Management Science and Engineering 管理科学与工程, 2020, 9(4), 266-276
Published Online December 2020 in Hans. http://www.hanspub.org/journal/mse
https://doi.org/10.12677/mse.2020.94035

Hans汉斯

时间序列长度对基于Prophet的自动售货机销量预测影响研究

刘　晴，董平军

东华大学旭日工商管理学院，上海
Email: liuqing_950820@163.com

收稿日期：2020年11月29日；录用日期：2020年12月24日；发布日期：2020年12月31日

摘　要

本次采用prophet模型对自动售卖机进行不同时间序列长度的销量预测，对预测效果进行对比分析，同时考虑外部天气因素对销量的影响。并且对自动售卖机的异常值采用基于离差系数和小波分析相结合的方法进行识别处理。实验结果表明，随着时间序列长度的增加，总体预测的精度逐渐提高；对于单机预测，需要具体问题具体分析，不同机器对于时间序列长度的敏感性不同。这为企业应对市场需求变化提供了重要的科学依据。

关键词

Prophet，时间序列长度，离差系数，小波分解与重构

Research on the Influence of Time Series Length on Prophet-Based Vending Machine Sales Forecast

Qing Liu, Pingjun Dong

Xuri School of Business Administration, Donghua University, Shanghai
Email: liuqing_950820@163.com

Received: Nov. 29th, 2020; accepted: Dec. 24th, 2020; published: Dec. 31st, 2020

Abstract

This time, the prophet model is used to forecast the sales of vending machines with different time

文章引用: 刘晴, 董平军. 时间序列长度对基于 Prophet 的自动售货机销量预测影响研究[J]. 管理科学与工程, 2020, 9(4): 266-276. DOI: 10.12677/mse.2020.94035

series lengths, and the forecasting effects are compared and analyzed, and the influence of external weather factors on sales is considered. And the abnormal value of the vending machine is identified by a method based on the combination of dispersion coefficient and wavelet analysis. Experimental results show that as the length of the time series increases, the accuracy of the overall forecast gradually improves; for single-machine forecasting, specific problems need to be analyzed in detail, and different machines have different sensitivity to the length of the time series. This provides an important scientific basis for companies to respond to changes in market demand.

Keywords

Prophet, Time Series Length, Dispersion Coefficient, Wavelet Decomposition and Reconstruction

1. 引言

商业预测是商业智能研究的重要问题之一。对于现代成熟企业而言，准确的预测商品的销售趋势能够很好的把握市场需求，为企业提供库存及营销方案的参考。随着新零售的发展，自动售卖机成为目前饱受关注的新领域。时间序列分析是自动售卖机销量预测的重要工具，在实际应用中，由于数据获取的限制，只能获得有限长度的历史数据，在这种情况下，能否建立可靠的预测模型还有待探讨。如何解决上述问题成为目前亟待解决的需求。

在自动售卖机的销量预测上，目前已经有了一定的研究。Tomas [1]提出了监控系统获取产品的销售信息，通过不断对销售不佳的产品进行替换达到销售方案的优化。Feng-Cheng Lin [2]等人将产品根据起销售业绩进行聚类，并通过决策树提取受欢迎产品的关键属性，最后用贝叶斯网络过滤销售产品。Hidetaka Sakai [3]等人通过组合逻辑和多元回归预测模型来动态修正售货机销量预测的精度，已介于售货机制冷所消耗的能量。洪鹏，余世明[4]提出通过 ARMA 模型预测售货机受制约销量来补偿实际销量，再通过 RBF 神经网络进行销量预测。孙娜[5]等人通过考虑不同地理位置的影响，提出了基于灰色模型的销量预测研究。王庆阳[6]通过考虑间断性预测以及连续性预测对自动售卖机销量进行组合预测。

以上研究大部分是基于所有能获得的历史销售数据进行预测，而没有考虑影响自动售货机销量的外在因素以及不同时间序列长度对预测结果的影响，对于饮品自动售卖机而言，商品的销量往往受到强烈的温度、节假日、等因素影响，销量数据波动性较强。并且由于机器大小、故障等因素导致自动售卖机的销量会存在部分异常值。因此本文将利用某企业的销量时间序列数据以及获取到的外部环境数据，采用离差系数和小波变化结合的方法对异常值进行处理[7]，并且基于不同时间序列长度采用 prophet [8]模型对自动售卖机进行单机预测以及总体预测，并将不同时序长度的预测结果进行对比评估。

