

文章编号:1005-9679(2017)06-0025-04

银行 ATM 设备业务总量的时序特征分析及预测

韦金香, 张建同

(同济大学 经济与管理学院, 上海 200092)

摘要: 本文旨在分析银行 ATM 设备业务总量的时序特征, 并据此对其进行预测。首先, 本文以十分钟为间隔, 绘制了银行 ATM 设备业务总量的 30 天趋势图, 发现其以日为单位, 呈现出显著的周期性、扰动性和多峰性, 因此本文建立了以日为周期的 ATM 设备业务总量时序分布模型。在求解模型的过程中, 本文利用模拟退火算法将每日银行系统 ATM 设备业务总量按其特征分为八段, 消除了业务总量时间序列的多峰性。在此基础上, 建立了 Holt-Winters 模型对业务总量进行预测, 最后用第一时段进行验证, 得到 95% 置信区间内的预测值。本文的研究结果为银行数据监控中心判断设备运行状态提供了依据。

关键词: 时序分布模型; 模拟退火; Holt-Winters 模型; 运行状态

中图分类号: F 832 **文献标志码:** A

Research and Prediction on Time-Sequence Characteristics of the Total Banking Automatic Teller Machine Business

WEI Jinxiang, ZHANG Jiantong

(School of Economics and Management, Tongji University, Shanghai 200092, China)

Abstract: This paper aims to analyze the time-sequence characteristics of the total banking ATM equipment business and predict the number of the business. Firstly, we map the banking ATM equipment 30 days of total trend diagram by taking the data of ten minutes interval. It is found that it has a significant periodicity, perturbation and multi-peak. Then we set up a time-sequence distribution model of total ATM equipment business base on the data of day interval and solve the mode by simulated annealing algorithm (SA). We divided the total amount of ATM equipment of the daily banking system into eight segments according to its characteristics, and eliminating the multi-peak of the total time series of the total business. Finally we forecast the total volume of the business through the Holt-Winters model and use the first period data for verification. And we get the 95% confidence interval of the first period business. The results of this paper provide a basis for judging the operation status of the equipment in the bank data monitoring center.

Key words: time-sequence distribution model; SA; Holt-Winters model; operation status

关于时间序列应用问题, 多用于以下方面: (1) 降维, 寻找序列特点; (2) 聚类, 在聚类结果上进行分析; (3) 预测。银行 ATM 设备业务总量数据具有时间序列特点, 序列形状复杂但具有规律, 本文运用时

间序列的思考方式对银行 ATM 设备业务总量的特征进行分析, 并预测其正常状态下的数值, 从而对各行自助设备运行情况判定。

收稿日期: 2017-10-12

作者简介: 韦金香(1993—), 女, 广西人, 硕士研究生, 研究方向为数据挖掘; 张建同(1966—), 教授, 博士生导师, 研究方向为应用统计。

E-mail: zhangjiant@163.com。

1 ATM 设备业务总量的时序建模

1.1 样本数据基本特征

因为银行 ATM 设备业务量是时间序列,因此,首先应判断该序列是否能够进行预测,才能够进一步通过建立模型判定设备运行状态。因此本文建立业务量的时间序列自回归方程,利用 MATLAB 进行平稳性检验,得到其显著水平为 0.1%,因此通过 5% 的显著性水平检验,判定业务量时间序列符合平稳性的特征。

随后,利用 R 软件初步画出样本数据的趋势图,由此判断业务量时间序列的趋势性和周期性,如图 1 所示。

其中第一行为原业务量图,第二行为幅值表征趋势变化图,第三行为周期性,第四行为随机扰动。从图 1 中可以清晰看到数据以天为周期,呈现明显的周期性和趋势性,扰动分布规律不明显。

