Liquor Sales and Unemployment Data in Iowa(2012-2021)

EMRE ÖZZEYBEK, Middle East Technical University, TR
ORÇUN S. TANDOĞAN, Middle East Technical University, TR

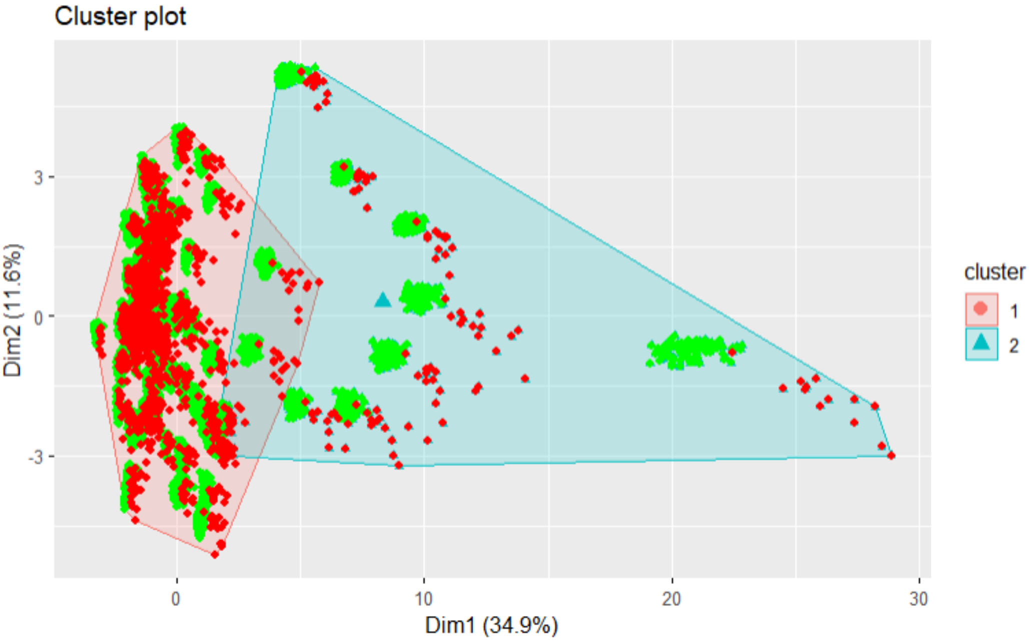


Fig. 1. The real clusters in our presentation

In this article we will be trying to visualize and find clusters of a dataset with representing liquor sales and unemployment information for the state of Iowa for 2012-2021 which will focus on the affect of the pandemic. We initially set out to find 2 clusters and we presented using that assumption. But afterwards we realized there are actually 99 clusters in the data.

CCS Concepts: • **Computing methodologies** → Classification and regression trees; **Cluster analysis**; **Dimensionality reduction and manifold learning**; **Anomaly detection**.

Additional Key Words and Phrases: dataset, projection, clustering, embedding, dimension reduction

ACM Reference Format:
Emre Özzeybek and Orçun S. Tandoğan. 2022. Liquor Sales and Unemployment Data in Iowa(2012-2021). In . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

CENG 574 Final Paper, January 22, 2021, Ankara, TR

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCING THE DATA

This dataset has 2 tables: The first one shows the state of employment, liquor sales and some census data for counties in the state of Iowa from January 2012 to April 2021. The second one is the summation of the the data such that each month has a single data point. The second table is only relevant to our research for the time_index's for the months which are pretty generic.
The main table we created consists of 33 attributes and 11071 points. Almost all of the data are numerical. Those (month-year, county) that aren't have corresponding integer values in the data itself.

2 PRE-PROCESSING

2.1 Handling Missing Values

We first started with replacing some $NULL$ values with 0s where necessary. Some of the $NULL$ valued data points had to go. After eliminating the missing values we were left with 10972 points.

2.2 Handling Outliers

We initially removed the what we thought were outliers. After the presentation we realized that these points could be in fact the post-covid data. So we decided not to remove them.

2.3 Normalizing

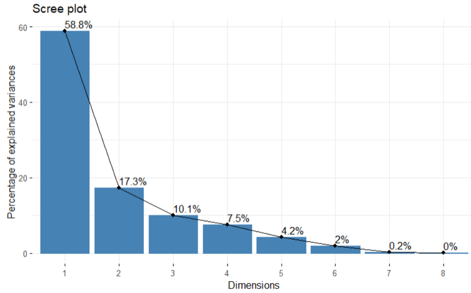
We normalized the data making each of the points to be between 0 and 1 by dividing the values to the maximum value of the column which was easy since we had mostly numeric data. We added $log(1.1)$ to all the values in order not to get errors if we were to sample a all-0 column.

