TDA for Time Series Clustering and Predictive Model Generalization

Asgard Mendoza-Flores[1], Diego Gutiérrez-Vargas[1], Daniel Arana-Bodart[1], Diego Mijares-Ledezma[1], Mauricio Valencia-González[1], AlejandroUcán-Puc[2], LiliaAlanís-López[2]

[1]Undergraduate, Data Science and Mathematics Engineering, Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Ave. Eugenio Garza Sada 2501, Monterrey, N.L., México, 64849.
[2]Professor, Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Ave. Eugenio Garza Sada 2501, Monterrey, N.L., México, 64849.

Contributing authors: A00572566@tec.mx; A01285421@tec.mx; A01741202@tec.mx; A01722421@tec.mx; A01234397@tec.mx; alejandro.ucan-puc@tec.mx; lilia.alanislpz@tec.mx;

Abstract

Steel consumption in Mexico has experienced significant fluctuations in recent years, influenced by various economic and social factors. As a vital industry for infrastructure development and manufacturing, steel production faces the ongoing challenge of balancing supply, storage, and sales in an increasingly volatile and competitive market. The lack of precise tools for demand forecasting has led to uncertainty, causing economic losses, shortages, or overproduction, which directly impact profitability and long term sustainability for steel producers nationwide. This project proposes a novel approach to predicting steel consumption trends using topological data analysis (TDA) by mapping multiple steel-related time series as well as non-steel related time series with Mapper functions. By identifying similar time series within critical historical data, we aim to cluster patterns and develop predictive models based on these groupings creating robust clusterizations and forecasting models to optimize the production strategies and mitigate instability into the future behavior of steel production in Mexico.

Keywords: Topological Data Analysis, Steel, Time-series, Mapper, Clusterization, Forecasting Model

1 Introduction

The steel industry is internationally regarded as a vital and strategic sector because of its relevance within every supply chain as well as most construction projects, and more recently as a financial investment across national and international markets. According to INEGI[1] the steel industry in Mexico generated an estimated production value of more than $179 million MXN and almost contributing 2% to the total GDP in of the country.

One of the biggest challenges in the steel industry is to accurately forecast future demand of steel-related products, in part because steel production processes are slow and the storage of these products is expensive and complicated[2]. An accurate steel demand forecasting is vital to adjust production, manage inventories and make better decisions regarding future investments and areas of opportunity in the market.

However, understanding the factors that drive steel demand is a complex task with multiple correlating causes. As explored by Basu and Dlotko[3] where, intense correlation was found with other non-traditionally related commodities as well as futures contracts and futures returns yielding. There's an increasing amount of literature. Another complex key piece to understanding the market price of commodities in modern economies arises from studying the market hedging pressure, speculative pressure, and spot market volatilities, as pointed out by Shanker (1980)[4].

In the mentioned case, the study of co-movements as well as commercial hedging pressure (CHP), futures contracts as well as roll yield across multiple non-traditionally related commodities in the US market during a 20-year span captured accurately key periods of transformation within the market value of the commodities, as well as heavy correlation-based dissimilarities within quarterly periods of any given year. The use of this TDA based methodology has proven to be useful in determining strong linkages between time periods as well as different commodities that exist in a given financial market.

In the other hand there's a rich literature regarding more commonly used approaches for future decision making in smart manufacturing which can be reviewed in Uray et al.[5] study of state of the art methodologies for industry demand forecasting. In this study the authors analyze possible applications of Mappers, persistent homology (PH) and uniform manifold approximations within the supply chains of industry 4.0.

The author validates the relevance of these techniques given that an adequate geometrical representation of the data is made and allows for the topological properties within the data to be manipulated accordingly.

