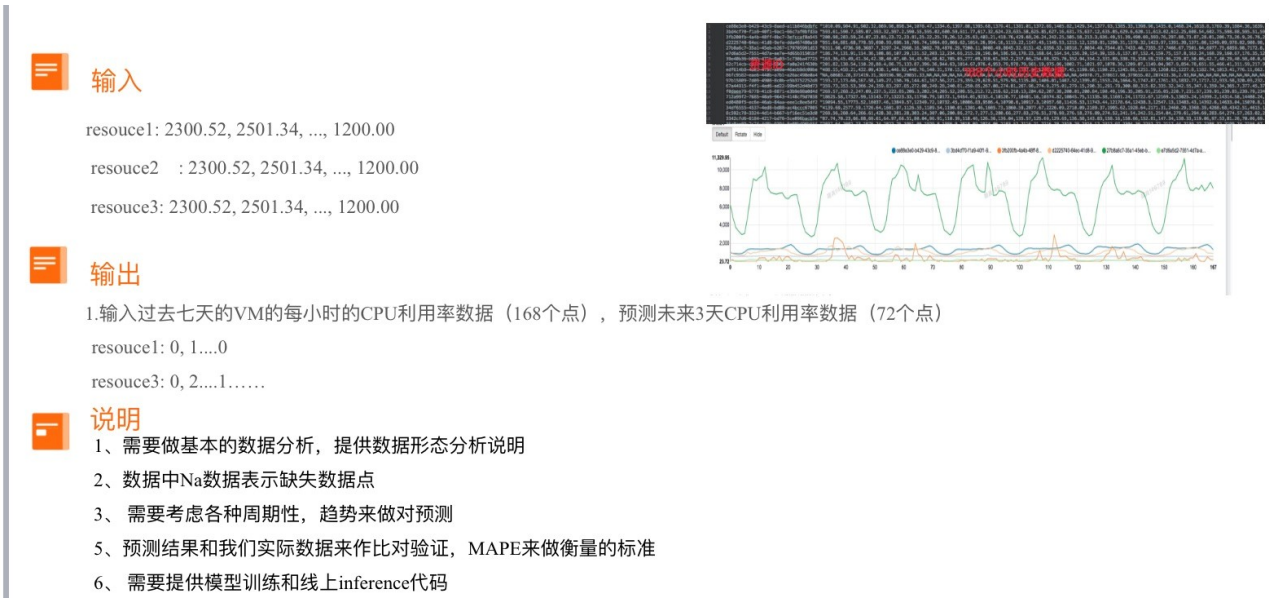
弹性计算高校挑战赛-智能容量规划

兰州大学2016级基础数学本科生 黎婕

赛题要求



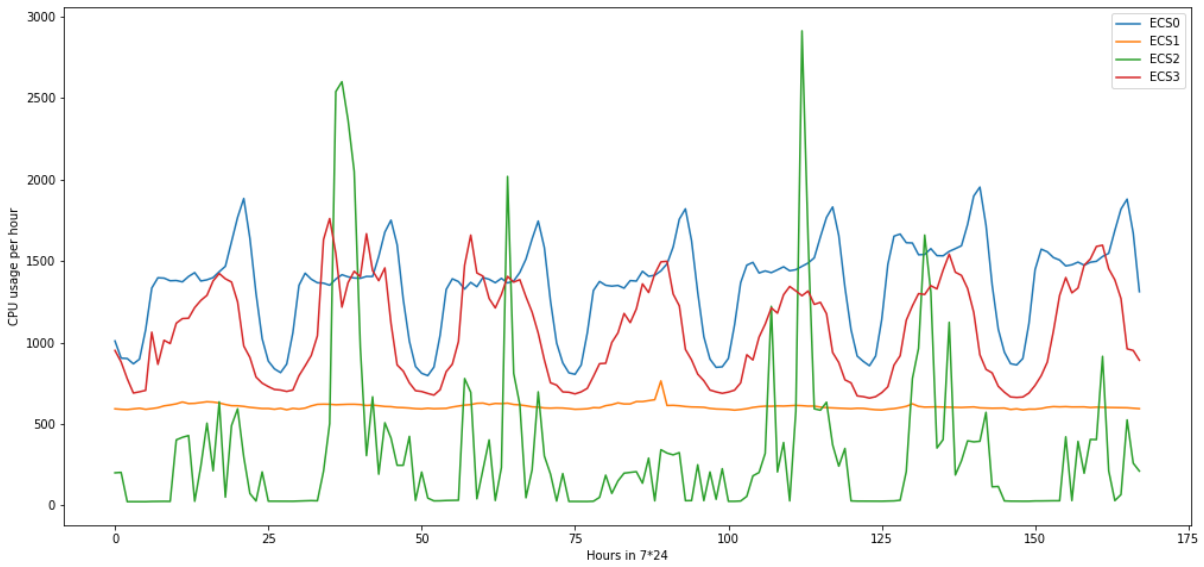
赛题数据分析

- 数据集基本信息

给定数据集中共有120个样本，每个样本有7*24(168)个数据点，需要对每个样本预测3*24(72)个数据点。

- 实际背景分析

样本为VM过去七天内CPU每小时利用率序列。不同样本间我认为是**独立**(例如:华北用户a实例的CPU利用率序列与华东用户b实例的CPU利用率序列可以认为是独立的)。通过对数据的可视化分析，不同样本间确实会出现形态差异极大的情况。



VM过去7天CPU利用率序列可视化

- 特征分析

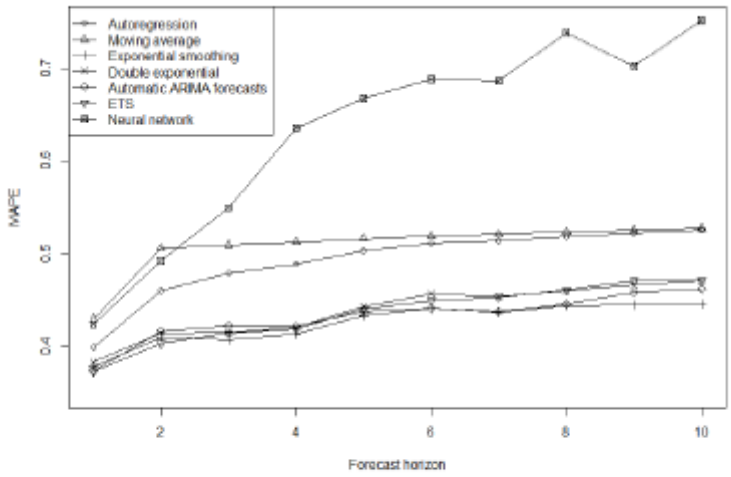
根据云服务器产品的性质，**周期**与**趋势**是其重要特征，结合MA(Moving average,移动平均)模型，这里我选择**24h**作为普适性的周期。

MA模型下以24h作为时间窗口能得到该CPU利用率序列平滑的趋势线

模型分析

比较模型

根据Time Series Forecasting of Cloud Data Center Workloads for Dynamic Resource Provisioning论文中对谷歌云CPU每小时使用率序列预测的MAPE误差分析如下图所示



(a) Rolling forecast origin cross-validation for Google cluster traces-1 hour

