Application of Multivariate Data Analysis

in Biopharmaceutical Production

Volumes: 1

Submitted by

Elspeth Kathryn Ritchie

on

03 June 2016

for the qualification of

Engineering Doctorate

in **Biopharmaceutical Process Development**

Biopharmaceutical Bioprocessing Technology Centre,

School of Chemical Engineering and Advanced Materials,

University of Newcastle upon Tyne

Abstract

In 2004, the FDA launched the Process Analytical Technology (PAT) initiative to support product and process development. Even before this, the biologics manufacturing industry was working to implement PAT. While a strong focus of PAT is the implementation of new monitoring technologies, there is also a strong emphasis on the use of multivariate data analysis (MVDA). Effective implementation and integration of MVDA is of particular interest as it can be applied retroactively to historical datasets in addition to current datasets. However translation of academic research into industrial ways of working can be slowed or prevented by many obstacles, from proposed solutions being workable only by the original academic to a need to prove that time invested in developing MVDA models and methodologies will result in positive business impacts (e.g. reduction of costs or man hours).

The presented research applied MVDA techniques to datasets from three scales typically encountered during investigations of biologics manufacturing processes: a single product, dataset; a single product, multi-scale dataset; a multi-product, multi-scale, single platform dataset. These datasets were interrogated in multiple approaches and multiple objectives (e.g. indictors/causes of productivity variation, comparison of pH measurement technologies). Individual project outcomes culminated in the creation of a robust statistical toolbox.

The toolbox captures an array of MVDA techniques from PCA and PLS to decision trees employing *k*-NN. These are supported by frameworks and guidance for implementation based on interrogation aims encountered in a contract manufacturing environment. The presented frameworks ranged from extraction of indirectly captured information (Chapter 4) to meta-analytical strategies (Chapter 6). Software-based tools generated during research ranged from translation of high frequency online monitoring data as robust summary statistics with intuitive meaning (Appendix A) to tools enabling potential reduction in confounding underlying variation in dataset structures through the use of alternative progression variables (Chapter 5). Each tool was designed to fit into current and future planned ways of working at the sponsor company.

The presented research demonstrates a range of investigation aims and challenges encountered in a contract manufacturing organisation with demonstrated benefits from ease of integration into normal work process flows and savings in time and human resources.

Dedication

For my parents, Douglas and Karen Ritchie.

I wouldn't be who I am today without your love and care.

Thank you.

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

- T.S. Eliot, "The Rock" (1934)

Acknowledgements

The research presented in this thesis was made possible through the support of the Engineering and Physical Sciences Research Council and sponsorship from Lonza Biologics in Slough.

Nothing on these pages would have been possible without the guidance of Dr. Colin Jaques and Dr. Andy Racher, my industry supervisors, and Professor Elaine B. Martin, my academic mentor, or the help of the Lonza Cell Culture Process Development scientists, who not only me loose with their data and files but also put up with me presenting the results.

Table of Contents

Table of Figures

Table of Tables

Chapter 1. Introduction

The only useful function of a statistician is to make predictions and thus provide a basis for action. — William Edwards Deming

In a 1994 press release, Andrew J. Guarriello, then chief operating officer of AT&T Power Systems, stated that *"the roots of today's Total Quality Management can be traced to the work of three AT&T scientists and quality pioneers--Walter Shewhart, W. Edwards Deming, and Joseph Juran."* [1] In 1924, Shewhart presented a single page document at a meeting at Western Electric, CA, USA. One-third of the page was given over to what would now be called a *Shewhart control chart*. This document is often seen as the start of statistical process control (SPC) as a separate field of study blending engineering, quality control, and statistics. Further developments by Shewhart created the basis for his 1931 book *Economic Control of Quality of Manufactured Product*. Deming and Juran became interested in Shewhart's work and promoted the use of SPC, in particular Shewhart's Plan-Do-Check-Act cycle.

SPC initially focussed on univariate analysis. However univariate analyses do not allow interactions between variables to be easily identified or tested, as observed in Figure 1 where readings for a sample appear normal when considered in a univariate manner but is clearly unusual when considered in a multivariate manner. Identification of such multivariable interactions and evaluation of the impact of those interactions can be used to improve process robustness [2], efficiency [3], and safety [4]. The tools of SPC have expanded to include a wide array of multivariate techniques, including decision trees [5], principal component analysis (PCA) [6,7], partial least squares (PLS) [6,7], artificial neural networks [8,9], self-organising maps [10,11], structural equation modelling [12], and even multivariate adaptations of Shewhart's original control charts among others.

Multivariate data analysis (MVDA) techniques had been suggested before Shewhart's work, however the computational power necessary to complete the associated equations limited their use when relying on manual computation. As greater computational power became available through the development of computers, ever more intensive multivariate techniques could be applied to ever larger datasets. In recent years, the speed at which analyses can be completed has started to become less of an obstacle than the volume, quality, and diversity of data able to be brought together from analysis.

The variety of data sources and MVDA techniques available for process understanding is matched by the range of industries they have been used in and aims achieved. A multiple

case study review by Miletic et al. [13] captured four adaptations of PCA and PLS algorithms in various degrees of complexity in execution. The simplest was the use of PCA-based control charts to identify when a specific type of fault was about to occur for a continuous slab caster, resulting in a 50% reduction of faults over a 6 year period for one caster, a reduction of 4 faults for a second caster, and increased operator confidence at higher levels of production. This methodology was adapted for a second area of operation to track batch evolution of a sulphite pulp digester in real-time.

More complex methodologies were required in two further examples. An adaptive PLS-based automatic control system for the control of chemical reagents in a desulphurisation process for a liquid metal required extensive supervision during tuning and initial operating period. Investment into this more sophisticated control system yielded a 50% reduction in the root mean square error of sulphur content of the final output. Additional benefits were reductions in reagent use, most notably a 70% reduction in the addition rate for a second reagent and a 25.5% reduction in purchased reagent quantities.

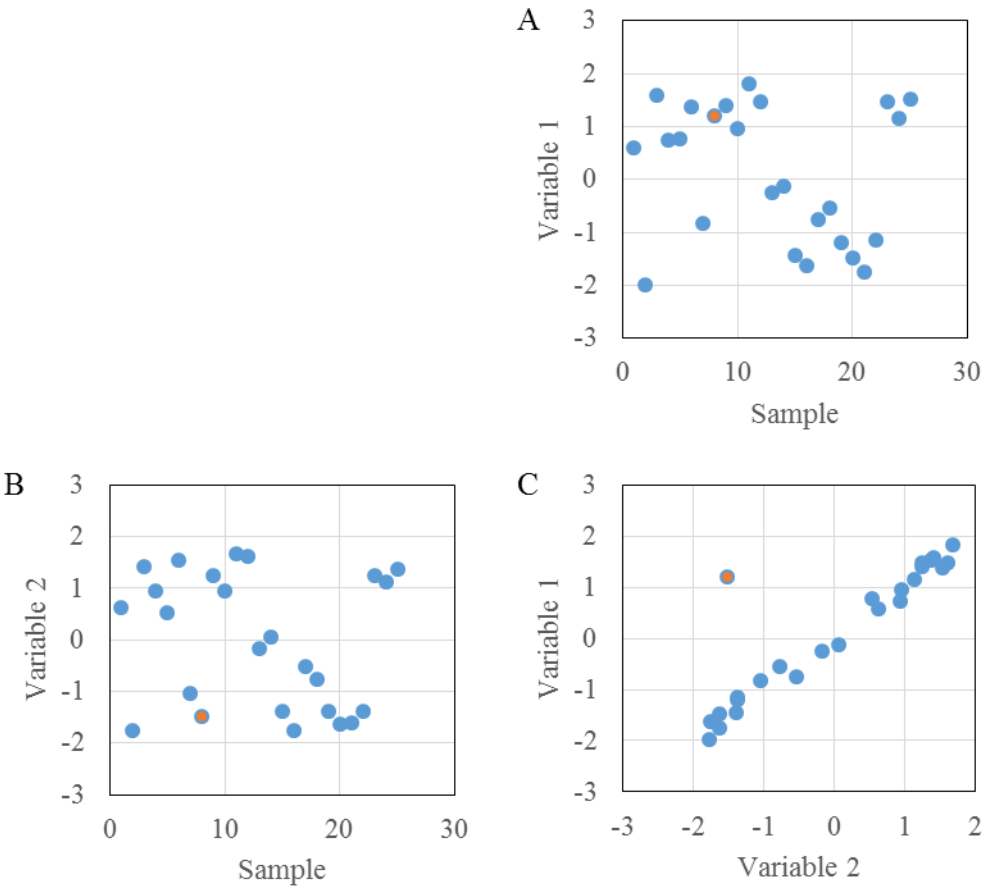


Figure 1. Variable 1 and Variable 2 recorded for 25 samples. The highlighted sample appears to show normal behaviour when Variable 1 and Variable 2 were viewed separately (A and B), however multivariate analysis (C) shows there is an unusual interaction for the highlighted sample.

The final example concerned a paperboard manufacturing process. The original aim was to identify process faults through the application of PCA, similar to the case studies detecting faults in continuous caster machines and desulphurisation processes. However, multiple grades of paperboard were manufactured using the same machinery, which caused severe confounding in the dataset. Due to the large number of paper grades produced, it was not economically feasible to create and maintain models for each grade of paperboard. Furthermore, this view would be strongly focussed on the final product and not the underlying process or machinery. Instead, a partial least squares-discriminant analysis (PLS-DA) model was used to identify differences in paper grades. The PLS-DA model residuals were then used to create a PCA model for fault detection, as was originally intended. The final result was a 60% reduction in processing variability.

Beyond the academic novelty of re-purposing one model's residual to create a second model, the final case study demonstrated a more holistic approach to process data analysis than is typically observed in more academically driven papers. Specifically, the cost of maintained use and appropriateness for the intended area of application (here, a flexible, multi-product manufacturing platform) is in direct contrast to the more traditional academic emphasis on a single, fixed analysis workflow developed for a specific use with at best a limited view towards adoption by industries.

In each of Miletic et al.'s case studies [13], MVDA was applied retroactively to datasets and incorporated into pre-existing data flows. Ideally, multivariate statistics can be employed from the very start of development to plan a deliberate experimental design space. A commonly employed technique is Design of Experiments (DOE), outlined in Ronald A. Fisher's 1935 book of the same name [14] or a derivative thereof. In DOE, a scientist selects multiple variables of interest to be tested multiple levels as a series of experiments designed to test multivariate interactions. This methodology can be used to create a Design Space for a process, e.g. a multidimensional/multivariate combination of input variables and process parameters for which quality is assured through product and process understanding [15].

SPC and MVDA grew to become fundamental tools in a variety of manufacturing industries during the 20th century, at the time they were relatively unused by the bioprocessing industries, in particular the biopharmaceutical and biologics industries.

Biologics are a category of medical treatments and therapies derived from living organisms. The US Food and Drug Administration (FDA) uses the term biological

product, which is defined as *"a virus, therapeutic serum, toxin, antitoxin, vaccine, blood, blood component or derivative, allergenic product, or analogous product, or arsphenamine or derivative of arsphenamine (or any other trivalent organic arsenic compound), applicable to the prevention, treatment, or cure of a disease or condition of human beings."* [16] The European Medicines Agency (EMA) uses the term biological medicinal product, defined as *"a medicinal product whose active substance is made by or derived from a living organism."* [17]

According to a market survey published in 2014, there were 230 approved biologics on the market in 2012 with global sales of US$124.9 billion [18]. While a wide variety of products are allowed under the term "biologic", the majority of biologics are recombinant proteins and monoclonal antibodies (Figure 2) with approximately 50% of sales attributable to 10 block buster drugs (Figure 3).

The use of established cell lines allows biologics production processes to benefit from the use of platform processes and platform technologies. A platform can be defined as *"a set of stable components that supports variety and evolvability in a system by constraining the linkages among the other components."* [19]. A definition more specific to biopharmaceutical manufacturing platforms is *"[t]he approach of developing a production strategy for a new drug starting from manufacturing processes similar to those used by the same applicant to manufacture other drugs of the same type (e.g., as in the production of monoclonal antibodies using predefined host cell, cell culture, and purification processes, for which there already exists considerable experience)"* [15].

The use of process platforms allow a contract manufacturing organisation (CMO) to benefit from a wide variety of savings both within the company itself and through interactions with other companies, e.g. reductions in development times or improved resource use efficiencies [20]. Within the company, the use of platform process allow for improved efficiency in scale-up of projects and site-transfer/technology transfer of projects. These improved efficiencies can also be achieved when transferring a technology or product with other, external companies.

In biologics, two commonly encountered platforms are expression platforms (host cell line and expression system) and process platforms (including culture operating conditions and feed strategies).

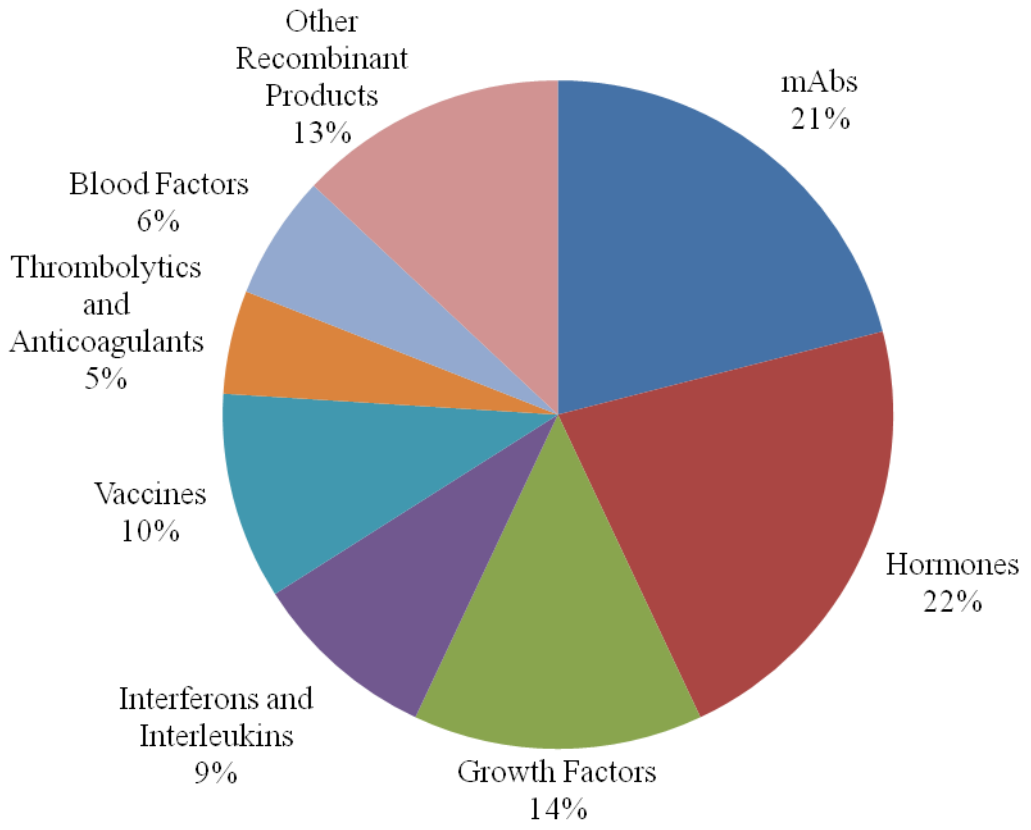


Figure 2. Distribution of 230 approved biologics by compound class [18]. Monoclonal antibodies make up 21% with other recombinant products making up a further 13% of approvals.

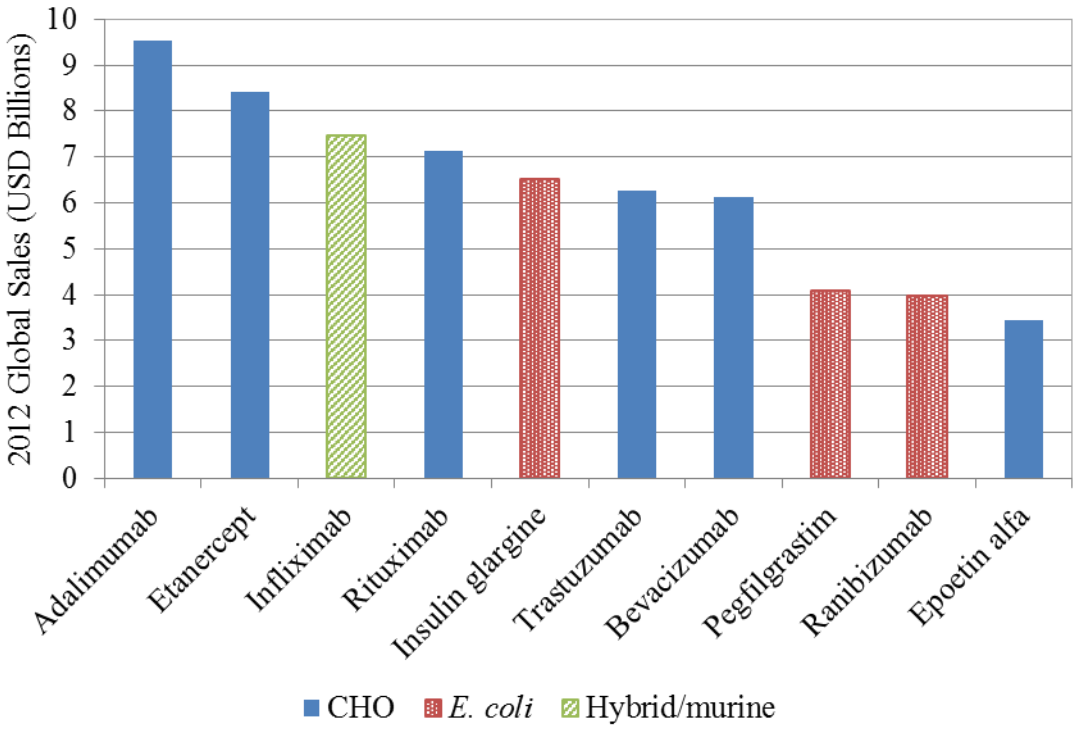


Figure 3. Top ten biologics which represent approximately 50% of all global sales of biologics in 2012 [18]. Six were produced using a CHO cell line, three using an *E. coli* cell line, and one using a hybrid/murine cell line.

In a typical protein or monoclonal antibody (mAb) production process, the genetic code for the product of interest is inserted into the DNA of a host cell. The cell then produces the protein as part of its normal metabolism. Although mammalian cell lines have been a part of biopharmaceutical development since the 1900s [21], wide spread usage of mammalian cell lines in protein or mAb production was limited by numerous obstacles and perceived obstacles, from shear sensitivity of mammalian cells in suspensions [22] or complex media requirements [23,24]. Instead, large scale processes for much of the 20th century were reliant on non-mammalian cell lines [25].

The first human recombinant protein licensed was the recombinant insulin Humulin (Genentech) in 1982, produced using the bacteria *Escherica coli*. *E. coli* was and continues to be widely used as *E. coli* typically grows quickly and robustly in large-scale manufacture [26,27]. However despite greater robustness and lower costs when compared to mammalian cell lines, fundamental drawbacks regarding product safety and efficacy can exist when using bacterial hosts.

Host cell lines perform post-translational modifications to the expressed protein or mAb, and this affects which host cell line can be safely used. A key example is glycosylation, the process by which oligosaccharides are attached to asparagine (*N*-linked), serine (*O*-linked), and threonine (*O*-linked) side chains [28]. Oligosaccharides are often essential for recognition of the protein by the patient's immune system. Post-translational modifications or lack thereof can cause a variety of undesired outcomes from decreased efficacy [28–30] to side effects caused by patient immune systems attacking a drug as an infection [31–35]. For these reasons, the ability to glycosylate proteins in a manner similar to glycosylation by human cells can lead a manufacturer to select a mammalian cell line, which do glycosylate proteins, over a bacterial cell line, which do not glycosylate proteins.

Two commonly used mammalian cell lines are Chinese hamster ovary (CHO) and murine myeloma. Of the top ten biologics in 2012, representing approximately 50% of all global sales of biologics for that year, six were produced using a CHO cell line and one was produced using a hybrid/murine cell line. [18]. In terms of new drug approvals, CHO and murine myeloma accounted for 31% and 11% of approved biologics in 2012 respectively [18]. Additional mammalian cells used include human, hybridomas, and baby hamster kidney (BHK) (Figure 4). Research continues be conducted to identify, isolate, and evaluate mammalian cell lines able to express protein and mAb products with

economically viable titres and with appropriate post-translational modifications as differences in glycoforms produced by a mammalian host cell and those produced by human cells can still lead to undesirable side effects. The murine-derived cetiumab (Erbitux®) was shown to trigger anaphylaxis in a subset of patients with pre-existing antibodies that attacked a sugar residue seen on products derived from CHO and murine hosts which is absent from post-translational modifications by human cells [36].

Of note are studies that demonstrated that several of the issues which had prevented wide-spread use of mammalian cell lines were simply perceived issues. In particular several studies stating that mammalian cell lines can survive the shear forces encountered in suspension cultures [37–39] stand in stark contrast to older studies [40,41] that indicated mammalian cells were not suited to suspension cultures.

When the genetic information for a mAb is introduced to the host cell, the vector typically contains additional genetic information. This additional information may serve to promote growth of transfected cells or to improve stable integration of the new genetic information into the host cell genome.

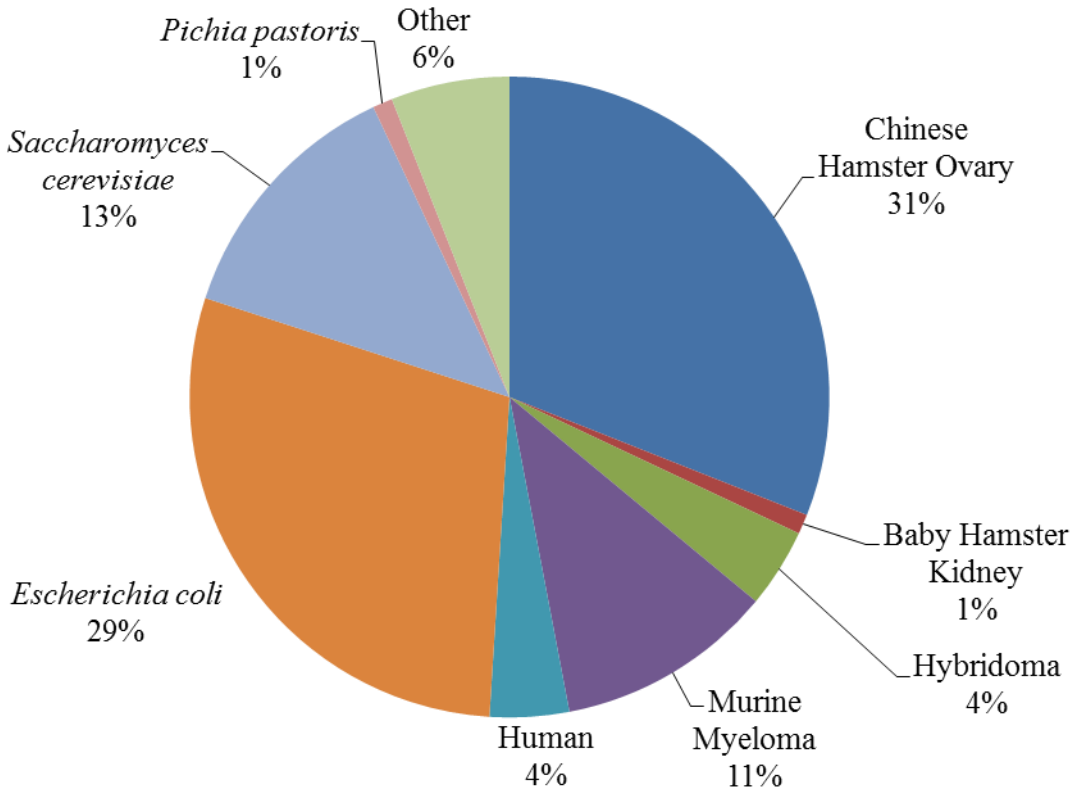


Figure 4. Distribution of host cells used in industrial biologics based on number of licensed biologics using cell line until 2012 [18].

A common addition to transfected vectors are genes that add some new metabolic activity, such as the ability to synthesise some metabolite A which is crucial to cell survival but would normally need to be made available in culture media. These genes allow for selective survival of cells based on whether the gene vector was successful transfected and integrated into the cell's DNA. Continuing the previous example, only successfully transfected cells would be able to survive in media deficient in metabolite A, assuming the cells were provided with the necessary precursors.

This complexity of interactions from the very beginning of the biologics manufacturing system, from drug discovery to cell line and selection system choice to expression, isolation, formulation, and storage through to the interaction of final drug with the patient makes biologics manufacturing an ideal industry for improvements through MVDA. Decision trees have been used to select local optimum operating conditions for cultures from tested conditions [42] and combined with the PLS algorithm for use in control decisions for cultures [43]. PCA has been used to monitor batch performances, to "finger print" media using spectral datasets, and identify genes of interest [44–46]. The field of chemometrics in particular has readily adopted MVDA to develop "electronic noses" for the detection, classification, and measurement of multiple chemicals by sensors [47].

Drivers for manufacturers to adopt MVDA range from improved safety of the end product, reductions in manufacturing costs, and reduced time to market. Reduced time to market may be achieved through techniques such as DOE decreasing development times. However following the release of the Process Analytical Technology (PAT) Initiative by the US Food and Drug Administration (FDA) in 2004 [48], time to market may also be reduced by using PAT to support submissions to regulators. Two PAT-supported biologics approved through the FDA's expedited process for breakthrough therapies were Genentech's Gazyva™ (obinutuzumab) [49] and Genentech's Perjeta™ (pertuzamab) [50].

The FDA's PAT Initiative guidelines were intended as a way for industry to meet three aims [48]:

1. Improve the scientific basis for establishing regulatory specifications.
2. Promote continuous improvement.
3. Improve manufacturing while maintaining or improving current product quality.

The main concept behind PAT was that *"quality cannot be tested into products; it should be built-in or should be by design."* [48] Three key areas are covered by the PAT framework: process understanding, analytical principles and tools, and strategies for implementation. As a flexible framework, PAT permits the use of many different tools and technologies to be used, from the introduction of brand new equipment such as spectral readers to the re-examination of existing historical databases with multivariate data analysis to knowledge management systems.

While the PAT Initiative is often spoken in terms of FDA documentation, the PAT Initiative is supported by many regulatory bodies around the world, including the European Medical Authority (EMA) and the Japanese Pharmaceuticals and Medical Devices Agency (PDMA). As part of global harmonisation efforts, the FDA, the EMA, and the PDMA co-operated as the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) to create joint guidance with the aim *"[t]o promote a mutual understanding of regional harmonisation initiatives in order to facilitate the harmonisation process related to ICH Guidelines regionally and globally, and to facilitate the capacity of drug regulatory authorities and industry to utilise them"* [51]. Three ICH documents particularly relevant to the presented work are ICH Q8 "Pharmaceutical Development", ICH Q9 "Quality Risk Management", and ICH Q10 "Pharmaceutical Quality System" [52].

In addition to PAT usage guidelines, these documents include guidance on the use of Quality by Design (QbD), a systematic approach to product/process development, including life-cycle management [53], and the use of Design Spaces to support regulatory decisions. Variation within the Design Space and related effects on the product quality are considered to be understood, whereas variation outwith the Design Space is considered a change requiring additional regulatory approval and could lead to destruction of the product. As an example, a process design space is created and tested for a pH range of 6.5 to 7.5 and temperature range 30°C to 35°C with the desired process conditions pH 7.0 and 32.5°C as the centre point. The process conditions pH 6.8 and 31°C are within that design space and the effects of this change considered understood with appropriate actions for based on these effects. However the process conditions pH 6.4 and 29°C are outwith the design and effects of this change on the product are untested and not considered understood.

Guidance from regulatory bodies can also be seen as an attempt to avoid repeating the failures to introduce MVDA and SPC experienced by other industries. When techniques such as Kaizan and Lean Design were first introduced to American manufacturing industries, in particular American automotive manufacturing, they were rarely as successful as hoped [54]. The poor results were often attributed to poor implementation due lack of understanding and appreciation for the tools. When the biopharmaceutical industry as a whole began implementing PAT tools and technologies in the 1990s and early 2000s, it was with awareness of known obstacles. Several such obstacles highlighted by Miletic et al. [13] are:

- Poor acceptance by shop-floor personnel
- Lack of know-how in model implementation, maintenance, or interpretation
- Difficulties developing and tuning monitoring systems for full operating range
- Handling process drifts and changes to processes over time
- Lack of provision for on-going operation and system maintenance

In addition to these obstacles, the biopharmaceutical industry faces challenges from the complexity of the data generated by the biologics manufacturing process [55,56]. However the continuing development of data collection equipment such as the electronic noses and the MVDA techniques required to interrogate these datasets continue to provide new opportunities to enhance process understanding. It was with these barriers in mind that the contents for the statistical toolbox representing a core outcome of the presented work were selected and developed for use by the mammalian cell culture research and development department of the biopharmaceutical contract manufacturer, Lonza.[1] It is intended that the toolbox be directly utilised, adapted for use, or act as a foundation for the development of a new toolbox by other departments, sites, and industries.

The toolbox included outcomes from several sub-projects that focussed on different questions posed by the host company and explore different aspects of MVDA highlighted in this introduction. Each chapter is prefaced with more detailed information on the relevant areas of investigation to better contextualise the research and conclusions.

[1] An outline of the final toolbox and contained tools and guidelines can be found in §7.5.

Chapter 2. Materials and Methods

Three datasets form the foundation of the research discussed within the thesis. Each dataset was formed from a different combination of host cell lines and expression systems to produce a variety of protein products (Table 1). This chapter is broken down into four sections detailing the materials and methods used to generate the data:

1. Development and selection of mAb-producing cell lines
2. Overview of platform processes at Lonza
3. Description of data collected
4. Description of data collection methods

Dataset	Host Cell and Expression System	Number of Cultures	Note	Identifier
1	DHFR-CHO	48	Single product	Culture ID
2	GS-NS0	99	Single product	AXXX
3	GS-CHO	185	Multiple products	Pro_XXX_XXX

Table 1. Summary of three datasets used in the course of EngD research. DHFR – dihydrofolate reductase deficient. GS – glutamaine synthetase. CHO – Chinese hamster ovary. NS0 – non-secreting murine myeloma.

2.1 mAb-Producing Cell Lines

Three different cell types were used in the work presented in this thesis: DHFR-CHO, GS-CHO, and GS-NS0. These were developed using two selection systems (dihydrofolate reductase (DHFR) and glutamine synthetase (GS))and two base host cell lines (Chinese hamster ovary (CHO) and non-secreting murine myeloma (NS0)).

2.1.1 Selection Systems

Glutamine synthetase (GS) is an enzyme required for the synthesis of the amino acid glutamine from glutamate and ammonia [57,58]. If the host cell lacks endogenous GS activity, successful transfection with a vector containing the GS enzyme allows the host cell to survive through exogenous GS activity in a glutamine-deficient media where glutamate and ammonia are available [59]. Culturing in glutamine-deficient media allows successfully transfected cells to be isolated because cells not successfully transfected would not survive. For cells possessing endogenous GS metabolism, selection pressure can be applied through the addition of methionine sulphoxine (MSX) which inhibits GS metabolism [60,61]. The GS cell lines in the presented research used Lonza's GS gene expression system, which includes the genetic sequences for the selectable marker (Patent W087/04462) and the associated hCMV promoter (Patent W089/011036).

A second commonly used selection expression system is based on the enzyme dihydrofolate reductase (DHFR). DHFR catalyses the conversion of folic acid to tetrahydrofolate, which is required to produce glycine, purines, and thymidylic acid for cell growth and proliferation [62]. According to Racher and Birch [59], the main role of the DHFR gene is to improve vector amplification when culturing cells in a folic acid-deficient media. Selection pressure can be applied during amplification by inhibiting endogenous DHFR activity through the addition of the folate analogue methotrexate (MTX) [63].

2.1.2 Host Cell Lines

Mammalian cell lines are generally the preferred host cell for monoclonal antibody production. This is due to a variety of post-translational modifications mammalian cells perform, in particular the glycosylation of proteins. The glycosylation profile of a protein plays a role in protein recognition by a patient's immune system, which in turn can affect drug efficacy and the likelihood of side effects [28,64–69].

Two industrially important mammalian cell lines are non-secreting murine myeloma (NS0) and Chinese hamster ovary (CHO) [70,71]. The NS0 cell line is a non-glutamine secreting subclone of the murine myeloma cell line, NS-1, which was isolated and identified in 1976 [72,73]. In industry, CHO typically refers to one of several cell lines derived from a single clone isolated in 1957 by Dr. Theodore T. Puck [74]. Three derivatives frequently encountered in industry are: DUXB11, DG44, CHOK1SV [75]. The DUXB11 and DG44 cell lines are DHFR-deficient cell lines developed by Columbia University [76]. The CHOK1SV cell line was developed by Lonza [59].

DHFR deficiency in CHO cell lines occurs due to either mutation or deletion of the *dhfr* alleles. Unlike the GS expression system, the competitive inhibitor MTX is required during cell line selection to isolate successfully transfected cells.

Both CHOK1SV and NS0 cell lines can be cultured using the GS system [77–79]. In the case of CHOK1SV, this is due to the natural GS activity of the CHO cells being greatly reduced through gene silencing and the use of the GS inhibitor MSX to improve selection [80,81]. The NS0 cell line lacks endogenous GS activity and therefore does not require MSX selection [82,83]. However MSX may be used during GS-NS0 cell line selection as increased levels of MSX have been associated with increased GS gene copy number in GS-NS0 [77]. A 1985 study by Bebbington and Hentschel found that "[t]he amount of protein product of transfected genes is often found to be roughly proportional to the

number of functional copies of the gene present" [84]. There is some debate as to whether this is true only in specific cases as studies have been published both supporting [85] and discrediting [80] this theory. MSX is not used in GS-NS0 selection at Lonza unless specified by the client.

2.2 Platform Processes and Technologies

Three GS-CHO platform processes were offered by Lonza. In each platform process, operating conditions for temperature control, pH control, dissolved oxygen tension (DOT) control, gassing strategies, feed strategies, and medium compositions are specific to the platform version. When comparing GS platform versions 6, 7, and 8 (Table 2), it can be seen with increasing version number that platform developments have led to nutrient feeds and operating conditions more tailored to culture performance. Key developments are the increasing use of variable feed rates (rates based on some measure of biological performance, e.g. cell mass) and the use of a planned change in pH setpoint at a defined point during a culture in Version 8. If no platform process meets evaluation requirements (e.g. low return on investment), a bespoke process may be developed.

Parameter	**Version 6**	**Version 7**	**Version 8**
pH Setpoint	Constant	Constant	Planned change
pH Control Boundary	Wide	Narrow	Narrow
DOT	15%	40%	40%
Medium	CM42 CD-CHO*	CM54 CD-CHO*	CM76 Proprietary
Nutrient Feeds	SF40 FCR, FV (4 to 5 days)	SF50 CF, VR FCD	SF76 CF, VR FCD
	SF41 CF, VR	Glucose CF, VR FCD	Glucose CF, VR FCD
Nutrient Feed Bolus Additions	N/A	SF52, SF53, SF54 Days 5, 8, 11	SF71, SF72, SF54 Days 3, 5, 8, 10

Table 2. Lonza platform processes for GS expression system [86]. *Invitrogen owned. CF – Continuous Feed. VR – Variable Rate. FCR – Fixed Continuous Rate. FCD – Full culture duration. CR –Continuous Rate. FV – Fixed Volume. BA – Bolus addition

Bioreactor design and scales must also be selected in addition to operating conditions. The two bioreactors designs offered at Lonza were continuous stirred tank reactors (CSTR) and airlift reactors (ALR). Both CSTR and ALR designs have established use in biopharmaceutical manufacturing. The primary difference between CSTR and ALR designs is how the bioreactor culture is agitated and gassing introduced. In an ALR, a vertical baffle divides the interior space with space at the top and bottom to allow continuous circulation of the culture. Gases are introduced on one side of the baffle and drive both culture circulation and the distribution of gases and nutrients throughout the culture. An important point in ALR operation is ensuring the culture volume adequately clears the baffle top and bottom to allow thorough mixing. A second important point in ALR operation is ensuring the gassing strategy provides an appropriate physical force for circulation throughout the culture duration. In a CSTR, agitation is driven by an impeller in the bioreactor. The impeller is driven by an external motor and hence agitation strategy can be designed independent of gassing strategy.

The bioreactor scales offered at Lonza considered in the presented research are 10L, 130L, 2000L, and 5000L. The scale used for a culture reflects the development stage of a project. The 10L scale bioreactors were used for research and development activities following Good Laboratory Practice (GLP), such as evaluating adaptation of a cell line to process platforms or experiments to test the effects of potential deviations. The 130L scale bioreactor, also referred to as “pilot scale”, was used to evaluate non-experimental culture performance at a larger scale using Good Manufacturing Practice (GMP). The 2000L and 5000L scale bioreactors were used for full scale production using GMP procedures.

During research and development activities, a cell line might be cultured in both CSTR and ALR to determine which provided a more suitable environment for growth or to evaluate the effects of using a different design from what had been previously used, e.g. Client A wishes to transfer from a CSTR process to an ALR process. Above the 10L scale, all higher scales used the design selected at the 10L scale.

2.3 Data Collected

The development of new process platforms is dependent on identifying areas for improvement, such as favouring a particular metabolism pathway through altering culture temperature. As it is not feasible to monitor all possible variables, online monitoring and offline daily monitoring of cultures centred on a core set of variables (briefly outlined in Figure 5). The biological relevancy of the core set of variables are described here. Due to the variety of methods in which these variables were monitored, variable monitoring methods are described in §2.4.

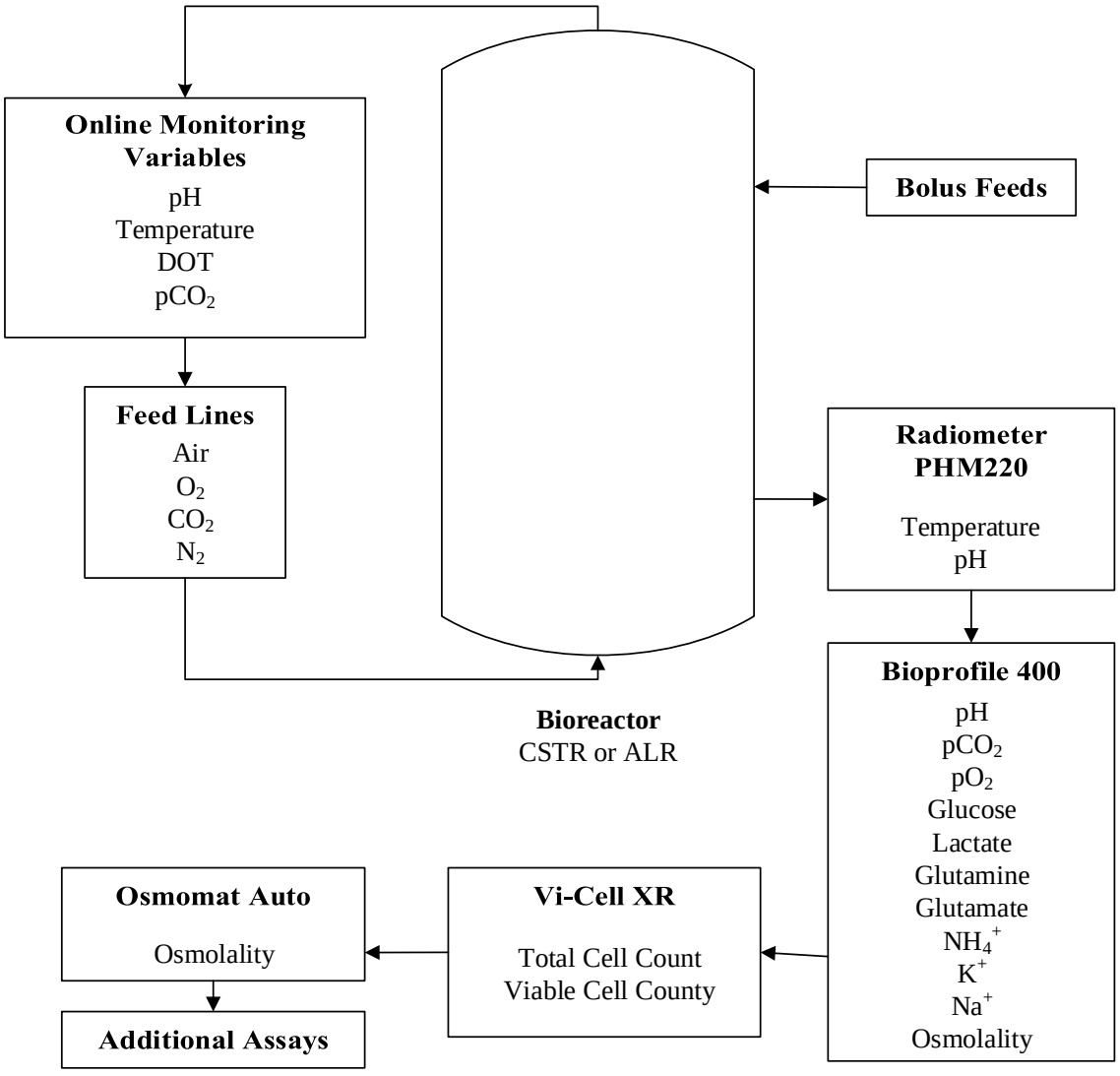


Figure 5. A simplified view of the variables monitored through online and offline measurements. The sequence of Radiometer PHM220 to Additional Assays shows the typical process a culture sample underwent. The biological significance of these variables are described in §2.3. Variable measurement methods are described in §2.4.

2.3.1 Temperature

For mammalian cells, the effects of deviation from normal physiological temperature of 36.5°C are dependent on the magnitude and direction of the deviation, in addition to other environmental conditions. For a DHFR-CHO cell-line producing a humanised mAb, a temperature shift from 37°C to 31°C at pH 6.8 during the stationary phase of the culture gave a 2.3-fold increase in mAb concentration [87]. Similarly, a shift from 37°C to 32°C during the stationary phase resulted in a 3-fold increase in the production of a recombinant IgG_4 mAb by a GS-CHO cell line [88].

It is important to note that these effects do not hold true for all cell lines or products as shown through two studies by one research group. In the first study, CHO-DHFR$^-$ producing erythropoietin (EPO) cultured at 33°C showed a 4-fold increase in productivity when compared to a 37°C culture [89]. When the same cell line and CMV promoter were used to produce anti-4-1BB antibody, low culture temperatures did not result in enhanced productivity [90]. Through these two studies, the research group demonstrated that the degree of enhanced productivity from lower temperatures (if any) was affected by the product itself or the integration site of the vector.

2.3.2 pH

Culture pH affects cell growth [91–94], metabolism [95], product quality [96], and production rates [97]. Due to the ability of pH to affect biochemical characteristics [87], deliberate changes in the pH operating setpoint may be made as part of an experimental study or as part of the standard operating platform [98].

The behaviour of the cell culture itself can affect pH levels as accumulation of metabolites such as lactate can alter culture pH. Furthermore, effects from changes in pH can interfere with effects from other variables [87]. As noted above, in one study changing the setpoint of a bioreactor from 37°C to 31°C during the stationary phase of the culture resulted in a 2.3-fold increase in mAb concentration for a CHO-DHFR$^-$ cell line [87]. This increase was observed at a pH of 6.8 but did not occur at a pH of 7.0.

2.3.3 Dissolved Oxygen Tension

Mammalian cell lines, such as CHO and NS0, are aerobic and require oxygen to survive [25,41]. While the exact effects of dissolved oxygen tension (DOT) depends on the cell line in question [64], it is known that hypoxia can affect growth rate [99–101], metabolism [102], and glycosylation of the mAb product [64,99]. Additionally, there is evidence that oscillating DOT levels can also affect product glycosylation without

obviously affecting growth [103]. By affecting glycosylation patterns, particularly without notably affecting growth, control of DOT can pose a major concern to biopharmaceutical producers.

Studies suggest that hyperoxia is better tolerated than hypoxia in certain aspects of metabolism [99,104]. However DOT setpoints used at Lonza are typically in the range of 15% to 40%. These levels are low enough that hyperoxia conditions would likely only occur in the event of equipment failure, e.g. a DOT probe giving a false reading.

2.3.4 Glucose

Glucose is an important nutrient for mammalian cell metabolism [105–107]. Insufficient glucose levels can result in decreased rate of growth [107–109], decreased specific rate of productivity [108], and incorrect glycosylation of products [110]. Glucose depletion may also cause cell metabolism to shift from lactate production to lactate consumption [102].

2.3.5 Lactate

Mammalian cells are known to produce lactate as a part of normal metabolism and in response to stress conditions [107,111–113]. Accumulation of lactate may in turn affect culture performance, e.g. growth [114] or expression levels [109]. Normal lactate metabolism may also be affected when expressing the product of interest. Effects including altered lactate consumption [115–117] or increased lactate expression [118,119].

2.3.6 Glutamine, Glutamate, and NH_4^+

Glutamine is an amino acid which serves a wide variety of functions in cells [77]. These include, among others, roles in the synthesis of proteins, pyrimidines, and purines, degradation of amino acids, acting as a nitrogen source, and the ability to function as a source of carbon and energy [77]. In the absence of glutamine, glutamine can be synthesised by the enzyme glutamine synthetase (GS) using glutamate and NH_4^+ present in the culture medium.

As stated previously, NS0 cells exhibit very low levels of endogenous GS activity, hence exogenous GS activity can be used to identify successfully transfected cells in a glutamine-free medium during cell line construction [58,61,85]. As CHO cells exhibit a degree of endogenous GS activity, MSX or a similarly competitive inhibitor is used to apply selection pressure during cell line construction and isolation [61,120,121].

In addition to acting as substrates for GS, glutamate and NH_4^+ serve other functions in cell metabolism. Glutamate can be consumed in the synthesis of the Kreb's cycle intermediary compound α-ketoglutarate [122]. NH_4^+ has been shown to effect protein glycosylation patterns [123].

2.3.7 Na^+ and K^+

Cell metabolism produces acid equivalents as a by-product, accumulation of which leads to acidification of the cytosol [124]. To maintain an appropriate physiological intracellular pH, these acid equivalents are transported through the cell membrane by the transmembrane enzyme sodium-potassium adenosine triphosphatase (Na^+/K^+-ATPase) in exchange for Na^+ ions.

Na+ and K+ uptake and release by cells is also driven by the activity of Na^+/K^+-ATPase during regulation of intracellular osmotic pressure, cell volume, and signal transduction. Na^+/K^+-ATPase activity is related to cell life cycle stages, in particular the transition between the G1 growth phase and the S senescence phase [125].

2.3.8 pO_2

Oxygen is an essential input for aerobic metabolism. The partial pressure of oxygen in in solution in the culture medium (pO_2) is a measure of oxygen in the culture medium available for use by cells. It is monitored for the same reasons given for monitoring of DOT.

2.3.9 pCO_2

CO_2 is a by-product of several cell metabolism pathways for the production of compounds such as pyrimidines, purines, and fatty acids [96,126,127]. pCO_2 is the partial pressure of dissolved CO_2 gas in culture medium. Accumulation of CO_2 gas in the culture medium causes increased pCO_2 levels which can have a range of effects on protein-producing mammalian cell cultures depending on product and cell type. These include inhibition of cell growth [114], decreased rates of glucose consumption [126], decreased rates of lactate production [126], and altered glycosylation [96,128,129].

Increased accumulation of CO_2 in culture medium is often observed during scale-up of bioreactors and may be caused by insufficient stripping of CO_2 from the culture (e.g. by nitrogen), poor culture mixing, or as a side-effect of controller action if sparging with CO_2 gas is used as part of pH control [126]. In particular, build-up of CO_2 at a constant

pH is associated with increased osmolality and subsequent knock-on effects such as increased intracellular pH and increased Na^+/K^+-ATPase activity [127].

2.3.10 Culture Osmolality

Osmolality is the concentration of solutes in a sample measured in osmoles of solute per kilogram (Osm/kg) of solvent [130]. This provides information on how much material is in a sample, however it does not identify or specify quantities of individual components present in the sample. The indiscriminate nature of osmolality, its impact on statistical analyses, and a proposed solution are further addressed in Chapter 4.

The osmolality of a culture can effect a culture in many way including effects on cell growth rate [114,131], specific productivity [131], product quality, e.g. affecting product glycoform [128] or polysialylisation [96], and cell mechanical properties such as bursting force and cell diameter [132]. The extent of these effects vary with cell type and product, e.g. greater inhibition of cell growth from elevated osmolality has been observed in hybridomas than in CHO cells grown in the same medium [91,113,126,127] and product glycoforms may be robust to changes in osmolality, as in the case of CHO-derived tissue plasminogen activator [133].

Culture osmolality can be altered by a wide range of causes. For example, increased pCO_2 levels can result in increased osmolality [134] as can accumulation of products or by-products in the culture medium.

2.4 Data Collection Methods

A wide variety of data are collected when a bioreactor culture is performed. Data can be subdivided into three general classes based on origin and data type: meta-data, online monitoring data, and daily monitoring data. Four types of probe were used during the online monitoring and daily monitoring data collection: resistance temperature detector, potentiometric, amperometric, enzyme-immobilised amperometric.

A resistance temperature detector (RTD) is based on the resistance of a metal element as a function of temperature for a given operating range [135]. The metal element is held by a glass or ceramic core, and the full assembly is sheathed in a protective housing that allows it to be safely inserted into a reactor or bioreactor. Platinum is the most commonly used metal due to its high accuracy and resistance to corrosion with the platinum-coiled Pt100 probe design found across many industries [136].

Potentiometric probes are ion selective probes, where the ion of interest (typically hydrogen) is sensed by a probe membrane [137]. This results in a change in the membrane potential from which the ion concentration can be determined using the Nernst Equation (Eq. 2.1).

$$E = E_o + 2.303 \left(\frac{RT}{nF}\right) \log a_o \qquad \text{Eq. 2.1}$$

where E is the total potential developed between sensing and reference electrodes (mV), E_o is the standard potential of the electrode (mV), R is the Universal Gas Constant, (8.314 J K^{-1} mol^{-1}), T is the temperature (K), n is the moles of ion in the sample, F is the Faraday constant, and a_o is the activity of the ion in solution.

In an amperometric probe, a permeable membrane covers the electrode, allowing the ion of interest to pass through the membrane. The ions initiate some reaction that produces an electrical current from which the ion concentration can be determined. For example, oxygen ions initiate an oxidation-reduction reaction [138].

For an amperometric probe with an enzyme immobilised membrane, the immobilised enzymes produce measureable by-products in the presence of the substrate of interest and any other necessary reagents. For example, conversion of the substrate in the presence of oxygen may produce hydrogen peroxide (H_2O_2) [138]. H_2O_2 is then oxidised at the anode resulting in a change in current charge proportional to the concentration of the substrate of interest in the sample.

2.4.1 Meta-Data Collection

Meta-data is a term capturing a wide variety of data and data sources, such as project name, culture identifier, media batch numbers, bioreactor station identifier, operator/scientist, and version numbers for spreadsheet calculators for feedrates. Effective utilisation of meta-data is a challenge due to both the scale of the meta-data available in biopharmaceutical process and the manner in which it is captured, i.e. predominantly written records or print outs [139]. Even if preserved electronically by scanning, the resulting files are often difficult to search as an information databased.

2.4.2 Online Monitoring Control and Data Collection

At the 10L scale, online monitoring was achieved by the use of probes inserted into the bioreactor and flowrate meters on gas lines into the bioreactor. These probes were connected to an Applikon i-Control unit (Applikon Biotechnology, UK). The unit recorded values at a five minute intervals. Nutrient feeds were not controlled using this system. Instead nutrient feeds were controlled manually using peristaltic pumps and electronic balances. Hence nutrient feeds were treated as part of daily monitoring activities.

At larger scales, one probe was used for control with two additional probes connected for monitoring. This allowed two-against-one arguments to be used to identify faulty probes, in addition to providing redundancy in the event of the control probe failing. Multi-probe arrangements could also be used to monitor and identify gradients within the bioreactor as replicate probes were located in different positions in the bioreactor, e.g. top, middle, and base levels.

During a bioreactor culture, the signals and readings informing the controller are sampled at a set interval, e.g. 5 minutes. The bioreactor's online monitoring record can be exported to .csv files whenever desired, e.g. during a culture or after harvesting.

2.4.2.1 pH

Online monitoring and control of pH was achieved through the use of a Mettler-Toledo 405-DPAS-SC-K8S/425 potentiometric pH probe (Mettler-Toledo, UK). If the recorded value was outside an operating deadband (e.g. ±0.02), automatic corrective action was taken by the control system through CO_2 addition or base addition.

pH measurements of samples collected during offline monitoring were compared to control system readings to identify and correct probe drift. If the discrepancy was > 0.02,

an offset was made to the online probe to bring online readings into agreement with the offline reading. If the discrepancy was ≤ 0.02, no adjustment was made.

2.4.2.2 Temperature

Culture temperature was measured using a Pt100 temperature probe with Lemo connector (Electrolab Biotech, UK). At the 10L scale, passive cooling was used with additional heat supplied by a thermal/heating pad jacket. At larger scales, heating and cooling requirements were met by circulating water through vessel jackets.

Temperature measurements were made on daily offline samples, however these were not used to make external adjustments as part of daily monitoring. If bioreactor temperature or temperature control was question, an Almeno temperature probe was used to verify culture temperature and determine if corrective action was required.

2.4.2.3 Dissolved Oxygen Tension

Culture DOT was monitored using a P52201015 DOT probe (Mettler-Toledo, UK). DOT levels were controlled by increasing and decreasing air and oxygen gas flowrates.

2.4.2.4 Carbon Dioxide, Air, and Oxygen Flowrates

Carbon dioxide (CO_2) gas is an acidic gas that was used to correct pH in cultures when pH measured > pH setpoint.

CO_2, air, and oxygen gas flowrates into bioreactors were controlled and monitored using flowrate meters. Gas flowrate and composition when exiting the bioreactor were not recorded.

2.4.2.5 Level

At the 10 L scale, bioreactor fill levels were monitored visually. Above the 10 L scale, bioreactor fill levels were monitored using level probes within the bioreactor.

2.4.3 Daily Monitoring Data Collection

As part of normal operation, cultures were sampled approximately every 24 hours. At the 10L scale, samples underwent the sequence described below. At scales larger than 10L, a similar sequence was followed with some differences. This differences are noted as required.

2.4.3.1 Bioreactor Conditions

Bioreactor temperature, bioreactor pH, and bioreactor dissolved oxygen tension (DOT) at the time of daily sampling were recorded from the online control unit.

2.4.3.2 Radiometer PHM 220

A Radiometer Analytical PHM220 meter with a Mettler-Toledo potentiometric pH probe was used to measure sample temperature and sample pH. The sample pH measurement would be compared to the bioreactor pH measurement to determine if the online probe required adjustment due to drift.

2.4.3.3 NOVA Bioprofile 400

A NOVA Bioprofile 400 was used to measure multiple variables, ranges and accuracies for which are presented in Table 3. A second offline pH measurement, referred to as the NOVA pH, was made with a potentiometric probe. Enzyme-immobilised membrane amperometric probes were used to measure sample concentrations of glucose, lactate, glutamine, and glutamate. Ion selective electrodes were used to measure sample concentrations of ammonium ions (NH_4^+), potassium ions (K^+), and sodium ions (Na^+). Membrane amperometric electrodes were used to measure sample partial pressure of oxygen (pO_2) and partial pressure of carbon dioxide (pCO_2).

Component	Range	Accuracy	Probe Type
pH	5.00 to 8.00	± 0.01%	Potentiometric
Glucose	0.2 to 15.0 g/L	± 5.0%	Enzyme-immobilised membrane amperometric
Lactate	0.2 to 15.0 g/L	± 5.0%	
Glutamine	0.2 to 6.0 mmol/L	± 5.0%	
Glutamate	0.2 to 6.0 mmol/L	± 5.0%	
NH_4^+	0.2 to 25.0 mmol/L	± 5.0%	Ion selective electrode
K^+	1.0 to 25.0 mmol/L	± 3.0%	
Na^+	40 to 220 mmol/L	± 1.5%	
pO_2	0 to 800 mmHg	± 5.0%	Membrane amperometric
pCO_2	3 to 200 mmHg	± 5.0%	

Table 3. Summary of daily monitoring offline sample pH, chemical concentrations, and partial pressures measured using a NOVA Bioprofile 400, including ranges, accuracies, and probe types [140].

2.4.3.4 Vi-CELL™ XR

Trypan Blue permeates the membranes of non-viable cells but is excluded by the membranes of viable cells. When added to a sample, the dye stained only non-viable cells [141]. A Vi-CELL™ XR (Beckman Coulter) was used to count viable cell numbers and total cell numbers. From this, viable cell concentration (VCC, 10^6 cells/mL) and total cell concentration (TCC, 10^6 cells/mL) were determined. Viability was then calculated as the percentage ratio of viable and total cell counts.

$$Viability\ (\%) = \frac{Viable\ Cell\ Count}{Total\ Cell\ Count} \qquad \text{Eq. 2.2}$$

The integral of viable cell concentration of the sample was calculated as the area beneath the viable cell concentration profile with the units 10^6 cells.h/mL as shown in Figure 6).

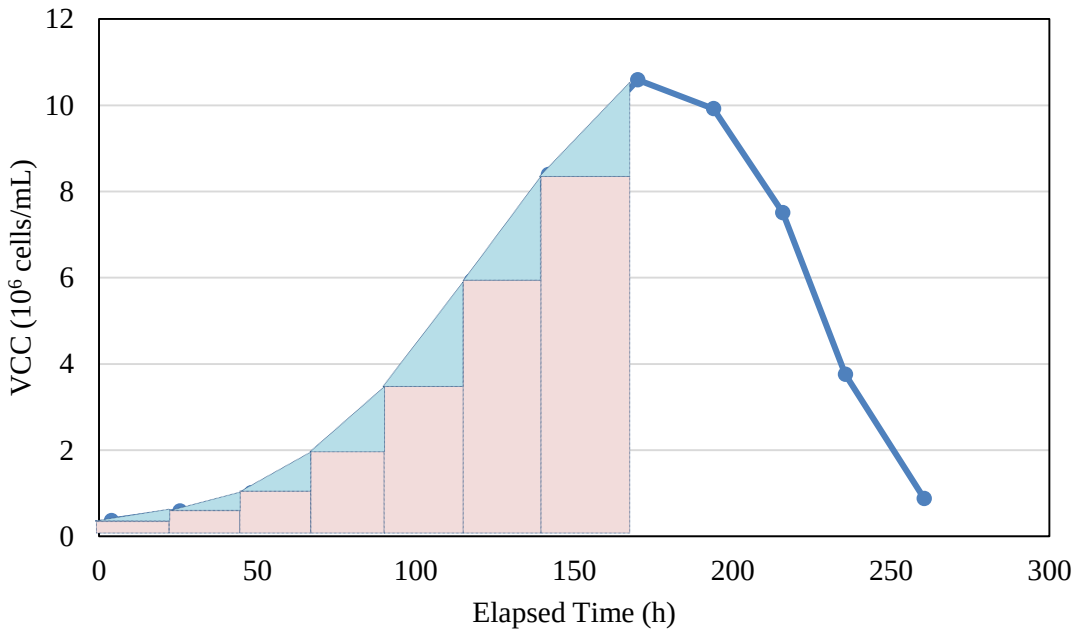


Figure 6. Viable cell concentration (VCC) for the culture Dataset α ProA_001. The pink rectangles and blue triangles indicate the areas used to calculate the integral of viable cell concentration (IVC) for a sample.

2.4.3.5 NOVA Bioprofile 400 and Osmomat Auto

Two different methods were used to measure sample osmolality. Each osmolality measurement method was treated as producing a different variable. This allowed for greater traceability regarding data origin and identification of potential equipment biases or errors.

The first measurement was made using a NOVA Bioprofile 400 using a component calculator. A component calculator is an equation estimating a sample's osmolality based on known concentrations of sample components and those compounds' effects on osmolality. The benefit of this method is that it does not typically require additional equipment, and it is technically possible to use a component calculator to generate osmolality values for an historical dataset. However it is important to note that component calculator values are not true osmolality measurements as the calculations can only take into account the effect of components that are directly monitored and does not take into account unmonitored components.

The second osmolality measurement was based on freezing point osmometry (FPO) using an Osmomat Auto (Gonotec Gmbh, Germany). In FPO, the osmolality of a sample is determined by comparing the difference between the freezing point of the sample and the freezing point of water as the freezing point of a liquid is depressed when another compound is added. FPO is a rapid, inexpensive method appropriate for small sample sizes of low viscosity, non-colloidal solutions. For this reason, it is the preferred method for most biological applications [142], however an FPO osmolality measurement was not part of routine daily monitoring at scales larger than 10 L.

2.4.3.6 Additional Assays

Following the above sequence, daily monitoring samples were also submitted for more detailed protein assays and metabolite tests, which were not treated as part of the daily monitoring data set for the purposes of the presented research.

Chapter 3. Statistical Methods

Multiple statistical tools are employed throughout this thesis. The techniques relevant to multiple chapters are presented here. Additional statistical background is presented where appropriate for the presented work.

3.1 Multiple Linear Regression

A simple form of regression is multivariate linear regression (MLR). Given the inputs $(x_1, x_2, \ldots, x_k)$, a response (y) can be modelled as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_n x_k \qquad \text{Eq. 3.1}$$

where $\hat{y}$ is the predicted response, $\hat{\beta}_0$ is a constant, k is the number of independent variables, and $\hat{\beta}_1$ to $\hat{\beta}_k$ are coefficients for the inputs x_1 to x_k respectively. The difference between a response predicted by a model $(\hat{y})$ and the actual measured response (y) is termed a model error or residual (ε), e.g. for a sample i

$$\varepsilon_i = y_i - \hat{y}_i \qquad \text{Eq. 3.2}$$

The coefficients are calculated assuming that model residuals are normally distributed and the sum of errors minimised. Distribution of residuals can indicate if the model is being distorted by outliers or that behaviours in the response data are not captured in the developed model, .e.g. heteroscedasticity (skew) in the modelled dataset.

Residuals can also be used to evaluate the extent to which a model can be generalised by using the predictive error sum of squares (PRESS) statistic. The model is fitted against every subset of observations excluding a sample i. Residuals for each subset are calculated (ε_{-i}), then squared (ε_{-i}^2). The PRESS statistic is then calculated as:

$$PRESS = 1 - \frac{\sum_{i=1}^{n} \varepsilon_{-i}^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \text{Eq. 3.3}$$

where $\bar{y}$ is the mean average of all responses. The PRESS statistic is analogous to the coefficient of determination of a model with a range of 0 to 1 with 1 indicating perfect predictive accuracy for the tested observations.

