Will Davis

Tricia Dudley

Olivia Sylvester

Time Series Analysis of US Industries

In the modern world, business decisions are being made with more and more emphasis on insights from data. The United States Census Bureau has compiled decades of economic data from hundreds of domestic industries to track trends in the US economy. In this study, we attempt to analyze this vast amount of data by through the use time series graphs and various methods of data clustering. This process included designing and implementing an algorithm that combs through these large datasets and reconfigures them based on statistical measures such as the mean, median, and standard deviations of the various measurements included in the data. The results of this statistical analysis are then displayed as time series, clustered by their similarity in our selected statistical measures.

The data sets being used in this project come from the United States Census Bureau, our sponsor, specifically historical data on US industries. There are 6 different sets of data, each with a different economic measure, for the over 80 industries for which information is available. These measurement categories are as follows: Shipments, New Orders, Unfilled Orders, Total Inventories, Inventories to Shipments, and Unfilled Orders to Shipments [2]. The task at hand is to compare these industries based on their recorded values for these varying measures from 1992 to 2021, and to attempt to observe any previously unknown similarities between industries.

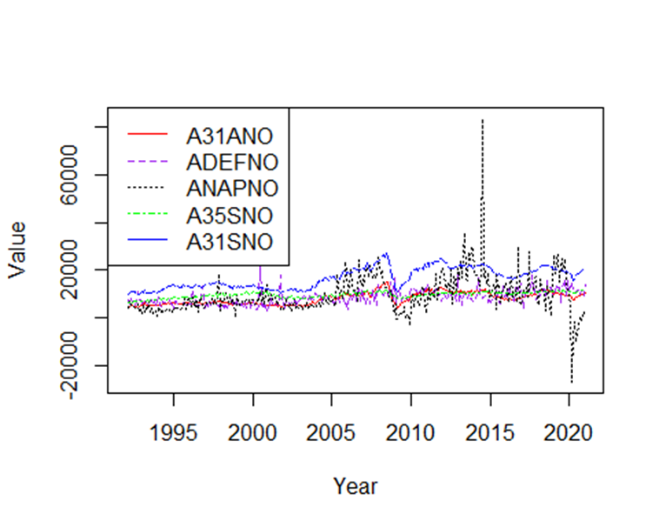
To analyze and work with the data, it is necessary to wrangle it. In other words, we need to format the data to be easily manipulated in the software program we are using for this project, RStudio. First, we downloaded all six excel sheets from the United States Census Bureau’s historical data archive. Second, as the industry names and methods of measurement are classified based on a unique ID code, we created an array that stores each ID code as an element and initialized a data frame to store the finalized wrangled data. Then, we used the industry ID array to filter data specific to each unique industry. Next, we stored the filtered data in the initialized table with each industry as a column, and data listed underneath. After that, we converted the type of data to numeric for each column. Last, a final column was added to include dates for each data entry, concluding the data wrangling process.

Our next focus was finding an optimal clustering methodology for our newly wrangled data. Clustering, the process of grouping similar sets of data based on certain similarities and criteria, is used to both cut down on the amount of data being worked with as well as serving as a means of analyzing said data [4]. This is almost always done using an algorithm due to the size of the data sets being far too vast for it to be done by hand. While there are several different ways to cluster data, we have chosen to cluster based upon three statistical measures: mean, median, and standard deviation.