In a similar way to Basu & Dlotko[3] this project studies the topological relations of steel production in Mexico with historical data regarding steel production, steel imports, steel exports, as well as national apparent consumption of steel (CNA in Spanish) within and outside the country, as well as historical time series from 2012 to 2024 of the prices of other non-traditionally related commodities in Mexico and across the world that were hypothesized to have an impact on historical steel demand fluctuations and its topological relation with the historical behavior of multiple commodities both related and unrelated to steel. With these insights, the time series were clustered using topological mapping tools and a wide array of processing functions for

the original time series, and these clusters were used to enrich future predictions of the demand for critical steel-related products that were of particular interest.

This approach of mapping historical data comprehending the development of a phenomena in different geographical spaces was heavily influenced by Chen & Volic 2021[6] where they were able to prove the value of using topological data mapping to reconstruct and cluster the development of the COVID-19 pandemic across all of the US capturing growth rates, regional prominence and proliferation of cases and the development of hot-spots, that traditional ways of data visualization aren't able to identify while being based solely in the geometry and topological properties in the data. In the paper, other applications of Mapper algorithms for understanding how socioeconomic factors can correlate with important historical events of crisis are presented such as Dlotko & Rudkin 2020[7]. Such applications of Mapper as a topological clusterizations tool adapt better to the clusterizations that have been developed for the identification of the dynamics that define the demand and consumption of steel across Mexico. To finalize this project, an evaluation of possible higher-dimensional cyclical tendencies within the clustered time series was made utilizing the topological technique of Takens' Embeddings and sliding windows to comprehend if the behavior can be reliably forecasted for long time periods.

This methodology has been previously used in Goel, Paricha & Mehra (2020) [8] topological data analysis for investment decisions in which an application of the Takens' Embedding theorem as a reconstruction of time series into higher dimensional spaces can preserve the intrinsic topological properties within the time series and can then be identified utilizing persistent topological patterns which can enrich the comprehension of the desired time series better than traditional statistical methods which are also commonly limited by the requirement of a priori assumptions regarding the distribution of the data, whereas Topological Data driven approaches such as Takens' Embeddings are designed to manage point clouds not uniformly distributed and can preserve even in non-linear geometries without a trivial topology.

2 Objective

The proposed analysis consists of making agglomerations with the time series utilizing the Mapper algorithm over different data transformations such as ROI, Beta Coefficient, CV squared, and Autocorrelation Function (ACF).

After analyzing and determining the best agglomerations over the data, a series of different models will be proposed to find the best model prediction adjustment over that cluster of time series. Finally, using the technique of Takens Embeddings, it will be determined if the cluster of time series presents any type of cyclic behavior to validate if any model adequate to the future predictions of that specific cluster could have medium or long-term usage.

3 Database pre-processing

The data provided for this project can be divided into three different groups: steel-related products production data from a large company focused on selling processed and raw steel, including national apparent consumption, and interest variables for

them such as prices and demand; importations and exports of general steel products in Latin America (particularly Mexico, Brazil, Argentina, Chile, Colombia and Peru); and lastly, time series on the price fluctuation from diverse commodities "unrelated" to steel.

This data as a whole had some adjustments that needed to be made; for instance, there was not a base type of chart, and some format changes had to be done manually in order to standardize the database so cleaning it and applying transformations were possible through programming. Due to the great volume of data available in each group presented earlier, this standardization process still had to be split. Once this was done, each group was analyzed normally with a statistical approach.

Null data was not immediately deleted or imputed. Every time series with two or more null data records in a row was instantly deleted, since there are no reliable ways of imputing them. However, if there was only one null data record in a row, this missing value was imputed through a 3-point moving average MA(3), with this several time series did not have to be deleted.

Once the data was cleaned, all three groups were merged into a database containing all the information that would be later worked using topological data analysis. The total dimension of the database was 205 columns (time series) and 144 rows (months), resulting in 12 years of information regarding steel consumption and prices from different commodities.

