Data Science Techniques, Assumptions, and Challenges in Alloy Clustering and Property Prediction

Madison Wenzlick, Osman Mamun, Ram Devanathan, Kelly Rose, and Jeffrey Hawk

Submitted: 21 October 2020 / Accepted: 24 October 2020 / Published online: 5 January 2021

Data analytics methods have been increasingly applied to understanding materials chemistry, processing due to the manufacturing approach, and uni-axial and cyclic property relationships in the highly complex space of alloy design. There are several benefits to applying data analytics to this space, including the ability to manage non-linearities in the responses of the alloy attributes and the resulting mechanical properties. However, key difficulties in applying and understanding the results of data analytics include the often lack of reported assumptions and data processing steps necessary to improve interpretation and reproducibility in derived results. In this work, the methods used to generate clustering and correlation analyses for experimental 9% Cr ferritic-martensitic steel data were investigated and the resulting implications for mechanical property predictions were assessed. This work uses principal component analysis, partitioning around medoids, t-SNE, and k-means clustering to investigate trends in composition, processing and microstructure information with creep and tensile properties, building on work done previously using a smaller version of the same dataset. The initial assumptions, preprocessing steps and methods are investigated and outlined in order to depict the fine level of detail required to convey the steps taken to process data and produce analytical results. The variations in the resulting analyses are explored due to the influence of new and more varied data.

Madison Wenzlick, National Energy Technology Laboratory, 1450 Queen Avenue SW, Albany, OR 97321; and Leidos Research Support Team, 1450 Queen Avenue SW, Albany, OR 97321; **Osman Mamun** and **Ram Devanathan**, Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, WA; **Kelly Rose** and **Jeffrey Hawk**, National Energy Technology Laboratory, 1450 Queen Avenue SW, Albany, OR 97321. Contact e-mail: Jeffrey.Hawk@netl.doe.gov.

Keywords clustering, data analytics, materials informatics, metadata, 9% Cr ferritic-martensitic steel

1. Introduction

Data science, including machine learning (ML), has become an increasingly useful tool for assessing and modeling data. However, one commonly described difficulty in improving the scope and application of data science in physical sciences is the lack of training of data scientists in domain science (Ref 1, 2) and the increasing need for domain scientists to be well-versed in data science (Ref 3). Additionally, the challenges in communicating and interpreting the results of large datasets and assumptions made during data analysis of these sets are a barrier for widespread adoption and application of data science methods and tools. In published manuscripts, the level of detail in the reported methods and assumptions required to make the results interpretable as well as reproducible and usable is difficult and often much nuanced. Sufficient detail about the assumptions, methodologies, and actual data used is often not included in the body of the work or in the supplemental information for the reader to draw the given conclusions. In materials development research articles, descriptions of processing, heat treatment, testing and analysis standards, and other significant attributes are often insufficient as reported or not provided at all (Ref 4). In terms of data science, many assumptions and processing steps are required before the resulting analysis can be generated and visualized. Details of these steps may be assumed by the author or left out entirely, or at best inferred or discovered from the previous work. This lack of requisite detail creates challenges in interpreting and reproducing the analytical results.

Through the US Department of Energy Fossil Energy (DOE FE) eXtremeMAT (hereafter XMAT) program, a database of physical and mechanical properties of engineering materials for fossil energy power generation (but applicable to many engineering fields) has been under creation since 2018. The database initially was developed to assist internal FE research (National Energy Technology Laboratory, hereafter NETL) both in the design of advanced heat resistant 9% Cr ferritic-martensitic steels for use at temperature up to 650°C and in determining the end of life for these steels in creep. It was subsequently realized that while information on these materials was very mature, the variability in the data was insufficient to show broad trends or differentiate trends due to minor chemistry or processing changes. Thus, the initial 9% Cr ferritic-martensitic steel database was expanded to include not only 9% Cr up to the 9% Cr ferritic-martensitic steels as well as ferritic steels with Cr content greater than 12% (by weight). More recently, the XMAT program began designing advanced austenitic stainless steels, based on both chromia formers (traditional FE tubing and piping alloys used in the boiler section of the power plant) as well as new aluminum oxide forming compositions (to achieve higher operating temperatures based on the greater stability of the oxide). As such, the need for a broader range of alloy information has resulted in addition of gamma matrix alloys (i.e., a face centered cubic (FCC) matrix crystal structure) to be gathered into the database. Given that temperatures greater than 700°C are envisioned for many of the components in these power plants, expansion of the database to include precipitate strengthened nickel superalloys is currently underway. Each iteration of the database to include new materials has resulted in changes to the initial methodology of what should be included in the database and the information that should be tested for. This continues to be

an ongoing and evolving process as internal testing has continued to add property information to the database as well as ongoing efforts to pull data into the database from less accessible external sources. This database has been used in many analytical efforts to draw conclusions using data on the effect of composition (actual chemistry of the major elements as well as the minor ones down to parts per million), properties (static and dynamic where existing), and general microstructure features on material behavior for general alloy classes (Ref 5-11). To improve understanding and prediction based on the given data, the results of these past analyses, therefore, should be included and expanded upon in later works. As such, an emphasis was placed on leveraging the analytical work performed using prior editions of the XMAT database to develop an understanding of challenges in reproducibility of data science analyses.

Specifically, several research initiatives have used the initial (smaller) versions of the XMAT 9% Cr ferritic-martensitic steels creep and tensile dataset to produce observations of trends relating composition and processing to performance. Additionally, this information has been examined to identify the critical attributes of this alloy class. These analyses were produced both by in-house researchers as well as external teams and used various data preprocessing methods, clustering analyses, and linear (as well as nonlinear) regression techniques (Ref 5-11). As part of the XMAT data science task, it was essential to explore and include the results and insights of prior analytical efforts on similar (but expanded) data as an initial point of reference for future work. Additional clustering techniques were investigated for application to these analyses as well to gain new insights into the data, including other methods of interpreting reduced dimension space and applying clustering methods to the resulting data (Ref 12). Additionally, these analyses form a comparison metric for subsequent analyses supplemented with new data, as the 9% Cr ferritic-martensitic steel dataset has continuously been updated to include test data generated since the previous analyses were completed. This allows the previous assumptions used in producing those conclusions to be investigated and to ultimately find if these initial conclusions are generalizable to larger datasets with additional data. Incorporating the results of these prior works also enables analytical efforts to gain the most insight into the data and the techniques used to analyze and interpret the data.

