所在组别	2021 年中国高校大数据挑战赛	参赛编号
研究生		bdc211037

智能运维中的异常检测与趋势预测

摘要

KPI（key performance indicator）异常检测是智能运维的一项底层核心技术。本文针对KPI异常检测技术展开研究，对KPI数据进行可视化分析，并对KPI异常进行了提取与分类，最后利用多种时间序列预测方法对为来3天的趋势进行了预测。

针对问题一的异常检测，本文首先对每一个小区的3个KPI指标进行单独提取，然后对数据进行统计性分析与可视化绘图，并根据其是否服从正态分布，提出了依据固定分位数、同比（与历史周期对比）、环比（与相邻时间对比）等指标判定异常点的方法。总共得到小区内平均用户数指标的异常点为155个、小区PDCP流量指标的异常点为347个、平均激活用户数指标的异常点为305个。最后给定3小时为异常周期选择标准，判定异常点距离小于等于3小时内的异常值即为一个异常周期，剩下的异常值即为异常孤立点，3个KPI指标的异常检测汇总见表2。

针对问题二的异常预测，本文首先对3个KPI指标进行相关检验，其Pearson相关系数R均超过0.86，存在明显的线性相关性，因此本文使用三维指标联合预测，并使用极限学习机（ELM）对问题一提取出的异常点进行二分类预测，其3个KPI指标58个小区异常预测的F1值见表5，模型整体F1指标的平均值为0.910，标准差为0.075。

针对问题三的趋势预测，本文首先使用长短期记忆网络（LSTM）等神经网络方法对其中一个KPI指标进行单步预测，但是在多步预测过程中神经网络会存在累计误差，因此本文又尝试使用传统方法奇异谱分解（SSA）对数据进行多步预测，其趋势预测的MAPE值大多在10%左右，但这一方法对模型参数很敏感，一个参数不能适用于所有的小区，综上，本文提出了多元线性回归预测结果和季节性自回归滑动平均(SARIMA)模型预测结果进行加权融合，其精度最高，最终使用该方法预测未来3天的指标，预测结果已填写在支撑材料附件2中的预测值表格中。

关键词：KPI异常、异常预测、趋势预测、极限学习机、LSTM、SARIMA

一、问题一：异常检测

1.1 问题分析

问题一要求利用附件1中的KPI指标数据，对所有小区在小区内的平均用户数、小区PDCP流量、平均激活用户数这三个关键指标上检测这29天内共有多少个异常数值，其中异常数值包含两种情况：异常孤立点、异常周期。

本文首先对这个3个KPI指标按照58个小区分别进行提取，然后批量可视化绘图，对数据进行充分的分析，例如第32个小区的3个KPI指标如图1所示：

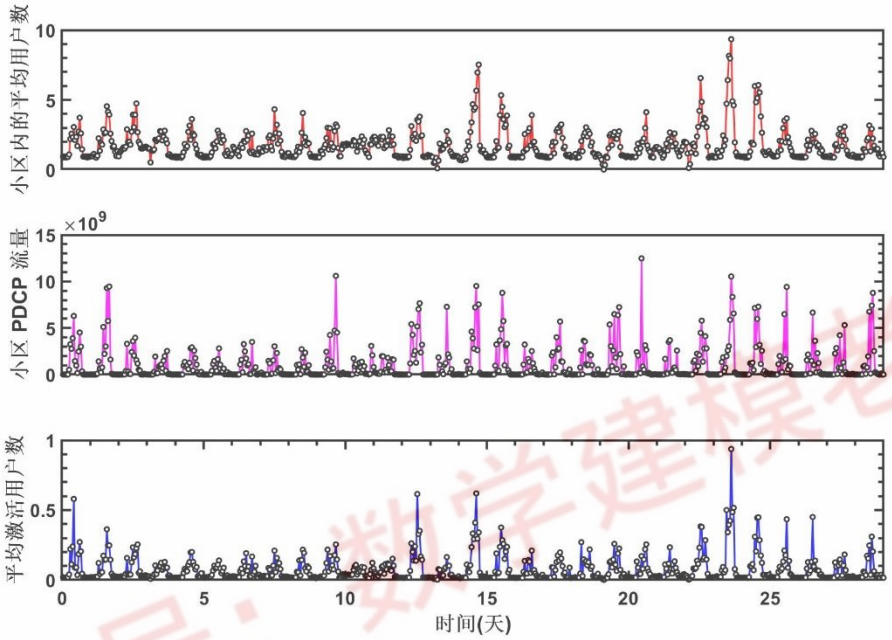


图1 KPI指标原始时间序列（第32个小区）

通过分析58个小区的3个KPI指标发现，不同小区的KPI指标的量级一般不同，但是每个小区的KPI指标的变化基本呈现一定的周期性。

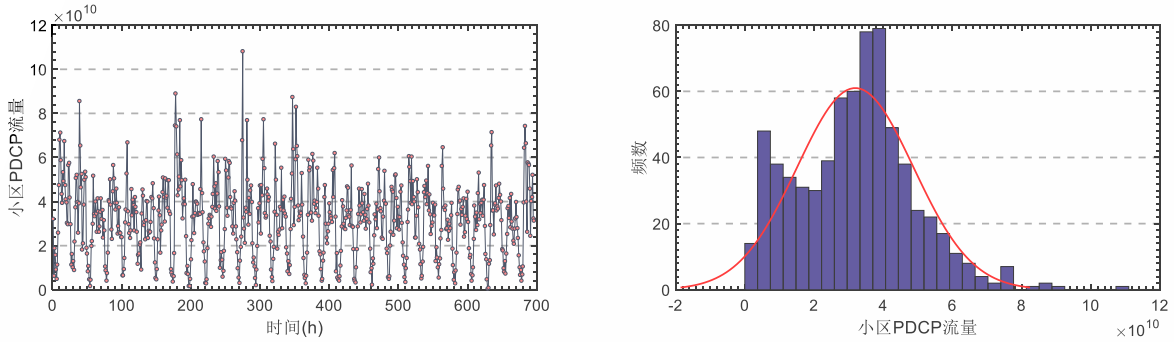
KPI序列可能呈现不同类型的异常模式[1]，主要有以下3种：

（1）全局异常：全局异常指在不考虑数据点的时间关系的情况下，某些超出了KPI序列正常范围的离群点。

（2）集合异常：集合异常指某些数据点的集合相对于整个KPI序列是异常的。集合异常的各个数据点本身可能不是异常的，但作为集合一起出现会被视为异常。

（3）局部异常：局部异常也可称为上下文异常，区别于全局异常，在某些情况下，虽然数据点仍在KPI序列正常范围内，但与邻居点存在很大差异。

针对单维KPI数据的异常检测，基于统计学的方法是一类经典方法，采用$k\sigma$、分位数等方法或基于预设定的阈值等进行异常检测。例如3σ准则，假定数据服从正态分布，如果某些值超过3倍标准差，那么可以将其视为异常点。3σ准则具有方法简洁、运算速度快等优势，但由于其存在较强的假设条件，在很多数据中表现不佳。我们对小区的3个KPI指标数据均进行了分布拟合，个别结果如图2所示。



(a) (b)

图2 小区1的PDCP流量(a)与分布直方图(b)

从图2中可以看到，由于数据本身是非负的，数据分布表现出明显的非对称性，与正态分布存在比较大的差异，不满足3σ准则使用的前提条件[2]，并且每个小区的KPI指标水平存在比较大的差异，也无法设定统一的阈值等进行异常检测。因此本文基于分位数指标来对异常数据进行处理，分位数是一种定序指标，对于异常数据不敏感，适用于含有异常数据的问题的处理。

基于此，本文对问题一中异常检测的整体思路如图3所示：

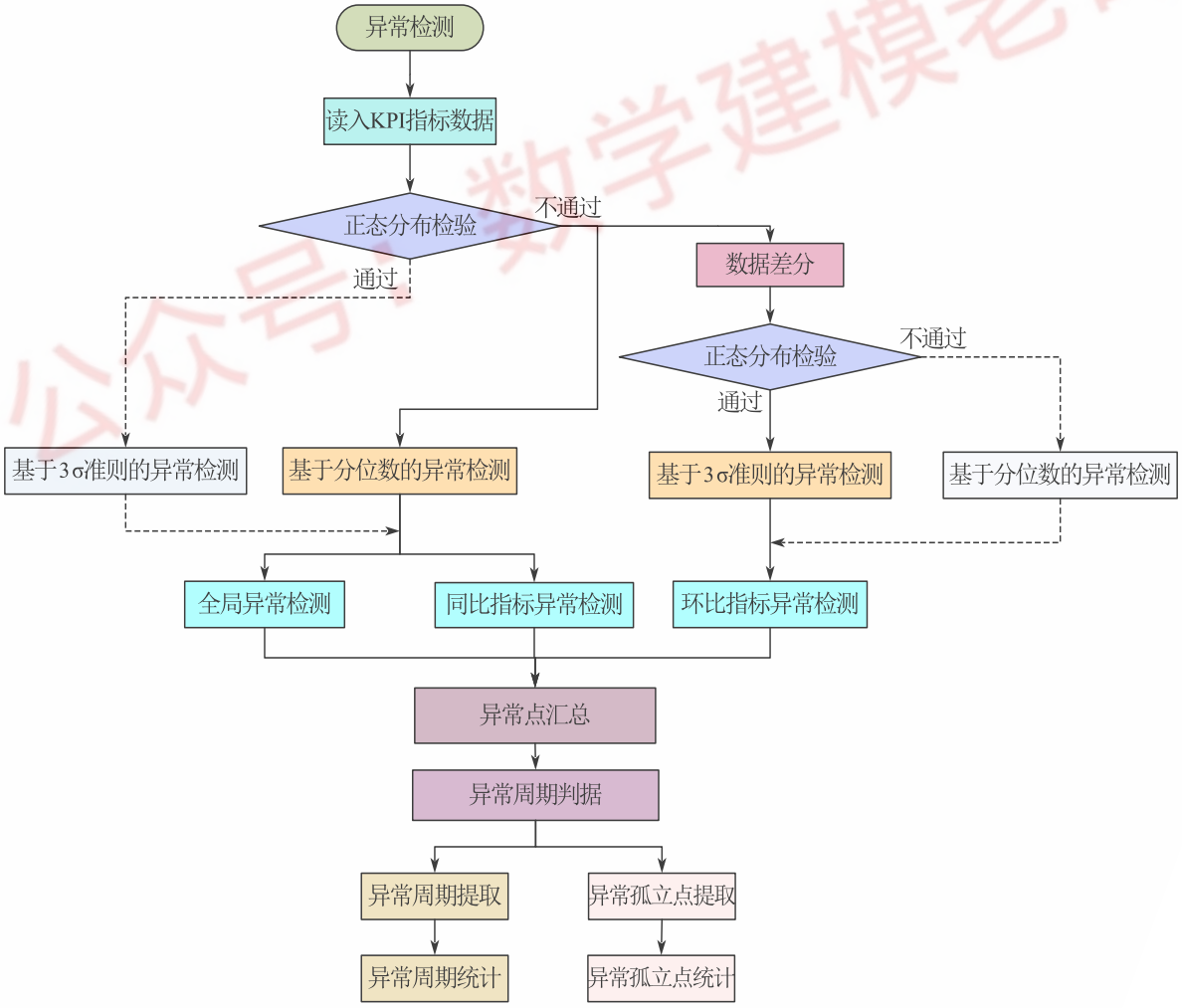


图3 异常检测整体思路

1.2 分位数全局异常检测

分位数（英语：Quantile），亦称分位点，是指用分割点（cut point）将一个随机变量的概率分布范围分为几个具有相同概率的连续区间。分割点的数量比划分出的区间少1，例如3个分割点能分出4个区间。常用的有中位数（即二分位数）、四分位数（quartile）、十分位数（decile ）、百分位数等[3]。

