Creating a Data-Driven Taxonomy

Randall Powers, Wendy Martinez, and Terrance Savitsky

Office of Survey Methods Research, U.S. Bureau of Labor Statistics

Powers.Randall@bls.gov

Abstract: The Monthly Labor Review (MLR) is published by the U.S. Bureau of Labor Statistics. Issues of the MLR often focus on a particular topic, and most articles are written by BLS staff. The need for a classification system of past MLR articles that can be used to label future articles has been recognized by the agency. To address this problem, we employed various unsupervised learning approaches to cluster MLR articles from 2000 to 2013. In this presentation, we will discuss the processes used to prepare the data set, the cluster approaches used, and the results.

Key words: clustering, R Shiny, text data, unsupervised learning, Monthly Labor Review

1. Background

The Bureau of Labor Statistics is one of the principal statistical agencies of the United States. The BLS is the principal fact-finding agency for the Federal Government in the broad field of labor economics and statistics. It is a politically independent agency within the Department of Labor that serves as the statistical arm for all labor related data. It is part of the federal statistical system that includes the Bureau of Economic Analysis (BEA) and the Census Bureau. Its mission is to collect, process, analyze, and disseminate essential economic information to support public and private decision-making.

BLS produces a Consumer Price Index, Producer Price Index, and both Import and Export Price Indices. It produces statistics related to workplace conditions, injuries, illnesses, and fatalities. BLS also produces employment and unemployment numbers at the federal, state, and local levels. Productivity statistics are available for the U.S. business sector, the nonfarm business sector, the manufacturing sector, and 18 groups of manufacturing industries. Further information on all BLS programs and data can be found here: www.bls.gov.

The *Monthly Labor Review* (MLR) is the principal journal of fact, analysis, and research published by the BLS. It has its own subpage on the BLS website. It can be found at https://www.bls.gov/opub/mlr/. Articles by economists, statisticians, and other experts from BLS and stakeholders provide a wealth of knowledge on subjects pertaining to a wide range of economic issues.

Cluster analysis is the grouping of data in a way such that data records assigned to the same group (or cluster) are more similar to each other than data in other groups. Clustering is often used for dimension reduction and to perform inference. It is an iterative process that can be achieved using a variety of algorithms [Martinez, et al., 2011].

2. Motivation-Goal

The BLS makes data, articles, and resources available to the public on their website and tries to make them easily discoverable and accessible. However, it is often hard to locate the required information. Thus, BLS created a working group, which was tasked to create a taxonomy from domain knowledge for the BLS website. This taxonomy was developed using subject matter experts and their knowledge of fundamental concepts. The BLS Office of Publications sought something similar for articles in the MLR. In other words, to create a taxonomy that could be used to categorize and tag MLR articles making them easier to locate.

The first issue of the *Monthly Labor Review* was published in 1915. It would be a long and difficult process to obtain, read, and tag all articles. Then there is the question of what tag, subject, or topic to assign to them. What should be used – the BLS website taxonomy, the various program titles, or something else? The authors decided to cluster the papers using statistical machine learning and natural language processing techniques. This would hopefully group documents in such a way that articles in each group cover a similar topic. This would just be a starting point in developing the taxonomy.

We used the clustering approach of Savitsky [2016], which was implemented in an R package called `growclusters`. The `growclusters` package is designed to estimate a clustering or partition structure for relatively high-dimensional multivariate data. Estimation is performed under a penalized optimization derived from Bayesian non-parametric formulations in the limit that the model noise variance of a hierarchical Dirichlet process (HDP) model goes to 0. It is called "Bayes" clustering, but it is not really Bayesian; it is just inspired by Bayesian models. An important aspect of Savitsky's methodology is that it estimates the number of clusters. It performs what we call single source and hierarchical clustering approaches. The single source method provides results similar to *k*-means clustering [Martinez, et al., 2011]. The hierarchical method allows for known sub-domains (e.g., years) in the data. We use the `growclusters` package as part of our clustering analysis for this project.

The MLR articles used in this study have been published between 2000 and 2013 and should have some common topics over that span of years. We could take all the articles and cluster them as if they were published at the same time, i.e., in the same year. This would be the single source approach. We could also cluster articles published in each year separately, which would be the hierarchical method. The hierarchical clustering approach finds global topics just as in single source, but accounts for possible dependence between journal years.

3. The Data Story and Our Workflow

As previously introduced, the MLR is the principal journal of fact, analysis, and research from BLS. MLR articles are available HTML and PDF formats and include a variety of standard sections, such as abstracts, introductions, references, footnotes, etc. The steps of our analysis are described next.

In this analysis, each data point is an unstructured text document corresponding to an MLR article. As usual in any analysis, we first had to clean the data. We removed stop words (uninformative words), short words less than 4 characters, long words more than 9 characters, infrequent words, and domain stop words. Still, the data were noisy and messy because of acronyms, typos, repeated words, etc.

Next, we used the term-document matrix to encode the text or to convert it to numbers. [Solka, 2008] This is also known as the bag-of-words approach. The rows of the data set correspond to words in the lexicon, the columns to articles. The lexicon is the set of unique words across all documents or articles in the data set. We essentially encode each article as a vector of words. We could count the number of times a word appears in the article, which is the raw frequency approach. In this approach, each element of the vector represents the number of times that word appears in the document. Or we could use a 1 if a word appears in the article or a 0 if it does not appear in the document. This is called the binary encoding.

