Data Drive Model for Outlier Detection

Melanie Santiago
Industry Pricing Branch Chief
Producer Price Index
U.S. Bureau of Labor Statistics

39th Voorburg Group Meeting
September 22 – 26, 2025

This presentation contains research materials and is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Original analysis contained herein is presented for statistical purposes. Findings do not represent official BLS statistical products.

Outlier detection in the US PPI

Use of tolerance thresholds

- Quality checks on reported microdata
- System flag for review
- Price verification by BLS economist
- Justification for price change logged

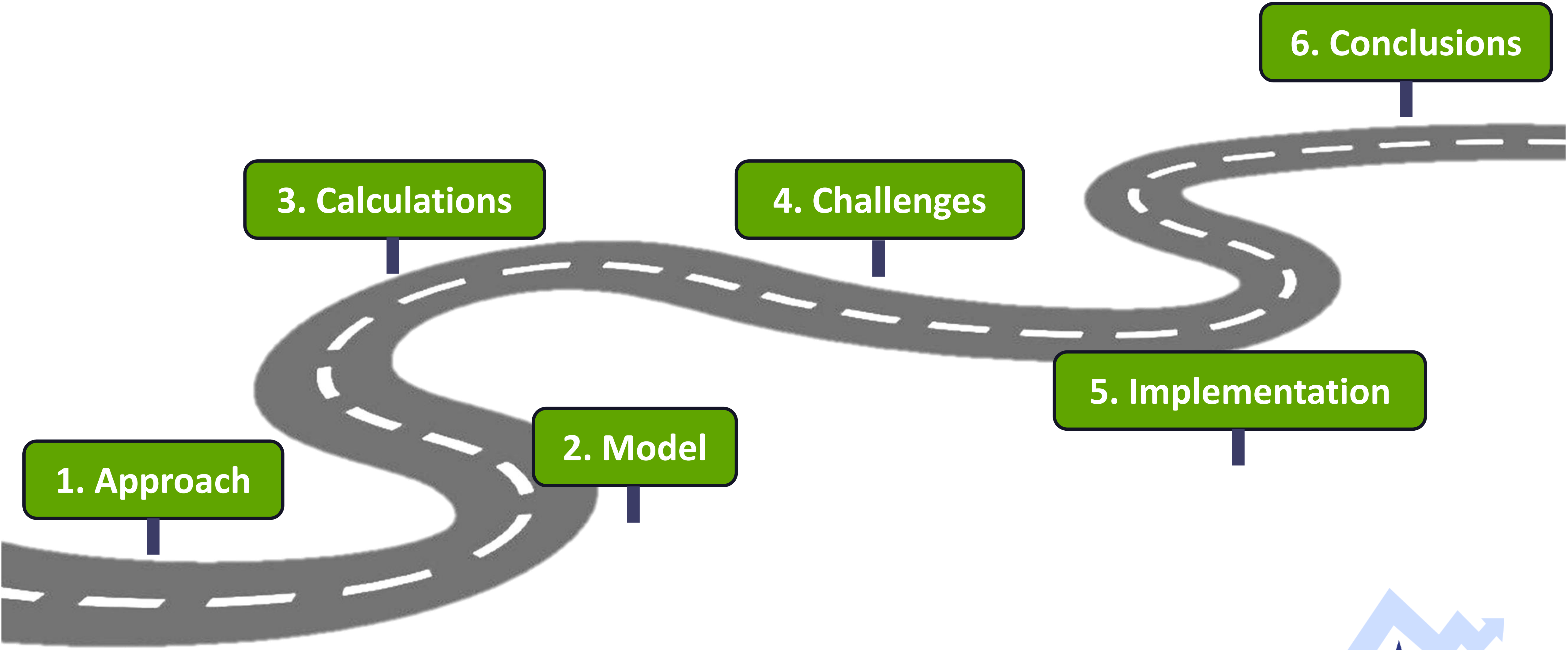
Traditional method for setting tolerance level

- Set when new industry sample introduced
- Values for positive and negative thresholds
- Static values in data processing system

