

# Automated Classification and Analysis of Non-metallic Inclusion Data Sets



MOHAMMAD ABDULSALAM, TONGSHENG ZHANG, JIA TAN,  
and BRYAN A. WEBLER

The aim of this study is to utilize principal component analysis (PCA), clustering methods, and correlation analysis to condense and examine large, multivariate data sets produced from automated analysis of non-metallic inclusions. Non-metallic inclusions play a major role in defining the properties of steel and their examination has been greatly aided by automated analysis in scanning electron microscopes equipped with energy dispersive X-ray spectroscopy. The methods were applied to analyze inclusions on two sets of samples: two laboratory-scale samples and four industrial samples from a near-finished 4140 alloy steel components with varying machinability. The laboratory samples had well-defined inclusions chemistries, composed of MgO-Al<sub>2</sub>O<sub>3</sub>-CaO, spinel (MgO-Al<sub>2</sub>O<sub>3</sub>), and calcium aluminate inclusions. The industrial samples contained MnS inclusions as well as (Ca,Mn)S + calcium aluminate oxide inclusions. PCA could be used to reduce inclusion chemistry variables to a 2D plot, which revealed inclusion chemistry groupings in the samples. Clustering methods were used to automatically classify inclusion chemistry measurements into groups, *i.e.*, no user-defined rules were required.

<https://doi.org/10.1007/s11663-018-1276-x>

© The Minerals, Metals & Materials Society and ASM International 2018

## I. INTRODUCTION

NON-METALLIC inclusions are an inevitable product of chemical reactions occurring during steel processing and they play an important role in defining the properties of steel.<sup>[1]</sup> If not controlled, they can often reduce ductility, fatigue resistance, and toughness of steels.<sup>[2–7]</sup> There has been a significant amount of research on inclusion control during liquid steel processing<sup>[8]</sup> with intentional efforts to control inclusion populations referred to as “inclusion engineering” with the resulting product called “clean steel.”

An enabling technology for inclusion engineering efforts on both the laboratory scale and the production scale is automated analysis in a scanning electron microscope equipped with energy dispersive X-ray spectroscopy (SEM/EDS).<sup>[9]</sup> While one of several analysis and quantification methods,<sup>[10]</sup> automated SEM/EDS can measure hundreds or thousands of individual inclusions per sample and obtain chemistry, size, shape,

and spatial distribution information in times on the order of hours or less. Several recent studies have advanced in back-scattered electron (BSE) imaging and EDS measurement parameters,<sup>[11,12]</sup> to improve the speed and accuracy of the method. Automated analysis is now a common tool among companies and researchers in the steel industry and it has enabled many developments in the scientific understanding of inclusions and in industrial process control.<sup>[10,13–18]</sup>

Inclusion chemistry and chemistry changes have been of primary interest. Considering many different grades and process routes, inclusions could be combinations of Al, Ca, Si, Mg, Mn, Ti, Zr, Ce, La, O, S, and N although the total number of possible elements is usually constrained by grade or processing conditions. Despite the constraints, often more than three variables are needed to fully represent inclusion populations and this creates a visualization challenge. Display of inclusion chemistry results is most typically done by Gibbs triangle ternary plots with inclusion compositions represented by cation or anion mole or mass fraction. An example plot is shown in Figure 1(a), where each inclusion chemistry in mole fraction is represented by a single data point. A proportional symbol plot (Figure 1(b)) shows the same data but with the symbol size proportional to the number of inclusions of that chemistry. The assumption is frequently made that there is little chemistry change during solidification, so the phases that comprise an

MOHAMMAD ABDULSALAM, TONGSHENG ZHANG, and BRYAN A. WEBLER are with the Materials Science and Engineering Department, Center for Iron and Steelmaking Research, Carnegie Mellon University, Pittsburgh, PA, 15213. Contact e-mail: mabdulsa@andrew.cmu.edu JIA TAN is with Nucor Castrip Arkansas LLC, Blytheville, AR, 72315

Manuscript submitted September 01, 2017.

Article published online May 14, 2018.

inclusion in the liquid steel can be inferred from its composition by overlaying the ternary phase boundaries for the system at steelmaking temperatures, as shown in Figure 1(c).

Representing inclusion chemistry by ternary diagram provides an easily interpretable view of inclusion chemistry distribution. However, they are limited in that only three variables can be plotted. For example, the Ca-containing inclusions in Figure 1 are CaS and this can only be verified by plotting the distribution on an Al-Ca-S diagram. Other data representation schemes have been proposed,<sup>[19]</sup> but still a diagram (or diagrams) must be produced for each sample. These diagrams provide engineers easily interpreted visual representations of inclusion chemistry distributions that are typically used to diagnose specific process issues.<sup>[18]</sup> Comparisons of many samples (*e.g.*, looking at trends in behavior over many heats of steel or sequential samples) are difficult because one or more diagrams must be compared. Overall averages do not capture the multi-phase nature of inclusions. User-defined classification rules have been developed based on expected chemistry ranges.<sup>[18]</sup> However, this requires assumptions about the expected inclusion chemistries and they must be consistently applied.

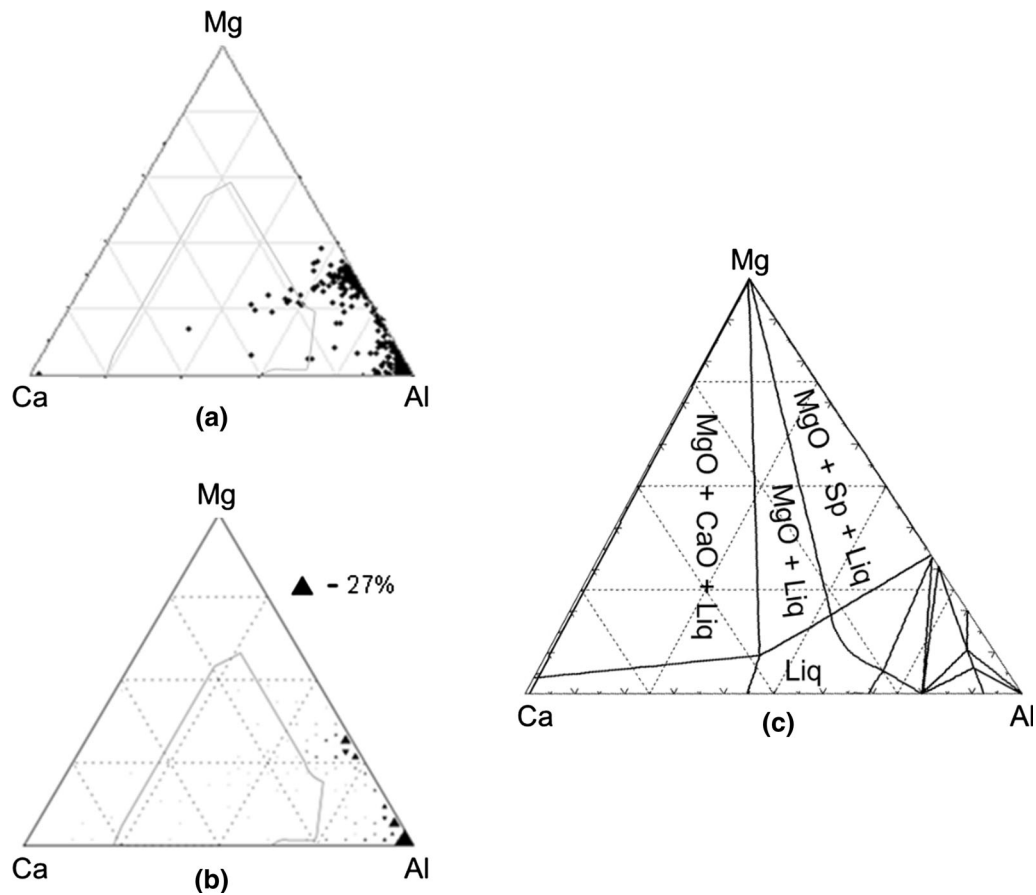


Fig. 1—Two representations of the same inclusion chemistry distribution (a) each measured chemistry (cation fraction) plotted as one data point, (b) the same data with symbol size proportional to the number of inclusions with that chemistry. (c) The ternary phase diagram of the MgO-CaO-Al<sub>2</sub>O<sub>3</sub> system at 1873 K (1600 °C).