分位数可以利用箱线图可视化汇总统计量。箱线图（Boxplot）也称箱须图（Box-whisker Plot），是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数与最大值来描述数据的一种方法，它也可以粗略地看出数据是否具有有对称性，分布的分散程度等信息。

箱线图提供样本数据的汇总统计量的可视化，如图4所示[4]，一般包含以下特性：

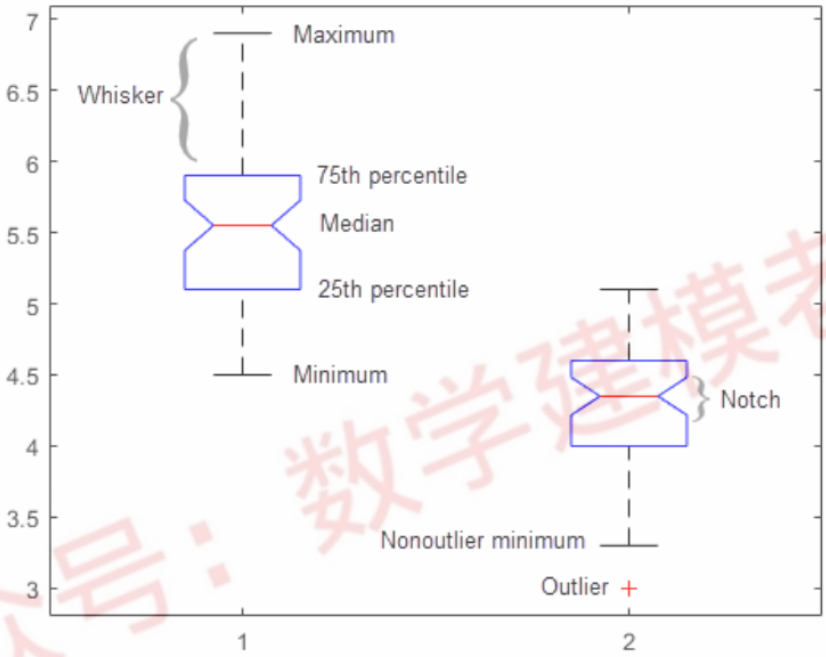


图4 箱线图简介

（1）每个箱子的底部和顶部分别表示样本的第25个百分位数q_1和第75个百分位数q_3。每个箱子的底部和顶部之间的距离表示四分位差。

（2）每个箱子中间的红线表示样本中位数q_2。如果中位数不在箱子的中心，则绘图显示样本偏度。

（3）须线是自每个箱子的顶部向上延伸和底部向下延伸的线条。须线从四分位差的端点延伸到须线长度内最远的观测值（相邻值）。

（4）超出须线长度的观测值标记为离群值，具体为大于$q_3+w\times(q_3-q_1)$或小于$q_1-w\times(q_3-q_1)$的值。默认情况下，离群值是距离箱子底部或顶部超过1.5倍四分位差的值w。不过，该参数可以在实际运用中实时调整调。离群值显示为红色+号。

（5）缺口显示样本间中位数的变异性。计算缺口的宽度，使得缺口不重叠的框在5%显著性水平上具有不同中位数。显著性水平基于正态分布假设，但对于其他分布，中位数比较也可合理地认为是稳健的，因此这就是基于分位数检测的优势。