Goals

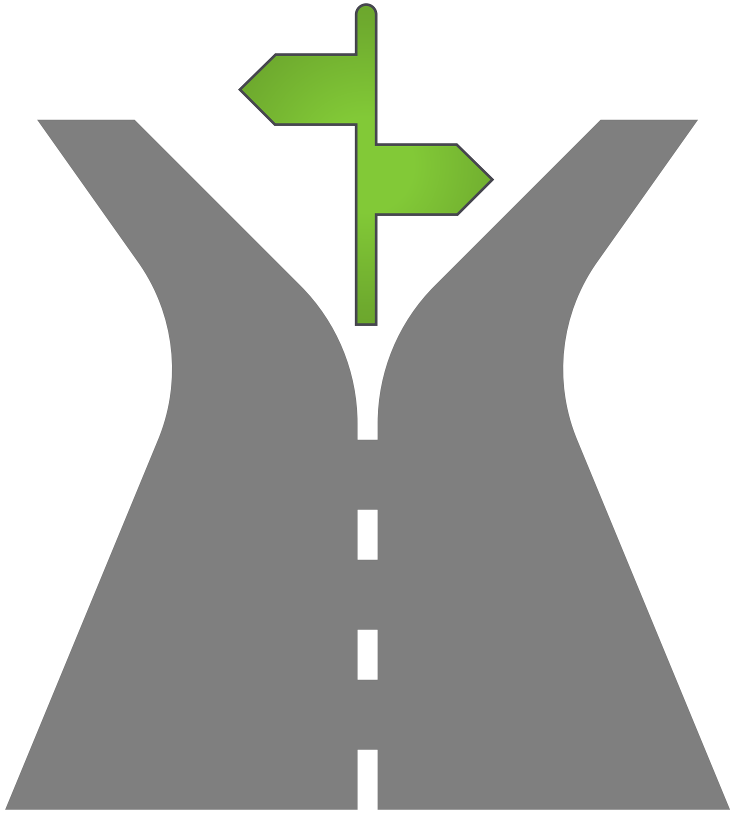
- Data-driven decision making
- Improved outlier detection
- Dynamic process
- Operational efficiencies

Roadmap


1. Approach
2. Model
3. Calculations
4. Challenges
5. Implementation
6. Conclusions

Approach

Develop ARIMA Model: Autoregressive Integrated Moving Average



Explore alternative modeling techniques

Modeling techniques

	Accuracy	Generalizability	Computational Efficiency	Interpretability
ARIMA: Autoregressive Integrated Moving Average	✓	✓	✓	✓
ETS: Error, Trend and Seasonality	✗			
Machine Learning				✗
Deep Learning			✗	

Model

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

Parameters

p: range of values allowed for autoregressive terms
d: range of values allowed for order difference
q: range of values allowed for moving average terms

P: range of values allowed for seasonal autoregressive terms
D: range of values allowed for seasonal order of differences
Q: range of values allowed for seasonal moving avg terms

m: seasonality

Seasonal model

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

- Up to 24 months of net price changes
- Up to 24 months of past forecast errors used to predict future errors
- Incorporate seasonal trends from up to 2 prior 12-month periods

Non-seasonal model

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

- Up to 12 months of net price changes
- Up to 12 months of past forecast errors used to predict future errors
- No seasonal trends used

Data preparation

Consolidate price change data into two price change values per month	
Median of top 10 percent of all positive net price changes in each industry – for positive outlier detection model	Median of bottom 10 percent of all negative net price changes in each industry – for negative outlier detection model

Calculation

1. Enter median of upper 10% of positive or lower 10% of negative net price changes into Auto-ARIMA
2. Select seasonal or non-seasonal model based on amount of historical data available
3. Use Auto-ARIMA to generate parameter options
4. Determine optimal parameters by AIC for ARIMA model for tolerance level predictions
5. Create error dataset (difference between predicted and actual value for all historical time periods)
6. Determine absolute value of median error
7. Create deviation dataset (median of differences between Median Error and actual error for each time period)
8. Calculate tolerance level for a given time period

Challenges

Structural changes (Combined or recoded industries)

