Computers and Chemical Engineering 140 (2020) 106881

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems

Oliver J Fisher[a], Nicholas J Watson[a], Josep E Escrig[b], Rob Witt[c], Laura Porcu[d,e], Darren Bacon[e], Martin Rigley[e], Rachel L Gomes[a,*]

[a] Food, Water, Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK
[b] i2CAT Foundation, Calle Gran Capita, 2 -4 Edifici Nexus (Campus Nord Upc), 08034, Barcelona, Spain
[c] Totally Brewed, Unit 8-9 Wholesale District, Meadow Lane, Nottingham, NG2 3JJ, UK
[d] Energy Innovation & Collaboration, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, UK
[e] Lindhurst Engineering Ltd., Midland Road, Sutton in Ashfield, Nottinghamshire, NG17 5GS, UK

ARTICLE INFO

Article history:
Received 13 August 2019
Revised 1 April 2020
Accepted 20 April 2020
Available online 14 May 2020

Keywords:
Data-driven models
Process resilience
Waste valorisation
Mathematical modelling
Machine learning
Industry 4.0

ABSTRACT

The increasing availability of data, due to the adoption of low-cost industrial internet of things technologies, coupled with increasing processing power from cloud computing, is fuelling increase use of data-driven models in manufacturing. Utilising case studies from the food and drink industry and waste management industry, the considerations and challenges faced when developing data-driven models for manufacturing systems are explored. Ensuring a high-quality set of model development data that accurately represents the manufacturing system is key to the successful development of a data-driven model. The cross-industry standard process for data mining (CRISP-DM) framework is used to provide a reference at to what stage process manufacturers will face unique considerations and challenges when developing a data-driven model. This paper then explores how data-driven models can be utilised to characterise process streams and support the implementation of the circular economy principals, process resilience and waste valorisation.

1. Introduction

Fundamental to manufacturing is mathematical modelling (Hangos and Cameron, 2001), which utilises analogies to help understand the behaviour of a complex system (Cross and Moscardini, 1985). Models translate theories of how the world functions into the language of mathematics. Once built, a model can be used to aid in decision-making, develop scientific understanding, communicate knowledge and/or make predictions (Schichl, 2004). Manufacturers have utilised mathematical modelling for four primary applications: (1) planning and design; (2) monitoring and control; (3) process optimisation and (4) risk mitigation (Perry and Green, 2008). There are two branches of manufacturing, discrete manufacturing and process manufacturing. Discrete manufacturing consists of a bill-of-materials that moves between a set of manufacturing equipment, as it is cut and assembled together (Brandl, 2007). Whereas in process manufacturing, raw or waste materials flow through the manufacturing plant undergoing thermal, chemical and/or biochemical conversion (Fisher et al., 2018). This fundamental difference between the two causes numerous disparities between the manufacturing practices/processes (Brandl, 2007). Therefore, when constructing a model of a system, both discrete and process manufacturing have a unique set of considerations, challenges and opportunities. This paper shall focus on those faced in process manufacturing environments.

Within the field of mathematical modelling, there are two distinct branches: first principles modelling (often referred to as mechanistic models) and empirical modelling. First principle models build a series of equations by examining the workings of the system's individual parts (Schichl, 2004). First principle models rely on system understanding to compensate for lack of data. Because of this, they have a greater potential for extrapolation compared to empirical models (Mathews, 2004). Empirical models are mathematical equations derived from the analysis of data; therefore, requiring less knowledge of the system (Solomatine et al., 2008). Empirical models rely on the assumption that the data is of

* Corresponding author.
E-mail address: rachel.gomes@nottingham.ac.uk (R.L. Gomes).

https://doi.org/10.1016/j.compchemeng.2020.106881

sufficient granularity and/or quantity to define the system. However, if the system is defined within the data, empirical models are valuable tools for characterising the system input-output relationships, especially when there is limited engineering-domain knowledge to characterise complex systems (Luo et al., 2016). For example, empirical models have proven efficiency at modelling novel configuration bioelectrochemical systems (BES), which otherwise would require detailed knowledge of complex interactions between physical, chemical and electrochemistry principles (Luo et al., 2016).

Empirical models are already well utilised in process manufacturing because of the volume of data manufacturers produce (Rasmuson et al., 2014). Recently the capabilities of empirical models have greatly expanded due to advances in the fields of computational intelligence and machine learning, these new approaches are encompassed into the field of data-driven modelling (Solomatine et al., 2008). Computational intelligence are nature-inspired computational approaches to problem-solving; for example, algorithms that mimic the behaviour of animals (e.g. swarm intelligence) (Saka et al., 2013) or algorithms that replicate the behaviour of how humans solve problems (e.g. artificial neural network, ANN) (Kim, 2017). Machine learning focuses on the development of algorithms that can access data and use it to learn for themselves (e.g. support vector machines or random forest) and algorithms can belong within each fields (e.g. ANN) (Coley et al., 2018). Data-driven models (DDMs) are able to find relationships between the system state variables (input and output) without prior knowledge of the system (Angria et al., 2018); although, the incorporation of prior may enhance DDM predictive capabilities (Lauer and Bloch, 2008). Data-driven models derive the system's relationships by training an algorithm (e.g. linear regression, ANN, Gaussian process) on manufacturing data (Kim, 2017). Although data-driven algorithms have long existed (Ojha et al., 2017), their use in industry has been limited because of data constraints (e.g. lack of data, data not stored in useable format) and lack of computational processing power (Ge, 2017). There is an unprecedented rise in the volume of manufacturing data being generated due to the adoption of cyber-physical systems, smart factories and the industrial internet of things (IIoT) (Sadati et al., 2018). In 2015, it was reported that manufacturers globally generated more than 1000 Exabyte of data and by 2025 data generation in manufacturing will increase 20-fold (Yin and Kaynak, 2015). This means that the volume of data available to build DDMs has never been greater. Data-driven models were not always considered suitable for enterprise-wide modelling due to the computational cost in modelling vast volumes of data (Boukouvala et al., 2016). However, with the introduction of cloud computing, manufacturers now have affordable access to the processing power required to model large data sets (Ge, 2017). Subsequently, DDMs are becoming prevalent across industry for modelling and monitoring of plant-wide industrial processes (Ge, 2017).

There are well-known methodologies that provide a structured approach to developing a DDM. These include data mining and knowledge discovery in databases (KDD) (Fayyad and Stolorz, 1997), the cross-industry standard process for data mining (CRISP-DM) (Shearer, 2000), and sample, explore, modify, model and assess (SEMMA) (Shafique and Qaiser, 2014). Out of these the CRISP-DM is the most widely used methodology for developing DDMs and considered the *de facto* standard by industry (Mariscal et al., 2010). The CRISP-DM is composed of six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. Business understanding focuses on defining DDM objectives and requirements from a business objective. Data understanding is the collection and exploration of the data. Data preparation is the selection, cleaning and transformation of the data, in order to format the data from the next phase modelling. Modelling concerns the selection and application of various modelling techniques. Evaluation is the evaluation of obtained models and how to use their results. Finally, deployment focuses on utilising of the obtained results, knowledge and/or models to benefit the business. Developing a DDM is iterative as knowledge gained during the process may redefine the objectives and modelling approach, as shown in Fig. 1.

Utilising a structured methodology to develop a DDM helps the developer avoid common data modelling mistakes that may result in models built that exhibit poor generalisation and overfitting problems. Unless properly addressed these are the two main sources of error in DDMs (Kim, 2017) and defined as:

- *Generalisation*: is the capability of a DDM to fit and make predictions of data that was not used during the development of the model (Kim, 2017). The challenge of making the model performance consistently between data used to develop the model and new input data is known as generalisation.
- *Overfitting*: is the generation of a model that corresponds too closely or exactly to the noise (error) within the dataset, which negatively impacts future predictions (Srivastava et al., 2014).

Process manufacturers can face some unique challenges when facing generalisation and overfitting problems, due to the difficulty in defining the process manufacturing system. A manufacturing system is defined as an input stream(s) which passes through a process changing their physical and/or chemical nature into an output stream(s), which may consist of multiple products, by-products and/or waste material. Because of characteristics specific to process manufacturing systems (e.g. feedstock and waste variability, non-linearity of processes, product specification) deciding on what data is required and knowing whether the system is accurately represented in the model's development data is a challenge. This problem is further compounded by a system where data availability is limited (e.g. batch production of multiple products, frequent changes to manufacturing practise, implementation of new technologies). This is because the system is likely to contain regions underrepresented by the data, which possibly restricts the model's capability to make predictions in that region. Models of said systems will have to undergo frequent retraining as more data becomes available. These considerations and challenges have often been neglected in previous process manufacturing DDM (Charte et al., 2017; Ning and You, 2018; Sadati et al., 2018).

The CRISP-DM approach has been successfully applied to process manufacturing scenarios where data is abundant and the challenge has been extracting useful information from the data (Arce et al., 2018; Atzmueller et al., 2017). There have been several attempts to update the CRISP-DM approach from a manufacturer perspective (Harding et al., 2006; Soroush Rohanizadeh and Moghadam, 2009), but these approaches have taken a general view of manufacturing as a whole. Process manufacturers whose systems are poorly defined will face challenges specific to process manufacturers from data availability and variability. Therefore, the aim of this paper is to present the considerations and challenges unique to process manufacturers. This knowledge may benefit future manufacturers who wish to develop models of their systems and avoid overfitting and generalisation problems. These points are presented in the context of the CRISP-DM framework to provide a structure to the modelling task, as shown in Fig. 1. This may be used as a guide by process manufacturers and modellers to facilitate the development of DDMs for their particular application.

Through two case studies, the considerations and challenges faced when modelling process manufacturing systems will be presented, before discussing what new opportunities arise from data-driven modelling. Each case study project was performed with a small and medium enterprise (SME). The aim of these projects was to increase the economic and environmental sustainability of

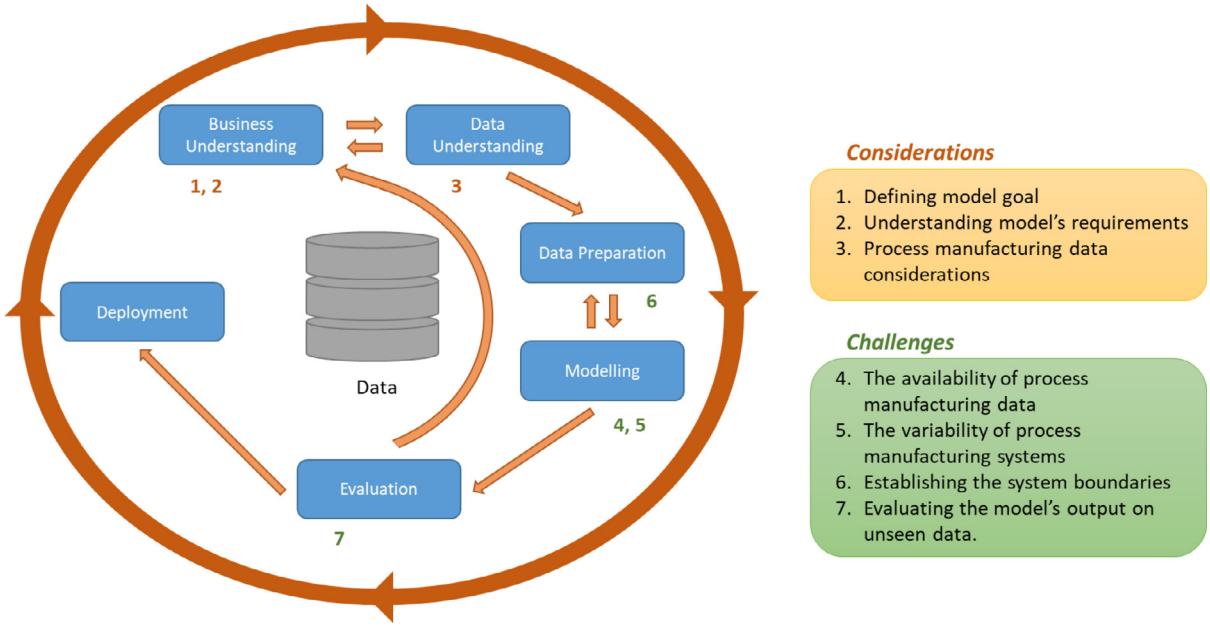

Fig. 1. Phrases of the CRISP-DM detailing updated to show the unique considerations and challenges faced when developing a data-driven model describing a process manufacturing system. Adapted from (Shearer, 2000).