对于本问题，绘制小区内的平均用户数、小区PDCP流量和平均激活用户数这三个KPI指标的箱线图，分别如图5、图6和图7所示。

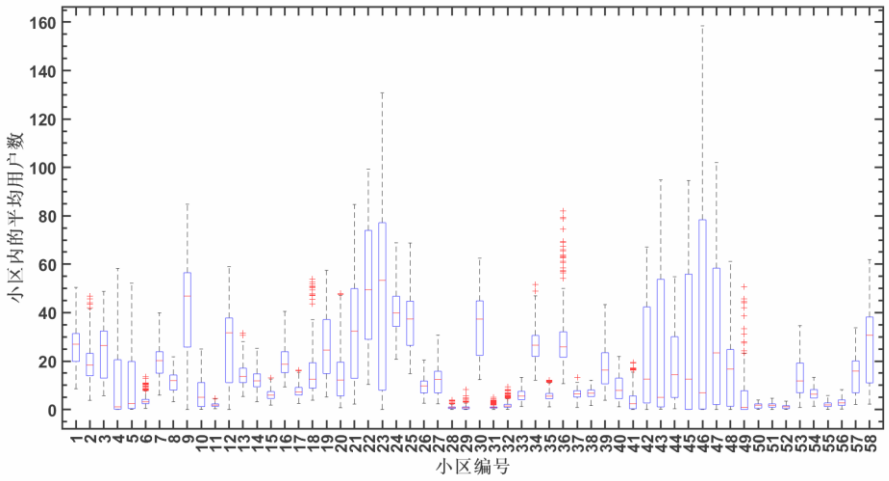


图5 小区内的平均用户数箱线图

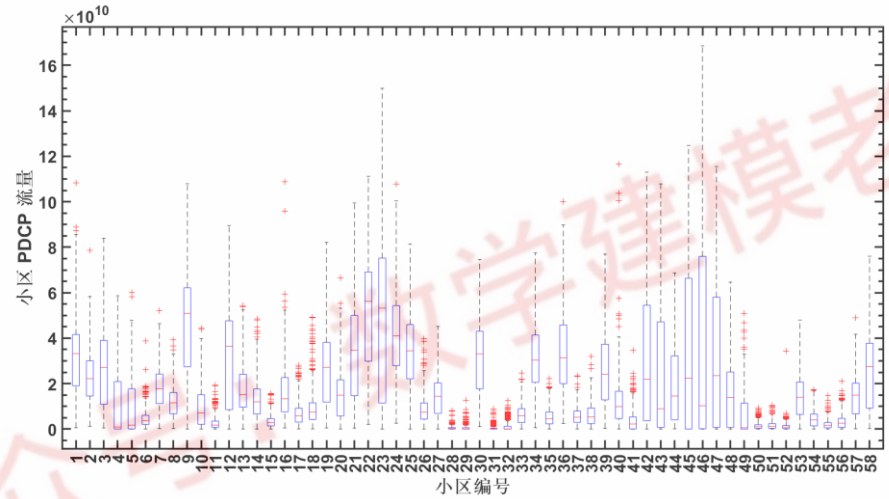


图6 小区PDCP流量箱线图

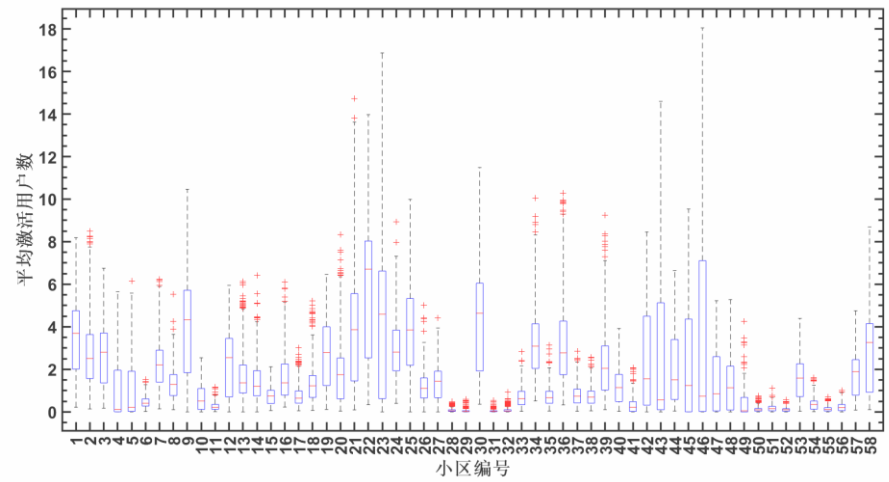


图7 平均激活用户数箱线图

从上述图5、图6和图7中可以看出，每一个指标在不同小区的分位数差异较大，从中可以找到较多异常点，但是有个别小区异常点过多，这显然是不符合实际的，因此需要调整

百分位数q_1和百分位数q_3的值以及乘数w。对于本问题，百分位数q_1=35%，百分位数q_3=85%，乘数w=1.5。基于分位数的异常点检测效果如图8所示：

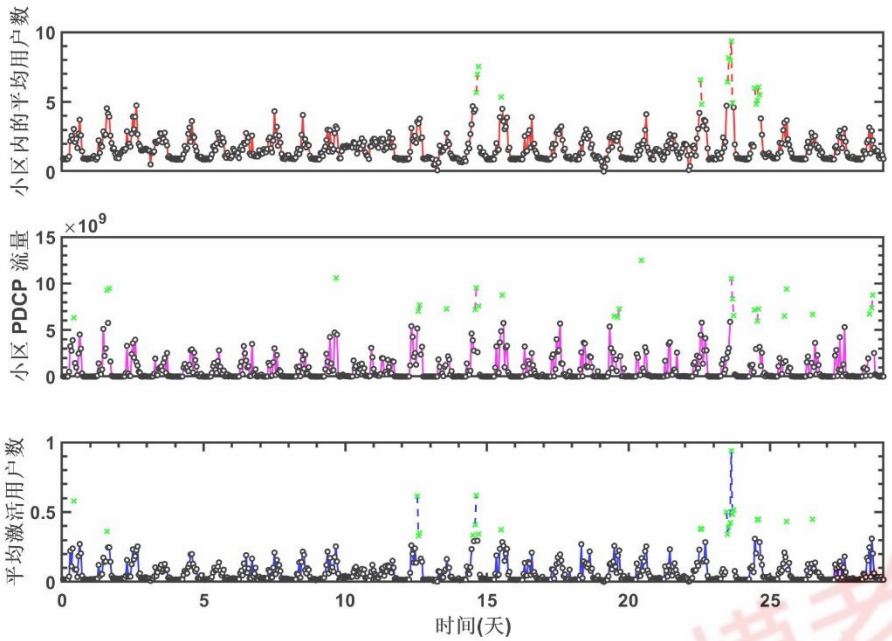
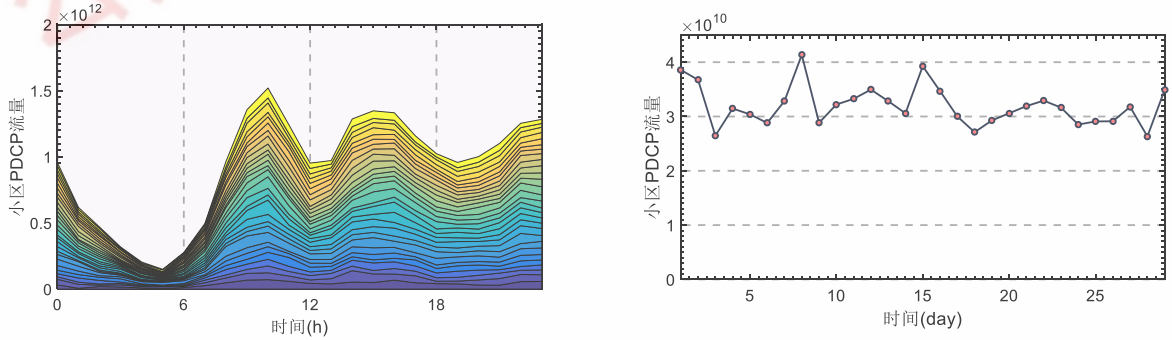


图8 基于分位数的异常点检测（第32个小区）

从图8中可以看出，绿色的×即为异常点，基于分位数的异常点检测效果很好，绝大部分异常点已经被检测出来。

1.3 同比指标检测异常

考虑到对小区全部数据计算分位数的方法无法处理集合异常与局部异常，本文提出了同比（与历史周期对比）、环比（与相邻时间对比）两个指标来判断这两种异常。本节进行同比指标检测异常，如图9所示。



(a) (b)

图9 小区1的PDCP流量29天数据的堆叠条形图(a)，每天的总流量时间序列图(b)

从图9(a)中可以看到，每天各个时刻的流量变化规律具有高度相似性，图9(b)则表明，每天的总流量整体上保持稳定，并且存在以一星期为周期的小幅度波动。这是进行同比、环比指标分析的理论依据。

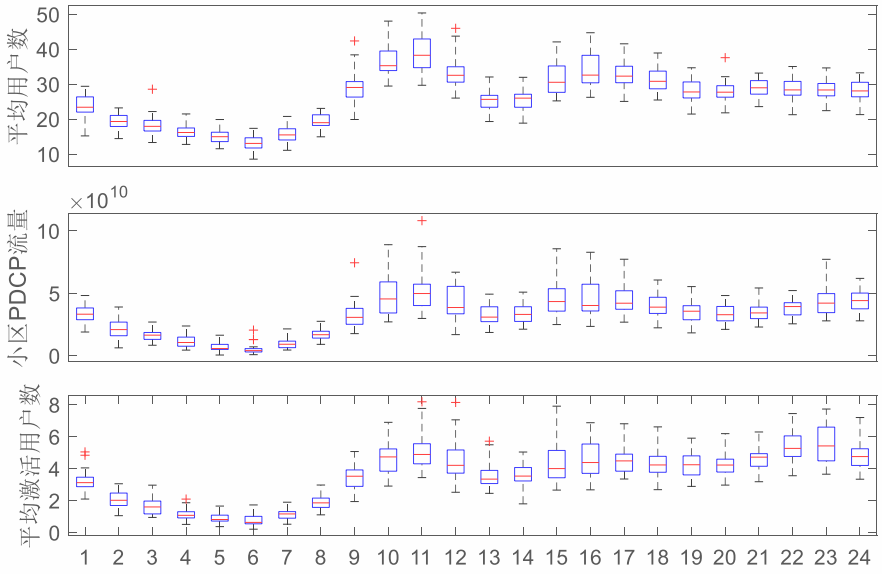


图10 小区1的平均用户数、PDCP流量、平均激活用户数的同比箱线图

图10给出了小区1的平均用户数、PDCP流量、平均激活用户数的同比箱线图。同比分析首先将小区29天的696 h的时间序列数据拆分为24个子序列，每天相同时刻的数据为1个序列，例如每日12点的平均用户数为1个序列。同比异常检测结果如图11所示。

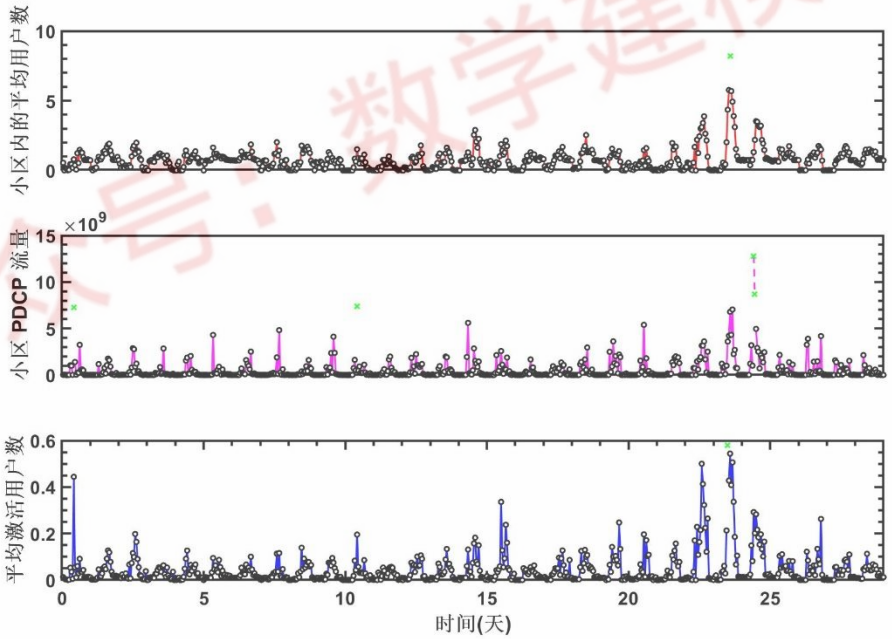


图11 同比异常点检测（第29个小区）

从图11中可以看出，通过同比指标监测出的异常点用绿色×表示，同比异常检测会检测出一部分新的异常点，而另一部分异常点是在基于分位数整体异常检测时已经检测到的点。

1.4 环比指标检测异常

环比指标主要是与相邻时间数据对比差异，这就需要对原始的数据序列进行差分处理，差分后的时序图如图12所示。

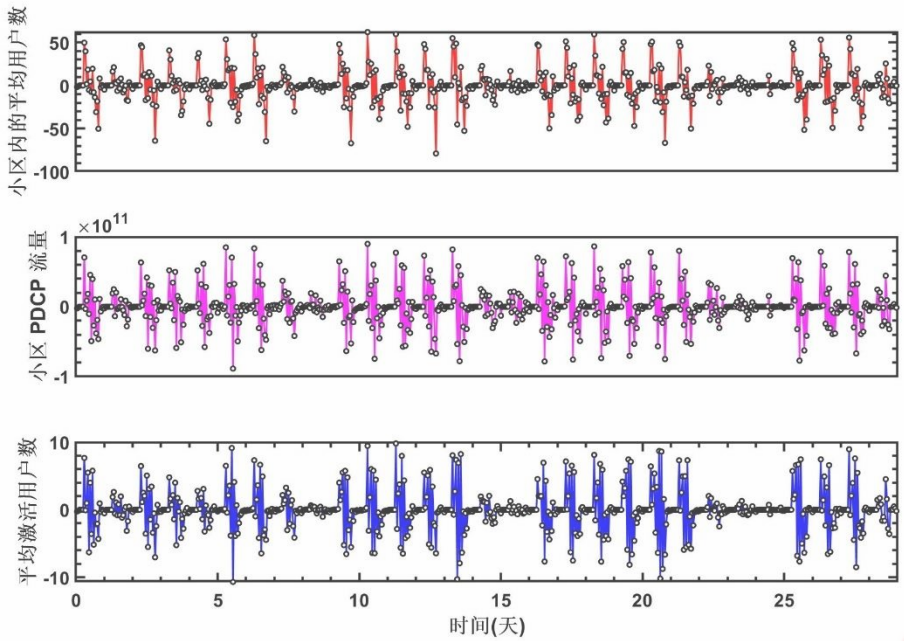


图12 KPI指标差分序列（第46个小区）

从图12中的差分序列数据可以看出，其数据分布表现出明显的对称性，因此可以考虑利用3σ准则进行异常检测，但是需要数据服从正态分布，正态分布有两个参数，期望值（或均值）μ和标准差σ，记为$N(\mu,\sigma^2)$：正态曲线的中心位置是由正态分布的期望值μ决定的，它代表了实验数据的平均水平；正态曲线的陡峭程度是由正态分布的标准差σ决定的，σ越小，正态分布的曲线就越“高而尖”；σ越大，曲线越“矮而平”[2]。正态分布的公式如(1.1)所示：

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1.1}$$

标准正态分布如图13所示：

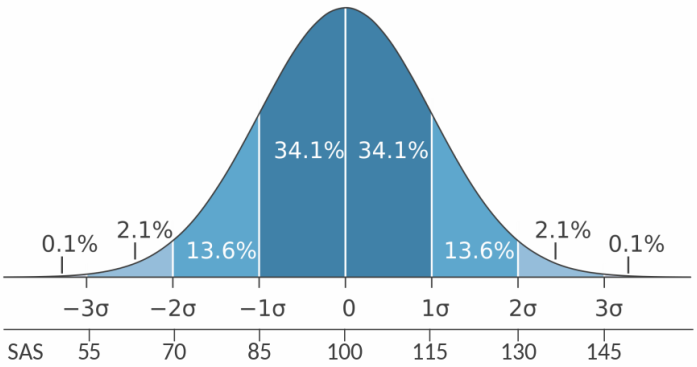


图13 标准正态分布图

如图13所示，该图表示某一个数据在$\mu\pm\sigma$范围内的概率是68.3%，在$\mu\pm2\sigma$范围内的概率是95.4%，$\mu\pm3\sigma$范围内的概率是99.7%。

