Creating a Taxonomy of Statistical Methods using Text Analysis

Wendy L. Martinez and Terrance Savitsky
U.S. Bureau of Labour Statistics, 2 Massachusetts Ave, NE, Washington, DC 20212

Abstract
The United Nations Economic Commission for Europe (UNECE) holds an annual workshop on Statistical Data Editing with a focus on official surveys. The 2017 workshop organizers formed subgroups who were tasked to come up with ideas to foster the implementation of good practices and international collaboration among the statistical offices of member countries. One proposal from the subgroups was to conduct a classification of existing methods for data editing and imputation based on papers presented in previous UNECE work sessions on data editing. Another idea was to create an indexed and searchable inventory of these papers using a taxonomy. This paper describes research addressing the first idea – to construct a taxonomy of topics addressed by the UNECE data editing group. To do this, we downloaded all papers from the annual work sessions, converted them to machine readable format, and applied text analysis approaches to create a taxonomy based on the papers. This paper will describe the process and tools used to create the taxonomy, so others can apply these same ideas to their document collections.

1. Introduction

The United Nations Economic Commission for Europe (UNECE) has held twelve (12) workshops on Statistical Data Editing since 2000. At the 2017 workshop, the organizers formed subgroups who were tasked to come up with ideas to foster the implementation of good practices and international collaboration among the statistical offices of member countries. One proposal from the subgroups was to conduct a classification of existing methods for data editing and imputation based on papers presented in previous UNECE work sessions on data editing. Another idea was to create an indexed and searchable inventory of these papers using a taxonomy [1].

This paper describes research addressing the first goal – to construct a taxonomy or classification of topics or methods addressed by the UNECE data editing group. To this end, we downloaded all papers from the twelve work sessions and applied text analysis approaches to cluster the papers, so papers grouped together have similar topics. This grouping can serve as a starting point for a taxonomy of data editing and imputation topics presented at the UNECE workshops.

We note that this work was also presented at the 2018 UNECE Working Group Session in Neuchatel, Switzerland [2]. We used the MATLAB Text Analysis Toolbox for most of the analyses presented in this paper, while R was used for the Bayes clustering approach.

2. Preparing the Data

2.1 Source of Documents

The body of documents (or corpus) includes papers downloaded from the UNECE wiki site that hosts information on the “Seminars and Workshops organized by UNECE in the area of

Statistical Data Editing." [3] There have been twelve (12) workshops focusing on this area starting in the year 2000. All of the content on the site is open and available to the public.

Papers of approximately 10 pages in length must be submitted by authors before the workshop, and a template [2, 3] is to be used for all papers. So, we had significant content to work with, unlike most workshops, where only abstracts are available. Most (if not all) attendees work for their respective country's government statistical agencies, so the focus is on official statistics. Furthermore, the workshops are organized around major themes (e.g., tools, systems, methods), and all centre on just one area – data editing and imputation. Given this, these papers represent a restricted domain of discourse (i.e., statistical data editing), and we expect to obtain reasonable results when clustering the papers.

The documents are posted to the wiki site in .pdf format. We downloaded all documents by hand rather than use a web-scraping tool. There were a total of 427 usable papers or documents. One paper did not use the template, and another had problems with the .pdf encoding, and we could not convert it to text.

2.2 Pre-processing the Documents

One always has to pre-process unstructured text documents before encoding them. We followed these steps to clean the data [4].

- We removed the header information based on the template. This included everything from "I. Introduction" and above. So the title and author information were not part of the final documents to be clustered.
- We deleted the reference section.
- We removed special characters and end-of-sentence punctuation (! @ $, . ! ? …).
- We deleted stop words, which are considered to be non-informative. These are words such as *and*, *the*, *but*, *for*, … We used a generic list of stop-words.
- We removed short words (3 characters or less) and long words (15 characters and more).
- We deleted infrequent words – those that occurred two times or less.
- Finally, we converted all text to lower case.

We tried to remove acronyms (e.g., BLS, NCHS, NASS, UNECE…) using pattern matching, but we were not successful. So, some acronyms might be in the lexicon. (The lexicon is the list of unique words in the corpus.) See Table 1 for summary information on the set of papers included in the corpus.

Table 1: Summary information on the set of 427 papers (documents) downloaded from the UNECE Statistical Data Editing website and included in the corpus.

