GEOMET
@QUEEN'S
UNIVERSITY

Understanding process performance with causal inference for continuous variables [1]

Sebastian Avalos (sebastian.avalos@queensu.ca)
Miguel Cuba (miguel.cuba@arcelormittal.com)
Melanie Bolduc (melanie.bolduc@arcelormittal.com)
Julian M Ortiz (julian.ortiz@queensu.ca)

Abstract

In mining operations, relations of cause and effect are not always clear and the presence of spurious correlations further confuses the analysis of causal inferences. In this article, we briefly review the principles of counterfactual analysis and the study of causal inference for continuous variables using the Kolmogorov-Smirnov test. The methodology is applied to study the impact of grade values and source proportions on a SAG mill energy consumption. The approach can be applied in any context, to understand the performance of a specific process as a function of the input variables, and draw causal relationships that can be validated with domain expertise.

1. Introduction

In mining operations, we often develop predictive models between a set of predictors to a single response or a set of responses. The quality of such models relies on the quality of the measured values of predictors and responses, and on the underlying relationships between them. It is well known that correlation does not necessarily mean causation, and the presence of spurious correlations further confuses the analysis of causal inferences. Thus, we can not rely on the predictive models to answer *what-if* questions related to causal effects, since the former exploit linear and non-linear correlations while the latter looks for counterfactual conditions.

The fields of counterfactual analysis and causal inference provide frameworks and principles to formulate causal problem from a statistical perspective. They have been applied in several other disciplines, such as pharmaceutical industry, sociological studies and epidemiology. It seems natural to transfer the previous frameworks into the mining context, extending the current tools of predictive modeling practitioners to deal with what-if questions.

In section 2 we present the fundamentals of counterfactual analysis. We focus the analysis to continuous variables. In section 3 we define a simple methodology to determine the presence or not of a cause-effect relation between the possible states of a response attribute conditioned to the state of a predictor attribute. The method is applied in a mining context in section 4 studying the impact on a SAG mill energy consumption of two grades and the proportion of three ore sources. A final discussion is presented in section 5.

[1] Cite as: Avalos S, Ortiz JM (2021) Understanding process performance with causal inference for continuous variables, Predictive Geometallurgy and Geostatistics Lab, Queen's University, Annual Report 2021, paper 2021-03, 22-32.

2. Background

Counterfactual theories of causation study conditional statements of the form *if A were true, then B would be true*, and any variations, such as *If A had not occurred, C would not have occurred* (Menzies and Beebee, 2020). The theoretical analysis of causation started with David Lewis's theory (Lewis, 1974). Since then, the field has expanded, refined and matured. Causal inference, or causal modeling, is the state-of-the-art branch of counterfactual theory, providing mathematical models and causal representations (Pearl et al., 2016).

The causal inference principle is that data alone is not enough to explain causality, and the story behind the data elements is required. The story is formally conveyed into a graph representation, with nodes representing the data elements, and edges the connections between them. The edges and their directions form a direct graph that represents the underlying story of the data. A detailed analysis on the graph structure and node's connections can be found at Pearl et al. (2016).

The data and its associated direct graph provide the framework to deal with what-if questions by means of *interventions* and *conditioning*. We intervene a variable when we fix its value. The edges flowing in the nodes are removed. This modifies the original graph, and often changes the value of subsequent nodes (Altinpinar and Ortiz, 2020). When conditioning on a variable, the graph does not change but we rather focus on a subset of the original data, satisfying the node condition. To draw reliable causal conclusions, the graph must be a valid representation of the studied phenomenon.

The effects of interventions are analyzed by means of probability theory and statistical metrics. Formulas such as Controlled Direct Effect (CDE), Average Causal Effect (ACE), and test of goodness-of-fit are some of the most used tools (Maldonado and Greenland, 2002). In the next section, we describe how to build the direct graph and a particular test of goodness-of-fit, in the context of mining and continuous variables.