2. 相关理论

2.1. 离差系数

离差系数能够反映样本的离散程度，为了体现距离中位数较大数据与较小数据的离散程度，本文提出上离差系数与下离差系数：

$$s_j = (-1)^{j-1}\sqrt{\frac{\sum\left(x_m - x_{ji}\right)^2}{n_j}}$$

$$cv_j = \frac{s_j}{x_m}$$

$$j = 1,2$$

其中 x_m 表示中位数，x_{1i} 表示大于中位数的数据，x_{2i} 表示小于中位数的数据，cv_1 表示上离差系数，n_1 表示 x_{1i} 的个数。

1) 阈值确定

根据上述公式计算 cv_1 以及 cv_2，设定安全系数 k_1、k_2，得到上下阈值 $k_1 * cv_1$，$l_2 * cv_2$。

2) 异常值识别

$$b_i = \left(x_i - x_m\right)/x_m$$

其中 x_i 表示原始序列的值，x_m 表示中位数，当 b_i 不在上下阈值范围内则为异常值。

3) 异常值处理

采用所在周(weekday)的饮品销量的平均值进行替换。

2.2. 小波变换

对于销量数据，其时间序列常具有周期性以及趋势性，只采用上下离散系数阈值方法进行处理只能识别极端异常值，而采用小波变换相结合得方法能够挖掘出时序数据的周期性以及趋势性。

对任意的 $x(t) \in L_2(R)$，称：

$$WT_x(\alpha,\tau) = \frac{1}{\sqrt{a}}\int x(t)\varphi\left(\frac{t-\tau}{a}\right)\mathrm{d}t$$

为一维信号 $x(t)$依赖于参数 a 和 τ 的一维连续小波变换。

小波逆变换：

$$X(t) = \frac{1}{c_\varphi}\int_0^{+\infty}\frac{\mathrm{d}a}{a^2}\int_0^{+\infty}WT_x(a,\tau)\frac{1}{\sqrt{a}}\varphi\left(\frac{t-\tau}{a}\right)\mathrm{d}\tau$$

经过小波变换和逆变换，可实现小波的分解与重构，从而达到去噪的目的。

1) 采用上述公式得到去噪后的时间序列

2) 阈值确定

采用局部计算上离差系数与下离差系数，利用左右相邻的 15 条数据进行离差系数的计算，当左右相邻数据一方不足 15 条时，则选取全部数据。中位数用重构后的序列替代。设定安全系数 k_1、k_2。

3) 异常值识别

$$c_i = \left(x_i - y_i\right)/y_i$$

其中 x_i 表示原始序列的值，y_i 表示重构后序列的值，当 c_i 不在上下阈值范围内则为异常值。

4) 异常值处理

采用小波分解与重构后得到的新时间序列对应的值替换。

2.3. Prophet 模型

Prophet 采用的也是时间序列趋势分解的模型，得到三个部分：trend (趋势项)，seasonality (周期性,

包括年周期，月周期，周周期以及日周期等)，holidays (节假日影响因子)，可以用以下式子表示：

$$y(t)=g(t)+s(t)+h(t)+\epsilon(t)$$

其中 $g(t),s(t),h(t)$ 分别表示趋势项、周期项、节假日项，最后一项表示整个模型中的不适应的残差部分。Prophet 也是对模型的三个部分分别建模，最后组合生成预测数据。

1) 趋势成分

趋势部分的实现主要应用两种模型，一种饱和增长模型，一种分段线性模型；

饱和增长模型公式：

$$g(t)=\frac{c(t)}{1+\exp\left(-\left(k+a(t)^{\mathrm{T}}\delta\right)\right)\left(t-\left(m+a(t)^{\mathrm{T}}\gamma\right)\right)}$$

分段线性模型公式：

$$g(t)=\left(k+a(t)^{\mathrm{T}}\delta\right)t+\left(m+a(t)^{\mathrm{T}}\gamma\right)$$

其中 k 为增长率，δ 为增长率的变化量，m 是偏置参数，在该模型中均为随时间 t 变化的函数。

2) 周期部分

Prophet 对周期性因子部分的处理采用的是傅里叶级数，并且可以手动调整傅里叶级数的阶数 N。

$$s(t)=\sum_{1}^{N}\left(a_n\cos\left(\frac{2\pi nt}{P}\right)+b_n\sin\left(\frac{2\pi nt}{P}\right)\right)$$

其中 P 表示时间序列的周期，当周期为年时，P 的值一般为 365.25，当周期为周时，P 值一般为 7。

3) 节假日部分

节假日部分采用自回归矩阵进行计算：

$$z(t)=\left[1(t\in D_1),\cdots,1(t\in D_L)\right]$$

$$h(t)=z(t)\kappa$$

$$\kappa\sim nomal\left(0,v^2\right)$$

其中 i 表示每一个节日，D_i 为每年这个节日的日期。V 表示节假日灵活度参数，其值越大，表示节假日对模型的影响越大。

4) 额外回归量

对于额外回归量，可以使用 Add_regressor()的方法将其添加到模型的线性部分。具有回归量的值的列需要添加到拟合和预测数据框中。Add_regressor()可以将另一个时间序列作为回归量，但是预测的值在另一个时间序列中必须是已知的，如果未知，可以先对该部分时间序列进行预测。