The specific methods investigated here include several commonly used dimensionality reduction, clustering, and correlation tools. Dimensionality reduction is used to simplify the number of parameters, or attributes, that are used to understand the response variable in a system (Ref 13). This is useful when there is a large number of attributes that are needed in order to model the dependent variable, which in this case is a specific mechanical property of interest. Mechanical properties that have been investigated include ultimate tensile strength (UTS), yield stress (YS), and the time to creep failure. (Note: The time to creep failure is a directly measured value, while many of the equations used to model creep of engineering alloys make use of physically based constants in a specific parametric equation to lump these physical features that are not typically easy to measure.) Dimensionality reduction can additionally help with comparisons between different groups of data points (representing alloy class subsets), or clusters, and can capture variation among several attributes in one reduced dimension, which makes the comparison easier to visualize.

Table 1 Attributes, units, and value ranges for variables included in the XMAT 9% Cr ferritic-martensitic steel dataset

Field	Units	Range	Field	Units	Range	Field	Units	Range
Alloy ID	...		Austenite grain size number	#		Low cycle fatigue test temperature	°C	23.9-700
Iron (Fe)	wt. %	78.611-90.2226	Austenite grain size	# grains/mm ²	0-0	Low cycle fatigue test stress concentration factor	...	1-1
Carbon (C)	wt. %	0.004-0.25	Tensile test temperature	°C	23.9-800	Low cycle fatigue test waveform (sinusoidal (1), triangular (2), etc.)	...	2-2
Chromium (Cr)	wt. %	7.915-17.55	0.2% yield stress	MPa	51-1227	Low cycle fatigue test frequency	Hz	0.06-0.6
Manganese (Mn)	wt. %	0.05-0.98	Ultimate tensile strength	MPa	107-1821.05	Low cycle fatigue test total strain amplitude	MPa	0.00074-0.01749
Silicon (Si)	wt. %	0.01-0.86	Elongation to failure	%	8.2-105	Low cycle fatigue test maximum strain amplitude	MPa	0.00001-0.01354
Nickel (Ni)	wt. %	0.01-2	Reduction in area	%	11.6-99	Low cycle fatigue test minimum strain amplitude	MPa	87-843
Cobalt (Co)	wt. %	0.008-8.226	Charpy Impact Value	J/cm ²	180-311	Low cycle fatigue test R-ratio	...	-1-1
Molybdenum (Mo)	wt. %	0.01-2.99	Vickers Hardness	Hv 196 N	155-312	Low cycle fatigue test A-ratio	...	0-0
Tungsten (W)	wt. %	0.005-3.93	Density	g/cm ³	7.675-7.734	Low cycle fatigue test cycles to crack initiation	...	0-0
Niobium (Nb)	wt. %	0.005-0.43	Elastic/Young's Modulus	kN/mm ²	133-239.6	Low cycle fatigue test cycles to failure	...	140-7200000
Aluminum (Al)	wt. %	0.001-0.057	Shear Modulus	kN/mm ³	84.98-103.56	High cycle fatigue test temperature	°C	24-600
Phosphorous (P)	wt. %	0.001-0.03	Poisson's Ratio	...	0.1569-0.2806	High cycle fatigue test stress concentration factor	...	1-3
Copper (Cu)	wt. %	0.002-0.97	Creep test temperature	°C	450-750	High cycle fatigue test waveform (sinusoidal (1), triangular (2), etc.)	...	1-1
Titanium (Ti)	wt. %	0.001-0.1025	Creep stress	MPa	10-517	High cycle fatigue test frequency	Hz	125-125
Tantalum (Ta)	wt. %	0.0003-0.361	Creep rupture time	h	0.6-297588.7	Maximum stress	MPa	117.5-390
Hafnium (Hf)	wt. %	0.16-0.16	Elongation to failure	%	0.676-146	Minimum stress	MPa	-1750-117.5
Rhenium (Re)	wt. %	0.0003-0.6	Reduction in area	%	0-98	Stress Amplitude	MPa	235-780
Vanadium (V)	wt. %	0.01-0.3	Minimum creep rate	%/h	0.00000107-1.6	Mean Stress	MPa	0-0
Boron (B)	wt. %	0.0003-0.051	Time to reach 0.1% creep strain	h	0.1-20067	High cycle fatigue test R-ratio	...	-1-1
Nitrogen (N)	wt. %	0.001-0.1651	Time to reach 0.15% creep strain	h	0.2-1300	High cycle fatigue test A-ratio	...	0-0
Oxygen (O)	wt. %	0.0014-0.072	Time to reach 0.2% creep strain	h	0.2-17806	High cycle fatigue test cycles to crack initiation	...	0-0
Sulfur (S)	wt. %	0.0003-0.023	Time to reach 0.25% creep strain	h	0.1-4400	High cycle fatigue test cycles to failure	...	51000-100000000
Zirconium (Zr)	wt. %	0.001-0.001	Time to reach 0.5% creep strain	h	0.2-44900	Stress relaxation test temperature	°C	500-650
Yttrium (Y)	wt. %	0-0	Time to reach 1.0% creep strain	h	1-86300	Stress relaxation test total strain	%	0.1-0.25
Homogenization heat treatment	0/1	0-1	Time to reach 2.0% creep strain	h	1-150000	Stress relaxation test initial stress	MPa	145-431
Normalization or austenization temperature	°C	815-1221	Time to reach 5.0% creep strain	h	2-163000	Stress relaxation test time running time	h	0.1-8760
Temper heat treatment 1	°C	550-860	Time to tertiary creep	h	1-189000	Stress relaxation test residual stress at specified running time	MPa	1-414
Temper heat treatment 2	°C	610-740						
Temper heat treatment 3	°C	630-690						
Temper heat treatment 4	°C	0-0						

Clustering is another useful method of understanding the structure of a dataset and is used to formulate subgroupings within the dataset of data points with similar attributes (Ref 14, 15). The resulting clusters can then be used for modeling the variable of interest. Clustering the dataset can aid in reducing the variation within a model, as the response variable is restricted to a segment of the dataset with the most similarity. The relationships between the independent and dependent variables can then be determined with reduced noise in the data. The end result is better understanding of the alloy subsets as defined by dataset clusters.

In this work, the techniques utilized include principal component analysis (PCA) (Ref 16), partitioning around medoids (PAM) or k-medoids clustering (Ref 15), t-distributed stochastic neighbor embedding (t-SNE) analysis (Ref 17), and k-means clustering (Ref 18). These clustering and dimensionality reduction tools were applied to the XMAT 9% Cr ferritic-martensitic dataset in order to evaluate (1) the assumptions inherent in each calculation and resulting analysis and (2) the change in results due to the addition of new data to the dataset. Further, in order to explore cluster extents in the larger dataset with increased variation, k-means analysis was applied to the reduced dimension space resulting from t-SNE. While the exact location in reduced dimension space is not directly correlated with the data attributes, the relative locations of the data are meaningful, which enables k-means analysis to be applied to the data to identify clusters.