当正态曲线基本对称，且呈现“钟形”分布时，基本可以认定该数据是服从正态分布的，但是根据直观观察得出的结论具有很大的主观性。因此，需要采用检测数据的方法来

判断数据的正态性。现有的常用的数据正态性检验方法有Jarque-Bera检验、Kolmogorov-Smirnov检验、Lilliefors检验和t-检验等，但是前三种数据检验方法多用于标准正态分布检验。因此在本任务中，我们使用t-检验法对数据进行正态性检验，该检验法又称为student's t-test。对于显著性水平，布尔变量h=0，表示该数据在总体上服从正态分布；否则h=1，表示数据总体不服从正态分布。具体的方法是将差分数据序列绘制数据分布直方图，如图14所示。然后使用MATLAB进行t检验，在0.2的显著性上通过了正态分布检验。

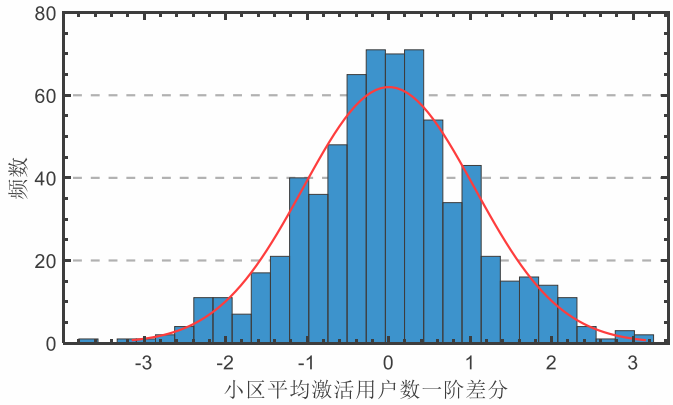


图14 数据分布直方图与正态分布拟合（第1个小区）

由于各差分序列数据近似服从正态分布，可以利用3σ准则找到差分数据的异常点，具体判据就是当前差分值大于3σ值而下一个差分值小于3σ值，找到差分值的索引后，再回到原始的时间序列找到异常点。利用环比检测的异常点如图15所示：

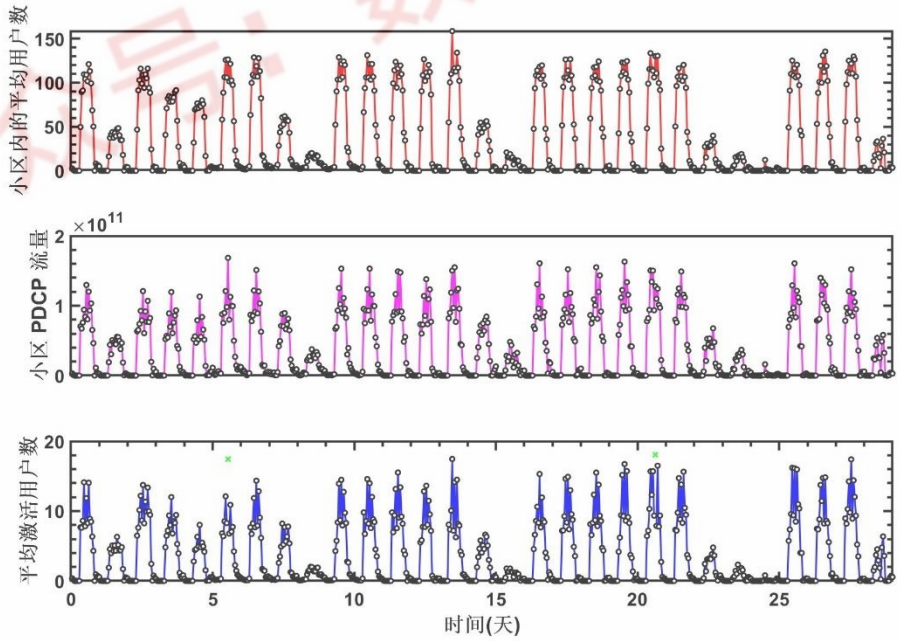


图15 环比异常点检测（第46个小区）

图15中展示了利用环比指标检测到的异常点，见平均激活用户数指标中绿色×的标识。但是环比检测到的数据相对比较少，在全部数据集中只出现了2个。

1.5 异常点汇总

上述1.2-1.4节分别利用分位数全局异常检测、同比指标异常检测和环比指标异常检测方法，检测到了三种类型的异常点，而这3种异常点可能会有交叉重复，因此需要找到所有异常点，然后求并集。将所有异常点进行汇总，汇总后的异常点数据如图16所示：

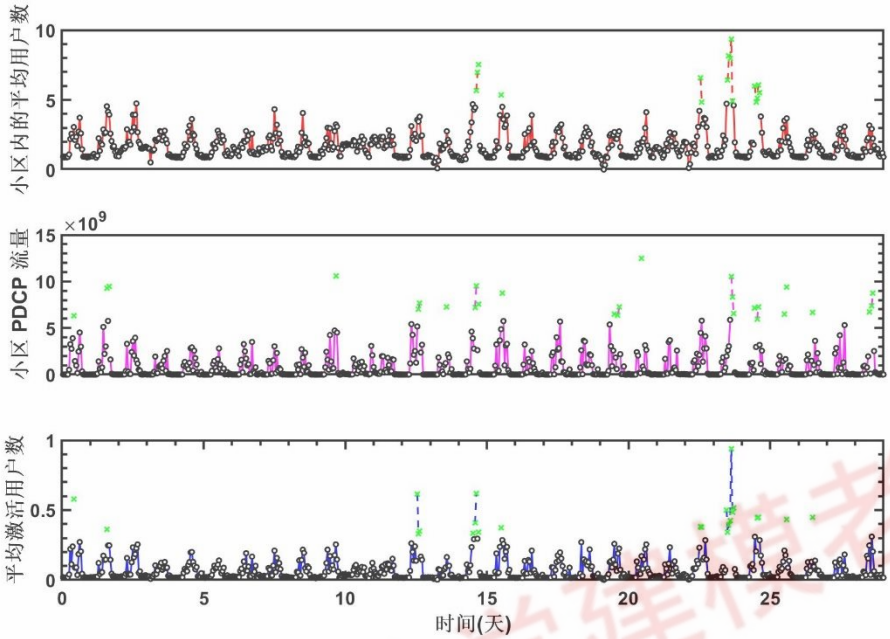


图16 3种异常点检测汇总（第32个小区）

找到所有异常点后，对3种KPI指标的三种异常点进行汇总，如表1所示。

表1 三种异常点检测汇总（所有小区）

KPI指标	异常点数量（个）
小区内的平均用户数	155
小区 PDCP 流量	347
平均激活用户数	305

1.6 异常周期和异常孤立点判断

异常周期指在一段时间内有多个异常值。首先需要指定异常周期的时间范围针对这一问题，本文分析了小区内平均用户数、小区PDCP流量、平均激活用户数这三个指标中，所有异常数据的时间间隔，并绘制了其分布直方图，如图17所示。

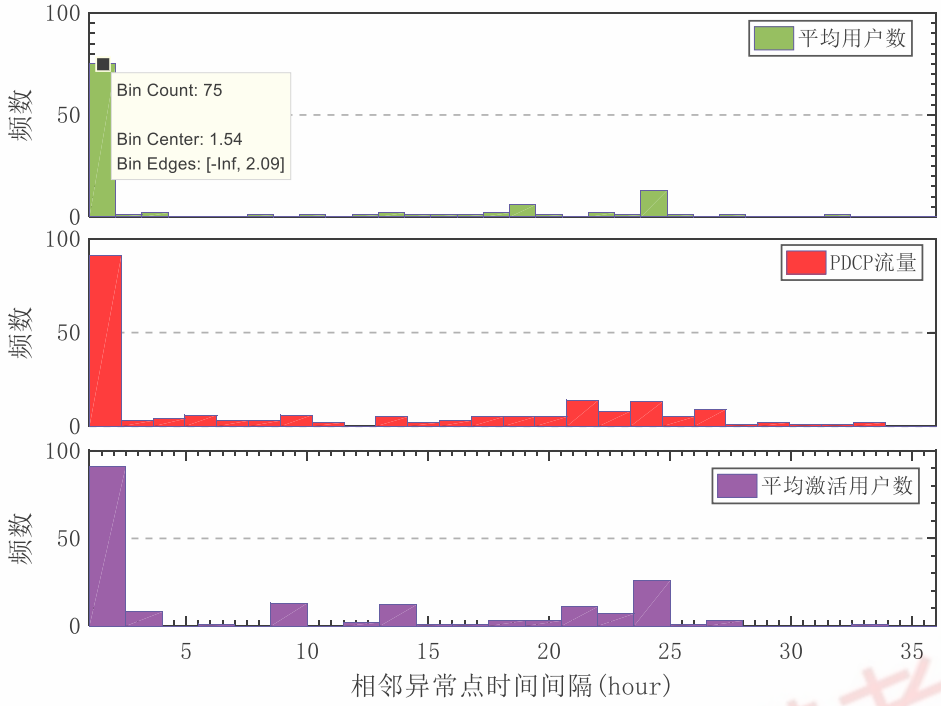


图17 三个指标KPI中，所有异常数据的时间间隔

从图17中可以看出，大部分相邻的异常点，其时间间隔都不超过2小时。因此，本文以3小时为分界距离，如果两个相邻的异常点的时间间隔小于等于3小时，认为它们属于同一个异常周期，如图18所示，反之则认为它们不属于同一个异常周期。异常周期判断完毕以后，剩下的异常点就是异常孤立点。