3. 数据准备

3.1. 实验数据

本文主要以国内某公司自动售卖机的销售数据为研究对象，涉及销售设备 154 台，覆盖该城市大部分区域，时间主要为 201801~201812 销售数据，该数据集覆盖地区较广，设备数目较多，数据量相对完整。原始数据集包含主要数据信息如表 1 所示。

由于自动售卖机的销量受到许多因素的影响，本次考虑的外部因素主要时气象数据，节假日等，因此本文爬取分析了可获得的外部环境数据，主要数据信息如表 2 所示。

Table 1. Main fields of vending machine sales data
表 1. 自动售货机销售数据主要字段

序号	字段	数据描述
1	付款时间	例：2019/9/17 19:44
2	商品 id	例：2914
3	商品类别	例：饮料，水等
4	设备 id	例：73c5fd40940e4e0eb25625a1092d1717
5	实际价格	例：5.5
6	经纬度	例：31.091778,121.520671
7	地铁线	例：1 号线，2 号线

Table 2. Main fields of external data of vending machine
表 2. 自动售货机外部数据主要字段

序号	字段	数据描述
1	节假日	例：休息日，节假日，工作日
2	气象数据	例：温度，降水量，湿度等

3.2. 数据预处理

由于目前该企业正处于上升阶段，自动贩卖机设备的数量处于不断上升的过程，所产生的数据由于各种原因会存在部分、整体缺失或者不合理的情况，无法直接用于分析与预测研究，因此需要对数据进行预处理操作，主要包含以下步骤：

1) 数据归约：

由于销量数据太过庞大，并且每条销量数据记录之间的间隔并不确定，因此需要对销量数据进行聚合处理。提取出销售数据中的关键信息，按照合适的时间粒度对初始数据进行整理，完成数据归约操作，得到售卖机销量数据表，本次选取的预测时间粒度为一天，反映了一天内用户购买的饮品数量。

2) 数据清洗：

对于接入的数据，很多都包含了缺失、异常或者类型不一致的字段数据，此时需要对数据进行数据清洗，数据清洗就是发现这些类型的数据并纠正这些类型的数据，其中包含数据一致性、无效值以及缺失值的处理，处理方法有补全、删除等。本次由于自动售卖机机器损坏或数据为接入等原因导致销量存在较多的缺失值与部分异常值；对于该部分缺失值与异常值军进行异常处理，采用离散系数和小波分析相结合的方式进行识别，识别后采用对应方法处理。

3.3. 数据分析

依据主观分析，温度对饮品的销量会产生较大的影响，此次采用散点图的形式判断温度与销量之间是否存在相关关系。

图 1 为 2018 年销量与温度对应的关系图表，其中纵坐标表示不同温度对应的平均日销量，从图中可以发现饮品的销量与温度之间存在明显的相关关系，尤其是当温度越高时，相关关系越明显。因此可以将温度作为销量预测的一个影响变量。

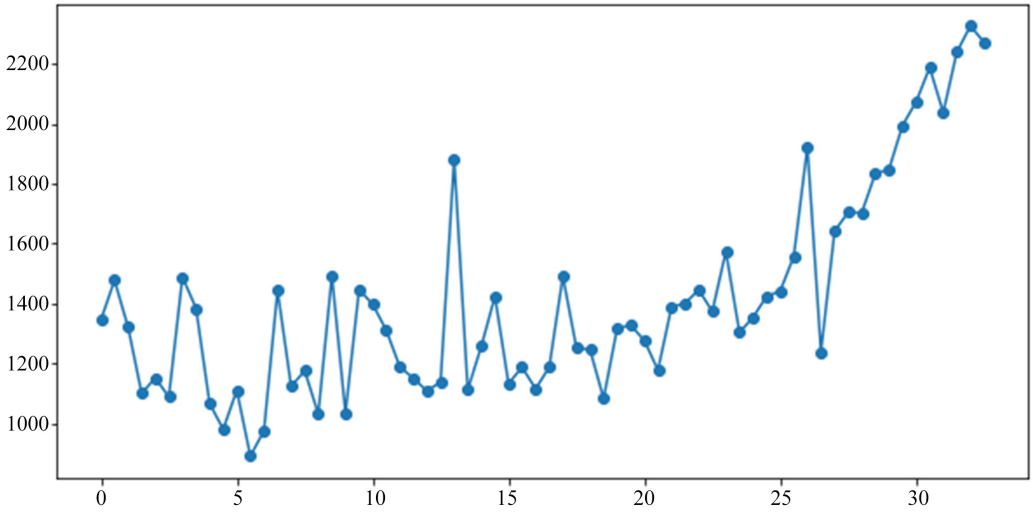

Figure 1. Relationship between sales and temperature
图 1. 销量与温度关系图

4. 实验分析

由于机器故障、机器容量大小等原因，部分机器缺失值过多，因此本次选取缺失值较少的 47 台机器进行实验分析。分别采用时间长度为一年，半年，四个月，两个月，一个月的销量序列作为基础时间序列，采用 prophet 模型进行未来 7 天的单机以及总体销量预测，同时将温度因素作为变量之一。

