One of These Things IS Like the Other: Pursuing a New Taxonomy of Industry for Improved Energy System Modeling

Preprint

Liz Wachs and Colin McMillan

National Renewable Energy Laboratory

Presented at the 2021 Summer Study on Energy Efficiency in Industry Virtual
July 12-15, 2021

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-6A20-80141
August 2021

One of These Things IS Like the Other: Pursuing a New Taxonomy of Industry for Improved Energy System Modeling

Preprint

Liz Wachs and Colin McMillan

National Renewable Energy Laboratory

Suggested Citation
Wachs, Liz and Colin McMillan. 2021. *One of These Things IS Like the Other: Pursuing a New Taxonomy of Industry for Improved Energy System Modeling: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-6A20-80141. https://www.nrel.gov/docs/fy21osti/80141.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-6A20-80141
August 2021

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Advanced Manufacturing Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

One of These Things IS Like the Other: Pursuing a New Taxonomy of Industry for Improved Energy System Modeling

Liz Wachs and Colin McMillan, National Renewable Energy Laboratory

ABSTRACT

Industrial processes drive the exchange of materials, energy, and currency throughout the economy. These processes are powered by electricity and direct combustion, with variations in their operation even within the same industry. This heterogeneity makes it difficult for large models, including the National Energy Modeling System (United States), to project their energy use while remaining tractable. Decarbonization and ensuing changes to the energy system require changes to industrial processes while offering opportunities for process innovation, but the extent and nature of changes are difficult to model with current classification schemes and corresponding data. The North American Industry Classification System (NAICS) is an economic taxonomy of industries, but its categories are less meaningful from an energy and material flow perspective. For example, a facility that makes steel from iron ore in a blast furnace/basic oxygen furnace is categorized under the same NAICS code as a facility that makes steel from scrap in an electric arc furnace despite the differences between the two facility types in terms of scale, the use of recycled scrap versus iron ore, and energy use. Exploratory analysis is performed on a large data set used for plant-level energy assessment to detect clusters that can aid in the better modeling of industry for energy analysis in an evolving system with breakthrough technologies.

Introduction

Industry is at the center of material and energy use in the U.S. economy. As such, it is a major contributor to greenhouse gas emissions, climate change, and other environmental pressures. Because energy use is directly tied to greenhouse gas emissions via direct combustion as well as via combustion-based electricity sources, initiatives to drastically reduce greenhouse gas emissions also require the study of energy use by industry. Unfortunately, the ability to model industry at the national scale is limited because of the heterogeneity of industrial processes, buildings, and related energy use. Most data issued by statistical agencies, such as the Manufacturing Energy Consumption Survey, are based on underlying classification systems where energy use and other environmental considerations are an afterthought. This is problematic because when energy use is dissimilar within a group, it is hard to model how changes will affect the group overall. This work is part of an attempt to study classification systems that are frequently used to describe industry to make changes that will allow better projections and better models for study.

The definition of *industry* determines which facilities and firms are included in analysis and can add confusion. *Industry* and *manufacturing* are often used synonymously. The lay definition of *manufacture* includes “something made from raw materials by hand or by machinery” and “the process of making wares by hand or by machinery especially when carried

on systematically with division of labor" (Merriam-Webster 2021b). *Industry* is defined as "manufacturing activity as a whole" as well as "systematic labor especially for some useful purpose or the creation of something of value" (Merriam-Webster 2021a). The broader definition of industry allows the exchange of labor for value, including service sectors that might not have a physical product. Manufacturing, however, requires physical goods and even a physical product. Still, that does not mean that all physical products are manufactured. Agriculture is sometimes considered to be outside of industry and generally considered to be outside of manufacturing because neither people nor machines affect the transformation of the raw materials into the finished good. Likewise, extractive industries have a physical product but are not usually considered to be part of manufacturing. Construction is generally excluded because the final good is not portable. Utilities, including the treatment of waste and water, are also often considered to be outside of manufacturing. This exclusion can create a separation of waste processing from its source.

This preliminary exploration of the definition of industry and manufacturing provides a background for what is included and excluded in this study. In this work, manufacturing refers to the production of physical, portable products by human or machine labor, excluding agriculture, extraction, construction, and waste treatment. Physical industry, however, includes all those sectors (agriculture, extraction, construction, and waste treatment); and industry includes the service sectors.

Energy use by U.S. manufacturing industries is heterogeneous. Electricity is widely used, but it represented only 13% of the first use of energy in manufacturing in 2018[1]. Many processes also employ direct combustion. Natural gas is the most commonly used fuel, followed by hydrocarbon gas liquids, and then a diverse group of fuels, including coal, bitumen, biomass, and others (US Energy Information Administration, 2018). Hydrocarbons used as feedstock rather than fuels accounted for more than 6,000 of the 19,436 trillion Btu of the first use of energy in manufacturing.

Energy use by industry can be considered at multiple scales—from a unit process level of a single component; to a single product, a single industry, or a single subsector; to manufacturing as a whole or even at the economy level. Methods of grouping the different types of energy use are important to enable modeling and projecting energy use as well as for offering solutions. To provide various degrees of resolution, different taxonomies or classification schemes have been proposed. On an international, inclusive industry level, the dominant scheme is the International Standard Industrial Classification (ISIC). On the product side, there is the Central Product Classification Version 2.1. In North America, the North American Industry Classification System (NAICS) and the North American Product Classification System (NAPCS) are used at the industry and product levels, respectively. There is a high degree of interoperability between these systems so that national statistics between countries can be compared; however, they are not as coordinated as the European systems, including NACE and ISIC.

Classification systems are important because they form the basis of data collection and reporting and hence modeling systems that are based on reported data; thus, they present a limitation for the data available to prioritize changes and track progress, which becomes limiting

[1] The Manufacturing Energy Consumption Survey issues Tables 1–5 on the "First Use of Energy for All Purposes," which refers to an accounting of all energy used by manufacturing facilities used for both fuel and nonfuel purposes. The total is calculated to avoid double counting where possible.

when discussing ambitious targets, such as halving or eliminating greenhouse gas emissions, or when discussing a circular economy. Current systems cannot show how technologies that can be used in many different sectors via processes can allow decarbonization because, for example, in the case of U.S. iron and steel mills, all are grouped under one six-digit NAICS code (331110), so although electrolytically produced hydrogen feedstock could allow for decarbonization in some resource-intensive facilities, it would not be relevant to the vast majority of facilities. The NAICS system masks the processes and equipment that can be used to produce similar products.

