Isolating the Impact of Rock Properties and Operational Settings on Minerals Processing Performance: A Data-driven Approach

Suriadi Suriadi[a,*], Sander J.J. Leemans[a], Cristián Carrasco[b,1], Luke Keeney[b], Patrick Walters[b], Kevin Burrage[a], Arthur H. M. ter Hofstede[a], Moe T. Wynn[a]

[a] *School of Mathematical Sciences and School of Information Systems*
Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia
[b] *CRC ORE, Pullenvale QLD 4069, Australia*

Abstract

Mining operations record a large amount of data from multiple sources (such as block model and online processing data) which is neither effectively nor systematically used to understand and improve operational performance. This paper proposes a generic semi-automatable data analytics method, the Integrated Analysis Method (IAM), that addresses the disconnection between disparate datasets. IAM enables evidence-based understanding of rock and machine parameters, laying the foundation for a potentially more sophisticated way to model and predict mining processes to deliver financial value. IAM systematically combines and analyses both rock characteristics and operational data to isolate the impact of the variability in rock characteristics and operational settings on key performances. Insights extracted from IAM allow one to narrow down key operating conditions, specific to a particular plant, that are correlated to, for example, significant differences in daily throughput while processing batches of ore with similar metallurgical characteristics. Such insights can be used for multiple purposes, for instance, to learn optimal processing recipes for a given set of rock properties. We applied IAM to a combined data set recorded at a Chilean ore deposit and evaluated our findings with domain experts.

Keywords: data mining, process mining, time series analysis, decision support systems, machine learning

1. Introduction

Increasing processing costs and declining ore body grades have called for the global mining industry to focus on improving productivity and energy efficiency in order to stay competitive and to meet the increasing demand for resources (Bearman, 2013; Carrasco et al., 2016b; Hesse et al., 2016; Napier-Munn, 2015; Prior et al., 2012). State-of-the-art operating mechanisms that aim at improving the efficiency of mineral processing, such as flexible circuits (Foggiatto et al., 2014; Powell et al., 2014) and Grade Engineering (Carrasco et al., 2016a; Walters, 2016), have been proposed and incorporated into some mine sites. Flexible circuits manipulate the design of comminution circuits such that they balance the comminution work across different comminution units as the distributions of the particle size of feeds change, allowing plants to handle rock variability. Grade Engineering maximises feed grade by removing low-grade materials as early as possible in the process such that resources (such as energy) are targeted towards processing valuable material as much as possible.

While these state-of-the-art techniques can, and should, be incorporated into practice, their outcomes can be further improved by making use of currently under-exploited knowledge buried within many sources of historical data collected in today's minerals processing plants. In particular, the *uniqueness of the setups of each plant and the geo-*

*Corresponding author

Email addresses: `s.suriadi@qut.edu.au` (Suriadi Suriadi), `s.leemans@qut.edu.au` (Sander J.J. Leemans), `cristian.carrasco@bhpbilliton.com` (Cristián Carrasco), `l.keeney@crcore.org.au` (Luke Keeney), `p.walters@crcore.org.au` (Patrick Walters), `kevin.burrage@qut.edu.au` (Kevin Burrage), `a.terhofstede@qut.edu.au` (Arthur H. M. ter Hofstede), `m.wynn@qut.edu.au` (Moe T. Wynn)

[1] Now present at BHP

metallurgical characteristics of materials being processed need to be taken into account. Gaining such information may be challenging due to uncertainties in the geo-metallurgical characteristics of materials being processed, variations in operating conditions, and a lack of knowledge on *how* different types of materials react to different operational settings.

Fortunately, we can attempt to learn this information by analysing historical data, from which we can extract patterns of various operating strategies and quantify their impacts on rock processing performance. Through the application of data analysis techniques, one can learn better operating strategies to maximise outcomes in light of various factors, such as energy, throughput[2], and materials characteristics such as the hardness of rocks. This extracted information can thus be applied in practice. For example, by understanding how different operational settings affect the throughput of a plant for a given rock metallurgical characteristics, site engineers in a comminution plant can adjust the processing parameters and operational settings to the values which, in the past, have shown to deliver higher throughput.

The main contribution of this paper is a generic semi-automatable data analytics method, called the Integrated Analysis Method (IAM) that can be used to isolate the impact of rock characteristics[3] and operational settings. In particular, IAM provides a systematic way to analyse traditionally-disparate data sets *combined*: short-term mine plan data (that includes information about the characteristics of rocks to be processed on a particular day) and processing data (collected from various sensor readings in a plant) - see Figure 1. By analysing these disparate data sets in an integrated manner, IAM is able to gain insights into the combined impact of materials variability and processing uncertainty on key performance indicators (KPIs) of a processing plant, such as tonnes per day (TPD).

A key benefit of IAM is its ability to learn better operating strategies, *adjusted* for a given characteristics of rocks to be processed. Given periods when rocks with similar characteristics were processed, IAM identifies those periods that produced optimal results and then learns the operational settings applied during those periods. This knowledge is valuable in enabling adjustments to machine settings that are *customised* to the geo-metallurgical characteristics of materials to be processed to achieve optimal KPIs.

Furthermore, IAM supports a better integration between scheduling and control (Harjunkoski et al., 2009a). Insights extracted from IAM can be used to inform short-term operating strategies of a plant, based on the expected characteristics of rocks that are to be processed within a certain time horizon (such as weekly or monthly) as dictated by short-term plan data.

We acknowledge that the application of data analytics for operational supports in the mining industry has been demonstrated before. For example, as early as the 1970s, there have been attempts to apply machine learning techniques to gain insights into the impact of operational settings on plant performances (Brittan and van Vuuren, 1973; Ge et al., 2017; Mackay and Lloyd, 1975). More recently, Marais and Aldrich (2011) and Aldrich et al. (2010) looked into how online images taken from a flotation process can be used to predict the recovery and grade of the extracted metal. Zhang et al. (2002) looked at how to apply a genetic algorithm to recorded sensor data in the design of a coal mill. Nevertheless, how the geo-metallurgical characteristics of materials being processed influences the choice of operating strategies was not addressed in these works. IAM aims to resolve this.

This paper is organised as follows. Section 2 details IAM. Section 3 describes the data set obtained from a processing plant related to a Cu porphyry deposit in Chile. Section 4 details the application of IAM on this data set to demonstrate its applicability. Section 5 raises issues related to the use of IAM, and Section 6 discusses related work. Conclusions are provided in Section 7.

2. Approach

The process of minerals extraction involves several stages, including blasting, comminution, and flotation (Wills, 2006). The notion of an 'optimal result' varies with the stage of the process. For example, during the comminution stage, an 'optimal' result typically includes high tonnes of rock processed per hour, while during the flotation stage, an 'optimal' result may refer to high recovery rate.

KPIs are influenced by several factors, such as geo-metallurgical properties and operational settings. The operational settings and rock properties both influence KPIs, but in order to study them,

[2] Throughput in tonnes per day (TPD)

[3] We refer to metallurgical properties of rocks with the term 'rock characteristics'.

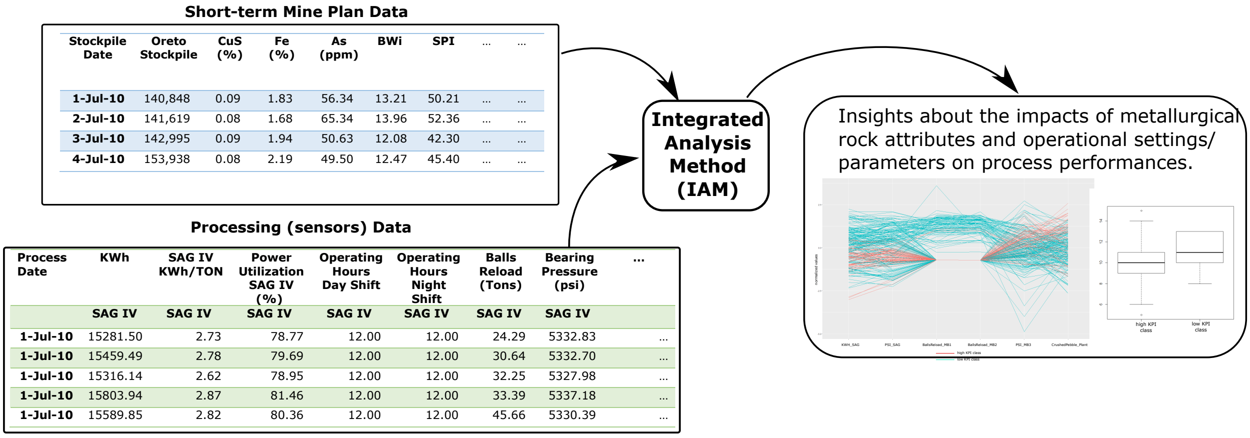


Figure 1: Data sources for the Integrated Analysis Method (IAM) proposed in this article. IAM facilitates integrated analysis of *both* rock properties and operating parameters to extract insights into the impacts of metallurgical rock attributes and operational settings/process parameters on process performances.

their influences on KPIs need to be analysed separately. To separate the influence of these factors, we introduce the Integrated Analysis Method (IAM), a methodology that aims to analyse the influence of these factors on KPIs. For instance, in Section 4, we illustrate IAM with an analysis of the influence of rock properties and operating strategies on the throughput per day of a comminution plant.

In the following sections, we will use the terms X and Y for variables related to rock characteristics and operational settings respectively.

2.1. A Pluggable Framework

IAM is designed as a pluggable framework whereby a wide range of analysis techniques can be included and applied in the analysis chain as long as the data sets and the chosen analysis techniques satisfy certain constraints (gradually elaborated throughout this section). As a consequence, IAM is agnostic to underlying domain-specific variations, such as the types of plants or mills. Furthermore, IAM is flexible to the KPI used: any KPI can be used as long as its corresponding values are available in the data set being analysed.