3.2 Significance Testing

The predictive ability of a MLR model may be negatively affected by over-parameterisation, the inclusion of too many factors in a model. When a model is over-parameterised, highly correlated variables compete to convey similar information

and noisy variables can reduce model efficiency. Significance testing can be used to manage these issues by identifying statistically insignificant variables that can then be removed. A two-sided t ratio (Eq. 3.4) is used to evaluate the variables in a MLR and is the ratio of the variable parameter estimate to the standard deviation of the variable:

$$t = \sqrt{\frac{R_{adj}^2(n-k-1)}{1-R_{adj}^2}} \qquad \text{Eq. 3.4}$$

where n is the number of samples, k is the number of independent terms in the model, and R_{adj}^2 is the adjusted Pearson's coefficient of determination. From the t score and a selected threshold value α (historically 0.05), a two-tailed p value for t is calculated as:

$$p = 2 * \mathrm{P}\left(t > t_{\left(\frac{\alpha}{2}, n-k-1\right)}\right) \qquad \text{Eq. 3.5}$$

If p<α, then the variable is said to be statistically significant. If p>α, the variable is said to be statistically insignificant. The variable with the highest p value above the chosen threshold, i.e. the least significant variable, is removed and a new model created. This is repeated until only statistically significant variables remain.

Significance testing in this manner is a form of stepwise backward elimination (Table 4) as the procedure begins with the full set of variables and with each step, the least informative variable is removed [143]. An alternative approach is stepwise forward selection (Table 4), where the most informative variable is determined and with each step, the next most informative variable is added to the set [143]. Each technique can be time consuming, however stepwise backwards elimination is generally both easier to implement and inspires greater personal confidence in the resulting model.

Stepwise Forward Selection	**Stepwise Backward Elimination**
Initial variable set:	Initial variable set:
{}	{A1, A2, A3, A4, A5, A6}
⇨ {A1 }	⇨ {A1, A3, A4, A5, A6}
⇨ {A1, A4}	⇨ {A1, A4, A5, A6}
⇨ {A1, A4, A6}	⇨ {A1, A4, A6}
Reduced variable set	Reduced variable set

Table 4. Comparison of stepwise forward selection and stepwise backward elimination for reduction of a variable set [143].

3.3 Principal Component Analysis

Principal component analysis (PCA) is a statistical dimensionality reduction tool suitable for use on a large dataset X composed of n samples and p factors, which may contain highly correlated factors. PCA captures variation in the dataset by creating new variables termed principal components (PC); these are eigenvectors calculated from the covariance matrix of X so that for a PCA model using k PCs:

$$X = T\mathrm{P}^{\mathrm{T}} + E \qquad \text{Eq. 3.6}$$

where X is the original nxp data matrix, T the nxp scores matrix, P^{T} the kxp loadings matrix, and E an nxp model residuals matrix [144].

Each PC is a new multilinear combination of the original variables and orthogonal, i.e. uncorrelated with other PCs. PCs capture variance in a cumulative manner, where the first PC captures the most variance in a single PC, the second PC the next most variance, and so on. Up to n or p PCs can be computed (whichever is smaller), however in practice the number of PCs retained in a model will usually be less than this. The number of PCs to be retained can be selected from a variety of ways. This may be a simple calculated threshold, e.g. a minimum cumulative variance is reached or the number of PCs giving the lowest root mean square error (RMSE) for a cross-validation set:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2} \qquad \text{Eq. 3.7}$$

When the loadings for PCs are plotted in two-dimensional space, correlated variables will cluster together. Negatively correlated variables will be at opposite points across the origin; variables with little or no correlation will be orthogonal. Similarly, when scores for samples are plotted in two-dimensional space, samples with similar behaviour will cluster together while samples with dissimilar behaviour will separate.

PCA can be performed on both covariance and correlation matrices. Both matrix types are essentially the same, in that relationships between variables are evaluated for independence, however different models will be returned depending on which matrix is used. Covariance is an unbounded, unlimited value. When the covariance matrix of a dataset is used, variables with high covariance values are be prioritised over variables with low covariance during model creation due to differences in value magnitude. Correlation is standardised to the limits [-1.0, 1.0] and so the magnitude of the original

values is eliminated as a potential bias. For this reason, correlation matrices were used in the presented work.

An extension of PCA is principal component regression (PCR). Here the scores generated in PCA are then regressed against a response *Y*. A weakness of PCR is that PCA is a technique for describing variance in the input dataset, which may have very little to do with variance in the response of interest.

3.4 Partial Least Squares/Projection to Latent Structures

In PCA, the purpose of the model was to capture variance in an X dataset without linking that behaviour to a response Y. While PCA is a valuable tool for dataset exploration, it is of limited use when a response of interest exists as the related technique, principal component regression (PCR), emphasises capture of the X dataset.

Partial least squares or projection to latent structures (PLS) are the same iterative algorithm whereby variance in an X dataset is captured based on the ability to describe variance in a response dataset Y[145,146]. The Non-Linear Iterative Partial Least Squares (NIPALS) algorithm uses an outer regression between X and Y (Eq. 3.8) and inner regressions for the input dataset X (Eq. 3.9) and the response dataset Y (Eq. 3.10).

$$Y = X\beta + \varepsilon \quad \text{Eq. 3.8}$$

$$X = T * P' + E \quad \text{Eq. 3.9}$$

$$Y = U * Q' + F \quad \text{Eq. 3.10}$$

where T and U are score matrices, P' and Q' are loadings matrices, and E and F are error/residuals matrices. This is shown in Figure 8 and the NIPALS algorithm is shown in Table 5.

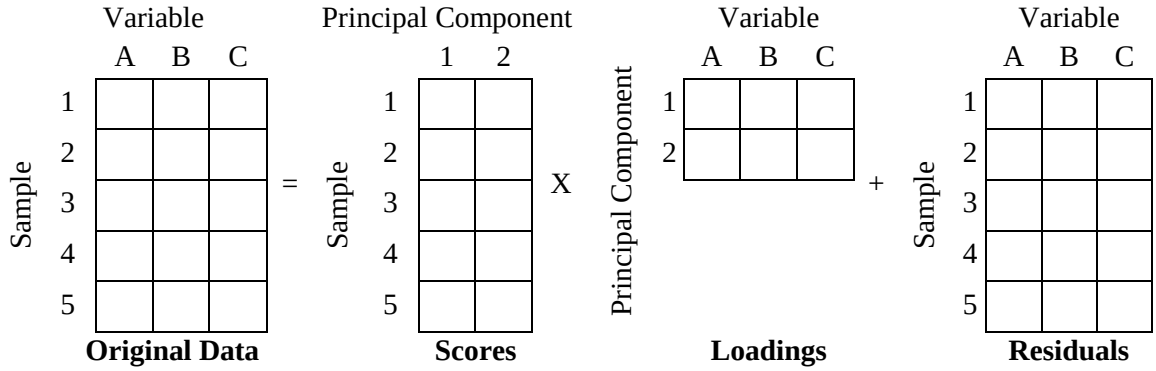


Figure 7. Visual representation of PCA architecture. New variables termed principal components (PC) are created. These are multivariate linear combinations capturing variance in the original dataset. In the PCA model, each sample is now represented by a single score for each PC retained plus a multivariate residual, which captures variance not captured in the PCA model. Here two PC are retained, therefore each sample has a score for PC 1, a score for PC 2, and a residual.

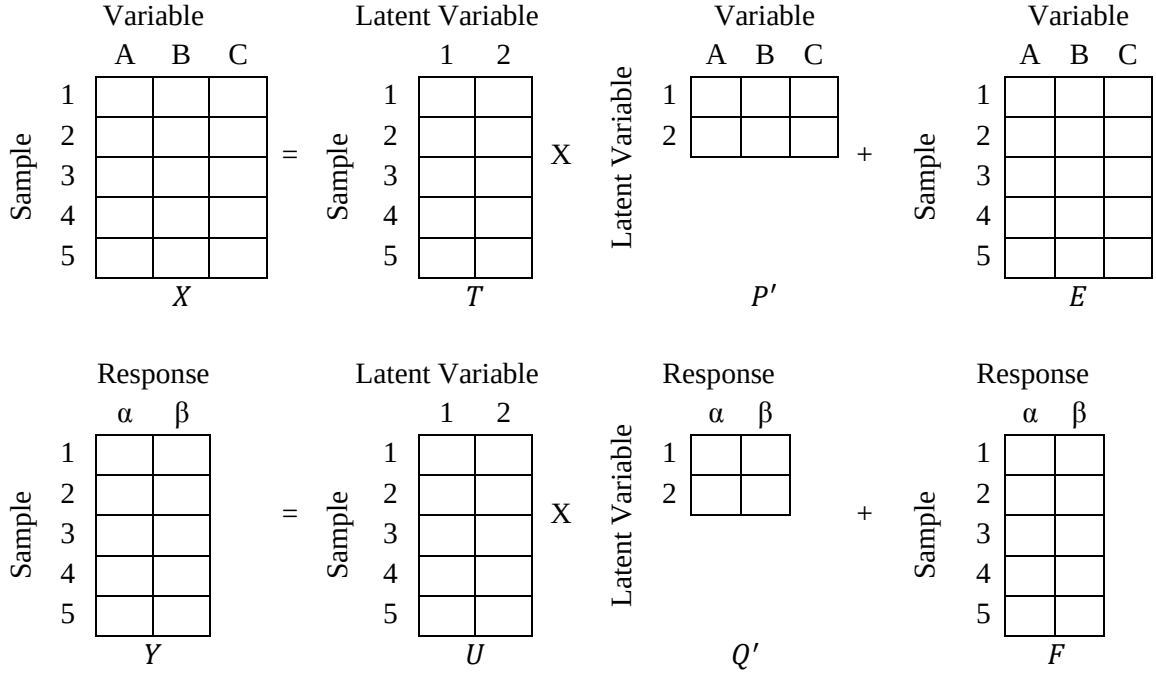


Figure 8. Visual representation of PLSR architecture. For input dataset X and response dataset Y respectively, T and U are scores matrices, P' and $Q'$are loadings matrices, and E and F are error/residuals matrices.

	X Regression
1.	Let $t =$ some x_j
2.	$$p' = \frac{t'X}{t't}$$
3.	Scale $$p' = \frac{p'}{\|p'\|}$$
4.	$$t = \frac{Xp}{p'p}$$
5.	If Step 2 $p' \neq$ Step 4 p', return to Step 2. If Step 2 $p' =$ Step 4 p', go to Step 6.
6.	Repeat Step 1 to 5 on residuals matrix E.

	Y Regression
1.	Let $u =$ some y_j
2.	$$q' = \frac{u'Y}{u'u}$$
3.	Scale $$q' = \frac{q'}{\|q'\|}$$
4.	$$u = \frac{Yq}{q'q}$$
5.	If Step 2 $q' \neq$ Step 4 q', return to Step 2. If Step 2 $q' =$ Step 4 q', go to Step 6.
6.	Repeat Step 1 to 5 on residuals matrix F^*.

	Single Algorithm for Improved Inner Relationship	
1.	Let $u =$ some y_j	
2.	$$p' = \frac{u'X}{u'u}$$	$$w' = \frac{u'X}{u'u}$$
3.	Scale $$p' = \frac{p'}{\|p'\|}$$	Scale $$w' = \frac{w'}{\|w'\|}$$
4.	$$t = \frac{Xp}{p'p}$$	$$t = \frac{Xw}{w'w}$$
5.	$$q' = \frac{t'Y}{t't}$$	If Y has only one variable, then $q = 1$ in Step 5 to 7.
6.	Scale to unit length $$q' = \frac{q'}{\|q'\|}$$	
7.	$$u = \frac{Yq}{q'q}$$	
8.	If Step 4 $t \neq$ previous Step 4 t, return to Step 2. If Step 4 $t =$ previous Step 4 t, go to Step 9.	
9.	Repeat Step 1 to 8 on residuals E and F^*.	

Table 5. Non-Linear Iterative Partial Least Squares (NIPALS) algorithms for outer regressions and combined algorithm for improved capture of the inner relationship [146]. In the Single Algorithm, both t (some x_j) and u (some y_j) are used to determine individual latent variable loadings(p' and q') in both the X regression and Y regression. All three algorithms use convergence as stopping criteria for proceeding to calculate the next latent variable from the residuals matrices E and F^*.

Table 6 and Table 7 are presented as a theoretical comparison of regressions using PCR and PLS for a dataset X and a response Y. In Table 6, PCA has been applied to the dataset X. Considering the X variance captured, it is seen that the 1st PC captures the greatest percentage variance (60%), the 2nd PC a smaller percentage (20%), and the 3rd PC smaller still (10%). Only three PCs are retained as these captured 90% of X variance which, in this example, has been deemed sufficient for proceeding to regression. The scores for this model are then regressed on the response Y. Considering then the Y variance captured, it is seen that only a 55% of Y variance is captured total, despite 90% of X variance being captured. Furthermore, the percentage of Y variance captured does not decrease with PC number as occurs for the percentage of X variance captured.

In Table 7 where PLSR has been employed, the situation is reversed. The 1st LV captures on 20% of the X variance but 60% of the Y variance, the 2nd LV 5% of X variance and 20% of Y variance, and the 3rd LV 30% of X variance and 10% of Y variance. Here a total 55% of X variance explains 90% of the Y variance.

The purpose of this comparison is to clearly demonstrate the difference between PCA-based and PLS-based techniques. PCA focuses on X variance regardless of any response, whereas PLS focuses on X variance only so far as it explains Y variance.

	X Variance Captured (%)		**Y Variance Captured (%)**	
PC	**This PC**	**Total**	**This PC**	**Total**
1	60	60	20	20
2	20	80	5	25
3	10	90	30	55

Table 6. Captured X and Y Variance Using PCR. A total of 90% X variance is captured but only 55% Y variance can be ascribed to this behaviour.

	X Variance Captured (%)		**Y Variance Captured (%)**	
PC	**This PC**	**Total**	**This PC**	**Total**
1	20	20	60	60
2	5	25	20	80
3	30	55	10	90

Table 7. Captured X and Y Variance Using PLSR. 90% of Y variance can be ascribed to only 55% of X variance.

The PLS algorithm is not restricted to regressions to predict a response dataset response dataset *Y.* If the response dataset is categorical classifications, e.g. pass/fail, positive/neutral/negative, the PLS algorithm is used perform discriminant analysis (PLS-DA). In PLS-DA, latent variables and variable weightings are calculated to provide the greatest level of separation between the response classes. Hence, as in PLSR, a structure is imposed on the dataset to capture variance differentiating classes.

As with PCA, the use of cross-validation analysis is recommended to prevent over-fitting of the model. In 'Leave One Out' cross-validation, the number of latent variables used is based on the PRESS statistic. This is repeated until each sample has been left out and a total PRESS is calculated. The 'best' model is selected based on the number of latent variables that gives the lowest total PRESS. The number of latent variables retained can also be based on alternative model performance criteria, e.g. correct classification of specific samples in PLS-DA, instead of overall classification accuracy of the dataset.

A benefit of PLS is that input variables identified by the algorithm as statistically significant for the response of interest are given greater weighting than those identified as statistically insignificant. This ability can improve signal-to-noise ratios in large datasets as statistically insignificant variables can be given effectively zero impact in prediction or analyses.

It is important to question the identified behaviours as the PLS algorithm imposes a structure which, in a suitably large dataset, can potentially create spurious models with high predictive accuracy. While cross-validation can help reduce the likelihood of such model being selected, it is also possible for more nuanced behaviours to be buried beneath "obvious" behaviours. PLS is generally recommended as a tool for extracting information from highly correlated multivariate dataset where there is a high signal-to-noise ratio whether noise is defined as true noise or unnecessary data [147].

3.5 Lack-of-Fit Statistics and Outliers

Lack-of-fit statistics are used to evaluate models produced through modelling algorithms such as PCA and PLS. Two key lack of fit statistics used in the presented work were Hotelling T^2 residual and Q Residuals.

Hotelling T^2 can be thought of as a multivariate version of univariate standardisation as it compares the cumulative variance of a sample's data to the model mean (Eq. 3.11).

$$T_i^2 = x_i^T \mathbf{P} \mathbf{\Lambda}^{-1} \mathbf{P}^{\mathrm{T}} x_i \qquad \text{Eq. 3.11}$$

where ${T_i}^2$ is the Hotelling T^2 for a sample, x_i is the sample data, ${x_i}^T$ is the transpose of the sample data, **P** is the loading matrix for input data, $\mathbf{P}^{\mathrm{T}}$ is the transpose of the loading matrix for input data, and **Λ** is the diagonal matrix of eigenvalues $diag(\lambda_1, \lambda_2, \dots, \lambda_k)$ for the selected number of components or latent variables [148]. An upper control limit T^2_{UCL} is calculated as shown:

$$T^2_{UCL} = \frac{k(n-1)}{n-k} F_{k,n-k;\alpha} \qquad \text{Eq. 3.12}$$

where T^2_{UCL} is the upper control limit value, k is the number of variables, n is the number of observations, $F_{k,n-k;\alpha}$ is the Fvalue for the limit [148]. Values above the calculated limit represent samples with statistically unusual values when compared to other samples in a dataset, however as the limit is calculated using a threshold (α), a percentage of samples should always lie on or above the limit.

As each individual variable's contribution to a sample's Hotelling T^2 residual is calculated, it is possible to identify variables contributing to high Hotelling T^2 values. Three causes for high Hotelling T^2 values encountered during the course of the presented research were:

- Simple decimal error, e.g. a dataset has a mean pH of 6.91 and standard deviation 0.02. A pH reading of 6.91 (σ = 0) is recorded as 69.1 (σ = 3109.5).
- A statistically 'unusual' but biologically irrelevant outlier, e.g. a dataset has 49 readings of 36.5°C and one reading of 36.6°C. When mean centred and normalised, the single 36.6°C reading has a standardised value of 7 σ.
- Samples with experimental operating conditions in a dataset comprising predominantly samples with ordinary operating conditions.

In the case of decimal error, technically all that is required is dataset cleaning. In the case of variables such as pH in a well-controlled process, decimal errors can be confidently identified and rectified. Ideally they are anticipated as part of data capture and the error can be flagged at the point of data entry (e.g. an error box appears when entered into the data monitoring spreadsheet). However decimal errors may not be caught during data cleaning and may have a detrimental effect on subsequent models. This is a particular risk for variables with a greater range of values or potential range of values such a pCO_2 which may range from close to 0 mmHg to over 300 mmHg during the course of a culture.

The statistically unusual but biologically irrelevant outlier is an issue that can be caused by over-cleaned datasets, differences in rounding used between data entries, and biased or unrepresentative sampling (e.g. 'blocking' effects). The following example is a simplification of an issue concerning differences in bioreactor scale (10 L v. 5000 L) and temperature control.

At the 10 L scale, temperature is more quickly and tightly controlled relative to temperature control at the 5000 L scale, where there is longer lag attributed to increased volume. Essentially, when a scientist records reactor temperatures, it is likely that greater variation in temperature will be observed at the 5000 L scale due to the greater lag in control. This statistical difference in control response is not necessarily biologically relevant. However, once this difference between scales is identified, it must be demonstrated that the difference does not impact on, for example, the efficacy of the product. Once this is done, there are several options:

1. If the variable has low significance in the model, remove the variable.
2. Replace the variable with a more robust measure, e.g. replace offline sampling/ daily monitoring temperature measurements with median temperatures calculated from online monitoring records.
3. Introduce more samples of this type to the dataset. In the example, the dataset is heavily biased towards capture of 10 L scale control. Reducing the number of 10 L samples or increasing the number of 5000 L samples would improve the balance between the two control schemes.
4. Rectify identifiable errors in input data if possible. Rounding errors, inconsistency in how many significant figures to record, different units used for measurements (mmol/L v. mg/L), and decimal points errors can affect the calculated Hotelling T^2 contributions.

A high Hotelling T^2 residual does not necessarily mean a sample is an outlier. A sample with a high Hotelling T^2 may represent more extreme values, however this does not necessarily make the sample an outlier if the correlations identified by the model are conserved. Conversely, a low Hotelling T^2 does not necessarily preclude a sample from being an outlier or non-representative sample.

These behavioural outliers can be identified through the use of Q Residuals[2] (Eq. 3.13) also called the squared prediction error (SPE), where the residuals matrix is used to calculate the agreement between a sample's *n*-dimensional location and the location according to the model.

$$Q_i = e_i e_i^T \qquad \text{Eq. 3.13}$$

where Q_i is the calculated Q Residual of sample i, e_i is the i^{th} row vector from the error matrix E, and e_i^T is the transpose of e_i. An upper limit for Q Residuals can be approximated using the Jackson-Mudholkar formula (Eq. 3.14).

$$Q_i = \theta_1 \left[1 - \frac{\theta_2 h_0 (1 - h_0)}{\theta_1^2} + \frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} \right]^{1/h_0} \qquad \text{Eq. 3.14}$$

with

$$\theta_i = \sum_{j=k+1}^{n} x_j^i, i \qquad \text{Eq. 3.15}$$

and

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \qquad \text{Eq. 3.16}$$

where k is the number of PCs retained, α is the significance level, x_j^i the value of variable j for sample i, and z_α is the standard normal value corresponding to the upper 1-α percentile [148]. Q Residual contributions are simply the row vector e_i.When comparing Q Residual contributions for two or more samples, one observation (A) can be set as a baseline. Relative Q Residual contributions are calculated for the remaining samples by subtracting e_A.from the error matrix E.

Q Residuals	Low Hotelling T^2 Residual	High
High	'Normal' values Behaviour does not fit model Ex. Culture with behaviour not captured by model but measurements are within univariate limits.	'Unusual' values 'Behaviour does not fit model Ex. Experimental conditions which affect culture in manner that does not obey model.
Low	'Normal' values Behaviour fits model Ex. Control cultures and "Golden Batches"	Unusual' values Behaviour fits the model Ex. Experimental conditions affecting culture in a manner that obeys model.

Table 8. A simple demonstration of interpreting lack-of-fit statistics for a multivariate model. Ideally, Hotelling T^2 and Q Residuals for samples lie within or near the lower left quadrant. Otherwise this indicates a sample's data, behaviour, or both are statistically different.

[2] This value is also known as DModX. The related limit is referred to as $DModX_{Crit}$.

3.6 Decision Trees

A decision tree is a classification technique which uses induction to determine key variables for partitioning observations into individual classes [143]. There are three key items in the structure of a decision tree:

1. Internal Node—A decision point using a predictor variable value as the decision criteria, e.g. "pH>7.0"
2. Leaf Node—During tree construction, this is the majority sample class. For cross-validation and testing dataset, this is the predicted class for a sample.
3. Branch—Pathway between internal nodes and leaf nodes based on attribute value.

A notable feature of decision trees is that they can be constructed from heterogeneous datasets, e.g. continuous, discrete, and categorical data can be analysed as a single dataset. An example from industry is the use of decision trees to identify most optimal growth conditions for *E. coli* from parameters including inoculum volume, substrate source, and culture conditions [42].

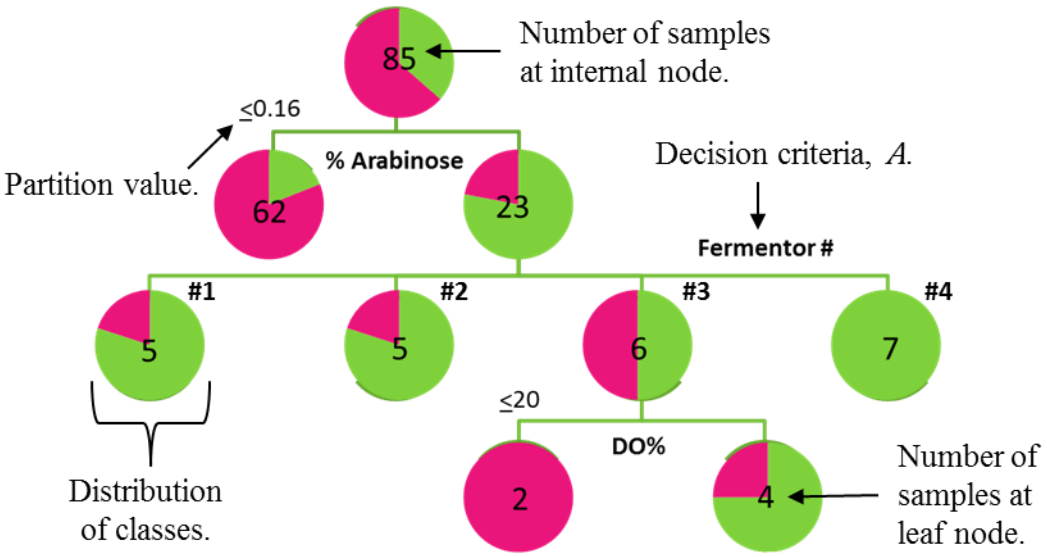


Figure 9. Decision tree to determine growth conditions leading to separation of 85 *E. coli* cultures as "High" or "Low" based on fluorescence using gain ratio [42]. Nodes are number sequential by layer and from left to right, e.g. node numbers in third layer would be 4, 5, 6, and 7 for fermentor numbers #1, #2, #3, and #4 respectively. In this visualisation, the distribution of the two classes can be seen at each internal node and each leaf node. The number of samples at each node is also displayed. It can be seen that the first decision criteria ($\% \, Arabinose \leq 0.16$) resulted in approximately 75% of samples (cultures) being grouped on a single node. This group is of high enough purity that further splitting of the sample subset is halted and the node is a leaf node.

Multiple algorithms exist for decision tree construction. Information Gain is one of the simplest decision tree criteria selection algorithms. The Information Gain algorithm calculates the information gained when choosing a decision criteria by comparing the expected information (entropy) needed to classify a sample (Eq. 3.17) to the information still needed to arrive at exact classification following the split (Eq. 3.18). This is shown in Eq. 3.19 to identify the decision criteria partition value of an attribute x that yields the greatest information gain.

$$Info(X) = -\sum_{i=1}^{w} p_i \log_2 p_i \qquad \text{Eq. 3.17}$$

$$Info_x(X) = \sum_{j=1}^{v} \frac{|X_j|}{|X|} \times Info(X_j) \qquad \text{Eq. 3.18}$$

$$Gain_x(X) = Info(X) - Info_x(X) \qquad \text{Eq. 3.19}$$

where p_i is the non-zero probability that a sample in dataset X belongs to a class, w is the number of distinct classes available, v is the decision criteria value, X_j is the subset of X with samples with attribute value $x_i \geq x_{Partition\,Value}$, $|X|$ is the purity of X, and $|X_j|$ is the purity of X_j [143].

Information Gain is biased towards attributes with many values. Gain Ratio is a variation of the Information Gain algorithm modified to reduce this bias through use of a normalisation term $SplitInfo_x(X)$ (Eq. 3.21). While bias towards many valued attributes is reduced, the Information Gain algorithm is itself biased towards unbalanced splits, i.e. selecting decision criteria which isolate a small proportion of the samples to be classified.

$$GainRatio_x(X) = \frac{Gain(X)}{SplitInfo_x(X)} \qquad \text{Eq. 3.20}$$

where

$$SplitInfo_x(X) = -\sum_{j=1}^{v} \frac{|X_j|}{|X|} \log_2 \frac{|X_j|}{|X|} \qquad \text{Eq. 3.21}$$

A third common algorithm is the Gini Index which selects decision criteria based on the impurity of dataset X at the decision node:

$$Gini(X) = 1 - \sum_{j=1}^{w} p_j^2 \qquad \text{Eq. 3.22}$$

For a partition based on attribute x, dataset X is split into X_1 and X_2 for which Gini Index can be weighted to the Gini Index for the split:

$$Gini_x(X) = \frac{|X_1|}{|X|} Gini(X_1) + \frac{|X_2|}{|X|} Gini(X_2) \qquad \text{Eq. 3.23}$$

A weakness of decision trees is that classes are determined by judging individual variables in a hierarchical manner. This is unlike other multivariate methods, such as PCA and PLS, where all variable dimensions are considered simultaneously. When using decision trees, a high degree of contextual information can be lost when cherry picking variables based on pure partitioning power. A number of related algorithms have been suggested to better cope with the lack of attribute independence when selecting decision criteria [149]. In the Relief algorithm, which is limited to two classes, $W[x]$, the quality estimation for all attributes x, is initially set equal to zero. A sample i is then selected from a total of m samples. The nearest neighbour of the same class (nearest hit, H) and the nearest neighbour of the different class (nearest miss, M) are identified using the cumulative Manhattan distance across all variables. $W[x]$ is then updated to take into account the difference in values for a variable (x) using the following iterative formula:

$$W[x] := W[x] - \frac{|x_i - x_H|}{\max(x_{1,m}) - \min(x_{1,m})} + \frac{|x_i - x_M|}{\max(x_{1,m}) - \min(x_{1,m})} \qquad \text{Eq. 3.24}$$

where x_i, x_M, and x_H are values for variable x for sample i, nearest miss M, and nearest hit H, and $x_{1,m}$ is the full range of values observed for variable$x$. A new sample is then selected and $W[x]$ is updated again. The number of samples $W[x]$ is updated against is userdefined parameter up to m.

The ReliefF algorithm (Eq. 3.25) is an adaptation allowing for more than two classes to be considered by taking a proportional average of the differences between the sample i and the nearest miss for each class. This is done using the prior probabilities of classes estimated from the training dataset, $P(C)$. ReliefF also allows an increase in the number of nearest neighbours compared to sample i to be increased to k.

$$W[x] := W[x] + G + Z \qquad \text{Eq. 3.25}$$

Where

$$G = \sum_{j=1}^{k} \frac{\left|x_i - x_{H,j}\right|}{\max(x_{1,m}) - \min(x_{1,m})} \qquad \text{Eq. 3.26}$$

and

$$Z = \sum_{C \neq Class(i)} \left[\frac{P(C)}{1 - P(Class(i))} * \sum_{j=1}^{k} \frac{\left| x_i - x_{M,j(C)} \right|}{\max(x_{1,m}) - \min(x_{1,m})} \right] \quad \text{Eq. 3.27}$$

A second expansion of the Relief algorithm called RReliefF allows use of the algorithm with regression trees through the incorporation of Bayes probabilities [149].

A further method for addressing to the univariate action of decision trees is to transform the dataset in question using MVDA techniques prior to applying a decision tree algorithm. For example, iterative PLS-decision trees calculate a latent variable to describe the differences between classes [150]. The latent variable becomes the first node attribute and the samples partitioned accordingly. The error matrix from the first latent variable is used to calculate a new latent variable to classify the samples on that node.

3.7 Summary

Research focussed on the application of multiple MVDA techniques detailed in this chapter. For each case study, technique choice was re-evaluated for suitability for the data to be analysed and for study aims. As data originated from an industrial research and development environment and analysis outcomes were to feed back into this environment, techniques were also selected based on interpretability and communication of results, in addition to flexibility and ease of implementation for future use.

Chapter 4. Comparison of pH Measurement Technologies and Extraction of Indirectly Captured Information

4.1 Introduction

The global biopharmaceutical industry's $90 billion worth [151] is dependent on the ability of cells to grow and produce the correct product. A fundamental parameter in cell cultures is pH as it affects cell growth [95] and production rates [97], which in turn affects culture product yield. pH can also affect product quality [96], potentially leading to rejection of the final product due to failure to meet release criteria. These issues could be addressed by improved pH understanding and control [152]. However to effectively apply a pH strategy, measurement technologies must give reliable and accurate readings.

The most commonly used sensor for pH measurement is the potentiometric pH electrode probe, comprising two electrodes: an indicator electrode with a glass membrane and a reference electrode [153]. pH is measured by immersing the probes into a sample to create a Galvanic cell with the difference between the electrode voltages denoting the potential. Using a modified form of the Nernst Equation, the sample's pH is then calculated [154]. All pH measurement technologies at Lonza for 10L and larger bioreactors used potentiometric pH electrode probes for online and offline monitoring.

An aspect of pH measurement frequently taken for granted is consistency between technologies. One possible scenario in industry is that different sites may use different technologies when manufacturing the same product. A similar situation may arise in a single laboratory when individual scientists prefer one of two available technologies over the other.

In Lonza's operating procedures, two offline pH measurements were made during daily offline monitoring of cultures. The first measurement was made with a Radiometer Analytical PHM220 pH meter (Radiometer Analytical, France) connected to a Mettler-Toledo pH probe (Mettler-Toledo, UK). The second measurement was made using a NOVA Biomedical Bioprofile 400 (NOVA Biomedical, MA, USA). It was assumed that each available pH measurement technology had an accuracy of ± 0.01 pH units [155] and that samples were sufficiently mixed to be considered homogenous at this measurement resolution. From this, the maximum allowable difference in readings between the two pH measurement technologies on the same sample due to pure instrument error was ± 0.02 pH units. However in the cell culture robustness study forming the basis of the

research presented in this chapter, nearly 60% of differences in pH reading by the two offline technologies fell outside the allowable error band (Figure 1). Multivariate data analysis (MVDA) was to be used to investigate this undesired area of variation. A notable constraint on study was to restrict analysis to a pre-existing dataset generated as part of standard cell culture robustness study and thus demonstrate if MVDA could be used to extract new information.

Analysis focussed specifically on potential effects from component concentrations in samples or sample temperatures. Effects from sample handling [156], sterilisation techniques [157], or variations specific to probe age or individual probes could not be considered as these are not captured in standard operation data collection. Offline probes were replaced as necessary during the robustness study; hence this study investigates differences between multiple random pairings of offline pH probes.

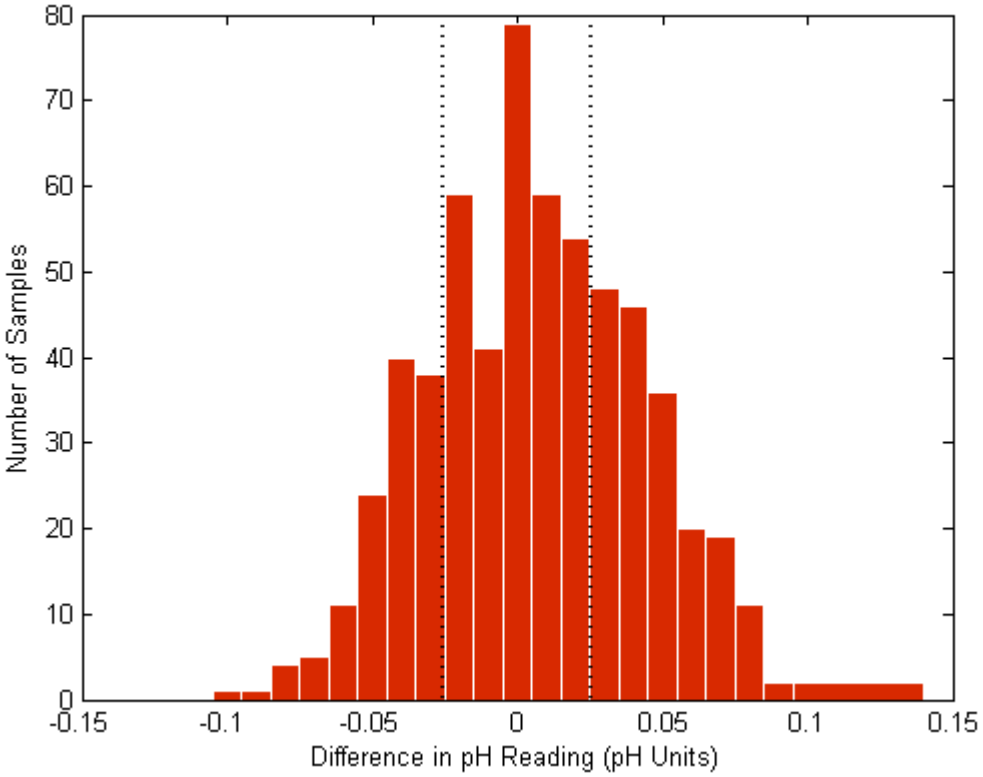


Figure 10. Histogram of Differences in Offline pH Readings by Two Offline Technologies. Dotted lines indicate boundaries for differences due to instrument error (±0.02 pH units).

4.2 pH and Temperature

The pH of a sample is not constant with respect to temperature. Without compensation, this temperature-pH relationship results in incorrect readings with the potential for the pH controller to force a change on a system which is at the pH setpoint in an extreme situation (e.g. technical error with temperature control feedback). This is further complicated as the pH-temperature relationship varies based on component concentrations [158]. Thus in 1947, when Rosenthal presented equations to be used by medical persons to determine the pH of human blood and plasma samples at physiological temperatures after samples had cooled to ambient conditions, three equations were required to describe the different pH-temperature relationships for whole blood, plasma, and derived plasma [159]. In contrast to this, an earlier study by Yoshimura and Fujimoto in 1937 found that blood and plasma samples from rabbits were found to have pH-temperature relationships similar enough to be addressed with a single equation, where they also note that the loss of CO_2 from the samples would also effect the measured pH [160]. Both studies assumed simple linear relationships, e.g. the mean change in plasma pH was calculated to be -0.118 per +1°C. While these simple linear relationships were deemed appropriate for the temperature range across which compensation was to be applied, pH-temperature relationships are not truly linear [161].

pH-temperature compensation could be performed manually, however it is more for pH measurement technologies have a built-in pH-temperature compensation function. In the presented study, a Mettler-Toledo pH probe (Mettler-Toledo, UK) physically paired with a temperature probe was connected to a Radiometer Analytical PHM220 pH meter (Radiometer Analytical, France). The system was calibrated daily using standards of known pH; the temperature of those standards was used to create a pH-temperature compensation. When an offline measurement was made, the sample's temperature was also taken and the compensation applied by the pH meter.

A different method was used by the NOVA Bioprofile 400 [162]. The sample's temperature (T) was entered at the user interface. The sample was then heated to 37 °C. The pH of the heated sample was measured then pH-temperature compensation was applied using the following equation:

$$pH_{corrected} = pH + [-0.0147 + 0.0065 * (7.400 - pH)] * (T - 37) \qquad \text{Eq. 4.1}$$

The concern with such compensation is the assumption all samples have the same pH/temperature relationship. In reality, different chemicals have different pH/temperature

relationships [159]. Variation in relative concentrations in a chemical mixture (e.g. supernatant) could affect the required pH/temperature compensation for a pH measurement technology. This has a long established concern which prompted Rosenthal to use his 1947 paper "to call attention to a misconception on the part of some who use commercial pH that are equipped with "temperature correction" controls" and that "Simply setting the pointer at 38° does not solve the problem of finding pH_{38} while the sample is at room temperature" [159]. Furthermore if two technologies assume different pH-temperature compensations, they are likely to give different readings for the same sample regardless of additional possible effects.

4.3 Data

The data forming the basis of this study were generated from the development of a bespoke fed-batch process for DHFR-CHO cell line. As part of the development process, a series of control and experimental conditions were tested to evaluate process limits, hence the study was termed a process limits evaluation (PLE) study. Data were taken from 10 L bioreactor cultures with 25 control cultures operating under normal process conditions and 23 experimental cultures operating with deviations introduced to normal process conditions.

The key feature of the process was a deliberate alteration in the setpoint for temperature control once a minimum viable cell concentration was reached. The control bioreactor cultures were initially maintained at pH 7.0 and 36°C. When a minimum viable cell concentration (determined by daily offline sampling) was reached, the pH and temperature setpoints were moved to pH 6.91 and 30°C respectively. Three different bolus additions were made to each bioreactor culture. Bolus A was added when the pH and temperature setpoints were adjusted, bolus B was added on Day 4 of the culture, and bolus C was added on Day 7 of the culture.

For the experimental bioreactor cultures, deliberate deviations from standard operation were introduced (summarised in Table 9 and Table 10). These included changes directly captured in daily monitoring data, e.g. increased operating temperature, decreased operating temperature, omission of the shifts in pH and temperature setpoints. Other changes not directly captured in the routine daily monitoring data outlined in Chapter 2 were collected as meta-data, e.g. use of expired medium, alterations to feedrates, or alterations to days on which media bolus feeds were added.

The cultures were monitored through daily offline samples as outlined in Chapter 2. In addition to cell concentrations and concentrations of key metabolites, measurements for pH are made using two offline pH measurement technologies:

- A Radiometer Analytical PHM220 with a Mettler-Toledo probe
- A NOVA Bioprofile 400

A total of 48 fed-batch cultures were cultured for an average of 15 days under the aforementioned range of conditions with several cultures reaching harvest criteria (severely declined viability and viable cell concentrations) earlier than others due to effects from experimental conditions. A total of 785 daily monitoring samples were collected and analysed during this time. For each sample, 12 daily sampling measurements were recorded. The data were collated into a single 785x12 matrix in Excel (Microsoft) and analysed using Minitab 15.

4.4 Removing Daily Adjustments to Online pH Reading

Offline pH technologies were calibrated daily. The online pH technology was calibrated before use. Daily adjustments to the online pH technology were made by comparing the online pH reading to the bench offline reading of a sample. If the difference in pH readings was equal to or greater than ± 0.02 pH units, i.e. outside instrument error, then the bench offline reading was used to adjust the online probe to give the same reading. If the difference in readings was less than ± 0.02 pH units, no adjustment was made.

It was necessary to know what the online reading would have been if no adjustments had been made. This 'true' reading could be calculated due to the capture of daily adjustments in paper records and then using Eq. 4.4 and Eq. 4.5.

$$P_i = P_{i-1} + D_{pH} \qquad \text{Eq. 4.4}$$

$$D_{pH} = R_i - (R_{i-1} + A_{i-1}) \qquad \text{Eq. 4.5}$$

where

P_i	'True' online pH reading for sample i
R_i	Online reading for sample i
D_{pH}	'True' change in pH
A_{i-1}	pH adjustment from previous sample *i-1*
T_i	Day sample taken – rounded down to nearest whole number Note: If $T_i = 0$, then $P_i = R_i$

Culture ID	Culture Conditions	Comments
3.1	Control	
3.2	Control	
3.3	Control	
3.4	Control	
4.1	Control	
4.2	Control	
5.1	Control	
5.2	Control	
5.3	Low Temp	
5.4	Low Temp	
5.5	Low pH	Removed for AS2 dataset – see §4.6
5.6	Low pH	Removed for AS2 dataset – see §4.6
5.7	High Temp	
5.8	High Temp	
6.1	Control	
6.2	Control	
6.3	Control	
6.4	Control	Contaminated – removed from all datasets
6.5	Control	Strong drift by online probe – removed from datasets
6.6	Control	

Table 9. Summary of cultures and culture conditions. Control culture initial operating conditions 36 ˚C and 7.0 pH units. When the viable cell concentration met a designated minimum, the temperature was reduced and the pH level lowered. Note on Culture ID: First value refers to round number, second value refers to culture number e.g. 3.1 is culture 1 of round 3.

Culture ID	Culture Conditions	Comments
7.1	Control	
7.2	Control	
7.3	Control	
7.4	Control	
7.5	High pH	Removed for AS2 dataset – see §4.6
7.6	High pH	Removed for AS2 dataset – see §4.6
7.7	High pH	Removed for AS2 dataset – see §4.6
7.8	Low pH	Removed for AS2 dataset – see §4.6
7.9	High Seeding	
7.10	High Seeding	
8.1	Control (T=32)	
8.2	Constant pH 7.0	Removed for AS2 dataset – see §4.6
8.3	Low DOT	
8.4	Low DOT	
8.5	High DOT	
8.6	High DOT	
9.1	Control	
9.2	Increased Feed	
9.3	Increased Feed	
9.4	Modified Feed Strategy I	Removed for AS2 dataset – see §4.6
9.5	Modified Feed Strategy II	Removed for AS2 dataset – see §4.6
9.6	Modified Feed Strategy III	Removed for AS2 dataset – see §4.6
C1	Low Generation Number	
C2	Low Generation Number	
C3	Low Generation Number	
A1	High Generation Number	
A2	High Generation Number	
A3	High Generation Number	

Table 10. Summary of cultures and culture conditions (continued). Control culture initial operating conditions 36 ˚C and 7.0 pH units. When the viable cell concentration met a designated minimum, the temperature was reduced and the pH level was lowered. Note on Culture ID: First value refers to round number, second value refers to culture number e.g. 7.1 is culture 1 of round 7.

4.5 Missing Data

Approximately 4% of data were missing (560 points out of a raw data total of 14130). The majority of missing data were due to issues with sensors used for variable measurement. Missing data could also be attributed to Lonza re-sampling policies, where a daily sampling may be repeated but only a subset of the variables monitored via daily monitoring are recorded.

The software used (Minitab 15) automatically excluded samples where variables were missing. This meant that as variables were removed through significance testing the number of samples used could potentially increase (Table 11). This variation was allowed as the maximum number of samples would be used with each iteration of statistical significance testing, potentially increasing the strength of the tests.

4.6 Division of Dataset

The change in temperature and pH set points provided a natural splitting point in the data. The data taken from before the shift and the data taken from after the shift reflect two different biochemical states. These different states are referred to here as 'operating conditions': 'All,' 'Before Shift' (BS), and 'After Shift' (AS).

A variety of changes were made to bioreactor operating conditions, referred to here as 'culture conditions'. Culture conditions included changes in temperature shift, pH shift, and other parameters, e.g. not undergoing the step change in temperature or operating a higher DOT setpoint as seen in Table 9 and Table 10.

The dataset AS was further subdivided into 'After Shift 1' (AS1) and 'After Shift 2' (AS2). AS1 contained AS data from all cultures. AS2 contained AS data from cultures where reactor conditions were directly captured in the inputs used only. As pH was excluded as an input due to concerns over bias, cultures not operating at the control pH set point were removed. Cultures using modified feed strategies were also removed.

In summary, the dataset 'All' contained all data from all samples, BS contained data from all samples taken before the change in operating conditions, AS1 contained data from all samples taken after the change in operating conditions, and AS2 contained data from all samples taken after the change in operating conditions but excludes samples where experimental conditions were not directly captured in the dataset.

I	Variable	A	B	C	D	E	F	G	H
Sample	1			●					●
	2	●					●		
	3						●		
	4								

Variable F removed as least significant.
Samples 1 and 2 excluded from new model creation.

II	Variable	A	B	C	D	E	F	G	H
Sample	1			●					●
	2	●					●		
	3						●		
	4								

Variable H removed as least significant.
Samples 1 and 2 excluded from new model creation.

III	Variable	A	B	C	D	E	F	G	H
Sample	1			●					●
	2	●					●		
	3						●		
	4								

Variable C removed as least significant.
Sample 2 excluded from new model creation.

IV	Variable	A	B	C	D	E	F	G	H
Sample	1			●					●
	2	●					●		
	3						●		
	4								

Variable A removed as least significant.
All samples included in new model creation.

V	Variable	A	B	C	D	E	F	G	H
Sample	1			●					●
	2	●					●		
	3						●		
	4								

Table 11. Effect of Missing Data on Samples Used to Model Response. Shading indicates exclusion from model creation. ● indicates a missing value. For I, model creation includes all variables A to H, causing the software (Minitab) to exclude samples 1 to 3. F is identified as the least statistically significant variable and removed from the model inputs. Sample 3 can now be included in the creation of the new model. This is repeated until only variables of the desired statistical significance remain in the model.

4.7 Development of New Variable: Osmolality Residuals

During the initial analysis to determine potential causes of differences in pH readings by offline pH measurement technologies, it was found that, for the four datasets tested, inclusion of osmolality resulted in an increase in the R^2 for a testing dataset by between 0.01 to 0.07 depending on data subset used (Table 12). A similar result was note when considering PLSR models to predict the difference between pH readings by the two offline technologies, with the R^2 for testing datasets improved by between 0.01 to 0.04 when osmolality was included as a variable (Table 13). This suggested that osmolality played a part in the difference in pH readings by different offline pH measurement technologies, even though the impact of that part varied. However, further interpretation of these results was problematic due to the indiscriminate nature of osmolality.

Osmolality is the concentration of solutes in a sample measured in osmoles of solute per kilogram of solvent [130]. It is an indiscriminate measurement as it does not specify how much of any specific component is present, i.e. an osmolality reading of 100 mOsm/kg does not specify whether there are 100 mOsm/kg of component A, B, D, or a mixture of all three. Due to the comprehensive nature of osmolality, the value contains information already captured (e.g. glucose concentration) as well as data not captured (e.g. background media components).

Dataset	**Osmolality Model R^2**	**No Osmolality Model R^2**	**ΔR^2**
All	0.36	0.29	0.07
Before Shift	0.35	0.32	0.03
After Shift 1	0.40	0.39	0.01
After Shift 2	0.42	0.39	0.03

Table 12 Effects of excluding osmolality in MLR on cross-validation R^2 during initial analysis. For each dataset, a MLR model was created using all available variables to predict the difference between pH readings by a Mettler-Toledo probe with Radiometer Analytics PHM220 and a NOVA Bioprofile 400 (B-N). Another MLR model was created to predict B-N with osmolality excluded from the dataset.

Dataset	**Osmolality Model R^2**	**No Osmolality Model R^2**	**ΔR^2**
All	0.24	0.22	0.02
Before Shift	0.24	0.23	0.01
After Shift 1	0.42	0.38	0.04
After Shift 2	0.44	0.40	0.04

Table 13 Effects of excluding osmolality in PLSR on cross-validation R^2 during initial analysis. For each dataset, a PLSR model was created using all available variables to predict the difference between pH readings by a Mettler-Toledo probe with Radiometer Analytics PHM220 and a NOVA Bioprofile 400 (B-N). Another PLSR model was created to predict B-N with osmolality excluded from the dataset.

It was theorised that contributions to osmolality by variables not directly monitored via standard daily sampling practices could be extracted if contributions to osmolality by those variables that were directly monitored were removed from the osmolality measurement. The extracted information could then be used to evaluate the impact of indirectly monitored variables in subsequent analyses, specifically the impact on differences in pH measurements by offline pH measurement technologies.

There are multiple ways of measuring osmolality. The two most suited to biological samples are freezing point osmometry (FPO) and vapour pressure osmometry (VPO) with FPO the industry preferred method [130]. The NOVA Bioprofile 400 is designed to measure glucose (Gluc), lactate (Lac), ammonia (NH_4^+), sodium (Na^+), and potassium (K^+) concentrations in addition to pH levels. This allows a component calculator to be included. The unit performs a simple linear combination using these measurements as seen in Eq. 4.2 [140].

$$Osmolality = 1.86([Na^+] + [K^+] + [NH_4^+]) + \frac{[Gluc]}{0.18} + \frac{[Lac]}{0.09} + c \qquad \text{Eq. 4.2}$$

where

$Osmolality$	Osmolality calculated (mOsm/kg H_2O)
Lac	Lactate concentration (g/L)
$Gluc$	Glucose concentration (g/L)
NH_4^+	Ammonium concentration (g/L)
Na^+	Sodium concentration (mmol/L)
K^+	Potassium concentration (mmol/L)
c	Calculated constant (mOsm/kg H_2O)

There are many components which may be present in a sample which are not measured by the NOVA Bioprofile 400. There are also components known to affect osmolality which are measured by the unit but not included. For these reasons, the component calculator is an unreliable measure of osmolality for a biological system and can only be counted as a general estimate.

4.7.1 Osmolality Model Residuals

Data collected from daily monitoring included freezing point osmolality (FPO) readings where an Osmomat (Gonotec) was used to directly measure sample osmolality. It was decided to create a component calculator to model the FPO measurement from the remaining daily monitoring data in an attempt to extract information concerning variables

not directly measured as part of daily monitoring procedure, e.g. bolus feed shot concentrations, metabolites, and media composition.

Multiple linear regression (MLR) was employed with all known component concentrations as input variables and the FPO reading as the response variable. The MLR model was refined through the use of statistical significance testing so that only variables with a *p*-value less than 0.05 were included. This was performed in an iterative manner removing one variable at a time (Figure 11). The final model (Eq. 4.3) accounted for 83.6% of variation in osmolality in the dataset.

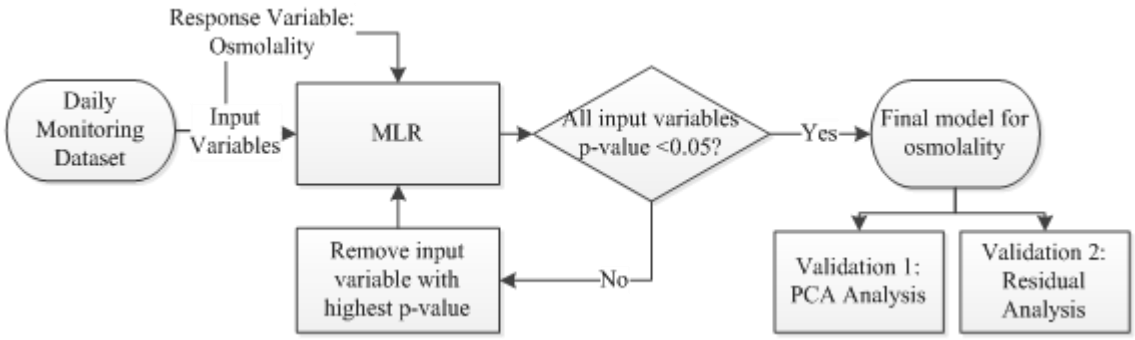


Figure 11. Component Calculator Creation using MLR and Iterative Statistical Significance Testing

$$Osmolality_M = 236 - 0.298pO_2 + 0.243pCO_2 + 74.4Gln - 14.6Gluc + 254NH_4^+ + 0.263Na^+ + 4.27K^+ + 7.95TCC \quad \text{Eq. 4.3}$$

where

$Osmolality_M$	Osmolality calculated by model (mOsm/kg H_2O)
pO_2	Oxygen partial pressure (mmHg)
pCO_2	Carbon dioxide partial pressure (mmHg)
Gln	Glutamine concentration (g/L)
$Gluc$	Glucose concentration (g/L)
NH_4^+	Ammonium concentration (g/L)
Na^+	Sodium concentration (mmol/L)
K^+	Potassium concentration (mmol/L)
TCC	Total Cell Concentration (10^6 cells/mL)

The purpose of the developed model was to explain osmolality values based on all available variables and therefore not a true *a priori* model. Several variables were included when creating the model for osmolality that would not typically appear in a

component calculator, for example viable and total cell concentrations. These were included as measurements related to cell growth and condition may act as a suitable substitute variable for unmonitored products of cell metabolism affecting sample osmolality. Temperature was rejected as a possible variable as FPO readings are made at freezing point, not sample temperature. Iterative significance testing with a $p < 0.05$ led to the removal of variables for one of two reasons:

1. The variable was statistically insignificant in the model.
2. Strong correlations between variables caused the variable in question to be rejected as statistically insignificant, i.e. multiple variables were competing to provide similar information to the model.

Competition between correlated variables is the likely reason for the rejection of lactate as statistically significant in the model despite lactate being known to affect osmolality as seen in Eq. 4.2. Specifically, in the datasets analysed lactate concentration was generally closely correlated with glucose concentration.

There is a notable difference between the presented model and established osmolality theory that requires further explanation. According to theory, the glucose coefficient should be positive as an increase in glucose concentration is an increase in the solute concentration and hence an increase in osmolality. In Eq. 4.3, the glucose coefficient is negative. This difference was due to the model capturing the behaviours observed in the dataset, e.g. as a typical culture progressed, osmolality increased while glucose concentration decreased. Furthermore, in MLR the sign of a variable does not necessarily indicate the direction of the relationship between the variable and the response. Coefficient direction can be altered depending on correlation with other variables.

The differences between the predicted osmolality and the recorded osmolality are the residuals (errors) for the model. The use of the osmolality model residuals as a variable in place of osmolality is an attempt to allow variation in uncaptured data to be evaluated. This could not be done with osmolality due to its high degree of correlation with other variables.

Two forms of evidence are presented to justify the inclusion of the osmolality residuals as a new variable. The first is based on a mathematical approach that uses principal component analysis (PCA). The second is a more heuristic, logical argument based on anticipated behaviour and knowledge of cell culture behaviours.

4.7.1.1 Validation of Osmolality Model Residuals using PCA

Principal component analysis (PCA) is a multivariate data analysis technique where dimensionality reduction is achieved through the creation of orthogonal linear combinations of variables termed principal components (PC)[3]. When PC variable loadings are plotted in two dimensional space, correlated variables will cluster together. Negatively correlated variables will be at opposite points across the origin; variables with little or no correlation will be at approximately right angles.

PCA was performed using daily monitoring data. Figure 12 and Figure 13 display loading for PC1 and PC2 when osmolality and osmolality residuals were included as inputs respectively. In both Figure 12 and Figure 13 the oxygen-based variables DOT and pO_2 were shown to be negatively correlated with respect to lactate concentration. This indicated that the PCA model created provided information that was correct from an understanding of the physical system. In Figure 12 it was seen that osmolality's location in the loading plot was in the centre between a loose cluster (Day, pCO_2, NH_4^+, K^+) and a second, more tightly defined cluster (Gln, Gluc, TCC, VCC, Na^+). This was due to the correlation with variables in both clusters in keeping with the relationships shown in Eq. 4.3. Osmolality also had a strong influence on scores due to high loading values.

In Figure 13, relative positions between variabes remained largely unchanged from those in Figure 12 apart for an inversion of the Y-axis. This inversion is a theoretical issue of negligble significance whereby principal components which are not unique can change signs. It can be seen that "osmolality residuals" ("Osmo Res") was located in a position of less correlation with the two clusters with reduced influence on score positions through reduction in loading values. This is in keeping with the theory that the variable osmolality residuals contains information not available in (i.e. uncorrelated with) other variables.

[3] See Chapter 3. Statistical Methods for a more detailed description.

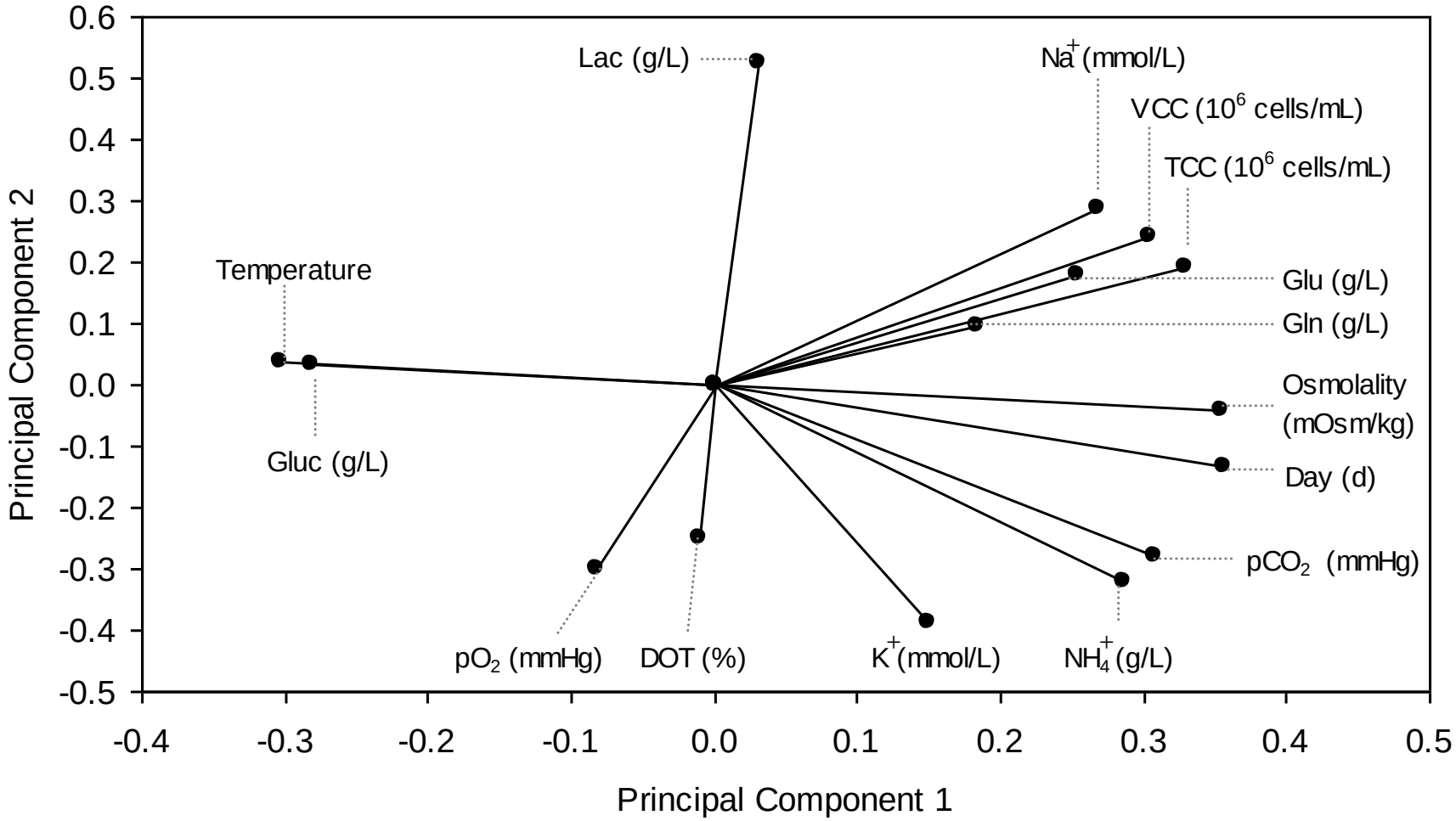


Figure 12. PCA loading plot for PC1 and PC2 when osmolality is used as a variable. Clustering of variables indicates positive correlations between those variables. The position of osmolality indicates positive correlations with multiple variables, however not all those variables are positively correlated with each other.

