Mapping the Shape of the U.S. Economy: A Topological Data Analysis Approach with BallMapper

Ryan Johnson[*]

[?] Department of Mathematics, University of Alaska Anchorage, Anchorage, AK

Student: *johnson.ryan1019@gmail.com*[*]
Mentor: *scook25@alaska.edu*

KEYWORDS

Topological Data Analysis; BallMapper; Data Science; Economics; Macroeconomics; Topology;

ABSTRACT

Topological Data Analysis is (TDA) is a new data analysis method which gained popularity starting in the early 21st century. Currently, a large body of TDA research utilizes the traditional Mapper algorithm. We aim to expand the body of literature on BallMapper, a new, Mapper adjacent algorithm. Applying BallMapper to widely used, U.S. macroeconomic data from the Bureau of Economic Analysis (BEA), the Federal Reserve Bank (FRB), and Bureau of Labor Statistics (BLS), we do this is three parts. Specifically, we show an example of BallMapper's utility in exploratory data analysis which also provides an example of how to interpret BallMapper's output; analyze our topological graphs to notable historic economic events; and test the stability of BallMapper's output and the topological graph's features. Results show. [...]

INTRODUCTION

Topological Data Analysis (TDA) is a new and emerging field of data analysis that is increasing in popularity. At a high level, the traditional applications of TDA use two tools: an algorithm called Mapper that produces a network-like graph and and analysis technique called Persistence Homology. Mapper can be thought of carrying out a statistical analysis with just pictures, it might confirm a hypothesis but is not enough to say something is statistically significant. Persistence Homology on the other hand is more akin to doing an analysis of variance (ANOVA). It provides the statistical, in our case mathematical, backing to what we might see if we were to only plot the data. Moreover, the combination of Mapper and Persistence Homology is what forms the central argument for TDA: data contains an underlying shape.[1],[2]

For this analysis we will be focusing on the graph creation portion of TDA. Specifically, we are examining various macroeconomic indicators of the United States of America (U.S.) using a Mapper-adjacent algorithm called BallMapper(BM).[3] BallMapper is of particular interest to us because it significantly reduces the parameters to create a topological map. It simplifies traditional Mapper by reducing the need for the user to pass data through multiple functions, each with their own parameters.[4] One benefit to this reduction in parameters is that it removes some of the barriers to learning Mapper (or Mapper-like algorithms), and more generally, TDA. Those who have a mathematical background, introductory course in Topology, or conceptual knowledge of data science methods and algorithms will fare much easier. However, if you are not equipped with any of the aforementioned tools, TDA might seem unnecessarily complicated for data analysis.

Beyond the main objectives of this analysis, we hope that this paper is serve as an on-ramp for anyone interested in TDA but feels inundated with jargon upon early stage researching. Out objectives are to add to the small body of literature whose main focus is applications with BallMapper.[5] We go about this by first showing BallMapper's use in exploratory data analysis (EDA). Using our

[1] https://arxiv.org/pdf/2504.09042
[2] Chazal F and Michel B (2021) An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. Front. Artif. Intell. 4:667963. doi: 10.3389/frai.2021.667963
[3] https://arxiv.org/pdf/1901.07410
[4] add reference
[5] add Mapper survey paper refrence here

results from EDA, we then will compare BallMapper's outputs (topological graphs) to notable historic economic events [and maybe less known?]. Lastly we will test the stability and consistency of BallMapper's output. The conclusion will talk broadly about results and other considerations of note, and the discussion will elaborate on future work and questions for research.

and is increasingly being applied to From the small body of literature we have reviewed, TDABM has removed many of the parameters that are required for more traditional methods of TDA. Generally speaking, traditional TDA contains four broad steps, each with multiple user-defined parameters. Conversely, TDABM has reduced this process to an the user selecting their data, a coloring variable and an epsilon value – details on this follow below. What should be noted, although TDABM reduces the number of steps needed to produce similar outcomes of traditional TDA methods, we lose control of being able fine tune out outputs. We also find that interpreting results becomes more difficult. However, TDABM is still in its infancy, so there is not a large body of research on interpretation. (Dłotko 2019) One major motivation for this paper is a paper written by Dłotko, Rudkin, and Qiu (2019) that applied TDABM to a global macroeconomic dataset to compare how countries have evolved over time, their transformation from the Great Depression Era, and various views on wealth and inequality. We could not find other TDA literature found specifically focused on the macroeconomic economy of singular countries. Hence, our topic of choice.