3. Methodology

3.1. Causal model representation

The mining operational context and the expert knowledge on the expected relationship between variables must be reflected in the direct graph. First, all variables in the data must be created as nodes. If one or more unmeasured variables must be considered, additional nodes to the graph should be added. Then, connections must be drawn between nodes to represent the reality based on expert knowledge, process flowcharts, and/or feasible cause-effect relations. The graph representation and the applied interventions are the critical elements during causal analysis.

Assuming an adequate graph representation and that all variables have continuous values, we proceed to *conditioning* on the nodes and not *intervening* on the graph by removing them. The conditioning could be on a single variable, $\mathbf{x}$, or a set of them, $\mathbf{x}_1, \mathbf{x}_2, ...$, by means of inequalities over a specific value. It is reflected into conditional probabilities in the graph, such as $\mathcal{P}(\mathbf{x}_2 \geq 0.7)$, that translates into selecting a subset of the original data where $\mathbf{x}_2$ meets the condition. We study the impact of a variable $\mathbf{x}_1$ on a variable $\mathbf{x}_3$ when $\mathbf{x}_1$ meets a certain criteria conditioning to an additional variable $\mathbf{x}_2$. In other words, we study the difference between $\mathcal{P}(\mathbf{x}_3|\mathbf{x}_1)$ and $\mathcal{P}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)$. The principle can be extended to more than two variables at a time.

3.2. Two-sample Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test for goodness-of-fit (Massey Jr, 1951) is applied to the conditional cumulative distributions functions on $\mathcal{P}(\mathbf{x}_3|\mathbf{x}_1)$ and $\mathcal{P}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)$, represented as empirical distributions F_1, n and F_2, m, respectively, with n and m being the amount of samples in $\mathcal{P}(\mathbf{x}_3|\mathbf{x}_1)$ and $\mathcal{P}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)$, respectively. The Kolmogorov–Smirnov statistic is defined as:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{1}$$

The null hypothesis states that samples on F_1, n and F_2, m are drawn from the same global distribution. The null hypothesis is rejected at a level α when:

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n \cdot m}} \tag{2}$$

where $c\,(\alpha)$ is computed as:

$$c\,(\alpha) = \sqrt{-0.5 \cdot \ln\left(\frac{\alpha}{2}\right)} \cdot \tag{3}$$

Back into the causal analysis, whenever the null hypothesis is rejected between $\mathcal{P}(\mathbf{x}_3|\mathbf{x}_1)$ and $\mathcal{P}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)$, we say that $\mathbf{x}_2$ has a cause-effect on x_3, conditioned to x_1.

4. Case of study

4.1. Context

The previous methodology is applied in a simple open pit mining scenario. The run-of-mine ore is fed to the processing plant from three different sources: A (sA), B (sB), and C (sC). The ore is sent to a primary crusher, resulting in a blended material. This crushed blend is moved into a SAG mill, through a conveyor belt, where the ore is further blended and reduced in particle size. A contextual scheme is shown in Figure 1.

The SAG mill consumes a high amount of energy. As a driver for better short-term ore scheduling, the decision makers are interested in measuring the impact that grade 1 (g1) and grade 2 (g2), along with the proportion of sources in the blended ore, have on the SAG mill energy consumption (EC).

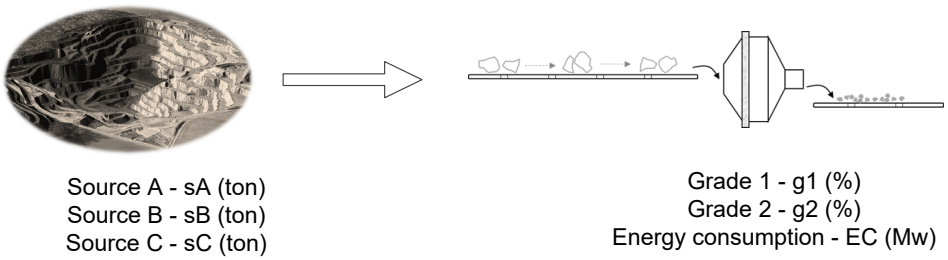


Figure 1: Contextual scheme and the location of measured data.