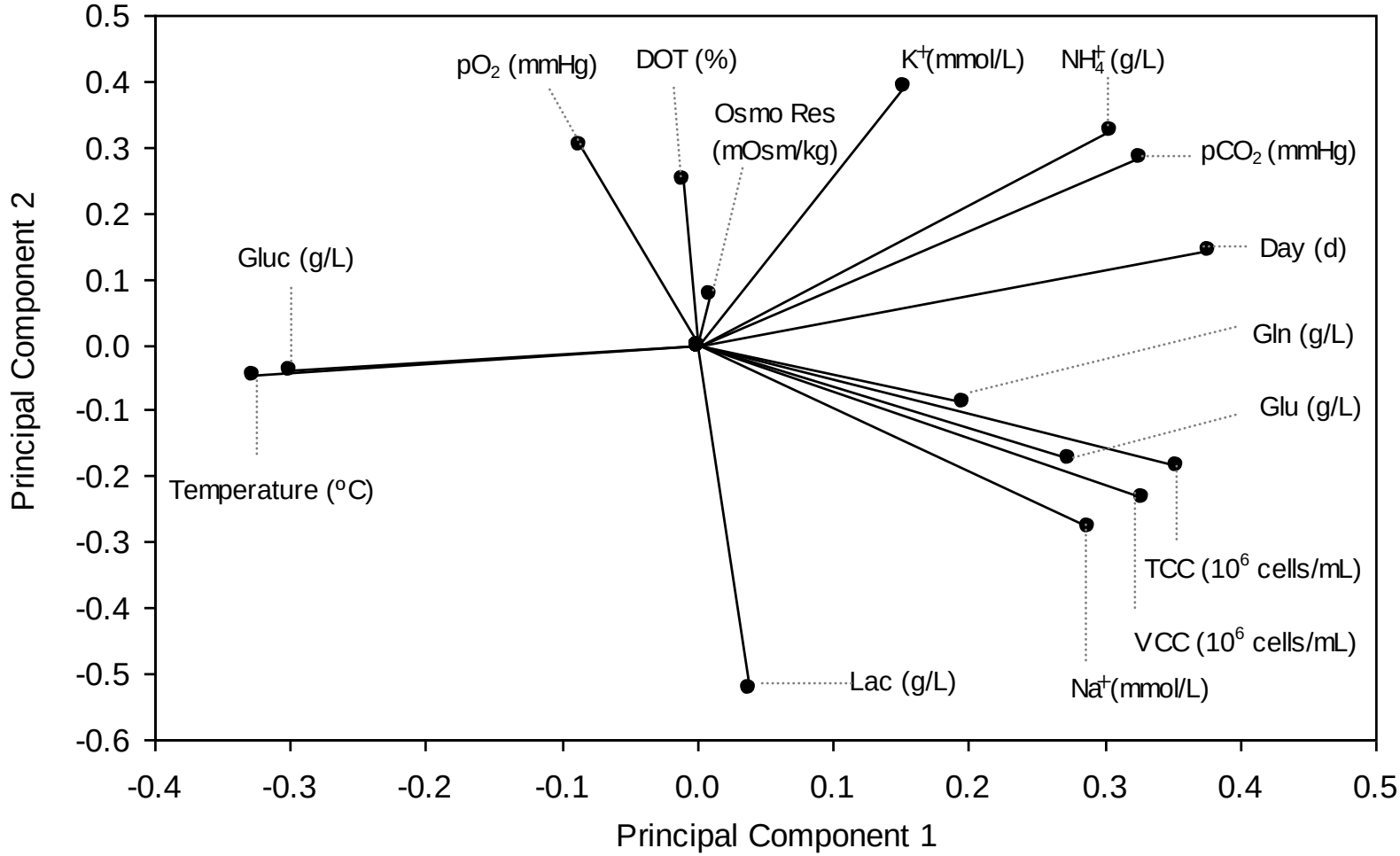


Figure 13. PCA loading plot for PC1 and PC2 when osmolality residuals ("Osmo Res") is used as variable in place of osmolality. The relative positions of the other variables are similar to the relative positions in Figure 12, save for a negligible inversion of the Y axis.

4.7.1.2 Validation of Osmolality Model Residuals Through Time Series Analysis

In the feed strategy employed, three chemically distinct bolus feed additions were added to the culture following a change in the reactor temperature and pH operating conditions. All reactors received bolus A on the day of the change in operating condition, bolus B on day 4, and bolus C on day 7. This pattern was identified when osmolality model residuals were plotted against elapsed time. Figure 14 and Figure 15 show the osmolality residuals against elapsed time for the cultures undergoing the setpoint change in operating conditions on day 3 and on day 4 respectively. Bolus additions are indicated with the following arrow colours: orange (bolus A), blue (bolus B), green (bolus C). Note that in Figure 14 and Figure 15 elapsed time is displayed as the day of the sample and not the precise elapsed time in hours. This is to allow a clearer comparison of general trends.

In the first days of culturing, osmolality residual values decreased in a manner thought to indicate consumption of medium by the culture during the exponential growth phase. As medium composition was not directly measured, the impact of medium composition on osmolality measurements could not be accounted for by the model. Therefore information on medium composition information would lie in the model's residuals.

The decrease in residuals was seen in all cultures in all days until the addition of bolus A. For the reactors shifted on day 3, there is a continued decrease after the addition of bolus A with a sharp increase following the addition of bolus B. For cultures shifted on day 4, both the continued decrease after bolus A and the sharp increase after bolus B were absent. This indicated that the effects of the boluses in the osmolality model cancelled to some extent.

For both sets of cultures, the osmolality residuals decreased following the addition of bolus B until day 7 when bolus C was added to cultures. A sharp increase in residual values occurred following the addition of bolus C, which is followed by a gradual reduction in the osmolality residuals over the remaining days of culture.

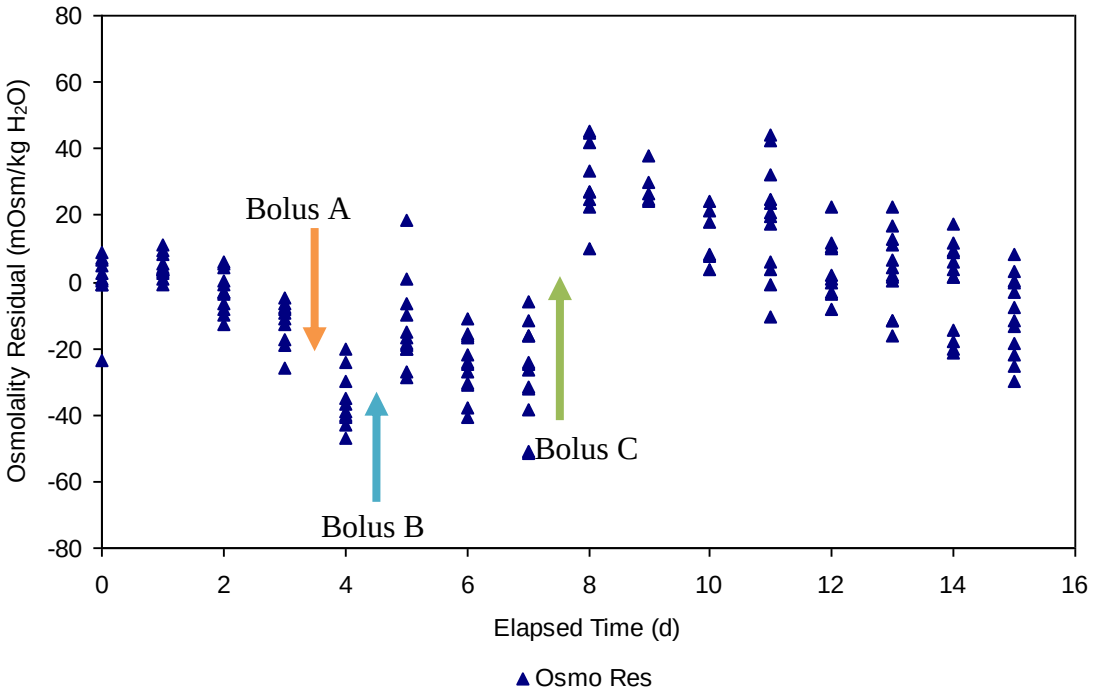


Figure 14. Osmolality model residuals for 15 cultures with setpoint changes on Day 3. Three chemically distinct bolus additions made after daily sampling are indicated with labelled arrows. A general decrease in residual values follows the addition of Bolus A. Increases in residual values follow the additions of Bolus B and Bolus C.

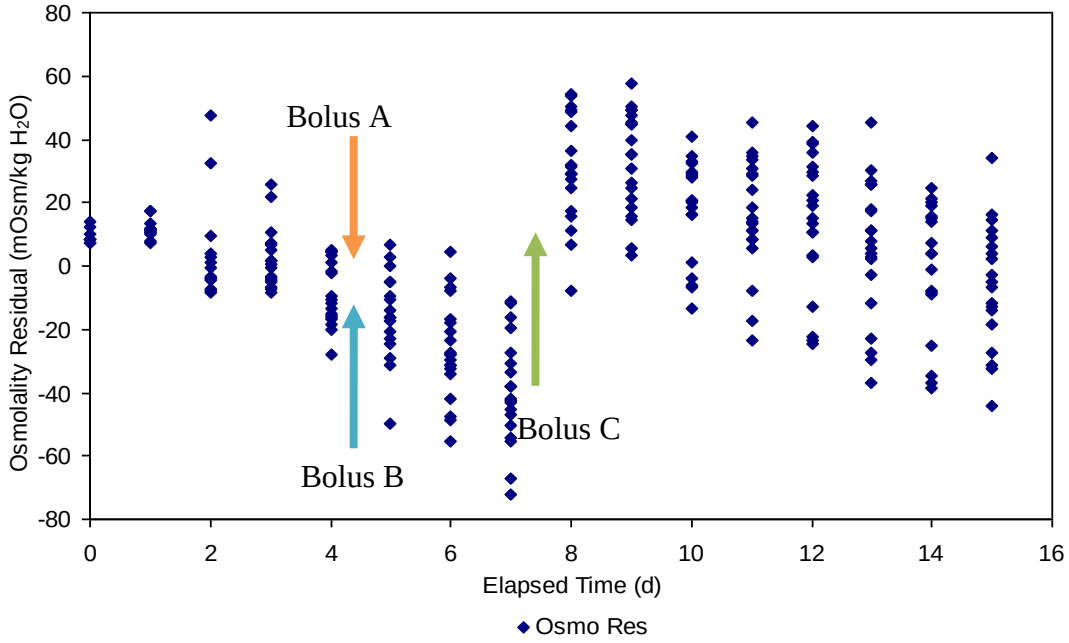


Figure 15. Osmolality model residuals for 24 cultures with setpoint changes on Day 4. Three chemically distinct bolus additions made after daily sampling are indicated with arrows. The decrease in residual values following the addition of Bolus A (seen in Figure 14) is absent, as is the increase that followed Bolus B, indicating the effects of Bolus A and Bolus B on osmolality residual have cancelled out. The increase in residual values following Bolus C remains present.

4.7.2 Limitations of Osmolality Residuals as a Variable

Osmolality measurements lack discrimination, i.e. it is not possible to identify which components contributed to the measurement nor to what extent. The calculated variable osmolality residuals is similarly indiscriminate.

Osmolality residual values capture a variety of components including Bolus A, bolus B, bolus C, culture medium components, and by-products of cell metabolism. The relative concentrations of these components contributing to the osmolality residuals are not constant. Due to this variation, it may be that osmolality residuals do not have a consistent impact, e.g. on the discrepancies in pH readings seen in 7.1. For example, a 40 mOsm/kg H_2O residual caused by bolus A may not affect readings to the same extent as a 40 mOsm/kg H_2O residual caused by bolus B. Extracting this information was not possible.

Furthermore, the osmolality model presented in Eq. 4.3 was specific to the dataset analysed. Due to the number of variable which might influence which are not directly captured, e.g. media composition or metabolism by-products, two projects cannot be assumed to have the same indirectly captured information. Hence, an osmolality model created using one project's data cannot be assumed transferable to another project.

It may be possible to create a more generalised model from a dataset with a suitably varied background, e.g. a single host cell line expressing different products using a common process platform would have variation in the indirectly captured osmolality contributions due to product-specific effects on metabolism by-products. While a suitably varied dataset was encountered during a process platform investigation using historical data, osmolality was not typically included during daily monitoring records at that time. Hence it was not possible to test whether a generalised osmolality model could be produced during the research period presented in this thesis.

Finally, osmolality residuals could be a more effective variable if the time between the NOVA measurements and the osmometer measurement could be taken into account. For example, two samples have the same composition when the NOVA measurements are taken. One sample is immediately measured using FPO; the other sample is measured 1 hour later. The osmolality for the second sample may not be the same as the osmolality of the first sample due to changes (e.g. degassing, metabolism) in that hour. In such a scenario, variation in time between sampling and osmolality reading may affect the consistency of statistical significance of a variable in an osmolality model.

4.7.3 Osmolality Model Residuals Conclusions

Additional knowledge concerning indirectly monitored variables was extracted from the pre-existing dataset. This was achieved through the comparison of recorded FPO measurements with osmolality values estimated through the use of a component calculator constructed from daily monitoring data. The extraction of known contributions to osmolality from unknown contributions allows the effects of those unknown contribution to be considered in the overall aim of determining possible causes of differences in readings by different pH measurement technologies.

4.8 Modelling Differences in pH Readings by Different Technologies

For each dataset, All, BS, AS1, and AS2, MLR was used to create a model to predict B-N, the difference in pH reading by the two offline pH measurement technologies. The first iteration of the model included Elapsed Time (h), temperature, DOT, pO_2, pCO_2, glutamine concentration, glutamate concentration, glucose concentration, lactate concentration, NH_4^+ concentration, Na^+ concentration, K^+ concentration, viable cell concentration, total cell concentration, and osmolality residuals as model inputs. The MLR model was then refined through the use of statistical significance testing so that only variables with a *p*-value less than 0.05 were included. This was performed in an iterative manner removing one variable at a time.

In the same method, models were created to the difference between the Radiometer PHM220 offline pH measurement technology reading and the true online pH reading (B-P) and difference between the NOVA Bioprofile 400 offline pH measurement technology reading and the true online pH reading (N-P).

4.8.1 Results and Discussion

Variables identified as statistically significant in the final reduced models are indicated by shading in Table 14. These models do not explain 100% of the differences between the two pH technologies; however they do indicate that sample composition and condition can affect agreement in pH readings by two pH measurement technologies.

A number of variables appear to be statistically significant only for specific combinations of dataset and response modelled. For example DOT was only significant for the response B-P when the ‘All’ dataset was used. Likewise ammonia was significant for the response B-P when either the ‘Before Shift’ (BS) or ‘After Shift 2’ (AS2) sets were used but not the ‘All’ set.

It is possible to rationalise several of these inconsistencies. First gas-based measurements have a tendency to be noisy due to degassing of the sample during handling. The value listed when the sample was analysed by the NOVA Bioprofile 400 may not be the same as when each pH reading was taken with the bench offline technology or the online technology. This could affect the variable's calculated statistical significance.

Second, some compounds were not present or were only present in low concentrations in the BS dataset. This could affect their statistical significance compared to the later AS2 or full All datasets.

Third, when iterative statistical significance testing is used for MLR model reduction, correlation between variables can cause variables to be rejected as statistically insignificant. As correlated variables provide similar information to the model, it can be advantageous to retain only one variable from a group of correlated variables. This variable may have an impact on the difference in pH measurement or simply be correlated with another variable which does have an impact.

The following conclusions were drawn from the analysis. The consistent significance of "osmolality residuals" indicates that certain components not directly monitored have a statistically significant impact on the difference in pH measurements. It is thought that osmolality residuals behaviour in the pre-bolus BS data were primarily caused by culture medium and that osmolality residuals behaviour in the AS data were primarily due to the bolus additions.

An unanticipated benefit of the calculated variable 'osmolality residuals' was a reduction in variation in usable sample numbers during iterative model generation. The inclusion of 'osmolality residuals' effectively acted as a filter, whereby a sample needed to have all variables used in the calculation of 'osmolality residuals', regardless of whether those contributing variables were retained in the subsequent model for predicting differences in pH reading. This benefit would only be active so long as 'osmolality residuals' is retained as statistically significant, the consistent significance of 'osmolality residuals' allowed greater user confidence in statistical results as sample size variation was reduced.

A second key variable identified as statistically significant was temperature. Before the change in temperature and pH operating setpoint, temperature did not have a statistically significant impact on differences in pH readings. After the reactor operating temperature was reduced, the effect of temperature was statistically significant. One possible explanation for this result is a difference in pH-temperature compensation methods used

by the technologies. As the degree of compensation required increases, any difference in compensated values will similarly increase.

Temperature was identified as statistically significant when the bench technology or the online technology were compared to the NOVA and the AS2 data is used. This was when the effects of different temperature compensations would be most emphasised as the shift in operation conditions includes a drop in temperature. Hence the gap between the temperature when the online reading (at culture conditions) or bench reading is taken (at or below reactor conditions) and the NOVA (37 °C) increases.

Finally it must be noted that the discrepancies in pH readings between the bench and NOVA technologies were more accurately modelled than the discrepancies between the online technology and either offline technology. The models comparing offline technologies tended to significantly outperform those comparing either offline technology to the online technology.

Model Set / Variable	ALL			BS			AS2		
	B-P	B-N	N-P	B-P	B-N	N-P	B-P	B-N	N-P
Constant									
Time									
Temp									
DOT (%)									
pO_2									
pCO_2									
Gln									
Glu									
Gluc									
Lac									
NH_4^+									
Na^+									
K^+									
TCC									
Osmo Res									
Model R^2	13.8%	38.9%	15.1%	19.6%	45.8%	39.1%	7.9%	46.5%	13.1%

Table 14. Variables identified as statistically significant ($p < 0.05$) when using iterative MLR and significance testing method (indicated by shading). Results are given for three dataset (ALL, BS, AS2) and three responses (B-P, B-N, N-P). It was observed that temperature was a significant variable when modelling data taken from after a change in operation temperature was introduced. The new variable Osmo Res was significant in all models comparing two readings by technologies using different compensation methodologies (B-N, N-P).

4.9 Conclusions

While much of what causes the discrepancies between the competing pH measurement technologies is still not understood, it has been shown that pH technologies are not necessarily interchangeable. If two different technologies are to be used in conjunction, e.g. the NOVA is used for offline measurements and a different pH technology is used for online measurements, differences in pH measurement could be caused by pure instrument error, drift by the online technology, sample composition and condition, or a mixture of all three. Eliminating sample composition and condition as possible causes of differences

It was possible to attribute part of the differences in pH measurement technologies to differing sensitivities to sample components. The pH technologies were considered as whole units, therefore it is not known if these differing sensitivities are caused by different probe designs or some other aspect of the technologies. However as both technologies operate on the same principles of potentiometrics, it is thought that the issue lies mainly with the built-in compensation methods.

Based on the project work and results, there are several recommendations to be made with regards to pH strategy.

1. Offline and online monitoring technologies should be as similar as possible, e.g. if a NOVA is used for online monitoring, then a NOVA should be used as the primary offline monitor and used to make adjustments.
2. Comparability of pH measurement technologies should be demonstrated across a variety of conditions within the culture design space including temperature.
3. The type of online and offline equipment used in a project should be recorded to ensure the same equipment is used at all scales of reactor.
4. Monitoring the unadjusted or ‘true’ online pH reading and comparing this value to the offline technologies may prove useful in identifying faulty or drift-prone probes.
5. Improved capture and analysis of individual pH probe performance over probe lifespan. This applies to both online and offline pH probes.
6. Using the same technologies for online and offline control throughout a project (i.e. initial lab testing to full production) will prevent the introduction of avoidable error. This will also aid consistency in corrective actions made by the scientist and the control systems.

.

Chapter 5. Productivity Investigation

5.1 Introduction

One of the most common hurdles to the introduction and implementation of statistical techniques is "Where do I begin?" In highly regulated environments, such as biopharmaceutical production, this is closely followed with "What is the correct technique?" The issue here is the presumption that there is a single best technique. The aim of the presented productivity investigation was to attempt to answer these two questions when identifying causes and indicators of poor productivity for a mAb-producing cell line (Project A). An additional aim was to design a more general approach for using historical data to identify weaknesses specific to the customer project and weaknesses general to the process.

5.2 Project Summary

Project A used a GS-NS0 cell-line to express a mAb which underwent a series of cultivations in 10L air lift reactors (ALR) to determine transferability of the cell-line from an external process to a bespoke Lonza process. After 15 cultures at the 10L scale, the cell-line was cultured at a 130L pilot scale. The 130L culture failed to reach an acceptable titre causing an intensive investigation of different parameters that could have affected the culture and three more 130L pilot scale cultures. In total, over 50 production-stage cultivations were performed at the Lonza Slough site and formed the main dataset for analysis.

Project A was then transferred to Lonza's production site in Portsmouth, New Hampshire, USA for four cultures at the 5000L scale. Underperformance at this scale and process validation requirements led to additional investigations in the US at the 10L scale, including a Design of Experiments study testing temperature, pH, DO, and feed parameters. A further four 5000L cultivations were performed. In addition to the change in physical location, US-sited cultures had modified seeding conditions and altered air/oxygen gas feed control parameters.

In total, Project A comprises data from 99 production-stage cultivations (49 UK, 50 US)[4]with eighty-seven 10L cultures, four 130L cultures, and eight 5000L cultures (Table 15).

[4] Cultivations halted early due to contamination or other known issues were excluded from analysis.

Data for Project A were drawn from two sources.

1. Daily monitoring records created by daily sampling
2. Online monitoring

The percentage of data missing cross the 99 cultures was strongly dependent on number of days of data considered and variables included, reaching ~25% when all possible days and monitored variables were considered. This was due to a variety of reasons including equipment such as NOVA Bioprofile 400 sensors not operating correctly during daily sampling, data collection for satellite/drop out cultures beginning on Day 4 of the main pilot, and not part of normal data collection at all scales (e.g. osmolality).

5.3 Aims

The overall aim of the productivity investigation was to test possible combinations of analytical options to identify suitable methods robust and adaptable enough to be transferred to other investigations. Methods were evaluated on relative ease in implementation and interpretability in addition to statistical power.

Initially, the specific aim of the productivity investigation was to identify key behaviours and related decision criteria leading to classification of bioreactor cultures classed as "High Producer" and "Low Producer", with particular focus on the UK-sited cultures A001 to A049. As the dataset increased through the addition of the US-sited cultures, the investigation was split into three key stages:

Stage 1. Initial Method Development

Given the UK cultures A001 to A049 and clear pass/fail criteria, to identify a core statistical method with additional consideration of data sources used, handling of data missing at random, and the use of data compression.

Stage 2. Improvements through Manipulation of Dataset Structure

Given the mixed UK- and US-sited cultures A001 to A099 and using key results from Stage 1, to develop a method for understanding causes of variation in Day 11 product concentration. The focus during Stage 2 was to improve model robustness through choice of progression variable and rigidity of the dataset sampling structure.

Stage 3. Media Batch Analysis.

In Stage 1 and Stage 2, media batch numbers were not included as a factor. The objective in the third stage was to determine whether variation in the media batches used were a contributing factor to variation in productivity.

UK-Sited Cultures				US-Sited Cultures			
ID	**Scale**	**ID**	**Scale**	**ID**	**Scale**	**ID**	**Scale**
A001	10 L	A026	10 L	A050[5]	5000 L	A075	10 L
A002	10 L	A027	10 L	A051[6]	5000 L	A076	10 L
A003	10 L	A028	10 L	A052[7]	5000 L	A077	10 L
A004	10 L	A029	10 L	A053[8]	5000 L	A078	10 L
A005	10 L	A030	10 L	A054	10 L	A079	10 L
A006	10 L	A031	10 L	A055	10 L	A080	10 L
A007	10 L	A032	10 L	A056	10 L	A081	10 L
A008	10 L	A033	10 L	A057	10 L	A082	10 L
A009	10 L	A034	10 L	A058	10 L	A083	10 L
A010	10 L	A035	10 L	A059	10 L	A084	10 L
A011	10 L	A036	10 L	A060	10 L	A085	10 L
A012	10 L	A037	10 L	A061	10 L	A086	10 L
A013	10 L	A038	10 L	A062	10 L	A087	10 L
A014	10 L	A039	10 L	A063	10 L	A088	10 L
A015	10 L	A040[2]	130 L	A064	10 L	A089	10 L
A016[1]	130 L	A041	10 L	A065	10 L	A090	10 L
A017	10 L	A042	10 L	A066	10 L	A091	10 L
A018	10 L	A043[3]	130 L	A067	10 L	A092	10 L
A019	10 L	A044	10 L	A068	10 L	A093	10 L
A020	10 L	A045	10 L	A069	10 L	A094	10 L
A021	10 L	A046[4]	130 L	A070	10 L	A095	10 L
A022	10 L	A047	10 L	A071	10 L	A096[9]	5000 L
A023	10 L	A048	10 L	A072	10 L	A097[10]	5000 L
A024	10 L	A049	10 L	A073	10 L	A098[11]	5000 L
A025	10 L			A074	10 L	A099[12]	5000 L

Table 15. Project A cultures analysed in the productivity investigation by location. 1 - First 130 L culture. 2 - Second 130 L culture. 3 - Third 130 L culture. 4 - Fourth 130 L culture. 5 - First 5000 L culture. 6 - Second 5000 L culture. 7 - Third 5000 L culture. 8 - Fourth 5000 L culture. 9 - Fifth 5000 L culture. 10 - Sixth 5000 L culture. 11 - Seventh 5000 L culture. 12 - Eighth 5000 L culture.

5.4 Stage 1: Initial Method Development

A variety of core statistical techniques were evaluated for ease of use, ease of interpretability, and suitability for data to be analysed. Black box and near black box techniques such as self-organising maps (SOM) and artificial neural networks were rejected due to difficulty of result and model interpretability. Discriminant analysis was rejected as the core statistical technique. This was due to the dataset containing only two classifications yet also containing a variety of conditions and potentially multiple paths to failure.

Decision trees were selected as the core statistical technique due to relative ease of interpretability and the ability to be applied to heterogeneous datasets. Although decision trees are generally not a computationally intensive algorithm when compared to PLS or SOM, they are not heavily promoted in statistical software. Four decision tree algorithms were considered: Gain Ratio, Gini Index, Information Gain, and ReliefF (see §3.6).

5.4.1 Stage 1: Data Selection

Online monitoring data were recorded by dataloggers at 5 minute intervals. For cultures reaching 11 days, this resulted in over 3,100 readings per variable monitored per culture. In comparison, from inoculation to harvest, variables monitored through daily monitoring samples had 12 readings per variable.

While each individual online monitoring point could be included as a variable, this would result in a highly unbalanced dataset when combining online and offline measurements. In order to reduce the volume of online monitoring data, new variables were created to capture online monitoring data through robust summary statistics termed "informative values" (these are discussed in greater detail in Appendix A). At the time of Stage 1 model development, Informative Values Version 1.0 was used (see Appendix A). In this version, data for each variable monitored online was split into "windows of activity" using offline sampling times. The average, standard deviation, gradient, and coefficient of determination was then calculated for each window of activity for each variable.

Each combination of dataset, dataset sources, and cultures dataset displayed in Table 16 were evaluated. At the request of the industry supervisor, the daily monitoring dataset was extended to included ratios and specific rates of change of biological interest. Detailed lists of variables in each dataset can be found in the appendices (Table 38 to Table 41).

5.4.2 Stage 1: Missing Data Handling

Four general approaches for the handling of missing data considered for use are briefly compared in Table 17 and are described in further detail in §5.4.2.1 to §5.4.2.4. Of the four approaches, three were employed: Fill In with Average, Rate Estimation, and iterative PCA. Two versions of Fill In with Average and Ratio Estimation were created, resulting in a total of five methods to be tested.

Dataset Name	**Dataset Source**	**Cultures in Dataset**
Daily Monitoring	Offline sampling	All
Online Monitoring	Online monitoring*	All
Daily Monitoring and Online Monitoring	Offline sampling Online monitoring*	All
Control	Offline sampling Online monitoring*	Cultures with non-control conditions excluded.
Daily Monitoring (Online Monitoring for Estimation)	Offline sampling Online monitoring**	All

Table 16. Summary of Dataset Combinations Tested. The primary difference between datasets was whether online monitoring and offline monitoring datasets were concatenated.

Method	**Description**	**Statistical Effects**	**Practicality**
Cut Down	Remove variables and/or samples with missing data.	Can result in few variables and/or few samples remaining.	Manually intensive.
Fill In with Average	Replace missing data with mean of available values.	Assumes mean as appropriate estimate. Reduces data spread.	Simple to implement
Rate Estimation	Ratios calculated between sampling points used to estimate values.	Assumes behaviour independent of other variables.	Extensive set-up work required.
Iterative PCA	PCA models created until estimated values converge.	Limit to missing data required to prevent spread reduction as seen in Fill In With Average.	Proprietary software.

Table 17. Summary of Missing Data Treatments Considered. “Cut Down” was rejected due to poor industrialisation potential. Two versions of Fill In with Average and Ratio Estimation were created, resulting in a total of five methods to be tested

5.4.2.1 Stage 1: Missing Data Handling: Cut Down

One method for the handing of missing data that is often considered an 'easy option' is the exclusion of variables with missing values or samples with missing values. As seen in Table 18 and Table 19, both of these approaches can vastly reduce data available for analysis. A combined approach is used in Table 20 to give the maximum retention of recorded data while excluding missing data.

Furthermore, while these approaches are seen as simple to implement, applying the final model to new data can become considerably more complicated if the dataset must be manually picked through for variable removal. Initial plans to demonstrate the negative effects of this approach on models produced were aborted due to the time-consuming manual work needed to apply the approach to datasets.

		Variable							
		A	B	C	D	E	F	G	H
Sample	1						●	●	
	2	●							●
	3		●						
	4								
	5				●	●			

Table 18. Removal of samples with missing data leaves one sample. ● – missing data. Shading – excluded sample. 8 values out of 33 remain (25%)

		Variable							
		A	B	C	D	E	F	G	H
Sample	1						●	●	
	2	●							●
	3		●						
	4								
	5				●	●			

Table 19. Removal of variables with missing data leaves one variable. ● – missing data. Shading – excluded variable. 5 values out of 33 remain (15.6%)

		Variable							
		A	B	C	D	E	F	G	H
Sample	1						●	●	
	2	●							●
	3		●						
	4								
	5				●	●			

Table 20. Removal of samples and variables to give best exclusion of missing data. ● – missing data. Shading – excluded variable. 15 values out of 33 remain (46.9%)

5.4.2.2 Stage 1: Missing Data Handling: Fill In With Average

While it is known that the "Fill In With Average" method of infilling missing data artificially reduces variance and variable distributions with potentially significant knock on effects on other summary statistics (Table 21), it remains a popular suggestion due relative ease of application to a dataset. Two versions of the method were tested to demonstrate the effects on a real dataset as a cautionary point of reference if suggested in future investigations.

'Mean Estimate' refers to where the mean of a variable was calculated using all available values for that variable. In this version, there is no filtering of the dataset to determine whether or not a sample's data is used, hence the calculated mean is influenced by all operating conditions represented in the dataset.

'Historical Mean' refers to where the mean of a variable was calculated only from cultures operating under normal control conditions. Ideally, use of an average value from control cultures only would result in a value representative of 'normal' behaviour.

-86	-10	11	67	-91	-22	-14	-60	-74	-7
-8	89	68	76	-28	67	39	86	76	71
83	56	21	37	-57	-45	-9	99	38	-31
-93	29	-82	-52	78	92	71	0	-38	-82
-11	-96	91	49	22	34	-87	-98	13	-64
26	23	-70	58	-24	-39	-50	80	22	61
6	-11	24	-36	61	88	-41	4	96	43
-62	50	-17	31	63	79	30	90	-63	-70
95	9	-89	-73	5	-74	-98	82	14	-35
-30	46	78	25	-65	-78	17	12	18	-63

Table 21. Dataset of 100 values generated using Microsoft Excel formula "=RandBetween(-100,100)" with mean = 4.66 and median = 11.5. 8% of the dataset (indicated by shading) is removed at random. The remaining dataset has mean = 3.65 and median = 12. This represents changes in mean and median of -22% and 4% respectively with the dataset median showing greater robustness to missing data than the dataset mean. If the missing values were replaced with the remaining dataset's mean 3.65, the filled in dataset would have mean = 3.65 and median = 3.83. When these statistics from the filled in dataset are compared to the original dataset, the changes in mean and median are -22% and -67% respectively. Hence the previously robust median is significantly altered.

5.4.2.3 Stage 1: Missing Data Handling: Rate Estimation

Rate estimation was the term used to describe a form of interpolation whereby observed rates of change of a variable were used to estimate missing data. For this method, a rate of change was calculated for each variable for each window of activity (i.e. between two consecutive sampling points) for each culture using:

$$\dot{x_{n,n+1}} = \frac{x_{n+1} - x_n}{t_{n+1} - t_n} \quad \text{Eq. 6.1}$$

Where $\dot{x_{n,n+1}}$ is the rate of change of a variable across the window of activity defined by sampling points n and $n+1$, $x_{n+1}$and x_n the recorded values for the variable of interest at those sampling points, and $t_{n+1}$and t_n the time of the samplings points as Elapsed Time (h). If either x_{n+1}or x_n was missing, no rate was calculated.

Two mean rates were calculated for each window of activity $n, n+1$ (Figure 16). The first mean rate was calculated using all available rates of change across all cultures and referred to as the 'Mean Ratio', $\dot{Mx_{n,n+1}}$. The second mean rate was calculated from a subset of control cultures and referred to as the 'Historical Ratio', $\dot{Hx_{n,n+1}}$. Missing values were estimated using Eq. 6.2 and Eq. 6.3 to plot the values in Figure 17.

$$\widehat{Mx_{n+1}} = \dot{Mx_{n,n+1}}(t_{n+1} - t_n) + x_n \quad \text{Eq. 6.2}$$
$$\widehat{Hx_{n+1}} = \dot{Hx_{n,n+1}}(t_{n+1} - t_n) + x_n \quad \text{Eq. 6.3}$$

where $\widehat{Mx_{n+1}}$ is the value predicted using the Mean Rate and $\widehat{Hx_{n+1}}$ is the value predicted using the Historical Rate. If a missing value was from the point of inoculation ($n = 1$), then Eq. 6.4 and Eq. 6.5 were used. This was the only instance in which extrapolation was permitted.

$$\widehat{Mx_1} = \dot{Mx_{1,2}}(t_1 - t_2) + x_2 \quad \text{Eq. 6.2}$$
$$\widehat{Hx_1} = \dot{Hx_{1,2}}(t_1 - t_2) + x_2 \quad \text{Eq. 6.3}$$

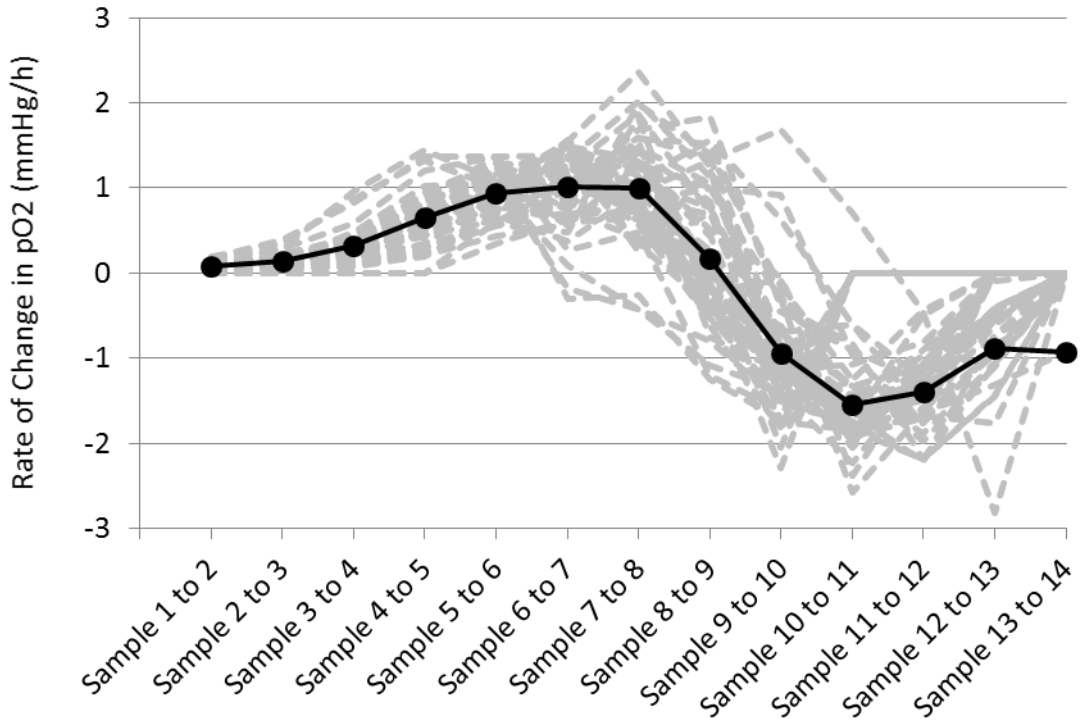


Figure 16. Rates of change in pO_2 between consecutive sampling points were calculated for each culture (---). Note that if any of the data points needed to create the ratio was missing, no value could be calculated. A mean average rate for rates of change in pO2 was calculated using all available rates (-●-). These mean rates were used to estimate missing values such as those in Figure 17.

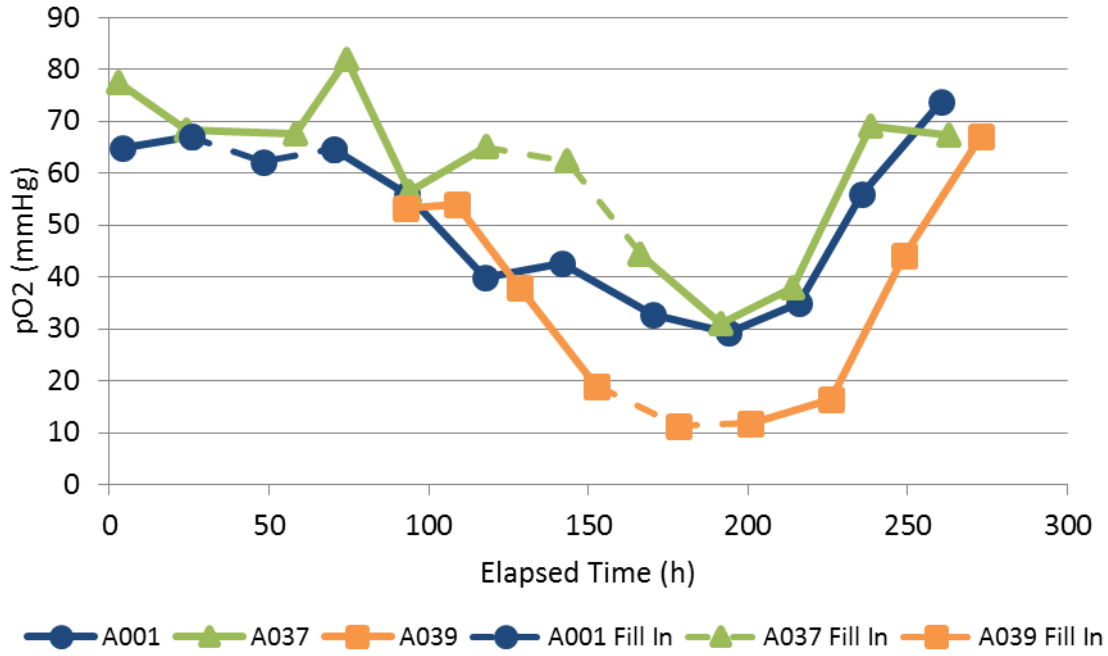


Figure 17. Three cultures with missing values for pO_2 were treated using ratio estimation. Dashed lines link the estimated values to the directly measured values. Note that estimated points do not lie in the same locations that simple linear interpolation would place them.

There are three main assumptions that must be made for this form of interpolation. First is the assumption that rates of change for a variable are independent of other variables in the dataset. Second is the assumption that the relationship is a simple proportionate relationship and not more complex, such as a quadratic relationship. A simple example of the importance to this assumption is a variable with a parabolic relationship with time. Simple linear interpolation may be adequate if (t_1, x_1) and (t_2, x_2) lie on the same side of the vertex, however it is obvious how simple linear interpolation is inappropriate if they lie opposite sides of the vertex. Finally, this form of missing data estimation assumes that the behaviour seen in other cultures (from which $\widehat{Mx_{n+1}}$ and $\widehat{Hx_{n+1}}$ are calculated) is the same as the behaviour exhibited by the sample in question.

5.4.2.4 Stage 1: Missing Data Handling: Iterative PCA

The methods presented thus far have been non-iterative calculations where a single replacement value is calculated. Iterative PCA (iPCA) is a form of inferential estimation which uses algorithmic modelling to estimate values for missing data based on the dataset's covariance matrix. Missing values are first replaced with the variable mean. A PCA model is generated to capture a set percentage of variance. Based on this model, new values are substituted for missing data with the new values selected based on consistency with the PCA model loadings. A new PCA model is generated and new values are substituted. This process is repeated until a suitable level of convergence is reached, i.e. the change when substituting new values drops below a threshold.

According to the Eigenvector Wiki "[u]sing PCA to replace data generally works better than using the mean of a variable because it uses the covariance in the data to estimate what the missing values should be." This method of missing data estimation was deemed to be of potentially high value as missing data is based on behaviour across all variables.

This form of missing data estimation was easily implemented as it is a feature of the PLS Toolbox (Eigenvector) and performed automatically during model creation. However as a software specific method, use of iPCA is dependent on software availability or significant time investment for an internally useable version.

5.4.3 Stage 1: Decision Tree Algorithm

Four decision tree algorithms were considered in this investigation: Information Gain, Gain Ratio, Gini Index, ReliefF. Detailed descriptions of how these algorithms function can be found in Chapter 2. To reiterate key differences according to Han *et al.* (2011), the Information Gain and Gini Index algorithms are biased towards attributes with a greater number of possible values when selecting decision criteria. The Gain Ratio algorithm is biased to unbalanced splits, e.g. one partition is much smaller than the others. The fourth algorithm considered was ReliefF, an adaptation of the Relief algorithm that includes a k-nearest neighbour function when selecting decision criteria. While ReliefF is only available with the data mining software Orange (University of Ljubljana), it was included to demonstrate whether investing in a more specialised algorithm could yield any benefit.

5.4.4 Stage 1: Data Transformation

Loss of context is a problem when using decision trees as single attributes are selected for decision criteria, particularly in the evaluation of a dataset in an unmodified state, i.e. values are as recorded, here referred to as "As Is". Furthermore, the reliance on a single variable reading can lead to spurious decision criteria and, consequently, a lack of robustness. The application of PCA as a data transformation step during pre-processing was investigated as a means of retaining contextual information and reduce spurious decision criteria selection.

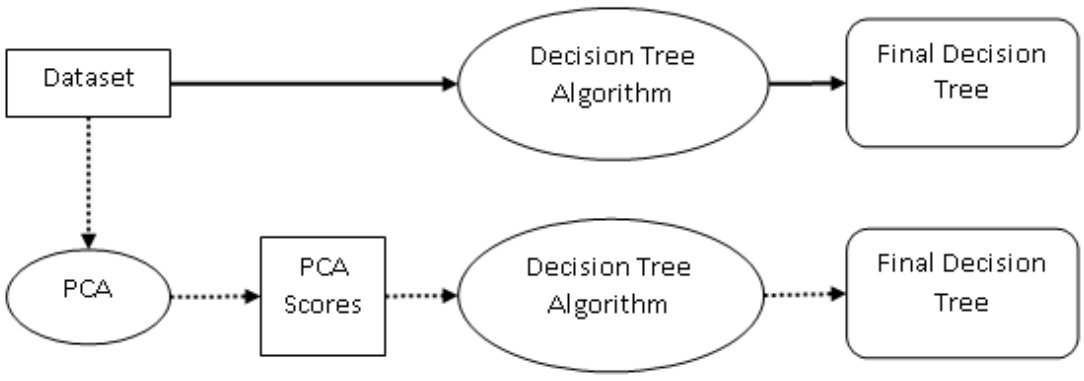


Figure 18. The two analysis pathways shown differ only in whether the dataset is passed directly to a decision tree algorithm or whether the PCA is applied to the dataset as a pre-processing step with the resulting PC scores then passed to the decision tree algorithm.

Following the process shown in Figure 18, PCA models were created for each combination of data source and missing data treatment using Eigenvector PLS Toolbox with random sampling as cross-validation (10 splits, 5 iterations). The first 10 PC scores for each dataset were extracted for decision tree creation. The number of PCs used was chosen to ensure a minimum of 90% of dataset variance was transfered to decision tree creation while maintaining a standardised method.

5.5 Stage 1: Method

The cultures were numbered as Pro_A001 to Pro_A049 and then classified as "High" or "Low" producers based on product concentration at harvest. The cut-off value was chosen to take into account a customer-defined breakeven point for economic viability (for confidentiality, this value cannot be stated). Of the 49 cultures analysed, 38 cultures were classed as "High" producers (pass) and 11 cultures were classed as "Low" producers (fail).

Decision trees were created for all possible combinations of options listed in Figure 19 using exhaustive binarisation for optimal split with leaf-splitting stopping criteria of 95% purity and m-estimate post-pruning (m=2). To prevent overfitting of models, 30% of 10L cultures were randomly selected as a validation dataset. The four 130L pilot cultures were excluded from calibration datasets as a final testing dataset. In total, 120 decision trees were evaluated for size, classification accuracy, and interpretability.

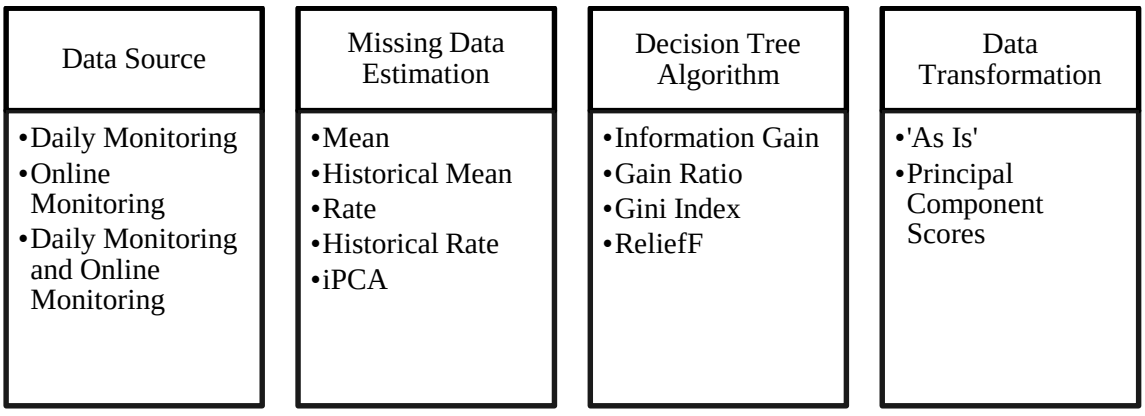


Figure 19. Summary of method options. Three combinations of data sources (daily monitoring, online monitoring, and combined daily monitoring and online monitoring) were treated with five different processes of missing data estimation. After missing data estimation, the datasets were then passed to four different decision algorithms. This was performed with the data "As Is" (i.e. with no additional pre-processing). The analyses were repeated using PCA as a pre-processing with the resulting PC scores passed to the decision tree instead.

5.6 Stage 1: Results and Discussion

Decision tree results and options were evaluated in two ways. In the first, a main effects plot was created with classification accuracy of the testing dataset as the response (Figure 20). From this it was seen that use of data from daily monitoring samples for all cultures generally gave higher classification accuracy. Inclusion of online monitoring data (here downsampled using Informative Values 1.0) had little effect on test set classification accuracy, however a general decrease was observed when using online monitoring data alone. Collating online monitoring and daily monitoring datasets for estimation of missing data followed by the use of only from daily monitoring had a negative effect on testing accuracy.

A decrease in test set classification accuracy was seen when cultures with deliberate experimental conditions were excluded from the daily monitoring dataset ("Control"). This demonstrated the effect of over fitting a model by calibrating using only control behaviours, particularly when a range of behaviours are to be considered. Deliberate inclusion of non-control conditions allowed multiple paths to failure and success to be identified in addition to a decreased likelihood of spurious decision criteria selection. In short, robust models cannot be calibrated from "Golden Batches" alone.

The missing data estimation method with the highest testing accuracy was the iPCA method included in the EigenVector PLS Toolbox. This was as expected as iPCA estimates values based on both the culture's behaviour across all variables and the behaviour of all other cultures in the dataset across those variables, whereas the other methods considered behaviour at a single variable at either a single sampling point or between two successive sampling points. Further evidence for this conclusion was the minor increase in classification accuracy when online monitoring data were included during iPCA estimation but excluded from model creation.

Similar conclusions to the above were made when method option results were evaluated on a case by case basis. In the first evaluation method, there appeared to be little benefit in creating a PCA model in order to use PC scores in place of the original dataset in terms of pure testing accuracy. However, discussion-based evaluation revealed that use of PC scores led to greater understanding of differences between high and low producers. This was due to the greater contextual information in decision criteria, i.e. values for splits were determined from behaviour across multiple variables and multiple days, not a single variable at a single time point (e.g. "Glucose (g/L) on Day 6").

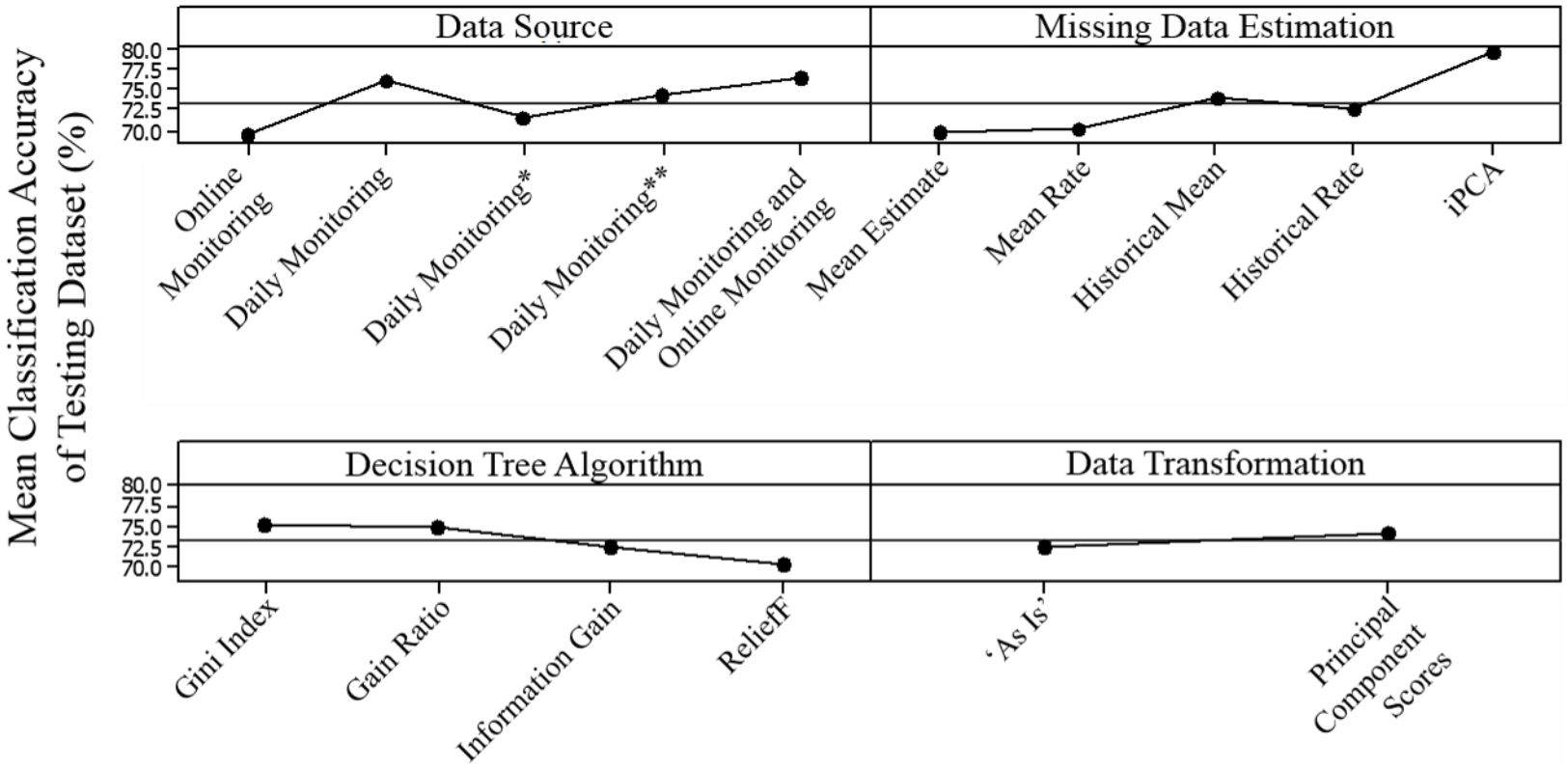


Figure 20. Main Effects Plot for Testing Accuracy. Four options were evaluated: data type (referring to datasets used), estimation method for missing data, decision tree algorithm, and treatment, i.e. whether data were compressed into PCA scores for decision classification or if data that had not first been summarised as PC scores ('As Is'). *Daily monitoring data from control cultures only. **Online monitoring data including during missing data estimation.

It was noted in both cases that interpretation was made difficult by the mixed cause/response nature of some variables (e.g. lactate concentrations) as compared to pure cause variables (e.g. temperature). As such, using the more context/information rich PC scores could provide a more informed basis for action.

Based on these results, the developed method was as follows:

1) Online monitoring data is summarised as informative values.
2) The informative values dataset and daily offline monitoring datasets are collated.
3) The collated dataset is unfolded into the profile (short and wide) configuration.
4) iPCA is used to estimate missing values.
5) The dataset is mean centred and scaled to unit variance.
6) A PCA model is created using random sampling and multiple iterations.
7) PC scores from the model are extracted as a new dataset.
8) A decision tree is created to classify cultures based on PC scores using the Gini Index.

The decision tree with the highest classification accuracy is shown in Figure 21. Analysis of loadings for PCs selected as decision criteria allowed overall trends to be analysed. Contribution analysis of scores allowed more specific behaviours to be further investigated. Applying this approach to the first failure at pilot scale (A016) indicated initial seeding conditions and subsequent glutamine behaviour as the main areas of deviant behaviour. While a specific cause for the altered glutamine behaviour could not be identified from the data analysed, it was suggested that batch-to-batch variation in media powders used could be a possible contributing factor.

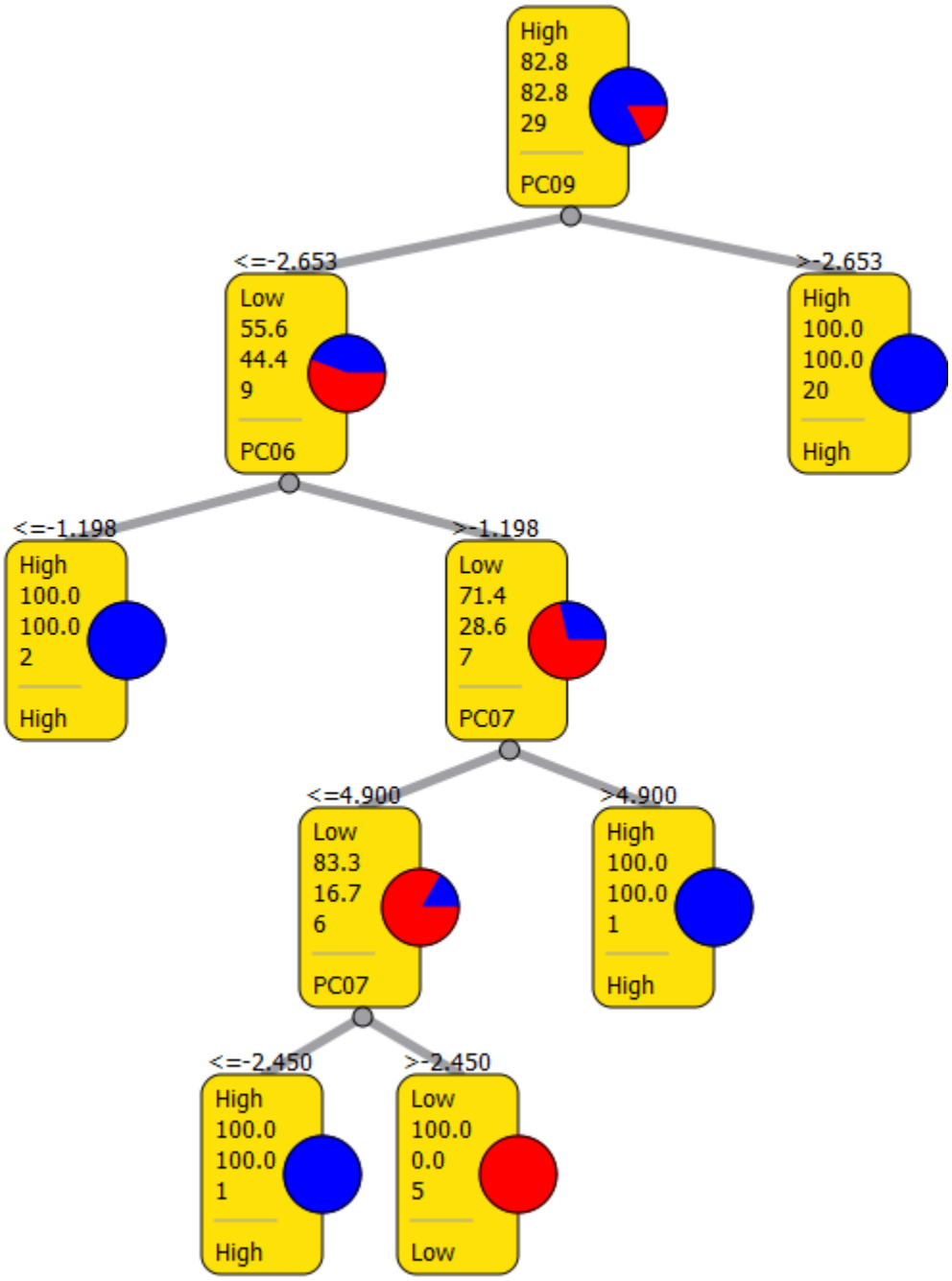


Figure 21. Decision tree with the highest classification accuracy and matching top options (DMDLG, PC, iPCA, Gini Index). Pie charts indicate the distribution of high-producing (blue) and low-producing (red) cultures at the node. Information written within the node indicates the majority class, majority class probability, target class probability, and the total number of instances on the node. Note that the values displayed are for the calibration dataset (29 cultures) and not the validation dataset.

5.7 Stage 2: Improvements through Manipulation of the Dataset Structure

In stage 1, a basic method for data analysis was suggested and locally optimised for available options. While this method met the stated objectives of classifying cultures A001 to A049 and identifying key indicators, there were several areas which could be further improved. Two areas for improvement addressed using the dataset A001 to A099 in Stage 2 were:

- The definition of progress measurement for a fed-batch culture.
- Rigidity in data collection with respect to culture progress.

PCA analysis of the mixed UK and US dataset indicated confounding caused by the differences in seeding. Due to this confounding and a change in investigation focus, the proposed alterations were not evaluated using decision trees and pass/fail criteria. Instead, the proposed alterations were evaluated when predicting Day 11 product concentration using PLSR.

5.7.1 Stage 2: Defining Progress and Progression Variables

Time is so heavily embedded in the concept of progress that it can be found in its definition: to improve or develop over a period of time [163]. However, while time must pass for progress to occur, should its *de facto* status as the yardstick for progress go unquestioned? When defining progress, should something other than time be plotted on the X axis?

A non-biological example of questioning measures of progress is academic performance of students and the question of "Is my child's intelligence developing normally?" In a study on intelligence development in school children [164], it was shown that, for intelligences evaluated using verbal and numerical tests, progress was better defined in terms of terms of time in education ("psycho-educational age") than in terms of absolute physical age ("biological age"). However, biological age tended to be a more appropriate when considering intelligences evaluated using figural tests (see Figure 66 and Table 43 in Appendix B).

The cell culture analogy to these measures are the absolute values of viable cell concentration and elapsed time (analogous to biological age") and the calculated integral of viable cell concentration, which measures growth since inoculation (analogous to psycho-educational age).

5.7.1.1 Stage 2: IVC as a Progression Variable

The integral of viable cell concentration (IVC) was proposed as an alternative progression variable for fed-batch and batch cultures, in place of Elapsed Time (h) or Elapsed Day (d). Evidence supporting IVC as an alternative progression variables are as follows:

1. From a conceptual perspective, IVC captures "culture history" in a single value. As IVC is calculated by summing the area beneath a viable cell concentration growth curve, the change in IVC between sampling points takes into account both the time between sampling points and the level of growth between sampling points. Hence while time still plays a role in measuring progress, using IVC to measure progress would allow for a more biology- and response-based yardstick to be employed.

2. An initial evaluation of correlation between events in air and oxygen profiles against different progression variables was performed for cultures A001 to A049. It was identified that there was comparable or greater correlation between when IVC was used in place of Elapsed Time (Figure 22). A follow up evaluation indicated that, in general, these correlations were stronger once experimental, non-control cultures were excluded.

3. PLSR models predicting product concentration were created using daily monitoring and online monitoring data from A001 to A049 for the following variable sets:
    a. **No Progression Variable** — "Obvious" progression variables VCC, TCC, Elapsed Time (h), and IVC excluded from dataset.
    b. **Elapsed Time** — VCC, TCC, and IVC excluded from dataset
    c. **IVC** — VCC, TCC, and Elapsed Time (h) excluded from dataset.

PLSR were created for each subset using the SIMPLS algorithm with random sampling and multiple iterations (10 splits, 10 iterations). The final models were selected based on maximum R^2 during cross validation. Full listings of the variables used and details of the final models can be found in Table 45, Table 46, and Table 47 in Appendix B, respectively. Data from US cultures A050 to A099 were then applied as a testing dataset.

Figure 23 shows the predicted product concentration against the measured product concentration for the three models, in addition to the calibration R^2, cross-validation R^2, and the US testing dataset R^2.

Considering only R^2 for UK cultures and UK predictions, it was seen that the highest values calculated for R^2 were reached in models where IVC was included as a measure of progress (Figure 23C) and lowest when no explicit progression variable was included. In addition to improved predictive power, increased linearisation of predicted value against measured value was noted when IVC was included in the model.

These improvements in model predictive accuracy and linearisation of residuals were preserved when considering data from the US-sited cultures A050 to A099. This was of particular note as it indicated that use of IVC as the progression variable could convey greater model transferability between sites and scale than Elapsed Time (h).

