Forecasting losses in food commodities using XGBoost and Bayesian Structural Time Series Models

Ankush Raut
University of Colorado Boulder
ankush.raut@colorado.edu

Tapas Das
University of Colorado Boulder
tapas.das@colorado.edu

Abhishek Mukundan
University of Colorado Boulder
abhishek.mukundan@colorado.edu

Abstract

The supply chain for food commodities demonstrates considerable efficiency across almost all geographies today; nevertheless, it continues to experience losses at each stage. This study delves into historical data to scrutinize losses within the supply chain of diverse food commodities across various countries. Initially, we analyze the extent to which losses across different stages are reflected in the historical data, examine the extent of losses across various stages, and elucidate their causes. We categorize these losses within specific commodities and countries based on distinct supply chain stages, creating separate time series. Additionally, we formulate hypotheses regarding observed trends and patterns. Ultimately, we present two forecasting frameworks—XGBoost and Bayesian Structural Time Series (BSTS) modeling—to model the trends in losses incurred in these commodities across different countries and supply chain stages. Our conclusion highlights that when year-on-year losses exhibit volatility, XGBoost outperforms BSTS and vice versa.

1. Introduction

The losses incurred within the food commodity supply chain are significant, necessitating a comprehensive investigation into their underlying trends and causes. It's evident that the prevalence of hunger in various regions worldwide emphasizes the importance of systematically analyzing food commodity losses as a crucial step toward alleviating global hunger. By leveraging historical trends in food commodity losses, we can establish a forecasting framework to estimate potential losses at different stages of the supply chain in the future. This analysis will form the basis for planning responsive actions to mitigate these losses effectively.

The specific causes of food losses vary worldwide, contingent upon unique conditions and local situations in each country. Generally, food losses are influenced by crop production choices and patterns, internal infrastructure and capacity, marketing chains, distribution channels, and consumer purchasing and food use practices. Regardless of a country's economic development level, efforts should aim to minimize food losses. Economically avoidable food losses directly and negatively impact both farmers' and consumers' incomes. Enhancing the efficiency of the food supply chain can reduce the cost of food for consumers, thus increasing accessibility. Given the magnitude of food losses, investing in reducing losses could be profitable, provided that financial gains outweigh the costs.

This study utilizes the Food Losses and Waste database maintained by the Food and Agriculture Organization of the United Nations, containing a historical log of food loss percentages since 1965 across various stages of food supply chains in different countries for a wide range of commodities. A dataset is sampled from this database, including only reliable estimates. Following cleaning and preprocessing, the study examines the nature of historical food loss wastages across different countries and commodities. Based on the observed characteristics of the data, two frameworks, XGBoost and BSTS, are utilized to forecast future losses in each time series. These forecasts, along with insights from the analysis, can inform strategies for planning loss minimization and optimizing supply chains.

2. Data

2.1 Data collection

The data was obtained from the Food Losses and Waste database maintained by the Food and Agricultural Organization of the United Nations. It contains historical data on loss percentages and loss amounts corresponding to different commodities across various countries and stages of the food supply chain. The data in this table was originally collected using different methods: expert opinion, case studies, controlled experiments, survey-based estimates, census data, and literature review. We initially excluded the data gathered from modeled estimates from our

estimation and utilized only the data obtained from deterministic methods. Due to the lack of information on the website regarding the specific quantitative characteristics of each method, we assume that all deterministic methods provide population-representative estimates of the loss percentage. The data encompasses multiple stages of the food supply but predominantly pertains to the whole supply chain, as depicted in Figure 1.

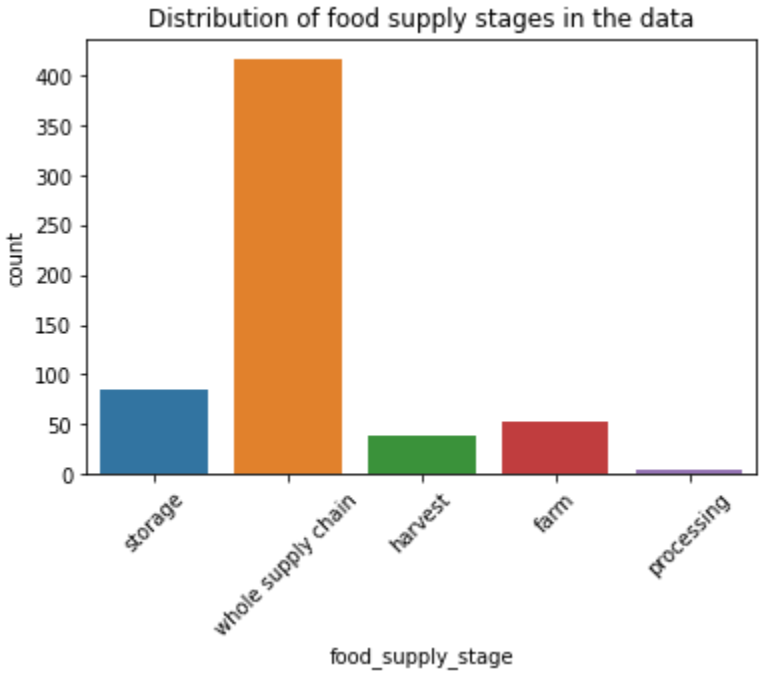

Figure 1: Distribution of instances of loss across different stages of the food supply chain in the cleaned dataset.

2.2 Main sources of bias

This dataset exhibits considerable inconsistency, largely attributed to the disparities in recording food losses, thereby highlighting the absence of adequate measures in supply chains to mitigate these losses. Consequently, the data harbors several potential sources of bias. Geographically, the dataset lacks consistency, resulting in some countries being underrepresented while others are overrepresented, complicating the comparative analysis of trends among nations. Moreover, inconsistencies across various commodities further hinder relative trend comparisons. Similarly, the irregular distribution of data across different stages of the supply chain poses challenges in drawing meaningful comparisons among them.

