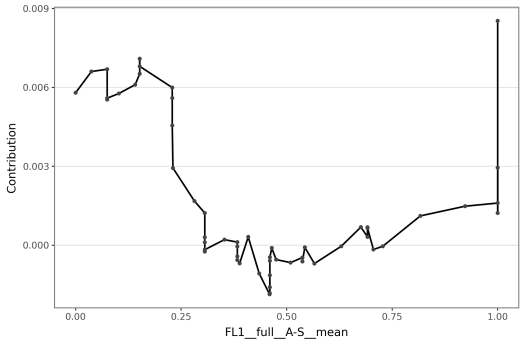
The ridge coefficient for the FL2 `B-C` early mean predictor is -0.0028 and all of the FL2 `B-AN` mean predictors have approximately -0.15 as their ridge coefficients, see Table B.2. Thus, higher values of these predictors are associated with lower predicted thickness log variance in the linear model.

In the ALE plots for the XGBoost model, we see the opposite trend where the `B-C` predictor has larger effect size than the `B-AN` predictors. The ALE plots similarly show a negative association with larger values of the predictors. Thus, the ridge regression coefficients and the XGBoost ALE plots point in the same broad direction, which suggests that higher values of these predictors are associated with lower thickness log variance.

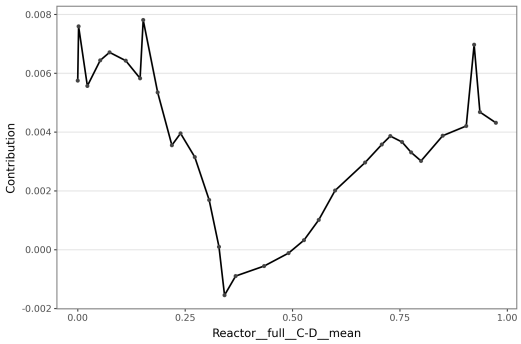
The predictor ranking partly agrees with the evidence-supported predictors. The highest ranked process variable is `A-S`, with a score of 26.2, followed by `B-C` with



(a) FL1 A-S ALE.



(b) FL1 A-S EBM.

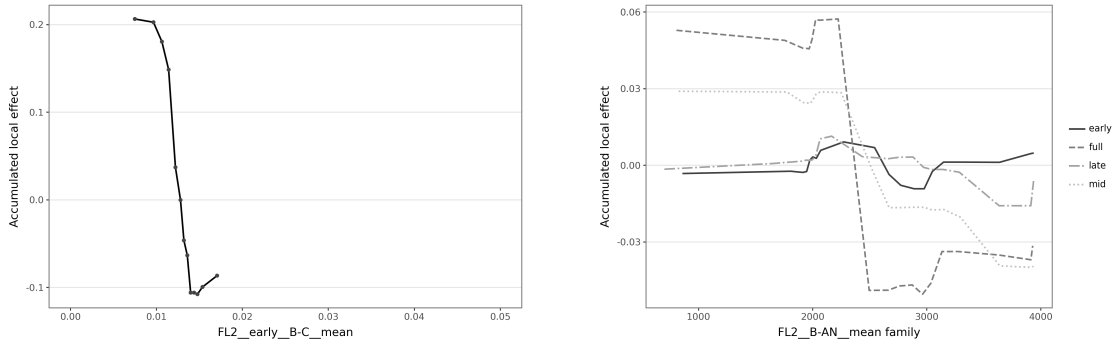


(c) Reactor C-D EBM.

Figure 4.11: ALE and EBM effect diagnostics for the evidence-supported process predictors retained for thickness mean in Table 4.2. The `FL1__full__A-S__mean` predictor is shown with both the XGBoost ALE curve and the EBM shape function, while `Reactor__full__C-D__mean` is shown with its EBM shape function because its evidence came from EBM and ridge regression rather than XGBoost. The corresponding ridge coefficients after min max preprocessing are -0.0091 and -0.0102, respectively.

17.9. The next ranked groups, B-AH, B-P, and B-AN, have lower scores between about 4.7 and 8.9.

Thus, the log variance interpretation points towards B-C and B-AN being important process variables on flow line 2, and also shows some support for A-S being an important process variable on flow line 1.



(a) B-C ALE. **(b)** B-AN ALE.

Figure 4.12: ALE diagnostics for the B-C predictor and the B-AN predictor family for thickness log variance. The FL2__early__B-C__mean ALE curve is zoomed from 0 to 0.05, and its ridge coefficient after is −0.0028. For the FL2__B-AN__mean family, the ridge coefficients are negative and similar across windows: early −0.1467, mid −0.1549, late −0.1488, and full −0.1501.

4.4 Thickness results for subset-data models

As described in the methods, the line- and product-specific thickness models were motivated by the residual diagnostics for the full-dataset thickness models. The subset-data models are not part of the main comparison, but they are a secondary follow-up analysis that explores whether more homogeneous subsets, such as limiting observations to one flow line, of the data can give better predictions and more specific interpretations.

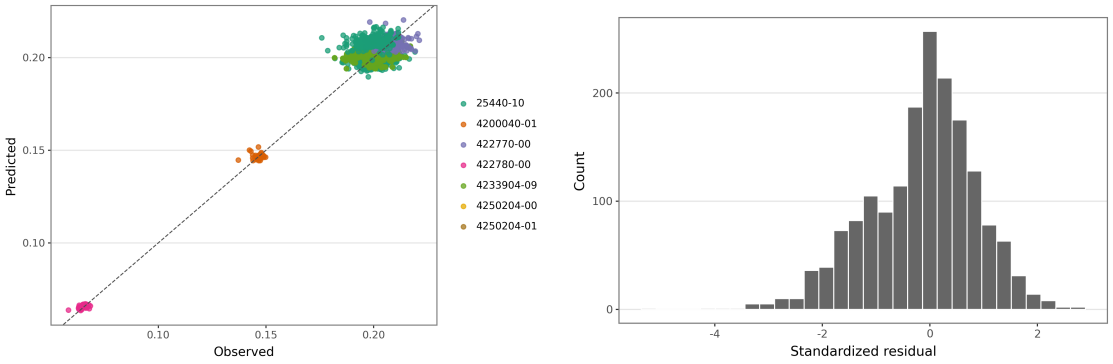
Due to time constraints, the subset-data models were only fitted using the SCADA-only XGBoost models, which were the best performing models for thickness mean and the second best for thickness log variance. The subset-data models are therefore not compared with other model families or feature sets.

Table 4.3 reports both development and held-out metrics. All combinations of line and product subsets were fitted, but almost all of these had negative development and held-out R^2 values, so they are not included in the table. The line-specific thickness-mean models remain strong, with test $R^2 = 0.932$ for line 1 and $R^2 = 0.966$ for line 2. The line 2 and product 422780-10 thickness mean model was the only product subset model with positive development R^2, but it had negative held-out R^2. The thickness log variance models had much weaker predictive performance, with only the line 1 model having positive R^2.