Figure 2 shows a data model for the data sets and configuration parameters that IAM needs, depicted using the Unified Modeling Language (UML) class diagram (Group, 2015). The basic building blocks for the data model are the two data types: `RockMetallurgy` which represents any metallurgical attributes of rocks, such as the Bond Ball Mill Work Index (BWi) and SAG Power Index (SPI), and `ProcessTag` which represents data items captured from a processing plant, such as bearing pressure, flow rates, and throughput rates. A name-value pair of rock metallurgical attribute and its corresponding value (for example, 'BWI = 10.89') is captured as a `RockCharacteristic` type. Similarly, `ProcessData` type captures a pair of process tag name and its value (for instance 'TPD = 6798 tonnes/day').

`ShortTerm_DataPoint` type is an abstraction of a single data point entry of short-term plan data (similar to Figure 1) that includes a timestamp representing the date that the batch of ore was added to the stockpile, another timestamp representing the *estimated* date/time the stockpile was processed[4]. A single entry of a short-term plan also includes a set of rock characteristics (captured in Figure 2 by the arrow from the `ShortTerm_DataPoint` type to the `RockCharacteristic` type). Furthermore, as detailed later in this section, applying IAM will result in the short-term plan data being augmented with the `rockCluster` attribute. In Figure 2, the forward slash '/' notation in front of the `rockCluster` attribute represents an attribute that is derived).

Similarly, `ProcessInformation_DataPoint` is an abstraction of a single data point entry of processing (sensor) data (similar to the one shown in Figure 1). This data point in-

[4]One could correct for stratification effects here. In our case, the time resolution of our data is daily, so we deemed correction not to be necessary. Issues that may arise surrounding the availability of this timestamp are discussed later in this article.

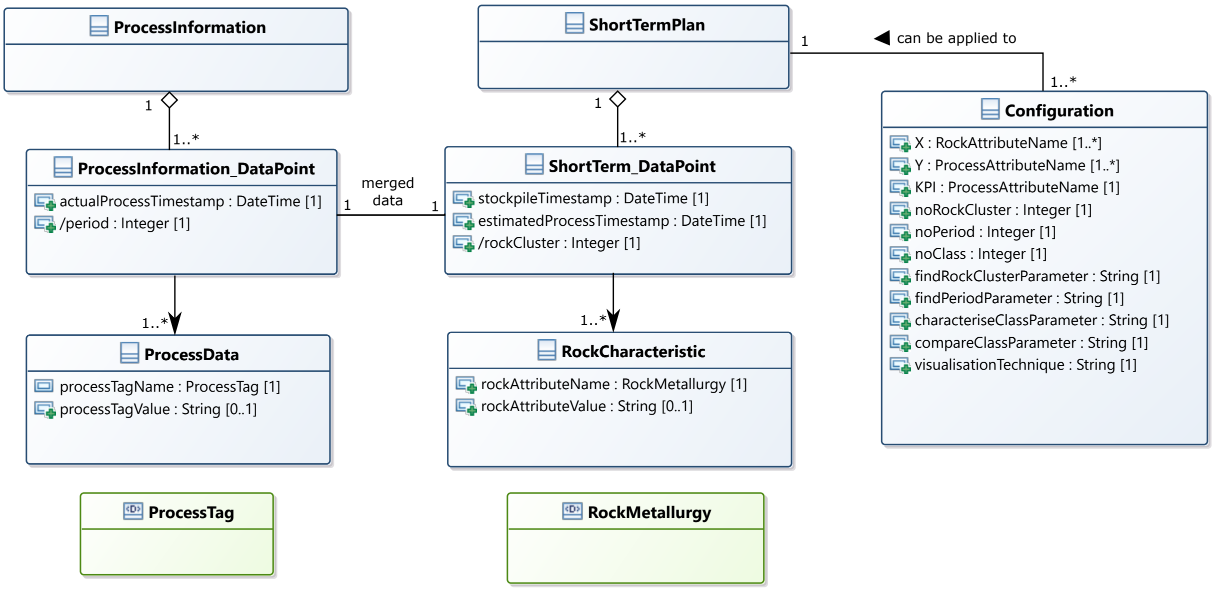


Figure 2: IAM data model expressed using a UML class diagram.

cludes a processing timestamp and a set of process data items (depicted by the arrow from the `ProcessInformation_DataPoint` type to the `ProcessData` type), representing sensors readings of operational conditions (also known as 'process tags') of the plant at that particular timestamp. Here, KPI values are assumed to be captured as part of a process information data point. During the application of IAM, this data set will also be augmented with a new attribute called *period* (detailed later in this article).

A `ShortTermPlan` data is thus a collection of one or more instantiations of the `ShortTerm_DataPoint` type, while a `ProcessInformation` data is a collection of one or more instantiations of the `ProcessInformation_DataPoint` type.

Finally, a set of configuration parameters is captured as a `Configuration` type. These parameters include the KPI, and the X and Y variables (representing rock characteristics and operational conditions/processing parameters respectively). Other parameters that IAM needs are described in the remainder of this section.

2.2. Overview of IAM

We first discuss the five high-level steps of IAM (as shown in Figure 3). We use the Business Process Modeling Notation (OMG, 2011) to express the sequence of analysis activities of IAM. The rounded-rectangles represent tasks; the plain and bold circles represent the beginning and the end of IAM respectively; the arrows represent the sequence of tasks execution; the rounded-rectangles with a '+' symbol inside them represent tasks that consist of further sub-tasks (as detailed later in Section 2.3); the diamond-shaped box with an 'X' inside it represents a split in the flow (meaning that one could take one of the two outgoing paths, but not both); and finally, the three vertical lines symbol ('|||') inside a task means that the task can be executed multiple times until the annotated completion conditions are satisfied. The 'folded' paper symbol represents a data object. Text annotations on outgoing arcs from the `Configuration` data object represent the reading of the value of the parameter name represented by the annotation. For example, the parameter `visualisationTechnique` is read by the 'Visualise results' task (Figure 3).

As a first step, from the input data `ProcessInformation` and a set of `Configuration` parameters (as modelled in Figure 2), we identify a number of *periods*. Each *period* consists of data points at a certain level of granularity (daily or hourly, for example), containing all data points of a certain time interval. The boundaries of the *periods* are chosen such that the distibution of KPI values in a *period* is statistically different from the distributions of KPI values in time-neighbouring

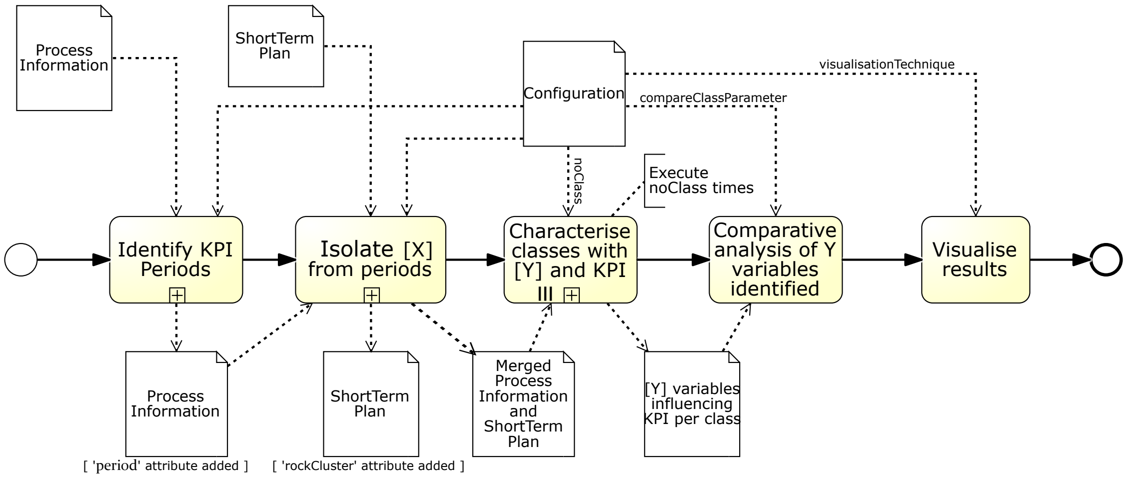


Figure 3: High-level view of IAM.

periods, according to a chosen statistical test (change point analysis). For example, in Figure 4, we can see that for days belonging to *period* '1', the throughput per day (TPD) ranges from 78,000 to 81,000 while for *period* '2', the TPD ranges from 97,000 to 99,500.

While both X and Y factors may strongly impact KPIs within each *period*, in order to isolate the influence of individual factors we need to 'neutralise', or control, the impact of one group of factors in the analysis.[5] Therefore, as a second step, we identify *rock clusters* that share similar values for the X variables (rock characteristics) from the `ShortTermPlan` data. For example, as shown in Figure 4, data points belonging to the *rock cluster* "3" have similar BWi and SPI values (around 10.00 and 40.00, respectively). Note that the identification of these *rock clusters* is independent of the *periods* identified earlier, that is, the *period* to which a data point belongs is *not* taken into account in the determination of its respective *rock cluster*.

Next, we merge the `ProcessInformation` and `ShortTermPlan` data to analyse the mapping between *rock clusters* and *periods*. Quite likely, we will see a one-to-many mapping from *rock clusters* to *periods*, that is, one *rock cluster* can be mapped to one or more *periods*. In other words, the membership of a data point to a *rock cluster* does not determine its corresponding KPI *period*. When we see such a mapping, it means that there were times where the processing of rocks with similar X variables (rock characteristics) achieved different KPI (as different *periods* represent different KPI values). We call the collection of data points with the same *rock cluster* and the same *period* memberships a *class*. Obviously, a *rock cluster* is made up of one or more *classes* - see Figure 4.

Thirdly, the *classes* within each *rock cluster* are characterised by their processing parameters and/or operational settings. As we have neutralised the impact of X variables at this stage, the characterisation exercise at this stage focuses on revealing the Y factors that strongly relate to the KPI in each *class*, by using supervised machine learning techniques such as regression analysis. This characterisation exercise is repeated for every *class* within an interesting *rock cluster* (that is, this task is performed `noClass`-times where `noClass` is a configuration parameter).