We then looked at word clouds across all articles to visualize and explore the content. The cloud of single words (see Figure 1) provides a summary of the highest frequency words and gives a sense of some topics covered. The highest frequency words are *industry*, *services*, *jobs*, *growth*, and *time*. This makes sense, as these are words that would likely be used across many articles and they pertain to topics common to BLS. The cloud of word pairs (see Figure 2) shows highest frequency word pairs or bigrams. Here we see "minimum wage," "wage salary," health care." and "retail trade." Again, these describe areas often discussed by BLS economists and stakeholders.

For our data set we had 574 articles containing 12,437 unique words, giving us a dimensionality of over 12,000. This is too high, making the data too noisy and the data matrix very sparse (lots of zeros). To address this issue, we used isometric feature mapping or ISOMAP to reduce the dimensionality [Martinez, et al., 2011]. We had to determine how many dimensions to use in our reduced space. We decided to use three dimensions for the raw encoding and four dimensions for the binary encoding. This is based on scree-like plots that are part of the ISOMAP output [Martinez, et al., 2011]..

To summarize our workflow, we encoded the documents in the two ways already described: raw and binary. Next, we visualize and explore the data through word clouds to get a sense of the content. We then reduce the dimensionality using ISOMAP. The final steps are to apply the clustering approaches available in the `growclusters` package and assess the results.

4. Cluster Analysis and Results

In Figures 3 through 6, we show the clustering results for the raw frequency encoded data set. In the scatterplot (see Figure 4), we can readily see some interesting clusters for the raw data. Each dot represents a document or article shown in the lower-dimensional ISOMAP space. Each color represents a cluster. Notice how well the dots in the different colors cluster together. This indicates that there is some structure (or groups) in the data.

For the parallel coordinates plot (see Figure 5), we can also see clear indication of groups or clusters. Each line represents a document, and each color is a cluster. We can clearly see patterns across each observation. These are shown as bundles of lines for each color (or cluster) that sort of stick together. Sometimes we prefer the parallel plot to the scatterplot because where the scatterplot shows us pairwise comparisons, the parallel allows us to see our results in all dimensions. In Figure 6, we highlighted a single cluster while simultaneously graying out the other clusters to give us a better view of the amount of clustering for a specific cluster.

We also looked at the binary encoded data set (see Figures 7 through 9). Here again we observe that the cluster grouping is reasonable. We have four variables because we reduced the dimensionality of the binary data to four. Again, just as we saw with the raw data, we can examine the clustering of individual clusters using the parallel plot feature of the application.

The cluster approach we used provides an estimate of the number of groups. Eight clusters were found in both encodings – raw and binary (see Figures 3 and 7). Visually, the groups seem reasonable. Recall that we are grouping or clustering documents in hopes of developing a taxonomy or classification system that can be used on all MLR articles. Now that we have the clusters, we need to assess the results. In other words, do the documents in a cluster cover a common topic or subject area? One way to help determine this is to look at the word frequency distribution of articles in the clusters, i.e., what are the high frequency words in each cluster?

In Figure 10, we see a word cloud for each of the eight clusters from the raw frequency encoded articles. We visualize word pairs, which provides more context, whereas the clustering itself was done on a single word. An initial look shows us that each cluster seems to focus on a different topic. “wage salary” “weekly benefit” “price index”, etc. In Figure 11, we see the word clouds for the binary encoded data. We notice that regardless of what type of encoding we use, we are getting similar results. This tells us we have some definite structure or clusters, not just random groupings.

We can see some interesting characteristics of the clusters by examining the word clouds in Figures 10 and 11. There are some common topics regardless of the type of encoding, such as prices, minimum wages, and weekly benefits. There are also some different topics discovered when we changed the encoding. These include Gulf War era veterans, social security, health insurance, and employment numbers. We can also see possible additional topics within each cluster.

5. Hierarchal Version of Bayesian Clustering- Account for Annual Issues

What we have described so far used single source clustering. In other words, we lumped all the articles into one pile, not considering the fact that these articles were published over the course of several years. So, we account for that aspect of our data in the next step of our analysis. We still find global clusters, but we allow for possible time dependence between the articles. The results of this method are shown in Figures 12 and 13. We found seven clusters with the binary encoding and eight clusters with the raw encoding.

We assigned topics or themes to each cluster (see Figure 13). We see similar topics found regardless of the encoding or the approach (single source or hierarchical). This is an indication that the clusters found are potentially real, i.e., they are not just grouped randomly. We also discover a few different topics. There were two large clusters found based on both encodings that are highlighted in Figure 12. We see that these clusters are the large ones (or the same, health insurance and employment) regardless of the encoding.

In Figures 14 and 15, we see bar plots of the global topics as distributed over the years. Note that the distribution of topics in each year are somewhat similar. This is an indicate of topic or cluster dependence over the years, and it is important that we account for this dependency in the clustering methodology. This result is not surprising since these are articles from the same journal and are on a common discipline area.

6. Future Work and Final Comments

We still have a lot of work to do. Our future work includes applying some model-based clustering to see what other themes or subjects we might discover. We will also subcluster the larger groups and look at the subclusters separately. We should verify the results with our economists, who are the subject matter experts. Finally, we plan to finalize the `growclusters` package and eventually publish the package on CRAN.

7. References

Martinez, Wendy L., Angel R. Martinez and Jeffrey L. Solka. *Exploratory Data Analysis with MATLAB*. CRC Press, 2011. Print.