Biases may also stem from the methodologies employed in data collection. If certain methods favored specific regions or stages within the supply chain, they could introduce bias. Temporal biases are evident in the dataset, manifested in the possibility of variability of data collection

methods or changes in reporting standards over time, implicitly affecting the interpretation of observed trends.

2.3 Characteristics of the data

Variable	Definition
Country	The country in which the loss is measured
Commodity	The losing food commodity
Year	Timestep of the record
Loss percentage	Percentage of loss in the commodity for the given setting
Loss amount	Amount of loss in the commodity for the given setting, unit varies significantly
Food supply stage	The stage of the supply chain from which the loss was recorded
Activity	Additional information regarding the stage of the supply chain from which the loss was recorded
Cause of loss	Information regarding the potential causes for the given loss
Method data collection	The method that was used to estimate the loss

Table 1: Variables in the dataset.

3. Methodology

3.1 Exploratory data analysis

As previously discussed, the dataset exhibits uneven representation across various countries, commodities, and years. The variable indicating loss amount suffers from inconsistency in units, encompassing measurements like kg, tons, metric tons, etc., in an unstandardized format. This makes the column challenging to utilize effectively. Consequently, we will utilize the loss percentage for our estimations. However, since percentage is a relative quantity, it cannot be aggregated across different countries, commodities, or food supply stages within a given year.

To address this, we will index the data based on country, commodity, and food supply stage, allowing us to observe losses at a combined level using these identifiers, combining to form a unique key. To ensure consistency, we calculated the mean of those values when multiple values

exist for loss percentages corresponding to a key in a given year due to different data collection methods. This approach ensures that only one value corresponds to each key for every year in the dataset, aligning with our assumptions regarding the nature of data collection methods.

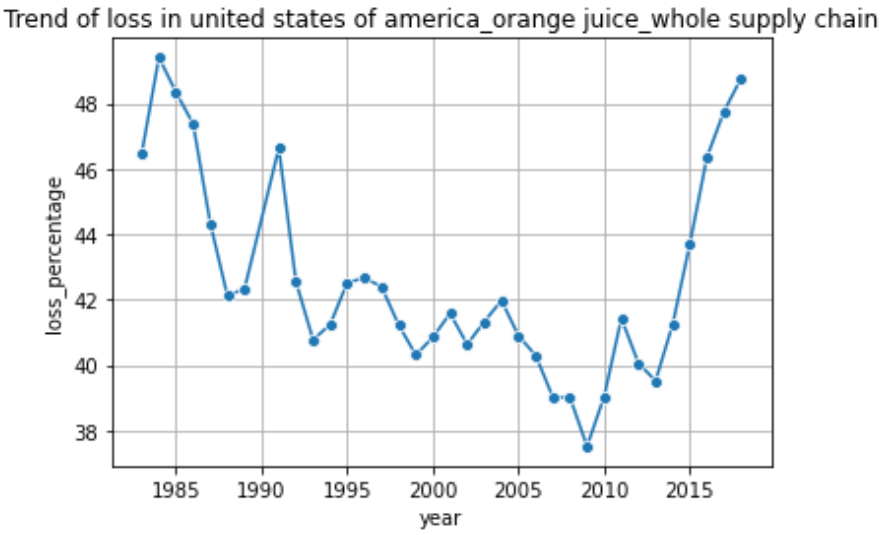

Figure 2: Historical loss percentage in the whole supply chain of orange juice in the US.

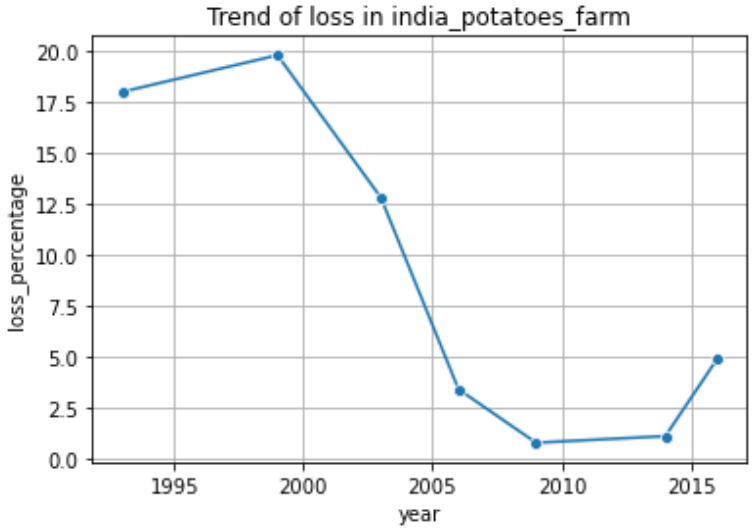

Figure 3: Historical loss percentage in the farm stage of the supply chain of potatoes in India.

To create a reliable model of the loss percentage, we exclusively utilize time series keys with data spanning over 5 or more years. Additionally, upon observing certain keys with zero variance, we opted to exclude them from the analysis due to potential errors in their records within the database.

