

# **Applications of latent variable methods in primary aluminium production**

**Thèse en cotutelle  
Doctorat en génie chimique**

**Petre Manolescu**

Université Laval  
Québec, Canada  
Philosophiæ doctor (Ph. D.)

et

Reykjavik University  
Reykjavik

# Applications of latent variable methods in primary aluminum production

Thèse

Petre Manolescu

Sous la direction de

Prof. Duchesne, Carl, directeur de recherche

Prof. Saevarsdottir, Gudrun, directeur de cotutelle

## Résumé

L'industrie de l'aluminium est actuellement confrontée à des défis sur plusieurs fronts. L'un d'eux est la qualité des matières premières et leur impact sur la stabilité du processus. Dans les alumineries modernes, les exigences environnementales et l'efficacité économique augmentent tandis que la qualité des matières premières diminue. La numérisation de l'industrie manufacturière, surnommée Industrie 4.0, ouvre de nouvelles voies d'exploration pour répondre aux exigences de la production d'aluminium du 21e siècle.

Dans cette recherche, des méthodes d'analyse des mégadonnées ont été utilisées pour déterminer la stabilité d'une ligne de cuves d'électrolyse de l'aluminium primaire. Un outil de détection d'anomalies a été construit à l'aide d'une analyse en composantes principales. Le prototype a montré que les cuves problématiques dans les opérations peuvent être correctement identifiées plus tôt qu'avec les outils existants, basées sur des analyses univariés.

Le prototype multivarié a ensuite été mis en œuvre en production chez Alcoa Fjardaal. À l'aide du modèle d'analyse en composantes principales, un système de classement a été développé comme méthode de priorisation des cuves de réduction problématiques. Les graphiques des contributions des variables ont été utilisés pour aider à identifier les problèmes dans les cuves problématiques. Les performances du modèle ont été évaluées avec l'aide d'opérateurs suivant les cuves dans la salle des cuves. Une comparaison avec les outils existants à l'usine a montré le potentiel d'une notification anticipée du début des anomalies de processus.

Grâce à un système de traçabilité des anodes, des données d'aluminerie ont été collectées pour estimer la consommation nette de carbone (NCC), pour des anodes individuelles, à l'aide de méthodes à variables latentes. Les propriétés des anodes et les données de performance des cuves d'électrolyse ont été collectées pour relier les propriétés individuelles des anodes cuites et crues à l'opération des cuves d'électrolyse. Avec la liaison des données de carbone et l'opération des cuves d'électrolyse, une analyse statistique multivariée a été réalisée. Ensuite, les prédictions du modèle ainsi développé ont été comparées à la formule existante de consommation nette de carbone développée par la compagnie R&D Carbon. La comparaison a démontré que le pouvoir de prédiction de la consommation nette de carbone, pour des anodes individuelles, a augmenté de manière significative en utilisant le système de traçabilité des anodes et les méthodes à variables latentes.

Les données provenant du système de traçabilité ont été divisées en plusieurs blocs déterminés par la séquence des étapes de fabrication. L'impact de chaque bloc sur la consommation nette de carbone des anodes individuelles, a été déterminé à l'aide de la projection des structures latentes multiblocs séquentiels. Cette méthodologie a permis d'établir un lien entre les propriétés des anodes crues et la consommation nette de carbone des anodes individuelles.

## Abstract

The aluminum industry is currently facing challenges on multiple fronts. One of them is the raw material quality and their impact on process stability. In modern aluminum smelter plants, there are increasing environmental and cost efficiency demands while raw material quality decreases. The digitalization of the manufacturing industry, dubbed Industry 4.0, opens up new avenues of exploration to meet the demands of the 21st century aluminum production.

In this research, big data analytics methods were employed to determine the stability of a potline. An anomaly detection tool was built using principal component analysis. The prototype showed that drifts in process operations can be correctly identified earlier than with existing univariate statistical process control tools.

The multivariate prototype was later implemented in production at Alcoa Fjardaal. Using the principal component analysis model, a ranking system was developed as the prioritization method of the problematic reduction cells. The variables' contributions plots were used to help identify the issues in the problem cells. The model performance was assessed with the help of operators attending to the pots on the shop floor. A comparison with existing plant tools showed the potential of advanced notice of the onset of process anomalies.

Through an anode tracking system, smelter data was collected to estimate the net carbon consumption (NCC) of individual anodes using latent variable methods. The properties of individual anodes were thus linked to their corresponding reduction cell performance. A multivariate statistical analysis of the linked carbon and pot operation databases, was compared to the existing net carbon consumption formula developed by R&D Carbon. The comparison demonstrated that the prediction for net carbon consumption, for individual anodes, significantly increased using the anode tracking system and latent variable methods.

The anode tracking data was split into several blocks determined by manufacturing steps sequence. The impact of each block onto the net carbon consumption was determined using the sequential multi-block partial least squares algorithm. This methodology helped establish a link between green anode properties and the net carbon consumption of individual anodes.

## Útdráttur

Áliðnaðurinn stendur nú frammi fyrir áskorunum á mörgum vígstöðvum. Ein þeirra er hráefnisgæði og áhrif þeirra á framleiðslu stöðuleika. Í nútíma álverum eru auknar kröfur um umhverfis- og kostnaðarhagkvæmni á meðan hráefnasgæðinn minnka. Stafræn væðing framleiðsluiðnaðarins, kölluð "Industry 4.0", opnar nýjar leiðir til rannsókna til að mæta kröfum 21. aldar álframleiðslu.

Í þessari rannsókn var notuð stórgagnagreining til að ákvarða stöðugleika kerskála. Með því að nota aðalhlutagreining, var smíðað töl til að greina álframleiðslu ferli frávik. Frumgerðin af þessari greiningu sýndi að hægt er að bera kennsl á framleiðsluferlis frávik fyrir en með fyrirbyggjandi skýrslur, byggð á einþátta tölfræðilegum aðferðum.

Fjölþáttagreiningar frumgerðin var síðar innleidd í framleiðslu hjá Alcoa Fjarðaáli. Forgangsroðunar kerfi fyrir vandamála ker var þróuð út frá aðalhlutagreiningarlíkaninu. Framlögðs skýringar gröf voru notuð til þess að bera kennsl á frávik í vandamál kerjum. Frammistaða líkansis var metin með aðstoð verkamanna, sem sinntu kerum út á gólfi í kerskála. Samanburður við núverandi verkfæri verksmiðjunnar sýndi fram á möguleika á aðvörun áður en vandamálið fer af stað.

Með því að nota kerfi til þess að rekja skaut í gegnum framleiðslu ferli var hægt að safna gögn til þess að áætla heildar kolefnisnotkun (NCC) einstakra skauta með því að nota fjölþátta tölfræðilegar greiningaraðferðir. Eiginleikar einstakra forskauta og voru þannig tengd samsvarandi kerum og rekstrargögnum þeirra. Fjölþáttagreiningar aðferðum var beitt á gagnasafnið voru þær bornar saman við núverandi formúlu sem var þróuð af R&D Carbon. Samanburðurinn sýndi fram á að, fyrir einstakra skaut jókst nákvæmni spár á NCC verulega með því að nota skautrakningar kerfið og fjölþáttagreiningu.

Rakningargögnum fyrir rafskautin var skipt í mismunandi blokkir ákvarðaðar af röðun skrefa framleiðsluferlins. Áhrif hvernar blokk á NCC voru ákvörðuð með því að nota SMB-PLS (sequential multi-block partial least squares) aðferðin. Þessi aðferð kerði kæyft að koma á tengslin milli eiginleika grænna rafskauta og NCC einstakra rafskauta.

# Contents

|                                                                                                                                            |             |
|--------------------------------------------------------------------------------------------------------------------------------------------|-------------|
| <b>Résumé</b>                                                                                                                              | <b>ii</b>   |
| <b>Abstract</b>                                                                                                                            | <b>iii</b>  |
| <b>Útdráttur</b>                                                                                                                           | <b>iv</b>   |
| <b>Contents</b>                                                                                                                            | <b>vii</b>  |
| <b>List of Tables</b>                                                                                                                      | <b>viii</b> |
| <b>List of Figures</b>                                                                                                                     | <b>x</b>    |
| <b>List of abbreviations</b>                                                                                                               | <b>xi</b>   |
| <b>Acknowledgements</b>                                                                                                                    | <b>xiii</b> |
| <b>Preface</b>                                                                                                                             | <b>xiv</b>  |
| <b>Introduction</b>                                                                                                                        | <b>1</b>    |
| Importance of aluminum production . . . . .                                                                                                | 1           |
| Aluminum production process . . . . .                                                                                                      | 1           |
| Current challenges of aluminum production . . . . .                                                                                        | 6           |
| Economic challenges . . . . .                                                                                                              | 6           |
| Anode raw material challenges . . . . .                                                                                                    | 8           |
| Process monitoring challenges . . . . .                                                                                                    | 12          |
| Research Objectives . . . . .                                                                                                              | 14          |
| Contributions . . . . .                                                                                                                    | 15          |
| <b>Chapter 1 On the use of Multivariate Statistical Methods to detect, diagnose<br/>and mitigate abnormal events in aluminium smelters</b> | <b>17</b>   |
| <b>Résumé</b>                                                                                                                              | <b>17</b>   |
| <b>Abstract</b>                                                                                                                            | <b>17</b>   |
| 1.1 Introduction . . . . .                                                                                                                 | 17          |
| 1.2 Data and methodology . . . . .                                                                                                         | 18          |
| 1.3 Introduction to PCA . . . . .                                                                                                          | 19          |
| 1.3.1 Building the PCA model . . . . .                                                                                                     | 20          |
| 1.3.2 Projections onto the PCA model . . . . .                                                                                             | 22          |
| 1.4 Conclusion . . . . .                                                                                                                   | 28          |
| <b>Chapter 2 Multivariate monitoring of reduction cell operations: model perfor-<br/>mance assessment</b>                                  | <b>29</b>   |
| <b>Résumé</b>                                                                                                                              | <b>29</b>   |

|                                                                                                                                          |           |
|------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>Abstract</b>                                                                                                                          | <b>29</b> |
| 2.1 Introduction . . . . .                                                                                                               | 29        |
| 2.2 Data Collection . . . . .                                                                                                            | 31        |
| 2.3 Methodology . . . . .                                                                                                                | 33        |
| 2.3.1 Algorithm . . . . .                                                                                                                | 33        |
| 2.3.2 Model Building (Phase I) . . . . .                                                                                                 | 34        |
| 2.3.3 Model Exploitation (Phase II) . . . . .                                                                                            | 35        |
| 2.3.4 Ranking system and performance metrics . . . . .                                                                                   | 37        |
| 2.4 Results . . . . .                                                                                                                    | 39        |
| 2.4.1 Performance assessment of the monitoring scheme . . . . .                                                                          | 39        |
| 2.4.2 Examples of anomaly detections . . . . .                                                                                           | 43        |
| 2.5 Conclusion . . . . .                                                                                                                 | 50        |
| 2.6 Future Work . . . . .                                                                                                                | 51        |
| <b>Chapter 3 Estimation of Net Carbon Consumption in Aluminum Electrolysis<br/>Using Multivariate Analysis</b>                           | <b>52</b> |
| <b>Résumé</b>                                                                                                                            | <b>52</b> |
| <b>Abstract</b>                                                                                                                          | <b>52</b> |
| 3.1 Introduction . . . . .                                                                                                               | 53        |
| 3.2 Anode manufacturing and smelter data collection . . . . .                                                                            | 55        |
| 3.3 Modelling of the dataset . . . . .                                                                                                   | 58        |
| 3.4 Results . . . . .                                                                                                                    | 59        |
| 3.4.1 Data cleaning and pre-processing . . . . .                                                                                         | 59        |
| 3.4.2 Analysis of the small dataset . . . . .                                                                                            | 60        |
| 3.4.3 Analysis of the large dataset . . . . .                                                                                            | 63        |
| 3.5 Conclusions . . . . .                                                                                                                | 66        |
| <b>Chapter 4 Net carbon consumption in aluminum electrolysis - impact of anode<br/>properties and reduction cell operation variables</b> | <b>68</b> |
| <b>Résumé</b>                                                                                                                            | <b>68</b> |
| <b>Abstract</b>                                                                                                                          | <b>68</b> |
| 4.1 Introduction . . . . .                                                                                                               | 68        |
| 4.2 Anode manufacturing and smelter data collection . . . . .                                                                            | 70        |
| 4.3 Modelling of the data . . . . .                                                                                                      | 73        |
| 4.4 Results . . . . .                                                                                                                    | 75        |
| 4.4.1 Data cleaning and pre-processing . . . . .                                                                                         | 75        |
| 4.4.2 SMB-PLS modelling . . . . .                                                                                                        | 76        |
| 4.5 Conclusion . . . . .                                                                                                                 | 85        |
| <b>Conclusion and future work</b>                                                                                                        | <b>87</b> |
| <b>References</b>                                                                                                                        | <b>89</b> |

|                                                                   |            |
|-------------------------------------------------------------------|------------|
| <b>Appendices</b>                                                 | <b>97</b>  |
| <b>Appendix A</b> <b>Variable names for material in Chapter 2</b> | <b>97</b>  |
| <b>Appendix B</b> <b>Supplementary material for Chapter 2</b>     | <b>99</b>  |
| <b>Appendix C</b> <b>Supplementary material for Chapter 3</b>     | <b>101</b> |
| <b>Appendix D</b> <b>Supplementary material for Chapter 4</b>     | <b>102</b> |
| D.1   SMB-PLS algorithm pseudo-code . . . . .                     | 102        |
| D.2   PLS metrics . . . . .                                       | 102        |
| D.3   SMB-PLS Model Latent Variables . . . . .                    | 103        |
| D.4   List of abbreviations for Figure 47 and Figure 48 . . . . . | 103        |

## List of Tables

|           |                                                                                                                                                                                                                                    |     |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Table 2   | Process variables categories . . . . .                                                                                                                                                                                             | 32  |
| Table 3   | Number of SMB-PLS components for each data block . . . . .                                                                                                                                                                         | 76  |
| Table 4   | Anode position category per pot . . . . .                                                                                                                                                                                          | 81  |
| Table A.1 | Variable names for figures 26 to 31 . . . . .                                                                                                                                                                                      | 97  |
| Table B.1 | Overall model performance. Note that the metrics were abbreviated to better fit the page formatting. Thus, Obs = Observations, Prec = Precision, Acc = Accuracy, Sens = Sensitivity, Spec = Specificity . . . . .                  | 99  |
| Table B.2 | Model performance metrics per 10-cell bins. Note that the metrics were abbreviated to better fit the page formatting. Thus, Obs = Observations, Prec = Precision, Acc = Accuracy, Sens = Sensitivity, Spec = Specificity . . . . . | 100 |
| Table C.1 | Distribution of coefficients for MLR parameters obtained for the MLR vs the R&D Carbon formula coefficients . . . . .                                                                                                              | 101 |
| Table D.1 | Pseudo-code for the SMB-PLS algorithm [101] . . . . .                                                                                                                                                                              | 102 |
| Table D.2 | Number of SMB-PLS latent variables (LVs) for each data block and corresponding SMB-PLS latent variables representing that block within the model . .                                                                               | 103 |
| Table D.3 | Variable abbreviations for Figure 47 . . . . .                                                                                                                                                                                     | 103 |
| Table D.4 | Variable abbreviations for Figure 48 . . . . .                                                                                                                                                                                     | 104 |

## List of Figures

|           |                                                                                                                                              |    |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 1  | Early Hall Cell at the Pittsburgh Reduction Company [7] . . . . .                                                                            | 2  |
| Figure 2  | Section of the early Héroult cell [8] . . . . .                                                                                              | 2  |
| Figure 3  | Reduction cell schematic representation [12] . . . . .                                                                                       | 2  |
| Figure 4  | Soderberg anode [10] . . . . .                                                                                                               | 4  |
| Figure 5  | Representation of the Net Carbon Consumption (adapted from [18]) . . .                                                                       | 5  |
| Figure 6  | Aluminum production world wide . . . . .                                                                                                     | 6  |
| Figure 7  | Aluminum demand by region in 2022 [14] . . . . .                                                                                             | 7  |
| Figure 8  | Electric energy use for aluminum electrolysis, by region and source for<br>2022 [13] . . . . .                                               | 8  |
| Figure 9  | Global oil production by type projection [22] . . . . .                                                                                      | 9  |
| Figure 10 | Global oil quality [22] . . . . .                                                                                                            | 9  |
| Figure 11 | Conversion projects by region 2023-2028 [23] . . . . .                                                                                       | 10 |
| Figure 12 | The three different coke types [19] . . . . .                                                                                                | 11 |
| Figure 13 | Loadings of the principal components 1 and 2 . . . . .                                                                                       | 21 |
| Figure 14 | Hierarchy of the analysis . . . . .                                                                                                          | 22 |
| Figure 15 | Heatmap of the model data. . . . .                                                                                                           | 23 |
| Figure 16 | Heatmap of the projected data. The excursion onset is roughly the begin-<br>ning of June. . . . .                                            | 24 |
| Figure 17 | Pot A110 example - timeseries of its $T^2$ values. . . . .                                                                                   | 25 |
| Figure 18 | Pot A110 example - $T^2$ contribution plot for samples 47 to 57. . . . .                                                                     | 25 |
| Figure 19 | Time-series of trace elements and feed control parameters in pot A110 . .                                                                    | 26 |
| Figure 20 | Potline Summary Statistics . . . . .                                                                                                         | 27 |
| Figure 21 | Potline Trace Elements . . . . .                                                                                                             | 28 |
| Figure 22 | PCA model percentage variance explained per Latent Variable (LV) . . .                                                                       | 39 |
| Figure 23 | Daily number of cells above the $T^2$ statistical limit for detecting abnormal<br>events . . . . .                                           | 40 |
| Figure 24 | Accuracy distribution of the MSPC scheme according to the rank bins.<br>The number of cells in each bin is displayed above each bar. . . . . | 42 |
| Figure 25 | Time series of Hotelling's $T^2$ and cell ranking for Cell 1, Cell 2 and Cell 3                                                              | 44 |
| Figure 26 | $T^2$ contributions for Cell 1. Only the most important contributors are<br>shown (absolute value of contribution $> 1$ ). . . . .           | 45 |
| Figure 27 | $SPE$ contributions for Cell 1. Only the most important contributors are<br>shown (absolute value of contribution $> 1$ ). . . . .           | 46 |
| Figure 28 | $T^2$ contributions for Cell 2. Only the most important contributors are<br>shown (absolute value of contribution $> 1$ ). . . . .           | 47 |
| Figure 29 | $SPE$ contributions for Cell 2. Only the most important contributors are<br>shown (absolute value of contribution $> 1$ ). . . . .           | 48 |
| Figure 30 | $T^2$ contributions for Cell 3. Only the most important contributors are<br>shown (absolute value of contribution $> 1$ ). . . . .           | 49 |
| Figure 31 | $SPE$ contributions for Cell 3. Only the most important contributors are<br>shown (absolute value of contribution $> 1$ ). . . . .           | 50 |
| Figure 32 | Schematic representation of the anode tracking system . . . . .                                                                              | 55 |

|           |                                                                                                                                                                                              |    |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 33 | Full Scale PLS data structure . . . . .                                                                                                                                                      | 57 |
| Figure 34 | Small Scale PLS data structure based on the R&D Carbon formula variables                                                                                                                     | 57 |
| Figure 35 | Variance explained ( $R^2$ and $Q^2$ ) comparison between the methods . . . .                                                                                                                | 60 |
| Figure 36 | $RMSE$ comparison between the methods . . . . .                                                                                                                                              | 61 |
| Figure 37 | Distribution of regression coefficients in the MLR Monte Carlo simulation.                                                                                                                   | 62 |
| Figure 38 | Scatter plots showing the line of perfect correlation (diagonal line) and<br>observed vs predicted values for PLS 6 variable dataset . . . . .                                               | 63 |
| Figure 39 | Scatter plots showing the line of perfect correlation (diagonal line) and<br>observed vs predicted values for PLS large dataset . . . . .                                                    | 64 |
| Figure 40 | PLS model on full dataset: Variables Important to Prediction (VIP) . . .                                                                                                                     | 65 |
| Figure 41 | Schematic representation of the green anode resistivity measurements used<br>in the analysis . . . . .                                                                                       | 71 |
| Figure 42 | Fused data structure. The number of variables included in each block is<br>indicated at their top right corner. The number of tracked anodes (observations)<br>is 1088. . . . .              | 72 |
| Figure 43 | SMB-PLS, block importance to prediction (BIP) . . . . .                                                                                                                                      | 78 |
| Figure 44 | SMB-PLS, block relative weight for each latent variable . . . . .                                                                                                                            | 79 |
| Figure 45 | SMB-PLS, cumulative block variance explained by each latent variable . .                                                                                                                     | 80 |
| Figure 46 | Super scores plots for the latent variables representing the green anode<br>block and alumina feed block . . . . .                                                                           | 82 |
| Figure 47 | Alumina feed block $T^2$ contribution plot, difference between Low NCC<br>group and High NCC group. The full names of the variables are given in Table<br>D.3, in the Appendix D.4 . . . . . | 83 |
| Figure 48 | Green anode block $T^2$ contribution plot. The full names of the variables<br>are given in Table D.4, in the Appendix D.4. . . . .                                                           | 85 |

## List of Abbreviations

|                |                                              |
|----------------|----------------------------------------------|
| <b>4PP</b>     | Four Point Probe                             |
| <b>AAC</b>     | Aluminum Association of Canada               |
| <b>ACD</b>     | Anode Cathode Distance                       |
| <b>ADQ</b>     | Alcoa Deschambault smelter                   |
| <b>API</b>     | American Petroleum Institute                 |
| <b>AP</b>      | Air Permeability                             |
| <b>AP-30</b>   | Aluminium Pechiney 30 technology             |
| <b>ARR</b>     | Air Reactivity Residue                       |
| <b>ATS</b>     | Anode Tracking System                        |
| <b>BIP</b>     | Block Importance to Prediction               |
| <b>BLS</b>     | Bottom Long Side                             |
| <b>BSS</b>     | Bottom Short Side                            |
| <b>CPC</b>     | Calcined Petroleum Coke                      |
| <b>CE</b>      | Current Efficiency                           |
| <b>CRR</b>     | CO <sub>2</sub> Reactivity Residue           |
| <b>DC</b>      | Direct Current                               |
| <b>DLS</b>     | Delta Long Side                              |
| <b>DSS</b>     | Delta Short Side                             |
| <b>DPLS</b>    | Discriminant Partial Least Squares           |
| <b>FN</b>      | False Negative                               |
| <b>FP</b>      | False Positive                               |
| <b>GPC</b>     | Green Petroleum Coke                         |
| <b>GDP</b>     | Gross Domestic Product                       |
| <b>HLS</b>     | Horizontal Long Side                         |
| <b>kA</b>      | Kilo Amper                                   |
| <b>KPCA</b>    | Dynamic Kernel Principal Component Analysis  |
| <b>KPI</b>     | Key Process Indicator                        |
| <b>LV</b>      | Latent Variable                              |
| <b>MLR</b>     | Multiple Linear Regression                   |
| <b>MHD</b>     | Magnetohydrodynamics                         |
| <b>NCC</b>     | Net Carbon Consumption                       |
| <b>PCA</b>     | Principal Component Analysis                 |
| <b>PLS</b>     | Partial Least Squares                        |
| <b>RMSE</b>    | Root Mean Square Error                       |
| <b>RMSECV</b>  | Root Mean Square of Cross-Validation         |
| <b>RMSEC</b>   | Root Mean Square of Calibration              |
| <b>RMSEP</b>   | Root Mean Square of Prediction               |
| <b>SMB-PLS</b> | Sequential Multi-Block Partial Least Squares |
| <b>SPC</b>     | Statistical Process Control                  |
| <b>SPE</b>     | Squared Predicted Error                      |
| <b>TC</b>      | Thermal Conductivity                         |
| <b>TN</b>      | True Negative                                |

|            |                                     |
|------------|-------------------------------------|
| <b>TP</b>  | True Positive                       |
| <b>TLS</b> | Top Long Side                       |
| <b>VIP</b> | Variable's Importance to Prediction |
| <b>VLS</b> | Vertical Long Side                  |

## Acknowledgments

My PhD journey has been long and it would not have been possible without support. I'd like to start by thanking my supervisors Carl Duchesne and Guðrún Sævarsdóttir for accepting to take me on as a student and for their patience to guide me. This work could not have been possible without Jayson Tessier, whom I first met in late 2015 while I was a process engineer at Alcoa Fjarðaal. He introduced me to the Alcoa collaboration with Laval University, we settled on a topic and put together a proposal for the plant's management to sponsor my PhD. I am very grateful to Ingólfur Þór Ágústsson and Smári Kristinsson for their decision to sponsor me and the support they have shown over the years.

My role within Alcoa has changed over the years and the level of responsibility increased, but I continued my PhD alongside it. This was possible thanks to the support of my new managers, first Ingólfur Tómas Helgason at Fjarðaal, then Geff Wood and Bruno Longchamps in Québec, Canada and now, Peter Rose. I'd also like to thank Julien Lauzon-Gauthier for his valuable insights, discussions and coordination.

I would also like to thank other students from Laval University and members of the REGAL community who have helped me settle in and provided good company throughout this endeavour. Thank you Ramzi Ishak, Geoffroy Rouget, Jean-François Boulanger, Adéline Paris, David LaJambe, Daniel Rodrigues and Cecillia Rodrigues.

Finally, I would like to thank my family for the encouragement and support they have shown me throughout the years. Most of all, I would like to thank my fiancée, Caroline, for being a steady rock that I could lean on, along with her understanding and support of spending long hours at my desk. I look forward to having more free time for our life projects.

## Preface

This thesis presents the outcome of my work throughout my PhD studies at Laval University in joint supervision with Reykjavik University. It includes 4 chapters, each representing an article that has been published. I was the first author in all articles responsible for the data collection, the analysis, results and the writing. The concepts and objectives for chapters 1 and 3 were defined in coordination with the coauthors Carl Duchesne, Gudrun Saevarsdottir and Jayson Tessier. The concepts and objectives for chapters 2 and 4 were defined in coordination with the coauthors Carl Duchesne, Gudrun Saevarsdottir and Julien Lauzon-Gauthier.

The Introduction presents an overview of the importance of the aluminum industry within Canada and Iceland along with the overall production process and the influences of raw materials quality onto it. Chapter 1 presents a process anomaly detection tool built using multivariate analysis. Chapter 2 takes this work further and presents the model performance in a production environment. Chapter 3 shows how an anode tracking system can be used to predict the net carbon consumption of individual anodes. The results are compared with the reference method, widely used in the industry. Chapter 4 takes this work further and shows how latent variable methods can be used to link individual green anode properties to the net carbon consumption.

The first chapter presents the conference article *On the use of multivariate statistical methods to detect, diagnose and mitigate abnormal events in aluminium smelters*, co-authored by Carl Duchesne, Jayson Tessier and Gudrun Saevarsdottir. The article was submitted on September 1, 2017 and published on January 31, 2018 in *Light Metals 2018*.

The second chapter presents the article *Multivariate monitoring of reduction cell operations - model performance assessment*, which was co-authored by Carl Duchesne, Gudrun Saevarsdottir and Julien Lauzon-Gauthier. The article was submitted on September 30, 2024 to JOM and published on February 7, 2025.

The third chapter presents the article *Estimation of net carbon consumption in aluminum electrolysis using multivariate analysis*, co-authored by Carl Duchesne, Jayson Tessier and Gudrun Saevarsdottir. The article was submitted on August 7, 2020 and published on November 12, 2020 in the Journal of Sustainable Metallurgy.

The fourth chapter presents the article *Net carbon consumption in aluminum electrolysis - impact of anode properties and reduction cell operation variables*, co-authored by Carl Duchesne, Julien Lauzon-Gauthier and Gudrun Saevarsdottir. The article was submitted on March 24, 2022 and published on June 27, 2022 in the Journal of Sustainable Metallurgy.

# Introduction

## Importance of aluminum production

In the Earth's crust, the third most abundant element is aluminum, or 8.3 % [1], and today is a widely used non-ferrous metal. Its production is an important part of the Canadian economy. According to the Aluminium Association of Canada (AAC) [2], the country is the world's 4th largest producer with one of the lowest carbon footprint thanks to the use of hydroelectric power. The total annual production is 3.2 million metric tonnes which accounts for 5 % of the global production. There are currently 9 smelters, 8 of them in Quebec and 1 in British Columbia. Together they account for 9000 direct jobs and 20000 indirect jobs in Canada, while generating exports of over \$CAD 12 billion.

In 2019, the Canadian government along with the provincial government of Quebec, Alcoa and Rio Tinto announced the launching of a new process for the production of aluminum that releases oxygen instead of the green house gases [3]. While this is a remarkable and revolutionary achievement, it will take a long time to commercialize it and potentially convert Hall-Héroult smelters to it. It is thus important to carry on searching for improvements in the existing technology.

In Iceland, the aluminum industry has grown and is now about a third of the export balance of goods. The annual production reached 866 thousand tonnes from 3 smelters in 2023 and, like Canada, the energy used is from renewable sources. The industry provides direct employment for about 4600 people which accounts for roughly 2.4 % of the total Icelandic workforce [4]. The value of the aluminum exported goods has increased from about 10 % in 1970 to about 40 % in 2022 [5].

## Aluminum production process

The aluminum production process is a complex one involving many different raw materials and high energy consumption. The primary raw material, the alumina, is dissolved in a molten salt bath (cryolite) and is electrolysed in the cell by passing a high DC current from the anode to the cathode. Depending on technology used the current can be from 100kA to over 600kA.

An early use of aluminum was in the form of alum (potassium aluminum sulphate) to stop bleeding in ancient Rome and to colour fabric. Its use spread throughout Europe with production increasing to 1500 tonnes a year in the papal states, during the middle ages [6].

Aluminum forms a strong bond to oxygen and thus it has been difficult to find a way to extract the metal. Early attempts involved extraction of impure aluminum from alum in 1809 by Sir Humphry Davy. In Denmark, 1825, Hans Christian Ørsted heated aluminum chloride with potassium metal to obtain an impure form of aluminum. Later, aluminum chloride was reduced with sodium by Henry-Étienne Sainte-Claire Deville in 1855, France. However, the cost was too high and the metal less than 95 % pure [6]. In 1845 Fredrich Wöhler was able to produce small amounts of pure aluminum and was credited as the discoverer of the metal [6].

The Hall-Héroult process was invented independently and almost simultaneously in 1886 by Charles M. Hall and Paul Héroult, each receiving a patent for the technology in the USA and France respectively. The early aluminum electrolysis cells are represented in figures 1 and 2, while a modern cell is presented in figure 3.

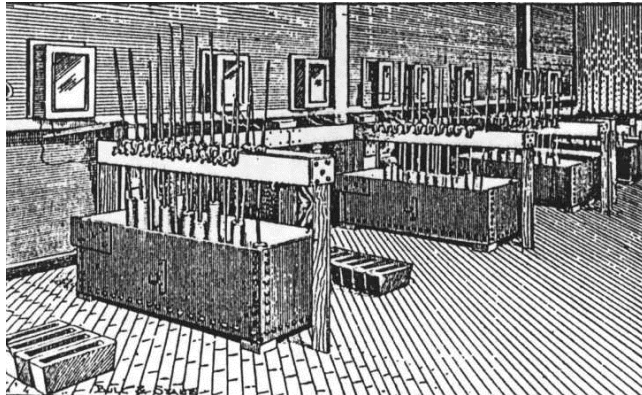


Figure 1: Early Hall Cell at the Pittsburgh Reduction Company [7]

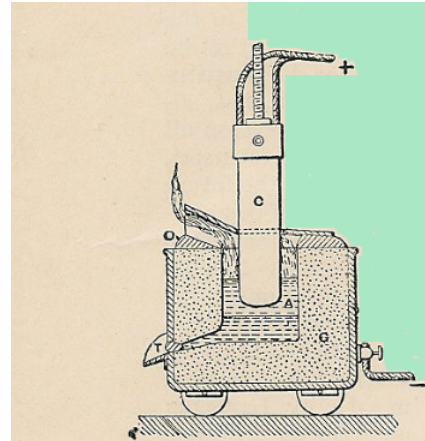


Figure 2: Section of the early Héroult cell [8]

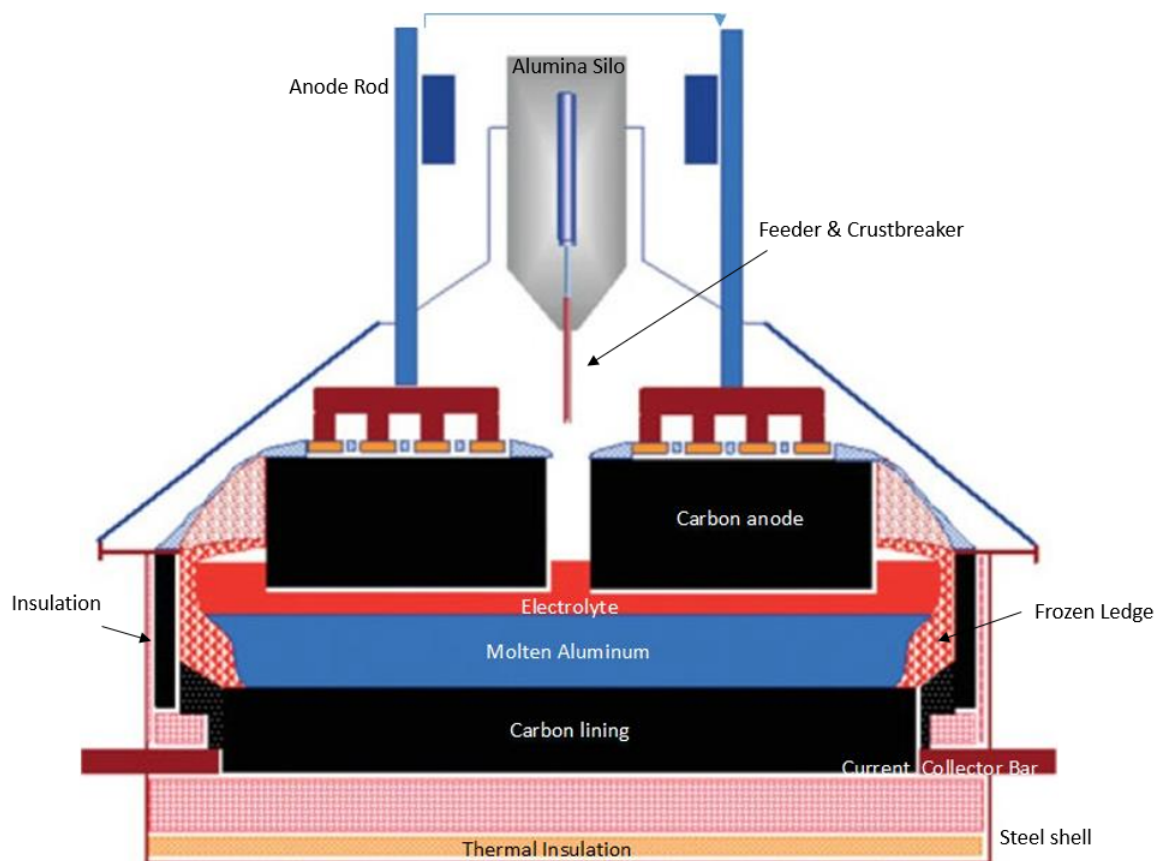
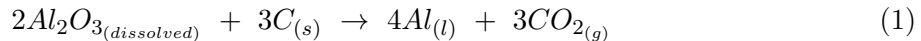


Figure 3: Reduction cell schematic representation [12]

The industry has come a long way from its early days with technology advancements allowing for increased production in a safer manner for both the operator and the environment. At its core, however, the process remains the same and is governed by the overall chemical reaction that can be written as:



As seen in equation (1) the primary raw material is alumina. It is produced from bauxite via the Bayer process. The only available electrolyte is based on cryolite ( $Na_3AlF_6$ ), which has a melting point of 1012 °C but various fluoride additives ( $AlF_3$ ,  $CaF_2$ ,  $LiF$ ) are used to reduce it and thus lower the operating temperature to around 960 °C and improve current efficiency. Most commonly, the bath contains 6-13 wt %  $AlF_3$  in excess of the cryolite composition, 4-6wt %  $CaF_2$ , 2-4 wt %  $Al_2O_3$  [11].

To obtain aluminum metal a DC electric current is passed through the electrolyte within a steel shell called a cell or pot. Heat transfer management plays an important part in operation. Molten cryolite is highly corrosive so, to prevent tap-outs, a side ledge of frozen cryolite around the walls of the cell must be present at all times.

The anode and cathode materials are made of carbon. The anodes are consumable and the preferred type are pre-baked. The second type, now rare, are Soderberg which are baked continuously. Similar to electric arc furnace technologies, briquettes (blended petroleum coke and pitch) are fed on the top of the electrode shell. As the bottom of the anode is consumed, the briquettes go down and bake to a solid anode by pyrolysis. They have a large surface area and usually contain almost double the pitch content of pre-bake anodes [9]. The electrical resistivity is higher in Soderberg anodes than in the pre-bake ones by approximately 30 %. Additionally, greater emissions come with the Soderberg technology due to difficulty in sealing the cells [12].

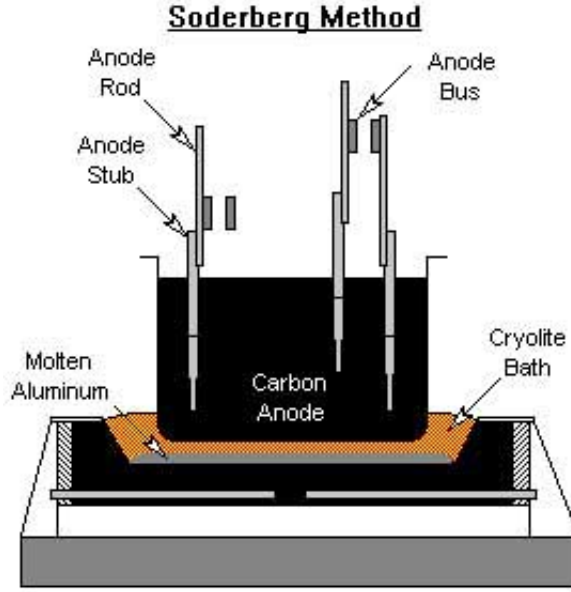
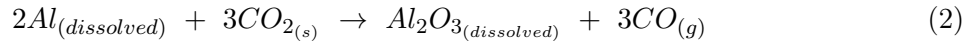


Figure 4: Soderberg anode [10]

The most common technology in use today is the pre-baked. Along with greater current efficiency, the cells are easier to seal and therefore more environmentally friendly. The number of anodes in each cell depends on the amperage used and the pot design at each location. In the present work, the cells operate at about 380 kA and contain 40 anodes each.

Even though technological strides have been made throughout the years, the process is not without losses. The main source of current efficiency loss comes from the back reaction, in which aluminum dissolved in the electrolyte comes into contact with carbon dioxide to form alumina and carbon monoxide, as shown in equation 2. The alumina must be electrolysed again thus reducing the current efficiency.