METHODS AND PROCEDURES

PROCEDURES

Data Selection

This paper relied on three publicly available data from US government sources. The Bureau of Economic Analysis (BEA), The Federal Reserve Board (FRB), and The Bureau of Labor Statistics (BLS). The data were gathered using R using two application programming interfaces (APIs): one for data from the BEA, and the other from the Federal Reserve Economic Data (FRED) API. FRED aggregates data from national and international sources, as well as public and private sources. We additionally used recession dates based on business cycle contractions and expansions provided by the National Bureau of Economic Research (NBER).

These agencies were selected because of they are authoritative sources for U.S. economic data. Their widespread use in both the private and public sector gives us high confidence in the accuracy and integrity of the data. [6]

These three transformations allow us to interpret Housing Starts as a leading indicator, PPI as a measure of Supplier Inflation (Cost-Push Inflation), and CPI as a measure of Consumer Inflation (Demand-PUll Inflation).

Data Preparation

When looking at *Year Start* and *Year End* in Table 1, we excluded Employment Cost Index (ECI) and Foreign Direct Investment (FDI) due to their limited availability of years. Of the remaining data series, New Privately-Owned Housing Units Started (Housing Starts) has the smallest range and will provide the initial base range of years used for our analysis (1960-2024).

```
Error in relocate(mutate(map_df(all_annual, .id = "series", ~tibble(year_start = min(.x$year), : could r

Error: object 'data_summary_table' not found
```

Table 1: *Housing Starts is considered a flow because it is provided as a Seasonally Adjusted Annual Rate (SAAR).

Table 1 also shows which data came in annual, quarterly, and monthly time frequencies. This analysis focuses on an annual time frame so some data transformation needed to be carried out. Additionally, BallMapper is less effective and inefficient to use when the input data variation there is large variation in the scales of each dimension of the dataset (columns). [7] To aide this fact, we transformed our

[6] Hughes-Cromwick, Ellen, and Julia Coronado. 2019. "The Value of US Government Data to US Business Decisions." Journal of Economic Perspectives 33 (1): 131201346. DOI: 10.1257/jep.33.1.131

[7] [find reference]

data to either a rate change or a percentage change from the previous year where appropriate. This ensures all dimensions of our data stay on a 0-100 scale and helps with the time complexity of running BallMapper.[8]

The first data series we made adjustments to was Personal Income and Its Disposition (Personal Income). [9] Personal Income is reported in nominal dollars. Thus, it does not account for inflation, so we first adjusted it using the Consumer Price Index (CPI) to get income data to Real Dollars. [10]

$$\text{Real Dollars} = \frac{\text{Nominal Dollars}}{\text{CPI}} * 100 \tag{1}$$

The output of Equation 1 was then used to take the log difference (Equation 2) to get a percentage change from the preceding year. This log difference helps to make our data more linear which aides BallMapper's use of the standard Euclidean Distance between each row of data. [If BM did not use the Euclidian distance between points, we might not need to take log differences. However topic is for another paper.] [11] [12]

$$\Delta \ln(\text{Level}) = [\ln(\text{Level}_t) - \ln(\text{Level}_{t-1})] * 100 \tag{2}$$

We additionally need to do some transformations on Housing Starts, Producer Price Index - All Commodities (PPI), and CPI. All three of these sources were given only on a monthly time frame, so to get an annual value for them we used a simple arithmetic mean.[13] Once we have these data in annual form, we then find the change from the previous year using Equation 2 for their final data in our analysis.

```
Error in select(relocate(mutate(arrange(map_df(analysisData, .id = "series", : could not find function "
```

```
Error: object 'analysis_data_summary_table' not found
```

Table 2: All years for analysis are 1960-2024.

We are including the differnt parts of gdp, c + g+ i + net exports

METHODS

BallMapper

The Algorithm

Technical Audience

Suppose we have a dataset

$$X$$

of

$$N$$

with

$$K$$

dimensions such that

$$K > 2$$

.

[8] reference about timing
[9] Table 2.1 from the BEA's Interactive Data, National Data section.
[10] https://www.dallasfed.org/research/basics/nominal
[11] https://econbrowser.com/archives/2014/02/use-of-logarithms-in-economics
[12] https://www.numberanalytics.com/blog/log-transform-econ-how-to-guide
[13] PPI tracks only physical goods such as Farm Products and Feed, Industrial Commodidities, and Other Commodidites such as Aircraft, Steel, and Lumber.