2. Methods

2.1 9% Cr Ferritic-Martensitic Steel Data

The XMAT database is comprised of alloy data from several sources, including NIMS, NETL in-house research, industry data, as well as data obtained from open-source resources and literature (Ref 19-32). The alloys included in this work specific to the database include ferritic-martensitic steels. An overview of the attributes included in the dataset is provided in Table 1. Additional attributes were included in the metadata for each source, including data on alloy designation, ingot weight, product form and dimensions, and the test method standards. This data included 121 unique compositions, with 140 unique composition and heat treatment temperature combinations. This is a significant increase in data from previous versions of the dataset which contained 79 unique compositions with tensile data (Ref 8) or 56 sample compositions (Ref 6).

2.2 Data Preprocessing and Assumptions

Analyses on this data are focused on determining the relationships between the elements in the steel chemistry, the manufacturing approach used to include melting, deformation processing and heat treatments, and general starting microstructure prior to creep and tensile response, as well as low and high cycle fatigue and stress relaxation. The analyses described herein focus specifically on the data processing and analysis steps used to determine several underlying trends in the 9% Cr ferritic-martensitic steel data, primarily using dimensionality reduction and clustering analyses.

There are several challenges in reassessing and improving analyses presented in the literature. In terms of data science, these challenges encompass initial specific data assumptions and methods used to curate and process data, including methods used to fill in missing data, address outliers, and exclude and include data points of varying parameters. Additionally, in terms of methods used to visualize and analyze data, the correct program and function are needed, as well as specific data processing for the function and the assumptions and specific parameters used within the functions. Correctly assembling all of these details is necessary in recreating the analyses as they are shown, since there are a significant number of decisions that the analyst must make in order to manipulate data to be ready for subsequent analysis and to perform the actual work.

The XMAT data are complex and require processing and cleaning before the dataset can be used in analysis. The 9% Cr ferritic-martensitic steel dataset used in this work encompasses continuous numerical, discrete numerical, and categorical data types. Therefore, the different variable types must be correctly processed in order to be in the correct format for analysis. For this analysis, the data were limited to data points that include the tensile test results. Additionally, the data were cleaned, and exceptions were removed. For example, there are some fields which have mixed numerical and categorical values, such as the ingot weight being designated as “CC” or continuously cast. These exceptions, as well as non-blank characters (i.e., “-”), were changed to NaN. After this step, missing composition values were replaced with 0, missing heat treatment values were replaced with 21°C, missing homogenization information was filled in with 0, and all other missing values were replaced with 0. These two steps ensured that each attribute had a uniform variable type as well as no missing values. Further, while the data are structured in tables, the data must be manipulated in order to generate tidy datasets which are optimized for use in clustering, regression, and classification (Ref 33). Manipulation steps to ensure this include making sure

Table 2 Data processing attributes and metadata

Filling in missing values	Was the data filled in using interpolation? 0? Or another value? Which attributes did this apply to?
Scaling	How was the dataset scaled, if at all?
Categorical data	Were categorical data included? Were they binarized or otherwise processed for use?
How to address outliers	What parameters (if any) were followed to remove outliers?
Errors	Which errors were identified? How were errors corrected, removed or otherwise addressed?
Metadata	How are metadata addressed? Was any special processing needed to connect metadata with the data? Was it used in analysis?

Table 3 Analysis and methods attributes and metadata

Functions	Which function and language was used to perform this analysis?
Hyperparameter settings	For each analytical step, how were the hyperparameter values determined? What values for hyperparameters were used in the analysis?
Sub-setting data	Which subsets of data were included or excluded in this analysis? Which attributes determined this?
Feature engineering	Were certain data points or attributes excluded for any reason, i.e., missing too much data? Was any feature engineering used?
Workflow	How were the attributes determined? How were the analytical steps connected? How was the workflow determined or optimized?

that each data point is entirely contained in one line, and adjusting the columns to contain one variable per column. Extra spaces were removed.

There are several pieces of information that are necessary to be conveyed for an external researcher to understand and replicate the analyses presented in manuscript form. These include the data mentioned in Tables 2 and 3.

Depending on the analysis, some or all of these assumptions and methods may be relevant. Some of these details may not play a large role in the resulting analytics. However, in general, the assumptions made in the data analytics process tend to shape the results, so the inclusion of these descriptors enables the interpretation of the work and gives context to the results.

2.3 Previous Work and Analytical Methods: Overview of Clustering Steps & Goals

In order to investigate trends and underlying patterns in the dataset, several visualization and clustering techniques were applied to the data. A correlation matrix was created using the *corrplot* 0.84 library in R v. 4.0.1 to determine the correlation strength and direction between each pair of steel attributes. Additionally, the *pairs.panels* function through the *psych* library in R was used to visualize the distribution of each of the tensile properties as well as the interactions between the properties (Ref 34).

The distribution of data within the attribute space was then investigated using clustering. As the available data is based generally on the results of successful alloys, the data are naturally biased toward certain regions of the possible ranges of attributes. Identifying these biases can indicate the limitations of the dataset, and further the bias inherent in alloy design within a specific class (or designation like P91 steel used for boiler piping). However, the biases in the data can also be used to improve analysis by limiting the particular regression, or classification, algorithm to a narrower region of the data that have similar attributes, which is achieved through clustering.

A variety of clustering and dimensionality reduction techniques were used on the XMAT data. These include unsupervised methods based directly on the 9% Cr ferritic-martensitic steel data, which can be compared to clustering methods that use domain knowledge regarding design practices to generate typical steel class groupings. The comparison between the data-driven cluster results with the domain knowledge of steel class groupings can validate domain knowledge, or find areas where other, unidentified trends, occur. The clustering and segmentation techniques applied include k-means clustering, t-distributed stochastic neighbor embedding (t-SNE) analysis, and

principal component analysis (PCA) followed by partitioning around medoids (PAM).

Principal component analysis was performed on the 9% Cr ferritic-martensitic steel data using the *prcomp* function in R and visualized using the R library *factoextra*. PCA was evaluated with several combinations of scaling, centering, and attribute groupings, and the resulting principal components were visualized. PCA decomposes the input data into several principal components, which are linear combinations of attributes which are uncorrelated, and optimized to explain the most variance (Ref 16). The number of possible independent variables is then reduced to the few principal components that explain the majority of the variance (Ref 16). Partitioning around medoids (PAM), or k-medoids, was then performed on the principal components using the *pam* function in the cluster library through R. PAM was applied to identify groupings of data as represented by their principal components. Both PAM/k-medoids and k-means clustering algorithms use distance metrics to assign data points to a predetermined number of clusters (Ref 15). The main difference is the choice of group center, which for k-means is any point within the cluster center, where PAM/k-medoids uses one of the data points as the cluster center from which distance is measured.

t-SNE analysis was performed to visualize the high-dimensional data in a two-dimensional space, using the *Rtsne* library and *Rtsne* function, the *dplyr* library, and visualized using *ggplot2* in R (Ref 17). t-SNE was performed as an alternate technique to PCA. In t-SNE, the dimensionality reduction and visualization depend on the perplexity parameter, which relates to the number of nearest neighbors included in creating the dimensionality reduction (Ref 17). Lower perplexity focuses on local effects, where higher perplexity results in longer range effects (Ref 5, 17). The algorithm selects a unique solution each run, so the position of the points in space may vary. However, the overall trend in clustering, or smoothing, should be preserved.

