A Business Analytics Approach to augment Six Sigma Problem Solving: A Biopharmaceutical Manufacturing Case Study

Will Fahey[1,2], Paul Jeffers[1], Paula Carroll[2][0000-0003-1029-1668]

[1] Pfizer Ireland Pharmaceuticals, Dublin, Ireland
[2] School of Business, University College Dublin, Ireland

william.fahey.1@ucdconnect.ie

Highlights

- **A business analytics methodology to complement the Six Sigma defect reduction framework.**
- **Adapt symbolic representation schemes for large biopharma manufacturing datasets**
- **A novel approach to identifying and characterising influential variables in manufacturing time series data.**
- **Translation of model findings into manufacturing rules which improve process robustness.**

Abstract

Biopharmaceutical manufacturers are required to collect extensive observational data sets in order to meet regulatory and process quality monitoring requirements. These datasets contain information that may improve the performance of the production process. Analytics provides a means of extracting this information and provides an approach for incorporating actionable insights into production practices.

We present a novel framework which combines Six Sigma and Business Analytics. This approach mines large volumes of inline and offline biopharmaceutical production data, allowing the entire production process to be analysed and modelled. The recommendations of the model are represented as manufacturing rules which give actionable insights to improve the performance of the process.
The integrated approach delivers promising results from synthetic experiments as well as being applied in practice to a cell culture process.

Keywords: Biopharmaceutical, Manufacturing, Analytics, Six Sigma, Time Series

1 Introduction

A biopharmaceutical is defined as a medicinal product with an active ingredient created by biological rather than chemical synthesis. Biopharmaceutical manufacturing involves a complex sequence of reaction, separation and concentration process steps. The biological components are highly sensitive to their environmental conditions, hence fluctuations in process performance are inevitable. Owing to the complexity, diversity of data types and non-linear relationships that characterise the biopharmaceutical manufacturing process, a new approach is needed to understand these fluctuations. The Business Analytics suite of data mining and machine learning algorithms is one approach proven to gain insight from highly complex datasets (Otsuka, Nagasaki, & Kataoka, 2012). This paper explores the feasibility of using Business Analytics to improve the performance of a biopharmaceutical manufacturing process.

Process Robustness (PR) is defined as the ability of a manufacturing process to demonstrate the finished product is of consistently acceptable quality while tolerating variability within process variables and raw materials. Ensuring this consistency is one of the most challenging issues faced by manufacturers of biopharmaceuticals (Rader & Langer, 2018). Critical Quality Attributes (CQAs) are defined as the most significant chemical, physical and biological attributes that can be measured to ensure the final product remains within acceptable quality limits (Rathore & Winkle, 2009) Unexplained CQA fluctuation poses a challenge for manufacturers attempting to satisfy clinical market demands in a stable supply chain in the most cost-effective manner. Six Sigma's problem-solving framework is the universally agreed approach within the manufacturing sector to mitigate this fluctuation.

Six Sigma statistical methods provide a structured approach for identifying the root causes of production defects (Schroeder, Linderman, Liedtke, & Choo, 2008). While Six Sigma has demonstrated its effectiveness to this end, the characteristics of the modern biopharmaceutical dataset require practitioners to adopt a new approach. First, the data normality statistical assumptions underpinning Six Sigma are restrictive. Second, the abundance of manufacturing data obtained by modern manufacturing control systems is difficult to analyse with the traditional Six Sigma approach. Manufacturing data is characterised by the Big Data four “V’s ” of volume, variety, velocity and veracity (Babiceanu & Seker, 2016). Although manufacturing datasets have grown in size and complexity, the capability to distil knowledge has remained static (Flath & Stein, 2018). Business Analytics has the potential to generate hypotheses from manufacturing datasets whereas Six Sigma traditionally uses this data to confirm hypotheses generated by process experts.

Rather than discard the traditional six sigma approach, the Business Analytics (BA) techniques for knowledge extraction offer an opportunity to build upon the existing Six Sigma data-driven foundation and map a pathway for biopharmaceutical businesses to the fifth V - value. Business Analytics is the exploration of an organisation's data to improve decision-making (Holsapple, Lee-Post, & Pakath, 2014). These approaches have the potential to transform the entire biopharma value chain, increasing flexibility, quality, and speed to market in an initiative known as Industry 4.0 (Zhong, Xu, Klotz, & Newman, 2017). Embracing this disruptive technology could unlock new sources of value and competitive advantage for manufacturers.

Due to the variability of the manufacturing processes, a structured approach called Quality by Design (QbD) is used to transfer a manufacturing process from the bench to large scale manufacturing (Rathore, Mittal, Pathak, & Arora, 2014). Adherence to the QbD process ensures that the quality of the final product remains within the Normal Operating Ranges (NOR) by maintaining the settings of the production variables. In the early stages of the manufacturing lifecycle, this prioritisation of process effectiveness over process efficiency is appropriate. However, as a plant or product matures, there should be a shift in focus from process effectiveness to process efficiency and BA techniques are a potential enabler for this. BA approaches such as Support Vector Machines and Random Forests have provided the means for characterising raw material and process variables and understanding their influence on the quality of the final product (Hill & Waner, 2017).

Strategic factors force manufacturers to challenge the status quo in terms of process performance. Biopharma manufacturing is highly competitive, with a rapid increase in manufacturing capability in developing countries through improved knowledge and technology transfer (Rader & Langer, 2018). The increased competition coincides with the demand for lower-cost alternatives from healthcare providers and consumers to existing drugs and products (Chiang, Lu, & Castillo, 2017). These changes in the market landscape have resulted in manufacturing companies turning to less traditional revenue sources, namely the production of biosimilars (generic versions of biopharmaceuticals) and niche products targeting rare diseases. The new demands of these shorter production lifecycles require that process scientists build knowledge of manufacturing processes faster than the statistical toolbox permits. Huang et al. argue that intangible assets such as tacit process knowledge can be harnessed to effect innovation and achieve an organisational advantage in the competitive global biopharmaceutical market (Huang, Lai, & Lin, 2011). They ask how biotechnology firms can influence their innovative capability and performance with intangible assets, which they categorise as human, information and organisation capital. Taking a resource-based view, they suggest that intangible assets transformed into skilled actions can provide firms with a competitive advantage and improved performance.