4.2. Database

A total of 250 daily measurements of the source tonnages (ton), grades (%), and energy consumption (Mw) have been collected. Table 1 displays the database main statistics. A visualization of the entire database time series is displayed in Figure 2. As we are interested in understanding the influence of source proportions and grades on the SAG mill energy consumption, scatter plots including their kernel density maps are shown in Figure 3. Note that we are not displaying the EC against the source proportion. No clear correlation is shown on the tonnage per source, but a slight positive correlation between EC and sC.

Figure 2: Time series visualization of sources, grades, and energy consumption.

Table 1: Database main statistics.

	sA (ton)	sB (ton)	sC (ton)	g1 (%)	g2 (%)	Energy consumption (Mw)
Min	2,479	0	0	0.26	0.24	6,000
Mean	25,968	33,744	33,048	6.98	10.59	34,161
Max	53,857	93,219	84,228	14.31	20.57	47,423
St Dev	10,792	16,429	20,160	2.55	3.56	6,940
Count	250	250	250	250	250	250

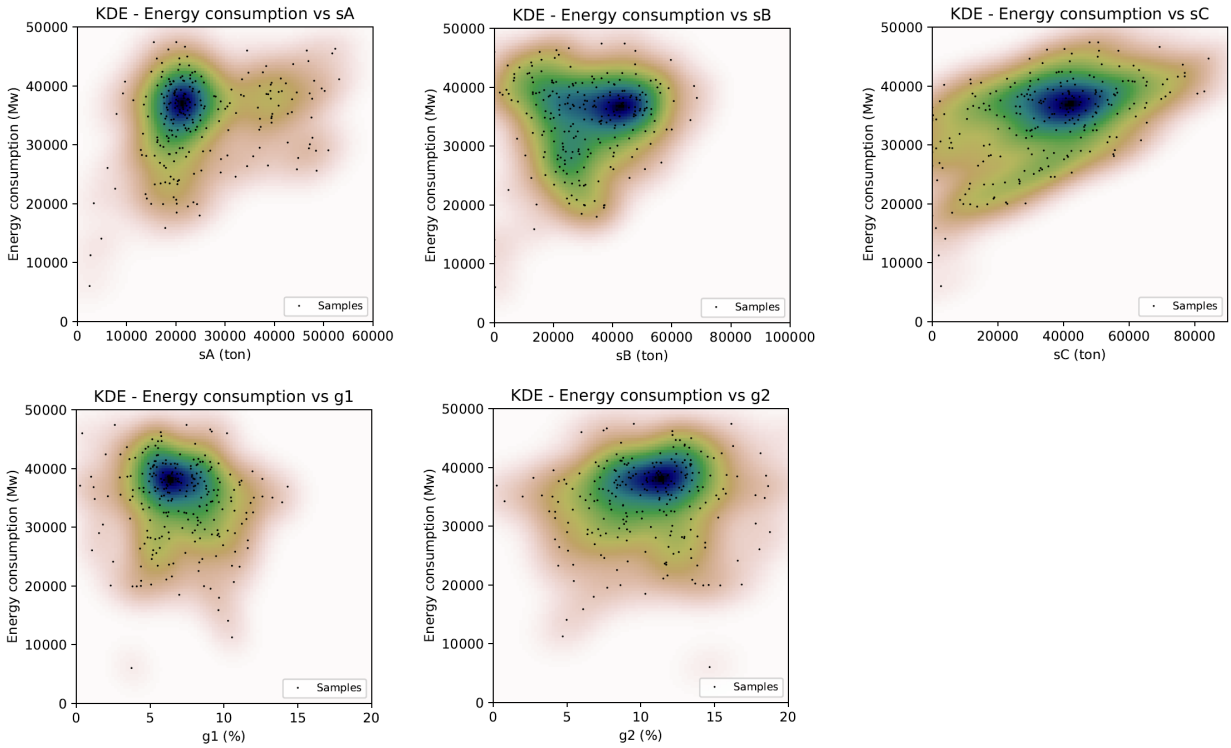


Figure 3: Scatter plots and kernel density maps of energy consumption against source tonnages and grades percentages.