Once t-SNE dimension reduction was performed on the data, domain knowledge determined labels were applied to the data. Specific 9% Cr ferritic-martensitic steel descriptions were used to segment the dataset and determine the overlap of the domain expertise-determined groupings with the t-SNE groupings. The 9% Cr ferritic-martensitic steel labels were determined based on the presence, absence, and quantity of alloying elements in the steel.

Previous cluster analyses of the 9% Cr ferritic-martensitic steel composition space were performed on initial versions of the dataset, with fewer unique steel class compositions. Due to the increased number of compositions, the data span more of the composition space and contain increased variability among

Fig. 1 Correlation matrices and scatter plots of test data. The scatter plots (left) indicate the substantial volume of data added to the database since analyses with previous editions of the database in Ref 11. Normal = normalization temperature, Temper1 = first tempering temperature

the composition attributes. This results in a reduced dimension space that may contain different groupings, and more outliers. Therefore, k-means clustering was additionally used after t-SNE analysis was performed to compare the resulting clusters from the t-SNE analysis and the domain knowledge labeling scheme with the clusters determined using k-means (Ref 12). K-means was performed using the *kmeans* function in R. This comparison improves interpretability and verifies the results of the t-SNE analysis. The t-SNE + k-means workflow can be iterated, and the results compared with domain-specific labels to optimize the clusters and to accommodate a higher degree of variation in the expanded dataset. The resulting composition ranges were evaluated based on the clusters identified.

3. Results

3.1 Pairwise Correlation

Correlations and trends in the data were investigated first, as demonstrated in Krishnamurthy et al. (Fig. 2; Ref 10), Krishnamurthy et al. (Fig. 5; Ref 9), and Romanov et al. (Fig. 2; Ref 11). Figure 1 shows the correlation matrix which highlights interdependencies between attributes, as well as scatter plots of the test attributes, distributions of each test attribute, and the Pearson correlation coefficient of each pair.

The resulting trends more closely align with the pairwise correlation plots and correlation coefficients in Romanov et al. (Fig. 2; Ref 11). This shows that there was a significant change

to the database during these iterations. The trends shown in Romanov et al. (Fig. 2; Ref 11) are for the most part present in Fig. 1 as well, especially the tensile test pairwise distributions, and the correlation trends between Fe and other elements. Additional data points can be seen in the *reduction in area* (TT_RA) plots as well as in the *temperature* (TT_Temp) plots, showing that the range of values for these attributes was expanded in the updated dataset. There are further additional outliers that do not appear to be present in the previous analyses. The correlation direction and shape between the tensile test attributes (test temperature, UTS, YS, elongation, and reduction in area) are generally preserved, but the strength of the correlation coefficients tends to decrease with more data as seen in the difference between Fig. 1 (left) and Romanov et al. (Fig. 2; Ref 11).

3.2 Principal Component Analysis and Partitioning Around Medoids

To investigate potential groupings and trends within the 9% Cr ferritic-martensitic steel compositions, PCA was performed on a previous version of the dataset, and clusters were analyzed using PAM, described in Fig. 7 in Krishnamurthy et al. Ref 9. The groupings, generated using the partitioning around medoids algorithm (PAM), are intended to show trends in composition relating to tensile properties including UTS. A similar analysis was completed in this work on the updated version of the database, shown in Fig. 2. In this case, only the non-homogenized steel variants were included. [Note: Internal research at NETL under the Advanced Alloy Development

Fig. 2 PCA with PAM clustering using 17 compositional element attributes and test temperature. The two PCA dimensions shown account for 32 and 22.8% of the overall variance, respectively

Fig. 3 Left, PCA & PAM, using all composition, heat treatment, homogenization and grain size attributes; right, using only composition attributes

(AAD) field work proposal (FWP) has shown that attaining *chemical* homogeneity in advanced heat resistant alloys improves overall high-temperature alloy performance (Ref 35, 36). As such heat-resistant alloys developed under the AAD FWP utilize a homogenization cycle whenever possible (dependent on constitution of elements in the alloy).] PCA was performed first, and the attributes were centered around zero

and scaled for unit variance using the *prcomp* function parameters before performing PCA. The attributes included in this analysis were 17 compositional elements (Fe, Cr, C, N, Mn, Ni, Si, V, Nb, Mo, W, Cu, Al, S, P, B, Co), as well as the UTS. The addition of the UTS results in the nearly vertical trend in points (Fig. 2). This effect is not seen in the clustering results of only composition (Fig. 3). PAM was performed using

Fig. 4 Left, k-means clustering of the 9% Cr ferritic-martensitic steel elemental composition data for non-homogenized variants, visualized using Fe vs. Cr content (wt.%). Right, possible new data added to the database highlighted in red

the first seven principal components, and groupings were created based on the Euclidean distance. Seven groups were chosen to visualize the data following (Ref 9). The results from the recreation of the clusters appear to be reversed along the horizontal and vertical axes, but a similar clustering pattern can be seen. Additional data points appear in the center of the plot to the purple cluster (cluster 6), showing the addition of data since the publication. These results suggest that the addition of non-homogenized steels to the dataset affected the clustering results less than the additional homogenized data.

Several assumptions are the key to plotting and interpreting principal components. For example, the choices for pre-processing the data include the choice to scale the data, determining which attributes to include, and whether or not to subset the data based on specific attributes. The choice of analyzing data points with, or without, test data further influences the results of the clusters. Including test data results in duplicate points with the same composition, processing and heat treatment data, with the only variation being in the test attributes. Further, depending on the algorithm and function used to determine and plot the principal components, the researcher can determine a set number of principal components to include, or set a specific percentage of variance explained and let the function automatically determine the number of components. Similarly, analytical choices in performing partitioning around medoids (depending on the choice of algorithm and function) include the choice of which distance metric to use, as there are calculation differences in using Euclidean distance versus the Manhattan distance between data points. Note that overlapping groups occur due to the visualization of the clusters in two dimensions rather than in the original 18-dimensional space or the reduced seven-dimensional principal component space.