Regulatory agency direction is also an essential driver of manufacturing change. Cox et al. highlight that regulatory agencies would welcome the increased process

understanding that results from a BA approach (Cox & Onraedt, 2012). The FDA guidelines for industry (FDA guidelines for industry, 2009) and more recent direction from the HPRA (HPRA, 2017) reinforce this conclusion. Jenzsch et al. goes so far as to assert that regulatory bodies could mandate BA methods as a means of building process understanding (Jenzsch et al., 2017).

This study identifies the limitations of Six Sigma when applied to the Biopharmaceutical dataset while showing how BA techniques can integrate with the existing Six Sigma framework rather than replace it. The research question asks if BA techniques can complement the Six Sigma approach to improve process understanding by extracting insights from biopharmaceutical manufacturing data.

The project also puts forward an amendment to the SAX algorithm that addresses the analysis of a multitude of skewed time series variables contained within a typical biopharmaceutical manufacturing dataset. Finally, it identifies an approach that translates the model findings to manufacturing rules that can be readily executed on the manufacturing floor.

2 Literature Review

A structured literature review was carried out in order to identify any previous attempts to combine Six Sigma and Business Analytics into a new framework. The only examples were where practitioners had substituted a step in the six sigma process with one from the business analytics toolkit. Examples include using manufacturing data instead of a Design of Experiments at bench scale (Ghosh & Maiti, 2014) and the replacement of linear regression with logistic regression (Mishra & Rane, 2019).

2.1 Business Analytics Framework

This project represents the convergence of two disciplines: Business Analytics and Biopharmaceutical manufacturing. In order to resonate with readers from either discipline, it is necessary, to begin with, a definition of key terms (Rader & Langer, 2017). In this paper, the term analytics is employed as a broad and general term for all data-driven analysis techniques and concepts across enterprise data and big data. Analytics is an umbrella term for machine learning, business intelligence, big data, advanced analytics and process analytics. Business Analytics is the application of these techniques to improve decision making in a business setting (Holsapple et al., 2014).

2.2 Cell Culture process

The cell culture approach to manufacturing proteins has been widely reported in the literature (Wurm, 2004). The manufacturing process at the centre of this study is a

typical large-scale mammalian cell culture process, where the active ingredient is a protein expressed by a specific cell clone.

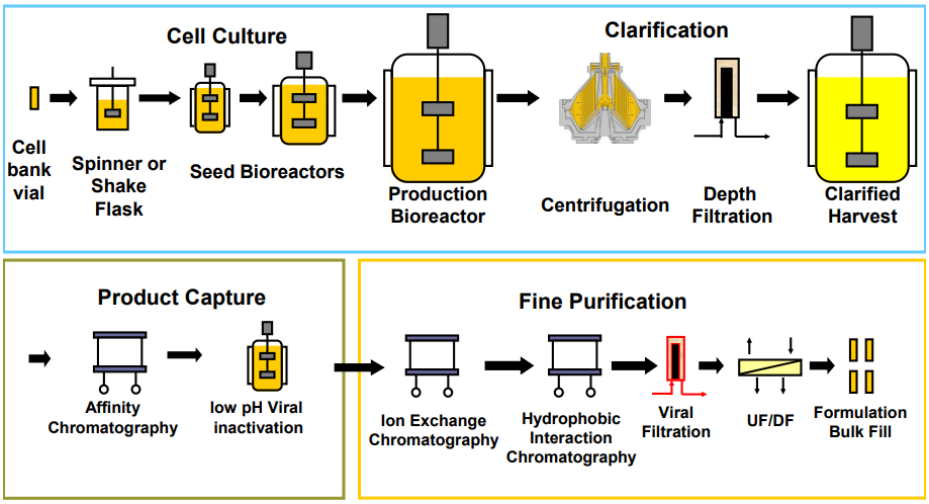

Figure. 1. The cell culture manufacturing process (Minow, 2012)

The manufacturing process can be divided into two stages, Cell Culture and Purification (Minow, 2012). The cell culture phase produces the target protein. Titre is defined as the amount of target protein recovered at the end of the process, usually measured in kilograms.

First, a 1 ml vial from the working cell bank is expanded through stainless steel vessels, called bioreactors. The first scale up is from wave bags of 1 ml to 50L. From there, the culture is transferred to ever increasing bioreactors from 100L up to 12,500L. At this stage, the batch is held under conditions that attempt to maximize cell density. Next, a sharp reduction in temperature triggers the release of the target protein. As the culture begins to produce the product, new media is continuously added in, while cell debris and other undesirable growth by-products are removed. At the end of this stage, the target protein and the cells are together in one blend. These components are separated during the purification stage.

The purification process is a sequence of filtration (Shukla & Gottschalk, 2013) and chromatography (Miller, 2005) steps. A combination of ultrafiltration and diafiltration (UF/DF) reduce the product to a specified concentration incrementally eliminating more cellular debris and particulate matter. The final phase of filtration focuses on inactivation and filtration of viruses. Chromatography is the second component of the purification phase. The target protein has a biological tag which enables the chromatography steps to capture it. The manufacturing process from inoculation to final puri-

fication takes 6-7 weeks. A large manufacturing plant will have multiple batches at various stages of the process running at any one time. Every process step described above has a built-in recipe containing explanatory variables that are automatically monitored and controlled within the Normal Operating Range (NOR) by the Manufacturing Control System (MCS). Examples of these explanatory variables (EV) are agitation rates, temperatures and transfer times.

The inadequacy of statistical methods to interpret the abundance of available time series data is a key driver for the proposed improvements to the existing Six Sigma method for time series analysis. Each of the 20 process steps described in figure 1 has fifteen inline explanatory variables monitored at five second intervals. The longest process step generates just under three million inline data points.

2.3 Data mining the biopharmaceutical dataset

The applications of Six Sigma and Business Analytics individually to resolve manufacturing issues are well documented (Chien, Wang, & Cheng, 2007; Fahey & Carroll, 2016; Johnston, Maguire, & McGinnity, 2009; Kamsu-Foguem, Rigal, & Mauget, 2013). Few published studies have explored how the frameworks can be combined to utilise all the available manufacturing data to generate a holistic model of the process. Explanatory variables can be categorised into inline measurement and offline measurement (Charaniya, Hu, & Karypis, 2008). Inline measurement generates time series data from sensors and is available in real time. Offline measurements require a sample of the batch to be taken and sent to the lab for analysis. Most offline measurements are taken once per batch. Other discrete explanatory variables such as the cell line used, and equipment name should also be included. All previous work has focussed on one of these categories in isolation, whereas this approach provides a method to combine both elements to build a holistic model that has the highest explanatory power possible.

