Correlation to causality

Rotari, Marta; Kulahci, Murat

Published in:
Quality Engineering

Link to article, DOI:
10.1080/08982112.2024.2372489

Publication date:
2025

Document Version
Peer reviewed version

QUALITY ENGINEERING
https://doi.org/10.1080/08982112.2024.2372489

Taylor & Francis
Taylor & Francis Group

QUALITY QUANDARIES

Correlation to causality

Marta Rotari[a] and Murat Kulahci[a,b]

[a]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark; [b]Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

ABSTRACT

Causality is important in many engineering applications for the optimization, robustness, and control of manufacturing processes. Randomized experiments have been the conventional tool for establishing and quantifying causal effects. With great advances in sensorics and information technology, the temptation to use an unprecedented amount of observational data for this purpose has been growing. Most classically trained data scientists are warned against such practice as jumping from correlation to causality is presented as a treacherous leap. Yet causal inference based on observational data has been of great interest in, for example, social and medical studies for which randomized experiments can be infeasible or even unethical. In this Quality Quandaries, we propose a compromise where observational data, with the help of process expertise, is used to establish the set of factors that will be further tested for causality. We demonstrate a practical application of our proposal through a case study in additive manufacturing.

Savoir pour prévoir, afin de pouvoir.

Auguste Comte

1. Introduction

The above quote often attributed to the founder of modern positivism Comte reflects his views on the goal of science of going beyond knowing for the sake of knowing. We find the direct translation somewhat challenging because of the verb "pouvoir," which we believe can be best translated using the modal verb "can" with no infinitive form. Therefore, we will instead use "to be able to." Hence, the rough translation of the quote would be "to know to predict in order to be able to." We find the sequence of these infinitives very relevant for the topic we cover in this Quality Quandaries. For the reader of this article steeped in data science, we refer to the sequence of descriptive models (savoir) that can lead to predictive models (prévoir) with the ultimate goal of developing prescriptive models (pouvoir). The age-old quandary on moving from correlation to causality lies in the connection between the predictive and prescriptive models.

We would be remiss however if the above introduction evokes the impression that correlation is only relevant in the pursuit of causality. Correlation, a.k.a. in certain circles association, is at the heart of many modeling approaches. Through correlations between the inputs and the output, we can build predictive models without necessarily requiring the relationship to be causal. Here is a small example we often use in our presentations to industry to emphasize the distinction. Imagine a dairy farmer with free range cows grazing in the fields during the day only to return to the barn every night. Over the years, the farmer has observed a peculiar yet very consistent pattern in his cows' behavior. When, on a given day, the cows arrive back at the barn before sunset; the next day happens to be sunny with clear skies. However, when the cows return to the barn after the sunset, the next day tends to be rainy and dark. This observation (or association/correlation) gives the farmer the ability to predict the next day's weather without knowing whether there is a causal relationship between his cows' behavior and the weather. The stronger this association is, the more reliable the farmer's prediction will be. How would this predictive ability be of use to the farmer, though? Consider, for example, the fact that the farmer is to throw an outdoor wedding party for one of his kids.

CONTACT Murat Kulahci muku@dtu.dk Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark.

If the day before the cows return to the barn late, he would most likely and somewhat confidently resort to setting up a big tent to allow the party to take place as planned. This is an example of the analyst using the observed association to minimize the perceivably negative impact of the predicted outcome. Indeed, predictive models are great tools, if for nothing else, certainly for *risk management*. Yet, in this example, it is not difficult to contemplate that the farmer would much prefer to secure nice weather on the wedding day for various reasons, the least of which is the cost of renting a tent. If that is true, could the farmer take advantage of the association he has observed over the years? Can he, for example, round his cows up personally before sunset on the day before the wedding and hence, according to the observed association, guarantee nice weather? In other words, can he use this association, however strong it may be, to actually cause the change of his choosing? Can he leap from correlation to causality since achieving a specific outcome rather than preparing oneself for that outcome requires more than just an observed association? This hypothetical often invokes an affirming smile in the audience, something much desired to achieve when giving a scientific talk for industry. But we also believe it highlights the main concern related to the instinctive urge to move from correlation to causality as depicted in Figure 1.

In our research group, we do at times work on projects involving financial and medical studies, yet a majority of our projects are in manufacturing. That is why we will admittedly take on a somewhat restrictive view in our discussion mainly focusing on data analytics concerning engineering applications, or simply production analytics. The need for determining the causal relationships then becomes evident as many such applications aim at optimization, robustness and control of the processes, all of which rely on achieving a desired level for the output of interest through manipulating the input variables that are causally related to that output. The causal inference in these applications has traditionally been made through randomized experiments. Yet with the recent push for digitalization in industry and advances in information technology and sensorics, an unprecedented amount of data is being generated and collected in manufacturing. Such data are classified as primarily observational in that it is collected with minimal interference to the system as in the case of placing a temperature sensor in a blast furnace. The sheer volume of this observational data has naturally led the practitioners to question the need for collecting even more (and often expensive) data through randomized experiments bringing us back to the earlier quandary: can we make causal claims through observational data?

In the next section, we discuss the difference between experimental and observational data and its implication on causal inference. We also discuss the considerable attention that has been given to causal inference in various fields such as social and medical studies, as well as our thoughts on the reasons for this focus. In "Causality in engineering applications" section, we delve into the details of data analytics in manufacturing and their connection to causal inference. We then present in "Proposed approach" section our preferred approach based on our experience and our outlook for data analytics in manufacturing, at least for the near future. To illustrate the application of our approach, we present in "Case study: additive manufacturing" section a case study in Additive Manufacturing (AM), which has garnered significant attention in recent years due to its potential to revolutionize manufacturing processes. Finally, in "Conclusion" section, we present some reflections and conclusions.