Martinez, Wendy L., and Terrance Savitsky (2019). Towards a Taxonomy of Statistical Data Editing Methods. Working Paper presented at the UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T_USA_Martinez_Paper.pdf

Savitsky, Terrance D. (2019). About `growclusters`. Documentation for `growclusters` package (to be published on CRAN)

Savitsky, Terrance D. (2016). Scalable Approximate Bayesian Inference for Outlier Detection under Informative Sampling, Journal of Machine Learning Research 17(225):1−49.

Solka, Jeffrey L. (2008). Text Data Mining: Theory and Methods, Statistics Surveys. 2:94–112, https://doi.org/10.1214/07-SS016

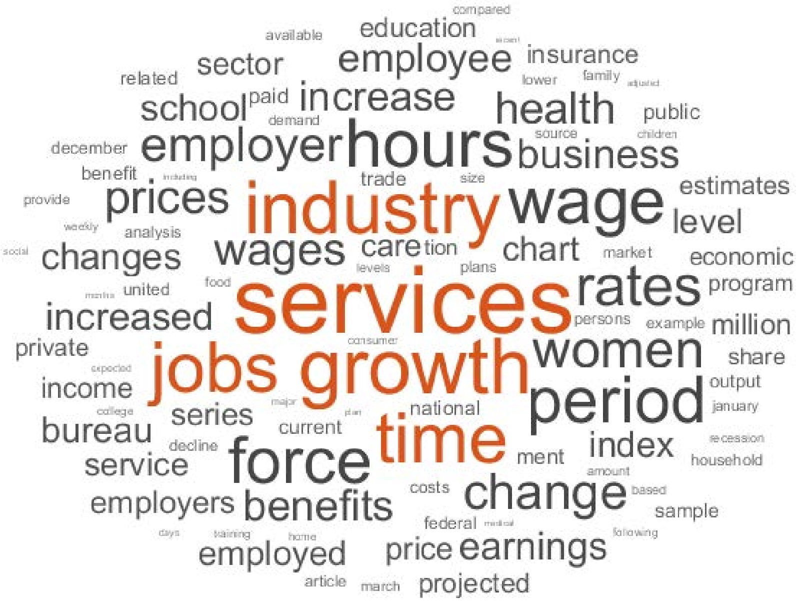


Figure 1. This is a screenshot of the single word cloud which provides a summary of the highest frequency words and gives a sense of some topics covered.

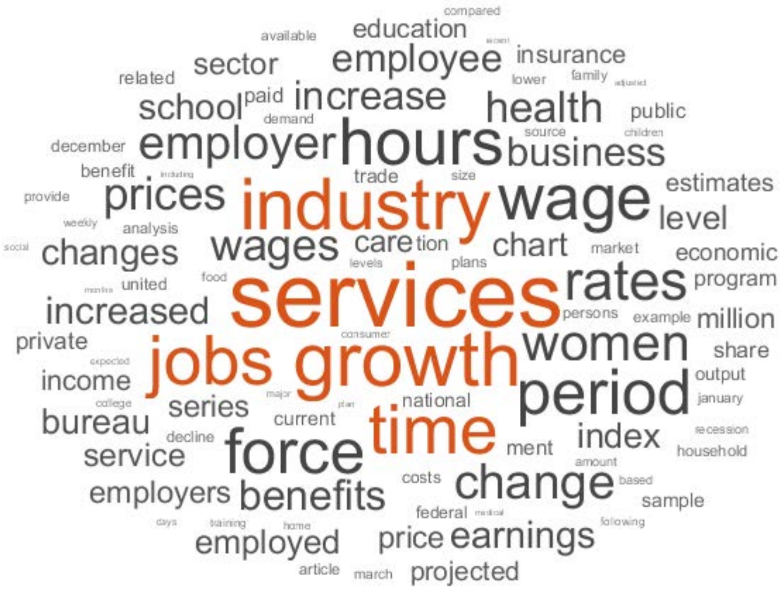
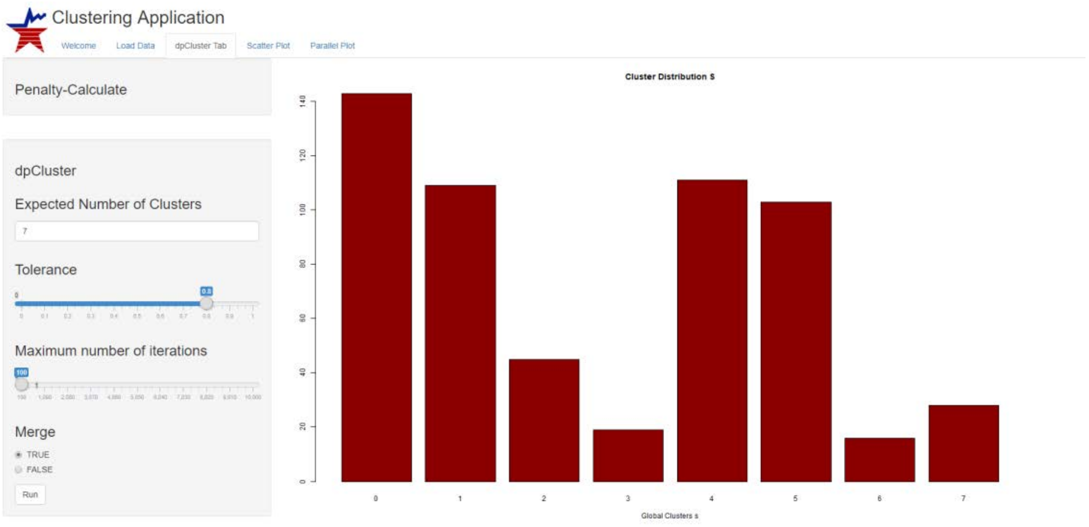


Figure 2. This is a screenshot of the word pairs word cloud which provides a summary of the highest frequency paired words and gives a sense of some topics covered.