通过观察发现，指数类模型在该序列上的MAPE最低，效果最佳。
但经充分考虑，我决定使用**季节性ARIMA**模型进行建模(在该图上表现非最优)，原因如下:

- 1 原文作者对CPU使用率序列通过KPSS检测后应用ARIMA模型，未考虑到**季节性**因素，导致效果非最优
- 2 原文作者仅考虑4种参数组合的ARIMA模型，即ARIMA(2,d,2),ARIMA(0,d,0), ARIMA(1,d,0), ARIMA(0,d,1)。(d通过KPSS检测确定)，未能充分进行超参数搜索。
- 3 通过改进超参数搜索流程，我能够将指数类模型纳入季节性ARIMA模型中，以提升其综合效果。
- 4 未采用热门的神经网络模型是考虑到120个样本数量较少，且形态差异较大(可能会出现欠拟合)。如果考虑神经网络模型，建议数值归一化并避免数值零点下溢。

季节性ARIMA

基本介绍

本题我选择**季节性 ARIMA**(Auto Regressive Integrated Moving Average)统计模型，其数学表示为 $ARIMA(p,d,q)(P,D,Q)s$。

参数介绍如下:

- $p(P)$是模型的自回归部分。回归即表明依据过去的值做决策。这类似于声明如果过去七天CPU使用率极高，那么未来三天可能依旧极高。
- $d(D)$是模型的差分阶数。对时间序列进行差分可以得到更加平稳(stationary)序列。

- $q(Q)$是模型的移动平均部分。可用于估计序列的总体趋势(如上图)。
- s是模型的周期性部分，根据以上分析，我们设置$s = 24$。

事实上，ARIMA是诸多时间序列模型的泛化，如

- ARIMA(0,1,1)是指数平滑模型
- ARIMA(0,2,2)是双指数平滑模型

  ...

