Yang Shu (Orcid ID: 0000-0002-4630-9287)

Observational process data analytics using causal inference

Shu Yang, and B. Wayne Bequette

Department of Chemical and Biological Engineering

Rensselaer Polytechnic Institute, Troy, NY

Abstract

Voluminous process data is available with the paradigm shift towards smart manufacturing. However, most historical data are observational, containing non-causal correlations due to confounders and mediators. Estimating causal effects from observational data remains a bottleneck in leveraging them for active applications such as optimization and control. This work aims to introduce a causal modeling framework for analyzing observational process data and extracting quantitative causal information. We demonstrate a real-world application in steel manufacturing where causal inference is used to analyze observational production data and improve the steelmaking process. Additionally, we propose a novel formulation for identifying critical process parameters from observational data, where causal inference is combined with variance-based methods to estimate corresponding risks of interventions to the manufacturing system. The proposed methods are compared with statistical ones to illustrate that causally interpreting statistical correlation leads to problematic results, while the provided workflow generates satisfactory strategies for process improvement.

Keywords: Smart Manufacturing, Causal Inference, Machine Learning

Prelude: Reflections of Wayne Bequette

I was fortunate to be a friend and colleague of Babatunde (Tunde) Ogunnaike for over 30 years. Once, when I visited him at du Pont, he noted that he was carrying a disk drive fully loaded with the content of his over 1200-page textbook with W. Harmon Ray, to be shipped to Oxford for publication. Tunde was a good-natured spirit who received a number of "awards" (notoriously presented at Chemical Process Control conferences). Given that most of control theory has been based on linear models, Tunde once presented the analogy that linear models were elephants and that nonlinear models would be the equivalent to all non-elephantine animals. Thus, at the 1996 CPC he received the "Wild Kingdom Award" for introducing the set of non-elephantine animals. As a follow-up, at the 2006 CPC, The Tunde Ogunnaike Systems Zoology Award was given to: (i) Frank Allgower for a presentation on (non) Elephantine Animals, (ii) Ron Pearson for using a set of geometrical cows, and (iii) to Jim Rawlings for Multi-model duck hunting.

Professor Ogunnaike received an MS in Statistics in addition to his PhD in Chemical Engineering at the University of Wisconsin, and the rigorous applications of statistics appear in many of his publications; indeed, he authored a comprehensive textbook on statistics [1]. His paper with Verma et al. [2] presents an application of causality analysis in the modeling of signaling patterns in liver lobules. We are very happy to provide our paper on a related topic to this special issue devoted to the memory of Tunde Ogunnaike.

Introduction

With the rapid advancement of the Industrial Internet of Things (IIoT), artificial intelligence, and cloud computing, the manufacturing industry is moving towards a more integrated and data-driven paradigm known as smart manufacturing [3–5]. In smart manufacturing systems, a massive amount of process data are generated, providing a great opportunity for process improvement [6,7]. As a result, process data analytics plays an increasingly important role in smart manufacturing, and many successful applications have been reported [8–10].

One of the major issues in process data analytics is causality [11]. Unlike statistical associations which are based on the observed input-output responses, causal relationships explore the underlying driving force and enable the prediction of system behavior under perturbations [12]. Causality is not necessary for passive smart manufacturing applications such as monitoring and soft sensors, but it is critical for active applications such as control and optimization [11,13]. In passive applications such as process monitoring, the goal is to model statistical relationships between process variables during normal operation, so deviations from normal operating conditions can be detected [5,14]. Active application on the other hand requires the inference of system behavior after applying perturbations to the system. Active alteration of process operation will change the underlying data-generating process, which changes the statistical associations in process data. Therefore, purely statistical models are insufficient to drive process improvement decision-making [15]. This idea is commonly known as “correlation does not imply causation”. Statistical associations in data are static “snapshots” of the system, while causal dependences capture the dynamic behavior of the system when interventions are introduced. Additionally, causality can facilitate the creation of explainable machine learning

models [16,17]. In smart manufacturing applications, decisions often need to be made where the rationale behind each analysis is transparent. Causality shows the causal and effect relationships between input and output and incorporated mechanistic understanding into the data-driven decision-making process. By improving model expandability, causality improves data-driven model acceptance in real-world smart manufacturing applications [18]. Causality is crucial for data-driven attempts at process improvement.

Traditional methods for developing causal data-driven models of manufacturing systems are experimental, where input variables are manipulated and the changes in outputs are recorded. Randomized experiments are the golden standard for extracting causal relationships in systems [19]. For active applications such as control, system identification techniques are commonly used to construct causal data-driven models of the dynamic system [20]. Excitation signals such as pseudorandom binary sequences are applied to the input variables as a designed intervention to extract causal information from the dynamic system [21].

However, in manufacturing systems experimentations may be impractical or even infeasible, and only observational data are available for analysis. Unlike experimental data which is based on designed perturbation of the manufacturing system, the analyst has no control over how data are collected in observational studies [22]. Without experimental design, observational data often contain spurious correlations that will skew our understanding when the statistical associations are interpreted in a causal way [23]. For instance, the logged data may not represent the target population, or spurious associations may exist due to unmodeled variables. Extracting causal relationships from observational data remains a bottleneck in leveraging process big data for process improvement [24].

Extracting causal information from observational data is a challenging issue with ongoing

research efforts [13,25]. Because observational data only contains static snapshots of the data-generating process, it alone is insufficient to answer causal queries where interventions are introduced and the underlying data-generating process is changed. Therefore, additional assumptions of the system are required to support causal queries [26]. Causal inference seeks to bridge this gap by providing a systematic framework to model causal assumptions and solve causal queries [27]. Observational studies are common in medical, social, and natural science, and causal inference has contributed to many fruitful pieces of research [28–33].

Applications of causality in manufacturing are limited, and the focus is mainly on causal discovery [34–39]. Causal discovery is a qualitative task, seeking to extract the causal structure of the system from observational data; while causal inference focuses on the quantitative side, seeking to quantify the strength of the causal effect between variables [40]. Additionally, current works are mainly based on Bayesian networks and Granger causality, which is unable to address spurious correlations in observational data [41,42]. Causality has been widely used in fault detection and diagnosis (FDD) to understand the propagation path of abnormality [43–47]. In manufacturing systems where field knowledge is unavailable or insufficient, data-driven causal discovery can help augment FDD efforts by identifying possible root causes and reducing alarm flooding [14,48–51]. Another common application of causal reasoning in manufacturing is to facilitate statistical models. Using causality, statistical machine learning models can learn more invariance and robust relationships and improve model explainability [25,52–54].

This paper seeks to introduce causal modeling tools that provide a generalizable framework for analyzing non-causally correlated observational process data, including confounding, mediating, as shown in Figure 1. Confounders are variables that affect both input and output

variables, which introduces spurious correlations between them. Mediators are variables are caused by independent variables which in turn cause the dependent variable. Mediators will cause indirect causal relationships.

The problem we seek to solve is "when given qualitative knowledge about the causal structure of the manufacturing system, can we generate an unbiased quantitative estimation of the strength of the causal relationships". Firstly, we present the popular Structural Causal Models (SCM) framework which provides a strong theoretical foundation for analyzing Spuriously correlated observational data. Secondly, we presented a real-world application where SCM is used to improve a steel manufacturing process. Thirdly, we presented a novel formulation for identifying critical process parameters (CPPs) from observational data using SCM, illustrated through a simulation case study.

The paper is structured as follows: In section 2, we provide a brief overview of the theoretical background for causal inference. In section 3, we discussed a motivating real-world example from a steel manufacturing plant, where causal inference was used to analyze observational process data to improve the manufacturing process. In section 4, we provided a novel causal framework for identifying the critical process parameters using observational data. The paper concludes with a discussion in section 5.

Theoretical Background: Causal inference and do-calculus

This section presents a brief overview of the SCM framework [55] which serves as the theoretical foundation for representing causal relationships and estimating causal effects. Causal assumptions are integrated into the SCM framework in the form of a causal graph. It should be noted that another widely used framework for causal inference in statistics is the "potential

outcomes" framework, which relies less on graphic representations [56]. The two frameworks are logically equivalent [57]. Because graph representations are common in manufacturing systems (such as P&ID), the SCM framework is selected in the manuscript as the theoretical foundation. This framework provides an unambiguous way to annotate causal assumptions and queries. Additionally, it provides a set of algorithms to combine causal assumptions with observational data to answer causal queries.

