Adjustments Made Easy: A Look at Clustering Time Series from the US Census Bureau Manufacturers' Shipments, Inventories, and Orders

US Census Bureau, SUNY Fredonia

Matthew Barton, Kampbell Howard, Allyson Hineman, Ryan Plumer

Spring 2022

Department of Mathematical Sciences, SUNY Fredonia
Industry Liaison: Dr. James Livsey
Supervisor: Dr. Lan Cheng

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)

1 Abstract

Our work was done in collaboration with Dr. James Livsey who specifically works in the center for statistical research and methodology for the US Census Bureau. We took time series data released from the Economic Directorate to analyze correlations between differing industries. Using correlations between series, we made suggestions on which series could be jointly modeled. This was done with the intention of selecting candidate series for multivariate seasonal adjustment, as opposed to the univariate way that data is adjusted now. Examining in a multivariate fashion allows better approximations for data, as well as takes other dynamics into account to relate industries to one another. The results outlined in this paper provide thirteen suggested clusters using dynamic time warping methodology that the Economic Directorate can utilize.

2 Introduction

A time series is a collection of repeated measurements at regular time intervals. Plotting a time series over a period of time allows for data to be visualized and can be further utilized by analysts for select purposes. Many industries use time series analytics to predict relationships between their products and for forecasting to see how their industry will grow with time. For example, the federal government collects monthly data on retail sales, consumer prices, employment, and gross domestic product to determine if there are trends within the data that industries can safely predict and prepare for. Time series data is a major component for stock market analysis, yield projections and inventory studies amongst other topics.

Figure 1: Sample clustered time series plot of unadjusted data from durable goods, manufacturing excluding defense, manufacturing with unfilled orders, manufacturing excluding transport, and durable goods excluding defense industries in millions of dollars.

The U.S. Census Bureau is one of the leading federal industries in collecting time series data and publishing for public consumption. As a part of the United States Federal government, the U.S. Census Bureau is famously known for conducting a national census every ten years where the data is ultimately collected to determine congressional seats, redistricting, and further economic activity. A lesser known, but still equally important data collection point, is the tracking of business and economic activity of industries. By recording the economic value of all industries monthly, the U.S. Census Bureau can provide historical data that shows the trends of data as the economy changes and hopefully grows. A piece of this collection deals with Manufacturers' Shipments, Inventories and Orders which tracks the economic conditions in the domestic manufacturing sector and provides a basis for further business trends. Some of the popular economic variables recorded in the survey include total inventory, new orders, value

of shipments, and unfilled orders. Additional data has been collected for other variables, however, they are out of the scope of this project but can be considered for future work. The industries that this data is collected from is indexed by the North American Industry Classification System (NAICS) (see appendix Figure 7).

Time series clustering is the method of partitioning specific time series data into groups based upon the similarities of their values and/or distance from one another. When clustering it is important to take into account: (1) similarities in time series data, (2) how to compress the data or reduce dimension, and (3) what algorithm to apply. A popular measure of similarity uses a Euclidean distance where length of line segments between points from different time series are calculated. By this method, data with the smallest distances between their respective points are grouped together to be clustered. The major fault in this approach is that it fails to produce effective similarity measurements when the data is not the same length or if there is a distortion in their axes. To amend this issue, dynamic time warping (DTW) is often utilized by data analysts. DTW solves the aforementioned issues, and also cleanly clusters data which is later represented in a dendrogram. By inspection and graphing clustered time series data onto a single plot, analysts are able to validate the correlation between time series data to be jointly modeled.

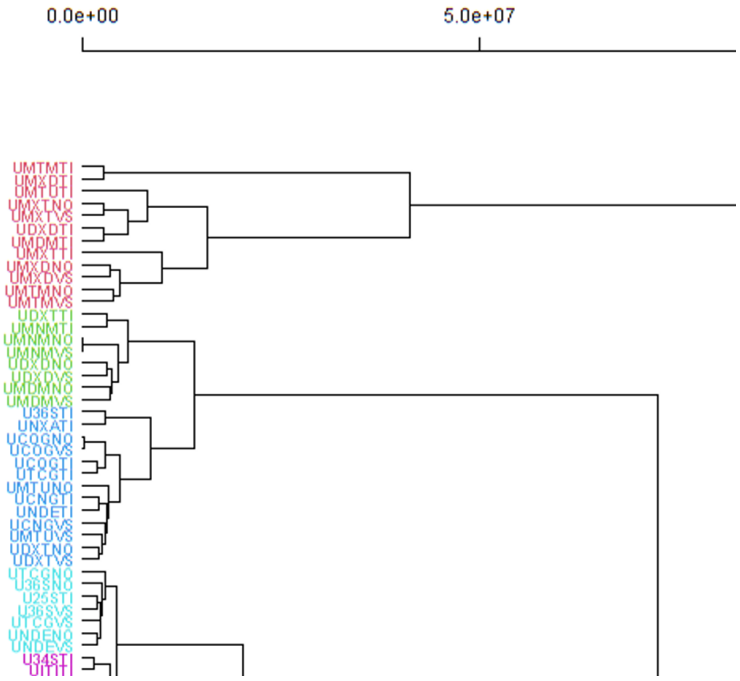

Figure 2: Section of completed dendrogram in millions of dollars from Dynamic Time Warping clustering method that suggests which time series should be jointly modeled by inspection.