Year of Workshop	*Number of Documents In Workshop*	*Average Number of Words per Document After Pre-Processing*
2000	**22**	**690**
2002	**36**	**799**
2003	**34**	**781**
2005	**51**	**851**
2006	**26**	**732**
2008	**38**	**803**
2009	**43**	**725**
2011	**44**	**794**
2012	**44**	**797**
2014	**30**	**673**
2015	**33**	**721**
2017	**26**	**700**

2.3 Encoding the Documents

Now that we have the set of cleaned and pre-processed papers or documents, the next step is to convert the text to a numerical format so we can compute with it. A widely used method is to encode the documents in a term-document matrix (TDM). This is also known as the bag-of-words approach. The *ij*-th entry in the TDM corresponds to the number of times the *i*-th word appears in the *j*-th document. Each column of the TDM represents a word frequency distribution for a given document or paper.

The TDM has p rows, which is the number of unique words in the lexicon. It has n columns corresponding to the number of documents in the corpus. To statisticians, each document is an observation, and each word is a dimension. As we will see in the next section, the number of dimensions is quite a bit larger than the number of observations, so $n << p$.

What we just described is the TDM with raw frequencies; i.e., the number of times a word appears in a document. Other types of term (or word) weighting schemes have been developed in the information retrieval literature [5]. A useful one when we have a restricted domain of discourse, which is what we have in this application, is the binary encoding. The binary representation records the presence or absence of a word. So, an entry of one (1) means the word is in the document, and a zero (0) indicates the word is not in the document.

Another approach sometimes used in information retrieval and text analysis is stemming [4, 6, 7]. With stemming, we reduce words to their stem or root word. For example, the words *editing*, *edited*, *edits* would be replaced by the word *edit*. We created a corpus based on stemming, but have not analysed these stemmed documents yet.

Here is a summary of where we are at this point in the process. We have $n = 427$ papers or documents to cluster. There are $p = 10{,}737$ unique words in the corpus after we remove the domain-specific stop words, as described in the next section. We have two types of encodings – raw frequencies and binary. We are not using the stemmed corpus in what follows.

2.4. Describing the Corpus

It is always a good idea to look at word frequency distributions for the corpus to get a sense of the content. Figure 1 shows the word frequency distribution for the set of papers in our corpus, and we see that the highest-frequency words are *data*, *editing*, and *imputation*. These are not informative in our application, and we should remove these and similar domain-specific stop words.

Word	Count
"data"	18074
"editing"	7542
"imputation"	5853
"variables"	5205
"survey"	4599
"values"	4539
"edit"	4061
"process"	3947
"statistics"	3212
"statistical"	3180

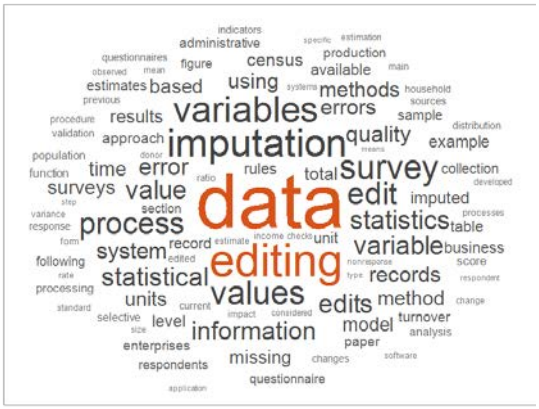

Figure 1: This shows a word frequency distribution for the set of 427 documents. The highest-frequency words *data*, *editing*, and *imputation* are not informative in this application. The domain-specific stop words were removed.

3. Exploratory Process

While some statistical machine learning methods can be used with high-dimensional data, we found that the data are just too noisy in these high-dimensional spaces. Thus, we first reduce the dimensionality before further analysis. There are several approaches for dimensionality reduction [8], such as singular value decomposition, multidimensional scaling, and non-negative matrix factorization. We have had success using ISOMAP or Isometric Feature Mapping [9], which tries to find an embedding for the observations on a lower-dimensional sub-manifold.

ISOMAP takes the interpoint distance matrix as inputs. It uses these distances to estimate the geodesic distances between points along a sub-manifold. These geodesic distances are used in turn as the inputs to classical multidimensional scaling to find an embedding in a lower-dimensional space.

We applied ISOMAP to our TDM (matrix with 10,737 rows and 427 columns). A reasonable value for the number of dimensions to use for the embedding appeared to be five (5). So, we finally end up with a data matrix containing 427 rows (or observations) and $d = 5$ columns (dimensions). One of our ISOMAP embeddings based on the binary-encoded TDM is shown below in Figure 2. The TDMs reduced to these 5-D ISOMAP spaces (one for binary encoded terms and one for raw frequencies) are clustered.