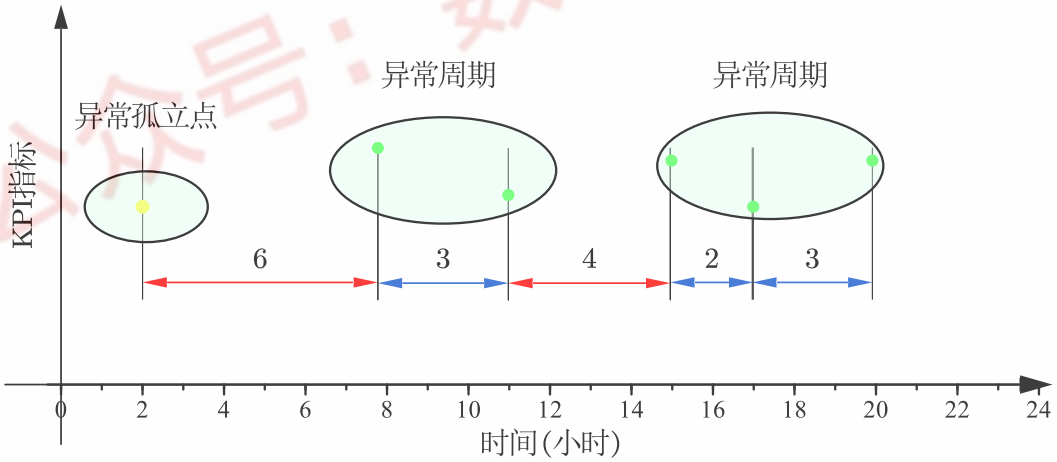


图18 异常周期判断示意图

对58个小区提取出异常周期和异常孤立点，以小区32为例，效果如图19所示：

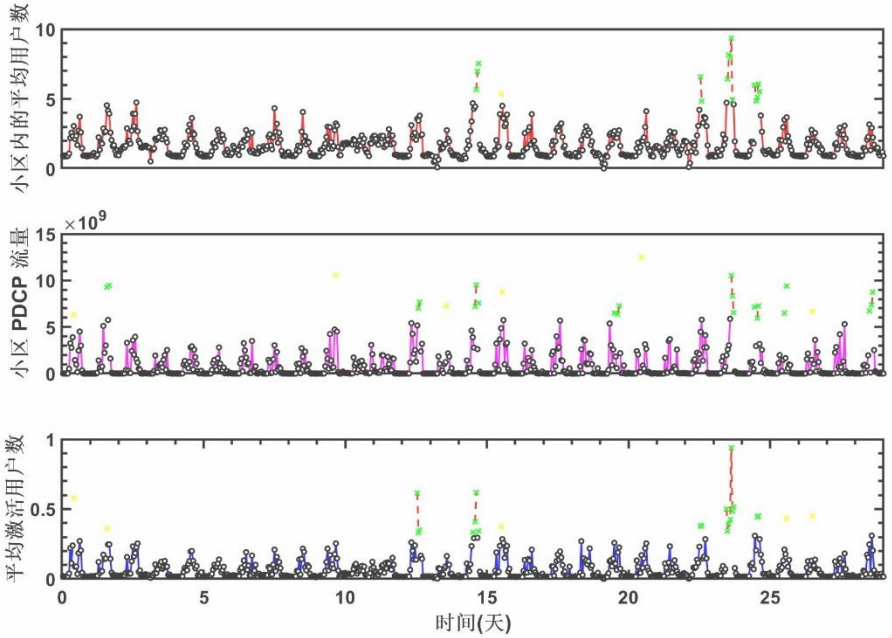


图19 异常周期和异常孤立点（第32个小区）

图19中，黄色×代表异常孤立点，绿色×代表异常周期内的点。最后对异常周期和异常孤立点进行汇总，如表2所示。

表2 异常检测汇总表

	时间周期选择标准	异常孤立点的个数	异常周期个数
小区内的平均用户数	3小时	48	31
小区PDCP流量	3小时	196	57
平均激活用户数	3小时	150	58

二、问题二：异常预测

2.1 问题分析

问题二异常预测要求针对问题一检测出的异常数值，建立预测模型，预测未来是否会发生异常数值。异常预测模型需要考虑两点：

1）模型输入的时间跨度，输入数据的时间跨度越长，即输入数据量越多，模型越复杂，会增加计算成本和模型鲁棒性，降低泛化能力，本模型重复考虑到问题一异常数值本身的特性选择了预测点前6 h-24 h一共18-72个特征数据作为输入进行单独建模得到F1。

2）模型输出时间跨度，即预测的时长。本文中为确保模型预测的精度选择仅预测下一个时刻的数据是否异常，实际上本文的模型亦可进行多步预测，不过预测精度会有一定下降。

2.2 三维指标联合检测

通过小区3个KPI指标的散点图，可以分析指标之间的相关性以及指标的分布特性。以第1个小区为例，图20是小区内的平均用户数、小区PDCP流量、平均激活用户数之间的散点图。可以看出，三个指标存在明显的线性相关性，其Pearson相关系数R均超过0.86，如表3所示，因此可以考虑通过3个指标进行异常联合预测。

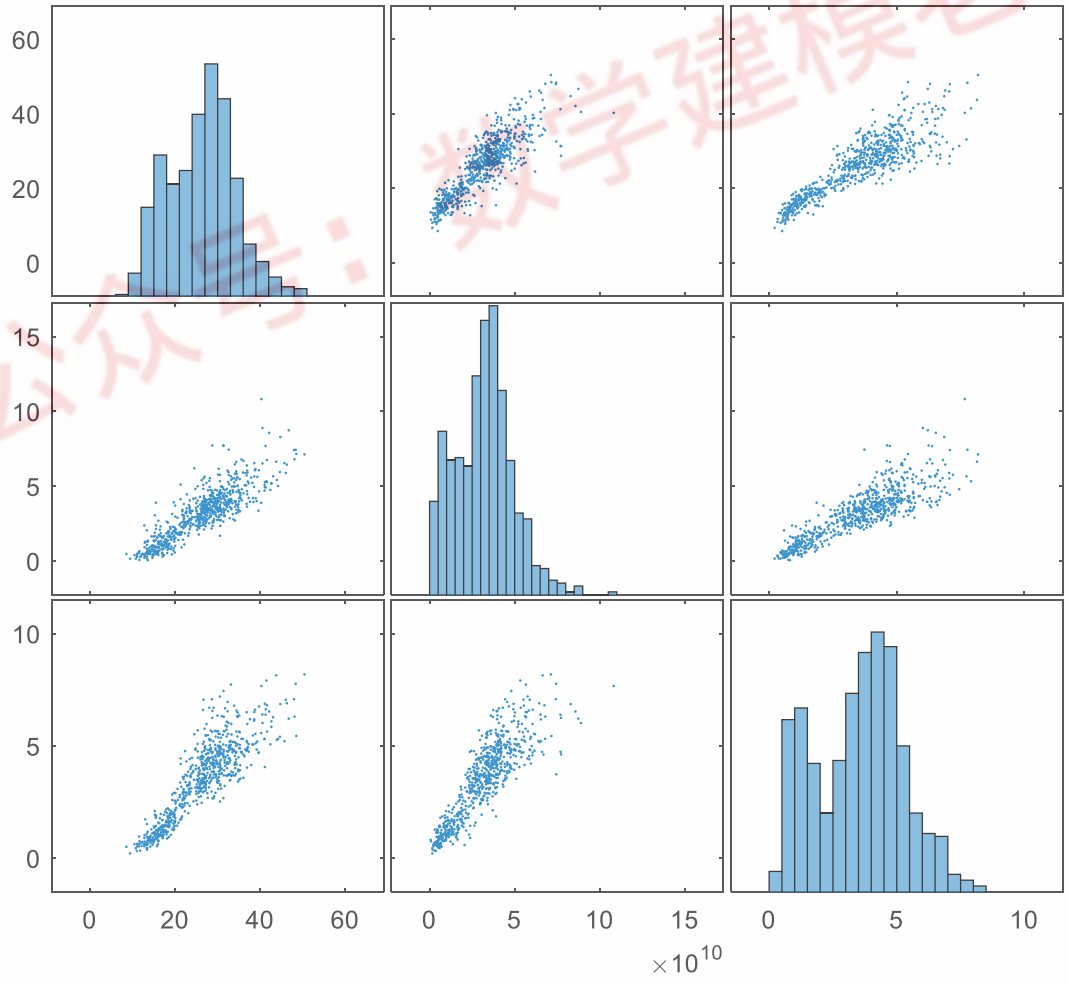


图20 小区1对应的3个KPI指标的散点图

表3 小区1的3个KPI指标的Pearson相关系数

R	平均用户数	PDCP 流量	平均激活用户数
平均用户数	1	0.86	0.87
PDCP 流量	0.86	1	0.86
平均激活用户数	0.87	0.86	1

2.3 极限学习机（ELM）异常预测

传统前馈神经网络具有训练速度较慢、易陷入局部最优、学习率的选择较敏感等缺点。南洋理工大学黄广斌教授[5]在对单隐藏层前馈神经网络[6]的研究过程中发现SLFN的学习能力不会受到网络的输入权值和隐藏层节点阈值的影响，提出了一种新的单隐藏层前馈神经网络，即极限学习机(Extreme Learning Machine, ELM)。ELM的所有参数都可以通过分析确定，而不是调整，因此从理论上讲该算法能以极快的学习速度下提供良好的泛化性能。

极限学习机网络结构如图 21 所示。假设有 N 个样本(X_i,t_i)其中$X_i=[x_1,x_2...,x_n]^T\in R^n$为模型输入样本数据,$Y_i=[y_1,y_2...,y_m]^T\in R^m$是模型的期望输出，图 21 的神经网络可以表示为：

$$\sum_{i=1}^{L}\beta_i G(W_i\cdot X_j+b_i)=y_j,j=1,...,N \tag{2.1}$$

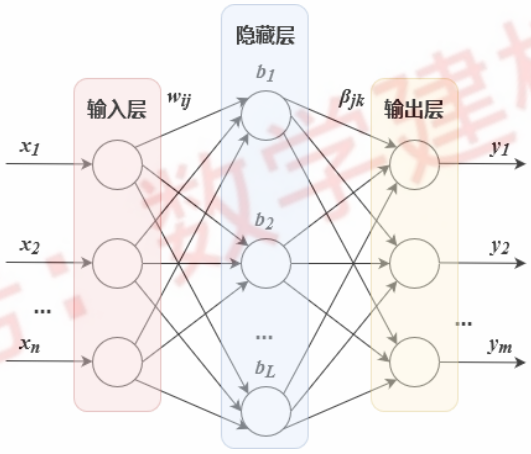


图21 极限学习机网络结构图

公式中W_i为输入权重，β_i为输出权重，b_i为偏置，G为隐藏层激活函数。训练的目标为使得输出误差最小[7]，这等价于最小化损失函数$E=\sum_{j=1}^{N}\left[\sum_{i=1}^{L}\beta_i G(W_i\cdot X_j+b_i)-t_j\right]^2$。在 ELM 神经网络中,一旦输入权重$W_i$和隐藏层偏置$b_i$被随机确定,则隐藏层的输出矩阵就被唯一确定,且输出权重β的最小二乘解是唯一的[8]。本文中神经网络通用参数为：隐藏神经元个数设置为 1000；采用 sigmoid 函数作为激活函数。

此处考虑到问题1中异常点的最早时刻为第7个小时，另一方面针对日常人类生活作息规律该输入的时长超过该点前24 h的数据所起到的训练作用有限。因此当异常点最早时刻小于24 h时采用异常点最早时刻减1 h长度的数据长度作为输入，当超过时统一采用预测点前24 h共72个数据作为输入。输出为用0和1表示的异常是否发生的单个输出。