Table 1
Summary of case studies (CSs).

	Waste/water management CS1	Brewery CS2
Manufacturing system	The H^2AD bioprocess is a circular economy technology, which treats wastewater to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy.	Fermentation is a critical process in beer production where the sugar in the wort is converted to alcohol.
Project's aim	To develop a model to analyse and predict the effect of wastewater variability (e.g. suspended solid content) has on the bioprocess performance. Utilise model to make predictions on bioprocess's ability to treat new feedstocks.	To develop a model to predict alcohol concentration from temperature and ultrasonic measurements during a fermentation process.
Feedstock characteristics	Farm waste which has variability due to changes in farm practices, the welfare of cattle, rainfall, seasons and more.	Raw ingredients are cereal grain, hops, water and brewer's yeast. Each ingredient contains inherent variation which can affect flavour, colour, carbonation, alcohol content and other subtle changes in the beer.
Product specification	Wastewater pollutants are reduced so water is fit for reuse. And biogas generated contains a minimum of 70% by volume of methane.	Slap in the Face beer containing 4% alcohol concentration.
Process description	The H^2AD is a semi-batch bioprocess, meaning that X volume of liquid enters and leaves the system at predefined time intervals. When entering the system the feedstock is heated to 30 °C. The process stream is recirculated through the system.	The main stages in the brewing process are: wort production, alcoholic fermentation and maturation. It is a batch process with the fermentation lasting between 3–6 days.
Model inputs	Water quality analysis parameters of feedstock (e.g. pH, dissolved oxygen, chemical oxygen demand, total suspended solids). H^2AD process conditions (e.g. temperature, system pressure).	Temperature, ultrasonic velocity and received ultrasonic signal amplitude, during fermentation.
Model outputs	Chemical oxygen demand (mg/L) and total suspended solids (mg/L) of the output stream. Daily biogas production (L) and methane concentration of biogas (% vol).	Alcohol concentration during fermentation.
Volume of data	Samples were collected once a week for a 1 year period, resulting in 52 data points.	Four datasets were collected from four batches of the Slap in the Face beer. Each data date set was made up of between 830–835 data points.

these companies through the development of DDMs which utilised data collected from each manufacturing process. Both projects presented a variety of considerations and challenges that the authors and SMEs had to overcome (Table 1).

1.1. Case study 1: waste/water management

SMEs are under pressure to reduce manufacturing costs and an effective waste/water management strategy is paramount. By 2050 there will be a 400% increase in demand for water by the manufac-

turing sector (OECD, 2012). Larger manufacturers have been utilising processes, like anaerobic digestion (AD), to valorise their waste to produce bioenergy and consider the resulting solid waste as fertiliser (Lin et al., 2013). The upfront capital costs of anaerobic digestion are a significant barrier for SMEs, which are generally more sensitive to additional financial costs (Rizos et al., 2015).

In the UK, Lindhurst Engineering Ltd. in partnership with the University of Nottingham has developed a technology called H^2AD Micro AD. The H^2AD bioprocess is capable of treating a variety of wastewaters, from sources including agriculture, brewing, soft drinks, foods, bio-manufacture residues, to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy. The H^2AD is a hybrid of anaerobic digestion and a microbial fuel cell, targeted at treating SME process manufacturer's wastewaters due to its modular, low-cost design.

Variations in the composition and characteristics of wastewater require management practices that can accommodate fluctuations in feedstock properties and process conditions, yet still deliver appropriate outputs without compensation in capital or operating costs. These wastewater variations, as well as potential environmental factors (e.g. seasonal), can decrease the efficiency of the bacteria to reduce pollutant load and generate biogas, which are the heart of the H^2AD bioprocess. This case study aims to develop a DDM aimed at understanding how variations in the input wastewater stream and H^2AD process conditions inform on the process performance. The performance is defined as the percentage removal of wastewater pollutants and volume of biogas generated. The model is trained on data collected from an operating H^2AD plant currently treating wastes produced by a 300 cattle dairy farm. The farm waste contains cattle slurry, bedding waste, waste milk, footbath, parlour washing and rainfall. The model will then be used to:

(1) Analyse and predict wastewater variations effect on the bioprocess performance and respond to energy challenges.
(2) Predict the bioprocess's ability to handle new feedstocks.

1.2. Case study 2: craft breweries

There are over 2000 craft breweries within the UK, most utilising equipment and processing techniques that have not changed for +30 years (Simmonds, 2017). This makes it an often unpredictable and inefficient process, especially regarding water overconsumption (Edmonds, 2016). A critical stage of the brewing process is fermentation, where yeast is added to the wort (the liquid extracted from the mashing process downstream) to convert sugar to alcohol (ethanol). The fermentation process is complete once the beer has reached the desired alcohol content and flavour profile. For SME breweries, this is currently determined by removing samples from the vessel and manually measuring the specific gravity with a hydrometer. Although the fermentation duration should be identical for each batch of a particular beer, this is rarely the case due to seasonal variability in ingredient (malts, hops, water) properties and natural fluctuations in process temperature. As specific gravity measurements are only taken every 4–10 h (or longer if no one is working overnight) this often leads to overfermentation, affecting product quality and resource utilisation.

Ultrasonic sensors can be used to monitor industrial processes such as equipment cleaning (Escrig et al., 2019) and multiphase flow (Al-Aufi et al., 2019). Previous research has shown that ultrasonic measurements can be used to monitor changes in ethanol volume concentration during beer fermentation processes (Krause et al., 2011; Resa et al., 2004). The authors of this paper are currently working on a project to develop a low-cost ultrasonic sensor designed for craft brewers. They are collaborating with an SME craft brewery in Nottingham, UK called Totally Brewed. This project records ultrasonic and temperature measurements during fermentation and uses supervised machine learning algorithms, such as artificial neural networks (ANNs), to predict the alcohol concentration from these measurements.

These case studies were chosen because they contain characteristics unique to process manufacturing systems (Table 1). Examples of the considerations and challenges resulting from these characteristics are referred to throughout this paper to support the topics discussed in each section.

2. Considerations for data-driven process manufacturing models

When developing a model, there are considerations, which must be made before the data collection and model building stages. Primarily, these considerations are

(1) Defining the model's goal;
(2) Understanding what is required from the model and,
(3) Data considerations (what data is already collected, what additional data points are required, how much data is needed, how the data is to be collected and how often).

Detailed discussions between the manufacturer and modeller must take place to address these considerations. The modeller (or team of modellers) may be an internal employee(s) or manufacturers may sub-contract data analysists and software specialists to model their processes. Digital companies, like Microsoft's Azure software, are now offering these services through the cloud (Microsoft, 2018). This service-orientated approach is often more affordable as they operate on a pay-as-you-go business model (Fisher et al., 2018). This means they are more readily available, especially for SMEs, who are understandably unlikely to have the required expertise within their current workforce or be able to afford the cost of developing, running and maintaining the models themselves.

2.1. Defining the model's goal

When constructing a mathematical model, the first stage is to identify and define the model's goal, as this will define the model's outputs (Cross and Moscardini, 1985). A model may have multiple goals; these may change over time and may come into conflict with each other. For example, in the waste/water management case study (CS1), the model's aim is to understand and predict how the system' variability affects the H^2AD bioprocess performance and make predictions how the H^2AD bioprocess will perform on new feedstocks. The performance of the process is defined as the percentage removal of wastewater pollutants and volume of biogas generated. The model's goal is to make predictions using the feedstock and H^2AD process data to determine the H^2AD process conditions that maximise the removal of key pollutants and volume of biogas generated. These two goals may come into conflict with one another; for example, maximising pollutant removal may adversely affect the rate of biogas production. Determining which goal to favour will be driven by current economic and environmental factors, customer demands and regulations. In the brewery case study (CS2), the model's aim is to predict when fermentation is complete and the beer has reached desired alcohol concentration and flavour profile. By doing so it will minimise over-fermentation, which wastes resources and results in an inferior product. This defined the model's goal of developing a model that could predict the alcohol concentration in real-time from affordable ultrasonic and temperature measurements.

Table 2
Examples of data-driven models requirements for the 2 case studies, waste/water management and brewery.

	Identify relationships	Make predictions
Waste/water management CS1	To understand how variations in both the feedstock's characteristics and process conditions affect the H^2AD bioprocess. In addition, identify which key variables in both strongly affect performance.	(1) To predict how variations in feedstock will affect the bioprocess performance and optimise the process conditions to maximise performance. (2) To predict the H^2AD's performance when treating new feedstocks to determine if they are financially viable
Brewery CS2	Models data to identify the relationship between temperature and ultrasonic measurements during fermentation to alcohol. Furthermore, to understand how variations in these measurements affect alcohol generation.	(1) Utilises the model to make a prediction of fermentation endpoint for the current batch. (2) To evaluate feedstocks (hops, malt, water, yeast) influence the fermentation time and adjust process conditions to accommodate.

2.2. Understanding the requirements from the model

A model's goal will inform on what function the model is to perform and what degree of accuracy is required in the model outputs. A DDM may perform two functions:

(1) **Fit data** (by regression, classification and/or clustering) to identify relationships between input, output, process and material properties.
(2) **Predict** process performance and resource use.

Dependant on the goal, a DDM may only be required to perform function (1) or it may perform both functions (1) and (2). Data-driven models are able to model highly complex, nonlinear systems to identify these relationships (Pasini, 2015). Dependant on the manufacturer's requirement, DDMs can also be utilised for predictions (Almeida, 2002). However, the complexity of the model may increase, which will be dependant on the required accuracy of the model's predictions (Almeida, 2002). Table 2 gives examples of both functions (1) and (2).