This work looks at classification systems of industry and how they are structured—particularly at how energy systems models classify industry. It is high risk in the sense that this is exploratory work and might not uncover better ways of classification. Still, it is useful because current limitations make modeling difficult, especially with large-scale transformations. A preliminary exploration of manufacturing data at the plant level is conducted to find patterns and clusters that suggest similarity in energy use trends. It is part of supporting evidence that looks at how reported data on industrial energy use might be better structured for energy modeling. Because the energy systems models do not have the bandwidth to model in detail many types of industrial processes, this work attempts to find groupings that might be overlooked by the NAICS codes and major process similarities. It looks to find a small number of groupings with relatively high correlation between energy use and the and the information available at the plant level (in the case of the IAC dataset this includes plant hours, number of employees, plant area and sales), with an eye to finding archetypes that can be investigated further. To do this, artificial intelligence approaches of unsupervised learning are employed. Unsupervised learning is one approach to this type of exploratory data analysis; it is more subjective and less goal oriented than approaches such as regression (James, Witten, Hastie, & Tibshirani, 2013). Clustering is a group of approaches that leverage the information in large data sets to search for groupings. It is frequently used for market segmentation—for example, to understand different groups within a clientele. Although customers might differ in terms of characteristics, they all consume a specific product, so understanding the clusters can help corporations market to them.

Classification Systems

Starting from the top level of aggregation, the United Nations has created ISIC. In the United States, the Office of Management and Budget worked with Mexico's Instituto Nacional de Estadistica y Geografia and Statistics Canada to develop the NAICS, which is generally compatible with the ISIC but with a higher level of detail. The hierarchy in NAICS 2017 consists of 20 sectors, 99 subsectors, 311 industry groups, 709 five-digit industries, and 1,057 national industries. Although the three North American countries share the same two-digit level classification, there is some divergence at the lower levels because of idiosyncrasies in the different countries and economies. In the NAICS, which is revisited every 5 years, each single physical "establishment" is assigned to a specific code according to its primary activity (i.e., the headquarters of a manufacturing company would have the management code even though the primary activity of the business is manufacturing). The activity is ideally chosen based on the costs incurred, but frequently other proxies are used instead, such as revenue or employment. In the case of vertical integration, the final process is used (except in special cases such as steel mills), and in the case of agriculture with joint production, a top-down classification is used. Production processes are the organizing principle for these systems, so they embody a supply-

side approach. NAICS codes also tend to differentiate between materials used. A demand-side approach is taken in the NAPCS (which is still under development) that classifies by product rather than industrial process. Note that the NAPCS is not the only product classification system in common use because international trade also requires classifying products, which is usually done via the international harmonized system. Although a single plant can be associated with only one NAICS code, if it produces multiple products, it will be associated with multiple NAPCS codes. Similarly, there is not one-to-one mapping between the aggregate economic activities covered in the supply-side systems and NAPCS. Instead, most NAPCS sections include services related to the products as well as the products themselves. The relationships between NAICS, NAPCS, and the closely matched ISIC and NACE systems are shown in Table 1.

Table 1: Table from International Labour Organization Department of Statistics with NAICS and NAPCS data (through inspection of the NAPCS codes in the published document NAPCS sections) added. The overall organization of the sectors in the classification systems is shown. Note that a single NAPCS code can include multiple economic activities, but this is not the case for NAICS or ISIC codes.

AGGREGATE ECONOMIC ACTIVITY			**SECTIONS ISIC-REV. 4**	**NAICS**	**NAPCS**
Geographic coverage			Global (United Nations), Europe (NACE)	North America: Canada, United States, Mexico	North America: Canada, United States, Mexico
Basis			Supply Side	Demand Side	Supply Side
Agriculture			A	11	67
Non-agriculture	Industry	Manufacturing	C	31-33	11, 14, 21, 24, 27, 37, 51, 67, 71
		Construction	F	23	17, 47,
		Mining and quarrying; electricity, gas and water supply	B, D, E	21, 22	61, 67
	Services	Market services (trade; transportation; accommodation and food; and business and administrative services)	G, H, I, J, K, L, M, N	42, 44-45, 51, 52, 53, 54, 55, 56, 72	24, 27, 31, 37, 41, 44, 47, 51, 54, 57, 61, 64, 67, 71, 74, 77, 81, 84,

AGGREGATE ECONOMIC ACTIVITY			SECTIONS ISIC-REV. 4	NAICS	NAPCS
		Nonmarket services (public administration; community, social and other services and activities)	O, P, Q, R, S, T, U	61, 62, 71, 81, 92	34, 87
Not elsewhere classified			X		

Some effort has been made to further bridge the gap between manufacturing processes and the economic classification systems. Todd et al. (1994) created a taxonomy that divides the processes into two categories: shaping and nonshaping. The first four levels of the division are shown in Fig. 1. The classification is based on the function of the process. This is helpful when what must be achieved is known but not how to achieve it because, presumably, it is possible to consider various alternatives that are already grouped together. Still, these taxonomies consider only a subset of manufacturing processes rather than all manufacturing in the physical industry or industry at large.