因此季节性ARIMA模型能够很好地捕捉时间序列模型的周期与趋势特征，对未来进行较准确的预测。

模型训练

超参数确定

如模型介绍所述，模型超参数共有7个，其中s=24根据周期性确定。待确定的超参数形式为(p,d,q,P,D,Q)，即使每个超参数仅考虑2种，也会出现组合爆炸(2^6种组合)。
通过对CPU使用率序列进行KPSS检验，仅有13个序列的p值超过0.05，其余均符合平稳序列假设。因此设置差分阶数d=D=1对所有序列进行平滑化处理，基本能够通过KPSS检验。
根据该链接对R语言中的auto.arima()介绍，其仅考虑以下4种组合:

- ARIMA(2,d,2)
- ARIMA(0,d,0)
- ARIMA(1,d,0)
- ARIMA(0,d,1)

提交方案中利用网格搜索优化，在综合**模型效果**与**搜索时间**考虑，在**36**种可能组合中枚举最优模型

```
ps = qs = range(0, 3)
Ps = Qs = range(0,2)
d = D = [1]
s = 24

#Create a list with all possible combinations of parameters
parameters = product(ps, d, qs, Ps, D, Qs)
```

模型拟合

确定模型超参数后，我以季节性ARIMA模型的AIC(Akaike Information Criterion)作为衡量模型效果的标准，

$$\mathrm{AIC} = -2\log(L) + 2(p + q + k + 1)$$

AIC越小，则模型效果越优异(类似极大似然估计)。
根据官网Properties介绍，ARIMA提供AIC,MAE,MSE等多种评估方案，限于比赛时间短且训练时间较

长，本文选择接受度最广的AIC。

优化代码

```
def optimize_SARIMA(series, parameters_list, s):
    """
        Return dataframe with parameters and corresponding AIC

        parameters_list - list with (p, d, q, P, D, Q) tuples
        d - integration order
        D - seasonal integration order
        s - length of season
    """

    best_aic = float('inf')

    for param in tqdm_notebook(parameters_list):
        try: model = sm.tsa.statespace.SARIMAX(series, order=(param[0], param[1], param[2]),
                                               seasonal_order=(param[3], param[4], param[5], s))
        except:
            continue

        aic = model.aic

        #Save best model, AIC and parameters
        if aic < best_aic:
            best_model = model
            best_aic = aic
            best_param = param

    return best_param, best_aic, best_model
```

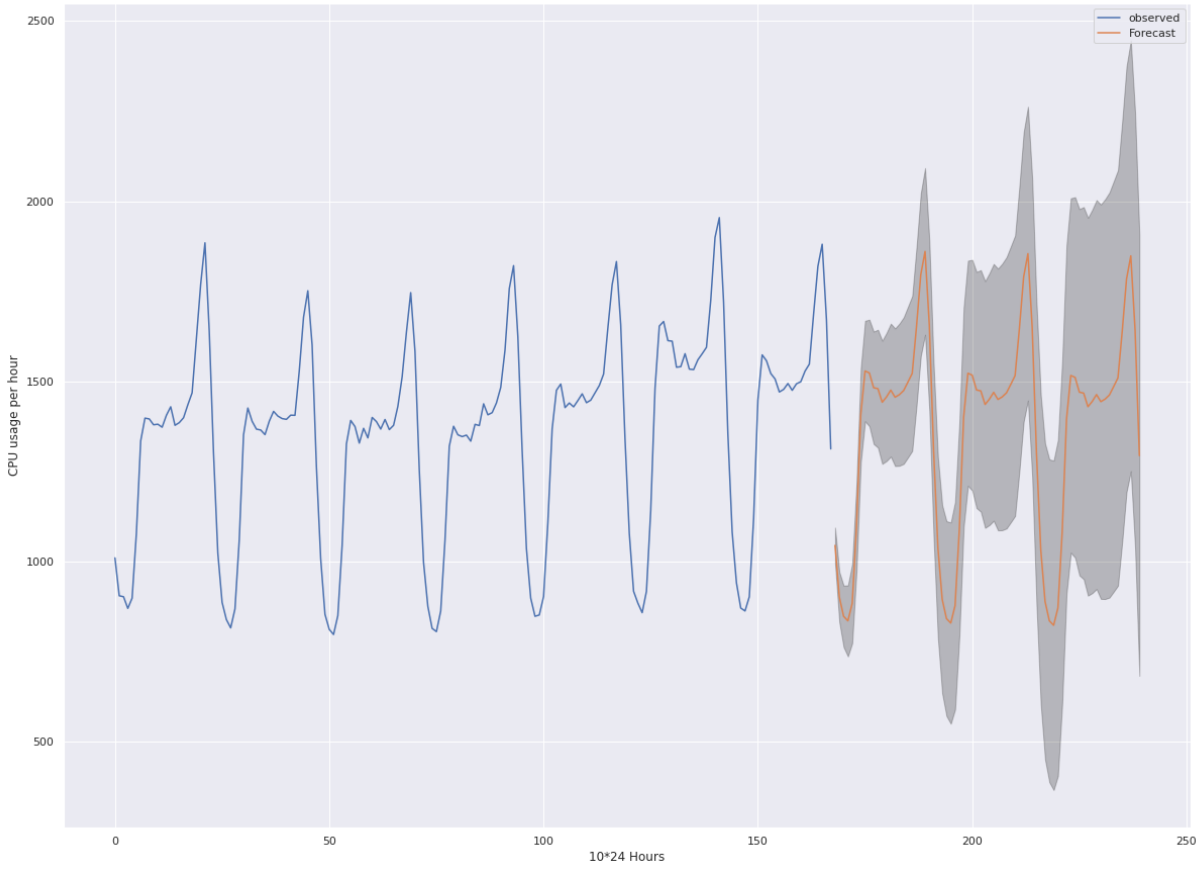
ecs.c5(2c4g)对单样本训练(网格参数搜索+模型拟合)时间为1min 27s ± 159 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)。