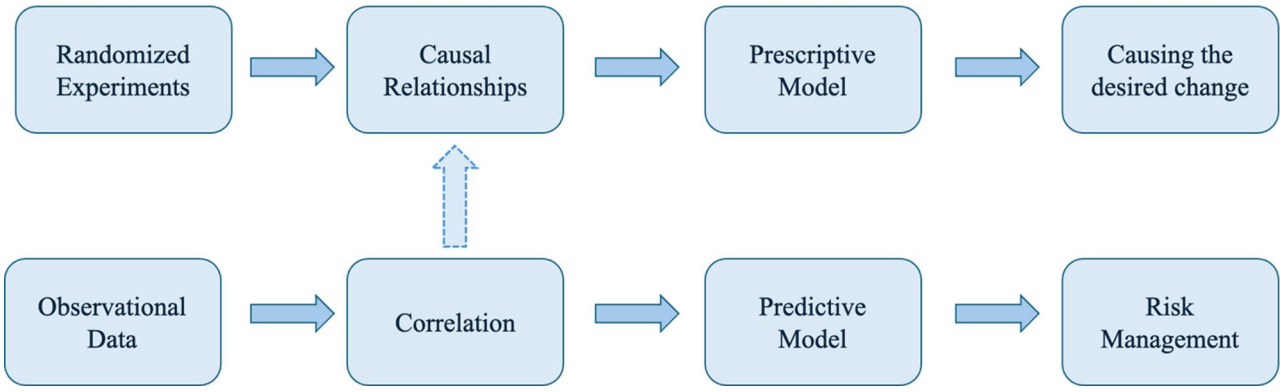

Figure 1. Observational and experimental data and their respective use. The dashed arrow represents the temptation for jumping from correlation to causality.

2. Causal inference

Causality and the meaning of causation have been a philosophical debate. This is however beyond the scope of this Quality Quandaries. The readers can find an abbreviated version of this debate, particularly in the context of data science in Holland (1986). We rather focus our attention on the means of achieving causality through randomized experiments, as most astute experimenters are taught. The simplest case can be made for experimenting with only one factor, during which while all other conceivable factors are kept constant, the factor of interest is changed so that any observed change in the response is causally attributed to that factor. The foundations of this approach are principally laid down by R. A. Fisher when he advocated factorial designs involving simultaneously varying multiple factor settings accompanied by the proper methods for the analysis of the data from these experiments to test for the (statistical) significance of the found relationships (Fisher 1925, 1935, 1956). Fisher's approach was certainly different from collecting observational data without any intervention as we discussed in the previous section. His views on the limitations of the inference that can be made from observational data are best illustrated in his staunch criticism of the studies of his time relating cigarette smoking to lung cancer. He was initially supported by other prominent statisticians. Still, he seemed to have remained as the main objector to the claims of a causal relationship between smoking cigarettes and lung cancer that are drawn from observational data. Cornfield et al. (1959) meticulously address all the justifiable criticisms laid down against the strong causal claims made in the reports by both American Cancer Society and Medical Research Council of Great Britain in 1957 and conclude that

> In the light of all the evidence on tobacco, and after careful consideration of all the criticisms of this evidence that have been made, we find ourselves unable to agree with the proposition that cigarette smoking is a harmless habit with no important effects in health and longevity. (Cornfield et al. 1959)

A greater expose of this conflict can be found in Chapter 18 of Salsburg's "The Lady Tasting Tea" (Salsburg 2001). Salsburg concluded that "You cannot prove anything without a randomized experimental design, said Fisher. Some things do not lend themselves to such designs, but the accumulation of evidence should prove the case, said Cornfield." We believe the latter is an exemplary case addressing the desire to make causal inferences through observational data and develop the methodological groundwork for it in certain fields such as social and medical studies. In these fields, it can indeed be challenging, infeasible, and even unethical to run randomized experiments. In the absence of perceivably the only tool for causal inference, using available observational data for this purpose has led to the development of approaches such as structural equation modeling and directed acyclic graphs.

In 1921, Sewall Wright introduced the "path analysis." This work later evolved into the Directed Acyclic Graphs (DAGs). Judea Pearl made substantial contributions to promoting DAGs in causal inference (Pearl 2009a, 2009b). DAGs are valuable instruments for comprehending and visualizing the complex causal pathways among variables (Dawid 2010). They can help establish causality by identifying potential causal pathways and excluding alternative explanations. They can also be used to identify possible sources of bias or confounding. DAG provides a visualization tool where the variables are represented as nodes and the causal relationships between variables are represented as directed edges or arrows. The direction of causality is indicated by arrows, with the arrow's tail representing the cause and the arrow's head representing the effect. DAGs also permit the representation of confounding variables as depicted in Figure 2. In that regard, DAGs offer a transparent and interpretable means to visualize causal relationships among variables, and they are valuable tools that help avoid causal inference pitfalls. Constructing DAGs could however be challenging, demanding a clear understanding of underlying causal relationships.