确定 ATM 设备业务量以天为周期,本文进一步画出以天为单位周期的图像,如图 2 所示。

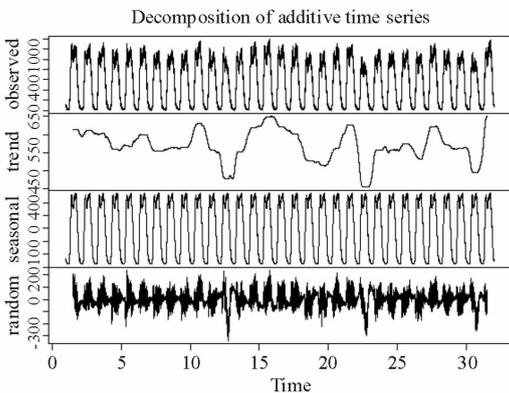


图 1 ATM 设备业务量趋势图

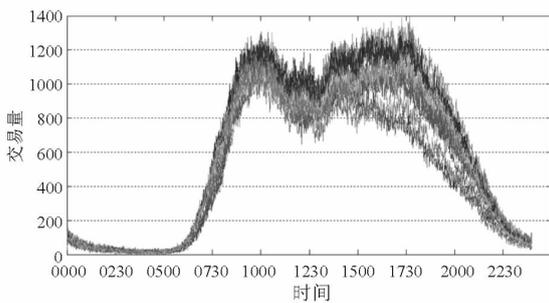


图 2 ATM 设备每天变化趋势图

从图 2 中可以看出,ATM 设备业务量在时序上呈现非均质特征,业务量所构成的时间序列具有多峰值和周期性。因此,需针对业务总量的时序特征,阶段性、周期性地建立准确的 ATM 设备业务访问预测模型。

1.2 ATM 设备业务总量的时序分布模型

上文提及 ATM 设备业务总量具有以天为周期

的性质,但一天内业务总量具有多峰性,因此应当进一步将其以天为周期,探究其阶段性特征。而分行上传的每个时刻的业务量是一个根据时间变化的源源不断的随机过程,所以在对其进行阶段性研究时应当将业务量作为一个时间序列进行分析,考虑业务量变化及正常范围时不应该将其与时间分离。对于 ATM 应用系统,将一个周期 T (天)内的时间分为 n 等长时段 $L(s, t) = \{S(t_1), S(t_2), \dots, S(t_n)\}$,其中 S, t 分别为分行业务量的业务量因子和时间因子。

本文研究其在不同时间段的特征,就是在研究按序(时间排列)将全部样品(业务量)截成几段,同类样品(业务量)次序互相衔接的问题,即可以归纳为有序分类问题,符合 Fisher 最优分割思想。

本文研究需要从 ATM 应用系统业务量时间序列 $L(s, t) = \{S(t_1), S(t_2), \dots, S(t_n)\}$ 中发现分行 ATM 应用系统业务量的模式 $T = \{T_1, T_2, \dots, T_k\}$,其中 $T_i (1 \leq i \leq k)$ 在时间上相连接。其中,由于以天作为周期,而不同日期具有不同特征,所以本文将样本数据研究天数(共 p 天)作为指标,记为 $S(t_m) = (d_{m1}, \dots, d_{mp}), 1 \leq m \leq n$ 。

得到一天 ATM 应用系统业务交易量的时间区间特征:

$$d(t) = \begin{cases} d_1, & 0 < t < T_1 \\ \vdots \\ d_k, & T_{k-1} < t < T_k \end{cases} \quad (1)$$

利用 Fisher 最优分割算法划分区间的模型构建是基于算法思想建立的,方法步骤如下:

(1)类的直径

假定分割区间为 K 类,某一类 i 业务量区间模式包含时序 $\{S(i), S(i+1), \dots, S(j)\} (j > i)$,记为 $G = \{i, i+1, \dots, j\}$ 。那么该业务量模式的向量均值为:

$$\tilde{S}_G = \frac{1}{j-i+1} \sum_{h=i}^j S(t_h) \quad (2)$$

设模式内业务量的直径为业务量集的利差平方和 $D(i, j)$,则有:

$$D(i, j) = \sum_{h=i}^j (S(t_h) - \tilde{S}_G) * (S(t_h) - \tilde{S}_G)^T \quad (3)$$

(2)分类的损失函数

设 $b(n, k)$ 为将有序的 n 个 ATM 应用系统业务量向量被分割为 k 个访问模式的一种分类方法^[2],已知分为 k 类有 A_m^{k-1} 种情况, $b(n, k)$ 可表达为:

$$\begin{cases} G_1 = \{i_1, i_1 + 1, \dots, i_2 - 1\} \\ G_2 = \{i_2, i_2 + 1, \dots, i_3 - 1\} \\ \vdots \\ G_k = \{i_k, i_k + 1, \dots, n\} \end{cases} \quad (4)$$