An important consideration for the manufacturer is assessing the required accuracy between the model's output and the true value. Is it important to the manufacturer whether the model is able to make predictions to 99.9% accuracy or is a model with reduced accuracy sufficient to complete the task? According to Dean Abbott, president of Abbott Analytics, it is more desirable to a manufacturer to build a model that meets the requirements for the task at a cheaper cost than to invest more money in unnecessary accuracy (Garment, 2014). The accuracy of a DDMs is widely defined by the norm of the differences ("residuals") between estimates and observations (Duarte et al., 2004). The required accuracy will be determined by any one or a combination of the following factors:

(1) **Product specification:** when DDMs are utilised to predict a product's specification, the model's accuracy is influenced by the acceptable tolerance around the product specification. The tolerance may be determined by regulations, customer demands and/or economic feasibility.
(2) **Regulation(s):** manufacturers are subject to regulations regarding the accuracy to which they state the composition of their product, co-product and waste streams. For example, beers of an alcoholic strength not exceeding 5.5% volume that are sold within the European Union are subject to a 0.5% tolerance (European Union, 2011).
(3) **Economic:** when predicting an output that has a strong correlation to the economic feasibility of the process. Dependant on how precarious the process is between profit or loss making will influence how accurate the DDM is required to be. For example, when predicting the product yield of yield-driven processes. Yield-driven processes are productions processes whose economic feasibility are substantially influence by-product yield.
(4) **Safety:** when making predictions that influence the safety of the manufacturing system the DDM's accuracy is influenced by the severity and likelihood of the hazard. For example, when predicting hazardous gas dispersion under complex terrain conditions a high level prediction accuracy is required (Wang et al., 2018).

2.3. Process manufacturing data considerations

Data-driven models can be harnessed to support the development and implementation of new technologies and processes (Qiao et al., 2003). However, this requires the generation of a set of data, which can be costly and time-consuming (Sadati et al., 2018). Ensuring the data collected is representative of manufacturing the system is paramount to the performance of DDMs (Batista et al., 2004). To achieve this there are four questions to consider when generating a set of data:

(1) What data already exists, to what granularity (the scale and level of detail in the data) and how representative of the system is the data?
(2) What volume of data is required for model development?
(3) How was the data collected?
(4) What, if any, additional data will the model require and how will it be collected?

Manufacturing companies today are collecting vast troves of process data but typically use them only for monitoring purposes and after the event analysis, and not as a basis for improving process resilience (Sadati et al., 2018). As part of the data consideration stage, the modeller will need to decide on the optimal number of input and output variables the model will require. This is achieved through feature selection and dimensionality reduction techniques, which improve the model's performance by identifying and removing noisy and/or irrelevant variables from the data (Sadati et al., 2018). Irrelevant variables are variables that have little to no impact on the model's output. Conversely, there exist process manufacturing systems that are not currently measuring/storing data to the required granularity to capture the system being modelled within the data. Therefore, a plan needs to be developed to decide which additional measurements and what volume of additional data is required, as well as how this data will be collected. For example, in CS1 the H^2AD bioprocess was already collecting 7 process performance variables but no data on the characteristics of the incoming or outgoing wastewater streams. Nineteen water quality variables were identified for initial collection, though with appreciation that not all of these would be truly relevant to modelling the output data. This made for a total of 26 input variables that was reduced to 9 by feature selection followed by principal component analysis. The volume of data is determined by three factors.

(1) The data needs to extend to, and preferably beyond, the system boundaries. Testing the model's capability to fit and pre-

dict beyond the system's boundaries is important to establish the confidence boundaries of the model.
(2) The duration period of data collection must be sufficient to capture any temporal variations observed in the system.
(3) There must be an equal distribution of the data within the system boundaries, to ensure the model is capable of fitting data and making predictions throughout the system.

If it is possible to collect data that achieves these three factors efficiently, then the volume of data required will be less. This can be done using Design of Experiments (DoE) techniques, such as central composite design and Box-Behnken design, to systematically collect data from across the system (Hamid et al., 2016). However, collecting data in industrial environments has numerous challenges (see Section 3) and a less intrusive method may be required, see (Section 3.3).

Data is collected either manually or automatically. Data collected manually is time-consuming and introduces the possibility of human error (Skoogh et al., 2012). Whereas, data collected automatically generally increases the volume of data recorded and removes human error (Skoogh et al., 2012) However there is a capital cost in purchasing and installing the required systems to collect the data automatically. Data-driven models often benefit from the automatic collection of data, as they can react in real-time to changes in the input variables. The DDM's response can feed into a larger online system creating a feedback loop. However, automatic data collection comes with challenges associated with handling large datasets. When handling a large dataset, the five V's are often employed to characterise the data (Addo-Tenkorang and Helo, 2016):

- **Volume:** refers to the size of the datasets.
- **Velocity:** is the speed to which data is generated, collected and analysed.
- **Variety:** is defined as the different types of data (e.g. time series data, image, audio, log files) incorporated into the datasets.
- **Value:** refers to the value extracted from the data.
- **Veracity:** the trustworthiness and quality of the data.

3. Challenges for data-driven process manufacturing models

Process manufacturers face unique challenges when collecting a set of data that is representative of the manufacturing system. The machine learning techniques utilised by DDMs have increased capabilities to fit data and make predictions when more data is provided (Kim, 2017). However, when ensuring there is sufficient data, and that it is representative of the system, there are a number of challenges in:

(1) The availability of process manufacturing data;
(2) The variability of process manufacturing systems;
(3) Establishing the system boundaries;
(4) Evaluating the model's output on unseen data.

3.1. Availability of manufacturing data

Data collection from real industrial processes faces a plethora of well-known challenges: sampling frequency, spatial representation of process environments, incomplete data, working conditions, sensor malfunction, communication exception or database shutdown, accuracy, etc. (Shang et al., 2014; Souza et al., 2016). This can result in poor quality data for model development that may require cleaning. Cleaning may be necessary to remove outliers within the data that are not representative of the system; however, the challenge is in knowing what data is and isn't representative of the system. Although DDMs developed from machine learning algorithms can overcome these challenges, large volumes of data are required (Qin, 2014). This is not always possible from process manufacturing systems, as they may be subject to additional challenges concerning the availability and variability of data. An established process manufacturing system may already have a wealth of historical data available for the initial training stage. However, this may not always be the case and the reasons for this may include:

(1) Modelling a new or adapted process, meaning no historical data representative of the current process exists;
(2) Variables (e.g. process conditions, external influences, new or changing feedstock) required for the model were not previously measured;
(3) Manufacturers may not have stored their data or not stored their data in a useable format;
(4) Manufacturers may produce a variety of products, meaning data on any one product is limited;
(5) Manufacturers may not keep potentially damaging data to the business if not required to.

Furthermore, industries that produce a variety of products from the same set of processing equipment (common in the pharmaceutical and food and drinks industries) will require the creation of a dataset for each of their products; as the process conditions, feedstocks and resources will be specific to that product. This data may not be immediately available, as these industries tend to produce a batch of one product then switch to another. This means the time between batches of the same product can be extensive delaying the development of the model. The CS2 faced these challenges, where craft breweries tend to ferment a range of different beers and a model would need to be developed/adapted for each beer. The beer in question for this study is currently only brewed once a month so this means only 12 data sets could be recorded each year, which may not be sufficient for the machine learning models. Equally, for CS1 a new dataset will be required for optimising the H^2AD bioprocess's performance on new feedstocks (e.g. a different farm's waste, food waste).

The challenge of process manufacturing data availability will be less to manufacturers that utilise industrial control systems as part of their manufacturing systems. Systems like SCADA (supervisory control and data acquisition) have existed since the 1970's and are used to monitor and control a plant in industries such as wastewater treatment, waste management, energy, oil and gas refining and food production (Qin, 2014). Industries that make a consistent product and have extensively utilised industrial control systems (e.g. wastewater treatment or oil and gas refining) are well placed to recover value by developing a DDM to discovery knowledge within their historic data (Qin, 2014).

3.2. Variability in manufacturing systems

All manufacturing systems contain inherent variation (e.g. fluctuations in process temperatures, pressure, and flowrates, human operators, leakages) which can affect the performance of a DDM. This variability may be overcome by collecting sufficient quantity of relevant data as the variability will be captured within the model (Kay et al., 1999). However, this is not always possible for process manufacturers (see Section 3.1). Process manufacturers face the further challenge from the variations present in the resource flow, particularly when the resource flow is a waste stream. The feedstock physical and compositional variability can have a significant impact on the biochemical and thermochemical conversion to the final product (Williams et al., 2016). This variation can occur for a number of reasons, as summarised in Table 3.

During CS1, the H^2AD bioprocess was affected from variations in the feedstock supply, caused by changes to farm practice. There is an inherent variation present in the feedstock, as shown by the fluctuations in concentrations of four key water quality parameters

Table 3
Causes of variation in process manufacturers' feedstock and some industry examples.

Cause of variation	Industry example
1. Processes upstream	Water is an essential brewing ingredient. Inherent variability in the water treatment methods employed upstream has a direct impact on the water's characterises, and thus the final beer's characteristics (Simate, 2015).
2. Changes in operating practices upstream	Farming practices vary throughout the year. For example, the practices of spreading the farm waste stored in a slurry tank increases between the months of April and September, as per the EU waste framework Directive 91/676/EEC (European Union, 1991). This affects the slurry's characteristics (including total suspended solids, metal content and nutrient and organic load) and impacts processes utilising it as a feedstock, e.g. the H^2AD bioprocess.
3. Changes in supplier	Anaerobic digestion is an attractive option for the valorisation of food waste and other wastes with high organic load and calorific value. However, instability of anaerobic digesters is a common problem that can be exasperated by changes to the feedstock (Fisgativa et al., 2016). Food waste characteristics vary hugely dependant on its source, as well as being inherently variable (Fisgativa et al., 2016). Introducing a new feedstock to a digester can reduce biogas generation as the system takes time to adapt (Zhang et al., 2014).
4. Local agronomic conditions	Geographic location affects feedstock characteristics through variations in local agronomic conditions (Williams et al., 2016). The structure of corn harvested in the US has been shown to have a stronger correlation to geographic location than genetic variety (Templeton et al., 2009).
5. Cultivation and harvesting practises	There is extensive exploitation of plants for active ingredients for drug development (Ncube et al., 2012). The active ingredient derives their therapeutic effects from secondary metabolites, which is influenced by numerous natural factors (Ncube et al., 2012). The timing of harvesting and/or handling of the plant material also has an impact on plant quality (Ncube et al., 2012).
6. Seasonal	Sewage sludge generated by wastewater treatment plants is often used for agriculture as it recycles nutrients and organic matter to land. However, there is an environmental threat from the heavy metal content in the sludge (García-Delgado et al., 2007). The heavy metal content has been shown to vary by seasons and understanding this variability is important to minimise any negative environmental impact from using the sludge as a fertiliser (García-Delgado et al., 2007).
7. Storage and transportation	Potatoes degrade after they are harvested, at a rate determined by the storage: temperature, relative humidity, air circulation and gas composition (Eltawil et al., 2006). During transportation, the potatoes' quality is further reduced by bruising (Eltawil et al., 2006). Food manufacturers receive potatoes from multiple suppliers. Therefore, classification is required to determine if a potato quality is sufficient for further food processing (Lopez-Juarez et al., 2018).