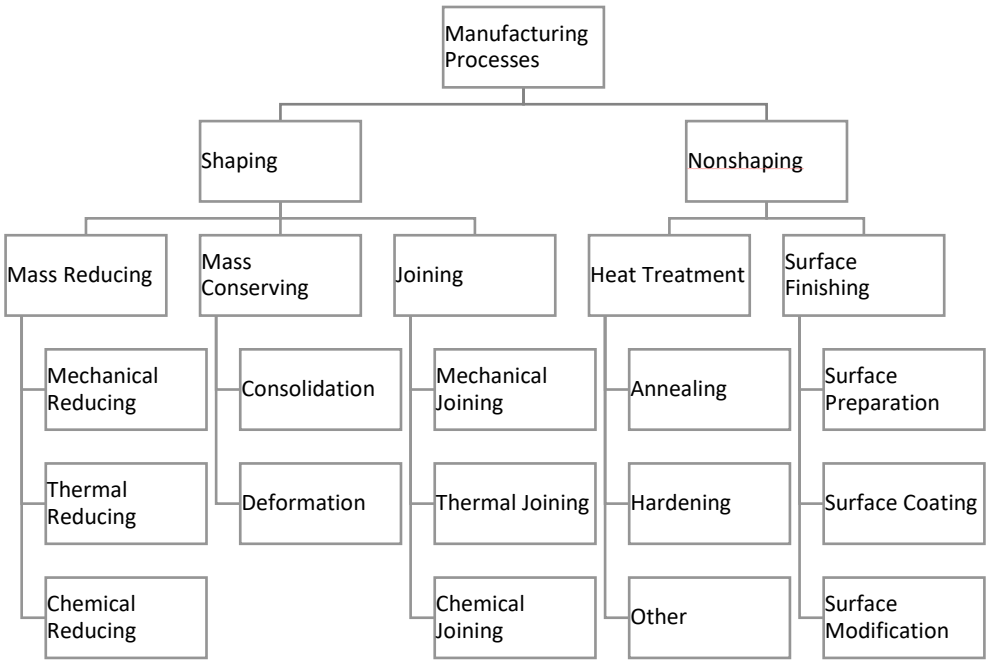

Figure 1: Taxonomy of the first four levels of the manufacturing processes (Todd et al. 1994)

Methods

Data

For this preliminary analysis, the data set used is publicly available plant-level data from the Industrial Assessment Centers (IACs). The data set is focused on medium-size plants whose electricity consumption is higher and combustion fuel use relatively lower than industry overall.

The data set is chosen to accord with an initial focus on electricity use, but it must be complemented with other data to provide a more representative look at industrial energy use.

The IACs are a U.S. Department of Energy program designed to allow medium-size plants to reduce costs and energy use by receiving customized audits while providing educational experience with real-life application to students (US Department of Energy, 2021). The program has operated since 1976 and provides additional services as well, such as waste audits, diagnosis of production issues, smart manufacturing opportunities, and cybersecurity awareness. Universities are affiliated with the program; they operate the affiliated assessment centers. Each assessment requires logging related data, which is made publicly available (https://iac.university). The IAC data set was chosen for this study because of its large size (at the time of writing, it includes 19,435 records, 5,027 since 2010 and 2252 since 2016) and the availability of plant-level data. The data represent a very small subset of industry every year. The maximum number of employees in the establishments included was 6,500, with an average of 177. According to the technical manual, facilities should have fewer than 500 employees, and since 2010 only 239 facilities included in the data set had more than 500 employees. The average plant area is 485,000 sq ft. The data set includes facilities from 45 three-digit NAICS codes, including all manufacturing subsectors. Ostensibly, the data set should include establishments with energy bills between $100,000 and $2,500,000 per year, but the data set includes values well outside of those bounds (the highest electricity cost in the data set is >$88 million; the lowest is $85). Most of the total records do not include NAICS but rather SIC codes because they were selected prior to the adoption of NAICS. Of the records that do include NAICS, more than 50% derive from 332 “Fabricated Metal Manufacturing” (~13%), 311 “Food Processing” (~12%), 326 “Plastics and Rubber Products Manufacturing” (~10%), 336 “Transportation Equipment Manufacturing” (~8%), 333 “Machinery Manufacturing” (~8%), and 325 “Chemical Manufacturing” (~7%). In more recent years, there has been a shift toward 221 “Utilities” as well.

Besides the plant area and number of employees, as mentioned, other predictor variables recorded include production level, hours, and sales. Production level is quantified in different units, however, so values are not always compatible. Two mass units are used, pounds and tons, which can be easily converted. Volumetric values are in thousands of gallons as well as bbl. Usually, bbl is a unit for barrels of oil, but in most cases in this data set, it signifies barrels of beer, which are equivalent to 31 gallons. Descriptive variables include state (location), fiscal year, and products. Response variables are 12 types of energy use, recorded in monetary and energy units, as well as water consumption and disposal and liquid, solid, and gas disposal in monetary and physical units. Of the total records, all but one include data for electricity usage, more than half include electricity demand, and 16,065 include natural gas usage. The other response variable types are sparsely populated. The only categories with more than 1,000 total records are water disposal (>3,000) and other solid nonhazardous disposal (>4,000).

According to the U.S. Bureau of Labor Statistics, in December 2020 there were 12,231,000 employees in manufacturing. The IAC data set included 63,923 employees in 2020, less than 1% of the total employment. The number of manufacturing industry establishments in 2020 was reported to be from 356,966–359,971 (see the dynamic nature of these estimates in Fig. 2). The IAC data set, which includes more than only manufacturing firms (NAICS 31–33), included 380 establishments assessed in 2020, so slightly more than 0.1%, hence an estimate can be made that from 2010–2020, each year the IAC assessed between 0.1%–0.15% of the total

establishments. Certain sectors are over-/underrepresented because of the criteria for IAC assessment, particularly those that use a large amount of energy or have a very large employee base. Electricity is a larger part of energy use than manufacturing and industry as a whole; thus, this initial exploration must be followed by more work on the industries that are not well covered by the IAC data set.

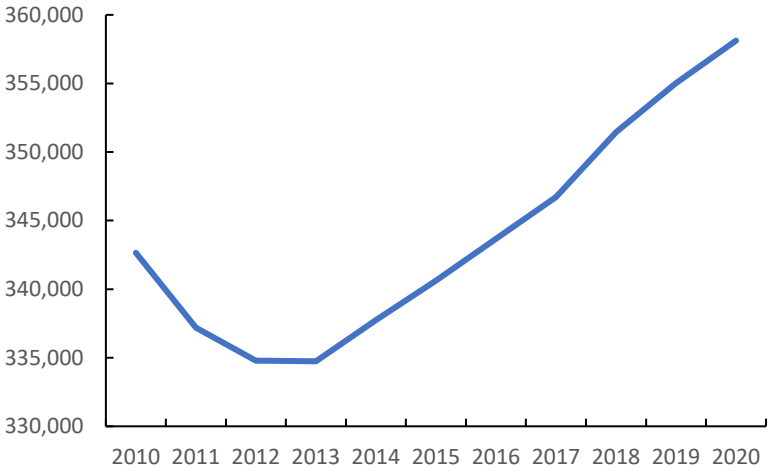

Figure 2: The number of establishments in NAICS 31–33 codes from 2010 to 2020 (preliminary figures for the first three quarters of 2020 are used) *Source*: (BLS, 2021).