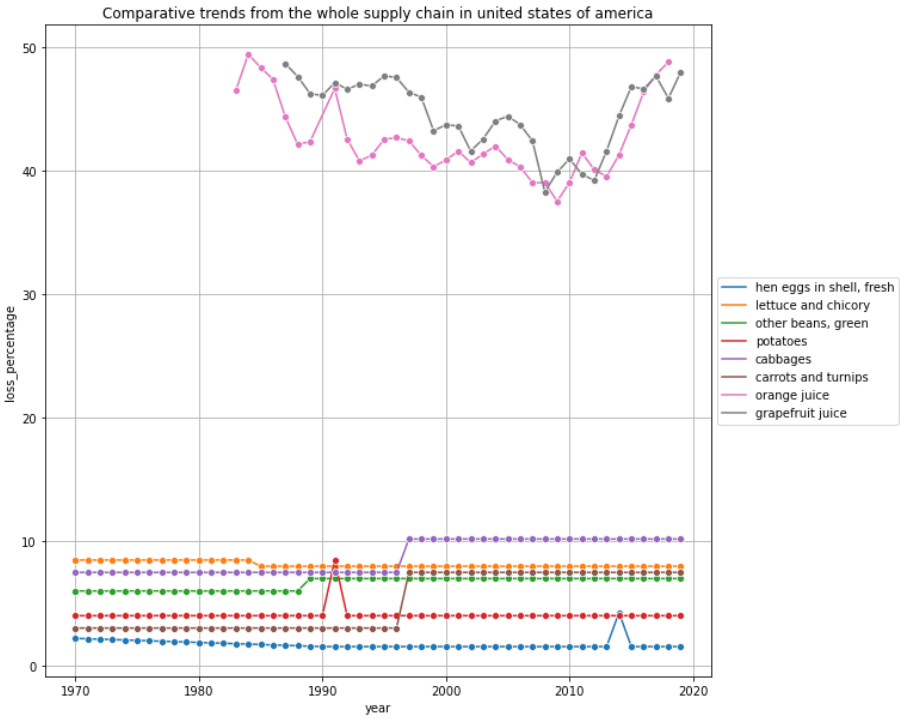

Figure 4: Comparing the historical loss percentages in the entire supply chain of various commodities in the US. It's evident that orange juice and grapefruit juice experience significantly higher losses than other commodities. The recorded loss percentages for the remaining commodities show minimal year-on-year changes.

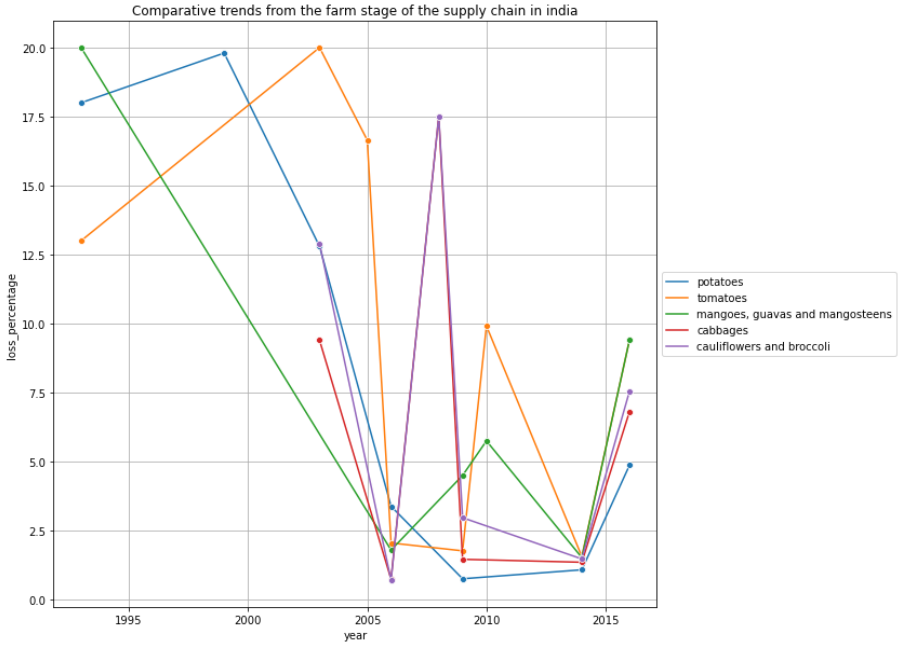

Figure 5: Comparing the historical loss percentages in the farm stage of the supply chain of various commodities in India. Observable declining trends are present in the loss percentage for every commodity. More recently, mangoes, guavas, and mangosteens have experienced the highest loss percentage.

Word-clouds were created on the causes of loss for different commodities. These highlight the major issues with the supply chain of different commodities.

Figure 6: Word-cloud for causes of loss in the supply chain of wheat across the world. This word cloud implies, albeit with low confidence, that losses in wheat usually occur during the harvesting and threshing phases, and when it's in the grain stage rather than the flour stage.

3.2 Hypothesis testing

In this section, we conduct a statistical analysis of the characteristics present in each time series within the dataset. Grouping the dataset based on unique combinations of country, commodity, and food supply chain stages, resulted in 44 distinctive time series keys. These keys exhibit irregular time steps and lengths spanning from 5 to 50. To effectively model and assess the properties of these time series, we employed the Augmented Dickey-Fuller test and Mann-Kendall test. These tests aim to ascertain two significant characteristics within the time series data: stationarity and trend, respectively. A stationary time series implies that the mean, variance, and autocorrelation structure remain constant over time. Essentially, such a time series displays unwavering patterns and trends throughout its observations. Stationary data is more amenable to modeling and forecasting due to its stable statistical properties.

Augmented Dicky Fuller Test

The unit root test is a statistical method used to determine if a time series is stationary or not. Stationarity refers to the property of a time series where the statistical properties, such as mean, variance, and autocorrelation, remain constant over time.

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon$$

Figure 6: Equation representing unit root, where Y(t-1) is lag1 of the time series and ΔY(t-1) first difference of the series at the time (t-1).

If the value of α is 1, then the time series is considered non-stationary. This is because a value of 1 indicates perfect correlation between the current and previous values of the series, extending infinitely backward. Consequently, it becomes impossible to differentiate between trends and random fluctuations in the data. The Dickey-Fuller (DF) test is a statistical method commonly used to examine the presence of a unit root in a time series dataset. The null hypothesis of this test suggests the existence of a unit root in the time series, signifying non-stationarity and the presence of a trend.

$$Y_t = c + \beta t + \alpha Y_{t-1} + \phi \Delta Y_{t-1} + \epsilon_t$$

Figure 7: Equation for Dicky Fuller test, where Y(t-1) is lag1 of the time series and ΔY(t-1) first difference of the series at the time (t-1).