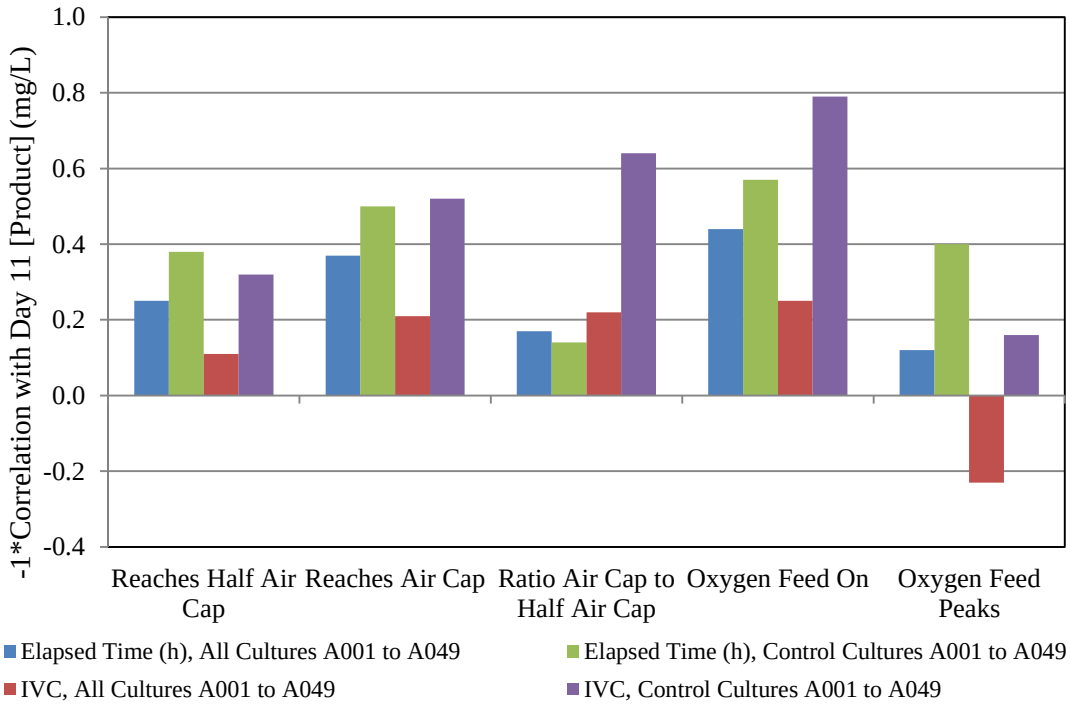


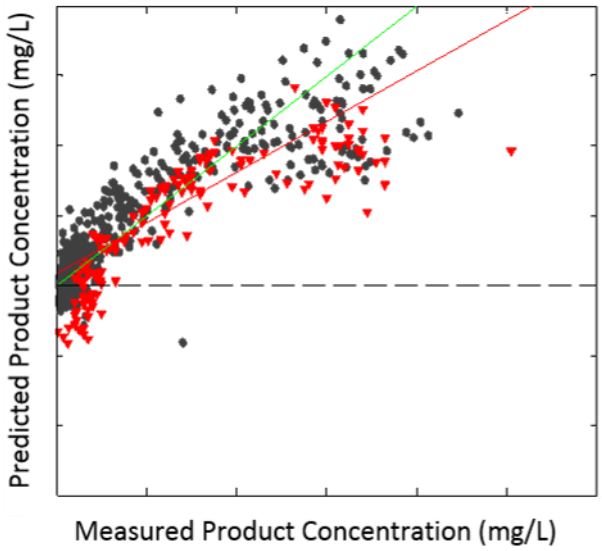
Figure 22. Correlations between Day 11 [Product] (mg/L) and specific events in air and oxygen feed profiles for cultures A001 to A049. It was observed that correlation between events and IVC was strongly affected by sample selection, with the dataset including all growth conditions having notably lower correlation with air and oxygen profile events than when the dataset was restricted to control conditions cultures only. Most notable is the strong correlation between IVC with the activation of the oxygen feed for control condition cultures. Also notable is correlation between IVC at half of air capped flowrate and the IVC at the air capped flowrate for control condition cultures. Note that correlations were multiplied by -1 for ease of viewing.

A. Progression Variable: "None"

UK Data Prediction
Cal R^2 = 0.76

UK Data Prediction
CV R^2 = 0.71

US Data Prediction
R^2 = 0.77

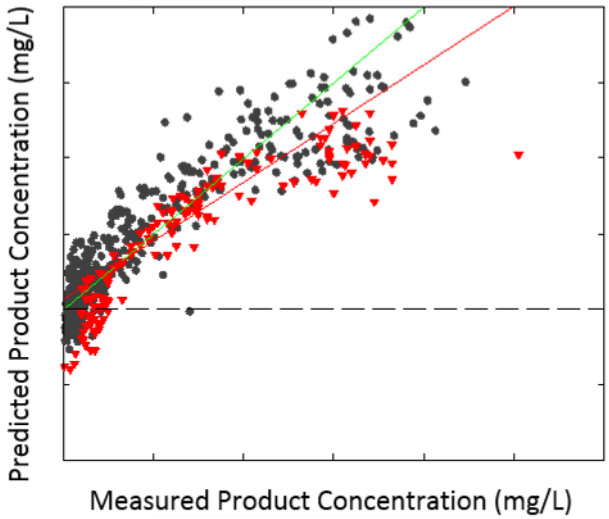


B. Progression Variable: Elapsed Time

UK Data Prediction
Cal R^2 = 0.86

UK Data Prediction
CV R^2 = 0.83

US Data Prediction
R^2 = 0.82



C. Progression Variable IVC

UK Data Prediction
Cal R^2 = 0.97

UK Data Prediction
CV R^2 = 0.95

US Data Prediction
R^2 = 0.94

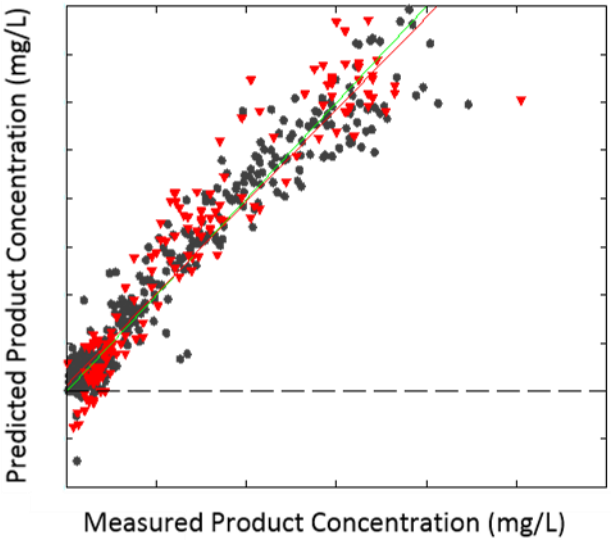


Figure 23. US data (red triangle ▼) applied to models calibrated from UK data (grey square ■). The green line shows the ideal 1:1 relationship between measured and predicted product concentrations. The red line shows the actual line of correlation between measured and predicted product concentrations. Note that values have been hidden due to confidentiality requirements.

5.7.2 Stage 2: Dataset Rigidity

Dataset rigidity is the adherence of sampling to a set interval size. Here, it is considered as a measure of a specific form of noise: random underlying temporal variation created by variation in sampling intervals for different cultures. In a dataset, this temporal variation is captured in addition to the variation of interest. Any model algorithms in subsequent analyses must be able to separate the undesired, underlying temporal variation from the variation of interest. If this cannot be done to a sufficient level, there is the risk of incorrect conclusions being drawn and used to justify further actions.

5.7.2.1 Hypothetical Example

Ten cultures grow identically. By chance a culture is always sampled slightly later than the other nine cultures. When metabolite and cell growth data are analysed, the culture will appear to be more advanced that the other nine by virtue of later sampling times. This could be fixed by including the sampling time as a variable in analyses, however this could lead to unhelpful results, e.g. time being identified a key predictor for product titre instead of viable cell concentration.

5.7.3 Stage 2: Dataset Realignment

The suggestion was made that reduction or elimination of the underlying temporal variation introduced by variation in sampling time could be achieved through appropriate manipulation of the dataset. More specifically, it was suggested that this could be achieved through interpolation of sampling data to user-defined values for a progression variable.

For concept clarity, dataset realignment is here described in terms of time as this is easily accessible conceptually and forms the basis of typical sampling procedures However, the concept of rigidity can be applied to any variable deemed to be the progression variable, e.g. IVC. While it would be very difficult to implement a sampling procedure based on such an approach[5], it may be possible to impose adherence after data are generated.

It is key to note that imposing rigid structure based on a given variable alters the distributions of all other variables (Figure 24.), which may be undesirable or unacceptable and potentially have an overall negative effect in subsequent analyses.

[5] A possible solution is the use of in-line probes to create alerts when a monitored progression variable reaches a pre-determined sampling point value.

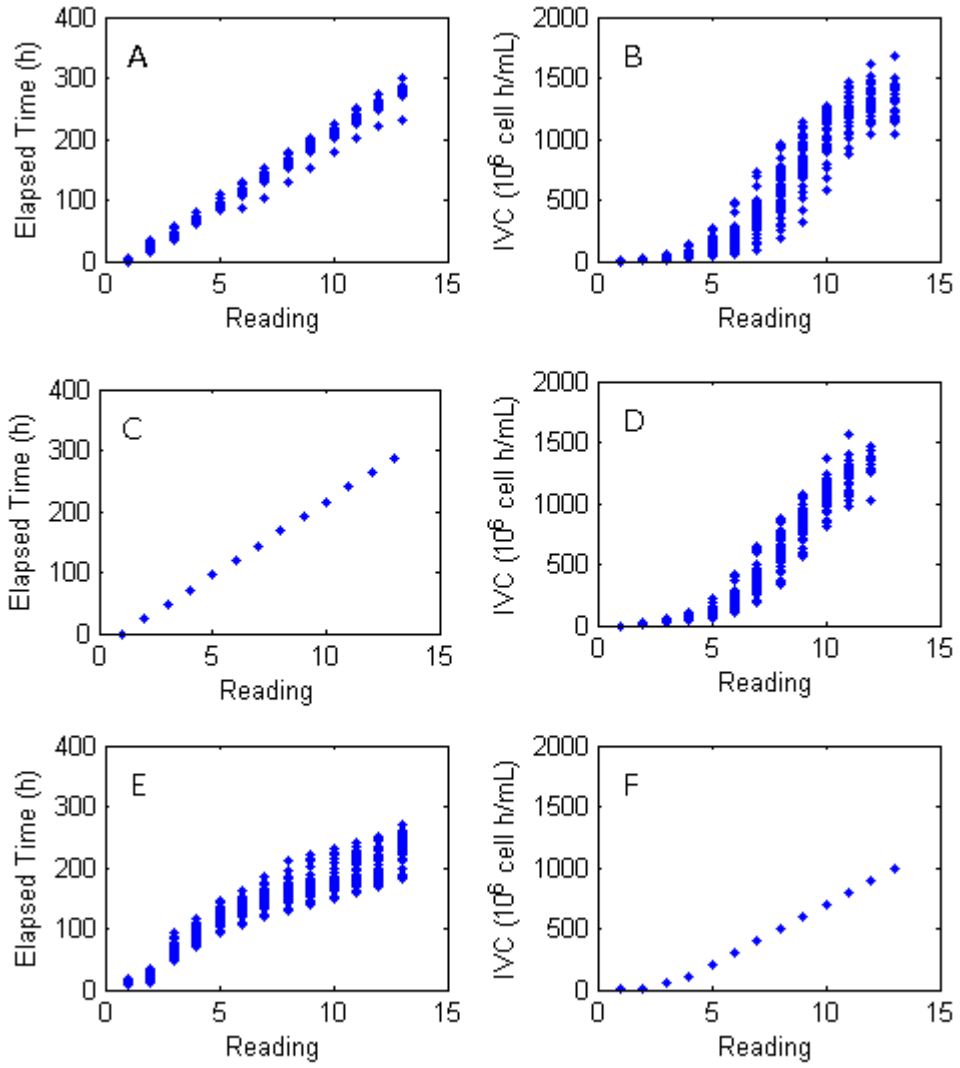


Figure 24. Comparison of distributions of values for readings of elapsed time and IVC for cultures A001 to A049 when using three different datasets: Natural (e.g. approximately 24 h sampling interval), Time Aligned (e.g. rigid 24 h sampling interval), and IVC Aligned (e.g. rigid structure imposed using select values for IVC).

A. Natural Data – showing variation in elapsed time for readings. Note that A016, whose data became offset due to both a non-standard interval and multiple data entries, can be visually identified as a temporal outlier.
B. Natural Data – showing variation in IVC for readings.
C. Time Aligned Dataset – showing lack of variation in elapsed time for readings.
D. Time Aligned Dataset - showing variation in IVC for readings.
E. IVC Aligned Dataset – showing variation in elapsed time for readings.
F. IVC Aligned Dataset - showing lack of variation in IVC for readings.

5.7.4 Stage 2: Progression Variable Selection and Alignment Effects

In Section 5.7.1 the potential benefits of alternative progression variables when analysing datasets with variation in sampling timepoints were demonstrated. The next part of the investigation focussed on the influence of imposing a rigid progression structure on a dataset by forcibly aligning data to specific progression points as outlined in Section 5.7.2.

Using A001 to A049 as the calibration dataset, three alignments were considered:

1. 'As Is'/unaligned allowing for natural variation in sampling time.
2. Realignment to user-defined values for Elapsed Time (h).
3. Realignment to user-defined values for IVC.

A further hypothesis was that realignment of data to user-defined values for the progression variable would allow the progression variable itself to be excluded from analysis. Hence a comparison was made between models where the progression variable was included and models where the progression variable was excluded.

As demonstrated in several publications [44,165,166], there are benefits to unfolding or reorientation of the dataset to treat each row as representing a single sampling of data during a culture ("Day by Day") or to aggregate serial observations of a culture as a single observation spanning the full duration of the culture ("Profile"). A Day by Day versus Profile comparison was made to determine whether dataset alignment effects, if any, were dependent on dataset orientation.

An area of concern was the suitability of the developed method for multiple responses of interest. In this study, the response of interest was Day 11 product concentration. However, future investigations utilising the developed methods may focus on maximising viable cell concentration or maintaining culture viability. For this reason, alignment and rigidity effects were evaluated for three responses: [Product] (mg/L), Viability (%), and Viable Cell Concentration (VCC) (10^6 cells/mL).

A final factor investigated was the effect of using only data originating from offline monitoring versus using data from both online and offline monitoring. To achieve this, informative values for online monitoring were recalculated to match the progression points used for realignment.

5.7.5 Stage 2: Method

Daily monitoring data were aggregated into a single dataset. Alignment points were selected using mean and median values across samples with numbers rounded to give standardised progression between interval points (Table 22 and Table 23).

In the original dataset of up to 14 samplings per culture, the maximum mean average for Elapsed Time (h) and IVC were 288.50 h and 1308.61 (10^6 cell/mL.h) respectively. As can be seen in Figure 25, few cultures reach these levels of progress. Maximum values chosen for time and IVC alignment points were 312 h and alignment point was 1000 x 10^6 cell/mL.h respectively.

Time realignment and IVC realignment were applied using the progression values listed in Table 22 and Table 23.

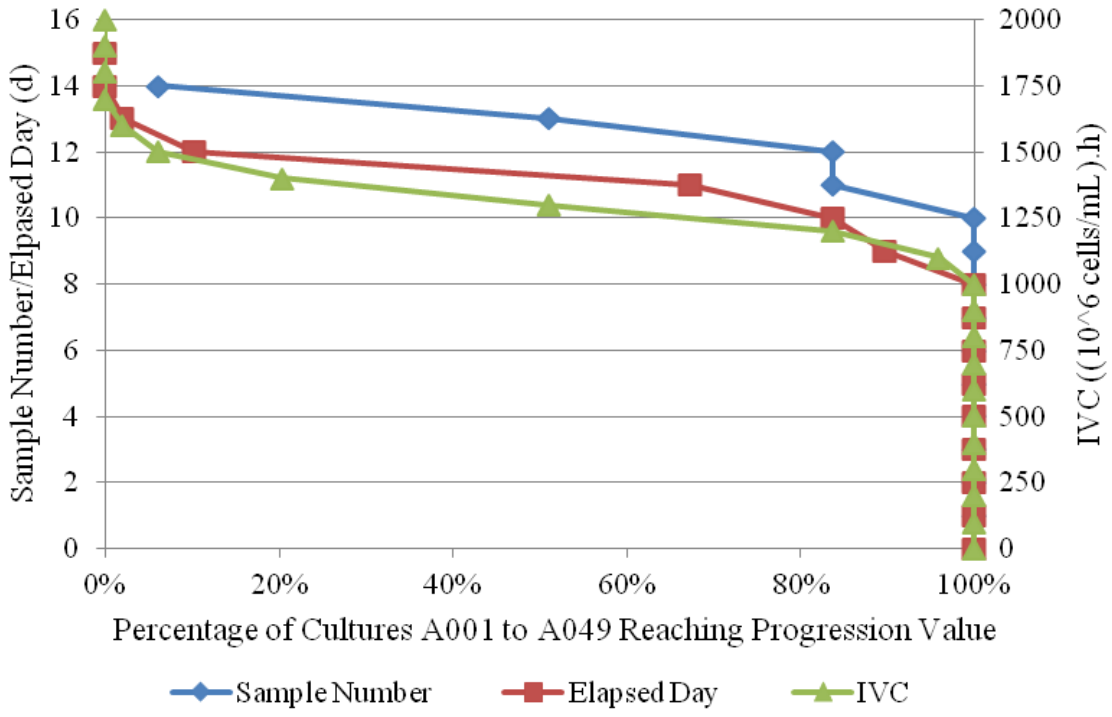


Figure 25. Percentage of cultures A001 to A049 reaching stated values for progression variables. These percentages were used to be determine the number of sampling points, the final timepoint, and the maximum IVC values used for realignment of datasets. Online monitoring data were translated to the robust statistics version of informative values. Informative values were calculated using natural sampling times, user-defined time points (as part of time realignment), and times at which user-defined IVC values were reached (as part of IVC realignment).

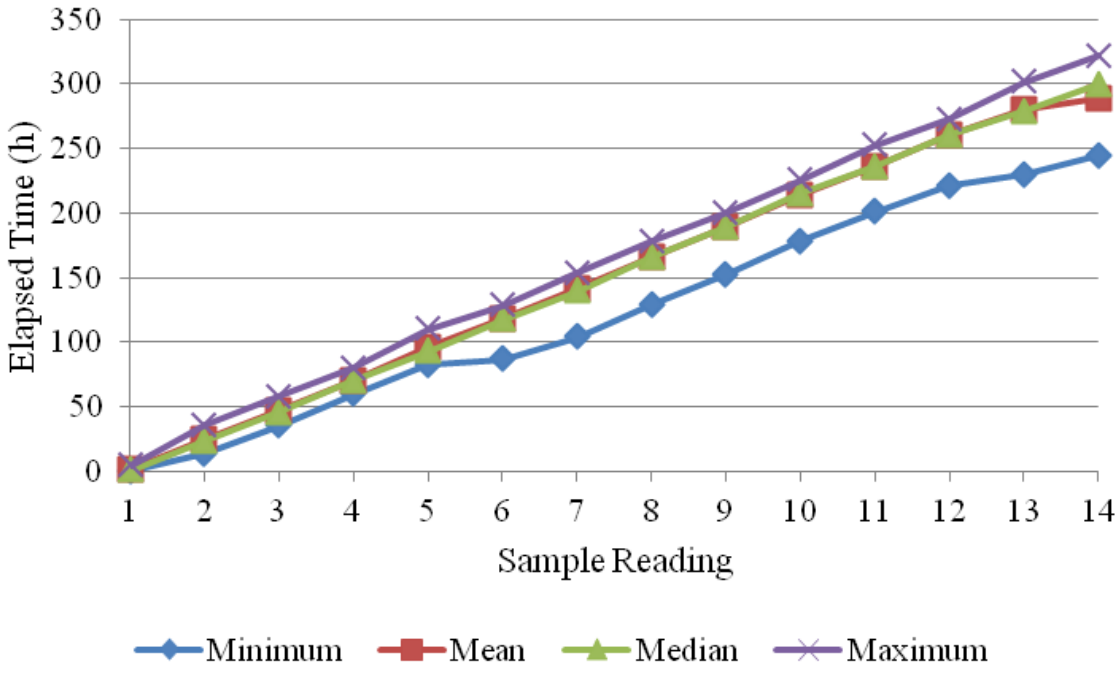


Figure 26. Distribution of values for Elapsed Time (h) at progressive sample readings. It was observed that resampling of a culture could notably offset recorded points, as seen in the change in progression in the minimum Elapsed Time before and after Day 6.

Original Sampling Point						**User-Defined Progression Points**	
#	**Min.**	**Mean**	**Median**	**Max.**		**#**	**Value**
1	0.00	1.38	0.97	4.58		1	0
2	13.83	25.03	23.42	35.42		2	24
3	35.18	46.91	45.95	58.08		3	48
4	59.83	70.03	70.50	79.87		4	72
5	82.50	95.55	93.08	110.00		5	96
6	86.92	117.84	117.50	128.58		6	120
7	103.83	141.75	139.67	153.33		7	144
8	129.08	165.64	166.00	177.78		8	168
9	152.58	189.46	188.83	200.60		9	192
10	178.17	213.90	214.92	225.60		10	216
11	200.92	236.36	235.63	252.33		11	240
12	221.00	260.41	260.67	273.13		12	264
13	229.52	280.50	279.12	301.50		13	288
14	244.25	288.50	299.83	321.42		14	312

Table 22. Evaluation of A001 to A099 recorded Elapsed Time (h) for selection of progression points. User-defined progression points were selected to give a 24h interval between samples.

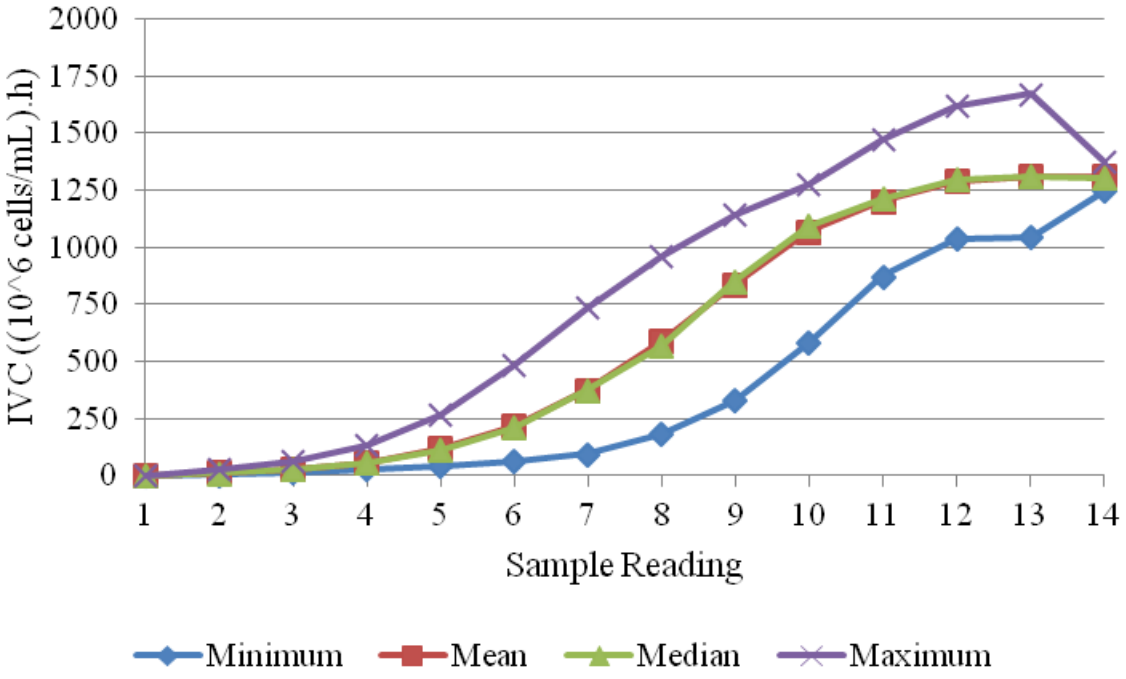


Figure 27. Distribution of values for IVC ($(10^6$ cells/mL).h) at progressive sample readings. It was observed that a wide range of values were recorded for each day. This was due to both variation in Elapsed Time for the samples and the effects from different growth conditions.

Original Sampling Point						User-Defined Progression Points	
#	Min.	Mean	Median	Max.		#	Value
1	0.00	0.50	0.33	1.65		1	0
2	4.05	12.01	10.50	26.25		2	10
3	13.55	27.92	25.38	64.45		3	50
4	27.33	56.88	56.73	132.22		4	100
5	43.70	120.54	111.57	265.62		5	200
6	62.65	217.28	209.29	485.60		6	300
7	95.05	375.41	374.80	734.41		7	400
8	181.66	590.22	568.52	958.95		8	500
9	327.12	838.31	851.75	1144.19		9	600
10	580.65	1068.51	1091.39	1277.62		10	700
11	873.56	1203.21	1213.89	1474.63		11	800
12	1039.97	1291.53	1295.17	1620.87		12	900
13	1047.14	1311.23	1310.35	1672.92		13	1000
14	1247.26	1308.61	1303.85	1374.72			

Table 23. Evaluation of A001 to A099 recorded IVC ((10^6 cells/mL).h) for selection of progression points. This was due to both variation in Elapsed Time for the samples and the effects from different growth conditions. The user-defined progression points were selected using the means and medians of the daily samples as a basic guide to appropriate intervals.

Daily monitoring datasets were combined with the appropriate complementary informative value datasets in both Profile and Day by Day arrangements. Data were imported to Matlab for analysis using the Eigenvector PLS Toolbox. iPCA was used to estimate missing values. PLSR models were created for all combinations of the options described below using random sampling for cross-validation. Due to differences in data sample numbers, Day by Day models used 10 splits with 10 iterations and Profile models used 7 splits with 5 iterations. Models were selected based on minimum RMSE during cross-validation. Tables fully detailing combinations and details of the models generated can be found in Table 48, Table 49, Table 50, and Table 51 in the Appendix B.

Arrangement

1. Profile - serial observations of a culture treated as a single sample.
2. Day by Day - serial observations as multiple samples.

Data Used

1. Daily Monitoring - data collected through daily monitoring samples.
2. Daily Monitoring and Online Monitoring - the Daily Monitoring dataset expanded to include data collected through online monitoring of cultures and summarised using a subset of Informative Values 7.0. (Table 42 in Appendix B).

Variables Used

1. No Obvious Indicators – Elapsed Time and IVC excluded from input dataset.
2. Elapsed Time – IVC excluded from input dataset.
3. IVC – Elapsed Time excluded from input dataset.

Output (Response to be Modelled)

1. Product Concentration (mg/L) – For models using the Day by Day arrangement, this refers to the product concentration recorded for each individual sample. For models using the Profile arrangement, this refers to the product concentration recorded for the Day 11 sample.
2. Viability – For models using the Day by Day arrangement, this refers to the viability recorded for each individual sample. For models using the Profile arrangement, this refer to the viability recorded for the Day 11 sample.
3. Viable Cell Concentration (VCC) – For models using the Day by Day arrangement, this refers to the VCC recorded for each individual sample. For models using the Profile arrangement, this refer to the VCC recorded for the Day 11 sample.

During model creation, the response to be modelled was excluded from the input dataset. When modelling VCC, Total Cell Concentration was also excluded from the input dataset as it was highly correlated with VCC. Detailed results of these models can be found in Appendix B (Table 50 and Table 51).

5.7.6 Stage 2: Results and Discussion

Effects of model options were evaluated by creating figures to visualise differences comparing model performance criteria. Each figure shows a different way of grouping results to focus on method options. Evaluations were made with respect to the following questions.

1. Does realignment to standardised progression values offer any real improvement in the ability to capture culture behaviours?
2. Does alignment to standardised progression values eliminate the need for the progression variable in the model?

5.7.6.1 Does realignment to standardised progression values offer any real improvement in the ability to capture culture behaviours?

Two main performance criteria were used to compare models' behaviour capture capabilities. The first was model predictive ability, which was evaluated using cross-validation R^2 (Figure 28). The second was model robustness, which was evaluated using the difference between calibration R^2 and cross- validation R^2 (Figure 29).

Figure 28 shows cross-validation R^2 for datasets where explicit progression variables have been excluded and IVC, Time, or No alignment has been applied to the dataset. Here it was observed that models generated from Day by Day arrangements strongly outperformed models generated from Profile arrangements. These results were thought to be due to a combination of the number of unique variables to be modelled per sample (an order of magnitude greater than Day by Day samples) and realignment introducing more noise to the dataset than it removed.

With regards to realignment effects, realignment typically results in higher cross-validation R^2 for all response modelled using data in the Day by Day arrangement. Realignment had a negative impact on cross-validation R^2 for all response modelled using data in the Profile arrangement. The extent to which alignment choice affected cross-validation R^2 was dependent on the data used and the modelled response.

Similar to the results of Stage 1 analyses, interpretation of models created from Profile arranged data was difficult as drill down analyses were complicated by the large number of variables from which latent variables were composed.

Figure 29 shows the difference between calibration R^2 and cross-validation R^2 for datasets where explicit progression variables have been excluded and IVC, Time, or No alignment has been applied to the dataset. The difference between calibration R^2 and cross-validation R^2 was used as an indicator of model robustness, with a lower value indicating greater model robustness. Similar to Figure 28 observations, it was observed that models generated from Day by Day arrangements had greater robustness during cross-validation. Again, this was believed to be due to the number of unique variables to be modelled.

Effects on model robustness from realignment could not be as easily generalised as effects on model cross-validation R^2. For models created from data in the Day by Day arrangement, realignment improved model robustness. For models created from data in the Profile arrangement, models created from IVC aligned datasets had consistently better robustness than time aligned datasets. Models created from un-aligned datasets had better robustness than IVC aligned datasets with the exception of models created from both Daily Monitoring and Online Monitoring datasets to predict product concentration. Similar to cross-validation R^2, the extent to which alignment choice affected robustness was dependent on the data used and the modelled response.

The introduction of more noise than was removed when data were aligned to standardised progression points was not entirely unexpected. Simple linear interpolation was used for realignment, which, as described in Section 5.4.2 on missing data estimation, is not typically an accurate representation of variable behaviour. Given the improvements seen in Day by Day models, further development of the equations used for alignment may yield better results.

Finally, it was noted in all comparisons that performance and robustness were dependent on the response being modelled. Models predicting product concentration and VCC consistently outperformed models predicting viability. This was likely due to the distributions of the measured response values. Specifically, a typical fed-batch bioculture's viability follows a distinct "hockey stick" shape – viability remains steady for the majority of the culture, ideally in the region of ~90% to ~100%. Viability declines rapidly in the last days of the culture. This gives a response dataset where the majority of

values lie in a limited region (~90% to ~100%) and a minority of values spanning the remaining range (0% to 90%).

5.7.6.2 Does alignment to standardised progression values eliminate the need for the progression variable in the model?

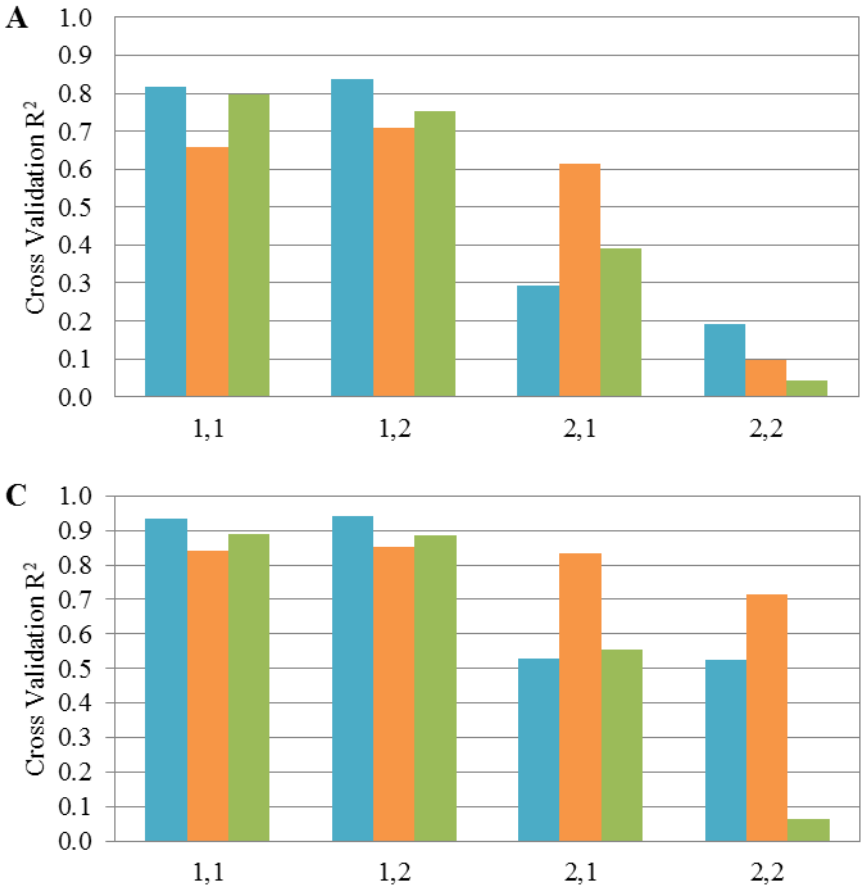
Figure 30 and Figure 31 compare the cross-validation R^2 and the difference between calibration R^2 and cross-validation R^2 respectively for models where the dataset has been realigned and the progression variable either included or excluded. In both figures it can be observed that inclusion or exclusion of the progression variables has few appreciable effects on either measure of model performance. Hence as a general rule, once data were aligned to standardised progression points, the progression variable could be excluded from input data with negligible loss in predictive power.

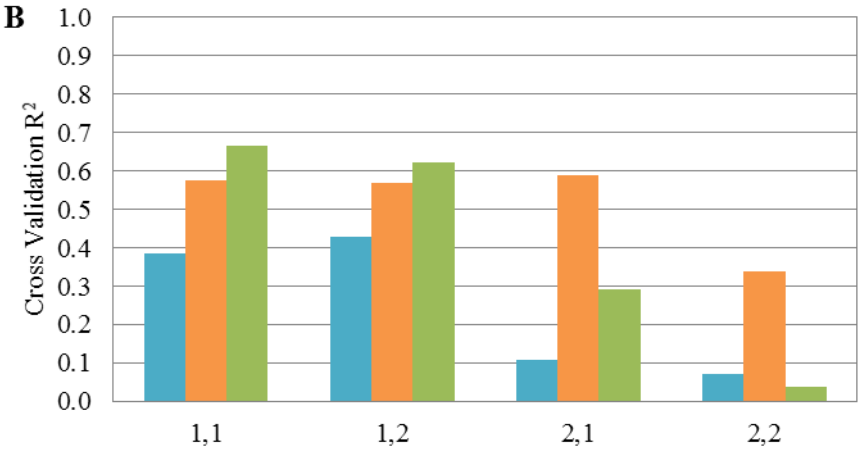
5.8 Stage 2: Conclusions

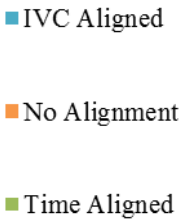
In Stage 2, it was demonstrated that in general IVC was a more robust indicator/progression variable than Elapsed Time when establishing a baseline for culture progress. This applied both when progression variables were included in model input datasets and when realigning data to standardised progression points.

For the datasets analysed, realignment to standardised progression points allowed for more robust models when considering data in the Day by Day arrangement even when the progression variable was excluded from the input dataset. However, realignment to standardised progression points led to decreased model robustness for models created from datasets in Profile arrangement. This was thought to be due to both the increased ratio of variables to samples caused by the Profile arrangement and the introduction of noise during the realignment process. Finally, the strength of these effects were dependent on the response to be modelled.

As a general conclusion, no generic "best" method could be identified. Instead it was demonstrated that the accuracy and robustness of generated models could be greatly altered by manipulating the original dataset in seemingly simple ways. It is recommended that multiple perspectives of the dataset are model during initial investigations, particularly as applying these manipulations required comparatively little effort after manipulation tools were created in Excel. From these multiple models, a number can selected based on robustness and interpretability for more in-depth analysis.

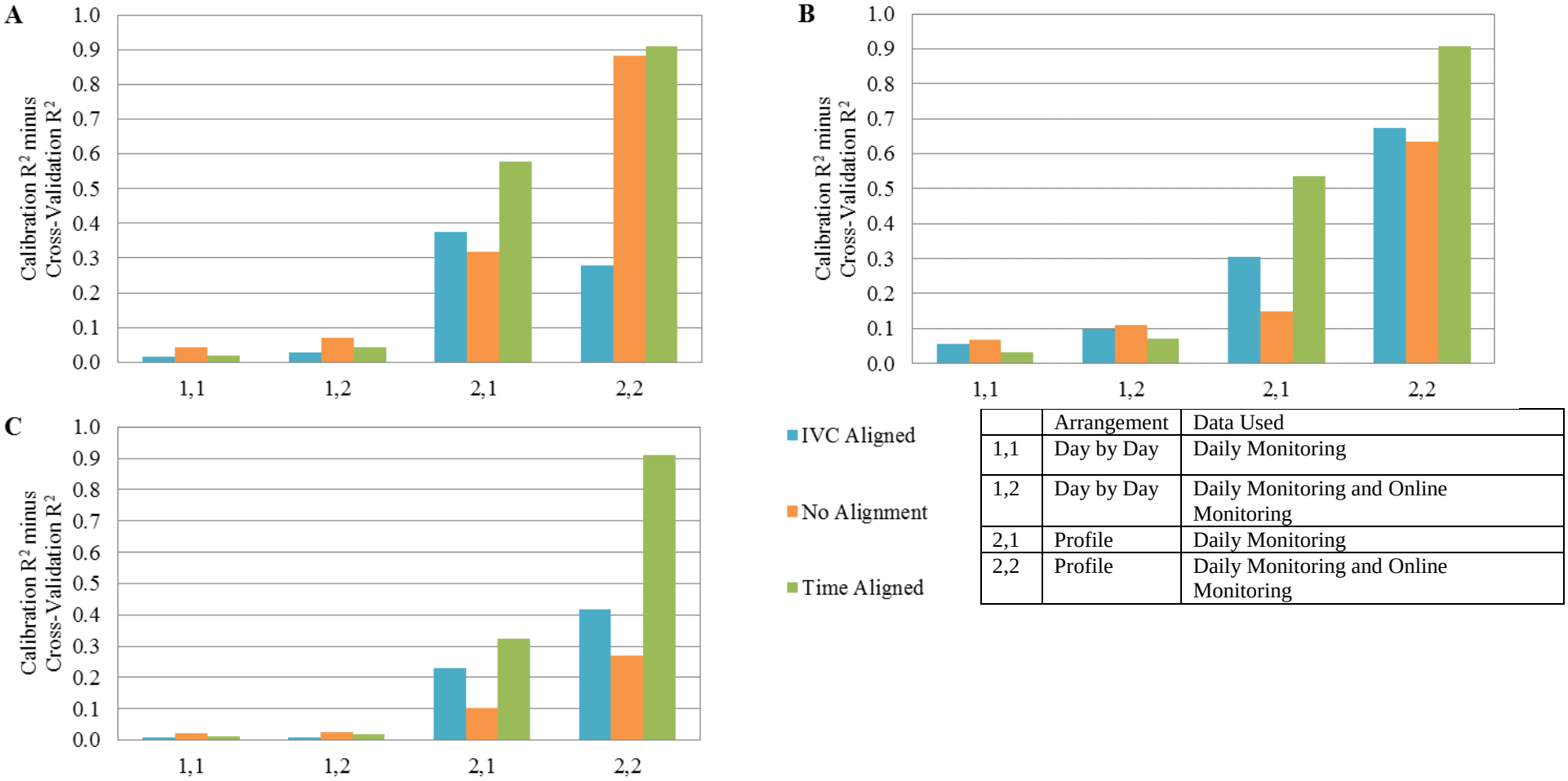






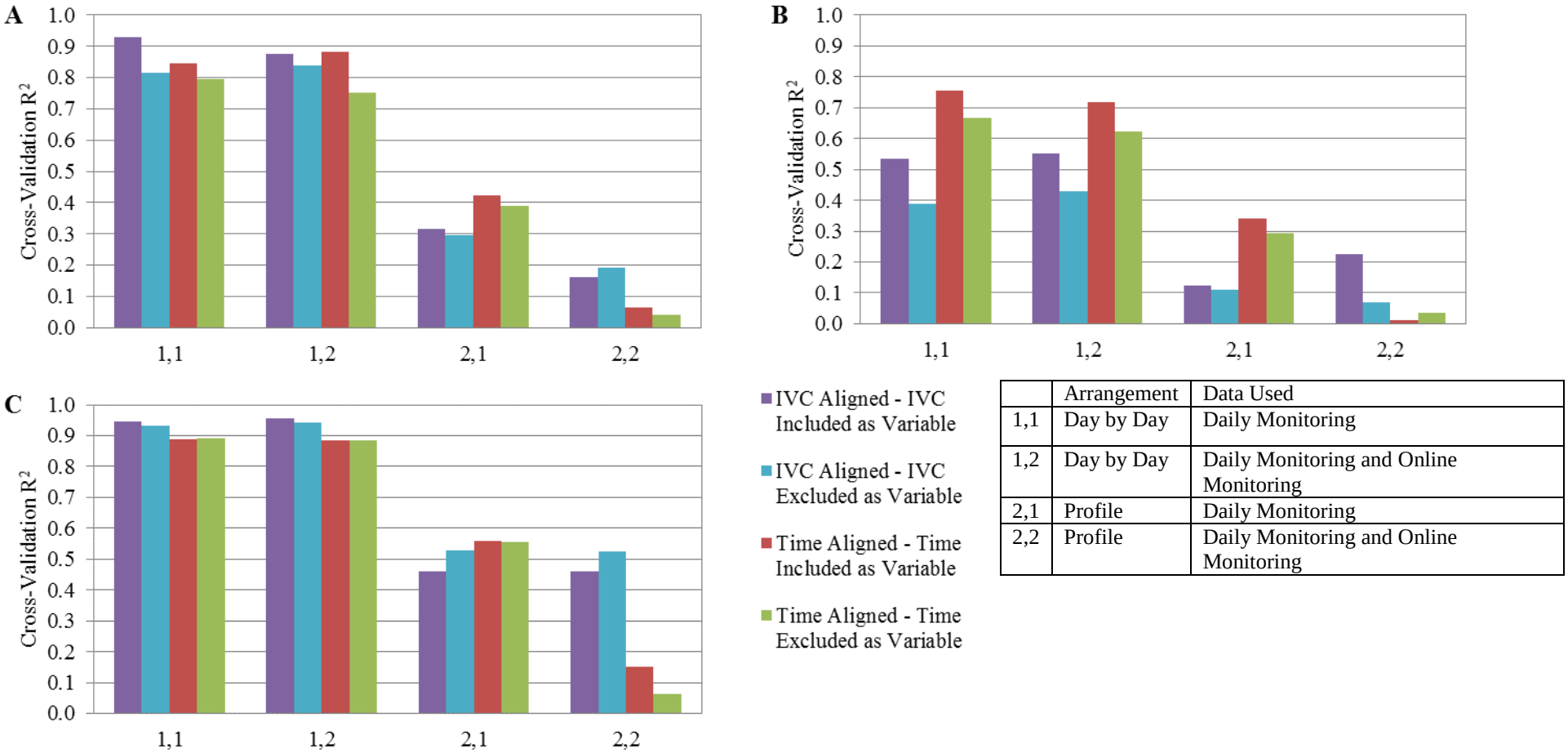
	Arrangement	Data Used
1,1	Day by Day	Daily Monitoring
1,2	Day by Day	Daily Monitoring and Online Monitoring
2,1	Profile	Daily Monitoring
2,2	Profile	Daily Monitoring and Online Monitoring

Figure 28. Evaluating effects of dataset realignment on model predictive accuracy through cross-validation R^2 when predicting product concentration (A), viability (B), and viable cell concentration (C). Overall, models created using datasets in a Day by Day arrangement had high R^2 during cross-validation. Effects of realignment were dependent on the response tested and dataset arrangement. In the profile arrangement, no alignment tended to give higher values for R^2 during cross-validation.



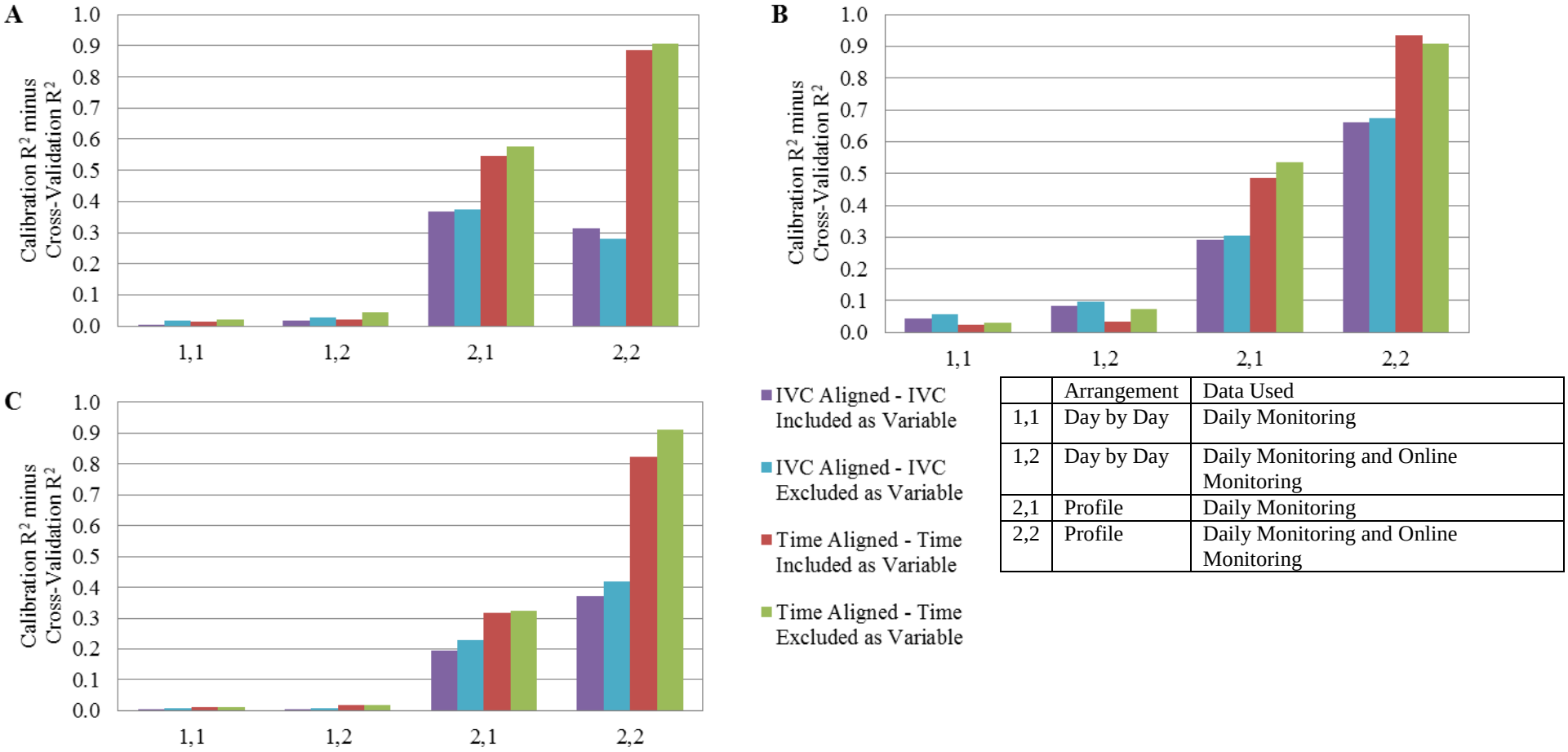
	Arrangement	Data Used
1,1	Day by Day	Daily Monitoring
1,2	Day by Day	Daily Monitoring and Online Monitoring
2,1	Profile	Daily Monitoring
2,2	Profile	Daily Monitoring and Online Monitoring

Figure 29. Evaluating effects of dataset realignment on model robustness through differences in calibration R^2 and cross validation R^2 when predicting product concentration (A), viability (B), and viable cell concentration (C). Here the lower the difference between R^2, the greater the model's robustness. Overall, models created from datasets in the Day by Day arrangement had good robustness.



	Arrangement	Data Used
1,1	Day by Day	Daily Monitoring
1,2	Day by Day	Daily Monitoring and Online Monitoring
2,1	Profile	Daily Monitoring
2,2	Profile	Daily Monitoring and Online Monitoring

Figure 30. Evaluating effects of inclusion of progression variable in a realigned dataset on model predictive accuracy through cross-validation R^2 when predicting product concentration (A), viability (B), and viable cell concentration (C). In nearly all cases, inclusion or exclusion of the progression variable does not notably alter the cross-validation R^2 for models where the cross-validation R^2 indicates a functioning model.



	Arrangement	Data Used
1,1	Day by Day	Daily Monitoring
1,2	Day by Day	Daily Monitoring and Online Monitoring
2,1	Profile	Daily Monitoring
2,2	Profile	Daily Monitoring and Online Monitoring

Figure 31. Evaluating effects of inclusion of progression variable in a realigned dataset on model robustness through differences in calibration R^2 and cross-validation R^2 when predicting product concentration (A), viability (B), and viable cell concentration (C). Here the lower the difference between calibration R^2 and cross-validation, the greater the model's robustness. In nearly all cases, inclusion or exclusion of the progression variable does not notably alter the robustness of the model.

5.9 Stage 3 Media Analysis

Due to the lack of a conclusive cause for variation in product titre during Stage 1 and Stage 2, the decision was made to focus on the media components used.

A major area of research in cell culture in the 1990s was the development of chemically-defined media and sera, which would eliminate or reduce the issues associated with animal-derived sera (Table 24). Chemically defined media and feeds can often be considered the most valuable and guarded asset of cell culture companies. However while these are thought of as set recipes, variation is still possible, including inherent variation in at the smallest measurement scales, e.g. nanomolar concentrations of trace elements. This variation may be inconsequential, have negligible effects, or could notably alter cellular behaviour. Hence, methods of determining whether variation in the base materials' composition may be a factor in an investigation of cell culture performance are a core factor in Quality by Design frameworks.

Advantages
— Binding and neutralisation of toxins. — Protease inhibition. — In agitated bioreactors, protection of cells from mechanical damage. — Buffer capacity of cell culture mixture improved. — Contain growth factors, hormones, and adherence factors.
Disadvantages
— Cost. — Lot-to-lot variability in composition (and potential impacts therefrom). — Negative impacts on both up- and down-stream processing, e.g. foaming in bioreactors or inference with columns, and associated increases in operating difficulties and costs. — Chemically undefined, e.g. unknown recipe, and undesired constituent chemicals such as growth and metabolism inhibitors. — Safety risk due to possible infection by viruses and other adventitious agents, such as those involved in Bovine spongiform encephalopathy [82].

Table 24. Summary of advantages and disadvantages of bovine foetal serum in cell culture media [167].

An acknowledged issue in the investigations performed in Stage 1 and Stage 2 was the data concerning media and feed compositions were not included during analysis. This due to both the categorical nature of the data and simple availability. Media batch numbers required transcription from handwritten records, representing a notable delay, particularly for an investigation with a large number of cultures where accompanying paper records are archived off-site for various reasons.

A total of 9 media components known to be key components in the process were chosen for analysis through identification of clustering. Only cultures A001 to A049 were considered for two reasons. First to determine if such analysis could have been used to identify and resolve issues before the site transfer was made. Second, media component batch numbers could not be accessed for the US-cited cultures A050 to A099.

5.10 Stage 3: Method

All available daily monitoring data for cultures A001 to A049 were arranged in the Day by Day orientation. As sample temperature and DOT were not recorded as part of daily monitoring at the 130 L scale, estimated values were created from online monitoring records by calculating the median between sampling timepoints. All remaining samples with missing data were eliminated.

The dataset was mean-centred and scaled to unit variance. A PCA model was generated in PLS Toolbox using random sampling (7 splits, 5 iterations). A 6 PC model capturing 83.54 % of variation was selected.

Scores plots were created for each combination of scores (e.g. PC3 v. PC6). Additional figures were created by plotting a PC's scores against time. Variations of these figures were created by plotting scores from only cultures operating with control conditions (Figure 32). Figures were then analysed for clustering by batch number for the 9 media components.

5.11 Stage 3: Results and Discussion

Several variations on the described method were attempted, e.g. use of PLSR in place of PCA, reorientation of the dataset from the Day by Day orientation to Profile orientation. Each method failed to improve understanding of product concentration variation for the same reason: Only media component batch numbers were recorded, not data pertaining to the actual physical differences between batches.

This issue was further compounded by the fact that the combinations of media component batches used were effectively unique to culture round[6]. Hence any clustering scene could only be attributed to the combination of media component batches used in that round and not to a specific component.

5.12 Stage 3: Conclusions and Recommendations

While recording of media component batch numbers ensures traceability, it does not allow for quantitative comparison of batch chemical composition. Batch number records can allow this data to be captured at a later date, however this assumes quantities of the component in question are still available and are of an identical state to when used in the cultures being investigated, e.g. no effects from aging or storage.

The use of technologies such as near-infrared spectroscopy to analyse media components before use would greatly benefit investigation such as the one presented in this chapter. In addition to being a potential screening step to prevent use of media with undesirable concentrations, the quantitative data generated by such techniques could be integrated into a single dataset with daily monitoring and online monitoring data. This could allow direct correlations between specific chemicals within media components to be correlated with observed biological behaviour.

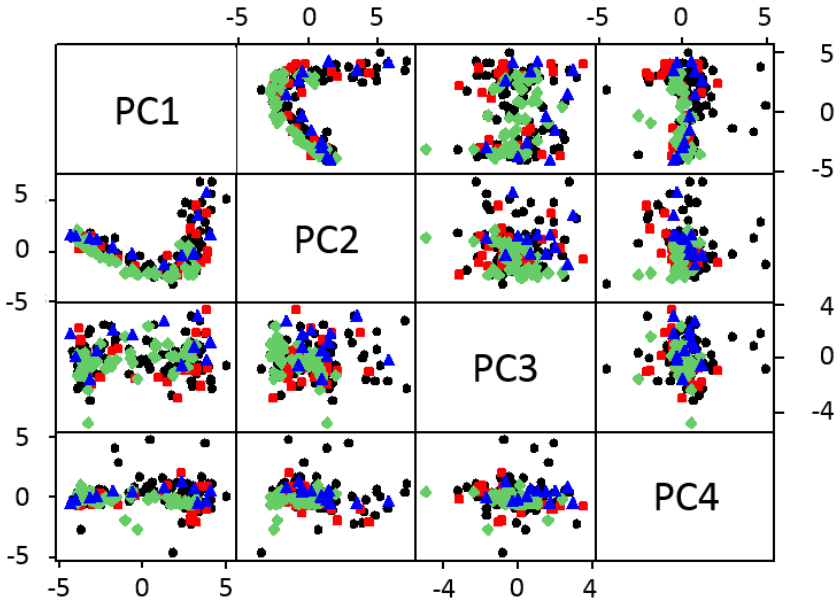


Figure 32. PC1 to PC4 scores for cultures operating at control operating conditions. Cultures are coloured according SF66 Choline Chloride batch number: ● 019K0066 ■ 070M0192V ♦ 119K0078 ▲ BCBD3356V.

[6] Culture round refers to groups of cultures being performed at the same time, which allows resources such as media to be prepared in bulk.

5.13 Productivity Investigation Conclusions

A wide number of statistical techniques and approaches were employed during the investigation in to variation in product concentrations at harvest for Project A. The productivity investigation comprised three stages.

In Stage 1 a basic framework was developed, incorporating several MVDA tools. For the developed framework, two datasets were available for analysis: online monitoring and daily monitoring. Due to differences in sampling frequency, these datasets could not be integrated as a single dataset as the online monitoring dataset far outsized the daily monitoring dataset. The development of robust summary statistics termed Informative Values (see Appendix A) allowed the two dataset to be integrated as a single, balanced dataset. In the developed framework, this allowed multidimensional pass/fail boundaries to be identified from the most comprehensive dataset possible.

Even with this reduction of the high frequency online monitoring dataset, the large number of total variables to be analysed posed a challenge in subsequent analysis and classification of cultures as high or low producing cultures by decision trees. To reduce the likelihood of spurious decision criteria selection and the loss of contextual information when a single variable is selected as a decision criteria, the integrated dataset was first dimensionally reduced using PCA. Scores from the resulting PCA model were then used in the subsequent decision tree in place of the unreduced dataset. In developing this framework, it was demonstrated that univariate methods for estimating missing data led to higher misclassification errors than when the multivariate iterative PCA was used.

Furthermore, it was demonstrated that a "Golden Batch" approach, whereby models are trained using only 'good' samples, resulted in less robust classification models when applied to data including a range of behaviour. This range of behaviours included both deliberate changes to test the effects of potential issues, in addition to samples exhibiting unusual behaviours of interest (e.g. control condition cultures with low product concentrations at harvest).

From this work, initial seeding conditions and subsequent glutamine behaviour were indicated as the main areas of deviant behaviour for the first pilot scale culture (A016). While a specific cause for the altered glutamine behaviour could not be identified from the data analysed, it was suggested that batch-to-batch variation in media powders used could be a possible contributing factor.

In Stage 2, the dataset was expanded to include data from cultures performed at Lonza's US site. An initial attempt to identify variables of interest through PLSR was unsuccessful due to a change in seeding protocol between sites. This acted as a confounding variables and caused US cultures to appear more mature than UK cultures with respect to time. To overcome this source of confounding, the default progression variable (Elapsed Time) was questioned as the most appropriate measure of culture progress and maturation. Integral of Viable Cell Concentration (IVC) was suggested as an alternative measure of progress. For models predicting daily product concentration created from the UK dataset, it was shown that models where IVC was used in place of Elapsed Time had lower prediction errors and better distribution of residuals. More importantly, models created with IVC used as progression variable in place of Elapsed Time had greater robustness when UK data were applied as a test dataset.

During Stage 2, it was shown that inclusion of an explicit progression variable yielded a higher cross-validation R^2, better residual distribution, and improved robustness to testing datasets than exclusion. It was thought that this was due to variation in sampling times introducing variation in sample progress value, which had to be accounted for in the models. It was theorised that realigning datasets to set progression point values would eliminate this underlying variation and hence progression variables could be excluded during model creation with little or no effect on model performances. Testing revealed that any benefits of realignment were dependent on dataset orientation and the response of interest.

Stage 2 had several learning points pertinent to improved future analyses of projects with site-transfers, simple sources of confounding (e.g. altered seeding conditions), or variation in sampling times. Specific results from Stage 2 models confirmed previous results from Stage 1, however no further understanding concerning variation in product concentration at harvest could be extracted.

The aim of Stage 3 was to investigate whether variation in product concentration at harvest could be attributed to variation in media batch composition. Here it was demonstrated that batch numbers do not convey any information concerning the chemical composition of the batch in question. This situation lends support to on-going PAT activities focussed on the introduction of spectral measurement devices such as Raman probes as a means of generating more comprehensive understanding of culture behaviours.

Chapter 6. Multi-Product Platform Process Analysis

6.1 Introduction

A general definition of a platform process or platform technology is "a common or standard method, equipment, procedure or work practice that may be applied to the research, development or manufacture of different products" [20]. Platform manufacturing can be defined as the "implementation of standard technologies, systems and work practices within manufacturing facilities, and their use for the manufacture of different products OR the approach of developing a production strategy for a new drug starting from manufacturing process similar to those used by the same manufacturer to manufacture other drugs of the same type" [20]. Due to the wide range of interpretations, the following definitions are used here:

Platform Process: A process where major operating parameters such as control system setpoints and deadbands, feed strategies, and media are pre-defined for the purpose of producing multiple products.

Platform Manufacturing: "implementation of standard technologies, systems and work practices within manufacturing facilities, and their use for the manufacture of different products" [20].

Platform Research: "the approach of developing a production strategy for a new drug starting from manufacturing process similar to those used by the same manufacturer to manufacture other drugs of the same type" [20].

The use of process platforms allows for standardisation of approaches and tools. Consequently, savings in time and money can be made through:

- Better utilisation of resources including equipment, materials, and personnel.
- Improved quality and/or greater consistency in product quality (e.g. less wastage due to intermediate or final product failing to meet quality criteria).

These benefits are common to any industry where platform technologies might be employed. In particular for pharmaceutical and biopharmaceutical process, there may also be time and money savings in regulatory applications if use of a platform process is supported by a sufficient level of evidence for platform understanding and robustness.

In the production of monoclonal antibodies (mAb) and other protein-based therapeutics by mammalian cells, successful transfer of a mAb-producing cell line to the platform process may require adaptation by the cell line. When comparing performances on the old and new process, such as the transfer of a mAb-producing cell line from one company's own process to a contract manufacturer's process platform, there may be differences observed such as altered harvest titre, differences in metabolism, or alterations to the product (e.g. glycan profile affected). These may be desired or undesired.

In biopharmaceutical production, ideally the process is robust to a wide variety of cell lines. Here, any variability is a function of the transfected cell line and potentially addressed through minor changes to the process, e.g. inoculating with a higher VCC. If undesired behaviours are observed across multiple products sharing a process, it suggests a common cause related to the process itself.

This occurred for a process platform utilised by Lonza. Sudden declines in culture viability (~80% to ~0% in 24 h) were observed in multiple projects for different products. All projects utilised the GS-CHO cell line with the Version 6 GS-CHO process platform. These crashes led to an intensive analysis in 2006 which took many man-months to accomplish.

6.2 The Dataset

The data used in this investigation originated from 17 customer projects performed using Lonza's GS-CHO Version 6 platform process. The 17 projects were selected in a list created by Lonza. The number of cultures included from each project ranged from 1 to 50 with an overall total of 185 cultures of multiple scale from wave bag to 5000 L.

6.3 Aims

The main aim was to identify indicators of sudden declines in culture viability and, if possible, determine potential causes. Additional aims were to build upon the strategies used in Chapter 6, with particular regards for result interpretability.

6.4 Obstacles

Six main obstacles were identified in the dataset that required resolution (Table 25). These obstacles are discussed in addition to the suggested resolution that was tested during the final presented method.

Obstacle	Summary of Solution(s)
1. Distribution of Crashes and Data Disparity	1. Removal of wave bag cultures from dataset. 2. Analysed data restricted to data collected through daily monitoring.
2. Definition of Crash Rate	1. Standardised tool for classification of samples according to maximum rate of decline
3. Definition of Crash According to Culture Stage	1. Rejustification of dataset to emphasise: a. Behaviour since inoculation b. Behaviour before crash/harvest c. Behaviour at peak VCC
4. Confounding by Expressed Product	1. Two-step scaling process (intrascaling) applied. 2. Pro_013 removed from dataset due to use of different cell line.
5. Interpretation of Multivariate Serial Observations	1. Use of purpose built models to identify either Days of Interest or Variables of Interest 2. Use of model hierarchies to better manage "information overload".
6. Robustness	1. Use of a multiple model confirmatory approach.

Table 25. Obstacles identified during platform process investigation including suggested solutions.

6.4.1 Obstacle 1: Distribution of Crashes and Data Disparity

The distribution of cultures by project and scale is given in Figure 33 and Figure 34. It can be seen that there was unequal representation of both projects and scales. As seen in Figure 33, the 10 L scale had greatest representation, accounting for 83% of all cultures (154 of 185) under consideration. Within the 10 L band (Figure 34), project Pro_014 had the greatest representation with 32% of cultures under consideration at that scale (50 out of 154) and 27% of all cultures (50 out of 185).

In addition to difference in distribution of crashes across projects, there was disparity in data collection across both scale and time. As noted in Chapter 5, osmolality was not recorded as part of daily monitoring at scales above 10L. Below the 10L scale, cultures were performed in roller bottles, wave bags, or shake flasks where daily monitoring captured only a limited subset of the variables recorded for cultures at scales of 10L and above. Furthermore no online monitoring data existed for cultures below the 10L scale.