2.3.1 Offline Measurement

Offline measurement can be defined as discrete variables such as the piece of equipment used in a process step or a lab measurement of the pH at the end of a reaction. Owing to the high number of explanatory variables, dimensionality reduction techniques such as Principle Component Analysis (PCA) and Partial Least Squares (PLS) regression have been extensively employed (Hong et al., 2018; Rathore et al., 2014). As useful and necessary as these techniques are, a different approach is required if time series data is to be analysed along with the variables classified as offline measurements (Ge, 2013).

Coleman et al. use a genetic algorithm to solve a multi-criteria optimisation problem in biopharmaceutical manufacturing by using a neural network (Coleman & Block, 2006). One of the criteria targeted for optimisation was profitability (referred to as a simplified economic factor). Interestingly, while the authors successfully improved

this measure by identifying that the optimised process required a reduced amount of costly raw material, they were unsuccessful in improving yield. A genetic algorithm was chosen for the optimisation problem due to its ability to deal with a sizeable multidimensional search space. The author defines this as below 20 explanatory variables, which is much lower than the biopharmaceutical dataset. Khalid et al. measure the ability of multiple techniques to predict the tensile strength of the finished product. Random Forest outperformed Neural Networks and genetic programming with the added advantage of improved interpretability (Khalid et al., 2016). Symbolic regression was used as a feature reduction technique. Random Forest has been used extensively in other high dimensional settings such as Semi-Conductor manufacturing and fraud detection (Do, Lenca, Lallich, & Pham, 2010; Li, Yan, Liu, & Li, 2018) The

The Random Forest algorithm was chosen for several reasons. First, the method was required to handle categorical and numerical variables. Random Forests (RF) have also been identified as out-performing other approaches (Delgado, 2014). The transparency by which the algorithm arrives at its conclusion is also a significant advantage within a regulated environment. The RF implementation in R can be translated into rules using the iTree package. In addition to identifying which variables are most influential, the approach also provides a target value within the normal operating range.

2.3.2 Inline measurement

The bulk of the cell culture manufacturing dataset is comprised of time series data collected by in-line sensors. In-line data is of particular importance in Cell Culture fed-batch where any anomaly or deviation from the NOR for critical parameters can have severe implications for Critical Quality Attributes (CQA) or product yield. There have been several approaches used to mine time series distributions in manufacturing including clustering (Singhal & Seborg, 2005), variations of wavelets (Kim, 2016) and Dynamic Time Warping (DTW) (Bork, Ng, Liu, Yee, & Pohlscheidt, 2013). These approaches consider the time series element of the dataset in isolation, neglecting the improved descriptive power available by adding in discrete variables such as the raw material lot used.

The time series element of the dataset is characterised by high levels of noise, outliers, and batch to batch variability. From a processing efficiency perspective, working with time series data in its raw format is very expensive due to the sheer volume. Further, if the data mining approach is applied to the raw time series, the algorithm could be identifying similarities in the noise rather than the similarity of the distribution itself (Keogh & Kasetty, 2003). As such, it is preferable to use a representation method to reduce the dimensionality of time series, while still preserving its fundamental characteristics (Wang et al., 2013). Wang et al. (2013) compare eight different representation methods using a measure called pruning power. The author concludes

that the effectiveness of the method is highly dependent on the nature of the subject dataset, but falls short of highlighting which representation methods should be used on a particular type of dataset.

Time Series analysis can be broken into two main subsections: time series representation and motif discovery. A motif or discord is defined as an unusual sub-sequence of points within a time series distribution (Keogh, Lin, & Fu, 2005). These discords are identified as the subsequence that is the most substantial distance from all other subsequences. Keogh observed that normalised subsequences have Gaussian distributions, which formed the basis for his work with Lin on symbolic time series representation with Symbolic Aggregate Approximation (SAX) (Lin, Keogh, Wei, & Lonardi, 2007).

SAX looks to describe the characteristics of a distribution by first using an intermediate representation called Piecewise Aggregate Approximation (PAA). PAA divides the series into equal sized subsequences and returns the average of the values within those subsequences. Leveraging the assumption that the normalised subsequences follow a Gaussian distribution, breakpoints are allocated that produce equal sized areas under the Gaussian curve. A symbol is allocated to the value of each PAA subsequence. The result is a string of symbols, called a word, which preserves the characteristics of the distribution while compressing it significantly.

One assumption pervading most of the SAX deployments in the literature is that the normalised time series data follow a Gaussian distribution. This assumption does not hold for the highly skewed biopharmaceutical manufacturing dataset, which is the subject of this study, which is why the algorithm was amended. Although rare, there are other examples of this observation in the literature such as an electrical load pattern grouping problem (Notaristefano, Chicco, & Piglione, 2013).

Once the SAX word has been generated, it can be passed to the multitude of Business Analytics algorithms to extract insight. Senin et al. take an interesting approach by first discretising the time series data and then passing the derived "words" to a vector space model (VSM) to rank the time series patterns according to their importance to their class (Senin & Malinchik, 2013). The methodology could potentially be used to associate specific characteristics of a time series distribution with either a low or high yield. Roda et al. take a similar approach to this project by proposing a method to analyse generic sensor data by decomposing the resulting time series distribution into rules which are referred to as "multi-variate temporal reasoning"(Roda & Musulin, 2014).

2.4 Six Sigma and CRISP-DM

The Six Sigma approach was designed by Motorola in the 1980s in order to reduce defects to below 3.4 parts per million defect rate by identifying and eliminating variation within their processes (Schroeder et al., 2008). Each Six Sigma project follows a

defined sequential path: Define, Measure, Analyse, Improve and Control (DMAIC). The Six Sigma methodology employs standard quality tools such as Process Mapping, Failure Mode and Effect Analysis (FMEA), cause and effect structured brainstorming, and statistical process control. The method's strengths are the use of statistical techniques for empirical verification of potential root causes, and the structure provided by the DMAIC sequential model (de Mast & Lokkerbol, 2012). Somewhat counterintuitively, the implementation of this rigid structure to problem-solving has been found to improve innovative approaches to manufacturing (Möldner, Garza-Reyes, & Kumar, 2018).

As with any framework, Six Sigma also has its drawbacks. Its generality as a methodology also means that it suffers in specificity, referred to as the power/generality tradeoff (Newell, 1969). A second limitation of the statistical methods is the higher prevalence of non-normal distributions. This is especially relevant in complex manufacturing processes such as biopharmaceutical manufacturing.