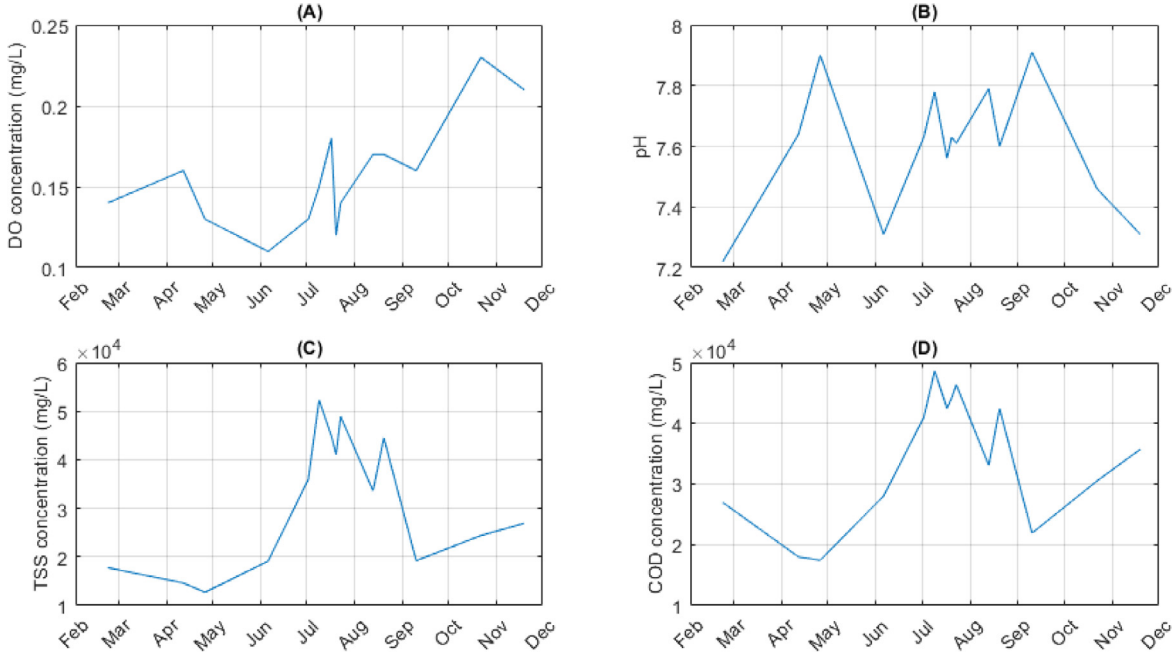

Fig. 2. Concentration of four water quality parameters of H^2AD feedstock during 2018: (A) dissolved oxygen (DO), (B) pH, (C) total suspended solids (TSS) and (D) chemical oxygen demand (COD).

in Fig. 2. The feedstock is stored in a slurry tank open to the environment, meaning during the winter months the total suspended solids (TSS) would decrease as the feedstock became diluted by increased rainfall.

There are further problems that arise from unexpected variability between batches of the same product. The brewery CS2 helps to illustrate the variability within supposedly identical processes. Fig. 3 shows the ultrasonic velocity, received ultrasonic signal amplitude and temperature as a function of fermentation time for four batches of a particular beer. Each of these measurements was recorded during different months in 2018, and are the input variables for a DDM predicting alcohol concentration during fermentation. The results indicate that the majority of variation within the signal is caused by temperature variations, which is typical of ultrasonic measurements. However, although the ultrasonic results show the same overall trends there are other variations caused by feedstock variability and marginally different brewing conditions (e.g. volume in the fermenter). These will affect the performance of the machine learning models and affect their accuracy. In Fig. 3, batch 4 shows a totally different temperature and speed of sound profile to the other batches. This was the result of a failure of the temperature control system during fermentation. Therefore,

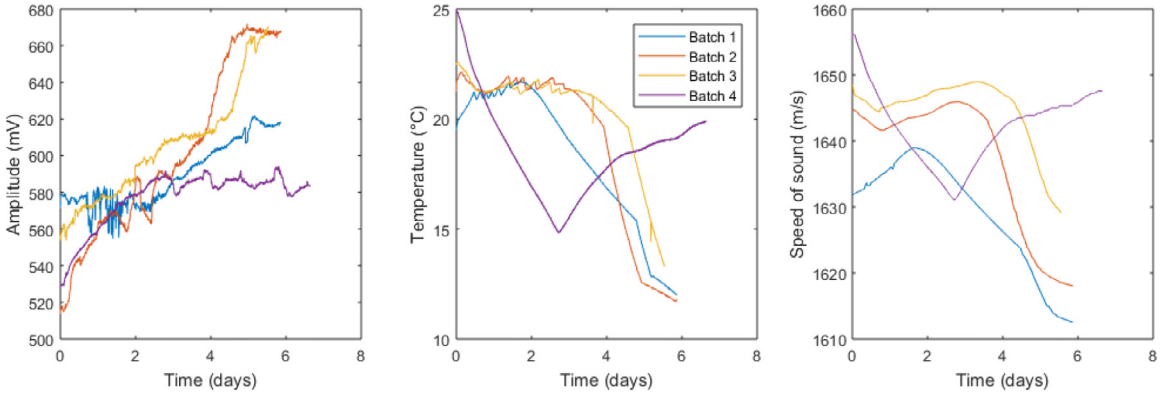

Fig. 3. Amplitude, temperature and speed of sound data during fermentation of four batches of Slap in the Face beer, collected from Totally Brewed in 2018.

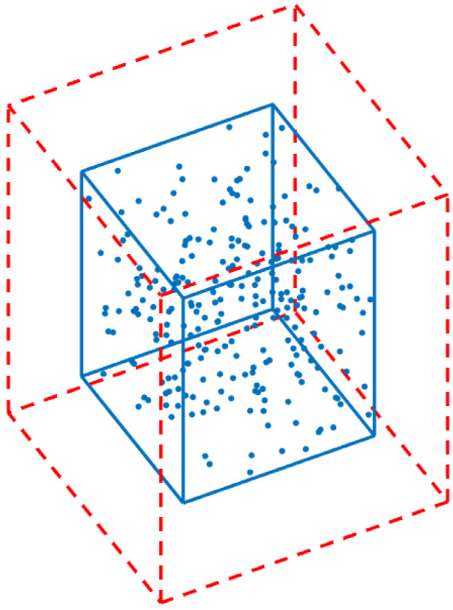

Fig. 4. Figure showing model development data that is only representative of one part of the whole manufacturing system. Where blue dots are model development data, the blue solid box represents model boundaries and the red dashed box represents system boundaries.

this batch was not representative of the system and is not suitable for training the model, further limiting the data available for model development.

3.3. Recognising the manufacturing system's boundaries

Predictions are a key feature of DDMs for process optimisation and intelligent decision-making. Data-driven models' predictive capabilities are generally strongest within the boundaries of the data used to develop them (Kim, 2017). If the data is only representative of a subsection of the manufacturing system, the model will fail at making accurate predictions of the entire system. Fig. 4 helps to visualise this problem. The model will be able to make accurate predictions for data inputted that is within the blue box but its prediction capability will likely decrease in accuracy the further away from the model boundaries. There are two methods to ensure the manufacturer can be confident the in model's results:

1. knowing the system boundaries and ensuring the model development data extends to these boundaries;
2. knowing the model boundaries and ensuring that only predictions made from this within the model development data range are acted upon.

Knowing the system boundaries of a process usually requires prior knowledge elicited from process industry practitioners (Shang et al., 2014). However, when modelling a new or changed system this knowledge may need reinforcement from trends identified within the model development data. Data visualisation plays an important role when understanding the space the system operates within (Lee and Ong, 1996). It is simple to visualise the boundaries' of a system of two or three dimensions, as demonstrated in Fig. 4. However, process manufacturing models typically contain data of higher dimensionality that will be impossible to visualise using conventional two-dimensional Euclidean space (Wang et al., 2004). There exist a number of approaches to plotting high dimensional data (Carr et al., 1987). Approaches like the scatter plot matrix can be utilised to plot the individual relationships between the variables alongside one another (Carr et al., 1987). The parallel coordinates system representation is particularly useful in visualising process manufacturing data in one plot (Wang et al., 2004). Wang et al. developed the Scan Circle algorithm that is able to identify regions of interest within the data that can be used to help describe the feasible region a DDM may operate successfully (Wang et al., 2002). Once the modeller is confident in the boundary conditions the model development data is selected from the historical data, ensuring that the model development data extends to these boundaries. A scatter plot matrix was used to understand the model boundaries for the feedstock characteristics in CS1, a section is shown in Fig. 5. The diagram clearly shows that TSS is between 10,000 and 20,000 mg/L for the majority of the training data, implying that predictions for when TSS is greater than 20,000 are likely to be less accurate. It also indicates that there is a strong positive correlation between TSS and chemical oxygen demand (COD).

If no historical data exists, then data needs to be collected and it is likely that the system boundaries are yet to be defined. This can be overcome by collecting data over a suitable number of batches and/or timeframe of a continuous process, to be confident that the system is modelled within this data. This may be a time-intensive approach, as data may need to be collected for up to or beyond a year. For example, the process may be influenced by seasonal variations. An alternative method is to conduct a DoE, to determine a robust set of experiments that will define the system limits and generate data to extend to these boundaries (Hamid et al., 2016). By using DoE the data collection time and volume of data may be greatly reduced; however, the manufacturer may face additional costs from process disruptions necessary for experiment execution to achieve a robust dataset (Sadati et al.,

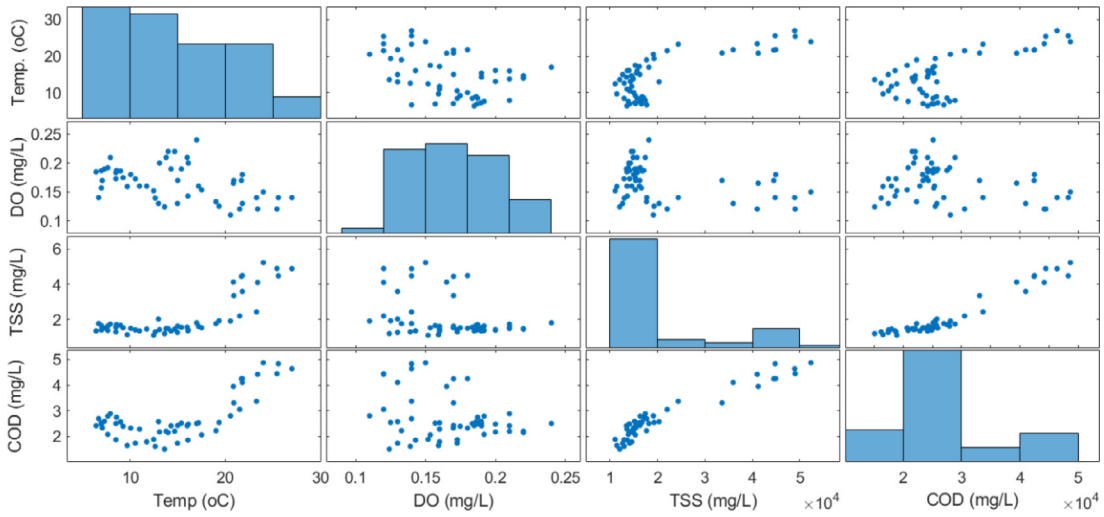

Fig. 5. Scatter plot matrix for four of case study 1′s feedstock's characteristics [Temp.: temperature; DO: dissolve oxygen; TSS: total suspended solids; COD: chemical oxygen demand].

