小批量物料生产安排的建模与分析

林进川　张威健
（漳州城市职业学院教师教育系，福建 漳州 363000）

摘要：本文讨论 2022 年“高教社杯”全国大学生数学建模竞赛 E 题“小批量物料的生产安排”可行性。本文在不同业务场景下对小批量物料生产安排建立周预测模型，提供了 TOPSIS 评价法从中选取六项建立周预测模型对物料的周需求量进行预测、ARIMA 时间序列预测和 XGboost 回归进行预测，比对模型预测结果，选择预测效果较高的方法。考虑到库存量与服务水平之间平衡性，进而建立二向平衡性周预测模型，可以得到较好的预测结果。

关键词：物料控制；TOPSIS 评价；ARIMA 模型；XGBoost 回归

一、问题背景

电子产品制造企业面临以下问题：在多品种小批量的物料生产中，事先无法知道物料的实际需求量。如果按照物料需求量的预测值来安排生产，可能会产生较大的库存，或者出现较多的缺货，给企业带来经济和信誉方面的损失。可以从需求量的预测值、需求特征、库存量和缺货量等方面综合考虑，以便更合理地安排生产。

二、问题分析

本文主要解决多品种小批量物料生产管理问题。要求依据所提供的物料需求数据，对库存、缺货情况、投用时间区间等进行分析，提出一套合理的物料生产计划 .

问题一对于数据的预处理，本文首先计算出频数、需求量、销售均值和销售总价，并基于此 TOPSIS 评价法从中选取六项建立周预测模型对物料的周需求量进行预测。在对附件数据进行筛选后，先使用 ARIMA 模型，利用 IQR 四分位距绘制箱型图观察数据离散程度，发现存在若干异常值，再基于随机森林 RF 将异常值剔除。之后摘取若干经过预处理后的数据设置训练集与测试集，并利用 ADF 自相关检验与 Ljung-Box 检验观察数据集是否平稳。然后通过观察自相关图与偏自相关图观察数据平稳性表现，利用 SPSSPRO 做 XGBoost 回归最后通过 ARIMA 时间序列预测与 XGboost 回归模拟结果的比对对模型做出评价，选择准确度较高的方法建立模型。

问题二利用 Python 选取物料 6004020918（序号 d）的前 100 周数据作为训练集，再根据题目设置步阶为 2，作时序数据滑窗转换，而后基于机器学习仿真 XGBoost 回归，预测第 101~110 周的生产计划。在保证服务水平大于或等于 0.85 的基础上，建立物料需求的周预测模型，就可以得到表 1 数据，再继续利用以上方法作另外 5 种物料的生产计划，取平均值作表 2 数据。

三、模型的建立与求解

（一）问题一的模型建立与求解

1. 数据预处理

（1）计算需求频数与需求量

本文首先借助 R 软件对物料数据进行筛选处理，使用 table 函数计算物料需求频数与需求量得到结果，部分数据如表 1、2 所示。此外本文还通过 Python 的 time 模块得到物料数据所在周数。且对物料数据进行了整理，得到各种物料频数、数量、销售单价均值与销售总价均值的具体内容，以确保所得数据的可靠性，如表 3 所示。

表 1 物料需求频数（部分）

	物料编码	频数
1	6004020503	1224
2	6004010256	955
3	6004020375	794
4	6004020918	620
5	6004020374	612
6	6004100008	573
7	6004020656	540
8	6004020418	531
9	6004020504	509
10	6004020622	468

表 2 物料数量（部分）

	物料编码	数量
1	6004010068	604
2	6004010116	153
3	6004010121	4
4	6004010134	4
5	6004010174	2601
6	6004010203	142
7	6004010205	93
8	6004010207	2064
9	6004010215	208
10	6004010217	189

表 3 各物料详细信息

物料编码	频数	数量	销售单价的均值	销售总价的均值
6004010068	6	604	892.9939231	86741.02775
6004010116	32	153	703.7524386	3363.353432
6004010121	3	4	757.6972287	1010.262972
6004010134	3	4	1832.359083	2439.511499
6004010174	418	2601	1296.900566	8102.245459
6004010203	69	142	1016.392385	2091.421305
6004010205	44	93	1082.228075	2287.436613
6004010207	153	2064	962.1260938	12953.95595
6004010215	40	208	664.9324013	3454.617889