The anodes are produced at designated plants, which in some cases are next to the smelter, as is the case of the Alcoa Deschambault smelter. Other plants, such as Alcoa Fjardaal, Iceland, receive their anodes from another location, in this case, Norway. The anode recipe consists of 55-65 % petroleum coke, 15-30 % recycled anode butts and about 15 % pitch. The paste is mixed and formed in vibrocompactors thus producing the green anodes. These are then placed in a baking furnace and baked at about 1100 °C.

Stoichiometry of equation (1) shows that 1.89 kg of alumina are required to produce 1 kg of aluminum. Since there are always some impurities in the raw material, the actual amount is closer to 1.93 kg of alumina per kg of aluminum produced. Theoretically, 0.33 kg of carbon are needed to produce 1 kg of aluminum and 1.22 kg of CO<sub>2</sub>. However, other reactions of carbon with oxygen and CO<sub>2</sub> result in an overall consumption of between 0.4 and 0.45 kg C per kg of

Al [9]. Thus the net carbon consumption (NCC) is always higher than the theoretical value. A schematic representation of the overall carbon consumption is presented in figure 5.

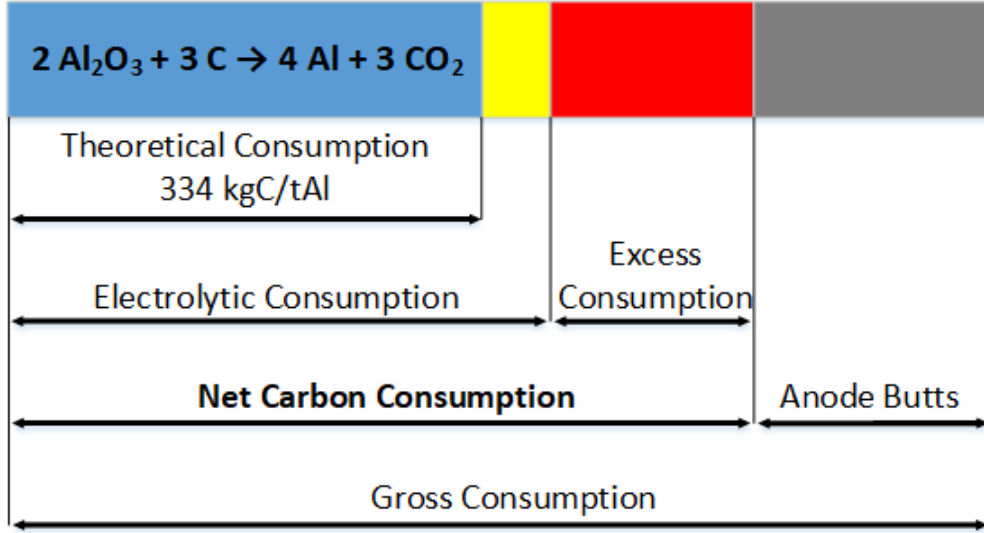
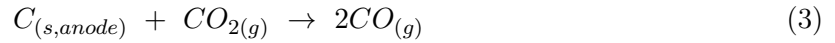
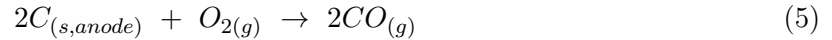


Figure 5: Representation of the Net Carbon Consumption (adapted from [18])

One of the contributors to excess carbon consumption is the Boudouard reaction:



If anode cover is incomplete, reactions may occur between air and carbon, while carbon monoxide is quickly converted to  $CO_2$ .



Still other carbon losses can take place in the form of unreacted dust, reoxidation of metal, pyrolysis and vaporization. Selective oxidation of binder coke can result in carbon losses via dusting. It takes place when the baked anodes do not contain the correct amount of pitch, they are under-baked or both [32]. Impurities in the carbon material can greatly influence the consumption. Elements such as Na, Ca, V and Ni act as strong oxidizing catalysts [9]. The main source of sodium enters the anode recipe through the crushed anode butt because of the frozen bath that remains attached to the butt when it is taken out of the cell. Proper cleaning of anode butts is required before sending the material to crushing [16]. Sulfur inhibits carbon air burn but the amount allowed in the anodes is limited due to the emission of sulfur oxide. These

limits vary from one region to another depending on local restrictions imposed by governments. Sulfur can also form COS, CS<sub>2</sub> and H<sub>2</sub>S. The chemical reduction of sulphate anion by carbon and metallic aluminum form sulphides and poly-sulphides [15]. Metal impurities such as nickel, iron and silicon impact the purity of the aluminum produced.

The anode properties can be traced to the quality of the raw materials namely, coke and pitch. They are a by-product of the oil industry and, for some time now, their quality has been declining while also increasing the variability of their properties [19]. This is proving to be a constant challenge for aluminum producers to maintain consistency within the anode production which will be further discussed in a subsequent section.

## Current challenges of aluminum production

### Economic challenges

It is well known that today's aluminum market faces challenges due to increased overcapacity. In particular, China has long surpassed the total production of the rest of the world combined. Recent data from the World Aluminium website [13] was extracted, that shows this trend in Figure 6.

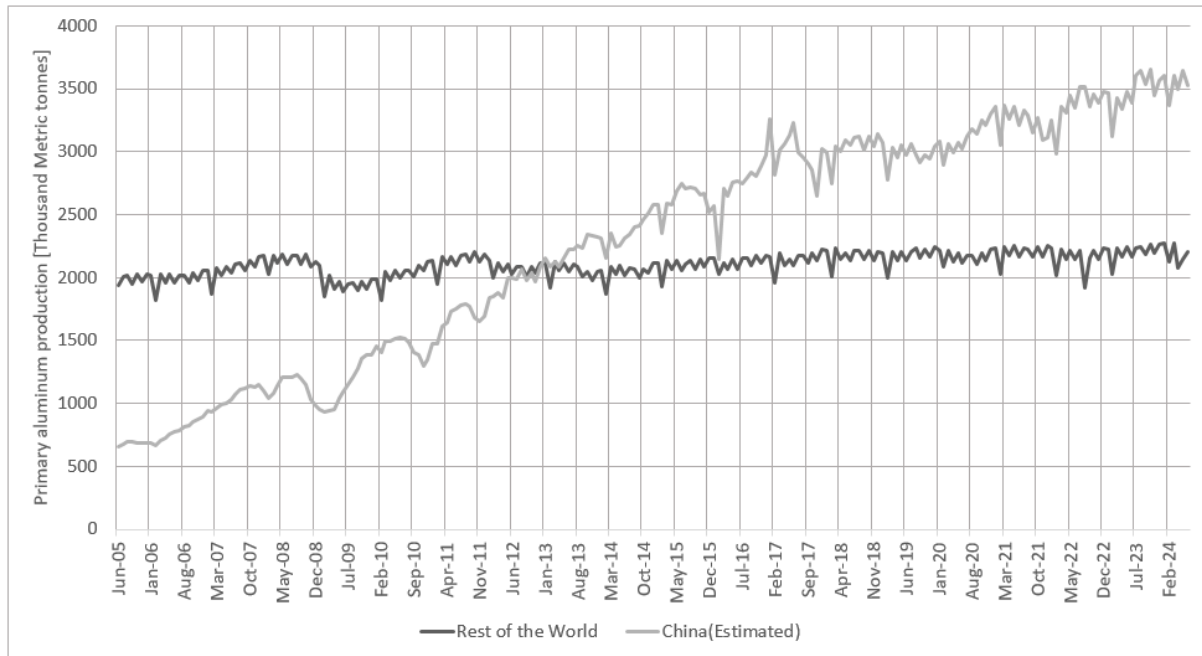


Figure 6: Aluminum production world wide

The world production, excluding China, has remained fairly constant within the past ten years while China is currently producing staggering numbers. These are almost double compared to the rest of the world combined. The Chinese are, however, also a large consumer of aluminum as shown in figure 7.

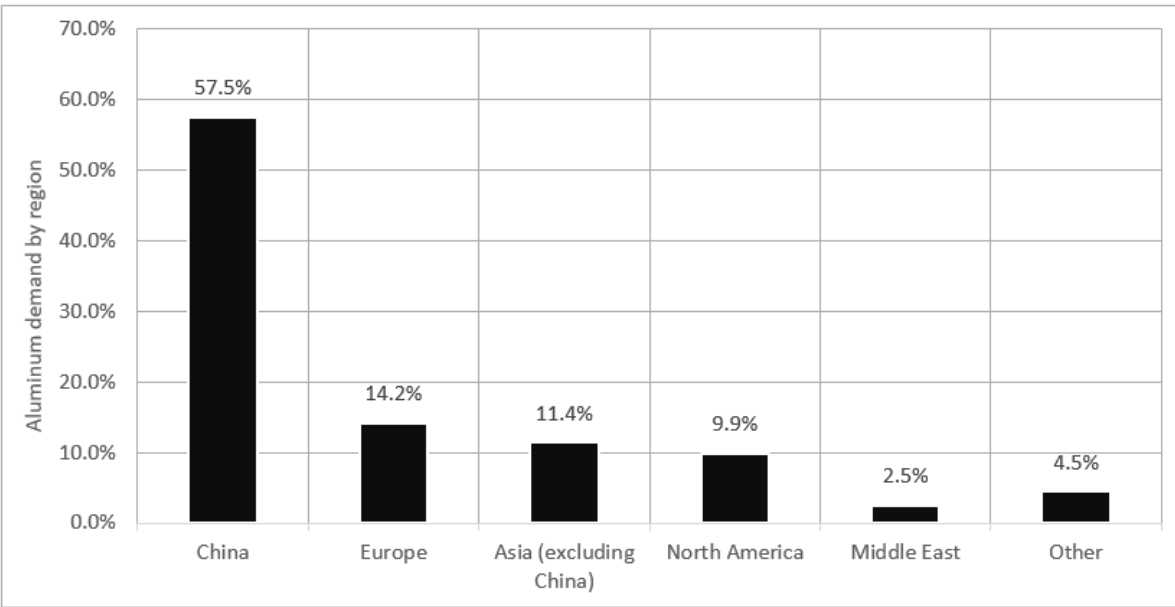


Figure 7: Aluminum demand by region in 2022 [14]

China has the highest unit cost for power [20] and relies heavily on coal as can be seen in figure 8. However, government subsidies keep the industry going, which would otherwise have difficulties to stay afloat. Some of these subsidies can be found as government purchases of aluminum stocks, lower alumina prices, subsidised electricity and coal, favourable lending terms, capital support and grants, logistical support and subsidised land and infrastructure [20]. Given China's high production rate, these government interventions are a major contributor to the overcapacity.

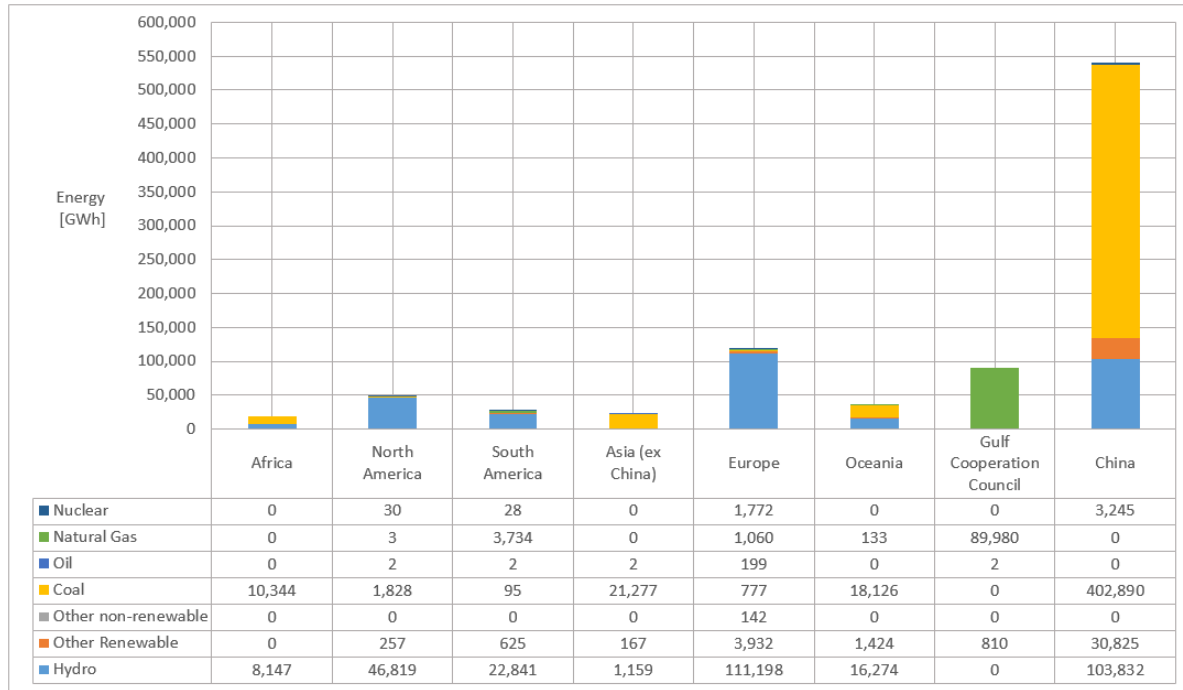


Figure 8: Electric energy use for aluminum electrolysis, by region and source for 2022 [13]

A summary of the pre-baked anode industry in China was published by Zhang et al. [21] where the annual production of aluminum increased from 0.555 million tonnes in 1985 to 4.32 million tonnes in 2002 and 32.5 million tonnes in 2016. With this the pre-baked anode production increased not only for domestic use but also with the scope of exporting. In 2002 it was 1.2 million tonnes and increased to a capacity of 23.6 million tonnes by 2016 from more than 130 carbon plants. The actual output for 2016 was 17.9 million tonnes, out of which more than 1 million tonnes were available for export. Variability in types and sizes of anodes make it difficult to standardize the product throughout the country.

The increasing competitiveness is driving most smelters to lower the cost of production and improve performance. Because of the existing infrastructure and raw materials on site, many smelters have ongoing projects of increasing amperage while maintaining or lowering the total energy consumption. These actions are aimed at increasing the overall production with minimal extra cost. Automation can improve daily operations efficiency and advanced analytics can be used to predict and mitigate production upsets. The use of big data analysis could provide additional capabilities to lower the cost and improve performance such as exploring individual anode consumption and their impact on reduction cell operation.

### Anode raw material challenges

The raw materials for anode production, petroleum coke and coal tar pitch, are based on hydrocarbons from crude oil residues. These provide the necessary properties desired in aluminum production: high chemical purity, high electrical conductivity, high mechanical strength and homogeneous with low air and CO<sub>2</sub> reactivity [9]. However, these materials are only residues from

the oil industry and account for approximately 2 % of their income. That provides low incentive for the oil refineries to produce high quality coke for use in aluminum smelting. Therefore, the carbon anodes are produced from raw materials having a high degree of variability [19]. Also, the coke and pitch are dependent on the type of oil processed. In recent years the raw materials have seen a declining quality and increasing variability. This increases the cost of anode production and results in purchasing from multiple suppliers and blending of the different qualities. It is thus increasingly difficult for smelters to receive anodes with consistent properties [24, 25].

Among the main crude oil quality parameters are the American Petroleum Institute (API) gravity and sulfur content. In the World Oil Outlook [22] the overall light grades of oil types are expected to decrease over time, as shown in figure 9. The average API gravity and sulfur content projections at a global level are shown in figure 10.

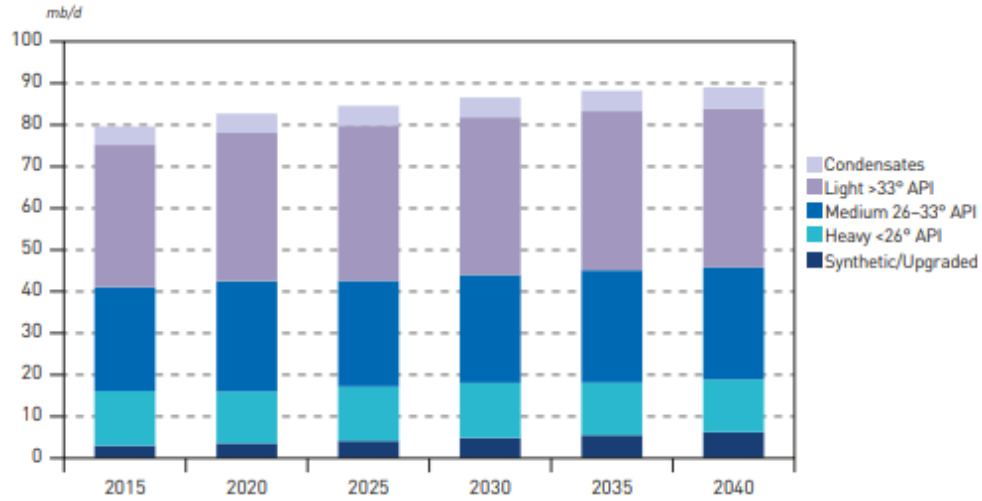


Figure 9: Global oil production by type projection [22]

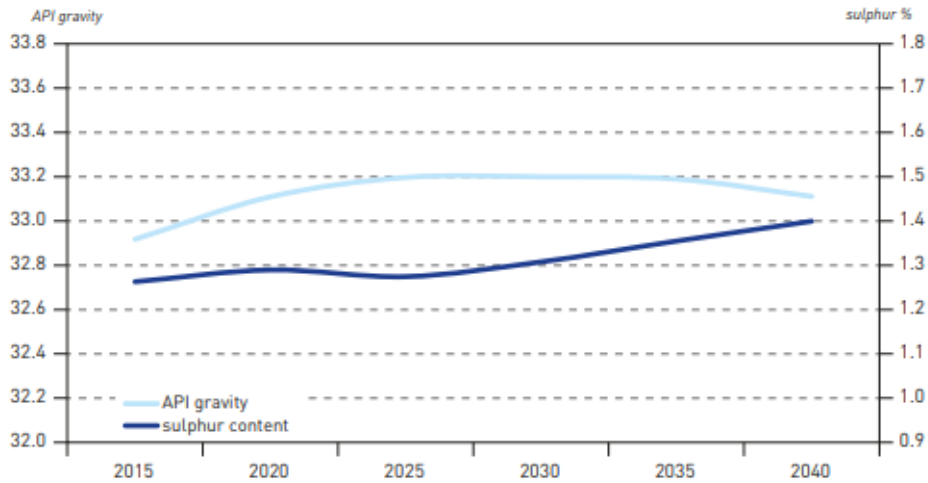


Figure 10: Global oil quality [22]

Coking additions are expected to take place during the period of 2017 to 2022 as shown in figure 11, where the y-axis represents the conversion units, in million barrels per day (mb/d). The coker throughputs are reduced due to the current lightening of the global crude slate. The introduction of global marine fuel standard for 0.5 % sulfur fuel in 2020 is expected to raise the throughput of cokers.

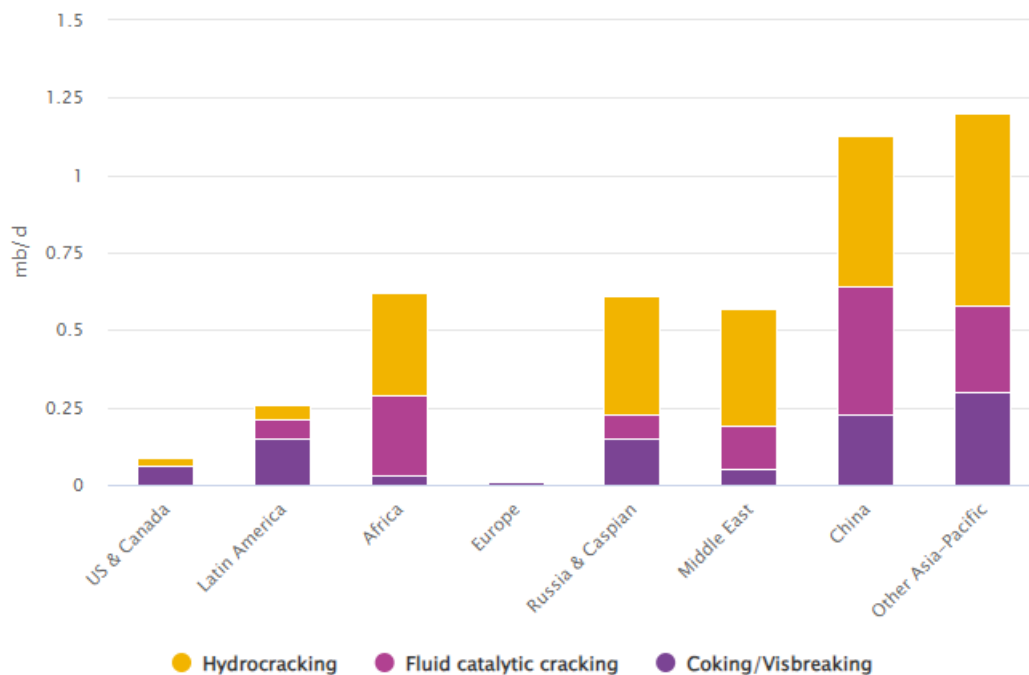


Figure 11: Conversion projects by region 2023-2028 [23]

Recently, methods have been developed to tap into reservoirs of shale oil, previously thought uneconomic. These fall into the category of light sweet crudes with very low sulfur and impurities content. However, the downside is that this type produces very little coke, even though it is of good quality. Thus, the coke production from shale oil is only 1 % of the crude oil barrel [26]. For the aluminum industry this means that coke blending will become more important to manage the qualities of the anodes.

To maintain competitiveness, modern smelters are under constant pressure to increase amperage, thus increasing production. However, pushing the cells towards their limit and beyond, means that the anode quality has a bigger impact on overall pot stability. Constant, good, anode properties are a prerequisite for stable cell operation. Today's variability in raw materials makes consistent anode production challenging. It is thus of high importance to find ways to predict both the impact of raw materials on the anodes and the effect of anode properties on the cell operation. In this way corrective action can be taken before production problems arise and impact the quality and/or amount of aluminum tapped from the pots.

The quality of the anodes is highly influenced by the raw materials used. The coke is produced from crude oil residues or tars from the coal industry. Therefore the properties of the coke

and pitch change according to the oil being processed. These are hydrocarbons which undergo phase transition from liquid to solid by cracking, dehydrogenation and polymerization [9]. The first stage of production is the green coke which has an amorphous, weak structure, high reactivity, poor electrical conductivity and 8 to 12 % volatiles. In the second stage, the green coke is calcined at a temperature of 1250 to 1350 °C. This step increases the carbon-hydrogen ratio and removes the volatiles. The grain strength is increased ensuring minimum shrinkage during anode baking while the pore structure is accessible to the binder. Calcining gives the coke the necessary thermal conductivity for the baking process [9].

There are three main types of coke that are produced. Asphaltene precursors produce mainly shot coke, which has a spherical shape. It is hard and has low porosity. High aromatic hydrocarbons feedstocks result in the formation of needle coke. It has an open porosity that allows for good pitch penetration during mixing resulting in a strong material after baking. It is the preferred raw material for anode production with a coefficient of thermal expansion of  $3.5 - 4.8 \cdot 10^{-6} K^{-1}$ . The third type is sponge coke which is a mixture of needle and shot coke, with an amorphous appearance [27]. These are shown below, in figure 12.

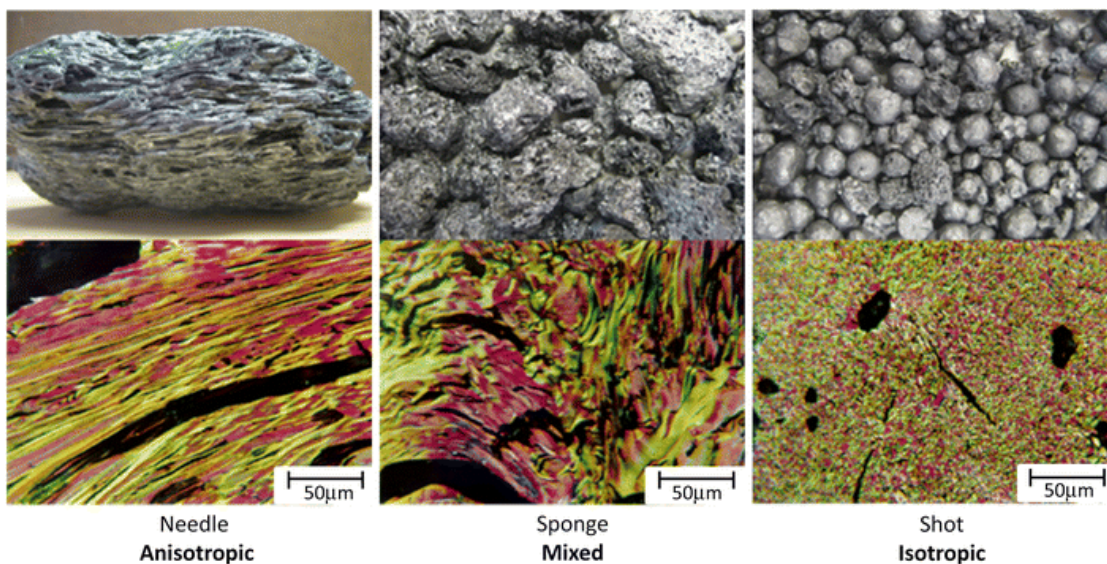


Figure 12: The three different coke types [19]

A thorough assessment of the impact of calcined petroleum coke on anode production for the aluminum industry is done by Edwards [19]. Depending on the recipe, the pre-baked anodes contain 55-65 % calcined petroleum coke (CPC), 13-15 % coal tar pitch binder and 20-30 % recycled butts.

The challenges faced by the industry are that it is harder to obtain green petroleum coke (GPC) with low sulfur and low metals. These are effects of changes in crude oil quality and refining economics. The  $SO_2$  emission limits determine the amount allowed in the anode recipe but the coke availability is low which drives prices up. Changes in the crude oil and refining economics have led to a higher impurity in anode raw materials.

Other authors have also expressed their concerns on the declining quality. Mannweiler et al. [25] assessed the availability of carbon products for the aluminum industry with respect to the growing demand of the metal production. Plants increase amperage to raise capacity often without the possibility of expanding the anode plant while some shut-down smelters are restarted without the anode plant. In some countries, like Iceland, anode production is not possible due to government restrictions. Soderberg smelters conversion to pre-baked require heavy investment in modifying the anode plant.

The petroleum coke quality is an issue since the crude oils processed contain more sulfur and metals. The calcining of this type of coke is best done in a shaft kiln as compared to a rotary kiln that is used widely today. The coal tar pitch quality is also an issue, with more than half of it coming from China. The main concern is the consistency of their properties.

Another important raw materials for the anode production is the coal tar pitch used for the binder matrix. An overview of the main properties is presented in the paper by Baron et al. [24]. Recent trends show that the market is becoming more dependent on the Chinese market of coal tar pitch which has different properties than in the rest of the world due to the methods used to produce it. The definition of good pitch properties varies around the world according to the existing supply. The authors also stress the importance of quality anode raw materials for the increasing amperage in current smelters.

With the growing concerns about the coke market Edwards et al. [28] looked at the use of shot coke in anode production. This particular type is isotropic and has low pitch penetration and a high coefficient of thermal expansion. The authors found that by blending shot coke in the aggregate (up to 20 %) increased the baked anode density and butt weight. There was no evidence of cracking due to the higher coefficient of thermal expansion and no effect on the net carbon consumption or cell operation problems.

### **Process monitoring challenges**

Two of the issues in today's aluminum industry are the sampling frequency of the data and the raw materials' quality. The anode properties are measured on samples drilled out of baked anodes. These, however, are taken from about 1 % of the population of manufactured baked anodes. Additionally, the samples may not be representative of the whole anode block since their size is only about 0.13 % of the whole anode [29]. The analysis of the samples takes about 5 weeks, at which time the anodes are already inside the electrolysis cell [48]. Because of this time lag, smelters are not able to use this information in a proactive fashion. The bad anodes, if any, would already be in operation within the cells.

As explained above, there is a considerable decline and increasing variability of the anode raw materials. These, in turn, affect the anode quality. It is thus important to assess the potential negative impact that these anodes have on cell operation. The pot performance can be looked at from different angles. One of them is determining the individual pot stability based on process

variables, equipment and operator input and interventions. Another performance method can be the excess carbon consumption. Both of these are explored in this work.

The technological improvements of the past two decades have greatly improved data collection and processing power for analytics. The emergence of advanced techniques allowed for reducing data complexity and identification of latent variables that could indicate process health [58, 61]. These were later shown to have potential in setting up process monitoring schemes and determine abnormal events in aluminum electrolysis cells [46, 49, 50]. New sensor development have paved the way to measure individual anode current. These additional measurements have recently been used in building new monitoring schemes to determine the onset of an anode effects [69] or anodic incidents [70]. However, these are not yet available for all smelters or even full potlines so they are not used for plant-wide monitoring schemes.

In today's production environment, individual pot health monitoring is usually done by combining process knowledge and experience with statistical process controls and reports. The smelters strive to detect process anomalies as early as possible, to mitigate their impact and maintain stable operations. While they collect an increasing amount of data, many key process variables are not measured continuously due to the harsh environment of aluminum electrolysis. The data quality and quantity available makes it challenging to develop advanced analytics tools for continuous process monitoring in production.

The excess carbon consumption, as mentioned in section 1.2, is the carbon consumed beyond the theoretical necessity to produce aluminum. This is influenced by both the process upsets in the electrolysis cell but also anode properties, tracing back to raw materials. Anodic incidents can manifest in the form of spikes or deformations that occur on the bottom of the anode, i.e. at the bath interface. These phenomena increase the localized current density and the rate of carbon oxidation, promote carbon dust formation and increased electrolyte resistivity. These result in the decrease of a cell's current efficiency and the increase in the net carbon consumption. When a pot is experiencing an anodic incident, the anode in question has to be replaced with a fresh one. Another option in this situation would be to scrape the deformities from the bottom of the anode.

Individual aluminum producers may use proprietary equations to predict the NCC, however, since the early 1990's the principal way of predicting the NCC has been by using a formula developed by R&D Carbon Ltd [17]. This is a linear model, equation (7), based on data from many different smelters, collected over a period of over 20 years up to the time of its publishing, and averaged on a potline basis over long periods of time. The technology involved contained currents of 180kA and lower.

$$NCC = C + \frac{334}{CE} + 1.2(BT - 960) - 1.7CRR + 9.3AP + 8TC - 1.5ARR \quad (7)$$

where NCC is the net carbon consumption, C the cell factor, which takes into account different technologies, CE the cell's current efficiency, BT the bath temperature of the pot, CRR the CO<sub>2</sub> reactivity residue, AP the air permeability, TC the thermal conductivity and ARR the air reac-

tivity residue. Pot operation behaviour reflects onto the amount of metal produced and therefore the CE is a good measure of the pots' health. The CRR and ARR are standard tests performed on the carbon samples to find the residue that remains after the block has been exposed to CO<sub>2</sub> and air respectively. As AP increases the residues decrease since there is less carbon that is consumed during the test. As higher residue is an indicator of lower reactivity, these parameters (CRR and ARR) have negative coefficients.

In the present literature there are a number of studies that have been performed in order to better understand the relationship between carbon consumption to the properties of baked anodes, raw materials and cell operation. However, in estimating the net carbon consumption, the R&D Carbon formula is still widely used. One such use can be found in the paper published by McClung et al, where they use the R&D Carbon equation to find the best combination of core sample properties for maximizing baked anode performance [30]. In another study performed at Alumar, Piffer et al., used the formula to optimize the final baking temperature and minimize the variations in the pits in the anode manufacturing process [31].

Practical NCC values today range between 400 and 450 kg of carbon per tonne of aluminum produced [32]. In order to reduce the NCC towards its theoretical value via process control adjustments, the industry needs to measure it frequently and on an individual anode basis. In the context of Industry 4.0, the presence of anode tracking systems at the smelters are becoming more frequent [33, 34]. Their development is driven by the need to extract useful information from the production process as a whole to better optimize it [35]. The declining quality of the anode raw materials [19] plays a big part in the need to optimize anode performance in the electrolysis cell.

## Research Objectives

This thesis seeks to show how latent variable methods can be applied in the aluminum production process with the overarching goal of reducing the greenhouse gas emissions. Thus the objectives of the thesis can be summarised as follows:

1. Demonstrate how latent variable methods can be used to detect aluminum reduction process abnormalities;
2. Show how aluminum reduction process abnormalities can be diagnosed using latent variable methods;
3. Demonstrate how a latent variable model can be used for process monitoring and help the operators prioritize the problematic cells;
4. Collect and analyse the data from an anode tracking system to estimate the net carbon consumption for individual anodes using latent variable methods;
5. Identify links between individual anode properties and their performance in the reduction cells.

## Contributions

This project has resulted in diverse applications for latent variable methods in aluminum electrolysis. These were used to both set up a new monitoring scheme in production at one of Alcoa's smelters and unlock new insights into relations between process variables. From this work three journal articles have been produced and published, along with one peer reviewed conference article.

Journal publications:

1. P. Manolescu, C. Duchesne, J. Lauzon-Gauthier and G. Saevarsdottir (2022). Net carbon consumption in aluminum electrolysis: impact of anode properties and reduction cell-operation variables. *Journal of Sustainable Metallurgy*, 8(3), 1167-1179.
2. P. Manolescu, C. Duchesne, J. Tessier and G. Saevarsdottir (2020). Estimation of net carbon consumption in aluminum electrolysis using multivariate analysis. *Journal of Sustainable Metallurgy*, 6, 739-750.
3. P. Manolescu, C. Duchesne, G. Saevarsdottir and J. Lauzon-Gauthier (2025). Multivariate monitoring of reduction cell operations: model performance assessment - *JOM*, 77, 3635-3650.

Conference article:

1. P. Manolescu, C. Duchesne, J. Tessier and G. Saevarsdottir (2018). On the Use of Multivariate Statistical Methods to Detect, Diagnose and Mitigate Abnormal Events in Aluminium Smelters. In *Light Metals 2018*, ed. by O. Martin (Cham, Springer, 2018), pp.475-483.

These articles form the main body of the thesis as discussed in the preface. Supplementary materials from the articles along with appendices are presented at the end of the thesis. The contributions of the thesis are presented below. These are shown in the order in which they appear in the text.

- **Contribution 1**

A latent variable model was built on operation data from the Alcoa Fjardaal smelter to detect and interpret an excursion at the smelter. A link between the trace elements and the anodic incidents was established using this methodology.

- **Contribution 2**

A generalisation stemming from the work in contribution 1 resulted in a latent variable model that was implemented in production at an Alcoa smelter to monitor potline operations. A new reduction cell ranking system was developed to help operators prioritize the daily problematic cells identified by the monitoring scheme along with contribution plots to help identify the potential problems.

- **Contribution 3**

Contribution 3 also stems from contribution 1, where the model performance of the monitoring scheme (contribution 2) was assessed and compared to existing tools at the plant. With the help of operators, individual cell investigations were performed on the shop floor

showing the potential of advanced notice using a latent variables model.

- **Contribution 4**

As anodic problems were identified as the source of abnormal operations in contribution 1, the decision was made to work on establishing a link between anode properties and cell performance. To do this, data from the historian was collected and fused to link individual anodes to pot operations using an anode tracking system at Alcoa's smelter in Deschambault. Thus the information collected for each anode contained green anode properties, baked anode properties and the pot operation data into which the anode was set. Using this database a latent variable model was built to predict the individual anode net carbon consumption. This multivariate approach was compared to the existing industry standard net carbon consumption method.

- **Contribution 5**

Contribution 5 stems from contribution 4, aimed at studying in more depth the contributions of different groups of variables on NCC. The anode tracking dataset was split into multiple, sequential blocks based on process flow diagram. The SMB-PLS algorithm was employed to identify the most important data blocks in the individual anode net carbon consumption. This led to establishing a link between green anode properties and the individual anode net carbon consumption.

# Chapter 1 On the use of Multivariate Statistical Methods to detect, diagnose and mitigate abnormal events in aluminium smelters

## Résumé

Ce travail vise à démontrer comment l'utilisation de méthodes à variables latentes permet de détecter et diagnostiquer l'apparition d'une situation anormale dans les cuves d'électrolyse dans la production d'aluminium. En utilisant des données récentes d'Alcoa Fjardaal, l'analyse en composantes principales (ACP) a été utilisée pour modéliser les variations typiques se produisant dans 336 cuves différentes. Une déviation dans le procédé a été correctement identifiée plus tôt qu'avec les outils de contrôle statistique traditionnelles. De plus, la concentration faible des d'éléments-trace dans le métal a été associée à une déviation de procédé. Une telle détection hâtive aurait pu contribuer à atténuer les effets des événements anormaux.

## Abstract

This work aims to demonstrate how latent variables methods can be used to detect and diagnose the onset of an abnormal situation in aluminium reduction cells. Using recent data from Alcoa Fjardaal, Principal Component Analysis (PCA) was used to model typical variations occurring in 336 different pots. A drift in process operation was correctly identified sooner than with traditional statistical control techniques. Additionally, low concentration in trace elements in the metal also corresponded to a drift in process operation. Such an early warning may have helped mitigate the impact of abnormal events.

## 1.1 Introduction

In aluminium smelting process variability is common due to the abundance of factors that affect one another. A common and constant factor in the operation of a pot line is to minimize the variability of process variables. In the case of the present study, it started with a few pots becoming hotter and hotter, reaching temperatures of over 1000 °C. The number of extra anode changes increased dramatically as did the carbon dust in the cells. By changing more than two pairs of anodes at any given point, within the same cell, the electrode plane is significantly altered resulting in uneven current distribution. Alternatively, hot anodes from an adjacent pot can be placed in the sick cell, so that the current load-up time would be minimized, and therefore, alterations to the anode plane and pot stability. However, a major drawback is the crane time necessary to perform the extra anode changes as well as scooping up the carbon dust. Treating one sick pot requires extended crane time which affects the regular flow of operations. In order to save the potline, multiple cells had to be stopped resulting in a significant loss of revenue. These types of abnormal behaviours occur throughout the industry, their impact on production varying according to the severity of the excursion. Statistical process control (SPC) systems to detect these anomalies are highly desired to help maintain a stable process operation.

The particular nature of aluminum reduction cell data needs to be taken into account when developing SPC schemes. Some key operation variables have a relatively low sampling rate such

as temperature and liquid level. There are a number of hurdles in improving measurements such as the molten bath and metal damaging immersed sensors and large operating area requiring distribution of equipment and tasks. Others are the constraint of taking measurements, typically, at the end of the cell, through the tapping hole, which may not be representative of the whole [36]. Other variables such as resistances and feeding are recorded automatically by the pot controller, with a higher rate (minutes/seconds or higher). This results in a large amount of data with different sampling rates. Often, these variables are highly correlated and it becomes difficult to interpret them individually, on univariate charts. It is thus of benefit to apply multivariate analysis in order to extract useful information from the pool of measurements. Principal Component Analysis (PCA) and Projections to Latent Structures (PLS) have been shown to be able to cope with large, industrial datasets. Contributions to their industrial applications have been made by [37–39, 42, 43] and [44].

The application of these statistical methods to the aluminium industry is quite new but a few others have touched on the subject. Tessier et al. [45, 46] have studied pot performance against alumina and anode properties as well as the influence of the cell start-up period on the overall lifespan of the individual pots. [47]. Majid et al. [49, 50] were able to detect anode effects and faults via multi-way PCA (MPCA). The approach relied on the resistance measurements during the underfeed-overfeed cycle between anode settings, which was treated as a batch process. They later proposed a new method based on Partial Least Squares - Discriminant Analysis (PLS-DA) to diagnose faults of anodic and non-anodic nature. A multivariate image analysis method was developed by Tessier et al. for determining the anode cover material composition [51–53].