模型预测

季节性ARIMA的预测结果包括均值及置信区间

```
pred_72 = best_model.get_forecast(steps=3*24)

# Get confidence intervals of forecasts
pred_72_ci = pred_72.conf_int()
```

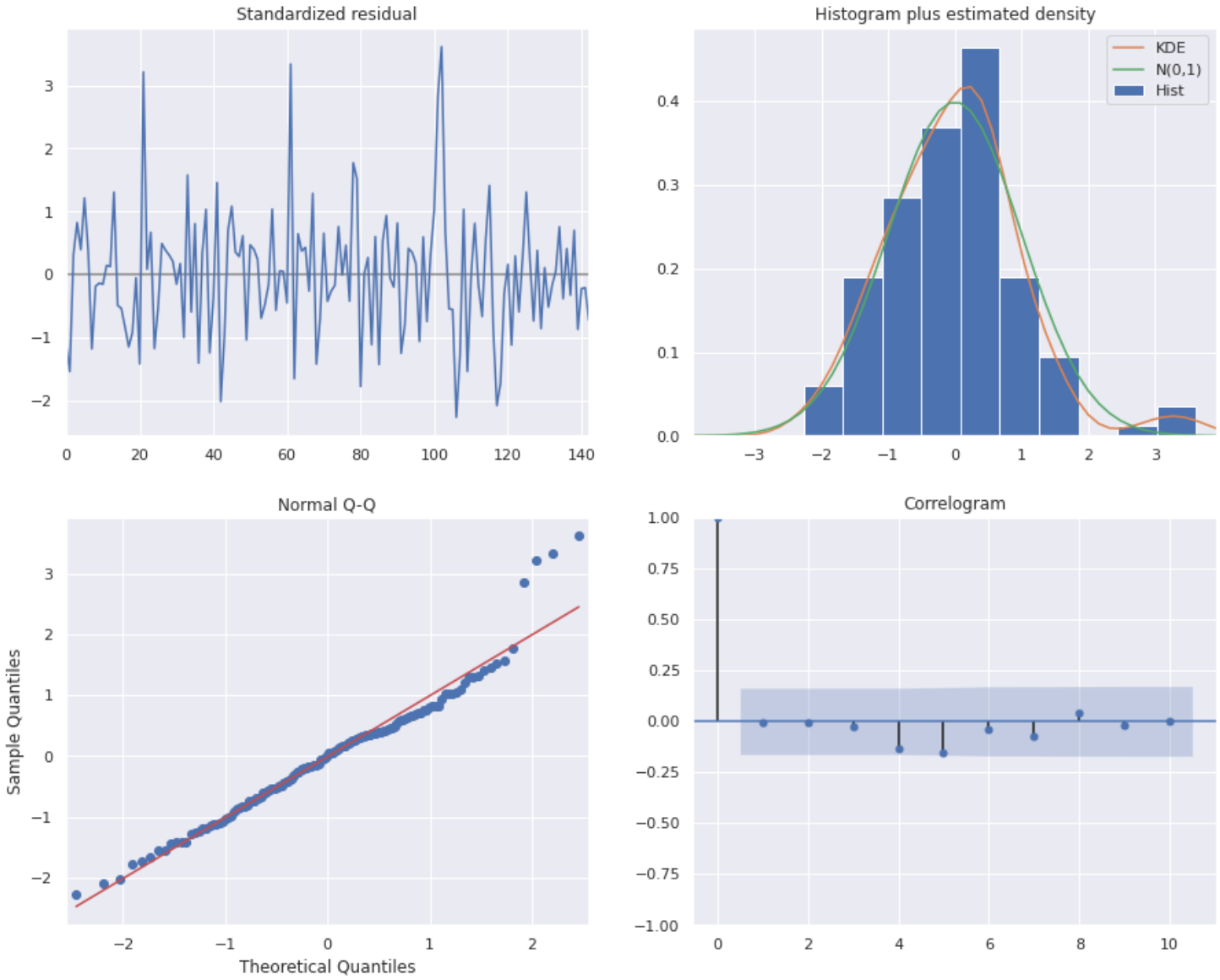


ce88e3e0-b429-43c9-8aed-a11b846bdbfc预测结果 蓝线-7天历史 黄线-3天预测 灰色95%置信区间

对于云服务业务，预测均值并不是最终目的，ARIMA提供的**置信区间**能够为是否扩容提供充分的参考性信息。

模型评估

残差分析



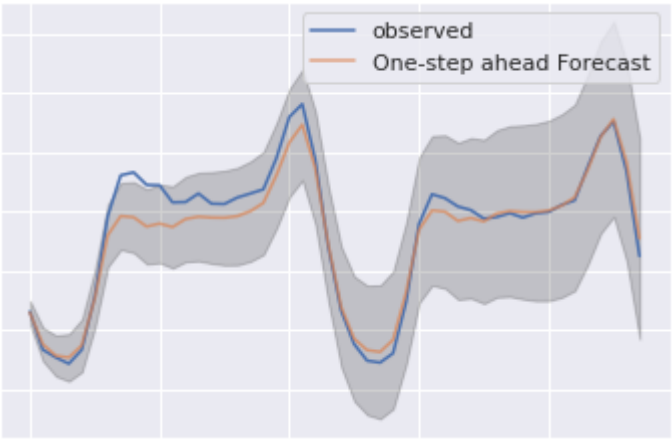
ce88e3e0-b429-43c9-8aed-a11b846bdbfc残差分析图