{*Ho*: The null hypothesis of the Dickey-Fuller test is that α=1, which implies that there is a unit root in the time series}
{*Ha*: The alternative hypothesis is that α<1, which implies that the time series is stationary and does not have a unit root}

The Augmented Dickey-Fuller (ADF) test expands upon the DF test by accommodating higher-order autoregressive processes and other influencing variables within the time series. The DF test operates through a regression analysis involving the initial difference of the time series against its lagged values. The resulting test statistic is then compared against critical values from a table to establish statistical significance. In contrast, the ADF test enhances the DF regression equation by incorporating additional lagged terms related to the first difference of the time series. This adjustment accommodates higher-order autoregressive processes, rendering the ADF test more robust in determining the stationarity of time series data.

Mann-Kendall Test

The purpose of the Mann-Kendall (MK) test is to statistically assess if there is a monotonic upward or downward trend of the variable of interest over time. A monotonic trend means that the variable consistently increases or decreases through time, but the trend may or may not be linear. There is no requirement that the measurements be normally distributed or that the trend, if present, is linear. The MK test can be computed if there are missing values. The assumption of independence requires that the time between samples be sufficiently large so that there is no correlation between measurements collected at different times. The MK test tests whether to reject the null hypothesis Ho and accept the alternative hypothesis Ha, where :

{*Ho*: No monotonic trend in the observed time series}
{*Ha*: Monotonic trend is present in the observed time series}

The initial assumption of the MK test is that this is true and that the data must be convinced beyond a reasonable doubt before it is rejected and accepted. Suppose there are missing data in the time series. For example, suppose that data are collected on the first day of each month, but the data for March 1st and July 1st have been lost. In that case, the MK test is conducted in the usual way using the smaller data set, reducing the value of n as appropriate.

Observations

The results of the ADF test indicate that a significant proportion—33 out of 44—of the time series in the dataset display non-stationarity. Similarly, interpretation of the MK test results revealed that almost 37 out of 44 series lack a monotonic trend. These individual time series results are summarized in Table 2. Additionally, we categorized the available time series based on their lengths and analyzed how the length affected stationarity and trend. An interesting observation emerged: time series with an average length of less than 15 did not exhibit any significant trend. Conversely, those with greater lengths showed a clear monotonic trend, either increasing or decreasing.

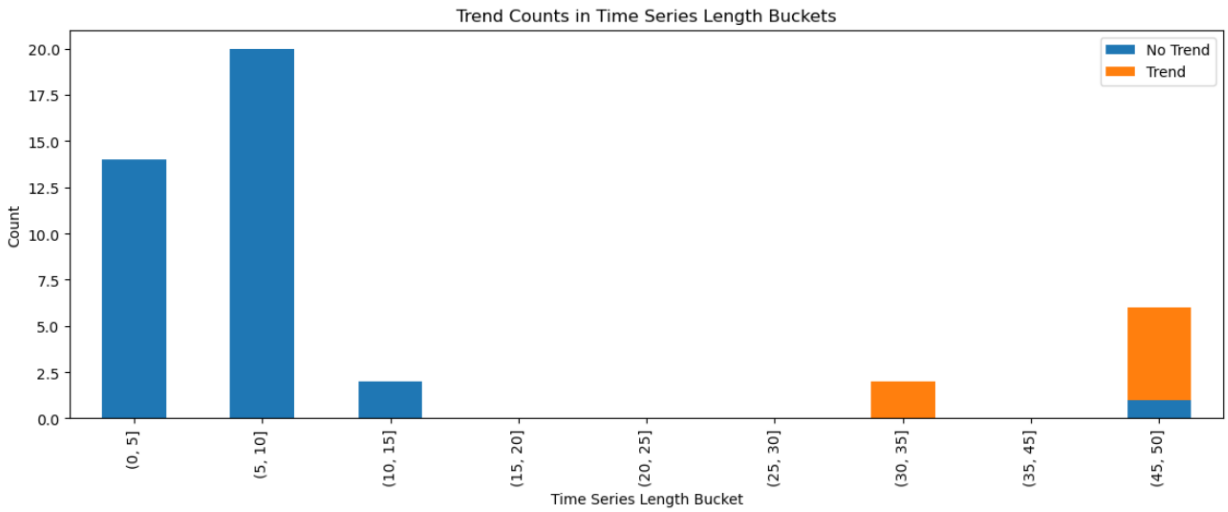

Fig 9: Trend vs no-trend counts in different buckets based on time series length.

There were no identifiable patterns in stationarity across different series lengths. Stationarity seems to be evenly diffused among the buckets, suggesting non-stationarity due to exogenous variables impacting the time series or the presence of non-monotonic trends.

These observations motivated us to adopt a 2-fold approach to modeling and compare various methods. The XGBoost approach addresses non-linear and inter-series relationships, especially in those time series lacking systematic trends or stationarity. The Bayesian Structural Time Series model caters to those series that exhibit a systematic trend and stationarity, but has been implemented on every time series in the dataset.