Raw Encoding, Single Source Bar Chart of Clusters



Penalty-Calculate

dpCluster

Expected Number of Clusters

7

Tolerance

0.8

Maximum number of iterations

100

Merge

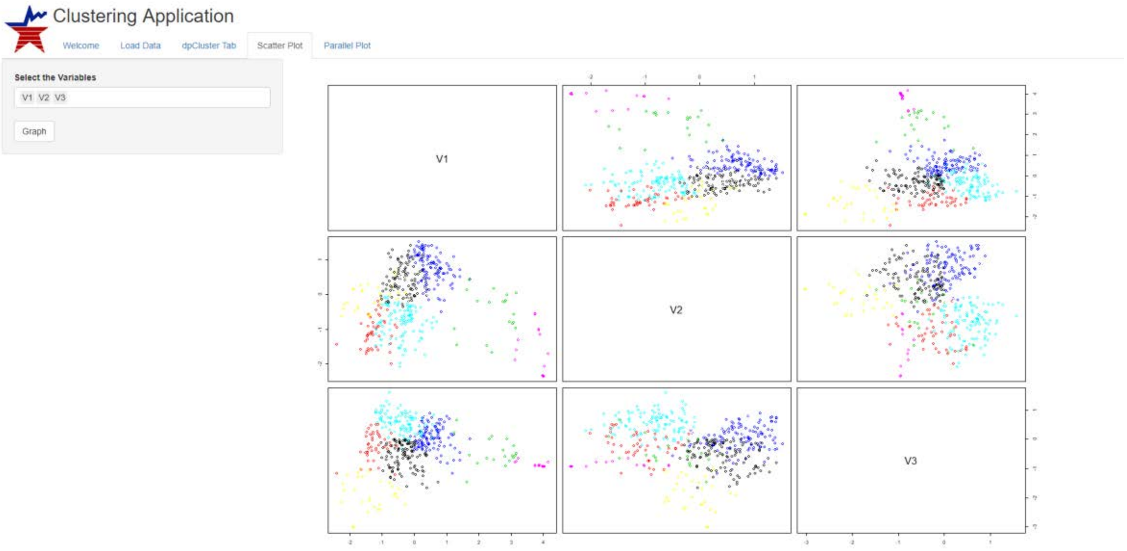
● TRUE

○ FALSE

Run

Figure 3. This is a screenshot of the bar chart of cluster counts for the raw encoding single source data

Raw Encoding, Single Source Scatterplot



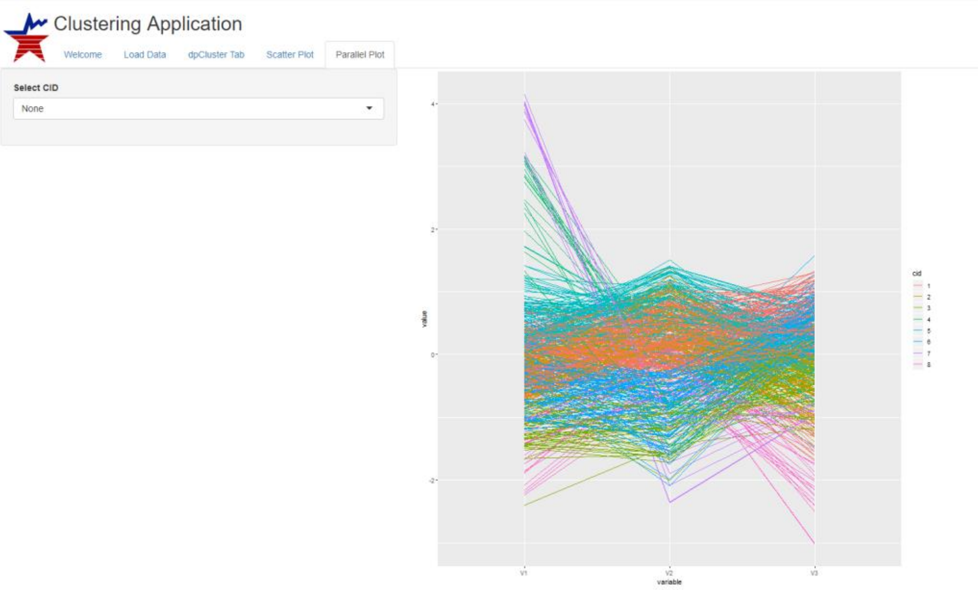
Select the Variables

V1 V2 V3

Graph

Figure 4. This is a screenshot of the scatter plot for raw encoding, single source data. Each dot represents a document, and each color represents a cluster.

