

## Paper Outline [Removed after Review]

- Abstract [x]
- Introduction
- Methods and Procedures
  - Procedures
    - \* Literature
    - \* Data
    - \* Mapper/BallMapper
  - Methods
    - \* Descriptive Analysis of BM Graph
      - Groupings
      - Statistical correlation and other known interpretations
      - difference epsilon testing
    - \* time travel - coloring by year
    - \* robustness check with randomization
    - \* other colorings?
  - Results
- Discussion
- Conclusion
- References

## 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](mailto:johnson.ryan1019@gmail.com)\*

Mentor: [scook25@alaska.edu](mailto: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 in three parts. Specifically we show BallMapper's utility in exploratory data analysis, compare the topological graphs produced to notable historic economic events, and test the stability topological graphs and their features. Results show, [...]

We aim to expand the body of literature using BallMapper. The latter is a sub-field of Algebraic Topology which provides mathematical certainty about what we are seeing from the Mapper algorithm. The former, is a visualization tool and can be used for exploratory data analysis.

# INTRODUCTION

Topological Data Analysis (TDA) is a new and emerging field of data analysis that is increasing in popularity. Conceptually, the traditional applications of TDA have two parts, a algorithm called Mapper and Persistence Homology. These two parts can be thought of graph creation and persistence of common features across many graphs.<sup>1</sup>

The first part, graph creation, is commonly carried out with an algorithm called Mapper which produces a topological map.<sup>2</sup> The second, persistence, creates many graphs and analyzes common features of each topological map to understand the underlying relationships in data. These underlying relationships in data form a central argument for TDA: data contains an underlying structure.

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).<sup>3</sup> 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.<sup>4</sup>

One benefit to this reduction in parameters is that it removes some of the barriers to entry to learning Mapper (or Mapper-like algorithms), and more generally, Topological Data Analysis (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 knowledge, TDA will seem unnecessarily complicated for data analysis.

One main objective of this paper is serve as an on-ramp for anyone interested in TDA but feels inundated upon early stage researching. We also want to add to the small body of research whose main analysis is with BallMapper. Our last goal is to improve literature on how to discern BallMapper graphs.

What follows is a summary of current literature and motivation for this paper. We then address our data sources and calculations for analysis. Third we will talk about BallMapper and Mapper theory, giving both technical and non-technical understanding. We then will touch on tools and test for [robustness] of BM. Our results comprise of: exploratory data analysis, changing BallMapper parameters, and the robustness checks. The discussion and conclusion will elaborate on where BallMapper fits into the TDA ecosystem, as well as talk about future work and applications.

## METHODS AND PROCEDURES

### *Data*

TDA's increasing popularity since the early 21st century and has led to a wide array of applications as well as development in software packages.<sup>5</sup>

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.<sup>6</sup>

Looking the *Year Start* and *Year End* in Table 1, we elected to exclude Employment Cost Index (ECI) and Foreign Direct Investment (FDI) due to their limited available years. Of the remaining data series, New Privately-Owned Housing Units Started (Housing Starts) has the smallest range and will provide the range of years used for our analysis - 1959-2024.

---

<sup>1</sup><https://arxiv.org/pdf/2504.09042>

<sup>2</sup>Need to explain what Topological map is?

<sup>3</sup><https://arxiv.org/pdf/1901.07410>

<sup>4</sup>add reference, seems appropriate here.

<sup>5</sup>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

<sup>6</sup>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

Series Name	Year Start	Year End	Data Frequency	Data Type	API
Gross Domestic Product	1930	2024	Annual	Flow	BEA
Personal Income & Outlays	1929	2024	Annual	Flow	BEA
Foreign Direct Investment	2014	2024	Annual	Flow	BEA
Federal Funds Rate	1955	2024	Annual	Rate	FRED
Employment Cost Index	2001	2025	Quarterly	Index	FRED
Consumer Price Index	1947	2025	Monthly	Index	FRED
New Privately-Owned Housing Units Started	1959	2025	Monthly	Flow*	FRED
Producer Price Index - All Commodities	1913	2025	Monthly	Index	FRED
Unemployment Rate	1948	2025	Monthly	Rate	FRED

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

Table 1 also shows that our data came in annual, quarterly, and monthly time frequencies. We are interested in annual data for this analysis so some data considerations are as follows because not all our data provided in Real dollars, were given as a rate (growth rate, proportion rate, or fixed rate), or on a similar scale (i.e., rate, millions of dollars, index). 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). <sup>7</sup> To aide BallMapper, we transformed our data to either a rate 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 BallMapper.