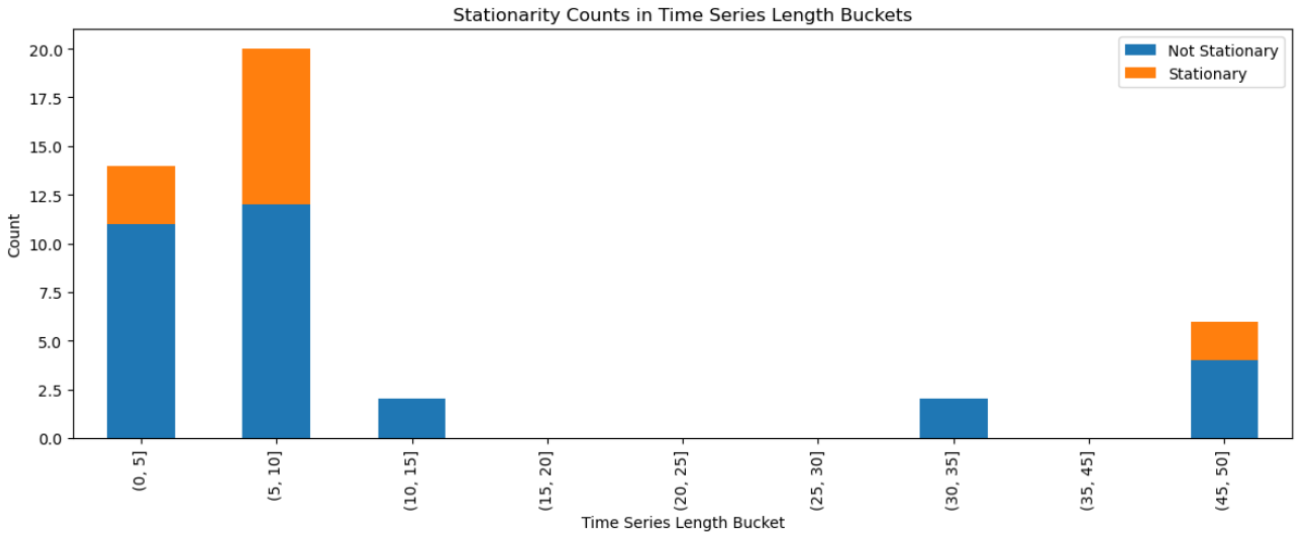

Fig 8: Stationarity vs non-stationarity counts in different buckets based on time series length.

Keys (country - commodity - food supply stage)	Stationarity	Trend
United States of America - lettuce and chicory - whole supply chain	No	Yes
United States of America - beans, green - whole supply chain	No	Yes
United States of America - cabbages - whole supply chain	No	Yes
United States of America - carrots and turnips - whole supply chain	No	Yes
United States of America - orange juice - whole supply chain	No	Yes
United States of America - grapefruit juice - whole supply chain	No	Yes
Kenya - maize(Corn) - storage	Yes	No
United States of America - hen eggs in shell - whole supply chain	Yes	Yes
United States of America - potatoes - whole supply chain	Yes	No
India - wheat - storage	Yes	No
Pakistan - wheat - farm	Yes	No
Indonesia - rice - processing	Yes	No
India - potatoes - farm	Yes	No
India - mangoes, guavas and mangosteens - farm	Yes	No
India - cabbages - farm	Yes	No
India - cauliflowers and broccoli - farm	Yes	No
India - soya beans - harvest	Yes	No
Azerbaijan - oats - whole supply chain	Yes	No
Azerbaijan - grapes - whole supply chain	Yes	No

Table 2: Time series with either a significant trend or observed stationarity.

3.3 Feature engineering

From the tests above, it is evident that many of the time series in the dataset do not exhibit stationarity or trend. While a structured time series model can work for the time series exhibiting stationarity and trend, a complex non-linear model will be required to model the loss percentage in other time series. For this, we begin with engineering features for the model. Table 3 explains these features.

Engineering strategy	Definition
One-hot-encoding country name, commodity name, and food supply stage	Using these one-hot-encoded vectors the model will be able to discern data from each time series. It will also be able to recognize overlapping patterns, if any.
5 lag features for loss percentage (previous value till previous 5th value)	Distinct lag features for each time series, containing previous values of that same time series. This will help the model in creating appropriate autocorrelation estimates. Here we return null if some value isn't present.
4 weighted moving average features for loss percentage (aggregating previous 2 values till previous 5 values)	Distinct weighted moving average for each time series using previous values of that same time series. This will add a moving average component to the model.
Year	Help the model understand different timestamps.

Table 3: Engineered features and their descriptions

3.4 XGBoost framework

This is a standard supervised learning framework using the XGBoost model, built on the Gradient Boosting Trees method. XGBoost excels in capturing complex, non-linear relationships within data due to its ensemble learning technique, combining multiple decision trees to model intricate patterns often missed by simpler models. Beyond its ability to handle non-linearity, XGBoost boasts advantages including regularization to prevent overfitting, inherent handling of missing values, and parallel processing for faster training. While its power is evident, factors like hyperparameter tuning and dataset nuances can influence its efficacy.

We utilize this model to predict the percentage of loss for each commodity across various countries in different years using the engineered features. We split the data sequentially into a training, validation, and test set. Specifically, the training set encompasses data until 2011, the validation set spans from 2012 to 2016, and the hold-out set progresses from 2017 until 2022. The validation set serves the purpose of stopping model iterations prematurely in case of minimal change in mean squared error with each new iteration (tree creation). The hold-out set remains concealed from the model. The objective is to evaluate the model's ability to accurately predict loss percentages in recent years using historical patterns.

3.4 Bayesian Structural Time Series (BSTS) framework

BSTS modeling merges Bayesian principles with structural modeling to dissect time series data, enabling adept handling of intricate and evolving temporal patterns. This method encompasses various components such as local level, linear trends, seasonality, regression effects, and custom elements, offering adaptability to diverse data structures. The parameters for these components are estimated using Bayesian inference. To estimate these parameters, Variational Inference (VI) is employed, casting approximate Bayesian inference as an optimization problem. VI seeks a 'surrogate' posterior distribution that minimizes the evidence lower bound (ELBO) loss, similar to KL divergence minimization with the true posterior. We utilize the Tensorflow Probability (TFP) library for defining the structural time series components and conducting VI. TFP facilitates gradient-based optimization as well as Hamiltonian Monte Carlo sampling-based optimization for VI. Figure 10 illustrates gradient-based VI.