Fourthly, the distributions of the identified Y variables (the operational settings/parameters) between *classes* within the same *rock cluster* are compared. The Y variables with statistically significant different distributions are then the processing parameters that may explain KPI differences between *classes* of a particular *rock cluster*.

Finally, the distributions of the Y variables identified in the previous step are visualised to communicate the results of the analysis to users.

2.3. IAM in Detail

Identifying KPI Periods. The first stage of IAM (identification of KPI *periods*) consists of several

[5] From the design of experiments (DOE) perspective, we are controlling the impact of X factor in the analysis such that the contribution of the Y factor on KPIs can be isolated. In other words, it is a fractional factorial DOE.

classes

Date	KWh SAG IV	KWh/TON SAG IV	Power Utilization SAG IV (%)	TPD	BWi	SPI	...	...	Periods	Rock Clusters
1/07/2010	10000.00	....	70.00	80000	10.00	40.00	...	...	1	3
2/07/2010	...	...	...	81000	9.8	39.87	...	...	1	3
3/07/2010	...	...	...	78500	10.1	40.12	...	...	1	3
4/07/2010	...	...	...	99500	9.5	38.89	...	...	2	3
5/07/2010	...	...	...	98000	10.8	40.86	...	...	2	3
6/07/2010	...	...	...	97000	9.6	...	...	...	2	3
7/07/2010	...	...	...	...	...	...	...	...	2	2

periods

rock clusters

Figure 4: Examples of *periods*, *rock clusters*, and *classes*. In this figure, the KPI of interest is the TPD, thus *periods* are determined based on the TPD values. Notice that there are two *periods* identified in the example above. The *rock clusters* in this example are determined based on the BWi and SPI values.

smaller steps (see Figure 5). As a first step, the variable that defines the chosen KPI (from the `ProcessInformation` data) is selected as per the `KPI` parameter set within the `Configuration` object (Figure 2).

Secondly, assuming each data point represents a good quality reading of the KPI (due to, for example, reliable data capture), the KPI values are clustered directly using a known clustering technique, such as k-means clustering. As a guideline, the number of *periods* to identify can either be determined through some statistical analyses, such as within-sum-of-square analysis (Thorndike, 1953), or through domain experts' knowledge. In this case, the number of *periods* to find is configured by the `noPeriods` parameter from the `Configuration` object (see Figure 2).

Alternatively, if the data points are suspected to be noisy for instance due to low-level granularity or environmental interferences, a single reading at that particular point in time will not give a true reflection of the KPI achieved. Instead, we need to look at *trends* over multiple data points such that continuous data points sharing similar KPI readings can be grouped together. We use change point analysis (CPA) techniques. Change point analysis partitions a data stream into several groups of adjacent data points where each group of data points shares similar values. CPA ensures that adjacent groups do not share similar distributions. However, non-adjacent groups may share similar distributions. Consequently, such non-adjacent groups need to be further grouped, yielding the required number of *periods* (as per the configuration parameter `noPeriod`) of similar KPI. As a guideline, a CPA technique that is able to detect change points from a data stream for both parametric and non-parametric data can be used. The `cpm` package written in R (Ross, 2015) has a collection of CPA algorithms that may be suitable in this case.

Currently, our IAM framework only considers a single KPI variable. Nevertheless, IAM can be extended to handle multiple KPI variables. In this case, the technique to identify *periods* needs to be adjusted to handle multiple variables distributions. If CPA is used, then multivariate CPA techniques, such as the `ecp` package in R (James and Matteson, 2015) can be used. If clustering is used, then many existing clustering algorithms can already handle multiple variables.

Thirdly, the identified *periods* are validated. That is, whether the identified *periods* actually contain distinct readings for the KPI of interest, using statistical tests. The choice of test should be guided by the distribution of the data: parametric tests, such as the t-test (Welch, 1947) should be used for normally distributed data. Otherwise, non-parametric hypothesis tests, such as the Mann and Whitney (1947) U-test can be used. Finally, each data point is labelled with the *period* to which it belongs.

In IAM, the combination of techniques applied to extract KPI *periods* described above (such as the clustering or change point analysis technique to use and the statistical test to apply) can be pre-configured by the `findPeriodParameter` in the

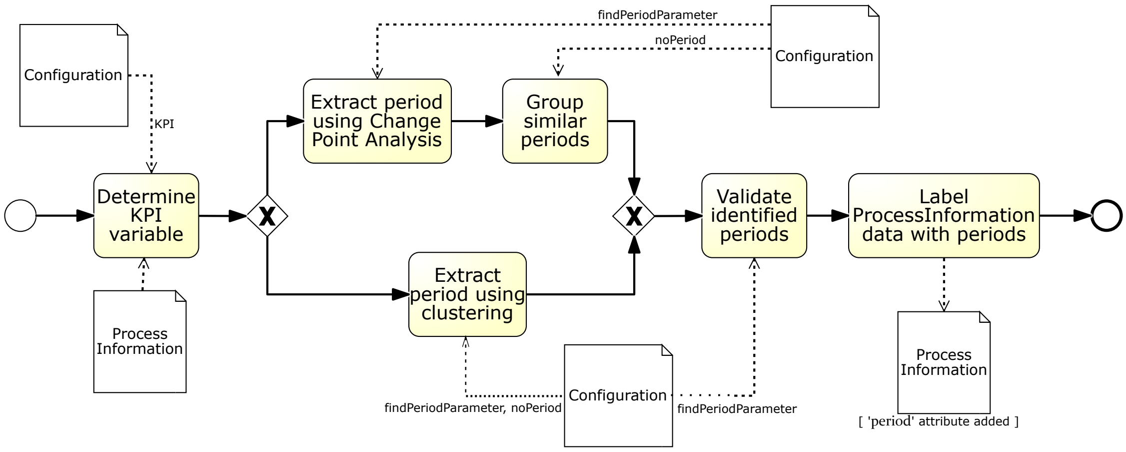


Figure 5: The KPI *period* extraction step of IAM.

`Configuration` object (see Figures 2 and 5).

Isolating Rock Variability. The second stage of IAM is to isolate the influence of X variables from Y variables (see Figure 6). Using the `ShortTermPlan` data as input, this stage first clusters data points that share similar X characteristics into *rock clusters*. Clustering X variables can be performed using one of the many clustering algorithms available, such as k-means (Hartigan and Wong, 1979) or fuzzy clustering (Bezdek, 1981). The number of *rock clusters* can be determined using, for instance, the within sum-of-squares analysis (Thorndike, 1953) or via domain experts' knowledge. In this case, the `noRockCluster` parameter within the `Configuration` object would have been set accordingly (see Figures 2 and 6).

Thirdly, statistical differences in the distributions of the values of each X variable across the *rock clusters* are asserted. Validation can also be conducted via visualisations, such as plotting each X variable, coloured according to their *rock cluster* memberships to show that data points within a *rock cluster* are close to one another while those in different *rock clusters* are far apart. Next, each data point's *rock cluster* is recorded in the data set.

In IAM, the combination of techniques to extract and validate *rock clusters* from a `ShortTermPlan` data can be configured by setting the `findRockClusterParameter` within the `Configuration` object (see Figures 2 and 6).

The next step establishes the mapping between *rock clusters* and KPI *periods*. To do so, we first need to merge both the `ProcessInformation` data and `ShortTermPlan` data (already augmented with *periods* and *rock cluster* attributes, respectively) by joining each data point on the *processTimestamp* attribute.

In this merged data, we denote the collection of data points from one *rock cluster* with the same KPI *periods* a *class*. As the distribution of KPI values amongst different *periods* is distinct, when we see, for example, two *classes* with the same *rock cluster* memberships but different KPI *periods*, one of these *classes* can represent periods (days) with *lower* performance, while the other can represent periods with *higher* performance. Furthermore, as data points within these *classes* share similar rock characteristics, the variations in the KPI achieved may be explained by the Y variables. These *classes* with a significant difference in performance values within the same *rock cluster* membership can thus be compared to see if there are indeed differences in the Y variables (next step of IAM).

Within IAM, the number of *classes* to be compared can be pre-configured by setting the `noClass` parameter within the `Configuration` object accordingly (see Figures 2 and 6).

Characterising Classes. During the third step of IAM, each *class* with different KPI performance as identified in the previous stage is characterised separately as shown in Figure 7. For example, we can characterise a higher KPI class and a lower KPI *classes* within a *rock cluster* separately using supervised machine learning techniques. In this instance, Y variables are independent variables, and the KPI is the dependent variable.

Depending on the richness and relevance of the Y variables in the data set, as an optional second step,

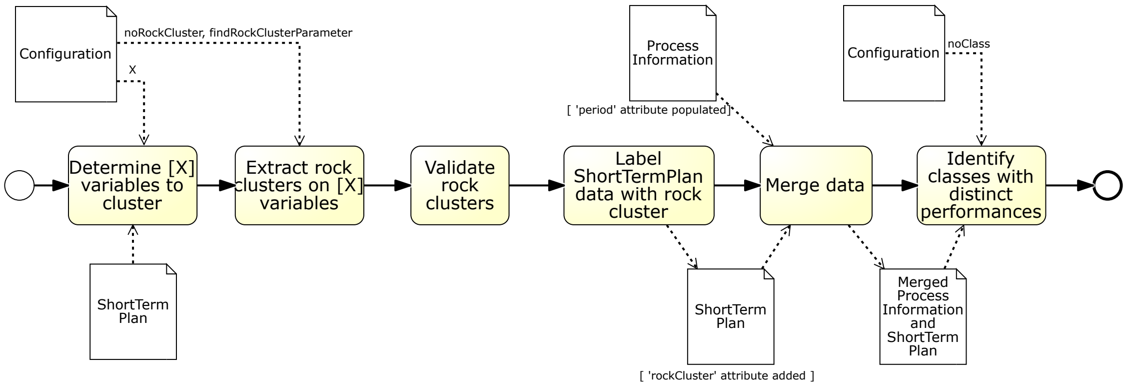


Figure 6: The rock variability isolation step of IAM.

one may enrich the Y variables by applying feature extraction, for instance, using a frequent pattern mining algorithm (Aggarwal and Han, 2014). Typically, feature extraction is performed to supply more discriminative features that may better explain the dependent variable, and may potentially reduce the number of independent variables to analyse and therefore might make the subsequent analyses easier.