式中分点 $1 = i_1 < i_2 < \dots < i_k < n = i_k - 1 (i_{k+1} =$

$m+1$); $b(n, k)$ 损失函数为:

$$\Gamma(b(n, k)) = \sum_{e=i}^k D(i_e, i_e - 1) \quad (5)$$

当周期内均分段数 n 和分割模式数 k 固定, 则 $\Gamma(b(n, k))$ 越小, 即各类模式离差平方和越小, 说明该模式的分类是合理的。因此, 寻找 A_n^{k-1} 种分法中使得 $\Gamma(b(n, k))$ 最小的最优情况为 $p(n, k)$, 此时, 满足:

$$\begin{cases} \Gamma(p(n, 2)) = \min_{2 \leq j \leq n} \{D(1, j-1) + D(j, n)\} \\ \Gamma(p(n, 2)) = \min_{k \leq j \leq n} \{ \Gamma(p(j-1, k-1)) + D(j, n) \} \end{cases} \quad (6)$$

则通过求解式(6)得到 $p(n, k)$ 对应的式(4)解, 即可得到 Fisher 最优分割思想下的最优分割点。从而得到式(1)解。

2 基于 SA 算法的 ATM 设备业务总量的时序模型求解

由于 ATM 应用系统业务量时间序列周期内具有 144 个时间点, 笔者利用求解精确解的思路在计算 $k=4$ 及以上时, 无法计算得到结果。由于业务量以天为周期进行阶段性的分析, 对于阶段个数没有精确要求, 因此本文采用启发式算法进行求解。由于模拟退火算法(SA)数学模型的描述: 在给定邻域结构后, 模拟退火过程是从一个状态到另一状态不断随机游动, 因此选择 SA 算法进行求解。

求解过程中, 分割阶段数的确定依据 Fisher 原理: 在 Fisher 最优分割过程中, 最小损失函数值随分割 k 的增加而减少。而当分割数增加到某一数值后, 最小损失函数值曲线将急剧变缓, 达到一定的平衡, 此时的 k 值为最佳分割值。

训练实验中, 对训练样本中业务量数据进行 3~12 次最优分割, 利用 MATLAB 软件分别得到各次分割下 SA 算法求得的最小损失函数值, 如图 3 所示。

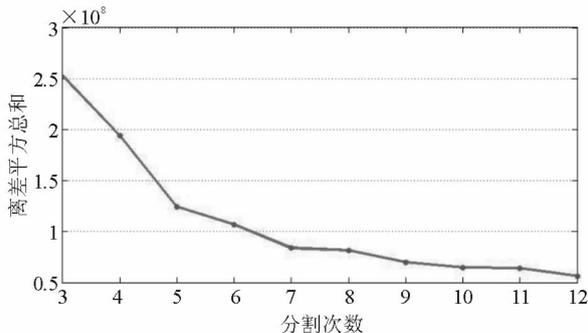


图 3 基于 SA 算法求解的最小损失函数值变化图

所以, 由图 3 可以得到 $k=7$ 时为最佳分割取值, 并且得到 $P(n, 7)$ 下的 ATM 应用系统业务量模式所在时间区间, 如表 1 所示。表 1 显示 k 值为 7

时各访问模式的离差平方和值较小。

表 1 各业务量模式所在时间区间的集合

T_k	时间段	交易量模式距离平方和
T_1	00:00 ~ 06:39	3.058 4e+06
T_1	06:40 ~ 07:29	4.765 7e+06
T_2	07:30 ~ 08:29	1.078 0e+07
T_4	08:30 ~ 11:59	1.531 8e+07
T_5	12:00 ~ 18:19	1.805 6e+07
T_6	18:20 ~ 19:29	2.620 7e+06
T_7	19:30 ~ 21:19	1.754 3e+07
T_8	21:20 ~ 23:59	2.042 3e+07