Under this framework, we use the local linear trend and the autoregressive integrated moving average (ARIMA) components to create the structural time series model. Given its speed and efficiency, we opt for the gradient-based VI to estimate Bayesian inference parameters for these components. VI can be used to estimate Bayesian credible intervals for parameters of any regression model to estimate the effects of various treatments or observed features on an outcome of interest. Credible intervals bound the values of an unobserved parameter with a certain probability, according to the posterior distribution of the parameter conditioned on observed data and given an assumption on the parameter's prior distribution.

We define the model using the aforementioned components and employ gradient-based VI to estimate parameters for each time series individually. Utilizing the historical time series data, we forecast the loss percentage for the 3 most recent timesteps in every time series.

Algorithm 1: Coordinate ascent variational inference (CAVI)

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$

Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

Initialize: Variational factors $q_j(z_j)$

while *the* ELBO *has not converged* **do**

  for $j \in \{1, \ldots, m\}$ **do**

    Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$

  end

  Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$

end

return $q(\mathbf{z})$

Figure 10: The variational inference algorithm.
Credits: https://leimao.github.io/article/Introduction-to-Variational-Inference/

4. Results

4.1 XGBoost results

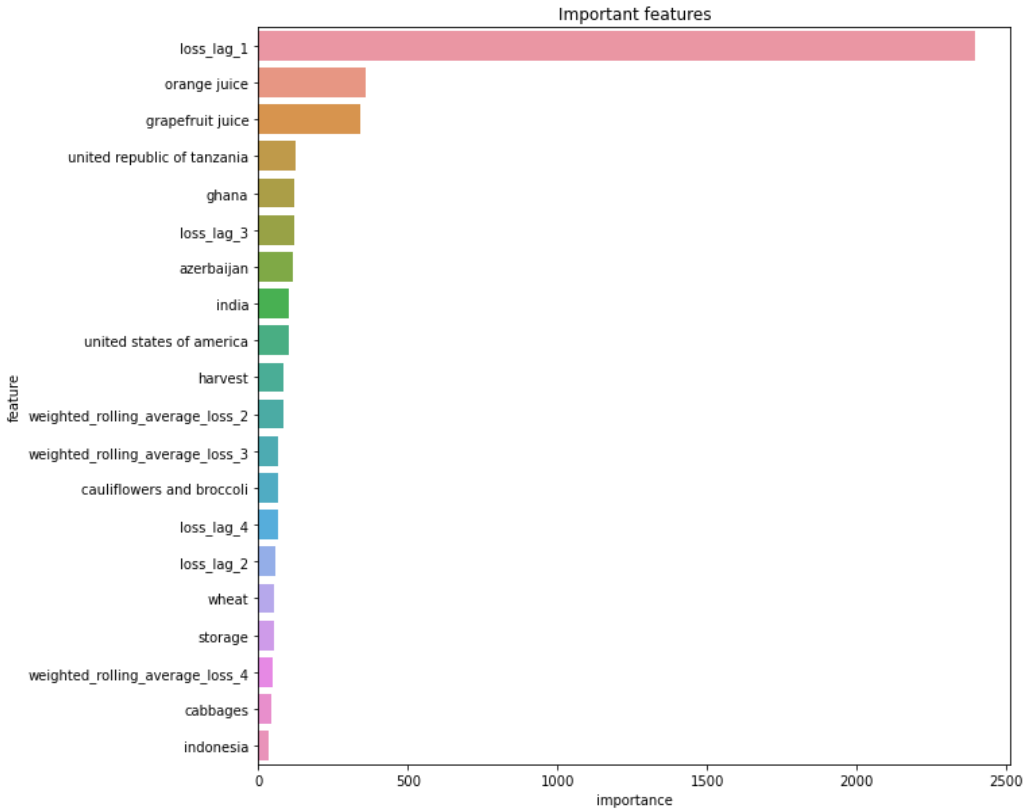

Figure 11: Features with highest gain values in the XGBoost model.

Data split	Root mean squared error (RMSE)	R-squared	Mean absolute percentage error (MAPE)
Train	1.95	0.97	18.59
Validation	7.54	0.33	52.49
Test	5.61	0.66	41.01

Table 4: Results from the XGBoost modeling framework.

It can be observed from the results as depicted in Figure 11 and Table 4 that the most crucial predictor variable for future loss percentage in the XGBoost modeling framework is the most recent loss percentage. The metrics worsen in both the validation and test (hold-out) sets, albeit to an acceptable degree. Figure 12, 13, and 14 illustrate the prediction behavior of the XGBoost model.

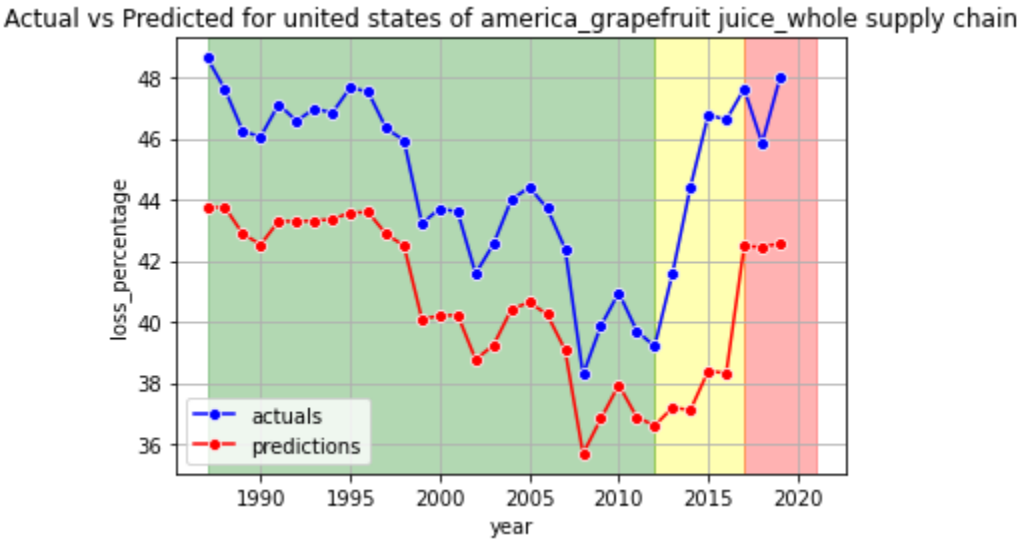

Figure 12: Actual vs XGBoost predicted loss percentage comparison for US - grapefruit juice - whole supply chain. The green region corresponds to the training set, yellow to the validation, and red to the hold-out. The model is capable of following the trend but consistently underpredicts for every year, particularly in the forecast horizon (hold-out set).