Raw Encoding, Single Source Parallel Plot

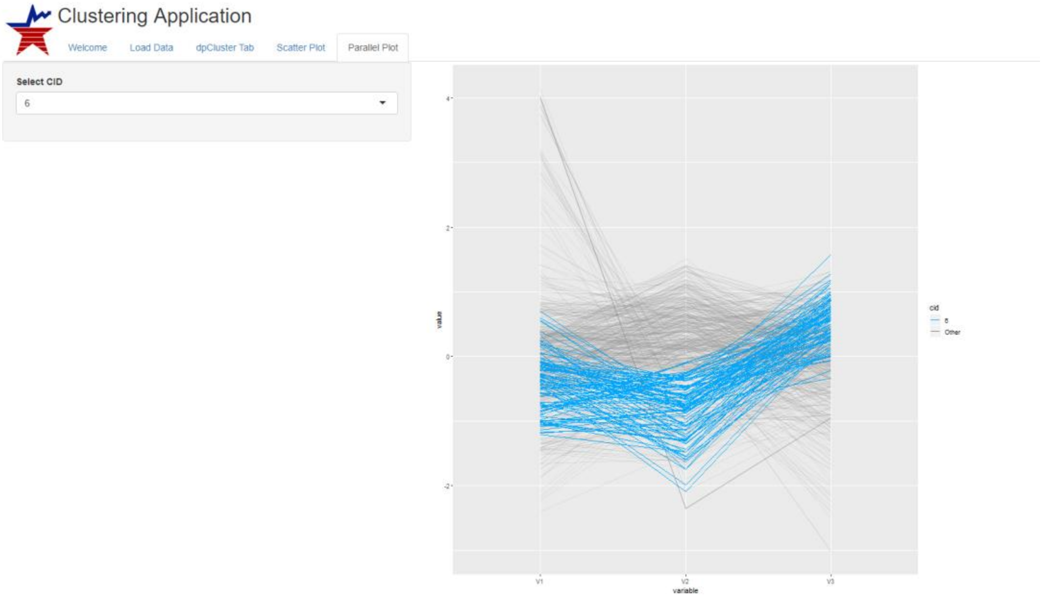


Select CID

None

Figure 5. This is a screenshot of the parallel plot for raw encoding, single source data. Each line represents a document, and each color is a cluster.

Raw Encoding, Single Source, Parallel Plot, CID=6

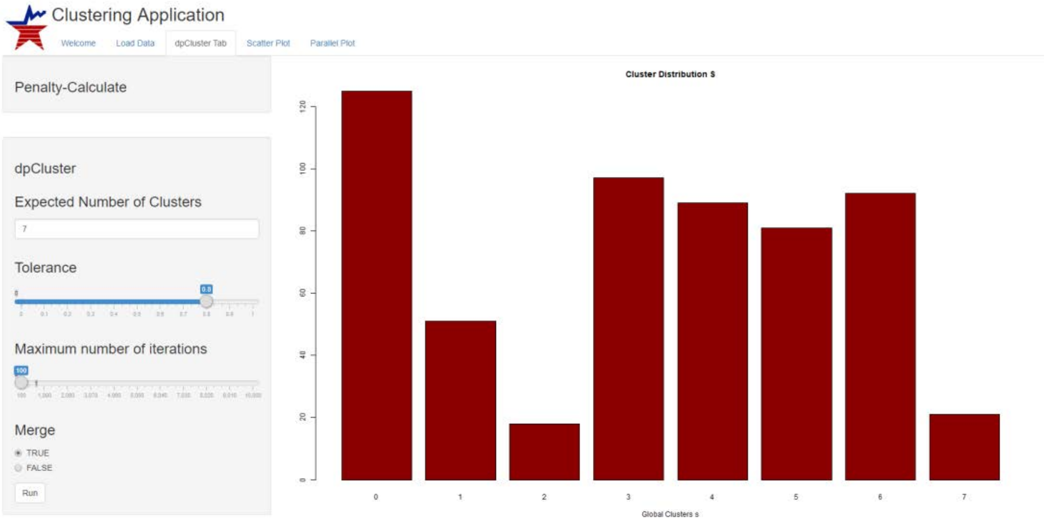


Select CID

6

Figure 6. This is a screenshot of a single cluster highlighted to illustrate the degree of clustering present, using raw encoded, single source data.

Binary Encoding, Single Source, Bar Chart of Clusters



dpCluster

Expected Number of Clusters: 7

Tolerance: 0.8

Maximum number of iterations: 100

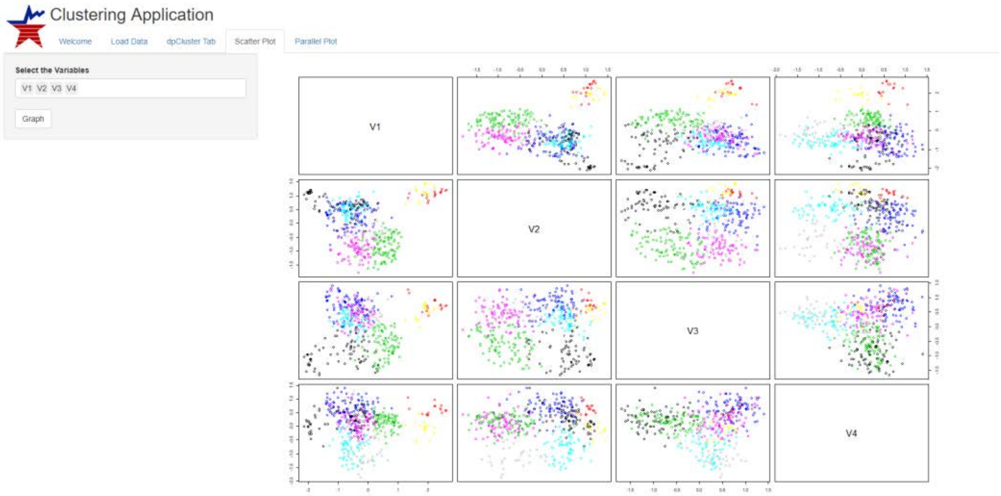
Merge

● TRUE

○ FALSE

Figure 7. This is a screenshot of the bar chart of cluster counts for the binary encoding single source data.