在机器学习中，二分类问题中常用的评估指标为准确率、查准率、召回率、F1值。如表4所示，在真实数据中按标签可以将数据分为正类样本和负类样本，将预测结果分为真类和假类。真正（TP）指的是将正类样本预测为真类的数量；真负（TN）指的是将负类样本

预测为假类的数量；假正（FP）指的是将负类样本预测为真类的数量，也称其为误报；假负（FN）指的是将正类样本预测为假类的数量，也称其为漏报。

表4 机器学习中的数据分类（二分类）

	正类（Positive）	负类（Negative）
真类（True）	真正（TP）	假正（FP）
假类（False）	假负（FN）	真负（TN）

准确率（Accuracy，ACC）是指所有预测正确的样本数（将正类样本预测为真类，将负类样本预测为假类）占总样本数的比例，如公式(2.2)所示。查准率（Precision，PRE）是指在所有分类为真类的数据中，确实为正类的比例，如公式(2.3)所示。查全率（Recall，REC）是指分类正确的数据中，原本为正类所占的比例，如公式(2.4)所示。F1值（F1_score，F1）是查准率与召回率的调和均值，如公式(2.5)所示。在衡量算法时，这四项指标越高，代表模型表现越好。

$$ACC=(TP+TN)/(TP+FP+FN+TN) \quad (2.2)$$

$$PRE=TP/(TP+FP) \quad (2.3)$$

$$REC=TP/(TP+FN) \quad (2.4)$$

$$F1=(2\times REC\times PRE)/(REC+PRE) \quad (2.5)$$

本模型对58个小区的三个指标进行了建模。模型预测结果的F1指标如表5所示，

表5 ELM模型预测结果的F1指标

小区编号	小区内的平均用户数 F1	小区PDCP流量 F1	平均激活用户数 F1
1		1	0.800
2	0.800	0.800	1
3			1
4			
5	1		1
6	1	0.833	0.912
7	0.909	1	1
8	1		0.857
9			1
10			1
11	0.857	1	0.900
12			
13			
14	0.857		0.857
15	1		0.909
16	1		1
17	0.857	0.857	0.900
18	1	0.968	0.974
19			1
20	0.896	1	0.857
21	0.727		

22			
23			
24	1		1
25			
26	0.909		0.857
27	1	1	1
28	0.857	0.857	0.930
29	0.846	0.800	0.807
30			
31	0.846	0.826	0.899
32	0.884	0.857	0.909
33	0.857		1
34	0.857	1	
35	0.727	0.857	0.909
36		0.824	1
37	0.889	1	0.889
38	0.909		1
39	0.923		
40			0.900
41	0.769	0.667	0.762
42			
43			
44			
45			
46	0.800		
47			
48			
49	0.841	0.889	0.889
50	0.800		0.818
51	0.857		1
52	0.889		0.857
53			1
54	0.786		0.889
55	0.857	1	0.857
56	1		0.875
57			1
58	1		

注：若某小区KPI指标无异常点，则代码输出F1的值为10。

考虑到某些小区的个别数据在29天中存在没有发生异常的情况，针对这种情况在此处未进行异常检测输入模型的训练及测试，表格中用斜下对角框线表示数据不存在异常。模型整体F1指标的平均值为0.910，标准差为0.075。

三、问题三：趋势预测

3.1 问题分析

问题三要求利用2021年8月28日0时至9月25日23时已有的数据，预测未来三天（即9月26日0时-9月28日23时）上述三个指标的取值，并填写附件2中的预测值表格。

该问题中，首先要建立有效的时间序列预测模型，利用已有的29天划分训练集和验证集对模型进行检验，然后再预测29天以后的数据，具体思路如图22所示。

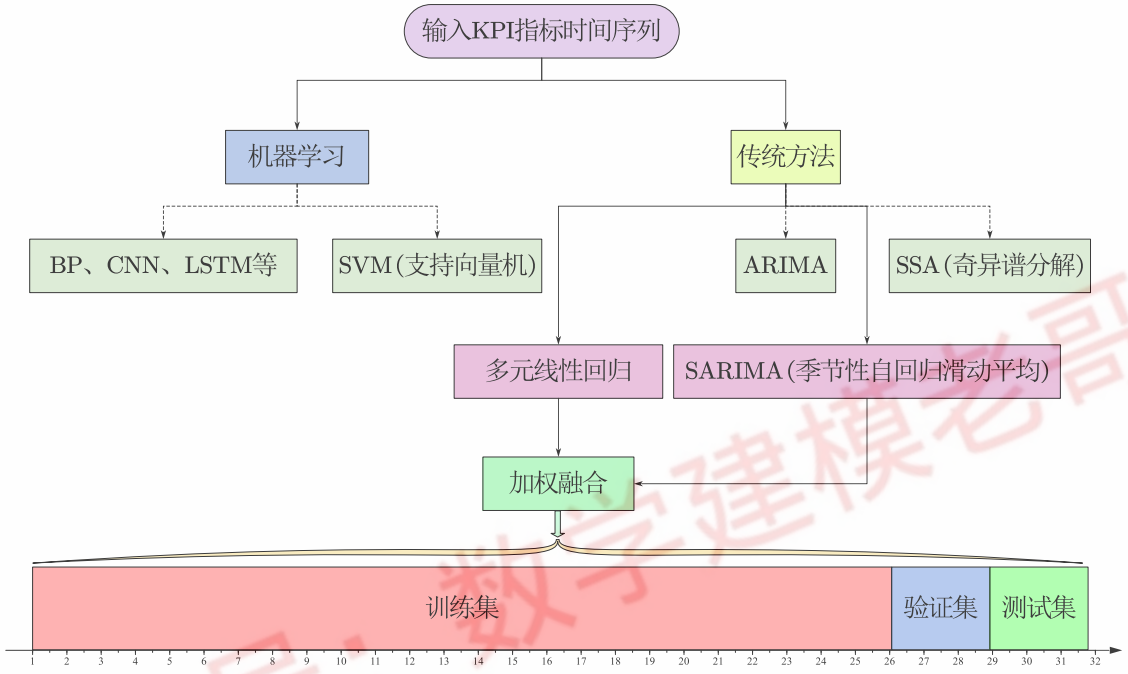


图22 趋势预测整体框图

对于时间序列预测方法可以分为经典方法和智能方法两大类。经典方法主要是基于各种统计理论的时间序列模型，如回归分析法、自回归(AR)，滑动平均(MA)，自回归滑动平均(ARMA)等模型预测方法、奇异谱分解 (SSA)或季节性自回归滑动平均(SARIMA)等，而智能方法包括人工神经网络(BP神经网络，CNN卷积神经网络，LSTM长短期记忆神经网络)、支持向量机(SVM)等方法。

3.2 LSTM 预测模型

本文首先尝试了LSTM长短期记忆网络进行时间序列预测，长短期记忆网络通常被称为LSTM，是一种特殊的RNN，能够学习长期依赖性[9]。由Hochreiter和Schmidhuber（1997）提出的，并且在接下来的工作中被许多人改进和推广。LSTM 在各种各样的问题上表现非常出色，现在被广泛使用。LSTM 被明确设计用来避免长期依赖性问题。长时间记住信息实际上是LSTM的默认行为，而不是需要努力学习的东西。

所有递归神经网络都具有神经网络的链式重复模块。在标准的RNN中，这个重复模块具有非常简单的结构，例如只有单个tanh层，如图23所示。

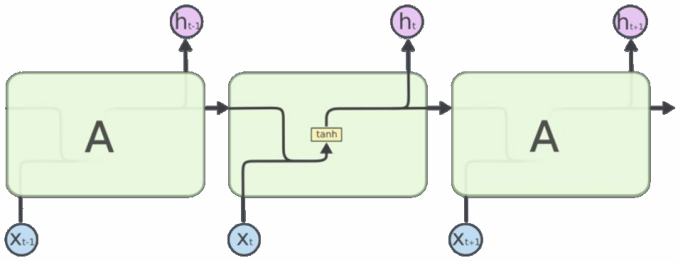


图23 标准的RNN网络

LSTM也具有这种类似的链式结构，但重复模块具有不同的结构。不是一个单独的神经网络层，而是四个，并且以非常特殊的方式进行交互，如图24所示。

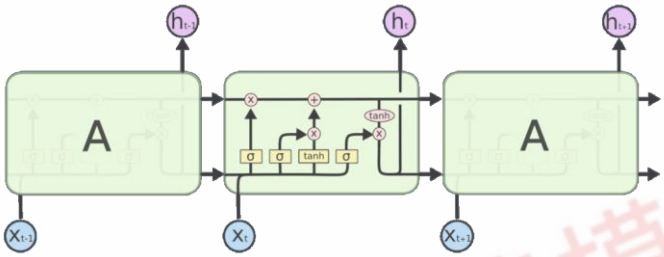


图24 LSTM网络结构

LSTM的关键是细胞状态，细胞状态类似传送带，如图25所示。它贯穿整个链条，只有一些次要的线性交互作用。信息很容易以不变的方式流过。

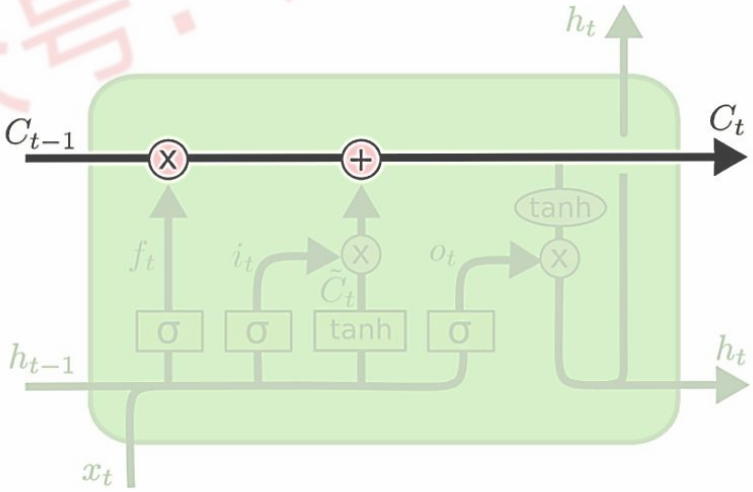


图25 LSTM网络结构的细胞状态

然后对第1个小区内的平均用户数的前26天作为训练集，后3天作为验证集进行训练LSTM神经网络，这里选择自回归阶数为6，也就是利用前6小时的输入作为预测，预测下一个小时的输出，以此来构建数据集。训练以后验证集的单步测试结果和单步预测误差如图26和图27所示，从图26和图27中可以看出，LSTM单步预测误差相对比较小，再次验证其他小区的数据，其MAPE的值在10%左右。