The most widely accepted framework to complete a Business Analytics project is CRISP-DM (Cross Industry Standard Process for Data Mining) (Wirth, 2000). The approach is independent of both industry and technology used. Hence, Newells findings can also be applied to the generality of the CRISP-DM approach with industry specific amendments to the framework emerging from the literature. One such example is Huber et al.'s customisation for engineering applications of Business Analytics (Huber, Wiemer, Schneider, & Ihlenfeldt, 2019).

In regulated sectors such as biopharmaceutical, medical device and food industries, the level of innovative improvements are constrained within the production process design parameters. BA can enable actions in that space by identifying process improvement options from insights extracted from data at hand. Although some processing is required to prepare the data for additional analytics exploration/mining, this offers an opportunity to address the Value opportunity of big data.
The approach detailed in this study is a means of increasing the specificity of the methodology, as the analytics model would be unique to the target process.

3 Methodology

3.1 Overview

In this study, a hybrid of Business Analytics and Six Sigma methods were applied to a CQA deviation on a cell culture product (Alt et al., 2016). The model was generated with 26 batches of data, each consisting of approximately 150 offline variables with 15-20 prioritised time series variables from inline sensors.

In the next section, the current Six Sigma process incorporating the suggested augmentations is described, broken down by phase. Next, the steps required to preprocess the data are detailed as this step has increased importance due to the relatively small number of batches available. The requirement to amend the SAX algorithm in order to represent the manufacturing time series data is explored. Finally, a modelling and cross-validation approach with direction on how to apply the model recommendations in practice is presented.

3.2 Integration of Six Sigma and Business Analytics approach

Previous literature has shown how the CRISP-DM framework can be mapped to the DMAIC structure (Fahmy, Mohamed, & Yousef, 2017). Instead, a practical phase by phase roadmap on how analytics can improve the Six Sigma process is described.
Figure 1 shows elements contributed to the hybrid approach by Six Sigma and CRISP-DM.

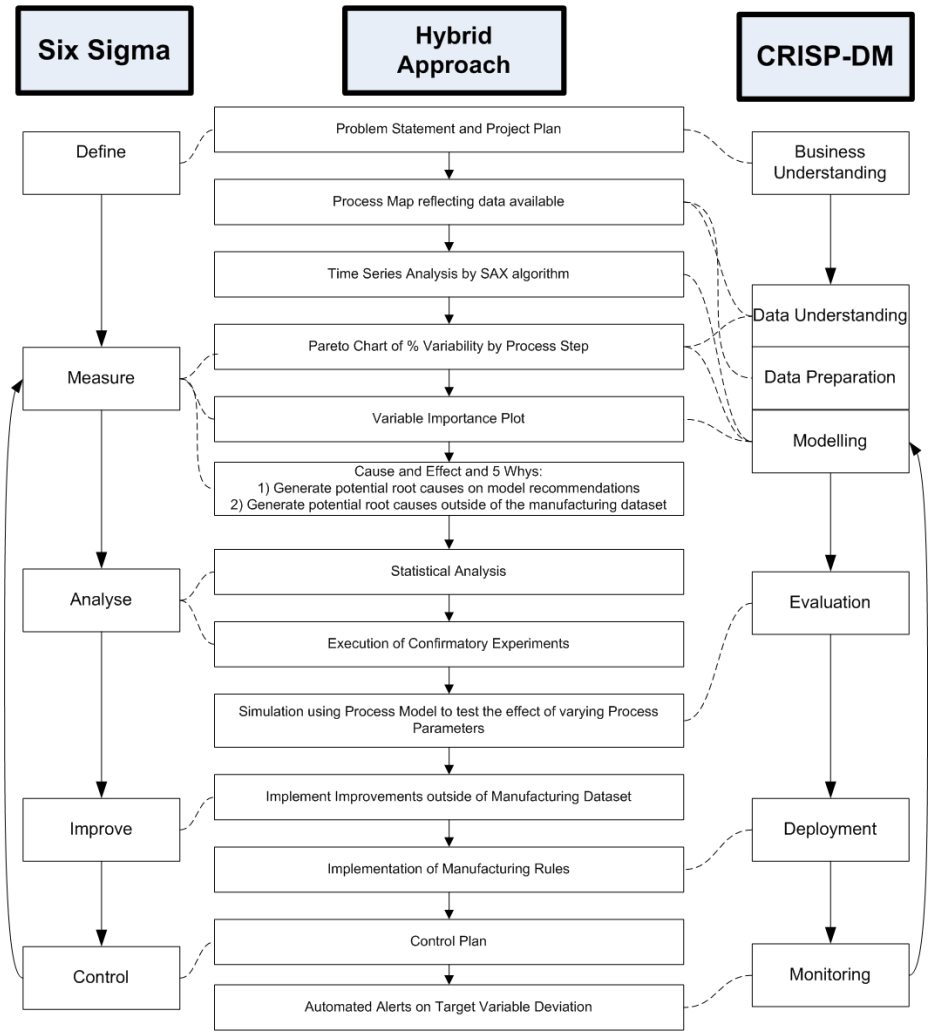

Figure 2: The Hybrid Approach combining elements of Six Sigma and CRISP-DM

3.3 Define/Business Understanding

The goal of the Define phase is to detail the target business problem and what the target state is.

3.3.1 Problem Statement and Project Plan

While there is no specific enhancement analytics has to offer when formulating a problem statement, specific criteria must be met to ensure that the issue requires analytics to be applied. The problem should be business critical and data rich. Also, the identification of a knowledgeable sponsor with an appetite for thoughtful risk-taking is beneficial.

3.3.2 Process Map reflecting the data available

The development of a process map is a critical step in any Six Sigma project. The activity involves the team mapping the process steps as they are executed rather than how it is recorded in the procedures. This exercise crystallises the team's understanding of the issue and highlights different approaches taken that may be potential sources of variation, essentially codifying the tacit knowledge referred to by Huang (Huang et al., 2011). Each step in the process map is then analysed to identify potential root causes to the issue. The Hybrid Approach suggests that this would be an appropriate stage to reflect the available data at each process step. This can help facilitate an initial prioritisation exercise around which data may be most likely to influence the target variable.

3.4 Measure/Data Preparation, Understanding and Modelling

The measure phase looks to visualise available data towards reducing the scope of the problem, referred to as the funnelling effect. Potential root causes to the issue are generated by brainstorming using a Cause and Effect Diagram. Two of the most popular visualisations of data are a Pareto Chart and a Time Series Plot.