For reliable results, a data set can be split into a training set and a test set where the former is used to learn the Y variables that may explain the KPI, while the latter is used to validate that the identified Y variables can explain the KPI values in the test set.

The Y variables that characterise each *class* can be parameterised in the `Configuration` object. Typically, the Y variables used should be as independent of one another as possible - for instance, a simple correlation matrix can be used to check dependence between various Y variables. Y variables that exhibit cross-correlation (i.e. correlation between two time series, each representing a particular Y variable, as a function of time lag) should be removed. The `ccf` tool[6] from R library can be used to identify potential cross-correlation in the data. The Y variables that are shown to have cross correlation and that are not meaningful for analysis may be removed.[7] Depending on the granularity of the timestamp, one may need to do time alignment for the Y variables to get valid regression results. For example, if the granularity of timestamp is at hour- or minute-level, one may need to shift the timestamp of those Y variables that refer to the later stages of ore processing so that each data point refers to the same 'group' of ore being processed.

Similarly, the exact supervised machine learning technique to be applied can also be parameterised by the `charateriseClassParameter` in the `Configuration` object. As a guideline, regression analysis should be used if the values of the dependent variable (the KPI) are continuous, while classification analysis should be used if the KPI values are discrete (i.e. categorical). Typical regression algorithms that can be used are simple linear regression techniques (Filzmoser, 2008), or more complex non-linear ones, such as Ridge and Lasso (Filzmoser, 2008), as well as random forest (Breiman et al., 1984) and feed forward neural network (Fine, 1999). Typical classification algorithm that one could use include the C4.5 algorithm (Quinlan, 1993), RIPPER (Cohen, 1995), as well as random forest (Breiman et al., 1984). These algorithms tend to produce simple classification rules (defined in terms of Y variables) that could be interpreted. Other classification algorithms, such as random forest (Breiman, 2001) (which can be used for both classification and regression analysis), are able to produce results with high accuracy. Furthermore, in combination with a contribution analysis, such as Palczewska et al. (2013), one could learn the influences of Y variables on the KPI variable from the obtained random forest model.

Alternatively, one could use classification method on the *low* and *high* KPI *classes* together (instead of characterising each *class* separately) to extract Y variables that influence the classification of each

[6] `https://www.rdocumentation.org/packages/tseries/versions/0.1-2/topics/ccf`

[7] Note that higher-order cross correlations may still not be detected.

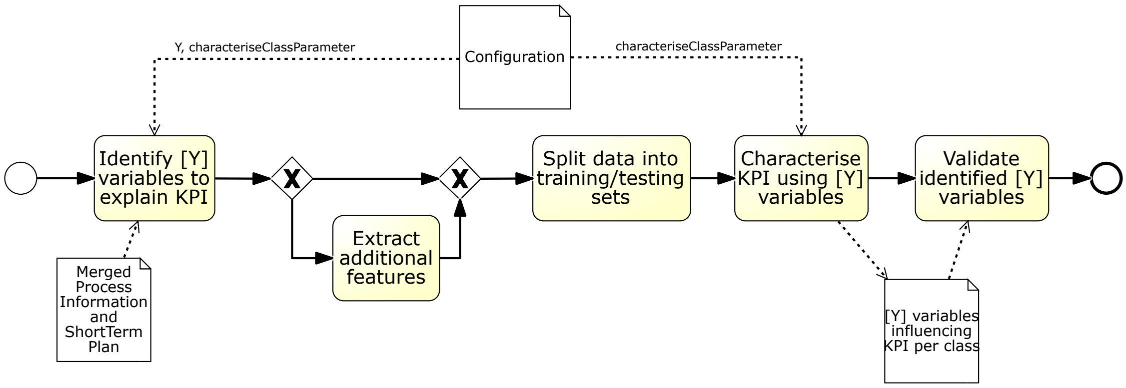


Figure 7: The *classes* characterisation step of IAM, during which key variables are identified.

data point into one of the two *classes*. In this case, the subsequent *comparative analysis* stage of IAM can be skipped.

Comparative Analysis. In the final step of IAM, the identified Y variables are compared across the *classes* of similar *rock clusters* to isolate those that are distinct between *classes*. While these Y variables may explain the KPI very well, it is not certain that they are always distinct across different *classes*. For example, there is likely a strong correlation between power draw and throughput. Thus, the power draw variable may be selected in our previous characterisation exercise as one of the Y variables that may explain the KPI. However, we do not yet know if the power draw between the *classes* is different; if they are the same, then we cannot say that power draw explains the differences in the KPI. Thus, we need to find those identified Y variables that have statistically-different distributions between the *classes* being compared.

The statistical test can be parameterised by the `compareClassParameter` within the `Configuration` object. The choice of the test method, of course, depends on the distribution of the data.

Visualise Results. From the previous statistical testing, we can thus isolate the independent variables that are statistically different. We then visualise these isolated variables through a number of techniques, such as parallel coordinates and/or box-and-whisker graphs to show visually the different range of values that these variables take across different *classes*.

Factors to consider in choosing a visualisation tool are the number of Y variables identified from the previous comparative analysis and the number of data points involved in the analysis. For a relatively small number of Y variables, and/or for a reasonable number of data points to visualise, using parallel coordinates may be effective as it can help one to see variations in the 'trends' between the values of Y variables across different *classes*. However, when there are too many Y variables to visualise or if there are a large number of data points, parallel coordinate visualisation may not work as the graph is likely to be clouded by too many intersecting lines. Visualisation techniques that can visualise the distribution of the values of the identified Y variables in a summarised manner, such as box-and-whisker graphs and bi-plots, may be more appropriate.

In IAM, the visualisation technique to be used can be pre-configured by setting the `visualisationTechnique` parameter within the `Configuration` object.

3. The Dataset

The goal of analysing the data sets taken from the comminution processes of a Cu porphyry deposit in Chile is to understand the impact of different operational settings (as captured by various operational and machine settings of the plant) on a certain KPI while processing materials with similar rock characteristics. The KPI being analysed in this instance is production throughput (TPD).

A data set related to the short-term plan and process sensors readings from the comminution plant were provided. This comminution process involved the operations of one SAG mill and three ball mills. The data set combines data from separate information sources, including data about the short-term mining plan (which includes the estimated rock properties that would be processed each day, such as the estimated BWi and SPI values), the daily summary of the comminution plant operations (such as energy used, total operating hours for each mill, total tonnes of balls being reloaded to each mill, and bearing pressure of the mills), and the daily summary of reconciliated flotation plant outputs (such as distributions of the particle size of materials being fed to the rougher, recovered minerals from rougher, first cleaner, second cleaner, as well as the overall tonnes of minerals recovered). The data set covered approximately 5 years of operation (July 2010 - January 2016, 2041 data points). We refer to this combined data set as the *daily data*.

The data set has some data quality issues, including incorrectly-recorded values and widespread missing values. Variables that contained many incorrectly-recorded values, for instance, negative percentage values for the particle size distribution of ore feed, were removed. Furthermore, data points in which values for the KPI (TPD) being analysed were missing were also removed. Data points where the values for rock properties (BWi and SPI) are obviously incorrect (for example, 0 or negative values) were also removed.

In the 5.5-year period covered in the daily data set, a change was introduced in the comminution plant whereby a fourth ball mill was installed in 2015. The data analysis conducted focused on the period before this change. Furthermore, because the analysis compares operational settings to TPD, a degree of 'similar processing capacity' is needed as a baseline. In our analysis, data points used are restricted to those where all mills (the SAG mill and the three ball mills) were operating at close to full capacity (>22 hours in a day). One could also average the TPD by the number of hours the plant is running in a day; however, we do not do so here because different equipment in the plant (there are one SAG mill and three ball mills) did not always run the same number of hours per day.

In total, the data set contains 1645 data points for the period before the introduction of the fourth ball mill. After filtering, 786 data points remained.

Finally, for the *daily data*, the short-term plan information needs to be aligned to the corresponding operational data so that the processing of different rock batches (with different characteristics) in the comminution and flotation plants can be traced. From site knowledge, there seems to be, on average, a 1-day delay between stockpiling ('stockpile date') and processing in the plant ('process date'). This delay is accounted for (Table 1).

Advances in *ore tracking* technology have resulted in richer data sets that include information about the original batch of rocks being fed to a particular mill at a given point in time. The availability of such information will allow a more precise alignment between batches of rocks (and their corresponding characteristics) and the data corresponding to their processing.

4. Analysis of Daily Data

One of the KPIs to be analysed that the stakeholders are interested in is the throughput of materials being processed daily (tonnes per day - TPD). The IAM configuration parameters set for this analysis are provided in Table 2. Obviously, there are other KPI variables that are just as important, such as grade and recovery. In these cases, similar analysis as demonstrated here can be applied, one simply needs to use the KPI variable of interest, instead of TPD.

4.1. Identifying KPI Periods.

As described in Section 2, the first step of our data analysis is to identify different KPI *periods*. While there is no strong consensus amongst domain experts,[8] some noise still exists, i.e. fluctuations in throughput happen at a daily level, thus a single reading of TPD is not sufficient to identify the *period* to which it belongs. Therefore, we use change point analysis (CPA) to find *periods* w.r.t TPD.

The CPA technique used is the Change Point Model (CPM) technique (Ross, 2015)[9]. The statistical test used to determine if a change has happened is the 'Kolmogorov-Smirnov' hypothesis test (Smirnov, 1948) with an *average run length* value of 10,000 (which means that, on average, there will be 10,000 observations before a false positive occurs). This analysis identified 10 change points, yielding 11 periods with distributions of TPD values (see Figures 8 and 9). Notice that non-adjacent time periods might have similar distributions, as visualised in a probability density function (PDF) graph (Figure 9). Therefore, we

[8]In this article, the domain experts are also co-authors of this paper.