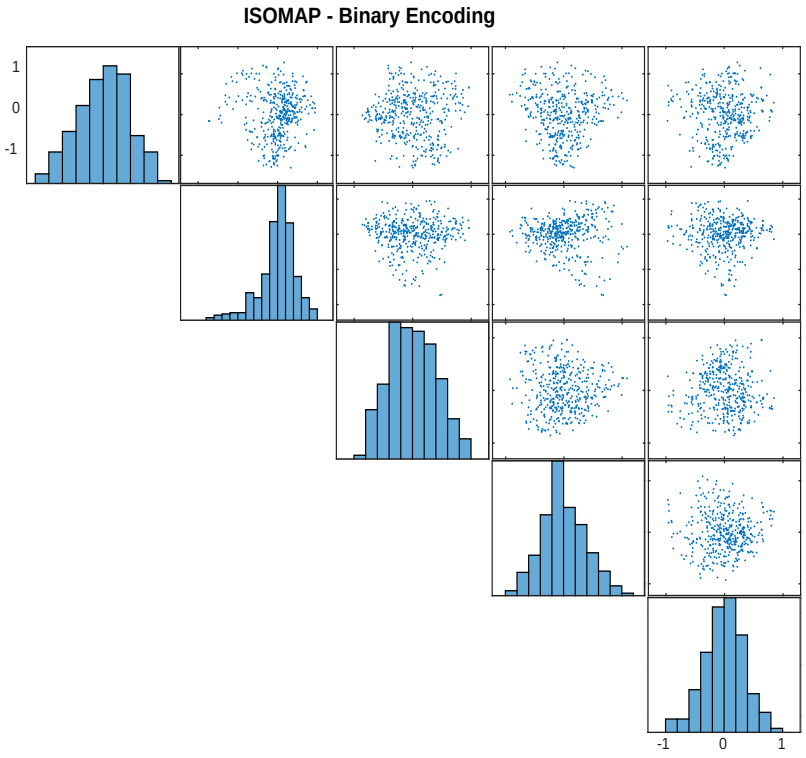

Figure 2: This scatterplot matrix shows the ISOMAP embedding in 5-D applied to the binary-encoded TDM. Each point is a document. The observations (documents) with coordinates in this 5-D space are clustered.

The next step in our analysis is to cluster the documents, such that documents with similar topics are grouped together. Clustering is an exploratory data analysis technique. Thus, it is a good idea to try different approaches or methods to explore our data and see what type of cluster structure we can find. This is especially true with clustering because the types of clusters one gets is dependent on the method used.

A key problem with clustering is determining the number of groups we might have in the data. After all, we might have only one group and no clusters! To make the analysis a little easier, we used two clustering approaches that also provide estimates of the number of groups or clusters based on the data. So, we do not have to specify the number of groups, as we would in *k*-means or agglomerative clustering.

These two methods are model-based clustering (MBC) [10] and Bayes clustering [11]. MBC obtains the clusters by modelling the data as a finite mixture of weighted multivariate normal probability density functions. Each component in the model (or multivariate normal density function) corresponds to a cluster. The estimated covariance for a component determines the shape of the cluster, and it allows for a very flexible structure. This is different from *k*-means, which tends to produce spherical clusters.

Bayes clustering provides a hierarchical version of *k*-means. It takes the limit of a hierarchical Dirichlet process model as the noise variance contracts to zero and employs penalized optimization to estimate the model parameters. The Carlinski-Harabaz statistic is used to select the penalty parameter, which in turn estimates the number of clusters. The advantage of

this approach and why it is particularly relevant to this application is it allows for global clustering with possible local dependencies among the individual workshops.

We could employ the following scenarios when clustering our set of documents or workshop papers. First, we could cluster the entire set of 427 documents. In this case, we consider the papers as belonging to one workshop and not taking into account that they came from different workshops over a span of years. So, this would consider each workshop to be independent of the others. We might also cluster the papers from each workshop separately. This option is beyond the scope of this paper, and we keep this for future work. Finally, we can take the middle view using Bayes clustering, where we can account for possible local clustering dependencies, which are linked to global clusters.

4. Results

We cannot show all results in this paper due to space considerations. So, we present results just for the following two cases:

1. Model-based clustering applied to raw frequency encoded documents
2. Bayes clustering applied to binary encoded documents

4.1. Model-Based Clustering

Using MBC with the raw frequency encoding in a 5-D ISOMAP embedding produced six (6) groups. The sizes of the clusters are 104, 82, 42, 112, 52, and 35. We show word clouds based on each cluster in Figure 3. These give us a sense of the content in the groups. These results indicate there is some coherent structure, and we could assign reasonable topics to the clusters. For instance, cluster 4 seems to be on accuracy and variance, while cluster 5 covers aspects of questionnaires.