The present work uses PCA to detect and interpret the excursion data. The next section focuses on explaining the data and methodology followed by a brief introduction to PCA. Subsequently, there is a description on how PCA model was built superseded by the interpretation of the projected data. Lastly, the conclusion summarizes the findings.

## 1.2 Data and methodology

Pursuing to determine the deviations from normal operation conditions, a latent variable model was built and the excursion data projected onto it. The main tool used was the PLS Toolbox (Eigenvector Inc.) multivariate package for Matlab [54]. A PCA model was built on pot operation data from the first three months of the year 2016 until the excursion starting in April. The selection was restricted to cells aged 60 days or older, since their operating parameters are different in the early life. The data is based on daily averages of cycle data along with liquid level measurements as well as StarProbe<sup>TM</sup> [55] information (temperature, fluoride content, alumina in bath, superheat) and impurities in the metal. Others are anode problems in the form of number of extra sets and anode lowers/raises, as well as bubble and metal noise, along with the age of the cell. All in all, data from 38 different variables were collected for each individual pot (of the 336 cells in the line) and analysed with PCA. The data was collected and placed in a matrix  $\mathbf{X}$  where each row contains all the variables for a given pot at a particular day. Thus, the matrix size was 38 columns by 17692 rows. In this dataset, however, only the days with StarProbe<sup>TM</sup> and liquid measurements were extracted. The reason lies within the organizational chart of the process, with measurements taken at regular intervals, resulting in gaps in the data. These blanks

in the metal sampling were filled in by using linear interpolation. It is worth mentioning that these values were calculated prior to linking the date of StarProbe<sup>TM</sup>/liquid level measurements. This method ensures minimal information loss.

### 1.3 Introduction to PCA

A good explanation of the method can be found in the statistics book by Kevin Dunn [56]. A brief description is provided here but the interested reader can find more details presented by [57–59] and [60].

PCA is a multivariate analysis tool that can effectively reduce all the variables to just two terms. These are the squared prediction error (SPE) and Hotelling’s  $T^2$ . The raw data is preprocessed, i.e. centered and scaled in order to have zero mean and unit variance in each column.

The PCA model can provide estimates for each observation, i.e, scores ( $\mathbf{T}$ ), which are linear combinations of the original variables using the loadings ( $\mathbf{P}$ ). These estimates depend, however, on the number of components used and results in a residual matrix also. Mathematically, this is:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \iff \mathbf{X} = \hat{\mathbf{X}} + \mathbf{E} \quad (8)$$

where  $\hat{\mathbf{X}}$  is the predicted  $\mathbf{X}$  matrix and  $\mathbf{E}$  is the residual matrix.

The loadings represent the correlation between the variables in the  $\mathbf{X}$  matrix. They are perpendicular to one another and they are oriented in the direction of greatest variance. The scores are a linear combination between the loadings and the data matrix, shown in equation (9). The loadings define a lower dimensional subspace and the scores are the projection of the data onto that subspace. They show how the observations cluster together.

The number of PCA components can be selected based on a scree plot. A description of the method is found in Westerhuis et al. [61]. Another method is to look at the root mean square of cross-validation (RMSECV) and calibration (RMSEC) plotted as a function of the number of components, used in this work. As the number of components increase the gap between these two metrics increases. The adequate choice is when the gap between these metrics is small. This approach can be used to validate the scree plot method. Selection of the number of components is thus subject to interpretation and is not a fixed number for every model. If the gap between RMSECV and RMSEC is big then the model would essentially be fitting noise, so one should be cautious in seeking to maximize the total  $R^2$  value.

The scores can be further summarised as the Hotelling’s  $T^2$ . The mathematical formulas are presented below.

$$\mathbf{T} = \mathbf{XP} \quad (9)$$

and

$$T_i^2 = \sum_{a=1}^{a=A} \left( \frac{t_{i,a}}{s_a} \right)^2 \quad (10)$$

where  $s_a$  is the standard deviation of score column  $a$ ,  $A$  is the number of columns in matrix  $\mathbf{T}$  and  $t_{i,a}$  is the score value for the corresponding row  $i$  and column  $a$ .

The  $SPE$  is the squared prediction error or, in other words, the distance off the model plane to the point, and is given by:

$$SPE = \sqrt{\mathbf{e}_{i,a}^T \mathbf{e}_{i,a}} \quad (11)$$

where  $e_{i,a}$  is the residual vector. The indexes represent the row number (i) and component number (a).

The  $SPE$  is the distance from the model plane to the point, while the  $T^2$  is the distance from the origin to the projection of the point on the plane. These two variables summarize the scores and residual matrices and show whether there are deviations from the correlation between the elements of the model. If one value of the  $SPE$  is much higher than the others, it means the relationship between the variables (of the pot it represents) is different than for model based on the historical data. If the  $T^2$  is high, it means that the values in matrix  $\mathbf{X}$  are higher or lower than the average operating conditions, since the loadings' vectors are fixed. The weight of each variable can be observed through the contribution plot of the Hotelling's  $T^2$ .

### 1.3.1 Building the PCA model

The model was built on normal operation data, with nine components, resulting in a 64.4 % of variance explained. The outliers were investigated and removed, which included 5 pots due to various operational anomalies as well as the days from the March 1st to 5th because of a power outage at the plant. Overall the model included 17692 data points. The loadings of the first two components are plotted and interpreted below.

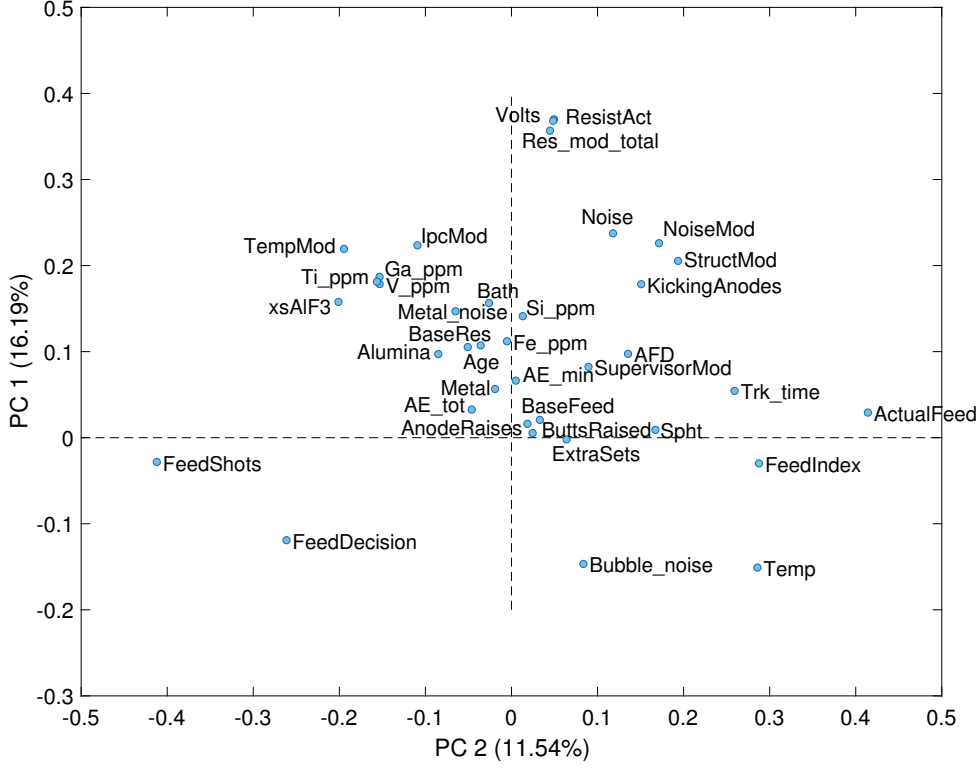


Figure 13: Loadings of the principal components 1 and 2

As mentioned above, in equation (8), PCA expresses the initial, centered and scaled, data as a linear combination of matrices  $\mathbf{T}$  (scores) and  $\mathbf{P}$  (loadings) plus an error matrix ( $\mathbf{E}$ ). Thus the loadings are in fact the weight of the contribution of each variable for that specific component. Loadings carrying the same sign are positively correlated while the loadings with opposite signs are negatively correlated. The absolute value of the loadings for each variable indicate the weight of the variable in the component. Thus the higher the absolute value of the loading, the higher the weight of the variable.

The first component explains variation in voltage and thus resistances along with noise and anode problems as well as temperature. There is a strong correlation between the noise, resistance, and thus voltage. If the pot is noisy, extra resistance (modifier) is added to the pot in order to increase the anode-cathode distance (ACD) and allow the pot to settle down. This in turn can affect the temperature. If there are anode problems they would be picked up by the anode fault metrics. Bubble\_noise and Metal\_noise are relative noise values and complementary metrics where, if one goes up the other goes down, which is shown by the plot as well. The trace elements, in particular Ga and V, can indicate anode issues as well [62]. The second component explains the alumina feeding along with temperature and anode problems.

### 1.3.2 Projections onto the PCA model

The approach used was to build an overview of the Hotelling's  $T^2$  for all the pots, i.e. a timeline of the  $T^2$  evolution, explained in more detail below. The  $SPE$  metric, given by PCA, was also investigated. However, the focus was not on it because it is more sensitive to variance in metrics that are usually very close to zero. One such metric is the number of extra sets which typically should be zero on a daily basis. If an anode spike occurs in the given cell, resulting in a premature anode change, the number of extra sets would increase to one. The PCA analysis would interpret this as a deviation from normal operation (i.e. zero extra sets) which would increase the absolute value of the  $SPE$  metric. During the excursion period, however, the extra sets numbers were higher than zero, especially for problem pots, which make the  $SPE$  spike out of bounds. It indicates the model works but it makes it difficult to pinpoint other problems with the pot.

The investigation method is shown schematically in figure 14, below. The scores are represented by the top block in figure 14, as a matrix of  $N$  (336 pots) by  $M$  (days - the time-frame of the model/projections). From that, the bad performing pots can be identified easily and their individual  $T^2$  timeline can be investigated. It is represented by the second block in diagram 14, where the matrix is now a vector since it represents only one pot. Its length is  $K$  (days) which can be different than the overall length  $M$ . That is because, in the vector, only available data is used while the  $N \times M$  matrix above it contains missing data as well, depending on the measurements cycle (explained further below). The points that fall outside the normal operation range can be further studied using the contribution plot tool which points out the weight of each variable in determining the  $T^2$ . A link can thus be established with the raw data and develop an interpretation.

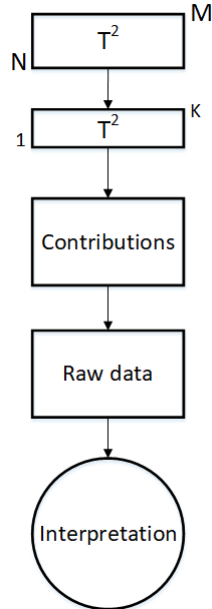


Figure 14: Hierarchy of the analysis

In order to have an overview of the behaviour of the whole potline heatmaps were created for the Hotelling's  $T^2$ , for each pot. The figure 15, below, shows the data prior to the excursion, i.e. the normal operation data. The pots are split accross 2 rooms, 4 quarters per room and 42 pots per quarter. Each pot is represented by a line in the plot. The cell number corresponding to the last one in each quarter was marked on the graph. Thus the chart starts with cell 1001 (Room 1, pot 1, unmarked) and ends with cell 2168 (Room 2, pot 168, marked). The x-axis shows the dates in an ascending order. The blue lines represent missing data on that particular day. The reader is reminded that only the days with StarProbe<sup>TM</sup> and liquid level measurements were included in the analysis. On February 9th the plant changed its organizational work schedule resulting in more frequent StarProbe<sup>TM</sup> measurements, for each pot. The green to orange colours represent pots operating within the range of normal behaviour, with a  $T^2$  between 0 and 1. The orange to red lines show the deviations, cells with a  $T^2$  greater than 1. The intensity of the colour indicates the severity of the drift. In figure 15, the model data, the continuous, horizontal, blue lines represent the pots that were left out of the model building. Similarly, in figure 16, the projections, the stopped cells appear as continuous blue lines from their shut-down onward.

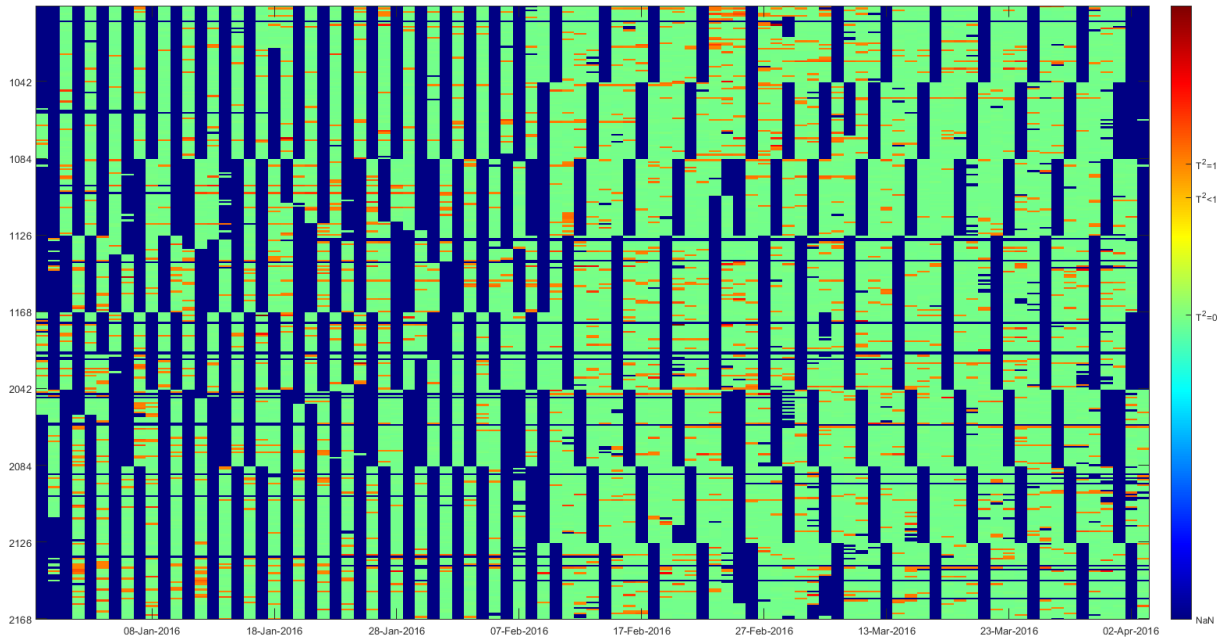


Figure 15: Heatmap of the model data.

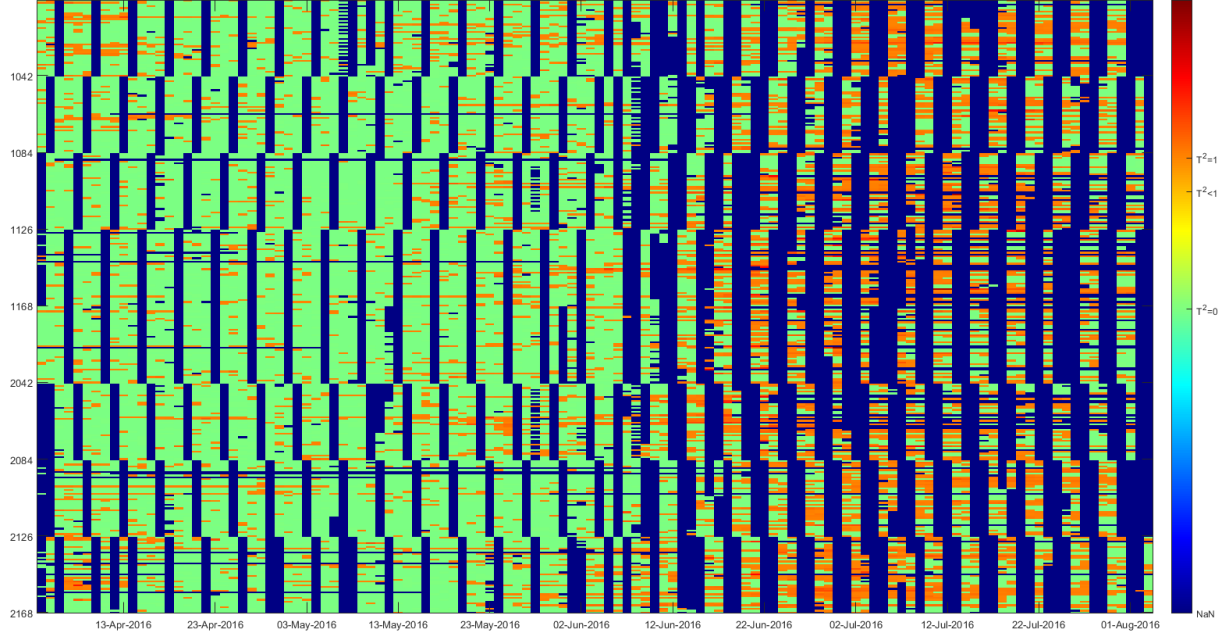


Figure 16: Heatmap of the projected data. The excursion onset is roughly the beginning of June.

The above figures show the  $T^2$  evolution, before and during the excursion, respectively. In figure 15 the data is generally green with some red areas. That is to be expected since, during normal operation, there is a certain degree of variability and cells can experience a few bad days. However, the interesting feature is that, generally, the cells become red but they return to green after a relatively short time. In contrast, the excursion data exhibits a lot more deviations for a prolonged period of time. It is easy to see how the potline shifts to red from June onward. In order to find an explanation for these drifts the pots were individually projected onto the model, i.e. pot by pot. As an example one particular pot's (A110) investigation is presented here. The pot's behaviour can be observed by looking at the timeseries of its  $T^2$ , in figure 17.

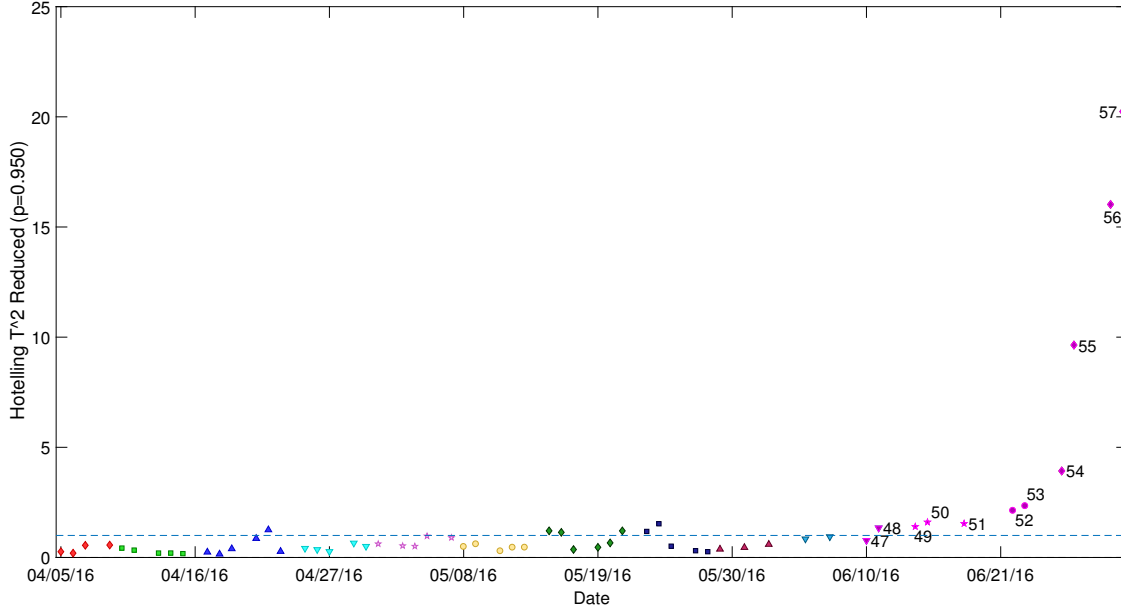


Figure 17: Pot A110 example - timeseries of its  $T^2$  values.

Around the 10th of June the  $T^2$  values start to rise, deviating from the model. Its contribution plot is shown below, in figure 18:

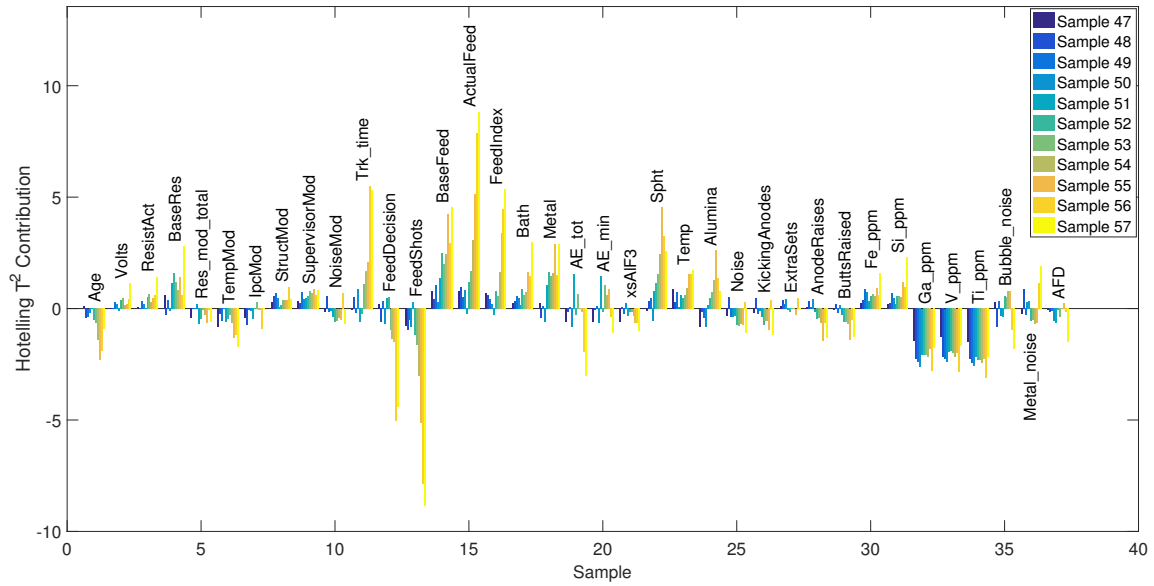


Figure 18: Pot A110 example -  $T^2$  contribution plot for samples 47 to 57.

From the contribution plot for the above cell, it is easily seen that trace elements such as V,

Ga and Ti have negative spikes. The samples are ordered by the date at which they occurred. This points out that the concentration of trace elements in the metal drops before the cell's feed control parameters, superheat, temperature, alumina concentration and anode fault metrics go out of bounds.

In their work, Dion et al. [62], postulated that trace elements in the metal sample can predict anode problems. There, the authors found that the presence of anodic incidents (spikes, deformations) increase the localized trace elements gaseous evolutions, resulting in a lower concentration of them in the metal.

During the excursion the plant experienced a significant increase in anode problems that resulted in poor pot performance. The feed control of a pot is a good indicator whether the cell is operating within its designated range. It was thus decided to plot the time-series of the trace elements along with key feed control parameters.

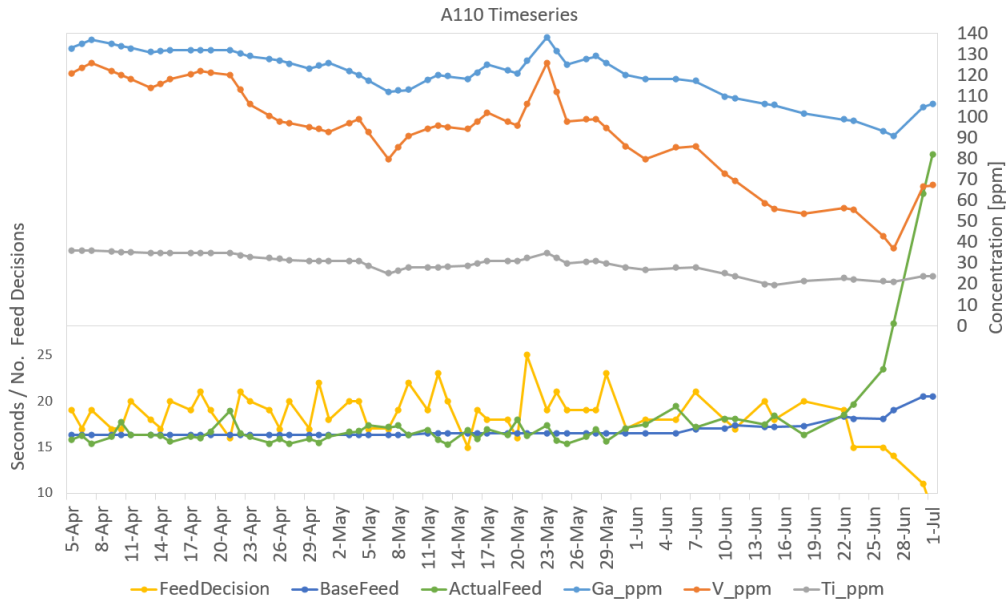


Figure 19: Time-series of trace elements and feed control parameters in pot A110

It can be seen that this pot exhibits drops in trace elements, most obvious, V and Ga, before the actual and base feed increase and the number of feed decisions drops. Looking at pot's V and Ga there is an overall trend downwards from the beginning of the time frame i.e. April, however, there is some variation, while the feed control displays good, normal, behaviour before the end of May. On May 23rd there is a peak in Ga and V followed by a sharp drop and a trend downward with a steep gradient. The feed control starts to pick up with a localised peak on June 5th and the base feed increases on the 6th of June. In this example the feed control lags the trace elements by about 10 to 12 days. On the 22nd of June, the pot becomes very difficult to operate leading to its shut-down on July 1st.

Other individual pots that were investigated showed similar symptoms. The analysis was performed on both cells that were stopped as well as on ones that continued operation after the

excursion. They follow a similar trend i.e. a decrease in trace elements followed by a detrimental response in feed control 10 to 12 days later.

A summary of the Hotelling's  $T^2$ , along with the SPE for the whole pot line, was plotted in figure 20 in order to see if there are any obvious trends specific to this time frame.

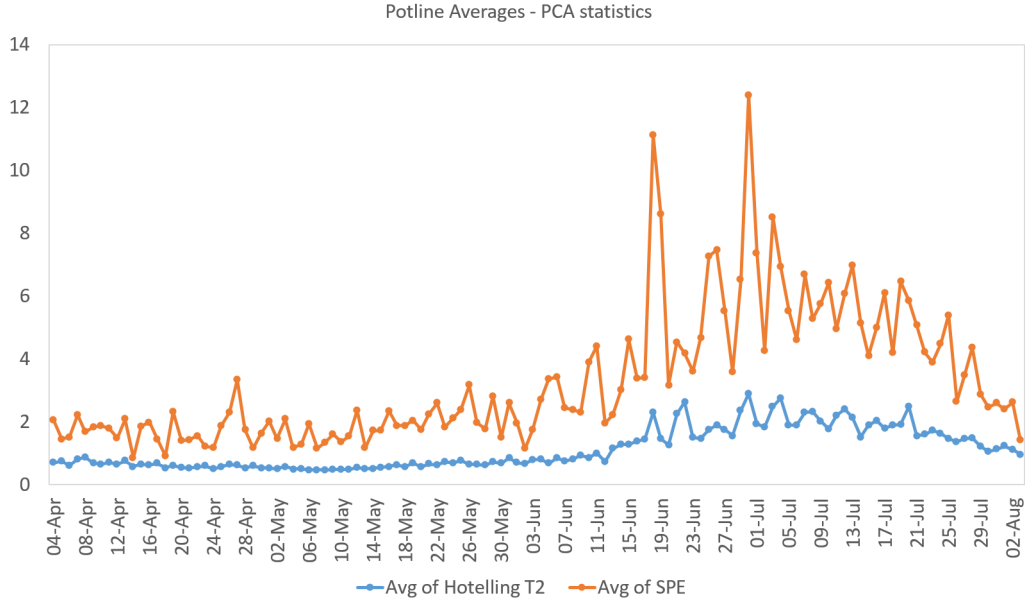


Figure 20: Potline Summary Statistics

From the above figure (20), it is easy to spot that the  $T^2$  starts to trend upwards in the second half of May. The same goes for the SPE. The potline averages for the trace elements were plotted below, in figure 21. They too, show a downward trend starting with the end of May followed by more variance between individual pots. This coincides with an increase in anodic incidents at the plant. The response in feed control starts on June 7th. This is consistent with the approximately 10 days lagging effect shown in the individual pot example above.

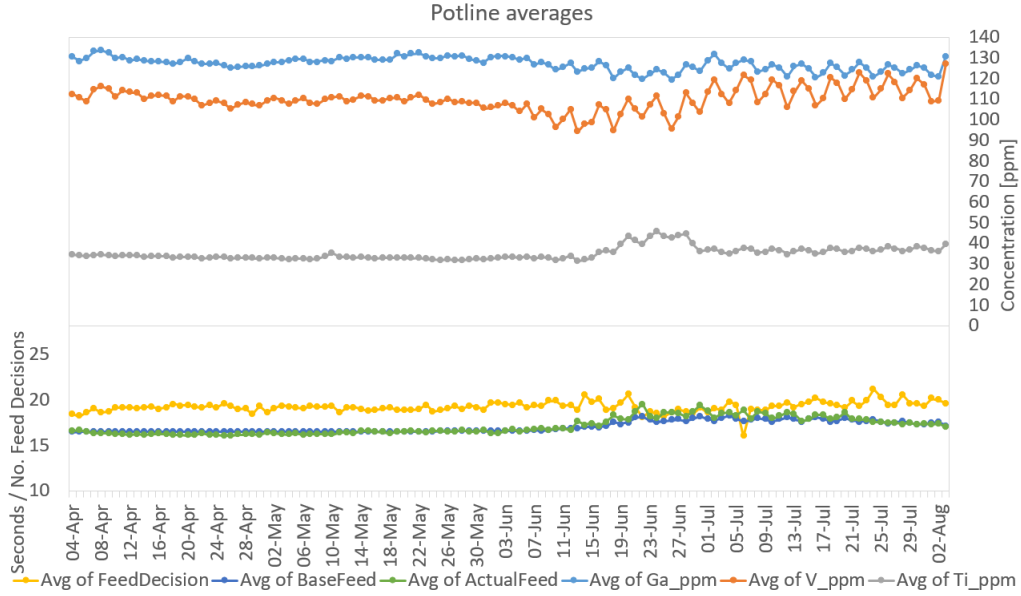


Figure 21: Potline Trace Elements

## 1.4 Conclusion

In this paper advanced multivariate statistical tools were used to determine and investigate problems that can occur in aluminium smelting. It is shown that by building a model on normal operation, with a good amount of variation, drifts can be observed on the testing data. In the example, it is shown that PCA picks up that trace elements in the metal can warn about the feed control and operation of a cell. Low concentrations of Ga and V in the metal are correlated with anode problems that can cripple production. A lag of approximately 10 days between low trace elements and detrimental feed control response is observed. Thus, trace elements could provide an early warning of problems to come. Other cells investigated for this work showed unstable pot operation due to fluctuations in bath level, which were also picked up by the PCA model.

These statistical tools can, therefore, be used to investigate and explain variability within the process that are not immediately obvious, especially considering the very large number of pots. An appropriate reaction plan can be set in motion as soon as the problem is identified.

## Chapter 2 Multivariate monitoring of reduction cell operations: model performance assessment

### Résumé

Les perturbations de procédé sont des phénomènes qui se produisent plus souvent que souhaité dans les opérations d'électrolyse de l'aluminium. Dans certaines conditions, elles peuvent causer des fuites de métal de la cuve. Dans des travaux précédents, les auteurs ont montré que les modèles à variables latentes, entraînés sur des données de procédé, peuvent détecter des situations anormales. Dans ce travail, un modèle d'analyse en composantes principales (ACP) a été construit sur des données de procédé des cuves d'électrolyse et mis en production dans l'aluminerie d'Alcoa Fjardaal. Un système d'ordonnancement des cuves, fondé sur les résultats du modèle, a été développé afin de prioriser les cuves problématiques. Les graphiques des contributions de l'analyse ACP se sont également avérés utiles pour identifier les problèmes potentiels. Les opérateurs ont examiné les cuves sélectionnées dans la salle des cuves, pour évaluer les performances du modèle. Une comparaison avec les outils existants dans l'usine a également été effectuée, montrant le potentiel de détection hâtive à l'aide d'un modèle à variables latentes. Des exemples d'anomalies détectées et de méthodes d'investigation possibles en utilisant le modèle multivarié sont également présentés.

### Abstract

Process upsets are phenomena that occur more often than desired in the aluminum electrolysis operations. In certain conditions these can lead to cell tap-out. In previous work, the authors showed that latent variable models, built on process data, can detect abnormal situations. In this work, a Principal Component Analysis (PCA) model was built on cell operation data and implemented in production at the Alcoa Fjardaal smelter. A cell ranking system was developed based on the model outcome, to enable a standard prioritization method of the problematic cells. The contribution plots of the PCA analysis were also shown to help identify the potential issues. Operators investigated selected cells on the shop floor, to assess the model performance. A comparison with existing tools at the plant was also done, showing the potential of advanced notice using a latent variables model. Examples of anomalies detected and investigation methods possible using the multivariate model are also presented.

### 2.1 Introduction

The Hall-Héroult process of primary aluminum production is an energy intensive process in which alumina is dissolved in molten cryolite, with aluminum metal accumulating at the bottom of the cell, via electrolysis of said alumina. A good reference covering the basic principles of aluminum electrolysis is found in [64]. Process stability is the result of a narrow balance between a multitude of variables. These include cell design parameters, raw materials composition and quality, chemistry within the cell, process control metrics, external factors such as ambient air flow and temperature along with manual, operator interventions. Minimizing process variability is thus the key in maintaining stable operations [36].

Various issues can arise in a potline that affect the overall stability. Some can be related to the

quality of raw materials such as alumina and fluoride additives as they can alter the chemical and thermal balance of the cell. The lack of alumina in the cell would result in the electrolysis of the cryolite bath itself, also called an anode effect. This increases the voltage of the cell along with greenhouse gas emissions, and could even damage the physical shell. Poor covering of the anodes can result in excessive carbon air burn, thus reducing the life of the anode while increasing the net carbon consumption. Low or high liquid levels are also detrimental to the process while the Silicon and trace elements levels also need to be monitored for metal quality and process health.

Yet other problems can be related to the anodes and their performance in the cells, such as spike formations. These are physical abnormalities on the bottom of the anode, which are protruding from the anode plane, thus bringing that part of the anode closer to the cathode. This, in turn, may create a short circuit, with the problematic anodes pulling in more current and disturbing the cell's current distribution. This would further cause thermal and chemical imbalance within the cell along with diminishing its overall current efficiency by almost 4 % [65].

Early detection of process upsets is highly important in aluminum electrolysis. For example, treatment of a problematic cell requires manpower and equipment availability. These resources would have to be diverted from other tasks within the potroom which impacts other pre-planned routine tasks and alters logistics. If there are many problematic cells in need of treatment at the same time, this could create a snowball effect resulting in more problems due to delays in routine tasks for normally operating cells. Through early detection of abnormal events, loss of current efficiency and higher emissions can be prevented. This also has a positive impact on cell lifespan and reduces the risk of cell tap-outs. These are particularly bad for a potline as the molten metal and bath could leak out of the pot. This poses health and safety risks for operators and can result in equipment damage and production loss [67].

With the technological advancements since the beginning of the 21st century, the aluminum production industry has been able to increase the amount of data analytics used in research and daily operations. Studies performed by Tessier et al. [46] showed that latent variable models can be used in aluminum smelting to establish multivariate specification regions for raw materials but also for plant-wide process monitoring.

Others have used multivariate latent variable models to build a cascading monitoring system based on the feed cycle variability to determine abnormal events such as anode spikes and anode effects [49]. The same authors also proposed a new statistical framework based on hierarchical diagnosis to distinguish between faults related to anodes and other faults. This framework uses discriminant partial least squares (DPLS) and treats the feeding cycle as a batch process to classify faults based on two hierarchical modules using two discriminant models [68].

In his PhD thesis, Yao [69] studied individual anode current measurements and built monitoring schemes for fault detection. The author compared linear and non-linear models using Principal Component Analysis (PCA) and Dynamic Kernel PCA (KPCA) to detect incoming anode effects within the electrolysis cell. While both models were able to detect the anode effect, the KPCA model required fewer components, returned a higher percentage of total variance explained and had fewer false alarms as opposed to the regular, linear, PCA model.

In another study, sensors placed on the anodes have enabled monitoring of individual anode

current signals. The authors built an anodic incident detection system using a PCA model based on the signals collected. The research showed that this method could detect anodic incidents before they occur. [65, 70]. Yet other researchers, have used multi-scale time series modelling based on hybrid deep learning to build predictive models of anode effects [71].

In the case of advanced analytics being deployed in production and used daily there are still gaps. Due to the difficult and corrosive nature of the electrolysis process many process variables are not continuously monitored. This induces gaps in the data collection along with uncertainty, making predictive analytics challenging to implement.

The authors have previously shown that a PCA model can be build on the daily averages of cell data and used to detect anomalies [72]. The model was built and tested on data extracted from the process data historian. Trace elements in metal samples were also seen to correlate to anode problems. The objective of this paper is to take the work further by deploying the model in production and assessing its performance in monitoring the cell operation using Hotelling’s  $T^2$  and Square Prediction Error (SPE) metrics [73]. Additional data transformations were done onto the training data and a ranking system based on the Hotelling’s  $T^2$  metric was developed. The model performance was assessed by observations on the shop floor and compared with the existing reporting toolset. The potential advance notice given by the model was also assessed and examples of problems identified are presented.

The chapter is organised as follows. First the data collection is presented followed by the methodology. This encompasses the algorithm used along with the ranking system developed. Performance of the model is shown in the results section along with the detection advance notice observed and examples of anomalies. This is then followed by a discussion and conclusion section.

## 2.2 Data Collection

The data collected for this work is from the Alcoa Fjardaal smelter that runs a line of 336 cells at about 380kA. The smelting industry has come a long way from its inception, with automation improving the process, data acquisition and operator safety. However, due to the particularly harsh conditions of an electrolysis cells, it is still difficult to obtain high frequency measurements of all process variables. Some of them, such as voltage and resistance, are registered by the cell controller at a high frequency. Using signal processing techniques, other metrics can be derived from them, such as noise. Many process variables still rely on manual measurements, for example temperature, superheat, bath chemistry, liquid level, and metal chemistry. These are limited to scheduled routine jobs by the operators every 24, 32 or even 48 hours, depending on the shift logistics of the plant. In addition, these measurements are limited to the tapping hole of the cell, which may not represent the whole cell. However, they do give a good indication of the state of the cell and, in this study, are assumed to represent the whole cell. As in any other industrial processes, bad data is also present, which adds another layer of complexity.

The variables extracted could be grouped into 6 main categories, listed in Table 2, along with the number of variables present in each one.