Another approach used in causal inference is the Structural Equation Modeling (SEM) (Bollen 1989; Weston and Gore 2006). SEM is a statistical technique employed to test and validate complex relationships between variables and estimate the magnitude and

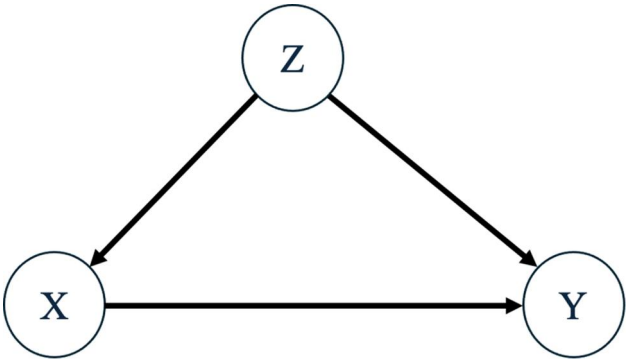

Figure 2. A simple DAG with the path between X and Y representing the causal effect. Z represents the observed confounder potentially causing the spurious correlation between X and Y.

direction of these relationships while also identifying the causal mechanisms that underpin them. In SEM, the variables are represented as either latent or observed variables, and their relationships are expressed through a set of equations that identify direct or indirect effects. In SEM, causality is used to establish the direction of the relationships between variables. Although SEM is a powerful tool for examining complex relationships between variables, it is crucial to recognize its inherent limitations. The challenges include larger models needing large data sets, assessing fit accuracy, computational difficulties, and subjective model interpretation requiring expert knowledge.

Finally, Instrumental variables (IV) have been introduced as a statistical technique for determining causal effects (Wu et al. 2022). It is primarily used when dealing with confounding variables. To control for unmeasured confounding, a third variable, named instrumental variable (IV), is introduced, which is a cause for the input variables and has no direct effect on the outcome. The instrumental variable is used to identify the true correlation between the input variable and the outcome. In machine learning, instrumental variables are used to address confounding bias in observational studies or experiments with non-randomized treatments for data mining and explanatory analysis purposes. However, there are limitations to using instrumental variables, such as the requirement of the strong assumption that there is no direct relationship between the instrumental variable and the outcome. In addition, the instrumental variable must be independent of any unmeasured confounding variables that may influence the outcome.

Constructing causal models mentioned above requires a clear understanding of the process or the system, including potential confounding factors and intermediate variables. In situations where the systems are inherently too complex involving many variables and their possible interactions, constructing accurate DAGs or SEM models becomes highly challenging. Building reliable causal models also becomes extremely problematic if the underlying processes are not well-defined or fully understood. Another issue is the assumptions made in modeling that require verification and often the involvement of expert knowledge (Hernán, Hsu, and Healy 2019), which may not always be readily available.

3. Causality in engineering applications

As mentioned in the Introduction, causality is also paramount for many engineering applications particularly concerning optimization, robustness and control of the processes. Many of the causal inference methodologies mentioned in the previous section have not been widely used in engineering applications. We believe there are several reasons for this. First, randomized experiments have been readily accepted and widely used in both R&D and production. We believe one of the most seminal papers in industrial experimentation by Box and Wilson (1951) has played a very crucial role in this. In that paper, the authors introduce the basic principles of what we now know as the response surface methodology (RSM). This represents one of the watershed moments that identifies the major differences in agricultural and industrial experimentation. The relative immediacy of the industrial experiments renders sequential experimentation a more realistic and preferred iterative approach for learning and discovery (Kulahci and Box 2003; Kulahci and Menon 2023). This is also very much in line with George Box' philosophy on the connection between statistics and science as depicted in Box (1971). RSM offers engineers the possibility to optimize their processes through simple yet highly effective tools.

The experimentation for robustness has its roots in agricultural experiments, see for example the book by Pearson and Wishart (1942) on the collected works of William S. Gosset (Student). Some industrial applications have also been proposed (Michaels 1964). Still, it can be argued that it is through Taguchi's work that the concept of robustness has gained its prevalence in industry (Taguchi 1992). The split plot designs or Taguchi's inner and outer arrays have been successfully used for this purpose. As for the control of the processes, the field of system identification is mainly concerned with modeling the relationship between the input(s) (manipulated variable) and the output (the controlled variable) to obtain the so-called transfer function, which can then be used for maintaining the output at a pre-determined level as in the case of feedback control. In this pursuit, engineers make extensive use of experimental design techniques to collect the needed data.

All these applications offer evidence for how ubiquitous the use of experiments is for engineers for causal inference. In that regard, feasibility and ethical concerns often plaguing experimental campaigns in social and medical studies can be of lesser concern in industrial experimentation. Yet the cost of experimentation both in R&D and particularly in the actual production has always been the major deterrent for industrial experiments as well. While more effective

experimentation methods have been proposed over the years, in many industrial applications, there has always been considerable pressure for using observational data even for causal inference, rather than steadfastly demanding running randomized experiments for the same purpose. As mentioned earlier, with steadily increasing capacity for collecting, storing and analyzing the process data, this pressure has been understandably growing.

In our projects, we found several reasons limiting the use of observational data for causal inference in industrial applications. Firstly, in retrospective studies involving historical data, there is a good chance that the information content in the existing data is questionable or irrelevant to the specific problem at hand. The inconsistencies in historical data are abundant due to changes in operating conditions, data collection schemes and measurement systems rendering often only a small and usually most recent part of the data relevant for analysis. We should also keep in mind that the historical data have usually been collected for documentation purposes to for example trace back a bad batch of products for a potential recall (Kulahci et al. 2020). In that regard, the historical data may not be directly usable for determining causal relationships. Secondly, in some more mature industries, the principle of "if it ain't broke, …" seems to be in effect. That is, the processes are run at well-established operating conditions that render them very stable but also generate process data with minimal variation. Any causal study would then require exciting these processes to generate variation in the operating conditions to establish causal models (Rønsch, Kulahci, and Dybdahl 2022). Furthermore, many processes operate in closed-loop or, at the very least, are controlled manually to ensure stability, contributing to the above-mentioned issue (Capaci et al. 2020).