Causal graphs and intervention

In the SCM framework, the causal relationships between variables are represented using directed acyclic graphs (DAGs), where the vertices represent process variables and directed edges represent the causal relationships. The graphs are assumed to be acyclic, which doesn't allow feedback loops. In causal inference, the causal graphs represent the causal assumptions of the system.

Under the SCM framework, interventions are annotated using the do-operator. $p(y|do(x = x_0))$ simulated a physical intervention where the variable x is manipulated to be a constant x_0, while the rest of the model remains unchanged. Since this intervention changes the underlying data-generating process, the causal query. $p(y|do(x = x_0))$ is different from the statistical query $p(y|x = x_0)$, which is based on the original data-generating process. The intervention $do(x = x_0)$ can be graphically represented using the graphical model as shown in Figure 2. Because of the intervention, variable x no longer varies with the changing z, instead, it was fixed to be a constant x_0, which is represented by removing the edge from Z to X. Using the new annotation system in SCM, we'll be able to systematically describe, identify and estimate causal effects.

Steps of causal inference

Because the simulated causal query $p(y|do(x))$ cannot be directly estimated from observational data, a procedure is developed for the SCM framework to reduce the causal queries into "do-free" expressions. Because the "do-free" query is statistical, it can then be estimated from observational data using statistical principles. If the do-operator can be successfully removed from the causal query, then the queried causal effect is identifiable [26,58].

The procedure for solving causal problems involves four steps, as shown in Figure 3 [59]. Firstly, in the definition step, the causal query is expressed using the do-operator. Secondly, in the formulation stage, knowledge of the underlying mechanism is organized to form the causal assumptions, in the form of a causal graph. Then in the identification process, based on the causal graph, the do-operators in the causal query are reduced using a set of analytical methods provided in the do-calculus [26]. If the causal query can be reduced into a "do-free" query, the resulting expression is the identified statistical estimand that answers the causal query. Lastly, in the estimation step, collected data is leveraged to estimate the resulting statistical estimand, which answers the causal query. In section 3, a real-world example will be discussed to illustrate this procedure.

Do-calculus provides three main analytical rules that can be leveraged for causal query identification [58]. Given causal graph G, with arbitrary nodes X, Y, Z, W, the following laws would hold for every data generating process that can be described by G.

$$p(Y|do(X), Z, W) = p(Y|do(X), W) \text{ if } (Y \perp Z \mid X, W)_{G_{\bar{X}}} \quad (1)$$

Where $G_{\bar{X}}$ represent the graph obtained by removing all edges going into X from G.

$$p(Y|do(X), do(Z), W) = p(Y|do(X), Z, W) \text{ if } (Y \perp Z \mid X, W)_{G_{\bar{X}\underline{Z}}} \quad (2)$$

Where $G_{\bar{X}\underline{Z}}$ represent the graph obtained by removing all edges going into node X and edges coming out from node Z from G.

$$p(Y|do(X), do(Z), W) = p(Y|do(X), W) \text{ if } (Y \perp Z \mid X, W)_{G_{\overline{XZ(W)}}} \quad (3)$$

Where $G_{\overline{XZ(W)}}$ represent the graph obtained by removing all edges going into node X and $Z(W)$ from G. $Z(W)$ represent the Z nodes that are not ancestors of any W-nodes in $G_{\bar{X}}$.

It should be noted that not all causal queries are identifiable. If through repeated application of the analytical rules of do-calculus, all do-operators can be removed from a causal query, then this query is estimable from observational data, where the final statistical estimand is the estimator. It has been proven that the do-calculus is complete to the identifiability of causal queries [60]. Because the identifiability of causal queries is only based on the causal structure and cannot be resolved by collecting more data [55], extra experimental efforts may be required to address nonidentifiable causal queries.

Motivating example: Steel manufacturing process improvement

This work was motivated by a data analytics task from a real-world steel manufacturing process, where observational process data was analyzed to improve the manufacturing process. Specifically, the process data seemed to suggest an improvement strategy that contradicts field understanding. After causal inference strategies were adopted, it was found that the contradiction is generated by interpreting statistical associations in a causal manner, and the

new strategy based on causal inference agreed well with field knowledge.

This project aims to address a common type of fault called clogging, which is caused by the undesired deposition of solids in the flow channel which restricted the flow of liquid steel [61]. Two years worth of real-world steelmaking data was collected from one of Cleveland Cliffs' plants [54]. There are 37 input features in this dataset, including compositions, ladle treatment, and temperature measurements, which can be roughly understood as the recipe of the steel. Each sample was labeled manually by operators, either as "faulty" or "normal". The objective of this project is to improve this manufacturing process and reduce clogging.

Initially, a data-driven predictive model is developed to predict clogging given the recipe of the steel [54]. The goal was to capture the statistical correlation in historical manufacturing data using a machine learning model, which can then be used as a surrogate to optimize the process. Based on historical batch recipe data, supervised-learning models including boosted decision trees and neural networks are implemented to map steel manufacturing recipes to the likelihood of clogging. Despite promising prediction accuracy, when the developed statistical model is used for process improvement, the resulting strategy contradicts the mechanistic understanding. One important reason for this contradictory result is that statistical correlations captured by the machine learning model should not be causally interpreted.

Following is an analysis of one critical process variable in steel manufacturing. Because scatter plots contain the most fundamental and uncompressed information of the statistical associations, it is selected to demonstrate that statistics alone is insufficient for solving such process improvement problems.

One important raw material in the steelmaking process is recycled steel named scrap. Besides

being an important raw material, it is often added as a coolant to the steelmaking process [62]. Scrap addition correlates negatively with clogging in the collected dataset, as shown in Figure 4 and Table 1. Based on the negative correlation, one may recommend increasing scrap addition in the steelmaking process to reduce clogging. However, this is opposite to the mechanistic understanding that scrap reduces steel cleanliness thus increasing the likelihood of clogging.

This seemingly contradictory result is due to the attempt to interpret statistical associations in a causal way. This interpretation is inexplicitly assuming that the process data is experimental, which is collected when scrap addition is independently manipulated while the other process variables are controlled. However, the collected data is observational where scrap addition was adjusted together with other process variables to follow prespecified recipes, resulting in spurious correlations in the dataset.

These spurious correlations can be adjusted for using the causal inference procedure introduced in section 2, where a mechanistic understanding of the causal mechanism is combined with historical process data to estimate the causal contribution of the scraps on clogging.

Firstly, the causal query is written using the do-operator notation $p(clogging|do(scrap))$.

Secondly, a mechanistic-based causal assumption is added to identify a statistical estimand that quantifies the causal contribution and removes the do-operator in the query. Based on mechanistic understanding, scrap was added as a coolant to adjust liquid steel temperature and meet the temperature setpoint in the steelmaking recipe. Together with scrap, aluminum was also used to adjust steel temperature. The oxidation of aluminum is an exothermic process that

increases the temperature of steel. Figure 5 shows the causal graph that represented this causal assumption, where z represents the temperature setpoint that remains hidden, X_1 represents scrap addition and X_2 represents aluminum addition, Y represents clogging.

Thirdly, based on the causal assumption in the form of the causal diagram, Do-calculus was used to address non-causal correlations in data and estimate the causal contribution of scrap to clogging based on the given causal graph. By applying the analytic rules from SCM, the causal query is identifiable and can be estimated by:

$$p(y|do(x_1)) = \sum_z p(y|x_1, x_2)p(x_2) \tag{4}$$

Therefore, instead of running logistic regression of y on x_1, which inexplicitly assumes $p(y|do(x_1)) = p(y|x_1)$, the logistic regression should be of y on both x_1 and x_2. The result is shown in Table 2. Note that the sign of the coefficient for scrap has flipped to positive, which agrees well with the mechanistic understanding. Causal inference addressed the spurious correlations by identifying the correct statistics to estimate from the data.

Lastly, the identified statistical estimand can be estimated using appropriate statistical and machine learning methods. Logistic regression is selected so the results can be directly compared with the previously discussed statistical results.

The mechanistic reason for the seemingly paradoxical result is the spurious correlation between aluminum and scrap. Due to the hidden factor of the recipe, in this observational study scrap correlates strongly with aluminum addition, which is a strong non-causal correlation. The oxidant of aluminum is one of the main components in the clog, so aluminum addition

significantly increases the likelihood of clogging [63]. Scrap and aluminum are negatively correlated due to their opposite effect on temperature. As a result, high scrap correlates strongly with low aluminum, which jointly associates with a lower likelihood of clogging, thus the spurious negative correlation between scrap and clogging.