Finally, due to the time span considered, pO_2 was only recorded for cultures following a change in daily sampling data collection.

6.4.1.1 Resolution

Wave bag cultures were eliminated due to the limited number of variables monitored when compared to scales above wave bag. Between 10 L and 5000 L, reactor design and monitored variables are highly conserved with limited differences in variables monitored. Due to issues with availability and discoverability[7] of online monitoring data, only data from daily offline sampling were collated (Table 26). Viable cell concentration (VCC), IVC, and total cell concentration (TCC) were collated but excluded from analysis due to high correlation with viability ('obvious indicator').

Finally, it was found that Pro_013 had used a GS-NS0 cell line and was hence excluded as use of a different host cell line meant a different platform process had been used.

Temperature	Viability	Lactate	NH_4^+	pCO_2
pH	IVC**	Glucose	K^+	pO_2*
DOT	VCC**	Glutamate	Na^+	Osmolality*
-	TCC**	Glutamine	NH_4^+	-

Table 26. Variables available from daily offline sampling. *Excluded due to inconsistent collection (e.g. not recorded at all scales). ** Excluded as 'obvious' indicator.

[7] Discoverability refers to the ease with which data can be found including navigation of file structures, access permissions, location maintenance, file extensions (e.g. if file requires specialist software to open), and file consistency (e.g. are all data exported as .csv with identical layout).

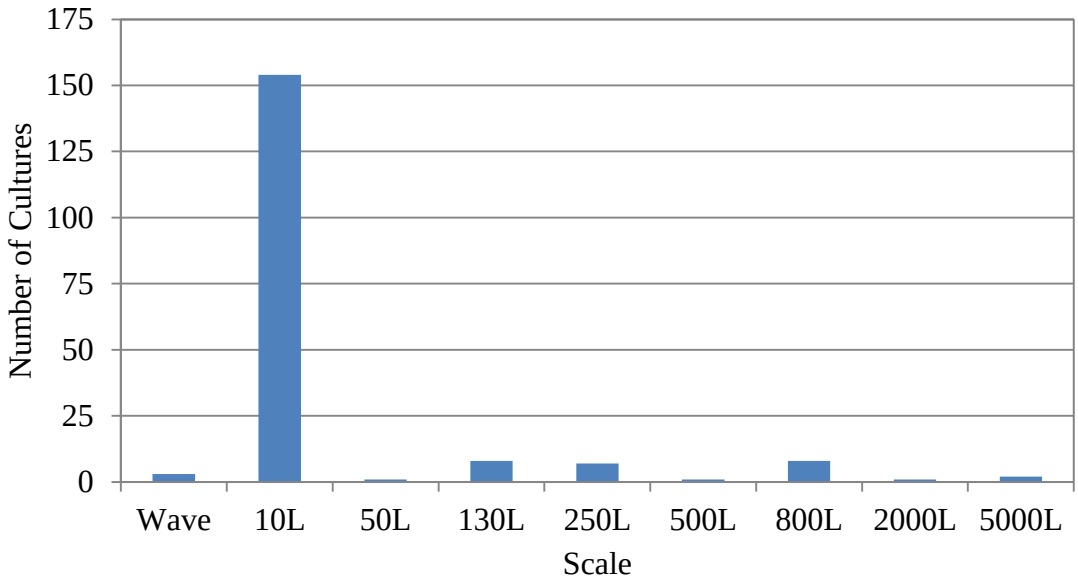


Figure 33. Distribution of Cultures by Scale. The majority of cultures available for analysis were performed at the 10L scale.

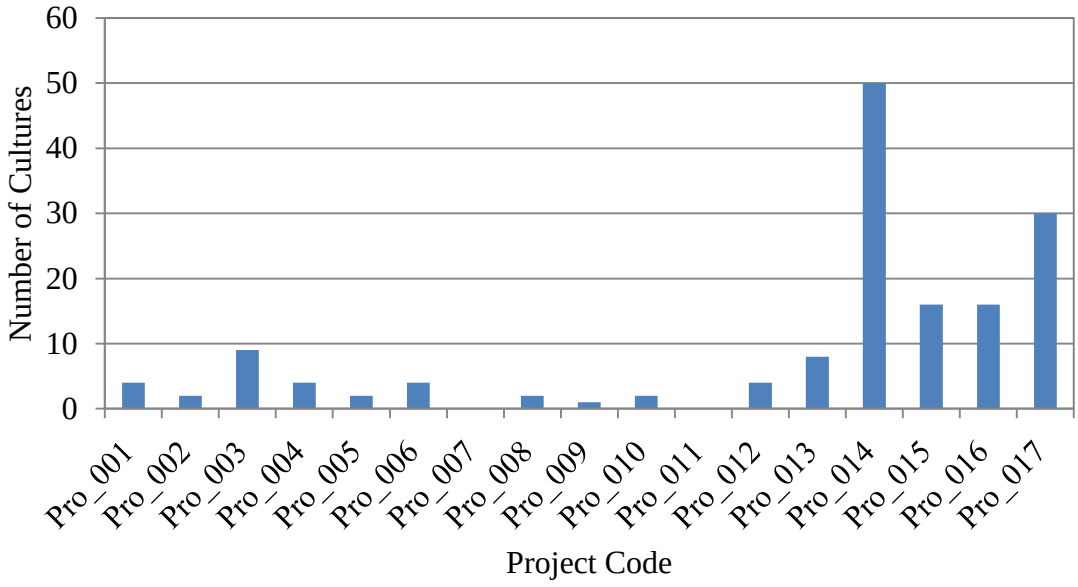


Figure 34. Distribution of Cultures by Project at 10 L Scale. There is unequal representation of projects with Pro_014 having the greatest number of cultures available for analysis.

6.4.2 Obstacle 2: Definition of Crash Rate

The definition of "crash" provided during an investigation kick-off meeting can be paraphrased as "culture viability quickly dropping over the course of 24 hours". There are three main issues with this initial definition.

First, there is no clear value for acceptable/unacceptable rates of decline. Second, lack of clear value raises questions on consistency of results including if all projects demonstrated the same rate of decline during a crash. Third, the definition of "crash" did not allow for contextual information and only uses values calculated for viability.

6.4.2.1 Resolution

Due to the large number of cultures to be analysed and a relatively quantifiable definition of pass/fail behaviour as decline in viability over twenty-four hours, it was considered neccessary to create a simple algorithm to assign classes in place of manually classifying cultures by viewing viability profiles. This had the benefit of removing subjective classification of pass/fail, adding an additional level of robustness to the process.

Once created, this algorithm allowed multiple limits for pass/fail classification to be easily assigned by changing cut-off limit value. Hence, each culture was classified as pass or fail for the following maximum decline in 24 hours:

10% 20% 30% 40% 50% 60%

There was no clear limit for pass/fail classification of cultures based on maximum calculated 24h decline in viability. A spectrum of pass/fail limits existed (Figure 35). At the more extreme limits considered (10% and 60%), there was very low representation of pass or fail cultures respectively. It was decided to test all limits (10% to 60%) to identify if identified indicators were consistent across all limits or if identified indicators were dependent on the pass/fail limit.

It was noted that in several cases, repeat cultures (i.e. two cultures of a single product with the same media, feeds, and stated operating conditions) did not have the same pass/fail profile. Closer examination revealed that this was generally due to the maximum decline in viability in 24 hours for the two cultures lying very near a classification limits, e.g. Pro_001_002 (-40.26%) and Pro_001_003 (-38.33%) at -40% or Pro_014_005 (-19.92%) and Pro_014_006 (-21.34%) at -20%.

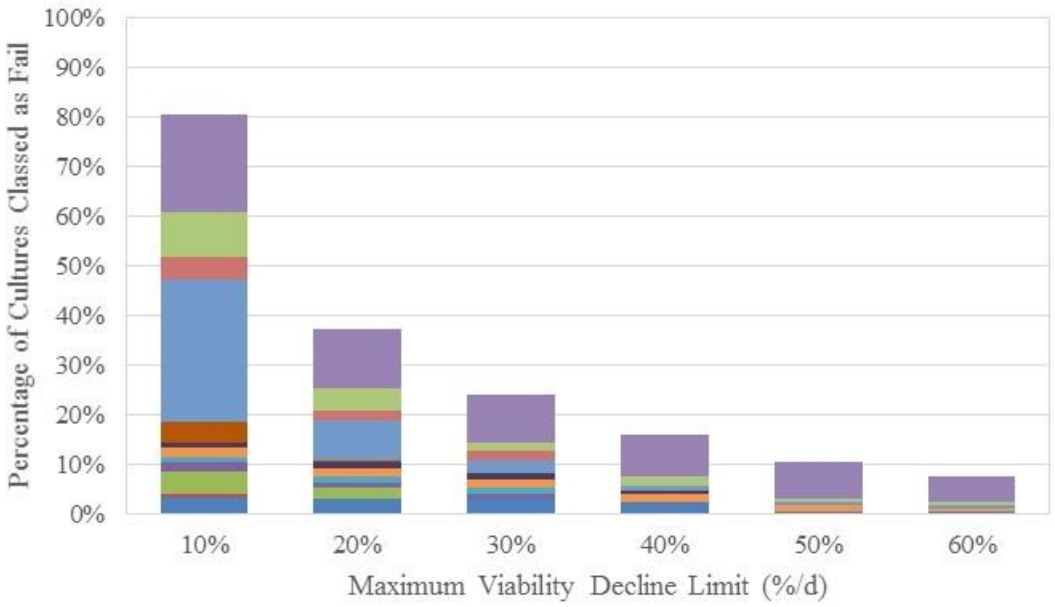


Figure 35. Percentage of cultures in dataset failing at decreasing limits for maximum calculated rate of viability decline in 24 h (excluding Pro_013, n = 174). Bars are coloured to show relative representation of projects. It can be observed that there is not equal representation of cultures at any of the maximum viability decline limits.

6.4.3 Obstacle 3: Definition of Crash According to Culture Stage

When data were collected, serial observations were arranged in what is considered a typical order: the first observation after inoculation is n=1, the second observation after inoculation is n=2, etc. As the number of observations for multiple cultures can be different due to crashes and/or meeting harvest criteria at different times, this can leave a 'rag' of missing data on the harvest side of a dataset (Table 27A).

This *de facto* arrangement complicates analysis in two ways. The first complication is how the 'rag' of missing data is dealt with, particularly as this data is not missing at random. The second complication is that this arrangement biases both model and interpretation to focus on behaviour since inoculation. However, behaviour before crash/harvest was the focus in this investigation.

6.4.3.1 Resolution

During the productivity investigation, efforts were made to remove variability in the chosen progression variable, i.e. to make the dataset sampling structure rigid. This was achieved by interpolation of data to set values for the progression variables. Set values were selected based on variable distribution to minimise the magnitude of adjustment made. In summary, dataset justification is a value-based manipulation of the dataset.

In contrast, dataset justification is a structure-based manipulation of the dataset. The purpose is to manipulate the behaviours identified during MVDA by altering the dataset structure to emphasise specific aspects.

In this investigation two rearrangements were considered. To improve model focus on behaviour preceding crash/harvest, data were arranged as shown in Table 27B so that the observations were counted n-1, n-2, n-3, etc.

The second rearrangement distributed the data according to maximum recorded VCC (Table 27C). Here, data are arranged so that the maximum VCC occurs in the same observation/column and the data 'rag' can be on either side of the dataset or distributed across both. Theoretically, this should bias model focus towards behaviour concerning this specific event.

A) Inoculation Justified

	Viable Cell Concentration (N= First Observation)								
Culture	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+9
A	2	25.4	50.7	74.6	89.9	99.7	96.5	90	
B	0.2	24.9	50.4	73.4	84.8	99.1	97.4	92.5	75.4
C	1	25	50.1	75.7	84.3	93.2	97.1	100	78.7
D	0.6	24.7	49.8	73.2	86.9	98.1	95.6		
E	1	25.3	50.8	72	82.9	95.4	99.9	91.4	

B) Harvest Justified

	Viable Cell Concentration (N = Final Observation)								
Culture	N-8	N-7	N-6	N-5	N-4	N-3	N-2	N-1	N
A		2	25.4	50.7	74.6	89.9	99.7	96.5	90
B	0.2	24.9	50.4	73.4	84.8	99.1	97.4	92.5	75.4
C	1	25	50.1	75.7	84.3	93.2	97.1	100	78.7
D			0.6	24.7	49.8	73.2	86.9	98.1	95.6
E		1	25.3	50.8	72	82.9	95.4	99.9	91.4

C) Peak Viable Cell Concentration Centred

	Viable Cell Concentration (N = Observation with Maximum VCC)								
Culture	N-6	N-5	N-4	N-3	N-2	N-1	N	N+1	N+2
A		2	25.4	50.7	74.6	89.9	99.7	96.5	90
B		0.2	24.9	50.4	73.4	84.8	99.1	97.4	92.5
C	25	50.1	75.7	84.3	93.2	97.1	100	78.7	
D		0.6	24.7	49.8	73.2	86.9	98.1	95.6	
E	1	25.3	50.8	72	82.9	95.4	99.9	91.4	

Table 27. Effects of Dataset Justification on Location of Missing Data. Darker shading indicates missing data. Lighter shading indicates the peak value for viable cell concentration recorded. Note: data artificially generated for demonstration of concept.

6.4.4 Obstacle 4: Confounding by Expressed Product

Confounding is when some variable, termed a confounder, acts a source of variability that interferes with or masks more subtle variation in a dataset. While the analysed data were generated from a single platform process, the data was also generated from the production of several different mAbs and mAb fragments. As the product being expressed can affect host cell metabolism in both a variety of ways and extents, confounding posed an obstacle.

Several methods for removing confounding have been suggested. One of particular note is the use of PLS-DA to first classify samples according to a known confounder (e.g. product type) to "draw out" the confounded information, giving a "deconfounded" residual matrix that can then be analysed [13].

While these techniques exist, they are dependent on the data to be analysed and may require significant input from an end user either in performing methods manually or in development and validation of software to perform such actions. Furthermore, they require strong understanding of the techniques to be confidently communicated with others. While this last point is a non-technical limitation in the purest of terms, the ability to confidently communicate statistical techniques employed during an analysis are important when discussing analyses with colleagues, clients, and regulatory authorities. Effective implementation of simpler tools is of more benefit to achieving Quality by Design than attempting complex, high-powered methods which are poorly understood.

6.4.4.1 Solution

The key to the implemented solution was that the statistical methods to be employed used the correlations between variables and not the recorded values themselves. Hence it was the explicit aim of any confounding reduction method applied was to allow the correlation structures of the product-specific subsets to be compared.

The suggested solution was to employ a two-step scaling method (Figure 36) where the multiproduct dataset is first split into single product subsets. Each single product subset is scaled using the desired scaling method(s). The scaled single product subsets are then recollated into a single dataset, which can then be further treated as desired.

A trial test of intrascaling during initial exploration of the dataset using PCA indicated that this approach could reduce confounding with little to no appreciable impact on captured variance (Figure 37).

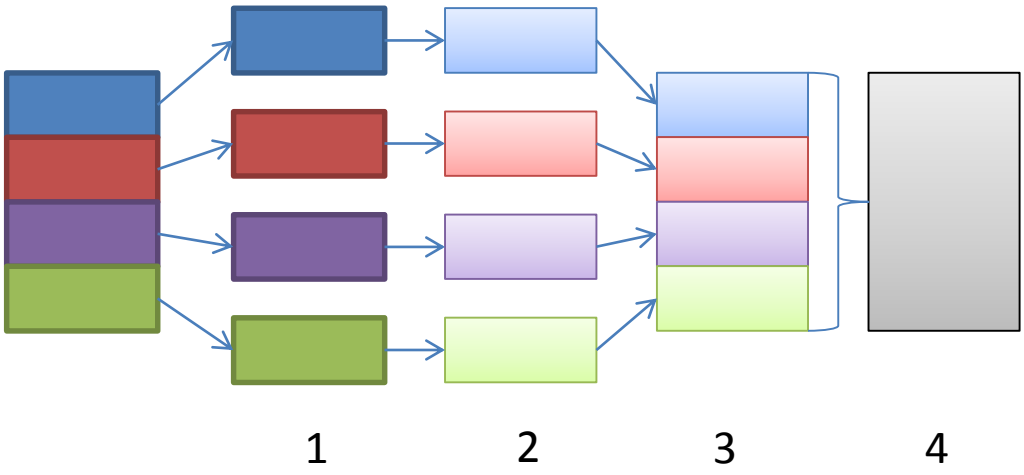


Figure 36. Simple Schematic of Intrascaling Process. The multiproject dataset is split into single project subsets (1). The single project subsets are scaled using a desired method (2). The scaled subsets are recollated into a single dataset (3). The recollated dataset is scaled using a desired method (4).

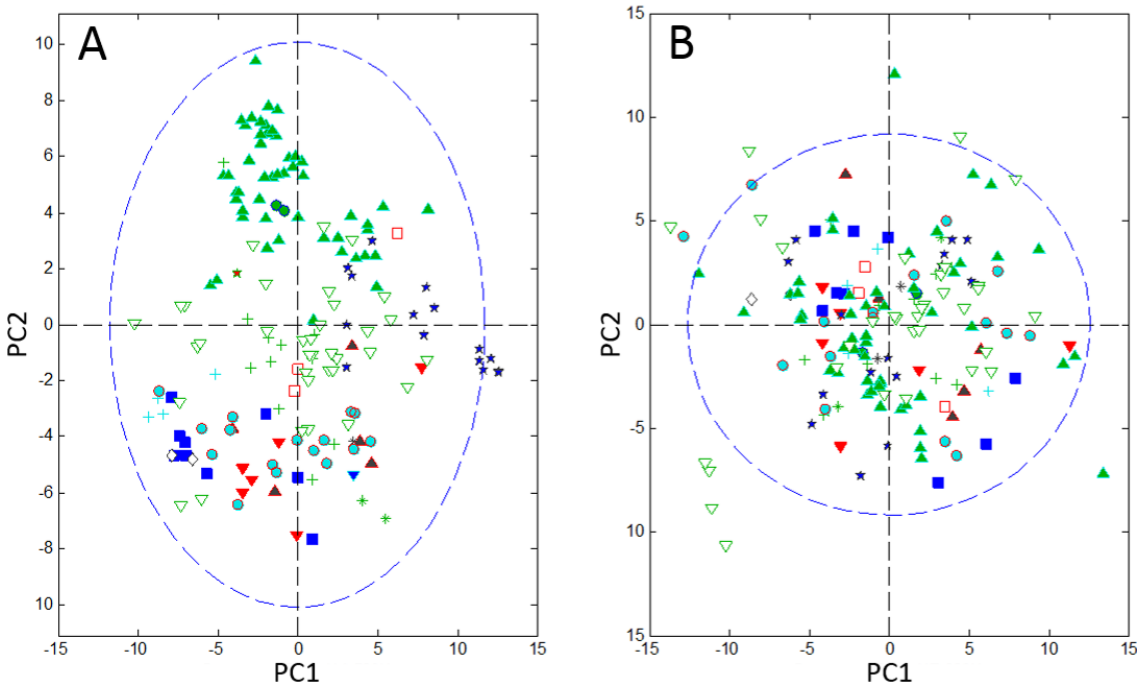


Figure 37. PCA scores plots generated using different scaling methods. Each point represents Day 1 to Day 12 for one culture. Shape and colour indicate project of origin. As cultures appeared to be more strongly cluster by project in A than in B, clustering in A was thought to be more heavily influenced by project-specific differences, whereas clustering in B was thought to be more strongly governed by culture behaviour irrespective of project.

A. Dataset mean-centred and scaled to unit variance. Total X variance captured: 25.33% (PC1 14.58%, PC2 10.75%).
B. Intrascaled dataset. Total X-variance captured: 26.22% (PC1 17.09, PC2 9.13%).

A second option based on this approach was initially considered. In the second option, variables that should not be affected by product types or vector integration site are excluded from the first scaling step. These variables are scaled only once the single project subsets are recollated into a single dataset. Variables excluded from the first scaling step include pH, DOT, and temperature. In this way, variation in behaviours related to hardware or control systems are preserved between projects.

The first option where all variables undergo the two-step scaling process was referred to as Intrascale A. The second option where only a subset of variables undergo the two-step scaling process was referred to as Intrascale B. Only Intrascale A was tested in the presented investigation.

6.4.5 Obstacle 5: Interpretation of Multivariate Serial Observations

An observation during the Chapter 5 investigation into variation in product concentration for a single product dataset was that interpretation was complicated by the number of variables to be considered, particularly when data were in the Profile arrangement.

6.4.5.1 Solution

Multilevel or hierarchical modelling is a form of regression where regression coefficients are a function of submodels representing another level of the data. For example, a top-level model predicting a child's academic performance may use regression coefficients calculated from socio-economic data.

According to Gelman [168], hierarchical modelling is an improvement over regression "to varying degrees; for prediction multilevel modelling can be essential, for data reduction it can be useful, and for causal inference it can be helpful". It is for these last two points – data reduction and improved causal inference – that hierarchical modelling was used as a template for the presented solution.

The presented solution is not a true hierarchical model as top-level model regression coefficients are not a function of a lower level of data. Instead, the presented solution is a hierarchy of models, where results from intermediary models feed into a top level model (Figure 38 and Figure 39).

The key similarity between hierarchical modelling and the model hierarchies developed was that the focus of the top-level model could be altered by altering the focus of the intermediary model, e.g. if the intermediary models focussed on behaviour on by observation number (Figure 38), then contribution analysis of the top-level model would

indicate days of particular interest. Similarly, if the intermediary models focussed on individual variable behaviour over the course of the culture (Figure 39), then contribution analysis of the top-level model would indicate variables of particular interest.

6.4.6 Obstacle 6: Robustness

Due to the breadth of behaviour to be considered including multiple products, no definitive pass/fail limit, and variation in "when" crashes occurred, robustness of models and results was a concern.

6.4.6.1 Solution

Instead of attempting to create a single model to capture behaviour and identify indicators of crashes, it was instead decided to create several simpler models and perform a meta-analysis from the results. By altering model focus as described in §6.4.3 (rejustification of dataset) and §6.4.5 (model hierarchies to emphasise days or variables of interest), the overall aim of meta-analysis was robustness through consistency of results. This meta-analysis approach also included multiple limits for pass/fail classification (§6.4.2).

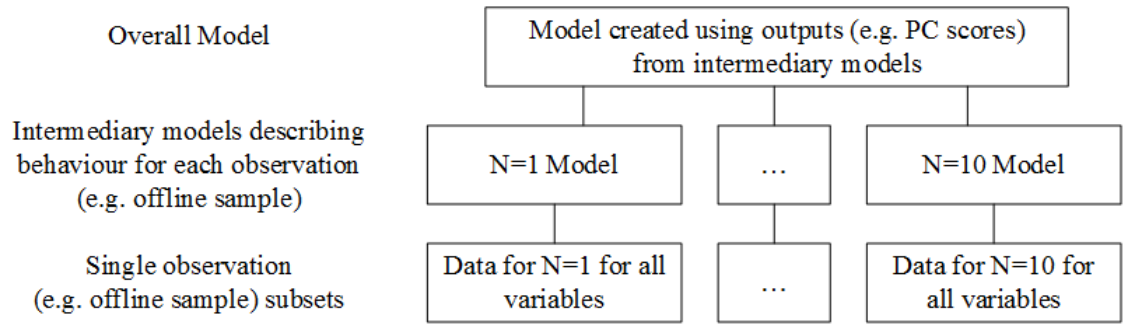


Figure 38. Model hierarchy structure producing a top level model focusing on Observation/Days of Interest.

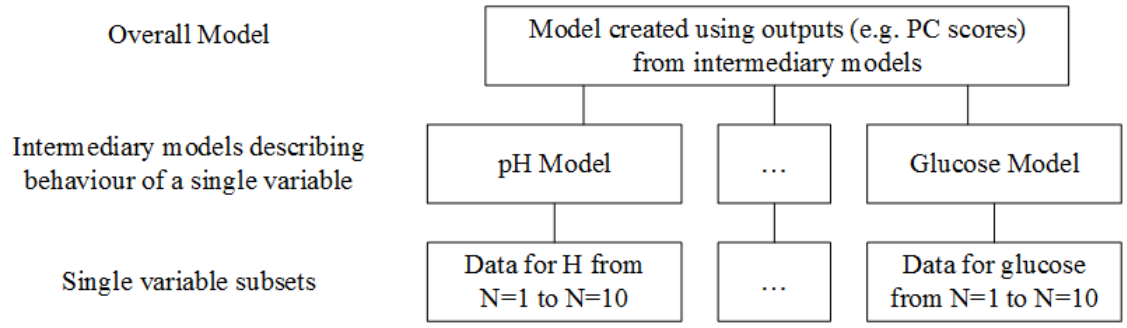


Figure 39. Model hierarchy structure producing a top level model focusing on Variables of Interest.

6.5 Method

A meta-analysis approach was created using the three analysis patterns described below and the variables listed in Table 26. The resulting models were compared for both consistency and discrepancy across the different data pre-treatment, statistical methods, and decline rate/classification limits tested. Models determined to be of particular interest then underwent more detailed results analysis. The time needed to complete the investigation would also be compared to a previous investigation undertaken by Lonza

6.5.1 Analysis Pattern 1

In Analysis Pattern 1 (Figure 40) data were arranged in Profile arrangement (1 row = all samples for 1 culture) and then Inoculation Justified, Harvest Justified, or Max VCC Centred. The dataset was scaled using two-step intrascaling or mean-centred and scaled to unit variance. Missing data were estimated using iterative PCA. A PCA model was created using random sampling (10 splits, 5 iterations). The number of PCs retained was made on minimum RMSE during cross-validation. Scores were used to classify cultures as pass or fail using PLS-DA and decision trees (Gini Index) for each maximum decline limit 10% to 60%.

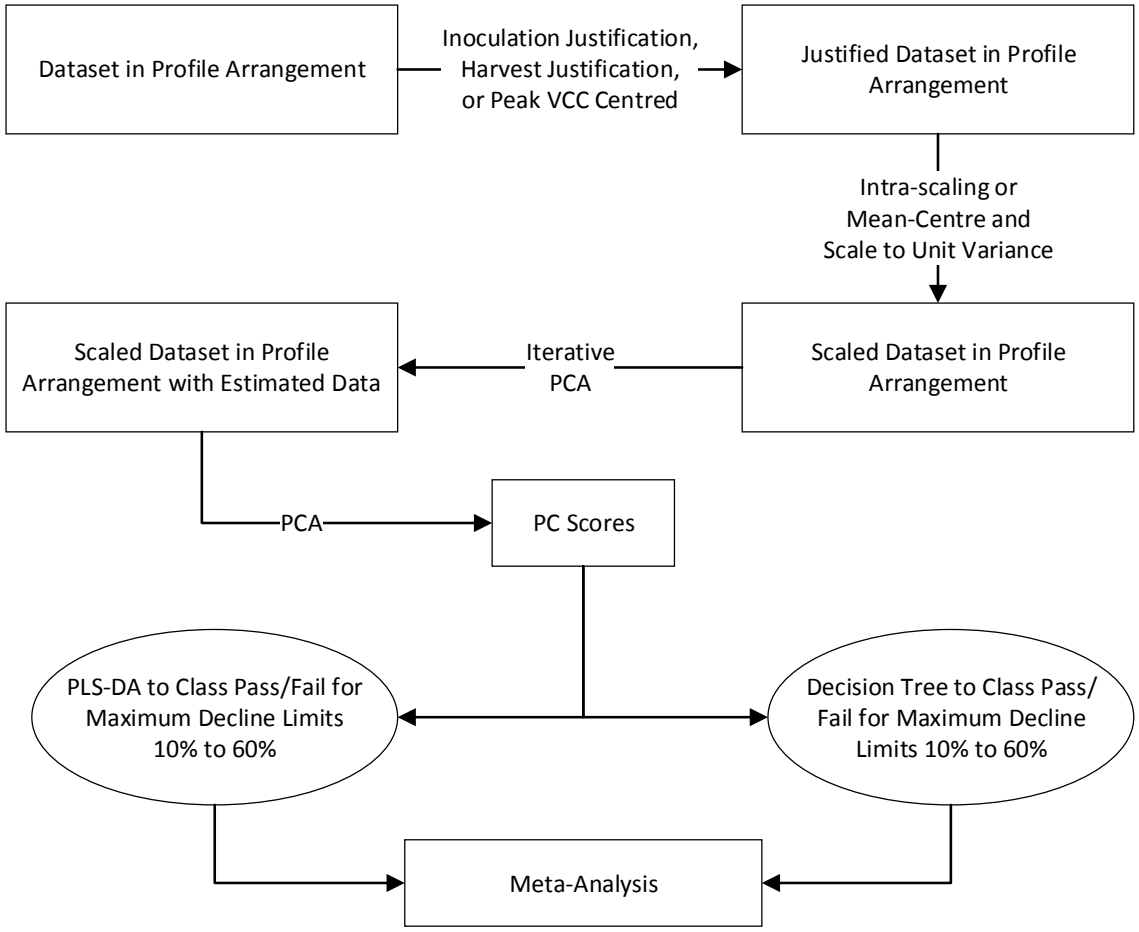


Figure 40. Analysis Pattern 1

6.5.2 Analysis Pattern 2

Analysis Pattern 2 was designed to identify specific variables of interest and followed the sequence shown in Figure 41.

Data were arranged in Profile arrangement (1 row = all samples for 1 culture) and the dataset then justified. The justified dataset was scaled using intrascaling or mean-centred and scaled to unit variance. Missing data were estimated using iterative PCA. The dataset was then subdivided into subsets by variable type (e.g. glucose). For each variable subset, a PCA model was created using random sampling (10 splits, 5 iterations).

Scores for PC1 and PC2 were extracted for each variable model and collated into a single dataset. This dataset was the used to classify cultures as pass or fail using PLS-DA and decision trees (Gini Index) for each maximum decline limit 10% to 60%.

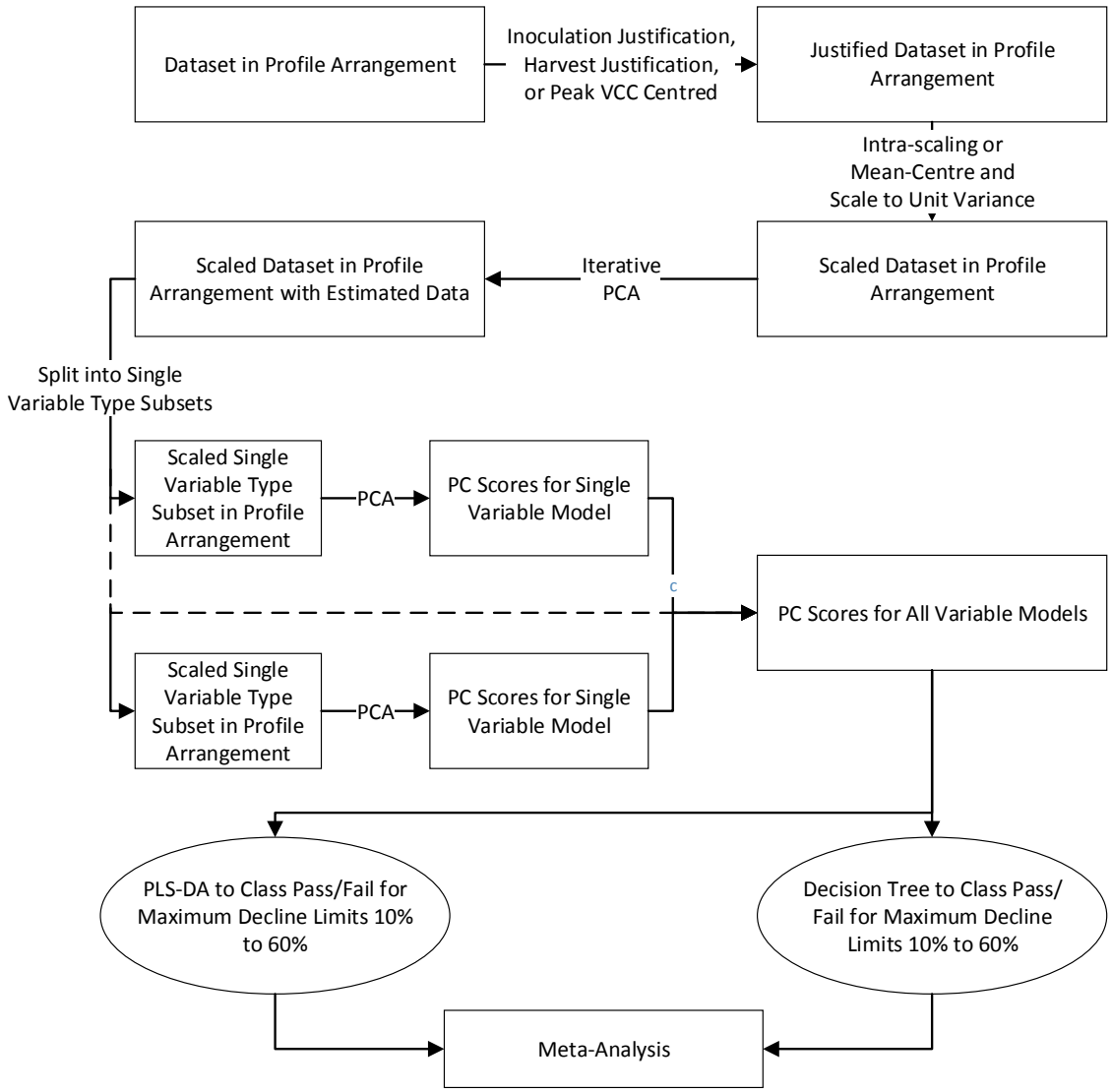


Figure 41. Analysis Pattern 2 - This analysis pattern was intended to identify specific variables of interest by first summarising data by variable type.

6.5.3 Analysis Pattern 3

Analysis Pattern 3 was designed to identify specific days or observations of interest and followed the sequence shown in Figure 42. Data were arranged in the Day by Day arrangement (1 row = 1 sample for 1 culture). The dataset was scaled using two-step intrascaling or mean-centred and scaled to unit variance. Missing data were estimated using iterative PCA. A PCA model was created using random sampling (10 splits, 5 iterations). PC1 and PC2 scores were extracted and arranged into Profile arrangement, i.e. cultures were now described as Day 1 PC1, Day 1 PC2, Day 2 PC1, etc.

The Profile scores dataset was then used to classify cultures as pass or fail using PLS-DA and decision trees (Gini Index) for each maximum decline limit 10% to 60%.

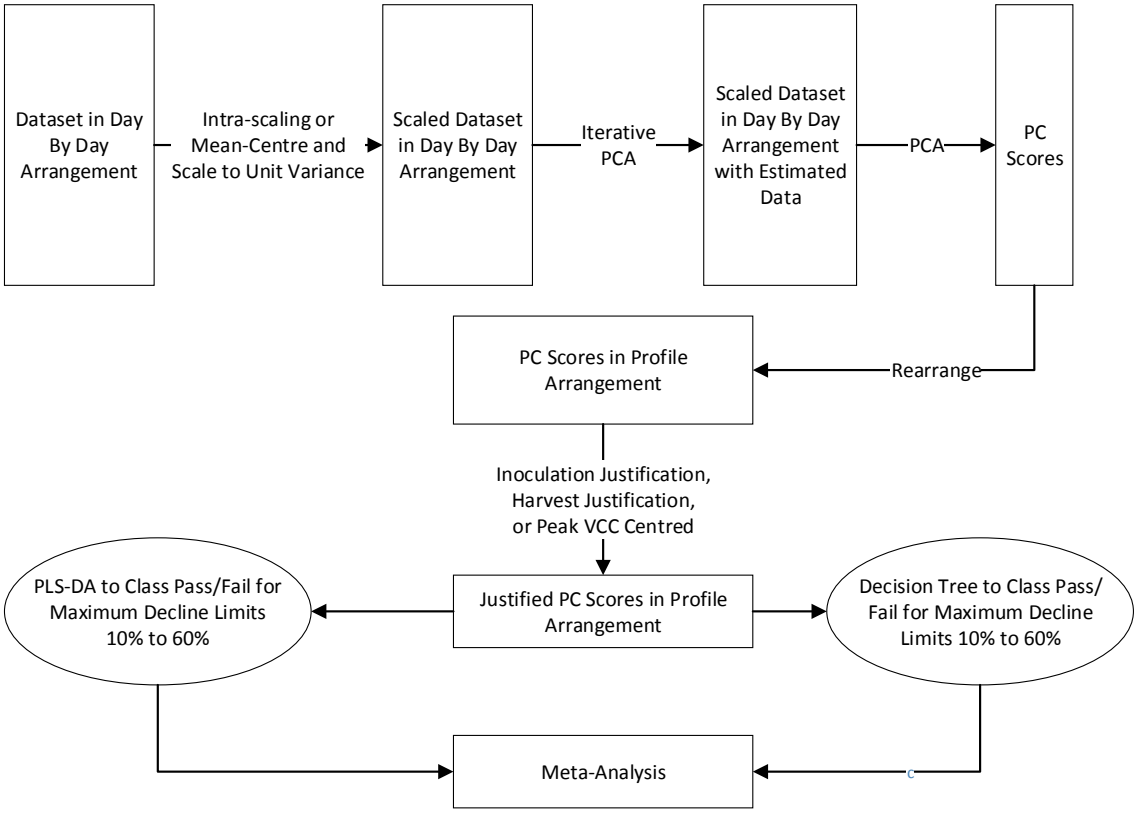


Figure 42. Analysis Pattern 3 – This analysis pattern was intended to identify specific days and observations of interest by first summarising data in ~24 blocks of information.

6.6 Results and Discussion

The developed method was assessed against an historical investigation of the same dataset conducted several years prior in keeping with soft aims related to ease of implementation and interpretation of results. This previous investigation had primarily relied on univariate or qualitative analysis and was human-resource intensive, requiring hundreds of man hours and department-wide involvement.

The method and results presented in this chapter were completed in approximately 200 man hours by a single person. Personal records allot ~50% of this time to data collation and cleaning (100 hours), ~30% to development of tools for dataset justification and intrascaling (60 hours), and ~20% to model creation and result interpretation (40 hours). The results of the meta-analysis were then presented to scientists involved in the previous investigation for comparison to previous results. Hence the final developed framework and supporting tools offer significant time savings in future large scale investigations.

For the top level meta-analysis, comparisons between models were restricted to the criteria noted in Table 28. The following areas were then addressed in a general manner:

- Decline limit choice
- Scaling option
- Statistical method used
- Analysis pattern used

Based on this top level evaluation was made, a select number of models were chosen for more in-depth analysis with particular regards to the following areas:

- Result interpretability
- Days of interest
- Variables of interest

Statistical Method	**Recorded Results**
Decision Tree	Top node decision criteria Tree size (number of nodes) Misclassification (%)
PLS-DA	X variance captured (% Y variance captured (%) Misclassification (%)

Table 28. Results recorded for meta-analysis. These values can be found in Table 54 to Table 57 in Appendix B.

A main effects plot was generated to identify general trends associated with misclassification values (Figure 43). These general trends can be summarised as:

1. Lower misclassification of cultures was more strongly associated with the use of decision trees for classification than with the use of PLS-DA.
2. Models tended to have lower misclassification error rates when decline limits of 30%/d or 40%/d were used.
3. Models were relatively insensitive to analysis pattern choice.
4. Models were relatively insensitive to dataset justification choice.

Regarding scaling, further analysis showed effects from choice in scaling were dependent on whether PLS-DA or decision trees were used. It was seen that for PLS-DA-based models there was a general split in misclassification based on scaling (Figure 43). PLS-DA models built from intrascaled datasets had higher rates of misclassification than models built from datasets to which Autoscaling (mean-centred and scaled to unit variance) had been applied. However, this clear split based on scaling was not observed for decision tree-based models. As PLS-DA models retain all variables while decision trees retain only decision criteria, decision trees are potentially more robust when the ratio of cultures used for model training versus the number of variables is low.

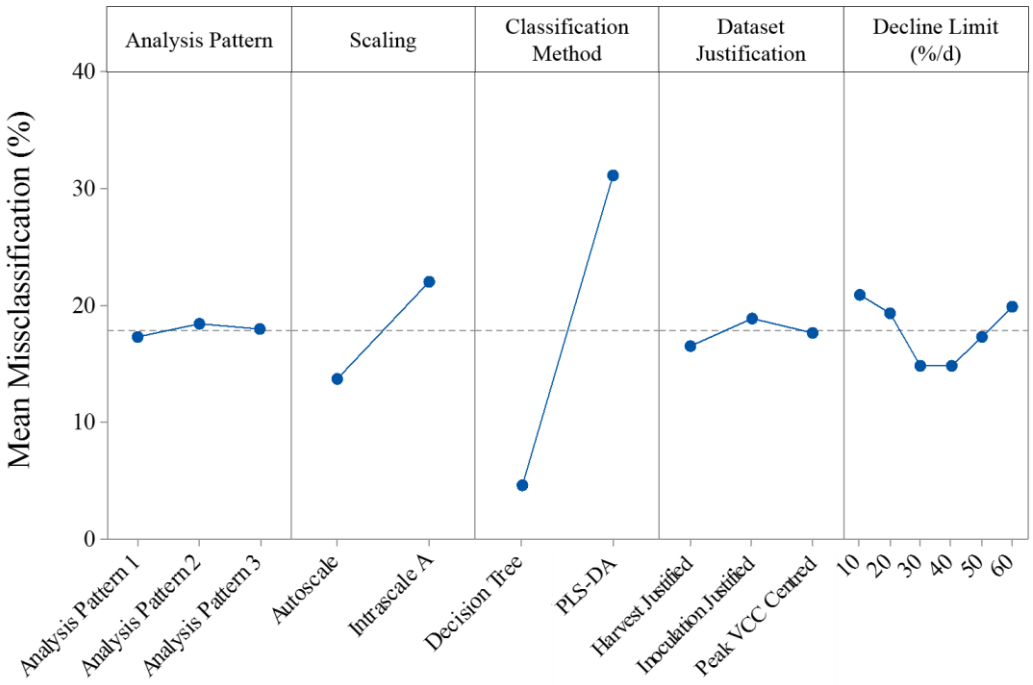


Figure 43. Main effects plot for identifying general outcomes based on misclassification. Regarding dataset justification, harvest justification offered minor improvements in classification accuracy when the maximum decline in viability used for pass/fail classification was greater than 20%/d.

As seen in Figure 35, pass/fail distribution and culture numbers were not equally distributed across projects. Hence the lower rate of misclassification observed when Autoscaling was used may be in part due to identification of projects with higher failure rates rather than identification of fail behaviour. This possibility was further supported by the different effects the scaling methods had on clustering when PCA was applied during initial data exploration (Figure 37).

As the meta-analysis was set out in a similar manner as Design of Experiments, Minitab was used for response surface analysis to identify an optimal model for minimising misclassification. The suggested optimal model was:

- Viability Decline Limit: 45.3535 5/d
- Analysis Pattern: Analysis Pattern 3
- Scaling: Intrascale A
- Statistical Method: Decision Tree
- Justification: Peak VCC Centred

This was interesting as it rejected the general correlation of Autoscaling with lower misclassification errors. From the Minitab optimiser results (Figure 44) was seen that there was in fact little difference in predicted model accuracy whether Intrascaling or Autoscaling was applied to the dataset.

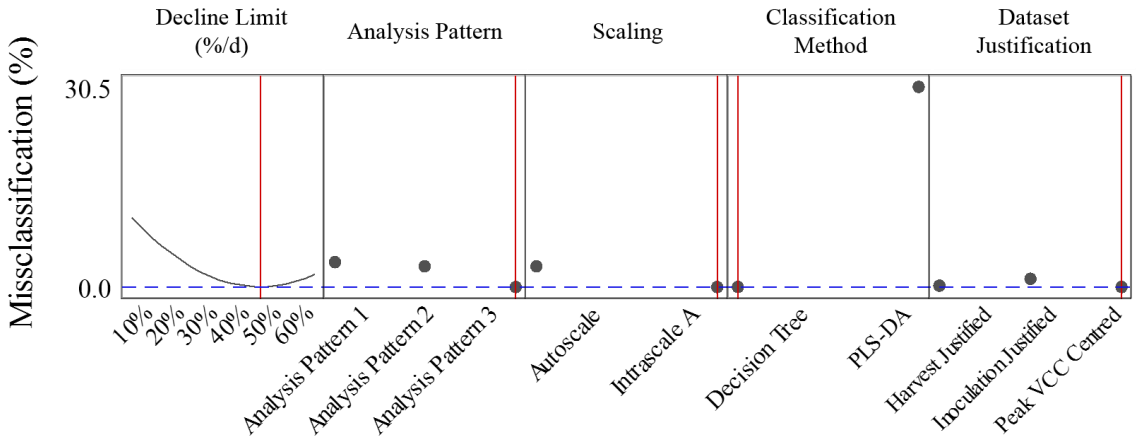


Figure 44. Minitab optimiser results using response surface model. The greater the height of an option, the higher the misclassification error by a model using that option with all other options remaining unchanged. Hence model misclassification is notably insensitive to choice of dataset justification, relatively insensitive to choice of Analysis Pattern or scaling applied, notably sensitive to decline limit, and highly sensitive to classification method used.

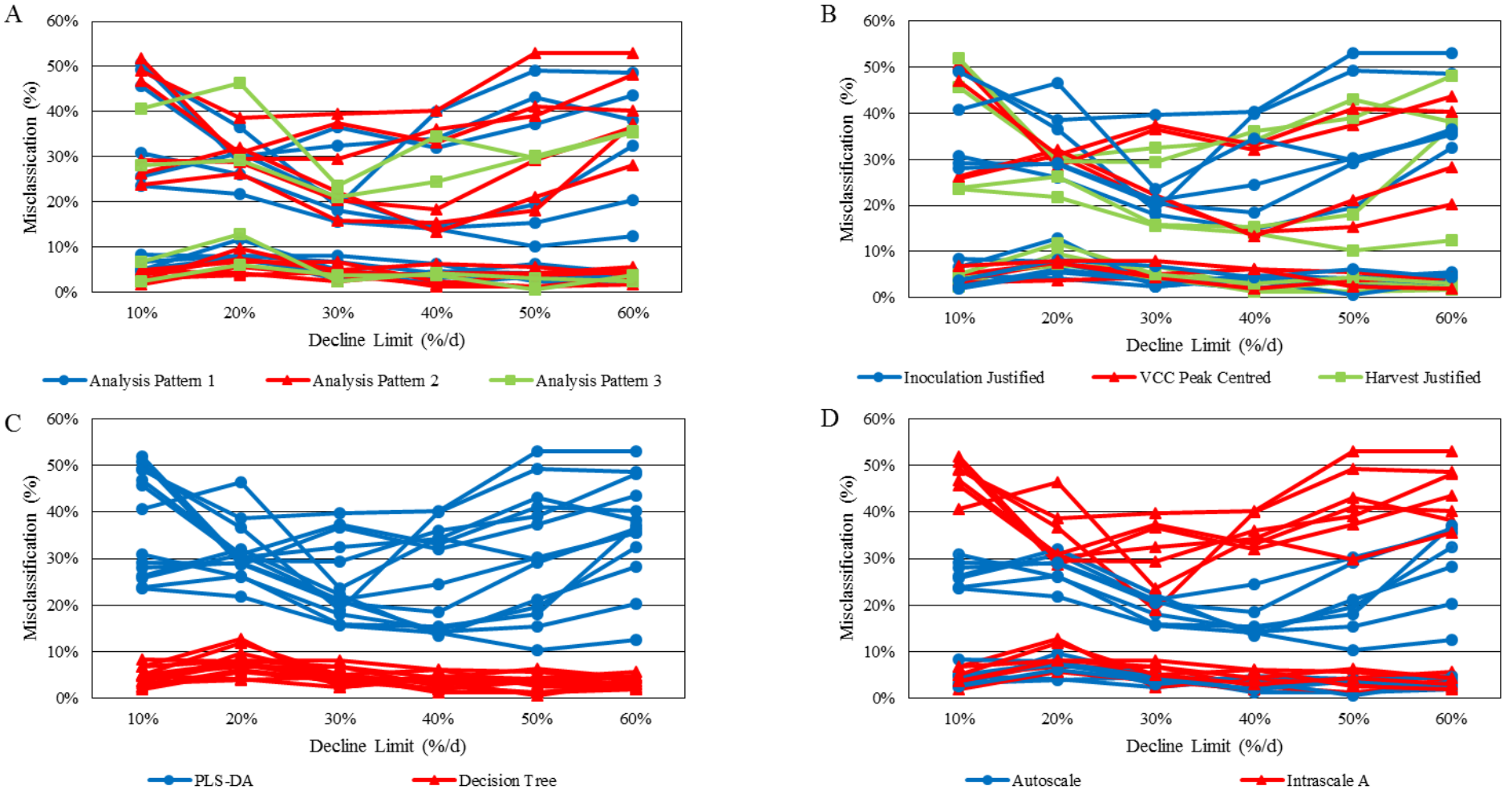


Figure 45. Four comparisons of misclassification of cultures. In A, series are coloured by analysis pattern used. In B, series are coloured by dataset justification. In C, series are coloured by whether PLS-DA or decision trees were used for classification. In D, series are coloured by scaling used.

6.6.1 Analysis Pattern 1 Results and Conclusions

An in-depth analysis was performed using Analysis Pattern 1 on inoculation justified, intrascaled data to classify cultures according to a 30%/d decline limit. Two classification methods were applied this dataset: PC-decision trees and PLS-DA.

When using PLS-DA for classification, the model statistic 'Variable Importance in Projection' (VIP) to be used to identify variables of interest. A variable's VIP scores is an indicator of the variable's importance in a PLS projection. Variables with VIP scores ≥ 1 are considered to be important in the model.

Figure 46 and Figure 47 show VIP scores from an Analysis Pattern 1 PLS-DA model classifying inoculation justified, intrascaled data according to a 30% decline limit. It can be seen that even when variables are grouped by variable type (Figure 46) or reading number (Figure 47), it is difficult to identify specific variables or days of interest due to the large number of variables declared important based on VIP number. While this result gives a very comprehensive overview of differences correlated with pass and fail classification suitable for less- or un-time-constrained analysis, it is data rich but information poor from an operating/manufacturing standpoint.

Similar interpretability issues were encountered when applying similar drill down analysis to models created using PC-decision trees, despite the observed improvement in classification accuracy.

6.6.2 Analysis Pattern 1 Conclusions

Differences in multiple monitored variables were observed between pass and fail cultures in the final samplings for those cultures. According to VIP plots from PLS-DA models, these differences typically manifested from the fourth or fifth sample onwards. However, further analysis would require intensive analysis by a biologist to separate variables of interest (e.g. glucose) from time periods of interest (e.g. activity between Days 3 and 4).

From these results, it was concluded that use of PCA-decision trees could be used to efficiently classify culture behaviour, however data were not in an easily interpretable form for onward analysis.

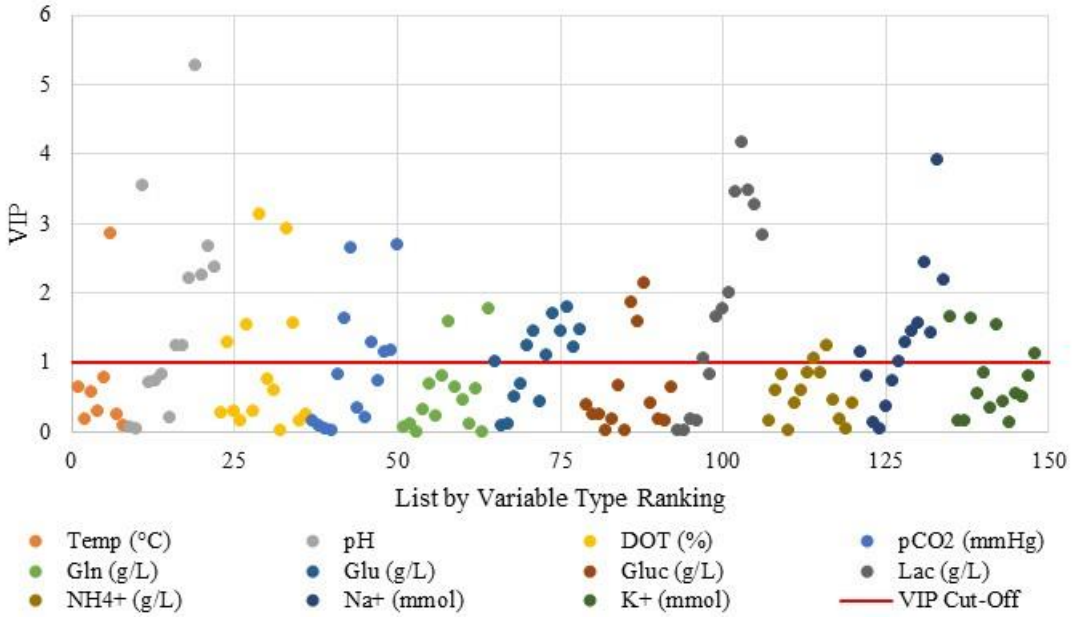


Figure 46. VIP scores from an Analysis Pattern 1 PLS-DA model classifying inoculation justified, intrascaled data according to a 30% decline limit grouped by variable type.

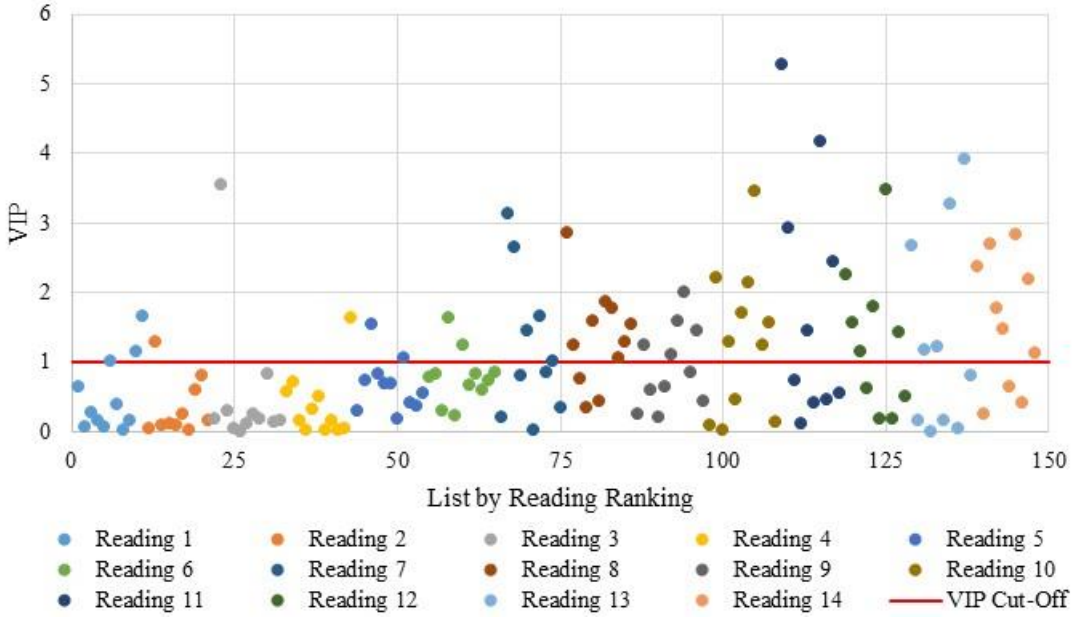


Figure 47. VIP scores from an Analysis Pattern 1 PLS-DA model classifying inoculation justified, intrascaled data according to a 30% decline limit grouped by reading.

6.6.3 Analysis Pattern 2 Results

The decision criteria for the top node of decision trees created using Analysis Pattern 2 and Analysis Pattern 3 were recorded. This information was then used to evaluate the most frequently selected decision criteria. Frequency evaluation was not extended to the lower levels of the decision trees for the results shown, however this a strong area for follow on investigation.

Table 29 and Table 30 display the frequencies of top node decision criteria for decision trees generated using Analysis Pattern 2, which was designed to identify key variables of interest. These frequencies were then broken down according to justification and scaling applied to the dataset during analysis to give pH PC1 as the top variable of interest (top decision criteria for 56% of decision trees) followed by Lactate PC1 (14%) and Glutamine PC2 (11%).

	All		**Inoculation Justified**		**Harvest Justified**		**Peak VCC Centred**	
Top Node	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**
pH PC1	20	56%	7	39%	6	33%	7	39%
pCO2 PC2	1	3%			1	6%		
Gln PC1	2	6%					2	11%
Gln PC2	4	11%	1	6%	3	17%		
Gluc PC1	1	3%	1	6%				
Lac PC1	5	14%	2	11%	1	6%	2	11%
Na PC1	2	6%	1	6%	1	6%		
K PC1	1	3%					1	6%

Table 29. Decision criteria for top node in decision trees using Analysis Pattern 2 with respect to dataset justification. Shading indicates that the variable was not a top node decision criteria for the dataset justification listed.

	All		**Autoscale**		**Intrascale A**	
Top Node	**Count**	**%**	**Count**	**%**	**Count**	**%**
pH PC1	20	56%	10	56%	10	56%
pCO_2 PC2	1	3%	1	6%		
Gln PC1	2	6%	1	6%	1	6%
Gln PC2	4	11%	1	6%	3	17%
Gluc PC1	1	3%	1	6%		
Lac PC1	5	14%	3	17%	2	11%
Na PC1	2	6%			2	11%
K PC1	1	3%	1	6%		

Table 30. Decision criteria for top node in decision trees using Analysis Pattern 2 with respect to scaling method applied. Shading indicates that the variable was not a top node decision criteria for the scaling method listed.

For the process analysed, lactate and glutamine were not controlled variables and were therefore pure indicators/symptoms of culture behaviour. pH was technically a controlled parameter, however due to the wide deadband used by the process, pH was also a possible indicator of natural cell behaviour within the deadband.

When considering nodes below the top node, in particular for decision trees beginning with pH PC1 as the top node decision criteria, decision criteria on the majority pass path were predominantly metabolite-based (e.g. K^+). Decision criteria on the majority fail path were predominantly control-based (e.g. pH or temperature) or closely tied to control strategies (e.g. pCO_2).

A detailed analysis was performed using the decision tree created from intrascaled, inoculation justified data using a pass/fail maximum viability decline limit of 40%/d (Figure 48). Culture progressions through the decision tree were determined and the PC scores from which decision criteria were selected were plotted. The original data from which the intermediary model were created were then located for the cultures at the decision node and classed as pass or fail, as determined by the decision rules. These data were then summarised as mean, mean + 2 standard deviations, and mean – 2 standard deviations for both pass and fail subset for each sampling point (n=1 to n=12). These calculated values were plotted against sampling point to visualise the general trends in the data. Figures presented in this chapter are limited to those immediately relevant to the presented results. The full set of figures can be found in Appendix B.

From Figure 48 it was seen that two primary failure pathways existed (Table 31). It should be noted both pathways were primarily described by PC1 values (i.e. main behaviour for the variable in question) and a limited number of variables.

Pathway 1		**Pathway 2**	
Node	Decision Criteria	Node	Decision Criteria
1	pH PC1 <= -0.408	1	pH PC1 <= -0.408
2	pCO_2 PC1 > -2.301	2	pCO_2 PC1 > -2.301
5	Na PC1 > 3.052	5	Na PC1 <= 3.052
		9	pH PC2 > -0.755
		13	Temperature PC1 > 0.178
		19	Temperature PC1 <= 0.695

Table 31. Main failure pathways for decision tree classifying intrascaled, inoculation justified dataset to a 40%/d viability decline limit.

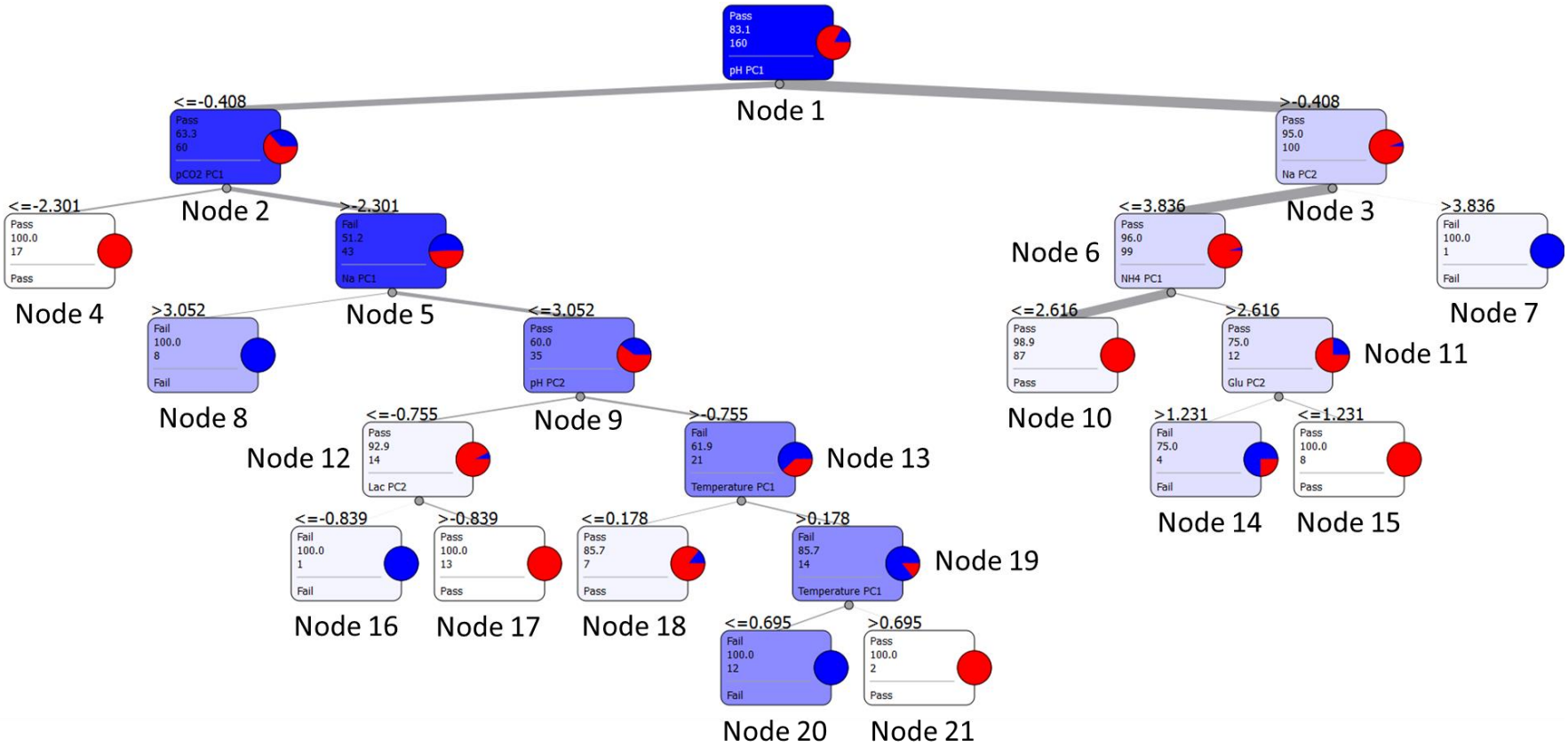


Figure 48. Decision tree classifying inoculation justified, intrascaled data according to a 40%/d viability decline limit. Each node displays the majority class (pass/fail) for samples at the node, the percentage of samples with that class, and the total number of samples on the node on the left-side of the coloured box. Pass/fail distribution at the node is also indicated by a pie chart on the right-side of the coloured box (red = pass, blue = fail). If the node is a decision node, then the decision criteria variable is listed in the bottom of the box. If the node is a leaf node, then the final class is listed at the bottom of the box. The values used for the decision criteria are displayed above the subsequent child node. Nodes are coloured by the percentage of fail samples at the node, i.e. the top node is coloured darkest blue as it holds 100% of fail cultures whereas leaf node with only pass cultures is completely white. The thickness of lines between nodes indicates the number of samples following that path.