Model requires two years of price data for each industry code

New industry codes have limited historical data

Manual process to link data from the prior and component indexes to new and combined industry codes

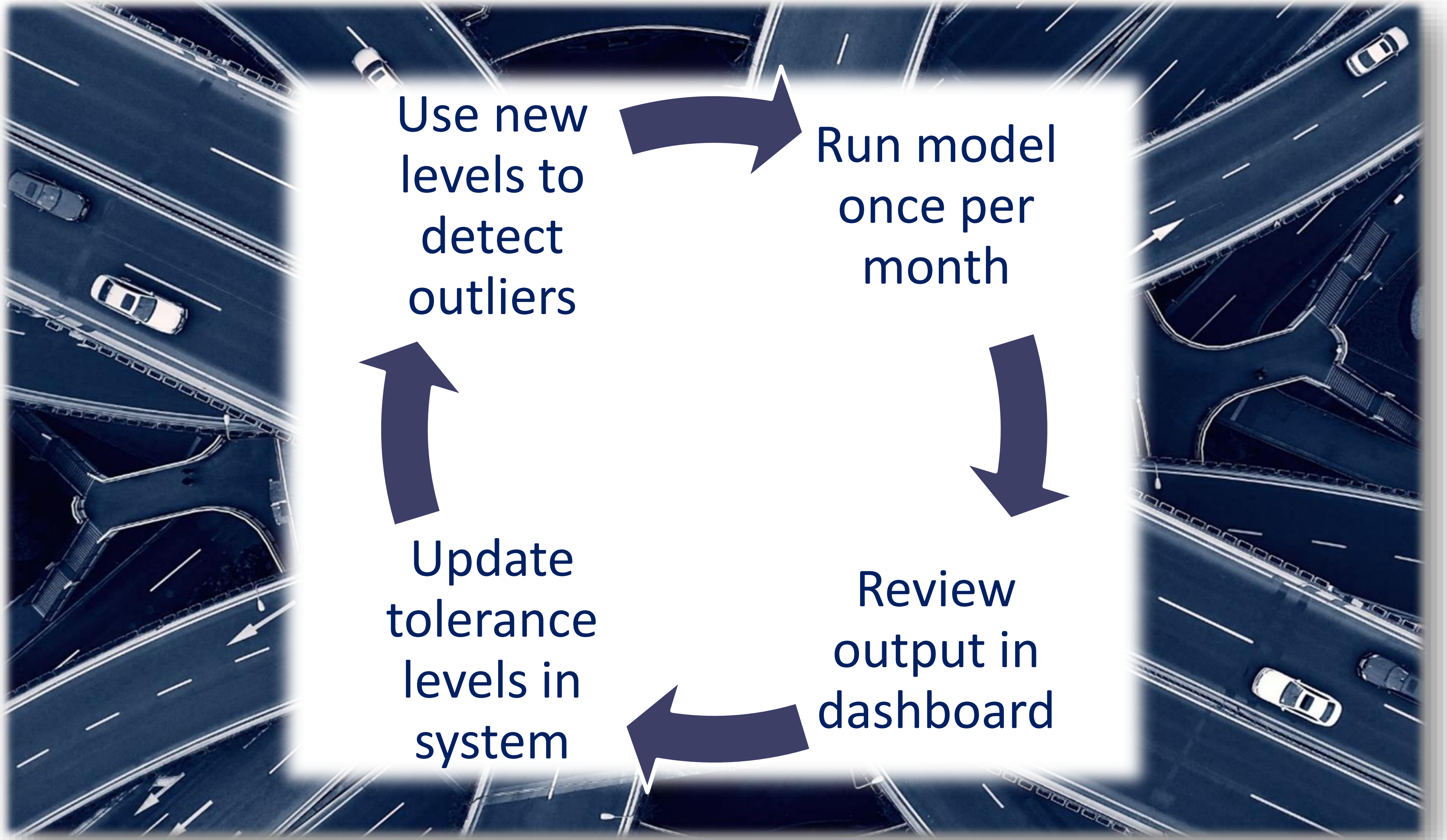
Product level tolerances

Different products produced within an industry may exhibit different price trends