为了更好的体现模型的优越性和实用性，本文采用平均百分比误差 MAPE 作为模型评估指标，定义如下：

$$\mathrm{MAPE} = \sum_{1}^{N}\left|\frac{y_i - \widetilde{y_i}}{y_i}\right| * \frac{100}{N}$$

4.1. 异常值识别

以 10 号站台 01 号机为例进行分析，如图 2。

1) 上下离散系数初步检测：

图 3 中超出红线阈值部分为异常值，图 4 为采用周均值替代异常值之后的销量时序图。

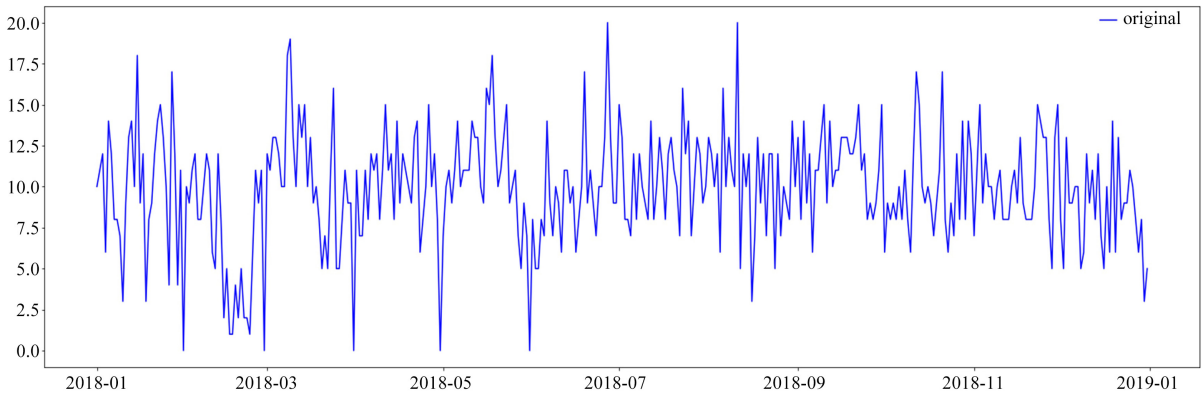

Figure 2. Time sequence diagram of sales volume of station No. 01 on platform 10
图 2. 10 号站台 01 号机销量时序图