通过观察该实例的残差分析图，1-1表明其残差呈白噪声分布，2-2进一步表明时间序列残差与自身的滞后版本具有较低的相关性。1-2很好地表明了残差呈正态分布，而2-1的Q-Q图进一步验证了这点。

MAPE评估

尽管我的季节性ARIMA模型接收7天历史数据进行训练，但其内置的预测方法可仅根据5天历史数据对未来两天进行**动态**预测，通过预测值与实际值进行比对，能够提供模型验证

```
pred = best_model.get_prediction(start=5*24,  dynamic=True, full_results=True)
pred_ci = pred.conf_int()
```

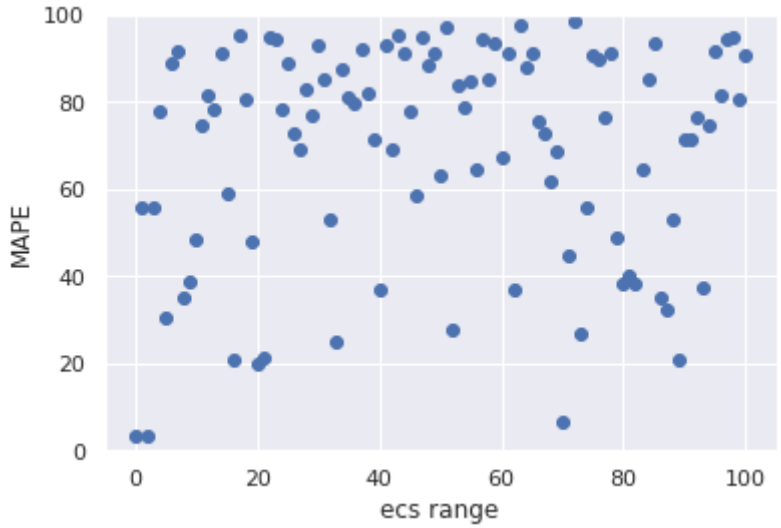


ce88e3e0-b429-43c9-8aed-a11b846bdbfc 6-7天预测与实际值比对

经计算，该实例的MAPE误差为3.6%，达到高精度预测的水平。

整体评估

为了评估具有普适性，我对115个完整CPU使用率进行如上MAPE评估(根据前5天的数据，预测后两天的数据)