Table 2: Process variables categories

| Category number | Category name         | Number of Variables | Examples                                                                   |
|-----------------|-----------------------|---------------------|----------------------------------------------------------------------------|
| 1               | Descriptive Variables | 2                   | Cell Number, Location                                                      |
| 2               | Age                   | 1                   | Cell Age                                                                   |
| 3               | Process controller    | 25                  | Cell Voltage, Cell Resistance, Alumina and Fluoride feed variables         |
| 4               | Process set points    | 3                   | Alumina base feed, Bath level target, Cell Base Resistance                 |
| 5               | Manual Measurements   | 12                  | Liquid levels, Alumina content in bath, Cell temperature, Metal impurities |
| 6               | Calculations          | 13                  | Lagging variables, Differences between targets and measurements            |

The first category relates to descriptive variables which were the cell name and its location in the line. These were not used in the model building itself, but rather to identify the observations. In the second category, the age of the cell was selected, measured in days. However, cells aged below 60 days were excluded from the model because different setpoints are used for them and they behave differently within the early stages of their life. This would make them stand out from the rest of the population and skew the model. The next category were the process controller variables. These were measurements taken automatically at a high frequency and reliability, such as cell voltage, amperage, resistance, noise, alumina feed control, anode effect (AE) time, bubble and metal noise for a total of 25 variables. In the fourth category, some process setpoints were included. The ones chosen were the bath level target, base alumina feed rate and base cell resistance. Other setpoints such as metal level target and excess aluminum fluoride target were not included because they did not have enough variability to make them meaningful additions. The fifth category contained manual measurements taken by operators. These accounted for 12 variables that included liquid level measurements, bath temperature, superheat and chemistry given by the StarProbe<sup>TM</sup> sensor [55] and trace elements within the metal samples. The silicon mass balance for each cell was also extracted from the process historian and added to the data matrix. This variable can give an indication of process upsets related to the different sources of silicon within the cell. In the final category, additional, new, variables were calculated and added. The first type were the differences between liquid level targets and measurements along with the intra-day difference of the copper in metal measurements. The second type of calculations were time-lags added for a subset of variables to account for their dynamic (transient) behavior. Autocorrelation analysis was used to identify the variables needing lags and their number. In this case, the variables chosen were one day lags of vanadium and gallium impurities in the metal samples, cell resistance, cell noise, alumina feed rate and bubble noise. These accounted for 13 variables.

Certain transformations were applied to the dataset prior to model building. Because of the nature of operations, the manual measurements were taken on a 32-hour schedule, inducing gaps

in the dataset. Since the MSPC scheme is run at a faster sampling period than 32 hours, and concentrations of trace elements are required to compute the PCA scores, a simple time-weighted linear interpolation was employed to impute missing values of slowly varying trace elements. For the rest of the manual measurements, alumina in bath, alumina fluoride in bath, temperature, superheat and bath and metal levels, a PCA-based imputation algorithm was employed. This was done because, for example, temperature can vary more quickly between measurements and simple linear interpolation would result in unreliable data.

The data was organized in a single matrix. The raw data matrix was thus composed of 54 variables (columns) and 12139 observations (rows). Each row represented daily averages for each of the cells in operation during the period in question. The data was gathered over a two month period consisting of common cause variations only (i.e., natural process variability, free of abnormal events) for building the PCA-based MSPC models. This period was identified using the existing process stability charts coupled with expertise from on-site process engineers. After cleaning the data, the number of observations dropped to 10649, onto which the model was built.

## 2.3 Methodology

The multivariate latent variable method used to model the data (i.e., PCA) is described first. Then, the two phases of a process monitoring scheme are presented: model building (phase I) and model exploitation (phase II). Finally, the metrics used to assess the performance of the process monitoring scheme are discussed.

### 2.3.1 Algorithm

Principal Component Analysis is a multivariate latent variable statistical method aiming at explaining the variance of large data matrices. The algorithm is described in detail in [57], while some of its applications to process monitoring are presented in [74–76]. The main idea is that the data studied is pooled into a matrix and each observation can be summarized with two new variables, namely Hotelling’s  $T^2$  and Squared Prediction Error ( $SPE$ ) [76]. The first one measures the magnitude of the Mahalanobis distance [40] of a multivariate observation from the mean of the dataset on the latent variable space of the PCA model. The  $SPE$  measures how the correlation structure between the variables deviate from the dataset used to build the model (i.e., distance off the PCA latent variable space).

The main steps of the algorithm are as follows. The data is collected into a single matrix  $\mathbf{X}(N \times K)$  with  $N$  rows (observations) and  $K$  columns (variables). Next, the data is mean-centered and scaled to unit variance. This ensures that the weight of each variable is equal when fitting the model. Next, the loadings ( $\mathbf{P}$ ) and scores ( $\mathbf{T}$ ) vectors/matrices are calculated via the Non-Linear Iterative Partial Least-Squares (NIPALS) algorithm [57]. The model equation is shown in equation (12). The correlation structure of the matrix  $\mathbf{X}$  is stored in the loadings matrix  $\mathbf{P}$ . The vectors within this matrix are orthogonal to one another and point in the direction of the greatest sources of variance in  $\mathbf{X}$ . The scores matrix  $\mathbf{T}$  is found via linear combinations of the original x-variables with weights stored in the loadings matrix  $\mathbf{P}$ .

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \quad (12)$$

where the initial  $\mathbf{X}$  matrix can be written as the sum between the predicted matrix  $\hat{\mathbf{X}}$  and the residual matrix  $\mathbf{E}$ .

The scores thus represent the new latent variables, given by equation (13).

$$\mathbf{T} = \mathbf{XP} \quad (13)$$

These can be summarised in a vector given by the Hotelling's  $T^2$  from equation (14)

$$T_i^2 = \sum_{a=1}^{a=A} \left( \frac{t_{i,a}}{s_a} \right)^2 \quad (14)$$

where  $A$  is the selected number of latent variables of the PCA model, which is equal to the number of columns in matrix  $\mathbf{T}$  and  $t_{i,a}$  is the score value for the corresponding row  $i$  and column  $a$ , while  $s_a$  is the standard deviation of score column  $a$ .

The second summary metric is the Squared Prediction Error, calculated from the residual matrix as :

$$SPE_i = \sqrt{\mathbf{e}_i^T \mathbf{e}_i} \quad (15)$$

where  $\mathbf{e}_i$  is the residual vector for observation  $i$ .

### 2.3.2 Model Building (Phase I)

The data was collected under normal operating conditions. To gather maximum variability within the dataset, a single model was built for the whole potline, instead of one model per cell. After the acquisition, the data was cleaned. First step was to look at the trajectory of each individual variable and remove any outliers. For example, the age of the cells was set to 60 days or higher, the temperature range was set between 960 °C and 980 °C, which is the normal operation range of the cells.

To select the appropriate number of PCA components, a cross validation procedure was employed, as described in [77]. Thus, the training data was split into several different folds, which, in this case, was set to 10, a common value used in the field. Each observation was randomly assigned a fold number. Then, a PCA model was built on 9 folds while testing the model with the remaining 10th. This was done until all 10 folds have been used as testing data. From the results, the explained variance in training ( $R^2$ ) and in prediction by cross-validation ( $Q^2$ ) were calculated and investigated. The  $R^2$  metric is cumulative and increases with the number of components. On the other hand, the  $Q^2$  is always lower than the  $R^2$  and it typically reaches a maximum value at a certain number of components, after which it decreases. This inflection point indicates that the new component added does not predict well the testing data, and is

generally an indication of overfitting. As such, the new component should not be considered, with the optimal number of components occurring before the inflection point.

A second cleaning step was done using multivariate filtering. A PCA model was built on the initial, univariately cleaned data. The  $T^2$  and  $SPE$  statistical limits for normal operating conditions were calculated and used to set the threshold of multivariate filtering. The  $T^2$  limits were calculated using equation 16 from the work of Jackson [78]:

$$T_{lim,\alpha}^2 = \frac{(N-1)A}{N-A} F_{\alpha,A,N-A} \quad (16)$$

Where  $N$  is the size of the training data,  $A$  is the number of principal components and  $F$  is the Fisher-Snedecor's F-distribution, with  $\alpha = 0.05$  as the level of significance.

The  $SPE$  limit was calculated based on the work by Jackson et al. [79]. The main formula is shown in equation (17)

$$SPE_{limit} = \theta_1 \left[ \frac{c\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right] \frac{1}{h_0} \quad (17)$$

where  $c$  is the 95 % percentile assuming normal distribution, and  $h_0$  can be found as:

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_3^2} \quad (18)$$

Additionally,  $\theta_1$  and  $\theta_2$  are given by (19)

$$\theta_i = \sum_{j=a+1}^A \lambda_j^i, i = 1, 2, 3 \quad (19)$$

where  $\lambda$  are the eigenvalues of the training matrix  $\mathbf{X}$ .

A final data cleaning step was put in place to remove entire cells from the dataset, that had  $T^2$  and/or  $SPE$  higher than the limits for more than 30 % of the observations. This was done to remove notoriously problematic plots from the model and avoid special cause (abnormal) variability. The objective is to capture normal process variability for good operating conditions and to create a benchmark for detecting abnormal events. In this case the data collected were 12139 observations. The data removed were 915 observations which represents 7.5 % of the observations.

### 2.3.3 Model Exploitation (Phase II)

Once the model is built, it is used on-line by projecting new data onto it. First the new observation has to be preprocessed, using the mean-centering and scaling coefficients from the matrix the model was built upon. Next, the new data,  $\mathbf{x}_{new}$  is projected onto the model to calculate the new scores, as shown in equation (20).

$$\mathbf{t}_{new} = \mathbf{x}_{new} \mathbf{P} \quad (20)$$

The predicted values are then calculated using equation (21)

$$\hat{\mathbf{x}}_{new} = \mathbf{t}_{new} \mathbf{P}^T \quad (21)$$

The new residual matrix is obtained using equation (22)

$$\mathbf{e}_{new} = \mathbf{x}_{new} - \hat{\mathbf{x}}_{new} \quad (22)$$

This leads to the calculation of the new  $SPE$  using equation (23)

$$SPE_{new} = \mathbf{e}_{new}^T \mathbf{e}_{new} \quad (23)$$

while the new Hotelling's  $T^2$  becomes (24)

$$T_{new}^2 = \sum_{a=1}^{a=A} \left( \frac{t_{newa}}{s_a} \right)^2 \quad (24)$$

The  $T^2$  metric is an overall measure of distance of a new observation from the mean of the data used in the model building phase. Since the  $T^2$  is a squared quantity, there will only be an upper limit, calculated using equation (16). The  $SPE$  measures how consistent the correlation structure between the variables in the new observation is, with respect to the training dataset. A break in the correlation structure could happen when major process changes occur, changing the dynamics between variables. Thus, using the limit calculated in (17) it can be determined whether the observation vector  $\mathbf{x}_{new}$  follows the model's correlation structure and is well fitted by the PCA model. As with the  $T^2$ , the  $SPE$  only has an upper limit. Thus, using these metrics and their limits, it can quickly be determined if a process is in control, which is when both the  $T^2$  and  $SPE$  are below their respective limits. A problematic cell is detected when one of these limits, or both, are violated.

To diagnose a problematic cell, the contributions of each variable were calculated, to the deviation in  $T^2$  metric from the model origin (mean of training data). The contribution of the  $k^{th}$  variable to the  $T^2$  is calculated using equation (25).

$$T_{contrib_{i,k}}^2 = x_{i,k} \sqrt{\sum_{a=1}^A \left( \frac{t_{i,a}}{s_a} \right)^2 p_{ka}^2} \quad (25)$$

Thus  $x_{i,k}$  is the value of the  $k^{th}$  variable in the  $i^{th}$  observation investigated,  $t_{i,a}$  its score within component  $a$ , while  $s_a$  is the standard deviation of the score column within component  $a$ , and  $p_{ka}$  is the loading of variable  $k$  within component  $a$ .

The *SPE* contribution for variable  $k$  and observation  $i$  can be calculated using equation 26.

$$SPE_{contrib_{ik}} = e_{ik}^2 \cdot sgn(e_{ik}) \quad (26)$$

where  $e_{ik}$  is the residual value for variable  $k$  and observation  $i$  obtained previously from equation 22. Note that the sign of the contributions on  $T^2$  and *SPE* indicate the direction of the deviation of the variables with respect to their mean values computed from normal operation data. A positive value indicate that a variable is higher than its mean value in normal operation, and it is lower when it is negative.

### 2.3.4 Ranking system and performance metrics

The day-to-day operations at the smelter are run in shifts, and at the end of each one, new data is projected onto the model. This results in new values of  $T^2$  and *SPE* metrics for each cell. The  $T^2$  allowed detection of most of the faults of interest and so the focus has been on this metric. The *SPE* measures the deviation from the model structure, and in this case, it is very sensitive to small changes to variables with little variability. Although both metrics are continuously monitored, it was decided to create a ranking system based solely on the  $T^2$  metric. The *SPE* was only used for interpreting deviations together with the  $T^2$ . This would add context by ordering the cells in a descending manner, from highest  $T^2$  to the lowest one. The worst cell in the line would have rank number 1, highest rank, with its corresponding  $T^2$  being the highest. On the other hand, the best cell in the line would have a rank number of 336 (or the total number of operating cells), and the lowest  $T^2$  value. This was done because the total number of cells in operation can vary, depending on the life-cycle of the pot-line. During times of cell relining and new cell start-ups, the total cells in operation would be lower than the maximum of 336. In addition, potential cells out of operation, due to tap-outs or other issues, would also affect the total number.

Throughout the timeframe in question, existing cell deviation reports were also extracted to assess the performance of the model. These reports are built on univariate data analysis. At the beginning of each shift a set of conditions are passed to the reporting tool to identify the problematic cells. These can be hot cells, high/low liquid measurements, cells with high silicon or other trace elements. The key point here, is that the logic is looking at individual variables and the report created as a database query. The cells on the report list are not ranked by the level of need of intervention. This is left to the technical staff and cell operators to decide which cells should be attended to first. Logistics of the shift and equipment availability would also play a role in the decision.

Due to the abundance of routine operational tasks, the shift operators have limited capacity to attend to problematic cells. During the monitoring scheme assessment time frame, the minimum number of cells that appeared on the existing reports was 60. This number was thus chosen as a threshold for the ranking system, that is, cells with a rank lower than 60 were more likely to have a problem. During the trial period the cells for investigation were chosen by the plant's potroom technical team and operators. Often these were cells that were ranking high on the list, but other rankings were also chosen, to be able to assess the performance of the model.

To assess the model performance, the data was classified in four categories, namely True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). The TP represent the observations correctly predicted as positive, in this case, a bad cell in need of assistance. The TN represent the observations that were correctly predicted as negative, that is, cells that performed well and did not need any assistance. The FP represent the observations that were falsely predicted as positive, that is, cells that were predicted to need assistance but, in reality, they were running within the operating window. This is also known as a Type I error. Lastly, FN, represent the observations that were falsely predicted as negative. These cells were predicted as not needing assistance but, they were running outside the operating window and would have benefited from manual interventions. This is also known as a Type II error. These classifications are the building blocs of a 2-by-2 confusion matrix, where each row of the matrix represents the actual condition while the columns show the predicted condition. Correct predictions are given by the diagonal elements TP and TN. The off-diagonal elements FP and FN represent the errors.

From this matrix, several performance metrics can be calculated, to assess the model performance [80]. The Precision is a classic performance metric which is defined as the proportion of the positive alarms detected. In other words, this is the positive predictive value. The Accuracy represents the proportion of correct predictions within the total number of observations. The Miss rate is the type II error, or the false negative rate. In this case the prediction indicates the absence of a problem which in fact is present, thus making the prediction incorrect. The Sensitivity represents the probability of detection, given by the true positive rate. The Specificity is the true negative rate, defined as the ratio between the true negative observations and all the negative ones. The last metric used is the Fall-out, which is the type I error, or false positive rate, represents the false alarm rate. The formulas for each metric are presented in equations (27) to (31)

$$Precision = \frac{TP}{TP + FP} \quad (27)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (28)$$

$$Miss = \frac{FN}{TP + FN} \quad (29)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (30)$$

$$Specificity = \frac{TN}{TN + FP} \quad (31)$$

$$Fall - out = \frac{FP}{TN + FP} \quad (32)$$

## 2.4 Results

### 2.4.1 Performance assessment of the monitoring scheme

During the data cleaning step, the investigation of individual variable trajectories and identification of normal operating window resulted in the number of observations to drop from 12139 to 11847. A PCA model was built on the latter set of observations and a second data cleaning step was placed onto the resulting  $T^2$  and  $SPE$ . The adjusted limits were placed higher than the calculated statistical limits for detecting abnormal events, to not remove too much of the process' natural variability. Thus, the statistical limit for  $T^2$  was found to be 14.08, but the adjusted limit was set to 25. Likewise, the  $SPE$  statistical limit was found to be 53.69, but the adjusted limit was set to 62.

After the multivariate cleaning set by the adjusted limits, there were 11272 observations left in the training dataset. In the third cleaning step, associated with abnormal process variability, 623 additional observations were removed. Thus the total number of observations for building the model were 10649.

The model achieved a total variance explained of 56.7 % with 7 components, also called latent variables (LV), chosen using the cross-validation procedure explained above. The  $Q^2$  decreases from latent variable (LV) 7 to 8, as shown in Figure 22. The  $Q^2$  increases again for LV 9, but it is still lower than LV 7. Thus, to avoid over-fitting, a model with 7 components was chosen.

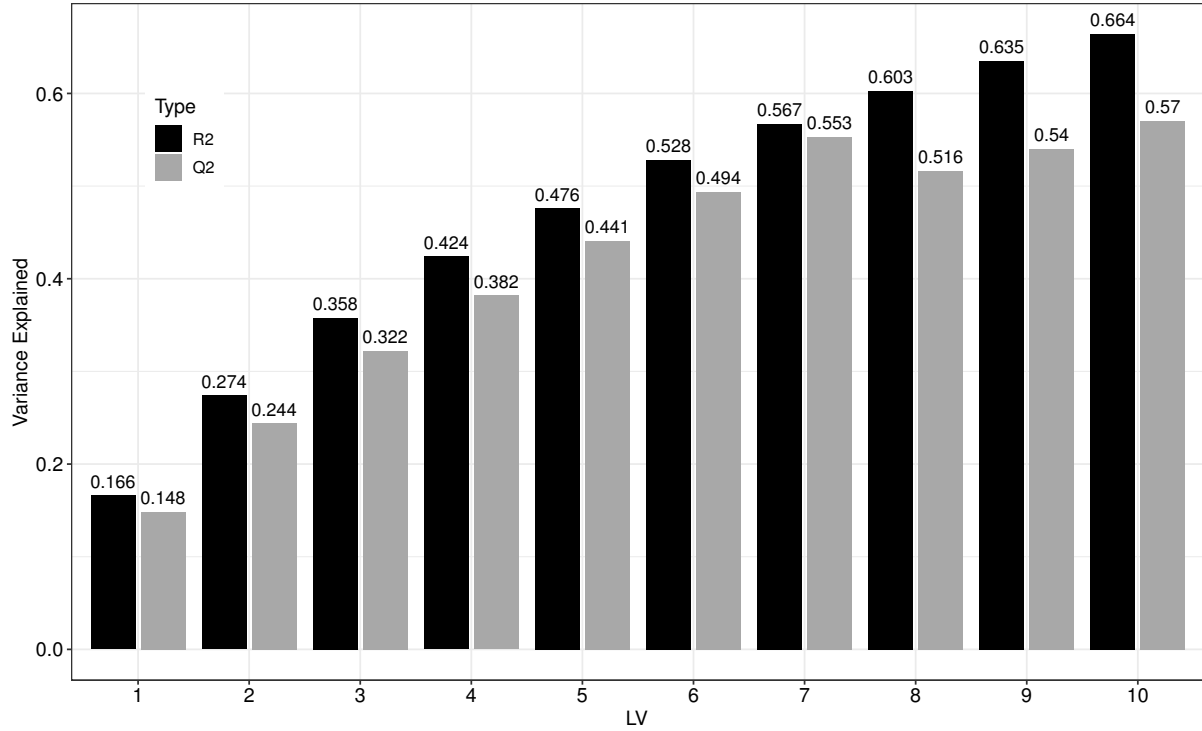


Figure 22: PCA model percentage variance explained per Latent Variable (LV)

As mentioned above, the  $T^2$  limit can be used to determine abnormal events. While creating

the  $T^2$  ranking system, a secondary threshold was placed limiting the worst performing cells to 60 in total. Figure 23 shows the daily number of cells that are above the  $T^2$  limits for each day, during real-time monitoring of the potline, that is using new data, projected onto the model (i.e., model exploitation phase II). It can be seen that most days the limit is surpassed by more cells that the operators could manually check. In mid July, a power interruption brings the number of cells above the limit to 208 (i.e. spike in Figure 23) before dropping to 103 the next day and back down to 63 the second day.

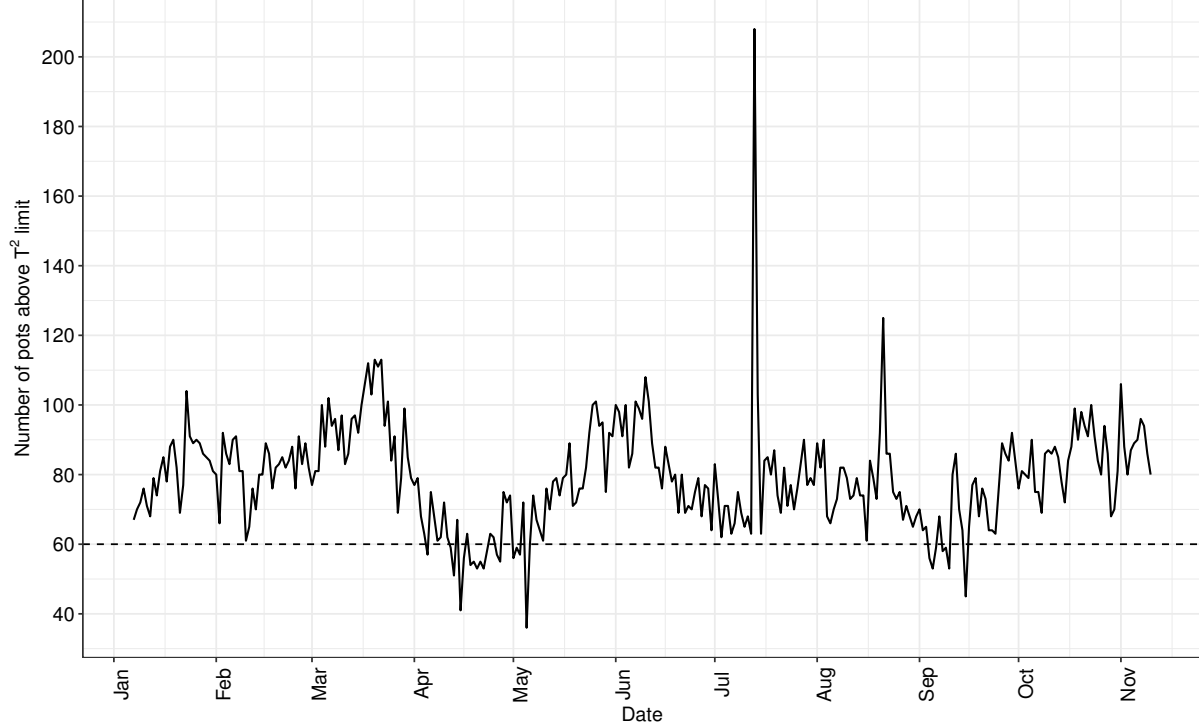


Figure 23: Daily number of cells above the  $T^2$  statistical limit for detecting abnormal events

The data on operators' manual checks was collected from March 2020 to November 2020 for a total of 331 observations, in which operators went out on the shop floor, inspected the cells in question and recorded their findings. These data were then used to assess the performance of the model. As discussed above, the operators have limited capacity regarding the number of problem-cells they can attend to daily. This limit is influenced by the equipment availability, shift manning, priorities of the day and overall potroom stability. On each shift there are dedicated operators for problematic cells, but their capacity depends on the amount of time it takes to diagnose and fix a given problem cell. Therefore, proper identification of the order in which the cells need attention can be crucial.

From the limited sample size, only a very small fraction could be allocated to this test. Thus, only 1 to 2 of the checked cells per day were part of the dataset during this test. This was further influenced by plant priorities, thus arriving at a number of 331 observations over an 8-month period.

The observations recorded by operators were assessed and manually checked against individual cell process variables for each given observation. This was done to ensure that the correct classification was assigned to each cell. From the 331 cells, manually investigated, 173 were identified as True Positives, or positive predictions that were correctly identified. In 90 instances the negative predictions were correctly identified, or True Negatives. In 11 cases there was an erroneous positive prediction, thus a False Positive. Lastly 57 instances were falsely predicted as negative, the cell having a problem was not identified by the system, i.e. a False Negative. Using the formulas (27) to (32) the metrics were calculated to assess the model performance.

The model was found to have a good Precision of 94 % (i.e. positive cases that were correctly identified as problem cells). 79.5 % of all cases are correctly identified as either a problem cell or a normal operating cell, representing the Accuracy. The type II error or Miss rate was present in 24.8 % of cases. The Sensitivity was 75.2 % probability of detecting a problem. The Specificity was 0.891, or, in 89.1 % of the negative cases the prediction was correct. Lastly, the false alarm rate was given by the Fall-out which was observed in 10.8 % of cases.

The distribution of Accuracy of the MSPC scheme is shown in Figure 24 as a function of the rank bins (bar chart). The number of cells in each bin is displayed above the bars. For example, for the bin containing the highest-ranking cells (interval from 0 to 10), the accuracy of the model is about 90 % based on 57 cells investigated in that bin. The number of cells in each bin vary due to prioritisation of problematic cells. The figure clearly shows a consistently very high accuracy for the 60-highest ranking cells, that is the threshold above which the cells likely affected by an abnormal event and set as priorities for operators to investigate. Below rank 60, accuracy is more variable and the number of cells investigated smaller in number. The overall performance for all cells investigated is shown in Table B.1 and the distribution of all classification metrics per bin is shown in supplementary material Table B.2.

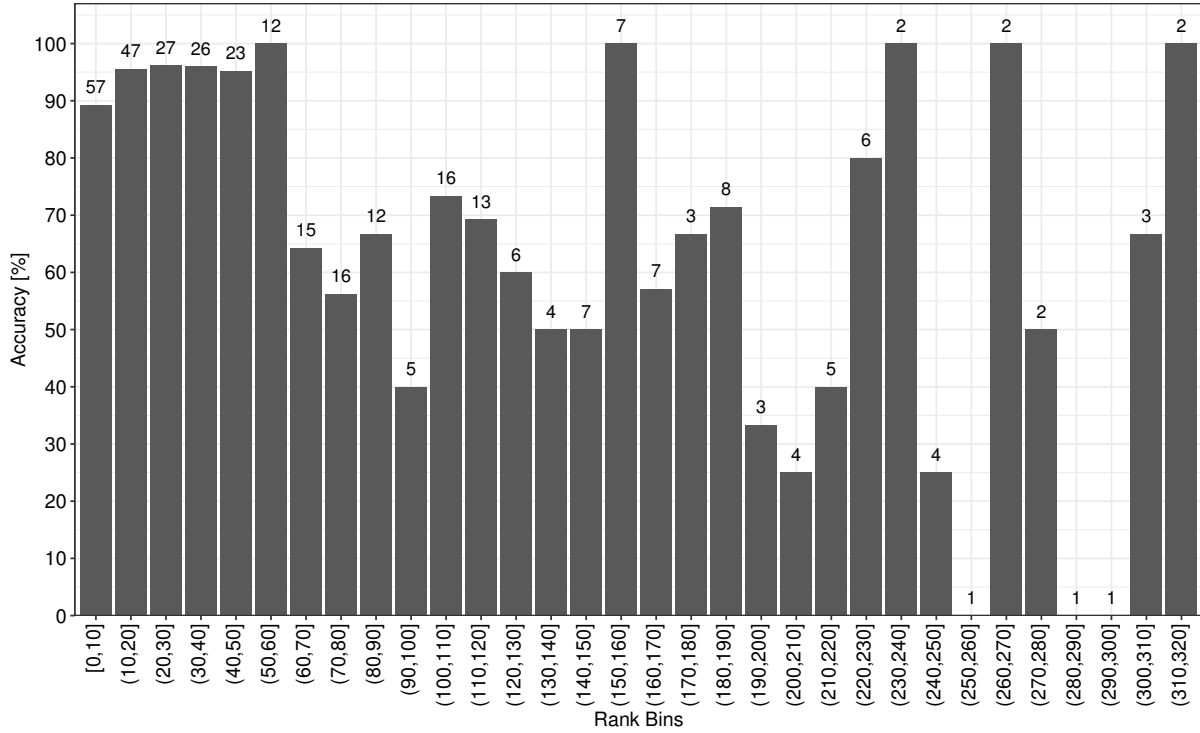


Figure 24: Accuracy distribution of the MSPC scheme according to the rank bins. The number of cells in each bin is displayed above each bar.

By comparison to the pre-existing deviation cells reports, it was possible to determine whether the PCA-based ranking system was able to detect the anomalies with an advanced notice. These reports were compared to the observations gathered to build the confusion matrix. Out of the 173 True Positive observations, 60 of them showed up on the  $T^2$  ranking system before appearing on the operators' standard system. From these, 42 had 1-day advance notice, accounting for 22.4 % of all true positive instances. Another 13 observations had 2 days notice, or 7.5 % of all true positive instances.

In potline operations, getting an advanced notice of process upsets is very important, since the optimal window for the electrolysis process is quite narrow. This can allow the operators to investigate and mitigate the problem before the cell goes out of control. For example, when an old cell experiences side ledge loss due to high superheat, it could make the cell vulnerable to side tap-out. This could happen if the side ledge protection is melting in an area with an already thinned out refractory wall. The steel shell could easily be exposed resulting in bath making its way out of the cell. If caught in time, localized cooling can be placed on the outside of the cell, in spots that glow red. In other cells, if anode problems could be identified sooner, they can be planned into the anode setting schedule of the next shift that is operating in that area. This can prevent cells from becoming unstable and developing additional issues.

### 2.4.2 Examples of anomaly detections

As the reduction cells age, they become more fragile and at risk of tapping out due to several factors. Older cells are typically run with lower excess aluminum fluoride percentage to reduce bath resistivity and the liquidus point. Material transformation over time, toe ridge and other phenomena increase the cathode voltage drop. With time, the carbon cathode becomes eroded which increases the potential of cell tap-outs, in which, metal starts flowing from the bottom of the cell.

Shortly before a tap-out, the metal sample analysis can show drastic increases in iron and copper, if the current collector bar is fitted with a copper insert. Iron alone is less reliable as an indicator since it can come from the anode stubs. If spotted in time, operators can try to locate these cathode holes by measuring the cathode current distribution and comparing them to the ones taken early in the cell's life. Another possible action is to measure the bar temperatures. If located in time, these holes can be patched up to keep the cell running. If these are not observed in time, the cell would tap-out.

Within older cells, the refractory sideling, often silicon carbide, can be eroded. This is rarely something that happens uniformly within the cell, leading to thinner spots. When the cell experiences a sudden surge in superheat, the side ledge melts exposing the refractories, thus thinning them. This is seen in the increase of the amount of silicon within the metal samples. Using mass balance, it is possible to estimate the amount of silicon coming from the side lining, but not pinpoint the location of the source. When the side lining is eroded and the cell experiences an excursion, meaning the side ledge melts, the molten bath can penetrate through the steel shell and spill. This would result in a side tap-out. The results can be quite devastating as the bath pours out along with the metal. If the location of the hole is high enough the cell could still run if the bath level remains below the exposure level.

Two examples of anomalies of cells above 2000 days old are presented in Figure 25, namely Cell 1 and Cell 2. In both cases the  $T^2$  metric trajectory rises well above the normal operation limit, with an upward slope over time. The rank of the cell puts the  $T^2$  value into perspective, in comparison to the rest of the potline. In this case Cell 1 experienced a sharp increase in rank while the Cell 2 had a high rank for a sustained period, spanning about 22 days. The  $SPE$  values were also interrogated, but they are prone to sharp variations with changes in variables that usually have little variation.

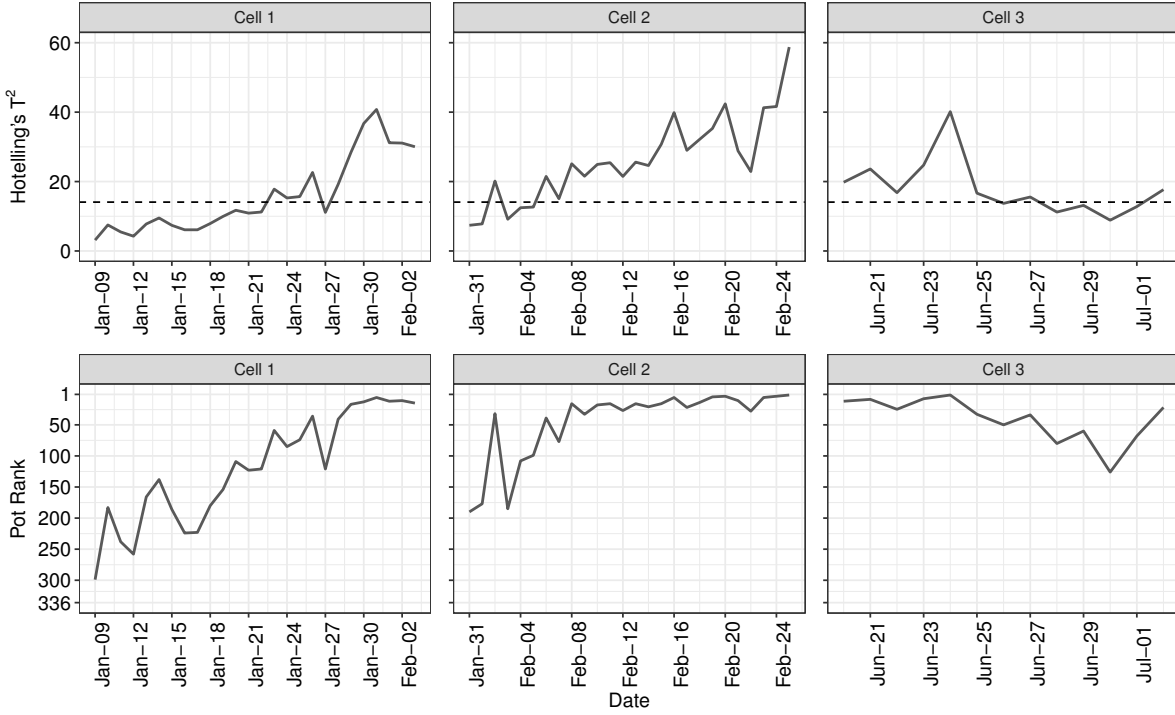


Figure 25: Time series of Hotelling's  $T^2$  and cell ranking for Cell 1, Cell 2 and Cell 3

The  $T^2$  and  $SPE$  contribution plots for Cell 1 are shown in Figures 26 and 27. The  $T^2$  contribution plot provides a measure of the distance of an observation with respect to the model center, or average, normal, operating conditions of good cells. The center is 0, in this case, since the data has been centered and scaled. The  $SPE$  contribution plots provide an indication of how well the correlation structure of the data used to build the PCA model, is respected. An  $SPE$  value of 0 means that the observation falls onto the model plane and, as such, follows the model's correlation structure exactly. These two plots can be used, in tandem, to determine anomalies within a cell. For all the contribution plots in this chapter, only variables with absolute contributions' values greater than 1 are shown. This was done in order to focus on the most important variables and reduce clutter on the chart. In addition, the variable names were shortened. The full names for the variables shown in Figures 26 to 31 are presented in Table A.1, listed in alphabetical order.

The contributions are shown for the last 6 days in the life of Cell 1 and Cell 2. The  $T^2$  and the  $SPE$  plots for Cell 1 are shown in Figures 26 and 27. The color codes represented the different variable groups explained in Table 2 to better connect to the Data Collection section. The same color code is used throughout all the examples shown hereafter.

For the first one, on the  $T^2$  contribution plot in Figure 26, it can be observed that the trace elements in the metal sampling decrease, indicating anode problems [72]. The peak on  $T^2$ , reached on January 30, (see Figure 25, Cell 1), was driven by significantly lower values of trace elements concentrations (i.e., Vanadium, Titanium and Gallium) compared with normal operations. It has been previously shown that anode spikes tend to accumulate trace elements and thus decrease

the amount passing through to the metal pad [62]. At the same time the cell temperature and superheat also increase. In the last three days, there is also an increase in voltage and resistance along with waves. The  $T^2$  contributions are consistent over the time period when the  $T^2$  is at its peak value, which increases trust in the interpretation/diagnosis made from these plots.

On the  $SPE$  contribution plot in Figure 27 there is a negative spike in the bath level target (BathTgt) and base alumina feed (BaseFeed) which could indicate a cell tuning opportunity. The contribution plot shows that they are lower than what they should be compared to normal operation. The sudden negative spike of the aluminum fluoride feed (AlF3add\_real) on Jan 31<sup>st</sup> could indicate a physical issue with the feeder. During the last 2 days there is an increase in resistance, through modifiers, along with waves and temperature.

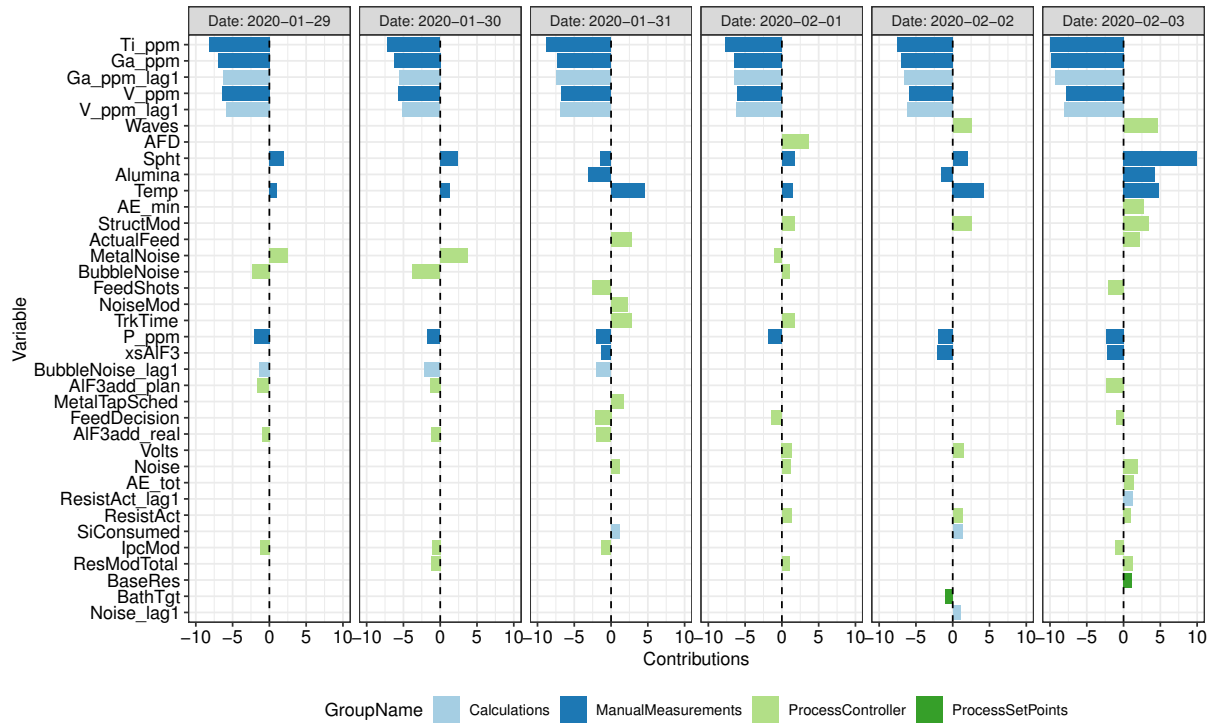


Figure 26:  $T^2$  contributions for Cell 1. Only the most important contributors are shown (absolute value of contribution  $> 1$ ).