Even though this example is from a steel manufacturing plant, the challenge of extracting causal information from observational data is universal [64]. Due to recipe designs, feedforward adjustment, and human intervention, non-causal correlations are very common in process data [6,25,65]. Without the scrutiny of causal inference, any found statistical association may lead to a faulty attempt at process improvement.

Applying the SCM framework to identify critical process parameters

This section presented a novel formulation for identifying critical process parameters (CPPs) from observational data using SCM. The problem we seek to solve is "given observational process data, can we identify the key process parameters so we can prioritize experimentation and process control effort?".

Background

Critical process parameters

Critical process parameters (CPPs) are the key parameters in manufacturing processes that most heavily affect product quality. As a result, experimentation and process control efforts should be prioritized in CPPs to maintain satisfactory process performance. Therefore, identifying CPPs is a crucial step to support many active tasks for smart manufacturing [66].

The definition of CPPs is intrinsically causal instead of associative. The goal of CPP identification is to facilitate active tasks such as process improvement and process control. The search of CPPs tries to infer the hypothetical outcome if interventions such as control are provided on certain process parameters. Because the hypothetical intervention will change the underlying data-generating process, the search for CPPs is intrinsically a causal query.

The identification of CPPs is similar to feature selection in predictive modeling. Feature selection seeks to identify the most relevant subset of variables for model construction. By eliminating redundant or irrelevant features, feature selection can improve model performance and robustness [67]. The general objective of feature selection is usually associational instead of causal, aiming to improve the predictive performance of the model. However, causality can also facilitate the feature selection process [68].

Because CPP identification is intrinsically a causal query, established methods for CPP identification are experimental, which are based on the Design of Experiment (DOE) principles upon either physical systems or computer simulations [69,70]. In established methods, the analysts usually have control over the way data are collected, where process variables can be manipulated independently. By efficient sampling from the design space and carefully designing the experiments, experimental identification of CPPs are widely accepted technique in process system engineering [70,71].

However, experiments on the physical system can be very expensive if feasible at all. Additionally, constructing high-fidelity process models may also be infeasible due to process complexity. Smart manufacturing technology has significantly increased the availability of observational process data. Taking advantage of the increasingly available observational process data to aid the identification of CPPs is a promising alternative.

Sensitivity analysis

Sensitivity analysis is the most widely used statistical tool for identifying CPPs[72]. The application of sensitivity analysis assumes experimental or simulation data. The Morris method and variance-based method are the most commonly used sensitivity analysis methods for identifying CPPs.

The Morris method is commonly used for preliminary candidate variable screening as a result of its computational simplicity. It calculates the "average slope" of the output on each candidate variable. The Morris method assumes an experimental dataset where only one variable is changed at each sampling step so that the contribution of each variable is calculated independently. After a random perturbation Δ_i is made on variable x_i, the resulting change in the output y is recorded. Then the elementary effect, which is a rough estimation of the slope, can be calculated as:

$$EE_i = \frac{Y(x_1, x_2, \ldots, x_i + \Delta_i, \ldots, x_K) - Y(x_1, x_2, \ldots, x_i, \ldots, x_K)}{\Delta_i} \quad (5)$$

After calculating elementary effects for all samples, the global sensitivity score for the system can be generated by combining all elementary effects:

$$\mu_i = \frac{1}{r}\sum_{j=1}^{r} EE_i^j , \quad (6)$$

$$\sigma_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}\left(EE_i^j - \mu_i\right)^2, \tag{7}$$

$$\mu_i^* = \frac{1}{r}\sum_{j=1}^{r}\left|EE_i^j\right|, \tag{8}$$

Where r is the number of trajectories for sample generation during experimental design. The μ_i metric measures the averaged elementary effect, while σ_i^2 represents the variance in elementary effects, quantifying the level of nonlinearity. Because the absolute value of elementary slopes quantifies the strength of influence, to avoid positive and negative elementary effects canceling out, μ_i^* is used to quantify the "averaged absolute slope". Usually, variables with higher metric values are considered more important.

Variance-based methods take a second-order approach, where the variance in the output variable is decomposed into contributions of each input variable, including individual and joint contributions [73]:

$$V(Y) = \sum_{i=1}^{k} V_i + \sum_{1\le i\le j\le k} V_{i,j} + \cdots + V_{i,j\ldots,k}, \tag{9}$$

$$S_i = \frac{V_i}{V(Y)}, \tag{10}$$

$$S_{T_i} = \frac{V_i + \sum_{i\neq j} V_{i,j} + \cdots + V_{i,j\ldots,k}}{V(Y)}, \tag{11}$$

where V_i is the contribution of variable i, $V_{i,j}$ is the contribution of the interaction between variable i,j. Correspondingly, S_i is the first-order sensitivity score, and S_{T_i} is the total

sensitivity score including the first-order effects and all interactions. The difference in S_i and S_{T_i} represents the degree of the interaction between different variables. The sensitivity scores are estimated through Monte-Carlo techniques.

$$S_i = \frac{V_{X_i}\left[(E_{X_{\sim i}}(Y|X_{\sim i}))\right]}{V(Y)}, \tag{12}$$

$$S_{T_i} = 1 - \frac{V_{X_{\sim i}}\left[(E_{X_i}(Y|X_{\sim i}))\right]}{V(Y)}, \tag{13}$$

where X_i represents the combination of all variables except i. Usually, variables with higher metric values are considered more important.

Methodology: Potential Variance Reduction

Our proposed methodology seeks to combine causal inference with the variance-based method to generate risk quantification using observational data. The variance-based analysis is one of the key process systems engineering techniques for CPP identification. By directly quantifying the risks in manufacturing, variance-based methods aid the identification of CPPs from a risk management perspective. Compared to the traditional methods where each variable can be manipulated independently in experiments, this proposed method can infer the relative importance of the process variables based only on observational data and causal assumptions. This proposed method can serve as a preliminary screening of candidate CPPs without the production disturbance required by experimental methods.

Because CPP identification is intrinsically causal, in order to apply sensitivity analysis to observational data, the do-calculus notation system should be introduced to create a generalized formulation. The element effect of the Morris method, for example, can be formulated as:

$$EE_i = \frac{E[Y|x_1, x_2, \ldots, do(x_i = \overline{x_i} + \Delta_i), \ldots, x_K] - Y(x_1, x_2, \ldots, x_i, \ldots, x_K)}{\Delta_i} \quad (14)$$

where $E[Y|x_1, x_2, \ldots, do(x_i = \overline{x_i} + \Delta_i), \ldots, x_K]$ represent the causal query of the expected value output Y if variable x_i is perturbed by Δ_i from the observed value $\overline{x_i}$. Similarly for variance-based methods, do-operator can be applied to represent hypothetical perturbations to the input variables.

$$V_i = Var_{\Delta i}(E_{x_{\sim i}}(Y|do(x_i = \overline{x_i} + \Delta i))) \quad (15)$$

where $\overline{x_i}$ represent the observed value of variable x_i. However, with the increasing number of interacting variables, the causal formulation of the variance-based method will easily become intractable. It should be noted that if the dataset is experimental, the do-operator can be directly removed and these formulations become the original statistical formulation.

This paper introduces potential variance reduction (PVR) as a new causal variance-based method. It is defined as the change of variance in the output if a certain process variable is manipulated. This causal query mimics the scenario of further process improvement, where candidate process variables are manipulated or controlled to improve the process quality attributes Y. Variables with a large PVR value can be chosen as a candidate CPP. Because PVR directly infers the post-intervention outcome without trying to analyze the individual strength of variable interaction, it is analytically more trackable for an observational study.

The definition of PVR can be local or global. Local methods study the impact of variable x_A around a nominal value $\overline{x_A}$:

$$pvr_{local}(x_A = \overline{x_A}) = Var(Y) - Var(Y|do(x_A = \overline{x_A})) \quad (16)$$

where x_A is a process variable of interest, and Y is the system output.

Global methods study the expected effect of variable x_A under altered new distribution $p(x_A)$.

$$pvr_{global}(x_A \sim p(x_A)) = E_{x_A \sim p(x_A)}[Var(Y) - Var(Y|do(x_A = \overline{x_A}))] \quad (17)$$

where $p(x_A)$ defines the distribution of variable x_A in the design space after the intervention. Uniform distribution can be a natural choice. Although the global formulation provides a more holistic view of the design space, it can create problems during the estimation stage. Because available samples may not fully populate the design space, the altered new distribution $p(x_A)$ may not be fully sampled. As a result, the estimation process may suffer from high uncertainty. Because the do-calculus framework provided a nonparametric methodology for solving the causal queries, and the derivation of PVR is kept nonparametric, any statistical model such as neural networks or Gaussian processes can be applied to estimate the statistics.