Finally, as a side product of automated data collection schemes and advances in sensorics, the process data are often multidimensional containing tens if not hundreds of process variables. This creates serious challenges in the analysis as such data would most likely be highly cross-correlated and autocorrelated depending on the sampling frequency and system dynamics. The complexity of the processes has also compounded with many studies moving from unit operations to the entire production process to wholistically consider every step of the production and their impact on the outcome. All these issues greatly impair the potential use of the approaches mentioned in the previous section. Yet we firmly believe that we should find ways to use observational data in causal studies in engineering applications as well due to their increasing availability and accessibility. One such strategy is to use multiple sources of evidence. It is possible to triangulate information from different sources to strengthen causal claims as in the case of Cornfield's rebuttal. Drawing upon diverse archival records and other primary sources allows for a more comprehensive understanding of the historical context and potential causal mechanisms.

Another strategy involves examining temporal precedence. Demonstrating that the cause precedes the effect in historical data can support causal arguments. Analyzing the timing and sequencing of events can offer valuable insights into potential causal relationships. The consistency of patterns across numerous historical cases is another crucial aspect of establishing causation. If similar causal relationships are consistently observed, it strengthens the plausibility of a causal connection. Identifying and highlighting these consistent patterns can further enhance the evidence for causation.

In situations where the input space is extensive, the use of observational data is necessary to narrow down the input space. In such situations, latent structures-based regression techniques such as Partial Least Squares (PLS) provide an appealing solution by analyzing the observational data in conjunction with the corresponding output data. These methods facilitate the inference of causality within a reduced space. Several approaches based on Latent Variable Regression Model Inversion (LVRMI) have been proposed in the literature (Liu et al. 2011; MacGregor et al. 2015; Palací-López et al. 2019).

4. Proposed approach

Panahi, Steiner, and De Mast (2023) discuss the determination of the cause of the variation in the output. Their discussion is based on a simple model with one input for illustration purposes and clearly depicts what can be concluded using experimental and observational data. We find their work to be very relevant to the topic of this Quality Quandaries and highly recommend the authors' articles on related topics as well (De Mast et al. 2023; Panahi, Steiner, and De Mast 2021).

In our projects with industry, we have also worked with both experimental and observational data and would propose combining both for causal inference. But we start with an explicit assumption, the violation of which renders the pursuit somewhat trivial. We assume that the process we deal with is highly

complex and generates high dimensional data with intricate interdependencies. Based on our collaborations with industry particularly in the last decade, we found this to be a reasonable assumption. The immediate consequence however is that the approaches for causal modeling we discussed in "Causal inference" section will most likely not be feasible. Yet direct use of randomized experiments even including a screening phase will most likely not be feasible either due to the sheer number of process variables. Therefore, perhaps due to our background in engineering, we would like to recommend a pragmatic approach and advocate aiming to answer two simple questions consecutively:

1. What do we have now?
2. What else can we do?

What do we have now? This question is *retrospective* and involves historical/available data. Despite the concerns we discussed in the previous section regarding the historical data in manufacturing, we implicitly assume the information content in the historical data is relevant for the causal study and we have the means to extract that information. The former is a strong assumption and can be confirmed once the initial analysis is performed. If there is evidence against this assumption, posing the second question is the most natural next step anyway. The latter however needs further elaboration. We do not necessarily espouse any specific modeling approach. We have in the past used statistical, chemometrics, machine learning, and deep learning methods as the situation required. In the context of this Quality Quandaries, what we expect is the retrospective study to narrow down the set of process variables that will be considered in the next stage for randomized experiments. That is, the first question implies a screening process and therefore the model used in this stage should be at a certain level of transparency and allow for the selection of a few variables that will be passed on to the next stage. The high dimensionality of the data brings several hinderance on this front. The implication of screening the process variables is that the modeling approach offers a way to determine the importance of the variables in the model. It should be noted that we used the word "importance" advisedly over "significance," a term more familiar to statisticians. In parametric models as in multiple linear regression (MLR) with imposed assumptions on the error term, hypothesis testing can be performed on the model coefficients using a pre-defined significance level. Further testing the model assumptions on the error term as well as the subjective selection of the significance level aside, this offers the analyst the means to screen out the process variables that contribute significantly to the prediction of the response. However, high dimensionality of the input data, which increases the likelihood of high cross-correlation among the variables, can easily hinder the performance of more basic models such as MLR.

In non-parametric approaches as in some machine learning models, the algorithms may provide the ranked list of importance of the variables based on an accuracy measure. However, it is usually left up to the analyst to decide on the final list of the variables that significantly contribute to the prediction. There have been some attempts to remedy this as in the case of random forest models by adding dummy variables to the data and only retaining process variables reported as more important than the dummy variables. But this cannot necessarily overcome another concern related to the importance measure. The potential high correlation among the input variables can lead to overestimating the set of important variables as some of these variables would simply be "tagging along" with a causal input with which they are correlated. Therefore, at this stage using at the very least semi-transparent predictive models that allow for a clear selection of a handful of variables that are deemed important/significant is paramount. We do not however expect nor propose this to take place in isolation using a purely data-driven approach. As recommended or even demanded in many causal modeling strategies, we believe it is extremely crucial to bring in the process expertise to decide on the validity of the selection of the important variables. We should as in many data analytics applications in manufacturing make use of the deep engineering understanding of our processes as well as past experiences. These may not be sufficient to construct the causal models we discussed in "Causal inference" section that can address the complexity of the processes fully but would be a great resource in interpreting the results from the data-driven predictive model.

What else can we do? This question involves a prospective study where the outcome from the retrospective study could, for example, be followed by the decision to collect more observational data if the results of the predictive model were ambiguous or the model fit was not satisfactory. The decision on what and how to collect the new data should be made in collaboration with the sensorics experts, IT personnel and engineers/operators who are running the process. The proper installment of the right sensors

accompanied by the automatic data collection and storage tools would be vital to avoid further cost and delay in the investigation. Engineering expertise would also assist in selecting the type, location and sampling frequency of the process variables to be considered. In all of this, the initial study should play an essential role in guiding what *else* can be done.