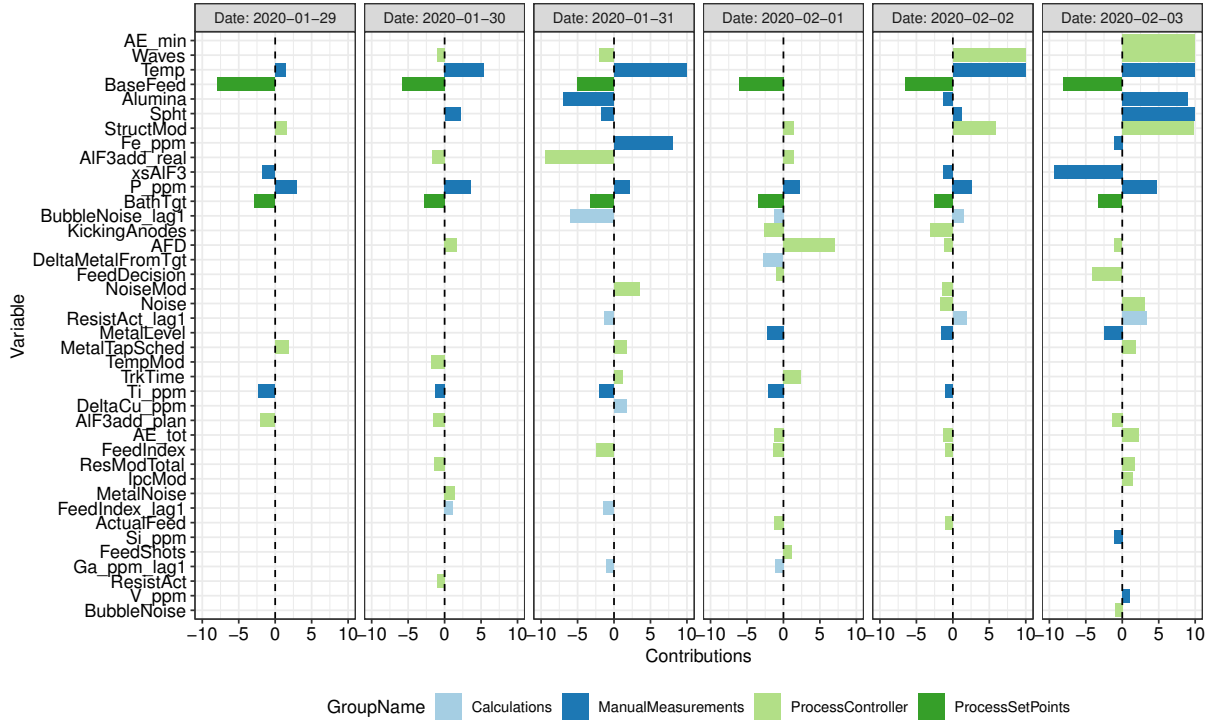


Figure 27: *SPE* contributions for Cell 1. Only the most important contributors are shown (absolute value of contribution > 1).

The  $T^2$  for Cell 2, in Figure 28, shows a similar deteriorating trend for the last six days of its life. The trace elements decrease, also indicating anode problems in the last days prior to its tap-out. The temperature and the superheat are also increasing. Additionally, in this cell the silicon consumed was increasing, suggesting that the cell was losing its side ledge, eroding the refractory bricks. Meanwhile, the number of alumina feed shots were also decreasing, which is often related to longer underfeeds, as indicated by the increase in the alumina feed variable (ActualFeed). The higher temperature increases the back reaction resulting in electrolyzing the alumina again and thus lower alumina feed rate.

On the *SPE* contribution plot for Cell 2 in Figure 29 cell tuning parameters use different set-points than the average of the model building data. In this case the bath level target and base feed rate were reduced while the base resistance target was increased. A lower base feed rate translates to a shorter time between feed shots and thus a higher amount of alumina to be fed to the cell. The feed rate modifier (Feed Index) was positive for Jan 21<sup>st</sup>, 24<sup>st</sup> and 25<sup>st</sup>. This was in opposite direction of the base feed rate, meaning the controller was trying to increase the time between feed shots and thus reduce overall alumina feed. This goes hand in hand with the higher temperature seen on the  $T^2$  contribution chart in Figure 28. The *SPE* contribution plot suggests that there were cell tuning opportunities for this cell at the time.

The model was able to detect anomalies within these cells prior to their respective tap-out. Older cells are more at risk so they could benefit from such a monitoring scheme to preserve their ongoing operation or to plan for their retirement to avoid a tap-out. A fast deteriorating

cell health in a scheme such as this, would raise alarms sooner than with traditional tools, and allow operators to take preventive actions.

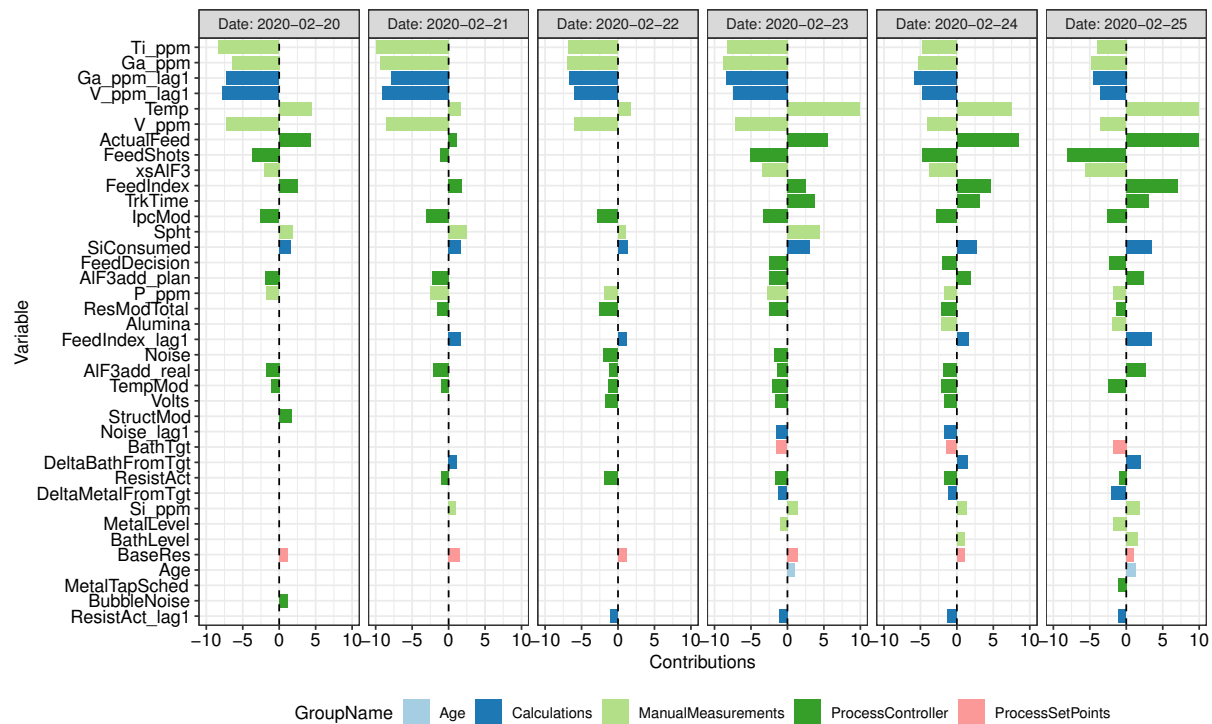


Figure 28:  $T^2$  contributions for Cell 2. Only the most important contributors are shown (absolute value of contribution  $> 1$ ).

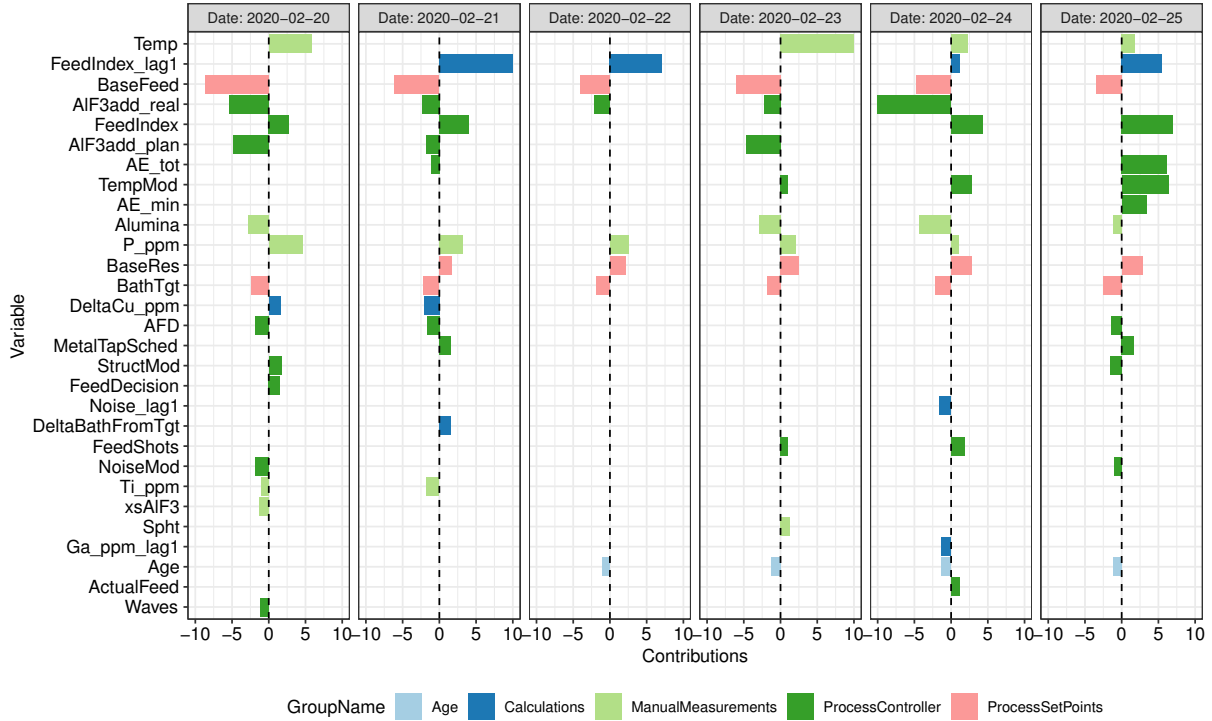


Figure 29: *SPE* contributions for Cell 2. Only the most important contributors are shown (absolute value of contribution  $> 1$ ).

Another example of a problem detected in Cell 3 is presented below. This time, Cell 3 returned to normal operation and did not tap out. A period of six consecutive days is shown, from June 21<sup>st</sup> to 26<sup>th</sup>, where the cell is experiencing process upsets. The ranking of the cell is illustrated in Figure 25.

The  $T^2$  contributions plots, are shown in Figure 30. The trace elements are shown to have negative contributions, indicating that there could be anode problems in the cell. The silicon was also spiking indicating potential loss of side ledge within the cell. Indeed, on June 23<sup>rd</sup> the operators found two problematic anodes and replaced them with new ones. On June 24<sup>th</sup> the cell experienced a greater process upset as indicated by the loss of feeding, increased superheat, noise, resistance, and liquid levels. At this time this cell was the second worst in the potline according to the  $T^2$  ranking system. The operators found another problematic anode and replaced it with a new one. Red spots on the shell were also found and air cooling placed to prevent the cell from a side tap-out. Additionally, bath transfer was performed to adjust the liquid levels. After all these actions were taken the cell improved and returned to stable operation which is shown by both the ranking of the cell decreasing but also the contribution plot of the variables for June 25<sup>th</sup> and June 26<sup>th</sup>.

On the *SPE* chart, in Figure 31, there are big spikes around the difference between daily measurements of copper in metal samples (DeltaC\_ppm). These could be alarming if the copper concentration in metal variable (Cu\_ppm) would also be spiking on the  $T^2$  chart. However, this is not the case. Since the copper concentration in metal is quite low and stable, around 2 ppm,

small changes could make a big impact on the *SPE* contribution. The *SPE* chart does show that there could be a cell tuning opportunity in this case. During the first two days the alumina feed rate setpoint (BaseFeed) was reduced, increasing the alumina feed. If the base feed is set correctly, the feed rate modifiers are not needed. Adjustements made to the base feed from June 21 to June 25 resulted in the decrease of the feed modifier (FeedIndex). The feed control is better adjusted for the cell condition, therefore less modifiers are needed. The variable kicking anodes represents the anodes with fast variations in voltage drop. These are seen in the spikes in the KickingAnodes variables on June 23<sup>rd</sup> and 24<sup>th</sup>, coinciding with the anode problems found by the operators.

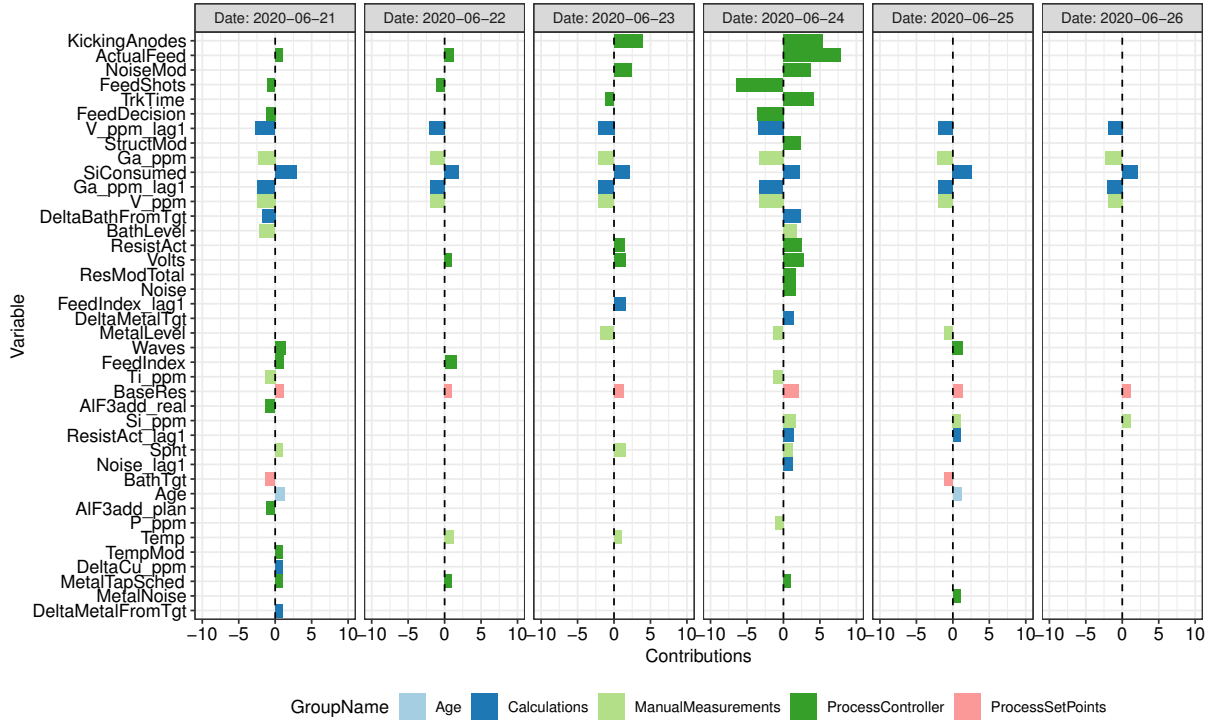


Figure 30:  $T^2$  contributions for Cell 3. Only the most important contributors are shown (absolute value of contribution  $> 1$ ).

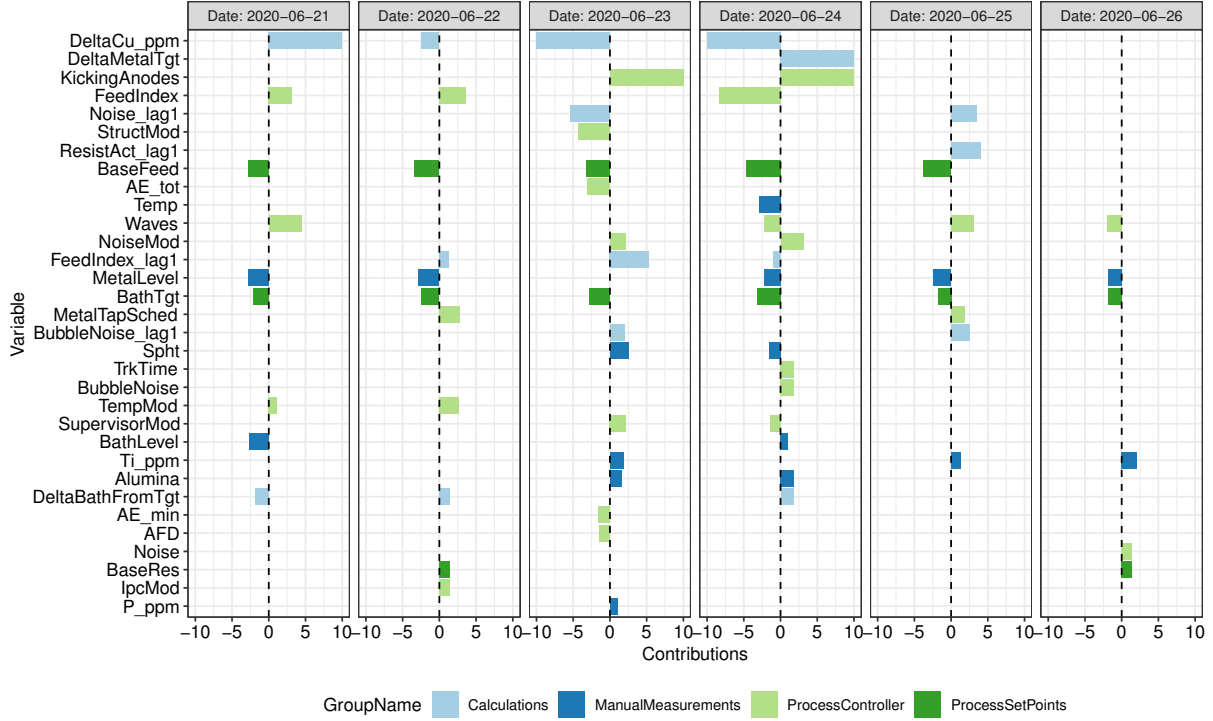


Figure 31: *SPE* contributions for Cell 3. Only the most important contributors are shown (absolute value of contribution  $> 1$ ).

## 2.5 Conclusion

In this paper the implementation of a multivariate monitoring scheme of cell operation at an aluminum smelter was described. A PCA model was built on the production data and a cell health ranking system was developed based on the  $T^2$  metric. The model performance was assessed by comparing existing reports with observations from the shop floor. Examples of anomalies detections were also presented. It was observed that process upsets can be correctly identified and can aid the operators in taking corrective action sooner than with traditional tools. Additionally, this monitoring scheme can standardize the order of the problematic cells. This would help operators prioritize the critical cells but also the logistics around the needed equipment availability. The contribution plots of the  $T^2$  and *SPE* can help identify the issues inside the problematic cells. A confusion matrix was built out of the cells that were manually investigated by operators. Due to logistics around cell operation the sample size was relatively small but it showed that the model returned a precision of 94.0 % with a 79.5 % accuracy, 10.9 % type I error, and 24.5 % type II error. Compared to existing reporting tools, the model provided a 1 day advance notice in 22.4 % of cases where a true positive was identified. Another 7.5 % of cases showed an advanced notice of 2 days. With an advanced notice of an issue in the cells, proactive action can be taken to ensure process stability.

## 2.6 Future Work

The current model works, but there is room for improvement in the model performance. One option could be to build one model per section of the potroom, instead of a single model encompassing all cells. The location of the cells can have an impact on their behaviour. For example, there is one section of the potroom that is notoriously low in bath levels. Thus, special adjustments have to be made in the process controller to account for this. Another section could have other particularities. An alternative option could be to build different models per age group. For example, old cells run with a different fluoride setpoint to open the anode-cathode distance and reduce noise. The variability in the dataset induced by these factors could be treated as anomalies when applied to younger cells. The sample size that this study was built on was limited. A good step forward would be to collect more data and corroborate these findings. One way to do it would be to only rely on this ranking system for a section of the potroom and compare stability key process indicators to a control section, for an extended period of time.

## Chapter 3 Estimation of Net Carbon Consumption in Aluminum Electrolysis Using Multivariate Analysis

### Résumé

Dans l'électrolyse de l'aluminium, la quantité de carbone consommée par unité de métal produite est toujours supérieure à la quantité théorique. La variabilité des matières premières, les propriétés des anodes; les conditions d'opération et la performance des cuves d'électrolyse peuvent influencer cet écart. Lier ces facteurs à chaque anode individuelle pourrait aider à combler cet écart, améliorant ainsi l'empreinte environnementale et réduisant les coûts de production. Dans l'environnement industriel actuel, un grand volume de données, à la fois structurées et non structurées, est collecté, autrement appelé mégadonnées. Celles-ci pourraient permettre de prédire la consommation de chaque anode au lieu de l'estimer au niveau de l'usine, comme cela a été fait jusqu'à présent. Cela peut être réalisé en utilisant un système de traçabilité pour relier l'opération des cuves aux propriétés des anodes individuelles, cuites et crues. La performance de la formule de consommation nette de carbone de R&D Carbon a été comparée à une approche statistique multivariée. Celle-ci prend en compte les variables de procédé recueillies à partir des bases de données liées susmentionnées. Compte tenu des données de procédé disponibles, le pouvoir de prédiction de la consommation nette de carbone pour les anodes individuelles est considérablement amélioré par rapport à l'utilisation du modèle existant.

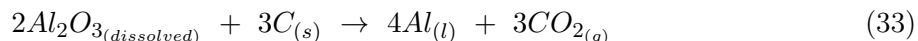
### Abstract

In the aluminum electrolysis the amount of carbon consumed per unit of metal produced is always higher than the theoretical one. Raw material variability, anode properties and reduction cell operation and performance can influence this discrepancy. Linking these factors to each individual anode could help close this gap thus improving the environmental footprint and reducing production cost. In today's industrial environment, large volume of data, both structured and unstructured, are collected, otherwise known as big data. These capabilities could enable predicting the consumption for each anode as opposed to estimating it at the plant level, as performed so far. This can be achieved using a tracking system to link the pot operation to individual baked and green anode properties. The performance of R&D Carbon Net Carbon Consumption formula was compared to a multivariate statistical approach that considers process data variables gathered from the aforementioned, linked, databases. Given the process data at hand, the prediction power of the net carbon consumption for individual anodes is significantly improved as opposed to using the existing model.

### 3.1 Introduction

The aluminum industry is an important part of the Canadian economy, the country being the fourth largest producer in the world [81], with one of the lowest carbon footprints thanks to the use of hydropower. It is produced via the Hall-Héroult process where alumina is dissolved in a molten salt bath (cryolite) and a high, DC, electric current is passed between the anode and the cathode in the cell. In existing technologies the current can be from 100 kA to over 600 kA. Both the anodic and cathodic electrodes are made of carbon, the anodes being consumable. The preferred technology employs pre-baked anodes, although continuously baking anodes still exist (Soderberg), but are now rare.

The aluminum production process is governed by the overall equation (1):



According to the stoichiometry of the electrochemical reaction shown in equation (33), 1.89 kg of alumina and 0.334 kg of carbon are needed to produce 1 kg of aluminum. Therefore, the theoretical net carbon consumption (NCC) per tonne of aluminum produced is 334 kg [9]. However, in practice, it is very difficult, if not impossible, to achieve this number. This can happen due to a number of factors that include raw materials properties, formulation parameters, anode properties (green and baked) and reduction cell performance. The carbon losses are typically caused by air burn, anodic incidents and dust formation within the cell. A number of mechanisms contribute to the excess carbon consumption, such as the Boudouard reaction, between anode carbon and primary CO<sub>2</sub> gas [9]. If the anodes are under-baked and/or the pitch content is not optimal there can be dusting through selective oxidation of binder coke [32]. Yet other carbon losses can take place in the form of unreacted dust, reoxidation of metal, pyrolysis and vaporization. Other contributors to consider are the impurities in the petroleum coke and coal tar pitch, used to manufacture the carbon anodes, that act as strong oxidizing catalysts, such as Na, Ca, V and Ni [9]. The presence of sodium is mainly due to the spent anodes (butts) used in the anode recipe. Thorough cleaning of the butts is required in order to minimize the amount of frozen bath attached to them, before crushing [16]. Sulfur plays the role of inhibitor towards oxidation but the amount in the recipe is limited by governmental regulations involving emissions of sulfur dioxide.

Since the early 1990's the principal way of predicting the NCC has been by using a formula developed by R&D Carbon [17]. The model, equation (34), is based on data from many different smelters, collected over a period of over 20 years up to the time of publishing, and averaged on a potline basis over long periods of time. The technology involved contained operating currents of 180 kA and lower.

$$NCC = C + \frac{334}{CE} + 1.2(BT - 960) - 1.7CRR + 9.3AP + 8TC - 1.5ARR \quad (34)$$

In the above equation, NCC is the net carbon consumption, C is the proposed cell factor, which takes into account different technologies, CE is the cell's current efficiency (ranging between 0

and 1), BT is the bath temperature of the pot, CRR is the CO<sub>2</sub> reactivity residue, AP is the air permeability, TC is the thermal conductivity and ARR is the air reactivity residue. The cell current efficiency is the ratio between the metal produced and the theoretical amount that could be produced at the same amperage. Due to the effects of the back reaction, where molten aluminum is re-oxidized, the CE values are never 100 %. With today's technology the CE can be up to 95 %. Pot operation behaviour reflects onto the amount of metal produced and therefore CE is a good measure of the pots' health.

The CRR and ARR are standard quality control tests, performed on the carbon samples to find the residue that remains after the block has been exposed to CO<sub>2</sub> and air respectively. An increase in AP, may result in higher dusting effect, since there is more surface area available for attack by air and CO<sub>2</sub>. A higher residue is an indicator of lower reactivity, and thus, these parameters (CRR and ARR) have negative effects on the NCC.

In the present literature there are a number of studies that have been performed in order to better understand the relationship between carbon consumption and the properties of baked anodes, raw materials and cell operation. However, in estimating the net carbon consumption, the R&D Carbon formula is still widely used. For example, McClung et al. [30], used the R&D Carbon equation to find the best combination of core sample properties for maximizing baked anode performance. In another study performed at the Alcoa Alumar in Brazil, Piffer et al. [31], used the formula to optimize the final baking temperature and minimize the variations in the pits in the anode manufacturing process.

Practical NCC values today range between 400 and 450 kg of carbon per tonne of aluminum produced [32]. In order to reduce the NCC towards its theoretical value via process control adjustments, the industry needs to measure it frequently and on an individual anode basis. The newest phase in the Industrial Revolution (Industry 4.0) focuses on improving traditional processes via automation, smart technology and machine learning. In this context, the presence of anode tracking systems in smelters are becoming more frequent [33, 34]. Their development is driven by the need to extract useful information from the production process as a whole to better optimize it [35]. The declining quality of the anode raw materials [19] plays a big part in the need to optimize anode performance in the electrolysis cell.

The present work explores new ways to estimate the net carbon consumption of individual pre-baked anodes using an anode tracking system (ATS) at the Alcoa Deschambault Quebec (ADQ) smelter in Canada, which links the different databases together. The widely used R&D Carbon formula provides a starting point and it is shown that a much larger number of measured variables is necessary for predicting the NCC at the individual anode level to account for most of the factors affecting it. The current paper is arranged as follows. First, the data collection is discussed followed by a brief introduction to the latent variable methods used. The results section is split into two parts, the first contains analysis of the dataset using the same variables as in the original R&D Carbon formula, using multi-linear regression and partial least squares (PLS). The second part contains a PLS analysis on the full dataset using around 200 variables collected from raw materials to pot operating parameters. A comparison between the performances of these models are discussed followed by a summary and future work.

### 3.2 Anode manufacturing and smelter data collection

The present work is based on the data collected from the Alcoa smelter at Deschambault. The plant has an anode production facility and 264 reduction cells with AP-30 technology, where each cell has 40 anodes. At the time of the data collection the operating amperage was around 380 kA. The plant generates a lot of data from the individual processes such as raw materials and green anode manufacturing, baked anode production and reduction cell performance. Anode manufacturing is a process that consists of a series of steps performed in a defined order. Only a limited quantity of anodes can be produced at once, making it a batch process. By contrast, the aluminum reduction cells are operating with a continuous process, as the aluminum oxide is reduced to molten aluminum in a constant flow, within each pot. Earlier plantwide studies used aggregated data since no anode tracking system was available at the time [48].

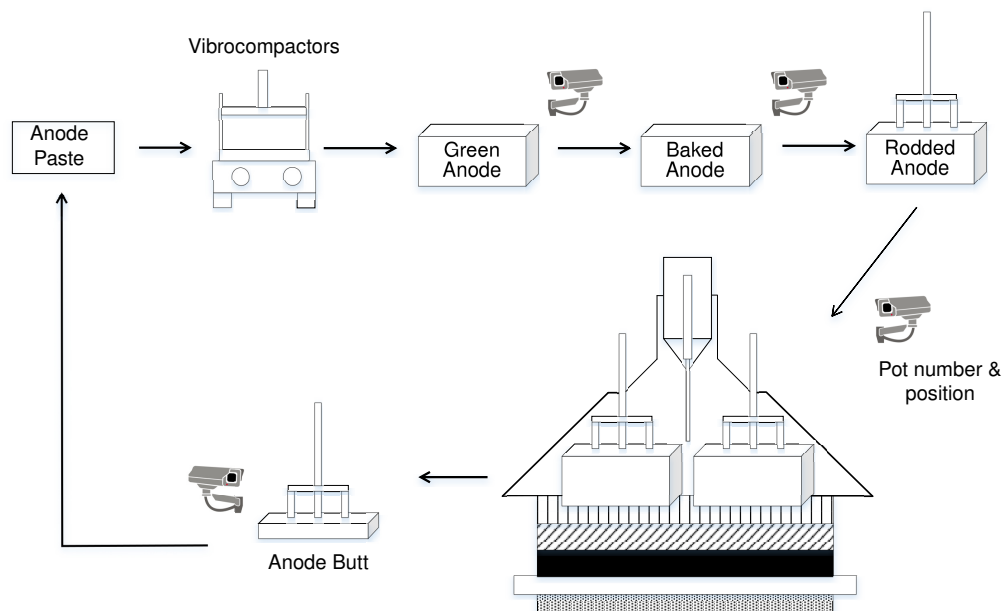


Figure 32: Schematic representation of the anode tracking system

A new method for tracking the anodes has been an enabler to connect and synchronize these three major databases on an individual anode level. The Anode Tracking equipment includes a set of cameras that are placed at strategic locations throughout the ADQ plant, as shown in Figure 32. The batch data collected from the anode production can thus be connected to the continuous data from the aluminum reduction cells operation, using the stamped, individual and sequential, anode number as a reference. The end-user can thus track individual anodes down to their position within a given cell, enabling investigation of the interaction between the anode properties and reduction cell performance.

The dataset was extracted from ADQ's production database over a period of 2 years from 2016 to 2018. Measurements were collected on an individual anode basis starting with the green anode properties. These included resistivity along with some formulation parameter and weight.

After baking, the anodes were weighted again and their measurements synchronized to the existing data block. Since only a small percentage of anodes are sampled, for quality control, the core samples properties for each week were averaged and attached to all the anodes produced in that given week. Therefore the assumption was made that all anodes produced within a given week had the same characteristics. The linking element was within the coding of the anode number.

By knowing where the anodes are placed in the potroom a link can be established with the reduction cell performance parameters over the lifespan, typically around 25 days, of the anode in that particular pot. Once the anodes are taken out of the cell, they are cleaned by removing the adhering bath and analysed through an image processing system. The information can then be related back to the cell performance and anode properties. There are sampling time differences between the pot variables such as voltage/resistance, measured every fraction of a second, and bath chemistry, liquid level measurements or metal purity which are done once a day or every other day. They were thus aggregated (i.e averaged) to daily values. The, time-series, of the reduction cell parameters (or trajectories), were represented by their average values and standard deviation over the anode set cycle after some outlier data were removed. The standard deviations of the metrics were also added to have an indication of their variance. To provide additional information, some variables were aggregated by summing them up, such as total alumina fed to the pot or the total amount of anode effect minutes. Others, were aggregated with a count function such as the number of anodic incidents experienced by the cell during that particular time frame.

The main response metric chosen was the NCC, which was calculated as the difference between the baked anode weight and its corresponding butt weight. The result was divided by the amount of metal produced from the cell in which it was placed, during the anodes lifespan. However, another assumption was made at this level. At any given time and for all reduction cells the full population of anodes is not known. Only individual anodes were captured by the ATS, spread randomly throughout the potline. Therefore, in order to calculate the NCC, it was assumed that the anode tracked was representative of the whole population within its cell and the metal produced was a result of all anodes behaving the same way. Without this assumption, it is impossible to know exactly how much metal each anode produces. Another assumption was made that the consumption was evenly distributed over the anode's lifespan inside the cell.

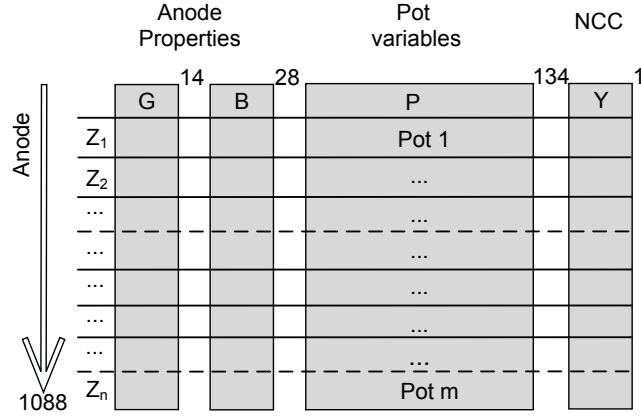


Figure 33: Full Scale PLS data structure

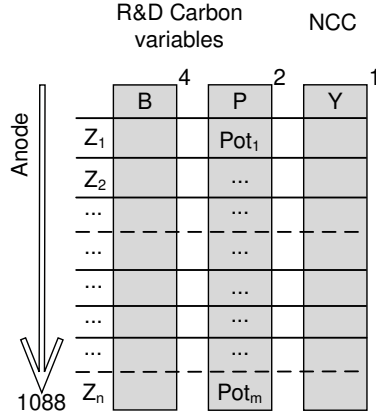


Figure 34: Small Scale PLS data structure based on the R&D Carbon formula variables

The final structure of the full dataset is presented in Figure 33, showing the three different data blocks used and their respective dimensions. These are, green anode properties (**G**), containing 14 variables, baked anode properties (**B**), with 28 variables, pot operation variables (**P**), with 134 variables, and lastly, the response block (**Y**), net carbon consumption, containing only 1 column. The number of observations is 1088 throughout all blocks, each representing an individual anode. For each anode number ( $Z_1$  to  $Z_{n=1088}$ ) its corresponding process variables were aligned. The number of variables is therefore different in each block. These blocks were concatenated into a single data matrix **X** of size 1088x176 to be used in the regression modelling work.

Out of this database a smaller subset was extracted containing only the variables present in the R&D Carbon formula, while keeping the same number of observations, shown in Figure 34. The **B** block contains the core sample variables namely, ARR, CRR, AP and TC while the reduction cell operation is captured by the **P** block, containing the bath temperature and current efficiency. In both cases the same NCC values were used. The database was split, randomly, into 75 % training and 25 % validation data, prior to training the models. Both the six variable dataset and the large one contained the same training and validation observations.

### 3.3 Modelling of the dataset

The small dataset, in Figure 34 was investigated using three different approaches in a comparative analysis. The first approach consists of using the R&D Carbon formula as published, without re-estimating the correlation coefficients, except for the C factor. This model establishes a benchmark for comparing the performance. The C factor in equation (34) had to be predetermined so that it would mirror the dataset and the potroom technology the data was obtained from. For the second approach, multiple linear regression (MLR) was used to re-estimate all the parameters using the small dataset.

The third approach used a partial least squares (PLS) regression model. This prediction method suits well for datasets where there is strong collinearity between the variables, as is in this case. The objective function is to maximize the covariance explained between two blocks of data,  $\mathbf{X}$  and  $\mathbf{Y}$  [59]. It performs dimension reduction on the two matrices, to remove noise and extract useful information, and builds linear relationships between the two spaces. The  $\mathbf{X}$  and  $\mathbf{Y}$  matrices are centered and scaled to zero mean and unit variance and projected onto a low-dimensional subspaces (planes or hyperplanes) so-called latent variable subspace. The method of selecting the number of latent variables (components) can vary between softwares but cross-validation is the most common one [59]. Equations (35) and (36) represent the models for  $\mathbf{X}$  and  $\mathbf{Y}$ , where,  $\mathbf{T}$  are the scores,  $\mathbf{P}$  and  $\mathbf{Q}$  the loadings, and  $\mathbf{E}$  and  $\mathbf{F}$  the residual matrices for  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively. Equations (37) and (38) represent the common latent variable subspace relating  $\mathbf{X}$  to  $\mathbf{Y}$ , where  $\mathbf{W}^*$  are the weights which relate the scores information to the original  $\mathbf{X}$  matrix and are calculated from the deflated weights,  $\mathbf{W}$ , and loadings  $\mathbf{P}$ . Full explanations on the algorithm can be found in reference [82].