2.4 Sampling

For tasks with higher computational complexity we used samples of 500, 1000, 2000 depending on the task.

3 PROJECTION OF THE DATA

3.1 PCA



We used PCA function from the $FactoMineR$ package. From the screeplot we decided to use $perc\_race\_multi$, n_liters, $unemployment\_rate$, $unemployment\_rate$, $median\_age$, $deaths$ and $cases$. This doesn't really make sense in terms of the screeplot. But the least varying data we had were $deaths$ and $cases$ and we did not want to lose them since we were trying to represent more of the COVID side of the data.

3.2 MDS

We used $cmdscale$ function from the $stats$ package to apply Classical Multidimensional Scaling on the distance matrices we created from the data to try to visualize our data in 2 dimensions. For the distance matrices we used euclidean and manhattan distances.

3.3 t-SNE

We used *Rtsne* function from the *Rtsne* package to try to see if there are any clusters in our data. We were not able to identify any when we were looking for 2 clusters. A suspicion we have is that the presence of 99 clusters may be apparent here.

3.4 IsoMAP

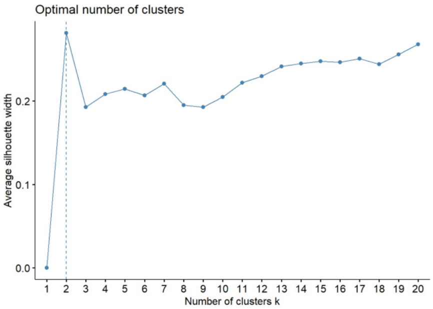
We used *isomap* function from the *vegan* package in hopes of finding clusters in our data. There were many lines. These could be the county clusters.

4 CLUSTERING OF THE DATA

4.1 Hierarchical Clustering

To see which works best we cluster using 3 different methods: Single Linkage(MIN distance between clusters), Average Linkage(AVG distance between clusters) and Complete Linkage(MAX distance between clusters). We used *hclust* function from the *stats* package. Since hierarchical gives a clustering anyway it was hard to interpret. When we saw 2 clusters of $\sim$ %10 and $\sim$ %90 we thought they meant the COVID months versus the other months because it corresponded similarly to the data. In hindsight it probably clustered the counties.

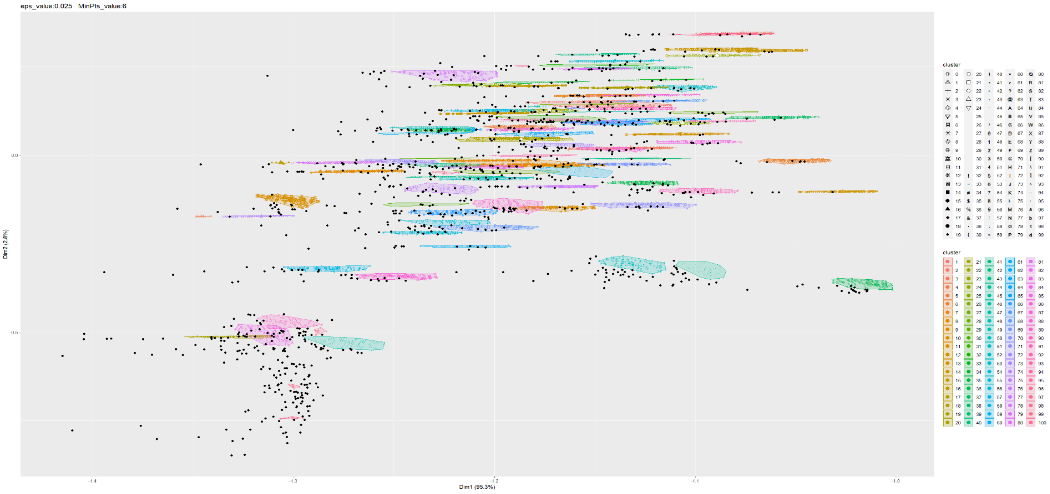
4.2 k-Means Clustering



We set out trying to find 2 clusters. We of course tried other k values; but not 99. For $k = (2 : 20)$ $k = 2$ had the best dunn index results. We used *kmeans* function from the *stats* package. We used different k values and for each k value tried different initializations. We selected the clustering with the best dunn score among these.