General Audience

Before going over the algorithm there are a few For the following algorithm some useful definitions are as follows: - We will use the word pointcloud instead of dataset. - Conventionally the Euclidean distance metric is used; let

$$d = dist() is used as our distance metric, denote$$

The core steps of the TDABM algorithm are as follows:

1. From your dataset (pointcloud) select a random point, α_n, and draw circle with radius epsilon,

   $$\epsilon$$

   .
2. For some distance metric where

   $$d$$

   is the distance from, any points inside the circle will be associated with point

   $$\alpha_n$$

   .
3. Repeat steps 1 - 3 until there are no more points to select and you have a set

   $$\alpha_1,...,\alpha_n$$

   .
4. Draw an edge between

   $$\alpha_i$$

   and

   $$\alpha_j$$

   if they share a point(s) for our set

   $$\alpha_1,...,\alpha_n$$

   .

Although the above list five steps, it should be noted that this is just the algorithm and is all done with one function call.[14] From the steps above, the only thing the user of TDABM is responsible for is creation of a pointcloud and choosing an appropriate

$$\epsilon$$

. - Summary of algorithm - correlation considerations

Since BM is a developing method, we have not found any standardized way to finding a epsilon(s) that describes our data well. So to ensure that we do not miss any potential interesting outputs, our strategy is carried out # stages. We first started by taking out standardized pointcloud and found the upper and lower bounds of our graphs, 0.3-1.4 respectively. To find the lower bound we test various epsilons such that every record (row) of data is its own node in our graph. To find the upper bound we do the opposite and find an epsilon so large that we reduce our graph to just a few nodes, usually connected. After finding these initial bounds we aim to compute around 100 graphs at even intervals from our lower to upper bound. From this range of graphs we observe a sub-range of epsilon values where our graphs are not too noisy nor hide too much information. For this step we found that epsilon values 0.4-0.75 gave us an interesting range of graphs. We look for components (connected nodes) appearing or disappearing, significant changes in coloration in the graph, [...]. Finally, after looking at our reduced range we decided on an epsilon of

$$0.511$$

.

[14] There are packages available both in R and Python under the names BallMapper and pyBallMapper, respectively.

Exploratory Data Analysis (EDA) Application

For our exploratory data analysis exploration we colored our graph by the

$$Year$$

. The coloration of each node representing the average of all the years within in a node. The size of our node is commensurate with the number of years it houses. Instead of using the graph output provided by BallMapper, we created graph output using other R packages to bring clarity to our analysis. We have not changed the core structure of the BallMapper output, but rather the presentation. The largest changes of note to the graphs are: spread out nodes to avoid overlapping; display the edge weights between nodes, representing the number number of shared values between two nodes; custom color scale for accessibility.

Looking at our graph, we first observe the overall structure of our graph. There is one large connected component, one small component, and the rest singleton nodes. For the largest connected component, we immediately are interested in the three large nodes (24, 4, and 26) with more overlap than the rest of the component. Additionally, we also notice that each of the large nodes have some single node flares. This structure tells us that for this component, there are three distinct groups and each flare representing similar data to the group but perhaps an outlier in specific ways.

The small component that we see is of interest because it tells us us that these related groups distinctly different from all other groups but similar enough to each other to be connected. Seeing these small components like this sometimes can indicate outliers in data. In our case this could be economic years which were recessions or of extraordinary growth. We also could interpret our singleton nodes in the same manner. Each on representing an outlier year for a singular variable or combination of variables for our data.

Notes

11/29/20205 - inital range 0.3 - 1.4 - ran seq to find interesting maps (bm_loop_00), found that .4-.75 seem to produce the most interesting graphs so we will run another sequence to see if there are any nuances we should be interested in (bm_loop_01) - starting at 0.0447, maps have 3 components - around 0.468, small components start to have more points included. We see this reflected in sizing between connected points

11/30/2025 - run new maps with new graph function - 0.537 & 0.538 have a three node component? - 0.546 & 0.547 have 4 components? - 0.61-0.625 2 components but the shape changes b/c of points moving around - 0.677-0.75 2 components and various combine of nodes - Most Interesting Epsilon Values: -

12/1/2025 - Choosing handful of maps to see which to use for analysis - c(0.468, 0.473, 0.478, 0.487, 0.511, 0.600, 0.605, 0.614, 0.642, 0.716) - 0.475-0.482 stable graph, no change - 0.495 start to see 2-simplex form, persists throughout - 3-7-11 –> 3-6-10 –> - Noticing developing three big groups of nodes, three “phases” of economy? - talk about differnt shapes: 2-simplex, mickey mouse, spurs, lolipop holders, - make edge strenth acording edge length? #discussion - later year v past year 2-simplex pattern?