Lower-level indexes are prone to more frequent structural changes

Structural changes result in a lack of required historical data

Implementation



Dashboard

PRICE TOLERANCE MODEL PREDICTIONS

MODEL PREDICTIONS | DATA DETAILS

Excel | CSV | Print | Show 50 entries | Search:

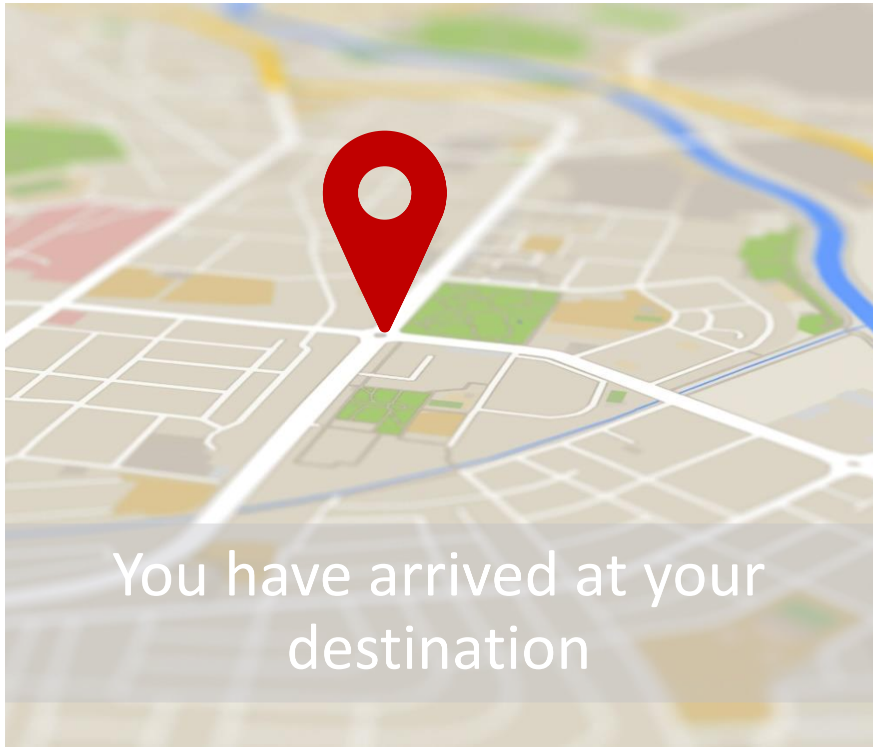
Tree ID	Tree Title	Price IRM	Current Upper Bound	Current Lower Bound	Predicted Upper Bound	Predicted Lower Bound	Assigned Analyst	Original Lower, Upper Bounds	Historical Predictions	Predictions Plot
All	All	All	All	All	All	All	All	All	All	All
111110	Soybean farming	2025-06-01	97	-97	115.5644	-113.0278		-97, 97	DOWNLOAD CSV	VIEW PLOT
111120	Oilseed, except soybean, farming	2025-06-01	75	-75	121.8703	-121.1752		-75, 75	DOWNLOAD CSV	VIEW PLOT
111130	Dry pea and bean farming	2025-06-01	53	-53	100.2426	-103.5933		-53, 53	DOWNLOAD CSV	VIEW PLOT
111140	Wheat farming	2025-06-01	62	-62	105.2235	-104.045		-62, 62	DOWNLOAD CSV	VIEW PLOT
111150	Corn farming	2025-06-01	52	-52	109.3015	-108.4475		-52, 52	DOWNLOAD CSV	VIEW PLOT
111199	All other grain farming	2025-06-01	75	-75	109.9039	-108.2486		-75, 75	DOWNLOAD CSV	VIEW PLOT

Showing 1 to 6 of 6 entries

Previous 1 Next

Conclusions

- Data driven decision making
  - Microdata review efficiency
  - Reduce workloads?
  - Improve resource allocation



Contact Information

Melanie Santiago
Industry Pricing Branch Chief
Producer Price Index
U.S. Bureau of Labor Statistics
www.bls.gov/ppi
202-691-7844
santiago.melanie@bls.gov