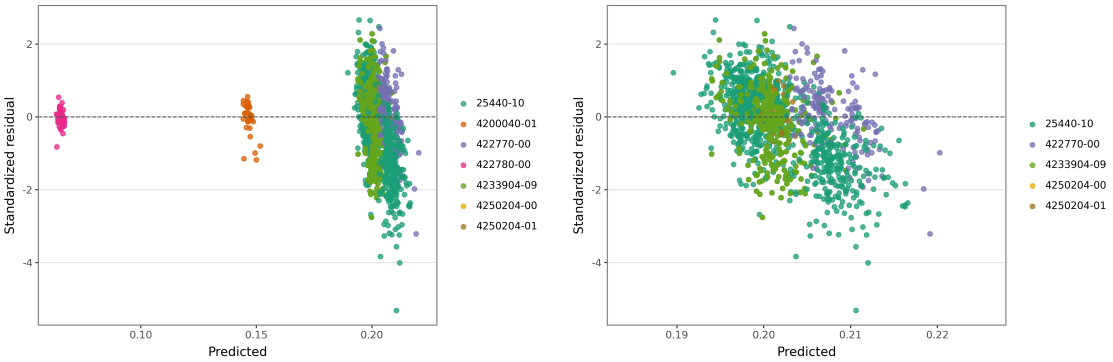
Table 4.3: Mean CV metrics on the development set and held-out test metrics for SCADA-only XGBoost thickness models using subsets of the data. Additional subset combinations gave negative R^2 and are therefore not included.

Response	Data subset	CV RMSE	CV MAE	CV R^2	Test RMSE	Test MAE	Test R^2	Test n
Thickness mean	Line 1	0.0074	0.0059	0.917	0.0066	0.005	0.932	1732
Thickness mean	Line 2	0.0046	0.0023	0.861	0.0028	0.0019	0.966	1619
Thickness mean	Line 2, product 422780-10	0.0023	0.0019	0.125	0.0027	0.0022	-0.062	633
Thickness log variance	Line 1	0.7469	0.5843	0.172	0.6368	0.5016	0.289	1732
Thickness log variance	Line 2	0.7320	0.5876	-0.093	0.9644	0.7808	-0.729	1619
Thickness log variance	Line 2, product 422780-10	0.8353	0.6733	-0.659	0.8339	0.6891	-0.802	633

The held-out diagnostics for the two line-specific thickness-mean models are shown in Figures 4.13 and 4.14. Both line specific models have lower held-out RMSE than the SCADA-only XGBoost model using the full dataset: 0.0066 for line 1 and 0.0028 for line 2, compared with 0.0071 for the full-dataset model. While the performance improved there still remains residual structure from different product regimes. The corresponding diagnostics for the thickness log variance and product 422780-10 models are provided in Appendix B.



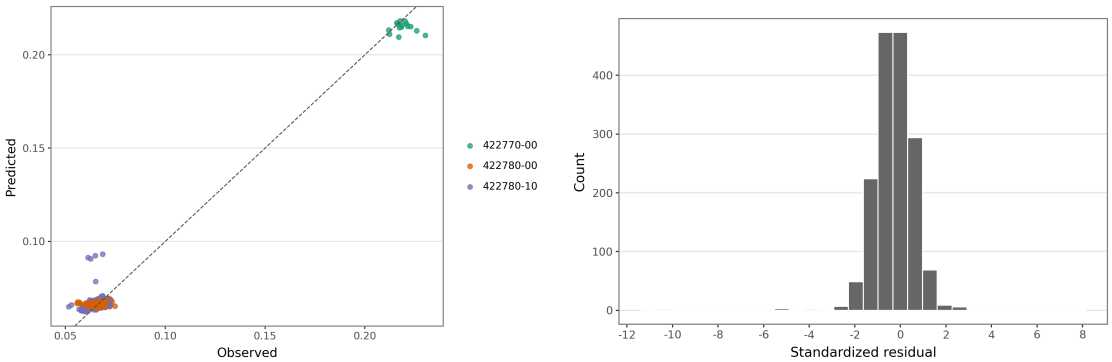
(a) Predicted-versus-observed by product. **(b)** Standardised residual distribution.



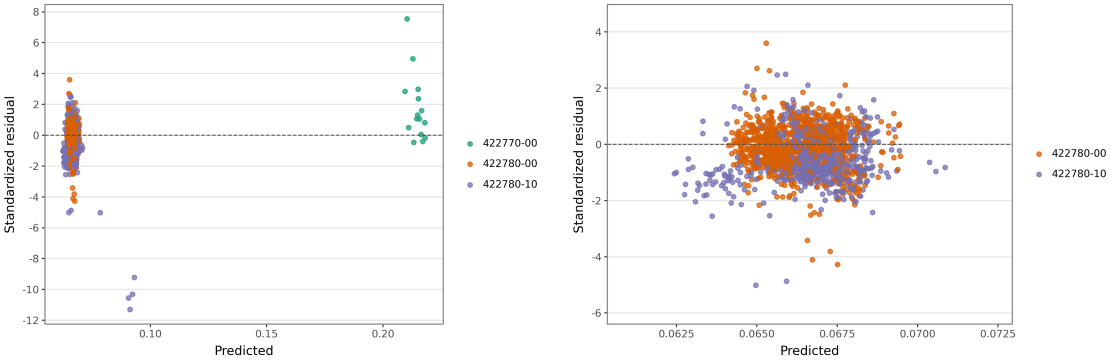
(c) Standardised residuals by product. **(d)** Standardised residuals by product, zoomed view.

Figure 4.13: held-out diagnostics for the Line 1 thickness mean XGBoost model using a subset of the data. The zoomed residual panel shows the region containing most observations.

Figure 4.16 and Table 4.4 show the predictor rankings and evidence-supported



(a) Predicted-versus-observed by product. **(b)** Standardised residual distribution.



(c) Standardised residuals by product.

(d) Standardised residuals by product, zoomed view.

Figure 4.14: held-out diagnostics for the Line 2 thickness mean XGBoost model using a subset of the data. The zoomed residual panel shows the region containing most observations.

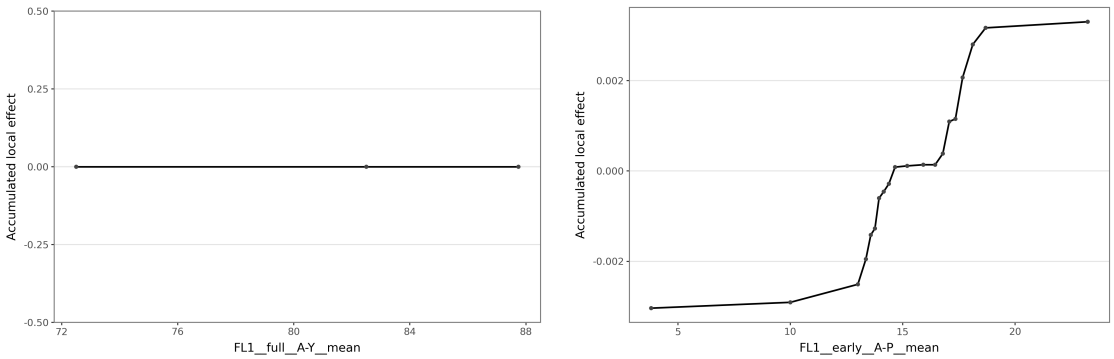
predictors for the subset-data models, respectively. Since the line 1 model has the strongest performance and is the only model which has high evidence predictors, namely `FL1__full__A-Y__mean` and `FL1__early__A-P__mean`, we only include ALE plots for these predictors and leave the other predictors to Appendix B.