This study applied methods of classifying, learning from, and representing large, multivariate data sets to inclusion chemistry data measured by automated SEM/EDS analysis. The methods investigated were (1) principal component analysis (PCA), and (2) cluster analysis. Method (1) is a technique for dimensionality reduction, *i.e.*, representing multivariate chemistries in a two-dimensional plot, and method (2) automatically groups inclusions by chemistry. These methods were applied to a laboratory- and an industrial-scale data sets, to illustrate how these methods might be used to condense inclusion chemistry presentation, *i.e.*, not require one or more ternary diagrams per sample. The industrial data set was gathered from analysis of two semi-finished components that exhibited differences in machinability. The objective of the analysis was to compare the inclusion populations to examine if they might be the cause of the machinability differences, as inclusions are well known to influence machinability.<sup>[20]</sup>

## II. ANALYSIS METHODS

### A. Principal Component Analysis (PCA)

PCA was investigated as a dimensionality reduction technique. It is a common exploratory data analysis tool

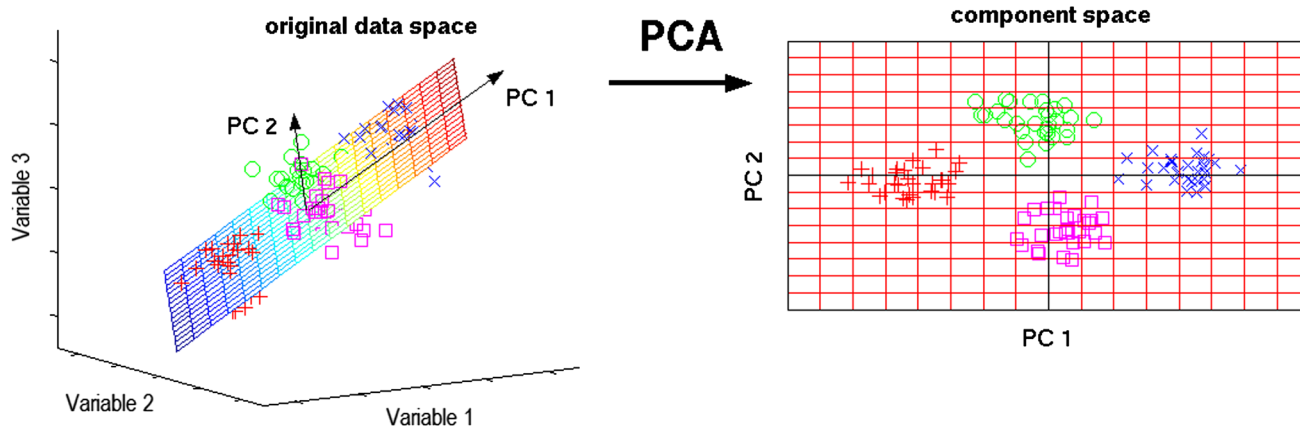


Fig. 2—Transformation of 3D data to 2D using PCA, reprinted from Ref. [22].

that is used to provide a visual relationship between observations for multivariate data sets. The aim of PCA is to extract the important information from multivariate data sets and express the output as a set of new orthogonal variables called principal components (PC).<sup>[21]</sup> The new PC variables are linear combinations of the original variables. The principal components are generated such that the greatest variance (*i.e.*, the greatest spread in the data set) between the observations lies on the first principal component, PC1, and the second greatest variance on PC2, and so on; in addition, the succeeding PC is computed on the constraint that it is orthogonal to the previous PC. The number of PCs computed is equal to the number of variables in the initial data set, but in most cases more than 80 pct of the variance between the observations is contained in the first two PCs, thus enabling the plotting of multivariate data sets on 2D scatter plots with minimal information loss.<sup>[21]</sup> Figure 2 shows an example of PCA, where high dimensional data, 3D, is transformed to 2D using the first 2 PCs, as a result the observations are viewed in simple 2D plots while losing minimal information, keeping in mind that PCs are linear combinations of the original variables.

### B. Cluster Analysis

Not to be confused with the physical clustering of inclusions, clustering is an automated method of arranging data into groups (referred to as clusters). The analysis breaks down a data set into a set of clusters, such that observations in one cluster are related to each other, in one way or another, more than observations in other clusters. Accordingly, the analysis referred to in this study pertains to the cluster of data, where a cluster is composed of inclusions with similar chemistries. Clustering itself is not one specific algorithm that can be applied to any set of data, rather it is a general term used that serves a specific task. There are numerous clustering models, each with its own algorithm on how to define and identify a cluster; as a result, the different models vary significantly depending on the input parameters and desired output.<sup>[23]</sup>

Because there are numerous models, an understanding of the various cluster models and the data sample is essential for selecting the appropriate clustering algorithm. The most basic method is k-means clustering,<sup>[23]</sup> a centroid-based clustering technique where the number of clusters is predetermined and the initial cluster centroids are randomly assigned. Then observations are assigned to the nearest cluster by minimizing the distance between the observations and the centroids of the various clusters. Distance here is a generic term. The physical distance between any two points in 2D space is  $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$  but this can be generalized to any variables (not just positions).

While fast and conceptually straightforward, two drawbacks of k-means clustering are that the number of clusters for a data set must be preselected and that equal size clusters are often produced. Other algorithms do not have these restrictions. One such algorithm is the expectation-maximization (EM) algorithm and this method was employed in this study. The EM algorithm is a soft clustering method. In contrast to hard clustering, where each observation is assigned to one specific cluster, soft clustering assigns probabilistic mix proportions for an observation to belong to each cluster. For example, if sample *X* has 2 clusters, the *i*th observation has certain probabilities of belonging to clusters 1 and 2, and the algorithm assigns the observation to the cluster with the highest probability. If the *i*th observation has a 95 pct chance of belonging to cluster 1, and therefore a 5 pct chance of being in cluster 2, it will be assigned to cluster 1. This form of clustering is beneficial for assessing the quality of cluster classification, and when observations are midway between clusters. The main assumption behind the EM algorithm is that each cluster has its own Gaussian distribution, and the whole data set is a Gaussian Mixture Model (GMM).