2018). Dependant on the potential economic benefit of the model, it may be worth the additional cost or it may prove more feasible to conduct pilot-scale work to generate the data.

Prediction made by DDMs will always contain a degree of uncertainty. Uncertainty quantification is concerned with understanding the impact from the uncertainties inherent within the input data on the model's outputs (Iskandarani et al., 2016). There are a wide spectrum of techniques to consider when undertaking uncertainty quantification including: Bayesian methods, Latin hypercube sampling, polynomial chaos expansions, stochastic finite-element methods, Monte Carlo (Owhadi et al., 2013). Owhadi et al. proposed an optimal uncertainty quantification framework, which may be used as a guide for process manufacturers when accessing a DDM's predictive performance across the manufacturing system boundaries.

When implementing a new model into a system, careful monitoring is initially required to ensure the model is reflective of the system. Particularly if there is variability in the feedstock and other external factors. Once successfully installed, the model requires monitoring to prevent a gradual degradation in performance. The decrease of the prediction quality is caused by the gradual and abrupt changes in the process (Kadlec and Gabrys, 2009). This can be avoided by retraining the model periodically as the availability and collection of data increases. Regular retraining often proves expensive and "online-learning" is an alternative approach to developing a DDM. Online learning is a different approach to machine learning where models continuously evolve as data becomes available sequentially in time (Chandrasekaran et al., 2012). This will be particularly effective for process manufacturers modelling data collected in real-time from control systems or IIoT.

3.4. Evaluating the model's output on unseen data

Model evaluation is necessary to determine how accurately the model reflects the system (Kim, 2017). Data is partitioned from the model development data to evaluate the model at different stages during its development. The model development data is partitioned into training, validation and testing datasets (Bishop, 2006). Training data is the data fitted to the model's algorithm, whilst validation data provides an evaluation of the model's fit to the training data and is used for tuning the algorithm's hyperparameters. A hyperparameter is an adjustable parameter that must be either manually or automatically tuned in order to obtain a model with optimal performance (Zeng and Luo, 2017). The testing data is used to evaluate of the model's fit on the model development data. While this may be sufficient when modelling systems whose boundaries are strongly defined, an additional evaluation will be required for systems whose boundaries are loosely defined. This will be likely for process manufacturers whose system may have challenges from data availability (Section 3.1) or from high levels of variability (Section 3.2). These systems will require additional evaluation using "unseen data", which is additional data not used during the development of a DDM. Unseen data is able to better evaluate the model's extrapolation capabilities beyond the model development data (Panerati et al., 2019). If the accuracy of the model does not meet the requirements defined by the manufacturer (see Section 2.2) the model must undergo redevelopment either through optimisation of the algorithm's architecture or the collection of additional data (Keviczky and Banyasz, 2015).

Process manufacturers that face data availability and variability challenges are more susceptible to overfitting and poor generalisation, as capturing these challenges increases the likelihood of that the model will exploit relationships within the data that do not describe the manufacturing system. Evaluating the model on unseen data is essential to avoid overfitting and ensure good generalisation, yet this stage has often been overlooked (Hamid et al., 2016). Unseen data has two sources:

1. **Partitioning model development data:** split data into model development data (training, validation and testing data) and unseen data.
2. **Experimental data:** an experiment is performed to generate unseen data that covers the range of the model boundaries. Design of experiments can be used to ensure the parameter range for the experiment is comprehensive.

Partitioning data is the most common approach (Liu and Cocea, 2017). The data partitioning can be performed randomly or by using a fixed method. However the data is partitioned, the unseen validation data must extend evenly to cover the boundaries of the model. This is to ensure the model is able to both fit and predict data through the manufacturing system. The unseen data should also aim to go beyond the limits of the model. By using data beyond the model boundaries the modeller gains knowledge of the model's capability to fit and predict outside of the training data.

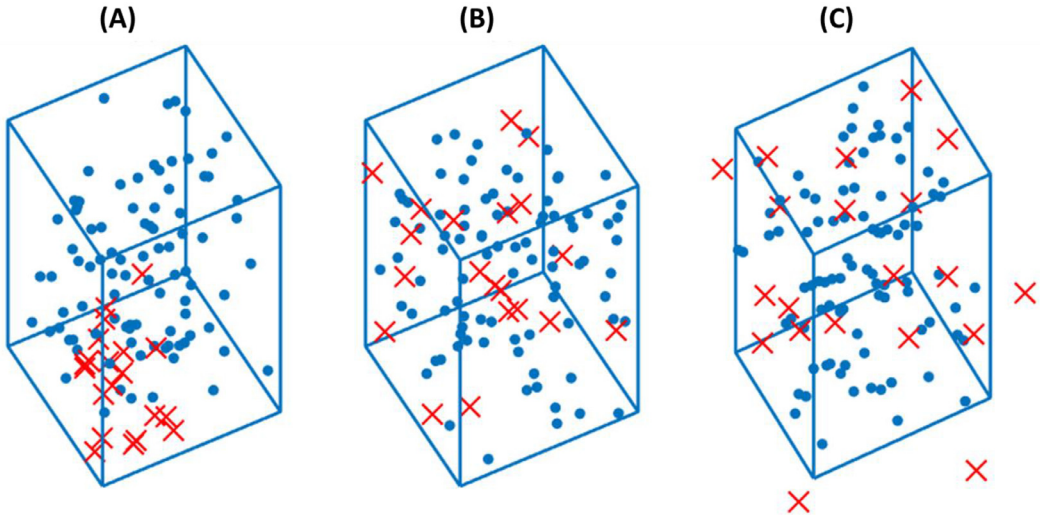

Fig. 6. (A) Example of unseen data only evaluating one region of the model, (B) example of unseen data evaluating the whole of the model, and (C) evaluating the whole of the model and the extrapolating capabilities of the model. Where blue dots are training data, red crosses are unseen data and blue box is the model boundaries.

An example of good and poor selection of unseen data is shown in Fig. 6.

4. Opportunities arising from data-driven modelling of process manufacturing environments

In an age where manufacturing is under pressure to reduce cost and environmental impact, while maintaining product quality, the utilisation of data-driven modelling is a promising tool for process manufacturers. When the Aberdeen Group surveyed 223 global manufacturing organisations, 47% believed they needed to become more data-driven to remain competitive (Geiger, 2017). Four areas have been identified as an opportunity for process manufacturers to strengthen their manufacturing systems through the utilisation of DDMs:

(1) Utilising data-driven models to improve process manufacturing models;
(2) Enabling affordable characterisation of process streams;
(3) Ensuring greater process resilience;
(4) Evaluating waste valorisation potential.

4.1. Utilising data-driven models to improve process manufacturing models

The application of DDMs to support and improve traditional modelling tasks has been well established. Examples exist of their use for design and planning, control and monitoring, optimisation and safety (Boukouvala et al., 2016; Krenczyk, 2012; Pan and Hu, 2016; Wang et al., 2011). Furthermore, techniques are being designed to integrate data-driven and first principal models into a hybrid model structure (von Stosch et al., 2014). Within process manufacturing hybrid models have been developed for various application including chemical reactors (Azarpour et al., 2017), polymerization processes (Fiedler and Schuppert, 2008), crystallisation (Nicoletti et al., 2009), metallurgic processes (Hu et al., 2011) and distillation columns (Caballero, 2015). Hybrid models are built by combining the predictions of a first principle model and a DDM into a single model. Hybrid models embrace the benefits of both techniques and overcome the disadvantages of both models (Azarpour et al., 2017). There are multiple methods by which the models may cooperate (von Stosch et al., 2014):

- **Proxy:** one model acts as a surrogate for the other;
- **Complement:** the solution is a combination of the two models;
- **Supplement:** one model provides a correction for the other model;
- **Embedment:** one model is embedded within the other model;
- **Integrate:** the output of one model serves as an input for the other model;
- **Inspiration:** the structure of one model is developed from knowledge provided by the other model.

Hybrid modelling is considered the state-of-the-art modelling techniques to model complex manufacturing systems (Barbosa and Azevedo, 2017). When compared to individual models, hybrid models tend to have a higher prediction accuracy, better calibration properties, enhanced extrapolation capabilities and better interpretability than DDMs (von Stosch et al., 2014). Process manufacturers that rely on existing first-principle models have an opportunity to improve the models results and prediction accuracy by developing and integrating a DDM into the existing model. However, manufacturers will face additional challenges in developing hybrid models. Developing hybrid models requires knowledge about different modelling techniques and flexibility from modellers to find a good fit between models (Barbosa and Azevedo, 2017). When deciding whether the development of a hybrid model is justified over a single model the manufacturer should considered if the model goal and requirements demands a hybrid modelling approach. If the model accuracy can be achieved through conventional single modelling techniques then the additional time and expertise demanded by a hybrid model is unjustified.

4.2. Enabling affordable characterisation of process streams

One of the main benefits of DDMs is their ability to enable real-time decisions by the collection of real-time manufacturing data (Chaturvedi et al., 1993). However, process manufacturing's feedstock, product and waste streams are not a set of specific discrete components. Instead, they are a multi-component and/or multi-phase matrix subject to inherent variation. Therefore, real-time online characterisation is expensive and sometimes unfeasible. This results in important process variables, like product quality, being measured infrequently offline (Yan et al., 2017). Because of the infrequency in which these measurements are performed, continuous direct monitoring of process streams' composition is not possible (Slišković et al., 2011).

Industrial processes are often described as *data rich but information poor*. There is often a large quantity of process data from conventional measurements (pressure, temperature, flow rate, etc.) but a lack of data on stream compositions. This is due to a lack of suitable and affordable technologies capable of performing these measurements non-invasively, online and in real-time. (Dong and Mcavoy, 1996). To overcome this, process manufacturers rely on soft sensors, which are able to model data collected from conventional measurements and used to predict key variables (Kleinert et al., 2011). As more novel IIoT technologies are introduced, the variety of data available to train soft sensors shall expand. One such example is a multi-sensor system designed to optimise the Clean-In-Place (CIP) process, which utilises both visual and ultrasonic data (Simeone et al., 2018, 2016). By utilising DDMs, soft sensors are able to compute not only a wider range of data but also unclean data (Qin, 2014). This will allow for the development of cheaper sensors, as the requirement for high-quality data shall decrease, instead being replaced by volume of data (Qin, 2014). However, the data must always be representative of the system.

Being able to track the relevant composition and physical characteristics of a process stream in real-time, will become even more important to process manufacturers as regulations may change so that manufacturers have more responsibility to record and share data for compliance. Combating fraud in certain process manufacturing industries is an ongoing issue, with a particular focus on the food and drinks sector (Manning and Soon, 2016). The introduction of a greater number of affordable soft sensors would help build trust across the supply chain and with consumers. If this is to be successful, ensuring that these sensors are not susceptible to malicious interference or corrupted with false data is paramount.