Seasonal adjustment is a method that tries to remove predictably cyclical patterns. For example, around December every year, holiday shopping spikes economic data in a predictable fashion. From this, adjustments can be made to accurately predict the trend and account for these "spikes". After noticing these trends and accounting for them, data becomes clearer and net trends are showcased. This is done by evaluating time series one at a time, in a univariate approach. This approach ignores cross-correlations, hence this study researches candidate series for a multivariate approach. Multivariate models borrow strength from all jointly modeled time series to improve estimation of trends and seasonal effects. Utilizing DTW and plotting the correlated data series with one another we are able to create multivariate time series clusters of more than two industries.

The overall goal of this project will be to take the time series data that is released from the Economic Directorate provided by the United States Census Bureau and produce suggestions to analysts about series that could be jointly modeled together.

3 Approach and Methodology

We analyzed manufacturers' shipments, inventories, and orders historical unadjusted time series data obtained from `census.gov`, released by the Economic Directorate. The data was stored in a series of 6 Excel spreadsheets, with thousands of rows of data to sort for analysis. Our team created a GitHub Repository to work in a collaborative coding environment (see Appendix Figure 8). Once R-Studio was integrated with Git, we were able to push updates to the repository as they were created. Throughout the course of this project, our team was simultaneously learning about time series while getting acquainted with the syntax of the R programming language. Many of us had taken introductory level courses that taught basic coding using R, but this project involved more complex programming. We were often able to debug with a few quick Google searches, but some issues proved to be more complicated than others.

As we got more accustomed to the coding structure, we started the data wrangling process, which involved manipulating the format of the data frames to convert them into time series objects. Initially, every row in the data frame represented one year between 1992 and 2022. We converted the data into time series objects and then transformed each data frame into wide format, where every observation for each time series is recorded in one row. Certain years were missing observations, which led us to filter only for the years 1993 to 2019, to exclude both the missing values from early years and the effects of COVID-19 on these industries. Additionally, we found the U34GVS series for the unadjusted value of shipments of Semiconductor and Related Devices Manufacturing was not completely published, and therefore, removed from our data frame. Upon completion of the data wrangling process, we began the analysis and clustering process.

Analysis and interpretation of related time series was our next task. We utilized DTW to cluster our data, with methods explained above. The two main R commands that we used to implement this technique included the distance matrix computation and hierarchical clustering. The distance matrix computation calculates and stores a matrix that computes the distances between the rows of a data matrix, using the specified DTW method.

The DTW algorithm uses the following equation to match two time series that can differ in variables including time, speed, and length. To begin, consider two time series x and x' of respective lengths n and m. Here, all elements x_i and x'_j are assumed to lie in the same p-dimensional space and exact timestamps at which observations occur are disregarded: only their ordering matters. In the following, a path π of length K is a sequence of K index pairs $((i_0, j_0), \cdots, (i_{K-1}, j_{K-1}))$. Therefore, DTW between x and x' is formulated as the following optimization problem:

$$DTW(x, x') = \min_{\pi \in \mathcal{A}(x,x')} \sqrt{\sum_{(i,j)\in\pi} d(x_i, x'_j)^2}$$

where $\mathcal{A}(x, x')$ is the set of all admissible paths ,i.e., the set of paths π such that: $\pi = [\pi_0, \cdots, \pi_K]$ is a path that satisfies the following properties:

- π is a sequence $[\pi_0, \cdots, \pi_{K-1}]$ of index pairs $\pi_k = (i_k, j_k)$ with $0 \leq i_k < n$ and $0 \leq j_k < m$
- $\pi_0 = (0, 0)$ and $\pi_{K-1} = (n-1, m-1)$
- for all $k > 0, \pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows
    - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
    - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

This DTW equation calculates the distance between each element in X and its nearest point in Y by taking the square root of the sum of the squared distances. [7, 6]

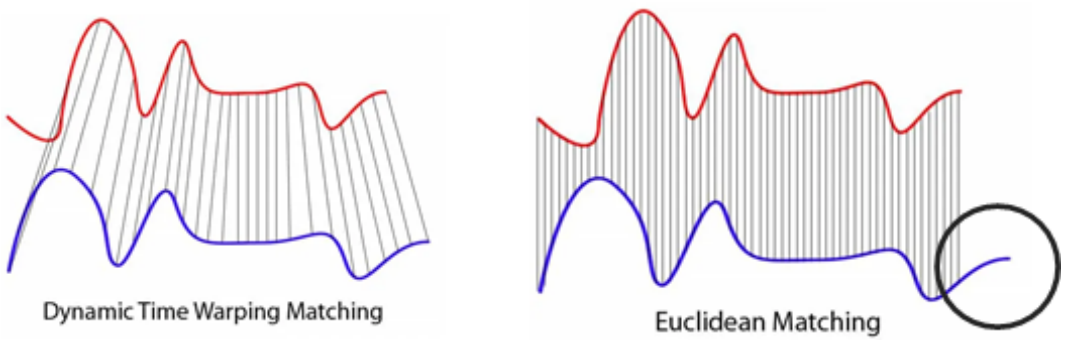

Figure 3: The image above displays two different matching methods for two time series curves, DTW vs Euclidean.

Notice that the blue curve is longer than the red curve. When applying the one-to-one Euclidean matching method, part of the data from the blue curve is not accounted for. To resolve this problem, DTW matching uses a one-to-many match, and thus utilizes the entire time series despite their differing lengths. Therefore, we applied the DTW matching method to take into account time series that follow similar trends but at different times, according to their distances. Hierarchical clustering creates meaningful clusters based on the distances that were previously computed. Once clusters have been created, it is best to display the results graphically.

Dendrograms are common graphical methods used to display hierarchical clustering, which show the hierarchical relationships between different time series objects. Thus, we see many stacked branches, which are called clades. These clades break down into smaller branches, where the lowest level represents individual time series, and the highest level represents large clusters of related time series. Looking at our data, the dendrogram was so large that interpreting results was a difficult task. To combat this problem, our group decided to make certain assumptions to narrow down the number of time series that we would be examining. Additional information regarding these assumptions and the corresponding analysis is given in the Results section.