The decision criteria at Node 1 was pH PC1. Plotting pass/fail subsets as shown in Figure 49, it was seen that two basic pH behaviours existed. In the 'pass' behaviour, the mean pH declines from ~7.00 to ~6.85 sometime between the fourth and fifth readings (i.e. between Day 3 and Day 4). The mean pH then gradually returns to the previous mean of ~7.00. In the 'fail' behaviour, mean H behaviour is similar to pass behaviour until the fourth reading. The mean 'fail' pH also declines to 6.85, however the recovery to a mean of ~7.00 was not observed. Both pH behaviours were within the permitted operating conditions, hence the differences in behaviour have originated from the cell culture behaviour, interactions between cell culture behaviour and the feed strategy, or a combination of the two.

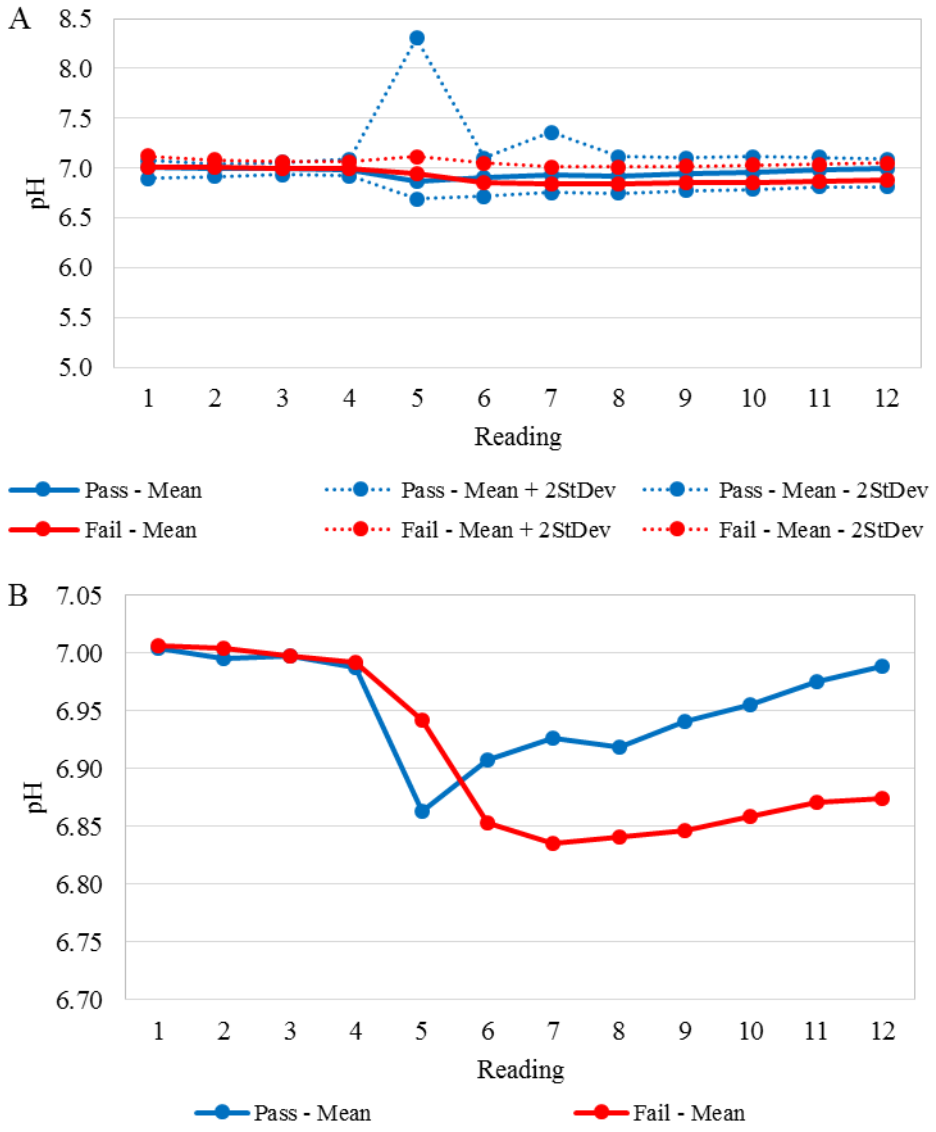


Figure 49. Pass/fail pH behaviour for cultures at Node 1. Note that pass/fail refers to the class applied by the decision tree at this node and not the final classification according to the decision tree rules or the class determined using the simple profile classifier. In B, the ±2σ have been removed to improve identification of mean trends.

The decision criteria at Node 2 was pCO_2 PC1. Plotting pass/fail subsets as shown in Figure 50, there appeared to be divergences in pCO_2 behaviour beginning at Reading5 (Day 4). At this point onwards, fail cultures had on average slightly higher pCO_2 readings. More notably, it was observed that there was greater variance in each pCO_2 reading for fail cultures than for pass cultures. These observations were interpreted as indicating pCO_2 level stability as an indicator of pass/fail behaviour and possible contributor to undesired behaviours.

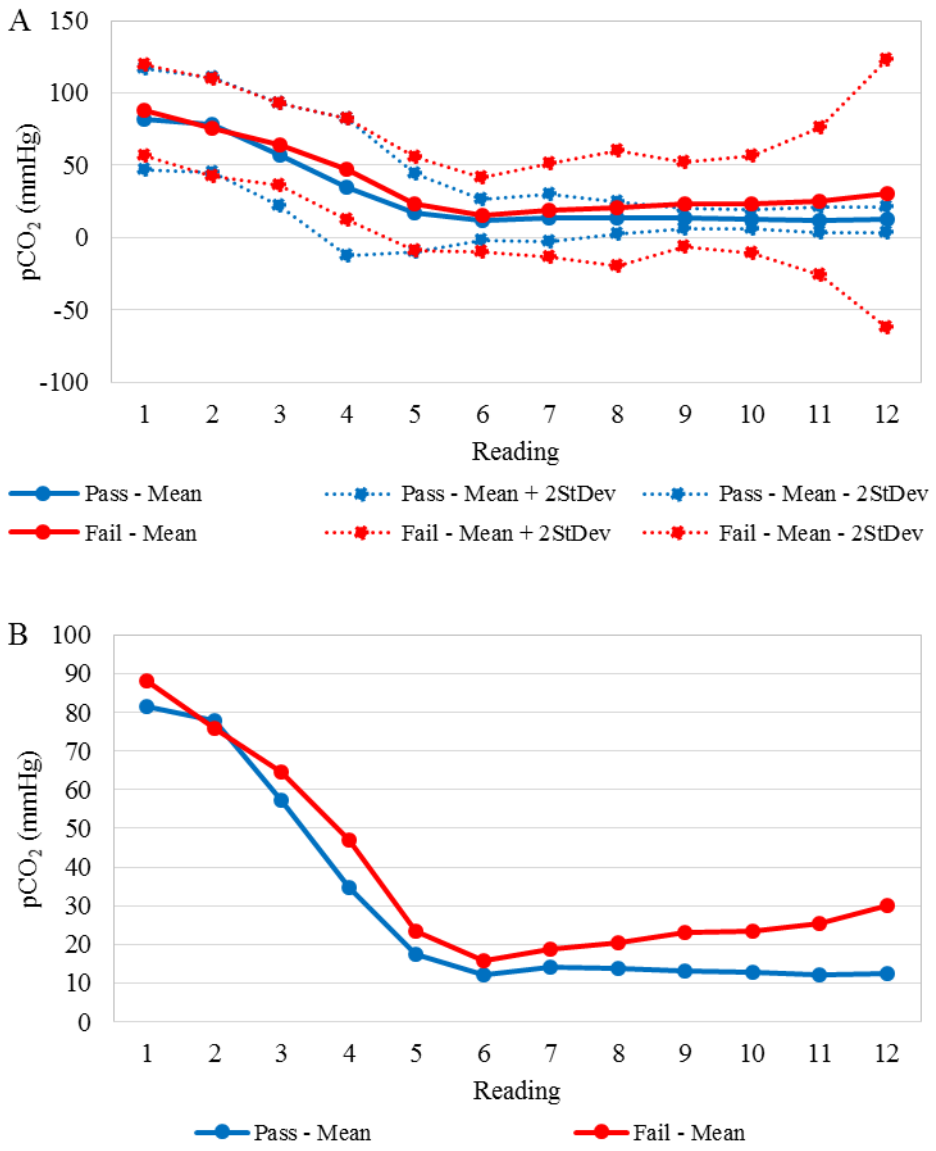


Figure 50. Pass/fail pCO_2 behaviour for cultures at Node 2. Note that pass/fail refers to the class applied by the decision tree at this node and not the final classification according to the decision tree rules or the class determined using the simple profile classifier. Also note the negative values plotted were due to the calculated standard deviation. In B, the ±2σ have been removed to improve identification of mean trends.

The decision criteria at Node 5 was Na PC1 Plotting pass/fail subsets as shown in Figure 51, no immediately obvious differences in Na^+ profiles could be seen. However it was observed that cultures classed as fail at this node typically had higher recorded values for Na^+ than cultures classed as pass at this node. This general divergence was observed as beginning after Reading 4 (Day 3).

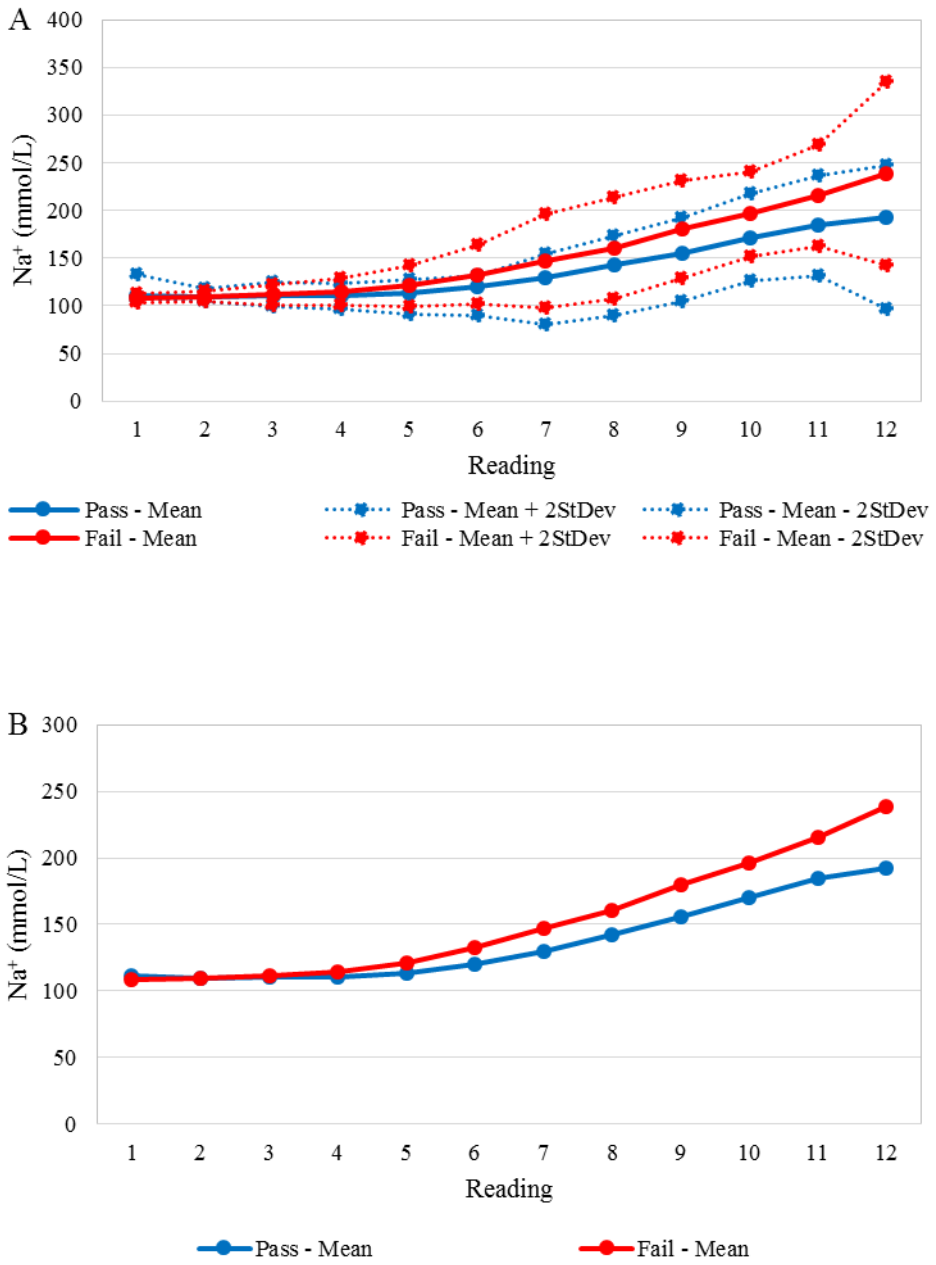


Figure 51. Pass/fail Na^+ behaviour for cultures at Node 5. Note that pass/fail refers to the class applied by the decision tree at this node and not the final classification according to the decision tree rules or the class determined using the simple profile classifier. In B, the ±2σ have been removed to improve identification of mean trends.

The decision criteria at Node 9 was pH PC2. Plotting pass/fail subsets as shown in Figure 52, it was observed that the primary difference between mean pass behaviour and mean fail behaviour was the rate of pH decline. Results at this node were considered a refinement of the behaviours generalised at Node 1.

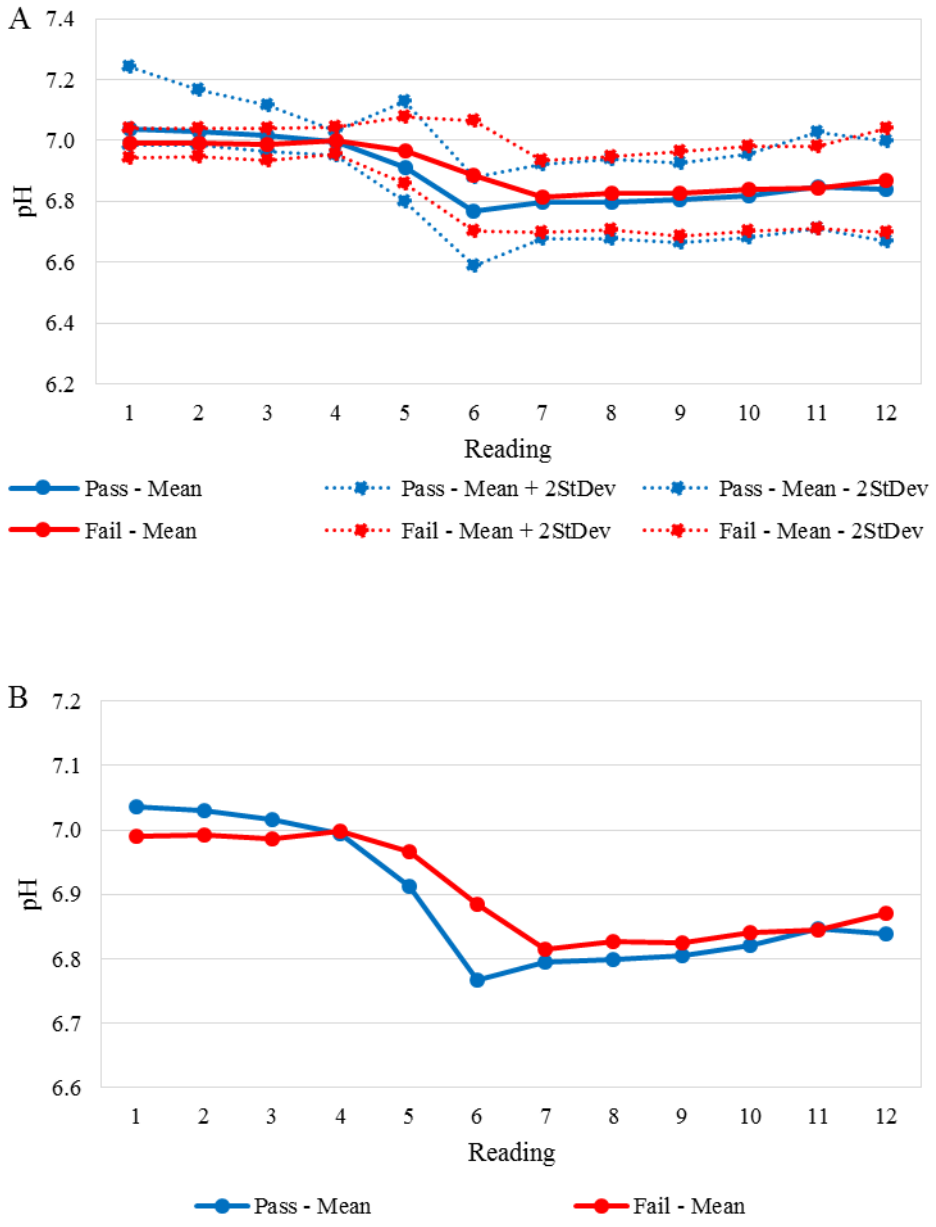


Figure 52. Pass/fail pH behaviour for cultures at Node 9. Note that pass/fail refers to the class applied by the decision tree at this node and not the final classification according to the decision tree rules or the class determined using the simple profile classifier. In B, the ±2σ have been removed to improve identification of mean trends.

The decision criteria at Node 12 was Temperature PC2. No appreciable practical differences were observed when plotting the original data for the cultures (Figure 53) with one exception (Pro_016_005). It was suggested that the dataset had become so reduced in terms of variance that spurious decision criteria were beginning to be selected. This suggestion was supported by the fact that temperature data recorded in daily monitoring had very limited variance (~ 0). This can be seen in Figure 53 where nearly all recordings for all cultures were 36.5 ℃. As discussed in Chapter 5 and Chapter 6, "too perfect" datasets heavily exaggerate the slightest differences when scaled. Hence, the selection of temperature as a decision criteria was taken as an indicator that all useful information had been extracted from the decision tree and that further interrogation could be halted.

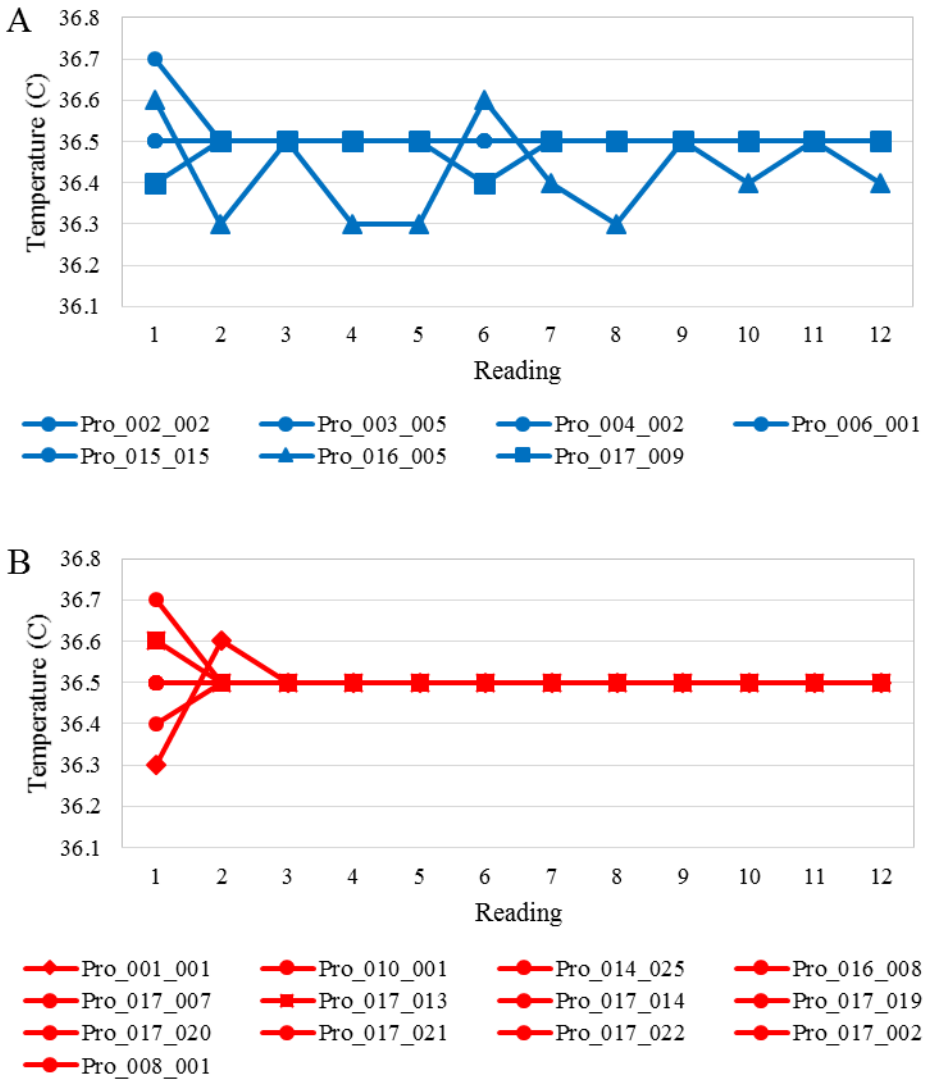


Figure 53. Original temperature measurements for cultures at decision node 12. In A, cultures were classed as "pass". In B, cultures were class as "fail". In both A and B, the majority of temperature readings were 36.5 ℃.

6.6.4 Analysis Pattern 2 Conclusions

pH behaviour was top node decision criteria in the majority of decision trees generated. Decision trees could then be generalised as a "pass" path described by metabolite concentrations and a "fail" path described by control-related measurements.

The strongest generalised pass/fail behaviours were observed for pH. Key differences between pass behaviour and fail behaviour were the magnitude of the decline in pH readings typically observed by Days 4 and 5 and whether the pH returned to pre-decline values or whether pH readings remained lowered.

Detailed analysis of the decision tree classifying the intrascaled, inoculation justified dataset according to a pass/fail limit of a 40%/d calculated decline in viability revealed several strongly interrelated variables as being of interest: pH, pCO_2, Na^+ concentration.

Two main potential causes of viability crashes were identified as areas for improvement:

- pH strategy – Adjustment of pH controller deadband to force pH readings nearer to the observed "pass" culture pH behaviour.
- Gassing strategies related to pCO_2 control.

6.6.5 Analysis Pattern 3 Results

Table 32 displays the frequency of top node decision criteria for decision trees generated using Analysis Pattern 3, which was designed to identify key days of interest. As Analysis Pattern 3 did not use justification, comparison between top node decision criteria was limited to dataset scaling process applied where the most frequent variables selected for top node decision criteria were Day 1 PC1 and Day 2 PC1.

From this trend, it appeared that behaviour very early in a culture's residence was correlated with the final pass/fail class and that behaviour on or just prior to these readings could be either related to the cause or a strong indicator of the cause of failure.

To determine the state of the cultures on Day 1 and Day 2 relative to the culture behaviour across all readings, contribution analysis was performed on the PCA models used to generate the scores summarising daily samples. The types of figures used to conduct this analysis are shown in a series of figures in Figure 54.

Figure 54A was the score plot of the PCA model created using intrascaled data in the day by day arrangement, where each point represents the data from one daily sample from a single culture. When a decision tree was generated to classify samples according to a 30%/d decline limit, the Day 2 PC1 score was selected as the top node decision criteria. Hence, scores in Figure 54A were limited to those related to Day 2 samplings.

Figure 54B shows the loading plot of the PCA model from which it was seen that the relative positioning of metabolites appeared to be in keeping with known GS-CHO metabolism behaviour during a typical 12-15 culture. While there was comparatively low variation in DOT (%) or temperature, as was expected for cultures running according a set process, pH and pCO_2 showed higher degree of variation.

	All		**Autoscale**		**Intrascale A**	
Top Node	**Count**	**%**	**Count**	**%**	**Count**	**%**
Day 1 PC1	4	33%	4	67%		
Day 2 PC1	4	33%	1	17%	3	50%
Day 6 PC1	1	8%	1	17%		
Day 7 PC1	1	8%			1	17%
Day 9 PC1	1	8%			1	17%
Day 10 PC1	1	8%			1	17%

Table 32. Decision criteria for top node in decision trees using Analysis Pattern 3. Shading indicates that the variable was not a top node decision criteria for the scaling method listed.

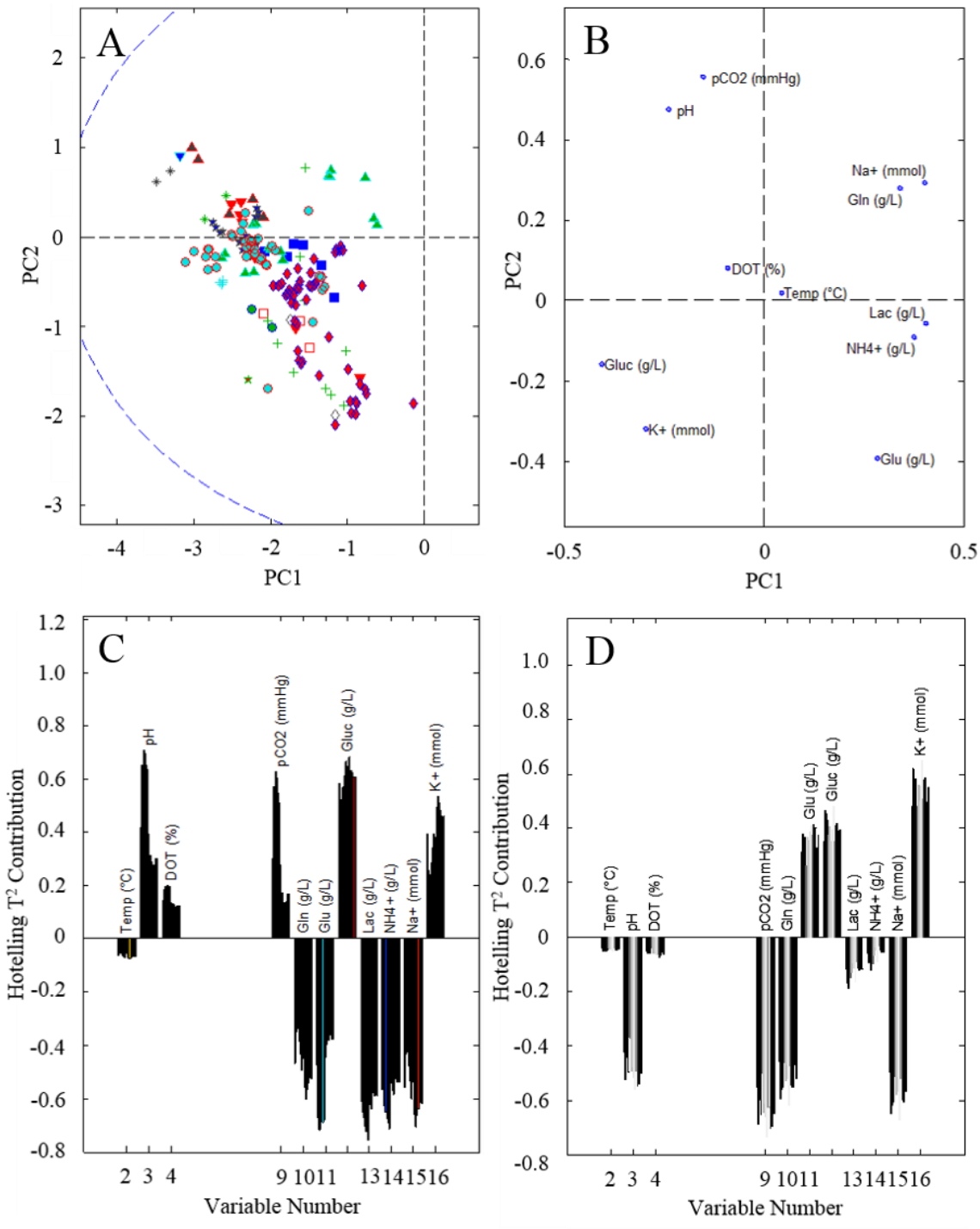


Figure 54. Figures used for contribution analysis of Day 2 readings to compare pass and fail cultures according to a 30%/d decline limit. The following graphs were generated using Analysis Pattern 3 with intrascaled data:

A. Score plot showing scores for Day 2 only. Point colour and shape indicated project.
B. Loadings for model.
C. Hotelling T^2 contributions for samples classed as fail. Gaps in variable number were due to variables excluded from analysis (e.g. osmolality, VCC).
D. Hotelling T^2 contributions for samples classed as pass. Gaps in variable number were due to variables excluded from analysis (e.g. osmolality, VCC).

Figure 54C and Figure 54D show Hotelling's T^2 contributions for two sets of Day 2 sample scores. Hotelling's T^2 contributions show the relative contribution by a value for a variable to a sample point's position within a multidimensional model. For the dataset under consideration, each point shows the relative contribution compared not only to other cultures for that daily sample but also the point compared to sampling pints throughout the cultures' span.

In Figure 54C, the contributions describe Day 2 samples scores for cultures classed as fail by the top node decision criteria (N.B. this is not necessarily the final class assigned by the decision tree). In Figure 54D, the contributions describe Day 2 samples scores for cultures classed as pass by the top node decision criteria. When these figures were compared, it was seen that contributions from pH, DOT, pCO_2, and glutamate were notably different, effectively opposite values in terms of negative/positive. A difference in the magnitudes of contributions by lactate and NH_4+ was also noted.

As stated previously, these contributions were relative to all readings for all cultures in the dataset. Hence differences in Hotelling's T^2 contribution were also affected by how measurements for Day 2 compared to culture behaviour before and after Day 2. Essentially, scaled readings would capture if pH readings were relatively constant throughout the culture span, if decline and recovery (as seen in Figure 49) occurred, etc. These comparisons were further emphasised by the use of intrascaling.

6.6.6 Analysis Pattern 3 Conclusions

pH, DOT, and pCO_2 were key indicators related to control strategies. Additionally cultures classed as fail by the top node showed lower concentrations of lactate, glutamate, and NH_4^+ than cultures classed as pass by the top node. These results were in keeping with conclusions from Analysis Pattern 2.

6.7 Final Results and Discussion

The lack of ease of interpretation for Analysis Pattern 1 stands in contrast to the ease of interpretation when separate models were used to identify and interpret variables and days of interest (Analysis Pattern 2 and Analysis Pattern 3 respectively). Here it was shown that, for the dataset in question, the development of several "simple" models focussed on only two dimensions of a three-dimensional dataset (culture x sample x time) was of greater use than the development of a single, fully comprehensive model using a more complex statistical method (e.g. PLS-DA).

The conclusions from the above analyses were presented to key figures in the original investigation by Lonza. The conclusions from the presented investigation were in keeping with conclusions from the historic investigation used to develop an improved platform process, primarily that adjustment of the pH control strategy to enforce behaviours similar to that more typical of 'pass' cultures. After a key turning point in pH behaviours, analysis became complicated by complex interactions between indicators and feed strategies. Therefore process development activities would need to take a minimum two-stage approach to first address pH behaviours and to afterwards address feed strategies.

While the presented investigation did not reveal any significantly new information compared to the historic investigation, it was shown that appropriate use of multivariate data analysis could allow similar conclusions to drawn with both reduced time and personnel requirements.

6.8 Conclusions

It was demonstrated, that with appropriate modifications to the method developed in Chapter 6, a meta-analysis approach of developing many models from a core dataset resulted in stronger identification and understanding of captured behaviours than would have been achieved by relying on a single model. In doing so, it was also demonstrated that the major perceived obstacle to conducting a meta-analysis was time required was untrue.

By developing technically simple spreadsheets, a number of analytical options for restructuring and pre-processing the core dataset were tested for a low cost when compared to the time need to develop a single model in terms of additional man hours. While harvest justification showed only minor improvements in classification accuracy for the presented investigation, it was shown that rejustification of the dataset to better reflect the response in question could be easily applied and tested.

Finally, it was demonstrated that the use of intrascaling reduced project-specific confounding in a manner appropriate for the dataset used and the investigation conducted. Specifically, in this investigation ensuring platform process behaviours were captured in place of project-specific behaviours was prioritised over pure classification accuracy. The two-step scaling process can be recommended as a tool for the following purposes:

1) Reduction of low levels confounding as demonstrated in the presented investigation.

2) Identification of high levels of confounding, i.e. confounding remains after intrascaling applied as the scaling step indicating a more powerful method is required, e.g. confounding reduction through the use of PLS-DA with subsequent analysis performed using PLS-DA residuals matrix [13].

3) Identification of whether confounding has any appreciable impact on analysis, i.e. confounding effects on data not appreciably related to response of interest.

6.8.1 Recommendation 1

Aspects of the developed meta-analysis framework can applied in a variety of ways for process development. For example, a "generic" model is developed for a process platform (e.g. CHO V8.0) from a multiproduct dataset. New projects using the host cell line and platform process are analysed using this model for a various reasons:

1. To provide an initial model for performance monitoring until sufficient data is available to create a project-specific model.
2. To identify past projects/products with similar behaviour and identify potential sensitivities or issues.

6.8.2 Recommendation 2

As the development time for models was shown to be minimal compared to time needed to collate and vet data, a meta-analysis/multiple model confirmatory approach is recommended as a normal action during investigations, at least during initial data exploration and model development.

6.8.3 Recommendation 3

In light of the growing adoption of electronic laboratory notebooks and recommendations 1 and 2, it may be of benefit to develop models to evaluate data at the point of capture. A 'dashboard' of models (generic process and/or product-specific) using the data could be displayed. With appropriate development, these models could be used to tracking culture progress or monitor known indicators of undesirable future behaviours. Furthermore, as data would be directly entered and potentially evaluated one sample at a time, this could reduce the time needed to vet data during a large scale analysis as common entry errors (e.g. decimal point errors) would be more likely to be spotted near the point of entry

Chapter 7. Conclusions

The research presented demonstrated multiple applications of multivariate data analysis (MVDA) to address four main projects. The conclusions and recommendations presented here prioritise the methods used, statistical tools generated. Additional space was also given to how the works performed met soft research aims, e.g. demonstrating how to link multiple statistical techniques in a cradle to grave (data generation to implementable result) workflow, considerations for adapting workflows based on available data and investigation aims, etc. Key project-specific results of analysis are reiterated, however for greater detail of project-specific results, readers are advised to consults the relevant chapter.

7.1 Comparison of pH Measurement Technologies and Extraction of Indirectly Captured Information

During the comparison of pH measurement technologies (Chapter 4), it was shown that it was possible to extract indirectly captured information from a pre-existing dataset. More specifically, the indirectly captured information were contributions to osmolality measurements by unidentified compounds, here termed *osmolality residuals*. It was recommended that osmolality be recorded at all scales to allow the extraction of this information and thus improve comparisons between scales.

Osmolality residuals were calculated through the creation of a multiple linear regression (MLR) model that was refined using statistical significance testing. Use of this newly created variable was backed by observations of data behaviour through use of time series analysis and principal component analysis (PCA). Use of MLR and statistical significance testing was then used to demonstrate a different purpose: the identification of variables correlated with differences in pH readings by different pH measurement technologies. Hence, the application of three MVDA methods was demonstrated: MLR, statistical significance testing, and PCA. This met soft aims by demonstrating flexibility of MVDA methods in both dataset preparation and final analysis.

The hard aim of the pH measurement technology comparison was the identification of variables correlated with differences in pH readings between the offline pH measurement technologies. Results indicated that samples' chemical compositions and physical condition (e.g. temperature) were correlated with differences in pH readings by different pH measurement technologies.

From this conclusion, it was recommended the company ensure consistency in pH measurement technology type used for activities across sites and between scales, e.g. if NOVA Bioprofile 400 units are used for offline pH measurement during R&D at 10L at one site, use NOVA Bioprofile 400 units for offline pH measurement at different scales and/or different sites.

Similarly, the company should ensure consistency in pH measurement technology type used across activities, e.g. if possible, ensure online and offline pH measurement units use similar technologies, e.g. both online and offline pH measurements are made using Radiometer pH probes with temperature compensation methods.

7.2 Productivity Investigation

An investigation into variation in product concentration for a single product project was described in Chapter 5. The investigation was broken into three distinct stages.

In the first stage, the dataset comprised 10L and 130L cultures performed at the Slough, UK site. From this dataset, a general method based on decision tree was created. During method creation, a variety of available options for data sources, missing data estimation, data pre-processing, and decision tree algorithms were tested.

In addition to identifying the "best" options to employ in terms of lowest misclassification error, it was shown that a choice in one area could cause knock on effects in other areas and in turn effect the accuracy of the overall method. The specific example was the use of iterative PCA for estimation of missing data when the dataset was restricted to data from daily monitoring of cultures or when the dataset was extended to include data from both daily monitoring and online monitoring of cultures. It was suggested that this was due to both greater availability of information to the model and improved estimation of missing data when using iterative PCA to estimate missing data.

In the second stage, the dataset was expanded to include cultures from 5000L and 10L cultures from a US site, which used an altered seeding criteria and altered harvest criteria. As these alterations prevented the use of the initially developed method of decision trees, it was suggested partial-least squares regression models to predict product concentration could be used to identify behaviours correlated with productivity instead.

The alterations in seeding criteria introduced a source of confounding; the US-sited cultures appeared to be more "mature" with respect to Elapsed Time (h) due to higher cell concentrations for transferred inoculum and associated metabolite differences. It was

suggested that this confounding could be reduced by realigning data from time-based sampling points (e.g. every approximately every 24 hours) to sampling-based on set values for the integral of viable cell concentration (IVC). To test this hypothesis, new datasets were created based on rigid sampling points for time and IVC by simple interpolation recorded data.

It was found that use of IVC in place of Elapsed Time (h) lead to improved predictive accuracy and better distribution of residuals for PLSR models predicting daily product concentration from that day's associated data. Use of IVC in place of Elapsed Time (h) also provided greater robustness when data generated from US cultures with altered seeding criteria were applied to models generated from UK culture data. However these improvements did not necessarily occur when analysing the dataset in profile orientation (e.g. 1 sample = all data for 1 culture) or when using a different response of interest (e.g. viability). These findings demonstrate additional considerations that must be taken into account each time an interrogation is made of a dataset, in particular if multiple interrogations are being made of a single dataset.

It was also shown that interpolating the dataset to realign sample to set progression values (e.g. 0.0 h, 24.0 h, 48.0 h, etc.) allowed the progression variable itself to be eliminated from the model with minimal negative effects on model accuracy and robustness. This was in keeping with expectations as there would be no variance between samples for these variables following realignment and hence no additional information to be captured by including the progression variable. This also demonstrated that the underlying temporal/progressional variation in the dataset had been effectively removed.

In the third and final stage, it was suggested that the underlying cause of variation in product concentration may have been linked to the batches of media used in the course of the study. In collating batch numbers for all UK cultures, it was shown that combinations of batch numbers for media components were effectively unique to each round of cultures performed. Hence, it was not possible to confidently tie any individual media component or specific lot of media component to undesirable behaviour. Even if specific components or lots could have been pinpointed as the underlying causes, it would not have been possible to identify the differences in composition leading to undesirable behaviours as lot compositions were not recorded, only lot number.

The hard aim of the productivity investigation was the identification of variables and factors correlated with variation in product concentration at harvest. Soft aims also existed

related to guidance on handling investigations and datasets of this nature, as well identification of areas requiring or benefiting from further attention. Additionally, a variety of tools were generated in the course of the work.

In Stage 1, several different means for handling missing data were employed. In addition to providing reasons for or against their use from a theoretical viewpoint, the presented work also demonstrated the real effects of using the methods for handling missing data. While the results were as expected, demonstrating knock-on effects during analysis, such as filling in missing values with variable means artificially reducing data spread and in turn leading to decreased classification accuracy by PC-decision trees, was within the soft aims of the EngD.

Stage 1 of the productivity investigation provided the drivers for the development of informative values for capturing behaviours in online monitoring and subsequently the development of the Excel-based tool EPIC-CAT. The work presented in Stage 1 demonstrated that the initial Informative Values 1.0 met two of the stated criteria:

1) Capturing behaviours in online monitoring data in a manner suitable for follow on use in MVDA
2) Allowing integration with data from daily sampling.

However Informative Values 1.0 (see Appendix A) did not meet the desired level of intuitive interpretability. Further development was deemed an appropriate use of time as informative values had been demonstrated as a useful variable set. This led to the most current version of informative values, Informative Values 7.0, which were used during Stage 2.

In Stage 2, another Excel-based tool was created for realignment of datasets to a variable of choice. Use of this tool was restricted to realignment to specific values for elapsed time or IVC, however the tool allows realignment to any variable which changes in a roughly proportional manner with time. For example, it is unlikely that the current version of the tool would be able to realign a typical fed-batch dataset according to lactate as this can follow an increase-decrease relationship with time caused by a period of accumulation in the culture followed by a period of consumption by cells. Further development could allow some analogous version to be applied.

In Stage 3 a key knowledge gap in the collected data was identified. It was shown that for this investigation batch numbers for raw materials were of little statistical use and

provided a strong example of a case where the use of spectral devices such as Raman probes might provide valuable information concerning raw material quality.

The capture and use of such spectral data is of growing interest in the biopharmaceutical industry. While the presented work did not deal with spectral data, key learning points could be carried over. As an example, in spectral data analysis, wavelengths or wave numbers are typically treated as variables measuring emission or absorption of light, depending of spectroscopy type. Many thousands of wavenumbers are recorded for a single sample leading to situations similar to those seen in Chapter 5 with high frequency data from online monitoring. Two approaches demonstrated in this thesis could be considered:

1. Summarising spectral data as sets of key values, analogous to informative values.
2. Compression of the spectral dataset using PCA and using the resulting scores in subsequent analyses, as demonstrated with the use of PC-decision trees.

7.3 Multi-Product Platform Process Analysis

Chapter 6 described an investigation conducted on multiple projects using a common host cell and process platform (GS-CHO Version 6) to identify variables correlated with crashes in culture viability. In regards to this hard aim, pH control and behaviour were identified as the top variable of interest. pCO_2 control was highlighted as an additional area for process improvement with Na^+ concentration as an additional indicator of interest.

General softer outcomes were identified concerning overall trends in the dataset and the way in which the analysis was perform. General trends included identifying a range for the ‘best’ pass/fail decline limits for the dataset, i.e. misclassification errors for models were lower when limits of 30%/d and 40%/d were used. Model optimisation using response surface methodology to minimise misclassification error indicated a local optimum decline limit of 45%/d.

Outcomes concerning the method of analysis included improvements to interpretability by employing model hierarchies (e.g. Analysis Pattern 2 and Analysis Pattern 3) and the creation of a confounding reduction method appropriate for the desired use and future implementation. Furthermore, it was demonstrated that classification methods should not be selected based on perceived statistical power alone. This was shown by both the lower misclassification error and increased interpretability for models using decision trees

algorithms when compared to models using PLS-DA, which would generally be regarded as the more statistically powerful.

During the platform process analysis, a variety of different tools and techniques were employed. The key obstacle faced was the sheer scale of the dataset to be considered as it stretched back several years and covered a wide variety of products with uneven distribution of culture numbers, scales, and instances of crash behaviour. Indeed, the first task undertaken was a quantitative definition of crash/non-crash behaviour to allow efficient classification of the cultures in an acceptable time frame.

The scale of the dataset was addressed through the use of a many models meta-analysis which also serves as a response to the question posed at the start of Chapter 6, "Where do I begin?" The use of many models allowed general trends captured in the dataset to be identified and thence direct more focussed analyses based on those trends.

One example of the benefits of the meta-analysis approach was the use of multiple pass/fail limits for decline in viability over a 24 hour period to identify a 'natural' pass/fail division in the dataset. A second example of the benefits of the meta-analysis approach was the use of two different scaling approaches (Autoscale and Intrascale A) with three primary benefits.

1. When Autoscale was applied to the dataset, project-specific variance was allowed to influence the resulting multi-project model. This provided the person performing the analysis an opportunity to identify projects with behaviours notably dissimilar to other projects in the dataset.
2. The use of Intrascale A allowed the influence of project-specific variance in a multi-project dataset to be reduced.
3. Use of both scaling methods as part of a meta-analysis approach enabled determination of whether any confounding by product/project observed during an initial exploratory PCA affected classification accuracy.

Finally, the greatest hurdles to implementing a meta-analysis approach to test various options for data realignment, justification, scaling, etc. were identified as the initial collation of the dataset and data checking. These activities were measured in weeks, whereas the creation of several spreadsheet-based tools allowed many different options to be applied in seconds, generation of the multiple models was a matter of minutes, and initial conclusions were available within hours.

7.4 Informative Values

The aim of the research presented in Appendix A was improved analysis of high frequency datasets generated through online monitoring of cell cultures. Initially this was specifically for the purpose of the productivity investigation described in Chapter 6. Due to the high potential of the high frequency online monitoring dataset, it was decided to further build on the initial, project-specific work to create a robust tool designed for use with any high frequency online monitoring data.

Common obstacles in the use of online monitoring datasets in biopharmaceutical process are reliance on qualitative, univariate comparisons of monitored parameters and reliance on personal experiences. While techniques such as PCA and PLS-DA can be applied can be applied directly to online monitoring datasets, interpretation of results can become difficult due to both the high number of observations and the lack of similarly high-frequency variables of interest (e.g. product concentration). Typically these variables of interest or variables closely correlated with such variables of interest (e.g. cell growth may be closely correlated with product concentration) are included in offline monitoring dataset. However integration of online monitoring and offline monitoring dataset by simple extension of the dataset analysed is rarely possible as online monitoring data usually overwhelms offline monitoring data due to sampling frequency.

From these circumstances, the hard research aim was defined as the downsampling of the high frequency online monitoring dataset in a manner that captured behaviours in a quantitative form meeting the following criteria:

a) Appropriate for follow-on use in MVDA.
b) Retained a high degree of intuitive interpretability by scientists.
c) Allowed integration with offline sampling datasets.

These three criteria were met through the creation of a set of robust summary statistics termed *informative values*. The use of informative values allowed the use of PCA to identify unusual behaviours in online monitoring behaviours. These ranged from one-off events (e.g. temporary disconnection of a temperature probe) to differences in movement around temperature setpoint between reactors of the same scale.

Due to the summary statistics selected to make up informative values, these behaviours were communicated in a simple and efficient manner with a high degree of intuitive interpretability. In the case of the temperature probe disconnection, for the culture in

question in one 24 hour block the area below median temperature (ABM_{Temp}) and the total area away from the median temperature ($TAAM_{Temp}$) for temperature were notably high. However the area above median temperature (AAM_{Temp}) and median absolute distance (MAD_{Temp}) were within expected ranges for the 24 hours in question. Informative values for all other 24 hour blocks appeared within expected ranges. From this, it was known that an event had occurred with the following conditions:

1. The event was restricted to the 24 hours in question
2. The event did not last long enough to affect the median temperature or the median absolute distance around the median temperature within that 24 hour block.
3. The event increased area away from the median temperature in one direction only.

From these simple conclusions, it was determined that the recorded temperature measurement had dropped drastically for a short time. Referring back to the original data for the 24 hours under consideration, it was revealed that an error had occurred with the temperature probe and null readings were recorded.

The ability to integrate online monitoring and offline monitoring datasets following translation to informative values was demonstrated in the work presented in Chapter 5. There was seen that models generated from datasets integrating online monitoring data (as informative values) and offline monitoring data resulted in lower misclassification error by decision trees when classifying cultures as high producing or low producing. It was suggested that this was due to both greater availability of information to the model and improved estimation of missing data when using iterative PCA to estimate missing data.

Translation of high frequency online monitoring datasets into informative values initially required approximately 30 minutes per culture. By the conclusion of the research described in Chapter 5, processing time had reduced to approximately 30 seconds per culture. This was achieved through the development of the first purpose-built tool created from presented research, the Excel-based “Efficient Process Capture - Calculation and Alignment Tool” (EPIC-CAT).

7.5 Final Conclusions

During the course of the industry placement and through attendance at both academic and industry conferences, a wide variety of attitudes towards MVDA were encountered. These ranged from dismissal of MVDA techniques as unnecessary, somehow fallacious

representations of systems, or unsuitable for use outside laboratory conditions to strong devotion to a single statistical method or framework. A third extreme encountered was that a model can be created but that model must be perfect, entirely accurate, and the only model created.

These attitudes place a great deal of emphasis on pure predictive/classification accuracy and less on practical use of a model as a tool available to scientists. Furthermore, they centre on idealised scenarios, both that the model is appropriate for all interrogations being made of the dataset and that enough time is afforded to create such perfected models. There is also a tendency to rely on a single, inflexible process of analysis. The sum effect of these attitudes can lead to a general reluctance to apply MVDA until some form of crisis occurs.

Toolbox Division	**Tool**	**Type**	**Main Chapter (Related)**
1 – Core Technique	Core Technique Selection Guide	D	(3, 4, 5, 6)
2 – Dataset Adjustment Tools	Dataset Adjustment Tool Guide	D	(5, 6)
	Standardised Data Collection	S	6
	Dataset Collation	S	5
	Dataset Reorientation	S	5
	Dataset Realignment	S	5
	Dataset Re-Justification	S	6
	Data Intrascaling	S	6
	Data Source, Sample, and Variable Selection Guide	D	(3, 4, 5, 6)
	Missing Data Handling	D	5 (6)
3 – Complementary Tools	EPIC-CAT	S	Appendix A (5)
	EPIC-CAT Collator	S	Appendix A (5)
	Osmolality Residuals Guide	D	4
	Simple Profile Classifier	S	6
4 – Frameworks	Result Interpretation Guide	D	(4, 5, 6)
	Analysis Schemas	F	(4, 5, 6)

Table 33. Summary of statistical toolbox contents. The toolbox is split into four main divisions containing written documents (D), process flow documents (F), and spreadsheet-based tools (S). Where appropriate, the chapter from which the item originated is listed. Additional chapters where the item was used are indicated. If no specific chapter is indicated, the item was based upon the research body as a whole.

The presented research approached the application and implementation of MVDA in biopharmaceutical processes as the creation of a robust statistical toolbox (Table 33). This toolbox included not only the tools, but also guidance on how to use tools, warnings of how not to use tools, and flexible frameworks demonstrating how multiple statistical tools could be chained together based on user requirements.

In the four projects described, a variety of ways of using pre-existing historical datasets and new datasets generated from current standard data monitoring has been demonstrated. These included:

1. Comparison of discrepancies in readings by supposedly interchangeable technologies.
2. Extraction of indirectly captured information.
3. Translation of a high frequency dataset for improved user interpretability and integration with a lower frequency dataset.
4. Demonstration of effects of options for missing data estimation on both dataset spread and during subsequent MVDA.
5. Evidence for an alternative measure of culture progression in place of time
6. A means of removing underlying variability in a dataset from variation in a chosen progression variable (e.g. time or IVC).
7. Identification of a knowledge gap and how that gap may be resolved including integration into the presented MVDA frameworks.
8. A means of removing a well-known, if not necessarily well-understood, source of confounding to allow analysis of an underlying shared process.
9. Evidence of potential benefits in re-justifying physical structuring of samples to better reflect model purpose/responses.
10. Demonstration of a meta-analytical approach to define general trends and improved problem definition.

Demonstration of the benefits of employing multiple ‘simpler’ models with good user interpretability in place of a single, more comprehensive model with lower

As all necessary data were generated through normal company activities, the only additional company resources required were a workspace and access to the scientists involved in data generation, who represented the intended future users of the research outcomes.

In short, it has been shown that many areas for improvement in biopharmaceutical processes can be addressed by allowing an appropriate investment of time and freedom to fail in order to explore which statistical methods do or do not work, where they work, and to develop supplementary tools to enable use of statistical methods as a normal action.

It is hoped that the frameworks and tools developed and demonstrated in this thesis will be used to further support the implementation of the PAT Initiative and statistical process control in the biologics and other related industries

References

[1] AT&T Corporation, AT&T Unit is First U.S. Manufacturer to Capture Japan's Top Quality Prize, (1994).

[2] S. Charaniya, H. Le, H. Rangwala, K. Mills, K. Johnson, G. Karypis, Mining manufacturing data for discovery of high productivity process characteristics, Journal of Biotechnology. 147 (2010) 186–197.

[3] A.J. Porter, A.J. Racher, R. Preziosi, A.J. Dickson, Strategies for selecting recombinant CHO cell lines for cGMP manufacturing: improving the efficiency of cell line generation., Biotechnology Progress. 26 (2010) 1455–1464.

[4] N. Jenkins, Modifications of therapeutic proteins: Challenges and prospects, Cytotechnology. 53 (2007) 121–125.

[5] P. Utgoff, C. Brodley, An incremental method for finding multivariate splits for decision trees, in: M. Kaufmann (Ed.), Proceedings of the Seventh International Cpnference on Machine Learning, Austin, TX, 1990.

[6] S. Wold, Personal memories of the early PLS development, Chemometrics and Intelligent Laboratory Systems. 58 (2001) 83–84.

[7] H. Martens, Reliable and relevant modelling of real world data: A personal account of the development of PLS Regression, Chemometrics and Intelligent Laboratory Systems. 58 (2001) 85–95.

[8] R. Kamimura, K. Konstantinov, G. Stephanopoulos, Knowledge-based systems, artificial neural networks and pattern recognition: applications to biotechnological processes., Current Opinion in Biotechnology. 7 (1996) 231–234.

[9] F. Moatar, F. Fessant, A. Poirel, pH modelling by neural networks. Application of control and validation data series in the Middle Loire river, Ecological Modelling. 120 (1999) 141–156.

[10] J. Flanagan, Self-organisation in Kohonen's SOM, Neural Networks. 9 (1996) 1185–1197.

[11] M. Ignova, G.A. Montague, A.C. Ward, J. Glassey, Fermentation seed quality analysis with self-organising neural networks, Biotechnology and Bioengineering. 64 (1999) 82–91.

[12] J. Hox, T. Bechger, An introduction to structural equation modeling, Family Science Review. 11 (1998) 354–373.

[13] I. Miletic, S. Quinn, M. Dudzic, V. Vaculik, M. Champagne, An industrial perspective on implementing on-line applications of multivariate statistics, Journal of Process Control. 14 (2004) 821–836.

[14] M. Proust, R.A. Fisher, Design of Experiments, 2nd ed., Oliver and Boyd, Edinburgh, 1937.

[15] ICH, Guidance for Industry Q11 Development and Manufacture of Drug Substances, (2012).

[16] FDA, 42§262 Regulation of Biological Products, USA, 1999.

[17] EMA, Glossary (terms and abbreviations), 2010.

[18] A. Kantardjieff, W. Zhou, Mammalian cell cultures for biologics manufacturing., 2014.

[19] A. Gawer, ed., Platforms, Markets, and Innovation, Edward Elgar, Cheltenham, UK, 2009.

[20] European Biopharmaceutical Enterprises, Platform Manufacturing of Biopharmaceuticals: Putting Accumulated Data and Experience to Work, Brussels, 2013.

[21] L. Bren, The Road to the Biotech Revolution, FDA Consumer Magazine. (2006).

[22] D.C. Augenstein, a J. Sinskey, D.I. Wang, Effect of shear on the death of two strains of mammalian tissue cells., Biotechnology and Bioengineering. 13 (1971) 409–418.

[23] W.S. Hu, J.M. Piret, Mammalian cell culture processes., Current Opinion in Biotechnology. 3 (1992) 110–114.

[24] A. McQueen, J.E. Bailey, Influence of serum level, cell line, flow type and viscosity on flow-induced lysis of suspended mammalian cells, Biotechnology Letters. 11 (1989) 531–536.

[25] L. Chu, D.K. Robinson, Industrial choices for protein production by large-scale cell culture., Current Opinion in Biotechnology. 12 (2001) 180–7.

[26] M. Butler, Animal cell cultures: Recent achievements and perspectives in the production of biopharmaceuticals, Applied Microbiology and Biotechnology. 68 (2005) 283–291.

[27] J. Shiloach, R. Fass, Growing E. coli to high cell density - A historical perspective on method development, Biotechnology Advances. 23 (2005) 345–357.

[28] H. Li, M. d'Anjou, Pharmacological significance of glycosylation in therapeutic proteins, Current Opinion in Biotechnology. 20 (2009) 678–684.

[29] A. Sinclair, S. Elliott, Glycoengineering: the effect of glycosylation on the properties of therapeutic proteins, Journal of Pharmaceutical Sciences. 94 (2005) 1626–1635.

[30] N. Jenkins, E.M.A. Curling, Glycosylation of recombinant proteins: Problems and prospects, Enzyme and Microbial Technology. 16 (1994) 354–364.

[31] A.S. De Groot, D.W. Scott, Immunogenicity of protein therapeutics, Trends in Immunology. 28 (2007) 482–490.

[32] A.S. De Groot, Immunomics: Discovering new targets for vaccines and therapeutics, Drug Discovery Today. 11 (2006) 203–209.

[33] J.C. Tong, E.C. Ren, Immunoinformatics: Current trends and future directions, Drug Discovery Today. 14 (2009) 684–689.

[34] A.J. Chirino, M.L. Ary, S. a. Marshall, Minimizing the immunogenicity of protein therapeutics, Drug Discovery Today. 9 (2004) 82–90.

[35] M.D.F.S. Barbosa, E. Celis, Immunogenicity of protein therapeutics and the interplay between tolerance and antibody responses, Drug Discovery Today. 12 (2007) 674–681.

[36] C.H. Chung, B. Mirakhur, E. Chan, Q.-T. Le, J. Berlin, M. Morse, et al., Cetuximab-Induced Anaphylaxis and IgE Specific for Galactose- α-1,3-Galactose, New England Journal of Medicine. 358 (2008) 1109–1117.

[37] R.S. Senger, M.N. Karim, Effect of shear stress on intrinsic CHO culture state and glycosylation of recombinant tissue-type plasminogen activator protein., Biotechnology Progress. 19 (2003) 1199–209.

[38] R. Godoy-Silva, M. Mollet, J.J. Chalmers, Evaluation of the effect of chronic hydrodynamical stresses on cultures of suspensed CHO-6E6 cells, Biotechnology and Bioengineering. 102 (2009) 1119–1130.

[39] A.W. Nienow, Reactor engineering in large scale animal cell culture, Cytotechnology. 50 (2006) 9–33.

[40] C. Born, Z. Zhang, M. Al-Rubeai, C.R. Thomas, Estimation of disruption of animal cells by laminar shear stress, Biotechnology and Bioengineering. 40 (1992) 1004–1010.

[41] R.G. Werner, F. Walz, W. Noé, A. Konrad, Safety and economic aspects of continuous mammalian cell culture., Journal of Biotechnology. 22 (1992) 51–68.

[42] K.K.S. Buck, V. Subramanian, D.E. Block, Identification of critical batch operating parameters in fed-batch recombinant E. coli fermentations using decision tree analysis., Biotechnology Progress. 18 (2002) 1366–76.

[43] G.A. Montague, E.B. Martin, C.J. O'Malley, Forecasting for fermentation operational decision making, Biotechnology Progress. (2008) 1033–1041.

[44] E. Martin, A. Morris, Batch process monitoring for consistent production, Computers and Chemical Engineering. 20 (1996) 599–604.

[45] J.C. Gunther, J.S. Conner, D.E. Seborg, Fault Detection and Diagnosis in an Industrial Fed-Batch Cell Culture Process, (2007) 851–857.

[46] S.H.G. Khoo, M. Al-Rubeai, Metabolic characterization of a hyper-productive state in an antibody producing NS0 myeloma cell line, Metabolic Engineering. 11 (2009) 199–211.

[47] J.H. Cho, P.U. Kurup, Decision tree approach for classification and dimensionality reduction of electronic nose data, Sensors and Actuators, B: Chemical. 160 (2011) 542–548.

[48] FDA, Guidance for Industry. PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance, Rockville, MD, 2004.

[49] FDA, FDA approves Gazyva for chronic lymphocytic leukemia, (2013).

[50] FDA, FDA approves Perjeta for neoadjuvant breast cancer treatment, (2013).

[51] ICH, ICH Vision, (2015).

[52] M.E. Peña-rodríguez, Statistical Process Control for the FDA-Regulated Industry, 2013.

[53] EMA, Workshop on Process Analytical Technologies for Biologicals, 2007.

[54] J.K. Liker, The Toyota Way, McGraw-Hill, New York, NY, 2004.

[55] S. Charaniya, W. Hu, G. Karypis, Mining bioprocess data: opportunities and challenges, TRENDS in Biotechnology. 26 (2008) 690–699.

[56] C. Ündey, S. Ertunç, T. Mistretta, B. Looze, Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control, Journal of Process Control. 20 (2010) 1009–1018.

[57] G. Milman, L.S. Portnoff, D.C. Tiemeier, Immunochemical evidence for glutamine-mediated degradation of glutamine synthetase in cultured Chinese hamster cells, J Biol Chem. 250 (1975) 1393–1399.

[58] A. Vernon, The GS Gene Expression System, PHARMATECH 2004. (2004) 1–3.

[59] J.R. Birch, A.J. Racher, Antibody production, Advanced Drug Delivery Reviews. 58 (2006) 671–685.

[60] M.I. Cockett, C.R. Bebbington, G.T. Yarranton, High level expression of tissue inhibitor of metalloproteinases in Chinese hamster ovary cells using glutamine synthetase gene amplification, Biotechnology (N Y). 8 (1990) 662–667.

[61] M.C. de la Cruz Edmonds, M. Tellers, C. Chan, P. Salmon, D.K. Robinson, J. Markusen, Development of transfection and high-producer screening protocols for the CHOK1SV cell system., Molecular Biotechnology. 34 (2006) 179–90.

[62] J.J. Cacciatore, L.A. Chasin, E.F. Leonard, Gene amplification and vector engineering to achieve rapid and high-level therapeutic protein production using the Dhfr-based CHO cell selection system, Biotechnology Advances. 28 (2010) 673–681.

[63] Z. Jiang, Y. Huang, S.T. Sharfstein, Regulation of recombinant monoclonal antibody production in chinese hamster ovary cells: a comparative study of gene copy number, mRNA level, and protein expression., Biotechnology Progress. 22 (2006) 313–8.