4.3. Ensuring greater process resilience

As with all suppliers, process manufacturers are required to meet certain targets to be commercially successful and adhere to regulations. Paramount amongst these, from a manufacturing process viewpoint, is meeting minimum product specification (e.g. yield and quality) (Gani, 2004). Ensuring consistency for these targets is challenging as process manufacturers are subjected to a higher level of variability than other industries (Gani, 2004). For example, in the biopharmaceutical industry yields can vary from 50% to 100% for no immediately discernible reason (Sadati et al., 2018). In case study 2, the unexplained variation between batches, shown in Fig. 3 (Section 3.2), results in disparities in the final alcohol concentration in the product. By taking observational data (meaning data routinely collected by a manufacturing process) and developing a DDM, the variables that are having a significant effect on the industry yields can be identified and optimised (Sadati et al., 2018). The advantage of using data-driven techniques this way is that the DDM can intelligently identify significant variables responsible for the variability. This is achieved by using unsupervised machine learning techniques for exploratory data analysis to find hidden patterns or grouping in data (Bishop, 2006).

Although it is possible to model and limit the impact of some variability within a processing plant (Sadati et al., 2018), a substantial challenge for process manufacturers is limiting the impact of variables outside of an engineer's control. Paramount of these variables is the variability in the process's feedstock (see Section 3.2). This variability can have a significant impact on product quality and yield. For example, the production of anti-malarial drugs which is reliant on the active component artemisinin harvested from a plant called *Artemisia annua* (Pilkington et al., 2014). The plant's artemisinin concentration varies dependant on when it is harvested during its lifecycle and the storage conditions (Pilkington et al., 2014). By fitting historical harvesting, transportation and process data, a DDM would determine the optimal harvesting and storage point to maximise artemisinin recovery. The model could also be used to predict the feedstock's potential to produce artemisinin and whether to adjust the price for the supplier as not all feedstock is created equally.

Variation in the feedstock is not the only concern for a process manufacturer. The waste produced by process manufacturing is equally variable, in regards to the volume produced, the concentration of valuable components and pollutants (Parlikar et al., 2016). Governments and regulatory bodies around the world set maximum limits on what manufacturers may emit to the environment (European Union, 2010). Manufacturers may choose to treat their waste onsite, so it is within these limits, pay a third party to handle their waste or partially treat their waste to reduce cost. There is a cost involved with all options and this cost may fluctuate with waste characteristics. Being able to predict and accommodate these variations ahead of time may mean measures can be taken to reduce this cost. Process manufacturing waste is unique from discrete manufacturing in that there exists the potential for waste valorisation (Arancon et al., 2013). However, one of the barriers to waste valorisation is the detrimental effect variation in the waste has on the technologies in terms of performance. Case study 1 is an example of a waste valorisation technology that utilises a DDM to increase the resilience of the bioprocess. By collecting a set of training data over a 1-year period and fitting it to a model, it was possible to capture the variability of the feedstock due to seasons, changes in farm practices and inherent variation. The model is able to identify which key feedstock characteristics variations were detrimental to the bioprocess's performance. The bioprocess's process conditions were also varied throughout the year, in order to investigate their effect on performance. By combining and fitting this data to a model, it was possible to predict how to optimise the bioprocess conditions dependant on the incoming feedstock characteristics.

4.4. Evaluating waste valorisation potential

Increased competition for access to critical resources is a major concern for the manufacturing industry. In 2017, the EU expanded its list of critical raw materials (CRM) to 27, defined as materials considered to be of high importance to the EU economy and of high risk to their supply (European Commission, 2017). Greater focus has been placed on developing solutions towards implementing the circular economy into manufacturing (Lieder and Rashid, 2016). The Waste and Resources Action Program charity offers a clear definition of the circular model: "*A circular economy is an alternative to a traditional linear economy (make, use, dispose) in which we keep resources in use for as long as possible, extract the maximum value from them whilst in use, then recover and regenerate products and materials at the end of each service life.*" (Waste and Resources Action Program, 2018). The circular economy is being pursued by Governments, with Europe leading the way by implementing the 2018 Circular Economy Action Plan to create a set of measures to help manufacturers to adopt circular economy systems (Springer and Schmitt, 2018). Encouraging manufacturers to stop viewing waste as a *waste* but instead as *co-product* is an important step towards this goal.

Process manufacturing generates a wide range of solid, liquid and gas waste streams that have excellent potential for waste valorisation within a circular economy. Waste valorisation refers to industrial processing activities aimed at reusing, recycling, or recovering useful products or sources of energy from waste (Kabongo, 2013). There are currently three waste valorisation strategies, geared towards (1) production of fuel and/or energy to replace fossil fuels; (2) production and/or extraction of high-value chemicals from residues; and (3) production of other useful material(s) (Arancon et al., 2013).

Process manufacturers have a variety of options when evaluating which waste valorisation route to take. Choosing which route to take depends on a number of factors: composition, volume and variability of the waste stream, capital and operational costs of onsite treatment versus valorisation by third party, pre-treatment costs, other waste being produced by nearby manufacturers, handling and transportation costs. Due to the large number of factors involved, it can be time and cost-intensive for a manufacturer to employ someone to evaluate the most sustainable route. There exists a number of Enterprise Resource Planning (ERP) programs, that use DDMs, capable of simulating the manufacturing process to intelligently determine the most sustainable waste valorisation route (The Access Group, 2018). These platforms are available via the cloud and systems like IBM Watson IoT are capable of integrating the manufacturer's existing ERP programs into their software (International Business Machines, 2016). However, while these platforms are capable of planning the logistics and costs, it has not been the focus to predict the performance of these different waste valorisation technologies at treating a waste stream. There is a great opportunity to have models incorporated into the process to take a holistic view, considering waste minimisation and/or valorisation alongside the key product objectives.

5. Conclusion and further work

The use of DDMs is on the rise across manufacturing primarily due to the following three points:

(1) There is an increasing amount of data generated in manufacturing as IIoT technologies become more widespread.
(2) Manufacturers have easier access to the computational power required to harness this data due to cloud computing enabling parallel processing.
(3) New machine learning techniques are enabling the utilisation of new sources of data, such as texts, image, audio, video, log files.

Before attempting to develop a DDM there are certain points the manufacturer must consider. The model's goal must be clearly defined and should aim to solve the manufacturing problem that has been identified. The manufacturer should also consider what is required from the model. Is it enough simply to fit the data, or will the model be required to make predictions? Being able to make predictions increases the model's value but also its complexity. The accuracy in which the model is required to perform these functions should be defined by the product specification, regulations, economic value and safety requirements.

Paramount to the success of DDMs is the generation/collection of a representative set of data used to develop the model. This data is split into model development data, used to train the model, and unseen validation data to evaluate the prediction capability of the model. Before generating/collecting this data the manufacturer should consider what data is already available and if it is representative of the system. Challenges arise when ensuring a representative data set is used for model development. Process manufacturers face unique challenges regarding this point because of the limited availability and variability of industrial data. Even when data is available, knowing if it representative of the system and relevant to the model can be challenging. Furthermore, the boundaries of a process manufacturing system are hard to classify. The difficulty occurs in knowing when enough data has been collected to be representative of the whole system. Once a manufacturer is able to overcome the considerations and challenges involved when modelling a process manufacturing system, DDMs will allow the process manufacturers to:

(1) **Improve process manufacturing models:** By integrating DDMs with existing first-principle process manufacturing models into a hybrid model, which have been demonstrated to have greater prediction accuracy and extrapolation capabilities.
(2) **Characterise process streams:** Through the development of soft sensors that utilise recent advances in machine learning able to utilise cheaper, unclean data. This will allow process manufacturers to develop a deeper understanding of their processes, providing optimisation opportunities in a cost-effective manner.
(3) **Enhance process resilience:** By fitting data to the DDM, manufacturers are able to make accurate predictions on the effect, variability across the system, has on their process and take measures to optimise them.
(4) **Evaluate waste valorisation routes:** Existing DDMs may be used as a tool to evaluate new waste streams compatibility with an existing waste valorisation technology. Alternatively, use DDMs as a tool for manufacturers to see which technology is the most sustainable option for their waste.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing and [grant number EP/P001246/1], Network Plus: Industrial Systems in a Digital Age.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Oliver J Fisher: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Nicholas J Watson:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition. **Josep E Escrig:** Methodology, Software, Data curation, Visualization. **Rob Witt:** Validation, Resources. **Laura Porcu:** Validation, Resources. **Darren Bacon:** Validation, Resources. **Martin Rigley:** Validation, Resources. **Rachel L Gomes:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

The author (RLG) would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing. The author (NJW) would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) Network Plus: Industrial Systems in a Digital Age [grant number EP/P001246/1], BREWNET: Intelligent Cloud Connected Sensors for Economic Small Scale Process Optimisation. The author (OJF) would also like to acknowledge the University of Nottingham Faculty of Engineering for his PhD scholarship.

References

Addo-Tenkorang, R., Helo, P.T., 2016. Big data applications in operations/supply-chain management: a literature review. Comput. Ind. Eng. 101, 528–543. https://doi.org/10.1016/J.CIE.2016.09.023.

Al-Aufi, Y.A., Hewakandamby, B.N., Dimitrakis, G., Holmes, M., Hasan, A., Watson, N.J., 2019. Thin film thickness measurements in two phase annular flows using ultrasonic pulse echo techniques. Flow Meas. Instrum. 66, 67–78. https://doi.org/10.1016/J.FLOWMEASINST.2019.02.008.

Almeida, J.S., 2002. Predictive non-linear modeling of complex data by artificial neural networks. Curr. Opin. Biotechnol. 13, 72–76. https://doi.org/10.1016/S0958-1669(02)00288-4.

Angria, S., L., Sari, Y.D., Zarlis, M., Tulus, 2018. Data-driven modelling for decision making under uncertainty. In: IOP Conference Series: Materials Science and Engineering, p. 12013.

Arancon, R.A.D., Lin, C.S.K., Chan, K.M., Kwan, T.H., Luque, R., 2013. Advances on waste valorization: new horizons for a more sustainable society. Energy Sci. Eng. 1, 53–71. https://doi.org/10.1002/ese3.9.

Arce, D., Lima, F., Orellana Cordero, M.P., Ortega, J., Sellers, C., Ortega, P., 2018. Discovering behavioral patterns among air pollutants: a data mining approach. Enfoque UTE 9, 168–179. https://doi.org/10.29019/enfoqueute.v9n4.411.

Atzmueller, M., Hayat, N., Schmidt, A., Kloepper, B., 2017. Explanation-aware feature selection using symbolic time series abstraction: approaches and experiences in a petro-chemical production context. In: Proceedings - 2017 IEEE 15th International Conference on Industrial Informatics, INDIN 2017. Institute of Electrical and Electronics Engineers Inc., pp. 799–804.

Azarpour, A., N.G. Borhani, T., R. Wan Alwi, S., A. Manan, Z., I. Abdul Mutalib, M., 2017. A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors. Chem. Eng. Res. Des. 117, 149–167. https://doi.org/10.1016/j.cherd.2016.10.024.

Barbosa, C., Azevedo, A., 2017. Hybrid simulation for complex manufacturing value-chain environments. Procedia Manuf. 11, 1404–1412. https://doi.org/10.1016/j.promfg.2017.07.270.

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6, 20–29. https://doi.org/10.1145/1007730.1007735.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Boukouvala, F., Li, J., Xiao, X., Floudas, C.A., 2016. Data-driven modeling and global optimization of industrial-scale petrochemical planning operations. In: 2016 American Control Conference (ACC), pp. 3340–3345.