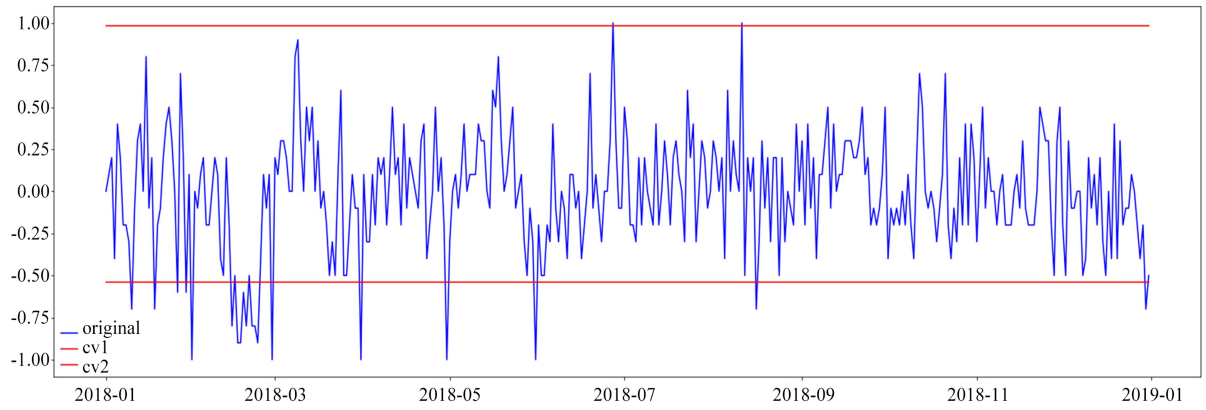

Figure 3. Initial detection of upper and lower thresholds
图 3. 上下阈值初步检测

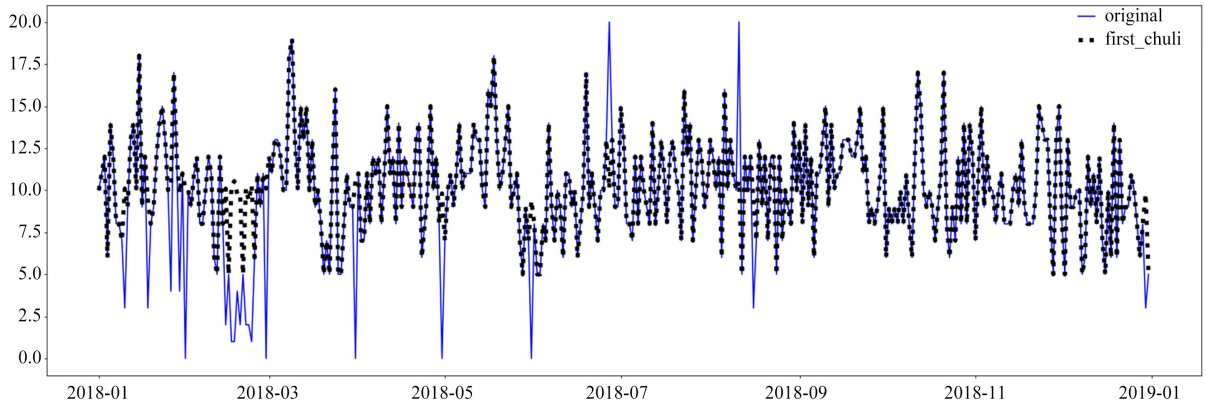

Figure 4. Sequence diagram after initial adjustment
图 4. 初步调整之后序列图

2) 小波分析再次检测

图 5 为基于初步异常值检测及调整后的时间序列进行小波变换后得到的去噪时间序列；图 6 为计算得到的上下阈值，超出阈值曲线部分为异常值；图 7 中黑线为初步调整之后的时序图，红色点线为最终调整异常值之后得到时序图。