[64] P. Hossler, S.F. Khattak, Z.J. Li, Optimal and consistent protein glycosylation in mammalian cell culture, Glycobiology. 19 (2009) 936–949.

[65] D.M. Sheeley, B.M. Merrill, L.C. Taylor, Characterization of monoclonal antibody glycosylation: comparison of expression systems and identification of terminal alpha-linked galactose., Analytical Biochemistry. 247 (1997) 102–110.

[66] C. Huhn, M.H.J. Selman, L.R. Ruhaak, A.M. Deelder, M. Wuhrer, IgG glycosylation analysis, Proteomics. 9 (2009) 882–913.

[67] Y.T. Jeong, O. Choi, H.R. Lim, Y.D. Son, H.J. Kim, J.H. Kim, Enhanced sialylation of recombinant erythropoietin in CHO cells by human glycosyltransferase expression, Journal of Microbiology and Biotechnology. 18 (2008) 1945–1952.

[68] Z. Wang, J.H. Park, H.H. Park, W. Tan, T.H. Park, Enhancement of recombinant human EPO production and sialylation in Chinese hamster ovary cells through Bombyx mori 30Kc19 gene expression, Biotechnology and Bioengineering. 108 (2011) 1634–1642.

[69] M. Butler, Optimisation of the cellular metabolism of glycosylation for recombinant proteins produced by mammalian cell systems, Cytotechnology. 50 (2006) 57–76.

[70] G.T. Yarranton, Mammalian recombinant proteins: Vectors and expression systems, Current Opinion in Biotechnology. 1 (1990) 133–140.

[71] H.E. Chadd, S.M. Chamow, Therapeutic antibody expression technology., Current Opinion in Biotechnology. 12 (2001) 188–94.

[72] G. Köhler, C. Milstein, Continuous cultures of fused cells secreting antibody of predefined specificity, Nature. 256 (1975) 495–497.

[73] C. Milstein, The hybridoma revolution: An offshoot of basic research, BioEssays. 21 (1999) 966–973.

[74] F.M. Wurm, D. Hacker, First CHO genome, Nature Biotechnology. 29 (2011) 718–20.

[75] S. Estes, M. Melville, Mammalian Cell Line Developments in Speed and Efficiency, Advances in Biochemical Engineering and Biotechnology. 139 (2014)

11–33.

[76] G. Urlaub, L.A. Chasin, Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity, Proceedings of the National Academy of Sciences of the United States of America. 77 (1980) 4216–4220.

[77] L.M. Barnes, C.M. Bentley, A.J. Dickson, Advances in animal cell recombinant protein production: GS-NS0 expression system., Cytotechnology. 32 (2000) 109–23.

[78] P.M. O'Callaghan, J. McLeod, L.P. Pybus, C.S. Lovelady, S.J. Wilkinson, A.J. Racher, et al., Cell line-specific control of recombinant monoclonal antibody production by CHO cells, Biotechnology and Bioengineering. 106 (2010) 938–951.

[79] E.M. Yoo, K.R. Chintalacharuvu, M.L. Penichet, S.L. Morrison, Myeloma expression systems., Journal of Immunological Methods. 261 (2002) 1–20.

[80] L.M. Barnes, C.M. Bentley, A.J. Dickson, Molecular Definition of Predictive Indicators of Stable Protein Expression in Recombinant NS0 Myeloma Cells, Biotechnology and Bioengineering. 85 (2004) 115–121.

[81] J.J. Trill, A.R. Shatzman, G. Subinay, Production of monoclonal antibodies in COS and CHO cells, Current Opinion in Biotechnology. 6 (1995) 553–560.

[82] J. Zhang, D. Robinson, Development of animal-free, protein-free and chemically-defined media for NS0 cell culture, Cytotechnology. 48 (2005) 59–74.

[83] F. Li, N. Vijayasankaran, A. Shen, R. Kiss, A. Amanullah, Cell culture processes for monoclonal antibody production, mAbs. 2 (2010) 466–479.

[84] C. Bebbington, C. Hentschel, The expression of recombinant DNA products in mammalian cells, TRENDS in Biotechnology. 3 (1985) 314–317.

[85] L.M. Barnes, C.M. Bentley, N. Moy, A.J. Dickson, S. Eden, P. Road, Molecular Analysis of Successful Cell Line Selection in Transfected GS-NS0 Myeloma Cells, 96 (2007) 337–348.

[86] Lonza, Media and Feeds, (2012).

[87] S. Oguchi, H. Saito, M. Tsukahara, H. Tsumura, pH Condition in temperature shift cultivation enhances cell longevity and specific hMab productivity in CHO culture, Cytotechnology. 52 (2006) 199–207.

[88] D.J. Galbraith, A.S. Tait, A.J. Racher, J.R. Birch, D.C. James, Control of culture environment for improved polyethylenimine-mediated transient production of recombinant monoclonal antibodies by CHO cells., Biotechnology Progress. 22 (2006) 753–62.

[89] S.K. Yoon, J.Y. Song, G.M. Lee, Effect of low culture temperature on specific productivity, transcription level, and heterogeneity of erythropoietin in Chinese hamster ovary cells, Biotechnology and Bioengineering. 82 (2003) 289–298.

[90] S.K. Yoon, S.H. Kim, G.M. Lee, Effect of low culture temperature on specific productivity and transcription level of anti-4-1BB antibody in recombinant Chinese hamster ovary cells, Biotechnology Progress. 19 (2003) 1383–1386.

[91] S.S. Ozturk, B.O. Palsson, Growth, metabolic, and antibody production kinetics of hybridoma cell culture: 1. Analysis of data from controlled batch reactors., Biotechnology Progress. 7 (n.d.) 471–80.

[92] J. Müthing, S.E. Kemminer, H.S. Conradt, D. Šagi, M. Nimtz, U. Kärst, et al., Effects of buffering conditions and culture pH on production rates and

glycosylation of clinical phase I anti-melanoma mouse IgG3 monoclonal antibody R24, Biotechnology and Bioengineering. 83 (2003) 321–334.

[93] E. Spens, Development of a protein-free fed-batch process for NS0 cells: Studies on regulation of proliferation, Royal Institue of Technology (Stockholm), 2006.

[94] S. Kuwae, T. Ohda, H. Tamashima, H. Miki, K. Kobayashi, Development of a fed-batch culture process for enhanced production of recombinant human antithrombin by Chinese hamster ovary cells., Journal of Bioscience and Bioengineering. 100 (2005) 502–510.

[95] G. Schmid, G.H. Blanch, C.R. Wilke, Hybridoma growth, metabolism, and product formulation in HEPES-buffered medium. II. Effect of pH, Biotechnology Letters. (1990) 159–165.

[96] J.A. Zanghi, A.E. Schmelzer, T.P. Mendoza, R.H. Knop, W.M.. Miller, Bicarbonate concentration and osmolality are key determinants in the inhibition of CHO cell polysilylation under elevated pCO2 or pH, Biotechnology and Bioengineering. (1999) 182–191.

[97] S.S. Ozturk, B.O. Palsson, Growth, metabolic and antibody production kinetics of hybridoma cell culture, Biotechnology Progress. 7 (1991) 481–494.

[98] J.J. Osman, J. Birch, J. Varley, The response of GS-NS0 myeloma cells to pH shifts and pH perturbations., Biotechnology and Bioengineering. 75 (2001) 63–73.

[99] A.A. Lin, W.M. Miller, CHO Cell Responses to Low Oxygen: Regulation of Oxygen Consumption and Sensitization to Oxidative Stress, Biotechnology and Bioengineering. 40 (1992) 505–16.

[100] A.A. Lin, R. Kimura, W.M. Millert, Production of tPA in Recombinant CHO Cells Under Oxygen-Limited Conditions, Biotechnology and Bioengineering. 42 (1993) 339–350.

[101] J.G. Khinast, Characterization of the Localized Hydrodynamic Shear Forces and Dissolved Oxygen Distribution in Sparged Bioreactors, Biotechnology and Bioengineering. 97 (2007) 317–331.

[102] J. Luo, N. Vijayasankaran, J. Autsen, R. Santuray, T. Hudson, A. Amanullah, et al., Comparative metabolite analysis to understand lactate metabolism shift in Chinese hamster ovary cell culture process, Biotechnology and Bioengineering. 109 (2012) 146–156.

[103] J.A. Serrato, L.A. Palomares, Heterogeneous Conditions in Dissolved Oxygen Affect N-Glycosylation but Not Productivity of a Monoclonal Antibody in Hybridoma Cultures, Biotechnology and Bioengineering. 88 (2004) 176–188.

[104] C.J. Fernandes, L. Rong, T. Tamura, K.D. Stewart, L.K. Rogers, H.W. McMicken, et al., Stable transfection of Chinese hamster ovary cells with glutamate-cysteine ligase catalytic subunit cDNA confers increased resistance to tert-butyl hydroperoxide toxicity., Toxicology Letters. 136 (2002) 107–20.

[105] B.C. Mulukutla, S. Khan, A. Lange, W.-S. Hu, Glucose metabolism in mammalian cell culture: new insights for tweaking vintage pathways TL - 28, Trends in Biotechnology. 28 VN - r (2010) 476–484.

[106] K.H. Moley, M.M. Mueckler, Glucose transport and apoptosis, Apoptosis. 5 (2000) 99–105.

[107] Y.S. Tsao, A.G. Cardoso, R.G.G. Condon, M. Voloch, P. Lio, J.C. Lagos, et al., Monitoring Chinese hamster ovary cell culture by the analysis of glucose and

lactate metabolism, Journal of Biotechnology. 118 (2005) 316–327.

[108] X. Sun, Y. Zhang, Glutamine cannot support recombinant CHO cell growth and maintenance in the absence of glucose, Process Biochemistry. 39 (2004) 717–720.

[109] M. Gagnon, G. Hiller, Y.T. Luan, A. Kittredge, J. Defelice, D. Drapeau, High-End pH-controlled delivery of glucose effectively suppresses lactate accumulation in CHO Fed-batch cultures, Biotechnology and Bioengineering. 108 (2011) 1328–1337.

[110] J.I. Rearick, A. Chapman, S. Kornfeld, Glucose Starvation Alters Lipid-linked Oligosaccharide Biosynthesis in Chinese Hamster Ovary Cells, Journal of Biological Chemistry. 256 (1981) 6255–6261.

[111] F. Chen, Z. Ye, L. Zhao, X. Liu, L. Fan, W.S. Tan, Correlation of antibody production rate with glucose and lactate metabolism in Chinese hamster ovary cells, Biotechnology Letters. 34 (2012) 425–432.

[112] S. Ozturk, M. Riley, B. Palsson, Effects of ammonia and lactate on hybridoma growth, metabolism, and antibody production, Biotechnology and Bioengineering. (1992).

[113] N. Kurano, C. Leist, F. Messi, S. Kurano, A. Fiechter, Growth behavior of Chinese hamster ovary cells in a compact loop bioreactor. 2. Effects of medium components and waste products, Journal of Biotechnology. 15 (1990) 113–128.

[114] Z. Xing, Z. Li, V. Chow, S.S. Lee, Identifying inhibitory threshold values of repressing metabolites in CHO cell culture using multivariate analysis methods, Biotechnology Progress. 24 (2008) 675–683.

[115] N. Ma, J. Ellet, C. Okediadi, P. Hermes, E. McCormick, S. Casnocha, A single nutrient feed supports both chemically defined NS0 and CHO fed-batch processes: Improved productivity and lactate metabolism, Biotechnology Progress. 25 (2009) 1353–1363.

[116] B.C. Mulukutla, M. Gramer, W.-S. Hu, On metabolic shift to lactate consumption in fed-batch culture of mammalian cells., Metabolic Engineering. 14 (2012) 138–49.

[117] H. Le, S. Kabbur, L. Pollastrini, Z. Sun, K. Mills, K. Johnson, et al., Multivariate analysis of cell culture bioprocess data - Lactate consumption as process indicator, Journal of Biotechnology. 162 (2012) 210–23.

[118] A. V. Carvalhal, I. Marcelino, M.J.T. Carrondo, Metabolic changes during cell growth inhibition by p27 overexpression, Applied Microbiology and Biotechnology. 63 (2003) 164–173.

[119] K.K. Frame, W.S. Hu, Comparison of growth kinetics of producing and nonproducing hybridoma cells in batch culture, Enzyme and Microbial Technology. 13 (1991) 690–6.

[120] L. Fan, I. Kadura, L.E. Krebs, C.C. Hatfield, M.M. Shaw, C.C. Frye, Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells, Biotechnology and Bioengineering. 109 (2012) 1007–1015.

[121] F. Zhang, X. Sun, X. Yi, Y. Zhang, Metabolic characteristics of recombinant Chinese hamster ovary cells expressing glutamine synthetase in presence and absence of glutamine., Cytotechnology. 51 (2006) 21–8.

[122] K.G. Clarke, Bioprocess Engineering - An Introductory Engineering and Life Science Approach, Woodhead Publishing, 2013.

[123] M. Brys, D. Linzer, E. Papoutsakis, Ammonia effects the glycosylation patterns of recombinant mouse placental lactogen-1 by chinese hamster ovary cells in a pH-dependent manner., Biotechnology and Bioengineering. 43 (1994) 505–514.

[124] N. Demaurex, S. Grinstein, Na+/H+ antiport: modulation by ATP and role in cell volume regulation., The Journal of Experimental Biology. 196 (1994) 389–404.

[125] I.I. Marakhova, T.A. Vinogradova, E. V. Yefimova, Early and delayed changes in potassium transport during the initiation of cell proliferation in CHO culture., General Physiology and Biophysics. 8 (1989) 273–282.

[126] V.M. DeZengiotita, R. Kimura, William M. Miller, Effects of CO2 and osmolality on hybridoma cells: growth, metabolism and monoclonal antibody production, Cytotechnology. 28 (1998) 213–227.

[127] R. Kimura, W.M. Miller, Effects of Elevated pC02 and / or Osmolality on the Growth and Recombinant tPA Production of CHO Cells, Biotechnology and Bioengineering. 52 (1996) 152–160.

[128] E. Pacis, M. Yu, J. Autsen, R. Bayer, F. Li, Effects of cell culture conditions on antibody N-linked glycosylation-what affects high mannose 5 glycoform, Biotechnology and Bioengineering. 108 (2011) 2348–2358.

[129] G.E. Grampp, T.K. Blumen, K. Kelly, P. Derby, L.A. Sleeman, D. Hettwer, Environmental control of sialic acid composition in glycoproteins secreted by mammalian cells, in: Cell Culture Engineering IV Meeting, San Diego, CA, USA, 1994.

[130] Advanced Instruments, AI University, (2014).

[131] S. Brady, The Effect of Hyper-Osmotic Conditions on the Growth, Metabolism, and Specific Antibody Productivity of a GS- NS0 Cell Line, University of Maryland, 2003.

[132] Z. Zhang, J.M. Blewett, C.R. Thomas, Modelling the effect of osmolality on the bursting strength of yeast cells, Journal of Biotechnology. 71 (1999) 17–24.

[133] R. Kimura, W.M. Miller, Glycosylation of CHO-derived recombinant tPA produced under elevated pCO2, Biotechnology Progress. 13 (1997) 311–317.

[134] a E. Schmelzer, V.M. DeZengotita, W.M. Miller, Considerations for osmolality measurement under elevated pCO(2): comparison of vapor pressure and freezing point osmometry., Biotechnology and Bioengineering. 67 (2000) 189–96.

[135] Omega, RTD Resistance Temperature Detectors, (2012).

[136] RTD Products, Industry Standards, (2012).

[137] W. Boyes, Instrumentation Reference Book, 3rd ed., Elsevier, 2003.

[138] H.C. Vogel, C.L. Tadaro, Fermentation and Biochemical Engineering Handbook - Principles, Process Design, and Equipment, 2nd ed., William Andrew Publishing/Noyes, 1997.

[139] P. Du, J.A. Kofman, Electronic Laboratory Notebooks in Pharmaceutical R&D: On the Road to Maturity, Journal of Laboratory Automation. 12 (2007) 157–165.

[140] Nova Biomedical, BioProfile® Automated Chemistry Analyzers For Cell Culture and Fermentation, Waltham, MA, USA, 2011.

[141] K.S. Louis, A.C. Siegel, Cell Viability nalysis Using Trypan Blue: Manual and Automated Methods, Methods in Molecular Biology. (2011) 7–12.

[142] Gonotec GmbH, Gonotec, (n.d.).

[143] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann Publishers, Waltham, MA, USA, 2012.

[144] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems. 2 (1987) 37–52..

[145] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems. (2001) 109–130.

[146] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chimica Acta. (1986) 1–17.

[147] P. Druilhet, A. Mom, PLS regression: A directional signal-to-noise ratio approach, Journal of Multivariate Analysis. 97 (2006) 1313–1329.

[148] Y. Zhang, T.F. Edgar, Multivariate Statistical Process Control, in: M.A. Boudreau, G.K. McMillan (Eds.), New Directions in Bioprocess Modeling and Control: Maximizing Process Analytical Technology Benefits, ISA, 2007: pp. 247–286.

[149] M. Robnik-Šikonja, I. Kononenko, Theoretical and Empirical Analysis of ReliefF and RReliefF, Machine Learning. 53 (2003) 23–69.

[150] T. Byrne, S. Wold, Data mining using PLS-trees and other projection methods, in: Advanced Semiconductor Manufacturing Conference (ASMC), 2011 22nd Annual IEEE/SEMI, 2011: pp. 1–5.

[151] A. Bourgoin, Biologics Market : An Overview, (2012) 1–29.

[152] E. Trummer, K. Fauland, S. Seidinger, K. Schriebl, C. Lattenmayer, R. Kunert, et al., Process Parameter Shifting : Part I . Effect of DOT , pH , and Temperature on the Performance of Epo-Fc Expressing CHO Cells Cultivated in Controlled Batch Bioreactors, Biotechnology and Bioengineering. 94 (2006) 1033–1044.

[153] Radiometer Analytical, The making of a combined pH electrode – the inside story, (2005).

[154] J.J. Barron, C. Ashton, L. Geary, S.F. Zone, C. Clare, The Effects of Temperature on pH Measurement, Shannon Free Zone, Co. Clare, Ireland, 2006.

[155] Radiometer Analytical, Conductivity Theory and Practice, 2004.

[156] B.H. Evans, T. Larson, Dealing with Disparity in On-line and Off-line pH Measurements, (2006) 1–4.

[157] V. Saucedo, B. Wolk, A. Arroyo, C.D. Feng, Studying the drift of in line pH measurements in cell culture, Biotechnology Progress. 27 (2011) 885–890.

[158] H. Yoshimura, Effects of anticoagulants on the pH of the blood, Journal of Biochemistry. 22 (1935) 279–295.

[159] T.B. Rosenthal, The Effect of Temperature on the pH of Blood and Plasma In Vitro, Journal of Biological Chemistry. 173 (1948) 25–30.

[160] H. Yoshimura, T. Fujimoto, Is the Hydrogen Gas Electrode Not Applicable to the Determination of the pH of Oxygenated Blood? Studies on the blood pH estimated by the glass electrode method. VI., Journal of Biochemistry. 25 (1937) 493–518.

[161] Mettler-Toledo AG, A Guide to pH Measurement, Mettler-Toldeo AG, Schwerzenbach, Switzerland, 2003.

[162] Nova Biomedical, COMPUTER INTERFACE MANUAL- Bioprofile Basic Analyzers, Waltham, MA, 2003.

[163] Merriam-Webster, Merriam-Webster, (2014).

[164] S. Cahan, N. Cohen, Age versus Schooling Effects on Intelligence Development, Child Development. (1989) 1239–1249.

[165] P. Nomikos, J. MacGregor, Multi-way partial least squares in monitoring batch processes, Chemometrics and Intelligent Laboratory Systems. 30 (1995) 97–108.

[166] T. Kourti, P. Nomikos, J.F. MacGregor, Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS, Journal of Process Control. 5 (1995) 277–284.

[167] C.H. Leist, H.P. Meyer, A. Fiechter, Potential and problems of animal cells in suspension culture, Journal of Biotechnology. 15 (1990) 1–46.

[168] A. Gelman, Multilevel (Hierarchical) Modeling: What It Can and Cannot Do, Technometrics. 48 (2006) 432–435.

[169] B. Horvath, M. Mun, M.W. Laird, Characterization of a monoclonal antibody cell culture production process using a quality by design approach, Molecular Biotechnology. 45 (2010) 203–206.

[170] J. Glassey, G. Montague, P. Mohan, Issues in the development of an industrial bioprocess advisory system., TRENDS in Biotechnology. 18 (2000) 136–41.

[171] M.J.T. Carrondo, P.M. Alves, N. Carinhas, J. Glassey, F. Hesse, O.W. Merten, et al., How can measurement, monitoring, modeling and control advance cell culture in industrial biotechnology?, Biotechnology Journal. 7 (2012) 1522–1529.

[172] B. Lennox, K. Kipling, J. Glassey, G. Montague, M. Willis, H. Hiden, Automated production support for the bioprocess industry, Biotechnology Progress. 18 (2002) 269–275.

[173] T.P. Frandsen, H. Næsted, S.K. Rasmussen, P. Hauptig, F.C. Wiberg, L.K. Rasmussen, et al., Consistent manufacturing and quality control of a highly complex recombinant polyclonal antibody product for human therapeutic use, Biotechnology and Bioengineering. 108 (2011) 2171–2181.

[174] B.H. Junker, H.Y. Wang, Bioprocess Monitoring and Computer Control: Key Roots of the Current PAT Initiative, Biotechnology and Bioengineering. 95 (2006) 226–261.

[175] FDA, FDA Guidance concerning demonstration of comparability of human biological products, including therapeutic biotechnology-derived products, 1996.

[176] EMEA, Specifications test procedures and acceptance criteria for biotechnology/biological products. ICH Topic Q6, 1999.

[177] FDA, Quality By Design and Dissolution, in: The Advisory Committee for Pharmaceutical Science, 2004.

[178] H. Matsumoto, R. Masumoto, C. Kuroda, Feature extraction of time-series process images in an aerated agitation vessel using self organizing map, Neurocomputing. 73 (2009) 60–70.

[179] G. Barreto, Time series prediction with the self-organizing map: A review, Perspectives of NeuralSymbolic Integration. (2007) 135–158.

[180] Q. Wang, V. Megalooikonomou, A dimensionality reduction technique for efficient time series similarity analysis, Information Systems. 33 (2008) 115–132.

[181] F. Gullo, G. Ponti, A. Tagarelli, S. Greco, A time series representation model for accurate and fast similarity detection, Pattern Recognition. 42 (2009) 2998–3014.

[182] A. Chitra, S. Uma, An Ensemble Model of Multiple Classifiers for Time Series

Prediction, International Journal of Computer Theory and Engineering. 2 (2010) 454–458.

[183] Y.M. Sebzalli, X.Z. Wang, Knowledge discovery from process operational data using PCA and fuzzy clustering, Engineering Applications of Artificial Intelligence. 14 (2001) 607–616.

[184] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, B. Walczak, Robust statistics in data analysis—a review: basic concepts, Chemometrics and Intelligent Laboratory Systems. 85 (2007) 203–219.

[185] P.J. Rousseeuw, M. Hubert, Robust statistics for outlier detection, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 1 (2011) 73–79.

[186] A.M. Bianco, M. Garcia Ben, V.J. Yohai, Robust estimation for linear regression with asymmetric errors, Canadian Journal of Statistics. 33 (2005) 511–528.

Appendix A. Downsampling of Online Monitoring Data as Informative Values

A.1 Introduction

A key concern in the production of therapeutic proteins such as monoclonal antibodies (mAb) is whether the product has the correct critical quality attributes (CQA) [169–174]. For a glycosylated product such as a mAb, glycoform profiles are among the CQAs regulatory agencies require biopharmaceutical companies to characterise and maintain [175,176]. To ensure CQAs and acceptable levels of growth and productivity are met, critical process parameters are monitored to allow identification of potential issues [177].

It is easy to state that a bioreactor will operate at 36.5 °C with a pH of 6.92 and dissolved oxygen tension (DOT) of 40 %. However achieving and maintaining stated conditions can be a complex challenge with typical control systems reliant on high frequency online monitoring. Similarly challenging is making full use of the high frequency data generated. A common practice is to overlay data from different bioreactors on a single figure. This style of comparison is inadequate as it is potentially highly qualitative and subjective, and does not meaningfully tie information from online monitoring data to biological data from offline sampling.

The FDA PAT initiative is often used to justify introducing newer, more comprehensive measurement equipment (e.g. non-analyte-specific spectra capture with Raman probes). However the initiative also promotes the use of new measurement techniques to understand the impact of process controls on performance. This includes techniques for improved interrogation of data from earlier technologies. One area of interest is moving beyond statistical process control and “In control/out of control” alarm limits to linking bioreactor control system behaviour to cell culture biological behaviour.

A wide variety of statistical techniques have been applied to online monitoring data for this purpose, including self-organizing maps [178,179], data similarity measures [180,181], and ensemble methods [182]. The listed techniques focus heavily on direct application to high frequency data and lack easy interpretability or intuitive meaning for users who are not statistical experts or are not afforded the luxury of time required for an exhaustive drill down. While academically interesting and promising for future implementation, the majority of these techniques are not amenable to immediate or effective implementation in industry

Online monitoring data analysis, in particular the analysis of data from online monitoring of bioreactors, is highly reliant on contextual information. For many users, an analysis that identifies the 1000th pH reading out of several thousand (e.g. ~3 minutes out of many days) as a variable of interest is not informative and lacks intuitive meaning.

It was theorised that interpretation could be improved if the data were summarised as intermediary statistics termed "informative values" and subsequent analyses were performed using these informative values. Informative values are defined as calculated variables that quantitatively summarise online monitoring data behaviours across meaning full windows of activity, such as between offline sampling points. Most importantly, informative values representing the original data hold intuitive meaning for the end-user. These informative values can then be used in subsequent analyses in place of the original online monitoring data.

The aim of the presented study was to create these informative values. Furthermore, the informative values were also required to be robust to transfer between processes and scales as well as to imperfections from non-ideal, real processing data.

Principal component analysis (PCA) [144] was used to demonstrate that the selected informative values effectively captured online behaviours when applied to a manufacturing dataset [183].

A.2 Materials

The data used to develop the presented research were taken from the online monitoring of mAb-producing cell cultures grown in 10 L and 130 L air lift reactors at the Lonza Slough site. The dataset included monitoring for a wide variety of conditions. A more detailed description of the dataset can be found in Chapter 6. An additional, artificially generated dataset (cultures P001 to P100) was created for demonstration purposes. This was because a variety of behaviours may occur over the course of a culture; the second dataset was to more clearly demonstrate individual behaviours.

Online monitoring of pH, temperature, and DOT was achieved by the use of probes inserted into bioreactors. Online monitoring of air flowrate, O_2 flowrate, and carbon dioxide flowrate was achieved by flowrate meters on lines into the bioreactors. Probes and flowmeters were connected to control units which recorded values at a set interval.

Variables were classed according to variable purpose. Temperature, pH, and DOT were classed as *steady state* because control systems were designed to keep these variables at

defined setpoints. Flowrates for air, O_2, and CO_2 were classed as *dynamic*. Values for these variables were dependent on biological behaviour and were used to control several steady state variables. The main behaviours of interest for steady state profiles are adherence to setpoint, magnitude of movement around setpoint, and drift. Two further behaviours encountered were perturbations and shifts (changes) in setpoint (Figure 55). Perturbations were periods of up to four hours where readings remained outside acceptable noise limits in a given direction. A variety of causes for perturbations exist from incorrect controller action to probe connection issues. However, the focus in this study was capture and identification of perturbations through the use of informative values.

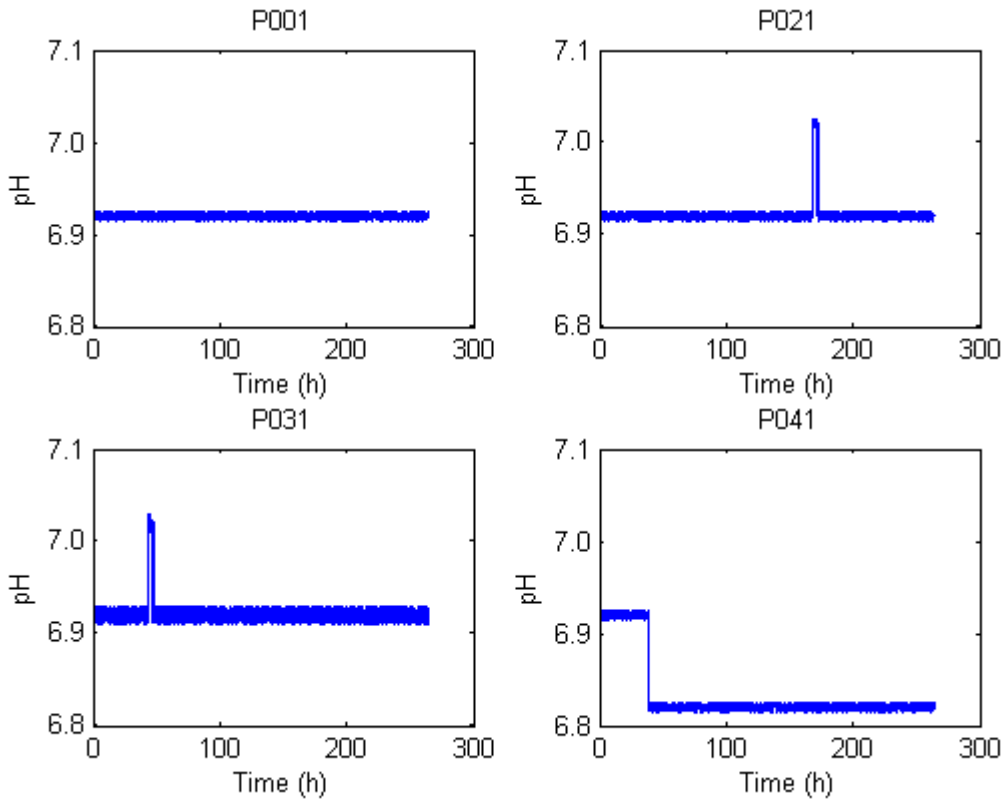


Figure 55. Artificially generated online monitoring data of pH for four theoretical cultures showing various behaviours and noise. P001: ideal, setpoint ± 0.005. P021: high perturbation, setpoint ± 0.005. P031: high perturbation, setpoint ± 0.010. P041: shift in setpoint, setpoint ± 0.005.

DOT was controlled using air, O_2, and N_2 feeds. Online DOT behaviours could typically be split into four sections:

1. An initial low noise period following inoculation.
2. When the cell mass consumes more O_2 than the airfeed can supply, an O_2 feed is activated, leading to a period of increasing noise lasting until O_2 demand peaks.
3. A period of decreasing noise as O_2 demand and hence O_2 flowrate lessen.
4. A final period of low noise where O_2 demands are by the air feed only.

A.3 Analysis of High Frequency Data in Native State

Two analyses were performed to determine whether the use of summary statistics could provide improvements in either model performance (e.g. X variance captured by model) or model interpretation by a user. The first analysis was performed using data in its native, high frequency state. The second analysis was performed using summary statistics generated from the original dataset.

A.3.1 Method

Online monitoring data from 49 cultures performed in 10 L and 130 L bioreactors at Lonza's Slough site were collated into a single, longitudinal dataset. In this orientation, there was a single observation per culture representing all data for all sampling points (Table 34). Analysed data were restricted to 1916 samplings per variable per culture (~160 h) and 11 cultures were removed, including all 130 L cultures, due to reach this criteria. Additionally, due to data loss and data validity issues related to transfer of data from datalogger to computer, CO_2 flowrate was removed as a variable.

Variable	**pH**					**Temperature (°C)**					**DOT (%)**				
Sample	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
A	$x_{A,1,1}$	…	…	…	$x_{A,1,5}$	$x_{A,2,1}$	…	…	…	$x_{A,2,5}$	$x_{A,3,1}$	…	…	…	$x_{A,3,5}$
B	$x_{B,1,1}$	…	…	…	$x_{B,1,5}$	$x_{B,2,1}$	…	…	…	$x_{B,2,5}$	$x_{B,3,1}$	…	…	…	$x_{B,3,5}$
C	$x_{C,1,1}$	…	…	…	$x_{C,1,5}$	$x_{C,2,1}$	…	…	…	$x_{C,2,5}$	$x_{C,3,1}$	…	…	…	$x_{C,3,5}$
D	$x_{D,1,1}$	…	…	…	$x_{D,1,5}$	$x_{D,2,1}$	…	…	…	$x_{D,2,5}$	$x_{D,3,1}$	…	…	…	$x_{D,3,5}$
E	$x_{E,1,1}$	…	…	…	$x_{E,1,5}$	$x_{E,2,1}$	…	…	…	$x_{E,2,5}$	$x_{E,3,1}$	…	…	…	$x_{E,3,5}$

Table 34. A longitudinal arrangement of a three-dimensional dataset (culture x sample [time] x variable). The data have been arranged so that an observation comprises all the data for a culture A to E. This data is then grouped by variable pH (1), temperature (2), DOT (3). Within in the single variable grouping, data are ordered chronologically by sample number, 1 to 5. The hierarchy of the arrangement is captured in the subscripts for values, $x_{Culture,Variable,Sample}$. An alternative arrangement could by to first group by sample numbers and then order data by variable, e.g. pH (1), temperature (2), DOT (3). This would be captured in subscripts as $x_{Culture,Sample,Variable}$.

Data were imported to Matlab for analysis with the Eigenvector PLS-Toolbox. Data were mean-centred and scaled to unit variance. A PCA model was created using random sampling (10 splits, 5 iterations) for cross-validation. A two PC model was retained, which captured 36.11% of variance in the dataset.

A.3.2 Results and Discussion

Model analysis was performed by first understanding the systematic behaviours captured through model loadings and then by performing a comparative drill down analysis on two cultures.

Figure 56 shows model loadings for PC1 and PC2. Loadings were coloured according to variable type (temperature, DOT, pH, air flowrate, O_2 flowrate). Clear structures were observed for both PC1 and PC2 loadings for the dynamic variables air flowrate and O_2 flowrate. These structures showed that the previous discussed dynamic behaviours were captured by the model.

It was also observed that PC1 and PC2 loadings for pH appeared to be sharply split into loadings > 0 and loadings < 0. This was interpreted as the model capturing changes in pH setpoint several cultures in the dataset underwent. PC1 and PC2 loadings for temperature and DOT variables were interpreted as the model capturing movement around the temperature and DOT setpoints. It is key to note that this movement could not be determined as random or systematic at this level of model interrogation.

Cultures A006 and A047 were selected for the comparative drill down analysis. A006 and A047 were selected based on their positions on the model score plot (Figure 17A) posed a greater analytical challenge because differences in behaviour would be more nuanced than statistically more obvious differences captured in PC1 behaviours (e.g. differences in setpoints). Furthermore while A047 was within 95% limits for Q Residual and Hotelling T^2 (Figure 17B), A006 was outside the Q Residual limit.

Q Residual contribution analysis (Figure 17C) indicated that the greatest difference in contributions was due to difference in pH behaviour. This was due to the change in pH setpoint for A006. This was in keeping with Q Residuals as comparison of a culture's behaviour to the mean behaviour of the model as the A006 pH setpoint change was not typical of the dataset.

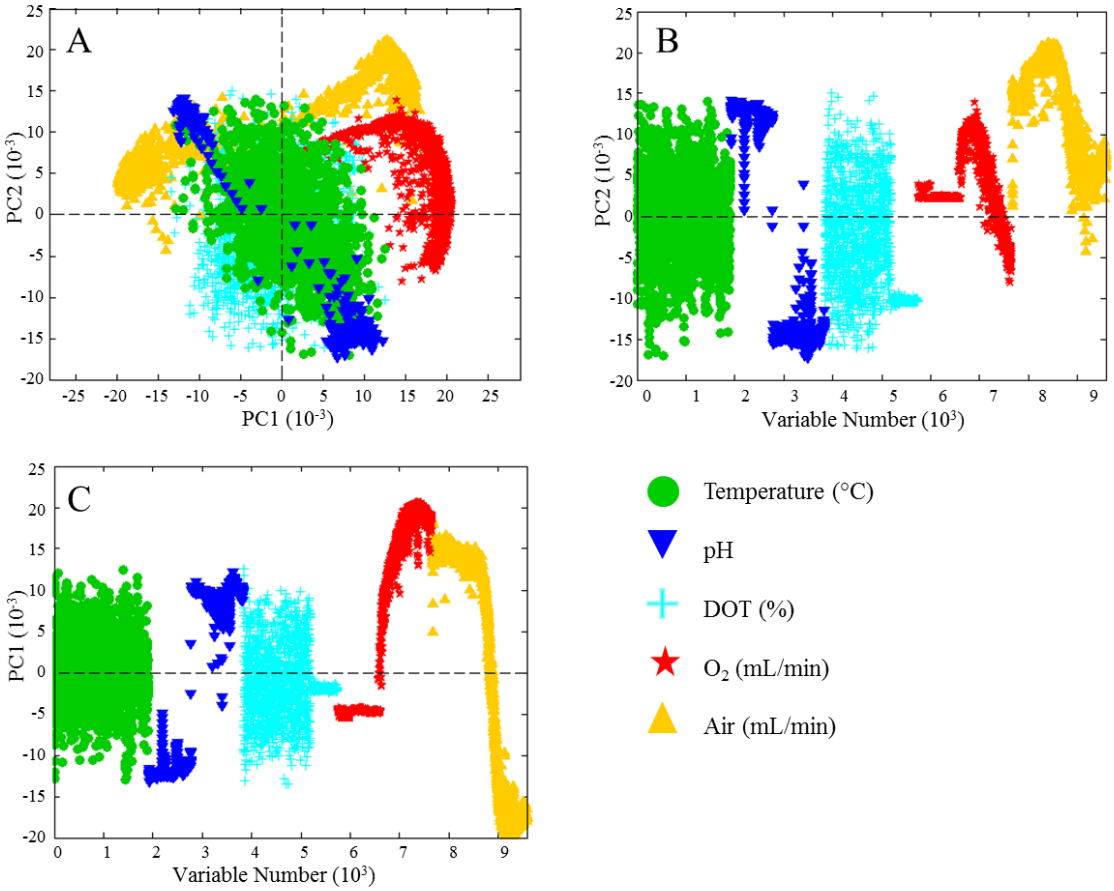


Figure 56. Loading plots for a PC model created from online monitoring data in its native state. The variables can be split into five distinct 'variable blocks' roughly every 2000 variable numbers: temperature (1 to ~2,000), pH (~2,000 to ~4,000), DOT (~4,000 to ~6,000), air (~6,000 to ~8,000), O_2 (~8,000 to ~10,000). Across the three loadings plots shown, interpretation of the captured model was problematic due to the high number of variables. In all three loadings plots, air and O_2 were observed to have a strongly conserved behaviour during the course of a culture, whereas behaviours for the setpoint controlled variables temperature and DOT appeared to capture movement around the setpoint. Behaviour for the setpoint controlled variable pH captured that several cultures underwent a change in setpoint while the majority cultures did not.

A) Loadings plot for PC1 (22.68%) and PC2 (13.43%).
B) Loadings plot for PC2 (13.43%).
C) Loadings plot for PC1 (22.68%).

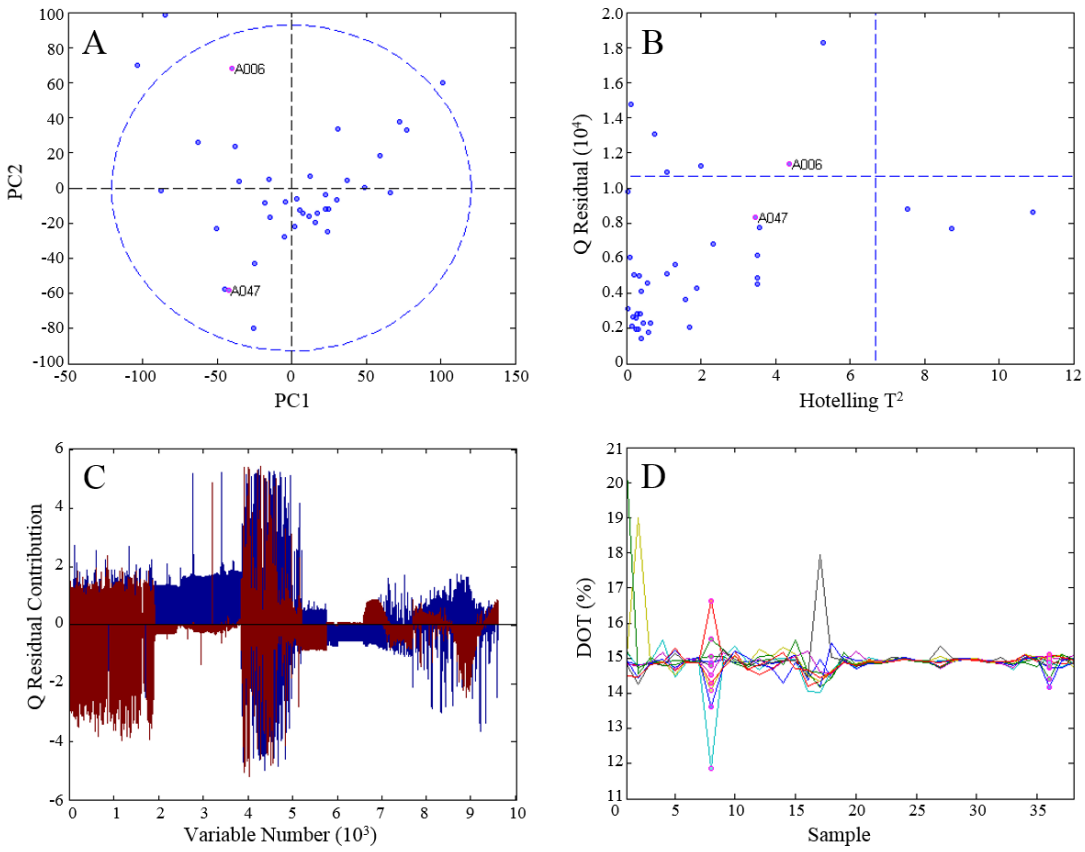


Figure 57. Drill down analysis of cultures A006 and A047 in a PC model created from online monitoring data in its native state.

A) Score plot for PC1 (22.68%) and PC2 (13.43%) with 95% Hotelling's T^2 interval (--).

B) Plot of cultures Q residual (63.89%) against Hotelling T^2 (36.11%).

C) Comparison of Q residual contributions for A006 (blue) and A047 (red). The variables can be split into five distinct 'variable blocks' roughly every 2000 variable numbers: temperature (1 to ~2,000), pH (~2,000 to ~4,000), DOT (~4,000 to ~6,000), air (~6,000 to ~8,000), O_2 (~8,000 to ~10,000). A clear difference was observed for the pH block due to the change in pH setpoint for A006 while the greatest contributions to Q residuals for both A006 and A047 come from the DOT block.

D) The recorded values for DOT for variable numbers 4551 to 4559 indicated greater noise in DOT for A006 (blue) than A047 (red). However it must be stated that the range of DOT values compared here are 12 to 17 for A006 and 14 to 15 for A0047. While statistically distinctive, this result was of little practical use, in part due to lack of context.

Differences in contributions from air and O_2 flowrate variables was also observed. An overlay comparison of the original air and O_2 data (Figure 18) showed three primary differences in air and O_2 behaviour.

1. A006 and A047 had difference caps for maximum air flowrate into the bioreactor, ~2.5 L/min and ~2.0 L/min respectively.
2. A006 reached the air flowrate cap approximately 40 hours earlier than A047.
3. A006 generally showed greater noise in both air and O2 flowrates than A047 across the ~160 hours under consideration.

It should be noted that these conclusions required a return to side-by-side analysis of online monitoring data.

It was also observed that both culture A006 and A047 had high Q Residual contributions from captured DOT behaviours. Drill down analysis to the original values in the dataset (Figure 17D) showed that the cultures appeared to have greater movement around the DOT setpoint than the majority of cultures. Selecting several consecutive readings for DOT indicated a greater range of movement was experience by A006 than A047.

While it was estimated that A006 experienced three times more movement around DOT setpoint than A047 based on these values, this conclusion suffered from being poorly defined and semi-quantitative. As with the comparison of air and O_2 flowrate behaviours, the stated conclusion required a drill down to the original values from online monitoring, then re-interpreting the original values up through the model.

Overall, it was possible to create a PCA model from online monitoring data in its native state. However interpretation of the resulting model was problematic due to both a potentially overwhelming number of variables and a lack of intuitive meaning or context available without resorting to a drill down analysis to the original data measurements. This made effective communication of results difficult. Although some differences could be stated quantitatively, there was a strong reliance on qualitative descriptions.

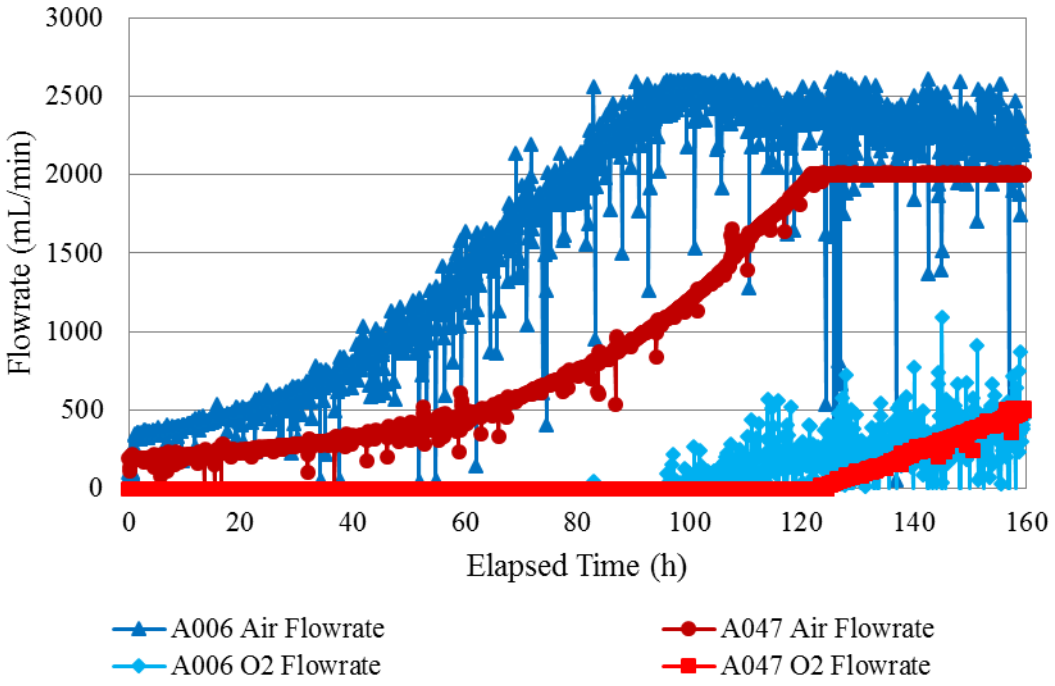


Figure 58. Air flowrate and O_2 flowrate for cultures A006 and A047. There were three main differences between gas profiles for the cultures:

1. A006 shows greater noise than A047 in both gas feeds.
2. A006 had a high air flowrate cap (~2.5 L/min) than for A047 (~2.0 L/min).
3. A006 reached the air flowrate cap approximately 40 hours before A047.

A.4 Analysis of High Frequency Data as Informative Values

In the previous analysis, data were analysed in their native high frequency state. A second analysis was then performed where the data were summarised as informative values and models were created from the resulting informative values instead.

Informative values were developed along two main lines based on variable behaviour type: Steady State and Dynamic. The ability of the selected informative values to capture variable behaviour in a manner with high intuitive meaning and interpretability for a human user was then tested using three datasets:

A.4.1 Informative Values for 'Steady State' Variables

The conversion of online monitoring data to informative values underwent several iterations (Table 35) to meet requirements for ease of calculation, ease of interpretation, robustness, and quality of information.

Initially, the same informative values were used to describe both maintained setpoint and dynamic profiles. Data were split into windows of activity $T_{n,k}$ according to offline sampling times T_1 (inoculation) to T_L (last offline sampling considered). For each window $T_{n,k}$, the mean and standard deviation were calculated for each variable. For daily calculations, $T_{n,k} = T_{n,n+1}$. For calculations over all days, $T_{n,k} = T_{1,L}$.

Two key issues were identified with these initial informative values. One, results could be unusable, e.g. divide by zero results. Two, calculated values were sensitive to noise, outliers (e.g. values resulting from disconnection of a probe), and spiking from chemical additions not indicative of reactor behaviour as a whole.

For example, a window of activity contains 59 readings of pH 6.89 and a single spike to pH 7.38 due to a concentrated bolus addition. The mean is 6.90 and standard deviation 0.06. If the desired setpoint is 6.90 with average movement of ±0.02, the mean does not reflect that the pH was consistently below setpoint and the range of noise appears to be three times the desired range.

If the data were to be summarised by a human, the spike would be dismissed from estimation of the average. Yet the human would still retain the information that a spike occurred. The developed informative values needed to replicate this split into overall information and special detail information.

Iteration	Values for Steady State Variables	Values for Dynamic Variables
1.0	Between Sampling Points: • Mean Average • Standard Deviation • Slope • Coefficient of Determination	Between Sampling Points : • Mean Average • Standard Deviation • Slope • Coefficient of Determination
2.0	Between Sampling Points: • Mean Average • Standard Deviation • Slope • Coefficient of Determination	Key Event Times for Air Key Event Times for O_2
3.0	Between Sampling Points: • Median • Median Absolute Distance	Key Event Times for Air Key Event Times for O_2
4.0	Between Sampling Points: • Median • Median Absolute Distance	Key Event Times for Air Key Event Times for O_2 Between Sampling Points: • Volume of Air • Volumes of O_2
5.0	Between Sampling Points: • Median • Median Absolute Distance	Key Event Times for Air Key Event Times for O_2 Between Sampling Points: • Volume of Air • Volumes of O_2
6.0	Between Sampling Points: • Median • Median Absolute Distance • Area Above Median • Area Below Median • Total Area Away from Median	Key Event Times for Air Key Event Times for O_2 Between Sampling Points: • Volume of Air • Volumes of O_2
7.0	Between Sampling Points: • Median • Median Absolute Distance • Area Above Median • Area Below Median • Total Area Away from Median	Key Event Times for Air Key Event Times for O_2 Key Event Times for CO_2 Between Sampling Points: • Volume of Air • Volumes of O_2 • Volumes of CO_2

Table 35. Development of informative values from first iteration to final seventh iteration.

To address the issues of sensitivity, the overall behaviours of steady state profiles were summarised using robust statistics [184,185]. These have been shown to handle asymmetric profiles, such as spiking, in a robust manner [186]. Mean was replaced with robust equivalent median ($m_{n,k}$), calculated:

$$m_{n,k} = \underset{i=\mathrm{n},\ldots,k}{\mathrm{median}}(x_i) \qquad \text{Eq. 5.1}$$

Standard deviation was replaced by median absolute deviation (MAD), which is calculated:

$$MAD_{n,k} = 1.483 \underset{i=1,\ldots,j}{\mathrm{median}}\left|x_i - m_{n,k}\right| \qquad \text{Eq. 5.2}$$

where x_i is the ith measurement in the window $T_{n,k}$, x_j is the last measurement in the window $T_{n,k}$, and 1.483 is a correction factor to make MAD unbiased at normal distribution [185].

To capture special detail behaviour, the informative value set was expanded to include calculations for the area between the measured value and the median. It was possible to consider areas above and below the median, and the total area away from the median. These areas were calculated using a simple algorithm, whereby when using Eq. 5.3 to calculate the area above median ($AAM_{n,k}$) for a window of activity containing online measurements i to j, samples with $x_i - m_{n,k} < 0$ were replaced with 0. Similarly, when calculating area below median ($ABM_{n,k}$), all samples with $x_i - m_{n,k} > 0$ were replaced with 0 when performing Eq. 5.3.

$$AAM_{n,k} or\ ABM_{n,k} = \sum_{i=2}^{j} (1.5 * \left(x_i - m_{n,k}\right) - 0.5 * (x_{i-1} - m_{n,k}))\ * (s_i - s_{i-1}) \qquad \text{Eq. 5.3}$$

where s_i is the time at which the ith sample measurement is made in hours. Note that the calculation is a backwards looking summation beginning at $s_i = 2$. If the sampling interval is constant, $s_i - s_{i-1}$ can be replaced with s_{int}, yielding:

$$AAM_{n,k} or\ ABM_{n,k}\ = \sum_{i=2}^{j} (1.5 * \left(x_i - m_{n,k}\right) - 0.5 * (x_{i-1} - m_{n,k}))\ * s_{int}) \qquad \text{Eq. 5.4}$$

The total area away from the median ($TAAM_{n,k}$) was calculated by adding the absolute values for $AAM_{n,k}$ and $ABM_{n,k}$ as seen in Eq. 5.5.

$$TAAM_{n,k} = \left|AAM_{n,k}\right| + \left|ABM_{n,k}\right| \quad \text{Eq. 5.4}$$

When calculating the MAD, AAM, ABM, and TAAM, the calculations could be completed using two different median values. These were:

- $m_{n,n+1}$*:* The median value of data captured between two sequential timepoints.
- $m_{1,L}$: The median of data captured between the first and last timepoints.

Use of different medians allowed the identification of culture where limited periods of operation were notably different from overall behaviour, e.g. values for area away from median and MAD calculated using $m_{n,n+1}$ are notably different when compared to the same values calculated using $m_{1,L}$.

Excluding perturbations lasting 6 hours or more (>25% of data points in a 24 hour window), $m_{n,n+1}$ and $m_{1,L}$ for a perturbation should not notably differ. In the event of a shift in setpoint, the daily median will be notably different from the overall median for days on the minority side of the shift, e.g. a culture spends 3 days at pH setpoint A and 12 days at pH setpoint B, assuming no major issues the overall median is B ± noise. Subsequently informative values calculated for the first 3 days using the overall median will be notably different than when using daily medians.

The artificially generated dataset (P001 to P100) is re-used here to demonstrate how these informative values can be used to analyse online monitoring of pH for four cultures. The cultures demonstrate a range of behaviours (described in Table 36 and shown in Figure 55) and lasting 11 days with online monitoring of pH (5 minutes sample interval x 264 h = 3168 measurements per culture).

P001 was an ideal pH profile with random noise ± 0.005 (Figure 55A). P021 had random noise ± 0.005 with a high perturbation from 168 h to 172 h (Figure 55B). P031 had random noise ± 0.010 as well a high perturbation from 43 h to 48 h (Figure 55C). P041 had random noise ± 0.005 and underwent a change in setpoint at 39 h (Figure 55D). For daily calculations, $T_{1,2}$= 1 h and all other $T_{n,n+1}$= 24 h.

The shift in setpoint for P041 was apparent when the median values for each window of activity ($m_{n,n+1}$) were plotted (Figure 3A). The perturbations for P021 and P031 could not be identified at this point.

Calculations for $MAD_{n,n+1}$ were completed for each window of activity using daily and overall medians (Figure 59B and Figure 59C). In Figure 59B, it can be seen that similar values were calculated for P001 and P021, which showed normal noise. For P031, MAD were higher than that of P001 or P021.

For P041, values $MAD_{n,n+1}$ calculated when using $m_{n,n+1}$ were of a similar magnitude as for P001 and P021. However the shift in P041 was clearly identifiable when $MAD_{n,n+1}$ was calculated when using $m_{1,L}$(Figure 59C). This was the only culture where replacing $m_{n,n+1}$ in the calculation with $m_{1,L}$caused such a change in $MAD_{n,n+1}$.

While median and MAD calculations could be used to differentiate between high and low noise as well as normal and shift behaviours, it was not possible to differentiate between normal and perturbed behaviours.

To differentiate between normal and perturbed behaviour, $AAM_{n,k}$, $ABM_{n,k}$, and $TAAM_{n,k}$ were used. It can be seen in Figure 3D that there was an increase in $TAAM_{n,n+1}$ for P021 for $T_{7,8}$ (168 h to 192 h) and P031 for $T_{2,3}$ (24 h to 48 h). This increase was seen when both $m_{n,n+1}$ and $m_{1,L}$ are used, indicating that behaviour in these windows was unusual.

Shift behaviour could also be identified using $AAM_{n,k}$, $ABM_{n,k}$, and $TAAM_{n,k}$. When using $m_{n,n+1}$, a higher area was seen for P041 in the window $T_{2,3}$ (24 h to 48 h). This value captured the shift in pH setpoint which occurred at 39 h. In Figure 59E where $m_{1,L}$ was used, all area values calculated prior to 48 h are increased, indicating pre-shift and post-shift data.

Culture	**pH Behaviour**	**Noise**	**Noise Range**	**Event Time**
P001	Ideal	Acceptable	Setpoint ± 0.005	N/A
P021	High Perturbation	Acceptable	Setpoint ± 0.005	168 h to 172 h
P031	High Perturbation	High	Setpoint ± 0.010	43 h to 48 h
P041	Setpoint Shift	Acceptable	Setpoint ± 0.005	39h

Table 36. Summary of pH profiles compared.

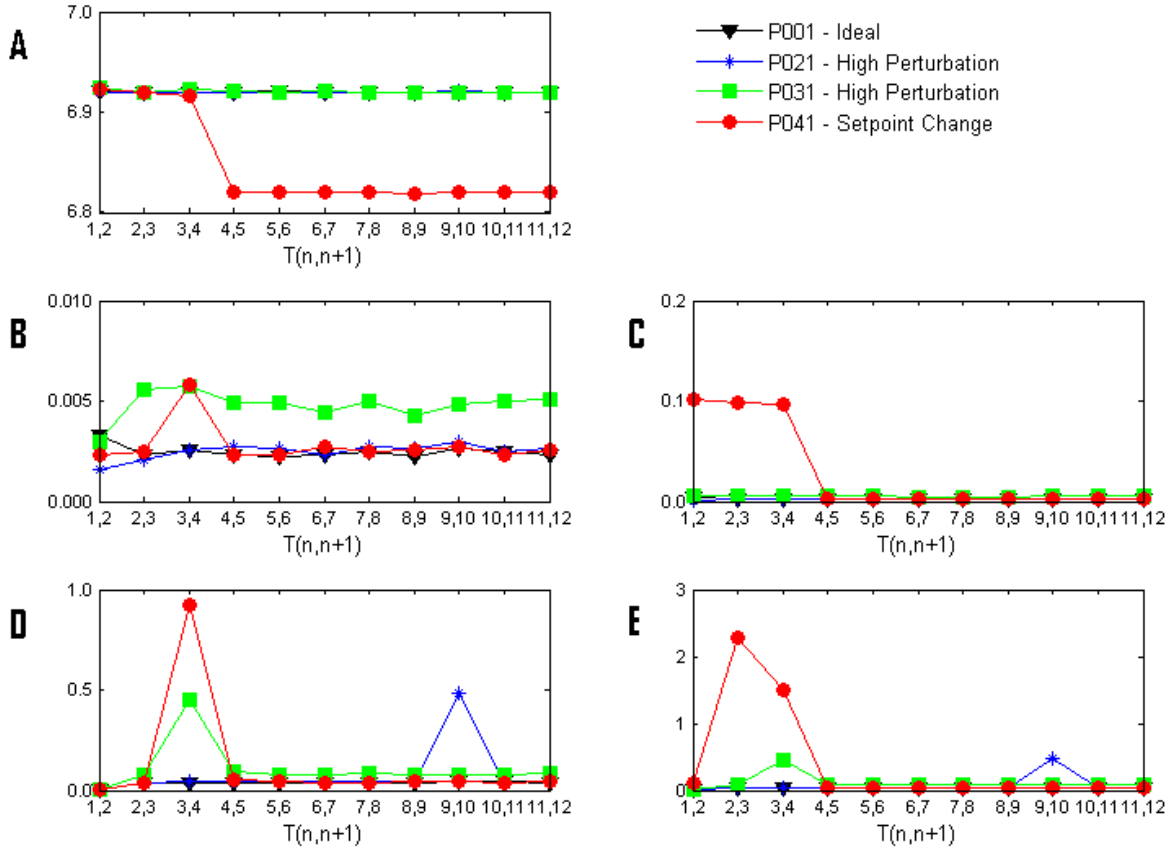


Figure 59. Informative values calculated from artificially generated online monitoring data of pH for four theoretical cultures showing various behaviours and noise. P001: ideal, setpoint ± 0.005. P021: high perturbation, setpoint ± 0.005. P031: high perturbation, setpoint ± 0.010. P041: shift in setpoint, setpoint ± 0.005.

A) $m_{n,n+1}$
B) $MAD_{n,n+1}$ using $m_{n,n+1}$
C) $MAD_{n,n+1}$using $m_{1,L}$
D) $TAAM_{n,n+1}$ using $m_{n,n+1}$
E) $TAAM_{n,n+1}$ using $m_{1,L}$

A.4.2 Informative Values for 'Dynamic' Variables

Dynamic variables are variables which are strongly dependent on culture performance, e.g. the flowrate of O2 into the bioreactor is determined by culture oxygen demands whereas pH level is affected by culture behaviour but is primarily dependent on pre-set operating setting points. Dynamic profiles were summarised using volumes of gas entering the bioreactor between sampling timepoints and when key events occurred. The key events for air, O_2, and CO_2 feeds are demonstrated with references in Figure 60. When the fermentation begins, O_2 is provided through the air feed, which is increased until a capped value is reached (A). Once the air feed reaches the capped value, the O_2 feed is activated to meet any further O_2 demands (B).

The O_2 flowrate increases until the viable cell concentration reaches a maximum. At this point the O_2 feed is at a peak value (C). As the number of viable cells in the culture decreases, the O_2 demand also decreases. The O_2 feed is reduced until it becomes effectively zero (D). At this point, the air feed begins to decrease (E).

In some control arrangements, O_2 and air could be considered a single mixed feed. As stated, air is increased to meet increasing demand for O_2 until a capped value is reached. Further demands for O_2 are met by increasing the proportion of O_2 in the feed. The air feed is decreased so that the flowrate remains at the capped value. This creates a distinctive 'dip' in the air flow (F).

In other controllers, air and O_2 could be considered separate feeds. The air feed was maintained at the capped flowrate while the O_2 flowrate increased and declined as necessary. Several additional events may be noted of interest: when the air feed reaches half of the capped flowrate (G), the time between the air flowrate capping and the O_2 feed activating, the time the air feed was at the capped flowrate, the time the O_2 feed was active, and the peak O_2 flowrate.

Due to the concentrated nature of feeds and bolus used, these additions could sometimes be identified by analysing the profiles of the variables intended to control pH, such as the CO_2 profile. CO_2 gas is an acidic gas used to correct pH in cultures when pH measured > pH setpoint by sparging. In Figure 60, distinct jumps in CO_2 flowrate can be seen at H, I, and J.