3.4.1 Pareto Chart and Variable Importance Plots

A Pareto chart is a bar chart that sorts by frequency (Wilkinson, 2006). The motivation is to compare categories to narrow the scope of the problem to the significant few. Analytics offers an improvement on this approach by introducing a Variable Importance Plot (VIP), which visualises the most influential variables in a Pareto format. The VIP was vital to demonstrating to the manufacturing teams that this approach could provide meaningful process insight, as the team immediately began to generate hypotheses as to why these variables could have an influence. For confidentiality reasons the variables are masked, but typical examples can include Dissolved Oxygen (DO), pH or the Oxygen flowrate. A VIP can also be used to identify the variability contributed to the target variable by each process step using the Random Forest % variance. Once the Random forest model is built, explanatory variables as-

sociated with each process step can be removed or added sequentially, and the effect on the explanatory power of the model observed. If successful, this could reduce the scope of the investigative effort considerably.

3.4.2 Time Series Analysis

During the measurement phase, Six Sigma directs a process scientist to visually review individual time series plots to identify discrepancies on batches with a significantly high or low target variable value. However, when a discrepancy is observed, it may be purely coincidental.
The application of Symbolic Aggregate Approximation (SAX) in conjunction with the Random Forest algorithm allows the identification of the subsequences, which correlate with the target variable. While this may not always equate to a causative relationship, it is still an improvement on the incumbent method. This improves the funnelling effect of the measure phase by allowing the simultaneous analysis of a larger number of time series.

3.4.3 Symbolic Aggregate Approximation (SAX)

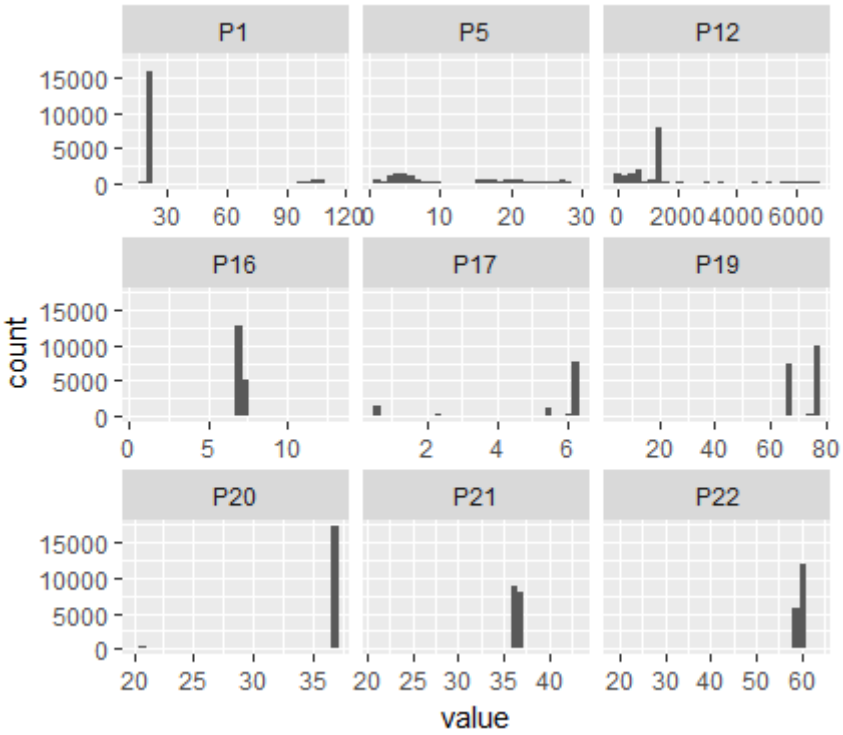

Figure 3. Probability distributions of the Time Series

As stated, one of the critical limitations of the Six Sigma statistical methods is that the assumption of data normality does not hold for the biopharmaceutical dataset. This is also the case for the SAX algorithm, which uses this assumption to allocate breakpoints between symbols of an equal area under the curve. As demonstrated by Figure 3, the manufacturing data was highly skewed and, in some cases, even bi-modal. The amended SAX algorithm first identifies the most closely fitting probability function and then allocates breakpoints based on this function.

SAX transforms a time series T of length m into a string derived from an alphabet size of K. There are four high-level steps:

1) Z-Normalisation of data: Comparing time series only makes sense if they are in the same range. This is done in 2 steps:
    a. Subtract the mean of the time series from every value
    b. Divide every value by the standard deviation

   The resulting series will have a mean of zero and a standard deviation of one.

2) Identification of the best fitting probability distribution to the data: The code compares the distribution of the normalised data to a supplied list of distributions. Some fitness statistics and criteria are considered when making this decision:
    a. Kolmogorov-Smirnov statistic
    b. Cramer-von Mises statistic
    c. Anderson-Darling statistic
    d. Akaike's Information Criterion
    e. Bayesian Information Criterion

   Akaike's Information Criterion (AIC) is prioritised as the most critical measurement.

3) Represent the data using Piecewise Aggregate Approximation (PAA): PAA is the first step in SAX, where each subsequence is represented by a constant. This approximation is the average of the values captured by the span of the subsequence size.

4) Convert the PAA representation to a symbolic string (SAX), picking k equal sized areas under the identified curve. The x co-ordinates of these lines are called breakpoints. The area under the identified curve between breakpoint i (β_i) and breakpoint i + 1 (β_{i+1}) has to be 1/k where k is the alphabet size. Now all the mean values given by the PAA transformation are converted to the symbol corresponding to the interval in which they reside. So all points of the original time series, whose windows mean value is below $\beta 1$ are now represented by the first symbol in the alphabet, here a. All points which were represented by mean values from $\beta 1$ to $\beta 2$ are now represented by the second symbol, here b, and so on. The resulting list of symbols is called a SAX "word".

The setting of the SAX parameters involves a compromise between the acceptable distance between the representation and the original series and the available computing resources. If the representation were to match the original series completely, there would be no data reduction.

3.4.4 Data Reduction

An important consideration is the temporal alignment of time series for batches where the durations of different batch phases may differ. This can occur for many reasons such as disturbances to the materials or environmental conditions (e.g. temperature of chilled water or cooling air) introducing changes in the magnitude of the driving forces behind the evolution of the process, which then changes the total time it takes to finish a given process step. Electronic batch records provide the means to do this by bookending the time phase with a manual prompt, which is answered by an operator (García-Muñoz et al., 2011). This method avoids potentially valuable information being lost through Dynamic Time Warping – such as when different phases started. Other work has ignored the differences in length of the time series and still achieved reasonable results (Gunther, Baclaski, Seborg, & Conner, 2009).