12/2/2025 - Narrowing down more which graph for analysis -

Small Comparative Analysis

- “New Deal”
- 08 financial crisis
- Covid
- Flash Crash 87
- Tech Boom
- AI Boom
- Oil Embargo/ high unemployment

Robustness and Stabilization

- Shuffle Dataset –> Run BallMapper
  - Look at:
    *

nodes

    * coloring consistency
    * look at financne application and brexit vote

CONCLUSION

DISCUSSION

REFERENCES

Old Paper

Old Intro

Dłotko demonstrates that, mathematically, TDABM produces a near similar output to that of traditional TDA methods via a fuctor map (Dłotko 2019). TDA BallMapper and closely follows other papers that have applied TDABM. We will begin with a general overview of TDABM: current published literature and useful concepts to know. We then will describe our data and reasoning behind their selection. Our Methodology and Results will touch on more theoretical concepts behind TDABM and then adress a couple of findings, respectively.

Procedures

Pointcloud Construction

A primary challenge we encountered when selecting variables for analysis was the variability of time period coverage. Some data spanned from Depression Era 1929 all the way up to 2024. Looking at our final nine variables (see) and the range of years available by by source (), we limited our years to 1960 - 2024 due to Housing Starts data having the smallest range of years.

Some of the data was only available by month. In these cases we calculated annual changes by finding the difference between January of the current year and January of the previous year. Note, this reduced the number of years in our data set by one. We also standardized the data by converting non-rate based data to percent changes from the previous year to ensure consistency across all variables (). One feature of TDABM is that it can be sensitive to data with large scale differences. This not only reduces how precise you can be when selecting a parameter, but can also significantly increase computation time.

Looking at our correlation table (), we observe moderate to strong correlations between the pairwise combinations of GDP, PCE, Ig, and Imports. We also see a strong correlation between CPI and Interest Rates. Since these correlations are moderate to high, it's worth noting that the shape of our output may be reflected by these correlations. They also could express themselves with similar coloring of TDABM graphs (Dłotko, Rudkin, and Qiu 2019).

Cloud Building

In our Data section, we noted that some calculations were needed before our analysis. Additionally, for the Federal Funds Rate, the Unemployment Rate, and CPI we had to take extra steps and considerations while constructing out pointcloud. Since the Federal Funds Rate and the Unemployment Rate are already a rate-based measurements, instead of taking a percent change from the previous year, we instead took the standard non-weight average rate by each year. For CPI, this data was given on a monthly-basis, we calculated it using our General Percent Change Formula with $k = 12$.

Results

Interpretation of TDA BallMapper Graphs, An Abridged Version

In we are presented with what looks like a graph, in the mathematical sense; a display of points (nodes) and edges –some connected while others are not connected at all (satellites). If we do not change the poincloud or ϵ of our TDABM graph, we will get the same "shape" of the graph regardless of the coloration we choose.

Nodes are colored depending on the users interest. In , we are focusing color by year. Coloration calculated by taking the average of the data that is "inside" each node, in our case the average value of the years. Further, if there is only one data point in a node (i.e. a satellite points), the node will appear small and the coloration will reflect as seen on the legend. Conversely, if the node is large, this indicates more data is "inside" the node, and the average value could be reflecting the mean values of a large or small variance.

General Remarks

We initially notice a large connected component[15] (component A) on the right side of our graph, as well as a smaller connected component (component B) on the lower left side. Noting the coloration, there are three main sections: the right, bottom, and upper-left. On the bottom and sweeping upward to the left we see there are multiple satellite points. These can be of interest because they may indicate outliers in our data.[16] Focusing back towards component A, we notice that the two largest nodes, 28 and 22, which seem to represent the late 2010s and the dot-Com boom, respectively. We also note that there is a arm coming off the upper-right portion of component A, as well as smaller, lollipop, features emerging at the bottom and top of component A.[17]

Time Travel

Looking at , we see that our graph is colored by year.[18] In this section we will only highlight a few observations in detail due to the nature of this paper. Additional areas for investigation can be found in the Conclusion.