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (35)$$

$$\mathbf{Y} = \mathbf{TQ}^T + \mathbf{F} \quad (36)$$

$$\mathbf{T} = \mathbf{XW}^* \quad (37)$$

$$\mathbf{W}^* = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1} \quad (38)$$

Two metrics of the multivariate model were used to check for outliers and make sure the new data is consistent to those used to build the model, shown in equations 39 and 40. The first one is  $T^2$ , which represents the distance from the origin to the projection of the point on the plane. It is calculated using the score value ( $t_{i,a}$ ), where  $i$  is the row number (observation) and  $a$  the component number of the model using  $A$  components, and the standard deviation of score column  $a$ . The squared prediction error, SPE, is the distance off the model plane to the point. It is calculated using the residual vector,  $e$ , between the actual and predicted observation,  $i$ , using a model with  $A$  components.

$$T_i^2 = \sum_{a=1}^A \left( \frac{t_{i,a}}{s_a} \right)^2 \quad (39)$$

$$SPE_i = \sqrt{\mathbf{e}_{i,A}^T \mathbf{e}_{i,A}} \quad (40)$$

In order to predict  $\mathbf{Y}$  values when a new  $\mathbf{X}_{\text{new}}$  matrix is passed through the model, where each  $x_{\text{new}}$  must be scaled using the training average and standard deviation. The following relation is used:

$$\hat{\mathbf{Y}}_{\text{new}} = \mathbf{X}_{\text{new}} \mathbf{W}^* \mathbf{Q}^T \quad (41)$$

The large dataset was composed of 176 variables with high collinearity and, given that this is production data, contains noise. In order to extract only relevant information of the interaction between all the variables the PLS method was applied. One feature of PLS is that the variables' importance to prediction (VIP) can be calculated [83]. In the case of the larger dataset this is a useful tool in determining the contributing factors to NCC. The VIP calculation is shown in the following equation:

$$VIP_j = \sqrt{\frac{k \sum_{a=1}^A \left[ SS_a \left( \frac{w_{aj}}{\|w_a\|} \right)^2 \right]}{\sum_{a=1}^A (SS_a)}} \quad (42)$$

where

$$SS_a = q_a^2 t_a' t_a \quad (43)$$

i.e. the  $\mathbf{Y}$  loadings ( $q$ ) and scores ( $t$ ) for each component from the calculated PLS model. Thus, for each variable  $j$ , its overall importance to the model is determined from the total number of variables,  $k$ , the explained sum of squares for each component ( $SS_a$ ) and the importance of variable  $j$  to the overall weights for each PLS dimension  $a$ . If all variables would have the same contribution to the model, they would all have a VIP of 1 [83]. Hence, variables with values higher than 1 are considered the most important ones.

When comparing the performance of the different methods, two metrics were employed. The first one is the variance explained of  $\mathbf{Y}$  data given by  $R^2$ , the model's coefficient of determination, for the training dataset. The greater the value the more variance is explained. For the validation dataset the variance explained of  $\mathbf{Y}$  data is given by  $Q^2$  calculated from the residual and total sum of squares of the testing dataset. The second metric is the root mean squared of prediction error (RMSE) and is an absolute measure of prediction accuracy.

## 3.4 Results

### 3.4.1 Data cleaning and pre-processing

The outliers were identified and cleaned up in three steps. Firstly, the anodes containing bad data were removed. Secondly, the anodes that experienced extreme conditions within the cell, during their lifespan, were removed. These were, for example, prolonged periods of temperature higher than 990 °C, frequent and severe anode effects or pots with high instability lasting a few days. Thirdly, while developing the PLS model and using the  $T^2$  and SPE metrics, the data-points outside the main cluster were thoroughly investigated and thus the remaining outliers removed on a multivariate level. This ensures that only legitimate data was used to build the model. The data was also corrected for metal inventory variations in order to minimize the margin of error when calculating the total current efficiency. This was done by taking the difference between

the metal level measurements on the day the anode was placed in the pot and on the day it was removed from it. The equivalent amount of metal per cm was then added or subtracted from the tapped amount of the day the anode was taken out of the pot.

Lastly, the cell factor was considered. The potline within the current dataset includes a few variations of pot lining design with the same technology. The predominant designs were selected and marked as binary indicator variables within the dataset in order to estimate the MLR regression coefficient for each individual design. This method ensures that the intercept in the equation is replaced by the cell factor corresponding to each design. Overall, once the data was cleaned, 1088 anodes were used for the analysis.

### 3.4.2 Analysis of the small dataset

The prediction power of the different statistical methods employed on the small dataset, is presented in Figure 35.

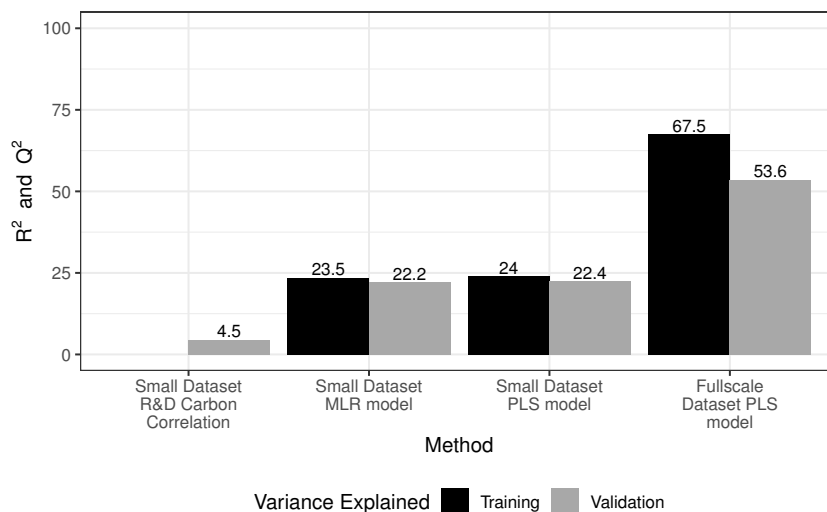


Figure 35: Variance explained ( $R^2$  and  $Q^2$ ) comparison between the methods

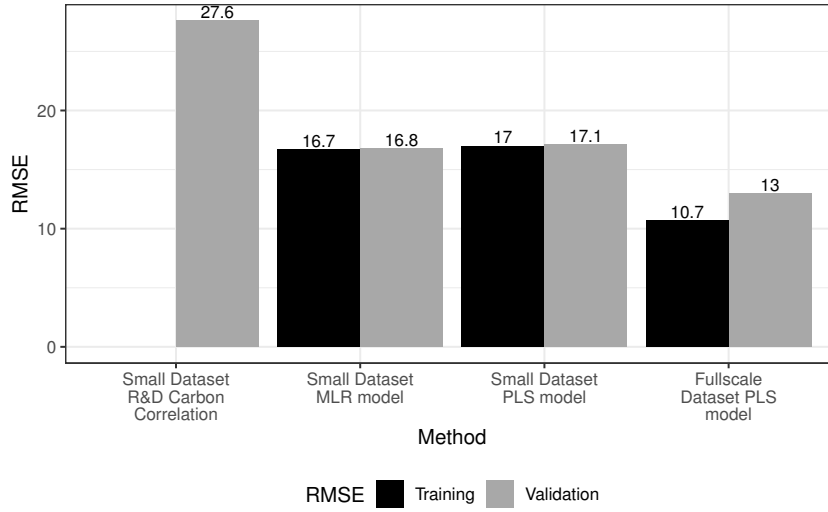


Figure 36: *RMSE* comparison between the methods

In the first step a reference was established by applying the R&D formula directly on the current dataset, which was split into 75 % for training the model and 25 % for testing. In order to establish a distribution for the NCC predictions, the observations to include in training and test were resampled at random 500 times and a model trained and tested for each individual run. The cell factor was calculated on the training set, each time and its value used in the equation, on the testing data, to predict the NCC. However, on an anode basis it provided a poor prediction of NCC, which exhibits the lowest variance explained and highest RMSE, as can be seen from Figure 35 and Figure 36. Here, the median value of the 500 runs was selected to be represented on the barchart.

MLR was employed to re-estimate the coefficients of each parameter based on the data at hand, to try to improve the model performance. The method was applied within the same 500 iterations presented above, in order to establish distributions for the parameters. The prediction power of the model increased thus to an  $R^2 = 23.5$  %, for the training dataset, while the root mean square error decreased, showing that the improvement is significant. Here, again, the barchart represents the median value of the runs.

To identify the driving factors behind the MLR analysis, the data was centered and scaled so that the regression coefficients would give an indication of the contribution of each variable, independent of their original units. These are represented in the boxplot in Figure 37. The raw units coefficients are presented in Appendix C. As it can be seen, the highest coefficient is for the current efficiency of the cell, making it the main driving factor behind this correlation.

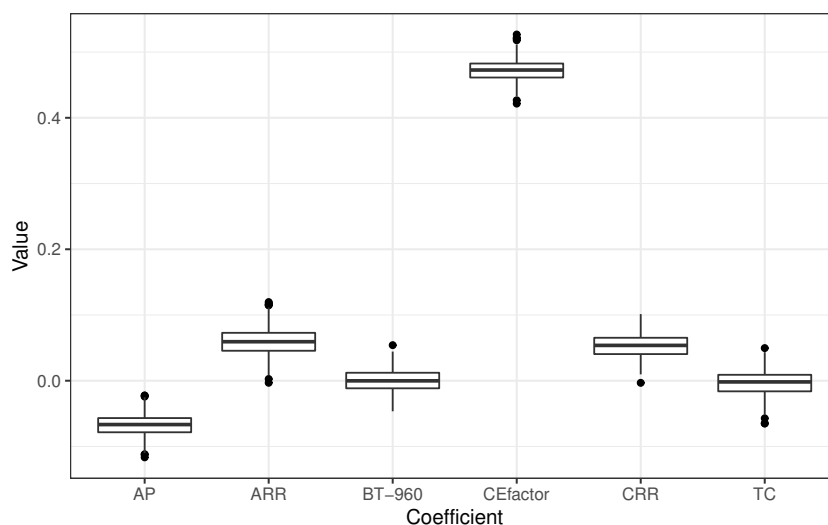


Figure 37: Distribution of regression coefficients in the MLR Monte Carlo simulation.

An obvious difference from the original correlation, is the sign of the AP and the residue reactivities (ARR and CRR). While they are in accordance to physics, that is, AP having opposite sign to the ARR and CRR, their contribution to the NCC is reversed. An increase in air permeability should result in a lower residue reactivity and therefore more net carbon consumption. However, this discrepancy could be explained by the the nature of the dataset the models are based on. This is production data on an anode basis, containing no data from carefully designed experiments. The time-frame spanned for each anode is 24, 25 or 26 days depending on the set cycle of the given anode. The data is therefore quite noisy and these methods only look at the correlation structure of the dataset in question. The data is not causal because of the high level of collinearity between the variables. Air burns, pot covering practices and carbon dust formation could contribute significantly to the NCC, but that information is not captured in this dataset.

A second factor could be the nature of the initial data extraction, which was, on an anode basis. The green anode properties were linked to the baked weight of the anode but the baked anode properties were averaged from the cores extracted during each week, and applied to all anodes for that specific week. Thus, the assumption was made that all anodes produced in a given week had the same baked properties. Further along the overall smelting process, these anodes were then linked to the individual pots, where they were placed in. Looking at the small scale dataset that would result in many anodes having the same baked anode properties but different CE, bath temperature and NCC. As it turns out, the bath temperature and thermal conductivity are insignificant for this model, as the coefficients are nearly zero in Figure 37. Restricting the MLR model to the correct signs of the variables, as in the R&D Carbon formula, did not allow finding a better solution for this dataset. This points to the fact that the small dataset (i.e. variables used in R&D carbon formula) does not contain sufficient information for predicting the NCC at the individual anode level.

A PLS model was created on the same six variables used in the MLR model. This was done in

order to check if a different statistical method, having the ability to cope with highly collinear data, would be able to better predict the NCC. However, as was shown in Figures 35 and 36, the model is similar to MLR. The number of principal components in this case is 2, selected by cross-validation, where, like in MLR, the main driver is the current efficiency. This points to the fact that the dataset does not contain sufficient information to predict the NCC at the individual anode level.

### 3.4.3 Analysis of the large dataset

Figure 38 shows a comparison between the observed versus predicted values of NCC kgC/tAl given by the individual PLS models for the small, 6 variable dataset and the large one, full-scale dataset, in Figure 39. The improved prediction is emphasized in Figure 39 by the close alignment of the data-points to the 45 °diagonal line, which represents the line of perfect correlation. The variance explained in the NCC increases to 67.5 % and the RMSE drops to 10.7, in the training dataset. The full-scale dataset model was built with 9 principal components selected by cross-validation.

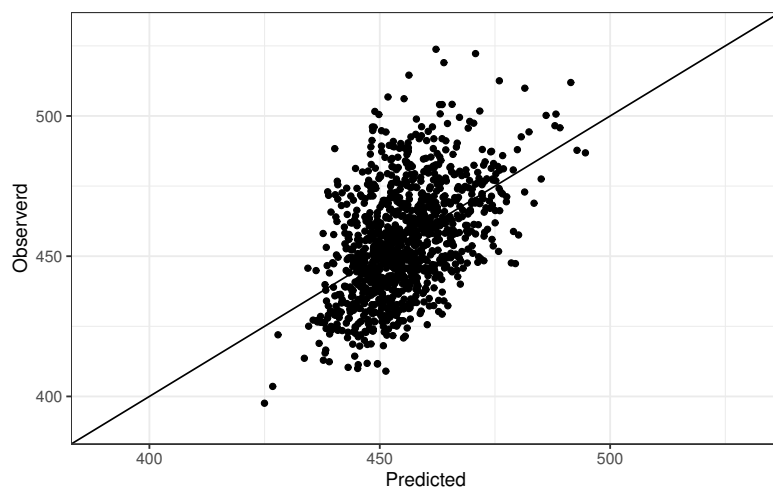


Figure 38: Scatter plots showing the line of perfect correlation (diagonal line) and observed vs predicted values for PLS 6 variable dataset

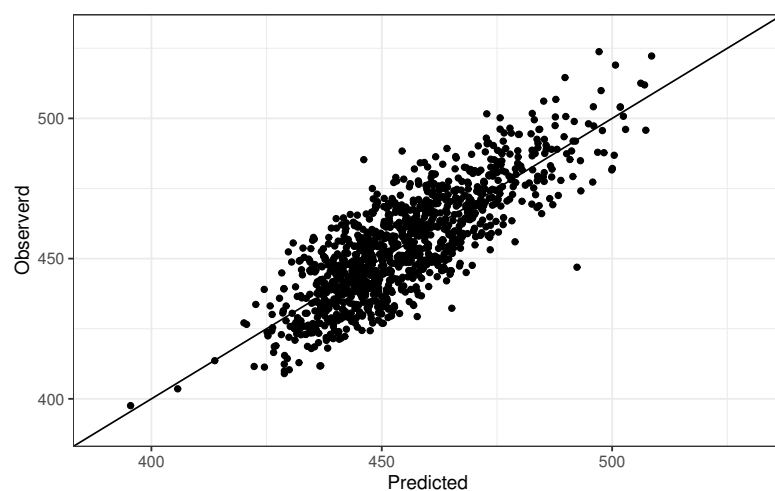


Figure 39: Scatter plots showing the line of perfect correlation (diagonal line) and observed vs predicted values for PLS large dataset

The model's variables important to prediction (VIPs) were calculated and shown in Figure 40. These are discussed further below.

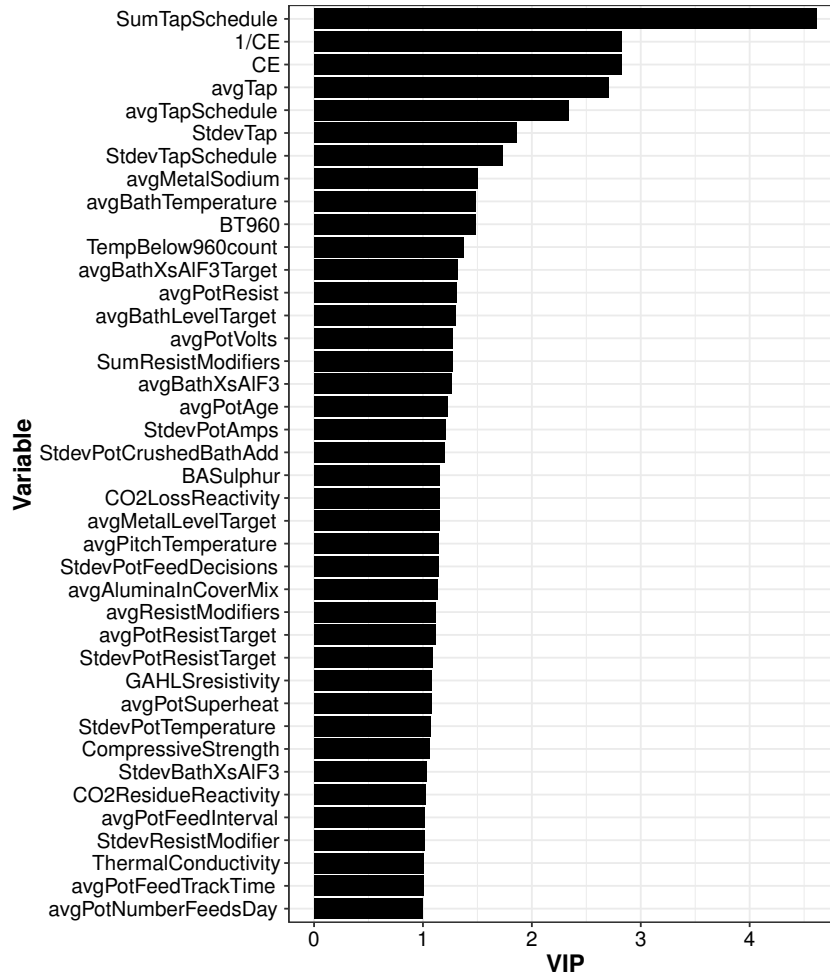


Figure 40: PLS model on full dataset: Variables Important to Prediction (VIP)

Because of their direct linkage to the calculation of NCC (i.e. kg carbon consumed per t of Al produced) the production variables show up as the most important ones as expected. As such, the alumina feeding parameters also show up on the chart, since it is directly related to production. In the aluminum electrolysis process the current is carried through the cathodic boundary layer by the sodium ions. The magnetohydrodynamic (MHD) metal-bath interface stability has an important effect onto the amount of sodium in the metal samples. Thus, in a cell with a uniform anode current distribution, there will be a highly stable interface, resulting in a high amount of sodium in the samples [84]. Since stable pots will have a higher CE, the sodium in metal samples shows up as an influencer in the prediction of NCC.

When a pot experiences periods of instability the anode-cathode distance is increased to stabilize the cell which results in excess heating. Also, elevated temperature, in itself, can increase the carbon consumption since it promotes dust and spike formation, while the pot's CE would also diminish thus increasing NCC. Variations of the ohmic resistance of the bath are strongly linked to the temperature and pot stability and therefore on the production rate. Common practice in the reduction cell controller is to decrease the anode-cathode distance (ACD) when the pot

is hot (temperatures higher than 980 °C) and increase the ACD when the cell is noisy, which increases temperature.

As the pots get older the operating parameters, such as the cell set-point resistance, bath chemistry and liquid levels, are set differently in order to protect the shell from premature tap-outs. However, these set-points have a detrimental effect to their individual production and thus will affect the overall carbon consumption. The anode reactivity will be affected because the anode block will sit higher above the deck plate, resulting in the carbon being more easily exposed. Also, lowering the excess  $\text{AlF}_3$  content in the bath, the cell temperature will increase which will result in a higher back reaction rate.

The alumina in cover mix is a metric of the composition of the anode cover material which is a mixture of crushed bath and alumina. This is used to cover the anodes after their setting to prevent air burn and it usually consists of 30-40 % alumina. If the percentage of alumina is lower, it would create a weaker crust that would have the tendency to melt down into the bath, exposing the anode. As previous studies have shown, the composition and granulometry of the cover mixture also contributes to the heat balance and fluoride fumes absorption in the cell [85,86].

The sulfur content of the baked anodes is shown in the plot, and, as studies have demonstrated [87,88], it improves their resistance to air burn. In terms of carbon dioxide reactivity, both the residue reactivity and loss reactivity show up on the plot. These metrics are strongly influenced by the sodium content in the anodes, which comes from the recycled anode butts within the recipe. The excess carbon consumption has been previously shown to be correlated to the sodium content in the anodes which acts as a catalyst to air and  $\text{CO}_2$  reactivities [89,90]. The compressive strength of the anode plays a role in its ability to withstand the stress imposed by the routine anode operation and occasional cell disturbances. Lastly, in accordance to the original R&D Carbon correlation, the thermal conductivity shows up as an influential factor of the baked anode properties.

It is interesting to see that some green anode properties also show up on the chart. High softening point pitches are thought to have a beneficial impact on the anode properties because of their increased coking value, given that they can properly penetrate the petroleum coke in the green anode [91]. The resistivity measurements of the green anodes can give information on their quality before entering the baking furnace. Exploration within the same research group has shown that green anode resistivity measurements could be used to detect post-baking anode agglomeration, which could indicate over/under pitching of the anode and thus, anode sticking potential [92]. The impact of these measurements on the overall NCC will be further investigated in future work.

### 3.5 Conclusions

Technology advancements in data acquisition and storage open the door to new types of analyses that have the potential to reduce the plant's environmental footprint and also reduce operating costs in the aluminum smelting industry. One of these is the possibility of tracking the net carbon consumption on an individual anode basis. As such, new analysis methods have to be

employed as opposed to the existing tools. The R&D Carbon correlation relates key baked anode properties to overall reduction cell performance and in recent studies it was used to optimize the anode manufacturing process. Using the anode-by-anode dataset, captured via an anode tracking system, the prediction performance of the existing correlation was compared to latent variable methods.

The methodology proposed is to use a PLS model built on the industrial data collected from the anode tracking equipment. In that way, a more detailed picture of the NCC can be build, down to the individual anode. These proven statistical methods provide insight into which variables contribute the most to the overall consumption and could thus allow for preventive actions to be taken in order to bring it closer to the theoretical value.

In this paper, an attempt was made to recalibrate the original NCC correlation to the data on hand, using MLR and PLS. While the prediction power improved, it does not provide enough detail at the individual anode level. Using PLS on the full dataset collected from the anode tracking the NCC could be predicted with an  $R^2$  of 66.6 %.

Because of the abundance of variables, coming from different databases and the sequential nature of the aluminum production process further investigation can be done. The dataset could be split into several different blocks and the relation between each of them investigated in the latent variable space. This methodology could dig deeper into the contributions of individual variables and overall importance of the CE to the model. It could therefore, point the plant staff to the corrective action to undertake in order to optimize overall performance.

It is important to mention that the R&D Carbon formula was intended to be used at a line level not at an individual anode level. The multivariate methodology has not been applied to the aggregated data, at a line level. This could also be explored further in future work.

## Chapter 4 Net carbon consumption in aluminum electrolysis - impact of anode properties and reduction cell operation variables

### Résumé

Dans le procédé de production d'aluminium Hall-Héroult, la quantité de carbone consommée par tonne d'aluminium produite est une mesure importante. Les écarts par rapport aux conditions idéales dans les cuves d'électrolyse contribuent à l'écart entre la consommation théorique de carbone et la consommation nette de carbone (NCC) réelle. Les travaux antérieurs des auteurs se sont concentrés sur la prédiction de la consommation nette de carbone pour des anodes individuelles à l'aide des moindres carrés partiels (PLS), sur la base de données collectées par un système de traçabilité des anodes. Dans cet article, l'importance de chaque étape de la production d'aluminium sur la consommation nette de carbone (NCC) est identifiée et un lien avec les propriétés des l'anodes vertes est établi. Les données collectées ont été analysées à l'aide de l'algorithme des projection des structures latentes multi-blocs séquentiels (SMB-PLS). Les données ont été divisées en différents blocs, déterminés par la séquence du procédé de fabrication. La modélisation a montrée que les trois blocs les plus importants étaient les paramètres d'alimentation en alumine, les propriétés de l'anode cuite et les propriétés de l'anode verte. L'homogénéité locale de l'anode verte a été définie comme des mesures de variabilité de la résistivité, calculées à partir des mesures de résistivité de l'anode verte. Cette homogénéité locale de l'anode verte s'est avérée plus grande pour les anodes à faible NCC.

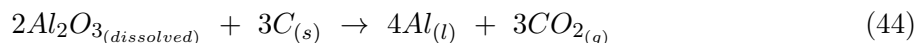
### Abstract

In the Hall-Héroult aluminum production process, the amount of carbon consumed per tonne of aluminum produced is an important metric. Deviations from ideal conditions in the reduction cells contribute to the discrepancy between the theoretical carbon consumption and the actual Net Carbon Consumption (NCC). Previous work from the authors, focused on predicting the net carbon consumption for individual anodes using Partial Least Squares (PLS), based on data collected by an anode tracking system. In this paper, the importance of each aluminum production step to the NCC is identified and a link to the green anode properties is established. The data collected was analysed using the Sequential Multi-Block Partial Least Squares (SMB-PLS) algorithm. The data was split in different blocks, determined by the manufacturing process sequence. The modelling showed that the top three most important blocks were alumina feeding metrics, baked anode properties and green anode properties. Local green anode homogeneity was defined as resistivity variability metrics, calculated from the green anode resistivity measurements. This local green anode homogeneity was found to be greater for anodes with low NCC.

### 4.1 Introduction

Aluminum production plays a key role in the Canadian economy providing around 10500 jobs and contributes \$4.7 billion to the country's GDP. In 2020 the output was of about 3.12 million tonnes which placed Canada in fourth place in terms of world production, while Iceland was in 10th place with 840 thousand tonnes of aluminum produced [14]. However, both countries use renewable energy sources, thus having among the lowest carbon footprint in the industry

[93,94]. This is an ever increasing advantage, due to the ongoing fight against climate change and the energy intensive nature of the process [94]. Aluminum is produced through electrolysis of aluminum oxide using the Hall-Héroult process. The alumina is dissolved in molten cryolite while passing a high DC current between the carbon anodes and cathode of the reduction cell. The anodes are consumable and are predominately pre-baked, and must thus be replaced at regular intervals [9]. In modern reduction cells the current can be from 100 kA to over 600 kA. At the Alcoa Deschambault smelter, where this study is performed, the amperage is in the range of 380 kA, and the anodes are changed every 24 to 26 days. The overall reaction in the production process can be expressed by equation 44:



From the stoichiometry of the above equation, it can be seen that about 0.33 kg of carbon are needed for each 1.89 kg of alumina in order to produce 1 kg of aluminum metal. However, in practice, between 0.40 and 0.45 kg of carbon are consumed, mainly due to aluminum re-oxidation and other reactions of carbon with oxygen and CO<sub>2</sub> [9]. Dust formation can also occur based on the anode properties by the selective oxidation of the binder matrix [32]. This effect is promoted when the cell operation temperatures is high. Excess air burns can occur when the carbon anode is exposed to ambient air, either by being poorly covered with anode cover material or when the pot experiences low liquid levels [95]. Prolonged periods of high cell instability can significantly reduce the amount of metal produced, which, in turn, would increase the overall carbon consumption.

The NCC is an important key process indicator in aluminum smelting but it is very difficult to monitor in real-time since it would require in-line measurements of the metal produced and of the anode's consumption rate during their set cycle. Usually, it is calculated using mass balance from monthly averages of plant wide measurements. The most common tool used to predict the NCC is the formula developed by Fischer et al. [17], which links baked anode properties to reduction cell operation key process indicators.

Technology advancements, in the data acquisition and analysis space, have enabled the development of new tools for process improvement. Anode tracking systems are set up at smelters to follow individual anodes throughout the operations chain [33,34,96]. The data collected from the different processing steps can then be fused and analyzed jointly to provide insights into which areas of the process can be improved.

In a recent study by the authors, the net carbon consumption was estimated at the individual anode level, using an anode tracking system implemented at the Alcoa Deschambault smelter [97]. This was accomplished by linking the different production databases together and employing multivariate analysis methods. The algorithm of choice was Partial Least Squares (PLS) regression, which could explain 66.6 % of variance in NCC while providing insight into the contributions of each variable to the predicted carbon consumption. However, using PLS, it was not possible to quantify the relative importance of the green and baked anode properties and that of each manufacturing step in the prediction of NCC. This is because with PLS one cannot distinguish between correlated or redundant information, contained in the variables measured in the differ-

ent manufacturing steps, from the new information brought by each of them. PLS regression does not impose a block ordering sequence. For example, if variations in green anode properties introduce fluctuations in downstream materials (i.e., baked anodes) and processes (i.e. reduction cells), the blocks of variables for these materials and processes will be correlated to some extent (i.e. contain redundant information), and PLS will identify them as important contributors, regardless of which block the fluctuations originate from. The sequential block ordering and orthogonalization procedures of the Sequential Multi-block Partial Least Squares (SMB-PLS) remove such redundancy between the blocks and highlights the new information each block brings into explaining variations in the responses variables (i.e., NCC).

The current chapter has two main objectives. The first one is to quantify the relative importance of the data collected from green and baked anodes, and reduction cells in explaining NCC variations. This is done by splitting the data into several different blocks representing each manufacturing step. The second objective is to improve the data interpretation by analyzing the correlations within the blocks. To accomplish these goals, the SMB-PLS algorithm was employed, because it removes redundant information between the data blocks, after arranging them sequentially, according to the process flow diagram.

This chapter is structured as follows. First, the data collection methodology is presented along with an introduction to the modelling approach used. The results section is focused on the interpretation of the model and correlations found between predictors and NCC.

## 4.2 Anode manufacturing and smelter data collection

Data science and advanced analytics are currently implemented throughout the industry, and contribute to improving process operations and product quality. Each plant collects great amounts of data which can be processed using various techniques with the objective of developing new tools to improve the production. Within this scope, an anode tracking system (ATS) was set up at the Alcoa Deschambault smelter to identify individual anodes throughout the whole production process [97]. Cameras were placed at strategic locations along the production chain, starting with the, newly formed, green anode. Another camera captured the anodes just before going in the baking furnace and yet another after baking. The individual anode identification number was transferred to the stem, after rodding, and camera equipped cranes, were able to identify the pot number and position that the individual anodes were placed in. At the end of their cycle in the potroom, the butts were inspected by a camera at a final checkpoint. This traceability facilitates linking the different production databases together, in order to better understand the impact of individual anode properties on the reduction cell performance and use that knowledge to further optimize the process. A total of 1088 anodes were tracked using the ATS over a period of 2 years from 2016 to 2018. For each anode, the system allowed collecting data on 176 variables, including properties of green and baked anodes, as well as the reduction cell operating conditions prevailing during their set cycle. NCC was finally calculated for each anode block.

In this work more emphasis was placed on the anode properties collected in the context of the multi-block approach. The data in the green anode block includes resistivity measurements for the green anode along with its height and weight. There are also some anode recipe parameters such as pitch quinoline insolubles, pitch and paste temperatures. The green anode resistivity

measurements were collected using the four point probe apparatus (4PP), developed by Alcoa [98]. Their locations are shown in Fig. 41, where the names of each measurement are represented by their location such that *TLS* represents the top long side, *VLS* is the vertical long side, *HLS* is the horizontal long side, *BLS* is the bottom long side and *BSS* is the bottom short side. A known current is applied at the extremities of each double-headed arrow and the voltage drop is measured. The resistivity is then calculated using Ohm's law and a shape factor. More detailed explanations of the apparatus and its applications are provided in reference [92] and in [99].

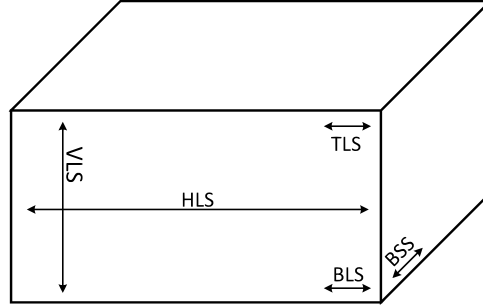


Figure 41: Schematic representation of the green anode resistivity measurements used in the analysis

To better assess the importance of the green anode resistivity measurements, two calculations were introduced in the dataset. These resistivity variability metrics, denoted by *DLS* (Delta Long Side) and *DSS* (Delta Short Side), are given by equation 45 and 46. The naming convention identifies which resistivity is being subtracted from the BLS metric. The DLS and DSS represent a proxy for the local green anode homogeneity [98], but not representative of the complete block. The lower the *DLS* and *DSS*, the greater the local homogeneity would be for the given green anode.

$$DLS = BLS - TLS \quad (45)$$

$$DSS = BLS - BSS \quad (46)$$

The ATS has limitations for the baking process. At this stage, individual anode number readings were performed before the anode entered the baking furnace and upon its exit. The anodes are not tracked inside the baking furnace and thus no data was collected during the baking process for the individual anodes in this study. In addition, the baked anode properties are limited to the core samples' lab analysis, which are available for about 1-2 % of the anodes produced in each week. Thus the weekly measurements were averaged and all anodes produced within the same week were assumed to have their properties equal to the average of the week. These include mechanical and chemical properties from lab analysis, that are standard across the industry. The baked weight of each anode was also added to the dataset.

Using the ATS, the cell number and position of the anode within it, was known, thus enabling the link between anode properties and pot operation. The typical lifespan of the anodes inside

the pot was between 24 and 26 days. Cell operation variables were collected as daily averages for the period of time each anode spent in its respective pot. Summary statistics were computed from these daily measurements such as mean and standard deviation of each cell variable to be used in the analysis. In order to gain more insights, some variables were also summed up such as the total number of anode effect minutes or total amount of alumina feed. Yet other variables were also aggregated using a count function such as the number of times the pot was above 980 °C or the number of anodic incidents.

The dataset was split into 12 predictor blocks and 1 response block (NCC), arranged sequentially, corresponding to the aluminum production process flow.

Figure 42: Fused data structure. The number of variables included in each block is indicated at their top right corner. The number of tracked anodes (observations) is 1088.

and amperage were placed in one block (P2). All parameters regarding feeding were put in the following block (P3). The reasoning behind placing the resistance first was that the amperage and voltage determine the behaviour of the feeding. The controller measures the change of the ohmic resistance within the electrolyte as the alumina dissolves and is consumed. The feed control is based on this measurement, reacting when the pots are approaching a predetermined resistance threshold.

Bath chemistry and crushed bath (dry) additions were placed in the next block (P4). These are adjustments made to the composition and volume of the electrolyte. The aluminum fluoride content in the bath lowers its melting point and affects the overall temperature of the cell. The temperature itself was placed in the next block (P5) where it was combined with other metrics that are known to impact the anode consumption. These metrics include the superheat, alumina in cover mix, liquid level measurements (metal and bath) and the time it takes to cover the anode after its setting in the cell. All variables related to pot stability were placed in the next block (P6). These are pot noise metrics and the anode effects time and frequency. They are usually metrics that respond to a condition within the pot and thus they were placed at this stage. The anode effects, for example, occur when the amount of alumina in bath is insufficient or the anode surface is covered by gas bubbles, hindering the passage of the current and increasing localized resistance.

All the trace elements measurements from metal sampling were pooled into one block (P7). Trace elements were placed downstream in the block sequence since they are influenced by the pot variables included in blocks P1-P6. They are mostly impurities coming from raw materials (anodes and alumina) but also from anode rodding, iron, or pot shell side wall insulation, silicon, while copper can come from the cathode current collector bar.

Each cell's anode problems, registered by the operators, such as spikes, deformations, burn-offs, were placed in the P8 block. These are not issues with the tracked anode but rather any other anode in the cell that experienced a problem during that time.

All the metrics related to the metal tapped are placed in the second last block, (P9), since these are influenced by all the operating conditions captured by the previous blocks. Lastly, comes the current efficiency (CE) (P10) block which is derived from the actual metal tapped and the theoretical amount at the given amperage.

Overall, 176 variables were used in the modelling presented in this chapter. Out of these, 175 were used as predictors. The number of individual anodes (rows in the data matrices), were 1088. The dataset was divided, randomly, into 75 % training and 25 % testing.

### 4.3 Modelling of the data

The Partial Least Squares (PLS) algorithm is a multivariate regression technique used for analyzing large and highly collinear datasets and for making predictions. It seeks to maximize the covariance explained between one block of variables ( $\mathbf{X}$ ) and the predicted variable ( $\mathbf{Y}$ ) [37,59]. This method was also used by the authors in previous work [97].

The PLS modelling has limitations which come from pooling the dataset into one common latent

variable space. When dealing with very large number of variables this can make interpreting their relationships difficult [37,83]. While being able to indicate which variables are important, PLS does not eliminate redundancies between them. For example, because of the way the NCC is calculated, the metal amount tapped is a metric that is highly correlated with the NCC. With the traditional PLS algorithm, the metal amount tapped would appear as a very important contributor because it carries the influence of pot operation variables' variation onto NCC (redundant information) and the effect of the tapped metal amount (new information). Thus, it's influence would be over-estimated. Removing the redundancy using SMB-PLS results in the metal tapped having a less important contribution reflecting only the new information it brings in explaining additional variations in NCC.

Given the nature of the aluminum production process, the manufacturing steps are performed in a certain, predefined order. That is, the anode needs to be formulated, using different raw materials, baked and placed in the reduction cell. This enables the use of the Sequential Multi-Block PLS (SMB-PLS) algorithm [101], where the dataset can be split into several different blocks ordered sequentially, according to a process flow sheet. The variance of the response variable ( $\mathbf{Y}$ ), explained by this method, is essentially the same as in the PLS algorithm, but the strength of SMB-PLS is shown through easier interpretation of the data. Using the smaller data blocks, both correlated (i.e. redundant) and orthogonal (new) information can be identified and extracted. Another interesting aspect is that the production steps (leading to data blocks) can be ranked based on their contribution to explaining variance in ( $\mathbf{Y}$ ), in this case being NCC.

SMB-PLS uses a hierarchical structure composed of a lower, local block, level and an upper, super level, which is similar to the wide spread multi-block PLS (MB-PLS) algorithm [83,100]. The pseudo-code of the SMB-PLS algorithm is presented in the appendix D.1. SMB-PLS has two main features that improve interpretability over older MB-PLS methods. One is that, in SMB-PLS, the number of latent variables are selected for each individual block while MB-PLS uses the same number of latent variables for all blocks. This ensures that only useful information is extracted from each block, thus avoiding overfitting. The second is that, once the algorithm computes the block weights and scores of the first block, it regresses the subsequent blocks onto the first block scores, shown by steps 2.4. to 2.7. in Table D.1 in Appendix D.1. This way, all information correlated to the first block is pooled into the same component. Upon convergence, orthogonalization is achieved through super-score deflation. Thus, the next component will use only new information to further explain variance in the response variable. In this way, the algorithm provides the ability to distinguish between the variables' influence in one block onto subsequent ones while also determining their influence on the response variable. Another advantage is that, at the block level, metrics can be computed for monitoring each individual block. Therefore abnormal events can be traced back to the blocks creating them. Thus, using SMB-PLS, a single framework is used to capture all available information.

The Block Importance to Prediction (BIP) metric can be calculated to show the contributions of each individual block to the overall prediction and are given by equation 47.

$$BIP_b = \sqrt{B \frac{\sum_{a=1}^A \left( R_Y^2(a) w_{T,b}^2(a) \right)}{\sum_{a=1}^A R_Y^2(a)}} \quad (47)$$

where  $b$  is the block index,  $B$  is the total number of blocks,  $A$  is the total number of components that the model is built on,  $R_Y^2(a)$  the variance of  $\mathbf{Y}$  explained by the  $a^{th}$  component and  $w_{T,b}^2(a)$  the super-weight of block  $b$  (given by step 2.9 in the SMB-PLS pseudo-code in the appendix). The block contributions are not scaled to a contribution of 1 for the average block. Rather, their magnitude should be taken at face value, each one representing the block's relative importance to the model.