（2）TOPSIS 评价法

进行结果评定时，各个影响因素的权重大小都是不一致的，

为了使各个指标间信息差效果最大化，本文首先基于熵权法对各项进行赋权，确定比重。熵权法作为一种客观性比较显著的赋权手段，在实践过程中，可以基于给定指标数据的离散程度，利用信息熵计算出各指标的熵权，再根据各指标对熵权进行一定的修正，就可以得到较为客观的指标权重，其公式如下所示。

$$H_j = -\sum_{i=1}^{m} f_i \cdot \ln f_i \qquad (1)$$

借助 SPSSPRO 计算各指标数据结果如表 4 所示。

表 4 各指标数据权重

项	信息熵值 e	信息效用值 d	权重（%）
W1	0.806	0.194	34.558
W2	0.896	0.104	18.436
W3	0.942	0.058	10.298
W4	0.794	0.206	36.707

D+ 和 D– 值分别代表评价对象与最优或最劣解（即 A+ 或 A–）的距离，也被称为欧式距离。评价对象与最优或最劣解的距离值越大说明距离越远，研究对象 D+ 值越大，说明与最优解距离越远；D– 值越大，说明与最劣解距离越远。D+ 值越小同时 D– 值越大。

综合度得分 C 值，C=（D–）/（D++D–），计算公式上，分子为 D– 值，分母为 D+ 和 D– 之和；D– 值相对越大，则说明该研究对象距离最劣解越远，则研究对象越好；C 值越大说明研究对象越好。

而后基于这些权重，我们分析后认为 TOPSIS 优劣解距离法优于成分分析法，故选取使用。TOPSIS 法是一种常用的综合评价方法，其能充分利用原始数据的信息，其结果能精确反应各评价方案之间的差距，其公式如下所示。

$$\frac{x-\min}{\max-\min} = \frac{x-\min}{(\max-x)+(x-\min)} \qquad (2)$$

借助 SPSSPRO 基于 TOPSIS 评价法进行计算，结合观察需求量与销售单价趋势图部分结果如表 5。从中选取综合表现最好的前 6 位， 即 6004020503、6004010256、6004010252、6004020918、6004010321、6004021055 作为重点关注物料，并将其命名为（n_）a~（n_）f。

表 5TOPSIS 评价结果（部分）

索引	正理想解距离（D+）	负理想距离（D–）	综合得分指数	排序	命名
6004020503	0.176876533	0.306550683	0.634119622	1	a
6004010256	0.225019322	0.216145892	0.48994319	2	b
6004010252	0.275000403	0.245983544	0.472151868	3	c
6004020918	0.223396375	0.182629633	0.449797869	4	d
6004010321	0.277706622	0.213611385	0.434772147	5	e
6004021055	0.247274475	0.187821289	0.431678045	6	f
6004010174	0.240659581	0.176340692	0.422879081	7	–
6004010372	0.264494737	0.185824705	0.41265086	8	–
6004020375	0.266645822	0.170009763	0.389345216	9	–
6004020656	0.267779799	0.139034615	0.34176423	10	–

2. 建立周预测模型

（1）处理异常值

建立以周为单位的时间预测模型首先要对数据当中的异常值进行处理。经过反复尝试，我们发现 InterquatileRange 四分位距与随机森林（RandomForest）能够更好地忽略异常值，较少的受其影响，因此选取进行异常值的处理。

InterquatileRange 四方位距，即 IQR、四分位差，是描述统计学中的一种常见方法。与方差、标准差相同，用于表示统计资料中各变量分散情形。也通常被用于构建箱形图，并进一步对概率分布的简要图表概述，其公式如下所示。