Brandl, D., 2007. 1. Manufacturing control. Des. Patterns Flex. Manuf..

Caballero, J.A., 2015. Logic hybrid simulation-optimization algorithm for distillation design. Comput. Chem. Eng. 72, 284–299. https://doi.org/10.1016/j.compchemeng.2014.03.016.

Carr, D.B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S., 1987. Scatterplot matrix techniques for large N. J. Am. Stat. Assoc. 82, 424–436. https://doi.org/10.1080/01621459.1987.10478445.

Chandrasekaran, M., Muralidhar, M., Krishna, C.M., Dixit, U.S., 2012. Online machining optimization with continuous learning, in: computational methods for optimizing manufacturing technology: models and techniques. pp. 85–110.

Charte, F., Romero, I., Pï¿½rez-Godoy, M.D., Rivera, A.J., Castro, E., 2017. Comparative analysis of data mining and response surface methodology predictive models for enzymatic hydrolysis of pretreated olive tree biomass. Comput. Chem. Eng.. https://doi.org/10.1016/j.compchemeng.2017.02.008.

Chaturvedi, A.R., Hutchinson, G.K., Nazareth, D.L., 1993. Supporting complex real-time decision making through machine learning. Decis. Support Syst. 10, 213–233. https://doi.org/10.1016/0167-9236(93)90039-6.

Coley, C.W., Green, W.H., Jensen, K.F., 2018. Machine learning in computer-aided synthesis planning. Acc. Chem. Res. 51, 1281–1289. https://doi.org/10.1021/acs.accounts.8b00087.

Cross, M., Moscardini, A.O., 1985. Learning the Art of Mathematical Modelling, Ellis Horwood Series in Mathematics and Its applications. Statistics and Operational Research. John Wiley & Sons, Inc., Chichester [Chichestershire] : New York, NY, USA.

Dong, D., Mcavoy, T.J., 1996. Nonlinear principal component analysis - based on principal curves and neural networks. Comput. Chem. Eng. 20, 65–78. https://doi.org/10.1016/0098-1354(95)00003-K.

Duarte, B., Saraiva, P.M., C., P.C., 2004. Combined mechanistic and empirical modelling. Int. J. Chem. React. Eng.. https://doi.org/10.2202/1542-6580.1128.

Edmonds, D., 2016. Before the taps run dry: incentivizing water sustainability in America's craft breweries. Georg. Washingt. J. Energy Environ. Law 7, 164–176.

Eltawil, M.A., Samuel, D.V.K., Singhal, O.P., 2006. Potato storage technology and store design aspects. Agric. Eng. Int. CIGR J. 8, 1–18. https://doi.org/10.1093/hmg/ddw317.

Escrig, J., Woolley, E., Rangappa, S., Simeone, A., Watson, N.J., Escrig, Escrig, J., Woolley, E., Rangappa, S., Simeone, A., Watson, N.J., 2019. Clean-in-place monitoring of different food fouling materials using ultrasonic measurements. Food Control 104, 358–366. https://doi.org/10.1016/J.FOODCONT.2019.05.013.

European Commission, 2017. 2017 List of critical raw materials for the EU.

European Union, 2011. Regulation (EU) No 1169/2011 on the provision of food information to consumers. Off. J. Eur. Union 18–63. https://doi.org/10.1109/60.911397.

European Union, 2010. Directive 2010/75/EU of the european parliament and of the council of 24 November 2010 on industrial emissions (integrated pollution prevention and control), OJ L 334.

European Union, 1991. Directive 91/676/CEE concerning the protection of waters against pollution caused by nitrates from agricultural sources. Off. J. Eur. Commun. 1–8. https://doi.org/10.1017/CBO9781107415324.004.

Fayyad, U., Stolorz, P., 1997. Data mining and KDD: promise and challenges. Futur. Gener. Comput. Syst. 13, 99–115. https://doi.org/10.1016/s0167-739x(97)00015-0.

Fiedler, B., Schuppert, A., 2008. Local identification of scalar hybrid models with tree structure. IMA J. Appl. Math. (Institute Math. Its Appl. 73, 449–476. https://doi.org/10.1093/imamat/hxn011.

Fisgativa, H., Tremier, A., Dabert, P., 2016. Characterizing the variability of food waste quality: a need for efficient valorisation through anaerobic digestion. Waste Manag. 50, 264–274. https://doi.org/10.1016/J.WASMAN.2016.01.041.

Fisher, O., Watson, N., Porcu, L., Bacon, D., Rigley, M., Gomes, R.L.R.L., 2018. Cloud manufacturing as a sustainable process manufacturing route. J. Manuf. Syst. 47. https://doi.org/10.1016/j.jmsy.2018.03.005.

Gani, R., 2004. Chemical product design: challenges and opportunities. Comput. Chem. Eng. 28, 2441–2457. https://doi.org/10.1016/J.COMPCHEMENG.2004.08.010.

García-Delgado, M., Rodríguez-Cruz, M.S., Lorenzo, L.F., Arienzo, M., Sánchez-Martín, M.J., 2007. Seasonal and time variability of heavy metal content and of its chemical forms in sewage sludges from different wastewater treatment plants. Sci. Total Environ. 382, 82–92. https://doi.org/10.1016/J.SCITOTENV.2007.04.009.

Garment, V., 2014. 3 Ways to test the accuracy of your predictive models [WWW Document]. URL https://www.kdnuggets.com/2014/02/3-ways-to-test-accuracy-your-predictive-models.html (accessed 11.15.18).

Ge, Z., 2017. Review on data-driven modeling and monitoring for plant-wide industrial processes. Chemom. Intell. Lab. Syst. 171, 16–25. https://doi.org/10.1016/J.CHEMOLAB.2017.09.021.

Geiger, D., 2017. Data-driven manufacturing – monetizing the analytical edge [WWW Document]. URL https://www.aberdeen.com/opspro-essentials/data-driven-manufacturing-monetizing-analytical-edge/ (accessed 11.7.18).

Hamid, H.A., Jenidi, Y., Somerfield, C., Gomes, R.L., 2016. Predicting the capability of carboxylated cellulose nanowhiskers for the remediation of copper from water using response surface methodology (RSM) and artificial neural network (ANN) models. Ind. Crops Prod. 93, 108–120. https://doi.org/10.1016/J.INDCROP.2016.05.035.

Hangos, K.M., Cameron, I.T.(Eds.), 2001. 1 - The role of models in process systems engineering. In: Process Modelling and Model Analysis, Process Systems Engineering. Academic Press, pp. 3–18.

Harding, J.A., Shahbaz, M., Srinivas, Kusiak, A., 2006. Data mining in manufacturing: a review. J. Manuf. Sci. Eng. Trans. ASME. https://doi.org/10.1115/1.2194554.

Hu, G., Mao, Z., He, D., Yang, F., 2011. Hybrid modeling for the prediction of leaching rate in leaching process based on negative correlation learning bagging ensemble algorithm. Comput. Chem. Eng. 35, 2611–2617. https://doi.org/10.1016/j.compchemeng.2011.02.012.

International Business Machines, 2016. IBM Watson internet of things (IoT) [WWW Document]. URL https://www.ibm.com/internet-of-things (accessed 1.10.19).

Iskandarani, M., Wang, S., Srinivasan, A., Carlisle Thacker, W., Winokur, J., Knio, O.M., 2016. An overview of uncertainty quantification techniques with application to oceanic and oil-spill simulations. J. Geophys. Res. Ocean. 121, 2789–2808. https://doi.org/10.1002/2015JC011366.

Kabongo, J.D., 2013. Waste Valorization. In: Idowu, S.O., Capaldi, N., Zu, L., Gupta, A.Das (Eds.), Encyclopedia of Corporate Social Responsibility. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 2701–2706.

Kadlec, P., Gabrys, B., 2009. Soft sensors: where are we and what are the current and future challenges?, IFAC Proceedings Volumes (IFAC-PapersOnline). IFAC. https://doi.org/10.3182/20090921-3-TR-3005.00098.

Kay, J.W., Titterington, D.M., Kay, S.L.S.J.W., 1999. Statistics and Neural Networks: Advances at the Interface, Royal Statistical Society Lecture Notes Series, 5. Oxford University Press.

Keviczky, L., Banyasz, C., 2015. 10. Process Identification, in: Two-Degree-of-Freedom Control Systems - The Youla Parameterization Approach. Elsevier, pp. 309–315.

Kim, S., 2017. MATLAB Deep Learning With Machine Learning, Neural Networks and Artificial Intelligence, 1st ed. Apress https://doi.org/10.1007/978-1-4842-2845-6.

Kleinert, T., Schladt, M., Muehlbeyer, S., Schocker, A., 2011. Combination of process analytical technology with soft sensors for online process data and advanced process information. TM-TECHNISCHES Mess 78, 589–602. https://doi.org/10.1524/teme.2011.0165.

Krause, D., Schöck, T., Hussein, M.A., Becker, T., 2011. Ultrasonic characterization of aqueous solutions with varying sugar and ethanol content using multivariate regression methods. J. Chemom. 25, 216–223. https://doi.org/10.1002/cem.1384.

Krenczyk, D., 2012. Data-driven modelling and simulation for integration of production planning and simulation systems. Sel. Eng. Probl. 119–122.

Lauer, F., Bloch, G., 2008. Incorporating prior knowledge in support vector machines for classification: a review. Neurocomputing 71, 1578–1594. https://doi.org/10.1016/j.neucom.2007.04.010.

Lee, H.-.Y., Ong, H.-.L., 1996. Visualization support for data mining. IEEE Expert. Syst. their Appl. 11, 69–75. https://doi.org/10.1109/64.539019.

Lieder, M., Rashid, A., 2016. Towards circular economy implementation: a comprehensive review in context of manufacturing industry. J. Clean. Prod. 115, 36–51. https://doi.org/10.1016/J.JCLEPRO.2015.12.042.

Lin, C.S.K., Pfaltzgraff, L.A., Herrero-Davila, L., Mubofu, E.B., Abderrahim, S., Clark, J.H., Koutinas, A.A., Kopsahelis, N., Stamatelatou, K., Dickson, F., Thankappan, S., Mohamed, Z., Brocklesby, R., Luque, R., 2013. Food waste as a valuable resource for the production of chemicals, materials and fuels. Current situation and global perspective. Energy Environ. Sci. 6, 426–464. https://doi.org/10.1039/c2ee23440h.

Liu, H., Cocea, M., 2017. Semi-random partitioning of data into training and test sets in granular computing context. Granul. Comput. 2, 357–386. https://doi.org/10.1007/s41066-017-0049-2.

Lopez-Juarez, I., Rios-Cabrera, R., Hsieh, S.J., Howarth, M., 2018. A hybrid non-invasive method for internal/external quality assessment of potatoes. Eur. Food Res. Technol. 244, 161–174. https://doi.org/10.1007/s00217-017-2936-9.

Luo, S., Sun, H., Ping, Q., Jin, R., He, Z., 2016. A review of modeling bioelectrochemical systems: engineering and statistical aspects. Energies 9, 111. https://doi.org/10.3390/en9020111.

Manning, L., Soon, J.M., 2016. Food safety, food fraud, and food defense: a fast evolving literature. J. Food Sci. 81, R823–R834. https://doi.org/10.1111/1750-3841.13256.