The predictor rankings in Figure 4.16 show that subset-data models can give different predictor rankings than the full-dataset models. For example, the line 1 thickness mean model has a very different ranking than the full-dataset thickness mean model, with the top predictors coming from `A-AJ`, `A-P`, and `A-Y` instead of `A-S` and `C-E`. This shows that the subset-data models are capturing different associations than the full-dataset models, which could be due to them fitting specific associations within more homogeneous subsets of the data or because the model uses another process variable to predict the response which is related to the ones used in the full-dataset models.

Table 4.4: Evidence-supported predictors selected for interpretation in the subset-data thickness models. All rows come from XGBoost SCADA-only models using subsets of the data.

Response	Scope	Predictor	Source	Evidence level
Thickness mean	Line 1	`FL1__full__A-Y__mean`	SCADA	XGBoost high
Thickness mean	Line 1	`FL1__early__A-P__mean`	SCADA	XGBoost high
Thickness mean	Line 2	`FL2__early__B-BA__mean`	SCADA	XGBoost medium
Thickness mean	Line 2, product 422780-10	`FL2__early__B-BA__mean`	SCADA	XGBoost medium
Thickness log variance	Line 1	`FL1__early__A-I__mean`	SCADA	XGBoost medium
Thickness log variance	Line 1	`FL1__late__A-BL__mean`	SCADA	XGBoost medium

Figure 4.15 shows the ALE plots for the two high evidence predictors in the line 1 thickness mean model. Although `A-Y` showed high evidence of association with thickness mean, its ALE plot is constant, highlighting the uncertainty of importance evidence based on a single model family. In contrast, the ALE plot for `A-P` indicates a positive association at higher predictor values.



(a) A-Y ALE. **(b)** A-P ALE.

Figure 4.15: ALE curves for the high evidence Line 1 thickness mean predictors in Table 4.4. The curves summarise the XGBoost fitted effects for `FL1__full__A-Y__mean` and `FL1__early__A-P__mean`. Corresponding ridge regression and EBM effect analyses were not produced for these XGBoost only subset data analyses.

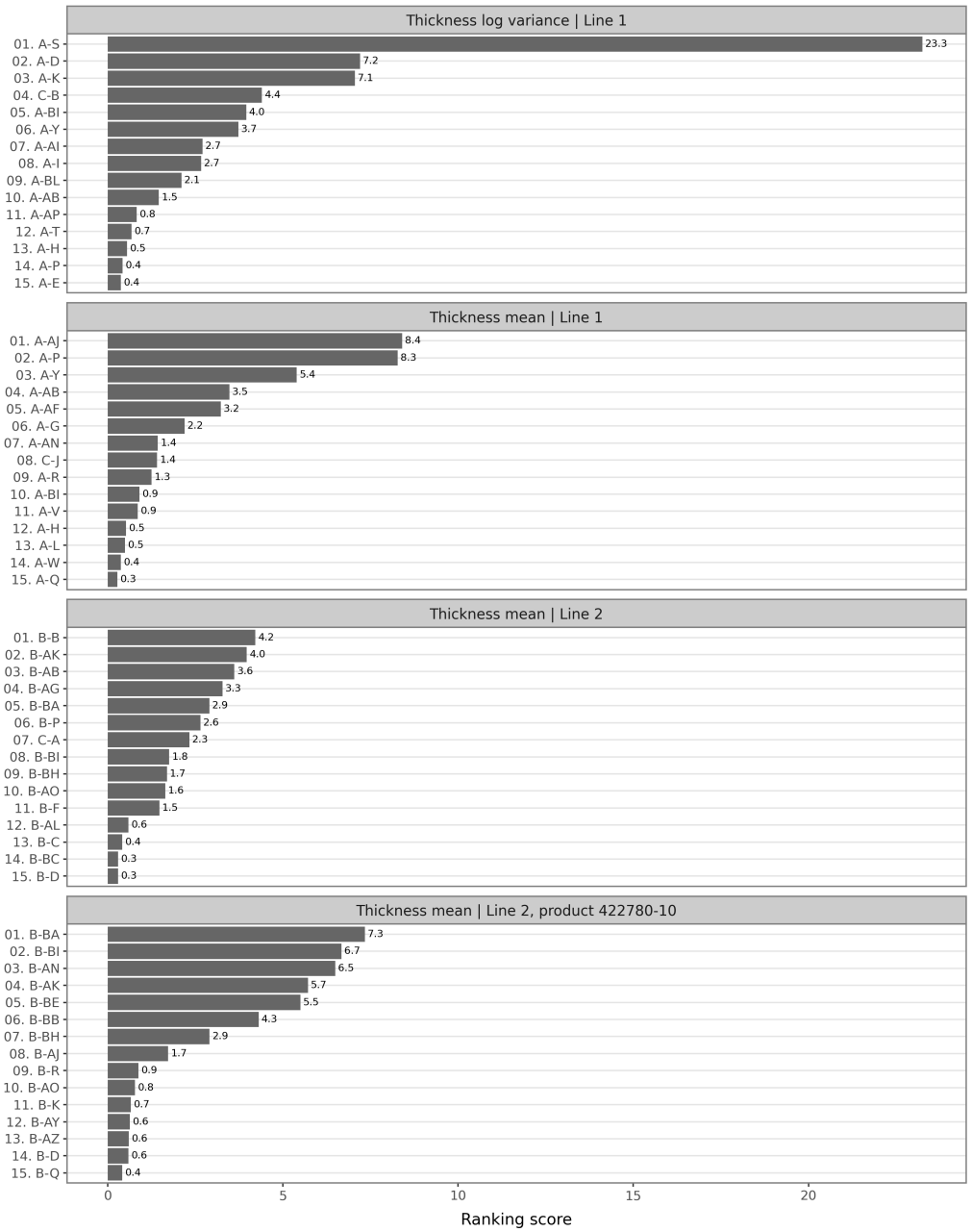


Figure 4.16: Top 15 grouped predictor rankings for the selected SCADA-only XGBoost thickness models using subsets of the data. The ranking highlights which simplified SCADA predictor families recur within the line and product specific thickness models.

5

Discussion and Conclusions

5.1 Main findings

The first research question asked how predictable the four responses were from the available data. The answer depends strongly on the response. Thickness mean was predicted very well from the SCADA-only all mean summaries. Thickness log variance was harder to predict, but it was still meaningfully predictable. In contrast, Ag [mg/g] was not usefully predictable, and Ag [mg/cm^2] only showed moderate development-period signal that did not transfer to the later held-out period.

The second research question asked which data sources contained the most predictive information. For thickness, the useful signal mainly came from SCADA summaries. For Ag [mg/cm^2], the fitted structure was mainly linked to QC profile variables, density, and roll position. The SCADA signal for this response was much weaker. For Ag [mg/g], no data source gave reliable predictive performance.