4.3 DBSCAN

We used *dbscan* function from package *dbscan* to cluster the data. We were skeptical using this clustering method since the Post-COVID data was very irregular. But it ended up being useful showing the 99 clusters for the counties since they are so tightly clustered. We used $kNNdistplot$ function from the *dbscan* package to find the suitable initial $eps_{v}alue$ to start looking and then adjusted this value.



5 VALIDATION

5.1 Validation metrics

Using the *clValid* function from the package *clValid* we looked at at fit of possible clusters.

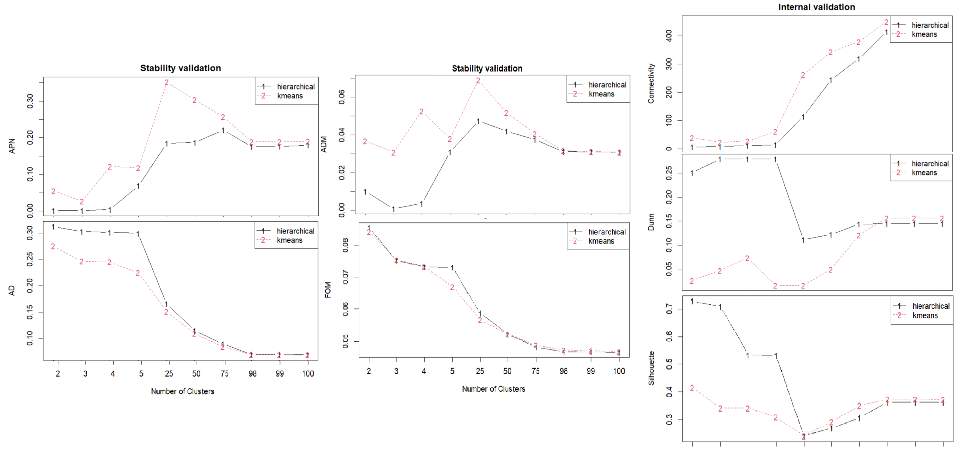


Fig. 2. Validation results for clustering

5.2 Shepard(MetaMDS)

6 RESULTS

6.1 Manually Changing the Data

We first thought the data would have 2 clusters:Pre-Covid,Post-Covid. We later realized that it is actually 99 clusters which is coincidentally the number of counties in the data.

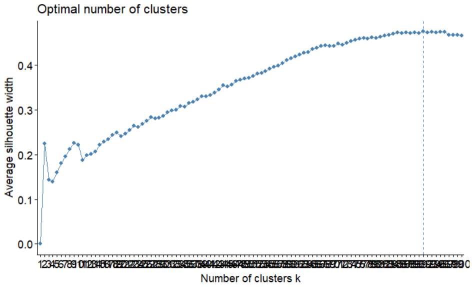


Fig. 3. Evidence of county information domination

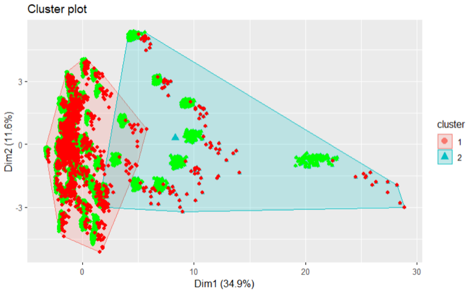


Fig. 4. The real labels of the 2 clusters

So we tried to manually get rid of the county influence in the data by eliminating some columns ourselves(educational, racial, age profile of residents, some redundant employment information, geographical information) but the results did not change in the way we expected them to. This issue is most probably because the data was constructed using different data for liquor sales, COVID information and county information and the county information turned out to be the most variant and identifying.

7 OTHER WORKS USING THE SAME DATA

There is 1 work(We Might be Wrong About What's Causing Pandemic Drinking) using this data. They are the ones that generated this data using 3 real life data sets. In it they are trying to find a correlation between unemployment and liquor sales by comparing the data that has both pre and post COVID data.

REFERENCES

[1] Rafal Ablamowicz and Bertfried Fauser. 2007. *CLIFFORD: a Maple 11 Package for Clifford Algebra Computations, version 11.* Retrieved February 28, 2008 from http://math.tntech.edu/rafal/cliff11/index.html

[2] Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. https://doi.org/10.1145/1188913.1188915

[3] Sten Andler. 1979. Predicate Path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN symposium on Principles of Programming Languages (POPL '79).* ACM Press, New York, NY, 226–236. https://doi.org/10.1145/567752.567774

[4] David A. Anisi. 2003. *Optimal Motion Control of a Ground Vehicle.* Master's thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.