Case study

A case study based on the steady-state operation of the Van de Vusse reactor, which exhibits interesting nonlinear relationships, is presented to illustrate the effectiveness of the proposed method [74]. The simulation results are collected as observational process data. This generated observational dataset will be analyzed using the proposed methodology, giving full knowledge of the causal mechanism in the form of a DAG. Results using the proposed method are compared against the simulated ground truth. Additionally, sensitivity analysis is applied to this dataset and compared against the proposed method. Without a causal perspective,

sensitivity analysis lacks information on how the underlying data-generating process changes after an effort to improve the process.

Van de Vusse Reactor

The Van de Vusse reactor is a classic example of a nonlinear dynamical system in chemical processes [74]. It comprises two parallel reactions from the raw material A, to desired product B, undesired by-product C, and D:

$$A \overset{k_1}{\rightarrow} B \overset{k_2}{\rightarrow} C, \tag{18}$$

$$2A \overset{k_3}{\rightarrow} D \tag{19}$$

Because components C and D do not influence the dynamics and states of this system, they are neglected in the model. Dilution rate $u = f/v$ is the manipulated input variable to regulate the system. Then, the concentration of A and B can be described by the following ordinary differential equation:

$$\frac{dC_A}{dt} = -k_1 C_A - k_3 C_A^2 + (C_{A_f} - C_A)u, \tag{20}$$

$$\frac{dC_B}{dt} = -k_1 C_A - k_2 C_B + C_B \cdot u, \tag{21}$$

Van de Vusse reactor is a challenging nonlinear process with input multiplicity. Additionally, this system exhibits nonminimum-phase behavior [75]. Therefore, this is commonly used as a

benchmark case study. The parameters for this reactor are given in Table 3.

To mimic real-world process operations which commonly involve disturbances and human intervention, variations in raw material concentration as well as human feedforward compensation were added to this case study. Specifically, the raw material attribute varied from batch to batch, which was assumed to follow a uniform distribution:

$$C_{A_f} \sim U(12,18) \tag{22}$$

As a response to the variations in raw material attributes, a human operator was assumed to be monitoring the process and manually manipulating the feed rate to compensate for the variations in raw material attributes. Human intervention similar to the simulated case study is common and necessary for maintaining the safe operation of manufacturing systems [65,76]. For higher concentration feeds, the operator would lower the feed rate to reduce the variation in the concentration of product B:

$$u \sim N\left(2 + \frac{15 - C_{A_f}}{10}, 0.5\right) \tag{23}$$

Through simulation 2000 samples were generated mimicking the observational process data (C_{A_f}, u, C_B). It should be noted that through the simulated human intervention, the two main process variables were correlated in the observational data, introducing spurious correlations in the dataset. Spurious correlations similar to this are one of the main reasons why a causal framework should be adopted.

The objective of this case study is to identify the relative importance of process parameters, based on which variable contributes more to the undesired variations (risk) in the output C_B. If the feed concentration was selected as the CPP, the process would be improved by using higher-grade raw material. If the feed rate was selected as the CPP, a better flow controller would be implemented to replace the human operator.

Study Design

This case study aims to compare the proposed methodology with sensitivity analysis, which is the most widely used process system engineering tool for identifying CPPs [70].

For the sensitivity analysis techniques, firstly, the generated training data was used to train a surrogate model to represent the input-output relationships. Gaussian process models were used to build the surrogate models because they are nonparametric and Bayesian, which will reduce possible biases in hyperparameter tuning [77]. Secondly, the design space was selected to be within one standard deviation from the nominal operating point where enough samples were available to support the estimation effort. Then both Morris and variance-based sensitivity analysis method was applied using the surrogate model and results were collected.

For our proposed framework, firstly the causal structure was extracted based on a mechanistic understanding of this process, as shown in Figure 6. Then this causal graph was incorporated into the causal inference framework to identify the statistical estimand that answers the causal PVR query.

For the feed rate u,

$$V\big(C_B \big| do(u)\big) = E[C_B|do(u)] - (E[C_B|do(u)])^2 \tag{24}$$

where the estimand for $E[C_B^2|do(u)]$ and $E[C_B|do(u)]$ can be derived from the SCM using the do-calculus:

$$\begin{aligned} E[C_B|do(u)] &= \int_{C_B} C_B\, p\big(C_B\big|do(u)\big) \\ &= \int_{C_B} C_B \int_{C_{A_f}} p\left(C_B\left|u, C_{A_f}\right.\right)\, p\left(C_{A_f}\right) \\ &= \int_{C_{A_f}} E\left(C_B\left|u, C_{A_f}\right.\right)\, p\left(C_{A_f}\right) \end{aligned} \tag{25}$$

$$\begin{aligned} E[C_B^2|do(u)] &= \int_{C_B} C_B^2\, p\big(C_B\big|do(u)\big) \\ &= \int_{C_B} C_B^2 \int_{C_{A_f}} p\left(C_B\left|u, C_{A_f}\right.\right)\, p\left(C_{A_f}\right) \\ &= \int_{C_{A_f}} E\left(C_B^2\left|u, C_{A_f}\right.\right)\, p\left(C_{A_f}\right) \end{aligned} \tag{26}$$

Notice that $E\left(C_B\left|u, C_{A_f}\right.\right)$ and $E\left(C_B^2\left|u, C_{A_f}\right.\right)$ are essentially statistical regression models.

The PVR for feed concentration C_{A_f} was also derived based on the causal graph:

$$V\left(C_B\middle|do\left(C_{A_f}\right)\right) = E\left[C_B^2|do(C_{A_f})\right] - \left(E\left[C_B|do(C_{A_f})\right]\right)^2 \tag{27}$$

Where similarly, the estimand for $E\left[C_B^2|do(C_{A_f})\right]$ and $E\left[C_B|do(C_{A_f})\right]$ can be identified as

$$E\left[C_B|do(C_{A_f})\right] = \int_{C_B} C_B\, p\left(C_B\middle|do\left(C_{A_f}\right)\right) \tag{28}$$

$$= \int_{C_B} C_B \; p\left(C_B \middle| C_{A_f}\right)$$
$$= E\left(C_B \middle| C_{A_f}\right)$$

$$E\left[C_B^2 | do(C_{A_f})\right] = \int_{C_B} C_B^2 \, p\left(C_B \middle| do\left(C_{A_f}\right)\right)$$
$$= \int_{C_B} C_B^2 \; p\left(C_B \middle| C_{A_f}\right) \quad (29)$$
$$= E\left(C_B^2 \middle| C_{A_f}\right)$$

Similarly, $E\left(C_B \middle| C_{A_f}\right)$ and $E\left(C_B^2 \middle| C_{A_f}\right)$ are also regression models.

After identifying the statistical estimand that answers the PVR causal query, the statistical estimands were estimated using the collected observational data using a standard machine learning procedure. Specifically, gaussian process regression was used for its non-parametric flexibility. As a result of the identification stage, all data can be used during the estimation stage without requiring further treatment such as matching.

Results of both the classic sensitivity analysis and the proposed PVR model were compared against the simulated ground truth. The ground truth benchmark was generated by simulating the subsequent process improvement effort given the identified CPP. To represent identifying feed concentration as CPP, variations in the feed concentration were removed and C_{A_f} becomes a constant. To represent identifying feed rate as CPP, variations in the feed rate were removed and u becomes a constant. The selection of CPP that more effectively reduces variations in the system output was considered the ground truth.

Results

The results of using sensitivity analysis for identifying CPPs in the Van de Vusse reactor case study are shown in Table 4. Both Morris method and variance-based method suggested that

feed rate is more critical. If this result is causally interpreted, process improvement efforts would prioritize selecting high-grade raw materials to reduce variations in the feed concentration.

The result of using the proposed methodology is shown in Table 5. Contrary to the results based on sensitivity analysis, the proposed method identified feed rate as the CPP, because it provided a higher potential decrease in output variation. Process improvement efforts would prioritize installing a flow controller.

The ground truth was calculated by simulating the process improvement effect, as shown in Table 6. The corresponding intervention was provided to the simulation based on the identified CPP. It suggested that installing a flow controller is a more effective process optimization effort.