The third research question asked which predictors had the strongest and most stable influence on the fitted models. For thickness mean, the clearest evidence points to line and product context together with recurring Flow Line 1 and reactor SCADA summaries, especially `FL1__full__A-S__mean` and `Reactor__full__C-D__mean`. For thickness log variance, the evidence was weaker, but `FL2__early__B-C__mean` and the `FL2__B-AN__mean` family provide useful hypotheses for process review. These results should be read as stable model associations, not as causal effects.

The fourth research question asked whether thickness models using subsets of the data showed the same interpretation patterns as the full-dataset thickness models. The line-specific thickness mean models remained strong and gave different predictor rankings than the full-dataset model. This suggests that narrower line level models can reveal process associations that are partly hidden when both lines are modelled together. The product specific thickness mean model and the subset-data thickness log variance models were much less reliable, so those results should be treated as exploratory.

5.2 Practical implications

The weak Ag [mg/g] result should not only be seen as a negative finding. According to the production experts, silver concentration is mainly governed by the chemical process and is kept as consistent as possible. It was therefore not surprising that the

production steps studied here did not explain much of the variation in Ag [mg/g]. In this sense, the result supports the view that these production steps are unlikely to be the main drivers of silver concentration.

The Ag [mg/cm^2] result is different. The strongest predictors were QC thickness, density, and roll position variables. This is also not surprising, because operators already use QC information to understand what needs to be changed to reach the desired silver area-density. The model is therefore consistent with the fitted silver area-density signal being mostly connected to finished roll quality measurements. However, because the model did not transfer to the held-out period, it should not be used for future-period silver prediction.

The strongest practical result is the thickness modelling. Thickness mean could be predicted very well from SCADA summaries alone, and thickness log variance was also partly predictable. This means that the process data contains information about both the average thickness of the roll and the uniformity of the foam. Since thickness contributes to silver area-density, better understanding of thickness may also help process experts understand how silver area-density varies.

The log variance result is especially useful from a production perspective. A lower thickness log variance means a more uniform roll. If the process variables linked to lower log variance can be understood by process experts, this may help reduce variation, waste, and rework. The results are not enough to define control rules directly, but they give a smaller set of process variables and production contexts to review.

All practical implications should be interpreted cautiously. The SCADA tags are anonymised, many predictors are correlated, and this thesis has limited domain expertise compared with the production experts. The models should therefore be used as support for process review rather than as direct evidence that changing one variable will change the response.

5.3 Strengths, weaknesses, and limitations

A main strength of the thesis is that model selection, validation, and interpretation follow the intended prediction setting. The final test set is later in time than the development data, and splitting is performed at the roll level. This keeps related observations in the same split and gives a more realistic estimate of future period performance than random splitting.

Another strength is the comparative interpretation design. The thesis compares ridge coefficients, EBM shape functions, ALE curves, fold stability, permutation rankings, held-out confirmation, and shuffled-response baselines. This is useful in observational process data, where correlated predictors make isolated importance scores difficult to interpret. It also makes the interpretation less dependent on one model family and one data split.

One limitation is that the study is observational. The interpretation results show fitted associations, not causal effects. This is especially important because many predictors are correlated and may describe the same operating state. The anonymised

SCADA tags also mean that the results cannot be translated into physical process explanations without input from production experts.

The silver results have an additional limitation. The silver measurements shifted upwards in the held-out period, and the fitted models did not capture this shift. The silver models should therefore not be used for operational prediction. The Ag [mg/g] result is also limited by the fact that the available predictors may not describe the chemical factors that mainly control silver concentration.

The main thickness limitation is that even strong models retain line- and product-related residual structure. Flow Line 1 and flow line 2 SCADA summaries are kept as separate anonymised predictor families because their tag sets cannot be treated as one-to-one equivalents. Consequently, the model can identify important predictors for each flow line, but cannot directly compare their importance between the two flow lines.

5.4 Future work

Future work should first extend the production history. More time coverage would make it easier to test whether thickness predictors remain stable and whether the Ag [mg/cm^2] drift can be explained by variables not included here or by having more data. If silver prediction remains a goal, future work should also include more direct variables from the chemical preparation, if such data are available.

Second, thickness modelling should become more explicitly line aware. A useful comparison would be between the present full-dataset feature set, separate line-specific feature sets, and a smaller harmonised feature set containing only process groups with meaningful analogues on both flow lines. Product specific modelling should be delayed until enough observations exist within product or product family groups to support genuine held-out comparisons.

Finally, interpretation should be connected to process knowledge earlier. Because the process tags are anonymised, selected predictors should be reviewed with operators and process engineers before they are treated as possible control levers. This would help separate variables that only mark an operating state from variables that can actually be changed in a useful way.

5.5 Conclusions

The thesis shows that routinely collected production data is most useful for thickness modelling. SCADA summaries can predict thickness mean very well and thickness log variance to a useful degree. This gives process experts a practical starting point for reviewing which production conditions are linked to roll thickness and foam uniformity.

The silver results are more limited but still informative. Ag [mg/g] was not explained by the production steps studied here, which is consistent with silver concentration being mainly controlled by the chemical process. Ag [mg/cm^2] was mainly linked

to QC thickness, density, and roll-position information, but it was not reliable for future-period prediction.

Overall, the models are best viewed as tools for process understanding and process-review prioritisation. They do not prove which settings should be changed, but they identify responses, data sources, and predictor groups where further expert review is most likely to be useful.

Bibliography

[1] Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN 978-0387-31073-2.

[3] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[4] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009.

[6] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–158. ACM, 2012. doi: 10.1145/2339530.2339556.

[7] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 623–631. ACM, 2013.

[8] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[9] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles, 2019.

[10] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Christoph Molnar, Munich, Germany, 2 edition, 2022. ISBN 9798411463330.

[11] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. Adaptive Computation and Machine Learning. The MIT Press, 2023. ISBN 9780262048439.

[12] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability, 2019.

A

Software implementation

The computational work in the project was implemented in Python. The code was developed as a version-controlled project containing the modelling workflow, experiment definitions, exploratory analyses, and reporting utilities used to produce the results in the thesis. The purpose of this appendix is therefore only to document the general software setting, rather than to describe the implementation in detail.

The analysis relied on standard Python packages for tabular data processing, modelling, experiment tracking, model interpretation, and plotting. Data handling was mainly carried out with packages such as `pandas`, `numpy`, `pyarrow`, `joblib`, `tqdm`, and `pyyaml`. `scikit-learn` was used for the preprocessing pipelines, several model families, evaluation utilities, and permutation-importance diagnostics, together with packages including `xgboost`, `lightgbm`, and `interpret`. Additional model interpretation and figure generation used `shap`, `plotnine`, and `matplotlib`.

The experiments were defined so that the response variable, predictor set, preprocessing, model family, and hyperparameters were explicit for each run. The same preprocessing logic was used across the modelling workflow: transformations were fitted within the relevant training data before being applied to validation folds or the held-out test period. This was done to keep the development-set evaluation and the final held-out evaluation consistent.