4.2. Bayes Clustering

Using Bayes clustering with the binary encoding in a 5-D ISOMAP embedding produced nine (9) groups of sizes 28, 31, 29, 41, 87, 37, 47, 32, and 95. We show word clouds based on each cluster in Figure 4. We see similar coherent cluster topics as we had with MBC applied to raw frequency encoded documents. It is interesting to note that there are now two questionnaire clusters – one on electronic questionnaires (number 2) and one on paper questionnaires (number 3).

Recall that Bayes clustering allows for possible local clustering structure, which is connected to the global clusters. This seems a reasonable view for our application because we have twelve separate workshops (local structure) all addressing one theme – data editing and imputation (global structure). So, we might expect to see some dependence across the different workshops.

Bayes clustering is implemented in an R package, which has not been posted yet to CRAN [12]. The package produces a series of plots to visualize cluster results, one of which is a set of histograms. In our application, there is one histogram for each year or workshop. A single histogram shows the frequency distribution of workshop papers based on the global clusters. This plot is shown in Figure 5.

Figure 3: These are bigram (word pair) clouds for each cluster found by applying MBC to the raw frequency encoded documents. Starting from left to right and top to bottom, the cluster sizes are: 104, 82, 42, 112, 52, and 35.

Figure 4: These are bigram (word pair) clouds for each cluster found by applying Bayes clustering to the binary encoded documents. Starting from left to right and top to bottom, the cluster sizes are: 28, 31, 29, 41, 87, 37, 47, 32, and 95.