Lastly, in order to use Mapper to group the data, all of the time series need to be transformed into a single point so the mapping algorithm can separate them into different groups. The data was transformed through four different methods: Return on Investment (ROI), a financial metric which determines how profitable an investment is; the beta coefficient, which represents the change in the dependent variable for every one-unit change in the independent variable; the CV^2 factor, which is a way to measure the variability of data with respect to the mean; and the autocorrelation function (ACF) which quantifies the correlation between a time series and its lagged values, revealing patterns and dependencies within the data. An additional method was later construed using ROI, beta coefficient and CV^2, to explore if t.

4 Modelation and Validation

Once the data was preprocessed and transformed into a single point per time series (ACF required dimensionality reduction since it transformed the series into vectors) the first step in order to make a model is to use TDA to group time series with a similar topological behavior. A design of experiments was used to create several Mappers for each transformation method, resulting in a bunch of clusters containing different time series that were grouped depending on the similarity given by the transformation method used.

After the clusters were created, they had to be analyzed to determine which time series had a similar topological behavior. In order to accomplish this, a Python script was created which plots a given cluster, and to check if their behavior is similar, it scales every time series so each member of the cluster begins in the exact same point. With this, it was very easy to determine if the transformation method was a proper method to differentiate the time series.

Given a time series cluster, five different models were created on the first member of a cluster (the one with the lowest index). These models were a 3-point moving average (MA3), simple exponential smoothing (SES), a linear regression, Holt and ARIMA. In general, simple predictive time series methods were used on a single time series in a cluster. These models were trained with all but the last year of available data, and they were tested by predicting the last year of available data.

In order to determine the best predictive model, each model had its sMAPE and MAPE calculated, and the one with the lowest value would be the model that best describes the time series; however, since the objective is to predict different time series from the same cluster, then these models have to be escalated to fit into the other time series. Then, each model is escalated, and the one with the lowest sMAPE and MAPE in the last year is the one chosen as the model that best describes each member of the cluster.

Next, in order to determine if this whole process was worth it, each member of the cluster had the five models created for themselves; this was made with the purpose of determining how separate the models created exclusively for each member of the cluster were, compared to the models created for one member, but escalated. In order to determine this, each model had its MAPE and sMAPE compared with the lowest one from the original time series.

Lastly, a Takens Embedding was created for each time series of the cluster, in order to determine if the whole cluster had a cyclic behavior. This was made to know if the model created could be useful in a long-term period or if it should be adjusted more frequently as time passes and more data are involved.

5 Results

5.1 Clusters

The results given by the different clusters from the parameters of the mappers mostly were found well by the ACF method, followed by mappers utilizing ROI, beta coefficient and CV^2 combined into a single projection method. The groups extracted from the ROI, CV^2, and beta coefficient by themselves did not yield good results; thus, a combination of them was tested, which ended up yielding better results.

The differences among both projections were that ACF yielded better results than the combination of the other methods; however, since it required a dimensionality reduction to transform the vectors into points, so the mapper could group them. Then several time series, which did not have enough similarities among others, were instead deleted and not accounted for by the mapper. This means that ACF results were better but far fewer in comparison with the projection from combining beta coefficient, ROI and CV^2. This can be observed in the following figure:

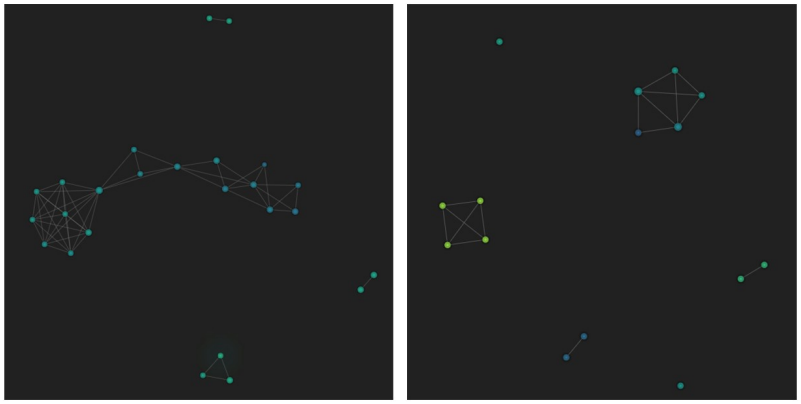


Fig. 1: Beta coefficient, ROI and CV^2 mapper (left) and ACF mapper (right)