Exploratory Data Analysis

Before performing the preliminary analysis, preprocessing steps were performed. First, the data set was reduced to records from 2010 and afterward because of the large scale of restructuring that took place in the manufacturing sectors after the economic crisis in 2009, as shown by the steep drop in employment, which has not recovered (Fort, Pierce, & Schott, 2018), as well as inconsistencies when using energy data during a longer time period. Data not needed for the analysis were removed; outliers were detected and removed; and inconsistencies in units were identified and resolved where possible and duplicates were removed.

The IAC data provide sales data in unadjusted nominal dollars. Sales data were adjusted according to the annual average producer price index (PPI) for manufacturing for consistency with 2019 prices (pre-COVID, final numbers) according to equation XY, where SA is adjusted sales, S is the original sales figure, and the PPI subscript refers to the annual average for the year corresponding to the observation:

$$S_A = S \times \frac{PPI_{2019}}{PPI_{Annual}}$$

Physical production data were omitted in this analysis because of the inconsistency in units (e.g., items of varying sizes, undisclosed units, both mass and volumetric data in the same subsector); therefore, quantitative data corresponding to predictor variables include sales, number of employees, hours of production, and square footage of plant area.

Because this work focuses on energy use estimates for industry, the observations labeled with NAICS codes in subsectors with values greater than 399 were studied manually. Although the IAC focuses on industry, this is loosely defined, and IAC centers are given discretion in choosing which assessments to perform. This means that 110 of the records since 2010 are in subsectors from 400+. Typically, for example, commercial space, such as distribution centers, would be included in commercial energy use estimates, as would schools, hospitals, hotels, and museums. In this analysis, three entries identified as 400+ were relabeled with industry codes, and with the exception of waste processing in code 562, the rest were excluded. NAICS codes less than 3XX included in the IAC data were mostly for wastewater treatment and water distribution, with other entries for construction, pipes, and agriculture. These were included in the analysis.

To include the sparsely populated response variables related to fuel combustion, a single entry for total combustion energy was calculated for each observation. The response variables used for the analysis were then Tc, total combustion; Ce, plant-level electricity consumption, and De, plant-level electricity demand, which corresponds to the peak load that must be planned for because of the facility's usage patterns.

The skewness for the seven variables used for the analysis was calculated, showing that data exhibit high skew and are not normally distributed for any variables. Because many procedures for outlier detection assume a normal distribution, the data were transformed to remove skewness. Removing skewness is also important because otherwise the largest values have an outsize influence on cluster formation, leading to unbalanced clusters. A Box-Cox transform was used for all variables except for hours, which was not skewed, and electricity demand and consumption, for which the log transform was used. Following these transformations, the StandardScaler function (Pedregosa, et al., 2011) was used to transform the data again to have a mean of zero and a variance of 1. Outlier detection was then performed by removing values outside 2.5 times the interquartile range. The processed data set comprises 4,248 observations.

To eliminate distortions resulting from size, ratios were calculated between each predictor and response pair; therefore, the data set used for the unsupervised exploration, described below in "Unsupervised Approaches", was the set of ratios of electricity consumption, demand, and fuel use to number of employees, operating hours, plant area, and adjusted sales.

Finally, for visualization, one additional transformation was used. Each variable in a data set represents a dimension; thus, the data set used was seven-dimensional. It is difficult to visualize seven dimensions and thus the ensuing groupings of observations. Principal component analysis (PCA) was performed to assist in visualization of the clusters and the data, although the results were not used for clustering. PCA involves finding the linear transformation of the collection of features that maximizes variance along with its orthogonal vector in two-dimensional space. Together, these two features capture most of the variance in the data set, allowing the data to be more easily visualized.

Correlations in the Data Set

First, the data were partitioned into three-digit NAICS codes, and correlation was measured between each predictor and the response variable, as shown in Fig. 3. This was used for validation of the clustering performed in the analysis—to determine whether the clusters

perform as well as the NAICS classification in terms of their correlation. Among the NAICS codes, visual inspection of the correlations showed the highest correlation between number of employees and electricity consumption. It also showed clear gradations between correlations among the different NAICS codes. In most cases, however, correlations between the predictor and response variables were only moderate. This correlation provided a baseline by using the NAICS codes, also showing which variables might be highly correlated to each other.