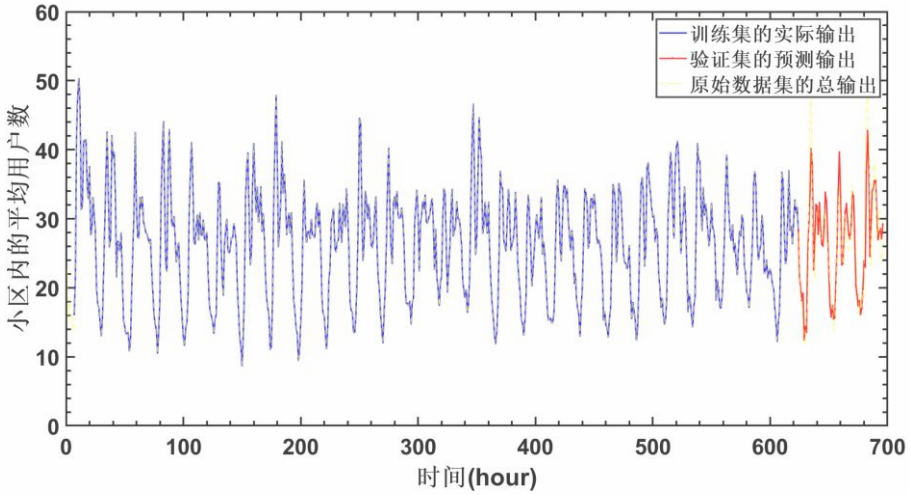


图26 LSTM单步预测（第1个小区）

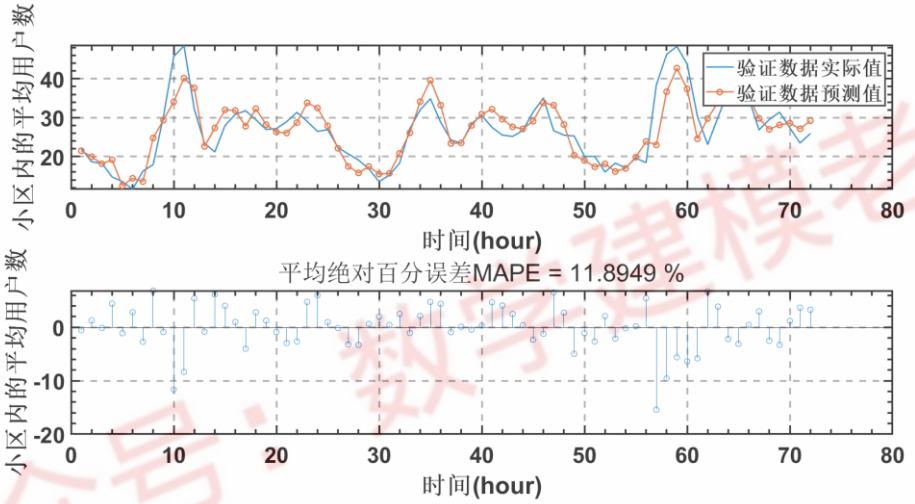


图27 LSTM单步预测误差（第1个小区）

接下来对上述训练好的模型进行多步预测，也就是利用单步预测结果来预测后面的输出，其多步测试结果和单步预测误差如图28和图29所示。从中可以看出，LSTM网络对多步预测的效果很差，当有一个预测无法很大时，会有累计预测误差，因此对于LSTM网络而言，其不太适用于多步时间序列预测。

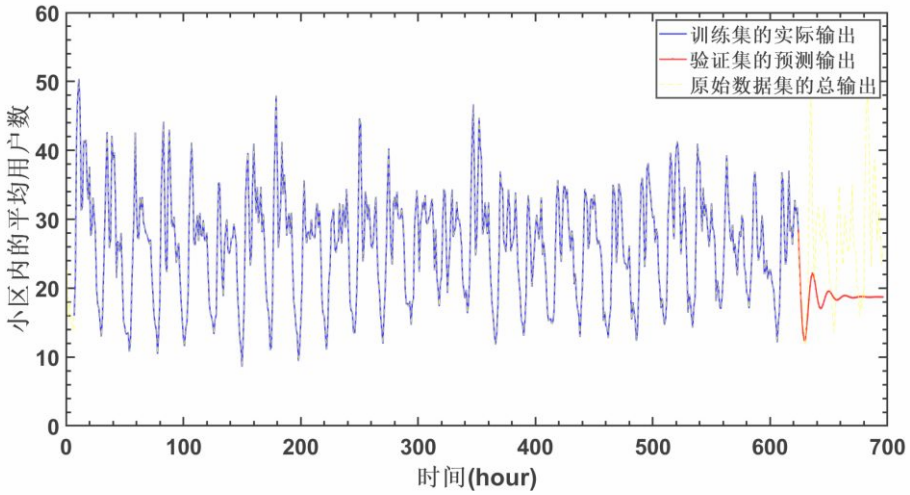


图28 LSTM多步预测（第1个小区）

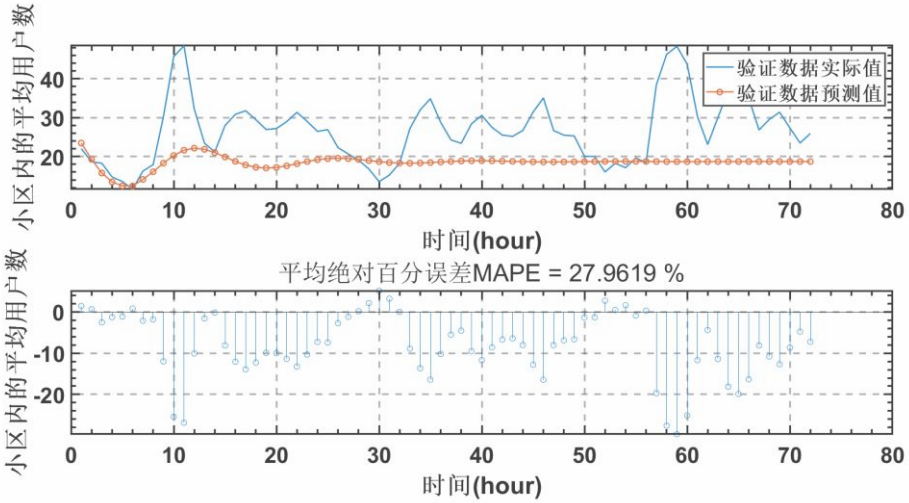


图29 LSTM多步预测误差（第1个小区）

3.3 SSA 预测模型

考虑到神经网络预测模型在多步预测中不是很友好，因此尝试一些传统方法，奇异谱分析（SSA）是近年来兴起的一种研究非线性时间序列数据的强大的方法。它根据所观测到的时间序列构造出轨迹矩阵，并对轨迹矩阵进行分解、重构，从而提取出代表原时间序列不同成分的信号，如长期趋势信号、周期信号、噪声信号等，从而对时间序列的结构进行分析，并可进一步预测[10]。

奇异谱分解的主要步骤分为4步：

（1）嵌入；

（2）SVD分解；

（3）分组；

（4）重构。

利用SSA训练模型的多步预测结果如图30所示，预测误差如图31所示。

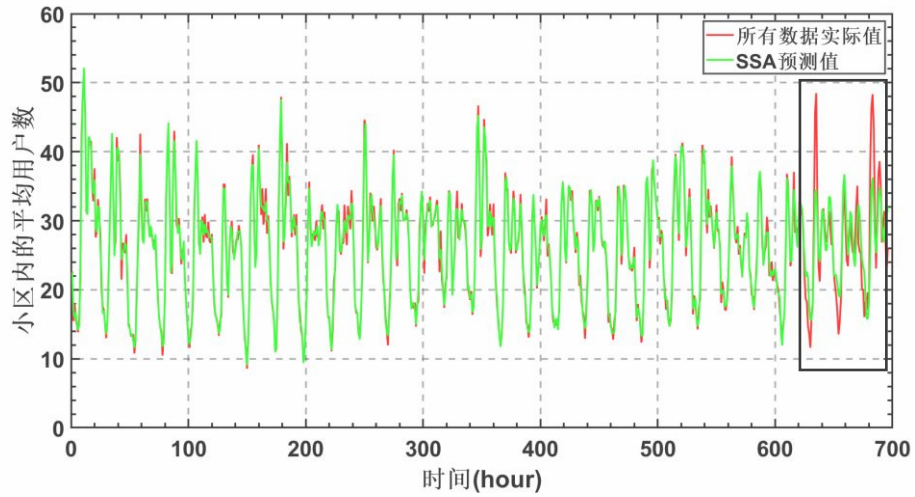


图30 SSA多步预测（第1个小区）

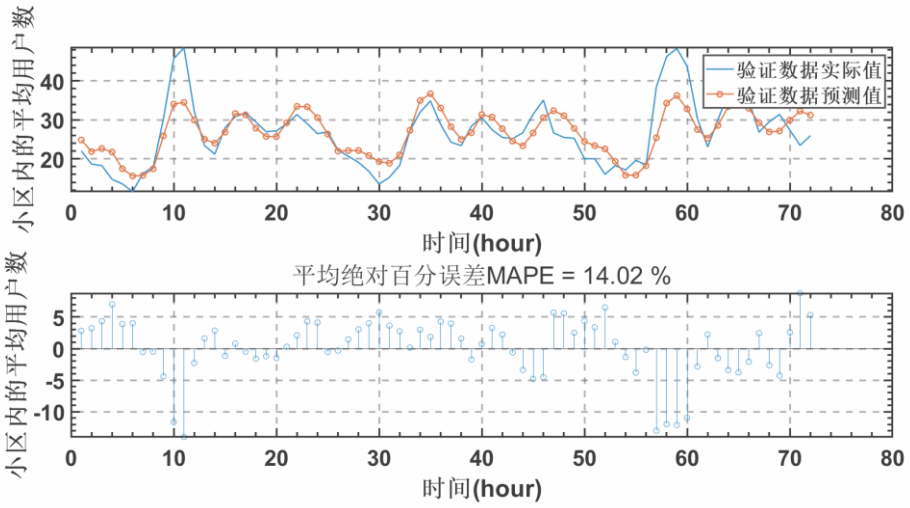


图31 SSA多步预测误差（第1个小区）

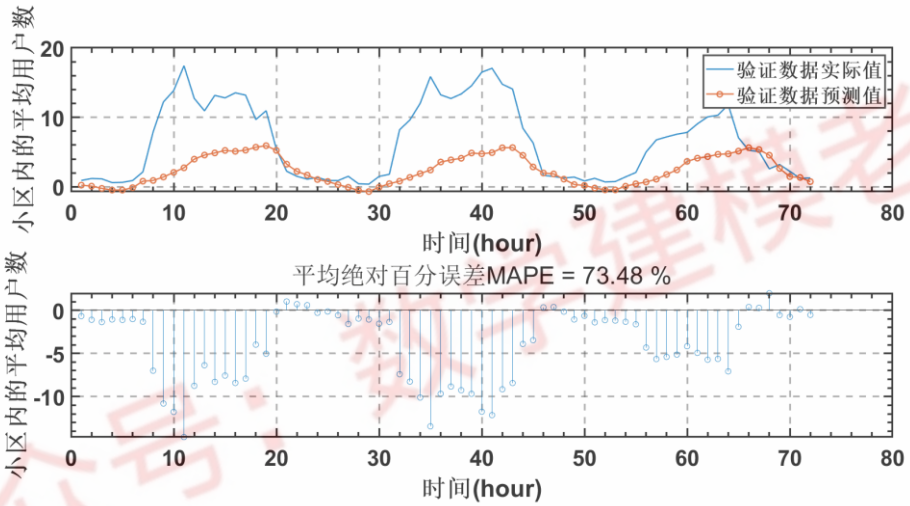


图32 SSA多步预测误差（第10个小区，SSA参数不同）

从图31中可以看出，SSA多步预测的整体趋势预测较好，其平均绝对百分误差MAPE为14.02%，但缺点就是对模型参数很敏感，同一个参数对所有数据可能不适用，其预测结果就会很差，如图32所示。这就需要调整重构分量输入数据序列的数量和数字划分季节值，这两个参数目前没有理论来计算出最优的参数，只能通过经验并不断测试，因此本文拟考虑其他预测方法。