Each cluster contains a certain number of time series that were grouped based on their similarity, depending on the projection method selected. As it can be observed in the following figure:

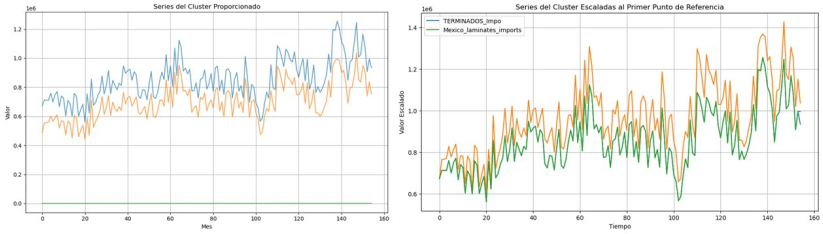
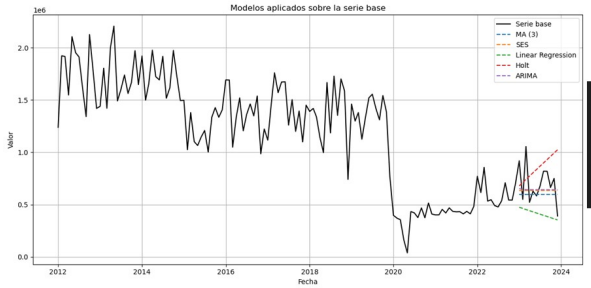


Fig. 2: A cluster with ACF projection

The figure 2 portrays one cluster from the ACF projection where the series finished product imports (blue), steel plans imports (orange) and Mexican sheets imports (green) can be seen. The left image contains the original time series, while the right image contains the same series, but escalated so they all have the same starting point. In the right image, it is easily observable that all three time series have an extremely similar behavior, which means that the mapper worked as intended.

After visually analyzing each cluster, a series was chosen to be the base of the cluster. This means that this time series will be the one where the models will be created so they can later be escalated to predict the remaining members of the cluster. An example of this can be seen in the following figure:



	Model	SMAPE (%)	MAPE (%)
0	ARIMA	20.993801	20.917462
1	SES	21.346911	20.929354
2	MA(3)	23.016402	21.479198
3	Holt	32.551428	41.351958
4	Linear Regression	47.813988	37.093457

Fig. 3: Different models in the series mineral coal ANC

The figure 3 shows the five different models tested on the base time series mineral coal ANC for a cluster, where it can be seen that the model that best predicts the data is ARIMA with a 20.91% MAPE. After this, all the models were adjusted by a factor so they could be used to predict the other models from the cluster. Resulting in the following:

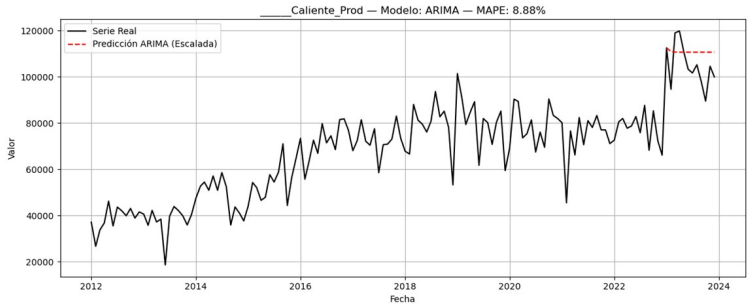
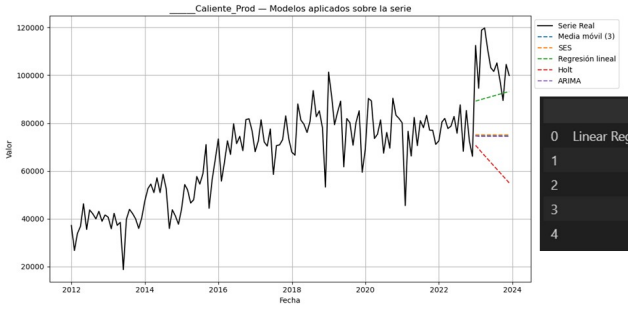