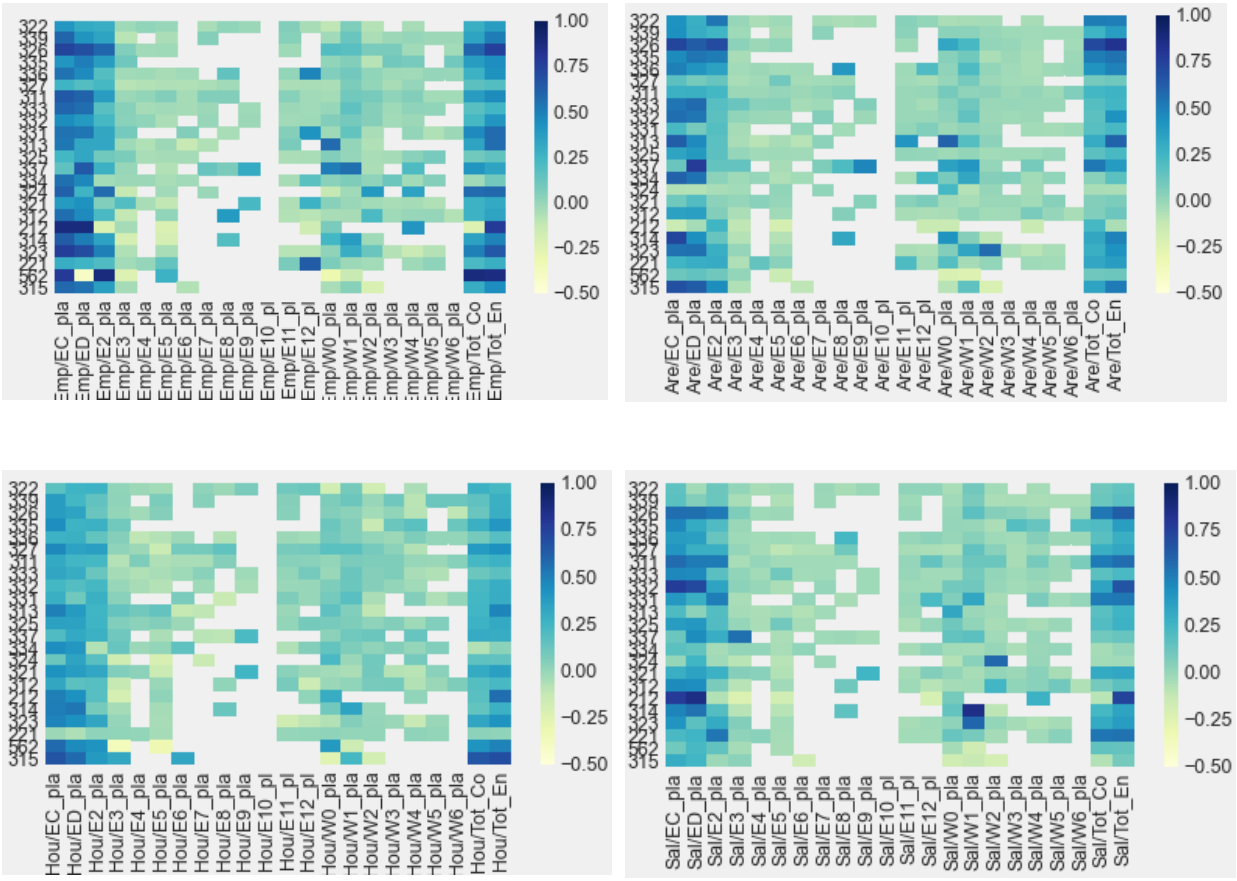

Figure 3: The Pearson correlation between values of each of the predictor variables (number of employees, plant area, hours, and sales) and all energy and material consumption variables was calculated in each 3-digit NAICS code, arranged here in heat maps where darker blues indicate high correlation, light green indicates no correlation, and yellow indicates negative correlation. Blank spaces are shown where there were not sufficient observations (< 20) in the NAICS code to calculate the correlation.

Unsupervised Approaches

The underlying goal of this exploration of plant-level data was to understand patterns in industrial facilities including but not exclusive to energy use. Correlation is a supervised method of exploration, where predictor and response variables are clear and can be specified. In unsupervised learning, patterns and clusters are sought from the data set without assumptions about the independence of variables. The most common unsupervised approach is a search for groupings, unrelated to cause and effect. This search is called clustering, and many approaches are available, although their suitability depends on characteristics of the data. In this work, hierarchical and k-means clustering approaches were used, as described next.

Hierarchical clustering is a deterministic unsupervised approach that calculates the clusters based on set criteria using a greedy algorithm; thus, a particular calculation approach always gives the same answer. The approach originally places each observation in its own cluster. Each step condenses the number of clusters until only one cluster contains all the observations. The visualization approach generally used is a dendrogram, which shows a tree with the clusters and their spacing. The speed of the approach is n^2 or n^3, so it is not particularly efficient, but it is quite transparent and allows for visualization of the entire tree. The deterministic characteristic also means that it does not need to be run multiple times.

The k-means algorithm is much faster than the hierarchical approach. The algorithm has the following steps:

1. Randomly identify k centroids for observations.
2. Assign each observation to the nearest centroid.
3. Recalculate the centroids.
4. If no observation changes the centroid, stop. Otherwise, repeat steps 1–3.

Its reliance on initially chosen centroids and its nondeterministic nature means that the algorithm must be run with many repetitions or iterations. Generally, algorithms such as the SciPy k-means have a built-in ability for the user to set the number of iterations, and the algorithm selects the clustering run with the best scores. In addition, the k value, or number of clusters, must be set by the user. The value for k can be set by partitioning the data and performing hierarchical clustering on a subset, or it can be set by calculating clusters for a range of k values and selecting the scheme that provides the most valuable results. In this work, the k-means algorithm was run 25 times on a range of clusters from 1–100. The random seed of 42 was used to allow replicability.

Results

A dendrogram of the hierarchical clustering is shown in Fig 4. Allowing the k-means to run for any number of clusters up to 100 resulted in an elbow score of 14. The silhouette score peaked at 20 clusters. Fig. 5 shows visualizations of the clusters of both k-means and hierarchical methods with 14 clusters over the PCA plot of observations. Fig. 6 shows the distribution of data in each cluster according to four of the variables studied using the k-means technique. Fig. 7 provides the same information using the hierarchical technique.