The contribution of individual variables to the model can also be calculated, using the model parameters calculated at the block level. This can be accomplished by comparing a selected group of observations to the center of the model (i.e. mean of the variables), which is zero when the data has been mean centered. Another way to do it is to select two groups of observations and compare their means, as shown in equation 48.

$$C_j = (\bar{X}_{jf} - \bar{X}_{ji}) \sqrt{\sum_{a=1}^A \left( \frac{T_f - T_i}{s_{ta}} \right)^2 w_{ja}^*} \quad (48)$$

where  $C_j$  is the contribution of variable  $j$ ,  $\bar{X}_f$  is the mean of the  $j^{th}$  variable in the final group (to group) and  $\bar{X}_i$  is the mean of the  $j^{th}$  variable in the initial group (from group) while  $T_f$  and  $T_i$  are the block scores of the respective groups, calculated in step 2.3 of the pseudo-code shown in the Appendix D.1.  $s_{ta}$  is the standard deviation of the block score for latent variable  $a$  and  $w_{ja}^*$  are the block weights of the variables, calculated in step 2.1. of the pseudo-code shown in Appendix D.1.

## 4.4 Results

### 4.4.1 Data cleaning and pre-processing

Process data collected was cleaned, mean centered and scaled to unit variance. Each variable was plotted and visually inspected to identify and remove outliers and bad data (missing data or erroneous entry of manual measurements) after confirmation based on expert knowledge of potroom operation. If any of the anodes experienced extreme operation conditions, during their time inside the pot, they were removed. These include elevated temperatures, 990 °C or higher, prolonged instability or severe anode effects requiring lengthy manual intervention to recover, and showing up as outliers in the dataset. This was done in order to capture only normal process variations. In this context, normal is considered to be stable operating conditions of AP-30 reduction cells running an amperage of about 380kA. After the univariate cleaning was performed, a second verification step was done at a multivariate level. The squared prediction error, SPE, (Appendix D.2, equation 11), and Hotelling's  $T^2$  (Appendix D.2, equation 18), were used to identify data points that are inconsistent with the bulk of historical data collected in this

work. The SPE is used to verify the correlation structure of the data, while the  $T^2$  focuses on the magnitude of the variables. The data analysed spanned only the lifetime of an individual anode, so adjustments were made to accommodate for metal inventory variations in order to minimize the margin of error when calculating the current efficiency. This was accomplished by taking the difference between the metal measurements on the day the anode was taken out of the pot and the day it was set. The metal height difference was converted to the equivalent weight in tonnes, which was then added or subtracted from the amount of metal tapped on the day the anode was taken out of the pot.

#### 4.4.2 SMB-PLS modelling

In this section, the results cover three main aspects identified using the SMB-PLS algorithm. First, the importance of each block was analyzed and is presented. Secondly, the correlation structure between blocks is shown. Lastly, the most important blocks were selected and analyzed deeper, at the variable level, to reveal new information about the process variables' relation to NCC.

After the initial data cleaning and pre-processing, the block ordering was established. This is an important step, prior to applying the algorithm, since SMB-PLS analyzes the data sequentially and thus the results depend on it. The ordering chosen is shown in Table 3, however, within the reduction cell aluminum is produced continuously and thus the block ordering is not unique. Different combinations were tested regarding the resistance and alumina feed blocks. The overall variance explained ( $R^2$ ) of the response variable was 63.1 %, which is very close to the initial PLS model of 66.6 % [97]. Interchanging the resistance and alumina blocks showed minimal changes in overall model performance. The total variance explained in NCC was less than 2 % lower if the alumina block was placed before the resistance, while the relative block importance to the model remained the same. Due to the reasons presented above, it was decided to place the resistance block before the alumina, since the line amperage and pot voltage were believed to be the driving factors for the rest of the pot performance parameters.

Table 3: Number of SMB-PLS components for each data block

| Block | Block Full Name                   | Block Abbreviation | Components |
|-------|-----------------------------------|--------------------|------------|
| G     | Green Anode Properties            | BL1_GA             | 3          |
| B     | Baked Anode Properties            | BL2_BA             | 3          |
| P1    | Anode position in pot and pot age | BL3_PosAge         | 2          |
| P2    | Pot Resistance                    | BL1_Res            | 2          |
| P3    | Alumina Feed                      | BL5_Feed           | 2          |
| P4    | AlF3 and Bath Feed                | BL6_AlF3Bath       | 3          |
| P5    | Anode Reactivity metrics          | BL7_Reactivity     | 2          |
| P6    | Instability metrics               | BL8_Instability    | 2          |
| P7    | ImpuritiesInMetal                 | BL9_Impurities     | 2          |
| P8    | Anode Problems                    | BL10_AnodeProb     | 2          |
| P9    | Tapped metal                      | BL11_MetalTap      | 2          |
| P10   | Current Efficiency                | BL12_CE            | 1          |

As described above, in the SMB-PLS algorithm each predictor block can have independent number of latent variables. Thus, the number of components per block is not fixed and can differ between blocks. Using cross-validation, the number of components per block were chosen as shown in Table 3, resulting in 26 latent variables for the model as a whole. The dataset was split into two subsets, one containing 75 % of the data for training the model while the remaining 25 % was placed in the testing set. The root mean square error of prediction RMSEP and the predicted  $\mathbf{Y}$  variance by the model on the test set,  $Q^2Y$  statistic, were computed and used to select the number of components for each block.

The BIP metrics obtained from the fitted SMB-PLS model are presented in Fig. 43. The plot shows the importance of each block to the NCC prediction by the model where the size of the bars should be taken at face value. Thus, the most important block is Alumina Feed followed by the Baked Anode properties and Green Anode properties. By contrast, the CE block has the least importance. The CE is a key process indicator, calculated from the amount of metal tapped from the pots. That is because the information contained in this block is already included in the previous ones. It is interesting to see that the Green Anode block is the third most important block. However, it was expected that the alumina feed block would have a big impact on the model. It is directly linked to the amount of metal produced and therefore the calculation of NCC. The Baked Anode properties is the second more important block, but, as explained in section 2, only the weight of the anodes was available for each individual block. Weekly averages of the laboratory analysis of core samples were added to each individual anode, based on their week of production. This induced a lot of uncertainty in the baked anode data block and therefore it was not possible to reach a clear conclusion on the contributions of the variables within this block. Hence, the Green Anode and Alumina Feed blocks were chosen to be presented in more detail further below.

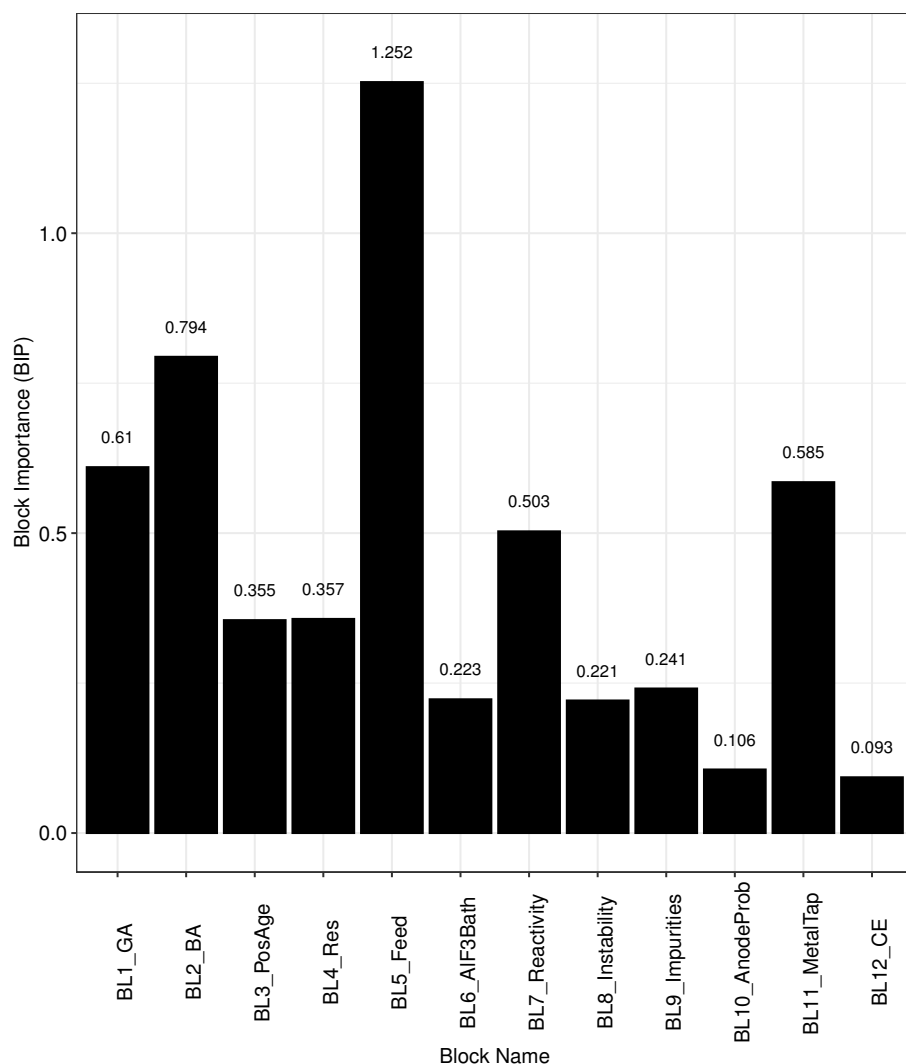


Figure 43: SMB-PLS, block importance to prediction (BIP)

As mentioned above, SMB-PLS can distinguish between the correlated information between blocks and the orthogonal, new, information contained in each of them. The bar chart below, in Fig. 44, shows how the information is distributed across the latent variables, more specifically, the correlation structure between blocks. The relative weight of each block and in each component is shown by dividing the super weight of each block by the norm of the super weight vector and expressing it in percentage. Since the algorithm is sequential, each component will first explain the variance within the current block, which is represented by the first (lowest) bar. The rest of the stacked bars represent the information in subsequent blocks that is correlated with the current block (redundant information), within that given component. For example, if two blocks are highly correlated within a component, the second block would be represented by a large bar stacked on top of the first one. If, on the other hand, they were orthogonal, the second block would have virtually no representation within that given component. When adding a component to the current block does not significantly improve prediction of  $\mathbf{Y}$ , the algorithm stops using this block and the modeling procedure continues with the next block in the sequence. The reader

is referred to [101] for further details on SMB-PLS.

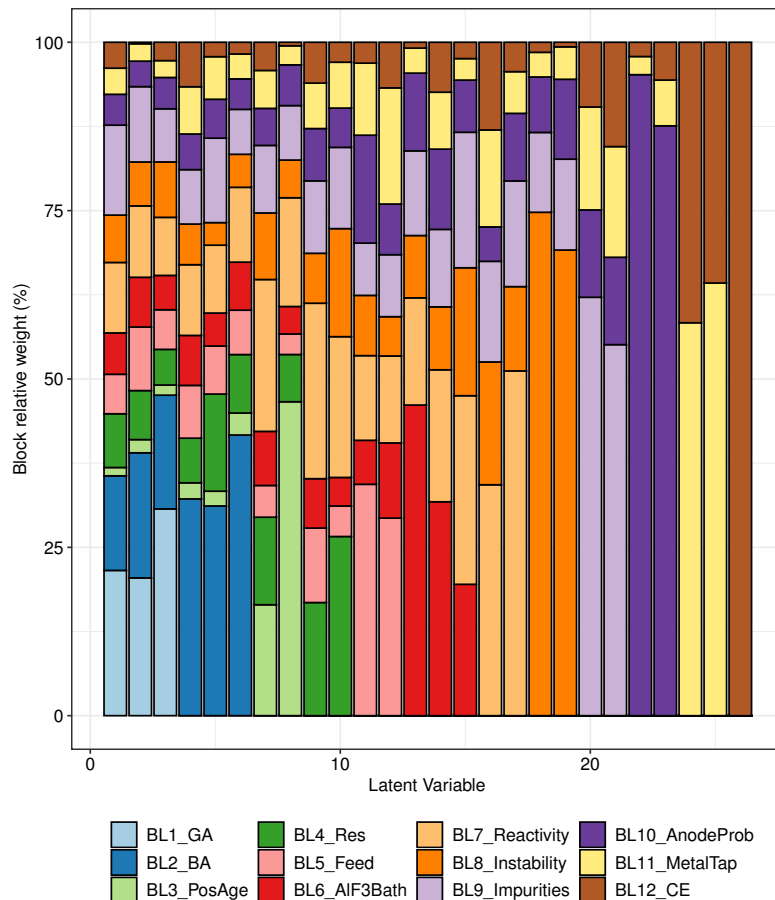


Figure 44: SMB-PLS, block relative weight for each latent variable

It can be seen that the first latent variable explains part of the variance in the green anode block, but it also explains variance in the baked anode block. This is to be expected because of the overall influence that mixing and forming will have on the final baked product. What is interesting to see, is that there is variance explained by the first component across all blocks in the model. This is the correlated information that is pooled into the first component. The relative weight of each block is represented by the size of the bar. Thus, it can be seen that the green anode properties have no influence on the position and age block, as it should be, since they are independent variables, as explained earlier. It does, however, explain more variance in the reactivity and impurities blocks. Some trace elements in the metal samples analysed, come from the raw materials of the anodes so, this representation makes sense. One interesting feature that can be observed at a quick glance, is that the position of the anode (block BL3\_PosAge) is influencing the reactivity block most, out of all the other ones, in component 7. This block contains age dependent variables, such as chemical, thermal and liquid level set-points which reflect on the operating temperature. This is something that was not captured by the previous built, PLS, model [97]. The dimension reduction was much greater, namely, 9 components from 176 variables, and therefore this information was hidden.

For each latent variable in the SMB-PLS model, the variance explained for each individual block is also calculated. This is shown in Fig. 45. As such, all the variance that can be explained within the green anode block reaches its peak at the third component. Similarly, all subsequent blocks, reach a maximum variance explained when the number of components for each individual block, defined by cross-validation, has been reached. The cumulative block variance explained is shown for all blocks and also for the response variable, NCC. It is interesting to see that the last block, CE, does not contribute much to explaining variance in NCC. The reason is that the CE is derived from amperage and actual metal tapped and these two metrics were placed in data blocks prior to the CE block. Thus, their impact onto the NCC had already been assessed, due to the sequential nature of the algorithm, by the time the CE block was reached. It is also seen from components 10 and 11, the alumina feed block, that its contribution to NCC is significant, something also seen in the BIP plot.

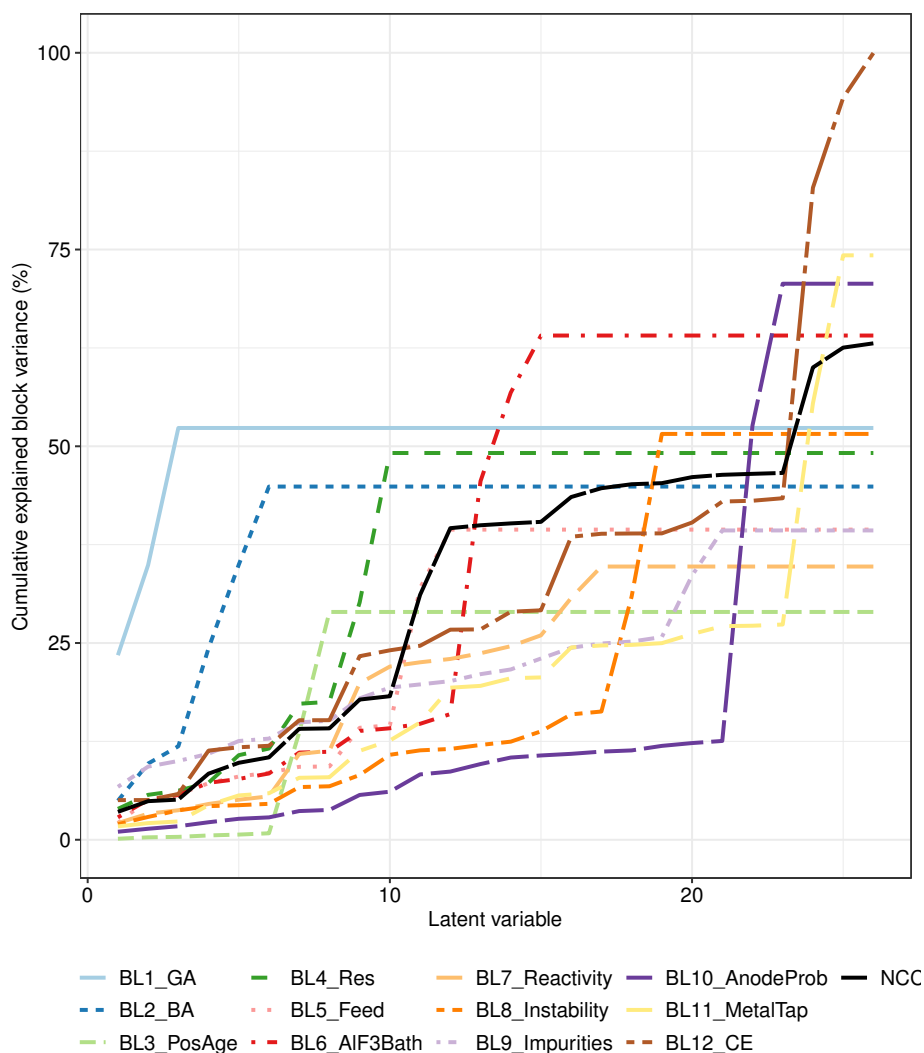


Figure 45: SMB-PLS, cumulative block variance explained by each latent variable

For the SMB-PLS model at hand, two groups of anodes were studied in more details. These

comprise of anodes that had a very low NCC versus the ones with a very high NCC. The response variable was divided in intervals of 5 kg C/t Al, from 405 to 525 kg C/t Al, based on the range within the dataset. The low NCC group includes 14 anodes distributed across the lowest 3 intervals, i.e. 405 to 420 kg C/t Al, while the high NCC group comprises of 15 anodes from intervals 495 to 525 kg C/t Al. The positions of the anodes in the two groups were split in 4 main categories as presented in Table 4. The high NCC anodes were consumed in 13 different cells, with 2 cells consuming 2 different anodes each, at separate times, which did not overlap. Hence, the anodes selected in the comparative analysis were consumed under a wide range of cell operating conditions. The low NCC anodes were consumed in 12 different cells, with 2 cells operating 2 different anodes with overlapping lifecycle period. However, these were not paired anodes. Additionally, the 2 groups of anodes were consumed in separate cells, that is, in a total of 25 cells. The anode positions were unique for each cell, i.e. no two anodes were consumed in the same location (pot and position) twice.

Table 4: Anode position category per pot

|          | <b>Tap hole</b> | <b>Feeder</b> | <b>Other positions</b> | <b>Duct end</b> |
|----------|-----------------|---------------|------------------------|-----------------|
| Low NCC  | 3               | 3             | 8                      | 0               |
| High NCC | 2               | 1             | 9                      | 3               |

The scores plots of each latent variable were studied but only 2 components are presented here. In Fig. 46, latent variables 1 and 11 are shown which represent the first component of the green anode block and the first component of the alumina feeding block, respectively. A clear separation between the high NCC and low NCC anodes is observed, emphasised by the diagonal line. This separation is greater on the vertical axis, LV 11, because the alumina feed block explains approximately 20 % of variance in NCC versus 5 % for the green anode block. The score plot suggests that green anode properties and variables in the alumina feed block need to be such that both LV1 and LV11 have negative values for the anodes to fall in the low NCC region.

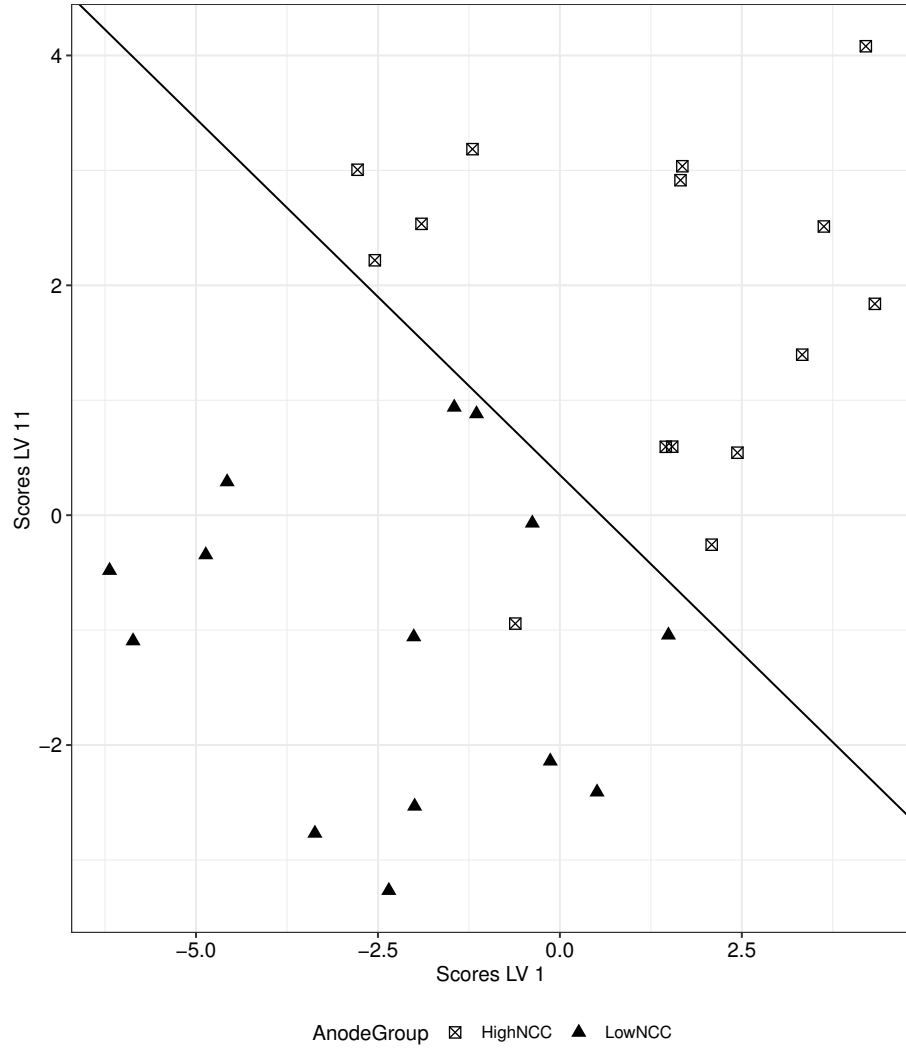


Figure 46: Super scores plots for the latent variables representing the green anode block and alumina feed block

Individual variables contribution are presented for the alumina and the green anode blocks. The former contains information that was expected to result from the model. The second one, contains new information, previously not discovered using the PLS algorithm [97]. This was identified by the introduction of the local green anode homogeneity variables (equations 45 and 46) and applying the SMB-PLS algorithm to the anode tracking data.

The variable contributions were calculated using the methodology presented above, in equation 48, where the initial (from) group was the High NCC one and the final (to) group was the Low NCC one. Thus, the contributions can be interpreted as the difference between the Low NCC anodes and High NCC anodes.

The Alumina feed block contains all metrics related with pot feeding and their contributions are shown in Fig. 47. A positive contribution means that the given variable has a higher magnitude for the low NCC anodes versus the high NCC anodes. In the opposite direction, the variables

with negative contributions have lower magnitudes for the low NCC anodes versus the high NCC anodes. The greatest contributor to Low NCC, in this block, is the sum of the total number of feeds. This was expected because of the way NCC is calculated, i.e. directly connected to aluminum produced and thus to the total amount of feeding to the pot. There is also indication that the Low NCC group of pots are running leaner, i.e. longer underfeeds because the average track time (underfeed time) is greater than for the High NCC group. This also makes sense, since leaner operation is correlated with better production at this smelter. The rest of the variables have a contribution below an absolute value of 0.5. Only the highest contributors were discussed for sake of conciseness of the article.

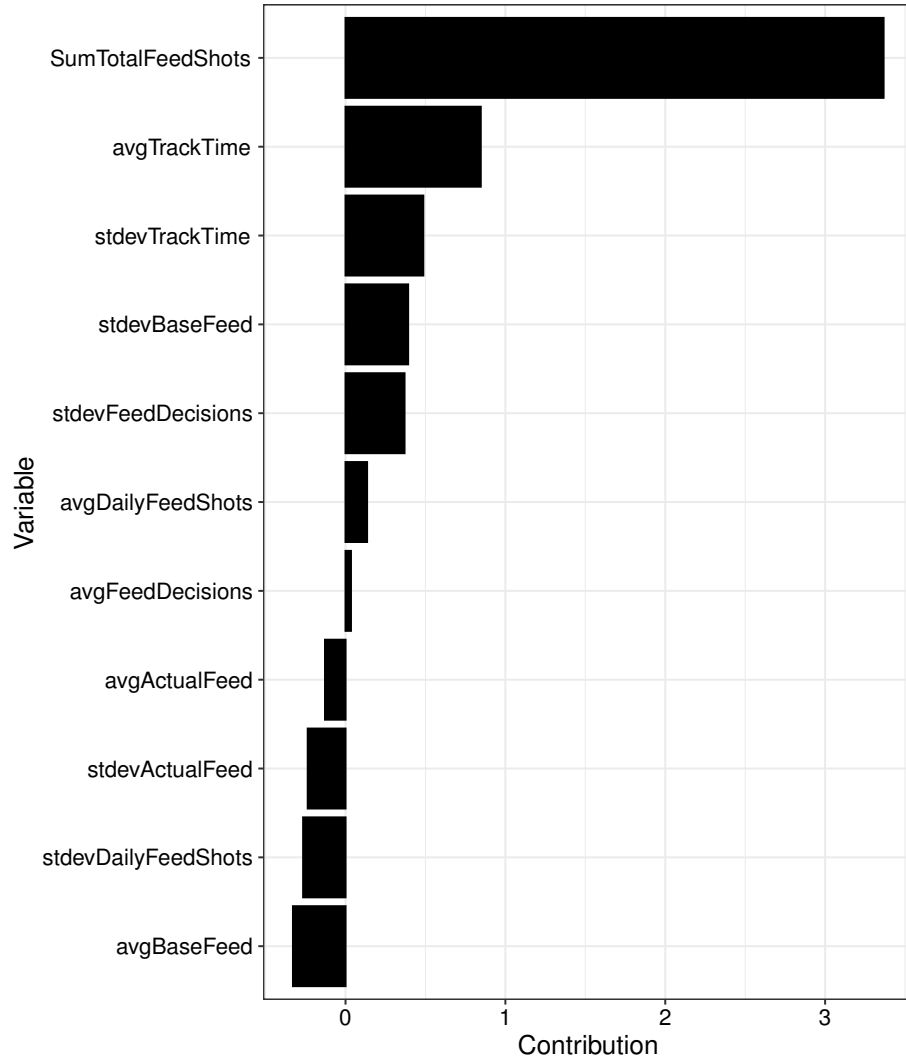


Figure 47: Alumina feed block  $T^2$  contribution plot, difference between Low NCC group and High NCC group. The full names of the variables are given in Table D.3, in the Appendix D.4

The contributions of the green anode block are shown in Fig. 48. It can be seen that the lower NCC anodes have a higher weight and height than the high NCC anodes. However, these variations are within the process limits. The anodes within this dataset all had similar features

and there were no design changes to shape or size within the timeframe spanned by the data studied. The difference in pitch temperature appears as an important contributor, being lower for the Low NCC anodes. However, further investigation is required to better understand this result. Another important difference is greater local green anode homogeneity of the low NCC anodes as opposed to the high NCC anodes. This is given by the difference in the resistivity measurements presented above, namely *DSS* and *DLS*. Green anode resistivity is known to be heavily determined by the internal structure of the anodes. This allows the current to be distributed in a more even fashion inside the anode, as opposed to finding the shortest path within a section of the anode. The lack of local green anode homogeneity could be attributed to multiple causes such as cracks around the probes' position, segregation, mixing intensity or temperature, uneven paste distribution in the mold or forming temperature. The resistivity and calculated local green anode homogeneity metrics (*DLS* and *DSS*) are the top contributors of the green anode block to NCC, suggesting that there is a relationship between them. Linking the local green anode homogeneity to NCC is important knowledge, not documented before, and highlights the need to conduct further work into the root causes. The other variables shown in 48 have very small contributions to the change between low and high NCC. Thus, they are not discussed further.

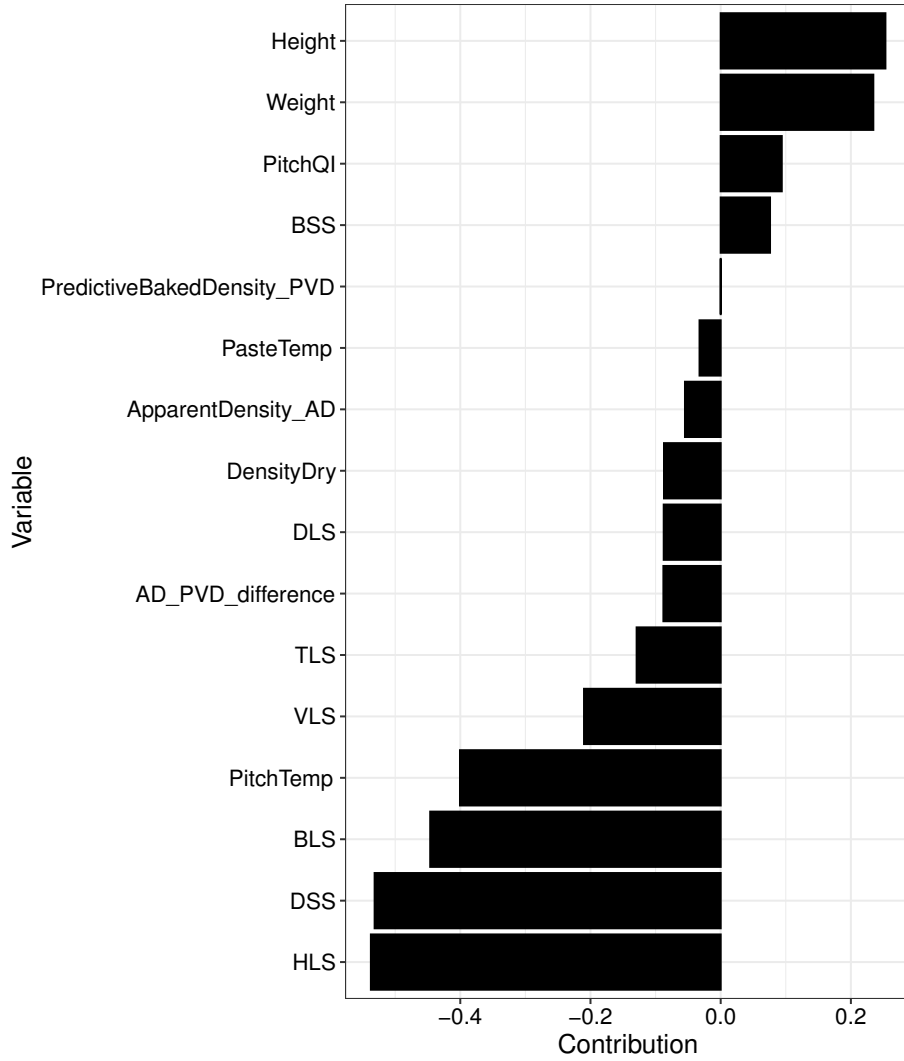


Figure 48: Green anode block  $T^2$  contribution plot. The full names of the variables are given in Table D.4, in the Appendix D.4.

## 4.5 Conclusion

In this chapter new insights were presented on the impact of green anode properties onto the net carbon consumption. This was enabled by applying the SMB-PLS algorithm to analyse the dataset collected from the anode tracking system at the Alcoa Deschambault plant. The data was split into 12 different blocks and ordered according to their sequence in the production process. A model was built using the NCC as the predicted variable resulting in 63.1 % of total variance explained. This methodology ensures that both correlated and orthogonal information is captured within a single framework. Thus the relative contribution of each individual block onto the NCC was revealed. The NCC interval was divided in intervals of 5 kg C/t Al produced and 2 subgroups of anodes were selected at the two extremes of the consumption spectrum (low NCC and high NCC). Individual variable contributions between the low NCC anodes and high NCC anodes were also calculated. An expected result was the high influence of the block

containing alumina feed information, upon the NCC. Another important observation was that the green anode block was the third most important block in the model. Further investigation into the green anode block showed that greater local green anode homogeneity of the green anode (calculated from the four point probe (4PP) resistivity measurements) corresponded to lower NCC values. This raises the hypothesis that local green anode homogeneity may play a significant role in anode over-consumption. However, this is a production data driven analysis and, at the time of this writing, there is no second dataset available to corroborate the findings. Further investigation into the resistivity measurements and their connection to anode quality and performance is needed to demonstrate this. In addition, building a monitoring scheme, based on the SMB-PLS model, to provide an early warning of potential issues with NCC, could also be investigated further.

## Conclusion and future work

Aluminum production is a complex task with challenges on multiple fronts. Economic challenges are pushing the smelters to continue to innovate in order to maintain competitiveness. Raw materials quality decrease as environmental constraints increase. The advantages are, that with improving technology, new insights into relationships between process variables can be uncovered. In addition, new process monitoring methods can be implemented in order to maintain process stability. The research presented in this thesis revolves around the application of latent variable methods to improve process monitoring and establish new links between process variables.

In chapter 1 a PCA model was built on smelting process data to detect and diagnose abnormal events. Normal operation data was used to build the model and thus drifts were observed when projecting new data onto the model. Using the PCA model it was able to identify that trace elements in the metal sampling can indicate onsetting abnormal events. These are associated with anode problems and a lag of 10 days was observed between decreasing trace element and detrimental feed control response of the cell.

In chapter 2 a PCA model was built and implemented in production at the Alcoa Fjardaal smelter. A reduction cell health ranking system was developed using the  $T^2$  metric. Its purpose was to standardize the order of the problematic cells and help operators prioritise the critical cells. Additionally, it could also help with the logistics around equipment availability. It was also shown how the contribution plots of both  $T^2$  and  $SPE$  can be used to identify issues with the problematic cells. With the help of the operators the model performance was assessed and compared to the existing reports. Though the sample size was small the model showed good potential with a precision of 94.0 %, 79.5 % accuracy, 10.9 % type I error, and 24.5 % type II error. In addition, an advance notice of one day was observed in 22.4 % of cases where a true positive was identified. Another 7.5 % of true positive cases showed an advance notice of 2 days. Having advanced notice of process deviations could enable proactive action in maintaining process stability.

In chapter 3 the data from an anode tracking system at the Alcoa Deschambault smelter was analysed. The individual anode properties, green and baked, were linked to their respective reduction cell performance during their operating life. It was shown how the data collected can be analysed and used to predict the NCC for individual anodes by employing the PLS algorithm. The model returned a variance explained of 66.6 % which is an improved prediction performance by comparing to the R&D Carbon formula at an individual anode level. This method can also provide insights into which variables contribute the most to the overall carbon consumption.

In chapter 4 the anode tracking data analysis is taken a step further. The data was split in sequential blocks, according to the process flow diagram and analysed using the SMB-PLS algorithm, with the NCC as the predicted variable. The variance explained in NCC was found to be 63.1 %, similar to the PLS model presented in chapter 3. This provided insights into the importance of each production step onto the NCC. However, the SMB-PLS algorithm provided insights into the relative contribution of individual blocks, with the green anode block being

the third highest. Deeper investigation into the green anode block showed that the green anode homogeneity metrics, calculated from the four-point-probe (4PP) resistivity measurements, were correlated to the NCC. Within this dataset, greater local green anode homogeneity corresponded to lower NCC values for individual anodes.

With respect to anode tracking data there are a few tracks to follow for future work. First option would be to look deeper into the green anode resistivities to unlock new insights into the green anode homogeneity. Recent developments at the plant allow for better green anode tracking and more reliable data collection. Experiments could then be carried out to track a series of green anodes in the baking furnace and assess potential correlations between green anode homogeneity and baked anode properties. Using newly developed non-destructive testing techniques for the baked anodes [102], the baked anode properties could be identified for a much larger sample size than with current core sampling methods.

A second option could be to improve the sample size in the anode tracking within the reduction cells. In this work only 1 or 2 pairs of anodes were known within each operating cell within the dataset. Therefore, assumptions had to be made regarding the rest of the anode population. Knowing the green and baked anode properties of all individual anodes in the reduction cells would improve the accuracy of individual NCC predictions. However, this is a challenging task due to the logistics required to track all the anodes.

A third option could be possible once a robust tracking system has been proven and deployed in production. Assuming options one and two have been carried out, multivariate specification regions for incoming raw materials could be defined combined with the optimization of operating conditions in the latent space of SMB-PLS, as proposed in [103, 104].

In terms of process monitoring further steps can be taken to improve the anomaly detection mechanism presented in this work. Increasing the sample size of pots verified and labeled by operators could have multiple benefits. Firstly, the larger dataset would allow for a more precise assessment of the model performance. Secondly, classifying the cells with different labels, would lay the groundwork towards a supervised model. A PLS classification model could be built [105], where conditional probabilities would be estimated and class membership assigned based on the maximum probability. An advantage of this method is its interpretability, as it retains the diagnostics characteristics, typical to regular PLS models.

Another avenue to explore, would be to look into data with higher granularity. The model in this work was based on daily aggregates of process operation data. At the process controller level, the data is captured at a higher frequency. This could be used to try to predict some of the manual measurements. These could then be fed into the daily model and be used in tandem to detect anomalies.