[9]Available in R; `https://www.r-project.org/`.

Table 1: A snippet of the daily data set.

Stockpile date	Processing date	Cu	SAG operating hours	SAG power draw	P80
01-07-2010	02-07-2010	1.58	24.00	15459.49	227
02-07-2010	03-07-2010	1.46	24.00	15316.15	246
03-07-2010	04-07-2010	1.25	24.00	15803.94	223
04-07-2010	05-07-2010	1.10	24.00	15589.85	210
05-07-2010	06-07-2010	1.22	24.00	15514.74	180
06-07-2010	07-07-2010	1.33	24.00	16165.57	204
...					

Table 2: IAM configuration parameters used in the analysis of the daily data.

Parameter	**Value - Daily Data**
`X`	[BWi, SPI]
`Y`	[Bearing Pressure, Power (KwH), Tonnes of Pebbles Crushed, ...]
`KPI`	Throughput (tonnes per day, TPD)
`noRockCluster`	4
`noPeriod`	3
`noClass`	2
`findRockClusterParameter`	"algorithm = k-means; validateMethod = XYPlot"
`findPeriodParameter`	"algorithm = CPA(test=Kolmogorov-Smirnov, ARL=10000); validateMethod = Mann-Whitney U test"
`characteriseClassParameter`	"machineLearning = Step-wise linear regression; isExtractFeature = false"
`compareClassParameter`	"Mann-Whitney U test"
`visualisationTechnique`	"Parallel Coordinates"

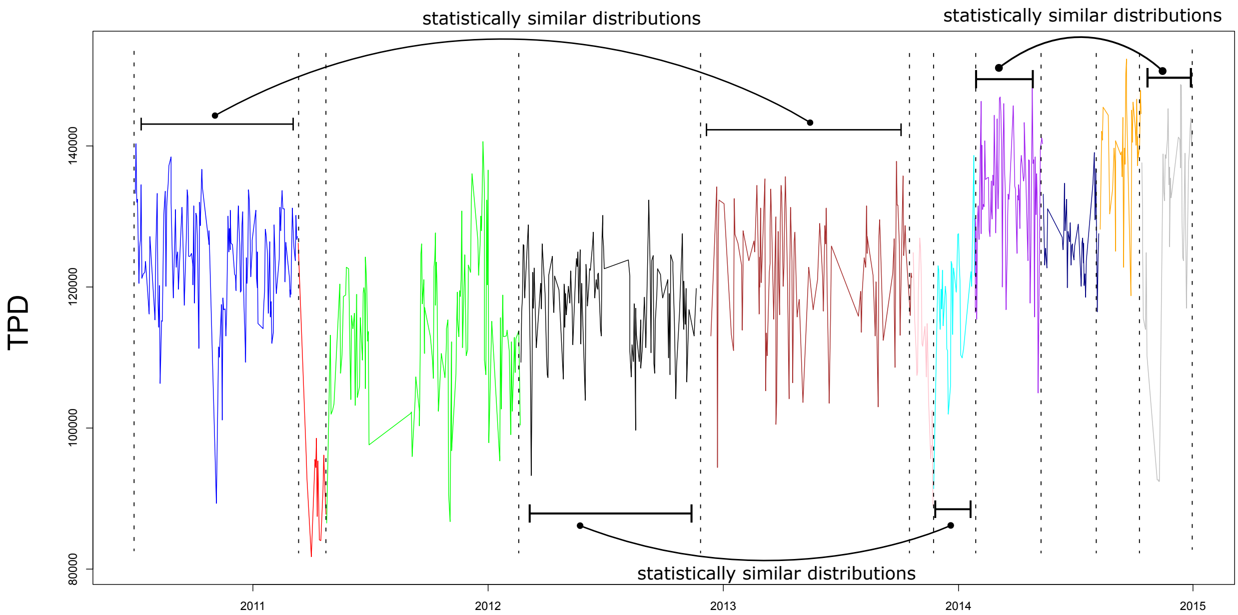


Figure 8: Identified change points (demarcated by vertical dotted lines). Non-adjacent time periods may be statistically similar.

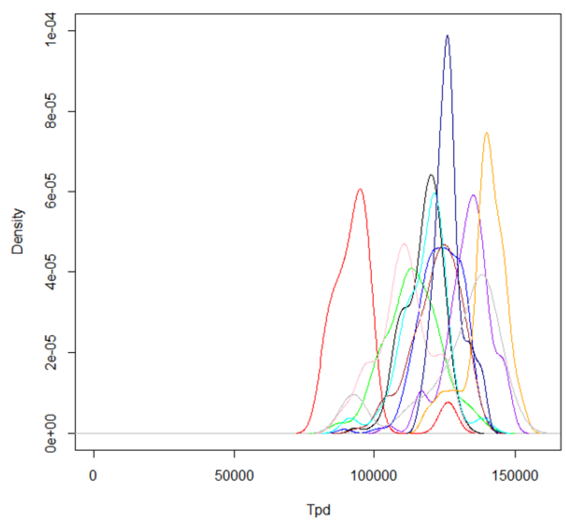


Figure 9: The probability density function (PDF) curves for the 11 time periods that were identified through change point analysis (best viewed in colour). Note that there are similar PDF curves among them. These similar distributions need to be further grouped together.

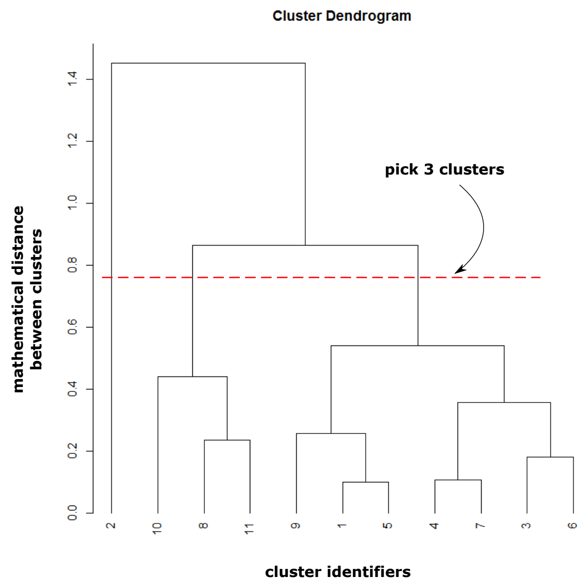


Figure 10: A dendrogram depicting the hierarchical clustering result of the original 11 *rock clusters*.

group similar time periods together to obtain similar KPI *periods*, using a hierarchical clustering approach with average Euclidian distances.

Figure 10 shows a dendrogram of the 11 *periods*. Each leaf represents a *period*, and the height of the connections captures the distance between *periods*. In this analysis, we decided to group the original 11 *periods* into 3 *periods*, based on the distinctness of distances and input from domain experts (Figure 11). The corresponding PDF graph (Figure 12) shows that these 3 *periods* are statistically different at 99% significance level using the Mann-Whitney-Wilcox rank sum test (Mann and Whitney, 1947).[10]

4.2. Isolating rock variability.

To learn optimal operational settings, we need to ensure that we are comparing KPI *periods* over similar rock characteristics. Therefore, we cluster the data on rock properties (SPI and BWi).

Many clustering techniques are available; in this case study, we applied the simple k-means clustering technique (Hartigan and Wong, 1979). This clustering technique groups data points into k clusters whereby data points within each cluster are 'close' to one another based on a mathematical distance function and are 'far apart' for data points across two different clusters. Thus, a necessity in using k-means clustering is to determine the number of clusters. We determined this number using the within groups sum of squares (WSS) (Thorndike, 1953) calculation using the SPI and BWi values. The graph of the WSS plot is shown in Figure 13a. From this figure, we can see that the decrease in the variation within cluster becomes less pronounced as we approach four *clusters*. We, therefore, decided to obtain four *rock clusters* using the simple k-means clustering technique, the result of which is visualised in Figure 13b.

4.3. Identifying classes with different performances.

At this stage, each data point in the data is labelled with both the KPI *period* and the *rock cluster* to which it belongs. The next analysis step is to verify whether there exist similar rock characteristics resulting in different KPI *periods*. In other words, we are identifying high and low KPI *classes*.

In an ideal situation where a *rock cluster* always produces a certain TPD range despite operational settings used, we should see an almost one-to-one matching between a *rock cluster* and a TPD *period*.

[10]Note that there are two 'small' peaks in the PDF for *periods* 2 and 3, which may suggest the existence of other distributions. We suspect this may be due to the chosen hierarchical clustering cut point where we may have aggregated too many *periods* together.

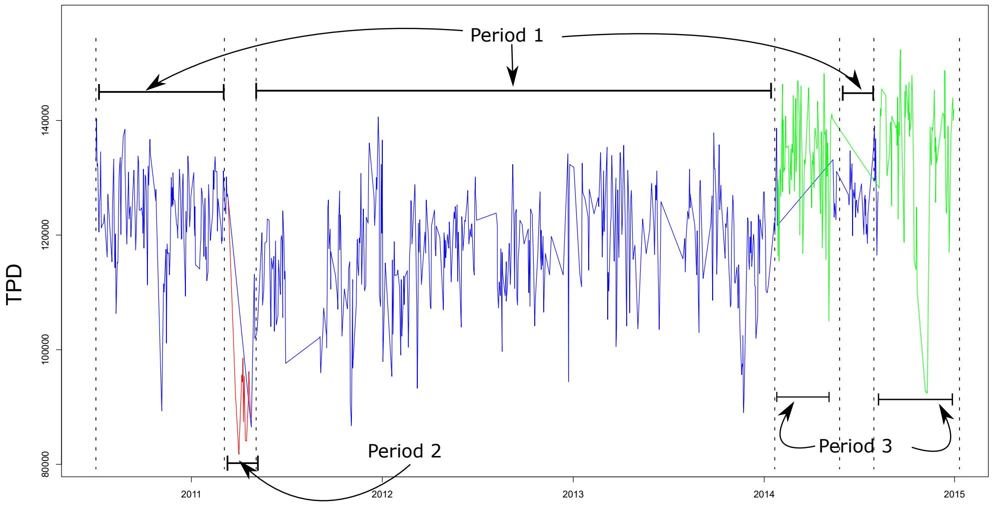


Figure 11: Identified *periods* after hierarchical clustering.