MLflow, a tool for tracking model runs, was used together with stored run outputs. For each run, the modelling configuration, performance metrics, predictions, and diagnostic outputs were recorded, allowing the reported tables and figures to be traced to the corresponding model specification.

Git was used for version control, and the code was stored in a Bitbucket repository accessible to Mölnlycke. The repository contains the implementation needed to inspect the modelling workflow and rerun the experiments on new data.

B

Supporting results

This appendix collects selected supporting material for the results. The figures shown here clarify the main findings or provide additional diagnostics without interrupting the results narrative.

B.1 Selected model hyperparameters

Table B.1 reports the explicit hyperparameters for the final selected models. The subset-data models used the same hyperparameter settings as the corresponding full-dataset models. All models used the common preprocessing and time-based split workflow described in the Methods chapter.

Table B.1: Explicit hyperparameters for the main reported models.

Response	Model	Feature set	Explicit hyperparameters
Ag [mg/g]	XGBoost	combined, all mean	`learning_rate=0.08`, `n_estimators=50`, `max_depth=7`, `subsample=0.8`, `colsample_bytree=0.8`
Ag [mg/cm^2]	XGBoost	combined, all mean	`learning_rate=0.03`, `n_estimators=150`, `max_depth=7`, `subsample=0.8`, `colsample_bytree=0.8`
Ag [mg/cm^2]	Ridge	combined, all mean	`alpha=200.0`
Thickness mean	XGBoost	SCADA-only, all mean	`learning_rate=0.05`, `n_estimators=300`, `max_depth=5`, `subsample=0.8`, `colsample_bytree=0.8`
Thickness mean	Ridge	SCADA-only, all mean	`alpha=0.316228`

Response	Model	Feature set	Explicit hyperparameters
Thickness mean	EBM	SCADA-only, all mean	`interactions=0`, `outer_bags=4`, `inner_bags=0`, `learning_rate=0.04`, `max_rounds=1200`, `max_leaves=4`, `validation_size=0.2`, `random_state=42`
Thickness log variance	XGBoost	SCADA-only, all mean	`learning_rate=0.03`, `n_estimators=150`, `max_depth=7`, `subsample=0.8`, `colsample_bytree=0.8`
Thickness log variance	Ridge	SCADA-only, all mean	`alpha=1000.0`
Thickness log variance	EBM	SCADA-only, all mean	`interactions=0`, `outer_bags=4`, `inner_bags=0`, `learning_rate=0.04`, `max_rounds=1200`, `max_leaves=4`, `validation_size=0.2`, `random_state=42`

B.2 Development set model screening

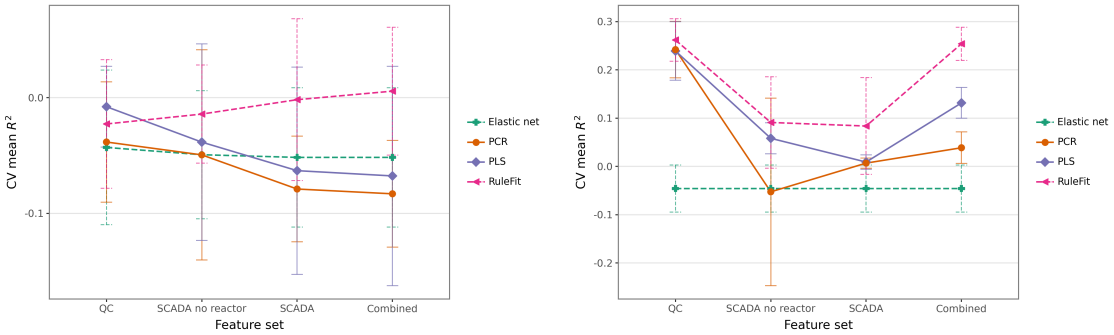
Figure B.1 collects the development set model comparison panels for additional candidate model families: elastic net, PCR, PLS, and RuleFit. These model and feature set combinations were evaluated during model development, but were not carried into the final interpretation because they performed similarly to simpler retained alternatives, did not add a clear predictive advantage, or were harder to interpret. The final reporting therefore focuses on ridge regression, random forest, XGBoost, and EBM.

B.3 Exploratory Ag area-density interpretation

The Ag [mg/cm^2] effect plots are treated as supporting material because the final Ag [mg/cm^2] models did not generalise well to the held-out test set. Although several QC-derived predictors met the evidence filters, their fitted effects are presented only as exploratory model associations and not as final process findings.

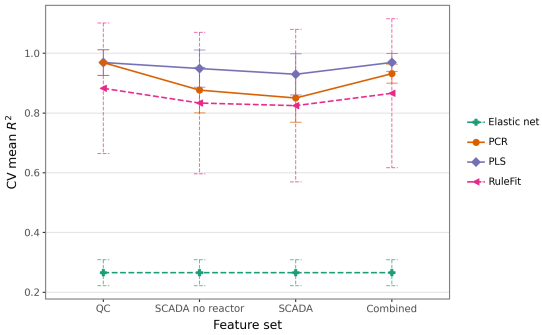
B.4 Selected ridge coefficients

The ridge regression evidence is reported numerically rather than as separate coefficient plots. The values in Table B.2 are coefficients after the common preprocessing pipeline. Since the final reported configurations use min-max scaling, these coeffi-

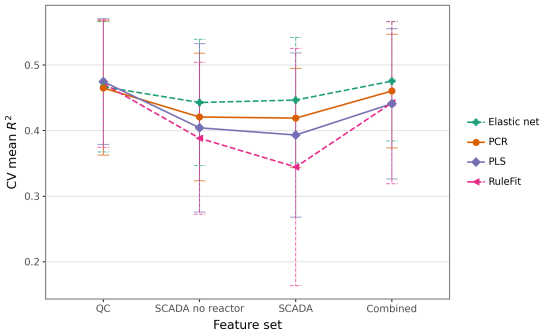


(a) Ag [mg/g].

(b) Ag [mg/cm^2].

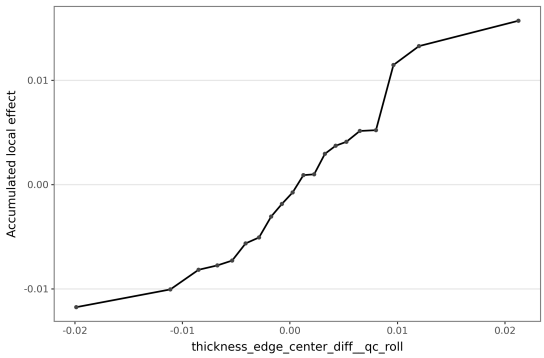


(c) Thickness mean.

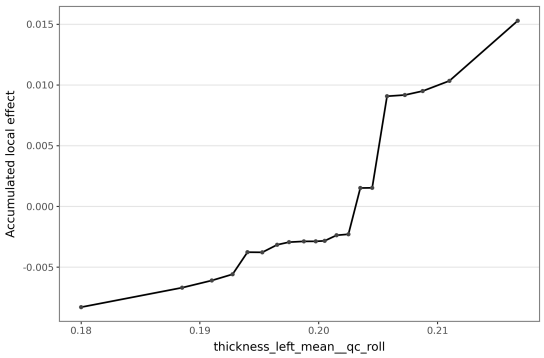


(d) Thickness log variance.