表 1 结果显示, 可以将一天内银行系统 ATM 设备业务总量分为八段, 在此基础上进一步预测银行 ATM 设备业务特征。

3 基于 Holtwinters 时间序列的业务量预测模型

上文将一天内的业务总量分为八段, 进一步观察分段之后业务总量具有的特点, 如图 4 所示。

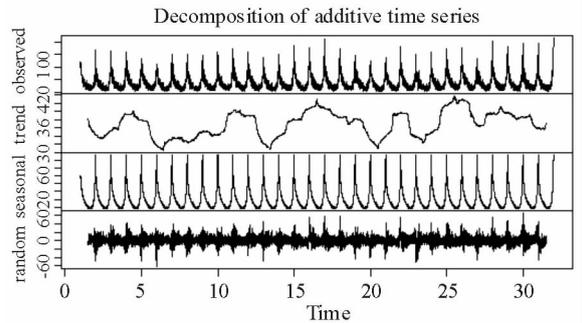


图 4 00:00~06:39 时段业务量月变化趋势

从趋势图中看到, 业务量变化具有很强周期性, 基于其具有的整体趋势变动性和周期性的二重变化特点, 采用加性 Holt-Winters 模型进行预测(HW 模型)。

沙川认为 HW 模型的基本思想是把具有线性趋势、周期变动和随机变动的时序进行分解研究, 与指数平滑法相结合, 分别对水平项 U_t (level)、趋势项 b_t (trend) 和周期项 F_t (seasonal component) 做出估计, HW 模型由以下三个基本公式组成:

$$U_t = \alpha \frac{d_t}{F_{t-L}} + (1-\alpha)(U_{t-1} + b_{t-1}) \quad (7)$$

$$b_t = \beta(U_t - U_{t-1}) + (1-\beta)b_{t-1} \quad (8)$$

$$F_t = \gamma \frac{d_t}{U_t} + (1-\gamma)F_{t-L} \quad (9)$$

其中, d_t 为该时段中的观测值, L 为该时段周期长度, α, β, γ 为对水平、趋势和周期项的平滑参数, 取值在 $[0, 1]$ 上, 并且使得历史数据满足:

$$\sum_{i=1}^n (d_i - \hat{d}_i)^2 = \min_{\alpha, \beta, \gamma} \sum_{i=1}^n (d_i - \hat{d}_i)^2 \quad (10)$$

以该月前十五天第一时段数据作为实验数据、后十五天的作为测试数据进行预测,检验 HW 模型对于 ATM 业务总量的预测效果。

得到 $\alpha=0.147$, 意味着当前预测基于较远的观测值; $\beta=0$ 表明趋势部分的斜率在整个时间序列上是不变的, 等于初始值, 符合图像的直观感受; $\gamma=0.199$ 表明当天该时段的预测基于较远的观测值。

如图 5 所示, 黑色部分为观测数据, 红色部分为 HW 模型求解得到的前 15 天实验数据, 可见 HW 算法成功预测了 ATM 设备业务总量的峰值。于是预测未来 2 天数据进行比较。

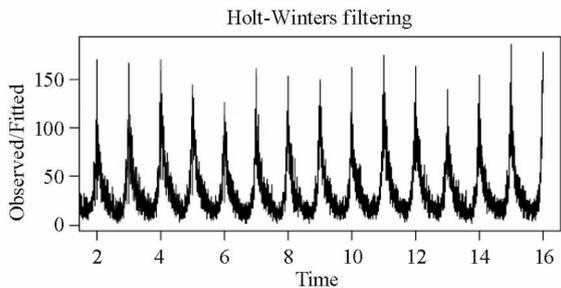


图 5 前 15 天 HW 模型求解结果

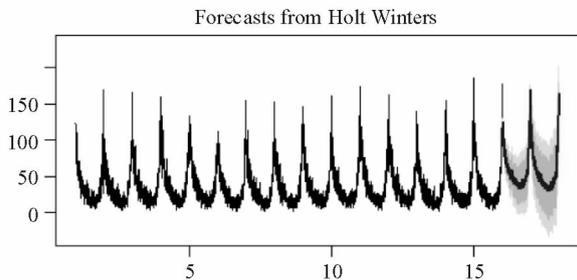


图 6 时段 1 内 ATM 业务总量预测

其中, 图 6 中蓝色线条表示预测值, 深灰色部分表示 80% 的置信区间, 浅灰色表示 95% 的置信区间。对比实际数据, 发现 99% 实际数据在 95% 置信区间内, 结果合理。因此, 可以利用 HW 模型进行预测, 判断未来时段 1~8 内 ATM 设备业务总量是否合理, 进而判断是否发生故障, 及时调整。