This case study demonstrated that causally interpreting statistical results without proper evaluation of the underlying causal assumption is problematic. As shown in Table 4, blindly applying sensitivity analysis to observational data without understanding the causal mechanism of the underlying data-generating process led to suboptimal attempts for process improvement.

Nonetheless, as shown in Table 5, using the proposed framework, statistical metrics can be identified which quantify the individual contribution of each of the entangled variables. If proper causal assumptions were incorporated into the proposed framework, the right CPPs can be identified based on observational data.

Discussion

The Smart Manufacturing Leadership Coalition (SMLC) defines smart manufacturing as the right data in the right form, the right people with the right knowledge, the right technology, and the right operations, whenever and wherever needed throughout the manufacturing enterprise [4]. Our proposed framework provided a systematic approach to incorporate the right knowledge in the right form to help identify the right data, which in turn drives the right decision. For observational process data, finding the right data to estimate the right statistics is a critical step for process improvement efforts. Unlike carefully designed experimental settings where the focus is to extract causal relationships in the manufacturing system, the majority of data collected from smart manufacturing systems are observational, and the true causal effect is hidden under the entangled statistical correlations. This proposed framework will help bridge the gap between the observational smart manufacturing data and the causal smart manufacturing objective.

For process data analytics tasks, the selection of dataset and statistical estimand relies heavily on experience using a trial-and-error approach [78]. Behind each data selection operation and statistics estimation effort, inexplicit causal assumptions are made. Without conveying the assumptions unambiguously and explicitly, the generated results cannot be systematically validated, and the robustness of the results cannot be analyzed. The proposed framework on the other hand explicitly represented the assumptions using graphs, allowing for easy validation from field experts and efficient robustness checks through sensitivity analysis of the assumptions. Additionally, graphic models of manufacturing systems are more accessible compared to observational studies in economics or social science [79,80], in the form of P&ID or

system flowsheet. As a result, the proposed framework can be integrated seamlessly into the current process data analytics workflow.

It should be noted that the feedback mechanism may introduce loops in the graph, violating the DAG assumption of the proposed framework. Additionally, not all causal queries are identifiable under the given causal assumption. As a result, experimental efforts are necessary to answer many causal queries.

Another challenge facing the proposed method, as well as many other data-driven methods are the lack of support. Due to the existing process design and control strategies, manufacturing systems are usually operated within a tight region in the design space. Therefore, most samples are restricted to the limited regions within the design space, and the majority of the design space remains unexplored. However, many data analytics tasks including CPP identification require queries outside the region where the majority of the samples lie. Especially in the case of causal inference, the methods may try to infer quantities far away from the training samples without the awareness of data availability. Therefore, Bayesian methods should be implemented to estimate the uncertainty of the predictions, such that limited confidence is assigned to out-of-training-distribution prediction [81].

Conclusion

In this work, we introduced a generalizable causal modeling framework for analyzing process data to extract quantitative causal information, in the context of smart manufacturing. Because most process data collected in smart manufacturing systems are observational, the datasets are commonly non-causally correlated and the true causal effects are hidden under convoluted statistical correlations. Based on the Structural Causal Models (SCM) framework, we provided

a workflow for analyzing spuriously correlated observational data to extract causal information.

Firstly, we demonstrated a real-world application where the causal inference framework is used to improve a steel manufacturing process. We illustrated that causally interpreting statistical associations in data will lead to faulty results which contradict mechanistic knowledge. Then we showed how to apply the SCM framework to understand the spurious correlation, and provided a workflow for data-driven causality-based decision support. Lastly, we demonstrated that by utilizing the provided workflow on the steelmaking dataset, the resulting process improvement strategy agreed well with field knowledge.

Secondly, we presented a novel formulation for identifying critical process parameters (CPPs) from observational data using SCM. The causal inference framework was combined with the variance-based method to allow process risk quantification using observational data. Through a simulation case study, we demonstrated that the proposed framework was able to identify the correct CPP, while directly applying sensitivity analysis methods without considering the causal structure led to problematic results.

Future work will include extending the proposed framework to dynamic systems and time-series data. Combining causal inference with system identification techniques may allow us to identify dynamic control models from observational data. Additionally, the proposed method can be integrated into a Bayesian experimental design framework. The first objective of experimental design is to explore the design space and maximize the information about the system. The second objective is to prioritize the experiments on the variables that are more likely to be CPPs. The exploration objective also improves the identification of CPPs. In future work, a Bayesian experimental design framework can combine these two objectives and adaptively balance them to suggest the next sampling point.

Acknowledgement

The authors are grateful to Dr. Ming Tang (ArcelorMittal Global R&D), Rudolf Moravec (ArcelorMittal Global R&D), Dylan Behrmann (Cleveland-Cliffs Inc.), Morgan Baird (Cleveland-Cliffs Inc.) for useful discussions. Financial support has been provided by Isermann Fellowships from the Department of Chemical and Biological Engineering at Rensselaer Polytechnic Institute, and by the US Department of Energy through the Clean Energy Smart Manufacturing Innovation Institute (CESMII), Award DE-EE0007613.

References

1. Ogunnaike BA. *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*. CRC Press; 2011.

2. Verma A, Antony AN, Ogunnaike BA, Hoek JB, Vadigepalli R. Causality analysis and cell network modeling of spatial calcium signaling patterns in liver lobules. *Front Physiol*. 2018;9. doi:10.3389/fphys.2018.01377

3. Davis J, Edgar T, Graybill R, et al. Smart manufacturing. *Annu Rev Chem Biomol Eng*. 2015;6(1):141-160. doi:10.1146/annurev-chembioeng-061114-123255

4. Edgar TF, Pistikopoulos EN. Smart manufacturing and energy systems. *Comput Chem Eng*. 2018;114:130-144. doi:10.1016/j.compchemeng.2017.10.027

5. Yang S, Navarathna P, Ghosh S, Bequette BW. Hybrid modeling in the era of smart manufacturing. *Comput Chem Eng*. 2020;140:106874. doi:10.1016/j.compchemeng.2020.106874

6. Kim YG, Donovan RP, Ren Y, et al. Smart connected worker edge platform for smart manufacturing: Part 1—Architecture and platform design. *J Adv Manuf Process*. 2022;n/a(n/a):10129. doi:https://doi.org/10.1002/amp2.10129

7. Korambath P, Malkani H, Davis J. Streaming workflows on edge devices to process sensor data on a smart manufacturing platform. In: *IEEE 15th International Conference on EScience, EScience 2019*. IEEE; 2019:621-622. doi:10.1109/eScience.2019.00088

8. Ning C, You F. Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming. *Comput Chem Eng*.

2019;125:434-448. doi:10.1016/J.COMPCHEMENG.2019.03.034

9. Qin SJ. Process data analytics in the era of big data. *AIChE J*. 2014;60(9):3092-3100. doi:10.1002/aic.14523

10. Shang C, You F. Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. *Engineering*. 2019;5(6):1010-1016. doi:https://doi.org/10.1016/j.eng.2019.01.019

11. MacGregor J, Cinar A. Monitoring, fault diagnosis, fault-tolerant control and optimization: Data driven methods. *Comput Chem Eng*. 2012;47:111-120. doi:10.1016/J.COMPCHEMENG.2012.06.017

12. Huang B, Zhang K, Zhang J, Sanchez-Romero R, Glymour C, Scholkopf B. Behind distribution shift: Mining driving forces of changes and causal arrows. *Proc - IEEE Int Conf Data Mining, ICDM*. 2017;2017-November:913-918. doi:10.1109/ICDM.2017.114

13. Wilson ED, Clairon Q, Henderson R, Taylor CJ. Dealing with observational data in control. *Annu Rev Control*. 2018;46:94-106. doi:10.1016/j.arcontrol.2018.05.001

14. Chiang LH, Russell EL, Braatz RD. *Fault Detection and Diagnosis in Industrial Systems*. London: Springer; 2001. doi:doi.org/10.1007/978-1-4471-0347-9

15. Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. *J R Stat Soc Ser B Stat Methodol*. 2016;78(5):947-1012. doi:10.1111/rssb.12167

16. Sivaram A, Venkatasubramanian V. XAI-MEG: Combining symbolic AI and machine learning to generate first-principles models and causal explanations. *AIChE J*.