3.4 SARIMA 预测模型

基于上文分析，本文进一步尝试采用季节性自回归滑动平均(SARIMA)模型进行数据预测。SARIMA模型源于单整自回归移动平均模型（ARIMA模型）。ARIMA模型适用于非季节时间序列短期预测。在某些时间序列中存在明显的周期性变化，这种周期是由季节性变化（包括季度、月度等变化）或其他一些固有因素引起的，这类序列称为季节性序列[11]。

预测步长为72小时，根据上文的分析，季节周期设置为24小时，同时考虑到数据的关联性，对数据进行多元线性回归，如图33所示。

首先，本文对不同阶数的SARIMA模型的预测效果进行了比较，发现SARIMA(1,0,0)的精度最高。基于SARIMA(1,0,0)模型，采用MATLAB的估计函数，估计SARIMA(1,0,0)模型

的参数，并根据AICBIC准则进行了模型检验。图34展示了通过另外两个指标的预测值进行线性回归得到的当前变量预测值，与当前变量自回归预测值，以及真实值的对比。

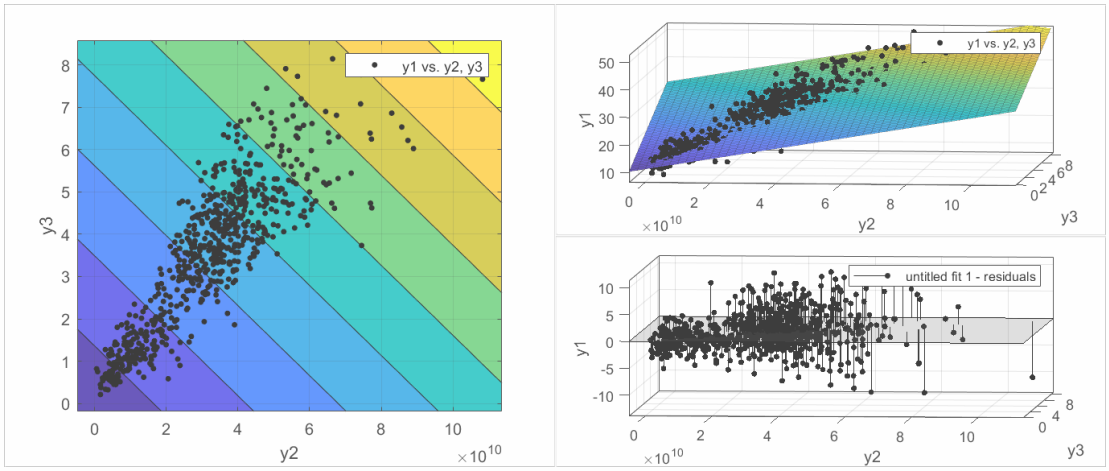


图33 三个KPI指标的多元线性回归结果

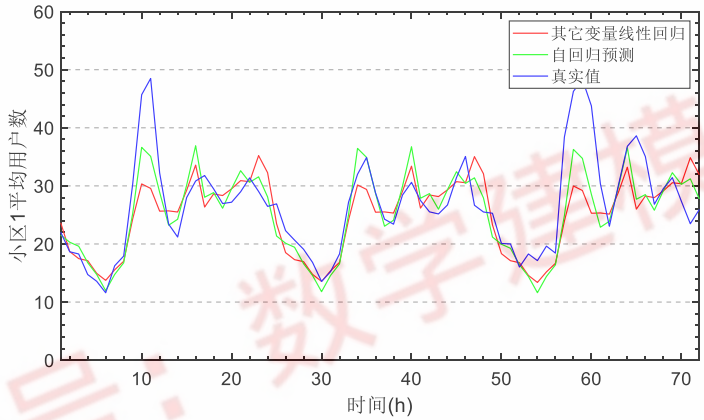


图34 多元线性回归模型、季节性自回归滑动平均、真实值的对比

然后在验证集上利用多元线性回归预测结果和SARIMA模型预测结果进行加权融合，例如，第1个小区的平均用户数指标的误差如图35所示，其平均绝对百分误差MAPE为11.34%，较LSTM和SSA等方法误差更小。

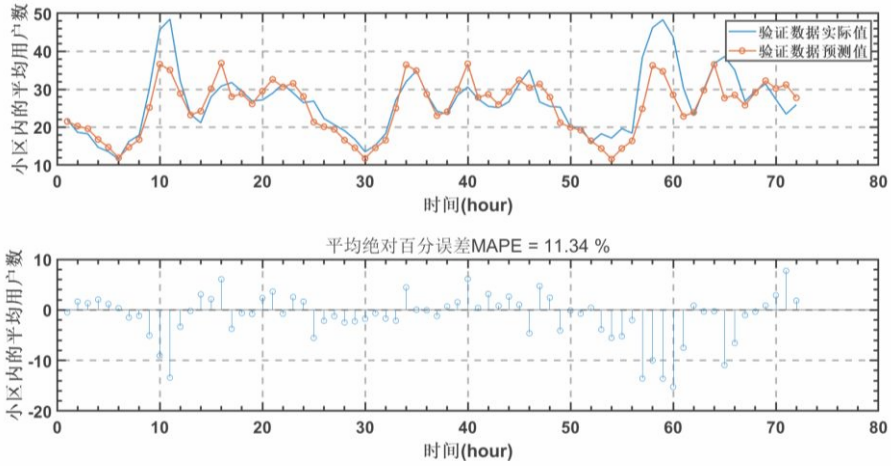


图35 多元线性回归预测结果和SARIMA模型预测结果进行加权融合误差（第1个小区）

最后，选用多元线性回归预测结果和SARIMA模型预测结果进行加权融合的方法对未来3天的数据进行预测，预测结果见支撑材料-附件2中的预测值表格。

参考文献

[1] 王速,卢华,汪硕,蔡磊,黄韬. 智能运维中KPI异常检测的研究进展[J]. 电信科学,2021,37(05):42-51.

[2] 费业泰. 误差理论与数据处理[M]. 机械工业出版社, 2004:105-114.

[3] https://zh.wikipedia.org/wiki/%E5%88%86%E4%BD%8D%E6%95%B0

[4] https://ww2.mathworks.cn/help/stats/boxplot.html

[5] Huang G B,Zhu Q Y,Siew C K．Extreme learning machine：theory and applications[J]．Neurocomputing,2006,70(1)：489-501．

[6] Gosso A .Training of a SLFN (Single Hidden-layer Feedforward Neural Network)[J].

[7] 缪希仁, 范建威, 江灏,等. 基站异常情况下基于改进极限学习机的超宽带室内定位方法[J]. 传感技术学报, 33(10):10.

[8] Kulmer J , Hinteregger S , Grosswindhager B , et al. Using DecaWave UWB Transceivers for High-accuracy Multipath-assisted Indoor Positioning[C]// IEEE ICC Workshop on Advances in Network Localization and Navigation (ANLN). IEEE, 2017.

[9] 方志强, 王晓辉, 夏通. 基于长短期记忆网络的售电量预测模型研究[J]. 电力工程技术, 2018, 037(003):78-83.

[10]温冬琴,王建东. 基于奇异谱分析的机场噪声时间序列预测模型[J]. 计算机科学,2014,41(01):267-270.

[11]陈玉霞. 基于SARIMA模型的贵州省季度GDP预测[J]. 经营与管理,2021(08):170-175.

附录

本文所有代码均由MATLAB软件编写，如需验证程序，请使用最新版本。

代码名称	代码解决的问题	代码电子版所在的位置
problem1.m	问题一异常检测的主程序	支撑材料中“问题一程序”文件夹
fun2Quantile.m	问题一分位数全局异常检测的子程序	支撑材料中“问题一程序”文件夹
fun1Quantile.m	问题一同比指标异常检测的子程序	支撑材料中“问题一程序”文件夹
abnormalCycle.m	问题一检测异常周期的索引值的子程序	支撑材料中“问题一程序”文件夹
abnormalCyclesAndOutliers.m	问题一检测异常周期和异常孤立点的子程序	支撑材料中“问题一程序”文件夹
problem2.m	问题二异常预测的ELM的主程序	支撑材料中“问题二程序”文件夹
elmtrain.m	问题二ELM训练的子程序	支撑材料中“问题二程序”文件夹
elmpredict.m	问题二ELM预测的子程序	支撑材料中“问题二程序”文件夹
Plot_Figure.m	问题一和问题二中部分画图程序	支撑材料中“问题一程序”和“问题二程序”文件夹
lstm_main_V1.m	问题三LSTM单步预测	支撑材料中“问题三程序”文件夹
lstm_main_Vmulti.m	问题三LSTM多步预测	支撑材料中“问题三程序”文件夹
SSA_V1.m	问题三SSA多步预测主程序	支撑材料中“问题三程序”文件夹
ssaformod.m	问题三SSA多步预测子程序	支撑材料中“问题三程序”文件夹
SARIMA_predict.m	问题三SARIMA模型预测趋势	支撑材料中“问题三程序”文件夹
AICBIC.m	问题三AICBIC检验确定模型阶次	支撑材料中“问题三程序”文件夹

完整代码较多，处于篇幅考虑，附录仅包含各问题中关键函数的完整代码，其他代码见支撑材料。