$$IQR=Q_3-Q_1 \qquad (3)$$

$$L_p = (n)(\frac{p}{100}) \qquad (4)$$

注：p= 四分位的百分比值，n= 样本总量

如果 L 是一个整数，则取第 L 和第 L+1 的平均值。

如果 L 不是一个整数，则取下一个最近的整数。（比如 L=2.2 则取 3）

通过构建并观察箱型图，如图 1 所示，可以发现：数据分布出现了零星的离散情况，这意味着数据当中出现了异常值，需要将其剔除。这里本文采用了随机森林处理异常值。

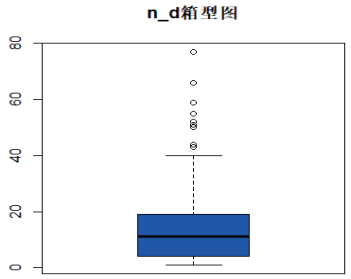

图 1 n_dIQR 箱型图

随机森林顾名思义，就是随机地建立一个包含很多决策树的森林，也就是一个包含多个决策树的分类器，其输出的类别是由个别树输出的类别的众数而定。其步骤如下所示。

Step1：用 N 来表示训练用例（样本）的个数，M 表示特征数目。

Step2：输入特征数目 m，用于确定决策树上一个节点的决策结果；其中 m 应远小于 M。

Step3：从 N 个训练用例（样本）中以有放回抽样的方式，取样 N 次，形成一个训练集（即 bootstrap 取样），并用未抽到的用例（样本）作预测，评估其误差。

Step4：对于每一个节点，随机选择 m 个特征，决策树上每个节点的决定都是基于这些特征确定的。根据这 m 个特征，计算其最佳的分裂方式。

Step5：每棵树都会完整成长而不会剪枝，这有可能在建完一棵正常树状分类器后会被采用。

以数据 n_d 为例，本文首先将给定的时间区间（2019 年 1 月 2 日至 2022 年 5 月 21 日）分解为 176 个周单位，由此可得需求变化的粗略趋势。而后通过 IQR 四分位距绘制箱型图得到若干个异常值并将其处理为缺失值，利用 R 语言的 mic 函数使用随机森林法将缺失值补全。此时可以发现异常值已经被剔除，可进行下一步操作。

（2）设置训练集与测试集

从共计 151 个数据中选取 1–130 行作为训练集以及 131–151

行作为测试集。

（3）平稳性检验

本文通过 ADF 检验与 Ljung-Box 检验进行平稳性检验。

ADF 检验用于检验序列中是否存在单位根，若存在单位根就表示序列为非平稳时间序列。单位根就是指单位根过程，若序列中存在单位根过程就不平稳，会使回归分析中存在伪回归。

Ljung-Box 检验即 LB 检验、Box 白噪声检验，用来检验 m 阶滞后范围内序列的自相关性是否显著，或序列是否为白噪声，Q 统计量服从自由度为 m 的卡方分布。若表现为白噪声，则该数据没有价值，可以放弃分析。其公式如下所示。

$$Q(m)=T(T+2)\sum_{i=1}^{m}\frac{r_i^2}{T-l} \qquad (5)$$

当 Q（m）$>x_a^2$ 时拒绝零假设，即认为序列中存在某些自相关。本文利用 R 实现 LB 检验计算程序会直接给出 P 值，此时 $p>a$ 时拒绝零假设，a 为显著性水平。

在进行 ADF 检验与 LB 检验后分别得到 P 值均明显小于 0.05，如图 2、3 所示，即 dn_d 训练集表现为平稳序列，非白噪声数据。

```
> adf.test(n_2train)

        Augmented Dickey-Fuller Test

data:  n_2train
Dickey-Fuller = -3.5176, Lag order = 5, p-value = 0.04327
alternative hypothesis: stationary
```

图 2 ADF 检验结果 P 值

```
> Box.test(n_2train, type = "Ljung-Box")

        Box-Ljung test

data:  n_2train
X-squared = 9.505, df = 1, p-value = 0.002049
```

图 3 LB 检验结果 P 值

（4）模型判断

对于模型判断我们采用了 AutocorrelationFunction 自相关函数与 PartialAutocorrelationFunction 偏自相关函数进行平稳性检验，其公式如下所示。

$$r_k=\frac{\sum_{t=1}^{N-k}(x_t-\bar{x})(x_{t+k}-\bar{x})}{\sum_{t=1}^{N}(x_t-\bar{x})^2} \qquad (6)$$

