Research on the Application of Statistical Anomaly Detection in Tobacco Sales Anomaly Detection

Sen Yang*, Mingsheng Zhu

China National Tobacco Corporation, Chongqing Company, Tongnan Branch,Chongqing, 500223, China.
**Corresponding author:pengmiwei1990@sina.com*

Abstract: *With the rapid development of a new generation of information technology, digitalization and informatization have become key factors in enhancing the core competitiveness of the tobacco industry. Detecting and monitoring abnormal tobacco sales behaviors play a vital role in combating illegal transactions and ensuring the healthy development of the tobacco industry. Manual identification of abnormal sales behaviors presents practical issues, such as low efficiency, strong subjectivity, and a reliance on accumulated experience. Utilizing statistical probability theory to help the tobacco bureau determine the threshold for defining abnormal sales behavior can effectively improve the efficiency of anomaly detection. This study applies relevant theories of statistical anomaly detection in combination with practical production needs, establishing four threshold criteria for identifying abnormal sales behaviors: the number of days without retail data, the ratio of daily sales volume to the sales volume of the previous seven days, the number of days of available sales, and the number of days with scanning activity per week. These criteria make it possible to automate and dynamically identify abnormal sales behaviors in subsequent sales situations.*

Keywords: *statistical anomaly detection, data analysis, tobacco sales.*

1. Introduction

With the rapid development of a new generation of information technology, digitalization and informatization have become key factors in enhancing the core competitiveness of the tobacco industry [1]. The government's understanding of the regularities of technological advancements driving productivity, along with its emphasis on building a powerful digital nation, has accelerated the tobacco industry's digital transformation.

In the current digital era, data has become a critical basis for corporate decision-making. By establishing data analytics platforms, tobacco companies can achieve comprehensive and precise analyses of markets, consumers, and supply chains, thereby supporting strategic planning and operational management. The tobacco industry is actively collaborating with universities and research institutions to jointly develop new technologies and processes, promoting industrial transformation and upgrading [2].

Moreover, the State Tobacco Monopoly Administration has issued a series of policy documents encouraging and supporting tobacco companies to strengthen their information infrastructure, advance digital transformation, and achieve high-quality development. However, the industry faces numerous

challenges in its digital transformation, including technological, organizational, security, and market-related issues. Machine learning, as an advanced technological tool, can help the tobacco industry address these challenges, realize intelligent operations, and maintain competitiveness [3].

To combat illegal transactions, ensure tax compliance, maintain market order, and promote the healthy development of the tobacco industry, the tobacco authorities have been committed to identifying and addressing abnormal sales behaviors among retailers. However, the manual identification of such abnormal behaviors is often inefficient, highly subjective, and relies heavily on accumulated experience. Therefore, applying statistical methods to screen for potentially abnormal sales behaviors, followed by manual review, can enhance detection efficiency. The most critical aspect of this process is defining what constitutes an abnormal sales behavior and establishing the threshold for determining such behaviors.

2. Relevant Technologies

2.1 Overview of Statistical Anomaly Detection

Statistical anomaly detection[4]is a method based on statistical principles, used to identify anomalies or outliers in a dataset. Outliers[5]refer to data points that significantly deviate from other observations in the dataset; these points may represent measurement errors, data entry mistakes, or special cases in the actual process. The principles of statistical anomaly detection mainly include the following aspects:

Data Distribution Assumptions: Statistical anomaly detection is typically based on certain assumptions about data distribution. For example, data may be assumed to follow a normal distribution, Poisson distribution, or other probability distributions. Parameter Estimation: After assuming the data distribution, the parameters of the distribution (such as mean and variance) need to be estimated. These parameters can be estimated from sample data, for example, by using the sample mean and sample variance. Construction of Anomaly Standards: Based on the estimated parameters, statistical criteria for detecting anomalies can be constructed. Adaptability and Robustness: In certain cases, anomaly detection methods need to adapt to changes in data distribution or maintain robustness against inaccuracies in initial parameter estimation. Visualization: Statistical anomaly detection is often combined with visualization tools, such as box plots, scatter plots, or residual plots, to visually display anomalies in the data.

The most common methods for constructing anomaly standards include: Threshold Method: Setting one or more thresholds, such as the mean plus or minus several standard deviations (e.g., $\pm 2\sigma$ or $\pm 3\sigma$). Data points outside these thresholds are considered anomalies. P-Value Method: Calculating the degree of deviation of each data point from the overall distribution and using the p-value to decide whether to reject the hypothesis (i.e., determine if the point is an anomaly). Anomaly Scoring: Assigning an anomaly score to each data point, which reflects the likelihood of that point being an anomaly. The scoring can be based on distance metrics (such as Euclidean distance, Mahalanobis distance) or probability models. Model Fitting: In some cases, more complex statistical models can be used to fit the data, and anomalies can be identified based on the residuals of the model. For example, the size of the residuals in a linear regression model can indicate whether a data point is likely to be an anomaly. Multidimensional Data: For multidimensional data, multivariate statistical methods such as Principal Component Analysis (PCA), Mahalanobis distance, or Canonical Correlation Analysis can be used to identify anomalies.

Statistical anomaly detection has a wide range of applications, including financial fraud detection, cybersecurity, industrial process monitoring, and medical diagnosis. In practical applications, the choice of anomaly detection methods depends on the characteristics of the data, its distribution, and business needs.

This paper primarily utilizes probability distributions such as the normal distribution and the chi-square distribution, both of which are very important in statistics. The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution. Its graph is bell-shaped and symmetric on both sides. The normal distribution is characterized by the equality of the mean, median, and mode, and the data distribution is symmetric around the mean. Many phenomena in natural and social sciences approximate a normal distribution, such as human height and weight.

The mathematical expression for the normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ\muμ is the mean and σ\sigmaσ is the standard deviation. The normal distribution is characterized by the equality of the mean, median, and mode, and the data distribution is symmetric around the mean.

The chi-square distribution, on the other hand, is a discrete probability distribution commonly used in statistical inference. Its graph is unimodal, and its shape depends on the degrees of freedom. The chi-square distribution is often used for hypothesis testing, such as independence tests and goodness-of-fit tests. When a random variable follows a standard normal distribution and there are multiple such independent random variables, the distribution of the sum of their squares follows a chi-square distribution.

The mathematical expression for the chi-square distribution is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

where E\text{E}E is the expected value, usually equal to the degrees of freedom multiplied by the variance.

There are some connections between these two distributions: if a random variable follows a standard normal distribution, it also follows a chi-square distribution with one degree of freedom. Additionally, the ratio of the sample mean to the sample variance in a normal distribution also follows a chi-square distribution. Both distributions are fundamental in hypothesis testing and parameter estimation in statistics.

Furthermore, the 3σ theory is utilized, which is a method in statistics for describing data distribution. The main idea is that if a process's output falls within the 3σ range, the process is considered acceptable, as only a small fraction of outputs will exceed this range. Data falling outside this range can be deemed abnormal. The specific rules are as follows:

The probability of a data point falling within ±1 standard deviation (σ) from the mean is approximately 68.27%.

The probability of a data point falling within ±2 standard deviations from the mean is approximately 95.45%.

The probability of a data point falling within ±3 standard deviations from the mean is approximately 99.73%.

2.2 Applications of Anomaly Detection in the Tobacco Industry

Anomaly detection technology is primarily applied in the tobacco industry in the following areas:

First, it is used to combat illegal transactions, ensure tax compliance, and maintain market order. By analyzing tobacco sales data, anomaly detection technology helps identify illegal activities such as illegal distribution, smuggling, or unauthorized sales. It also monitors tobacco sales and tax data to ensure tax compliance and prevent tax evasion. By identifying and analyzing anomalous patterns in sales dat[6], tobacco companies and regulatory agencies can maintain market order and prevent disruption by illicit tobacco products.

Second, anomaly detection technology can be employed in inventory management, consumer behavior analysis, risk management, and market trend analysis. It helps tobacco companies optimize inventory management by analyzing sales and inventory data to forecast demand and adjust inventory levels accordingly. It also analyzes consumer purchasing patterns to identify unusual buying behaviors or potential risk factors, such as changes in market demand or policy shifts, allowing companies to manage risks more effectively. Additionally, by analyzing anomalies in sales data, tobacco companies can gain insights into market trends and shifts in consumer preferences.

The application of anomaly detection technology helps the tobacco industry enhance operational efficiency, strengthen risk management, ensure compliance, and protect consumer interests. As data analysis technology advances, the use of anomaly detection in the tobacco industry will become more widespread and in-depth.

3. Experiments and Results

After thorough consideration of tobacco industry sales conditions and subsequent discussion and analysis, the following four criteria were selected for detecting anomalies: days without retail data, the ratio of single-day sales to sales over the past 7 days, sell-through days (the ratio of daily sales to inventory), and the number of weekly scans. The dataset used in this study was extracted from online store data by the tobacco bureau, resulting in a selection of 1,530 retail outlets and approximately 30,000 inventory records.

3.1 Data Analysis

Days without retail data: Typically, merchants should have consistent sales regardless of holidays or workdays. Therefore, if sales are zero for multiple consecutive days, it may indicate incomplete sales records or deliberate concealment of sales, suggesting that the merchant is experiencing anomalies. The specific method involves marking days with zero sales for each merchant as days without sales data. By examining the merchant's sales data, if the sales amount for a merchant is zero for N consecutive days, this period is flagged as a continuous no-sales data anomaly.

Ratio of single-day sales to sales over the past seven days: The daily sales volume of a merchant should generally be stable and not fluctuate significantly. Therefore, if the ratio of single-day sales to the sales over the past seven days is excessively high, it may indicate risks such as illegal reselling or

manipulation of sales data, suggesting anomalous sales behavior. This project has calculated the distribution of the ratio of single-day sales to the sales over the past seven days, as shown in Figure 3-1. The data overall conforms to a chi-square distribution.

Sell-through days: Sell-through days refer to the ratio of daily sales to inventory. Normally, merchants do not sell a large portion of their inventory in a single day but strive to maintain inventory stability. If the sell-through days are abnormally high or low, it may indicate issues with inventory management or discrepancies between sales data and actual inventory. Therefore, the ratio of daily sales to inventory should fall within a reasonable range, with deviations from this range considered anomalies. The specific distribution is shown in Figure 3-2, where the data overall conforms to a normal distribution.

Weekly scan days: If a retailer does not use the scanning system on certain days of the week, it may suggest that sales on those days were not recorded or other non-compliant behaviors may be present. In addition to identifying anomalies through days without retail data, it is necessary to use weekly scan days as an auxiliary criterion for anomaly detection. Days without retail data only consider consecutive days, while Figure 3-3 shows that there may be intermittent days of no sales data within a week. Thus, calculating weekly scan days helps to identify dispersed anomalies.

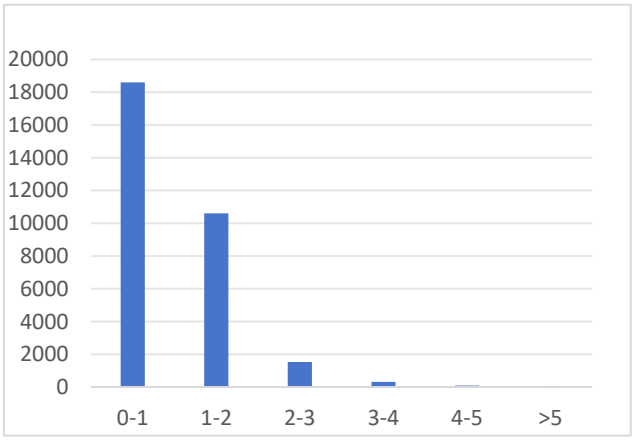

Figure 3-1 Distribution of Single-Day Sales to Average Sales Ratio

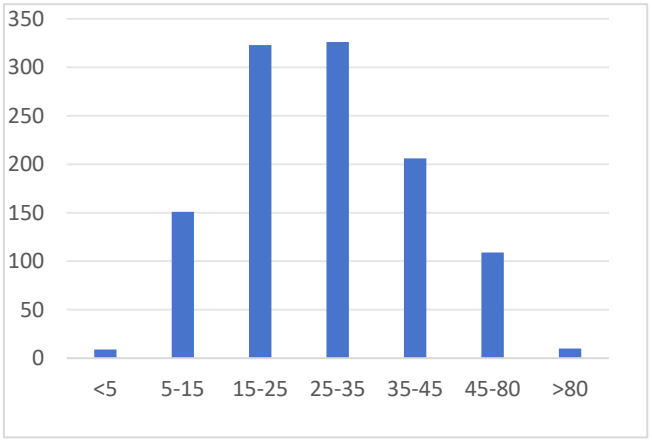

Figure 3-2 Distribution of Sellable Days

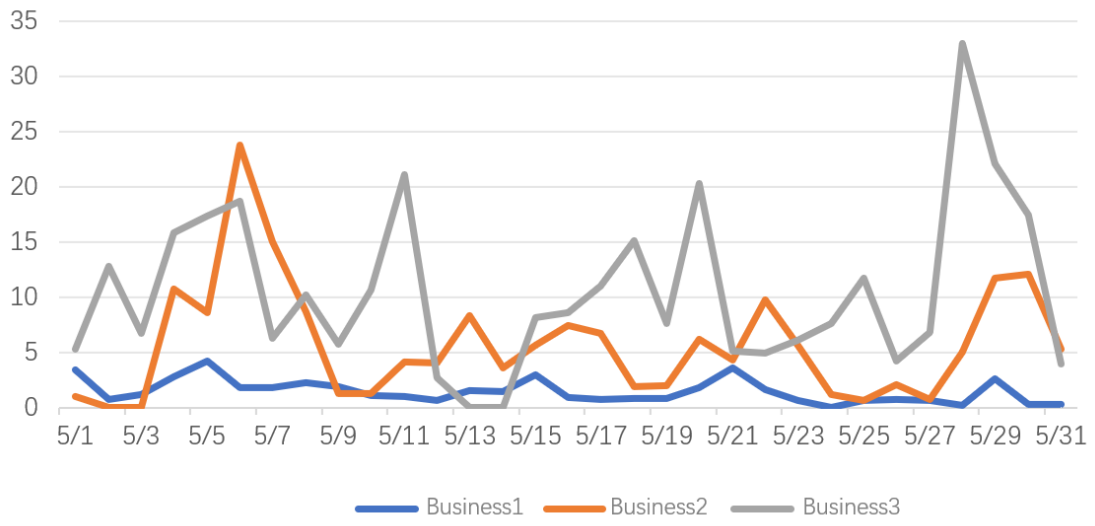

Figure 3-3 Line Chart of Sample Retailer Sales Volume

3.2 Experimental Results

Days with No Retail Data: Most of the data for days with no retail data is concentrated in the range of 0-2 days, with occurrences outside this range being rare. This results in identifying cases that differ from most merchants and are classified as anomalies. Calculating the probability based on the average for 254 recorded days over two years, the probability of having more than 3 consecutive days with no sales data is 0.0118, which fits the pattern of rare events in a normal distribution, defining it as anomalous.

Ratio of Single-Day Sales to the Average of the Past Seven Days: The ratio of daily sales to the average of the past seven days is mostly concentrated in the range of 0-2, with outliers being rare. Using percentiles to set the threshold for anomalies, the 99th percentile can be chosen as the normal range, with data points exceeding this range considered anomalous. As shown in Figure 3.1.2-1, a ratio greater than 3 is considered anomalous, which accounts for only 1.5% of the data.

Sellable Days: Based on the number of merchants, the 99th percentile is set as the normal threshold. Histogram analysis indicates that the sum of days greater than 80 and less than 5 accounts for 1.6%, with days greater than 80 accounting for 0.8%. By selecting appropriate percentiles, corresponding values are calculated as anomaly thresholds. Thus, sellable days less than 5 or greater than 80 are considered anomalous.

Weekly Scanning Days: Calculating the probability based on the average weekly scanning days for merchants over 48 weeks of the year, the probability of weekly scanning days being less than or equal to 6 for 3 weeks is 0.0625, while the probability of weekly scanning days being less than or equal to 5 for 1 week is 0.0208, aligning with rare events in a normal distribution. Therefore, weekly scanning days less than 5 are defined as anomalous.

4. Summary and Outlook

In this study, the application of statistical anomaly detection in monitoring unusual behavior in

tobacco sales was thoroughly explored, and a method for identifying anomalous sales behavior based on statistical probability principles was proposed. By aligning with actual production needs, four key indicators were identified: days with no retail data, the ratio of single-day sales to the average of the past seven days, sellable days, and weekly scanning days. These indicators provide a scientific basis for automated detection and dynamic identification of anomalous sales behavior. The effectiveness of statistical anomaly detection techniques in identifying unusual tobacco sales behavior was demonstrated, laying a foundation for further research and practice in this field. With ongoing technological advancements and innovations, it is believed that statistical anomaly detection will play an increasingly significant role in various domains.

References

[1] Chen, Yitong. Exploration of Digital Transformation in Tobacco Commercial Enterprises. Chinese Science and *Technology* Journal Database (Full Text Edition), Economic Management*, 2022(10):4.*

[2] Qu, Yanmei, Li, Xianneng, Sun, Shihang, et al. A Machine Learning-Based Intelligent Marketing System for the Tobacco Industry: CN202111646113.5. Patent CN202111646113.5 *[2024-06-02].*

[3] Fu, Qiuxuan. An Analysis of Major Issues and Solutions in the Digital Transformation of the Tobacco Industry in Province A. China Management Informationization*, 2022, 25(20):134-137.*

[4] Xiao, Huixu. Analysis of Statistical Data Anomaly Detection Methods in Campus Network Security. Integrated Circuit Applications*, 2023, 40(10):124-125.*

[5] Ge, Chengpeng, Zhao, Dong, Wang, Rui, et al. A Denoising Method for Segmented Point Clouds Based on Improved DBSCAN and Distance Consensus Evaluation. Journal of System Simulation*, 1-11 [2024-08-02].* https://doi.org/10.16182/j.issn1004731x.joss.24-0153.

[6] Xiao, Xiao, Feng, Pengcheng, Liu, Luni, et al. Tobacco Illegal Sales Early Warning Based on Time Series Prediction and Anomaly Detection. Journal of Guizhou Normal University (Natural Science Edition), 2023, 41(03):119-124.