[5] Sam Anzaroot and Andrew McCallum. 2013. *UMass Citation Field Extraction Dataset.* Retrieved May 27, 2019 from http://www.iesl.cs.umass.edu/data/data-umasscitationfield

[6] Sam Anzaroot, Alexandre Passos, David Belanger, and Andrew McCallum. 2014. Learning Soft Linear Constraints with Application to Citation Field Extraction. arXiv:1403.1349

[7] Lutz Bornmann, K. Brad Wray, and Robin Haunschild. 2019. Citation concept analysis (CCA)—A new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by two exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper. arXiv:1905.12410 [cs.DL]

[8] Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry).* Ph. D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171.

[9] Jacques Cohen (Ed.). 1996. Special issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996).

[10] Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalances among conjunctive aggregate queries. *J. ACM* 54, 2, Article 5 (April 2007), 50 pages. https://doi.org/10.1145/1219092.1219093

[11] Bruce P. Douglass, David Harel, and Mark B. Trakhtenbrot. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. https://doi.org/10.1007/3-540-65193-4_29

[12] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

[13] Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. https://doi.org/10.1007/3-540-09237-4

[14] Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In *Proceedings of the first USENIX workshop on Offensive Technologies (WOOT '07)*. USENIX Association, Berkley, CA, Article 7, 9 pages.

[15] Torben Hagerup, Kurt Mehlhorn, and J. Ian Munro. 1993. Maintaining Discrete Probability Distributions Optimally. In *Proceedings of the 20th International Colloquium on Automata, Languages and Programming (Lecture Notes in Computer Science, Vol. 700)*. Springer-Verlag, Berlin, 253–264.

[16] David Harel. 1978. *LOGICS of Programs: AXIOMATICS and DESCRIPTIVE POWER*. MIT Research Lab Technical Report TR-200. Massachusetts Institute of Technology, Cambridge, MA.

[17] David Harel. 1979. *First-Order Dynamic Logic*. Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY. https://doi.org/10.1007/3-540-09237-4

[18] Lars Hörmander. 1985. *The analysis of linear partial differential operators. III*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. viii+525 pages. Pseudodifferential operators.

[19] Lars Hörmander. 1985. *The analysis of linear partial differential operators. IV*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. vii+352 pages. Fourier integral operators.

[20] IEEE 2004. IEEE TCSC Executive Committee. In *Proceedings of the IEEE International Conference on Web Services (ICWS '04)*. IEEE Computer Society, Washington, DC, USA, 21–22. https://doi.org/10.1109/ICWS.2004.64

[21] Markus Kirschmer and John Voight. 2010. Algorithmic Enumeration of Ideal Classes for Quaternion Orders. *SIAM J. Comput.* 39, 5 (Jan. 2010), 1714–1747. https://doi.org/10.1137/080734467

[22] Donald E. Knuth. 1997. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms (3rd. ed.)*. Addison Wesley Longman Publishing Co., Inc.

[23] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY.

[24] Leslie Lamport. 1986. *LaTeX: A Document Preparation System*. Addison-Wesley, Reading, MA.

[25] Newton Lee. 2005. Interview with Bill Kinder: January 13, 2005. Video. *Comput. Entertain.* 3, 1, Article 4 (Jan.-March 2005). https://doi.org/10.1145/1057270.1057278

[26] Dave Novak. 2003. Solder man. Video. In *ACM SIGGRAPH 2003 Video Review on Animation theater Program: Part I - Vol. 145 (July 27–27, 2003)*. ACM Press, New York, NY, 4. https://doi.org/99.9999/woot07-S422 http://video.google.com/videoplay?docid=6528042696351994555

[27] Barack Obama. 2008. A more perfect union. Video. Retrieved March 21, 2008 from http://video.google.com/videoplay?docid=6528042696351994555

[28] Poker-Edge.Com. 2006. Stats and Analysis. Retrieved June 7, 2006 from http://www.poker-edge.com/stats.php

[29] R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[30] Bernard Rous. 2008. The Enabling of Digital Libraries. *Digital Libraries* 12, 3, Article 5 (July 2008). To appear.

[31] Mehdi Saeedi, Morteza Saheb Zamani, and Mehdi Sedighi. 2010. A library-based synthesis methodology for reversible logic. *Microelectron. J.* 41, 4 (April 2010), 185–194.