$$PCorr(Xt,Xt+k\mid Xt+1,...,Xt+k-1) \qquad (4)$$

首先利用自相关函数与偏自相关函数制作 dn_d 训练集的自相关图与偏相关图，如图 4、5 所示。观察发现 dn_d 需求量表现平稳，体现出明显季节性与周期性。

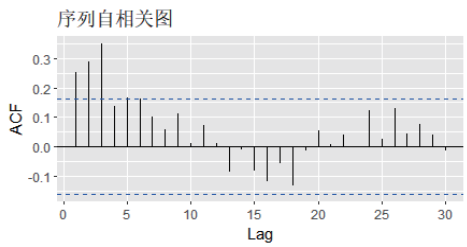

图 4 n_1$ 需求量自相关图

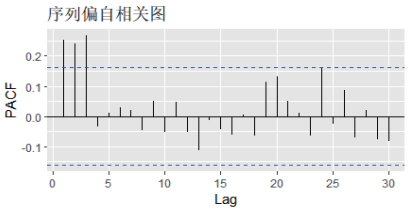

图 5 n_1$ 需求量偏自相关图

（5）ARIMA 时间序列预测

通过 Auto.arima 函数自动对训练集进行定阶，可以得到结果为 ARIMA（p，d，q）为（0，1，1），在此基础上进行 ARIMA 时间预测得到预测值，具体结果如图 6、7 所示。

```
> auto.arima(n_2train)
Series: n_2train
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.7698
s.e.   0.0633

sigma^2 = 164.7:  log likelihood = -579.7
AIC=1163.39   AICc=1163.48   BIC=1169.36
```

图 6 自动求阶结果

```
> Box.test(ARMAn_2$residuals, type = "Ljung-Box")

        Box-Ljung test

data:  ARMAn_2$residuals
X-squared = 0.0018129, df = 1, p-value = 0.966
```

图 7 残差的白噪声检验结果

此时对残差进行 box 白噪声检验可以得到 P=0.966 显著大于 0.05，判断为白噪声数据。这意味着有效数据已经被充分提取，剩余数据有效性较低，进行下一步操作。

（6）评价

本文在评价部分采用了两种方法：一种是单时序预测（ARIMA），另一种是通过将单序列问题转为多序列问题，使用时序数据滑窗转换对变量进行拆分，然后使用机器学习模型（XGBoost）进行预测（XGboost 回归）的时间序列预测方法，并对两种方法的预测结果进行比对。

ARIMA 模型（AutoregressiveIntegratedMovingAveragemodel）即差分整合移动平均自回归模型，是时间序列预测分析方法之一。ARIMA 算法事实上就是处理数据中具有趋势性（trend），季节性（seasonal），以及场景周期性（domaincycle）的规律，将其从原始数据中剥离出来，得到最后的噪声数据，理想状态下得到的是白噪声，在 ARIMA 模型中（p，d，q）中，AR 指的是“自回归”，p 即为自回归项数；MA 指的是“滑动平均”，q 即为滑动平均项数，d 为令其成为平稳序列的差分次数（阶数）。其公式及步骤如下所示。

$$y_t=c+\sum_{n+1}^{N}\varnothing_{t-n}y_{t-n}+\varepsilon_t+\sum_{n+1}^{N}\theta_{t-n}\varepsilon_{t-n} \qquad (8)$$

XGboost 回归是 GBDT 的一种高效实现，和 GBDT 不同，XGboost 给损失函数增加了正则化项；且由于有些损失函数是难以计算导数的，XGboost 使用损失函数的二阶泰勒展开作为损失函数的拟合。实现 XGboost 回归的步骤如下所示。Step1 通过训练集数据来建立。

Step2 通过建立的 XGBoost 来计算特征重要性。

Step3 将建立的 XGBoost 回归模型应用到训练、测试数据，得到模型评估结果。

注：XGBoost 无法像传统模型得到确定的方程，通常通过测试数据预测精度来对模型进行评价。

本文通过 R 语言实现 ARIMA 模型的建立，结果如图所示。对于 XGBoost 回归模型的建立我们通过 SPSSPRO 实现，部分预测结果如图 8、9 与表 6 所示（具体结果请参见附件）。

