Implementing a Data Driven Model for Automated Dynamic Outlier Detection During Data Validation

39th Meeting of the Voorburg Group on Services Statistics

Copenhagen, Denmark

September 2025

Chelsea Velic

Katie Frawley

Sarah Eian

U.S. Bureau of Labor Statistics

Paper approved for release September 04, 2025

This paper provides a summary of research results. The information is being released for statistical purposes, to inform interested parties, and to encourage discussion of work in progress. The paper does not represent an existing, or a forthcoming new, official BLS statistical data product or production series.

Background

The United States Bureau of Labor Statistics (BLS) requests monthly price updates from Producer Price Index (PPI) survey respondents for sampled product and service transactions. The monthly price changes for these sampled transactions are used to calculate the industry and commodity producer price indexes. BLS economists perform quality checks on microdata as they are reported. One of the quality checks is on the magnitude of the price change. When the magnitude of the one-month percent change for a transaction exceeds a certain level, the price is flagged for review as an outlier that exceeds "tolerance". When a price change exceeds tolerance, a BLS economist is required to verify the price change, determine the cause of the change, and, if necessary, adjust for any changes in quality.

Tolerance levels are set within the PPI processing system. Typically, each industry has a positive tolerance threshold for price increases at or above a certain level and a negative tolerance threshold for price decreases at or below a certain level. While most tolerance levels are set at the 6-digit North American Industry Classification System (NAICS) industry level, the system also allows tolerances to be set for detailed lower-level indexes.

Historically, tolerance thresholds were set at the time a new industry sample was introduced. The positive and negative tolerance thresholds entered in the system are static values, meaning they remain the same every month unless a user changes the value. Recently, however, BLS has been researching the use of data driven models to determine the tolerance thresholds for each industry in the PPI on a month-to-month basis. The goal of this work is to create operational efficiencies in outlier detection and microdata review by using dynamic, data-driven method for setting price tolerance levels.

Approach

The first step in this project was to develop a model that detects individual monthly price changes that deviate significantly from the rest of the price changes reported for a given industry in each month. The model uses historical data from PPI survey respondents to predict an expected range for current one-month price changes and to identify outliers. Time series outlier detection methods considered were an Autoregressive Integrated Moving Average (ARIMA) model; a class of models designed to account for error, trend and seasonality components (ETS); machine learning; and deep learning.

The team examined ARIMA statistical models using existing price change data to assess whether a data driven model would be feasible for identifying items with price changes outside of the expected range. The ARIMA models are a class of statistical models for analyzing and forecasting time series data. This class of statistical models captures the standard temporal dependencies that are unique to time series data. Alternative modeling techniques were explored and compared against the ARIMA model to ensure that the best model was selected for use in production. Accuracy, generalizability, computational efficiency, and interpretability were assessed in evaluating which method would be the best option.

1. Accuracy: Which model has the best tendency to minimize review of prices changes that fall within expected ranges and maximize review of outliers?
2. Generalizability: Which model performs best across multiple datasets that present different price distributions?
3. Computational efficiency: Time/computer resources needed to make the model. Which model needs the least resources?
4. Interpretability: When multiple models perform comparably in terms of accuracy and generalizability, which are the most interpretable to understand the causal mechanism by which a price change is determined to be put up for review?

Deep learning models were ruled out for the initial approach because they are computationally expensive. Machine learning models are often harder to interpret, and it is harder to fine tune their parameters, so they were not the initial point of focus either. ETS models are often comparable to ARIMA models but have subtle differences in performance depending on the data distribution. In some test cases for seasonal data, a comparison of the two suggested that ARIMA models performed slightly better. Therefore, for computational efficiency, interpretability, and accuracy purposes, this project used ARIMA models.

Model

Auto-ARIMA was used to generate each ARIMA model for this project. Auto-ARIMA is a means of identifying the optimal parameters for a particular ARIMA model using stepwise parameter optimization such as Akaike Information Criterion (AIC), a measure of model quality that combines model fit (explanatory ability) and model complexity (number of parameters).