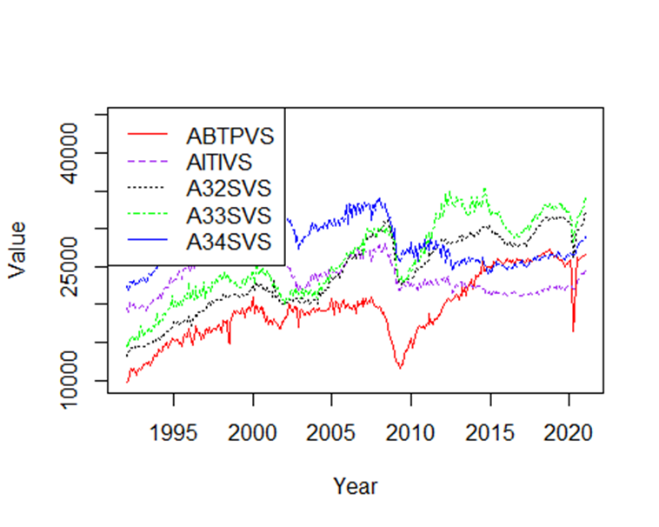
To cluster each industry many different R coding packages were used, including a package called “matrixStats.” From this package, the function “colMeans” was used to find the mean value for each industry column of the data frames. The industry means were then converted into a list and added as a new row to the bottom of their respective data frames. The columns of each data frame were then rearranged according to these mean values in ascending order. Additionally, the industries were separated into data classified as “adjusted” and “unadjusted”. “Adjusted” data is data that has been altered to lessen the effects of seasonal spikes and drops in economic behavior, such as the spike in sales of Christmas trees in December, or the drop in winter coat sales in July. “Unadjusted” data is raw data without any alterations. Next, a function was written to plot each industry column as a time series from the sorted data frames, clustering the industries into groups of five. Axis titles, a legend, different line types, and different colors were added to each cluster graphing to make the plots both easier to understand and nicer to look at.

The same methodology was utilized for clustering by the median and standard deviation values calculated for each industry within their respective tables. A new row was added to each data frame to include the median for each column, which was found using the “matrixStats” function “colMedians.” The same was done for standard deviation, utilizing the function “colSds” instead. The columns of each data frame were then put into ascending order based on their respective clustering types. These columns were again used to plot clusters of time series in groups of five, with the same elements included as with the plots based upon the mean values. Our result was many different clusters of time series reflecting all six of our economic measures, all with very similar mean, median, or standard deviation values depending on the selected statistical measure. It is important to note that there were instances where values were missing in many industries across the data sets used in this study. The code was written to include only the values that were ascertained. As such, the actual number of values used to calculate each statistical measure will vary.

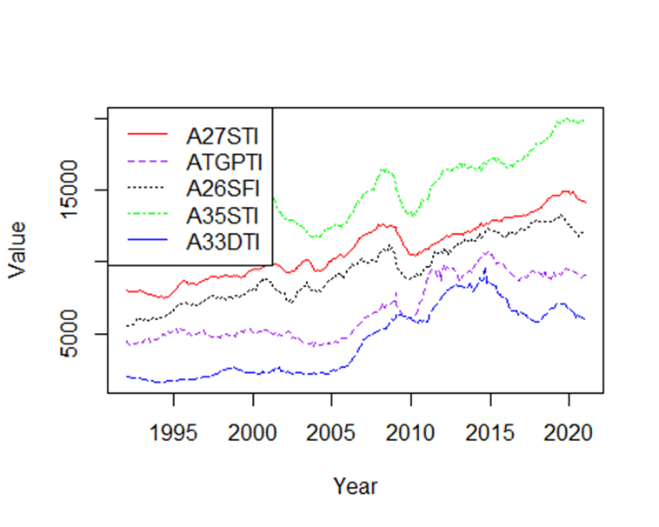
Included below are selected clusters that were particularly accurate in grouping industry time series by their respective measures:



This cluster is from the New Orders data set and includes the following industries in the order of the list in the legend, clustered by their mean value: Iron and Steel Mills and Ferroalloy and Steel Product Manufacturing from Purchased Steel, Defense Capital Goods, Nondefense Aircraft and Parts, Electrical Equipment, Appliances, and Components, Primary Metals. These industries all seem to involve steel products, which is likely the cause of their similar characteristics. Notice that excluding a few spikes and drops in the Nondefense Aircraft and Parts industry, the similar shape and proximity of the time series in the cluster is what we are looking to find.