The EM algorithm is an iterative clustering technique, involving two main steps, the expectation (E) and maximization (M) steps. First the initial parameters, the means and covariances of each cluster, are estimated, either randomly or using other statistical techniques such as utilizing the k-means algorithm to

initialize the EM. Then, using these parameters, the E step calculates the probabilistic mix proportions for observations to belong to each cluster. The M step then re-estimates the parameters using the values computed in the E step. The E and M steps are reiterated until convergence.

The EM algorithm requires the number of clusters to be an input. Determination of the number of clusters was made in the following way. First, the Bayesian Information Criterion (BIC)<sup>[25,26]</sup> is calculated assuming the data consists of one to nine clusters, and the evaluation is made to verify if a sample is composed of one cluster or more than one cluster. The BIC is the value for the maximized log likelihood, considering the observations, parameters, data dimensions, and number of clusters (Eq. [1]). It is computed for a range of numbers, and the optimal number of clusters should be given by the largest BIC value.

$$\text{BIC} = 2\log\mathcal{L} - df - \log(n), \quad [1]$$

where  $\mathcal{L}$  is the likelihood function,  $\mathcal{L}(\theta|x) = P(x|\theta)$ , a function of the parameters given the outcomes, which is equal to the probability of the outcomes given the parameters,  $\theta$  is the parameters of the model, the mean and covariance of each cluster,  $x$  is the outcomes of the model,  $df$  is the degrees of freedom, and  $n$  is the number of observations.

Usually, the BIC tends to over fit the number of clusters when several clusters exist in inclusion data sets. Therefore, if the optimal number of clusters evaluated using the BIC is one, then the data are assumed to be single clustered, and if it was found to consist of more than one cluster, the Silhouette index<sup>[24]</sup> was employed to determine the optimal number of clusters. This index measures how similar an observation is to its own cluster compared to observations in other clusters. The silhouette index is calculated using Eq. [2]:

$$\text{Silhouette} = \frac{\sum_{i=1}^n S(i)}{n}, \quad \text{Silhouette} \in [-1, 1], \quad [2]$$

where  $n$  is number of observations,  $S(i) = \frac{b(i)-a(i)}{\max\{a(i);b(i)\}}$  the Silhouette coefficient for observation  $i$ ,  $a(i)$  is the average distance between the  $i$ th observation and observations within the same cluster, and  $b(i)$  is the average distance between the  $i$ th observation and observations in its nearest neighboring cluster.

The index is computed for a range of number of clusters, and the maximum value defines the optimal number of clusters. This procedure was employed because the Silhouette index cannot be defined for one cluster.

### III. MATERIALS AND METHODS

#### A. Materials

The laboratory-scale samples were taken from a series of experiments studying calcium treatment and reoxidation. Details of the experimental methods can be

found in Reference 27. The samples, referred to in this paper as sample 1 and sample 2, were prepared in an induction furnace. Electrolytic iron was melted in an MgO crucible enclosed in a graphite crucible with no slag addition. The iron was deoxidized with Al, Ca treated, and reoxidized with Fe<sub>2</sub>O<sub>3</sub> powder. The first sample, 1, was taken just before reoxidation, and the second, 2, right after reoxidation. The purpose of these lab samples was to create samples with controlled inclusion distributions to demonstrate the capabilities of PCA and clustering.

The industrial-scale data set was composed of four (4) samples, referred to in this paper as A, B, C, and D. All samples were taken from the same location on a near-finished component. The components were machined from continuously cast 4140 bar stock that was melted *via* an electric arc furnace, Al deoxidized, and Ca treated. Table I presents an overview of the sample chemistries, analyzed by spark optical emission spectroscopy (OES). Samples A and B were noted by the supplier to have better machinability compared to that of samples C and D. It should be noted that this ranking of machinability was relative, but sufficient to suggest there were differences between the samples that could be identified.

#### B. Methods

Automated inclusions analysis was carried out on mechanically polished samples using an SEM, utilizing a back-scattered electron detector equipped with an EDS analyzer. For the laboratory samples, a Phillips XL-30 SEM was utilized, with the accelerating voltage set to 10 kV, and a working distance of 10 mm. For the industrial samples, an FEI/Aspex Explorer SEM was used, the accelerating voltage was held at 20 kV, and the working distance ranged from 13.7 to 14.3 mm. The output raw data were prescreened to remove obvious outliers. A filter excluded any readings with Fe content larger than 75 pct, and more than 90 pct of the original data were retained for all samples. For the industrial samples, at least 1140 inclusions were measured per sample and the minimum detected inclusion size was 0.3 μm<sup>2</sup>. A summary of the number of inclusions measured and scanned area per sample is shown for the industrial samples in Table II.

The variables considered for PCA and clustering were the inclusion chemistries only, which were limited to the Mg, Ca, and Al contents for the laboratory samples, and the Mg, Ca, S, Al, Mn, Si, and Ti contents for the industrial samples. Other variables investigated in this

**Table I. Relevant Steel Chemistry (Mn, S, Al, and Ca in Weight Percent) for Industrial Samples**

Sample	Mn	S	Al	Ca
A	0.84	0.030	0.021	0.0006
B	0.85	0.026	0.022	0.0006
C	0.88	0.020	0.030	0.0027
D	0.88	0.030	0.026	0.0012



study included the areas, maximum diameters, and aspect ratios, which represent the size and morphology of inclusions; however, these variables were examined post-clustering.

All the statistical analysis performed in this study was carried out in RStudio,<sup>[28]</sup> an interactive interface for R, a free and open-source programming language for statistical computing and graphics.<sup>[29]</sup> The “MCLUST” and “NBCLUST”<sup>[30,31]</sup> packages were utilized for the EM algorithm and the Silhouette index, respectively. In addition, the “GGTERN” package<sup>[32]</sup> was used to plot the ternary diagrams.

#### IV. RESULTS

The results of the inclusion analyses are presented below, first for the laboratory samples and then for the industrial samples. Initially, inclusion chemistry distributions on ternary diagrams are given. All ternary diagrams presented in this study pertain to the mole fraction of the respective elements. Thereafter, PCA was best used to help compare data from several samples on a 2D scatter plot, while clustering was utilized to automatically define groups in each sample. For PCA, principal components were computed for all samples combined in a single data set, to enable the plotting of observations from all samples on a single 2D plot.

It should be noted that the methods of PCA and clustering are not examined here simply to differentiate between samples. Differences in inclusion populations would be expected based on the processing history of these samples (as was known for the laboratory-scale

samples and can be inferred from the chemistry of the industrial samples). This work presents the methods of PCA and clustering in the context of automated inclusion analysis and connects them to relatively small data sets that can be interpreted *via* conventional means. Possible further applications of these techniques are presented in Section V.

##### A. Laboratory Samples

The laboratory samples investigated in this study had controlled chemistries, and the variables considered for the analyses were confined to the Mg, Ca, and Al contents only. An overview of the inclusion chemistry distributions is presented as ternary diagrams in Figure 3. Sample 1 inclusions were composed of MgO-Al<sub>2</sub>O<sub>3</sub>-CaO inclusions, while sample 2 displayed two distinct inclusion groups—spinel (MgO-Al<sub>2</sub>O<sub>3</sub>) and solid calcium aluminates.