4.3. Analysis

The contextual setting of the problem is translated into a causal model representation (Figure 4). Here, we interrogate the causal model about the influence on both grades and source proportion on the energy consumption.

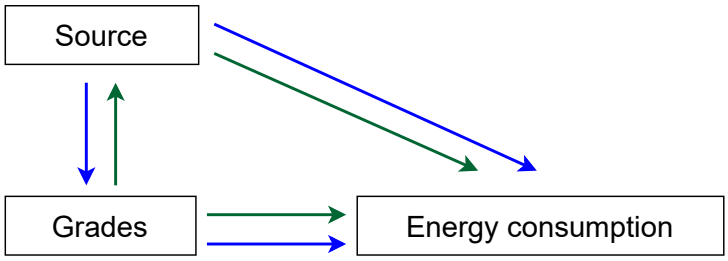


Figure 4: Causal model representation. Blue: impact of the grade on the energy consumption. Green: impact of the source proportion on the energy consumption.

Let pX be the proportion of a source in the blended material. For instance $p30$ of sA means that source A represents 30 % of the blended ore. Let pV be the V-percentile of a grade. For instance $p20$ of Grade 1 refers to the 20-percentile value of the Grade 1 distribution. Using the scheme of Figure 4 we have two possible pathways

Blue pathway Impact of grade on the energy consumption.

First, we select the subset of measurements in which a source is above a certain percentage. For instance, in vector representation we write - Source A above an X %, as $EC\,[sA > pX]$. We compute the conditional cumulative distribution function (ccdf) of the resulting subset, named ccdf_1.

Secondly, from the previous subset of measurements, we select an smaller subset of data with grade values above a certain percentile. For instance, in vector representation we write - Source A above an X % AND all Grade 1 above a percentile pV, as $EC\,[sA > pX, g1 > pV]$. We compute the ccdf of the resulting subset, named ccdf_2.

Green pathway Impact of the source proportion on the energy consumption.

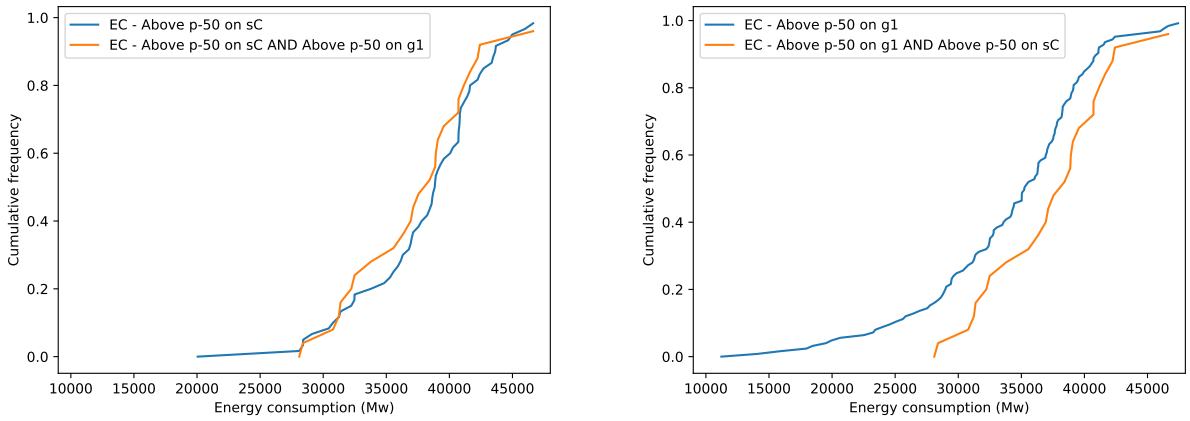
We start by selecting the measurements with grade values above a certain percentile. For instance, in vector representation we write - All Grade 1 above a percentile pV, as $EC\,[g1 > pV]$. We compute the ccdf of the resulting subset, named ccdf_1.