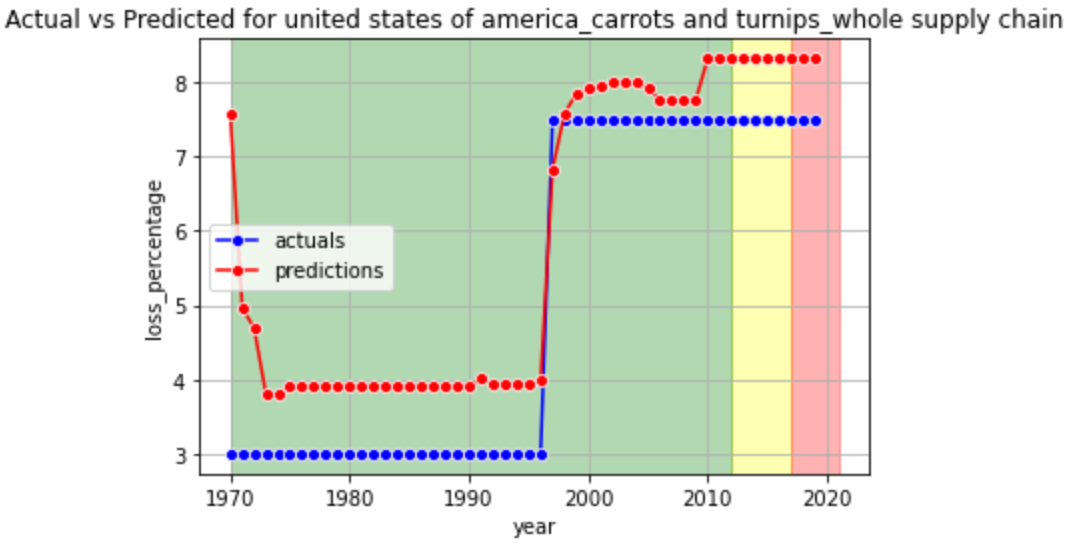

Figure 13 Actual vs XGBoost predicted loss percentage for US - carrots and turnips - whole supply chain. Despite the relatively low variance in the loss percentage, the model is unable to accurately predict.

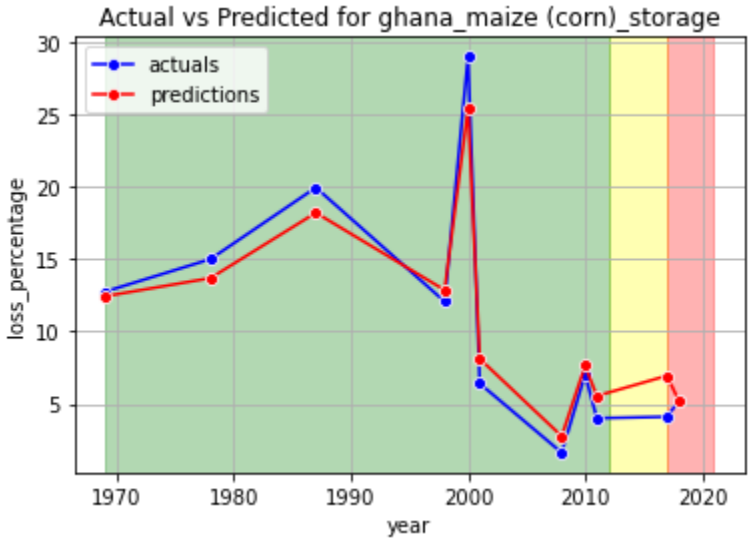

Figure 14: Actual vs XGBoost predicted loss percentage comparison for Ghana - maize (corn) - storage. The model does a relatively good job at predicting the loss percentage in every split. It's near perfectly following the trend.

4.2 BSTS results

Time series length	RMSE	R-squared	MAPE
Long (over 10 timesteps)	11.33	0.66	94.86
Short (less than 10 timesteps)	12.52	0.17	171.81

Table 5: Results from the BSTS modeling framework.