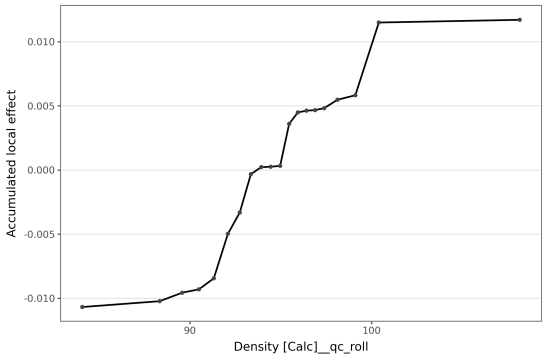
Figure B.1: Mean CV R^2 for additional candidate model families by response and feature set family on the development set using expanding window cross-validation. Error bars show approximate 95% confidence intervals for the fold mean. These comparisons document model screening rather than the final reported models.



(a) Thickness edge centre difference, XGBoost ALE.



(b) Left side thickness mean, XGBoost ALE.



(c) Density, XGBoost ALE.

Figure B.2: Exploratory ALE diagnostics for evidence-supported Ag [mg/cm^2] predictors in the combined XGBoost model: `thickness_edge_center_diff`, `thickness_left_mean`, and `Density`. Because the model showed poor held-out performance, these curves are exploratory model associations rather than final process findings.

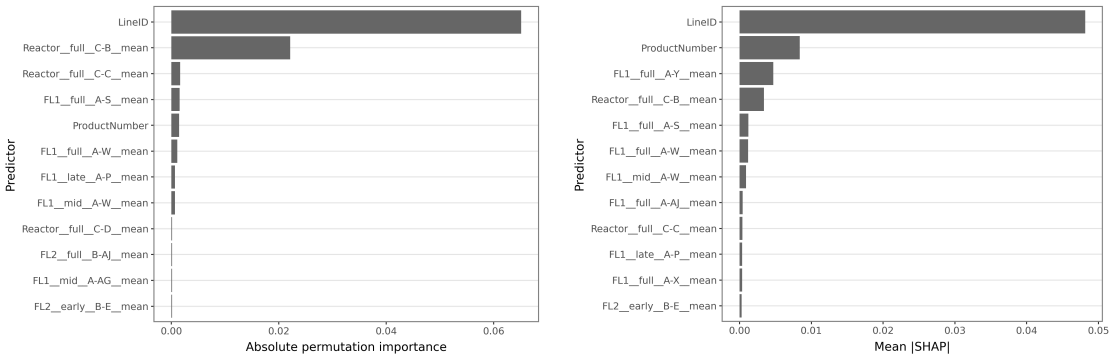
cients describe changes on the training-range-scaled predictor scale, conditional on the other predictors and the ridge penalty.

Table B.2: Selected ridge coefficients referred to in the Results chapter after preprocessing for the final reported ridge models.

Response	Predictor	Feature set	Coefficient	Sign consistency
Ag [mg/cm^2]	`Left1`	combined, all mean	+0.034	1.0
Ag [mg/cm^2]	`Density`	combined, all mean	+0.044	1.0
Ag [mg/cm^2]	`thickness_left_mean`	combined, all mean	+0.026	1.0
Thickness mean	`FL1__full__A-S__mean`	SCADA-only, all mean	−0.0091	0.8
Thickness mean	`Reactor__full__C-D__mean`	SCADA-only, all mean	−0.0102	0.6
Thickness log variance	`FL2__early__B-C__mean`	SCADA-only, all mean	−0.0028	1.0
Thickness log variance	`FL2__early__B-AN__mean`	SCADA-only, all mean	−0.1467	0.8
Thickness log variance	`FL2__mid__B-AN__mean`	SCADA-only, all mean	−0.1549	0.8
Thickness log variance	`FL2__late__B-AN__mean`	SCADA-only, all mean	−0.1488	0.8
Thickness log variance	`FL2__full__B-AN__mean`	SCADA-only, all mean	−0.1501	0.8
Thickness log variance	`Reactor__early__C-G__mean`	SCADA-only, all mean	−0.145	1.0

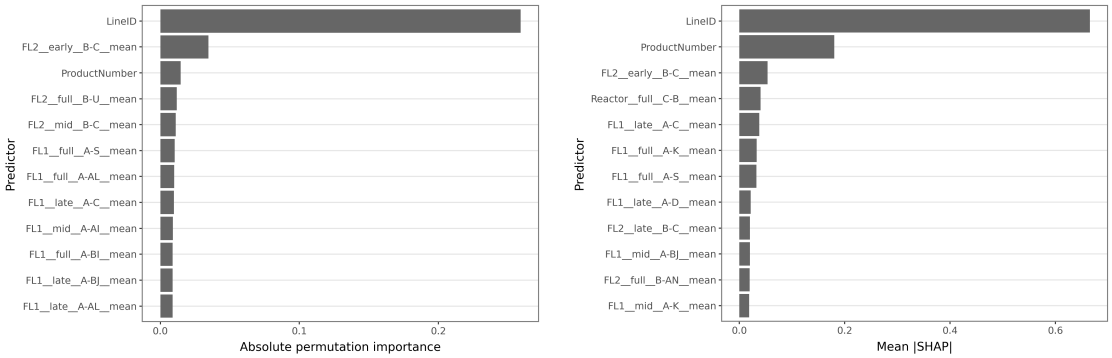
B.5 Thickness importance diagnostics

Figures B.3–B.8 give the top 12 XGBoost importance summaries for the full-dataset and subset-data thickness models. In each figure, held-out permutation importance is shown next to mean absolute SHAP importance from the fitted model. These supporting diagnostics show the model-specific importance rankings.



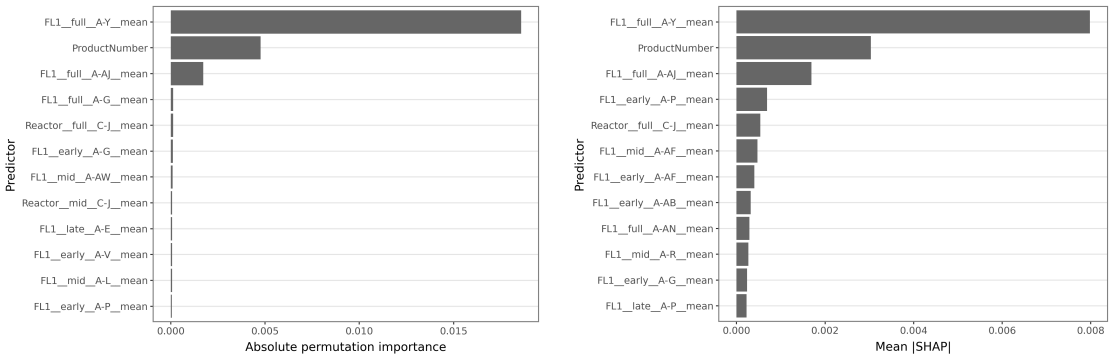
(a) Held-out permutation importance. **(b)** Mean absolute SHAP importance.