To get a visual comparison between samples on a single plot using PCA, both samples were combined into

**Table II. Industrial Sample Inclusion Count and Scanned Area**

Sample	No. of Inclusions	Scan Area (mm <sup>2</sup> )
A	6809	22.2
B	3309	11.1
C	1140	12.8
D	2686	11.9

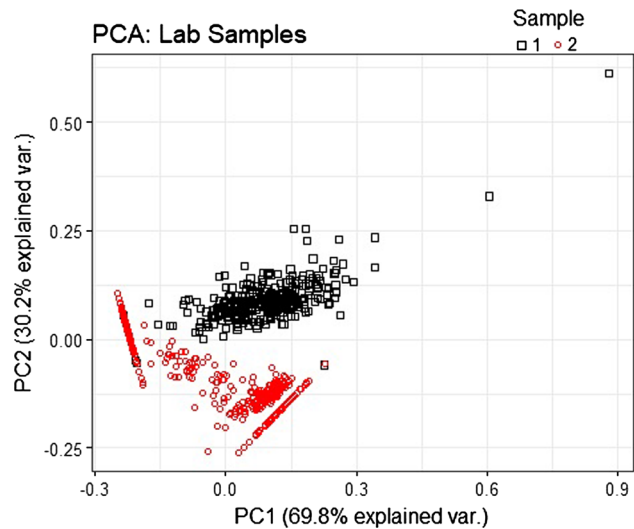


Fig. 4—Scatter plot of inclusions in laboratory samples, transformed to 2D using PCA.

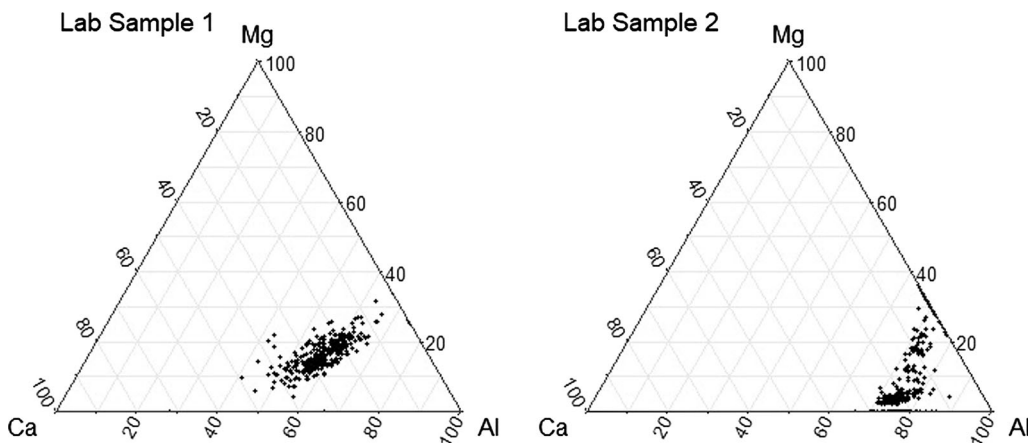


Fig. 3—Inclusion chemistry distribution of laboratory samples. Sample 1 on the left, and sample 2 on the right.

one data set. The transformation of the original data points to principal components was performed using the following equations:

$$PC1 = -0.585Mg - 0.201Al + 0.786Ca \quad [3]$$

$$PC2 = 0.570Mg - 0.791Al + 0.222Ca, \quad [4]$$

where the coefficients represent the relative loading of each of the original variables to each respective PC, which are unique to this data set. Figure 4 displays a 2D scatter plot of all the observations, where the horizontal axis is represented by the first principal component, PC1, and the vertical axis by the second, PC2, and observations are colored according to sample. As shown on the axes, PC1 represents 69.8 pct of the variance between the observations, and PC2 30.2 pct. Since only three variables were considered for PCA, the first two PCs were enough to represent the entire

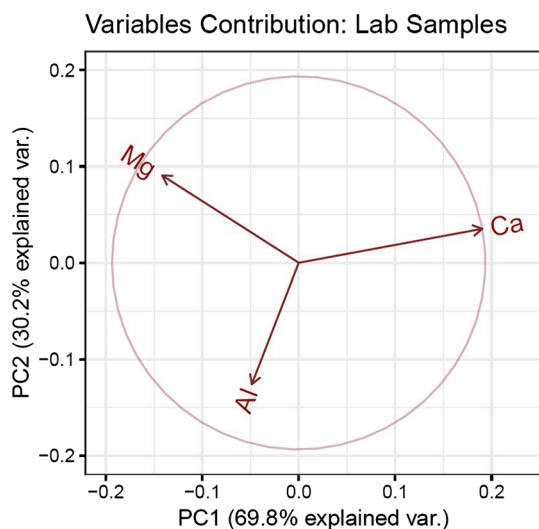


Fig. 5—Variable contribution to each PC, for lab samples.

variance in this data set (*i.e.*, the variance represented in PC3 was zero); however, this will not be the case when larger number of input variables are used. Observations from sample 1 are all relatively condensed in the same area, while sample 2 can be broadly broken down into 2 groups; in addition, overlap between samples is very minimal. This is expected from the lab samples, since the samples are composed of different inclusion types. Figure 5 provides an illustration of how the original variables contribute to the principal components. The arrows shown on the plot represent the contribution of the original variables to PC1 (by the magnitude in the horizontal direction) and PC2 (by the magnitude in the vertical direction).

To understand the distribution of data points, both Figures 4 and 5 need to be considered simultaneously. The red data points of sample 2 can be broadly broken down into 2 groups, one group with positive PC1 values (on the right), and the other with negative PC1 values (to the left of the plot). From Figure 5, it is clear that the former group contains higher Ca contents and the latter group has higher Mg contents, corresponding to solid calcium aluminate and spinel inclusions, respectively. Due to the nature of this data set, composed of only three input variables with some inclusions (*e.g.*, the spinels) comprising only two of the input variables, some of the data points form a straight line, as shown in Figure 4. These specific data points, or inclusions, having similar chemistries with slight deviations in cation mole fractions, correspond to the inclusions of lab sample 2 lying on the Mg/Al and Ca/Al axes of the ternary diagram (Figure 3); thus, they are not influenced by the third variable, and as a result they form a straight line on the scatter plot.

A summary of the cluster analysis, performed using the EM algorithm, is presented in Table III, for the lab samples. The table displays the average chemistry, the number and percentage of inclusions, and the average inclusion area, for each cluster. The table is ordered in decreasing cluster size, quantified by the number of inclusions per cluster.