The ARIMA model is defined as:

Equation 1:

$$ARIMA(p, d, q)(P, D, Q)[m]$$

where the parameters are as follows:

p: The range of values that are allowed for autoregressive terms.

d: The range of values that are allowed for order difference.

q: The range of values that are allowed for moving average terms.

P: The range of values that are allowed for seasonal autoregressive terms.

D: The range of values that are allowed for seasonal order of differences.

Q: The range of values that are allowed for seasonal moving average terms.

m: The seasonality.

After investigating which criteria would be best for setting parameter constraints, the amount of historical price change data available per industry was found to be the criteria that strengthened the model output most. The historical data set consists of monthly price changes collected per industry.

Seasonal Model

To test this finding, BLS selected a group of industries with at least 24 months of price data, which is the minimum amount of time to detect annual seasonality, so seasonal parameter boundaries are applied. The boundaries used by Auto-ARIMA during ARIMA model selection for these industries are:

Equation 2:

$$ARIMA(start\_p = 0, max\_p = 23, d = None, max\_d = 1, start\_q = 1, max\_q = 23, max\_P = 2, D = 0, max\_Q = 1, m = 12, seasonal = TRUE, stepwise = TRUE$$

When selecting the best-fit parameters for the seasonal ARIMA model, the Auto-ARIMA uses 24 months of net price changes (autoregressive terms) and past forecast error values used to predict future errors (moving average terms). To account for seasonality, the Auto-ARIMA incorporates seasonal trends in the autoregressive terms from the two prior 12-month periods.

Non-Seasonal Model

The second group of industries contains less than 24 months of data. Seasonality was not considered for this group because there are not enough time points to evaluate if seasonality is occurring. The boundaries used by Auto-ARIMA during ARIMA model selection for these industries are:

Equation 3:

$$ARIMA(start\_p = 0, \max\_p = 11, d = None, \max\_d = 1, start\_q = 1, max\_q = 11, seasonal = FALSE, stepwise = TRUE$$

When selecting the best-fit parameters for the non-seasonal ARIMA model, the Auto-ARIMA is allowed to use up to 12 months of net price changes (autoregressive terms) and past forecast error values used to predict future errors (moving average terms).

Data Preparation

To prepare the data for the ARIMA models, additional processing was necessary. For the model to detect time series data, all price change data for each 6-digit NAICS industry had to be consolidated into a single price change value per month. To accomplish this, a single monthly value was calculated by taking the median of the top ten percent of all positive net price changes for sampled transactions in each industry and the bottom ten percent of all negative net price changes for sampled transactions in each industry. The positive aggregate is then used for the positive outlier detection model, whereas the negative aggregate is used for the negative outlier detection model.

An outlier is detected for an industry if a price change value is substantially different from the value predicted by the ARIMA model. The difference between a predicted price change and a reported price change is the model error. For each historical month, ranging from when the industry was first introduced to the month prior to the current month, the error was recorded. The complete dataset of errors for an industry was then used to calculate a median and some degree of deviation. For this project, a 2.5 deviation from the median is considered an outlier. Standard deviations are used if the distribution of errors is normal, but that is an assumption that may not always be true. To avoid the normal distribution assumption, a non-parametric method is used. The Median Absolute Deviation (or MAD) was used as a replacement in this study, so that the distribution of errors is not sensitive to a few existing outliers.

Calculating Tolerance Levels for Identifying Outliers

Once the ARIMA models were set up and the data properly formatted, tolerance levels were calculated for each 6-digit NAICS industry using the following steps:

1. Median of upper 10% of positive, or of lower 10% of negative, net price change is given to Auto-ARIMA.
2. If x>23 months of price data, a seasonal model is selected. If x<24, a non-seasonal model is selected.
3. Auto-ARIMA stepwise process generates parameter options using either seasonal or non-seasonal model.
4. Akaike Information Criterion (AIC) measures quality of parameter options and determines best-fit for the ARIMA model used to predict tolerance levels.
5. Starting from the second time point in the data, predictions are made for each following month by giving all data prior to that month to the selected model. The difference between the predicted value and the actual value is recorded for all historical time points, this is the error dataset.
6. After all errors are collected into a dataset, the absolute value of the median error is determined. This is the |*Median Error*|.
7. The difference between the |*Median Error*| and the error for each month is collected into a dataset. The median of this dataset is determined. This is the |*Median Absolute Deviation*|.
8. Now a boundary for a given time point can be calculated. The formula is: (*ARIMA Predicted upper 10% median net price change)* + |*Median Error*| + *2.5** |*Median Absolute Deviation*| for the upper bound and (*ARIMA Predicted lower 10% median net price change)* - |*Median Error*|- *2.5** |*Median Absolute Deviation*| for the lower bound.

After identifying the positive and negative tolerance levels using the ARIMA model, BLS economists reviewed the tolerance levels for their assigned industries. Overall, the response was that using the data driven ARIMA models to determine the tolerance levels for each industry seemed advantageous for identifying outlier price changes in the microdata in an efficient manner. In general, the absolute value of tolerance levels increased based on historical trends, indicating that using the tolerance levels determined by the model would identify fewer outliers requiring review.

Further data will be compiled in the months following the October 2025 implementation of this new process.

Challenges

While the initial model worked well for identifying tolerance levels for most industries, there were some challenges with generating tolerances for combined industries (roll-ups), re-coded industries, and indexes below the 6-digit industry level.

Roll-up Industries

A roll-up industry occurs when two or more 6-digit NAICS industries are combined to form a new 6-digit industry. The NAICS code of the roll-up industry is often different from the NAICS codes of the previous industries. For example, with the 2022 NAICS revision Portfolio Management (NAICS 523920) and Investment Advice (NAICS 523930) were combined into a single industry Portfolio Management and Investment Advice (NAICS 523940).

To generate dynamic tolerance thresholds for the upcoming month, the ARIMA model uses historical monthly price data. Initially, a new roll-up industry has very few months of price history within the new 6-digit NAICS code. However, in many cases, the historical price data from the two or more industries that now comprise the roll-up industry can be used by the model to generate the tolerance levels for the upcoming period for the new roll-up industry. Eventually, the roll-up industry will have enough months of historical data so that data from the prior NAICS industries will not be needed by the model.

Running the ARIMA model in these situations requires a manual process to combine the available historical data from the new roll-up industry with the additional historical data from the prior industries that now comprise the roll-up to provide enough historical data for the model to generate tolerances for the upcoming period at the 6-digit level. This is a manual process that needs to occur for each roll-up industry to run the model each month.

In some instances, the prior industries that have been combined to form the roll-up industry may have experienced different price trends. In these situations, it may not be appropriate to run the model on the combined historical data at the 6-digit level to determine the tolerance levels for the newly formed roll-up industry. Depending on the lower-level structure of the roll-up, it may be more appropriate to run the model on price changes for index levels underneath the 6-digit, if available.

Given resource constraints, BLS will not use the dynamic model on roll-up industries in the PPI until there are at least two years – 24 months – of historical data for the new, rolled-up 6-digit NAICS code. Currently, over 80 percent of the industries for which BLS produces PPIs have the minimum amount of data for the model and that number is expected to increase over time. In the interim, BLS economists will decide the best approach for setting tolerance thresholds. Options include using the tolerance values from the pre-roll-up industries, using the tolerances from industries that exhibit similar price trends, using industry knowledge to determine the tolerance levels, or a combination of two or more of these approaches.

Recoded Industries

Recoded industries often represent the same or nearly the same industry definition as before but have experienced a change to their NAICS code. As with roll-up industries, the newly recoded industry has very few months of price history. Running the model with enough months of data requires manually attaching the price history from the previous NAICS code to the new NAICS code, which is currently not feasible due to resource constraints. As a result, for now, BLS will not use the dynamic model on recoded NAICS industries in the PPI until there are at least 2 years of historical data for the new NAICS industry code. Approaches similar to those discussed for roll-ups can be used to determine appropriate tolerance thresholds until two years of data are available.