4 Results

By aforementioned Dynamic Time Warping techniques, a full dendrogram including all the completed unadjusted times series data for one survey from the US Census Bureau Economic Directorate was created (see Appendix Figure 19). This allowed for thirteen clusters by inspection to be created that could be graphed with one another for potential correlation. As seen from the dendrogram, we decided to omit a very large chunk of time series data under the assumption that total inventory, value of shipments and new orders were relatively correlated. Thus, the rest of the paper will highlight 3 "strong" clusters for further analysis, including Cluster 2, Cluster 4, and Cluster 13. Refer to the Appendix for the remaining clusters.

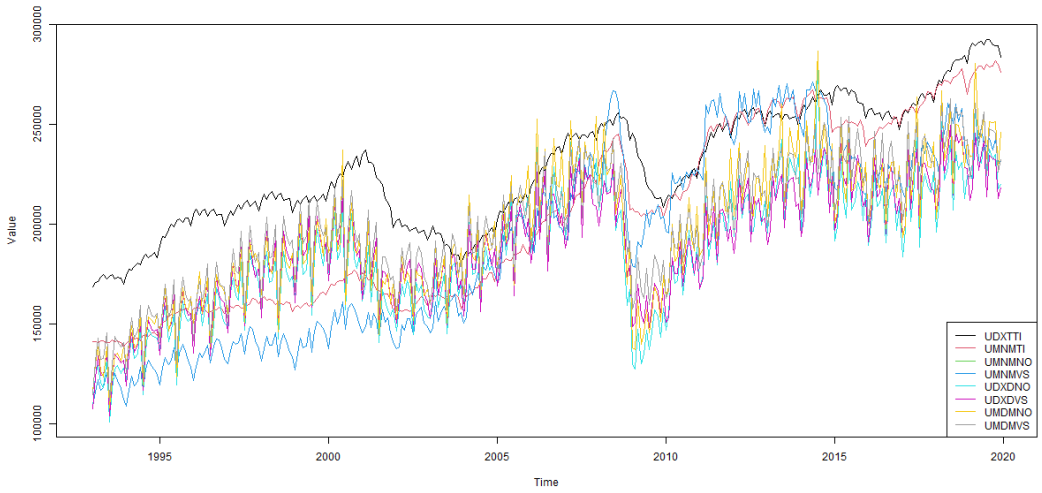

Figure 4: Cluster 2 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

Cluster 2 deals with durable goods excluding transportation total inventory, non-durable goods total inventory/new orders/value of shipments, durable goods excluding defense new inventories/value of shipments, and durable goods new orders/value of shipments. From the plot it is interesting to observe similar patterns with regards to non-durable and durable goods–that they both seem to steadily increase until a sharp drop around 2008, which can be attributed to the 2007-2008 economic recession in the United States. From economic data such as this it can also be observed that the economy did bounce back; however, the rate at which these industries were growing was observed to be much less around 2019 where it seemed to be stagnant.

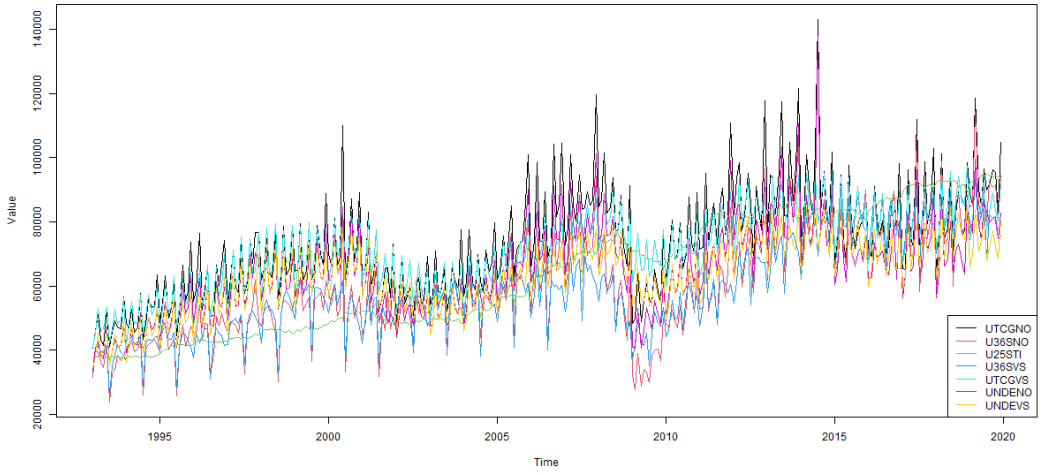

Figure 5: Cluster 4 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

Cluster 4 deals with capital goods new orders, transportation equipment new orders, chemical products total inventory, transportation equipment value of shipments, capital goods value of shipments, and non-defense capital goods new orders/value of shipments. This plot compiles some of the most "seemingly unrelated" industries including transportation equipment, capital goods, and chemical products. While these might seem obscure to cluster together, their graphical trends are undoubtedly very related and follow almost the exact same pattern. Again, the 2007-2008 economic recession can be observed with the sharp depression during that time, and the growth of the industries has seemed to lessen in recent years.