The next step of identifying the steel class compositions within these clusters is to match the clusters to known composition groupings, or steel class names. This was done in Fig. 4 from Romanov et al. (Ref 11). It was not clear which attributes were included in the analysis to generate the clusters in this figure. It is evident that test results were not included, due to the lack of the successive trend of “stacked” data points visible in Fig. 2. Several combinations of attributes were therefore used to cluster

the data and the resulting principal components and PAM clusters were visualized, and the clustering patterns were compared with Romanov et al. (Ref 11) (Fig. 3). In this case, the data were also centered around zero and scaled for unit variance using the center and scale options in the *prcomp* function, and the Euclidean distance was used to generate the PAM clusters. All of the principal components generated in the PCA were used to calculate the clusters. Clusters generated using all composition attributes (Fig. 3, right) and composition, heat treatment, homogenization, and grain size (Fig. 3, left) tended to be overlapping in the 2D principal component space, and resulted in clustering trends that did not match those seen in Ref 11. This could be due to Romanov et al. using a different combination of attributes to generate the clusters, or in differences in the clustering outcomes due to the addition of new data.

Romanov et al. further identified clusters using labels generated through a custom algorithm (Ref 11). Those are plotted and compared with the clusters identified using PCA/PAM. The labels are assigned based on the presence, or absence, of W, Co, V, Ta, B, Cu, and Mo.

These labels were recreated on the updated dataset by assigning a label based on the presence, or absence, of the alloying elements as listed above. However, the labels did not all apply to the new dataset. Upon assigning labels, only five were present out of the seven that were described by Romanov et al.: C111, C112, C122, C221, and C221 (Ref 11). Therefore, the labeling scheme for this technique could be refined by setting a minimum threshold for inclusion, or not, into each grouping. For example, indicating whether the absence of an element is indicated by 0, or if the value reported is less than some minimal value, i.e., < 0.01 .

While the technique of visualizing clusters using PCA and PAM can be useful to show clustering and underlying trends in the data, one downside is that the principal components cannot be directly traced back to a trend in a particular data feature. However, general trends can be observed by comparing resulting clusters with domain knowledge of typical compositions and processing methods. Due to the difficulty in interpreting the PCA results, other clustering techniques were investigated.

Fig. 5 t-SNE analysis of the 17 compositional elements for the non-homogenized in the 9% Cr database. The UTS value is the color of each point, highlighting the overwhelming trend in organization by UTS

3.3 K-Means Clustering

K-means clustering was performed on the 9% Cr ferritic-martensitic steel data considering the non-homogenized ones with 17 compositional elements as described in Krishnamurthy et al. Ref 9. The number of clusters was increased from 6 to 7 to account for added data and to improve the resulting visualization. The color of the point corresponds with the assigned cluster based on the 17-dimensional dataset. The number of random starts to randomly assign cluster centers was assigned using $nstart = 20$. This clustering technique successfully recreated the original clusters using the smaller database, while suggesting where new data had been added (Fig. 4). Suspected new data have been highlighted in red. The additional data shift the resulting clusters, and an additional cluster is now needed to show outliers from the overall trend, i.e., the points shown in light blue. The linear trend in Fe versus Cr indicates the trade-off between addition of the alloying element and reduction in Fe. However, for different clusters of steels, this trade-off involves other attributes in addition to Cr, indicated by different slopes in different clusters.

Implementing k-means clustering requires additional assumptions, and parameters must be set to perform the analysis. These again include decisions regarding data preprocessing, if scaling should be implemented, whether, or not, to subset the data, and whether to include test results, or just steel composition and processing. Additionally, the choice of the number of clusters as well as the distance metric used are important features, and there are several methods available to determine the optimal number of clusters, including the elbow method (Ref 37).

t-SNE analysis was performed to visualize the high-dimensional dataset in lower-dimensional space (see Fig. 8 in Ref 9). Compositional elements as well as the UTS are included, and only non-homogenized alloys were considered, with duplicates removed. In this case the, data were not centered or scaled. The analysis was performed in R (Fig. 5).

In this analysis, the alignment of the data points can be seen in the progression of increasing perplexity in Fig. 5. This overall trend was recreated from the updated datasets. Additionally, the data points were colored by a series of the data

Fig. 6 t-SNE analysis of the updated dataset, perplexity = 6. Circled data points indicate the alloys where a new label is needed

Table 4 Composition ranges for labels as assigned and visualized in Fig. 8

Group	Cr	Mn	Ni	Mo	W	Cu	Nb	V	B(a)
12Cr	11.64-12.70	0.32-0.83	0.08-0.53	0.02-0.21	0-0.05	0-0.210	0-0.180	0-0.040	0
9Cr1Mo	8.46-9.15	0.38-0.50	0.04-0.12	0.93-1.05	0	0.020-0.080	0	0	0
9Cr2Mo	9.04	0.50	0.11	2.03	0	0.020	0.006	0.010	0
9Cr1.5MoVNbB	9.11-9.29	0.05-0.06	0.08-0.09	1.54-1.57	0	0	0.055-0.059	0.260-0.290	82-88
10.5Cr1Mo1WVNb	9.94-11.00	0.41-0.59	0.67-0.78	0.97-1.10	0.95-1.07	0-0.090	0-0.050	0.170-0.210	0
12Cr1Mo1WCoVCu	11.20-12.90	0.55-0.98	0.64-0.90	0.91-1.22	0.9-1.13	0.010-0.150	0	0.210-0.300	0
9Cr1MoVNb	8.31-8.74	0.35-0.49	0.04-0.28	0.89-0.99	0	0-0.032	0.048-0.090	0.185-0.230	0
9CrMo1.8WVNbB	8.70-9.65	0.41-0.46	0.13-0.46	0.36-0.46	1.67-1.85	0-0.003	0.050-0.080	0.160-0.210	0-36
10.5CrMo2WVNbCuB	10.10-10.94	0.50-0.70	0.17-0.50	0.10-0.82	0.05-2.89	0-0.970	0.006-0.430	0.190-0.220	0-100
12CrMo1.8WVNbCuB	12.10	0.59	0.33	0.34	1.82	0.820	0.060	0.190	0.003
9Cr3W3CoVNbTaB	9.08-9.60	0.50-0.52	0.17-0.19	0.10-0.11	2.89-2.92	0.003-0.032	0.060-0.065	0.200-0.208	10-121
17Cr	16.20-17.55	0.40-0.69	0.16-0.36	0.01-0.09	0	0	0	0	0
CrMo0.5WCoVNbBTa	9.11-10.74	0.28-0.48	0.20-0.31	0.50-1.47	0.42-0.53	0.002-0.295	0.049-0.062	0.190-0.237	75-113
CrMoCo0.2VNbB	9.28-10.63	0.31-0.51	0.06-0.31	0.01-1.54	0-0.60	0-0.042	0.050-0.147	0.020-0.213	87-200

(a) Boron (B) entries are in parts per million (ppm)

attributes to discern the underlying trend. In this case, the UTS of the steel was used to generate the color scale, where dark blue represents low UTS and light blue represents high UTS. By increasing the perplexity past the initial 30 to 70, the t-SNE plot reveals the alignment of the data points completely according to UTS. While t-SNE analysis correctly identified the UTS as a critical variable, the resulting plot does not give new information about the design parameters of the steels as related to their tensile properties. This indicates that clustering data

without mechanical property test parameters (or results) may help the interpretation of the reduced dimensionality plot.