Binary Encoding, Single Source, Scatterplot



Select the Variables: V1 V2 V3 V4

Figure 8. This is a screenshot of the scatter plot for binary encoding, single source data. Each dot represents a document, and each color represents a cluster.

Binary Encoding, Single Source Parallel Plot CID=1

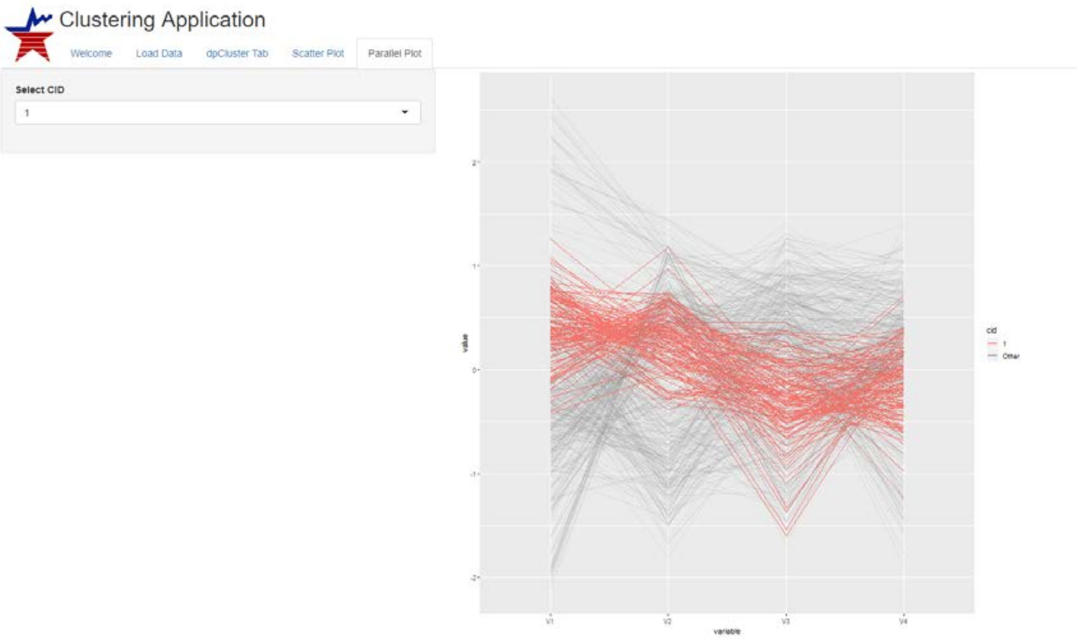


Figure 9. This is a screenshot of a single cluster highlighted to illustrate the degree of clustering present using binary, single source data.

Clusters – Raw Frequencies

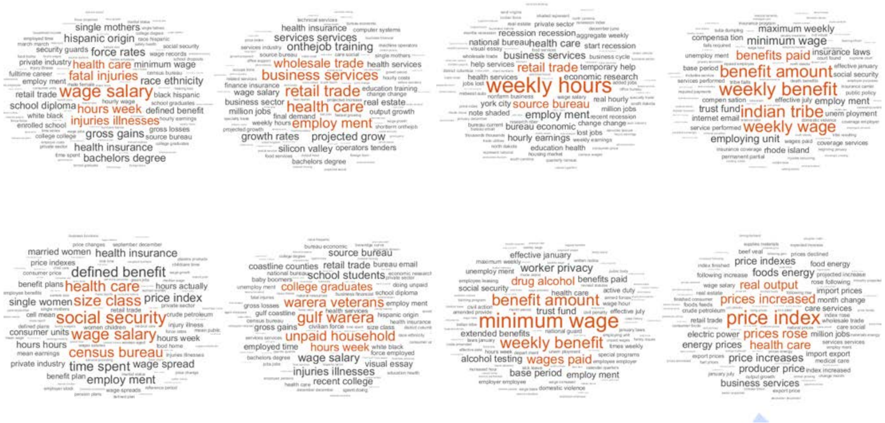
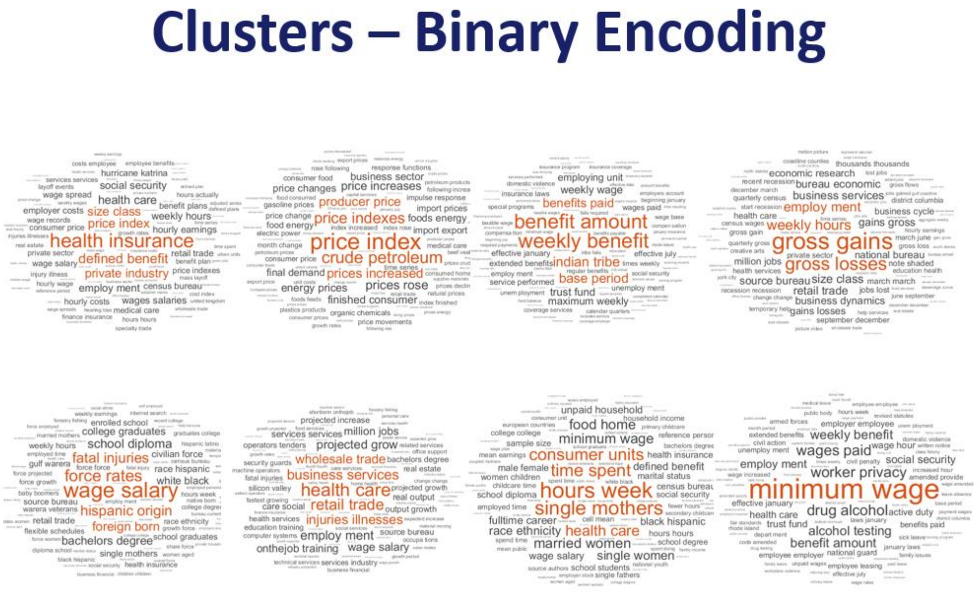
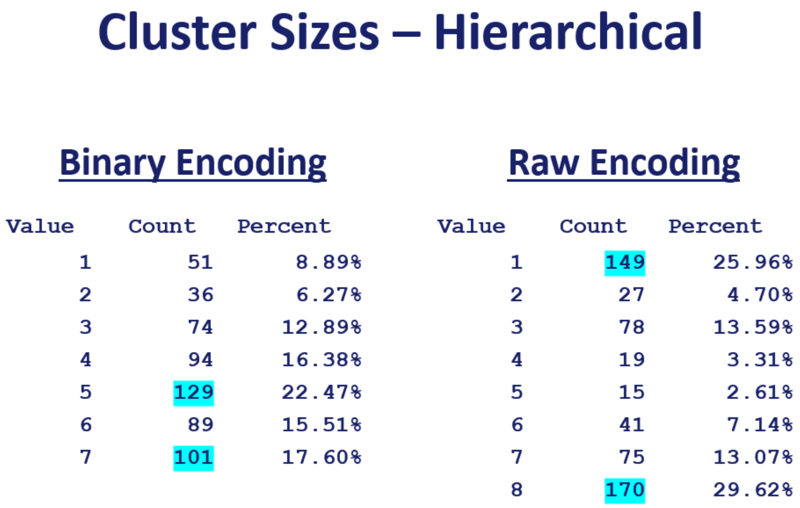