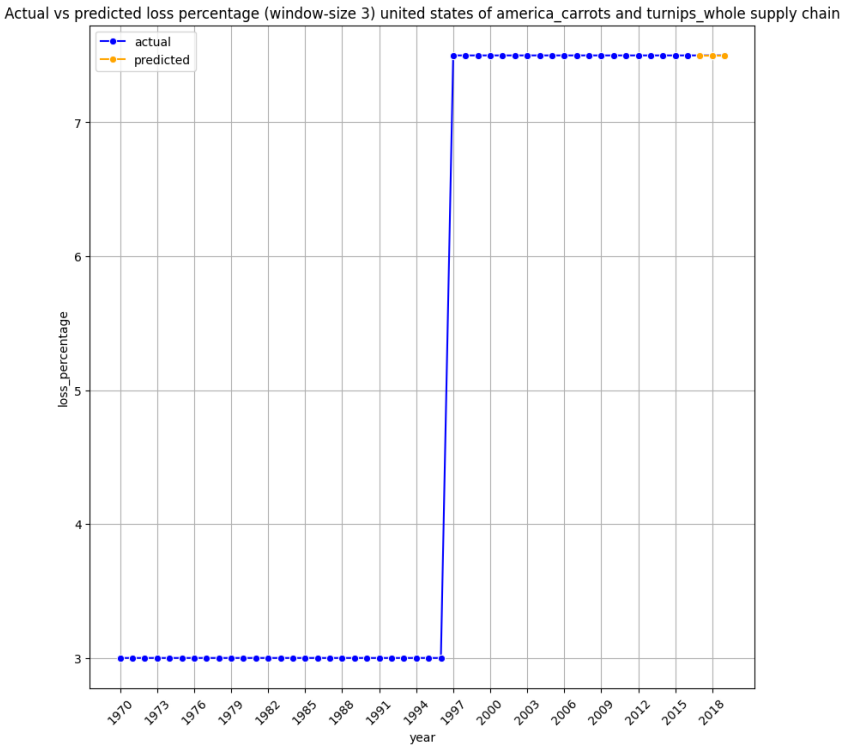

Figure 15: Actual vs BSTS predicted loss percentage comparison for US - carrots and turnips - whole supply chain. The model is able to accurately forecast the loss percentage in the last 3 timesteps using historical data.

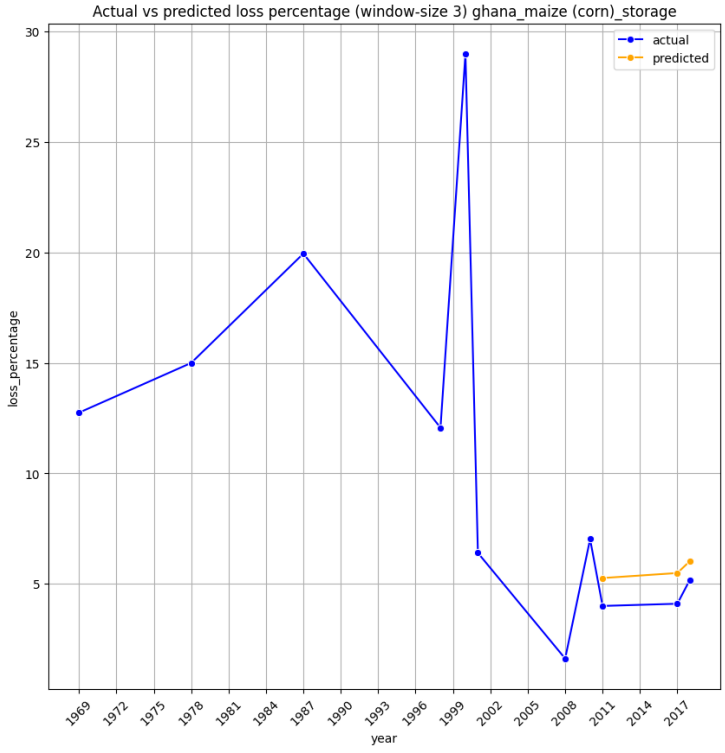

Figure 16: Actual vs BSTS predicted loss percentage comparison for Ghana - maize (corn) - storage. The last 3 forecasted values are quite close to the actuals, similar to what was observed in XGBoost results.

Despite the poor overall performance of BSTS, as indicated by the metrics in Table 5, it performs quite well in forecasting the most recent 3 loss percentage values using a small time series.

5. Conclusion

The XGBoost framework outperforms BSTS significantly in forecasting loss percentages, especially in scenarios with volatile historical trends. However, it isn't very accurate when dealing with time series characterized by low variability. Although its performance dips during the validation and hold-out periods, as indicated by the metrics in Table 4, the decline remains acceptable. Overall, XGBoost demonstrates proficiency in extrapolating from historical time series data and integrating behaviors observed in similar time series while forecasting.

The most important feature is the most recent loss percentage. Due to the overrepresentation of orange juice and grapefruit juice commodities, their encoded vectors exhibit high gain values; however, this may change with the inclusion of additional data. The encoded vectors representing various countries emerge as crucial features, underscoring distinct patterns in food commodity loss percentages across these nations.

The results from forecasting using the BSTS framework clearly indicate a lack of sufficient data. Its performance significantly declines when working with short time series compared to long time series, as demonstrated in Table 5. However, BSTS can accurately follow the trend in some time series, particularly those with low variability. In certain cases, as shown in Figure 16, BSTS doesn't lag far behind XGBoost in accurately forecasting the loss percentage. It's important to note that while XGBoost uses all past data for forecasting the current timestep, BSTS employs all past data solely for the first forecast in its 3-timestep window. For the remaining 2 timesteps, it relies on the previous timestep's data along with the estimate for the previous timesteps in the forecasting window.

From this analysis, it can be concluded that food loss has been increasing recently and can be estimated fairly accurately using XGBoost. BSTS also offers a good framework for this estimation, although it requires more data to generate reliable estimates for volatile time series. This study emphasizes the absence of a large and reliable data source for estimating losses in food commodities, highlighting the lack of established frameworks for monitoring and mitigating food losses. The future scope of this study would involve intensive data collection measures and case studies to form stronger hypotheses regarding trends in the loss of food commodities.

6. References

1. Food and Agricultural Organization of the United Nations database. https://www.fao.org/platform-food-loss-waste/flw-data/user-guide/en

2. Dickey, D. A., “Distribution of the Estimators for Autoregressive Time Series With a Unit Root”, Journal of the American Statistical Association (1979).

3. Kendall, M.G., “A New Measure of Rank Correlation”, Biometrika, Volume 30, Issue 1-2, June 193.

4. Chen, Tianqui; Guestrin, Carlos, “XGBoost: A Scalable Tree Boosting System”, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (August 2016).

5. Brodersen, Kay H.; Gallusser, Fabian; Koehler, Jim; Remy, Nicolas; Scott, Steven l., “Inferring Causal Impact using Bayesian Structural Time-Series Models”, The Annals of Applied Statistics (2015).

6. Blei, David M.; Kucukelbir, Alp; McAuliffe, Jon D., “Variational Inference: A Review for Statisticians”, Journal of the American Statistical Association (2017).