Therefore, another t-SNE analysis was performed on the compositional attributes without any test attributes. In this case, the analysis includes 19 compositional elements (Fe, C, Si, Mn, P, S, Cr, Mo, W, Ni, Cu, V, Nb, N, Al, B, Co, Ta, and O) and 2 heat treatment temperatures (normalization and tempering), following Verma et al. Ref 8. As no test data was included, duplicates were removed so that only unique steels remained.

Fig. 7 K-means analysis of Fig. 6, indicating the clusters that minimize the distance between each cluster center and the surrounding points

Fig. 8 t-SNE analysis repeated with new label applied to the cluster identified using k-means, appearing in Fig. 7 as the light green cluster toward the top of the figure. Note that the t-SNE analysis was repeated from Fig. 6, and the layout of the points is not preserved from run to run

Fig. 9 K-means analysis of updated steel grouping labels. The k-means groupings can be compared with the assigned labels, indicating where new labels can be created to fit the identified groups

In this case, each attribute was centered around zero and called for standard deviation using the *scale* function in R. Various low perplexity values were assessed (2-10), and perplexity = 6 (Fig. 6) showed a similar clustering trend to the t-SNE results in Fig. 1 from Verma et al. Ref 8.

In order to investigate the resulting clusters, the data points were colored using applied domain knowledge labels of 12 common steel groupings. This method was applied by Verma et al. Ref 5–8) and was adapted for this dataset. For example, the 17Cr data were not present in the previous version of the dataset and appear here as its own cluster. Clusters appear in the resulting reduced dimension plots for low perplexity (Fig. 6).

The resulting clusters are not as distinct as the t-SNE analysis on the initial dataset shown in Fig. 1 in Ref 8. This could be due to the addition of new data that does not fit into the same steel groupings as previous versions of the database. The differences could also be partially due to the inherent uncertainty in the t-SNE output, and the variation between runs. In Fig. (6b), the circled data points indicate clusters where the initially determined steel labels did not align with a single cluster. As in Verma et al. (Ref 8), a new label was assigned to this cluster in order to unify the grouping.

There are several assumptions that are necessary to produce a t-SNE visualization. These comprise what data attributes to include in the analysis, and whether, or not, to subset the data by a certain attribute. Additionally, there is the option of scaling and centering data during data preprocessing. There are several pre-built functions to generate t-SNE analyses, and they contain different options for setting hyperparameters and for visualizing the data, including functions written in R and Python.

3.4 Defining New Labels for the Full Dataset

To further investigate the distinct clusters generated by the t-SNE analysis, k-means clustering was performed. Calculating

the groupings according to the k-means process allows for the clusters to be seen according to their relative positions in the reduced dimension space. While the variation in distance between clusters in the t-SNE space is generally not relevant, as the relative locations in the space changes between runs, the assignment of points to a particular cluster in the space should be consistent in repeat analyses (Table 4).

The extent of the grouping where the new label should be applied was determined using k-means analysis. As seen in Fig. 7, the light green points encompass the greatest variation in steel class labels. Therefore, this grouping was relabeled with the steel label determined by Verma et al. (Ref 8) as the composition was similar to the grouping established in the previous work.

The updated t-SNE analysis of the 9% Cr ferritic-martensitic steels is shown in Fig. 8. Note that the variation in cluster location in the reduced dimension space is due to inherent variation in the t-SNE visualization process.

The range of composition for each label grouping is given in Table 2. Note: One group (9Cr1.5Mo1.3CoVNbB) is completely incorporated into the two newly labeled groups.

K-means analysis of the updated t-SNE visualization, with the two new labels, is shown in Fig. 9. This plot shows agreement in relative clusters with the first k-means analysis (Fig. 7). The composition ranges for these clusters are included in Table 5. Approximate domain knowledge labels are aligned with the k-means cluster numbers for comparison.

3.5 Tensile Test Results by Cluster

To investigate the trends in tensile properties based on the clustering results, the ultimate tensile strength was averaged for each temperature and cluster where four or more data points were available (Fig. 10). The resulting summarized UTS values

Table 5 Composition ranges for the clusters in Fig. 9.

Related Label	Group	Cr	Mn	Ni	Mo	W	Cu	Nb	V	B(a)
12Cr	7	11.65-12.70	0.56-0.83	0.09-0.22	0.02-0.15	0	0	0	0	0
12Cr	9	11.64-12.38	0.42-0.71	0.15-0.45	0.04-0.21	0	0.030-0.210	0	0	0
12Cr	4	11.65-11.94	0.32-0.81	0.08-0.31	0.03-0.08	0	0	0	0	0
9Cr1Mo + 9Cr2Mo	10	8.46-9.15	0.38-0.50	0.04-0.12	0.93-2.03	0	0.020-0.080	0-0.006	0-0.010	0
9Cr1.5MoVNbB	1	9.11-9.29	0.05-0.06	0.08-0.09	1.54-1.57	0	0	0.055-0.059	0.260-0.290	82-88
10.5Cr1Mo1WVNb + 9CrMo1.8WVNbB	13	8.70-10.50	0.41-0.55	0.13-0.77	0.36-1.05	0.95-1.85	0-0.003	0.041-0.080	0.160-0.210	0-36
12Cr1Mo1WCoVCu	6	11.00-12.90	0.55-0.98	0.64-0.90	0.91-1.22	0.90-1.13	0.010-0.150	0	0.210-0.300	0
9Cr1MoVNb	12	8.31-11.82	0.35-0.70	0.04-0.53	0.02-0.99	0-0.05	0-0.060	0.048-0.430	0.030-0.230	0
10.5CrMo2WVNbCuB + 12CrMo1.8WVNbCuB	5	10.65-12.10	0.59-0.63	0.32-0.36	0.33-0.38	1.82-1.97	0.820-0.970	0.050-0.060	0.190-0.220	24-39
9Cr3W3CoVNbTaB	3	9.08-10.10	0.50-0.52	0.17-0.19	0.10-0.101	2.89-2.92	0.003-0.032	0.006-0.065	0.200-0.208	100-121
17Cr	2	16.20-17.55	0.40-0.64	0.16-0.32	0.01-0.04	0	0	0	0	0
17Cr	11	16.20-16.35	0.64-0.69	0.33-0.36	0.08-0.09	0	0	0	0	0
CrMo0.5WCoVNbBTa	14	9.11-10.74	0.28-0.48	0.20-0.31	0.50-1.47	0.42-0.53	0.002-0.295	0.049-0.062	0.190-0.237	75-113
CrMoCo0.2VNbB	8	9.28-10.63	0.31-0.51	0.06-0.31	0.01-1.54	0-0.60	0-0.042	0.050-0.147	0.020-0.213	87-200

(a) Boron (B) entries are in parts per million (ppm)
The clusters are aligned with the domain knowledge labels visualized in Fig. 8, with some labels applying to two clusters, and some clusters containing multiple labels

indicate the best performing clusters and alloy properties at each temperature.