Table III. Summary of Cluster Analysis, for Lab Samples

Sample 1		1 cluster				
Cluster	Mg	Al	Ca	Inc/Clus		Avg Inc Area ( $\mu\text{m}^2$ )
	16%	57%	27%	352	100%	0.66
Sample 2		2 Clusters				
Cluster	Mg	Al	Ca	Inc/Clus		Avg Inc Area ( $\mu\text{m}^2$ )
	6%	75%	19%	293	57%	1.07
	31%	69%	0%	222	43%	0.42

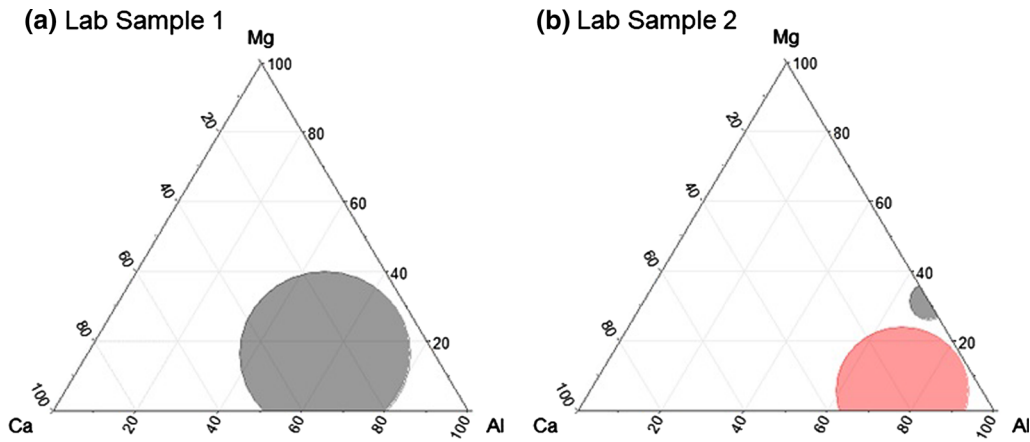


Fig. 6—Mg-Ca-Al cluster average ternary plots for laboratory samples. The center of the circles represents the cluster centroids, and the size of each circle is proportional to the total area fraction of inclusions in that particular cluster. (a) Sample 1, (b) Sample 2.

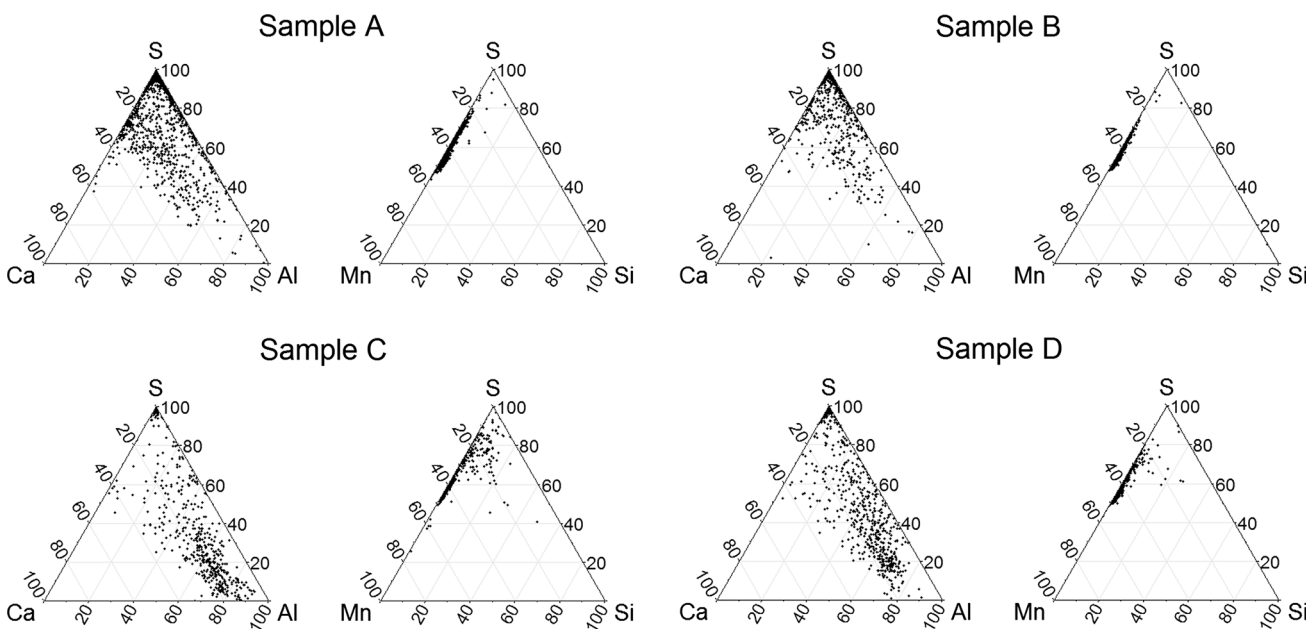


Fig. 7—Inclusion chemistry distributions of industrial samples, in the form of S-Ca-Al and S-Mn-Si Ternary Diagrams. Diagrams for Sample A shown on the top-left, sample B on the top-right, sample C on the bottom-left, and sample D on the bottom-right.

As stated earlier, initially the BIC was evaluated for each sample; if the BIC identified the optimal number of clusters as one, then the data were assumed to be single clustered; otherwise, the optimal number of clusters was computed using the Silhouette index. For the laboratory data set, the clustering algorithm identified one cluster was optimal for sample 1 and two clusters were optimal for sample 2.

The assignment of clusters can help simplify ternary diagrams, as shown in Figure 6. Here the average chemistry of each cluster is plotted as a semi-transparent circle, where the size of each circle is proportional to the cluster's total inclusion area. This was useful when comparing the relative amounts of inclusions in each cluster. Considering Figure 6(b) along with Table III, the two clusters identified in sample 2 pertain to calcium

aluminate and spinel inclusions, with calcium aluminate inclusions dominating the inclusion population.

### B. Industrial Samples

The inclusion chemistry distributions for the industrial samples are presented in Figure 7. Inclusions comprised Mn, S, Al, and Ca, with traces of Mg, Ti, and Si. Thus, analysis should strictly have accounted for all seven chemistry variables, but there could be a reasonable reduction to the four major components—Mn, S, Al, and Ca. This still required two ternary diagrams per sample, and in this case S-Ca-Al and S-Mn-Si ternaries were presented. Samples A and B were similar to each other and were predominately (Ca,Mn)S; inclusions in samples C and D were similar to

each other and had more calcium aluminate oxide inclusions.

As with the laboratory-scale samples, all four samples were consolidated into one large data set, for PCA. PC1 represents 88.1 pct of the variance, and PC2 represents 6.8 pct, and thus both PCs combined retain 94.9 pct of the original data's variance. The PCA transformation of the industrial data set was performed using the following equations:

$$\text{PC1} = -0.027\text{Mg} - 0.716\text{Al} + 0.002\text{Si} + 0.468\text{S} - 0.198\text{Ca} - 0.007\text{Ti} + 0.478\text{Mn} \quad [5]$$

$$\text{PC2} = 0.007\text{Mg} + 0.331\text{Al} - 0.008\text{Si} - 0.391\text{S} - 0.569\text{Ca} - 0.014\text{Ti} + 0.643\text{Mn}. \quad [6]$$

The coefficients represent the loading of each of the original variables to the PCs. PCA results are presented in Figure 8.

Figure 8 shows that samples A and B have similar chemistry distributions, and samples C and D display a similar trend, with sample C skewing more to the negative values of PC1. Samples C and D appear to differ from A and B because of the large number of data points at  $\text{PC1} < 0$ . Figure 9 provides an illustration of how the original variables contribute to the principal components. As expected, inclusion chemistries were dominated by Al, Mn, S, and Ca, while the Mg, Ti, and Si contents are negligible.

Thus, from the results presented in PCA, it can be inferred that samples A and B were similar to each other, and C and D were similar to each other.

Inclusions in samples A and B contained primarily Mn and S, and Ca and Al contents were higher in samples C and D.