If the initial predictive model reveals a proper set of process variables as significant and necessary in predicting the response, the next step would be a prospective study to determine the *vital few* among this *trivial many*. For this we implicitly assume that running randomized experiments is possible. At this stage, we would advocate a sequential approach based on factorial designs and their extensions as in the case of central composite designs, to go from screening to optimization if that is indeed the goal (Kulahci and Menon 2023). However, it is perfectly reasonable to consider more run-efficient, one-shot designs as in the case of optimal designs or definitive screening designs (Montgomery 2020).

As a final note, we would be remiss if we did not mention a research field within data science that seems to be at the crossroads of both experimental and observational data: active learning, which can be roughly seen as sampling in large data environments (Cacciarelli and Kulahci 2024; Cacciarelli, Kulahci, and Tyssedal 2024). It is often used in cases where obtaining the response (or labeling/supervision) is relatively more expensive than acquiring the input data. The input data are mostly unsupervised (unlabeled) meaning without any corresponding response value. Then the question is to figure out for which input values it is most crucial to obtain the corresponding response values. Often, the criterion used to determine the next labeling is related to the prediction accuracy of the model that is used to connect the inputs and the output. In that regard, it is also related to the so-called semi-supervised learning (Frumosu and Kulahci 2018). We believe more research is needed on active learning in manufacturing applications to develop proper sampling strategies.

5. Case study: additive manufacturing

In this section, we provide a case study in Additive Manufacturing (AM), to which the proposed approach is applied. AM, also known as 3D printing, is an innovative manufacturing process that enables the production of products with complex geometries and high levels of precision and efficiency. It is a revolutionary approach to manufacturing in which objects are constructed layer by layer using digital designs. While AM has demonstrated great promise, it remains relatively unexplored and faces numerous challenges. Many studies are conducted aimed at process understanding and optimization (Stichel et al. 2018; Yeh et al. 2024). One of the key problems is analyzing process parameters and identifying their optimal levels to ensure high and uniform quality standards.

In AM as in all production processes, a comprehensive understanding of the causal relationship between input variables and output quality is crucial in achieving process optimization and control. Multiple data sources can be identified within the AM process. The input space encompasses machine settings or *process parameters* established at the beginning of an order, representing the batch production of the products. Throughout the production chain, diverse sensors are strategically placed to gather data during production generating *process data* that encompasses a wide range of production-related information. The *output data* reflect the quality of the manufactured products, such as mechanical properties (Tensile strength and Young's Modulus) or defects such as dimensional defects, roughness, etc. The process we study involves industrial-sized AM equipment that can allow for batch production of 100s if not 1000s of products at once. The primary objective is to identify the process parameters that impact the quality, thereby establishing a causal relationship between these parameters and the output that can then be used for process optimization. For further details regarding the AM process, see Rotari and Kulahci (2024).

Emerging technologies often exhibit a greater complexity resulting in a large number of process parameters that must be configured at the outset of the process. The large number of process parameters renders it impractical to conduct an experimental design that covers the entire factor ranges and their combinations. An alternative approach is adopted to address this challenge, as elaborated in "Proposed approach" section.

The process data are obtained from 18 different batches of production. During the production process, data for 53 process variables, labeled as V1 to V53 for the sake of confidentiality, are collected. Our main objective is to investigate the mechanical properties of the final products, specifically the tensile strength, which represents the maximum tensile stress a material can endure before breaking or deforming. We employed a random forest model to establish the relationship between process variables and the output.

Subsequently, we used the mean decrease in accuracy measure to rank the variables based on their importance in the prediction process as shown in Figure 3. However, relying solely on this ranking made it challenging to determine the precise number and the identity of the variables that genuinely influenced the output. Furthermore, due to high correlation among the process variables, the ranking's reliability was compromised as discussed in the previous section. To address this issue and select the most relevant variables, we employed the Conditional Boruta model (Rotari and Kulahci 2024), which is specifically developed to handle correlated variables.

The result of the variable selection is depicted in Figure 4. Notably, three process variables have emerged as relevant: V36, V49, and V39. The next step involved the validation of these results using expert knowledge. After consulting with the process engineers, we concluded that each of the identified variables contributed to the specific aspects of the manufacturing process. Variable V36 is associated with creating each micro-layer generated during 3D printing. This variable is relevant to understanding the fundamental building blocks of the manufacturing process. Variable V49 represents the fusion temperature of each layer. Temperature plays a pivotal role in determining the cohesion between molecules, making it a crucial factor influencing the physical characteristics of the final product. Lastly, variable V39 pertains to light exposure, which gradually deteriorates.

The process engineers subsequently linked the selected process variables from the Conditional Boruta