Figure 10. This is a screenshot of the word pairing word clouds for each of the eight raw frequencies.



Clusters – Binary Encoding

Figure 11. This is a screenshot of the word pairing word clouds for each of the eight binary encoding frequencies.



Cluster Sizes – Hierarchical

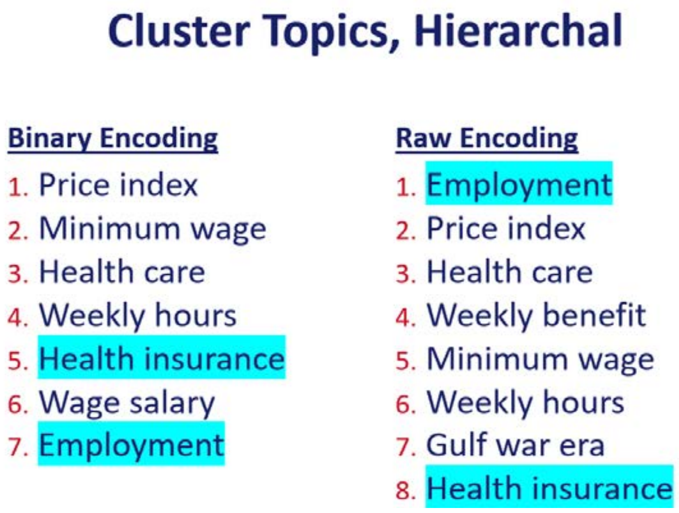
Binary Encoding

Value	Count	Percent
1	51	8.89%
2	36	6.27%
3	74	12.89%
4	94	16.38%
5	129	22.47%
6	89	15.51%
7	101	17.60%

Raw Encoding

Value	Count	Percent
1	149	25.96%
2	27	4.70%
3	78	13.59%
4	19	3.31%
5	15	2.61%
6	41	7.14%
7	75	13.07%
8	170	29.62%