The cluster analysis results for the industrial data set are presented in Table IV and Figure 10. Two clusters were identified for both samples A and B: one cluster with predominately MnS inclusions and the other with complex inclusions, (Ca,Mn)S + calcium aluminate oxides. The largest cluster in both samples corresponded to the MnS inclusions. In samples C and D, four and three clusters were identified, respectively. Again MnS

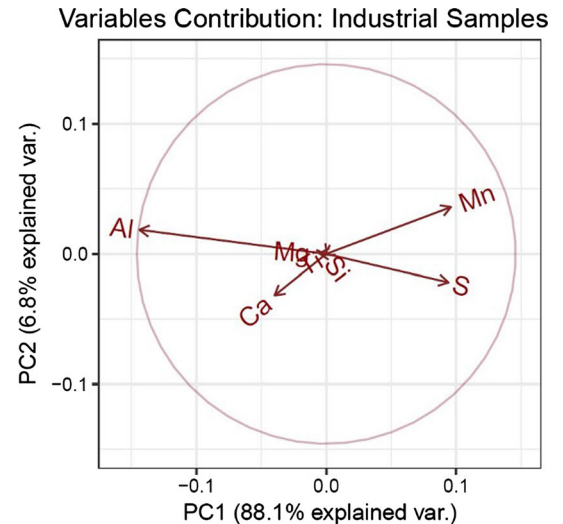


Fig. 9—Variable contribution to each principal component, for industrial samples.

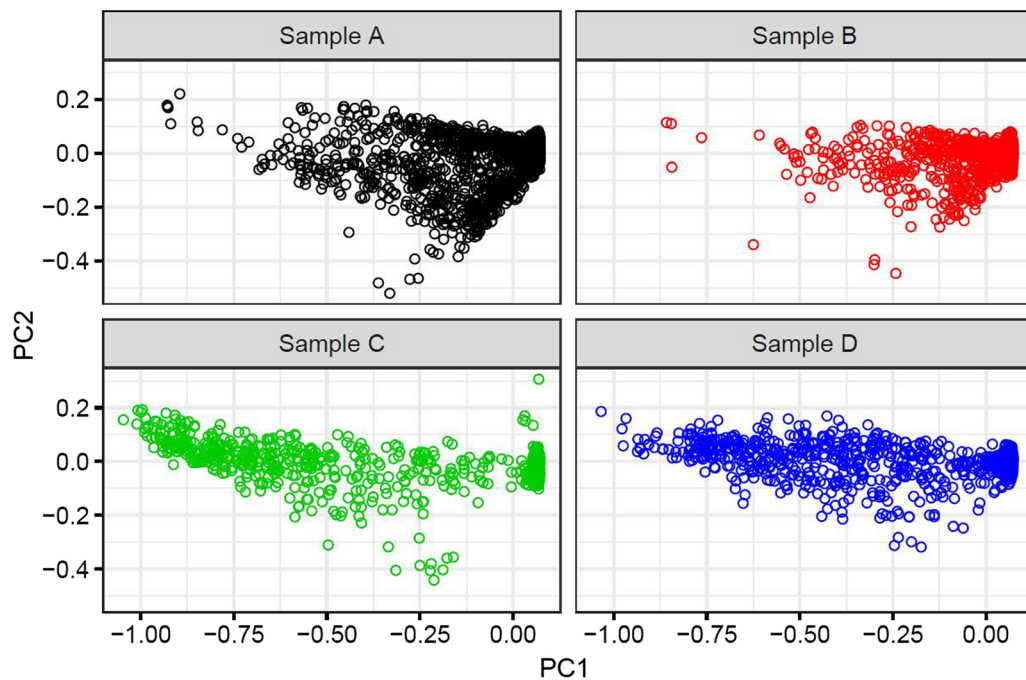


Fig. 8—PCA scatter plots of industrial samples. Sample A shown on the top-left panel, sample B on the top-right panel, sample C on the bottom-left panel, and sample D on the bottom-right panel.



Table IV. Summary of Cluster Analysis, for Industrial Samples

**Sample A 2 clusters**

Number Density: 307 /mm<sup>2</sup> Area Density: 899 ppm

Cluster	Mg	Al	Si	S	Ca	Ti	Mn	Inc/Clus		Avg Inc Area (μm <sup>2</sup> )
	0%	0%	1%	54%	0%	0%	45%	5687	83%	2.92
	2%	12%	0%	46%	8%	0%	32%	1122	17%	3.00

**Sample B 2 clusters**

Number Density: 299 /mm<sup>2</sup> Area Density: 1229 ppm

Cluster	Mg	Al	Si	S	Ca	Ti	Mn	Inc/Clus		Avg Inc Area (μm <sup>2</sup> )
	0%	0%	0%	56%	0%	0%	43%	2721	82%	4.28
	1%	10%	1%	48%	6%	0%	34%	588	18%	3.30

**Sample C 4 clusters**

Number Density: 89 /mm<sup>2</sup> Area Density: 432 ppm

Cluster	Mg	Al	Si	S	Ca	Ti	Mn	Inc/Clus		Avg Inc Area (μm <sup>2</sup> )
	1%	47%	0%	26%	13%	0%	13%	443	39%	7.07
	0%	0%	0%	58%	0%	0%	41%	400	35%	3.16
	0%	0%	0%	58%	0%	0%	41%	261	23%	2.71
	5%	33%	3%	27%	14%	7%	12%	36	3%	11.62

**Sample D 3 clusters**

Number Density: 227 /mm<sup>2</sup> Area Density: 757 ppm

Cluster	Mg	Al	Si	S	Ca	Ti	Mn	Inc/Clus		Avg Inc Area (μm <sup>2</sup> )
	0%	0%	0%	55%	0%	0%	43%	1157	43%	2.76
	0%	0%	0%	55%	0%	0%	43%	950	35%	2.69
	0%	34%	0%	33%	10%	1%	22%	579	22%	5.57

and complex (Ca,Mn)S + calcium aluminate oxide inclusions were identified. However, in this case the (Ca,Mn)S + oxide inclusion clusters were larger. There was also some overlap of clusters in these samples, suggesting there was opportunity for additional optimization of the analysis method.

**V. DISCUSSION**

In this work, PCA and cluster analysis were employed to analyze a controlled laboratory-scale data set and a relatively large industrial inclusion data set. The lab-scale data set was limited to two samples, where

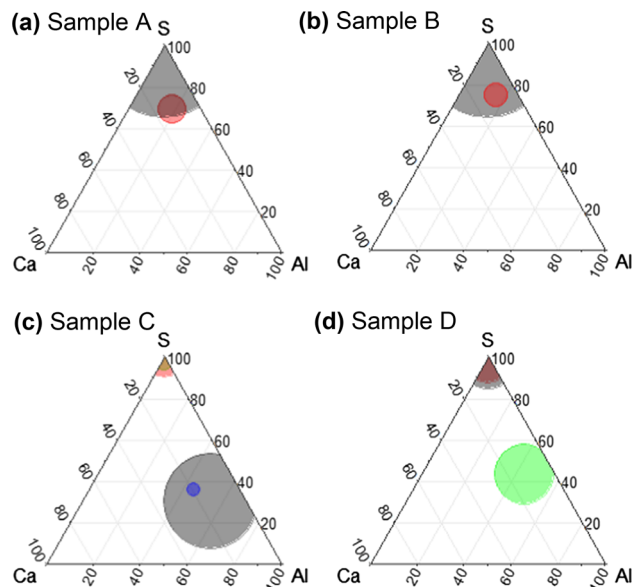


Fig. 10—S-Ca-Al cluster average ternary plots for industrial samples. The center of the circles represents the cluster centroids, and the size of each circle is proportional to the total area fraction of inclusions in that particular cluster. (a) Sample A, (b) Sample B, (c) Sample C, (d) Sample D.