Setting Tolerance Levels Below the 6-Digit Industry Level

Some industries have lower-level indexes that experience price trends or seasonal changes that are different from each other or from their 6-digit industry level index. For these industries, the ARIMA model can be used to identify separate tolerance thresholds at the 7-digit index level or lower. For example, if an industry consisted of three 7-digit lower-level indexes, the model could calculate separate tolerance thresholds for each 7-digit index. For a different industry, perhaps one 7-digit index moves differently from the remaining lower-level indexes. In this case, a separate set of tolerance thresholds could be calculated for the 7-digit index only and all other remaining indexes would be run together to determine their tolerance threshold.

Setting separate tolerances for lower-level indexes is beneficial for some industries. However, this approach faces challenges once implemented into production. One challenge is data quantity – 24 months of price data needs to be available at a detailed level for the ARIMA model to accurately predict tolerance levels for a lower-level index.

Another challenge is index restructuring. If the lower-level indexes are restructured, it may be challenging to identify the historical data that would be needed by the ARIMA model to generate tolerance thresholds for the new lower-level indexes. In this case, reverting back to calculating tolerances at the 6-digit level may be required until the new index structure contains 24 months of historical data to identify tolerance thresholds at a lower level. Alternatively, BLS economists could use industry knowledge to determine appropriate tolerance thresholds until the new index structure contains enough historical data to run the data-driven model.

In instances where the model is used to calculate separate tolerance levels for lower-level indexes, tolerance levels for the “other receipts” indexes also need to be determined. Other receipts indexes consist of all other products and services provided by industry firms that are not considered primary outputs of that industry. A wide variety of products and services with differing price trends can be included in the other receipts index for a given industry. For this reason, it may not be feasible or practical to run the model only on the price change data in the other receipts indexes. In this case, BLS economists may need to determine appropriate tolerance thresholds.

Given these challenges and resource constraints, the BLS is only running the model at the 6-digit NAICS industry level at this time.

Implementation

The initial implementation plan included automatically loading the new monthly threshold results from the ARIMA model into the PPI production system each month for outlier detection. However, this was not feasible due to resource constraints and the special nature of the reported microdata for some industries. In some cases, BLS economists prefer to review more outliers than the dynamic tolerance levels would detect, such as when a new sample is introduced or there is a change in the pricing methodology.

Instead, the ARIMA model is run every month, and the resulting tolerance levels are provided in a dashboard format created using R. BLS economists review the dynamic data-driven tolerance levels each month and use that information, along with their industry knowledge, to decide whether to revise the thresholds that are used to detect outliers in the production system.

The following is an overview of the process:

1. The model runs once a month.
2. Tolerance level thresholds for industry and/or detailed indexes are provided in an interactive dashboard application.
3. BLS economists review the ARIMA model results for that month and decide to:
    a. Leave the tolerances thresholds as is
    b. Set the tolerance levels based on the data driven model output
    c. Set the tolerance levels based on their own knowledge of the industry
4. Updated tolerance levels are entered into the production system for current month outlier detection.
    a. Updates made to tolerances immediately take effect in the production system.
    b. Any updates made to tolerances carry forward to future months
5. PPI production system detects outliers based on the current tolerance thresholds in the production system and flags them for review as monthly microdata are submitted.

Conclusion

The ARIMA model for dynamically setting price change tolerance levels and the production process described above are meant to create operational efficiencies. By using historical price change data to dynamically set tolerance levels for outlier detection, time can be better allocated to reviewing significant changes in monthly microdata rather than changes that would be considered typical for a particular industry. While the expectation is that the amount of microdata requiring review will decrease, it is too early in the implementation of this process to assess changes in workloads. Additionally, microdata identified for review for other reasons including changes to the transaction description or terms of transaction will persist. BLS economists will also continue to look for outliers in their review of calculated monthly PPIs. Going forward, BLS will study the effect this new tolerance-setting model has on microdata review efficiency, aggregated index data review, and ultimately, on resource allocation.