Fig. 4: Scaled model for a different series in the cluster

The figure 4 shows that by adjusting the models from the series mineral coal ANC, the one that best describes the new series is ARIMA with an 8.88% MAPE. The next thing needed to be done is to determine if this adjusted model is better than making a model exclusively for this time series.



	Model	SMAPE (%)	MAPE (%)
0	Linear Regression	14.100264	12.904714
1	SES	32.792507	27.958515
2	Ma (3)	33.352188	28.372579
3	ARIMA	33.532620	28.506599
4	Holt	50.022023	39.860215

Fig. 5: Models created specifically for the time series

The figure 5 shows how the best model has a 12.90% MAPE, which means that the adjusted model fits the data better than the models exclusively created for the time series, which demonstrates how using TDA techniques ended up yielding better results than not using them at all, not only because the model created describes the data better, but because it is way more efficient, since only one model is needed, which then only needs to be scaled to the other time series in the same cluster.

Lastly, once a good-enough model is created, it is important to know if the time series has a cyclical behavior; this would indicate that the model created could still be used in the long term. Following the example, the previous series Takens Embedding is the following:

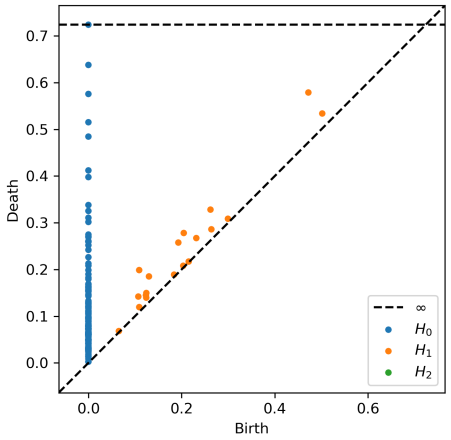


Fig. 6: Takens Embedding output for the time series

This Takens Embedding shows that the time series is not likely to be cyclical, which means that the model would need to be updated every once in a while.

The table 1 summarizes the whole project with those clusters and models for each method with the best results.

6 Discussion and Conclusions

The case study brought above were some of the most promising results using the previously detailed process, generating insights about seemingly unrelated phenomena. It has shown in some results outside the case study that making an adjusted model based on the intrinsic topological properties of the time series can improve the metrics on the models with the case study stated in the results 5, however, such results should be carefully studied as this proposed methodology does not guarantee an improvement on the metrics of the models as in some cases the traditional predictive algorithms may outperform a model based on sub-optimal projections to the time series.

The presented paper presents a methodology to cluster time series through the Mapper projection technique and different transformation functions for reducing the time series dimensionality, and then gives insight into choosing the best predictive models for each time series using simple techniques such as Linear Regression, Holt and Simple Exponential Smoothing as well as more complex predictive algorithms specially designed for time series forecasting such as ARIMA. Furthermore, the methodology sheds light in the medium to long term viability of the model based on the Takens Embedding theorem, thus, indicating the possible robustness of the model implementation as an decision making tool that helps understand the behavior of complex system based on the topological properties of influential phenomena that may influence such system in ways that are hard to comprehend.