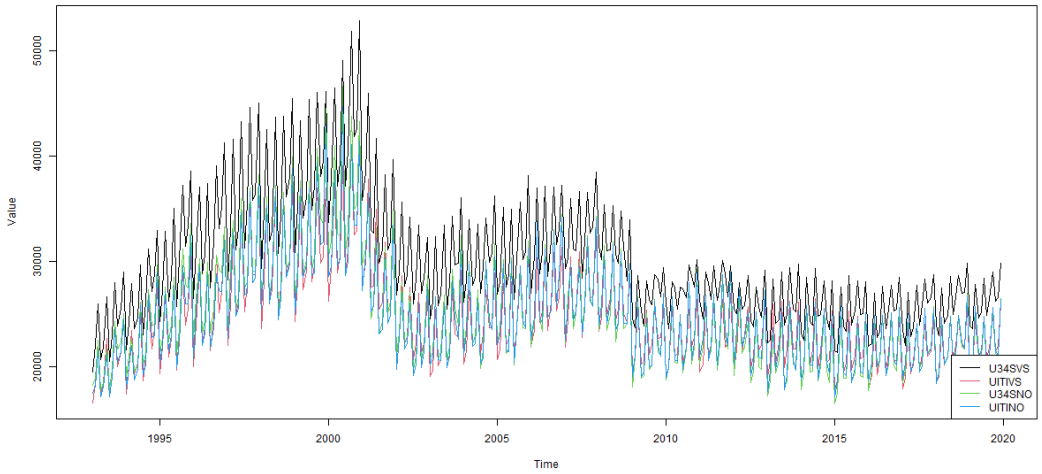

Figure 6: Cluster 13 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

Cluster 13 deals with computer and electronic products value of shipments/new orders, and information technology industries value of shipments/new orders. This plot follows a different trend compared to the last two and offers a clear reasoning with regards to the industries it deals with. All dealing with computers and technology, it would be safe to assume that the industries grow and fall together (as seen within the plot). As time increased until the year 2000 technology became more popular, and thus the industries grew. However, with rising prices and less demand the relevant industries fell and remained stagnant until the economic recession in 2007 where it dropped again.

5 Limitations to Solution

There are limitations to our solution for the Economic Directorate. When we were looking at data we windowed it to the Manufacturers' Shipments, Inventories, & Orders and more specifically, Total Inventories, Value of Shipments and New Orders. Therefore, we do not have full clusters for other data types. However, one could look into the other time series data that is released from the US Census Bureau using our methods. Another limitation to our solution is the fact that our hierarchical clustering results were not analyzed for the some of the time series that we allowed into our DTW algorithm. Therefore, the windowed data we analyzed was not fully clustered because they were all so close on the dendrogram (see Figure 19). One could possibly look into how to further cluster these time series.

6 Conclusion and Future Work

Through data analysis, data wrangling, and relevant coding in R-studio we have provided thirteen clusters by DTW methodology that the Economic Directorate can use to jointly model and forecast industries. A dendrogram was created that included every unadjusted data set in the US Census Bureau Economic Directorate and suggested clusters were chosen by inspection. From this, unadjusted multivariate models were created by plotting the clustered time series with one another.

No matter the industry being analyzed it was seen that the suggested clusters from the DTW dendrogram provided clear and reliable information. By plotting and verifying the time series data, it can be concluded that there are seemingly endless ways to manipulate the data and create relevant clusters just by inspection. Some possible areas for future work include manipulating these suggested clusters to create "stronger" associations, compiling more time series data from different directorates, and creating models for the data that analysts can utilize for forecasting. Future work could also include testing these unadjusted clusters against their adjusted counterparts to see how the clusters align, and if there are some alterations to the current clusters that could be made. Other possible areas for future work also include clustering using other methods such as Kernel-K means, KShape, or even other k-means clusters including Euclidean Matching or Soft-DTW.

7 Appendix

Figure 7: QR code to North American Industry Classification System (NAICS) to identify industry abbreviations.

Figure 8: QR code to Github repository where coding can be analyzed and reviewed.