A.4.3 Method

Online monitoring data from 49 cultures performed in 10 L and 130 L bioreactors at Lonza's Slough site were summarised as informative values. Informative values were calculated from inoculation to Day 10 using sampling times recorded in daily offline monitoring data.

For fairer comparability with the previous analysis of data in its native state, 11 cultures were removed including all 130 L cultures. Due to mixed O_2 gassing strategies and disparity in observations of key events for dynamic variables, informative values for air and O_2 flowrates were restricted to volumes added between sampling timepoints. All informative values based on CO_2 flowrate were removed as variables due to data loss and data validity issues related to transfer of data from datalogger to computer.

Data were imported to Matlab for analysis with the Eigenvector PLS-Toolbox. Data were mean-centred and scaled to unit variance. A PCA model was created using random sampling (10 splits, 5 iterations) for cross-validation. A one PC model was recommended based on lowest RMSE during cross-validation. As in the previous analysis, two PCs were retained. The model captured 26.24% of variance in the dataset.

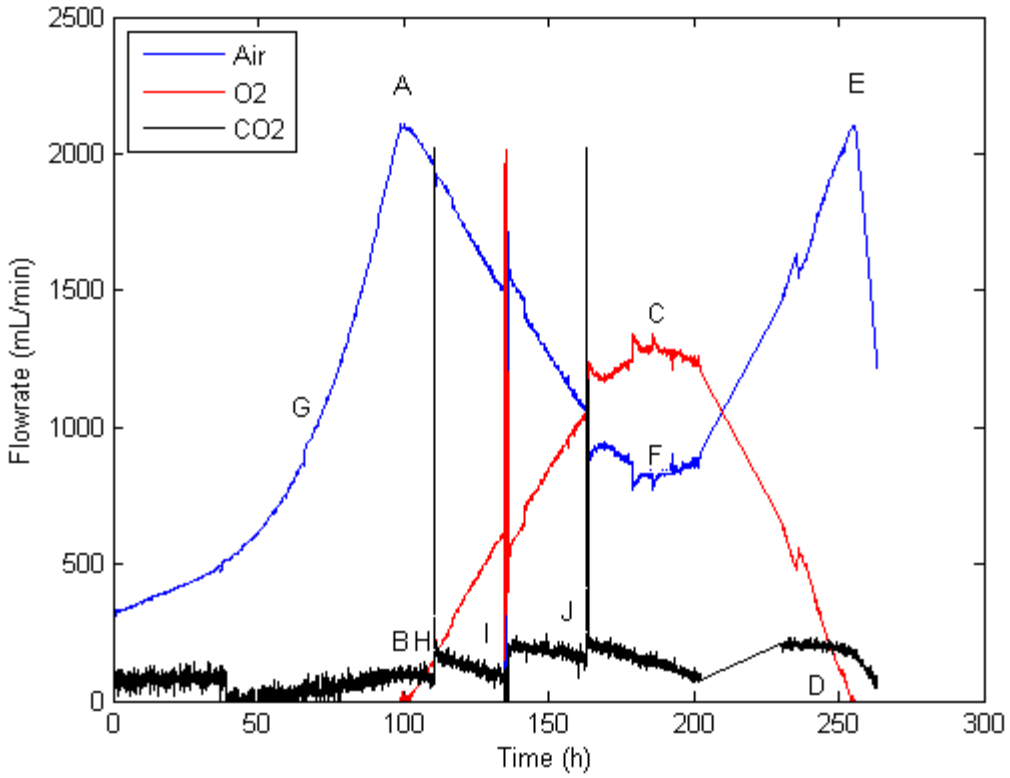


Figure 60. Sample profiles from process data from the online monitoring of air, O_2, and CO_2 flowrates for bioreactor culture A023. Key events described in the text are annotated A—J.

The informative values dataset was then extended to include the elapsed time at which sampling occurred, which was used when calculating the informative values. This extended dataset was mean-centred and scaled to unit variance. A PCA model was created using random sampling (10 splits, 5 iterations) for cross-validation. A one PC model was recommended based on lowest RMSE during cross-validation. As a single PC model would be difficult to visualise and interpret, two PCs were retained. The model captured 27.14% of variance in the dataset.

A.4.4 Results and Discussion

As there was less than a 1% difference in variance captured by the two models created, the second model, where offline sampling times were included in the modelled dataset, was excluded from further analysis.

Figure 61 and Figure 62 show the loadings for the final model. In Figure 61, the variables were coloured by the original variable from which the informative values were calculated (pH, temperature, DOT, air flowrate, O_2 flowrate). In Figure 62, the variables were coloured by the type of informative value calculated.

From the loading plots, it was concluded that the greatest sources of variation between cultures in the modelled dataset were DOT-based informative values. Two clusters of DOT-based informative values were observed in the upper-left and lower-right quadrants of Figure 61A. The upper-left quadrant cluster was comprised predominantly of three types of informative values describing DOT behaviour:

1. Area Below Median (Using Daily Median)
2. Area Below Median (Using Overall Median)
3. Daily Median

The lower-right quadrant comprised of six types of informative values describing DOT behaviour:

1. Area Above Median (Using Daily Median)
2. Area Above Median (Using Overall Median
3. Total Area Away from Median (Using Daily Median)
4. Total Area Away from Median (Using Overall Median)
5. MAD (Using Daily Median)
6. MAD (Using Overall Median)

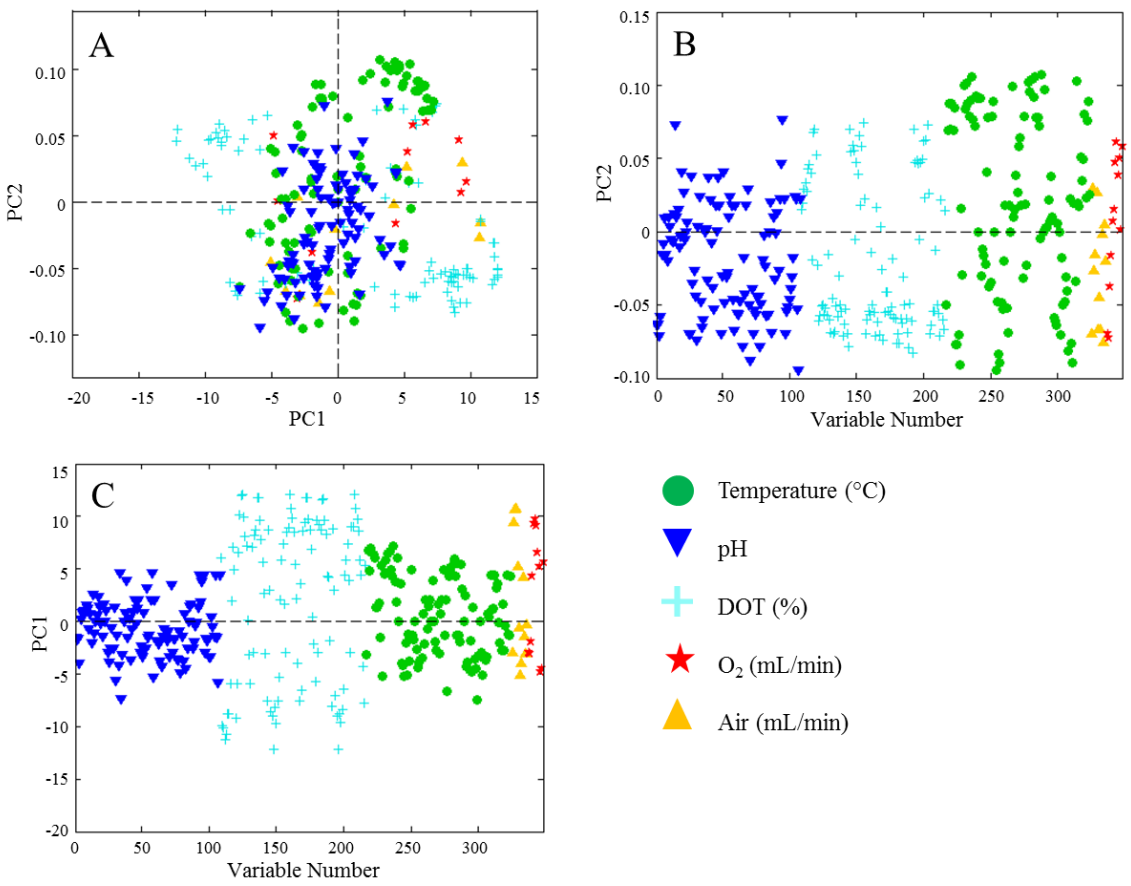


Figure 61. Loading plots for a PC model created from informative values calculated from online monitoring data. Variables were coloured by the original variable from which the new informative value variables were calculated. Note that Figure 61B and Figure 61C were arranged to match the axes of Figure 61A.

A) Loadings for PC1 and PC2. Two clusters of informative values calculated from DOT measurements were seen in the upper-left and lower-right quadrants.

B) Loadings for PC2. PC2 behaviour appeared to marginally greater defined by temperature behaviours. As PC2 was retained primarily for improved visualisation of PC1, the lack of clearly structured behaviour captured by the PC was expected.

C) Loadings for PC1. Informative values calculated from DOT had the greatest impact on overall behaviour captured in PC1. Air flowrate and O_2 flowrate had a greater impact of captured PC1 behaviour from inoculation to Day 3 and from Day 4 to Day 6, respectively, than on other days during cultures.

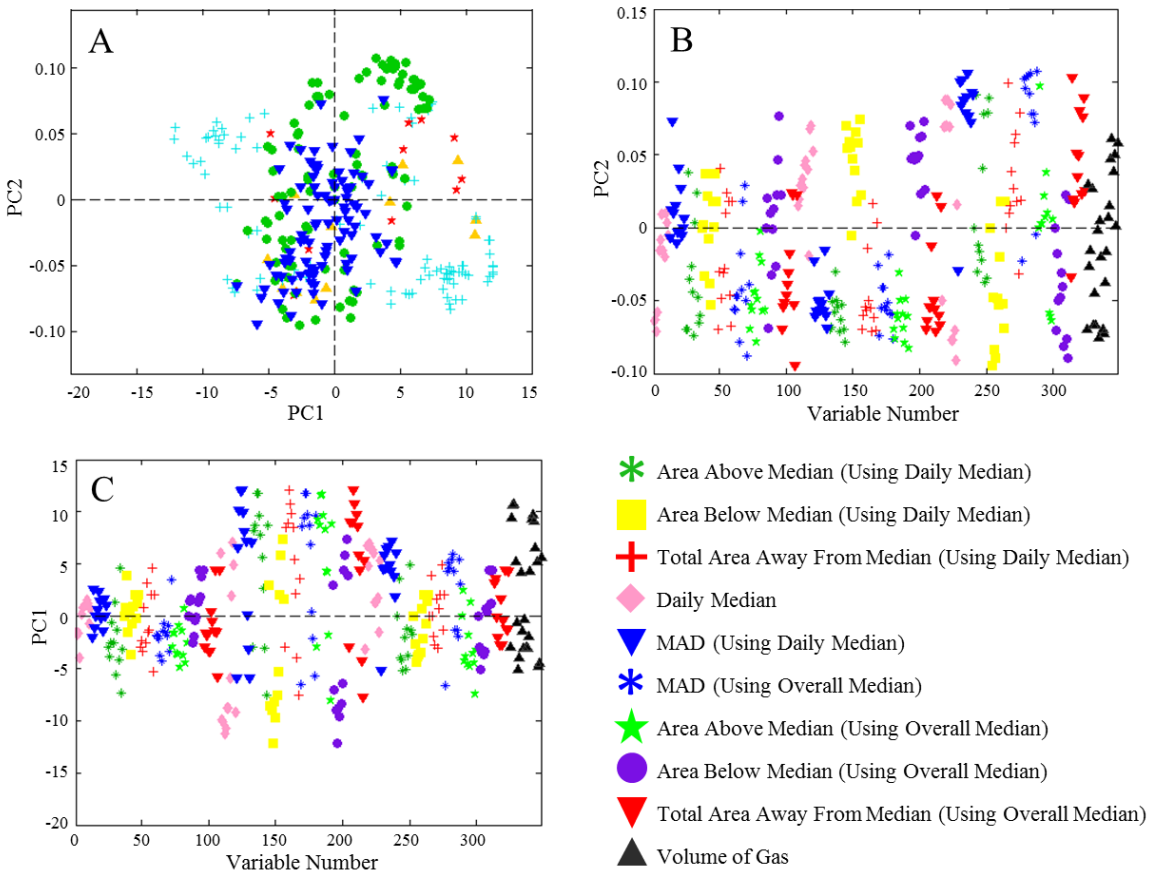


Figure 62. Loading plots for a PC model created from informative values calculated from online monitoring data. Variables were coloured by the type of informative value calculated. Note that Figure 62B and Figure 62C are arranged to match the axes of Figure 62A.

A) Loadings for PC1 and PC2. In Figure 61A, two clusters of informative values calculated from DOT measurements were observed in the upper-left and lower-right quadrants. Here it was observed that the cluster in the upper-left quadrant comprised of three main informative values for DOT: Area Below Median (Using Daily Median), Area Below Median (Using Overall Median), and Daily Median. The cluster in the lower-right quadrant comprised of Area Above Median (Using Daily Median), Area Above Median (Using Overall Median), Total Area Away from Median (Using Daily Median), Total Area Away from Median (Using Overall Median), MAD (Using Daily Median), and MAD (Using Overall Median).

B) Loadings for PC2. As PC2 was retained primarily for improved visualisation of PC1, a lack of clearly structured behaviour captured by the PC was expected. However, strong clusters for both MAD-based informative values for temperature were observed.

C) Loadings for PC1. Informative values calculated from DOT had the greatest impact on overall variance captured in PC1.

Air flowrate and O_2 flowrate had a greater impact of captured PC1 behaviour from inoculation to Day 3 and from Day 4 to Day 6, respectively, than on other days during cultures. This was in keeping with what was known about air and O_2 control strategies, i.e. the use of air flowrate cap and a supplementary O_2 feed. From this, it can be surmised that the activation of the O_2 feed on average occurred around Day 3/Day 4.

Overall, the loading plot analysis demonstrated that informative values had captured expert knowledge and process understanding in a form appropriate for MVDA in keeping with the stated project aims.

Figure 63 shows the scores for the final model. It was seen in Figure 63B that the cultures considered in the previous analysis with data in its native state, A006 and A47, both lie within the Q Residual limit and Hotelling T^2 limit. A006 and A47 showed a greater difference in Hotelling T^2 residual values than Q Residual value.

A Hotelling T^2 contribution analysis (Figure 63C) indicated that the majority of the difference in Hotelling T^2 value was due to informative values summarising DOT behaviour. From this contribution analysis, it appeared that A006 had greater movement around the DOT setpoint, approximately 5 to 6 times greater movement in terms of area-based informative values and MAD. As these differences were observed when using both daily median DOT and the median DOT across the full culture duration under consideration, it was known that this 5 to 6 fold increase in movement was sustained through the majority of culture duration. For completeness, this was confirmed by a further drill down to the calculated informative values (Figure 63D).

Additional differences observed in Figure 63C were the consistently higher contributions from air and O_2 volumes for A006 than for A047. This indicated that higher volumes of air and O_2 entered A006 than A047 throughout the majority of the culture duration under consideration.

In comparison to the analysis of online monitoring in its native state, a lower percentage of dataset variance was captured when the same number of variables were retained (36.11% v. 26.29%). However use of informative values reduced in the number of variables to be analysed from 9580 to 360. This in addition to the context-rich nature of the informative values used led to improved identification of behaviours of interest and quantitative communication of those differences.

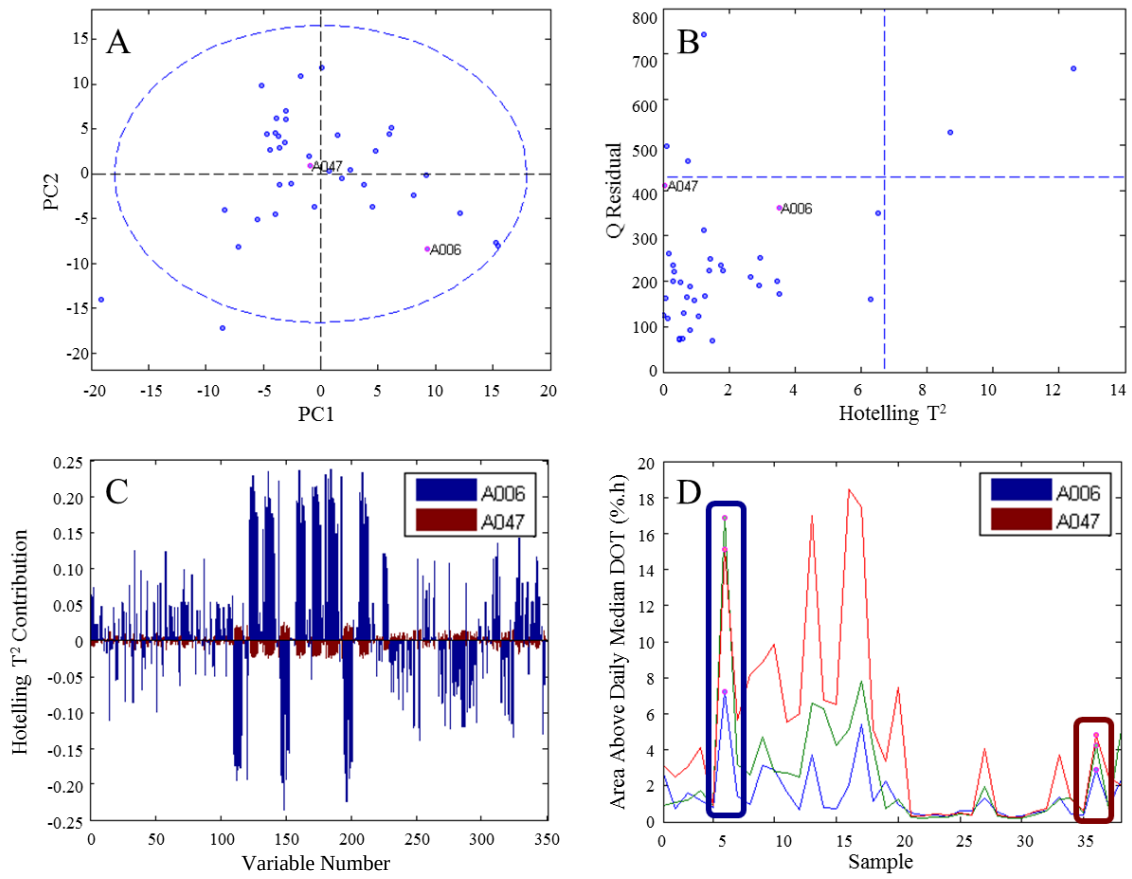


Figure 63. Drill down analysis of cultures A006 and A047 in a PC model created from informative values calculated from online monitoring data.

A) Score plot for PC1 (14.17%) and PC2 (12.07%) with 95% Hotelling T^2 limit (-).

B) Plot of values for Q residual (73.76%) against values for Hotelling T^2 (26.29%).

C) Comparison of Hotelling T^2 contributions for A006 (blue) and A047 (red). The variables can be split into five 'variable blocks': pH (1 to 108), DOT (109 to 216), temperature (217 to 324), air (325 to 335), O_2 (336 to 348). The greatest difference in contributions to Hotelling T^2 appear to come from the DOT block.

D) The calculated value for Area Above Daily Median for the sampling intervals Day 1 to Day 2, Day 2 to Day 3, Day 3 to Day 4 indicated greater degree of movement and a greater magnitude of movement around the daily median DOT for A006 than A047 over the 72 hours captured. From Day 2 to Day 4, the movement above the daily median for culture A006 was approximately three times that of culture A047.

A.5 Additional Demonstration of Informative Values

The previously created models were very comprehensive analyses of behaviour captured in online monitoring data, in that as they included many variables of interest (pH, temperature, DOT, etc.). While MVDA techniques can be used to create such comprehensive models, it may be desired to create simpler, more focussed models for a variety of reasons, e.g. variable-specific models for simpler interpretation. Such models and subsequent interpretation could also benefit through the use of informative values.

Temperature-based informative values were calculated from inoculation to Day 10 using the sampling times recorded in daily offline monitoring data. A total of 44 cultures were included in the dataset, including two 130 L cultures, A043 and A046.

Data were imported to Matlab for analysis with the Eigenvector PLS-Toolbox. Data were mean-centred and scaled to unit variance. A PCA model was created using random sampling (10 splits, 5 iterations) for cross-validation. A two PC model was retained that captured 60.68% of variance in the dataset.

It was seen that two cultures were outside the 95% confidence interval for PC1. (Figure 64A). These cultures were 130 L cultures – A043 and A046. Hotelling T^2 and Q Residual values were also above the calculated limits for the model (Figure 64B). It was decided to focus on the behaviour of A043 for the purpose of this demonstration.

Hotelling T^2 contribution analysis of A043 (Figure 64C) did not indicate any particularly unusual values compared to other cultures in the dataset. All values were within one standard deviation of dataset means, hence the high Hotelling T^2 calculated for A043 appears to be due to a cumulative effect. Q Residual contribution analysis for A043 (Figure 64D) revealed the separation was caused by unusually high contributions from the informative values listed in Table 37.

Interval for Informative Value	Informative Value
Day 2 to Day 3	Area Below Median Using Daily Median
	Area Below Median Using Overall Median
Day 3 to Day 4	Area Above Median Using Daily Median
	Area Below Median Using Daily Median
	Total Area Away from Median Using Daily Median
	Area Below Median Using Overall Median
	Total Area Away from Median Using Overall Median
Day 5 to Day 6	MAD Using Daily Median

Table 37. Informative values of interested identified through Q Residual contribution analysis.

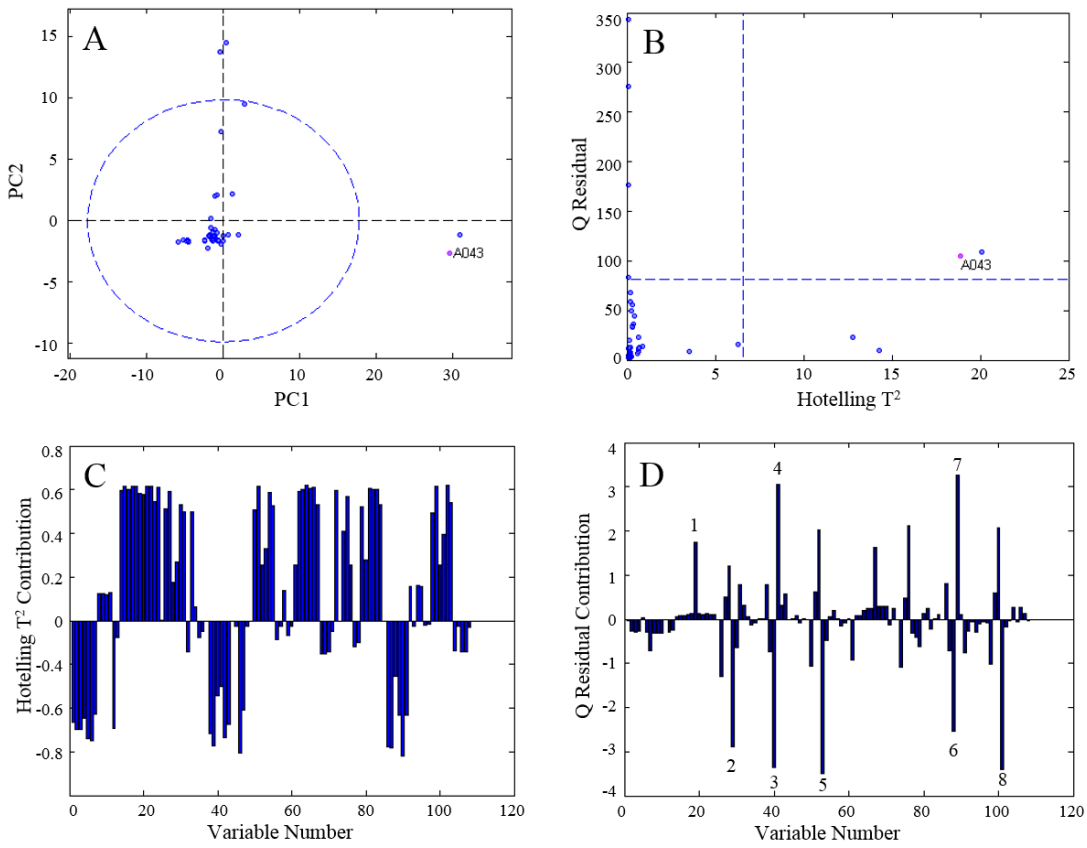


Figure 64. Results for a PCA model created using temperature-based informative value and drill down analysis for the indicated culture A043.

A) Score plot for PC1 (46.31%) and PC2 (14.37%).

B) Culture values for Hotelling T^2 (60.68%) and Q Residual (39.32%) with calculated limits. It was observed that A043 had unusual Hotelling T^2 and Q Residual values above the displayed limits.

C) Hotelling T^2 contribution analysis for culture A043. No informative values appeared to be of particular note as all informative values were within 1 standard deviation of the dataset mean.

D) Q Residual contribution analysis for culture A043. Eight informative values of interested are indicated with the numerals.
   1. MAD Using Daily Median for Day 5 to Day 6
   2. Area Above Median Using Daily Median for Day 3 to Day 4
   3. Total Area Away from Median Using Daily Median for Day 2 to Day 3
   4. Area Below Median Using Daily Median for Day 3 to Day 4
   5. Total Area Away from Median Using Daily Median for Day 3 to Day 4
   6. Area Below Median Using Overall Median for Day 2 to Day 3
   7. Area Below Median Using Overall Median for Day 3 to Day 4
   8. Total Area Away from Median Using Overall Median for Day 3 to Day 4

Informative values of interested identified through Q Residual contribution analysis described behaviour for three offline sampling intervals: Day 2 to Day 3, Day 3 to Day 4, and Day 5 to Day 6.

Unusual informative values for the offline sampling interval Day 2 to Day 3 were Area Below Median and the Total Area Away from Median when calculations were completed using either the median temperature for the sampling interval or the median temperature from inoculation to the Day 10 offline sampling. From this, it was known that an event had occurred with the following conditions:

1. The event was restricted to the ~24 hours in question.
2. The event did not last long enough to affect the median temperature or the median absolute distance for temperature for the ~24 hour block.
3. The event increased area-based informative values in one direction only.

From these conclusions, it was suggested that the recorded temperature measurement had dropped drastically for a short time. Referring back to the original data (Figure 65), it was revealed that errors had occurred in several readings which lead to the informative value calculator treating the readings as zero. Repeating the contribution analysis on A046 showed a similar issue, indicating a recurring equipment fault for the 130L bioreactor control system. This was confirmed through discussion with members of the UK pilot team.

This form of error can be easily hidden during visual analysis of raw data. When plotting the variable against time, missing data could be overlooked due to the graphing program rules in use, e.g. hold last known value or create a straight line connection to the next available value, or gaps being too small to notice amongst the 1000s of points. Through the use of informative values, this error was quickly captured and identified.

The second interval of interest was Day 3 to Day 4. All area-based informative values were unusual, except for the Area Above Median from when the median temperature from inoculation to the Day 10 offline sampling was used to complete the calculation. Referring back to the original data showed generally lower readings for temperature during this interval however the typical movement around the median (measured by MAD) was relatively unaffected. While the general decrease was relatively small, it was sufficient to capture that the interval was unusual compared to activity of other days and draw the user's attention to the transition.

The third interval of interest was Day 5 to Day 6. This interval contained the end of the general decrease in temperature. This 'return to normal' behaviour was captured in the MAD when using Daily Median.

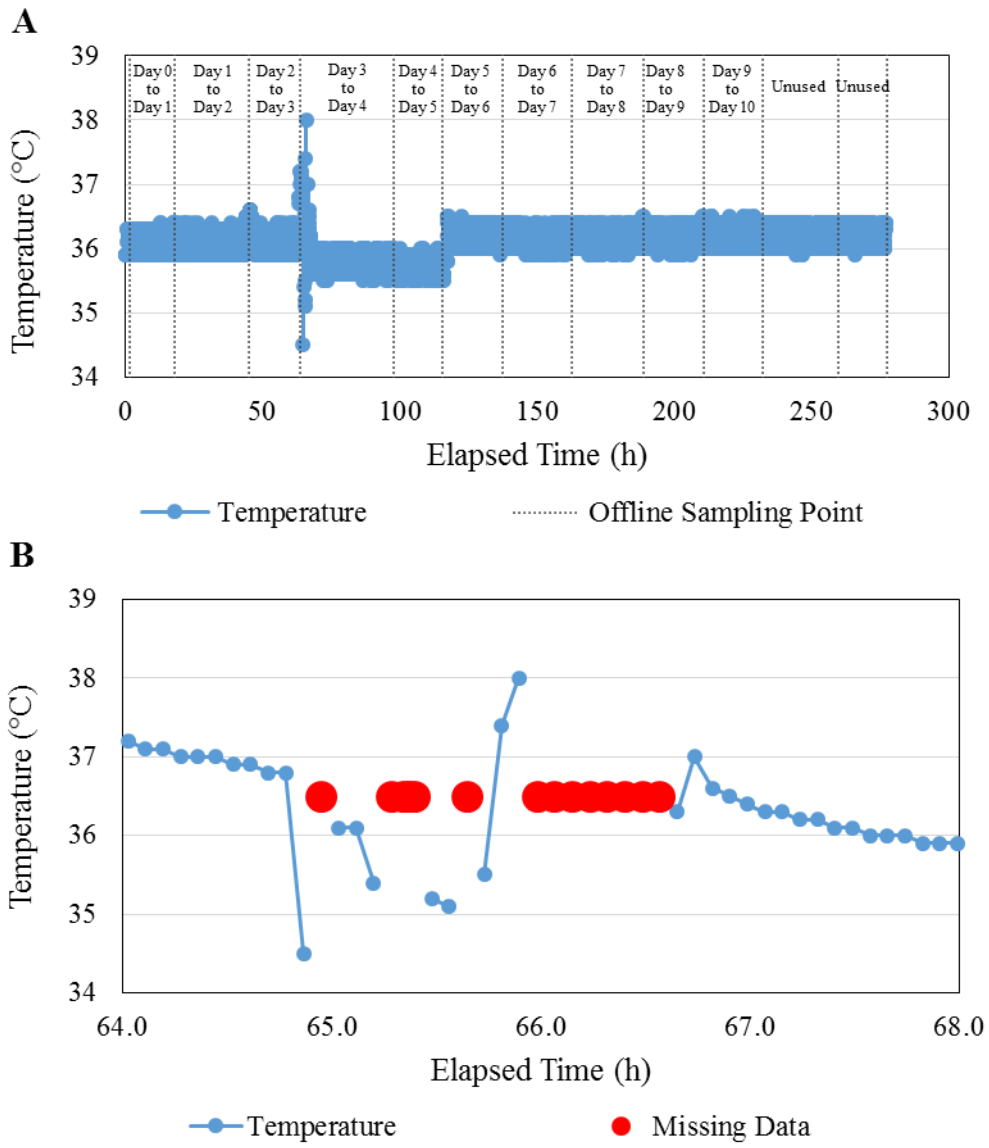


Figure 65. Original temperature measurements for online monitoring of culture A043. Figure 65A shows all measurements and it was observed that a shift in overall temperature measurements occurred and lasted approximately two days. Closer examination revealed missing data points (indicated by red markers at the temperature setpoint 36.5°C). Figure 65B shows a magnified view of measurements used to calculated informative values for the interval Day 2 to Day 3. When calculating informative values, these missing data points were treated as zero readings by the Excel-based calculator. This caused the increased values for Area Below Median and Total Area Away from Median, both when using the Daily Median and the Overall Median for culture temperature. Area Above Median was effectively unaffected.

A.6 Conclusions

The control systems used in bioreactors are dependent on measurements from online monitoring systems. These systems can create thousands of data points per variable in a matter of days. However the time series created are frequently analysed qualitatively and subjectively on an individual basis.

While statistical methods exist that can analyse high frequency online monitoring data with little to no human intervention, such analyses may be of limited use when considering non-major deviations and may prevent batch-to-batch comparisons of the data in its native form. Furthermore, it may benefit research and development or manufacturing departments to prioritise human interpretability over pure statistical power.

Online monitoring data were converted from univariate time series to informative values that provide a balance between both quantitative and qualitative understanding and possess intuitive meaning for users. The developed informative values were used to differentiate between common online monitoring data behaviours encountered during cultures of mAb-producing cells. This was achieved through the analysis of an online monitoring dataset, where online monitoring data were first summarised with informative values and then analysed using PCA. Behaviours in online monitoring data were adequately captured to be identifiable in the resulting PCA model. Interpretability during contribution analysis was increased as the informative values selected were developed to have intuitive and appropriate contextual meaning for human understanding. These conclusions were supported by an additional demonstration using a PCA model to analyse only temperature-based informative values.

The informative values presented were developed specifically for online monitoring data originating from the production of a mAb by a mammalian cell line and influenced by the verbal descriptions provided by scientists familiar with the process. Informative values could be developed for other high frequency process variables using similar logic.

The third aim of the presented research was to enable the interrogation of online monitoring and offline monitoring datasets in a single, balanced dataset. Achievement of this aim was tested through the productivity and culture viability investigations described in Chapter 5 and Chapter 6 respectively.

Appendix B. Additional Tables and Figures

Unique Code	[SPE] (mL/L)
ID	Vessel number
Cell	Inoc date
Cell line	Inoc time
Experiment ID	Post- Inoc volume (L)
Experimental Conditions	Stage and Round ID
Process	Working Cell Bank

Table 38. Routine Meta Data Collected in Project A.

Elapsed time (h)	Glu (g/L)
Elapsed time (days)	Gluc (g/L)
Bioreactor volume(L)	Lac (g/L)
Temp(°C)	NH_4^+ (g/L)
Bench pH	Na^+ (mmol/L)
DOT (%)	K^+ (mmol/L)
VCC (106/mL)	Osmolality (mOsm/Kg)
TCC (106/mL)	SF66 (g)
IVC (106 cell h/mL)	Gluc (g)
Product (mg/L)	SF66 (g/109 cell.hour)
Nova pH	Gluc Utilisation (g/10^9 cell.h)
pO2 (mmHg)	Viability (%)
pCO2 (mmHg)	Specific growth rate (h-1)
Gln (g/L)	Doubling time (h)

Table 39. Routine Daily Monitoring Data Collected in Project A.

d[Gluc]/d[Lac]	Osmolality - Theoretical Osmo (mOsm/kg)
d[K]/d[Na^+]	[K^+]/[Na^+] (mmol/mmol)
[Na^+] /pCO2 (mmol/mmHg)	d[Gln]/d[Lac]
[Na^+]/[Lac] (mmol/g)	Lac Production Rate (g/10^9 cell.h)
d[Na^+]/d[Lac] (mmol/g)	Antibody Accumulation Rate (mg/10^9 cell.h)
[Lac]/[NH_4^+] (g/g)	NH_4^+ Accumulation Rate (g/10^9 cell.h)
d[Lac]/d[NH_4^+] (g/g)	Glu Utilisation Rate (g/10^9 cell.h)
[Gln]/[NH_4^+] (g/g)	K^+ Utilisation (mmol/10^9 cell.h)
d[Gln]/d[NH_4^+] (g/g)	Na^+ Accumulation Rate (g/10^9 cell.h)
Theoretical Osmo (mOsm/kg)	Bench pH - Nova pH

Table 40. Additional Ratios and Rates Calculated for Stage 1 Extended DM Dataset.

Setpoint Variables	Dynamic Control Variables
pH Gradient	Air Gradient
pH R^2	Air R^2
pH Mean Average	Air Mean Average
pH StDev	Air StDev
DOT Gradient	O_2 Gradient
DOT R^2	O_2 R^2
DOT Mean Average	O_2 Mean Average
DOT StDev	O_2 StDev
Temp Gradient	CO_2 Gradient
Temp R^2	CO_2 R^2
Temp Mean Average	CO_2 Mean Average
Temp StDev	CO_2 StDev
N_2 Gradient	
N_2 R^2	
N_2 Mean Average	
N_s StDev	

Table 41. Informative Values Version 1.0 Used in Stage 1.

Setpoint Variables	Dynamic Control Variables
pH Median	Volume of Air Added
pH MAD	Volume of Oxygen Added
Area Above pH Median	True Volume of Oxygen Added
Area Below pH Median	Volume CO_2 Added
Total Area Away from pH Median	
Temp Median	
Temp MAD	
Area Above Temp Median	
Area Below Temp Median	
Total Area Away from Temp Median	
DO Median	
DO MAD	
Area Above DO Median	
Area Below DO Median	
Total Area Away from DO Median	

Table 42. Subset of Informative Values Version 7.0 Used in Stage 2.

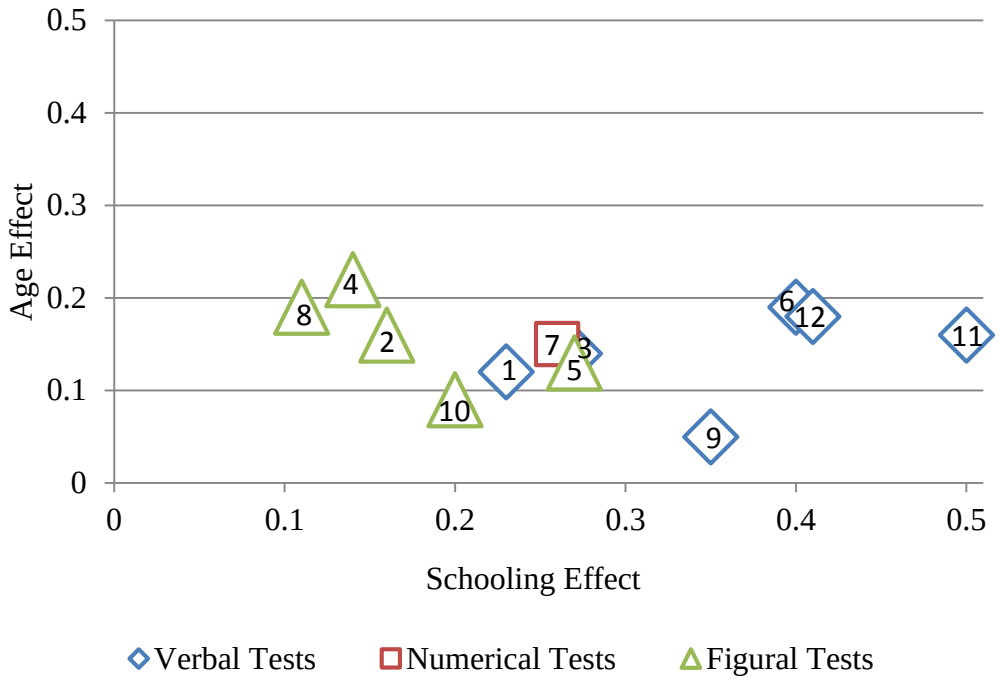


Figure 66. Comparison of schooling and age effects on performance in different tests to evaluate intelligence in students in grades 4 to 6 [164].

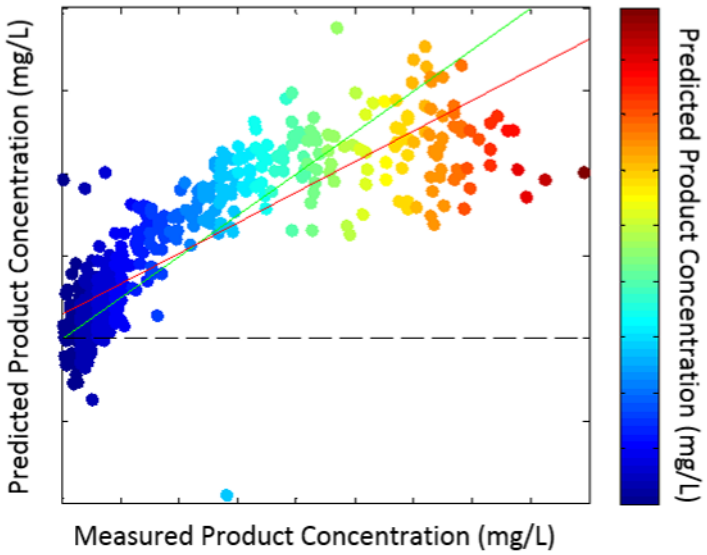
		Estimated Net Effect of 1 Year of		
Test Number and Name		**Age (A)**	**Schooling (B)**	**B/A (C)**
Verbal Tests				
1	Verbal Classification	0.12	0.23	1.9
3	Verbal Analogies	0.14	0.27	1.9
6	Vocabulary	0.19	0.40	2.1
9	Verbal Oddities	0.05	0.35	7.0
11	Arithmetic Problems	0.16	0.50	3.1
12	Sentence Completion	0.18	0.41	2.3
Numerical Tests				
7	Number Series	0.15	0.26	1.7
Figural Tests				
2	Figure Classification	0.16	0.16	1.0
4	Figure Analogies	0.22	0.14	0.6
5	Matrices	0.13	0.27	2.1
8	Figure Series	0.19	0.11	0.6
10	Figural Oddities	0.09	0.20	2.2

Table 43. Estimated effects of age and schooling on grade 4 to grade 6 student performance in standardised intelligence exams [164].

Factor	Correlation with Day 11 [Product]		
	Measurement	**All Cultures**	**Control Cultures**
Reaches Half Air Cap	Elapsed Time (h)	-0.25	-0.38
Reaches Half Air Cap	IVC	-0.11	-0.32
Reaches Air Cap	Elapsed Time (h)	-0.37	-0.50
Reaches Air Cap	IVC	-0.21	-0.52
Ratio Cap to Half Cap	Elapsed Time (h)	-0.17	-0.14
Ratio Cap to Half Cap	IVC	-0.22	-0.64
Oxygen Feed On	Elapsed Time (h)	-0.44	-0.57
Oxygen Feed On	IVC	-0.25	-0.79
Oxygen Feed Peaks	Elapsed Time (h)	-0.12	-0.40
Oxygen Feed Peaks	IVC	0.23	-0.16

Table 44. Correlations between Day 11 [Product] (mg/L) and events in air and oxygen profiles using Elapsed Time (h) and IVC as measures of progress.

No Progression Variable
Developed 04-Dec-2012 11:54:30.09
X-block: Project A Day by Day 475 by 32
Num. LVs: 6
CV: random samples w/ 10 splits 10 iterations



Model Performance	Cal	CV
RMSE	107.68	121.185
Bias	-8.83915e-012	1.11864
R^2	0.762318	0.713459

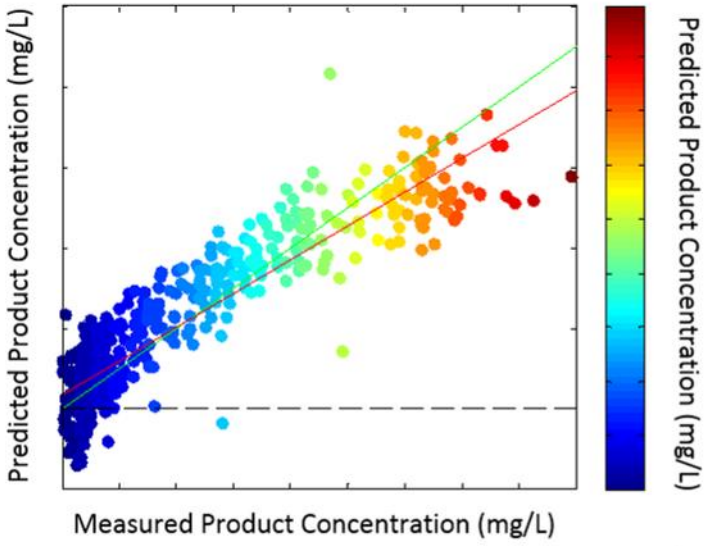
Variables		
Temp (C)	Volume of Air Added	Temp Median
Radiometer pH	Volume of Oxygen Added	Temp MAD
DOT %	True Volume of Oxygen Added	Area Above Temp Median
Nova pH	Volume CO2 Added	Area Below Temp Median
pO2 (mmHg)	pH Median	Total Area Away from Temp Median
pCO2 (mmHg)	pH MAD	DO Median
Gln (g/L)	Area Above pH Median	DO MAD
Glu (mmol/L)	Area Below pH Median	Area Above DO Median
Gluc (g/L)	Total Area Away from pH Median	Area Below DO Median
Lac (g/L)		Total Area Away from DO Median
NH4 (mM)		
Na (mM)		
K (mM)		

Percent Variance Captured by Regression Model

	---------X-Block----------		---------Y-Block----------	
Comp	This	Total	This	Total
----	-------	-------	-------	-------
1	19.18	19.18	55.48	55.48
2	7.38	26.56	12.71	68.19
3	5.86	32.42	4.56	72.75
4	5.11	37.53	2.21	74.96
5	4.12	41.65	0.99	75.95
6	3.91	45.55	0.28	76.23

Table 45. Details for PLSR model created using no explicit progression variable during Chapter 5.7.1

Elapsed Time
Developed 04-Dec-2012 12:11:34.43
X-block: Project A Day by Day 475 by 33
Num. LVs: 8
CV: random samples w/ 10 splits and 10 iterations



Model Performance	Cal	CV
RMSE	81.6259	90.8876
Bias	-2.10036e-011	0.030923
R^2	0.863422	0.830859

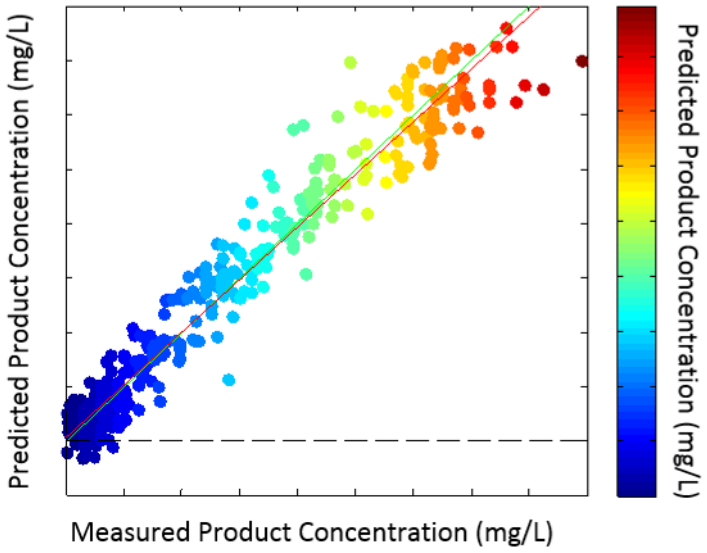
Variables		
Time	Volume of Air Added	Temp Median
Temp (C)	Volume of Oxygen Added	Temp MAD
Radiometer pH	True Volume of Oxygen Added	Area Above Temp Median
DOT %	Volume CO2 Added	Area Below Temp Median
Nova pH	pH Median	Total Area Away from Temp Median
pO2 (mmHg)	pH MAD	DO Median
pCO2 (mmHg)	Area Above pH Median	DO MAD
Gln (g/L)	Area Below pH Median	Area Above DO Median
Glu (mmol/L)	Total Area Away from pH Median	Area Below DO Median
Gluc (g/L)		Total Area Away from DO Median
Lac (g/L)		
NH4 (mM)		
Na (mM)		
K (mM)		

Percent Variance Captured by Regression Model

--------X-Block----------			----------Y-Block---------	
Comp	This	Total	This	Total
----	-------	-------	-------	-------
1	20.97	20.97	59.87	59.87
2	7.22	28.19	13.32	73.19
3	5.74	33.93	5.60	78.78
4	5.01	38.95	3.09	81.87
5	4.05	43.00	1.75	83.62
6	3.61	46.60	1.02	84.65
7	3.02	49.62	0.99	85.63
8	3.15	52.77	0.71	86.34

Table 46. Details for PLSR model created using Elapsed Time (h) as an explicit progression variable during Chapter 5.7.1

IVC
Developed 04-Dec-2012 12:19:7.25
X-block: Project A Day by Day 475 by 33
Preprocessing: Autoscale
Num. LVs: 8
CV: random samples w/ 10 splits and 10 iterations



Model Performance	Cal	CV
RMSE	40.63	47.18
Bias	-3.95062e-012	0.253164
R^2	0.966162	0.95457

Variables		
IVC	Volume of Air Added	Temp Median
Temp (C)	Volume of Oxygen Added	Temp MAD
Radiometer pH	True Volume of Oxygen Added	Area Above Temp Median
DOT %	Volume CO2 Added	Area Below Temp Median
Nova pH	pH Median	Total Area Away from Temp Median
pO2 (mmHg)	pH MAD	DO Median
pCO2 (mmHg)	Area Above pH Median	DO MAD
Gln (g/L)	Area Below pH Median	Area Above DO Median
Glu (mmol/L)	Total Area Away from pH Median	Area Below DO Median
Gluc (g/L)		Total Area Away from DO Median
Lac (g/L)		
NH4 (mM)		
Na (mM)		
K (mM)		

Percent Variance Captured by Regression Model

	---------X-Block----------		---------Y-Block----------	
Comp	This	Total	This	Total
----	-------	-------	-------	-------
1	20.59	20.59	64.50	64.50
2	7.33	27.92	15.36	79.86
3	5.59	33.52	6.61	86.48
4	5.06	38.58	3.71	90.19
5	4.06	42.64	2.36	92.55
6	3.53	46.17	1.74	94.29
7	2.89	49.06	1.61	95.90
8	3.57	52.64	0.72	96.62

Table 47. Details for PLSR model created using IVC as an explicit progression variable during Chapter 5.7.1

Model #	Data Used		Arrangement		Alignment			Include Variable		Response		
	Daily Monitoring	Informative Values	DbD	Profile	Natural	Time	IVC	Time	IVC	Viability	[Product]	VCC
1-3	X		X		X			X	X	1	2	3
4-6	X		X		X			X		4	5	6
7-9	X		X		X				X	7	8	9
10-12	X		X			X		X		10	11	12
13-15	X		X			X				13	14	15
16-18	X		X				X		X	16	17	18
19-21	X		X				X			19	20	21
22-24	X	X	X		X			X	X	22	23	24
25-27	X	X	X		X			X		25	26	27
28-30	X	X	X		X				X	28	29	30
31-33	X	X	X			X		X		31	32	33
34-36	X	X	X			X				34	35	36
37-39	X	X	X				X		X	37	38	39
40-42	X	X	X				X			40	41	42

Table 48. Dataset combinations, dataset arrangements, and variable selections test to evaluate impact of progression variable choice and dataset rigidity in analyses.

	Data Used		Arrangement		Alignment			Include Variable		Response		
Model #	**Daily Monitoring**	**Informative Values**	**DbD**	**Profile**	**Natural**	**Time**	**IVC**	**Time**	**IVC**	**Viability**	**[Product]**	**VCC**
43-45	X			X	X			X	X	43	44	45
46-48	X			X	X			X		46	47	48
49-51	X			X	X				X	49	50	51
52-54	X			X		X		X		52	53	54
55-57	X			X		X				55	56	57
58-60	X			X			X		X	58	59	60
61-63	X			X			X			61	62	63
64-66	X	X		X	X			X	X	64	65	66
67-69	X	X		X	X			X		67	68	69
70-72	X	X		X	X				X	70	71	72
73-75	X	X		X		X		X		73	74	75
76-78	X	X		X		X				76	77	78
79-81	X	X		X			X		X	79	80	81
82-84	X	X		X			X			82	83	84

Table 49. Dataset combinations, dataset arrangements, and variable selections test to evaluate impact of progression variable choice and dataset rigidity in analyses (con't).

Model Information				[Product]			Viability			VCC		
Arrange.	**Align.**	**Progression Variable**	**Cal/Val**	**LV**	**R^2 Cal**	**R^2 CV**	**LV**	**R^2 Cal**	**R^2 CV**	**LV**	**R^2 Cal**	**R^2 CV**
DbD	IVC	IVC	All	5	0.933	0.928	3	0.579	0.536	3	0.952	0.947
DbD	None	IVC	All	3	0.876	0.850	3	0.808	0.766	2	0.859	0.836
DbD	IVC	NOI	All	3	0.832	0.816	3	0.443	0.387	4	0.940	0.933
DbD	None	NOI	All	2	0.700	0.658	3	0.643	0.576	2	0.863	0.842
DbD	Time	NOI	All	3	0.816	0.796	3	0.698	0.666	3	0.901	0.891
DbD	None	Time	All	3	0.802	0.764	3	0.740	0.681	2	0.858	0.834
DbD	Time	Time	All	3	0.857	0.844	3	0.780	0.756	3	0.900	0.888
Profile	IVC	IVC	All	2	0.681	0.315	1	0.417	0.125	1	0.656	0.460
Profile	None	IVC	All	3	0.948	0.724	2	0.923	0.784	2	0.929	0.856
Profile	IVC	NOI	All	2	0.670	0.294	1	0.415	0.109	2	0.757	0.529
Profile	None	NOI	All	3	0.930	0.613	1	0.737	0.589	2	0.937	0.834
Profile	Time	NOI	All	4	0.967	0.390	2	0.828	0.292	2	0.877	0.554
Profile	None	Time	All	3	0.931	0.634	3	0.939	0.647	1	0.775	0.519
Profile	Time	Time	All	4	0.967	0.422	2	0.828	0.342	2	0.877	0.559

Table 50. Summary of PLSR Models and Results when Using Only Daily Monitoring Dataset

Model Information				[Product]			Viability			VCC		
Arrange.	**Align.**	**Progression Variable**	**Cal/Val**	**LV**	**R^2 Cal**	**R^2 CV**	**LV**	**R^2 Cal**	**R^2 CV**	**LV**	**R^2 Cal**	**R^2 CV**
DbD	IVC	IVC	All	4	0.893	0.876	6	0.634	0.550	5	0.962	0.957
DbD	None	IVC	All	7	0.914	0.890	7	0.873	0.831	2	0.863	0.842
DbD	IVC	NOI	All	7	0.867	0.838	6	0.527	0.429	5	0.950	0.943
DbD	None	NOI	All	6	0.778	0.709	5	0.680	0.569	3	0.876	0.851
DbD	Time	NOI	All	3	0.796	0.753	3	0.695	0.623	3	0.903	0.884
DbD	None	Time	All	7	0.838	0.795	7	0.794	0.744	2	0.861	0.840
DbD	Time	Time	All	3	0.903	0.883	3	0.752	0.717	3	0.903	0.886
Profile	IVC	IVC	All	1	0.474	0.160	4	0.888	0.227	2	0.833	0.461
Profile	None	IVC	All	1	0.681	0.111	5	0.984	0.339	9	0.997	0.649
Profile	IVC	NOI	All	1	0.472	0.192	2	0.743	0.070	4	0.944	0.525
Profile	None	NOI	All	5	0.979	0.097	5	0.973	0.338	5	0.984	0.714
Profile	Time	NOI	All	4	0.950	0.042	4	0.945	0.037	5	0.975	0.064
Profile	None	Time	All	5	0.977	0.983	5	0.976	0.367	10	0.998	0.675
Profile	Time	Time	All	4	0.950	0.065	4	0.945	0.011	5	0.975	0.152

Table 51. Summary of PLSR Models and Results when Using Integrated Daily Monitoring and Online Monitoring Dataset.

Informative Values Set	**Values for Steady State Profiles**	**Values for Dynamic Profiles**
1	Averages, Standard Deviations, Slope, Coefficient of determination (R^2)	Averages, Standard Deviations, Slope, Coefficient of determination (R^2)
2	Averages, Standard Deviations, Slope, Coefficient of determination (R^2)	Times of key events (air and O_2)
3	Median, Median Absolute Distance	Times of key events (air and O_2)
4	Median, Median Absolute Distance	Times of key events (air and O_2) Volumes added (air and O_2)
5	Median, Median Absolute Distance	Times of key events (air and O_2) Volumes added (air and O_2)
6	Median, Median Absolute Distance, Areas between reading and median (above, below, total)	Times of key events (air and O_2) Volumes added (air and O_2)
7	Median, Median Absolute Distance, Areas between reading and median (above, below, total)	Times of key events (air, O_2, and CO_2) Volumes added (air, O_2, and CO_2)

Table 52. Summary of Informative Value Datasets by Version

Scaling	Justification Used	Type	PCs Used*	X Variance Captured (%)
Autoscale	Harvest Justified	Data	2	25
		Single Variable Model PCs	1	19
	Inoculation Justified	Data	5	45
		Single Variable Model PCs	1	25
	Peak VCC Centred	Data	1	18
		Single Variable Model PCs	1	23
Intrascale A	Harvest Justified	Data	5	45
		Single Variable Model PCs	1	26
	Inoculation Justified	Data	4	38
		Single Variable Model PCs	1	20
	Peak VCC Centred	Data	5	48
		Single Variable Model PCs	2	42

Table 53. PCA and HPCA Models. *Number of PC used based on minimum RMSE during cross-validation.

Scaling Applied	Justification Used	Decline Limit (%/d)	Latent Variables Used	Variance Captured (%)		Misclassified (%)
				X	Y	
Autoscale	Harvest Justified	10	2	20	48	24
		20	2	19	50	22
		30	2	21	53	16
		40	2	20	53	14
		50	2	18	44	10
		60	2	17	31	12
	Inoculation Justified	10	2	25	43	31
		20	3	31	48	26
		30	2	25	45	18
		40	2	24	43	15
		50	2	23	35	20
		60	2	23	27	32
	Peak VCC Centred	10	2	24	37	26
		20	2	22	42	31
		30	2	24	45	21
		40	2	23	49	14
		50	3	30	47	15
		60	2	19	31	20

Table 54. PLS-DA results for CHO process platform investigation using Analysis Pattern 1 and Autoscaling (i.e. mean-centred and scaled to unit variance) applied.

Scaling Applied	Justification Used	Decline Limit (%/d)	Latent Variables Used	Variance Captured (%)		Misclassified (%)
				X	Y	
Intrascale A	Harvest Justified	10	2	19	21	46
		20	2	24	34	30
		30	2	23	33	32
		40	2	22	31	34
		50	2	21	32	43
		60	2	21	30	38
	Inoculation Justified	10	2	19	25	49
		20	2	21	37	37
		30	2	21	33	19
		40	2	20	31	40
		50	2	18	28	49
		60	2	18	30	49
	Peak VCC Centred	10	2	16	20	51
		20	3	32	44	29
		30	2	28	27	37
		40	2	27	25	32
		50	2	18	31	37
		60	2	12	37	44

Table 55. PLS-DA results for CHO process platform investigation using Analysis Pattern 1 and Intrascale A (two-step scaling process) applied.

Scaling Applied	Justification Used	Decline Limit (%/d)	Latent Variables Used	Variance Captured (%)		Misclassified (%)
				X	Y	
Autoscale	Harvest Justified	10	2	26	38	24
		20	4	44	44	26
		30	2	27	38	16
		40	2	26	38	15
		50	2	24	23	18
		60	2	22	12	37
	Inoculation Justified	10	2	35	32	29
		20	2	34	29	29
		30	2	35	35	20
		40	3	41	37	18
		50	2	33	19	29
		60	2	33	13	37
	Peak VCC Centred	10	2	31	29	26
		20	2	31	31	32
		30	2	33	35	22
		40	2	31	39	13
		50	2	28	27	21
		60	2	27	22	28

Table 56. HPLS-DA results for CHO process platform investigation using Analysis Pattern 2 and Autoscaling (i.e. mean-centred and scaled to unit variance) applied.

Scaling Applied	Justification Used	Decline Limit (%/d)	Latent Variables Used	Variance Captured (%)		Misclassified (%)
				X	Y	
Intrascale A	Harvest Justified	10	2	24	10	52
		20	2	36	30	29
		30	2	36	25	29
		40	2	34	22	36
		50	2	33	16	39
		60	2	32	14	48
	Inoculation Justified	10	2	24	12	49
		20	2	31	24	39
		30	2	31	20	40
		40	2	28	17	40
		50	2	26	10	53
		60	2	26	9	53
	Peak VCC Centred	10	2	18	13	47
		20	2	38	22	31
		30	2	38	19	37
		40	2	38	19	33
		50	2	35	11	41
		60	2	30	11	40

Table 57. HPLS-DA results for CHO process platform investigation using Analysis Pattern 2 and Intrascale A (two-step scaling process) applied.

Scaling Applied	Justification Used	Decline Limit (%/d)	Leaves in Tree	Misclassified (%)	Top Node
Autoscale	Harvest Justified	10	11	2	pCO2 PC2
		20	9	10	pH PC1
		30	11	5	Lac PC1
		40	11	1	pH PC1
		50	10	1	pH PC1
		60	10	2	Gln PC2
		70	6	1	Gln PC1
	Inoculation Justified	10	11	5	Lac PC1
		20	18	4	pH PC1
		30	14	2	pH PC1
		40	6	4	pH PC1
		50	11	4	pH PC1
		60	10	5	Gluc PC1
		70	8	2	Gluc PC1
	Peak VCC Centred	10	9	3	K PC1
		20	17	4	Lac PC1
		30	14	4	pH PC1
		40	8	4	pH PC1
		50	7	4	pH PC1
		60	5	4	Gln PC1
		70	2.	3	Gln PC1
Intrascale A	Harvest Justified	10	13	5	Na PC1
		20	13	7	pH PC1
		30	10	7	pH PC1
		40	10	2	pH PC1
		50	11	1	Gln PC2
		60	9	3	Gln PC2
		70	3	2	Gln PC1
	Inoculation Justified	10	13	2	Na PC1
		20	19	6	Lac PC1
		30	13	4	pH PC1
		40	11	2	pH PC1
		50	7	4	pH PC1
		60	7	6	Gln PC2
		70	7	1	Gln PC2
	Peak VCC Centred	10	14	4	pH PC1
		20	19	7	pH PC1
		30	15	5	pH PC1
		40	9	6	pH PC1
		50	10	6	Lac PC1
		60	9	4	Gln PC1
		70	4	2	Gln PC2

Table 58. Decision tree results for CHO process platform investigation using Analysis Pattern 2 and Intrascale A (two step scaling process) applied.

Scaling Applied	Decline Limit (%/d)	Leaves in Tree	Misclassified (%)	Top Node
Autoscale	10	11	2	PC1 Reading 7
	20	15	6	PC1 Reading 2
	30	12	4	PC1 Reading 2
	40	11	4	PC1 Reading 2
	50	12	1	PC1 Reading 2
	60	7	4	PC1 Reading 3
	70	6	2	PC1 Reading 1
Intrascale A	10	12	7	PC1 Reading 8
	20	13	13	PC1 Reading 10
	30	13	2	PC1 Reading 3
	40	10	4	PC1 Reading 3
	50	9	3	PC1 Reading 3
	60	7	2	PC1 Reading 11
	70	6	2	PC1 Reading 3

Table 59. Decision tree results for CHO process platform investigation using Analysis Pattern 3 and Intrascale A (two step scaling process) applied.