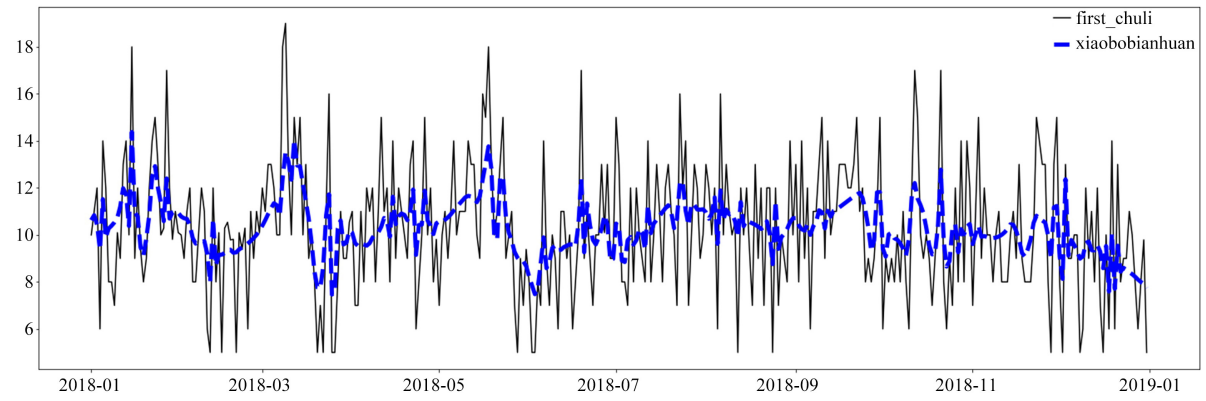

Figure 5. Sequence diagram after wavelet transform
图 5. 小波变换后时序图

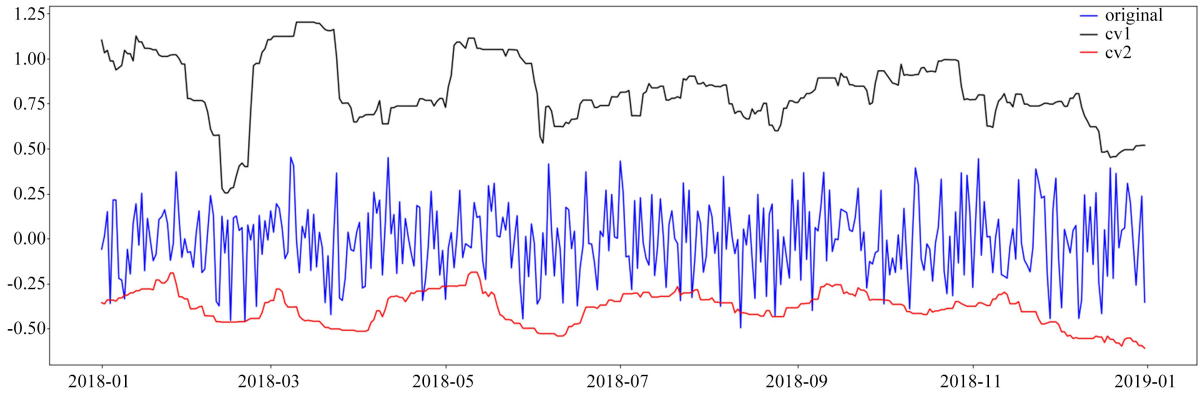

Figure 6. The upper and lower thresholds are tested again
图 6. 上下阈值再次检测

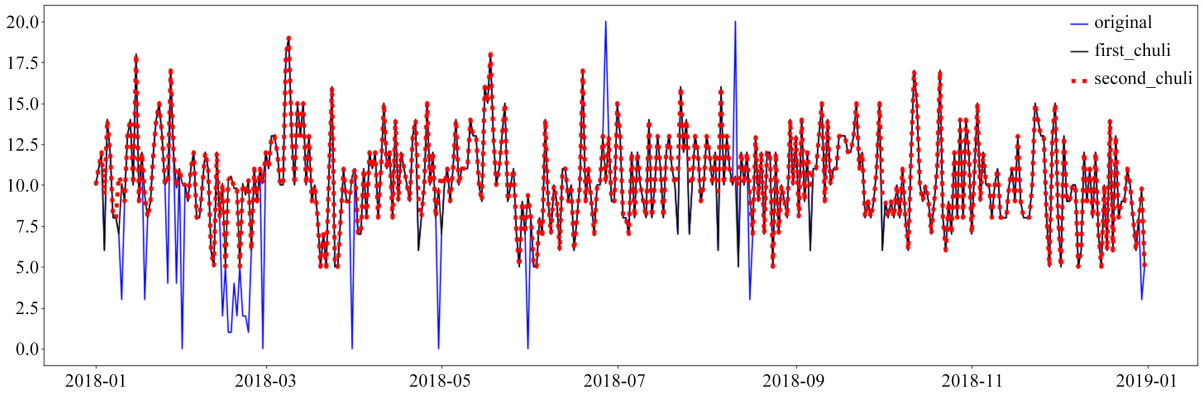

Figure 7. Timing chart after adjusting again
图 7. 再次调整之后时序图

4.2. 异常值调整前后时间序列预测评估

对 47 台机器进行异常值调整前后 prophet 预测结果进行指标评估，时间序列长度采取一年。

图 8 中橙色线图为异常值调整之前对 47 台机器预测的 MAPE 指标评估结果，蓝色波点线图为异常值调整之后对 47 台机器预测的 MAPE 指标评估结果，从图中可以发现异常值调整之后大部分机器的预测结果评估明显优于异常值调整之前。并且不同机器的评估指标值相较于之前较为平稳。

将 47 台机器的单台预测结果相加，得到 47 台机器总的预测结果。异常值调整前 47 台机器总预测结果 MAPE 评估指标为 5.48%，调整之后 MAPE 为 4.58%。

综上，不管是单机预测还是总体预测，异常值调整之后的指标评估结果均优于异常值调整之前，说明异常值处理结果合适。

4.3. 不同时间序列长度预测评估

1) 单机预测结果

对 47 台机器分别采用不同时间序列长度的 prophet 预测。MAPE 评估指标结果如下表 3。

从表 3 中我们可以发现对于不同的机器，最优的时间序列长度都不一样，因此对于单台机器需要具体分析，通过计算 47 台机器的平均 MAPE 指标值，一年、半年、四个月、两个月、一个月分别平均 MAPE 值为 20.7%, 19.68%, 19.31%, 21.47%, 23.52%，并且对单机不同时间序列长度对应 MAPE 值比较结果进行

综合分析，当时间序列长度为 4 个月或者 6 个月时，预测效果较好。

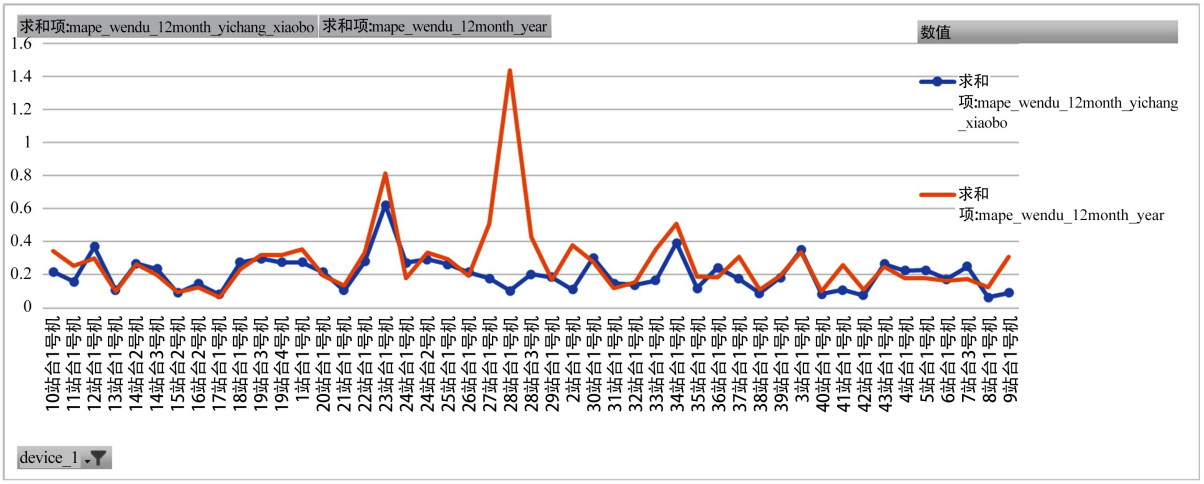

Figure 8. Time series prediction MAPE evaluation results before and after outlier adjustment
图 8. 异常值调整前后时间序列预测 MAPE 评估结果