Batch	EV	SAX V1	SAX V2	SAX V3	SAX V4	SAX V5	SAX V6	SAX V7	SAX V8	SAX V9
Batch 1	pH	a	a	a	b	b	g	c	a	f
Batch 2	pH	a	a	a	g	d	a	a	b	d
Batch 3	pH	a	a	c	c	b	g	c	e	h
Batch 4	pH	a	a	a	e	d	f	a	b	d
Batch 5	pH	a	a	a	h	d	f	b	b	a
Batch 6	pH	a	b	a	b	b	g	c	e	h
Batch 7	pH	a	a	a	b	d	f	a	b	d
Batch 8	pH	c	a	a	a	d	a	a	b	d
Batch 9	pH	a	a	a	a	d	f	a	b	a
Batch 10	pH	a	a	a	c	b	g	b	b	h

Figure 4. SAX representation of time series across multiple batches

The SAX algorithm representation of the time series data divides the distribution into an arbitrary number of subsequences. This approach does increase the number of variables and the associated risk of an overfitted model, and so data reduction techniques are required to reduce the number of explanatory variables to the significant few. SAX subsequences that had little variability were eliminated, where variability is defined as the percentage of distinct values out of the total number of samples. At a setting of 10%, the windows highlighted above would be removed.

3.4.5 Modelling approach

The Random Forest algorithm parameters, such as mtry and ntree, were optimised by the caret package in R based on a grid search method. A range of values for each parameter was supplied to the algorithm and the best ones identified to minimise the Root Mean Squared Error (RMSE) of the model.

3.4.6 Cross-Validation

Due to the high dimensionality of the dataset, vigilance must be taken at every step of the data mining process to ensure overfitting does not occur. Many steps can be taken during the pre-processing phase, such as removing variables that do not vary beyond an arbitrary threshold (Figure 2) and limiting tree depth in the RF model. Lastly, the depth of the tree is limited to ensure the model can be generalised to a new dataset. One agreed approach to identify overfitting is by using repeated cross-validation (Tusar, 2016).

3.4.7 Cause and Effect and 5 Whys

A Cause and Effect Diagram provides structure to a brainstorming session by providing six categories to generate potential root causes to the issue (People, Equipment, Method, Measurement, Environment and Materials). The Hybrid Approach suggests two applications for the Cause and Effect session:

1) Generate potential root causes as to why the random Forest model highlights variables as significant. For example, some SAX subsequences highlighted as significant were associated with times when additions were made to the bioreactors.
2) Generate potential root causes to the issue that may not reside in the manufacturing dataset used to build the model. For example, variation contributed by the laboratory method or the sampling method.

The potential root causes are prioritised

3.5 Analyse/Evaluation

The output of the Analyse phase are experiments that rule in or out the potential root causes generated during the measure phase. Some potential root causes can be analysed using statistical methods such as Regression and Hypothesis testing. However, some may require physical experiments and in some cases manufacturing runs to confirm the potential root cause. Hence, this phase of the Six Sigma process is the most time consuming and costly.
The Hybrid approach proposes that if the fitness parameters indicate the RF model created during the Measure phase describe the behaviour of the manufacturing process accurately, the model can be used to simulate the effects of changing explanatory variable settings on the target variable. This “Digital Twin” approach would be faster and cheaper than running experiments on the actual process (Tao et al., 2018).

3.6 Improve/Deployment

The Improve phase looks to implement improvements to the confirmed root causes. For Six Sigma this can include error proofing techniques that provide an engineering fix that make the deviation impossible.

3.7 Control/Monitoring

The Control phase of a Six Sigma project monitors the target variable to confirm the improvements have been effective. If so, there is a requirement to develop a control plan to ensure the improvements are sustained once the project team disbands.
The analytics toolkit could enable automated monitoring of multiple target variables simultaneously. The R package "mail" can generate an email if specified criteria are met, such as a control limit breach on a target variable.

4 Results and Discussion

In this section, the results of the application Business Analytics approaches to the Six Sigma framework are explored, using the Cell Culture process as a case study.

A Random Forest algorithm was used to identify the most significant manufacturing process steps. By applying this approach, the CQA dataset could be prioritised using the Pareto principle to 6 out of 20 process steps – describing 85% of the variation observed.

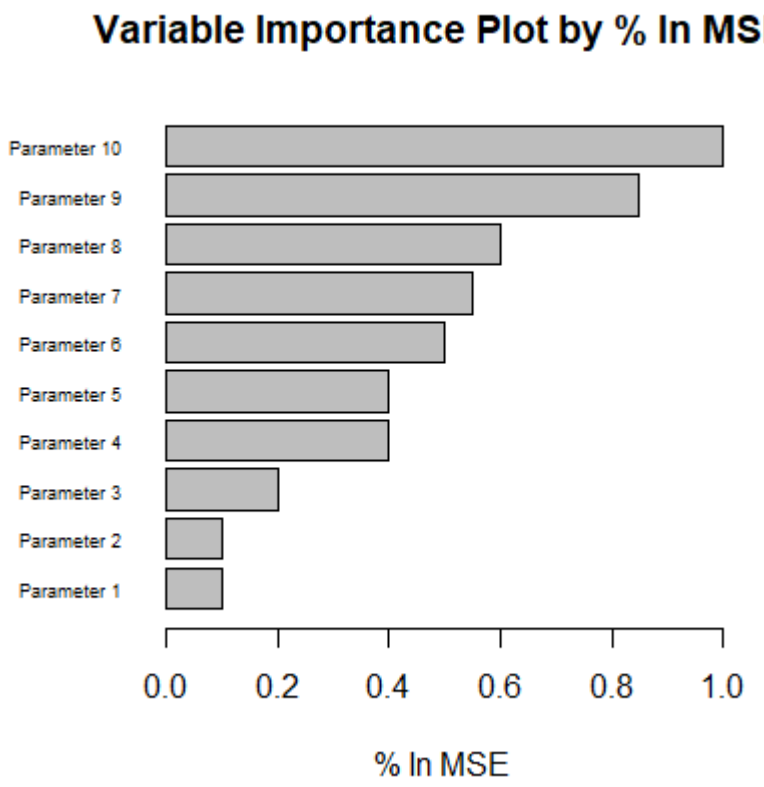

Figure 5: Variable Importance Plot (VIP) for Titre model

The VIP was vital to demonstrating to the manufacturing teams that this approach could provide meaningful process insight, as the team immediately began to generate hypotheses as to why these variables could have an influence. For confidentiality reasons the variables are masked, but examples include Dissolved Oxygen (DO), pH or the Oxygen flowrate.