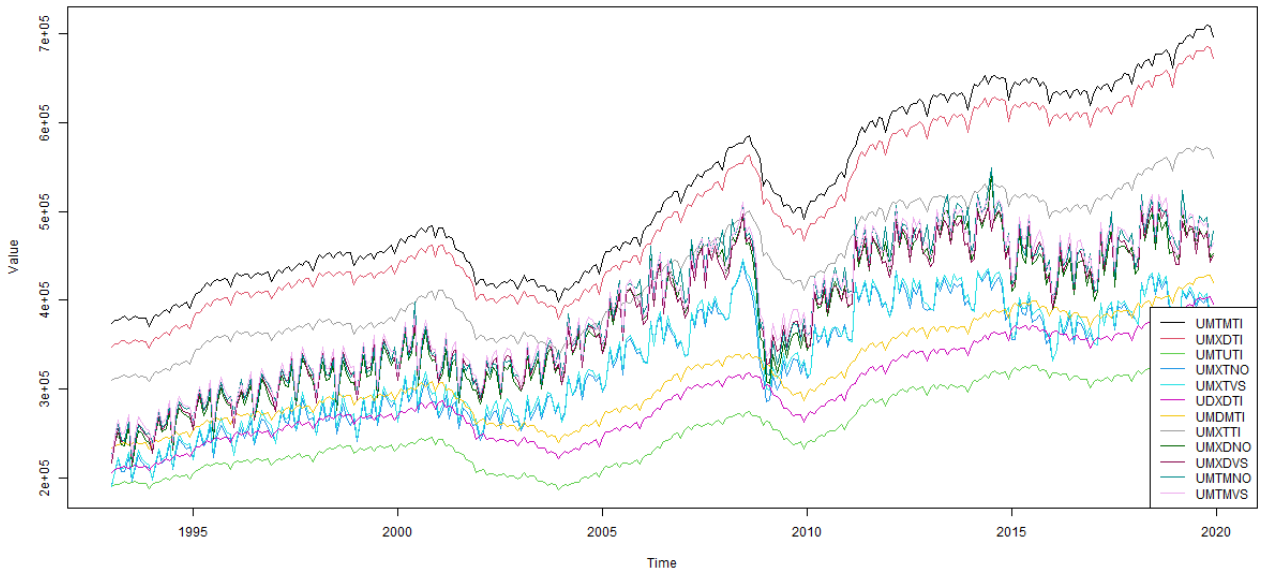

Figure 9: Cluster 1 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

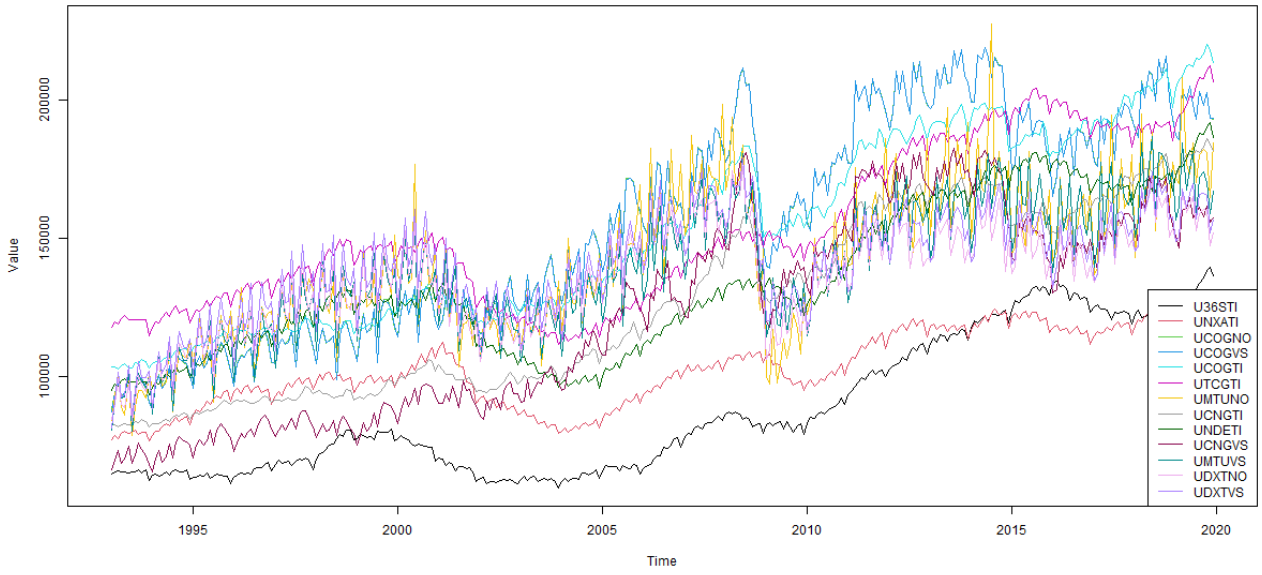

Figure 10: Cluster 3 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

Figure 11: Cluster 5 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

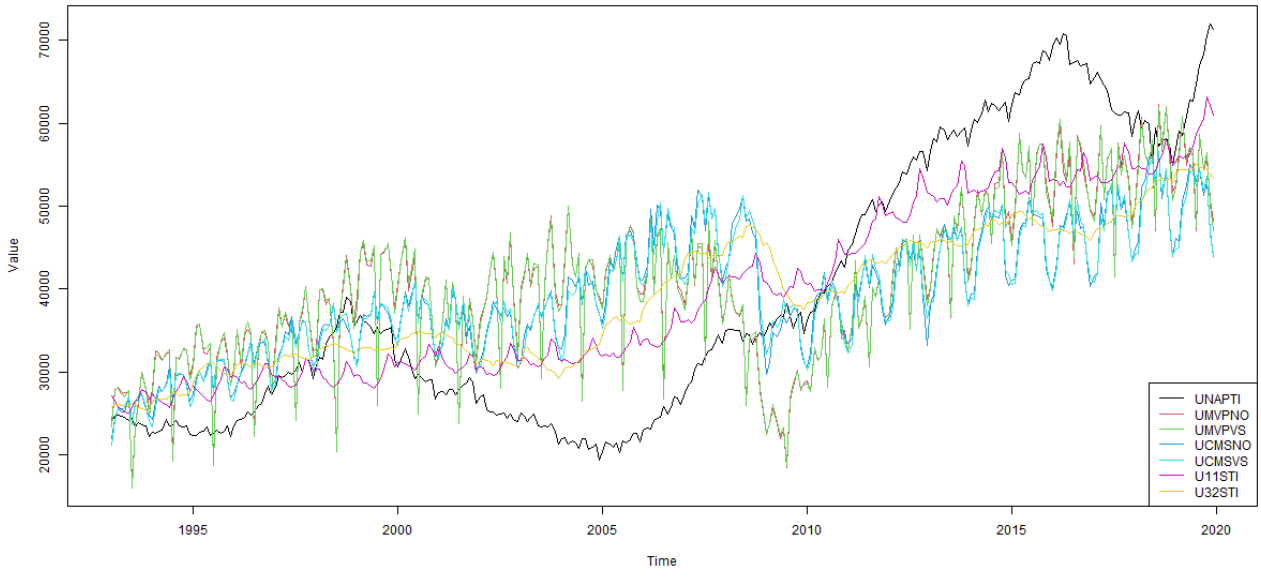

Figure 12: Cluster 6 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

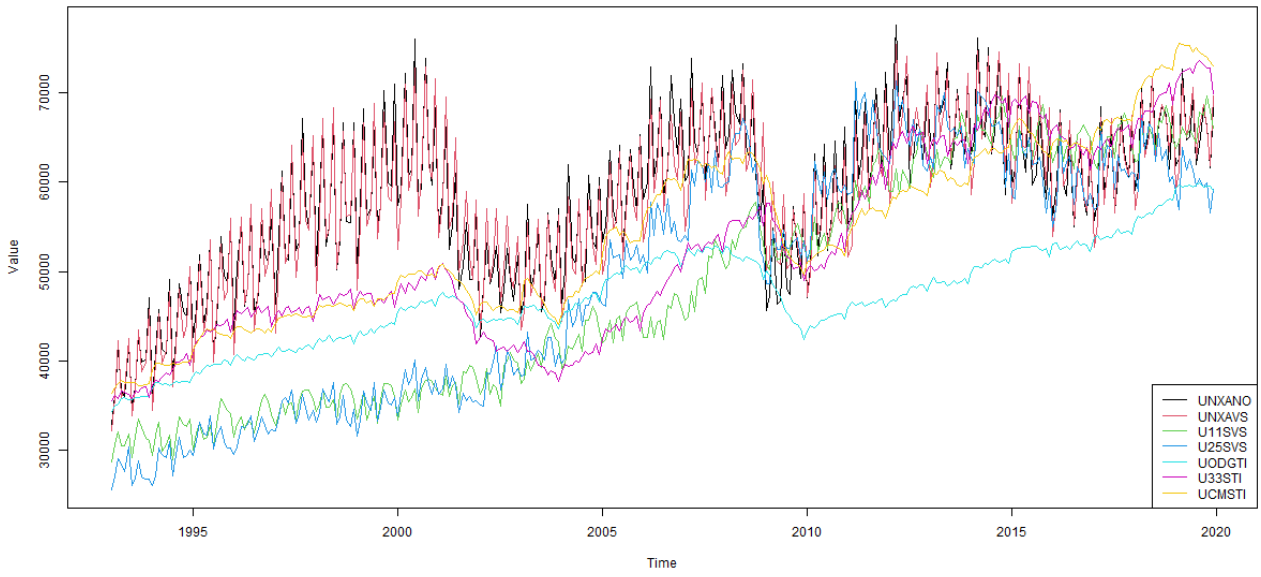

Figure 13: Cluster 7 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

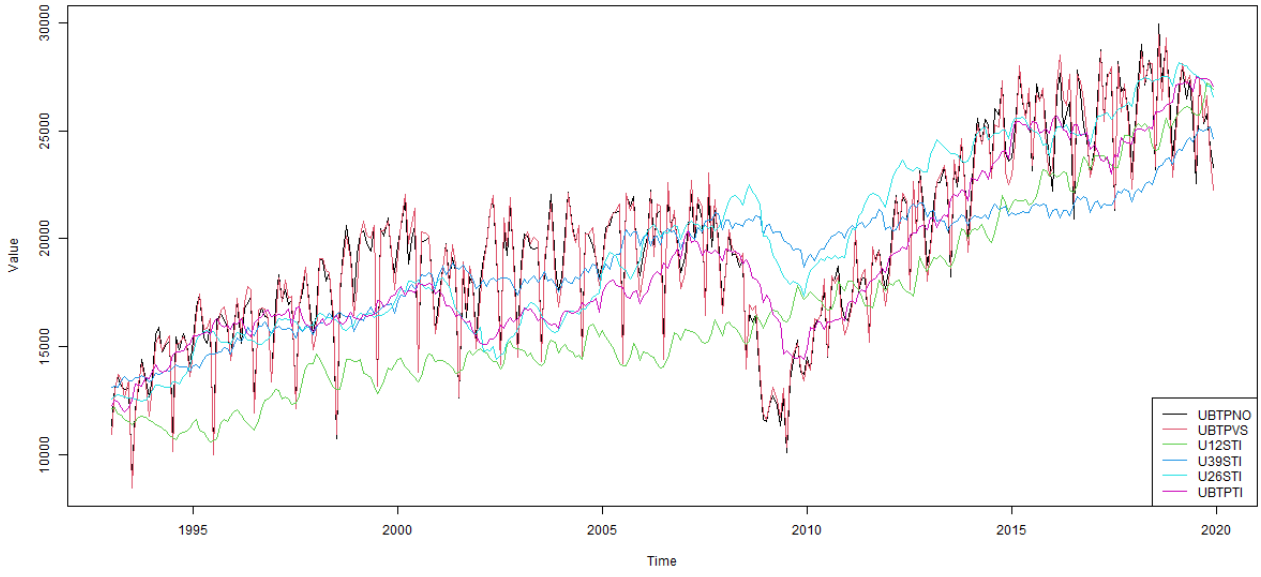

Figure 14: Cluster 8 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

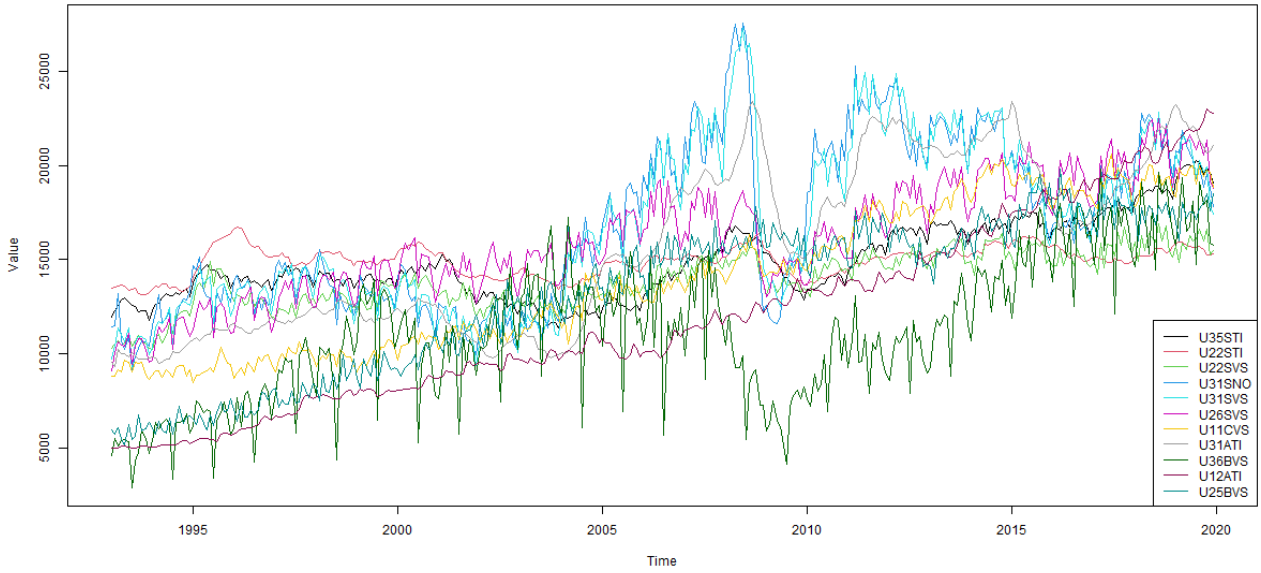