Table 3. Evaluate the MAPE value of the prediction results of a single machine under different time series lengths
表 3. 不同时间序列长度下单台机器预测结果评估 MAPE 值

设备	mape_12month	mape_6month	mape_4month	mape_2month	mape_1month	mape_12month_cvbg
1 站台 1 号机	27.60%	27.66%	31.70%	33.39%	33.87%	28.28%
2 站台 1 号机	11.21%	9.68%	8.75%	11.78%	23.02%	16.55%
3 站台 1 号机	34.97%	28.16%	28.84%	31.27%	42.82%	35.33%
4 站台 1 号机	22.47%	24.58%	22.56%	18.83%	22.92%	22.39%
5 站台 1 号机	22.75%	13.76%	12.24%	12.97%	18.18%	21.73%
6 站台 1 号机	17.25%	18.91%	19.24%	20.33%	23.05%	17.26%
7 站台 3 号机	25.01%	22.41%	26.09%	27.23%	21.80%	24.08%
8 站台 1 号机	6.28%	5.78%	6.82%	9.88%	9.52%	13.15%
9 站台 1 号机	9.08%	10.20%	10.35%	15.81%	17.00%	8.89%
10 站台 1 号机	21.89%	26.67%	25.28%	29.09%	40.32%	21.93%
11 站台 1 号机	15.85%	18.55%	21.68%	23.78%	31.34%	15.85%
12 站台 1 号机	37.08%	26.94%	27.66%	30.15%	30.45%	36.98%
13 站台 1 号机	10.72%	9.58%	10.62%	12.63%	16.09%	10.39%
14 站台 2 号机	26.82%	27.98%	24.11%	32.68%	39.72%	26.21%
14 站台 3 号机	23.51%	29.61%	24.28%	19.65%	22.60%	23.58%
15 站台 2 号机	9.06%	9.95%	9.22%	12.69%	12.25%	8.90%
16 站台 2 号机	14.64%	14.81%	13.86%	11.50%	9.69%	14.90%
17 站台 1 号机	8.15%	2.74%	7.55%	10.74%	9.28%	6.88%
18 站台 1 号机	27.47%	25.00%	26.54%	30.85%	32.06%	27.60%
19 站台 3 号机	29.94%	27.21%	36.72%	43.44%	32.61%	29.68%
20 站台 1 号机	21.65%	25.37%	19.37%	18.86%	19.66%	21.06%

Continued

21 站台 1 号机	10.67%	13.63%	12.92%	15.13%	20.12%	11.01%
22 站台 1 号机	28.11%	21.92%	19.30%	21.46%	26.73%	29.94%
23 站台 1 号机	62.33%	39.88%	35.53%	47.78%	50.46%	56.99%
24 站台 2 号机	29.23%	31.60%	29.64%	33.64%	36.26%	29.84%
19 站台 4 号机	27.65%	30.74%	32.19%	35.73%	32.40%	30.57%
25 站台 1 号机	26.39%	17.71%	16.60%	16.43%	20.65%	34.57%
26 站台 1 号机	21.49%	29.03%	28.76%	29.07%	29.20%	20.26%
27 站台 1 号机	17.76%	33.51%	24.79%	26.36%	26.23%	16.18%
28 站台 3 号机	20.35%	26.68%	19.47%	20.30%	19.54%	21.64%
29 站台 1 号机	18.47%	19.23%	17.73%	16.16%	15.99%	14.93%
30 站台 1 号机	30.31%	20.74%	17.51%	22.15%	20.41%	29.72%
31 站台 1 号机	14.96%	8.16%	8.15%	14.81%	17.90%	13.23%
32 站台 1 号机	13.79%	15.10%	17.36%	17.97%	20.50%	13.79%
33 站台 1 号机	16.62%	16.39%	18.98%	19.39%	16.68%	16.16%
34 站台 1 号机	39.37%	43.38%	41.02%	50.18%	51.33%	42.64%
35 站台 1 号机	11.85%	6.54%	7.39%	8.05%	11.23%	21.14%
36 站台 1 号机	24.43%	19.71%	17.91%	19.77%	26.12%	26.99%
28 站台 1 号机	10.31%	12.30%	10.28%	10.82%	14.79%	19.95%
37 站台 1 号机	17.58%	16.97%	18.28%	24.59%	16.29%	29.54%
38 站台 1 号机	8.91%	9.03%	10.02%	9.85%	13.04%	7.98%
24 站台 1 号机	27.21%	19.66%	24.25%	24.37%	21.47%	23.28%
39 站台 1 号机	18.30%	14.44%	14.71%	16.18%	14.14%	17.75%
40 站台 1 号机	8.43%	8.71%	9.14%	9.18%	29.01%	8.92%
41 站台 1 号机	10.92%	11.86%	11.05%	10.89%	11.16%	22.61%
42 站台 1 号机	7.50%	7.26%	8.20%	9.74%	10.32%	10.85%
43 站台 1 号机	26.62%	25.33%	23.01%	21.38%	25.17%	28.32%