both samples had well-defined inclusion compositions. The analyses were carried out on the controlled samples in order to examine their applicability. PCA enabled the visualization of the entire inclusion population from both samples on a single plot, and it is clear from this plot that inclusions in sample 1 have relatively similar chemistry, while inclusions in sample 2 can be roughly divided into two inclusion types. This was also asserted by the cluster analysis results, identifying one cluster in sample 1 and two clusters in sample 2. Based on average cluster composition, sample 1 formed predominantly liquid  $\text{MgO-Al}_2\text{O}_3\text{-CaO}$  inclusions, and the two clusters in sample 2 were identified as spinel and solid calcium aluminate inclusion clusters. The samples were specifically selected to demonstrate whether the cluster analysis would identify one and two clusters for samples 1 and 2, respectively, using the BIC and Silhouette index, and the results presented verified this presumption.

With respect to the industrial samples, the data were measured from four samples, A and B, which exhibited good machinability, and C and D, which exhibited poor machinability. A complete representation of the inclusion population required numerous ternary diagrams, while PCA and clustering could condense this information. The results showed that the two sets of samples were indeed different. Samples A and B contained predominately MnS inclusions and C and D had larger amounts of complex  $(\text{Ca,Mn})\text{S}$  + calcium aluminate oxide inclusions. Below some observations on the data analysis methods and the results are discussed.

Using PCA, an initial comparison between samples and inclusion populations was visualized in one 2D plot. The ability to visualize high dimensional data is a useful application of PCA. PCA has also been used as a method to explore data sets and reveal the existence of

potential clusters of data points. For the purpose of this study, observations from all samples were consolidated into a single data set, since in both cases, the laboratory and industrial data sets, samples were processed to produce the same steel chemistry. Otherwise, using PCA to compare disparate samples on the same plot would not necessarily be beneficial, since a scatter plot of the first 2 PCs will only display a dissimilarity between observations from each sample, which was already expected. For example, performing PCA on an Al-killed sample and a Si-Mn killed sample would be trivial since these samples are known to be different and the PCA results will only highlight the expected difference. PCA would be most useful in reducing the number of diagrams needed for visualization, particularly if sulfide or nitride inclusions would be considered along with oxide inclusions. If sufficient data are available, PCA could be a process monitoring tool,<sup>[33,34]</sup> where new inclusion populations could be benchmarked against an existing baseline.

The cluster analysis performed was successful in condensing the large data sets into a few points that were representative of the inclusion chemistries. In some cases, clusters were not uniquely defined in a sample and sometimes two different clusters had similar chemistries, such as the red and green clusters of sample C and the black and red clusters of sample D. This was a consequence of the automated nature of the analysis. The clustering algorithm was generally successful at recognizing and classifying the inclusion chemistry groups. In addition, the algorithm performed well in assessing the possibility of having single-clustered data, as shown in the results presented for lab sample 1.

Once the industrial data were grouped by chemistry, additional analyses were performed to examine the relationship between chemistry, average inclusion area, and average inclusion aspect ratio. To examine these correlations, each cluster from Table IV was classified as either a MnS cluster or complex  $((\text{Ca,Mn})\text{S} + \text{oxide})$  inclusion cluster based on its chemical composition. A summary of the average inclusion area and aspect ratio per cluster is provided in Figure 11, along with representative BSE images of inclusions in each cluster.

The MnS inclusions in all samples had similar average areas. The average inclusion area of the complex oxides was greater for samples C and D. The average aspect ratios (ratio of the longest 2D dimension to its perpendicular width) of the MnS inclusions were higher in samples C and D, indicating that they were slightly elongated. Although elongated, these inclusions were still small and so would not likely have the detrimental effect of stringer MnS on mechanical properties.

In summary, all four samples contained essentially two different inclusion populations—MnS and complex  $(\text{Ca,Mn})\text{S}$  + calcium aluminate oxide inclusions. In samples C and D, the complex inclusions were more numerous and they were larger than the complex inclusions in samples A and B. The MnS inclusions were globular in samples A and B, while they were more elongated in samples C and D. These differences in inclusion population likely affected the part

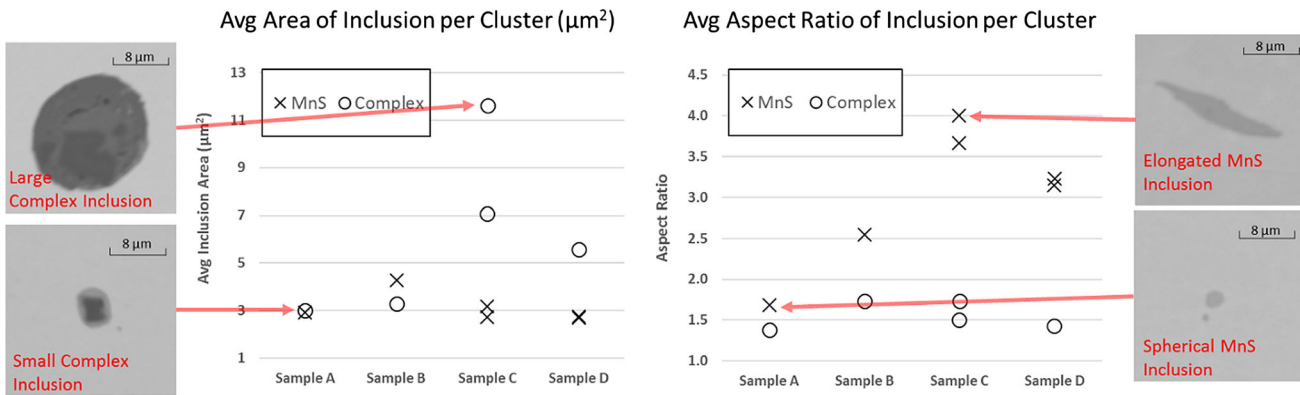


Fig. 11—Relationship between average inclusion areas and aspect ratios in each cluster. The BSE images represent what an average inclusion would look like in the respective cluster.

machinability, as it is well known that inclusions play an important role in machinability.<sup>[35]</sup>

Numerous studies have been conducted on the relationship between inclusions, steel processing, and machinability (for a recent review see Ref. [20]). It was not possible to develop many connections between the inclusions, processing, and machinability in this work. The steels in this work were 4140 alloy steel with approximately 300 ppm S and all samples were Al deoxidized and Ca treated. Generally, this composition and processing should lead to good castability, machinability, and properties.<sup>[36,37]</sup> The samples for this study were provided in the partially finished state and details of the refining and Ca treatment processes were not available; furthermore, significant changes to inclusion populations can occur upon the solidification of the samples.<sup>[38]</sup> The main purpose of this study was to introduce the use of PCA and clustering methods to analysis of inclusion distributions.