[32] Mehdi Saeedi, Morteza Saheb Zamani, Mehdi Sedighi, and Zahra Sasanian. 2010. Synthesis of Reversible Circuit Using Cycle-Based Approach. *J. Emerg. Technol. Comput. Syst.* 6, 4 (Dec. 2010).

[33] Joseph Scientist. 2009. The fountain of youth. Patent No. 12345, Filed July 1st., 2008, Issued Aug. 9th., 2009.

[34] Stan W. Smith. 2010. An experiment in bibliographic mark-up: Parsing metadata for XML export. In *Proceedings of the 3rd. annual workshop on Librarians and Computers (LAC '10, Vol. 3)*, Reginald N. Smythe and Alexander Noble (Eds.). Paparazzi Press, Milan Italy, 422–431. https://doi.org/99.9999/woot07-S422

[35] Asad Z. Spector. 1990. Achieving application requirements. In *Distributed Systems* (2nd. ed.), Sape Mullender (Ed.). ACM Press, New York, NY, 19–33. https://doi.org/10.1145/90417.90738

[36] Harry Thornburg. 2001. *Introduction to Bayesian Statistics*. Retrieved March 2, 2005 from http://ccrma.stanford.edu/~jos/bayes/bayes.html

[37] TUG 2017. *Institutional members of the TeX Users Group*. Retrieved May 27, 2017 from http://wwtug.org/instmem.html

[38]]CTANacmart Boris Veytsman. [n. d.]. *acmart—Class for typesetting publications of ACM*. Retrieved May 27, 2017 from http://www.ctan.org/pkg/acmart

A RESEARCH METHODS

A.1 PCA

Principal component analysis is a linear projection method for feature selection and extraction. It allows the feature which has the highest contribution to changes in the data. This method takes principal components of the data and carry these components to a new coordinate space so features which direct variation of the data most, can be selected for dimension reduction. This dimension reduction allows us to deal with only the most necessary and meaningful features.

A.2 MDS

Multidimensional Scaling(MDS) is used to visualize multidimensional data in 2 dimensions. It uses the distances(in some cases dissimilarities) between pairs of elements to build a model.

A.3 t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data.

A.4 Shepard(Meta-MDS)

Shepard diagram is used to validate of projection methods. For example it can be used for t-SNE and MDS. Shepard works based on comparing distances between two points before and after the transformation which is applied. It checks how data points are apart from each other. The reason why it can be called validation is that Shepard diagram is used for goodness of fit of the applied projection methods by plotting the points of how they are apart from each other.

A.5 isoMap

Isomap is a non-linear projection method for dimensional reduction. It is a type of metric MDS methods, but in fact, it preserves distances and allows for misleading results. For example, there are some data points which seem closer to the nearest cluster in euclidean distance, but they may not be. The real clusters can only be shown by using geodesic distance. While euclidean distance ignores the shape of the dataset, as it only cares about shortest path, Isomap‘s geodesic distance method considers adjacent data points of the referenced points. So related data points can be clustered together without being connected to less relevant points just because it is closer to them in euclidean distance.

A.6 Hierarchical Clustering

Hierarchical Clustering is used to cluster the data in a way that there are layers from which you can choose to cluster your data depending on the specificity of your choice. It can either be agglomerative or divisive. Different inter-cluster distance metrics can be used to achieve different cluster specifications.

A.7 k-Means Clustering

k-Means Clustering is used to cluster numerical data. It takes uses the mean values of the clusters' members as distances. It starts with random k points, where k is given, and keeps adding to the clusters and adjusting the clusters until the best(lowest SSE) is achieved. Because the initial points are randomly selected it is not definitive and should be run multiple times to achieve the best result.

A.8 DBSCAN

DBSCAN is a spatial clustering algorithm. As spatial implies, this algorithm connects points taking into account two hyperparameters: epsilon and minimum points value. It tries to find closer points in certain areas and cluster them. DBSCAN repeats this process for each point to find the optimum clustering because not all points can satisfy the hyperparameters. For example, consider we are working on only for 2D, there is a (x,y) point. DBSCAN looks for the closest values in limit of epsilon value and tries to cluster them. When the (x,y) point has minimum points closer to itself in limit of epsilon value, DBSCAN assigns it as a cluster.

B ONLINE RESOURCES

https://www.rdocumentation.org/
https://www.displayr.com/goodness-of-fit-in-mds-and-t-sne-with-shepard-diagrams/
https://www.r-bloggers.com/2017/09/goodness-of-fit-in-mds-and-t-sne-with-shepard-diagrams/
https://en.wikipedia.org/wiki/Multidimensional_scaling
https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html