4. Discussion

Working to re-evaluate the results of clustering and other analytical techniques on updated datasets indicates the influence of new data on the correlations and similarities of the data groups. New composition groups have been identified. Most clearly, the group for 17Cr can be seen as a distinct cluster from the rest of the composition groups.

The k-means clusters give fairly close alignment with the clusters identified through t-SNE with domain labels assigned, as indicated in the correct identification of several t-SNE groups using k-means. Increased variation in the alloying elements resulted in a wider range of steel compositions represented in reduced dimensional space as visualized through t-SNE. Additionally, the increased overlap between domain labels in a single cluster indicates that other trends have been identified in the composition beyond the major alloying elements for those clusters (Fig. 6). One notable difference between k-means and the t-SNE labels is the 12Cr group. This label group consistently appears close in the t-SNE analyses, but was identified by k-means as having several sub-clusters. These sub-clusters can be investigated in subsequent works to see if the sub-clusters give better correlations with the mechanical properties than the unified 12Cr group.

Summarizing the UTS by cluster over the range of tensile test temperatures allows trends to be identified within the t-SNE groupings and relationships between the clusters and the mechanical properties of interest to be determined. As shown in Fig. 10, the UTS vs. test temperature space can be aggregated into several clustering trends by label. Certain clusters result in uniform trends (CrMo0.5WCoVNbBTa, 9Cr1.5MoVNbB, 12Cr, 9Cr1MoVNb, 9Cr1Mo) while others incorporate a wider range of UTS (CrMoCo0.2VNbB, 10.5CrMo2WVNbCuB). While for temperatures between 100 and 550°C, the CrMoCo0.2VNbB cluster obtains the highest UTS, at 600–650°C the UTS drops beneath that of other clusters. At 650°C, the highest UTS is achieved by the CrMo0.5WCoVNbBTa cluster and the 9Cr3W3CoVNbTaB cluster. Where data are available at 700–750°C, the highest UTS is seen in the 9Cr1.5MoVNbB cluster. All three clusters have similar amounts of Nb and V, and between 9.08 to 10.7 wt.% Cr. As data are aggregated for tensile properties, improved correlations between alloy attributes and tensile behavior at high temperatures can be made.

The resulting clusters will be further used in regression and classification analysis to reduce the variation in the dataset and to evaluate differences among similar groups of steels. This approach should improve the effectiveness of ML algorithms in predicting the mechanical properties of different alloy types and in determining which factors are significant. The similarity of the data within the clusters will more distinctly highlight the influence of attributes contributing to mechanical performance.

4.1 Assumptions and Reported Attributes Needed in Data Analysis

During the process of examining the results of the work done previously on earlier iterations of the dataset, several insights were gained into the data processing and analysis

Fig. 10 Average UTS per temperature, where four or more data points available per cluster per temperature

methodologies. In terms of producing data visualizations and interpreting the results of data analytics, there are numerous options available and several inherent assumptions that are made in each analysis. Often data analytics are difficult to interpret, as details on these assumptions and methods may not be included in the published work. In this effort several categories of data analytics assumptions were identified that can be included in the discussion of the methods that will enable data scientists and domain experts to gain more insights into the stated and visualized results. Clarity in these areas will hopefully improve the trustworthiness of data analytics results.

5. Conclusions

Data analysis and design efforts in the area of heat resistant alloy composition–processing–property relationships are highly complex and the application space is broad. Several analyses discussed in this work have been performed on earlier iterations of the XMAT dataset, which therefore should be represented and incorporated into future analytics to improve efficiency of research and to broaden understanding of the results.

To leverage prior work, these analyses were reproduced by following the descriptions in the published manuscripts. The specific correlation and clustering methods that followed indicated that some results were easier to generate than others, depending on the level of detail supplied in the manuscript. These processes enabled key components of data processing

and analysis to be identified as necessary in recreating the analyses represented in the literature.

The assumptions inherent in data analytics are often contained in the data preprocessing and cleaning, as well as in the clustering or dimensionality reduction calculations themselves. There are few standards guiding the reporting of methodologies, but adherence to detail in reporting the methods used in this area has the potential to improve analytical reuse and reapplication in furthering the understanding from results.

Additionally, through this process, the variation in the correlations and clustering analyses were identified with respect to the addition of new data with increased variation in the alloy (specifically 9% Cr ferritic-martensitic steel) attribute space. This work indicates that strong correlations identified with small datasets can become weaker when more data are collected. Therefore, a certain degree of restraint is called for in drawing conclusions from small datasets. Further, new techniques were applied to the updated and expanded XMAT dataset. Applying k-means to the t-SNE space allowed the identification of new steel groupings not present in previous iterations of the database and will enable improved property prediction based on these new clusters.

Specifically, investigation of the average UTS per cluster over the range of temperatures tested indicated that while the alloys in the CrMoCo0.2VNBb cluster achieve the highest UTS up to 550°C, above this temperature from 600 to 650°C the highest UTS is achieved by the CrMo0.5WCoVNBbTa cluster and the 9Cr3W3CoVNBtAb cluster. These clustering trends

highlight the variation in mechanical properties over temperature regimes and indicate specific alloy properties correlated with improved UTS.

Acknowledgments

This work was supported by the NETL Crosscutting Research Program, Briggs White, NETL Technology Manager, and Regis Conrad and Robert Schrecengost, DOE-FE HQ Program Managers. This work was executed through the eXtremeMAT National Laboratory Field Work Proposal (NETL: FWP-1022433, LANL: FWP-FE85017FY17, ORNL: FWP-FEAA134, Ames: FWP-AL-17-510091, LLNL: FWP-FEW0234, INL: FWP-B000-17016, PNNL: FWP-71133). Research performed by Leidos Research Support Team staff was conducted under the RSS contract 89243318CFE000003.