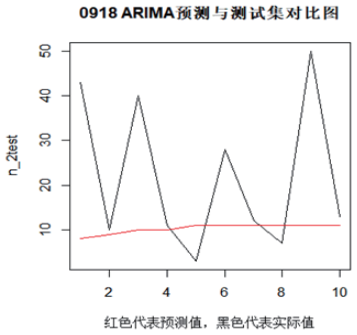

图 8 ARIMA 预测结果图

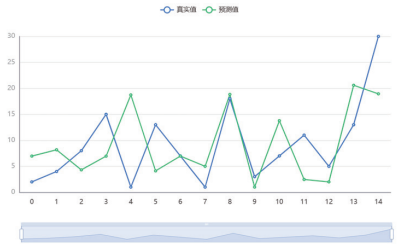

图 9 XGboost 预测结果图

表 6 XGBoost 预测结果表（部分）

预测结果 Y	时序变量转换_Y	时序变量转换_X1	时序变量转换_X2
27.26235580444336	11	9	4
38.48487091064453	7	4	11
13.154473304748535	30	11	7
12.015934944152832	12	7	30
39.37858200073242	14	30	12
2.5944950580596924	63	12	14
27.59190559387207	17	14	63
16.669025421142578	12	63	17
9.810935974121094	17	17	12
21.404890060424805	6	2	17
17.713777542114258	50	17	6
16.049400329589844	61	6	50
15.8612642288208	64	50	61
15.8612642288208	65	61	64
15.8612642288208	12	64	65

但是经过观察后发现 ARIMA 观测结果拟合度较低，与预期不符，数据不具有实际价值。故本文采用 XGboost 回归进行模型的评估，结果如表 7 所示。得到训练集 R2 值为 0.997，显著接近 1，因此判断评估结果准确度较高。

表 7XGboost 评价结果表

	MSE	RMSE	MAE	MAPE	R^2
训练集	0.575	0.758	0.26	3.513	0.997
测试集	178.351	13.355	9.211	230.548	-0.159

（二）问题二的模型建立与求解

首先我们提取 d 物料 1 至 100 周的数据作为训练集，添加时序数据滑窗转换——步阶 =1，2，做 XGboost 回归，结果显示 R2=0.9371316403090997 显著接近 1，说明结果准确度较高。

然后利用周预测模型可以得到 101 至 110 周生产计划、实际需求量、库存量以及缺货量的具体信息。最后计算服务水平，即，结果如表 8 所示。

表 8 6 种物料的综合结果（1）

物料编码	平均生产计划数（件/周）	平均实际需求量（件/周）	平均库存量（件/周）	平均缺货量（件/周）	平均服务水平
a	22.24	22.53	17.80	0.00	100.00%
b	10.92	11.55	5.84	0.08	99.27%
c	34.13	34.13	0.00	0.00	100.00%
d	10.39	11.31	0.92	0.92	95.67%
e	144.27	144.27	0.00	0.00	100.00%
f	38.53	38.54	0.04	0.00	99.99%

四、模型的分析

（一）问题一的误差分析

该问题在模型建立与求解阶段已经进行过预测值与实际数据的对比，发现 ARIMA 时间序列预测结果产生误差较大，而 XGBoost 回归预测结果产生误差较小，

（二）问题二的误差分析

本模型使用 XGBoost 回归来对 6 种物料的需求作预测。观察表 8 我们可以发现，实际值与预测值的误差数值如表 9 所示，可以发现其误差最大值为：0.92。

表 9 问题二误差值

物料编码	平均生产计划数（件/周）	平均实际需求量（件/周）	差值
a	22.24	22.53	0.29
b	10.92	11.55	0.63
c	34.13	34.13	0
d	10.39	11.31	0.92
e	144.27	144.27	0
f	38.53	38.54	0.01

参考文献：

[1] 薛毅，陈立萍 . 统计建模与 R 软件 [M]. 北京：清华大学出版社，2007：89-95.

[2] 肖风华 . 零基础 Python 从从入门到精通 [M]. 广州：广东人民出版社，2019：204.

[3] 王斌会 . 多元统计分析及 R 语言建模（第五版）[M]. 北京：高等教育出版社，2020：322-340.

[4] 杨磊，汪辉辉，许涛，蒋健伟，常琪 . 实时化电池热失控检测方法、系统、装置及介质 [P]. 上海市：CN114779085A，2022-07-22.

[5] 司守奎，孙兆亮 . 数学建模算法与应用 [M]. 北京：国防工业出版社，2021：418-421.

[6] 芮少权，匡安乐 . 高速公路月度交通量 ARIMA 预测模型 [J]. 长安大学学报（自然科学版），2010，30（04）：82-85+91.

基金项目：2021 年度福建省中青年教师教育科研项目（科技类），项目编号：JAT210935.