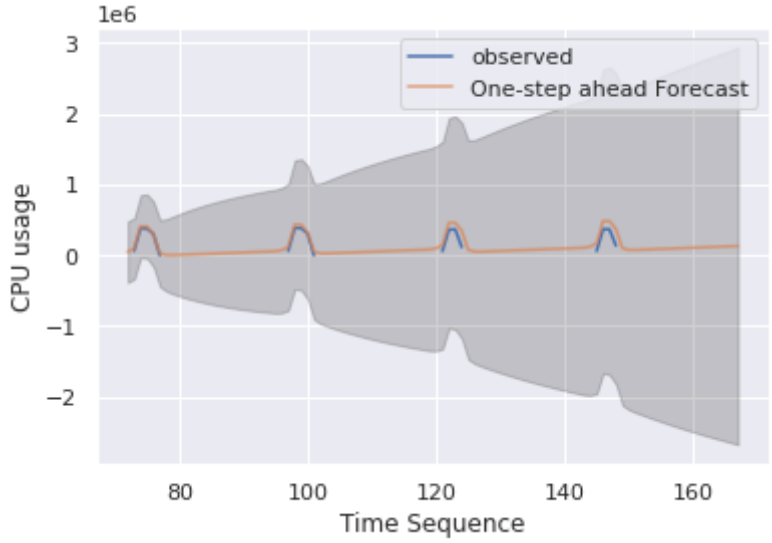
剔除离群值后，101个CPU使用率序列MAPE误差分布图

经计算，该101个CPU使用率序列的MAPE平均值为68.5%，能够达到对CPU使用率良好预测的效果。

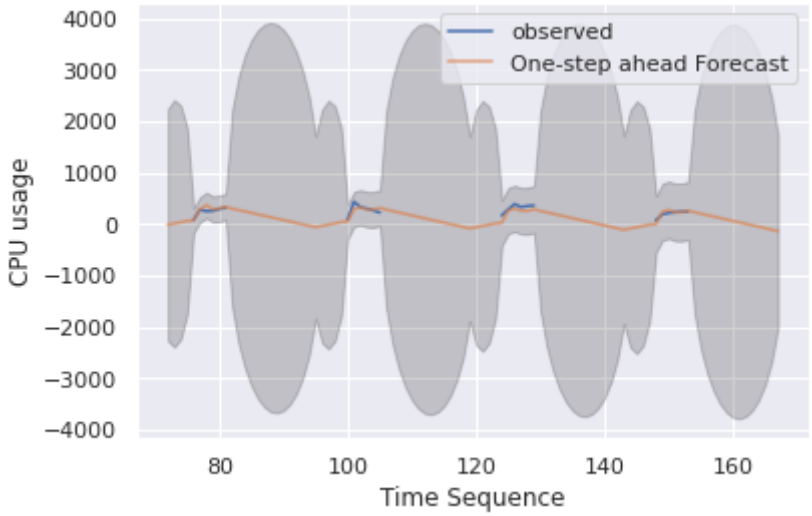
缺失值处理

对于时间序列预测中原始数据存在缺失值的情况，我设置差分阶数为0。其官方建议如下:
对于在缺失数据集上估算的模型，AIC选择ARMA（1,0,1）

出于搜索最优结果考虑，仍然对36种组合模型进行搜索选择最优AIC，但规避差分带来的损失。



86fc9582-eae6-440b-a7b1-a26ac498e0a4预测评估



77bc4464-b5fa-4e47-bfd1-ae4e4fcd5199预测评估

缺失真的太严重了...最终提交结果时对预测序列应用relu函数处理负数预测。

展望

季节性ARIMA模型属于统计模型范畴，因此其预测结果为均值形式，对于本题以预测结果正确率作为评分标准并不算占优，但其提供置信区间对于实际业务(是否需要扩容，预警通知)有着重要意义。在提交方案中本人也就超参数选择、模型损失函数标准进行详细的讨论，为各位技术人员用季节性ARIMA对CPU时间序列预测建模与改进方面提供思路。