## Pre Analysis Calculations

The first data series we needed to make adjustments for was Personal Income and Its Outlays; specifically, Personal Income and Its Disposition (Personal Income). <sup>8</sup> Personal Income is reported in nominal dollars, which does not account for inflation, so our first calculation is to adjust Personal Income for inflation using Consumer Price Index (CPI) data.

$$\text{Real Personal Income} = \frac{\text{Personal Income}}{\text{CPI}} * 100 \quad (1)$$

We then take the output os Equation 1 and take the log difference and multiply it by 100 to get a percentage change for the preceding year.

$$\Delta \text{Real Personal Income} = \ln(\text{Real Personal Income}_t) - \ln(\text{Real Personal Income}_{t-1}) \quad (2)$$

that came in monthly frequencies, two different calculations wered done based on the type of data (see Table 1). Data labeled type *Flow* <sup>9</sup> and was not already given as a percent growth, we took a two step calculation: first we found the annual average by calculating the non-weighted arithmetic mean. After finding the

Some of our data was only available in monthly time frames [insert data names] and were initially interested in quarterly and annual data. So to get representative data on our time periods of interest we took the arithmetic mean of their respective time intervals see in **equation\_quarterly** and **equation\_annual**.

Data for analysis is on an annual change. Converted data which are not a rate (fed funds rate and unemployment) or changes year over year (gdp), to reflect percent change from year to year. For persona income and it's disposition, also need to calculate the real amounts instead of the reported nominal [insert reference]. For this we will take the nominal value/cpi X 100 for the real value, and then take the log difference year to year to measure changes by year. <sup>10</sup>

PPI\_All only include Farm products, processed foods and feeds, Industrials, and other commodities such as rubber, lumber, machinery, and aircraft since 1959. Most all services only have data series from 2009 onwards.

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

<sup>7</sup>[find reference]

<sup>8</sup>Table 2.1 from the BEA's Interactive Data, National Data section.

<sup>9</sup>This data is measuring some quantity over time such as Personal Income

<sup>10</sup><https://econbrowser.com/archives/2014/02/use-of-logarithms-in-economics>

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 our 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. ([Pawel Dłotko 2019](#)) One major motivation for this paper is a paper written by Paweł 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.

## TDA BallMapper

TDABM, and in general TDA, believes that data contains a “shape” to it. It pulls ideas from the branch of Mathematics, Topology. At a high-level, Topology focuses on the properties of geometric shapes when you deform them without breaking them, e.g. bend, twist, scrunch. The following is a basic overview of TDABM <sup>[11](#)</sup>.

## BallMapper Algorithm

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  $d = \text{dist}()$  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,  $\alpha_n$ , 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  $\alpha_1, \dots, \alpha_n$ .
4. Draw an edge between  $\alpha_i$  and  $\alpha_j$  if they share a point(s) for our set  $\alpha_1, \dots, \alpha_n$ .

Although the above list five steps, it should be noted that this is just the algorithm and is all done with one function call.<sup>[12](#)</sup> From the steps above, the only thing the user of TDABM is responsible for is creation of a pointcloud and choosing an appropriate  $\epsilon$ .

## 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 [?@fig-corr](#)) and the range of years available by source ([?@fig-yearspan](#)), 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 (Equation 3). 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 ([?@fig-corr](#)), 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 ([Pawel Dłotko, Rudkin, and Qiu 2019](#)).

---

<sup>11</sup>Some very good examples and explanations can be found in ([Dłotko, Qiu, and Rudkin 2022](#)) and ([Pawel Dłotko, Qiu, and Rudkin 2019](#)) which touch on the finer details of TDABM algorithm. A very good illustrative example can be seen in the pre-print ([P. Dłotko, Qiu, and Rudkin 2021](#)).

<sup>12</sup>There are packages available both in R and Python under the names BallMapper and pyBallMapper, respectively.

## 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 our 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  $k = 12$ .

The analysis reportes here are guided by Paweł Dłotko, Rudkin, and Qiu (2019)'s approach. We will be referencing our TDABM output in Figure 1 for the remainder of this section.

## Results

### Interpretation of TDA BallMapper Graphs, An Abridged Version

In Figure 1 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  $\epsilon$  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 Figure 1 Remarks

We initially notice a large connected component<sup>13</sup> (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.<sup>14</sup> 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.<sup>15</sup>

### Time Travel

Looking at Figure 1, we see that our graph is colored by year.<sup>16</sup> 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 ?@tbl-covidplus 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.<sup>17</sup>

Shifting 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

<sup>13</sup>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.

<sup>14</sup>We have not found any literature yet on whether this observation is empirically true.