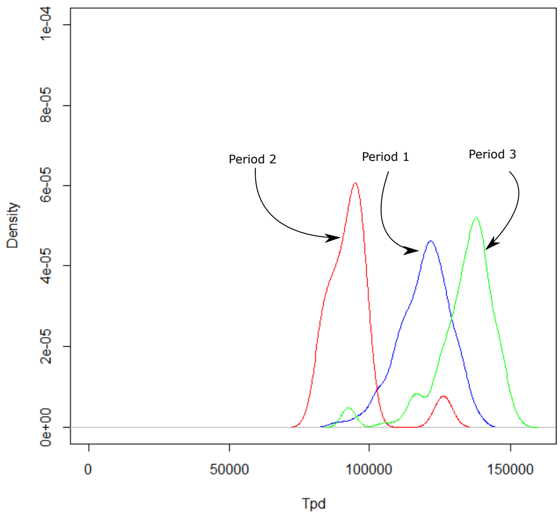


Figure 12: The probability density function graph for the three KPI *periods* extracted through change point analysis and hierarchical clustering.

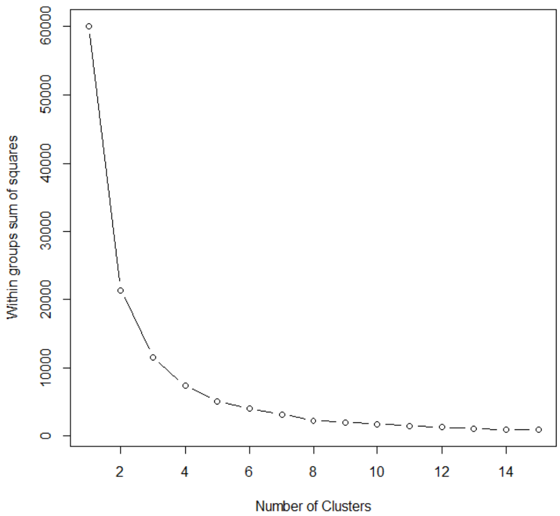
Since each *rock cluster*, by definition, contains data points with similar rock characteristics, a one-to-many mapping between *rock clusters* and TPD *periods* would suggest the existence of differences in the operational settings that contribute to varying TPD achievements.

A summary of the identified *classes* is provided in Figure 14 where we can see that for rock cluster '4', there were 170 days where the throughput belongs to the low KPI *class* (Period 1 - average TPD of 120322.86 tonnes), while for 62 days, the throughput belongs to the high KPI *class* (Period 3- average TPD of 131700.90 tonnes). Similarly, for rock cluster '1', we see that for 44 days, the throughput of the rocks processed can be classified as belonging to the low KPI *class*, while there were 87 days when it managed to reach a high KPI *class*.

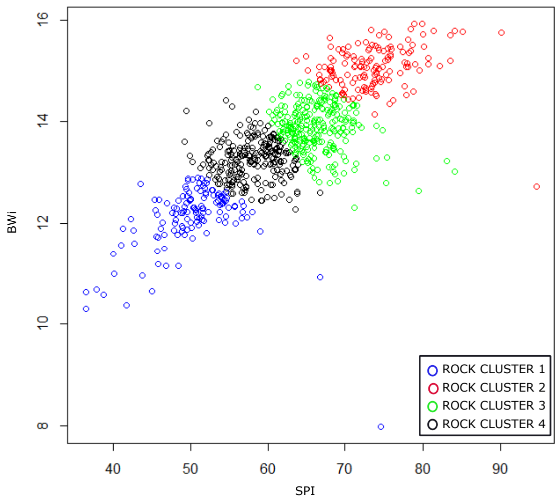
As each rock cluster has more-or-less similar rock characteristics, the existence of multiple *classes* within each rock cluster is likely to be caused by other factors, such as operational setting differences. In the remainder of this chapter, we compare low and high KPI *classes* for rock cluster 4. Extraction of these operational setting differences might aid users in learning better ways of processing various types of rocks.

4.4. Characterising classes.

The first step in characterising the two high and low KPI *classes* is to determine the variables or operational settings that one needs to compare. In a comminution plant, there may be hundreds of operational setting variables being recorded. Attempting to compare all variables may be intractable. A more targeted approach can be achieved by identifying the variables that are important, that is, variables that are critical in explaining the values of the chosen KPI. We can do so by performing a supervised machine learning analysis, such as a regression analysis.



(a) WSS plot for rock properties (SPI and BWi)



(b) A BWi and SPI plot, showing *rock cluster* membership (identified by colours) - daily data.

Figure 13

Recorded process-related variables in our daily data include daily kilowatt-hour used by each mill, the bearing pressure for the mills, the tonnes of balls being reloaded to each ball mill, the tonnes of crushed pebble generated by the plant, and P80. While these variables may not be the actual operational settings themselves, they are reflections of operational settings, and can thus be used as proxies. This does not reduce the relevance of our approach, as regression analysis and the subsequent comparative analysis can be conducted on actual operating parameters, once such information is available.

In this case study, we applied the stepwise regression analysis package (Venables and Ripley, 2013) available in R to automatically select the combination of variables that can best explain the TPD. For the high KPI *class*, our stepwise linear regression resulted in the selection of 7 variables (detailed in Table 3. The coefficients of these 7 variables are statistically significant, and the regression model provided an R-squared value of 0.6308. The stepwise regression for the low KPI *class* also picked up similar variables, although this time, the tonnes of balls reloaded for the first ball mill was also detected as an important variable. However, the R-squared value for this model is much lower (0.3395). Note that rows of data where the values for the chosen variables are missing were removed from the analysis.

4.5. Comparative analysis for optimal operating strategies.

The relatively low R-squared values in our stepwise regression analysis indicate that the variation in the TPD could not be fully explained by the variables in the data set. A more powerful regression algorithm, such as random forest (Breiman et al., 1984) in combination with contribution analysis (Palczewska et al., 2013) may be used to improve the results. However, the fact that the coefficients of our regression analysis are statistically significant indicates that these variables can explain the direction of the trends of TPD. Thus, an understanding of their differences would shed light onto how they influence TPD.

We applied the Mann-Whitney-Wilcox test (Mann and Whitney, 1947) to identify which of the variables that were picked up by our regression analysis are different between the two *classes*. Out of the 11 picked-up variables, 6 variables were found to be statistically different.

For the purpose of visualising the differences in these variables, we used a parallel coordinates graph, with normalised values, as shown in Figure 15.

From the statistical tests and visual clues, we conclude that the distributions of the values of these 6 variables are indeed distinct. From our earlier

Rock Cluster	KPI Period	Data point Count	Average TPD	Average BWi	Average SPI
1	1	44	123248.06	12.26	50.81
	2	6	92688.84	12.34	51.28
	3	87	135251.97	12.04	49.93
2	1	146	117632.57	15.06	73.52
3	1	251	119278	13.89	66.05
	2	7	91232.99	13.20	74.23
	3	10	131381.98	13.79	64.23
4	1	170	**120322.86**	13.32	57.89
	2	3	101929.72	13.15	58.22
	3	62	**131700.90**	13.05	58.01

Low samples (ignore)

low throughput for rock cluster 4

high throughput for rock cluster 4

Figure 14: Identifying high and low KPI *classes* per *rock cluster* - daily data.

Figure 15: A parallel coordinates graph depicting the differences in the variables that impact the achievement of high TPD. The green lines represent data points for the low KPI *class*, while red lines represent data points for the high KPI *class*. The y-values have been normalised to ensure that values from different variables are visually comparable.

High KPI class		
Variables	**Coefficients**	**Significance (p-value)**
(Intercept)	-3.601e+05	0.003398
KWH_SAG	-5.483e+00	0.007800
PSI_SAG	6.155e+01	0.004625
PSI_MB1	-6.127e+01	0.001064
KWH_MB2	2.339e+01	0.000241
BallsReload_MB2	9.047e+02	0.021933
PSI_MB3	4.652e+01	0.021348
CrushedPebble_Plant	6.805e-01	0.022067
Multiple R-squared	0.6308	
Adjusted R-squared	0.582	
Low KPI class		
Variables	**Coefficients**	**Significance (p-value)**
(Intercept)	2.005e+05	4.60e-09
P80	5.458e+01	0.144291
PSI_SAG	-8.963e+00	0.039356
KWH_MB1	-2.038e+00	0.136500
BallsReload_MB1	3.263e+02	0.000151
PSI_MB1	1.443e+01	0.008833
PSI_MB2	-2.548e+01	6.42e-06
KWH_MB3	2.174e+00	0.060155
CrushedPebble_Plant	5.384e-01	0.013033
Multiple R-squared	0.3395	
Adjusted R-squared	0.3062	

Table 3: Stepwise regression results for both the high and low KPI *classes*. KWH_SAG, KWH_MB1, KWH_MB2, KWH_MB3 refer to the power draw (in kilowatt-hour) of the SAG mill and the three ball mills in the plant; PSI_SAG, PSI_MB1, PSI_MB2, and PSI_MB3 refer to the bearing pressure (in PSI) for the SAG mill and the three ball mills in the plant; BallsReload_MB1 and BallsReload_MB2 refer to the tons of balls reloaded for the first and second ball mills; CrushedPebble_Plant refers to the tons of pebbles crushed in the plant; P80 refers to the P80 values produced in the plant.

regression analysis, we have also established that these variables are statistically significant in terms of explaining the trends of TPD. Therefore, we can be reasonably confident that the values seen for the high KPI *class* are indicators of optimal operating strategies for rocks with properties captured in rock cluster 4.