This cluster is from the Shipments data set and includes the following industries in the order of the list in the legend, clustered by their median value: Motor Vehicle Bodies, Trailers, and Parts, Information Technology Industries, Fabricated Metal Products, Machinery, Computer and Electronics Products. These industries seem to be centered around electronic goods and the means of transporting them. While the time series in this cluster may not be very near each other, the display a very similar shape.



This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their standard deviation value: Nonmetallic Mineral Products, Turbines, Generators, and Other Power Transmission Equipment, Plastic and Rubber Products, Electrical, Equipment (Finished Goods Inventories), Appliances, and Components, Mining and Oil and Gas Field Machinery Manufacturing. While these industries do not have very similar values in their time series, the series themselves exhibit a similar shape. This could indicate that these industries are connected through what they produce, with all five experiencing demand and the lack thereof for this mysterious product.

While the clusters above were selected because of their interesting properties, there was no discernable connection across measurements or method of clustering for any of these industry clusters. However, there were certain clusters that exhibited a strong connection across the Shipments, New Orders, Unfilled Orders, and Total Inventories tables no matter the clustering method. These clusters included, but were not limited to,

Cluster #1: Nondefense Capital Goods Excluding Aircraft, Durable Goods (Materials and Supplies Inventories), Durable Goods (Work in Process Inventories), Consumer Nondurable Goods, Nondefense Capital Goods.

Cluster #2: Total Manufacturing (Work in Process Inventories), Capital Goods, Consumer Goods, Total Manufacturing (Materials and Supplies Inventories), Total Manufacturing (Finished Goods Inventories).

Cluster #3: Nondurable Goods, Durable Goods Excluding Transportation, Manufacturing with Unfulfilled Orders, Durable Goods Excluding Defense, Durable Goods.

Note that the various time series graphs associated with these clusters have been included in the Appendix, and that the ordering of these clusters is not the same for each measurement. It may vary slightly.

Not only were these clusters consistently the similar, both in shape and proximity, but they were always among the last clusters generated by the code, meaning that these industries had some of the largest mean, median, and standard deviation values. Why? The key lies in the fact that these industries are not really industries at all. They are measures of all the industries included as a whole. Many of these “industries” are even the same measurement minus a key element. For example, in Cluster #2 there are three different versions of “Total Manufacturing” dependent upon different aspects of the measure. You may then ask why we are still interested in these clusters if they do not give insight into the connections of standalone industries. While this is true, the main reason is that these clusters provide proof that our algorithms for creating clusters work well on a broad scale. When we began this research, we had no idea that there were data points that were intended to track economic activity across more than one industry, let alone many of them. It was only after we implemented our methodologies that we found this connection, which was exactly the point of clustering this time series data in the first place. This is the true power of the cluster, the ability to take large data sets that cannot be combed through by hand and find these connections that may not be obvious from the start.

While our clustering led to a more direct evaluation of the industries included in our data, one may argue that the amount of time series clusters that were created in this study may still lead to a large amount of time necessary to ascertain the usefulness of a given cluster. Our solution has not been implemented here but would involve fitting a linear regression model to each cluster and calculating an statistic to examine the variability of each model created. We would then filter out clusters that do not meet our minimum variability requirements, which could be determined by a client for example, leaving us with only a few truly optimal clusters. These “optimal clusters” would be selected by the algorithm, eliminating the necessity of manual selection at the end of our clustering process.

To conclude, we consider the methods of clustering we employed to be successful. There were many clusters that we did not include in this report, as they did not present any useful information about the industries they contained. This is to be expected, as not all industries will have hidden connections waiting to be found. Still, we did find some clusters that were able to show us the characteristics we were looking for and allowed us to make inferences about those connections. While the methods we employed in this study were rudimentary, there are more and more advanced methods of clustering being developed every day to find even more deeply hidden connections across the vastness of our national economy. As the efficiency of these algorithms increases, the ability to make effective business and economic policy decisions increases as well. These decisions and their consequences will allow us to strengthen our economy across all industries and build toward a better tomorrow.