Secondly, from the previous subset of measurements, we select the smaller subset of data in which a source is above a certain percentage. For instance, in vector representation we write - All Grade 1 above a percentile pV AND Source A above an X %, as $EC\,[g1 > pV, sA > pX]$. We compute the ccdf of the resulting subset, named ccdf_2.

Then, regardless of the pathway, we compute the pValue of the two-sample Kolmogorov-Smirnov test for goodness-of-fit between ccdf_1 and ccdf_2. If the obtained pValue is above the critical value $\alpha : 0.05$, the null hypothesis is rejected, and therefore, the element (grade or source proportion) has an impact on the energy consumption. We won't compute the pValue if either the ccdf_1 or ccdf_2 have less than 10 samples.

4.4. Results

We begin by presenting the resulting ccdf_1 and ccdf_2 when analyzing both pathways (see Figure 5). The distributions are conditioned on sC above $p50$ (50% in blended ore) and g1 above $p50$ (grade values above the 50-percentile). When applying the K-S test, the null hypothesis is accepted on Figure 5a and rejected on Figure 5b. This means that the test indicates that in the former case, the values are drawn from the same distribution, while in the latter case, they come from different distributions. In other words, the grade does not have a significant effect on the energy consumption, while the source does.



(a) Impact of grade on the EC. *(b) Impact of the source proportion on the EC.*

Figure 5: Conditional cumulative distribution functions. Conditioning on sC above p50 and g1 above p50.

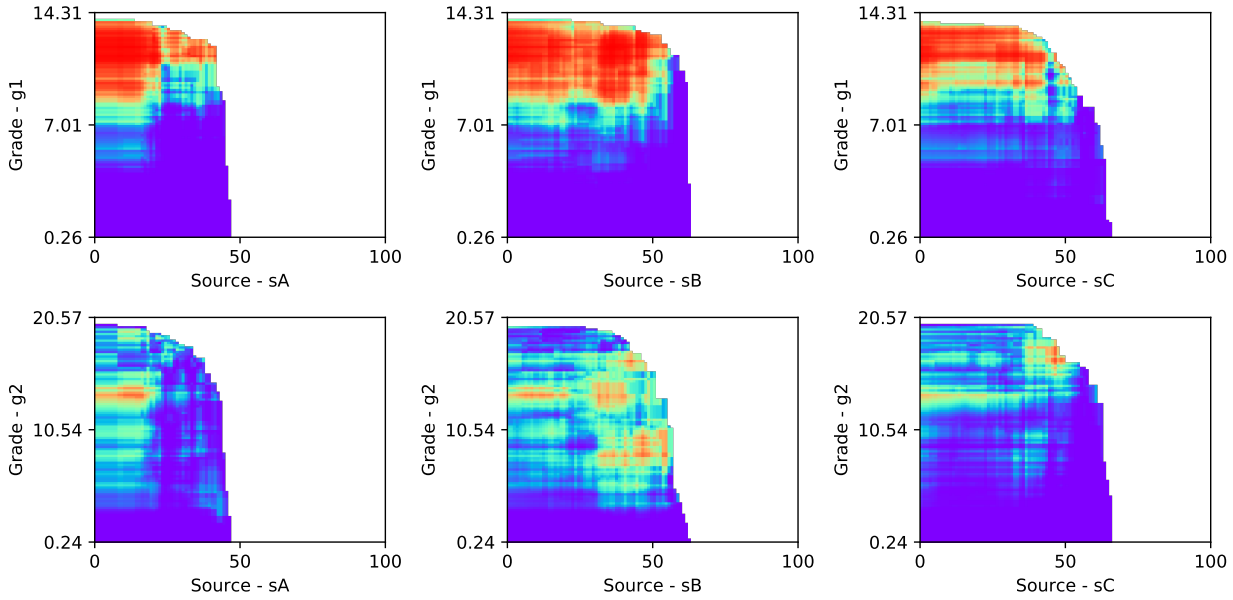


Figure 6: Impact of grade on the energy consumption. Grade 1 (top) and Grade 2 (bottom). Displaying the pValue of the two-sample Kolmogorov-Smirnov test.