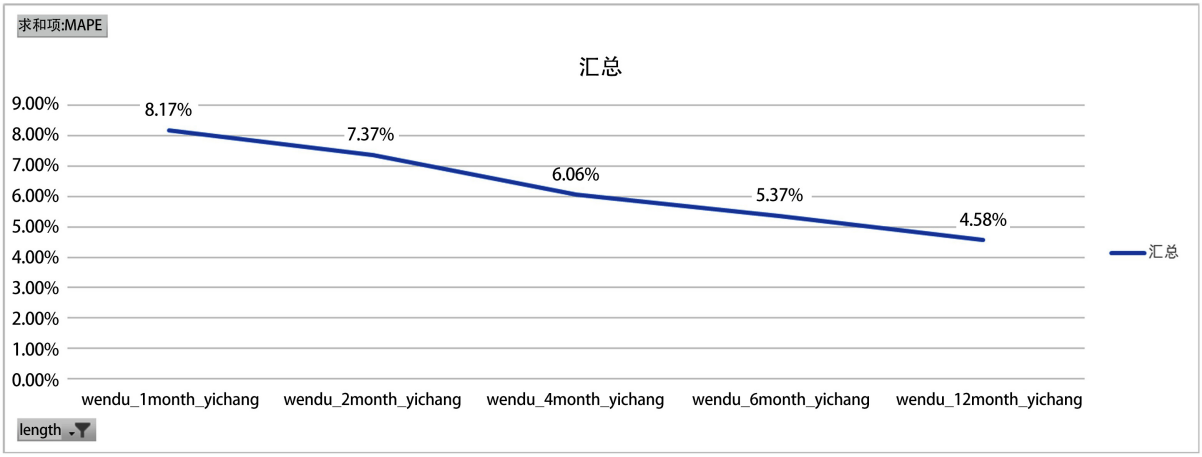

Figure 9. Cumulative forecast result MAPE indicator evaluation chart
图 9. 累计预测结果 MAPE 指标评估图

2) 累计相加总体预测结果

对47台机器的预测结果进行累计相加，得到总体预测值。预测结果评估MAPE值如图9。

图9为不同时间序列长度对应的MAPE指标评估结果值，可以发现当时间序列长度为一年时，预测效果最好，因此对于总体预测，时间序列长度采取一年。

5. 结论

对于时间序列的预测而言，不同时间长度的销量序列与不同粒度的预测分析对预测结果都会产生一定的影响。对于本次的自动售卖机的销量预测而言，采用时间长度为一年的销量序列作为基础时间序列进行时，进行总体预测的预测精度是最优的，但是对于单机预测而言，需要单台机器进行具体分析，对于不同的机器可以适当的调整基础时间序列的长度来进行短期预测，但就单台机器的整体预测效果而言，当时间序列长度为4个月或者6个月时，预测精度最优。

参考文献

[1] Murray, T. and Jansson, E.R. (2010) Methods and System for Managing Vending Operations Based on Wireless Data. US Patent No. 2010031261A1.

[2] Lin, F.-C., Yu, H.-W., Hsu, C.-H., *et al.* (2011) Recommendation System for Localized Products in Vending Machines. *Expert Systems with Applications*, **38**, 9129-9138. https://doi.org/10.1016/j.eswa.2011.01.051

[3] Sakai, H., Nakajima, H., Higashihara, M., *et al.* (1999) Development of a Fuzzy Sales Forecasting System for Vending Machines. *Computers & Industrial Engineering*, **36**, 427-449. https://doi.org/10.1016/S0360-8352(99)00141-2

[4] 洪鹏, 余世明. 基于时间序列分析的自动售货机销量预测[J]. 计算机科学, 2015, 42(S1): 122-124.

[5] 孙娜, 潘振华, 于金秀. 基于灰色预测模型的自动售货机商品销售量研究[J]. 商场现代化, 2020(2): 6-7.

[6] 王庆阳. 关于自动售货机间断需求预测方法的研究[D]: [硕士学位论文]. 南京: 南京大学, 2019.

[7] 张华, 李佳, 万毅. 基于离差系数和小波分析的取水量异常值研究[J]. 水利水电技术, 2020, 51(10): 35-40.

[8] 赖慧慧. 基于时间序列Prophet模型的乘用车消费税预测[J]. 税收经济研究, 2020, 25(1): 34-39.