4 总结

银行 ATM 设备总业务量具有周期性和时序性, 利用该特征能够判断银行 ATM 设备运行状况, 保证设备正常运行。本文建立银行 ATM 设备业务总量的时序分布特点, 建立时序分布模型, 随后利用模拟退火算法对模型进行求解, 将业务总量时序分布分为 8 个时段, 进而体现出业务总量分布的多峰值特征; 在此基础之上建立简单的加性 Holt-Winters 模型进行业务量预测, 得到 95% 置信区间内业务总量, 当实际数据低于该区间或者高于该区间, 说明银行 ATM 设备存在故障的可能, 能够给予银行维护人员警示, 保障银行 ATM 设备交易的正常运行。

现有网络故障诊断方法一般都采用模糊逻辑或概率分析的方法, 能够完成不确定性条件下的推理决策, 但对于银行网络交易数据而言, 其具有较强的时序特征, 因此利用常规的网络诊断方法缺乏针对性。本文通过对 ATM 设备业务总量时间序列进行建模, 按照序列特点采集其特征, 并进行业务量预测, 从时序特征的角度完成了银行 ATM 业务总量状态的诊断问题。

在未来研究中, 首先, 由于模拟退火算法求解结果的精确度较低, 因此探究如何准确求解时序模型, 将更有利于寻找时序特征。其次, 因为数据内容的限制, 本文没有考虑 ATM 设备的状态和地域设置对业务总量的影响, 所以未来可集合 ATM 设备业务总量的时序性和地域性, 从而进一步探究银行 ATM 设备业务总量精确的时空规律, 建立更加高效的 ATM 设备资源时空分配策略。

参考文献:

- [1] 冯峥. 基于粗糙集理论的银行卡故障诊断系统研究与实现[D]. 上海: 上海交通大学, 2008.
- [2] 徐定杰, 郑笑天. 基于 CMDB 的银行故障管理优化实现[J]. 黑龙江科技信息, 2010(6): 70.
- [3] ESLING P, AGON C. Time-series data mining[J]. ACM Computing Surveys, 2012, 45(1): 12.
- [4] 张俊, 殷坤龙, 王佳佳, 等. 基于时间序列与 PSO-SVR 耦合模型的白水河滑坡位移预测研究[J/OL]. 岩石力学与工程学报, 2015, 34(2): 382-391.
- [5] BAGNALL A, DAVIS L, HILLS J, et al. Transformation based ensembles for time series classification [C]//Proceedings of the 2012SIAM International Conference on Data Mining (SDM 2012), 2012: 307-318.
- [6] BRANDMAIER A M. pdc: permutation distribution clustering[J]. Psychological Methods, 2015, 18(1): 71-86.
- [7] RAKTHANMANON T, KEOGH E. Fast shapelets: a scalable algorithm for discovering time series shapelets. // Proceedings of the 13th SIAM International Conference on Data Mining (SDM 2013). Austin, USA, 2013: 668-676.
- [8] 吴华意, 李锐, 周振, 等. 公共地图服务的群体用户访问行为时序特征模型及预测[J/OL]. 武汉大学学报(信息科学版), 2015, 40(10): 1279-1286+1316.
- [9] 杜涛, 熊立华, 江聪. 渭河流域降雨时间序列非一致性频率分析[J]. 干旱区地理, 2014, 37(3): 468-479.
- [10] 钟锦源, 张岩, 文福拴, 等. 基于时间序列相似性匹配的输电系统故障诊断方法[J]. 电力系统自动化, 2015, 39(6): 60-67.
- [11] 沙川. Holt-Winters 时间序列模型参数估计和预测[D]. 南京: 南京大学, 2011.