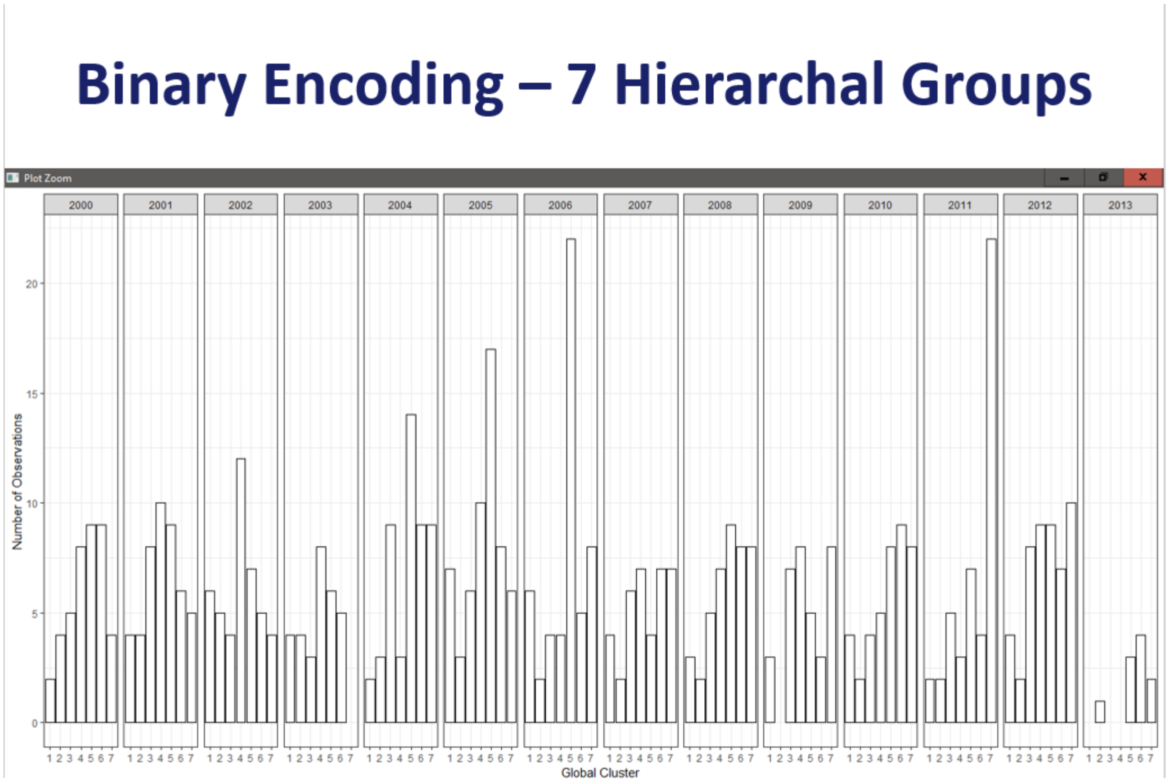
Figure 12. This is a table of the cluster size counts for each of the types of encoding for hierarchal data. Each had two large clusters, highlighted in blue. These highlighted groups have the same topics regardless of the encoding as shown in the next figure.



Cluster Topics, Hierarchal

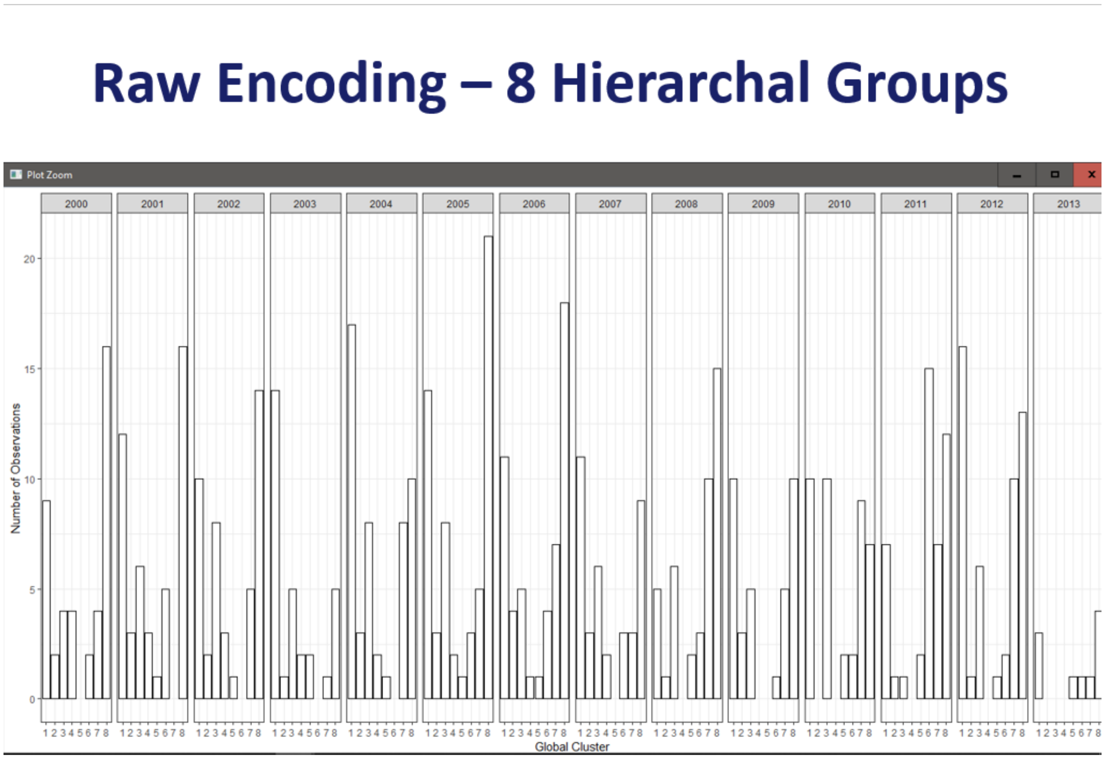
Binary Encoding	Raw Encoding
1. Price index	1. Employment
2. Minimum wage	2. Price index
3. Health care	3. Health care
4. Weekly hours	4. Weekly benefit
5. Health insurance	5. Minimum wage
6. Wage salary	6. Weekly hours
7. Employment	7. Gulf war era
	8. Health insurance

Figure 13. This is a table of the cluster topics for both binary and raw encoding, hierarchal data.



Binary Encoding – 7 Hierarchal Groups

Figure 14: Bar plot of the global topics as distributed over the years for binary hierarchal data. There are articles in the global topics in each of the years (except for 2013).



Raw Encoding – 8 Hierarchal Groups

Figure 15: Bar plot of the global topics as distributed over the years for raw hierarchal data. We see some differences compared to the previous figure.