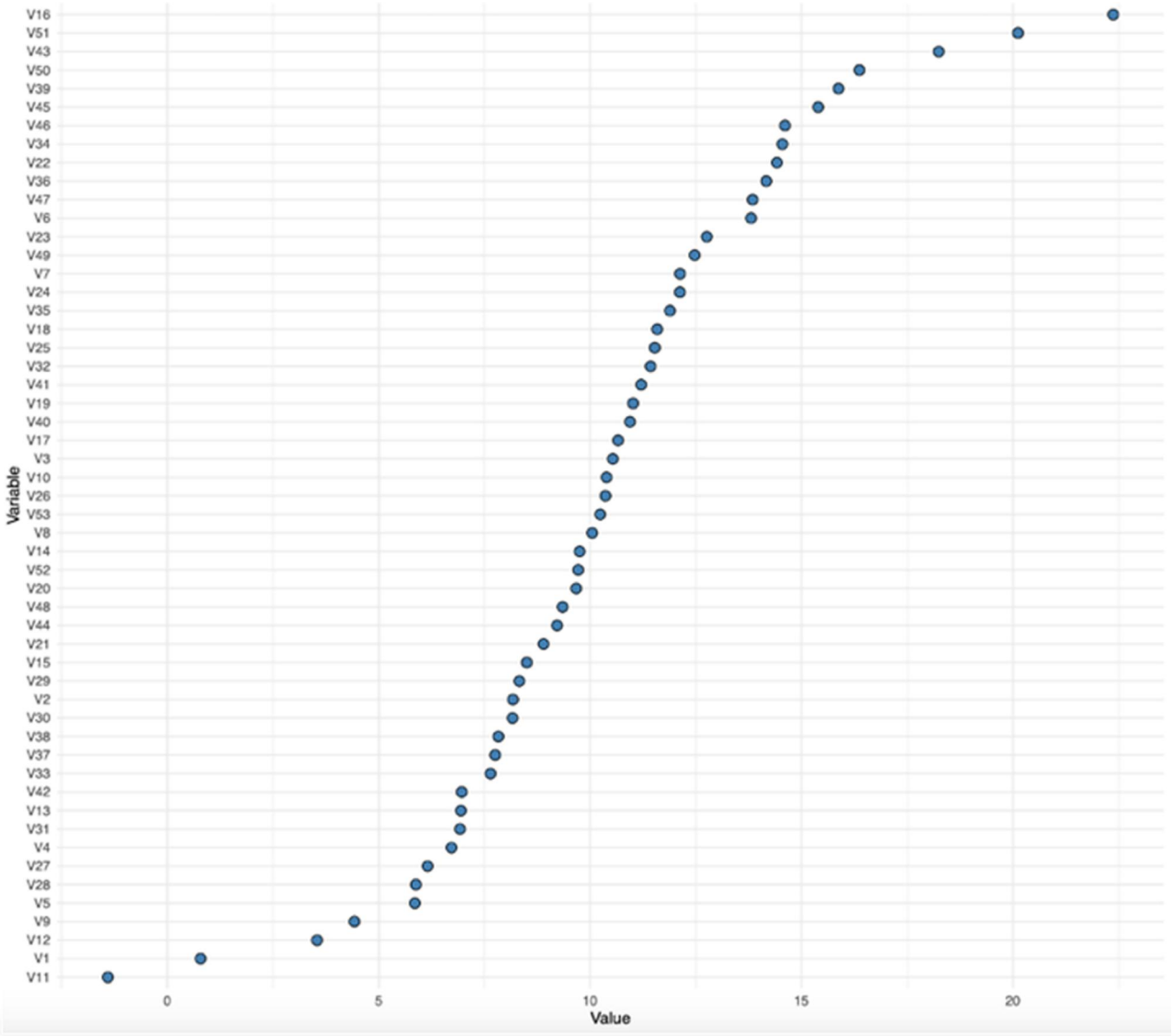

Figure 3. Variable importance ranking from the random forest model.

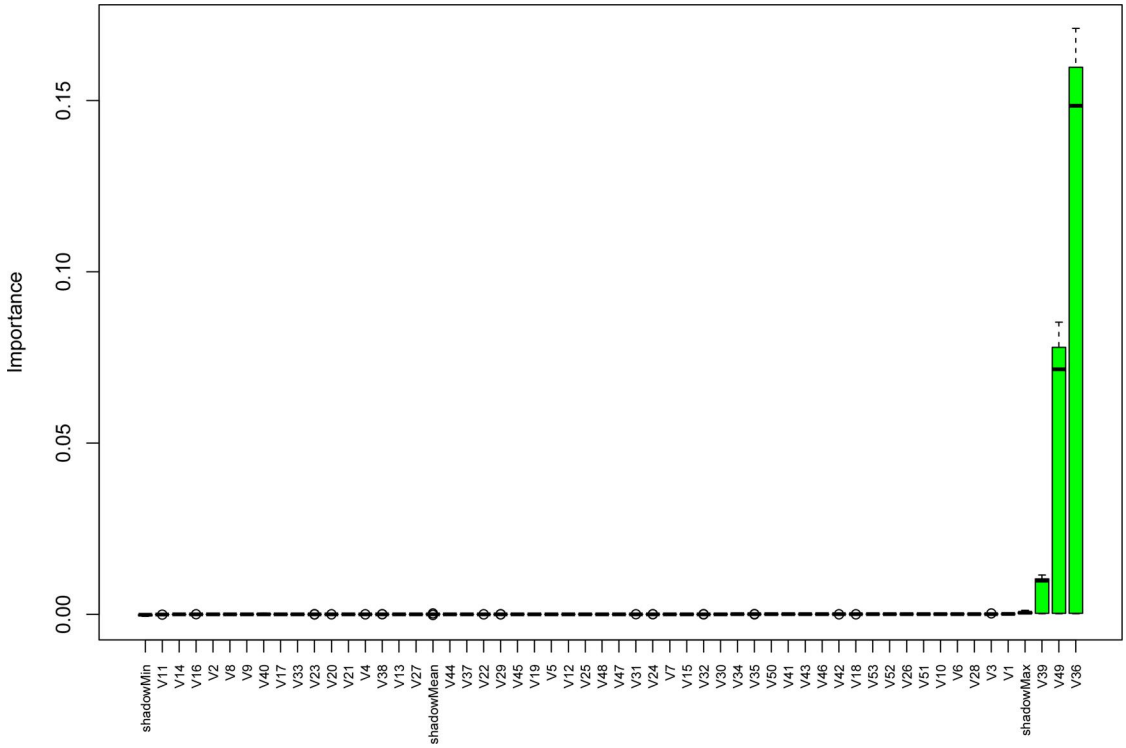

Figure 4. Variable selection using conditional Boruta algorithm.

model to the process parameters of the AM equipment. The process parameters thus identified are denoted as Factors 1, 2, and 3 due to confidentiality reasons. We then moved to the experimentation phase, where we aimed to investigate the causal relationship between the three factors and the outcome, the tensile strength. To accomplish this, we conducted a 2^3 full factorial design consisting of eight experimental runs without any replications due to the cost of experimentation.