Mentioned above, we observed the satellite points in the lower half of the graph. If we look at the upper-right quadrant, we see that it consists of nodes $\{1, 16, 25, 27\}$. Looking at we see that it covers the 2020 Pandemic and the Financial Crisis and its aftermath. Seeing a general coloration from the recent 10 - 15 years we would expect to see some of the recent outlier economic events here. Indeed, node 25 consists of two of the most recent economic downturns, 2009 and 2020. However, what is interesting is that node 27 consists of three years closely following the 2008-2009 Financial crisis. However, the years following the 2020 COVID Pandemic can be found scattered across nodes $\{6, 8, 24, 28, 29\}$. Looking at node 27 we find that the common thread is high unemployment. In contrast, the years following the Pandemic are not the same year to year.[19]

Shfiting our focus towards component A, it consists of data most similar to the economic years between the 1990s - 2010s. Meanwhile, looking at component B and the general lower area of our graph, we see that it represents the Great Inflation period of the 1970s - early 1980s [ivestopedia article]. Looking at nodes $12 - 20$, we get exactly ten years of data[20] where there was know high inflation, high unemployment, and in general know to be bad time economically for the US (). Something of interest to note is that nodes 14 and 15

[15] In graph theory, a connected component is one in which there exists a path from a node to every other node for a set of nodes and edges.
[16] We have not found any literature yet on whether this observation is empirically true.
[17] The observations mentioned above good starting places for interpreting BallMapper graphs.
[18] Supplementary graphs displaying coloration based on different variables can be found in Supplement_1.pdf.
[19] *2021:* An outlier, we see higher levels of unemployment and inflation. We see high PCE, Ig, and imports; *2022:* Nodes $\{8, 28\}$ we see that there is higher than normal inflation, export, and imports; *2023:* Nodes $\{6, 24\}$, we see low inflation and low unemployment.
[20] 1974 - 1984

consist of 1976 - 1978. This three year period based on data seems to show more normal economic conditions. On the other hand, we see that all the other years in this grouping (nodes $12 - 20$) show signs of a struggling economy in some way.

Discussion

This paper serves as a proof-of-concept of the usefulness of TDABM as a methodology for exploratory data analysis. By no means is this a totally comprehensive method to gain insight into data, but instead TDA BallMapper proves itself to be useful tool wehen employed in addition to other traditional analyses.

In our section we briefly went over two ways to interpret TDABM graphs. As seen in many other, longer papers written by Dlotko and colleagues, there is a lot of room to expound our TDABM graph here,though that is beyond the scope of this paper. In general, though, we can see that TDBM can give insight into our data that we might not see otherwise. Seeing that there is data that is not like the others gives an indication that there is something to investigate. Consider our comparison of the 08' - 09' financial crisis and the 2020 COVID Pandemic: We see learned that the years following the financial crisis were similar economically. Another way to view this is, clustering revels that the financial crisis was a singular problem that spurred the collapse of the financial system. Conversely, the COVID pandemic was a singular issue but exposed many different weaknesses in our current world.

For the future development of TDABM, and TDA in general, there are many possibilities of future research and one glaring downside. The downside to TDA is that it is a high barrier to entry to understand the methodologies inner workings. Those with a technical background will have an easier time, but nonetheless a higher barrier than methods such as linear regressions. Subject wise, TDA has an advantage in some of the social sciences because it can be somewhat of a bridge between qualitative and quantitative analyses. TDABM has been used to analyze the Brexit Vote data, seeking to understand quantitative and qualitative motivations behind its outcome Rudkin et al. (2023). TDA also has the advantages of dimension reduction which could be useful to the life-sciences. Fields such as genetics, chemistry, and biology have many fields of data, and being able to consolidate it into an understandable form could benefit the body of knowledge greatly.

References

Data Source

- Bureau of Economic Analysis
  - Real Exports and Imports
  - New Foreign Direct Investment in the United States
  - Real GDP
  - Personal Income and Its Disposition
- Federal Reserve Bank
  - Selected Interests Rates
- Bureau of Labor Statistics
  - CPI
  - Employment Cost Index
  - PPI
  - Housing Starts
  - Unemployment Rate

Dłotko, Paweł. 2019. "Ball Mapper: A Shape Summary for Topological Data Analysis." https://arxiv.org/abs/1901.07410.

Dłotko, Paweł, Simon Rudkin, and Wanling Qiu. 2019. "Topologically Mapping the Macroeconomy." https://arxiv.org/abs/1911.10476.

Rudkin, Simon, Lucy Barros, Paweł Dłotko, and Wanling Qiu. 2023. "An Economic Topology of the Brexit Vote." *Regional Studies* 58 (3): 601–18. https://doi.org/10.1080/00343404.2023.2204123.