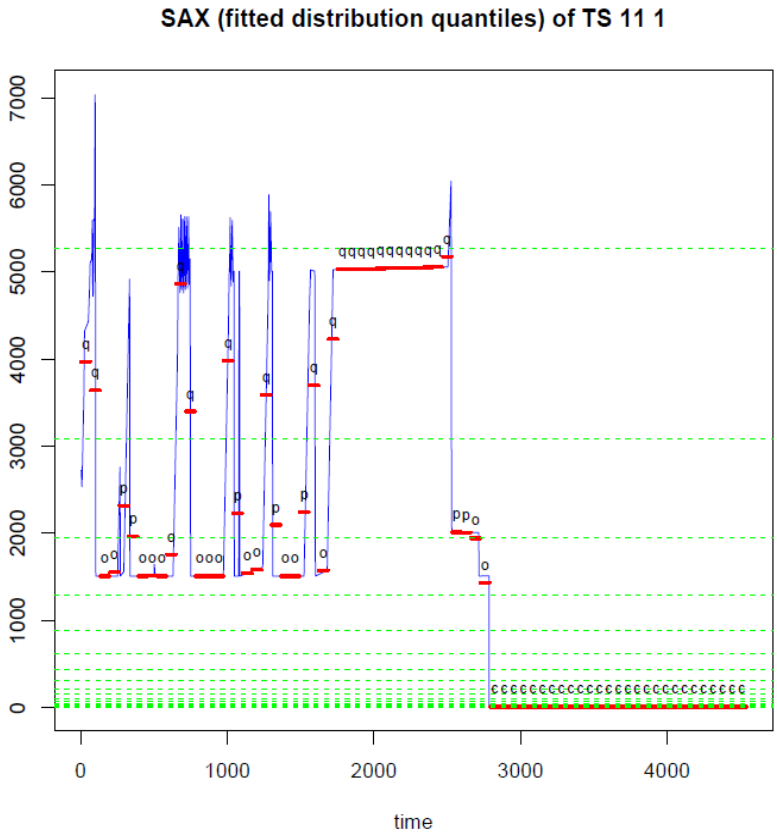

Figure 6: Example of SAX representation of a time series distribution

The initial step where the process teams identified the variables most likely to provide variability proved essential to the model accuracy. As stated in the methodology, this is an iterative process. If the manufacturing process is analysed in its entirety, the model recommendations tend to be made up of spurious correlations rather than any causative effect. This phenomenon is explored in more detail by Fan (Fan, Han, & Liu, 2014).

length	frequency	error	condition
1	0.389	0.000824	Parameter 1. V37 %in% c('d','f','j')
1	0.556	0.000804	Parameter 2.Day.5.5.g.L.<=4.535
1	0.444	0.00085	Final Bioreactor.Day.6.Parameter 3..mmol.L.<=4.54
1	0.222	0.001019	Parameter 4.V28 %in% c('c','f','h')
1	0.556	0.000704	Parameter 5. V24 %in% c('','b','e','h','i','j')
1	0.5	0.000956	Parameter 5.V28 %in% c('d','f','h','i')
1	0.333	0.000347	Parameter 6.V44 %in% c('b','d','g','h')
1	0.778	0.000766	Parameter 7.V33 %in% c('b','d','e','f','i','j')
1	0.333	0.000458	Parameter 8.V44 %in% c('e','g')
1	0.111	0.000625	Parametrer 9.V30 %in% c('','f')
1	0.5	0.000556	Final Bioreactor.Day.6.Parameter 10..mmol.L.>4.65035

Table 1. Example of manufacturing rules derived from the Random Forest model

The rules to be applied were chosen by fitness criteria such as importance to the model, the predicted result if the rule is satisfied and the error rate. As the rules increase in complexity, they become more challenging to implement, so unless the rule is exceptionally significant to the model the length should be limited to one explanatory variable as detailed in Table 3 above. Upon inspection, the prioritised SAX subsequences included by the algorithm in the manufacturing rules were found to correspond with the timings of events such as additions and pH adjustments, the timing of which are crucial. One example of a manufacturing rule is as follows:

If pH < 6.1 and the vessel temperature on day 3 (Subsequence 21) is a “d” or an “f” this correlates with a favourable value of the CQA.

One concern raised by the process scientists was that the model would recommend a setting that was outside the Normal Operating Range (NOR). This outcome is highly unlikely as the model is built using data which falls within the NOR, but this should be assessed on a case by case basis.

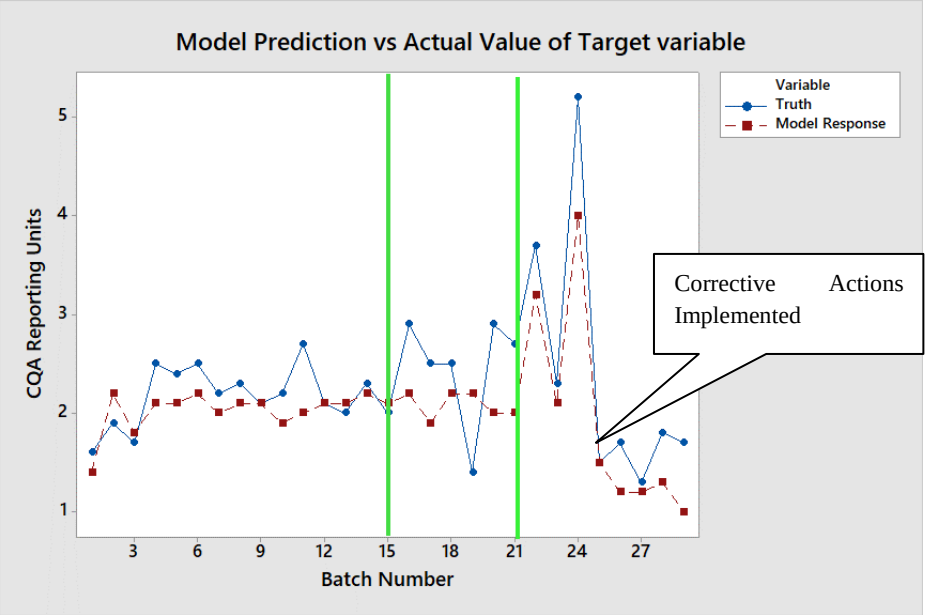

Figure 7. Model predictions for CQA model

Cross-validated performance measures:

Root Mean Squared Error	0.41
Model R-squared	90.4

Table 2. Model fitness parameter for CQA model

For the first fifteen batches, there is close alignment between the model predictions and the actual CQA values. The close agreement of the model prediction and actual value highlights the high probability that the root causes of the variability reside in the manufacturing dataset. This is also the case for the unfavourably high values for batch 22 and 24. Hence the manufacturing rules can be applied to reduce this variability.