Figure 15: Cluster 9 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

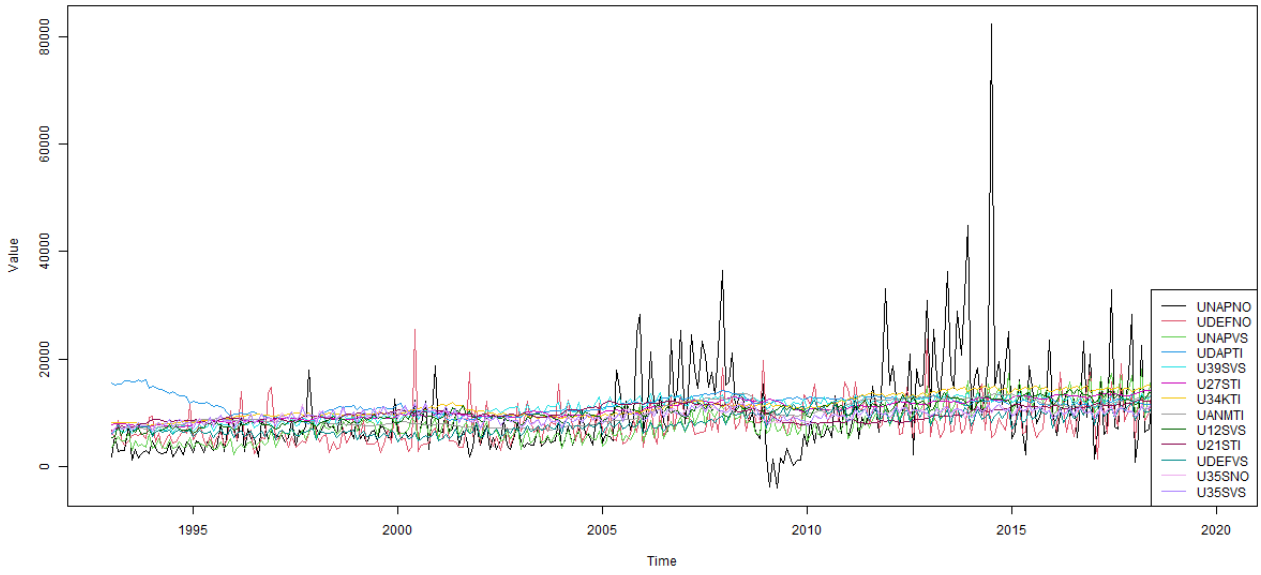

Figure 16: Cluster 10 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

Figure 17: Cluster 11 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

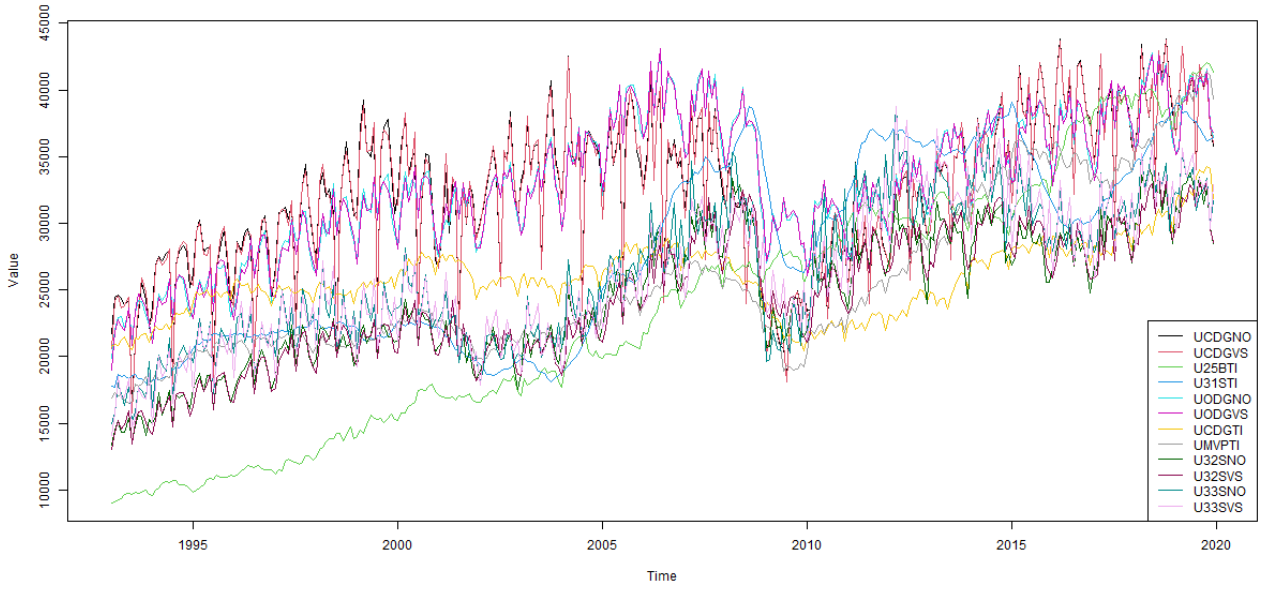

Figure 18: Cluster 12 of US Census Bureau Economic Directorate time series data by Dynamic Time Warping methodology and inspection in millions of dollars. Reference to NAICS code identification document for specific industries.

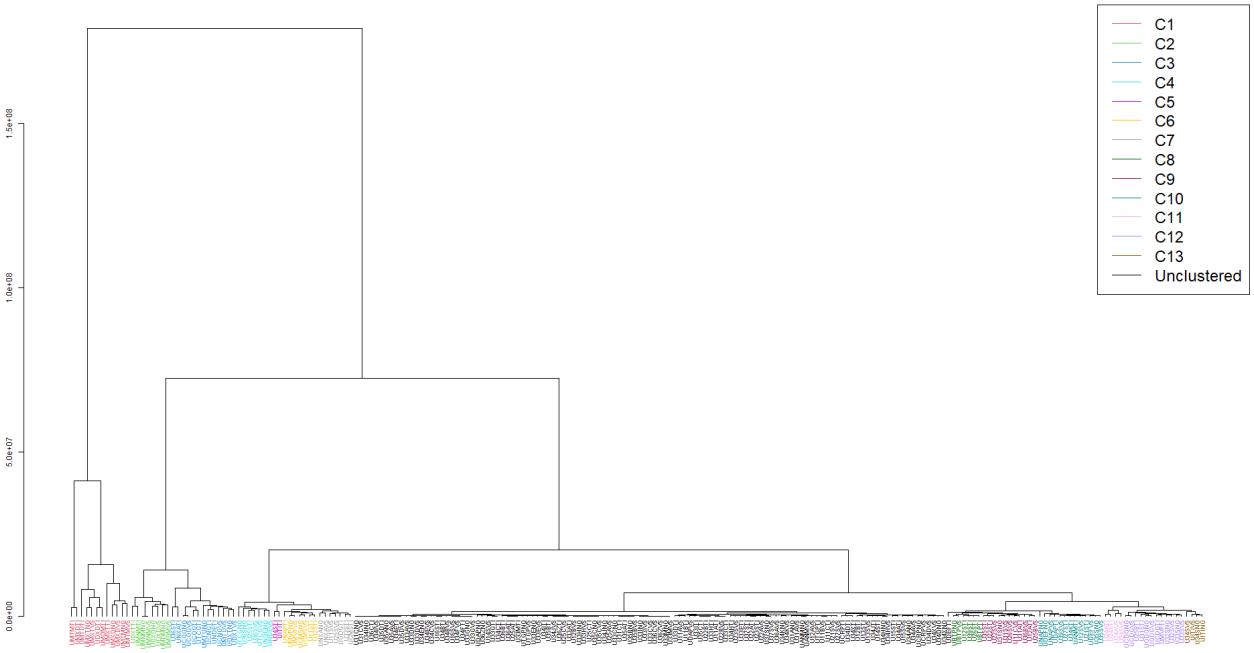

Figure 19: Completed dendrogram by Dynamic Time Warping methodology to create clustered time series data that can be correlated to each other.