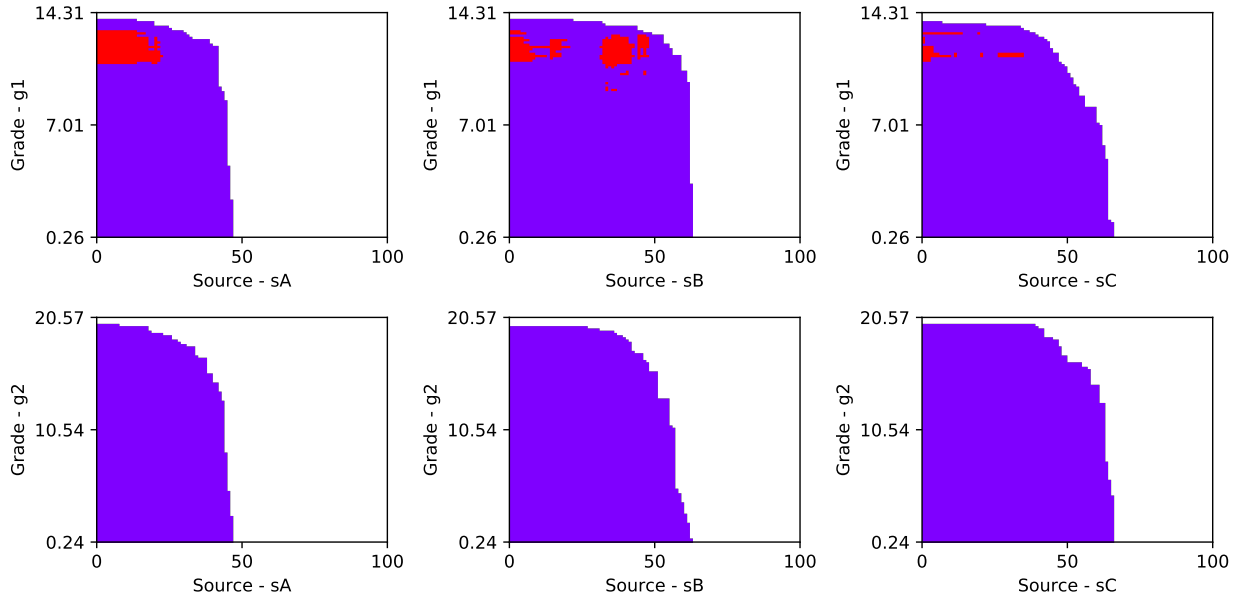


Figure 7: Impact of grade on the energy consumption. Grade 1 (top) and Grade 2 (bottom). Thresholding the pValue of the two-sample Kolmogorov-Smirnov test. Purple zone: accepting the null hypothesis. Red zone: rejecting the null hypothesis.

We extend the analysis to each grade and source. Grades are conditioned by percentiles while the source proportions goes from $[0, 1, ..., 99, 100]\%$. Figure 6 shows the pValue of the two-sample K-S test of the grade impact on the energy consumption for g1 (top) and g2 (bottom). The white areas correspond to insufficient amount of samples to compute the ccdf. By thresholding the maps with $\alpha : 0.05$ wherever the pValue is greater or equal to α, the Figure 7 is obtained.