2022;68(6):e17687. doi:https://doi.org/10.1002/aic.17687

17. Ting JYC, Barnard AS. Data-driven causal inference of process-structure relationships in nanocatalysis. *Curr Opin Chem Eng*. 2022;36:100818. doi:https://doi.org/10.1016/j.coche.2022.100818

18. Wang J, Li Y, Gao RX, Zhang F. Hybrid physics-based and data-driven models for smart manufacturing: Modelling, simulation, and explainability. *J Manuf Syst*. 2022;63:381-391. doi:https://doi.org/10.1016/j.jmsy.2022.04.004

19. Zhang K, Schölkopf B, Spirtes P, Glymour C. Learning causality and causality-related learning: some recent progress. *Natl Sci Rev*. 2018;5(1):26-29. doi:10.1093/NSR/NWX137

20. Lennart L. *System Identification: Theory for the User*. Upper Saddle River, NJ, United States; 1999.

21. Sjöberg J, Hjalmarsson H, Ljung L. Neural networks in system identification. *IFAC Proc Vol*. 1994;27(8):359-382. doi:10.1016/s1474-6670(17)47737-8

22. Sadati N, Chinnam RB, Nezhad MZ. Observational data-driven modeling and optimization of manufacturing processes. *Expert Syst Appl*. 2018;93:456-464. doi:https://doi.org/10.1016/j.eswa.2017.10.028

23. Pearl J. An introduction to causal inference. *Int J Biostat*. 2010;6(2):Article 7. doi:10.2202/1557-4679.1203

24. Rathinasabapathy R, Elsass MJ, Josephson JR, Davis JF. A smart manufacturing methodology for real time chemical process diagnosis using causal link assessment.

AIChE J. 2016;62(9):3420-3431. doi:https://doi.org/10.1002/aic.15403

25. Schölkopf B. Causality for machine learning. November 2019. http://arxiv.org/abs/1911.10500.

26. Tian J, Pearl J. A general identification condition for causal effects. In: *National Conference on Artificial Intelligence*. AAAI; 2002.

27. Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York, NY: Cambridge University Press; 2009. doi:10.1017/CBO9780511803161

28. Eide ER, Showalter MH. Methods matter: Improving causal inference in educational and social science research: A review article. *Econ Educ Rev*. 2012;31(5):744-748. doi:https://doi.org/10.1016/j.econedurev.2012.05.010

29. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian Networks to Analyze Expression Data. *J Comput Biol*. 2000;7(3-4):601-620. doi:10.1089/106652700750050961

30. Grimmer J. We are all social scientists now: How big data, machine learning, and causal inference work together. *PS Polit Sci & Polit*. 2015;48(1):80–83. doi:10.1017/S1049096514001784

31. Kamiński M, Ding M, Truccolo WA, Bressler SL. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biol Cybern*. 2001;85(2):145-157.

32. Mania H, Jordan MI, Recht B. Active learning for nonlinear system identification with guarantees. 2020. https://arxiv.org/abs/2006.10277. Accessed February 25, 2022.

33. Schölkopf B, Hogg DW, Wang D, et al. Modeling confounding by half-sibling regression.

Proc Natl Acad Sci. 2016;113(27):7391-7398. doi:10.1073/pnas.1511656113

34. Blinowska KJ. Review of the methods of determination of directed connectivity from multichannel data. *Med Biol Eng Comput*. 2011;49(5):521-529. doi:10.1007/s11517-011-0739-x

35. Engle RF, Granger CWJ. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*. 1987;55(2):251-276. doi:10.2307/1913236

36. Li G, Qin SJ, Yuan T. Data-driven root cause diagnosis of faults in process industries. *Chemom Intell Lab Syst*. 2016;159:1-11. doi:10.1016/j.chemolab.2016.09.006

37. Liu Z, Ge Z, Chen G, Song Z. Adaptive soft sensors for quality prediction under the framework of Bayesian network. *Control Eng Pract*. 2018;72:19-28. doi:10.1016/J.CONENGPRAC.2017.10.018

38. Marcot BG, Holthausen RS, Raphael MG, Rowland MM, Wisdom MJ. Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *For Ecol Manage*. 2001;153(1):29-42. doi:https://doi.org/10.1016/S0378-1127(01)00452-2

39. Vuković M, Thalmann S. Causal discovery in manufacturing: A structured literature review. *J Manuf Mater Process*. 2022;6(1). doi:10.3390/jmmp6010010

40. Hagedorn C, Huegle J, Schlosser R. Understanding unforeseen production downtimes in manufacturing processes using log data-driven causal reasoning. *J Intell Manuf*. 2022:1-17.

41. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci*.

2016;113(27):7345-7352. doi:10.1073/pnas.1510507113

42. Hund L, Schroeder B. A causal perspective on reliability assessment. *Reliab Eng Syst Saf*. 2020;195:106678. doi:https://doi.org/10.1016/j.ress.2019.106678

43. Chiang LH, Kotanchek ME, Kordon AK. Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Comput Chem Eng*. 2004;28(8):1389-1401. doi:https://doi.org/10.1016/j.compchemeng.2003.10.002

44. Wang J, Li H, Huang J, Su C. A data similarity based analysis to consequential alarms of industrial processes. *J Loss Prev Process Ind*. 2015;35:29-34. doi:https://doi.org/10.1016/j.jlp.2015.03.005

45. Wunderlich P, Niggemann O. Structure learning methods for Bayesian networks to reduce alarm floods by identifying the root cause. In: *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. ; 2017:1-8. doi:10.1109/ETFA.2017.8247692

46. Chen H-S, Yan Z, Zhang X, Liu Y, Yao Y. Root Cause Diagnosis of Process Faults Using Conditional Granger Causality Analysis and Maximum Spanning Tree. *IFAC-PapersOnLine*. 2018;51(18):381-386. doi:https://doi.org/10.1016/j.ifacol.2018.09.330

47. Bauer M, Cox JW, Caveness MH, Downs JJ, Thornhill NF. Finding the Direction of Disturbance Propagation in a Chemical Process Using Transfer Entropy. *IEEE Trans Control Syst Technol*. 2007;15(1):12-21. doi:10.1109/TCST.2006.883234

48. Bequette BW. Process control practice and education: Past, present and future. *Comput Chem Eng*. 2019;128:538-556. doi:https://doi.org/10.1016/j.compchemeng.2019.06.011

49. Landman R, Jämsä-Jounela SL. Hybrid causal analysis combining a nonparametric multiplicative regression causality estimator with process connectivity information. *Control Eng Pract*. 2019;93. doi:10.1016/j.conengprac.2019.104140

50. Ma L, Dong J, Peng K, Zhang K. A novel data-based quality-related fault diagnosis scheme for fault detection and root cause diagnosis with application to hot strip mill process. *Control Eng Pract*. 2017;67:43-51. doi:https://doi.org/10.1016/j.conengprac.2017.07.005

51. Rashidi B, Singh DS, Zhao Q. Data-driven root-cause fault diagnosis for multivariate non-linear processes. *Control Eng Pract*. 2018;70:134-147. doi:https://doi.org/10.1016/j.conengprac.2017.09.021

52. Bengio Y, Deleu T, Rahaman N, et al. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In: *International Conference on Learning Representations*. ; 2020. https://openreview.net/forum?id=ryxWIgBFPS. Accessed May 21, 2020.

53. Comas A, Zhang C, Feric Z, Camps O, Yu R. Learning disentangled representations of videos with missing data. *Adv Neural Inf Process Syst*. 2020;33:3625-3635. https://github.com/Rose-STL-Lab/DIVE. Accessed August 5, 2021.

54. Yang S, Rebmann A, Tang M, et al. Process monitoring using causal graphical models, with application to clogging detection in steel continuous casting. *J Process Control*. 2021;105:259-266. doi:10.1016/J.JPROCONT.2021.08.006

55. Pearl J. Causal inference in statistics: An overview. *Stat Surv*. 2009;3:96-146. doi:10.1214/09-SS057

56. Rubin DB. Causal inference using potential outcomes. *J Am Stat Assoc*. 2005;100(469):322-331. doi:10.1198/016214504000001880

57. Bollen KA, Pearl J. Eight myths about causality and structural equation models. In: Morgan SL, ed. *Handbook of Causal Analysis for Social Research*. Dordrecht: Springer Netherlands; 2013:301-328. doi:10.1007/978-94-007-6094-3_15

58. Pearl J, Bareinboim E. External validity: From do-calculus to transportability across populations. *https://doi.org/101214/14-STS486*. 2014;29(4):579-595. doi:10.1214/14-STS486

59. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-688. doi:10.1093/BIOMET/82.4.669

60. Shpitser I, Pearl J. *Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models*. eScholarship, University of California; 2006.