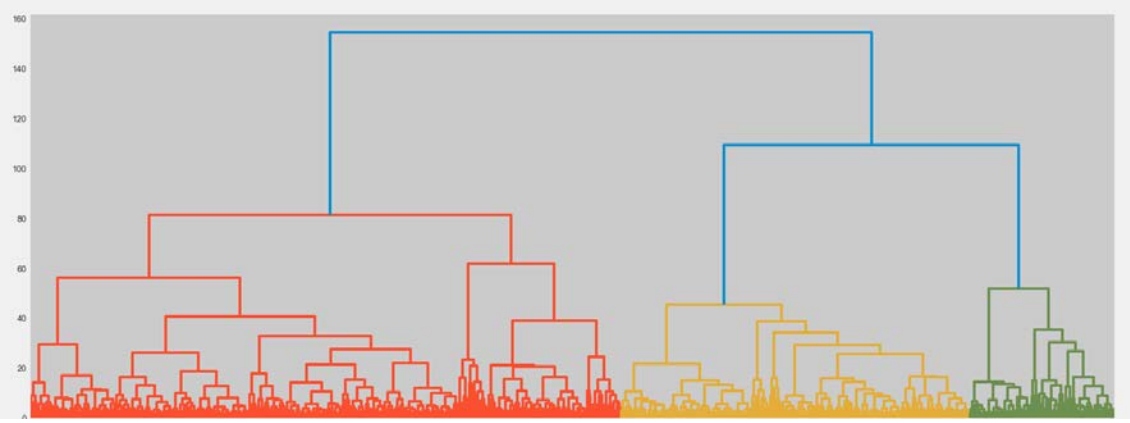

Figure 4: Dendrogram of the hierarchical clustering

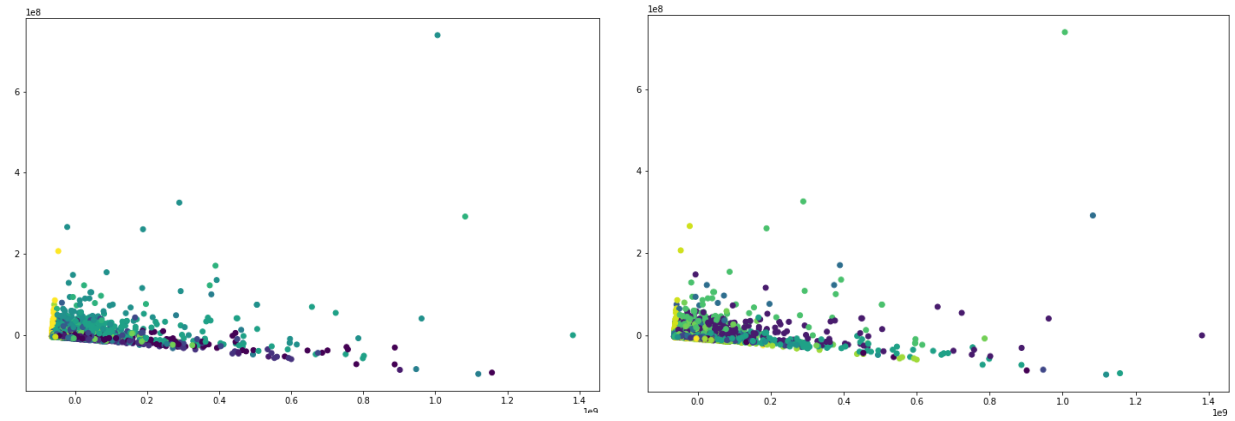

Figure 5: Hierarchical clusters are shown on the left; k-means clusters are shown on the right. Both were set at 14 clusters. The axes represent the two principal components.

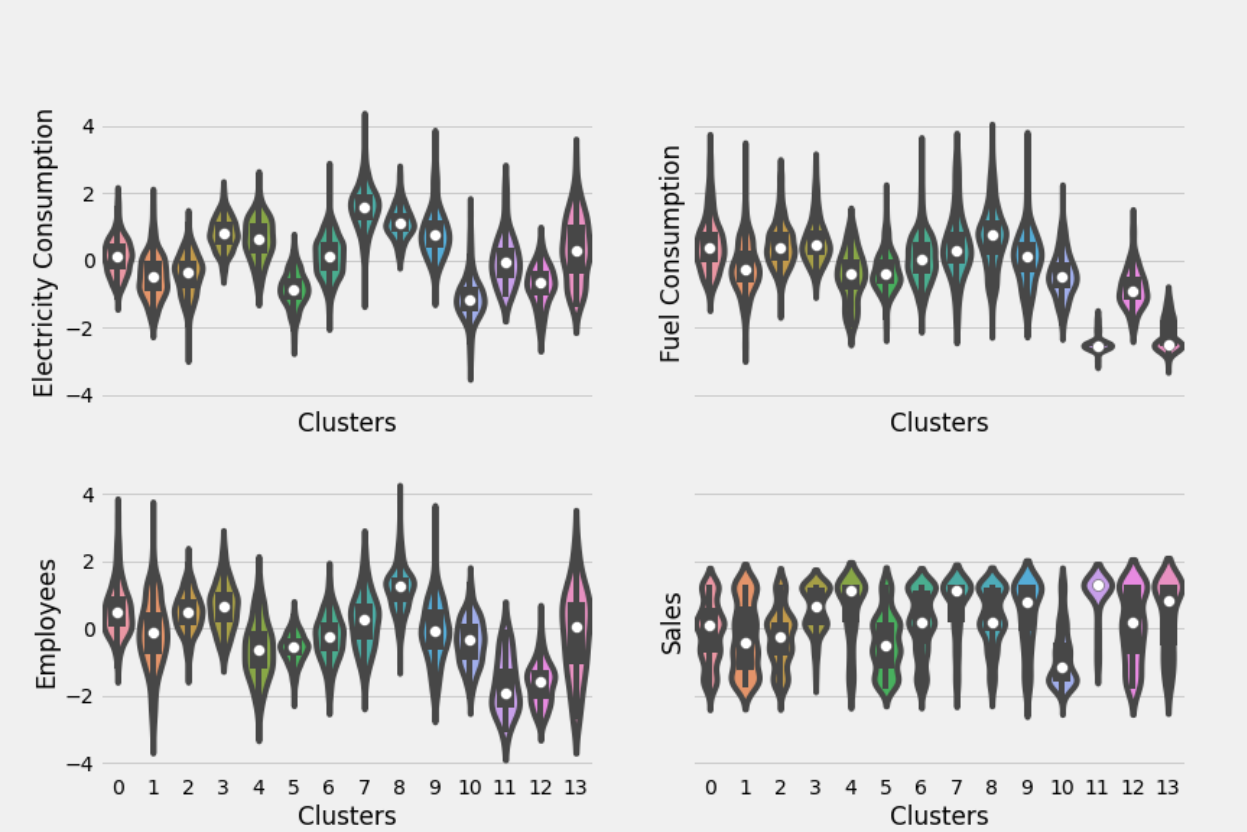

Figure 6: Violin plots of the 14 clusters formed hierarchically for electricity consumption, combustion fuel consumption, employees, and sales. Values are standardized, so violin plots show the distribution around the mean of zero in each cluster for the variables shown.