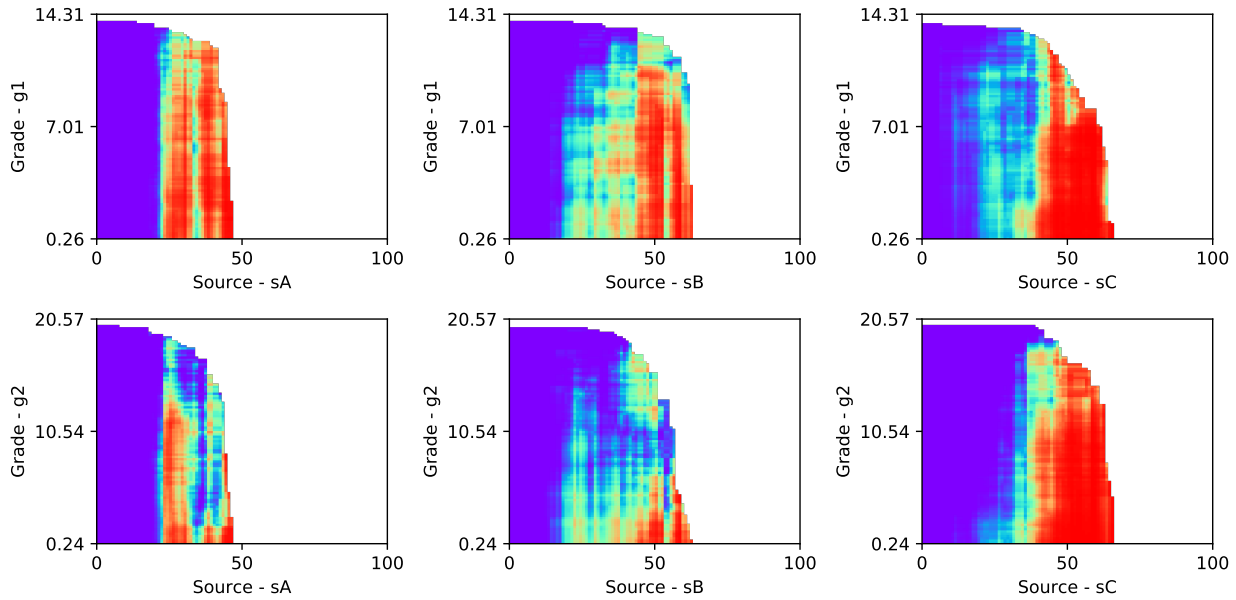


Figure 8: Impact of the source proportion on the energy consumption. Grade 1 (top) and Grade 2 (bottom). Displaying the pValue of the two-sample Kolmogorov-Smirnov test.