61. Ikäheimonen J, Leiviskä K, Ruuska J, Matkala J. Nozzle clogging prediction in continuous casting of steel. *IFAC Proc Vol*. 2002;35(1):143-147. doi:10.3182/20020721-6-ES-1901.01175

62. Kumar SD, Prasad G, Ghorui PK, Ranjan M. Coolant strategies for BOF steelmaking. *Ironmak \& Steelmak*. 2008;35(7):539-544. doi:10.1179/174328108X335159

63. Liu R, Laus L, Lee Y, et al. Understanding the clogging phenomenon during high-Al TRIP steel casting. *AISTech - Iron Steel Technol Conf Proc*. 2015;2:2443-2453.

64. Landman R, Jämsä-Jounela SL. Hybrid approach to casual analysis on a complex industrial system based on transfer entropy in conjunction with process connectivity

information. *Control Eng Pract*. 2016;53:14-23. doi:10.1016/j.conengprac.2016.04.010

65. Ghosh S, Bequette BW. Process systems engineering and the human-in-the-loop: The smart control room. *Ind Eng Chem Res*. 2020;59(6):2422-2429. doi:10.1021/acs.iecr.9b04739

66. Metta N, Ghijs M, Schäfer E, et al. Dynamic flowsheet model development and sensitivity analysis of a continuous pharmaceutical tablet manufacturing process using the wet granulation route. *Processes*. 2019;7(4). doi:10.3390/pr7040234

67. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng*. 2014;40(1):16-28. doi:https://doi.org/10.1016/j.compeleceng.2013.11.024

68. Jian W, Zhu L, Xu Z, Chen X. A variable selection method for soft sensor development through mixed integer quadratic programming. *Chemom Intell Lab Syst*. 2017;167:85-95. doi:10.1016/J.CHEMOLAB.2017.05.011

69. Thiry J, Krier F, Evrard B. A review of pharmaceutical extrusion: Critical process parameters and scaling-up. *Int J Pharm*. 2015;479(1):227-240. doi:10.1016/j.ijpharm.2014.12.036

70. Wang Z, Escotet-Espinoza MS, Ierapetritou M. Process analysis and optimization of continuous pharmaceutical manufacturing using flowsheet models. *Comput Chem Eng*. 2017;107:77-91. doi:10.1016/J.COMPCHEMENG.2017.02.030

71. Terrazas-Moreno S, Flores-Tlacuahuac A, Grossmann IE. Simultaneous cyclic scheduling and optimal control of polymerization reactors. *AIChE J*. 2007;53(9):2301-2315. doi:https://doi.org/10.1002/aic.11247

72. Saltelli A, Ratto M, Andres T, et al. *Global Sensitivity Analysis: The Primer*. John Wiley \& Sons; 2008. doi:10.1002/9780470725184

73. Dellino G, Meloni C. *Uncertainty Management in Simulation-Optimization of Complex Systems*. Springer; 2015.

74. Bequette BW. *Process Control: Modeling, Design, and Simulation*. Upper Saddle River, NJ, United States: Prentice Hall Professional; 2003.

75. Sistu PB, Bequette BW. Model predictive control of processes with input multiplicities. *Chem Eng Sci*. 1995;50(6):921-936. doi:10.1016/0009-2509(94)00477-9

76. Nguyen HH, Matschek J, Zieger T, Savchenko A, Noroozi N, Findeisen R. Towards nominal stability certification of deep learning-based controllers. In: *Proceedings of the American Control Conference*. Vol 2020-July. IEEE; 2020:3886-3891. doi:10.23919/ACC45564.2020.9147564

77. Damianou AC, Lawrence ND. Deep Gaussian Processes. In: Carlos M. Carvalho, Pradeep Ravikumar, eds. *Proceedings of Machine Learning Research*. Scottsdale, Arizona, USA: PMLR; 2013:207-215. http://proceedings.mlr.press/v31/damianou13a.html. Accessed September 7, 2020.

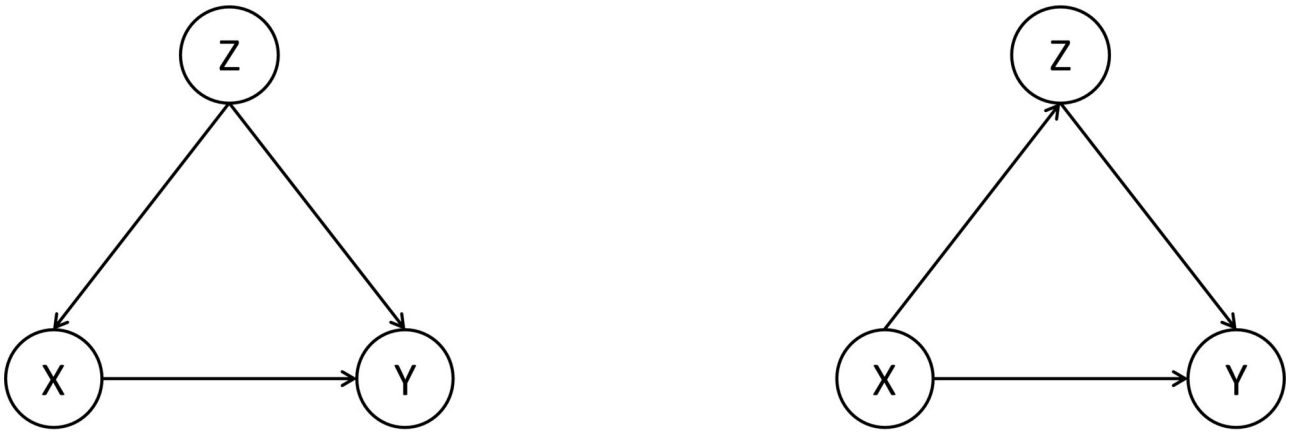
78. Ramírez-Gallego S, Krawczyk B, García S, Woźniak M, Herrera F. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*. 2017;239:39-57. doi:https://doi.org/10.1016/j.neucom.2017.01.078

79. Gangl M. Causal inference in sociological research. *Annu Rev Sociol*. 2010;36(1):21-47. doi:10.1146/annurev.soc.012809.102702

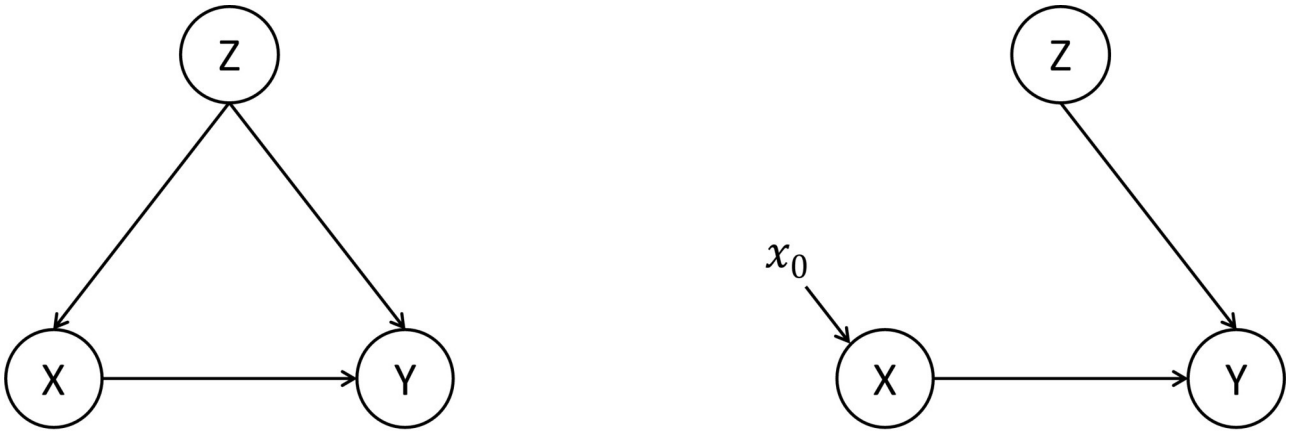
80. Varian HR. Causal inference in economics and marketing. *Proc Natl Acad Sci*. 2016;113(27):7310-7315. doi:10.1073/pnas.1510479113

81. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Precup D, Teh YW, eds. *Proceedings of the 34th International Conference on Machine Learning*. Vol 70. Proceedings of Machine Learning Research. PMLR; 2017:1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