The objective of this study was to demonstrate alternative methods for classifying and analyzing inclusion populations. The method of PCA was shown to produce a 2D representation of multivariate data, from which differences between samples could be noted. Cluster analysis provides an automated method for classifying inclusion compositions without user-defined rules. The clusters can then be analyzed individually. They are amenable to much larger data sets than the four samples investigated here and future efforts will be devoted to application of these and other techniques to improve the insights that can be gained from automated non-metallic inclusion analysis data.

## VI. CONCLUSIONS

In this study, the differences in non-metallic inclusion population between two sets of samples were analyzed with PCA and cluster analysis. The first data set, a pair of laboratory-prepared samples with well-defined and controlled inclusion chemistries, was utilized to validate the statistical techniques. And the second data set, four industrial-scale samples, grade 4140, Al deoxidized, and Ca treated: two samples (A and B) were noted to have

better machinability, while the other two samples (C and D) were noted to have poorer machinability. The following observations were noted:

- PCA reduced several chemistry variables to a 2D plot where differences were noted between samples.
- Cluster analysis, using the BIC and Silhouette index to define the number of clusters and the expectation-maximization algorithm to perform the clustering, automatically classified inclusions into groups. Essentially two groups: MnS and (Ca,Mn)S + calcium aluminate oxide inclusions, for the industrial samples. This was done without any user-specified composition ranges.
- For the laboratory-scale samples, both PCA and clustering generated promising results. PCA enabled the visualization of the multivariate inclusion data set on simple 2D plots, providing an initial comparison between inclusion distributions of various samples. And cluster analysis was successful in identifying the inclusion groups present in each sample.
- Considering the industrial data set, once the inclusions were clustered it was found that samples C and D contained a higher number of larger (Ca,Mn)S + oxide inclusions compared to samples A and B, which contained higher numbers of smaller MnS inclusions. The differences in machinability were attributed to the differences in the inclusion populations.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the member companies of the Center for Iron and Steelmaking Research and the use of the Materials Characterization Facility at Carnegie Mellon University, supported by Grant MCF-677785.

## REFERENCES

1. L. Zhang and B.G. Thomas: *ISIJ Int.*, 2003, vol. 43, pp. 271–91.
2. H.V. Atkinson and G. Shi: *Prog. Mater. Sci.*, 2003, vol. 48, pp. 457–520.

3. W.M. Garrison and A.L. Wojcieszynski: *Mater. Sci. Eng. A*, 2007, vol. 464, pp. 321–29.
4. W.M. Garrison and A.L. Wojcieszynski: *Mater. Sci. Eng. A*, 2009, vol. 505, pp. 52–61.
5. A. Gupta, S. Goyal, K.A. Padmanabhan, and A.K. Singh: *Int. J. Adv. Manuf. Technol.*, 2015, vol. 77, pp. 565–72.
6. J. Lankford: *Int. Met. Rev.*, 1977, vol. 22, pp. 221–28.
7. Y. Murakami: *Metal Fatigue: Effects of Small Defects on Non-metallic Inclusions*, Elsevier, Amsterdam, 2002.
8. A.W. Cramb: in *High Purity, Low Residual, and Clean Steels*, C.L. Briant, ed., Marcel Dekker, New York, 1999, pp. 49–91.
9. S.R. Story, T.J. Piccone, R.J. Freuhan, and M. Potter: *Iron Steel Technol.*, 2004, vol. 9, pp. 163–69.
10. P. Kaushik, J. Lehmann, and M. Nadif: *Metall. Mater. Trans. B*, 2012, vol. 43, pp. 710–25.
11. H.P. Lentz, M.S. Potter, and G.S. Casuccio: in *ISTech Proceedings*. 2017.
12. D. Tang and P.C. Pistorius: in *AISTech Proceedings*. 2015.
13. A. Harada, N. Maruoka, H. Shibata, M. Zeze, N. Asahara, F. Huang, and S. Kitamura: *ISIJ Int.*, 2014, vol. 54, pp. 2569–77.
14. N. Verma, P.C. Pistorius, R.J. Fruehan, M. Potter, M. Lind, and S.R. Story: *Metall. Mater. Trans. B*, 2011, vol. 42, pp. 711–19.
15. N. Verma, P.C. Pistorius, R.J. Fruehan, M. Potter, M. Lind, and S.R. Story: *Metall. Mater. Trans. B*, 2011, vol. 42, pp. 720–29.
16. J.H. Shin, Y. Chung, and J.H. Park: *Metall. Mater. Trans. B*, 2017, vol. 48, pp. 46–59.
17. E.B. Pretorius, H.G. Oltmann, and B.T. Schart: in *AISTech Proceedings*. 2013.
18. S.R. Story and R.I. Asfahani: *Iron Steel Technol.*, 2013, vol. 9, pp. 86–99.
19. M. Harris, O. Adaba, S. Lekakh, R. O'Malley, and V.L. Richards: in *AISTech Proceedings*. 2015.
20. N. Anmark, A. Karasev, and P.G. Jonsson: *Materials*, 2015, vol. 8, pp. 751–83.
21. H. Abdi and L.J. Williams: *WIREs Comp. Stat.*, 2010, vol. 2, pp. 433–59.
22. M. Scholz: Approaches to analyse and interpret biological profile data. PhD thesis. 2006, Potsdam University: Potsdam.
23. D. Binu: *Expert Syst. Appl.*, 2015, vol. 42, pp. 5848–59.
24. P.J. Rousseeuw: *J. Comput. Appl. Math.*, 1987, vol. 20, pp. 53–65.
25. C. Fraley, and A.E. Raftery, Technical Report No. 597 University of Washington, Seattle, 2007.
26. C.M. Bisho: *Pattern Recognition and Machine Learning*, Springer, New York, 2006, pp. 216–18.
27. J. Tan and B.A. Webler: *AIST Trans.*, 2017, vol. 14, pp. 172–79.
28. R. Team, RStudio: Integrated Development Environment for R, (RStudio, Inc.Place, Published, 2016), <http://www.rstudio.com/>.
29. R.D.C. Team, R: A language and environment for statistical computing, (R Foundation for Statistical Computing Place, Published, 2008), <http://www.R-project.org>.
30. C. Fraley, A.E. Raftery, L. Scrucca, T.B. Murphy, and M. Fop, Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation, Package 'mclust', 2016, <http://www.stat.washington.edu/mclust/>. Accessed 23 May 2016.
31. M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs: *J. Stat. Softw.*, 2014, vol. 61, pp. 1–14.
32. N. Hamilton, An Extension to 'ggplot2', for the Creation of Ternary Diagrams, Package 'ggtern', 2016, <http://www.ggtern.com>. Accessed 21 June 2016.
33. T. Kourti and J.F. MacGregor: *Chemom. Intell. Lab. Syst.*, 1995, vol. 28, pp. 2–21.
34. J.V. Kresta, J.F. MacGregor, and T.E. Marlin: *Can. J. Chem. Eng.*, 1991, vol. 69, pp. 35–47.
35. R. Kiessling: *J. Met.*, 1969, vol. 21, pp. 47–54.
36. L.E.K. Holappa and A.S. Helle: *J. Mater. Process. Technol.*, 1995, vol. 53, pp. 177–86.
37. R.V. Väinölä, L.E.K. Holappa, and P.H.J. Karvonen: *J. Mater. Process. Technol.*, 1995, vol. 53, pp. 453–65.
38. L. Zhang: *JOM*, 2013, vol. 65, pp. 1138–44.