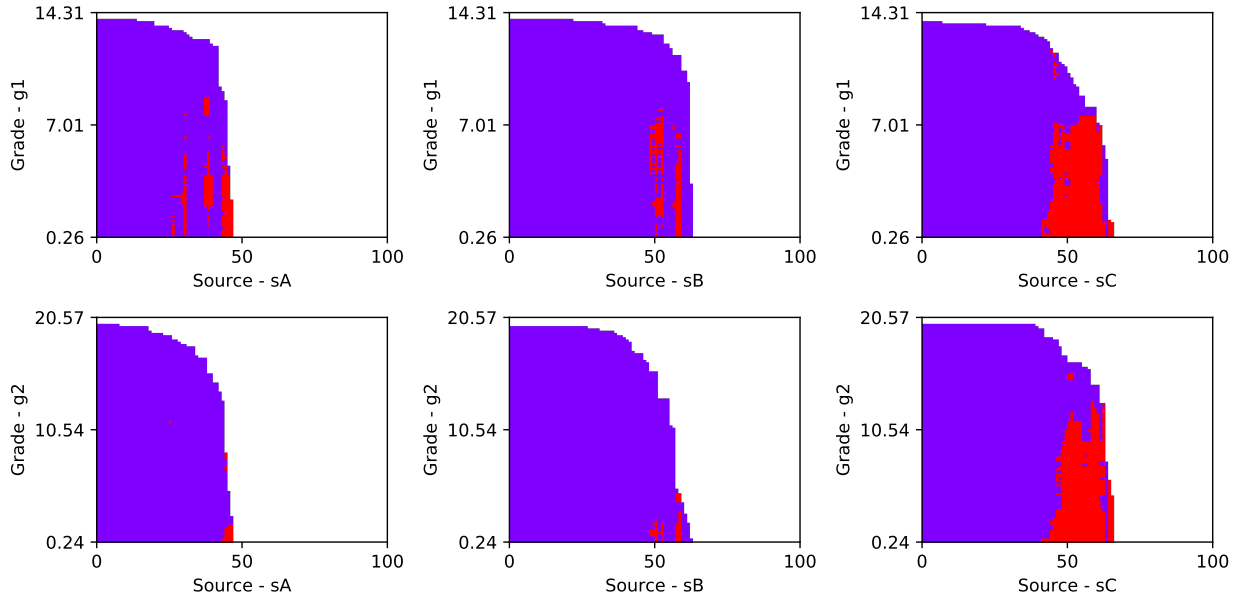


Figure 9: Impact of the source proportion on the energy consumption. Grade 1 (top) and Grade 2 (bottom). Thresholding the pValue of the two-sample Kolmogorov-Smirnov test. Purple zone: accepting the null hypothesis. Red zone: rejecting the null hypothesis.

The analysis is repeated on the impact of source proportion on the EC. Figure 8 shows the pValue of the corresponding two-sample K-S test for sA (left), sB (middle), and sC (right). The white areas correspond to insufficient amount of samples to compute the ccdf. By thresholding the maps with $\alpha : 0.05$ wherever the pValue is greater or equal to α, the Figure 9 is obtained.

4.5. Final takeaways

The maps of Figure 7 and Figure 9 summarize the *causalities* of grades and sources proportions, respectively. From them, the following key results are derived:

1. The Grade 1 (g1) **influences** the SAG mill energy consumption *if and only if* the value of g1 is above the 80-percentile (10.9 %) AND the proportion of sA or sB are below 24 % and 22 %, respectively, regardless of the proportion of sC.
2. The Grade 2 (g2) **has no influence** on the SAG mill energy consumption, regardless of the proportion of sources in the fed ore.
3. The proportion of Source A (sA) **has no influence** on the SAG mill energy consumption, regardless of the grade values of g1 and g2.
4. The proportion of Source B (sB) **has no influence** on the SAG mill energy consumption, regardless of the grade values of g1 and g2.
5. The proportion of Source C (sC) **influences** the SAG mill energy consumption *if and only if* the proportion is above 46 %, regardless of the grade values of g1 and g2.

The previous key results can be transferred into operational decisions either to avoid falling into areas where g1 or sC influence on the EC, or to expect EC variations when the conditions of points 1 and 5 are met.

5. Conclusions

Causal inference analysis has been widely applied in other disciplines, such as pharmaceutical industry and sociological studies. Transferring the theoretical background and acquired knowledge into mining operations is a fruitful area for applied research.

Data alone is not enough to explain causality. The story behind the variables is fundamental for the causal analysis. It translates into a graph representation. The graph representation and the interventions and/or conditioning must meet the real phenomenon and the what-if question being asked.

The graph representation can be as simple as the case study shown in the article or much more complex, when several processes and/or variables are considered. In addition, when testing the null hypothesis for goodness-of-fit, we have used the Kolmogorov–Smirnov test but other methods can be applied as well, such as Chi-squared test.

In the case study, and extrapolated to any causal inference analysis, causal models may indicate the impact on the energy consumption of a certain setting between grade value and source proportion but they do not describe if the change is positive or negative, meaning an increase or decrease in the consumed energy.

6. Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2017-04200 and RGPAS-2017-507956. We also thank the support of ArcelorMittal Mining Canada G.P.

7. Bibliography

Altinpinar, M., Ortiz, J.M., 2020. Review of causal inference and modeling. Predictive Geometallurgy and Geostatistics Lab, Queen's University, Annual Report 2020, 130–146.

Lewis, D., 1974. Causation. The Journal of Philosophy 70, 556–567.

Maldonado, G., Greenland, S., 2002. Estimating causal effects. International Journal of Epidemiology 31, 422–429.

Massey Jr, F.J., 1951. The kolmogorov-smirnov test for goodness of fit. Journal of the American Statistical Association 46, 68–78.

Menzies, P., Beebee, H., 2020. Counterfactual Theories of Causation, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Winter 2020 ed.. Metaphysics Research Lab, Stanford University.

Pearl, J., Glymour, M., Jewell, N.P., 2016. Causal inference in statistics: A primer. John Wiley & Sons.