Disclaimer

This work was funded by the Department of Energy, National Energy Technology Laboratory, an agency of the United States Government, through a support contract with Leidos Research Support Team (LRST). Neither the United States Government nor any agency thereof, nor any of their employees, nor LRST, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Data Availability

Part of the data used for the analysis is the property of DOE-NETL, an agency of the United States Government, and is confidential. The rest of the data comes from NIMS, Japan, and the Materials Algorithm Project, Cambridge. The NIMS data can be accessed directly from https://mits.nims.go.jp/index_en.html. Specific details of spreadsheets from NIMS are included in references (Ref 17–27). The Materials Algorithm Project data can be accessed directly at <http://www.phase-trans.msm.cam.ac.uk/map/data/materials/creeprupt-b.html>. Reference information is provided in Ref 16.

References

1. E. De Guire et al., Data-Driven Glass/Ceramic Science Research: Insights from the Glass and Ceramic and Data Science/Informatics Communities, *J. Am. Ceram. Soc.*, 2019, **102**(11), p 6385–6406
2. R. Stevens et al., *AI for Science*. 2020, Argonne National Lab.(ANL), Argonne, 2020
3. L. Cao, Data Science: Challenges and Directions, *Commun. ACM*, 2017, **60**(8), p 59–68
4. M. Wenzlick et al., Data Assessment Method to Support the Development of Creep-Resistant Alloys, *Integr. Mater. Manuf. Innov.*, 2020, **9**(1), p 89–102. <https://doi.org/10.1007/s40192-020-00167-3>
5. A. Verma, et al., *Mapping multivariate influence of alloying elements presentation*. in, *MSLE*. 2018. Carnegie Mellon University
6. A.K. Verma et al., Mapping Multivariate Influence of Alloying Elements on Creep Behavior for Design of New Martensitic Steels, *Metall. Mater. Trans. A*, 2019, **50**(7), p 3106–3120. <https://doi.org/10.1007/s11661-019-05234-9>
7. A.K. Verma et al., Predictions of Long-Term Creep Life for the Family of 9–12 wt.% Cr Martensitic Steels, *J. Alloys Comp.*, 2020, **815**, p 152417. <https://doi.org/10.1016/j.jallcom.2019.152417>
8. A.K. Verma et al., Screening of Heritage Data for Improving Toughness of Creep-Resistant Martensitic Steels, *Mater. Sci. Eng., A*, 2019, **763**, p 138142. <https://doi.org/10.1016/j.msea.2019.138142>
9. N. Krishnamurthy et al., 9Cr Steel Visualization and Predictive Modeling, *Comput. Mater. Sci.*, 2019, **168**, p 268–279. <https://doi.org/10.1016/j.commatsci.2019.03.015>
10. Krishnamurthy, N., et al., *Data analytics for alloy qualification*. 2018, National Energy Technology Lab.(NETL), Pittsburgh, PA, and Morgantown, WV
11. V.N. Romanov et al., Materials Data Analytics for 9% Cr Family Steel, *Stat. Anal. Data Min. ASA Data Sci. J.*, 2019, **12**(4), p 290–301
12. A. Juras et al., Mitochondrial Genomes Reveal an East to West Cline of Steppe Ancestry in Corded Ware Populations, *Sci. Rep.*, 2018, **8**(1), p 1–10
13. L. Van Der Maaten, E. Postma, and J. Van den Herik, Dimensionality Reduction: A Comparative, *J. Mach. Learn. Res.*, 2009, **10**(66–71), p 13
14. D. Xu and Y. Tian, A Comprehensive Survey of Clustering Algorithms, *Ann. Data Sci.*, 2015, **2**(2), p 165–193
15. P.J. Rousseeuw and L. Kaufman, *Finding Groups in Data*, Wiley Online Library, Hoboken, 1990, p 1
16. S. Wold, K. Esbensen, and P. Geladi, Principal Component Analysis, *Chem. Intell. Lab. Syst.*, 1987, **2**(1–3), p 37–52
17. Lvd Maaten and G. Hinton, Visualizing Data Using t-SNE, *J. Mach. Learn. Res.*, 2008, **9**, p 2579–2605
18. A.K. Jain, Data Clustering: 50 years Beyond K-means, *Pattern Recogn. Lett.*, 2010, **31**(8), p 651–666
19. F. Brun, T. Yoshida, *Data Library MAP_DATA_CREEP RUPTURE*, in *Materials Algorithms Project*. University of Cambridge: Cambridge, U.K
20. NIMS, *Fatigue Data Sheet Series*, no. 30. 1982, National Institute for Materials Science: Tsukuba-shi
21. NIMS, *Fatigue Data Sheet Series*, no. 29. 1982, National Institute for Materials Science: Tsukuba-shi
22. NIMS, *Fatigue Data Sheet Series*, no. 35. 1983, National Institute for Materials Science: Tsukuba-shi
23. NIMS, *Creep data sheet*, no. 13B. 1994, National Institute for Materials Science: Tsukuba-shi
24. NIMS, *Creep data sheet*, no. 19B. 1997a, National Institute for Materials Science: Tsukuba-shi
25. NIMS, *Creep data sheet*, no. 44. 1997b, National Institute for Materials Science: Tsukuba-shi
26. NIMS, *Creep data sheet*, no. 10B. 1998, National Institute for Materials Science: Tsukuba-shi
27. NIMS, *Fatigue Data Sheet Series*, no. 88. 2002, National Institute for Materials Science: Tsukuba-shi
28. NIMS, *Creep data sheet*, no. 46A. 2005, National Institute for Materials Science: Tsukuba-shi
29. NIMS, *Creep data sheet*, no. 51A. 2013a, National Institute for Materials Science: Tsukuba-shi
30. NIMS, *Creep data sheet*, no. 52A. 2013b, National Institute for Materials Science: Tsukuba-shi
31. NIMS, *Creep data sheet*, no. 43A. 2014, National Institute for Materials Science: Tsukuba-shi
32. NIMS, *Creep data sheet*, no. 48B. 2018, National Institute for Materials Science: Tsukuba-shi

33. H. Wickham, Tidy Data. *J. Stat. Softw.*, 2014. <https://doi.org/10.18637/jss.v059.i10>
34. *pairs* - R. 3.6.2:[Available from: <https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/pairs>]
35. P.D. Jablonski and C.J. Cowen, Homogenizing a Nickel-Based Superalloy: Thermodynamic and Kinetic Simulation and Experimental Results, *Metall. Mater. Trans. B*, 2009, **40**(2), p 182–186. <https://doi.org/10.1007/s11663-009-9227-1>
36. P.D. Jablonski and J.A. Hawk, Homogenizing Advanced Alloys: Thermodynamic and Kinetic Simulations Followed by Experimental Results, *J. Mater. Eng. Perform.*, 2017, **26**(1), p 4–13. <https://doi.org/10.1007/s11665-016-2451-3>
37. T.M. Kodinariya and P.R. Makwana, Review on Determining Number of Cluster in K-Means Clustering, *Int. J.*, 2013, **1**(6), p 90–95

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.