Supplementary information. GitHub repository where you can find the code for each algorithm implemented. [redacted]

References

[1] INEGI: Perfil de la Industria del Hierro Y del Acero en México, 2016 edn. (2016)

[2] Strasser, S., Tripathi, S.: Forecasting Steel Demand: Comparative Analysis of Predictability across diverse Countries and Regions. Procedia Computer Science **232**, 2740–2750 (2024) https://doi.org/10.1016/j.procs.2024.02.091

[3] Basu, D., Dlotko, P.: The Four Seasons of Commodity Futures: Insights from Topological Data Analysis. SSRN Electronic Journal (2019) https://doi.org/10.2139/ssrn.3506780

[4] Bessembinder, H.: Systematic Risk, Hedging Pressure, and Risk Premiums in Futures Markets. Review of Financial Studies **5**(4), 637–667 (1992) https://doi.org/10.1093/rfs/5.4.637

[5] Uray, M., Giunti, B., Kerber, M., Huber, S.: Topological Data Analysis in smart manufacturing: State of the art and future directions. Journal of Manufacturing Systems **76**, 75–91 (2024) https://doi.org/10.1016/j.jmsy.2024.07.006

[6] Chen, Y., Volić, I.: Topological data analysis model for the spread of the coronavirus. PLoS ONE **16**(8), 0255584 (2021) https://doi.org/10.1371/journal.pone.0255584

[7] Dlotko, P., Rudkin, S.: Visualising the Evolution of English Covid-19 Cases with Topological Data Analysis Ball Mapper. arXiv (Cornell University) (2020) https://doi.org/10.48550/arxiv.2004.03282

[8] Goel, A., Pasricha, P., Mehra, A.: Topological data analysis in investment decisions. Expert Systems with Applications **147**, 113222 (2020) https://doi.org/10.1016/j.eswa.2020.113222

Mapper Function	Takens Findings	Agglomerated time series	Multiplicative Factor for Model Forecast	Best Prediction Model	MAPE
ACF	Highly unlikely cyclicity	**Cast iron ANC**	0.38	Linear regression	10.35%
		Oxygen Steel Furnace Gas Production	0.48	3-point moving average	8.71%
		Cast iron production	0.37	Linear regression	5.94%
ACF	High likelihood of weak cyclicity	**Semi-finished Products ANC**	1.13	ARIMA	13.87%
		Galvanized Metal Sheets Exports	0.05	ARIMA	8.15%
		Semi-finished products Imports	0.79	ARIMA	19.17%
		Steel planks imports	0.77	ARIMA	17.71%
Beta	High likelihood of weak cyclicity	**Flats 2 ANC**	1.01	Linear regression	6.31%
		Hot Roll ANC	0.28	3-point moving average	10.71%
		Coated sheet ANC	0.27	Linear regression	11.07%
		Galvanized sheet ANC	0.25	Linear regression	6.24%
		Flats 2 imports	0.49	Linear regression	16.36%
		Hot Roll Production	0.24	ARIMA	19.19%
Beta	High likelihood of weak cyclicity	**Tinned sheet ANC**	0.79	ARIMA	38.74%
		Seamless pipes export	4.53	Holt	12.47%
		Semi-finished product imports	42.84	Holt	12.8%
		Plank imports	41.31	Holt	13.22%
		Coated sheet imports	9.13	ARIMA	18.7%
		Hot Galvanized sheet imports	6.89	ARIMA	13.04%
CV^2	High likelihood of weak cyclicity	**Cold rolling import**	0.66	Holt	26.84%
		Coated sheet import	1.07	Holt	14.5%
		Galvanized sheet import	0.81	Holt	9.91%
ROI, Beta and CV^2	High likelihood of weak cyclicity	**Wire Production**	1.00	ARIMA	6.36%
		Wire ANC	1.02	3-point moving average	7.04%
		Wire Rod Production	0.81	ARIMA	6.36%
		Wire and Derivates Production	1.00	ARIMA	6.36%
ROI, Beta and CV^2	High likelihood of weak cyclicity	**Mexico Laminates Imports**	1.00	Holt	14.13%
		Finished Steel Goods	0.0010	Holt	13.64%
		Semi-finished product imports	0.0013	Holt	14.44%

Table 1: Cluster, Takens Embeddings, and models results