Figure B.3: Top 12 XGBoost importance diagnostics for the SCADA-only thickness mean model using the full dataset. Permutation importance is computed on the held-out test set, while SHAP importance summarises mean absolute feature contributions from the fitted model.



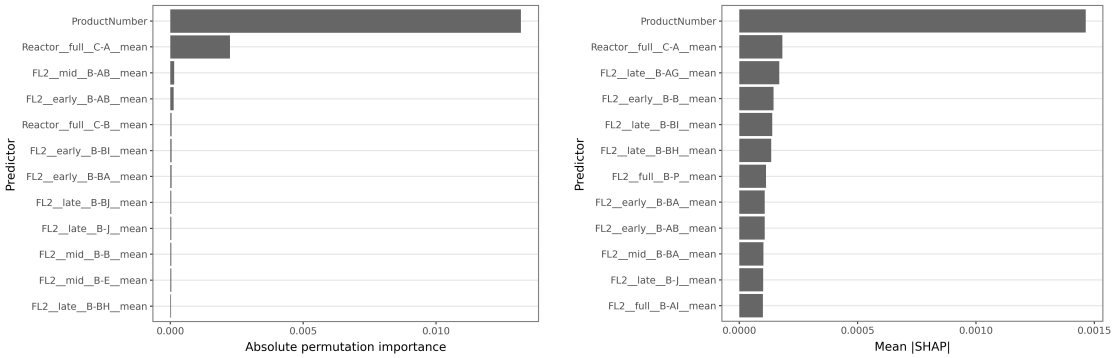
(a) Held-out permutation importance. **(b)** Mean absolute SHAP importance.

Figure B.4: Top 12 XGBoost importance diagnostics for the SCADA-only thickness log variance model using the full dataset. Permutation importance is computed on the held-out test set, while SHAP importance summarises mean absolute feature contributions from the fitted model.



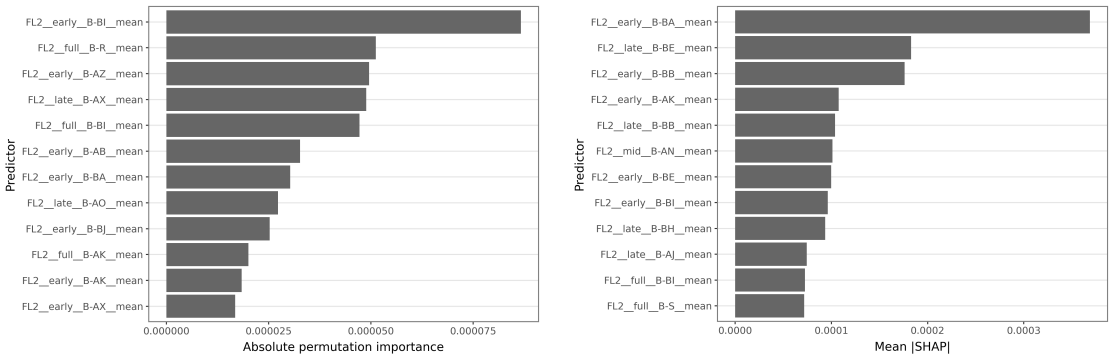
(a) Held-out permutation importance. **(b)** Mean absolute SHAP importance.

Figure B.5: Top 12 XGBoost importance diagnostics for the Line 1 SCADA-only thickness mean model using a subset of the data. Permutation importance is computed on the held-out test set, while SHAP importance summarises mean absolute feature contributions from the fitted model.



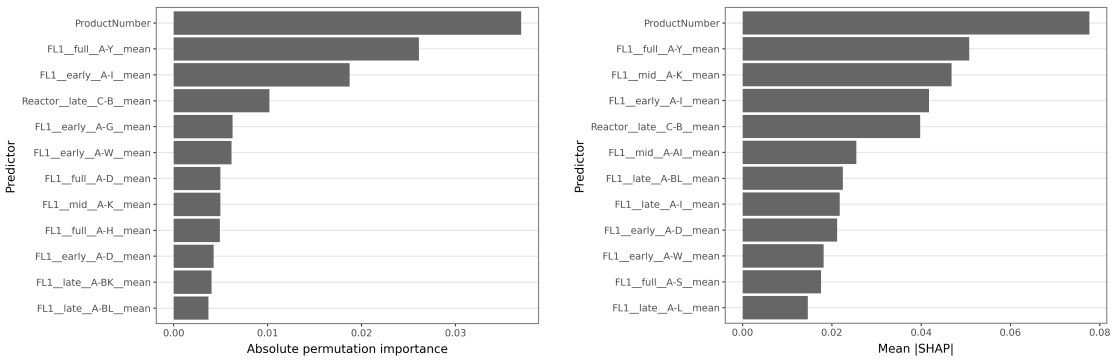
(a) Held-out permutation importance. **(b)** Mean absolute SHAP importance.

Figure B.6: Top 12 XGBoost importance diagnostics for the Line 2 SCADA-only thickness mean model using a subset of the data. Permutation importance is computed on the held-out test set, while SHAP importance summarises mean absolute feature contributions from the fitted model.



(a) Held-out permutation importance. **(b)** Mean absolute SHAP importance.

Figure B.7: Top 12 XGBoost importance diagnostics for the Line 2 selected product SCADA-only thickness mean model using a subset of the data. Permutation importance is computed on the held-out test set, while SHAP importance summarises mean absolute feature contributions from the fitted model.

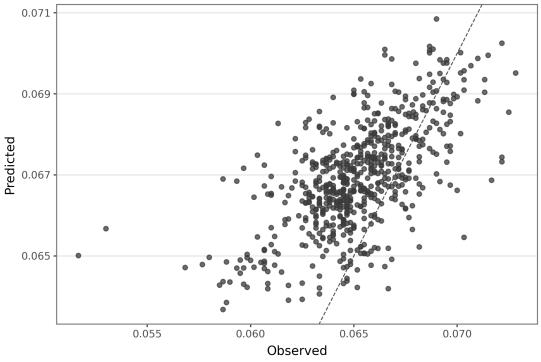


(a) Held-out permutation importance. **(b)** Mean absolute SHAP importance.

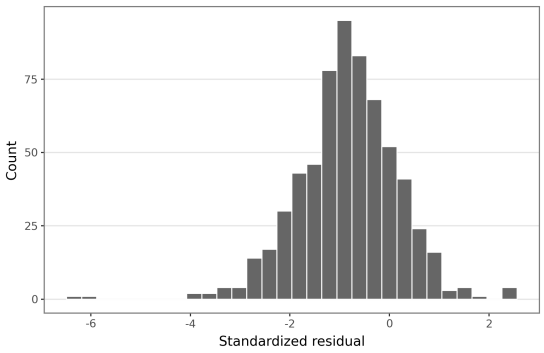
Figure B.8: Top 12 XGBoost importance diagnostics for the Line 1 SCADA-only thickness log variance model using a subset of the data. Permutation importance is computed on the held-out test set, while SHAP importance summarises mean absolute feature contributions from the fitted model.