<sup>15</sup>The observations mentioned above good starting places for interpreting BallMapper graphs.

<sup>16</sup>Supplementary graphs displaying coloration based on different variables can be found in Supplement\_1.pdf.

<sup>17</sup>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.

1980s [investopedia article]. Looking at nodes 12 – 20, we get exactly ten years of data<sup>18</sup> where there was know high inflation, high unemployment, and in general know to be bad time economically for the US (@tbl-greatinflation). Something of interest to note is that nodes 14 and 15 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 Results 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.

## Conclusion

### Endnotes

General Percent Change Formula

$$\Delta p_n = \frac{p_n - p_{n-k}}{p_{n-k}} \text{ for some } p \in P \text{ and } n > 0 \quad (3)$$

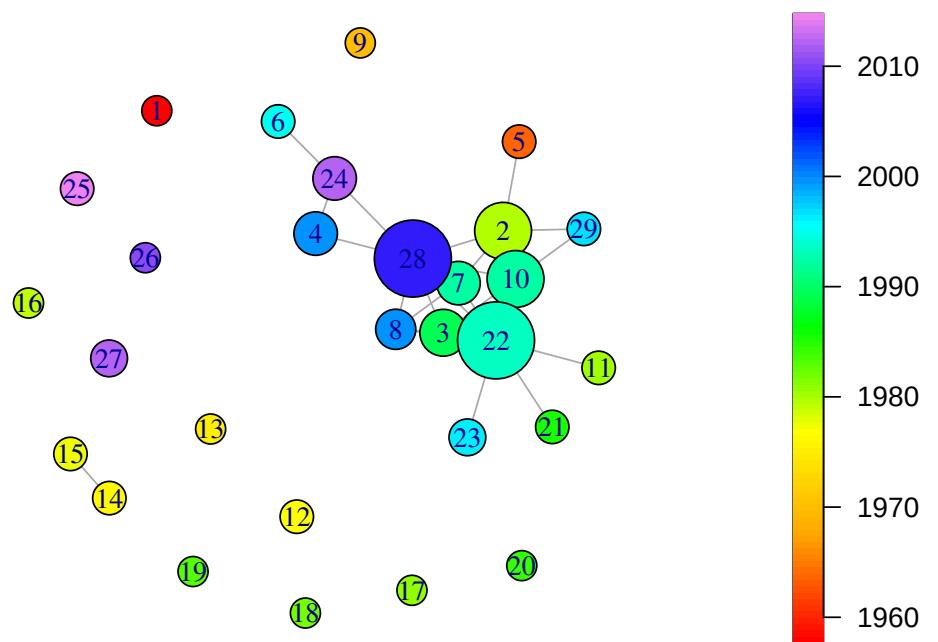
where  $P$  is a column of our data with  $x_1, \dots, x_n$  observations and  $k$  is a lag integer such that  $k > 0$ .

## Appendix

---

<sup>18</sup>1974 - 1984

### BallMapper output Colored by Year



Epsilon: 0.45

Figure 1: Data is normalized on a [0,1] scale due to all varialbe not being normally distributed See Supplement\_2.pdf.

## References

- Dłotko, Paweł, Wanling Qiu, and Simon Rudkin. 2022. “Topological Data Analysis Ball Mapper for Finance.” <https://arxiv.org/abs/2206.03622>.
- Dłotko, Paweł. 2019. “Ball Mapper: A Shape Summary for Topological Data Analysis.” <https://arxiv.org/abs/1901.07410>.
- Dłotko, Paweł, Wanling Qiu, and Simon Rudkin. 2019. “Financial Ratios and Stock Returns Reappraised Through a Topological Data Analysis Lens.” <https://arxiv.org/abs/1911.10297>.
- Dłotko, Paweł, Simon Rudkin, and Wanling Qiu. 2019. “Topologically Mapping the Macroeconomy.” <https://arxiv.org/abs/1911.10476>.
- Dłotko, P., W. Qiu, and S. T. Rudkin. 2021. “Financial Ratios and Stock Returns Reappraised Through a Topological Data Analysis Lens.” *The European Journal of Finance* 30 (1): 53–77. <https://doi.org/10.1080/1351847x.2021.2009892>.
- 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>.

## Data Sources

- 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 Interest Rates
- Bureau of Labor Statistics
  - CPI
  - Employment Cost Index
  - PPI
  - Housing Starts
  - Unemployment Rate

## Old Intro Material

Dłotko demonstrates that, mathematically, TDABM produces a near similar output to that of traditional TDA methods via a functor map (Paweł 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 address a couple of findings, respectively.