## References

- [1] Greenwood N. N., Earnshaw A. (1997), Chemistry of the Elements, Second Edition, Butterworth-Heinemann
- [2] <https://aluminium.ca/en/planet-aluminium/canada> - Accessed on August 18 2024
- [3] <https://www.elysis.com/en> - Accessed on August 18 2024
- [4] [www.samal.is/is](http://www.samal.is/is) - Accessed on August 17 2024
- [5] [www.samal.is/is/frettir-1/hvada-ahrif-hafa-alver-a-hagsveiflur-a-islandi](http://www.samal.is/is/frettir-1/hvada-ahrif-hafa-alver-a-hagsveiflur-a-islandi) - Accessed on August 17, 2024
- [6] Emsley J. (2001), Nature's building blocks and A-Z guide to the elements, Oxford University Press
- [7] Richards J. W. (1892), "Aluminium - the metal of the future", The cosmopolitan magazine
- [8] <http://electrochem.cwru.edu/encycl/art-a01-al-prod.htm> - Accessed on August 17, 2024
- [9] Grjotheim K., Kvande H. (1993), Introduction to Aluminum Electrolysis, 2nd Edition, Aluminium-Verlag, Dusseldorf
- [10] <http://heattreatconsortium.com/metals-advisor/anode-baking/> - Accessed on August 18, 2024
- [11] Kolåsa S., Størea T. (2009), "Bath temperature and  $AlF_3$  control of an aluminum electrolysis cell, Control Engineering Practice, Vol 17, Issue 9, pp. 1035-1043
- [12] Kvande H., Drabløs P. A. (2014), The Aluminum Smelting Process and Inovative Alternative Technologies, Journal of Occupational and Environmental Medicine, Vol 56, No. 5S, doi:10.1097/JOM.0000000000000062
- [13] World Aluminium the website of the International Aluminium Institute, <http://www.world-aluminium.org/statistics/primary-aluminium-production/> -Accessed on August 19 2024
- [14] <https://natural-resources.canada.ca/our-natural-resources/minerals-mining/mining-data-statistics-and-analysis/minerals-metals-facts/aluminum-facts/20510> - Accessed on August 19 2024
- [15] Fellner P., Ambrová M., Híves J., Korenko M., Thonstad J. (2005), "Chemical and Electrochemical Reaction of Sulphur Species in Cryolite Melts, Light Metals 2005, ed. by H. Kvande, (Warrendale,TMS,2005), pp. 577-581
- [16] Jassim A., Mofor E., Eddin J. W., Polle S., Whitfield D. (2018) Effect of Changes in Anode Top Cover Composition on Anode Butt Quality, Light Metals 2018, ed. by O. Martin (Cham,Springer,2018), pp. 1243-1251
- [17] Fischer W., Keller F., and Perruchoud R. (1991), Interdependence between anode net consumption and pot design, pot operating parameters and anode properties, Light Metals 1991, ed. by E. Rooy (TMS,1991), pp 681-686

- [18] Fischer J. C. (2012), The Impact of Anode Quality of Smelter Performance, 27th International Aluminium Conference, Metal Bulletin Events, Moscow, Sept. 14, 2012
- [19] Edwards L. (2015), The History and Future Challenges of Calcined Petroleum Coke Production and Use in Aluminum Smelting, JOM, Vol 67, No.2, DOI: 10.1007/s11837-014-1248-9
- [20] Anti-Dumping Commission (2016), Analysis of Steel and Aluminium Markets Report to the Commissioner of the Anti-Dumping Commission, August, 2016 [www.adcommission.gov.au/adsystem/referencematerial/Documents/Steel.Aluminium.Report.31.August.2016.pdf](http://www.adcommission.gov.au/adsystem/referencematerial/Documents/Steel.Aluminium.Report.31.August.2016.pdf) - Accessed on August 22, 2024
- [21] Zhang S., Dong W. (2018), The Current Status and Development Trend of the Prebaked Anode Market in China, Light Metals 2018, ed. by O. Martin (Cham, Springer, 2018), pp. 1315-1320
- [22] World Oil Outlook 2040 (2017), Organization of the Petroleum Exporting Countries (OPEC), ISBN 978-3-9003936-4-4, [www.opec.org/opec\\_web/flipbook/W002017/W002017/assets/common/downloads/W00%202017.pdf](http://www.opec.org/opec_web/flipbook/W002017/W002017/assets/common/downloads/W00%202017.pdf) - Accessed on August 24, 2024
- [23] World Oil Outlook 2045 (2023), Organization of the Petroleum Exporting Countries (OPEC), <https://woo.opec.org/index.php> - Accessed on August 24, 2024
- [24] Baron J. T., McKinney S.A., Wombles R. H. (2009), Coal tar pitch - past, present and future, Light Metals 2009, ed. by G. Bearne (Warrendale, TMS, 2009), pp. 935-939
- [25] Mannweiler U., Fischer W.K., Perruchoud R. (2009), Carbon products: a major concern to aluminum smelters, Light Metals 2009 ed. by G. Bearne (Warrendale, TMS, 2009), pp. 909-912
- [26] Bartholomew K. D. (2013), Changes in global refining and its impact on anode quality petroleum coke, Light Metals 2013 ed. by B. A. Sadler (Cham, Springer, 2013), pp. 15-20
- [27] Rohani A., Sharifi K., Golpasha R. (2014), Calcinations of Petroleum Coke, 3rd International Conference on Chemical, Ecology and Environmental Sciences, March 19-20, Abu Dhabi, pp. 118-120
- [28] Edwards L., Vogt F., Robinette M., Love R., Ross A., McClung M., Roush R.J., Morgan W. (2009), Use of shot coke as an anode raw material, Light Metals 2009 ed. by G. Bearne (Warrendale, TMS, 2009), pp. 985-990
- [29] Sinclair K. A., Sadler B. A. (2009), Which Strategy to use when Sampling Anodes for Coring and Analysis? - Start with how the data will be used, Light Metals 2009 ed. by G. Bearne (Warrendale, TMS, 2009), pp. 1037-1041.
- [30] McClung M., Ross J. A., Chovanec G. (2000), A Method To Determine The Optimal Baking Level Of Carbon Anodes, Light Metals 2000 ed. by R.D. Peterson (Warrendale, TMS, 2000), pp. 473-479

- [31] Piffer V., Miotto P., Kato C., Meier M., Perruchoud R., Sulger P. (2007), Process Optimization In Bake Furnace, Light Metals 2007 ed. by S. K. Das, W. Yin (Warrendale,TMS,2007), pp. 959-964
- [32] Khaji K., Qassemi M. Al (2016), The Role of Anode Manufacturing Processes in Net Carbon Consumption, Metals, vol 6, 128. doi:10.3390/met6060128
- [33] [https://www.outotec.com/company/media/news/2013/outotec-artstm-anode-and-rod-tracking-systemfor- primary-aluminum-smelters/](https://www.outotec.com/company/media/news/2013/outotec-artstm-anode-and-rod-tracking-systemfor-primary-aluminum-smelters/) - Accessed March 2nd 2020
- [34] Musallam Al-Marzooqi M., Riasat T., International patent application WO 2016/097 968, "Anode Rod Tracking System for Electrolysis Plant"
- [35] Sahu T. K., Jabri N. Al, Rabba S. A., Banjab A. J., Marzouqi A. Al, Marzouqi A. Al, Almarzooqi M., Ndjom B., Palavar P., Kaur R., Mofor E. (2017), Improving Net Carbon Consumption at EGA with the Support of Anode Tracking and Butt Analyser, Travaux 46, Proceedings of the 35th International ICSOBA Conference, Hamburg, Germany
- [36] Bearne G. P. (1999) , The development of aluminum reduction cell process control. Aluminum smelting. JOM 51(5):16-22.
- [37] Kourti, T. MacGregor J.F (1995), Process analysis, monitoring and diagnostics, using multivariate projections methods. Chemometrics and Intelligent Laboratory Systems, 28:3-21.
- [38] Kourti, T. (2002), Process analysis and abnormal situation detection: from theory to practice, IEEE Control Systems 22(5):10-25
- [39] Kourti T. (2005), Application of latent variable methods to process control and multivariate statistical process control in industry. International Journal of adaptive control and signal processing 19(4):213-246
- [40] Malahanobis P. C. (1936), On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India, 2(1):49–55
- [41] Dunn K. (2023), Process Improvement Using Data, <https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/algorithms-to-calculate-build-pca-models> Accessed Nov 12 2024
- [42] Duchesne C, Kourti T, MacGregor J. F. (2002) Multivariate SPC for startups and grade transistions, AIChE Journal, 48(12):2890-2901
- [43] Miletic I, Quinn S., Dudzie M, Vaculik V, Champagne M. (2004), An industrial perspective on implementing on-line applications of multivariate statistics, Journal of Process Control, 14:821-836
- [44] MacGregor J.F., Yu H., Munoz S.G., Flores-Cerrillo J. (2005), Data-based latent variable methods for process analysis, monitoring and control. Computers and Chemical Engineering, 29:1217-1223

- [45] Tessier J., Duchesne C., Tarcy G.P., Gauthier C., Dufour G. (2008) Analysis of a potroom performance drift, from a multivariate point of view, *Light Metals 2008* ed. by D. H. DeYoung (Warrendale,TMS,2008), pp. 319-324.
- [46] Tessier J., Duchesne C., Tarcy G.P., Gauthier C., Dufour G.(2012) Multivariate analysis and monitoring of the performance of aluminum reduction cells, *Industrial & Engineering Chemistry Research* 51(3):1311-1323.
- [47] Tessier J., Duchesne C., Tarcy G.P., Gauthier C., Dufour G.(2010) Increasing potlife of the Hall-Héroult Reduction cells through multivariate on-line monitoring of preheating, start-up and early operation *Metallurgical and materials transactions B*, 41(3):612-624
- [48] J. Tessier, T.G. Zwirz, G.P. Tarcy, R.A Manzini (2009) Multivariate statistical process monitoring of reduction cells, *Light Metals 2009* ed. by G. Bearne (Warrendale,TMS,2009), pp 305-310
- [49] Majid N. A. A., Taylor M. P., Chen J. J.J, StamM. A., Mulder A., Young B. R. (2011) Aluminium process fault detection by Multiway Principal Component Analysis. *Control Engineering Practice* 19:367-379
- [50] Majid N. A. A., Taylor M. P., Chen J. J.J,, Young B. R. (2011) Multivariate statistical monitoring of the aluminium smelting process. *Computers and Chemical Engineering* 35:2457-2468
- [51] Tessier J., Duchesne C., Gauthier C, Dufour G. , Estimation of alumina content of anode cover material using multivariate image analysis technique. *Chemical Engineering Science* 63:1370-1380
- [52] Tessier J., Duchesne C., Gauthier C, Dufour G. (2007) Anode cover material estimation using image analysis in primary aluminum production, *Preprints of the 12th IFAC Symposium on Automation in Mining, Mineral and Metal Processing*, Quebec City, QC, Canada, August, 2007:483-489.
- [53] Tessier J., Duchesne C, Gauthier C, Dufour G. (2008) Image analysis for the estimation of anode cover material composition, *Light Metals 2008* ed. by D. H. DeYoung (Warrendale,TMS,2008), pp.293-298
- [54] Eigenvector Research, Inc., [eigenvector.com](http://eigenvector.com)
- [55] Wang X., Hosler B., Tarcy G. (2011) *Alcoa STARprobeTM*, *Light Metals 2011* ed. by S. J. Lindsay (Cham,Springer,2011), pp. 483-489
- [56] Kevin Dunn, *Process Improvement Using Data*, 2016
- [57] Wold S., Esbensen K., Geladi P. (1987), Principal component analysis. *Chemometrics and intelligent laboratory systems* 2:37-52
- [58] Wold S., Sjöström M., Eriksson L. (2001), PLS-Regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58:109-130

- [59] Geladi P., Kowalski B.R. (1986) , Partial least squares regression: A tutorial. *Analytica Chimica Acta* 186:1-17
- [60] Höskuldsson A. (1988) , PLS regression methods. *Journal of Chemometrics* 2:211-228
- [61] Nomikos P., MacGregor J. F., (1995), Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37:41-59
- [62] Dion L., Kiss L. I., Lavoie D., Arvisais J.-P. (2014) Developing a New Process Indicator Based on the Relationship Between an Electrolysis Cell Impurity Balance and its Incidents. *Light Metals 2014* ed. by J. Grandfield (New Jersey, Wiley, 2014), pp. 829-834
- [63] Westerhuis J.A., Gurden S.P, Smilde A.K. (2000), Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and intelligent laboratory systems*, 51:95-114
- [64] Warren E. H. (2017), Principles of aluminum electrolysis, *Essential Readings in Light Metals 2017* ed by G. Bearne, M. Dupuis, G. Tarcy (Cham, Springer, 2017) Vol. 2, pp. 3-11
- [65] Martel A. (2018), Spike Detection Using Advanced Analytics and Data Analysis, *Light Metals*, ed. by O. Martin (Cham, Springer, 2018), pp 485-490
- [66] Dechent M., Taylor M. P., Meier R., Tiedemann L., Meier M., Friedrich B. (2021), Carbon Dust - Its Short-Term Influence on Potroom Operations During Anode Change, *Light Metals 2021* ed. by Linus Perander (Cham, Springer, 2021), pp. 382-392
- [67] Tabereaux A. T. (2022), The Survivability of Aluminum Potlines After Lengthy Electrical Power Outages, *Light Metals 2022* ed. by D. Eskin (Cham, Springer, 2022), pp. 448-457
- [68] Majid N. A. A., Taylor M. P., Chen J. J.J, Wei Yu, Young B. R. (2012) Diagnosing faults in aluminum processing by using multivariate statistical approaches, *Journal of Materials Science*, Vol 47, pp 1268
- [69] Yao Y. (2017), Process monitoring, modelling and fault diagnosis in aluminum reduction cells, PhD Thesis, 2017, p. 143. <https://doi.org/10.26190/unsworks/3102>
- [70] LaJambe D., Poulin É., Duchesne C., Tessier J. (2020), Anodic Incident Detection through Multivariate Analysis of Individual anode Current Sensors, *Light Metals 2020* ed. by A Tomsett, (Cham, Springer, 2020), pp 535-542
- [71] Qiang K., Li J., Zhang J., Li J., Ran L., Zhang H. (2024), A Method for Anode Effect Prediction in Aluminum Electrolysis Cells Based on Multi-scale Time Series Modeling, *Light Metals 2024* ed. by S. Wagstaff (Cham, Springer, 2024), pp. 436-444
- [72] Manolescu P., Duchesne C., Tessier J., Saevarsdottir G. (2018), On the use of Multivariate Statistical Methods to Detect, Diagnose and Mitigate Abnormal Events in Aluminium Smelters, *Light Metals 2018* ed. by O. Martin (Cham, Springer, 2018), pp. 475-483.
- [73] Kresta J. V., Macgregor J. F., Marlin T. E. (1991), Multivariate statistical monitoring of process operating performance, *The Canadian journal of chemical engineering*, vol 69 (1), pp 35-47

- [74] Mastrangelo C. M., Runger G. C., Montgomery D. C. (1996), Statistical Process Monitoring With Principal Components, *Quality and Reliability Engineering International*, vol 12, pp 203 [https://doi.org/10.1002/\(SICI\)1099-1638\(199605\)12:3<203::AID-QRE12>3.0.CO;2-B](https://doi.org/10.1002/(SICI)1099-1638(199605)12:3<203::AID-QRE12>3.0.CO;2-B)
- [75] Martin E.B., Morris A.J. (1996) An overview of multivariate statistical process control in continuous and batch process performance monitoring, *Transactions of the Institute of Measurement and Control*, vol 18 (1), pp 51 <https://doi.org/10.1177/014233129601800107>
- [76] Kourti T. (2005), Application of latent variable methods to process control and multivariate statistical process control in industry, *International Journal of Adaptive Control and Signal Processing*, vol 19(4), pp 213
- [77] Wold S., Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models, *Technometrics*, Vol 20(4), pp 397
- [78] Jackson J. E., (1991) *A User's Guide to Principal Components* Wiley: New York, 1991, p. 19
- [79] Jackson J. E., MudHolkar G. S. (1979), *Technometrics*, Control Procedures for Residuals Associated with Principal Component Analysis, Vol 21 (3), pp 341
- [80] Tharwat A., Classification Assessment Methods, *Applied Computing and Informatics*, Vol 17 (1), pp 168 <https://doi.org/10.1016/j.aci.2018.08.003>.
- [81] <https://aluminium.ca/en/planet-aluminium/canada> - Accessed: Apr. 30 2020.
- [82] Dunn K. (2016), *Process Improvement Using Data*, 6.7 <https://learnche.org/pid/> - Accessed: March 10 2020
- [83] Chong I.G., Jun C.H. (2005), Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, pp 103
- [84] Tabereaux A.T. (1996), The Role of Sodium in Aluminum Electrolysis: A Possible Indicator of Cell Performance, *Light Metals 1996* ed by W. Hale (Warrendale,TMS,1996), pp 319-326
- [85] Wijayarathne H., Hyland M., Taylor M., Grama A., Groutso T. (2011) Effects of composition and granulometry on thermal conductivity of anode cover materials, *Light Metals 2011* ed. by S. J. Lindsay (Cham,Springer,2011), pp. 399-404
- [86] Zhang Q., Taylor M., Chen J., Cotton D., Groutzo T., Yang X. (2013), Composition and thermal analysis of crust formed from industrial anode cover, *Light Metals 2013* ed. by B. A. Sadler (Cham,Springer,2013), pp 675-680
- [87] Feng-qin L., Ye-xiang L., Mannweiler U., Perruchoud R. (2006) Effect of Coke Properties and its Blending Recipe on Performances of Carbon Anode for Aluminium Electrolysis, *J. Cent. South Univ. Technol.* 13(6), 647
- [88] Fellner P., Ambrová M., Híves J., Korenko M., Thonstad J. (2005) Chemical and Electrochemical Reaction of Sulphur Species in Cryolite Melts, *Light Metals 2005* ed. by H. Kvande, (Warrendale,TMS,2005), pp. 577-582

- [89] Batista J. S., Silveira B. I. (2008) Influence of the Sodium Content on the Reactivity of Carbon Anodes, *Materials Research*, 11(3), 387
- [90] Fischer W., Perruchoud R. (2016) Interdependence between properties of anode butts and quality of prebaked anodes, *Essential Readings in Light Metals, Electrode technology for aluminum production*, Vol 4, Springer, Cham, pp 267-270
- [91] Rhedey P. (2016) A Review of Factors Affecting Carbon Anode Consumption in the Electrolytic Production of Aluminum, *Essential Readings in Light Metals, Electrode technology for aluminum production*, Vol 4, Springer, Cham, 4:611-622
- [92] Paris A., Duchesne C., Poulin E., Lauzon-Gauthier J. (2020) Development of a Soft Sensor for Detecting Overpitched Green Anodes, *Light Metals 2020*, ed. by A Tomsett, (Cham, Springer, 2020), pp. 1176-1182
- [93] <https://www.nrcan.gc.ca/our-natural-resources/minerals-mining/aluminum-facts/20510>  
Accessed: May 16, 2022
- [94] Saevarsdottir G., Kvande H., Welch B. (2020) Aluminum Production in the Times of Climate Change: The Global Challenge to Reduce the Carbon Footprint and Prevent Carbon Leakage, *JOM* 72:296-309
- [95] Wilkening S., Reny P., Murphy B. (2005) Anode Cover Material and Bath Level Control, *Light Metals* ed. by Halvor Kvande, (Warrendale, TMS, 2005), pp. 367-372
- [96] Mahieu P., Genin X., Bouché C., Brismalein D., Pédroli H. (2019) Carbon Block Tracking Package Based on Vision Technology, *Light Metals 2019* ed by C. Chesonis (Cham, Springer, 2019), pp. 1221-1228
- [97] Manolescu P., Duchesne C., Tessier J., Saevarsdottir G. (2020) Estimation of Net Carbon Consumption in Aluminum Electrolysis Using Multivariate Analysis, *Journal of Sustainable Metallurgy* doi:10.1007/s40831-020-00310-6
- [98] Ziegler D.P., Secasan J. (2016) Methods for determining green electrode electrical resistivity and methods for making electrodes. US Patent 9:416-458. 16 August 2016
- [99] Lauzon-Gauthier J., Secasan J. (2021) Development and Applications of the Four Point Probe (4PP) Electrical Resistivity Measurements for Anode Process Optimization, *Light Metals 2021* ed. by Linus Perander (Cham, Springer, 2021), pp. 951-956
- [100] Westerhuis J.A., Coenegracht PMJ (1997) Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares, *Journal of Chemometrics* 11:379-392.
- [101] Lauzon-Gauthier J., Manolescu P., Duchesne C. (2018) The Sequential Multi-block PLS algorithm (SMB-PLS): Comparison of performance and interpretability, *Chemometrics and Intelligent Laboratory Systems* 180:72-83
- [102] De Araujo Costa Rodrigues D., Non-Destructive Evaluation of Baked Carbon Anodes for Process & Quality Control using Modal & Acousto-Ultrasonic Testing, Ph.D. Thesis, Université Laval, 2023

- [103] Paris A., Duchesne C., Poulin E., Adjusting plant operating conditions to widen multivariate specification regions for incoming raw materials – An optimization framework, *Chemometrics and Intelligent Laboratory Systems*, Vol 244, 2024
- [104] Paris A., Duchesne C., Poulin E., Defining a three-zone multivariate specification region for incoming raw materials, *IFAC-PapersOnLine*, 2024, Vol 58, Issue 14 <https://doi.org/10.1016/j.ifacol.2024.08.436>
- [105] Stocchero M., De Nardi M., Scarpa B. (2021) PLS for classification, *Chemometrics and Intelligent Laboratory Systems*, 216 <https://doi.org/10.1016/j.chemolab.2021.104374>

## Appendix A Variable names for material in Chapter 2

Table A.1: Variable names for figures 26 to 31

| Abbreviation          | Variable Name                                          |
|-----------------------|--------------------------------------------------------|
| AE_tot                | Number of Anode Effects                                |
| AE_min                | Length of the anode effects in minutes                 |
| Variable abbreviation | Variable full name                                     |
| ActualFeed            | Actual alumina feed rate                               |
| AFD                   | Anode Fault Detection                                  |
| Age                   | Cell age                                               |
| AlF3add_real          | Actual alumina fluoride feed shot additions            |
| AlF3add_plan          | Planned alumina fluoride feed shot addition            |
| BaseRes               | Cell base resistance                                   |
| BathTgt               | Cell bath level target                                 |
| BathLevel             | Measured cell bath level                               |
| BubbleNoise           | Cell bubble noise                                      |
| BubbleNoise_lag1      | Cell bubble noise lagged by 1 day                      |
| DeltaMetalTgt         | Changes in metal target between two consecutive days   |
| DeltaMetalFromTgt     | Difference between measured metal level and its target |
| DeltaCu_ppm           | Difference between consecutive Copper ppm measurements |
| DeltaBathFromTgt      | Difference between measured bath level and its target  |
| FeedShots             | Total daily alumina feed shots to the cell             |
| FeedIndex_lag1        | Total alumina feed rate modifier lagged by 1 day       |
| FeedIndex             | Total alumina feed rate modifier                       |
| FeedDecision          | Total alumina feed decisions                           |
| Ga_ppm                | Gallium ppm in metal sample                            |
| Ga_ppm_lag1           | Gallium ppm in metal sample lagged by 1 day            |
| IpcMod                | Cell resistance modifier                               |
| KickingAnodes         | Cell kicking anodes (shorts)                           |
| MetalTapSched         | Metal amount scheduled to be tapped                    |
| MetalLevel            | Cell metal level                                       |
| metal_noise           | Cell metal noise                                       |
| Noise                 | Cell noise                                             |
| Noise_lag1            | Cell noise lagged by 1 day                             |
| NoiseMod              | Cell resistance modifier due to noise                  |
| P_ppm                 | Phosphorus ppm in metal sample                         |
| ResistAct_lag1        | Cell actual resistance lagged by 1 day                 |
| ResModTotal           | Cell total resistance modifier                         |
| StructMod             | Manual resistance modifier set on the cell             |
| Spht                  | Cell superheat                                         |
| SiConsumed            | Silicon consumed from the side wall                    |
| Si_ppm                | Silicon ppm in metal sample                            |

Table A.1: (continued)

| Abbreviation | Variable Name                                |
|--------------|----------------------------------------------|
| Trk Time     | Alumina feed track time                      |
| Ti_ppm       | Titanium ppm in metal sample                 |
| TempMod      | Cell Temperature resistance modifier         |
| Temp         | Cell temperature                             |
| Volts        | Cell voltage                                 |
| V_ppm_lag1   | Vanadium ppm in metal sample lagged by 1 day |
| V_ppm        | Vanadium ppm in metal sample                 |
| Waves        | Waves                                        |
| xsAlF3       | Cell measured excess aluminum fluoride       |

## Appendix B    Supplementary material for Chapter 2

### In reference to section: Ranking system and performance metrics

The overall model performance (all observations) is presented in Table B.1

Table B.1: Overall model performance. Note that the metrics were abbreviated to better fit the page formatting. Thus, Obs = Observations, Prec = Precision, Acc = Accuracy, Sens = Sensitivity, Spec = Specificity

| TP  | TN | FP | FN | Obs.<br>Count | Prec. | Acc.  | Miss  | Sens. | Spec. | Fall<br>Out |
|-----|----|----|----|---------------|-------|-------|-------|-------|-------|-------------|
| 173 | 90 | 11 | 57 | 331           | 0.940 | 0.795 | 0.248 | 0.752 | 0.891 | 0.108       |

The model performance was calculated for the bins shown in Figure 24 and presented in table B.2. The Precision and Accuracy are the same, up to and including bin (50,60]. That is because TN and FN are both 0 so the Precision, equation (27), and Accuracy, equation (28), calculations become identical. It is seen that the accuracy decreases as the rank increases, however this is based on a smaller sample size. The main takeaway, from presenting the data this way, is that the problematic cells get flagged consistently.

Table B.2: Model performance metrics per 10-cell bins. Note that the metrics were abbreviated to better fit the page formatting. Thus, Obs = Observations, Prec = Precision, Acc = Accuracy, Sens = Sensitivity, Spec = Specificity

| Bin       | TP | TN | FP | FN | Obs | Prec. | Acc.   | Miss | Sens. | Spec. | Fall Out |
|-----------|----|----|----|----|-----|-------|--------|------|-------|-------|----------|
| (0,10]    | 50 | 0  | 6  | 0  | 56  | 0.893 | 0.893  | 0    | 1     | 0     | 1        |
| (10,20]   | 43 | 0  | 2  | 0  | 45  | 0.956 | 0.956  | 0    | 1     | 0     | 1        |
| (20,30]   | 25 | 0  | 1  | 0  | 26  | 0.962 | 0.962  | 0    | 1     | 0     | 1        |
| (30,40]   | 24 | 0  | 1  | 0  | 25  | 0.96  | 0.96   | 0    | 1     | 0     | 1        |
| (40,50]   | 20 | 0  | 1  | 0  | 21  | 0.952 | 0.952  | 0    | 1     | 0     | 1        |
| (50,60]   | 11 | 0  | 0  | 0  | 11  | 1     | 1      | 0    | 1     | NA    | NA       |
| (60,70]   | 0  | 9  | 0  | 5  | 14  | NA    | 0.643  | 1    | 0     | 1     | 0        |
| (70,80]   | 0  | 9  | 0  | 7  | 16  | NA    | 0.5625 | 1    | 0     | 1     | 0        |
| (80,90]   | 0  | 8  | 0  | 4  | 12  | NA    | 0.667  | 1    | 0     | 1     | 0        |
| (90,100]  | 0  | 2  | 0  | 3  | 5   | NA    | 0.4    | 1    | 0     | 1     | 0        |
| (100,110] | 0  | 11 | 0  | 4  | 15  | NA    | 0.733  | 1    | 0     | 1     | 0        |
| (110,120] | 0  | 9  | 0  | 4  | 13  | NA    | 0.692  | 1    | 0     | 1     | 0        |
| (120,130] | 0  | 3  | 0  | 2  | 5   | NA    | 0.6    | 1    | 0     | 1     | 0        |
| (130,140] | 0  | 2  | 0  | 2  | 4   | NA    | 0.5    | 1    | 0     | 1     | 0        |
| (140,150] | 0  | 3  | 0  | 3  | 6   | NA    | 0.5    | 1    | 0     | 1     | 0        |
| (150,160] | 0  | 5  | 0  | 0  | 5   | NA    | 1      | NA   | NA    | 1     | 0        |
| (160,170] | 0  | 4  | 0  | 3  | 7   | NA    | 0.571  | 1    | 0     | 1     | 0        |
| (170,180] | 0  | 2  | 0  | 1  | 3   | NA    | 0.667  | 1    | 0     | 1     | 0        |
| (180,190] | 0  | 5  | 0  | 2  | 7   | NA    | 0.714  | 1    | 0     | 1     | 0        |
| (190,200] | 0  | 1  | 0  | 2  | 3   | NA    | 0.333  | 1    | 0     | 1     | 0        |
| (200,210] | 0  | 1  | 0  | 3  | 4   | NA    | 0.25   | 1    | 0     | 1     | 0        |
| (210,220] | 0  | 2  | 0  | 3  | 5   | NA    | 0.4    | 1    | 0     | 1     | 0        |
| (220,230] | 0  | 4  | 0  | 1  | 5   | NA    | 0.8    | 1    | 0     | 1     | 0        |
| (230,240] | 0  | 2  | 0  | 0  | 2   | NA    | 1      | NA   | NA    | 1     | 0        |
| (240,250] | 0  | 1  | 0  | 3  | 4   | NA    | 0.25   | 1    | 0     | 1     | 0        |
| (250,260] | 0  | 0  | 0  | 1  | 1   | NA    | 0      | 1    | 0     | NA    | NA       |
| (260,270] | 0  | 2  | 0  | 0  | 2   | NA    | 1      | NA   | NA    | 1     | 0        |
| (270,280] | 0  | 1  | 0  | 1  | 2   | NA    | 0.5    | 1    | 0     | 1     | 0        |
| (280,290] | 0  | 0  | 0  | 1  | 1   | NA    | 0      | 1    | 0     | NA    | NA       |
| (290,300] | 0  | 0  | 0  | 1  | 1   | NA    | 0      | 1    | 0     | NA    | NA       |
| (300,310] | 0  | 2  | 0  | 1  | 3   | NA    | 0.666  | 1    | 0     | 1     | 0        |
| (310,320] | 0  | 2  | 0  | 0  | 2   | NA    | 1      | NA   | NA    | 1     | 0        |

## Appendix C      Supplementary material for Chapter 3

Within the each Monte-Carlo simulation, an MLR was fitted to the raw unit data also. The distribution of the coefficients is presented below, in Table C.1. The original formula coefficients are also shown for comparison.

Table C.1: Distribution of coefficients for MLR parameters obtained for the MLR vs the R&D Carbon formula coefficients

| <i>Parameter</i> | <i>1st Quartile</i> | <i>Median</i> | <i>3rd Quartile</i> | <i>R&amp;D Carbon values</i> |
|------------------|---------------------|---------------|---------------------|------------------------------|
| $\frac{1}{CE}$   | 318.108             | 324.578       | 330.236             | 334                          |
| $(BT - 960)$     | -0.060              | -0.003        | 0.052               | 1.2                          |
| $CRR$            | 0.809               | 1.028         | 1.231               | -1.7                         |
| $AP$             | -0.737              | -0.628        | -0.532              | 9.3                          |
| $TC$             | -2.826              | -0.003        | 2.488               | 8                            |
| $ARR$            | 0.220               | 0.288         | 0.357               | -1.5                         |

As in the scaled version of the MLR model, the signs of the coefficients are not consistent with the original formula. The magnitude of the variables are not in the same range either, but as in the scaled MLR model, the main driver is the CE since it is the only parameter with a p-value lower than 0.05. The discrepancy could, again, be explained by the fact that this is production data on an individual anode basis. The CE and NCC values are calculated for a relatively short period of time, i.e. an anode set cycle, and thus the data is noisy. Several anodes can have the same baked anode properties but different CE, bath temperatures and NCC. This shows that for calculating the NCC on an individual anode basis more information is needed.

## Appendix D Supplementary material for Chapter 4

### D.1 SMB-PLS algorithm pseudo-code

Table D.1: Pseudo-code for the SMB-PLS algorithm [101]

---

|                                                                                                                                            |                                                      |
|--------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------|
| For $b = 1, 2, \dots, B - 1$                                                                                                               |                                                      |
| 1. Set $\mathbf{u}$ to any column of $\mathbf{Y}$                                                                                          | % initialisation                                     |
| 2. Begin convergence loop                                                                                                                  |                                                      |
| 2.1. $\mathbf{w}_b = \mathbf{X}_b^T \mathbf{u} / (\mathbf{u}^T \mathbf{u})$                                                                | % compute $\mathbf{X}_b$ block weights               |
| 2.2. $\mathbf{w}_b = \mathbf{w}_b / \ \mathbf{w}_b\ $                                                                                      | % normalize weight vectors                           |
| 2.3. $\mathbf{t}_b = \mathbf{X}_b \mathbf{w}_b$                                                                                            | % compute $\mathbf{X}_b$ block scores                |
| For $k = 1, 2, \dots, B - b$                                                                                                               | % do for all subsequent blocks to $\mathbf{X}_b$     |
| 2.4. $\mathbf{X}_{b+k}^{corr} = \mathbf{t}_b (\mathbf{t}_b^T \mathbf{t}_b)^{-1} \mathbf{t}_b^T \mathbf{X}_{b+k}$                           | % extract correlated information with $\mathbf{X}_b$ |
| 2.5. $\mathbf{w}_{b+k}^{corr} = \mathbf{X}_{b+k}^{corr T} \mathbf{u} / (\mathbf{u}^T \mathbf{u})$                                          | % compute weights for subsequent blocks              |
| 2.6. $\mathbf{w}_{b+k}^{corr} = \mathbf{w}_{b+k}^{corr} / \ \mathbf{w}_{b+k}^{corr}\ $                                                     | % normalize weights for subsequent blocks            |
| 2.7. $\mathbf{t}_{b+k}^{corr} = \mathbf{X}_{b+k}^{corr} \mathbf{w}_{b+k}^{corr}$                                                           | % compute scores for subsequent blocks               |
| End                                                                                                                                        |                                                      |
| 2.8. $\mathbf{T} = [\mathbf{t}_b \mathbf{t}_{b+1}^{corr} \dots \mathbf{t}_B^{corr}]$                                                       | % concatenate block scores in $\mathbf{T}$           |
| 2.9. $\mathbf{w}_T = \mathbf{T}^T \mathbf{u} / (\mathbf{u}^T \mathbf{u})$                                                                  | % compute super weights                              |
| 2.10. $\mathbf{w}_T = \mathbf{w}_T / \ \mathbf{w}_T\ $                                                                                     | % normalize super weights                            |
| 2.11. $\mathbf{t}_T = \mathbf{T}^T \mathbf{w}_T / (\mathbf{w}_T^T \mathbf{w}_T)$                                                           | % compute super scores                               |
| 2.12. $\mathbf{q} = \mathbf{Y}^T \mathbf{t}_T / (\mathbf{t}_T^T \mathbf{t}_T)$                                                             | % compute $\mathbf{Y}$ loadings                      |
| 2.13. $\mathbf{u} = \mathbf{Y} \mathbf{q} / (\mathbf{q}^T \mathbf{q})$                                                                     | % compute $\mathbf{Y}$ scores                        |
| Loop until convergence on $\mathbf{t}_T$ or $\mathbf{u}$ . Go to step 3 when converged.                                                    |                                                      |
| 3. For $k = b, b + 1, \dots, B$                                                                                                            |                                                      |
| 3.1. $\mathbf{p}_k = \mathbf{X}_k^T \mathbf{t}_T / (\mathbf{t}_T^T \mathbf{t}_T)$                                                          | % compute $\mathbf{X}_k$ blocks loadings             |
| 3.2. $\mathbf{E}_k = \mathbf{X}_k - \mathbf{t}_T \mathbf{p}_k^T$                                                                           | % deflate $\mathbf{X}_k$ blocks                      |
| End                                                                                                                                        |                                                      |
| 4. $\mathbf{F} = \mathbf{Y} - \mathbf{t}_T \mathbf{q}^T$                                                                                   | % deflate $\mathbf{Y}$ data                          |
| 5. Store all vectors at the block and super levels as new columns in matrices.                                                             |                                                      |
| 6. To compute the next LV, replace $\mathbf{X}_k$ by $\mathbf{E}_k$ ( $k \geq b$ ) and $\mathbf{Y}$ by $\mathbf{F}$ and go back to step 1. |                                                      |
| 7. When the relevant information in block $\mathbf{X}_b$ is depleted, increment $b$ and start at step 1.                                   |                                                      |
| End                                                                                                                                        |                                                      |
| 8. For $b = B$ , fit a regular PLS model to $\mathbf{E}_B$ and $\mathbf{F}$ .                                                              |                                                      |

---

### D.2 PLS metrics

Within the PLS and SMB-PLS algorithms two metrics can be calculated. One is the Hotelling's  $T^2$  and it represents the distance from the origin to the projection of the point onto the latent variable space, shown in equation 49 [82]. That is to say, it is a measure of the variables' magnitude. It is the sum of the squared ratio of the score value  $t_{i,a}$  to the standard deviation  $s_a$  of the score vector, where  $i$  is the observation number, and  $a$  the component number of the model using  $A$  components.

$$T_i^2 = \sum_{a=1}^A \left( \frac{t_{i,a}}{s_a} \right)^2 \quad (49)$$

The second metric is the Squared Predicted Error (SPE) and it represents the distance from the model plane to the observation in high dimensional space, calculated by equation 50 [82]. In other words, it is related with the correlation structure of the variables for each observation, as compared to the model. To calculate it the residual vector  $\mathbf{e}$  between the actual and predicted observation is used, where  $i$  is the observation number and  $a$  is the component number out of the  $A$  components of the model.

$$SPE_i = \sqrt{\mathbf{e}_{i,A}^T \mathbf{e}_{i,A}} \quad (50)$$

### D.3 SMB-PLS Model Latent Variables

Table D.2: Number of SMB-PLS latent variables (LVs) for each data block and corresponding SMB-PLS latent variables representing that block within the model

| Block | Block Full Name                   | LVs per block | Corresponding SMB-PLS LVs |
|-------|-----------------------------------|---------------|---------------------------|
| G     | Green Anode Properties            | 3             | 1-3                       |
| B     | Baked Anode Properties            | 3             | 4-6                       |
| P1    | Anode position in pot and pot age | 2             | 7-8                       |
| P2    | Pot Resistance                    | 2             | 9-10                      |
| P3    | Alumina Feed                      | 2             | 11-12                     |
| P4    | AlF3 and Bath Feed                | 3             | 13-15                     |
| P5    | Anode Reactivity metrics          | 2             | 16-17                     |
| P6    | Instability metrics               | 2             | 18-19                     |
| P7    | ImpuritiesInMetal                 | 2             | 20-21                     |
| P8    | Anode Problems                    | 2             | 22-23                     |
| P9    | Tapped metal                      | 2             | 24-25                     |
| P10   | Current Efficiency                | 1             | 26                        |

### D.4 List of abbreviations for Figure 47 and Figure 48

Table D.3: Variable abbreviations for Figure 47

|                              |                                             |
|------------------------------|---------------------------------------------|
| <i>SumTotalFeedShots</i>     | Sum of the number of feeds                  |
| <i>avgTrackTime</i>          | Average daily track time                    |
| <i>stdevTrackTime</i>        | Standard deviation of daily track time      |
| <i>stdevBaseFeed</i>         | Standard deviation of daily base feed       |
| <i>stdevFeedDecisionsDay</i> | Standard deviation of daily feed decisions  |
| <i>avgDailyFeedShots</i>     | Average daily number of feeds               |
| <i>avgFeedDecisions</i>      | Average of daily feed decisions             |
| <i>avgActualFeed</i>         | Average of daily actual feed                |
| <i>stdevActualFeed</i>       | Standard deviation of daily actual feed     |
| <i>stdevDailyFeedShots</i>   | Standard deviation of daily number of feeds |
| <i>avgBaseFeed</i>           | Average of daily base feed                  |

Table D.4: Variable abbreviations for Figure 48

|                                   |                                                                  |
|-----------------------------------|------------------------------------------------------------------|
| <i>Height</i>                     | Green anode height                                               |
| <i>Weight</i>                     | Green anode weight                                               |
| <i>PitchQI</i>                    | Pitch quinoline insolubles                                       |
| <i>BSS</i>                        | Green anode bottom short side resistivity                        |
| <i>PredictiveBakedDensity_PVD</i> | Green anode predictive baked density                             |
| <i>PasteTemp</i>                  | Paste temperature                                                |
| <i>ApparentDensity</i>            | Green anode apparent density                                     |
| <i>DensityDry</i>                 | Green anode density dry                                          |
| <i>DLS</i>                        | Green anode delta long side resistivity difference               |
| <i>AD_PVD_difference</i>          | Difference between Apparent density and Predictive baked density |
| <i>TLS</i>                        | Green anode top long side resistivity                            |
| <i>VLS</i>                        | Green anode vertical long side resistivity                       |
| <i>PitchTemp</i>                  | Pitch temperature                                                |
| <i>BLS</i>                        | Green anode bottom long side resistivity                         |
| <i>DSS</i>                        | Green anode delta short side resistivity difference              |
| <i>HLS</i>                        | Green anode horizontal long side resistivity                     |