The Normal plot of the effects is shown in Figure 5. The main effects of Factors 2 and 3 are tentatively identified as significant. A model with only these main effects is fitted and the resulting coefficients with the corresponding p values are summarized in Table 1. The identification and the quantification of the causal effects were very well received by the engineers. For further analysis, it is recommended to expand the original factorial design by adding center runs to check for curvature. If the curvature is present, the design can further be expanded to a central composite design by adding some axial runs and additional center runs. This is expected to allow for maximizing the products' tensile strength.

Keen data scientists among the readers can undoubtedly critique our choices of the predictive model and the experimental design, and suggest alternatives. But this is beyond the point we want to make

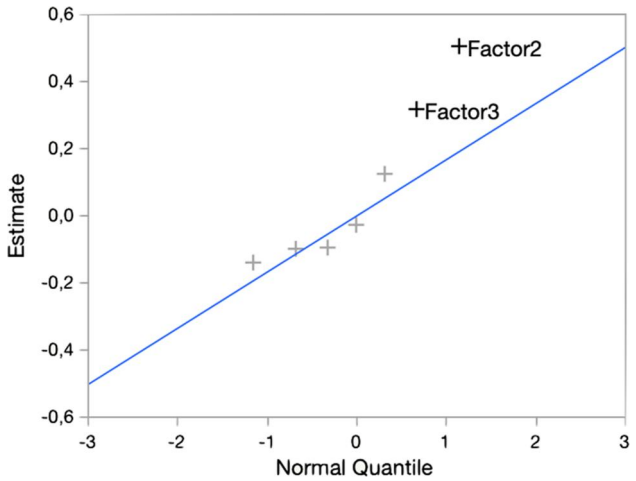

Figure 5. Normal plot of the effects from the 2^3 design.

with our approach. We propose the efficient use of abundant observational data and process expertise to set the scene for proper experimentation for causal inference that would otherwise be overwhelmed by the sheer number of variables to consider. In that regard, any predictive model that can allow for variable selection in the retrospective study and any experimental design that can identify and quantify the causal effects in the prospective study would be

Table 1. Coefficient estimates from the linear regression model for the tensile strength.

Term	Coefficient estimate	Standard error	*t* ratio	*p* Value
Intercept	37.801	0.105	358.35	<.0001
Factor 2	0.503	0.105	4.77	.005
Factor 3	0.315	0.105	2.99	.031

perfectly acceptable. Those we believe are simply the "dealer's choice."

6. Conclusion

Establishing causal relationships in manufacturing is essential for process optimization, robustness and control. The standard approach of randomized experiments for causality may not be feasible for complex manufacturing processes involving many process variables. These processes also produce unprecedented amounts of observational data that can potentially be used for casual inference as in social and medical studies. Yet the complexity of these processes also hinders the use of causal inference tools developed in these studies. We suggest using both observational and experimental data with the former used in a predictive model that provides a set of candidate factors that will be tested using the latter to obtain a prescriptive model. That way following Comte's view on science, we can establish the desired learning process of sequentially moving from the descriptive model to predictive model to finally achieve the prescriptive model.

Our recommended approach is based on the constraints we have encountered over the years related to the type and cost of data that can be collected from manufacturing processes and the complexity of these processes. This should however not be interpreted as indifference toward the causal inference approaches developed and employed in other areas such as social sciences and medical research. On the contrary, we attempted to deal with reality on reality's terms and propose a possibility of making use of different sources of data in manufacturing enhanced with expert knowledge in the pursuit of causality. We nonetheless hope that this Quality Quandaries can help raise awareness on the efforts and resulting methods in other fields of applications, which can then inspire the researchers and practitioners of data science in manufacturing to liberally but not recklessly seek and adopt these methods.

About the authors

Marta Rotari received her PhD in Statistics and Data Analysis from the Technical University of Denmark. Her research interests include statistical process control, statistical modeling, and supervised and unsupervised machine learning methods.

Murat Kulahci is a professor at the Technical University of Denmark and Luleå University of Technology in Sweden. His research currently focuses primarily on large data analytics for descriptive, inferential and predictive purposes. Many of his research applications involve high dimensional, high frequency data demanding analysis methods in chemometrics and machine learning. He has been collaborating with various industries in many industrial statistics projects and digital manufacturing.

Acknowledgments

The authors would like to extend their gratitude to Dr. Anil Menon and Dr. Bo Friis Nielsen for their corrections and comments on the initial draft of this Quality Quandaries.

References

Bollen, K. A. 1989. *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Box, G. E. P. 1971. Science and statistics. *Journal of the American Statistical Association* 71 (356):791–9. doi: 10.2307/2286841.

Box, G. E. P., and K. B. Wilson. 1951. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 13(1): 1–38. doi: 10.1111/j.2517-6161.1951.tb00067.x.

Cacciarelli, D., and M. Kulahci. 2024. Active learning for data streams: A survey. *Machine Learning* 113(1):185–239. doi: 10.1007/s10994-023-06454-2.

Cacciarelli, D., M. Kulahci, and J. S. Tyssedal. 2024. Robust online active learning. *Quality and Reliability Engineering International* 40(1):277–96. doi: 10.1002/qre.3392.