Mariscal, G., Marban, O., Fernandez, C., 2010. A survey of data mining and knowledge discovery process models and methodologies. Knowl. Eng. Rev. 25, 137–166. https://doi.org/10.1017/S0269888910000032.

Mathews, P., 2004. DOE language and concepts. In: Design of Experiments With MINITAB. ASQ Quality Press, Milwaukee, pp. 93–142.

Microsoft, 2018. Azure machine learning studio [WWW Document]. URL https://azure.microsoft.com/en-gb/services/machine-learning-studio/ (accessed 10.24.18).

Ncube, B., Finnie, J.F., Van Staden, J., 2012. Quality from the field: the impact of environmental factors as quality determinants in medicinal plants. S. Afr. J. Bot. 82, 11–20. https://doi.org/10.1016/J.SAJB.2012.05.009.

Nicoletti, M.C., Jain, L.C., Giordano, R.C., 2009. Computational intelligence techniques as tools for bioprocess modelling, optimization, supervision and control. Stud. Comput. Intell.. https://doi.org/10.1007/978-3-642-01888-6_1.

Ning, C., You, F., 2018. Data-driven stochastic robust optimization: general computational framework and algorithm leveraging machine learning for optimization under uncertainty in the big data era. Comput. Chem. Eng.. https://doi.org/10.1016/j.compchemeng.2017.12.015.

OECD, 2012. OECD environmental outlook to 2050, OECD environmental outlook. OECD Publishing. https://doi.org/10.1787/9789264122246-en

Ojha, V.K., Abraham, A., Snášel, V., 2017. Metaheuristic design of feedforward neural networks: a review of two decades of research. Eng. Appl. Artif. Intell. 60, 97–116. https://doi.org/10.1016/j.engappai.2017.01.013.

Owhadi, H., Scovel, C., Sullivant, T.J., McKerns, M., Ortiz, M., 2013. Optimal uncertainty quantification. SIAM Rev 55, 271–345. https://doi.org/10.1137/10080782X.

Pan, Y., Hu, M., 2016. A data-driven modeling approach for digital material additive manufacturing process planning. In: 2016 International Symposium on Flexible Automation (ISFA), pp. 223–228.

Panerati, J., Schnellmann, M.A., Patience, C., Beltrame, G., Patience, G.S., 2019. Experimental methods in chemical engineering: artificial neural networks–ANNs. Can. J. Chem. Eng. 97, 2372–2382. https://doi.org/10.1002/cjce.23507.

Parlikar, U., Bundela, P.S., Baidya, R., Ghosh, S.K., 2016. Effect of variation in the chemical constituents of wastes on the co-processing performance of the cement kilns. Procedia Environ. Sci. 35, 506–512. https://doi.org/10.1016/J.PROENV.2016.07.035.

Pasini, A., 2015. Artificial neural networks for small dataset analysis. J. Thorac. Dis. 7, 953–960. https://doi.org/10.3978/j.issn.2072-1439.2015.04.61.

Perry, R.H., Green, D.W., 2008. Perry's Chemical engineers' Handbook. McGraw-Hill, New York.

Pilkington, J.L., Preston, C., Gomes, R.L., 2014. Comparison of response surface methodology (RSM) and artificial neural networks (ANN) towards efficient extraction of artemisinin from Artemisia annua. Ind. Crops Prod. 58, 15–24. https://doi.org/10.1016/J.INDCROP.2014.03.016.

Qiao, G., Riddick, F., McLean, C., 2003. New manufacturing modeling methodology: data driven design and simulation system based on XML. In: Proceedings of the 35th Conference on Winter Simulation: Driving Innovation, WSC '03. Winter Simulation Conference, pp. 1143–1148.

Qin, S.J., 2014. Process data analytics in the era of big data. AIChE J 60, 3092–3100. https://doi.org/10.1002/aic.14523.

Rasmuson, A., Andersson, B., Olsson, L., Andersson, R., Olisson, L., Andersson, R., 2014. Empirical Model building, in: Mathematical Modeling in Chemical Engineering. Cambridge Univerisity Press, Cambridge, pp. 40–52.

Resa, P., Elvira, L., de Espinosa, F., 2004. Concentration control in alcoholic fermentation processes from ultrasonic velocity measurements. Food Res. Int. 37, 587–594. https://doi.org/10.1016/j.foodres.2003.12.012.

Rizos, V., Behrens, A., Kafyeke, T., Hirschnitz-Garbers, M., Ioannou, A., 2015. The circular economy: barriers and opportunities for SMEs. CEPS Work. Doc. 1–25 https://doi.org/0006-8993(94)90176-7 [pii] ET - 1994/02/04.

Sadati, N., Chinnam, R.B., Nezhad, M.Z., 2018. Observational data-driven modeling and optimization of manufacturing processes. Expert Syst. Appl. 93, 456–464. https://doi.org/10.1016/j.eswa.2017.10.028.

Saka, M.P., Doğan, E., Aydogdu, I., 2013. Analysis of swarm intelligence–based algorithms for constrained optimization. Swarm Intell. Bio-Inspired Comput. 25–48. https://doi.org/10.1016/B978-0-12-405163-8.00002-8.

Schichl, H., 2004. Models and the history of modeling. In: Josef, K. (Ed.), Modeling Languages in Mathematical Optimization. Springer US, Boston, pp. 25–36.

Shafique, U., Qaiser, H., 2014. A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). Int. J. Innov. Sci. Res. 12, 217–222.

Shang, C., Yang, F., Huang, D., Lyu, W., 2014. Data-driven soft sensor development based on deep learning technique. J. Process Control 24, 223–233. https://doi.org/10.1016/j.jprocont.2014.01.012.

Shearer, C., 2000. The CRISP-DM model: the new blueprint for data mining. J. Data Warehous. 5, 13–22.

Simate, G.S., 2015. Water treatment and reuse in breweries. Brew. Microbiol. 425–456. https://doi.org/10.1016/B978-1-78242-331-7.00020-4.

Simeone, A., Deng, B., Watson, N., Woolley, E., 2018. Enhanced clean-in-place monitoring using ultraviolet induced fluorescence and neural networks. Sensors 18. https://doi.org/10.3390/s18113742.

Simeone, A., Watson, N., Sterritt, I., Woolley, E., 2016. A multi-sensor approach for fouling level assessment in clean-in-place processes. Procedia CIRP 55, 134–139. https://doi.org/10.1016/J.PROCIR.2016.07.023.

Simmonds, J., 2017. Number of breweries in the UK breaks through the 2,000 barrier for the first time since the 1930s [WWW Document]. URL https://www.uhy-uk.com/news-events/news/number-of-breweries-in-the-uk-breaks-through-the-2000-barrier-for-the-first-time-since-the-1930s/ (accessed 12.19.18).

Skoogh, A., Perera, T., Johansson, B., 2012. Input data management in simulation – Industrial practices and future trends. Simul. Model. Pract. Theory 29, 181–192. https://doi.org/10.1016/J.SIMPAT.2012.07.009.

Slišković, D., Grbić, R., Hocenski, Ž, 2011. Methods for plant data-based process modeling in soft-sensor development. Automatika 52, 306–318. https://doi.org/10.1080/00051144.2011.11828430.

Solomatine, D., See, L.M., Abrahart, R.J., 2008. Data-driven modelling : concepts, approaches and experiences. In: Abrahart, Robert J., See, Linda M., Solomatine, D.P. (Eds.), Practical Hydroinformatics. Water Science and Technology Library. Springer, Berlin, pp. 17–31.

Soroush Rohanizadeh, S., Moghadam, M.B., 2009. A proposed data mining methodology and its application to industrial procedures. J. Ind. Eng..

Souza, F.A.A., Araújo, R., Mendes, J., 2016. Review of soft sensor methods for regression applications. Chemom. Intell. Lab. Syst. 152, 69–79. https://doi.org/10.1016/j.chemolab.2015.12.011.

Springer, N.P., Schmitt, J., 2018. The price of byproducts: distinguishing co-products from waste using the rectangular choice-of-technologies model. Resour. Conserv. Recycl. 138, 231–237. https://doi.org/10.1016/J.RESCONREC.2018.07.034.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Templeton, D.W., Sluiter, A.D., Hayward, T.K., Hames, B.R., Thomas, S.R., 2009. Assessing corn stover composition and sources of variability via NIRS. Cellulose 16, 621–639. https://doi.org/10.1007/s10570-009-9325-x.

The Access Group, 2018. Access ERP systems for all industries [WWW Document]. URL https://www.theaccessgroup.com/supply-chain/industries/ (accessed 1.11.19).

von Stosch, M., Oliveira, R., Peres, J., Feyo de Azevedo, S., 2014. Hybrid semi-parametric modeling in process systems engineering: past, present and future. Comput. Chem. Eng.. https://doi.org/10.1016/j.compchemeng.2013.08.008.

Wang, J., Chang, Q., Xiao, G., Wang, N., Li, S., 2011. Data driven production modeling and simulation of complex automobile general assembly plant. Comput. Ind. 62, 765–775. https://doi.org/10.1016/J.COMPIND.2011.05.004.

Wang, K., Salhi, A., Fraga, E., 2002. Cluster identification using a parallel coordinate system for knowledge discovery and nonlinear optimization. 12th European Symposium on Computer Aided Process Engineering (ESCAPE-12).

Wang, K., Salhi, A., Fraga, E.S., 2004. Process design optimisation using embedded hybrid visualisation and data analysis techniques within a genetic algorithm optimisation framework. Chem. Eng. Process. Process Intensif. 43, 657–669. https://doi.org/10.1016/J.CEP.2003.01.001.

Wang, R., Chen, B., Qiu, S., Zhu, Z., Wang, Yiduo, Wang, Yiping, Qiu, X., 2018. Comparison of machine learning models for hazardous gas dispersion prediction in field cases. Int. J. Environ. Res. Public Health 15. https://doi.org/10.3390/ijerph15071450.

Waste and Resources Action Program, 2018. WRAP and the circular economy WRAP UK [WWW Document]. URL http://www.wrap.org.uk/about-us/about/wrap-and-circular-economy; (accessed 11.15.18).

Williams, C.L., Westover, T.L., Emerson, R.M., Tumuluru, J.S., Li, C., 2016. Sources of biomass feedstock variability and the potential impact on biofuels production. BioEnergy Res. 9, 1–14. https://doi.org/10.1007/s12155-015-9694-y.

Yan, W., Tang, D., Lin, Y., 2017. A data-driven soft sensor modeling method based on deep learning and its application. IEEE Trans. Ind. Electron. 64, 4237–4245. https://doi.org/10.1109/TIE.2016.2622668.

Yin, S., Kaynak, O., 2015. Big data for modern industry: challenges and trends [Point of View]. Proc. IEEE 103, 143–146. https://doi.org/10.1109/JPROC.2015.2388958.

Zeng, X., Luo, G., 2017. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Heal. Inf. Sci. Syst. 5. https://doi.org/10.1007/s13755-017-0023-z.

Zhang, C., Su, H., Baeyens, J., Tan, T., 2014. Reviewing the anaerobic digestion of food waste for biogas production. Renew. Sustain. Energy Rev. 38, 383–392. https://doi.org/10.1016/J.RSER.2014.05.038.