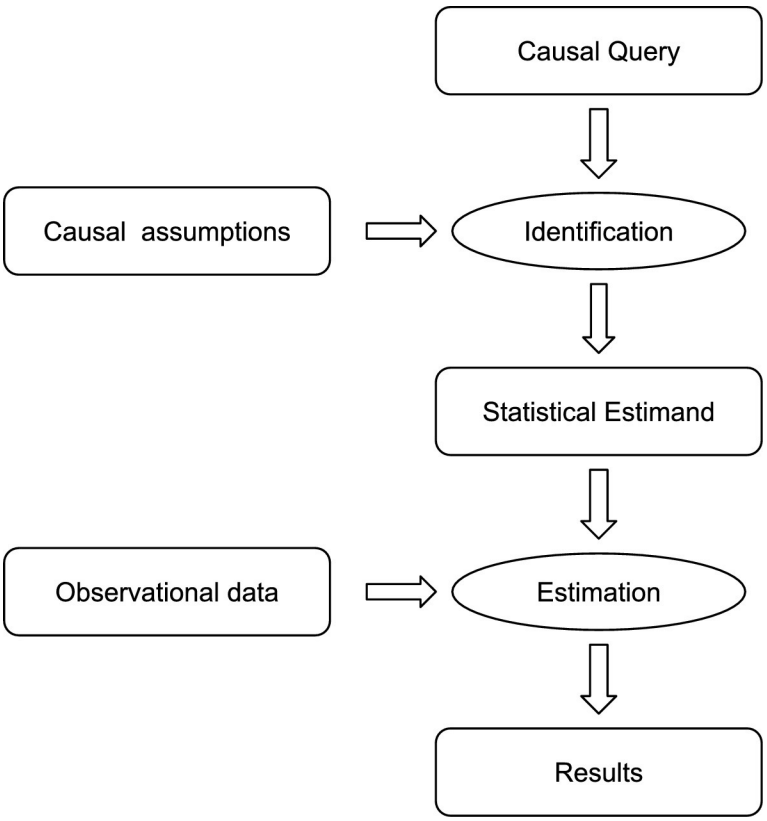
List of Figure Captions



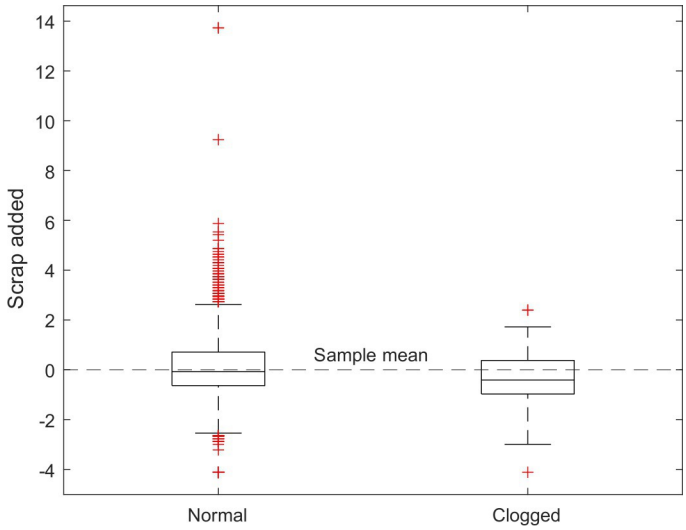
aic_17986_fig1.eps



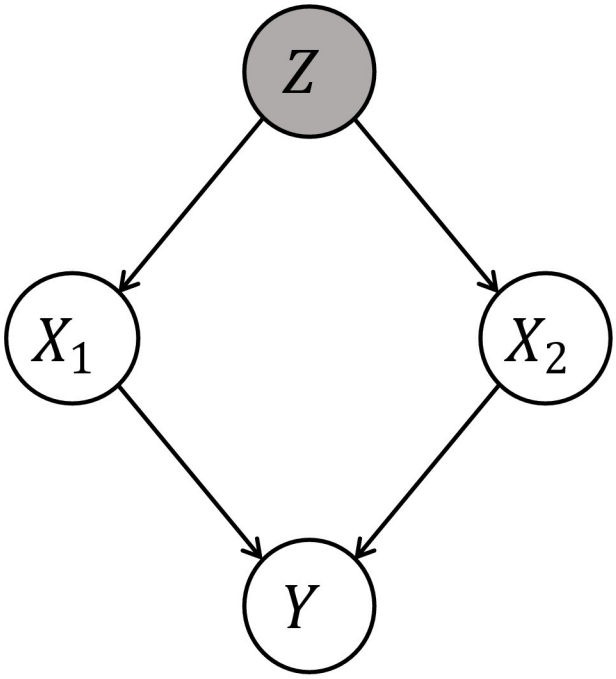
aic_17986_fig2.eps



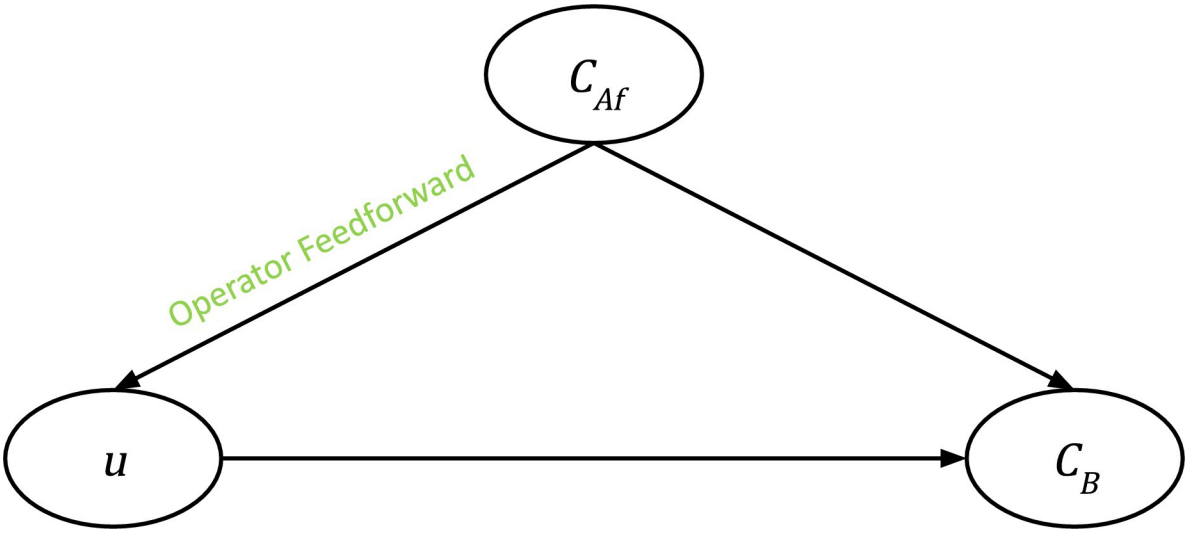
aic_17986_fig3.eps



aic_17986_fig4.eps



aic_17986_fig5.eps



aic_17986_fig6.eps

Table 1. Logistic regression of clogging on scrap addition. This table suggests a negative correlation between scrap addition and clogging.

Parameter	Coefficient	P-value
Scrap	-0.035	0.0474

Table 2. Logistic regression of clogging on scrap addition and aluminum addition. This table suggests a positive correlation between scrap addition and clogging.

Parameter	Coefficient	P-value
Scrap	0.075	0.0376
Al	0.233	0.09

Table 3. Process parameters of the Van de Vusse reactor for the MPC case study.

Parameter	Value
k_1	5/6 min^{-1}
k_2	5/3 min^{-1}
k_3	1/6 $mol/(l \cdot min)$

Table 4. Sensitivity analysis of the surrogate model of the Van de Vusse reactor, trained using observational data.

Parameter	$\boldsymbol{\mu^*}$**(Morris)**	$\boldsymbol{S_1}$**(Sobol)**	$\boldsymbol{S_T}$**(Sobol)**
u	0.192	0.113	0.127
C_{A_f}	0.515	0.876	0.885

Table 5. Potential variance reduction estimated from the observational process data using causal inference, based on full knowledge of the underlying causal graph.

Parameter	**PVR**
u	0.034
C_{A_f}	0.024

Table 6. Result of manipulating different process variables. The manipulations are determined by the identified CPP, where process control is provided to the CPP to reduce the variance in the process output.

Parameter	**Process improvement measures**	$\boldsymbol{V(C_B)}$
u	Flow control ($u = 2$)	0.0254
C_{A_f}	Feed concentration control ($C_{A_f} = 15$)	0.0354