5. Discussion and Future Work

The IAM method proposed in this paper focuses on methods for data analysis. While IAM is flexible in the sense that it supports the extraction of insights about the impact of any chosen rock variables (the X variables) and operational settings (the Y variables) on any KPI of interest, data used as input to IAM should contain enough information of reasonable quality for IAM to deliver quality results (garbage-in-garbage-out).

For example, when we assess the interplay between metallurgical rock characteristics, operational settings, and throughput, the ability to understand how different rock characteristics respond differently to operational settings is critical to optimising processing plants. Therefore, the linkage between characteristics of each batch of rocks being processed, the processing parameters used to process those batches, and the resulting throughput for each batch are crucial to extracting correct insights. While it has traditionally been difficult for operators to know exactly what types of rock are being processed at any given point in time, recent advances in, and adoption of, ore tracking technology (such as SmartTag$^{\text{TM}}$)[11] will gradually allow the collection of better data that allows a better estimation of the types of rock entering a comminution plant at various points in time and consequently, better ways to link processing parameters, rock characteristics, and the resulting processing outputs (such as throughput on a particular day).

Obtaining accurate information about rock characteristics poses another challenge. The nature of rock characteristics data (which is a combination of lab-tested data and interpolated data) calls for IAM to cater for possible uncertainties in rock characteristics. Consequently, the clustering of rock characteristics may inadvertently create an artificial separation between groups of rock that is not meaningful. To address these issues, IAM could be extended with other statistical measurements to inform users about uncertainties in the data sets, such as a non-parametric confidence interval for the values of rock properties indicating the lower and upper boundary of rock hardness measures per cluster. Alternatively, advanced clustering techniques, such as fuzzy clustering (Ferraro and Giordani, 2015) may also be used to communicate the degree of uncertainty in the mapping of rock characteristic information to various *rock clusters*. Such an approach will allow room for domain experts to have the final say in deciding the cluster membership of a particular batch of rocks.

Finally, IAM only defines the sequence of analysis steps needed to extract the 'optimal' operational

[11] http://www.metso.com/services/ore-tracking/

strategies. Data analysts who apply IAM have the freedom to choose the most appropriate statistical or machine learning algorithm, based on the nature of the data. For example, to perform the clustering of rock characteristics, users can choose more advanced clustering algorithms, such as DBScan (Ester et al., 1996). This allows us to perform similar data analysis multiple times, each using different, yet still suitable, algorithms. The advantage of doing so is that results from these multiple analyses can be compared to check if consistent insights are extracted, allowing a form of results validation.

We have further demonstrated the *generality* of IAM by having successfully applied IAM to *another data set* from the same plant but at a different level of granularity (hourly data). The results are not presented here but will be incorporated as part of our future work to compare different types of insights that one could extract by applying IAM on data with different level of granularity. Such a comparison may lead to a better understanding of the different data collection strategies that one can employ on site.

Other work involves the improvement of the types of analysis algorithms that can be used in IAM to incorporate those that deal with inherent uncertainties in the data, especially in clustering of rock characteristics. We also envisage the use of more advanced feature extraction techniques (for example, frequent pattern mining techniques (Aggarwal and Han, 2014)) such that one can study the impact of a combination of variables (such as the combination of a number of processing parameters) on a KPI of interest (such as throughput). Finally, it is also important that IAM is further applied on data sets from different processing plants to further assess its generality. Eventually, through further application of IAM, we seek to further strengthen the guidance for the choice of various analysis methods and the settings for the various parameters that IAM needs.

6. Related Work

Other research has been conducted in looking at how one could improve the process of mineral extraction.

Analysing plant operational data to gain insights about a plant's performances has been around for some time. Mackay and Lloyd (1975) applied regression analysis to identify factors affecting residue in a gold mining plant. Similarly, Brittan and van Vuuren (1973) applied multilinear stepwise regression to identify variables influencing gold recovery. However, these work tend to 'mix' the influence of ore characteristics with operational settings on the performance of a plant. A good overview of the use of data mining and other analytics methods in the process industry is summarised in the work by Ge et al. (2017). The IAM methodology proposed in this paper is therefore different from existing work in this area in that, in this article, we propose a step-by-step data analysis *method* to separate the influence of rock characteristics from operational settings on plant's performance. In other words, IAM guides users in *how* to analyse disparate data sets collected in today's plant operation to extract key variables influencing the performance of a plant in the processing of ore with certain characteristics.

Within the monitoring and control systems domain, such as Houseman et al. (2001), process data is heavily used. However, not all needed variables might be available at all times. This problem is addressed in Slišković et al. (2012) for the manufacturing industry: several models are described that estimate necessary process variables using measured on-line process variables. IAM lifts and specialises this to the field of resource extraction by separating the influence of rock properties and processing parameters.

In the field of resource extraction, recorded sensor data has been used to design or optimise plants. For instance, in Lestage et al. (2002) and Pan (2013), a grinding circuit is optimised using Linear Programming or other optimisation techniques. In Steyn et al. (2013), on-line real-time data is used to optimise SAG mill control systems. In Zhang et al. (2002), a genetic algorithm is applied to the recorded sensor data of a coal grinding mill in a UK power station. Finally, in the work by Marais and Aldrich (2011) and Aldrich et al. (2010), they showed how froth images can be used for the development of advanced control systems for platinum flotation processes, if features of these images can be linked to KPIs. IAM could aid in establishing this link by choosing between these features, such as in the identification of processing periods (see Figure 3).

Genetic algorithm uses a repeated procedure of random crossover, mutation and selection to construct a model of the plant's operations based on the input data. Genetic algorithms were also used in Huband et al. (2005) and Mhlanga et al. (2011) to design a complete comminution circuit for an iron-ore processing plant in Australia, thereby improving over manually-designed plants. Genetic algorithms optimise towards some evaluation criteria, which makes choosing good criteria critical. Therefore, as in Mhlanga et al. (2011), the obtained designs were validated using simulation techniques. IAM has a different purpose and differs from these methods in providing systematic steps and taking rock properties into account.

Another way to optimise operations is to build one or more models of (parts of) the plant and simulate these models using various processing parameters. For instance, Discrete Element Simulation (DES) can be used to simulate the interaction between rock particles to aid the Computer Aided Design (CAD) of SAG mills (Morrison and Cleary, 2008; Narayanan, 1987) and to compare power efficiency between designs (Cleary and Morrison, 2011). While DES is suitable to simulate the interaction of particles in a single mill or another machine, simulating models spanning an entire plant would take infeasibly long. Other model-specific simulation models are proposed in Yang et al. (2004), in which a simulation of a hydrocyclone is validated using sample data, and in SimSAGe (2016), in which SAG mills are simulated using data from liner wear inspections.

Such machine-specific simulation models can be used to optimise single machines, however as all steps in resource processing plant might influence each other, simulations should take groups of machines into account (Duarte et al., 1998), and preferably the entire planning, scheduling and control systems (Harjunkoski et al., 2009b; Shobrys and White, 2002). For instance, commercial simulation approaches such as SimSAGe (2016), JKtech (2016), and Bear Rock (2016) simulate comminution, flotation and classification circuits, blasting and machine models, and these methods include data obtained by mass balancing, simulation and plant surveys. Furthermore, in Sosa-Blanco et al. (1999), a technique is proposed to simulate a grinding-flotation plant for lead-silver ore, in particular how changes in comminution influence flotation, thereby aiming to optimise flotation. In the work produced by the Collaborative Research Centre - Optimising Resource Extraction (2017), the concepts of all these separate simulations are combined into a configurable plant-independent simulation tool. Khalesi et al. (2015) propose a techno-economic simulation tool that allows the simulation to study the impact of circuit designs on economic indicators, such as Net Present Values.

The approaches described above are model-driven and tend to be 'generic' to any machines. On the other hand, IAM is data-driven, thus allowing a certain degree of customisation (as data used for analysis comes from the particular machines being used in the plant). More importantly, IAM takes rock properties into account.

Traditionally, models used in a simulation are calibrated using data extracted from expensive surveys. Such an approach is expensive as mill surveys involve the halting of the plant for a substantial amount of time. Furthermore, the information gathered from mill surveys only captures the state of the plant at one particular point in time (the time when the survey was taken). Longer term variability is not captured in this data. Using IAM, these simulations can be calibrated to improve the accuracy of the simulation models and consequently optimise operational settings, possibly extended with an error propagation analysis. That is, IAM could be used to validate the comparisons between models and to refine models using real-life on-line data from existing equipment, similar to Slišković et al. (2012), in addition to more expensive data such as the results of mine surveys. By learning from historical data, IAM eliminates the limitations of mine surveys as a large amount of data is already being collected in today's plants. IAM could use this data as a proxy to mill survey data, hence allowing us to better learn the behaviour of the system, including how different types of rocks respond differently to different processing parameters and operational settings.

IAM can use data from various sources, as shown by our analysis in Section 4, in which we used data obtained from process control systems (such as power draw, speed) as well as data from short-term planning (block models). Several other types of data could be obtained from mining operations, for example, commercial tools such as SPLIT Engineering (2016) provide optical image recognition technologies to measure particle size distributions of streams of rocks.

7. Conclusion

In this paper, we have presented IAM, a general data analysis method that can be used to isolate the impact of rock properties and operational settings on mineral processing KPI (such as throughput).

Key data set requirements, as well as the corresponding configuration parameters, that IAM needs have been generalised in the form of a class diagram. Furthermore, we have also provided some guidelines on factors to consider in the setting of the parameters as well as in the choice of various analysis methods that IAM requires.

The IAM method proposed in this paper has been evaluated using a data set from a comminution plant operation related to a Cu porphyry deposit in Chile. In this evaluation, we showed how one could extract optimal operating strategies to achieve optimal throughput for rocks with certain characteristics. The results of our analysis have been discussed with domain experts to validate their correctness. Equally important, we have discussed some of the challenges in applying IAM, especially the lack of a more accurate way to link rock characteristics data (for each batch of rocks being processed) and the corresponding processing parameters/operational settings applied. Nevertheless, recent advances in ore tracking technology make it possible that such information will be obtainable in the near future.