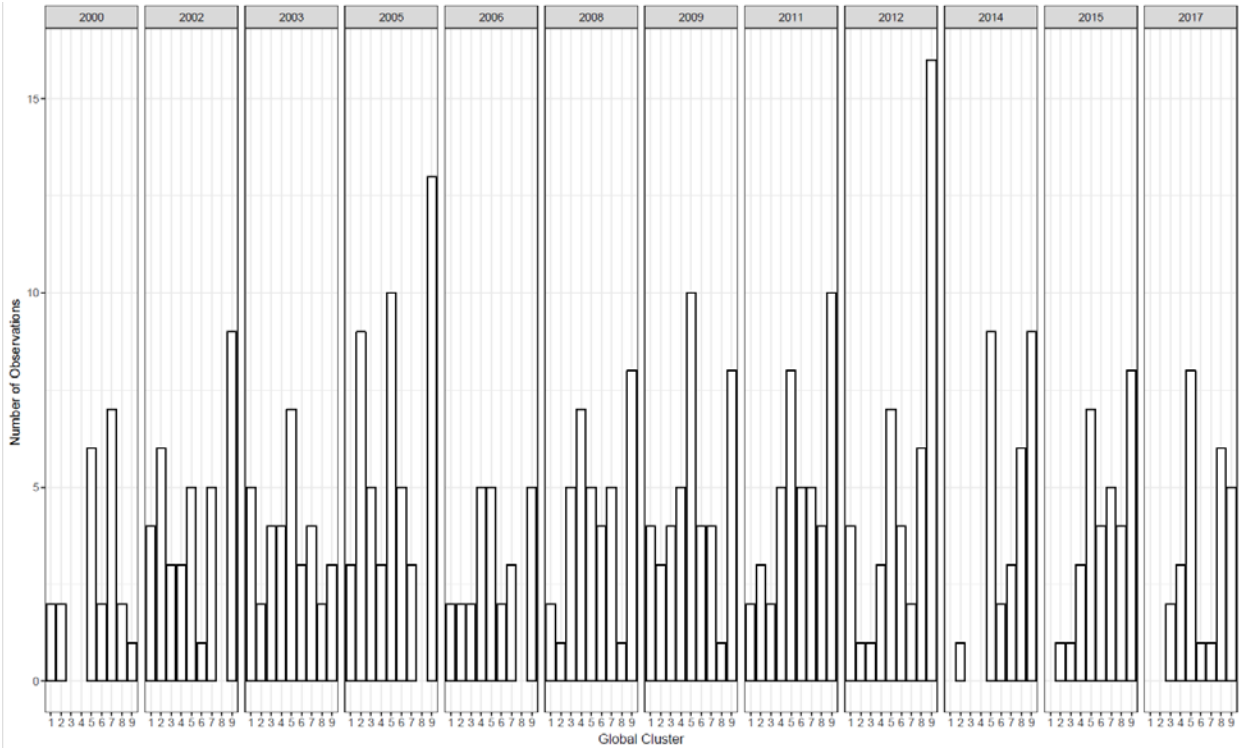

Figure 5: There is one histogram for each workshop. The horizontal axis of each histogram corresponds to the global cluster number. The vertical axis is the number of papers in that group for the given workshop. Some things to note are the similarity of the distributions or histograms over the workshops, and we see that each workshop contains papers in almost all clusters. This indicates some dependence in the topics addressed by the papers.

We can see a couple of interesting things in these histograms. First, the distribution of clusters is similar across the workshops. Second, for the most part, all nine clusters (or topics) are addressed in each of the workshops. This is an indication of local dependence, which is something we might expect for this application.

5. Discussion

Our goal in this analysis was to work toward a taxonomy of data editing and imputation methods addressed in UNECE working sessions. We used clustering to group documents from past workshops employing two document encoding schemes and two types of cluster approaches, both of which will estimate the number of groups in the data. When clustering data, one should always try different approaches and examine the results to make sure we are not discovery spurious structure and information. By looking at the word clouds for each cluster (Figures 3 and 4), we can see that there does appear to be some useful, informative and coherent clusters. There seems to be very little overlap in terms of the methods and concepts describing the clusters.

Several aspects of our preliminary results are worthy to note. First, we obtained a sensible number of clusters regardless of the method used, which makes sense given the source of the documents. Furthermore, the sizes of the clusters were reasonable. In our previous work clustering interviewer notes [13], we tended to have one or two large clusters with many smaller clusters, and some groups had the same themes or content. This was not the case here.

It is also interesting to compare the results from the two clustering approaches. The estimated number of groups was similar – 6 and 9, and we had comparable cluster themes or topics. The Bayes clustering method appeared to further divide clusters into sub-topics. For example, the questionnaire cluster found using MBC was subdivided into electronic and paper questionnaires with Bayes clustering.

We are presenting these results at the 2018 UNECE Data Editing Workshop to obtain feedback on the cluster structure we found and some guidance on our next steps. Some options that come to mind follow.

- We clustered the papers from each workshop separately, and we can explore the clusters as we did before. It would be interesting to see if topics changed over the years.
- The organizers of the workshops group papers into pre-defined categories. We could match the groups obtained through clustering with these categories. It might provide some insights.
- We could employ other types of exploratory data analysis schemes, such as different encodings and term weights, stemming, dimensionality reduction, and clustering approaches to see what structure we can uncover.

Finally, we could use the results presented in this paper as a starting point and start creating the taxonomy and summary of methods presented in the UNECE Data Editing Workshops.

References

1. https://www.unece.org/index.php?id=43887
2. Martinez, W. and T. Savitsky. 2018. "Towards a taxonomy of statistical data editing methods," UNECE Workshop on Data Editing, Neuchâtel, Switzerland, 18-20 September 2018.
3. https://statswiki.unece.org/display/sde/Workshops
4. Solka, J. 2008. "Text data mining: Theory and Methods," *Statistics Surveys*, volume 2, https://projecteuclid.org/euclid.ssu/1216238228
5. Berry, M. W. and M. Browne. 1999. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, SIAM
6. Martin F Porter. 1980. "An algorithm for suffix stripping," *Program*, 14(3):130–137.
7. Martin F Porter. 2001. "Snowball: A language for stemming algorithms," available at: http://www.snowball.tartarus.org/texts/introduction.html
8. Martinez, W. L, A. R. Martinez, and J. L. Solka. 2017. *Exploratory Data Analysis with MATLAB, Third Edition*, CRC Press.
9. Tenenbaum, de Silva, & Langford. 2000. "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290:2318-2323. http://web.mit.edu/cocosci/isomap/isomap.html
10. Fraley & Raftery, 2002. "Model-based clustering, discriminant analysis, and density estimation: MCLUST," *Journal of the American Statistical Association*, 97:611-631.
11. Savitsky, T.D. 2016. "Scalable Approximate Bayesian Inference for Outlier Detection under Informative Sampling," *Journal of Machine Learning Research*, 17:1-49.
12. https://cran.r-project.org/
13. Martinez, W. L. and L. Tan. 2015. "Categorizing sentiment using unstructured text," Joint Statistical Meetings, Seattle, Washington.