At this point, the predicted value and the actual value diverge, highlighted between the two green lines on the chart. This drove the team to execute a Cause and Effect session to generate potential root causes to changes that occurred during this period outside of the manufacturing process. A contributory root cause was identified as a change in the sample handling method, which resulted in increased results.

4.1 Application to the manufacturing process

The approach was applied to a large-scale manufacturing process resulting in improvements in the target variables. The CQA model proved valuable to the problem-solving effort and reduced the lead time of the investigation significantly. The method identified process related root causes but perhaps more importantly identified when variation originated outside of the manufacturing process, such as sample handling or variability within sample testing.

4.2 Potential barriers to implementation

The project is a demonstration of how the intangible assets of human, information and organisation capital identified in Huang et al. can be aligned to improve innovative capability incrementally (Huang et al., 2011). This approach fits with the general practice of the adoption of new techniques by the manufacturing industry. While the biopharmaceutical manufacturing industry currently collects an abundance of data, it has often lacked the required skills to extract business insight. Historically this skills shortage would have driven larger companies to enlist external consultants to provide technical input and support. However, intellectual property constraints can limit a manufacturing company's ability to share data.
There also may be cultural barriers to the adoption of BA techniques as a means of improving efficiency. Manufacturing professionals presiding over complex processes tend towards conservatism due to the potential for unforeseen consequences when implementing change. This approach can benefit the customer who can be assured as to the steady state of the manufacturing process, a consideration especially applicable to biologics, where the supply of a vaccine can mean the difference between life and death (Hill & Waner, 2017). The approach described here allows BA to be inserted into existing quality processes in an incremental rather than disruptive fashion.

The benefits of the application of BA from a management perspective need to be understood and articulated clearly. One such benefit is the pace of root cause identification compared to traditional methods such as small-scale experiments. Further, the approach could address the loss of process knowledge by employee attrition. Data mining techniques provide a method for addressing the reliance on tacit knowledge to troubleshoot a complex manufacturing process. This may be a potential solution to high levels of attrition observed in biotech clusters by codifying process knowledge (Cooke, 2002).

The implementation of the model recommendations will need to be implemented with a thoughtful approach to risk-taking. The ideal state is the model is built in tandem with domain experts as it may be challenging to understand the scientific rationale behind to rules identified. This can lead to potential roadblocks to the implementation of these techniques in a highly regulated industry. A senior sponsor who is familiar with analytics and its potential applications to manufacturing is very beneficial in negotiating these hurdles.

4.3 **Further research**

This project details how the Business Analytics toolkit can improve the Six Sigma defect reduction process. The next intuitive step would be to explore how the BA toolkit could fit into lean manufacturing where the objective is improved efficiency.

Also, optimisation tools could be applied to the SAX output to identify the "best" distribution with a specific target variable. The regression problem presented here is unbalanced. Further work could assess if the creation of virtual samples by upsampling/downsampling would improve the model accuracy (Le, 2007).

5 Conclusion

The project demonstrates how the emerging field of Business Analytics can be integrated into the Six Sigma problem-solving framework. The proposed methodology was applied to two complex business-critical problems which would typically be the substrate for the Six Sigma process: Titre (Yield) improvement and root cause analysis of a CQA deviation.
The concept of fully automated production systems in the biopharmaceutical industry is no longer a viable short or medium term vision (Bannat et al., 2011), which is why the goal of this work is to provide a data mining approach which works in tandem with process experts, distilling insight from unrefined manufacturing data that would otherwise remain hidden. Most manufacturing companies do not have the facilities to execute small-scale studies, and even if they do, there are scale-up considerations between small and commercial scale findings.

The project demonstrates how to achieve actionable insight from the data that can be readily implemented on a manufacturing floor. There may be advanced algorithms within the deep learning family that are customised to avoid overfitting, e.g. Recurrent Neural Network with Dropout, but performance is sacrificed for interpretability.

With the momentum behind initiatives such as Industry 4.0, this project represents a low-cost proof of concept, with a potentially significant impact on process robustness and process efficiency. The accuracy of the model could also potentially be improved by the leveraging additional sensors, such as Raman spectroscopy to improve the explanatory power of the model.
The non-linear behaviour described by the application of Business Analytics techniques to biopharmaceutical data are often unknown coming out of process transfer from small scale. The identification of these patterns within the manufacturing dataset is a way of giving the process a voice that is unique to that process alone, without looking through the lens of scale-up considerations or comparable processes.

One limitation of the approach observed during the pilot was the large number of rules that need to be implemented and tracked to have a meaningful impact on process performance. Also, the business may be unwilling to progress even if strong correlations are identified if there is no intuitive scientific reasoning behind. Due to the importance of continued supply, manufacturing science teams should be involved in the projects from the start so that scientific rigour can be applied to the model findings.
The most significant implication of the approach is its potential to augment the Six Sigma framework. The analytics techniques presented here could expedite the resolution of these complex data-rich problems by an order of magnitude, mitigating the need for lengthy charting and visual comparison by process scientists. This approach also proves that an anomaly observed in manufacturing explanatory variables is correlated to a degree of certainty with the target variable fluctuation.

6 Declaration of Interest

This project was funded by the Irish Research Council Employment Based Program reference number: EBPPG/2016/384.

7 Acknowledgements

The authors would like to acknowledge the contribution of the project team: Jamie O'Donnell, Tim O'Regan, Neil Duggan and Dave O'Donovan. The authors also would like to thank the following for their support and guidance: Gerald Kierans, Tony Walsh, Simon Hancock, Lynn Smullen, and Eamonn Nixon.

Six Sigma : Dermot Gavin, Pat McNally, Dave Clancy, Georg Bernhard We wish to thank the School of Business at University College Dublin for making the project possible. We would also like to thank Gareth Thornton for his significant contribution in proofreading this article.

8 References

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.