References

[1] Saeed Aghabozorg, Ali Seyed Shirkhorshidi, and Teh Ying Wah. *Time-series clustering– A decade review*. May 6, 2015. URL: https://wiki.smu.edu.sg/18191isss608g1/img_auth.php/f/fd/Time_Series_Clustering_A_Decade_Review.pdf.

[2] Ryan Greenway-McGrevy. *A Multivariate Approach to Seasonal Adjustment*. Apr. 2013. URL: https://www.bea.gov/research/papers/2013/multivariate-approach-seasonal-adjustment.

[3] *Manufacturers' Shipments, Inventories, and Orders - Historical Time Series - NAICS*. US Census Bureau. URL: https://www.census.gov/manufacturing/m3/historical_data/index.html (visited on 04/01/2022).

[4] Alexis Sardá-Espinosa. *Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package*. URL: https://cran.r-project.org/web/packages/dtwclust/vignettes/dtwclust.pdf.

[5] Pavel Senin. *Dynamic Time Warping Algorithm Review*. University of Hawaii at Manoa, Dec. 2008. URL: https://www.researchgate.net/profile/Pavel-Senin/publication/228785661_Dynamic_Time_Warping_Algorithm_Review/links/02bfe5100f11a7929f000000/Dynamic-Time-Warping-Algorithm-Review.pdf.

[6] Denyse Tan. *Time Series Clustering — Deriving Trends and Archetypes from Sequential Data*. July 28, 2021. URL: https://towardsdatascience.com/time-series-clustering-deriving-trends-and-archetypes-from-sequential-data-bb87783312b4.

[7] Romain Tavenard. *Machine Learning for Time Series – Notes from Lectures at ENSAI*. 2021.