Funding: This work was supported by the Cooperative Research Centre for Optimising Resource Extraction 2 (CRC ORE 2), project P4-001.

Aggarwal, C. C., Han, J. (Eds.), 2014. Frequent Pattern Mining. Springer.

Aldrich, C., Marais, C., Shean, B., Cilliers, J., 2010. Online monitoring and control of froth flotation systems with machine vision: A review. International Journal of Mineral Processing 96 (14), 1 – 13.

Bear Rock, 2016. Advanced sensors. `http://bear-rock.com/wordpress/portfolio/advanced-sensors/`, accessed: 2016-12-01.

Bearman, R., 2013. Step change in the context of comminution. Minerals Engineering 4344, 2 – 11, SI: Comminution.

Bezdek, J., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., et al., January 1984. Classification and Regression Trees, 1st Edition. Chapman and Hall/CRC, ISBN-10: 0412048418.

Brittan, M., van Vuuren, E., 1973. Computer analysis, modelling and optimisation of gold recovery plants of the anglo american group. Journal of the Southern African Institute of Mining and Metallurgy 73 (7), 211–222.

Carrasco, C., Keeney, L., Napier-Munn, T., Bode, P., 2016a. Unlocking additional value by optimising comminution strategies to process grade engineering streams. Minerals Engineering 103–104, 2–10.

Carrasco, C., Keeney, L., Scott, M., Napier-Munn, T., 2016b. Value driven methodology to assess risk and operating robustness for grade engineering strategies by means of stochastic optimisation. Minerals Engineering 99, 76 – 88.

Cleary, P. W., Morrison, R. D., 2011. Understanding fine ore breakage in a laboratory scale ball mill using DEM. Minerals Engineering 24 (34), 352 – 366, SI: Comminution.

Cohen, W. W., 1995. Fast effective rule induction. In: Proceedings of 12th International Conference on Machine Learning. Morgan Kaufmann, pp. 115–123.

Collaborative Research Centre - Optimising Resource Extraction, 2017. Integrated Extraction Simulator. `http://www.crcore.org.au/main/index.php/solutions/integrated-extraction-simulator-ies`, accessed: 2017-02-14.

Duarte, M., Sepúlveda, F., Redard, J., Espinoza, P., Lazcano, V., Castillo, A., Zorbas, A., Giménez, P., Castelli, L., 1998. Grinding operation optimization of the codelco-andina concentrator plant. Minerals engineering 11 (12), 1119–1142.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96. AAAI Press, pp. 226–231.

Ferraro, M. B., Giordani, P., 2015. A toolbox for fuzzy clustering using the R programming language. Fuzzy Sets and Systems 279, 1 – 16, Theme: Data, Audio and Image Analysis.

Filzmoser, P., 2008. Linear and nonlinear methods for regression and classification and applications in R. Tech. rep., Vienna University of Technology.

Fine, T. L., 1999. Feed–forward Neural Network Methodology. Springer.

Foggiatto, B., Hilden, M. M., Powell, M., 2014. Advances in the simulation of flexible circuits. In: Proceedings of the 12th AusImm Mill Operators' Conference 2014. No. 9/2014. AusIMM: Australasian Institute of Mining and Metallurgy, pp. 391–398.

Ge, Z., Song, Z., Ding, S. X., Huang, B., 2017. Data mining and analytics in the process industry: The role of machine learning. IEEE Access 5, 20590–20616.

Group, O. M., March 2015. OMG Unified Modeling Language.

Harjunkoski, I., Nystrm, R., Horch, A., 2009a. Integration of scheduling and control: Theory or practice? Computers & Chemical Engineering 33 (12), 1909 – 1918, FOCAPO 2008.

Harjunkoski, I., Nyström, R., Horch, A., 2009b. Integration of scheduling and control: Theory or practice. Computers & Chemical Engineering 33 (12), 1909–1918.

Hartigan, J. A., Wong, M. A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1).

Hesse, M., Popov, O., Lieberwirth, H., 2016. Increasing efficiency by selective comminution. Minerals Engineering.

Houseman, L., Schubert, J., Hart, J., Carew, W., 2001. Plantstar 2000: A plant-wide control platform for minerals processing. Minerals Engineering 14 (6), 593–600.

Huband, S., Barone, L., Hingston, P., While, L., Tuppurainen, D., Bearman, R., Sept 2005. Designing comminution circuits with a multi-objective evolutionary algorithm. In: 2005 IEEE Congress on Evolutionary Computation. Vol. 2. pp. 1815–1822.

James, N. A., Matteson, D. S., 2015. ecp : An R package for nonparametric multiple change point analysis of multivariate data. Journal of Statistical Software 62 (7).

JKtech, 2016. JKSimMet, JKSimFloat, JKMultiBal, JKSimBlast. http://jktech.com.au/jksimmet, accessed: 2016-12-01.

Khalesi, M. R., Zarei, M. J., Sayadi, A. R., Khoshnam, F., Chegeni, M. H., 2015. Development of a techno-economic simulation tool for an improved mineral processing plant design. Minerals Engineering 81 (Supplement C), 103 – 108.

Lestage, R., Pomerleau, A., Hodouin, D., 2002. Constrained real-time optimization of a grinding circuit using steady-state linear programming supervisory control. Powder Technology 124 (3), 254–263.

Mackay, J. G., Lloyd, P. J. D., 1975. Progress in assessing plant operating data. Recent advances in mineral dressing: papers presented at a symposium under the auspices of the National Institute for Metallurgy and the South Africa 76 (Special Issue), 162–170.

Mann, H. B., Whitney, D. R., 03 1947. On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics 18 (1), 50–60.

Marais, C., Aldrich, C., 2011. Estimation of platinum flotation grades from froth image data. Minerals Engineering 24 (5), 433 – 441.

Mhlanga, S., Ndlovu, J., Mbohwa, C., Mutingi, M., Dec 2011. Design of comminution circuits for improved productivity using a multi-objective evolutionary algorithm (MOEA). In: 2011 IEEE International Conference on Industrial Engineering and Engineering Management. pp. 1680–1684.

Morrison, R., Cleary, P., 2008. Towards a virtual comminution machine. Minerals Engineering 21 (11), 770–781.

Napier-Munn, T., 2015. Is progress in energy-efficient comminution doomed? Minerals Engineering 73, 1–6.

Narayanan, S., 1987. Modelling the performance of industrial ball mills using single particle breakage data. International Journal of Mineral Processing 20 (3), 211 – 228.

OMG, January 2011. Business Process Model and Notation (BPMN). Version 2.0 Edition.

Palczewska, A., Palczewski, J., Robinson, R. M., Neagu, D., 2013. Interpreting random forest models using a feature contribution method. In: 14th International Conference on Information Reuse and Integration (IRI). IEEE, pp. 112–119.

Pan, X., 2013. System integration of automated mine optimization system. IFAC Proceedings Volumes 46 (16), 148–154.

Powell, M., Foggiatto, B., Hilden, M., 2014. Practical simulation of flexicircuit processing options. In: Proceedings XXVII International Mineral Processing Congress IMPC 2014. pp. 219–228.

Prior, T., Giurco, D., Mudd, G., Mason, L., Behrisch, J., 2012. Resource depletion, peak minerals and the implications for sustainable resource management. Global Environmental Change 22 (3), 577–587.

Quinlan, J. R., 1993. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ross, G. J., 2015. Parametric and nonparametric sequential change detection in R: The cpm package. Journal of Statistical Software 66 (3).

Shobrys, D. E., White, D. C., 2002. Planning, scheduling and control systems: why cannot they work together. Computers & Chemical Engineering 26 (2), 149–160.

SimSAGe, 2016. SAG on-line charge simulators. http://simsage.com.au/projectdetail.php?pId=22, accessed: 2016-12-01.

Slišković, D., Grbić, R., Hocenski, Ž., 2012. Methods for plant data-based process modeling in soft-sensor development. AUTOMATIKA: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije 52 (4), 306–318.

Smirnov, N., 1948. Table for estimating the goodness of fit of empirical distributions. The annals of mathematical statistics 19 (2), 279–281.

Sosa-Blanco, C., Hodouin, D., Bazin, C., Lara-Valenzuela, C., Salazar, J., 1999. Integrated simulation of grinding and flotation application to a lead-silver ore. Minerals Engineering 12 (8), 949 – 967.

SPLIT Engineering, 2016. Split-online systems. https://www.spliteng.com/products/split-online-systems/, accessed: 2016-12-01.

Steyn, C., Keet, K., Breytenbach, W., 2013. Optimization and control of a primary sag mill using real-time grind measurement.
URL http://www.bluecubesystems.com/images/stories/pdf/Com12PrimarySAGMillGrindControlFinal.pdf

Thorndike, R. L., 1953. Who belongs in the family? Psychometrika 18 (4), 267–276.

Venables, W. N., Ripley, B. D., 2013. Modern applied statistics with S-PLUS. Springer Science & Business Media.

Walters, S., 2016. Driving productivity by increasing feed quality through application of innovative grade engineering technologies. Tech. rep., CRC ORE.

Welch, B. L., 1947. The generalization of student's problem when several different population varlances are involved. Biometrika 34 (1–2), 28–35.

Wills, B. A., 2006. Mineral Processing Technology - An Introduction to the Practical Aspects of Ore Treatment and Mineral Recovery. Elsevier.

Yang, I., Shin, C., Kim, T.-H., Kim, S., 2004. A three-dimensional simulation of a hydrocyclone for the sludge separation in water purifying plants and comparison with experimental data. Minerals Engineering 17 (5), 637 – 641, hydrocyclones '03.

Zhang, Y. G., Wu, Q. H., Wang, J., Oluwande, G., Matts, D., Zhou, X. X., Dec 2002. Coal mill modeling by machine learning based on onsite measurements. IEEE Transactions on Energy Conversion 17 (4), 549–555.