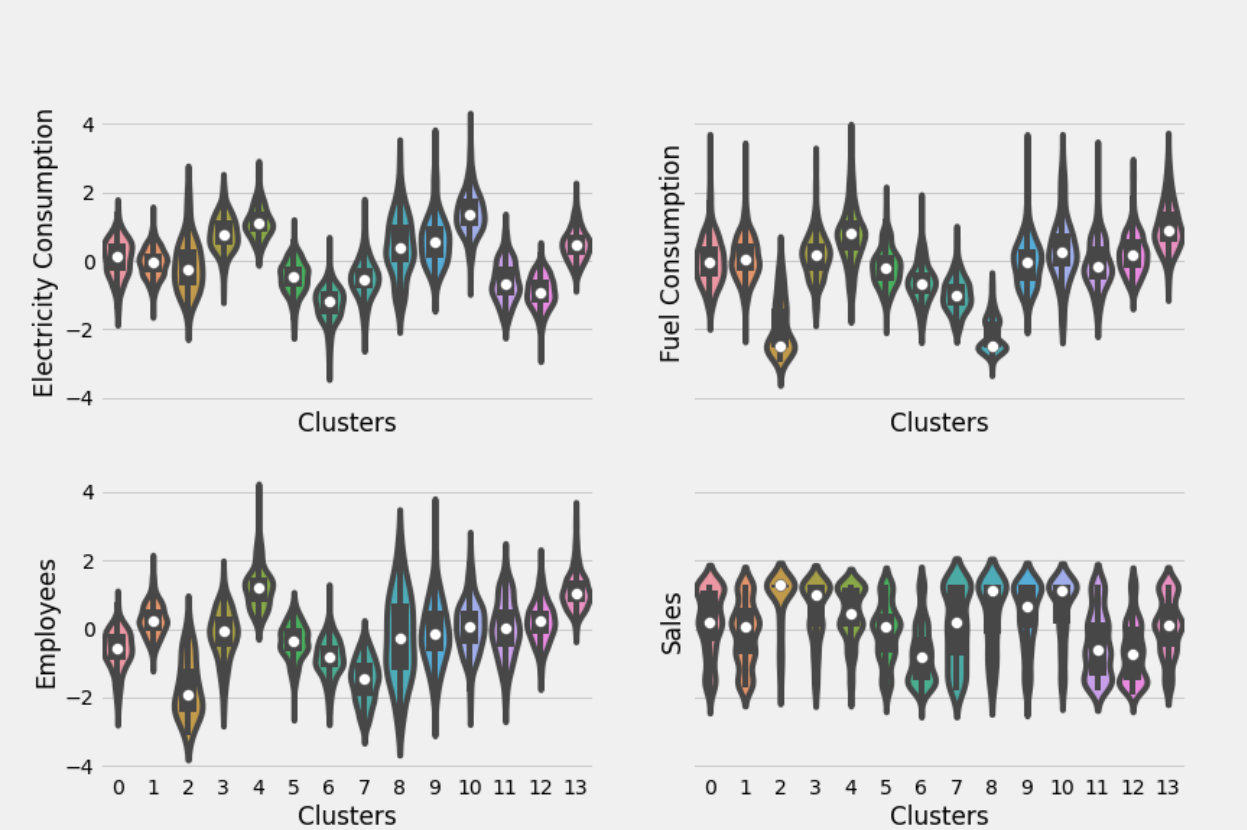

Figure 7: Violin plots of the 14 k-means clusters considered for electricity consumption, combustion fuel consumption, employees, and sales. Values are standardized, so violin plots show the distribution around the mean of zero in each cluster for the variables shown.

Discussion

The clusters formed by unsupervised exploration of the data were quite different from the NAICS codes. When both hierarchical and k-means were set at 14 clusters, even at the two-digit level, every cluster had at least three different two-digit NAICS codes (from a possible eight). Only 268 (~5%) of the 4,867 two-digit NAICS codes in the IAC data are 11, 21, 22, 23, or 56, and in the case of the cluster that has three different codes, they were 31, 32, and 33, the most represented codes. The violin plots show how the data are distributed in the clusters. The observations were grouped differently in the two methods.

The data set analyzed included 376 unique six-digit NAICS codes, and 32 three-digit NAICS codes. Correlation within three-digit NAICS codes was not particularly high (see Fig. 3). Validation of the approach was performed by measuring the correlation between the predictor and response variables using the 14 clusters shown as "best" by using the elbow test for the k-means algorithm (see Fig. 8 and Fig. 9). The much smaller number of clusters showed a similar level of correlation, as did the NAICS code classifications.

This is a promising step for identifying archetypes because the number of clusters is much fewer than for the NAICS three-digit industrial codes. Unsurprisingly, for both classifications, correlation was stronger between the electricity-related energy use and predictor variables than for the combustion energy use. This is probably because of the lower proportion of combustion as part of the total energy use in the data than in the larger industry. The two

approaches differed as to which predictor variable seemed most associated with electricity use and demand. In the case of hierarchical clustering, a higher correlation was shown between the plant area and electricity consumption and demand than the other variables, except for a limited number of clusters. In the k-means clusters, it was difficult to see a clear pattern from the heat map, but plant area and sales seemed to be more correlated than employees. This differed from the NAICS codes, where the number of employees was more closely correlated. Plant hours was not even moderately correlated in the case of either the NAICS codes or clustering. For now, hierarchical clustering seemed to be more successful in discovering clusters that show correlation between predictor and response variables.

This is only a first step in the exploration of archetypes for industrial data. Still, the IAC data represent the plants that are generally not part of energy models and for which little information is available to enable the estimation of loads. The work shows the need for plant-level data, as well as more detail in terms of physical production, to better characterize industrial energy use, particularly across a variety of production processes.