B.6 Thickness subset-data model diagnostics

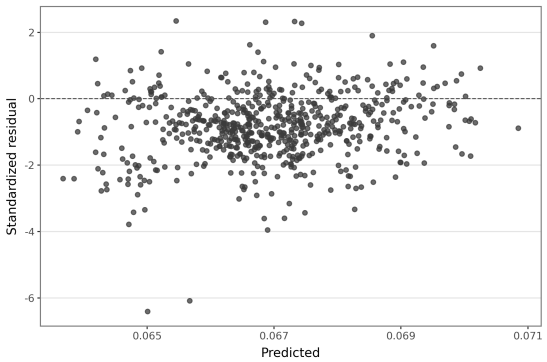
The subset-data metrics, predictor rankings, and line-specific thickness mean diagnostics are reported with the main subset-data results. The remaining held-out diagnostics cover the product-specific thickness mean model and the subset-data thickness log variance model. For each remaining subset-data model, the predicted-versus-observed plot, standardised residual histogram, and standardised-residual-versus-predicted plot are shown. Product labels are shown only when more than one product is visible in the plotted panel.



(a) Predicted-versus-observed.

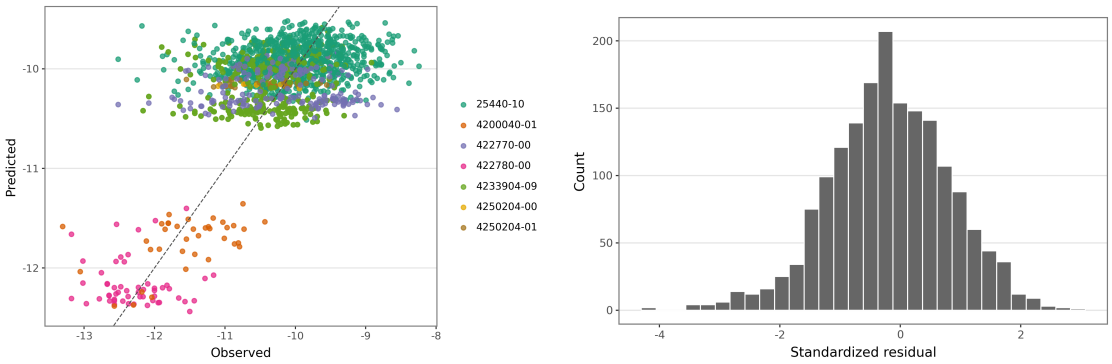


(b) Standardised residual distribution.

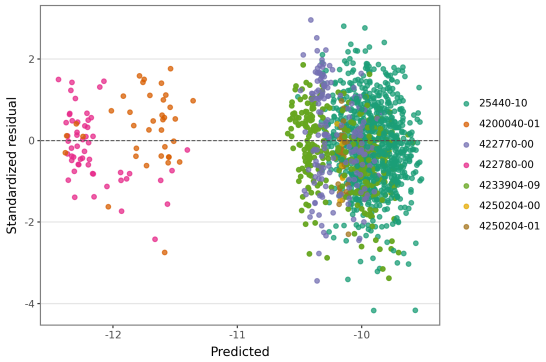


(c) Standardised residuals versus predicted.

Figure B.9: Held-out diagnostics for the Line 2 selected product thickness mean XGBoost model using a subset of the data.

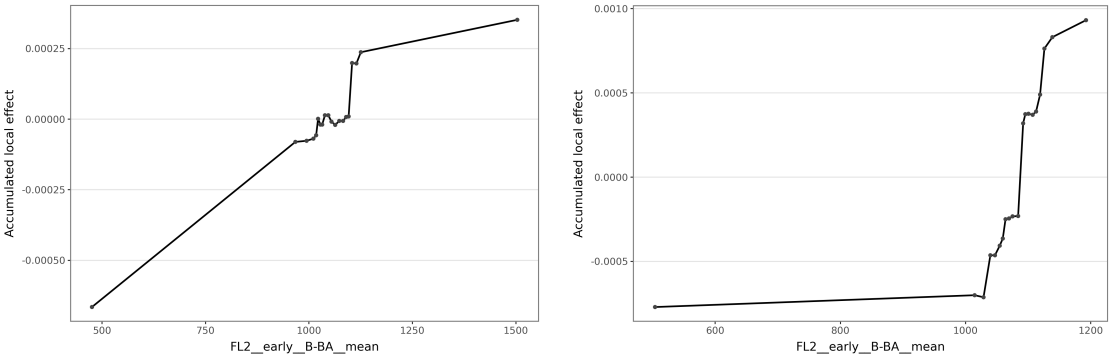


(a) Predicted-versus-observed by product. **(b)** Standardised residual distribution.



(c) Standardised residuals by product.

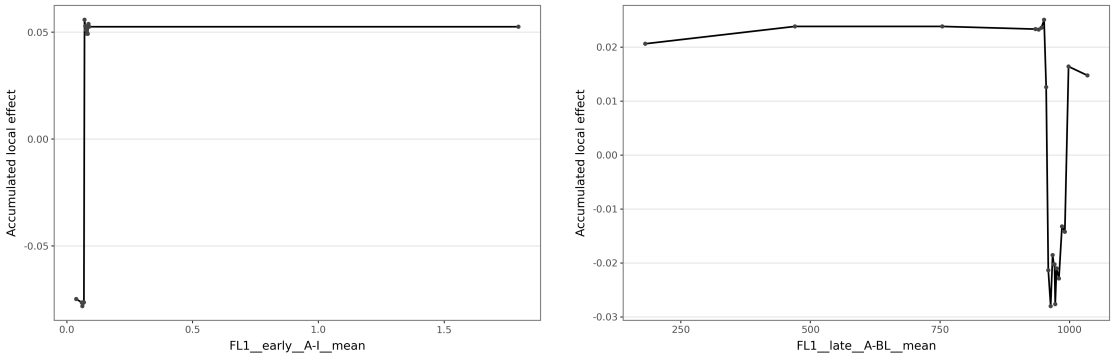
Figure B.10: Held-out diagnostics for the Line 1 thickness log variance XGBoost model using a subset of the data. The broader residual scatter is consistent with the weaker held-out performance of the subset-data thickness log variance model compared with the line-specific thickness mean models.



(a) Line 2 thickness mean, B-BA.

(b) Line 2 selected product thickness mean, B-BA.

Figure B.11: Medium evidence ALE diagnostics from the subset-data thickness mean interpretation. The same `B-BA` predictor appears twice because it was selected for interpretation separately for the Line 2 model and the Line 2 selected product model. These curves are supporting diagnostics; the product-specific curve should be read as exploratory because its held-out R^2 is negative.



(a) Line 1 thickness log variance, A-I. **(b)** Line 1 thickness log variance, A-BL.

Figure B.12: Medium evidence ALE diagnostics for the Line 1 thickness log variance XGBoost model using a subset of the data: early window `A-I` and late window `A-BL`. These effects are included for comparison with the stronger thickness mean results and should be interpreted cautiously because the subset-data thickness log variance model has weaker held-out performance.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY