Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(5):266-270

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Anomaly detection of cigarette sales using ARIMA on lunar calendar

Yang Xiao

School of Management Science and Engineering, Shandong University of Finance and Economics, China

ABSTRACT

Anomaly detection of cigarette sales amount and average price of total company, a specific brand and a specific salesman are important for detecting and preventing the illegal circulation of cigarette. By analyzing large amount historical sales records, a cigarette sales foresting model is built to predict the monthly and weekly sales status based on ARIMA model. And then an anomaly detective model is constructed on the predictive result. Although the ARIMA model has been analyzed extensively by researchers and used widely by forecasting practitioners due to its attractive theoretical properties and empirical evidence in its support, there are no empirical investigations have been conducted in the anomaly detection of cigarette sales. Result on the five years sales data from a Chinese city indicate that anomaly detection based on ARIMA on lunar calendar can be an effective way to improve forecasting accuracy and then improved anomaly detection accuracy is achieved.

Key words: Anomaly Detection, Cigarette Sales, Time Series Analysis, ARIMA, Lunar Calendar

INTRODUCTION

Cigarette market in China is a kind of monopoly market depends on the plan mechanism, and the sales of cigarette are planned and supervised by the government, in other words, cigarette sales supplement is fixed for each specific province, city and for each cigarette tailor. But people in different regions have different preference in cigarette brand, so some brand is much in demand in one region, but encounter poor sales in another. Since the supplement amount under supervision is not much reasonable, some tailors sell the unmarketable cigarettes to other tailors privately in which region the cigarettes are marketable. This kind of behavior disturbs the normal order of sales, and lead to the government cannot control the cigarette sales, so it is illegal in China. Therefore, besides the prediction of cigarette sales [1] to guide the cigarette company to stock and develop sales policies according to smokers' preferences, detecting and preventing this illegal circulation of cigarette is also important for government.

Currently, automatic anomaly warning mechanism is built in the cigarette business to detect abnormal sales data automatically by setting up some specific rules to define the abnormal brands or customers. For example:

(1) When the monthly sales volume of a specific band exceeds 50% over the amount of last three month sales, then the brand is identified as abnormal.
(2) When the monthly sales volume of a specific customer exceeds 200% over the last three month sales, then the customer is identified as abnormal.

The anomaly warning mechanism cannot reflect the truly abnormalities since the rules are too simple and tough. The warning mechanism produces excessive invalid warning especially when encounter holidays, which makes it difficult for the officers to find the real abnormal sales data from these huge amount of invalid warning data, which makes building an efficient anomaly detective mechanism is an urgent problem.
There are a large number of sales data accumulated during several years' cigarette sales, which makes it is possible using data mining on these data to solve anomaly detection problem.

SELECTION OF ANOMALY DETECTIVE MODEL

There are lots of literatures discussing the anomaly detection problem in several areas, and the techniques can be categorized into 6 types [2]. Among this, classification-based methods are considering the anomaly detection problem as a two-class classification problem, where samples are classified to normal and anomaly nodes. There are approach using training data with [3] or without anomalies [4] are provided. Clustering-based and density-based methods consider the outlier nodes as anomaly [5], while the statistical anomaly detection technique regard irrelevant observation as not generated by the stochastic model assumed, in which parametric [6] as well as non-parametric [7] techniques have been applied to detect anomaly, such as, histogram-based techniques are particularly popular in intrusion detection community [8].

Sales of cigarettes have shown obvious seasonal fluctuations. For instance, people will present gifts to each other in Chinese traditional holidays, and cigarette is one of welcomed gifts as a means of people interaction. Thus, cigarette sales are influenced much by the impact of holidays. Sales amount and the average price of the cigarettes from years 2008 to 2012 of a municipal cigarette corporation is shown as an example in Fig.1.

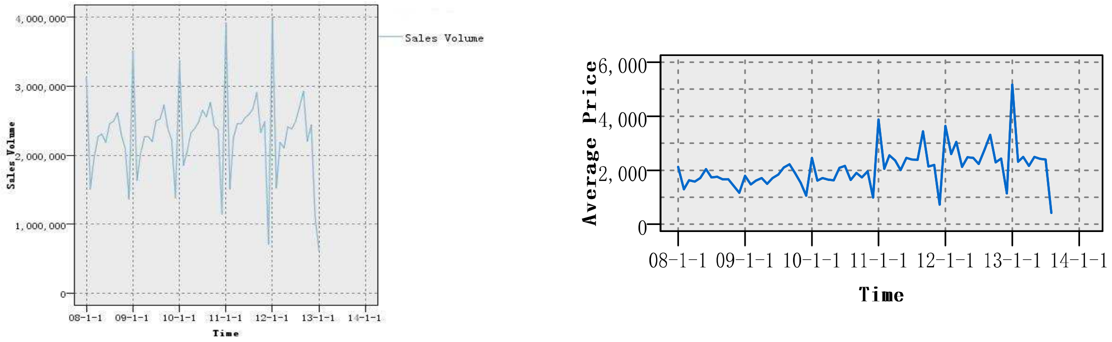

Fig.1. Sales amount and Average price of a city in years from 2008 to 2012

Due to the huge number of sales record, it is difficult to manually label anomaly record for each customer or cigarette and there not exists historical anomaly records, so it is important to automatically detect anomaly records without using any benchmark information. Since cigarette sales are affected extremely by holidays and other time-related factors, and the clustering or classification method cannot capture these time correlation, so we cannot choose classification-based or clustering-based method.

As can be seen from the cigarette sales amount and average price as Fig.1, we can find cigarette sales consistent with the typical characteristics of double trend. For such complex nonlinear system, the classic stochastic models such as moving average and regression are unable to get satisfactory results, so we choose the ARIMA (Auto Regression Integrated Moving Average) model, which is a parametric stochastic model to detect the cigarette sales anomaly.

ARIMA model is a widely applied time series model has been analyzed extensively by researchers and used widely by forecasting practitioners due to its attractive theoretical properties and empirical evidence. Although the model has been used for predicting cigarette sales amount [**Error! Bookmark not defined.**,9] and detect anomalies in many areas, such as Tsay et al. [10] detect anomalies in multivariate ARIMA model, there are no empirical investigations have been conducted to the anomaly detection of cigarette sales to our knowledge. This paper detect anomaly of cigarette sales using ARIMA model, and besides this, considering the sales of cigarettes are strongly affected by the holidays, especially the Chinese traditional holidays, the lunar calendar is used to construct the ARIMA model.

TYPES OF CIGARETTE SALES ANOMALY

There are several stations can be considered to be abnormal in cigarette sales. In this paper, we conclude four kinds of anomaly phenomenon often appeared in cigarette sales which can be detected from historical sales data.

(1) If the amount or the average price of the order is exceed much more/less than the predicted value, the order is called as “order anomaly”.

(2) Monthly total quantity and the average price about a specific customer if they are exceed much more/less than the predicted value, which called “customer anomaly”.

(3) Monthly sales amount of a specific brand is checked whether it is exceed much more/less than the predicted value, which called “brand anomaly”.

(4) Weekly sales amount and average price of all brands if they are exceed much more/less than the predicted ones, which called "entirety anomaly".

These anomalies give a hint of illegal circulation of cigarette. For example, when a customer with weak sales ability suddenly orders a large amount of cigarette, it should be considered a suspicious order. In this paper, we validate the proposed anomaly detective model for entirety anomaly as an example. The abnormal order, customer and abnormal brand can be solved by similar detective method. We predict the cigarette sales amount and average price using ARIMA model on lunar calendar, and then if the actual value is not in the predicted range, then it can be considered abnormal.

PREDICTION ON ARIMA USING LUNAR CALENDAR

Weekly sales volume and average price are needed to be predicted to check if the total sales volume and average price are abnormal. Using the sales data between years 2008 and the first half of 2013, we forecasted the sales amount and average price of the future 4 weeks. In some Chinese traditional festivals, especially Autumn Festival and Spring Festival, which are the most important traditional festivals to Chinese, people have custom to send each other gifts. High-grade cigarettes and wines are one of the important gift-giving items. Therefore, the sales of cigarettes will surge in these traditional festivals. This trend can be reflected in consumer sales fluctuations in Fig.1.
After deleting the null record, prediction based on ARIMA model is carried on in Clementine which is widely used data mining software. Taking entirety anomaly as an example, prediction on the solar calendar and lunar calendar are carried through respectively to test the effectiveness of time-series analysis model established on lunar time in this section. After differencing, the Clementine choose automatically ARIMA(1,0,2) for average price and ARIMA(2,0,7) for sales amount on solar time, and choose ARIMA (0,1,2) for average price and ARIMA(1,0,7) for sales amount on lunar time. The prediction results are shown in Fig.2 and Fig.3. From the figures, we can see intuitively that the ARIMA model can fit the data well.

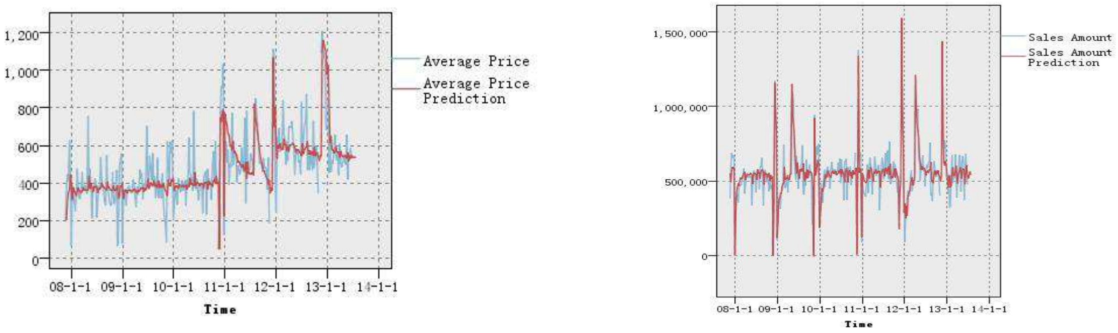

Fig.2. Prediction of Average Price and Sales Amount on lunar Calendar

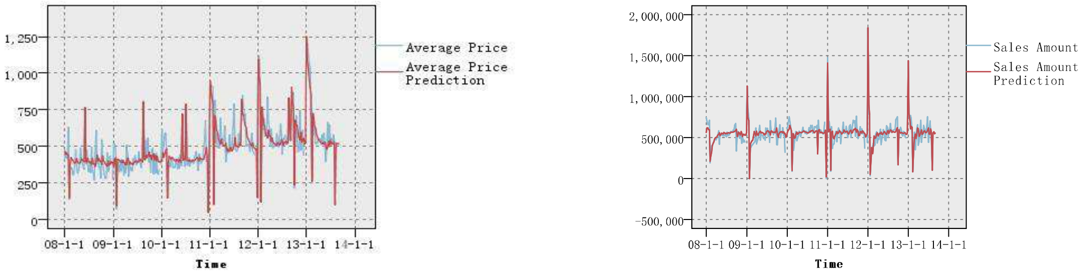

Fig.3. Prediction of Average Price and Sales Amount on Solar Calendar

To compare the predictive effect of weekly sales volume and average price on solar calendar and lunar calendar, the R2, RMSE, MAPE and MAE metrics are given in Table 1 and Table 2.

Table 1. Weekly prediction accuracy of sales volume on solar and lunar calendar

	on solar calendar	on lunar calendar
R2	0.8663	0.9307
RMSE	137055.1	176713.4
MAPE	4.945791	5.846084
MAE	96897.58	123969.8

Table 2. Weekly prediction accuracy of average price on solar and lunar calendar

	on solar calendar	on lunar calendar
R2	0.7760	0.9613
RMSE	420409.7	269789.5
MAPE	8.9541	5.1877
MAE	276129.8	176978.4

The MAPE of four models are all under 10, which illustrate the predictive accuracy on both lunar and solar calendar is relatively high. From the tables, we can see that the predictive accuracy based on lunar is better than on solar in general. Especially R2, which is a metric to measure the fitting degree of model, model on lunar calendar is much better than the one on solar calendar. And the RMSE, MAE and MAPE of sales volume on lunar calendar are lower than the solar one, which show the superiority of the model on lunar calendar. But the RMSE, MAE and MAPE of sales volume on lunar calendar are higher than the solar one. Through analyzing the data, we found that the lunar time is cannot be used directly in the model. For example, the month April and May in year 2012 is both convert to April, because April in lunar year 2012 is an intercalary month in the lunar calendar. Then the sales volume in April of year 2012 is summed by April and May in solar calendar, so it is extremely high. But the average price is not affected by the intercalary month in the lunar calendar, so all of the metrics is better than the solar time.

Based on the above discussion, the prediction on lunar time is used only when encounter lunar festival, and prediction on solar calendar is used for other data.

ANOMALY DETECTION

Having the predicted value, we can detect the anomaly sales data. If the observation value is larger than the upper bound of the Confidence interval or smaller than the lower bound of the Confidence interval, then the observation value is regarded as an anomaly value.

To verify the effectiveness of anomaly detection based on lunar calendar, we consider the entirety anomaly as an example in this paper, and then we used recent five years real sales records coming from a municipal cigarette commercial company. After data filtering and other checking, we get almost 300 weekly sales samples. The model of ARIMA with anomaly detection is constructed to forecast future 4 weeks' Cigarette sales.

Table 3. Anomaly detection of sales amount based on solar calendar and lunar calendar

Solar Time	Sales Amount	Lower Bound	Upper Bound	On Solar Is Anomaly	Lunar Time	Sales Amount	Lower Bound	Upper Bound	On Lunar Is Anomaly
2011/1/31	90688	-54686	256655	False	2010/12/28	90688	-76283	298071	False
2011/12/5	370437	397164	708408	true	2011/11/11	529406	319605	693959	False
2011/12/26	441873	390288	717429	False	2011/12/2	177875	-9302	365052	False
2012/1/2	1839542	1683080	1995985	False	2011/12/9	1593930	1396360	1791500	False
2012/1/9	792777	740106	1051469	False	2011/12/16	1025613	695827	1070181	False
2012/1/16	772801	627921	939173	False	2011/12/23	855134	513051	887405	False
2012/12/31	1437016	1272573	1599714	False	2012/11/19	1437016	1239446	1634586	False
2013/1/7	767790	639816	952721	False	2012/11/26	767790	645371	1019725	False
2013/1/21	770190	452694	763947	true	2012/12/10	770190	413228	787582	False
2013/1/28	736006	481797	794470	False	2012/12/17	736006	430908	805262	False

Table 4. Anomaly detection of average price based on solar calendar and lunar calendar

Solar Time	Average Price	Lower Bound	Upper Bound	On Solar Is Anomaly	Lunar Time	Average Price	Lower Bound	Upper Bound	On Lunar Is Anomaly
2011/1/31	125.241	-57.989	263.717	False	2010/12/28	125.241	-27.045	478.363	False
2011/12/5	443.631	328.52	649.994	False	2011/11/11	534.033	155.935	631.365	False
2011/12/26	158.442	-15.32	317.807	False	2011/12/2	188.367	176.592	651.965	False
2012/1/2	1077.618	934.827	1257.613	False	2011/12/9	613.698	121.894	602.671	True
2012/1/9	1118.744	816.979	1139.95	False	2011/12/16	1112.112	756.853	1378.42	False
2012/1/16	899.752	764.034	1086.596	False	2011/12/23	1087.952	419.481	983.902	True
2012/12/31	1096.064	1085.112	1418.239	False	2012/11/19	440.794	533.891	1070.332	True
2013/1/7	1207.174	925.113	1247.899	False	2012/11/26	1096.064	303.747	784.524	True
2013/1/21	1069.211	796.919	1119.481	False	2012/12/10	1207.174	784.13	1405.697	False
2013/1/28	1040.719	739.997	1062.328	False	2012/12/17	1069.211	869.45	1405.891	False

Table 3 and table 4 give the result if the sales amount and average price are anomaly on lunar and solar calendar respectively at the end of the Lunar New Year. From the table 3 and table 4, we can see that the anomaly detective result on lunar is not totally same as the result on solar calendar. For example, the result on solar time "2011/12/5" is "true" but on corresponding lunar time is "false", and we let the officer to check the result, and they also give the "false" as result.

CONCLUSION

ARIMA model consider trend variation, periodic variation and random disturbance comprehensively, and the model parameters can be quantified to remove people's subjective judgment to a certain extent, so it is widely used in prediction. But there is little literature to study the cigarette anomaly detection.

This paper used five years data to predict the sales amount and average price of the cigarette, and detect anomaly sales using the predicted values. This work can help officers to detect the anomaly and helping specifying cigarette market.

The conversion from solar time to lunar time is not much accurate, manual correction should be done to improve the predictive accuracy. The sales amount and average price can be affected by other factors, such as policy. The future work should include these factors.

Acknowledgments
The work presented in this article was supported in part by the research grant from National Science foundation of China (No.61201269), the Foundation of the Ministry of Education, Humanities and Social Sciences (No.10YJC630017), Shandong Doctor Foundation (No.BS2011DX026), Shandong Social Science Planning Fund Program (No.13CGLZ02), Shandong Provincial Natural Science Foundation, China (No.ZR2013GQ011), Shandong Province Higher Educational Science and Technology Program (No.J12LN33), Jinan High-tech Independent and Innovation (No.201303015) ZR2011FL020，Natural Science Foundation of Shandong Province, and Humanities and Social Sciences Project of Shandong Provincial Education Department (No.J13WG19).

REFERENCES

[1] B Luo, L Wan, WW Yan, JJ Yu, *American Journal of Operations Research*, **2012**, 2, 408-416
[2] Varun Chandola, Arindam Banerjee, and Vipin kumar, *Anomaly Detection: A Survey, ACM Computing Surveys*, Vol. 41, No. 3, **2009.**
[3] TAN, P.-N., STEINBACH, M., AND KUMAR, V. **2005.** *Introduction to Data Mining. Addison-Wesley.*
[4] John A. Quinn, Masashi Sugiyama, *Pattern Recognition Letters*, 40 (**2014**) pp.36–40.
[5] *An Outlier Detection Method Based on Clustering, Proceedings of the 2011 Second International Conference on Emerging Applications of Information Technology*, pp:253- 256, **2011**
[6] Grubbs, F. **1969.** *Technometrics* 11, 1, 1-21.
[7] Chow, C. and Yeung, D.-Y. **2002.** *Parzen-window network intrusion detectors. In Proceedings of the 16th International Conference on Pattern Recognition. Vol. 4. IEEE Computer Society, Washington, DC, USA*, 40385.
[8] Eskin, E. **2000**. *Anomaly detection over noisy data using learned probability distributions. In Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc*., 255-262.
[9] Lan LV, Yuesong TONG, Hongmei ZHANG, *Application of Seasonal Decomposition on Short-Term Forecast of Cigarette Sales in Xifeng,* **2012** *International Conference on Engineering and Business Management*, pp 792-796.
[10] Tsay, R. S., Pea, D., and Pankratz, A. E. **2000.** *Outliers in multivariate time series. Biometrika* 87, 4, 789-804.