The next steps are to study the clusters to understand what characterizes them and whether they can serve as the basis for archetypes. Improving the parameters to maximize correlation is also important. When the observations were used (predictor and response variables) for clustering, the correlations were very weak, with no correlation greater than 0.5 for any cluster or variable. This means that the ratio of energy use to employees, hours, plant area, and sales was necessary for the clustering to work. Further, this approach must be combined with further study of the industries not covered well by the IAC database.

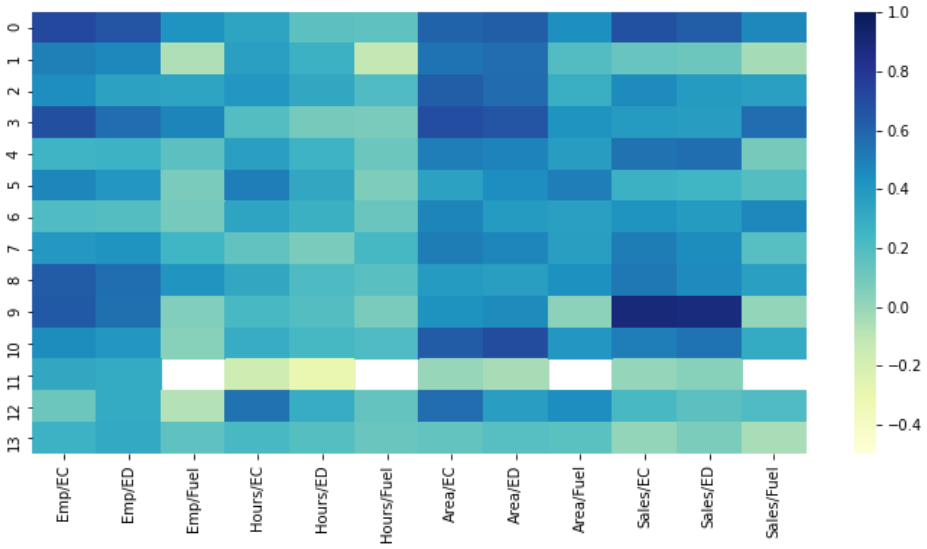

Figure 8: The Pearson correlation between predictor and response variables in the 14 clusters derived from the hierarchical clustering. Dark blue indicates high correlation, green indicates no correlation, and yellow indicates a negative correlation. White means that there were insufficient observations to calculate a correlation coefficient in the cluster. Clusters are shown on the y axis and the variables studied for correlation are shown on the x-axis.

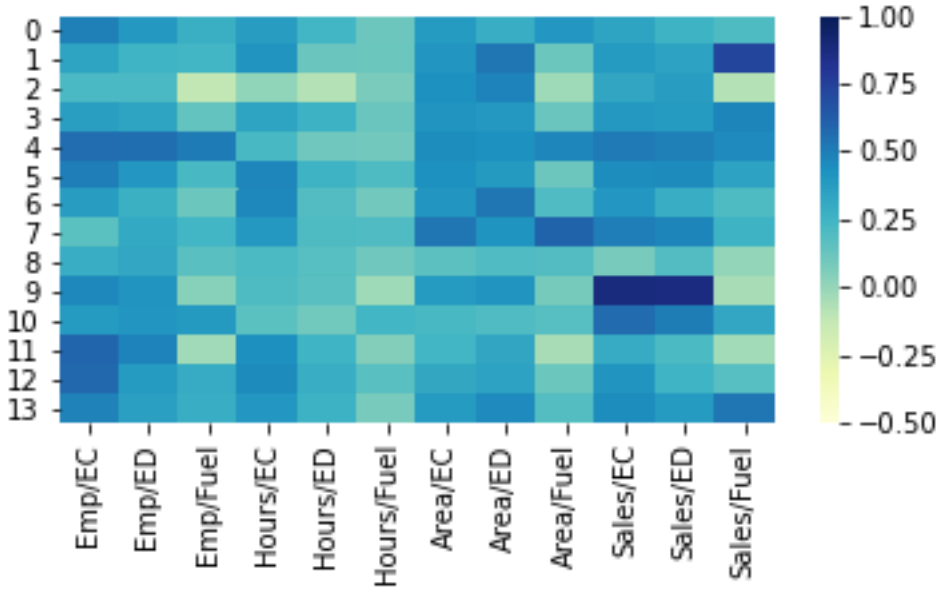

Figure 9: The Pearson correlation between predictor and response variables in 14 clusters from the k-means technique is shown in a heat map. Dark blue indicates high correlation, green indicates no correlation, and yellow indicates a negative correlation. White means that there were insufficient observations to calculate a correlation coefficient in the cluster. Clusters are shown on the y axis and the variables studied for correlation are shown on the x-axis.

Conclusions

Exploratory data analysis was conducted on the IAC database to distinguish groups with close associations between energy use and predictor variables. Unsupervised clustering algorithms allowed for grouping that has similar correlation values to NAICS despite much fewer clusters. Still, more work is needed to distinguish how to use the clusters to classify plants. Support vector machines will be examined as a classification method. The approach needs to be combined with more study of industries excluded in the data set to designate a group of industry archetypes.

Acknowledgments

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Advanced Manufacturing Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

BLS. (2021, May). Industries at a Glance - Manufacturing: NAICS 31 - 33. Retrieved from https://www.bls.gov/iag/tgs/iag31-33.htm

Fort, T. C., Pierce, J. R., & Schott, P. K. (2018). New perspectives on the decline of US manufacturing employment. Journal of Economic Perspectives, 32(2), 47-72.

"industry." Merriam-Webster.com. 2021a. https://www.merriam-webster.com (1 May 2021).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.

"manufacture." Merriam-Webster.com. 2021b. https://www.merriam-webster.com (1 May 2021).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2825-2830.

Todd, R. H., Allen, D. K., & Alting, L. (1994). Manufacturing processes reference guide. Industrial Press Inc.

US Department of Energy. (2021, March). Industrial Assessment Centers Dataset. Retrieved from https://iac.university

US Energy Information Administration. (2018). Manufacturing Energy Consumption Survey (MECS).

US Energy Information Administration. (2018). Model documentation report: Industrial demand module of the national energy modeling system. Washington.