Capaci, F., E. Vanhatalo, A. Palazoglu, B. Bergquist, and M. Kulahci. 2020. On monitoring industrial processes under feedback control. *Quality and Reliability Engineering International* 36(8):2720–37. doi: 10.1002/qre.2676.

Cornfield, J., W. Haenszel, E. H. Cuyler, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. 1959. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 22: 173–203.

Dawid, A. P. 2010. Beware of the DAG!. *NIPS 2008 Workshop on Causality, JMLR: Workshop and Conference Proceedings* 6:59–86.

De Mast, J., S. H. Steiner, W. P. M. Nuijten, and D. Kapitan. 2023. Analytical problem solving based on causal, correlational and deductive models. *The American Statistician* 77(1):51–61. doi: 10.1080/00031305.2021.2023633.

Fisher, R. A. 1925. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Fisher, R. A. 1935. *The design of experiments*. Edinburgh: Oliver and Boyde.

Fisher, R. A. 1956. *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyde.

Frumosu, F. D., and M. Kulahci. 2018. Big data analytics using semi-supervised learning methods. *Quality and*

Reliability Engineering International 34(7):1413–23. doi: 10.1002/qre.2338.

Hernán, M. A., J. Hsu, and B. Healy. 2019. A second chance to get causal inference right: A classification of data science tasks. *Chance* 32(1):42–49. doi: 10.1080/09332480.2019.1579578.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–60. doi: 10.2307/2289064.

Kulahci, M., and G. E. P. Box. 2003. Catalysis of discovery and development in engineering and industry. *Quality Engineering* 15(3):513–7. doi: 10.1081/QEN-120018049.

Kulahci, M., F. D. Frumosu, A. R. Khan, G. Ø. Rønsch, and M. P. Spooner. 2020. Experiences with big data: Accounts from a data scientist's perspective. *Quality Engineering* 32 (4):529–42. doi: 10.1080/08982112.2019.1686641.

Kulahci, M., and A. Menon. 2023. *How to develop drug products*. Morrisville: Datanumerix LLC,

Liu, Z., M. J. Bruwer, J. F. MacGregor, S. S. S. Rathore, D. E. Reed, and M. J. Champagne. 2011. Modeling and optimization of a tablet manufacturing line. *Journal of Pharmaceutical Innovation* 6(3):170–80. doi: 10.1007/s12247-011-9112-8.

MacGregor, J. F., M. J. Bruwer, I. Miletic, M. Cardin, and Z. Liu. 2015. Latent variable models and big data in the process industries. *IFAC-PapersOnLine* 48(8):520–4. doi: 10.1016/j.ifacol.2015.09.020.

Michaels, S. E. 1964. The usefulness of experimental designs. *Applied Statistics* 13(3):221–8. doi: 10.2307/2985488.

Montgomery, D. C. 2020. *Design and analysis of experiments*. 10th ed. New York: John Wiley and Sons.

Palací-López, D., P. Facco, M. Barolo, and A. Ferrer. 2019. New tools for the design and manufacturing of new products based on latent variable model inversion. *Chemometrics and Intelligent Laboratory Systems* 194: 103848. doi: 10.1016/j.chemolab.2019.103848.

Panahi, M., S. H. Steiner, and J. De Mast. 2021. Identifying dominant causes using leveraged study designs. *Quality Engineering* 33 (4):581–93. doi: 10.1080/08982112.2021.1938118.

Panahi, M., S. H. Steiner, and J. De Mast. 2023. Verifying a dominant cause of output variation. *Quality Engineering* 1–12. doi: 10.1080/08982112.2023.2253303.

Pearl, J. 2009a. Causal inference in statistics: An overview. *Statistics Surveys* 3:96–146. doi: 10.1214/09-SS057.

Pearl, J. 2009b. *Causality*. Cambridge, UK: Cambridge University Press.

Pearson, E. S., and J. Wishart. 1942. *"STUDENT's" collected works*. London: Biometrika Office, University College.

Rønsch, G., M. Kulahci, and M. Dybdahl. 2022. Real-time adjustment of injection molding process settings by utilizing design of experiment, time-series profiles and PLS-DA. *Quality Engineering* 34 (2):215–29. doi: 10.1080/08982112.2022.2033775.

Rotari, M., and M. Kulahci. 2024. Variable selection wrapper in presence of correlated input variables for random forest models. Journal of *Quality and Reliability Engineering International* 40(1):297–312. doi: 10.1002/qre.3398.

Salsburg, D. 2001. *The lady tasting tea*. New York: Henry Holt and Company.

Stichel, T., B. Geißler, J. Jander, T. Laumer, T. Frick, and S. Roth. 2018. Electrophotographic multi-material powder deposition for additive manufacturing. *Journal of Laser Applications* 30 (3):249–53. doi: 10.2351/1.5040619.

Taguchi, G. 1992. *Taguchi on robust technology development: Bringing quality engineering upstream*. New York, NY: ASME Press.

Weston, R., and P. A. Gore Jr. 2006. A brief guide to structural equation modeling. *The Counseling Psychologist* 34 (5):719–51. doi: 10.1177/0011000006286345.

Wu, A., K. Kuang, R. Xiong, and F. Wu. 2022. "Instrumental Variables in Causal Inference and Machine Learning: A Survey." http://arxiv.org/abs/2212.05778.

Yeh, H.-P., K. Meinert, M. Bayat, and J. Hattel. 2024. Part-scale thermo-mechanical modelling for the transfusion module in the selective thermoplastic electrophotographic process. *The International Journal of Advanced Manufacturing Technology* 131 (11):5419–35. doi: 10.1007/s00170-023-12300-5.