A Comparative Analysis of Statistical Methods for Detecting Emerging Chemical Signals Utilizing NIH Grant Data

Amirreza Nickkar , Ph.D.* & Ernest Kwegyir-Afful, Ph.D. **

* Visiting Scientist, Hazard Assessment & Analytics Branch, Office of Post-Market Assessment, Human Foods Program

** Acting Chief, Hazard Assessment & Analytics Branch, Office of Post-Market Assessment, Human Foods Program

Abstract

The Z-score-based method is particularly simple and useful in contexts where immediate identification of extreme changes is critical.

This research aims to assess and compare several statistical methodologies to determine which is most effective at detecting anomalies in the total number of projects and the volume of allocated grants that can be used as an indicator of an emerging chemical signal of interest to food safety regulation.

We hypothesized that the fluctuations in the number of funded grants and the total financial resources allocated to research topics can serve as an indicator of shifting priorities and heightened focus from scientists and decision makers in areas of public health. This study focuses on grants funded by the National Institutes of Health (NIH)* across various disciplines, evaluating the efficacy of three distinct statistical methods: the Z-score based analysis, the cumulative sum (CUSUM), and regression analysis. These methods were chosen for their differing approaches to identifying deviations from expected patterns.

The comparative analysis of these methods showed several similarities in their ability to detect shifts in funding patterns; however, the Z-score based method demonstrated a marginally better performance in capturing significant anomalies.

** The data was retrieved from the NIH RePORTER database in January 2024. Available at: https://reporter.nih.gov*

Materials and methods

The z-score based method approach involves the following steps:

Data Collection: Gathering projects costs and the annual number of projects involving a specific chemical.
Pre-processing: Cleaning and normalizing data to ensure quality and comparability.

Statistical Analysis: Using statistical metrics like mean, median, standard deviation, and interquartile ranges to model the distribution of the data.

$$\mu_{count\_diff} = \frac{1}{N-1}\sum_{t=1}^{N} count_{d_iff_t}$$

$$\sigma_{count\_diff} = \sqrt{\frac{1}{N-1}}\sum_{t=1}^{N}(count_{d_iff_t} - \mu_{count_{diff}})^2$$

$$\mu_{count\_diff}, \sigma_{count\_diff}, \mu_{cost\_diff}, \sigma_{cost\_diff}$$

Threshold Setting: Establishing bounds or thresholds, beyond which data points are considered anomalous. The threshold is set at the mean plus one standard deviation because under a normal distribution assumption, approximately 68% of the data falls within one standard deviation of the mean.

$$\text{Threshold}_{count} = \mu_{count\_diff} + \sigma_{count\_diff}$$

$$\text{Threshold}_{cost} = \mu_{cost\_diff} + \sigma_{cost\_diff}$$

Anomaly Identification: Comparing data points against the set thresholds to flag those that deviate significantly from the norm.

$$\text{Anomaly} = \begin{cases} 1, & if\,(count\_diff > \text{Threshold}count) \text{ or } (cost\_diff > \text{Threshold}cost) \\ 0, & otherwise \end{cases}$$

Importance Ranking: Assigning an importance index to each anomaly, quantifying its significance based on its deviation from the statistical norm. This step helps prioritize anomalies for further investigation or action.

$$Importance_t = \frac{1}{2}\left(\frac{|count_{d_iff_t} - \mu_{count\_diff}|}{\sigma_{count\_diff}} + \frac{|cost_{d_iff_t} - \mu_{count\_diff}|}{\sigma_{cost\_diff}}\right)$$

Review and Action: Analyzing the anomalies for insights.

The general process of the three methods Z-score based method, regression analysis and CUSUM have been provided in Figure 1.

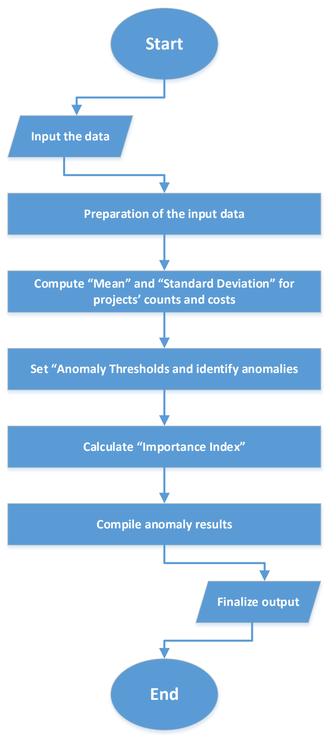

Fig. 1(a). Z-score based method

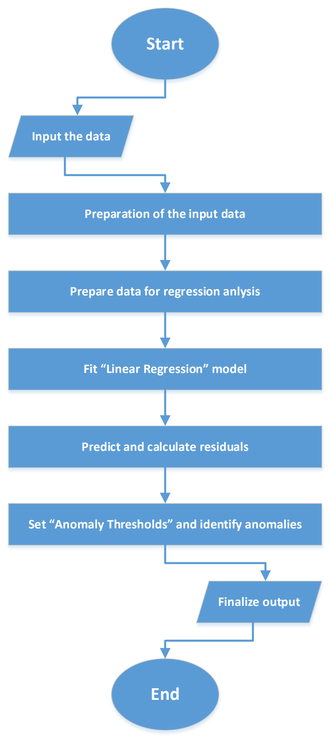

Fig. 1(b). Regression analysis

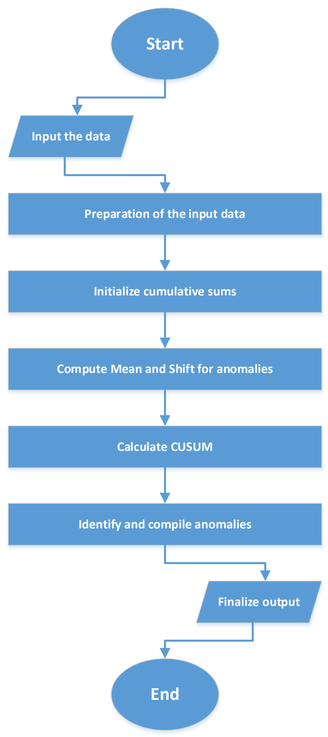

Fig. 1(c). CUSUM

Algorithm

1. Initialize an empty list called 'anomalies' to store results.
2. For each unique substance in the input dataframe:
   a. Sort the data by 'start_year'.
   b. Calculate 'count_diff' as the difference in 'project_count' from the previous year.
   c. Calculate 'cost_diff' as the difference in 'project_total_cost' from the previous year.
   d. Compute the mean (μ) and standard deviation (σ) for both 'count_diff' and 'cost_diff'.
   e. Set 'count_threshold' as μ_count + σ_count.
   f. Set 'cost_threshold' as μ_cost + σ_cost.
   g. Normalize the 'importance index' for the group by dividing by the maximum 'importance index' in that group if it is greater than zero.
3. Combine all data in 'anomalies' list into a single dataframe.
4. Merge the original dataframe with the anomalies dataframe to align the anomaly data with the original records.
5. Convert anomaly markers from numerical IDs (0 for normal, 1 for anomaly) to descriptive labels ("Not Signal" for 0, "Signal" for 1).
6. Remove any unnecessary columns from the final dataframe and return the result.

Results

The results of the three methods for two substances of CLOPIDOGREL and TITANIUM DIOXIDE have been provided in figures 2 and 3. These two substances were prime examples of a subset of substances which demonstrated significant anomalies which made them ideal candidates for testing our process.

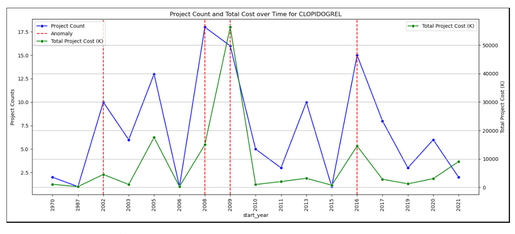

Fig. 2(a). Z-score based method

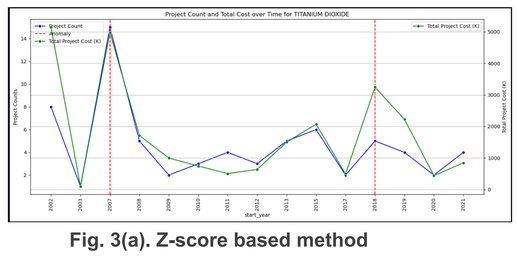

Fig. 3(a). Z-score based method

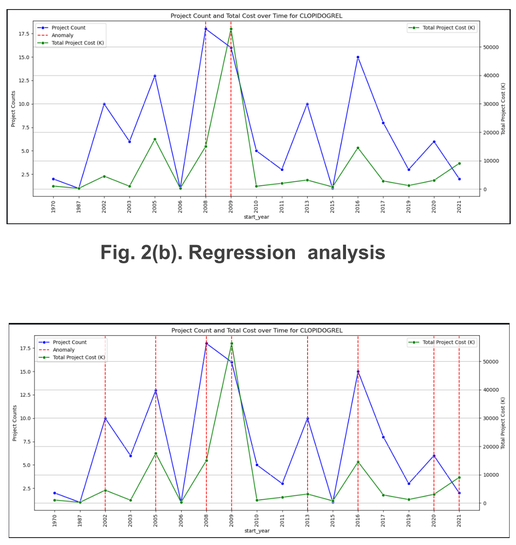

Fig. 2(b). Regression analysis

Fig. 3(b). Regression analysis

Fig. 2(c). CUSUM

Fig.2 . CLOPIDOGREL

Fig. 3(c). CUSUM

Fig.3. TITANIUM DIOXIDE

Conclusion

This z-score-based method of statistical anomaly detection is crucial for its ability to systematically identify, prioritize, and quantify anomalies in large datasets.

The contribution and importance of the statistical anomaly detection method used in the provided code are significant, particularly in contexts where monitoring changes over time is crucial.

- By calculating the mean and standard deviation, the method effectively normalizes the data, reducing noise and improving the accuracy of identifying what constitutes an anomaly versus a normal fluctuation. This is critical in environments where data can be volatile or highly variable.
- The method involves setting thresholds based on statistical measures. This thresholds can be adjusted using an optimization method to cater to different levels of sensitivity required.
- Calculating an importance index for each anomaly is a key contribution of this method. It quantifies the severity of each anomaly, not just flagging its presence.
- The method is not restricted by the scale of the dataset or its domain. It can be adapted and applied to various types of data across different sectors, making it a versatile tool for data analysts and scientists.
- By providing a systematic approach to detect and quantify anomalies, the method supports enhanced decision-making. Decision-makers can use these insights to make informed decisions about where to allocate resources, how to adjust strategies, and when to intervene in processes.

Acknowledgments

The authors would like to thank Dr. Kirk Arvidson, Chief of Regulatory Management Branch, Office or Post-Market Assessment for his insightful comments.