References

[1] “Hayes, A. (2022, February 8). *Understanding time series*. Investopedia. Retrieved March 24, 2022, from https://www.investopedia.com/terms/t/timeseries.asp”

[2] “US Census Bureau: Adriana Stoica (M3 Section Chief). (2009, September 30). *US Census Bureau Manufacturer' shipments, inventories, and orders main page*. United States Census Bureau. Retrieved March 29, 2022, from https://www.census.gov/manufacturing/m3/historical\_data/index.html”

[3] “*Data wrangling: What it is & why it's important*. Business Insights Blog. (2021, January 19). Retrieved March 29, 2022, from https://online.hbs.edu/blog/post/data-wrangling”

[4] “*What is clustering?* NVIDIA Data Science Glossary. (n.d.). Retrieved March 31, 2022, from https://www.nvidia.com/en-us/glossary/data-science/clustering/”

[5] “PIC Math is a program of the Mathematical Association of America (MAA) and the Society

for Industrial and Applied Mathematics (SIAM). Support is provided by the National Science Foundation (NSF grant DMS-1722275)”

Appendix

Included below are the graphs associated with Cluster #1, Cluster #2, and Cluster #3:

Shipments by Mean (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

New Orders by Mean (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Unfilled Orders by Mean (Adjusted)

Chart

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Total Inventories by Mean (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Shipments by Median (Adjusted)

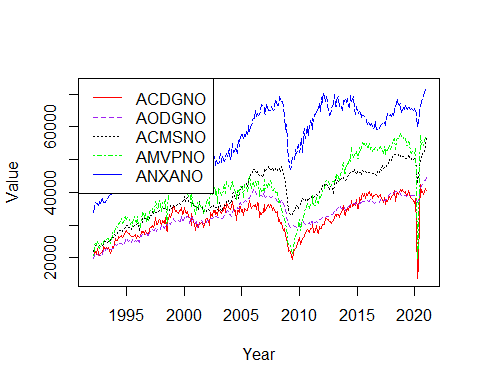
Chart, histogram

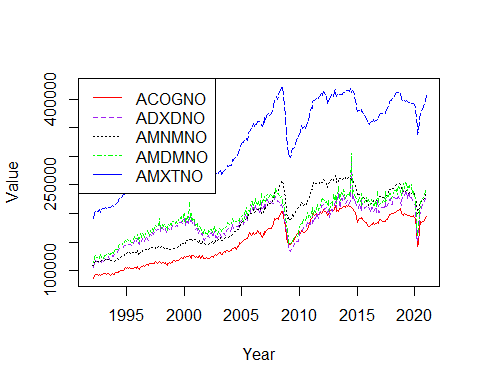
Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

New Orders by Median (Adjusted)

 Chart, histogram

Description automatically generated 

Unfilled Orders by Median (Adjusted)

Chart

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Total Inventories by Median (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Shipments by Standard Deviation (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

New Orders by Standard Deviation (Adjusted)

Chart, histogram

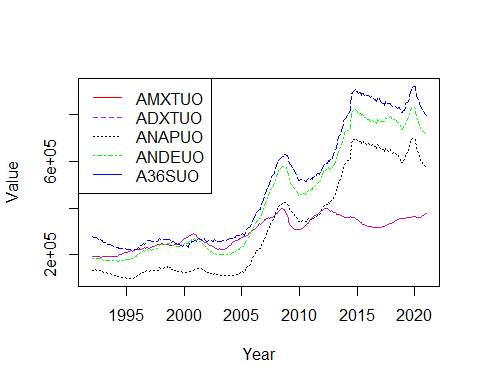
Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Unfilled Orders by Standard Deviation (Adjusted)

Chart, line chart

Description automatically generated  Chart, histogram

Description automatically generated

Total Inventories by Standard Deviation (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated