某大型汽车制造厂备件销售案例预测

赵赞豪　高　敏　胡　敏
（指导老师：肖枝洪）

重庆理工大学

1. 确定业务问题

1.1 案例背景

在“十一五”期间，我国的汽车行业得到了迅速的发展，面对全球性的金融危机冲击，在产业政策的作用下，我国汽车已经成为国民经济的重要支柱，2010 年我国汽车销量达到 1800 万台。在“十二五”期间，内需发展已经被提上重要日程，居民的收入水平得到加快发展，生活水平将迈上新的台阶。在近期召开的 2016 中国汽车论坛上，清华大学经济管理学院弗里曼经济学讲席教授李稻葵认为“十三五”将是中国汽车行业发展的最好的时期。由此预计未来 5-10 年间，整个汽车市场的潜在增长率还处在较高的水平，汽车工业仍然保持持续增长的态势，中国交通部部长翁梦勇估计到 2020 年中国的民用汽车保有量将达到 1.4 亿辆。汽车保有量持续的飞跃式增长必然引发汽车备件市场的繁荣。可以预见，在未来的一二十年中，汽车备件的消费将会持续的增长。

中国目前是全球第一大进出口国，第二大对外投资国，第一大吸引外资国家。未来五年中国经济增长速度应该平均接近 7%，甚至更高，今年第一季度中国经济出现了一个小阳春，增长速度趋于稳定，甚至各种指标比去年四季度还高一点。最可喜的实体经济增长从去年下半年 6.2%恢复到 6.5%，实体经济指的是去除金融业的贡献之外的狭义的实体经济增速。去年如果没有金融业的贡献增速就会跌到 6.2%，因为去年金融业占到了 10%左右附加值，金融业 15.9%的增长，所以今年一季度实体经济出现了比较好的回暖的态势。但无论如何发展，“十三五”期间最晚到 2018 年经济能够触底往上行。

在汽车相关的各产业中，汽车服务业占有重要的地位，其也是汽车产业链中稳定的利润来源。目前，中国的汽车服务业的发展还在逐步成熟中，伴随着我国汽车保有量不断的增加，我国居民在汽车消费观念上也越来越成熟，这就为汽车服务行业带来新的契机，特别是涉及其中的汽车金融、汽车售后、汽车物流、二手车交易等汽车服务行业都必定将加速发展。在 2009 年颁布的《汽车产业调整和振兴规划》中，已经明确提出要“发展现在汽车服务行业。加快发展汽车研发、

生产性物流、汽车零售和售后服务、汽车租赁、二手车交易、汽车保险、消费信贷、停车服务、报废回收等服务行业，完善相关的法规、规章和管理制度”。

一般来说，汽车由四大部分所组成，分别是发动机、底盘、车身和电器设备，这四大部分分别是由各个零部件的总成、各个零部件的分成和单个的零部件组合所装配起来。组成汽车的这些零部件总成、零部件分成和单个的零部件，以及在汽车使用中自然消耗和需要补充的或为维持汽车良好状态必需更换的零部件总成、零部件分成和单个的零部件，把它统称为汽车备件。

汽车售后备件物流是指将汽车售后备件从零部件供应商或者汽车主机处组织供应到汽车售后零部件消费者手中的全过程。配件的仓储、运输、信息集成等汽车备件的物流体系。由于汽车备件的型号繁多、种类复杂、且所需备件的备件数量和时间随机性较大、网店数量众多、终端需求量小等多种特点，使得面向汽车售后市场的备件物流系统的运作较为复杂，其物流运作的要求也远远高于整车物流。

而且随着中国汽车制造业逐渐进入微利时代和国内汽车售后市场的迅速成长，国内汽车备件售后物流也越来越收到大众的关注。

根据中国汽车工业协会的统计显示，从我国的汽车行业发展趋势来看，2001 年国产轿车价格降幅约为 3%-15%，2002 年国产轿车价格降幅达到 5%-6%，2003 年国产轿车整体价格水平降低了 8%-10%。到 2004 年以后我国的汽车价格大战不断加剧，行业竞争激烈，从而导致整车销售的利润降低。在整个汽车利润组成中，销售、配件、维修的利润比例约为 2:1:4,国内外各大汽车厂家也逐渐意识到整车销售并不是利润的最主要来源。配件和维修同属于售后市场的备件服务环节，因此，汽车售后市场备件服务的水平在很大程度上影响一个汽车生产商的利润。精于市场动态调查的丰田汽车公司早就发现：消费者一旦享受到某种品牌汽车的汽车厂商提供的最佳售后服务，就会对该品牌的汽车赞不绝口，在客观上为该厂商大力做广告宣传，就回带来很多潜在的消费者。由此可见，即便汽车各方面性能都很好，但如果却忽视了售后服务这一关键的环节，最终会对汽车的销售和品牌信誉造成负面的影响。反之亦然。

对于广大消费者来讲，最容易损害某种品牌汽车信誉的莫过于无法从销售商中买到该汽车的关键零部件。这就涉及到汽车售后备件物流的管理。世界汽车制造企业都十分重视其物流管理，把其看作是节约成本，拓宽利润空间的有效渠道。汽车物流的配送作业是需要供应各个环节必需衔接平滑的高技术行业，在国际物流领域，汽车物流被公认为是最复杂、最具专业性的领域。对于广大汽车制造厂商而言，无论是争取到公司产品生产的最大利益化还是提高公司在业界口碑和形象方面上，合理预测安排汽车售后备件物流是有重大作用的。因此，着手对我国汽车售后备件物流的预测进行讨论，具有重要的实际运用意义。

1.2 理解业务问题

某汽车备件销售规模不断增大，需建立企业级的备件销量预测管理平台，提高人员工作效率的同时，也能提升预测精准度，优化库存。此案例的主要需求是测试每个备件外推 4 个月的月度销量的预测准确性，同时实现产品的批量预测和分层预测。该问题主要是一个预测的问题，这种预测问题可以通过时间序列进行实现。在实现预测的同时，也要兼顾产品的批量预测与分层预测，那么我们可以用 SAS 的宏来实现批量预测与分层预测。时间序列预测完了之后，我们要对预测的结果与真实值进行比较，得到预测的准确度。

总结如下：1.为什么说是时间序列呢？从数据的收集是按照时间顺序获取的。2.要用到聚类分析。因为物料很多，应该分类。3. 按照分类预测。

1.3 提出业务问题

此案例的主要需求是实现每个备件外推 4 个月的月度销量的批量预测和分层预测，并给出预测的准确度。

1、将源数据导入 SAS 并进行清洗，如缺失值和异常值处理、产品替换等；

2、探索数据的特性，据此对产品进行分类，进一步构建相关的特征变量并检验特征变量的合理性；

3、利用 SAS 分层批量预测每个配件未来 4 个月的销售订单数量，时间粒度是月度，产品维度是物料编号，区域维度是应发库，预测变量是销售订单数量，可进一步添加事件或者自定义模型优化预测效果。

1.4 检查数据的可行性

表 1　部分原始数据的展示

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	创建日期	物料编号	物料描述	K3编号	产品组	客户物料号	销售组织	交货单号	是否拣配	交货单创建	交货单修改	交货单行项	销售大区	销售小区	发运渠道
2	2013.05.1	10482258-	冷媒_R134	20.07.01(	公共产品组	F007	X990	82099258	完全拣配	QCXSSD011	SP2WMS	3	深圳大区	广东一区	深圳中转
3	2013.05.1	10261297-	EG-37030	119.06.04(	公共产品组	F606	X990	82099261	完全拣配	QCXSSD011	SP2WMS	20	深圳大区	广东一区	深圳中转
4	2013.05.1	10134443-	483QA-10	119.01.11(	公共产品组	F601	X990	82099567	完全拣配	QCXSSD011	QCXSSD011	2	济南大区	山东一区	济南中转
5	2013.05.1	10195108-	纽扣电池_	123.04.10(	公共产品组	G304	X990	82099567	完全拣配	QCXSSD011	QCXSSD011	4	济南大区	山东一区	济南中转
6	2013.05.1	10261297-	EG-37030	119.06.04(	公共产品组	F606	X990	82099583	完全拣配	QCXSSD011	QCXSSD011	5	济南大区	山东一区	济南中转
7	2013.05.1	10134443-	483QA-10	119.01.11(	公共产品组	F601	X990	82099617	完全拣配	QCXSSD011	QCXSSD011	11	杭州大区	江西一区	长沙中转
8	2013.05.1	10746359-	汽油清净齐	29.08.01(	C6	G608	X990	82099625	完全拣配	QCXSSD011	SP2WMS	2	天津大区	山西一区	深圳中转
9	2013.05.1	10189123-	BYDQ831A	119.07.01(	公共产品组	F607	X990	82099692	完全拣配	QCXSSD011	QCXSSD011	4	杭州大区	江西二区	长沙中转
10	2013.05.1	10482258-	冷媒_R134	20.07.01(	公共产品组	F007	X990	82099709	完全拣配	QCXSSD011	SP2WMS	2	天津大区	山西二区	深圳中转
11	2013.05.1	10534567-	S6-52060	127.06.12(	公共产品组	S606	X990	82099709	完全拣配	QCXSSD011	SP2WMS	3	天津大区	山西二区	深圳中转
12	2013.05.1	10134443-	483QA-10	119.01.11(	公共产品组	F601	X990	82099711	完全拣配	QCXSSD011	SP2WMS	2	天津大区	山西二区	深圳中转
13	2013.05.1	10189123-	BYDQ831A	119.07.01(	公共产品组	F607	X990	82099711	完全拣配	QCXSSD011	SP2WMS	7	天津大区	山西二区	深圳中转
14	2013.05.1	10194030-	M6-11093(	25.08.03(	公共产品组	M608	X990	82099714	完全拣配	QCXSSD011	SP2WMS	5	天津大区	山西二区	深圳中转
15	2013.05.1	10143983-	LK-41162(	20.03.19(	公共产品组	F003	X990	82099718	完全拣配	QCXSSD011	SP2WMS	2	天津大区	山西一区	深圳中转
16	2013.05.1	10303138-	G3-37914(	17.04.32(	公共产品组	0	X990	82099766	完全拣配	QCXSSD011	SP2WMS	1	天津大区	山西一区	深圳中转
17	2013.05.1	10127489-	EG-41342(	19.05.36(	公共产品组	F605	X990	82099769	完全拣配	QCXSSD011	QCXSSD011	1	长春大区	内蒙古	沈阳中转
18	2013.05.1	10551879-	S6-29056(	27.03.01(	公共产品组	S603	X990	82099769	完全拣配	QCXSSD011	QCXSSD011	14	长春大区	内蒙古	沈阳中转
19	2013.05.1	10201891-	CQ02_气门	19.05.50(	公共产品组	F605	X990	82099774	完全拣配	QCXSSD011	QCXSSD011	6	济南大区	山东三区	济南中转
20	2013.05.1	10551879-	S6-29056(	27.03.01(	公共产品组	S603	X990	82099774	完全拣配	QCXSSD011	QCXSSD011	12	济南大区	山东三区	济南中转
21	2013.05.1	10201891-	CQ02_气门	19.05.50(	公共产品组	F605	X990	82099775	完全拣配	QCXSSD011	SP2WMS	1	济南大区	山东三区	深圳中转
22	2013.05.1	10189123-	BYDQ831A	119.07.01(	公共产品组	F607	X990	82099777	完全拣配	QCXSSD011	QCXSSD011	4	济南大区	山东一区	济南中转
23	2013.05.1	10261297-	EG-37030	119.06.04(	公共产品组	F606	X990	82099777	完全拣配	QCXSSD011	QCXSSD011	9	济南大区	山东一区	济南中转
24	2013.05.1	10134443-	483QA-10	119.01.11(	公共产品组	F601	X990	82099779	完全拣配	QCXSSD011	SP2WMS	10	天津大区	内蒙古一区	西安中转
25	2013.05.1	10278322-	发动机油_	20.07.01(	公共产品组	F007	X990	82099779	完全拣配	QCXSSD011	SP2WMS	25	天津大区	内蒙古一区	西安中转

根据上表 1 所显示的结果，原始数据具有该案例进行数据分析的数据基础。在原始数据的系统里，包括汽车备件的销售明细、汽车备件的仓库以及汽车备件的主数据。原始数据中的汽车备件的销售明细是日期数据，但是题目中的时间粒

度是月份，那么我们就需要将日期数据转化为月份数据再进行分析。而且销售明细里面具有产品纬度和区域纬度。

本案例重点需要指出的是：我们要用 32 个历史数据才能得到外推的 4 个预测数据。尽管整合后的数据集里面是包含着 36 个数据的，但是有一部分是训练集。并且，该原始数据的缺失情况低于五分之一。

经过一系列的数据可行性的判断之后，我们确定了原始数据可以用来进行数据分析的后续工作。

2. 数据准备

2.1 创建数据挖掘的环境

在数据分析的过程中，我们首先就是要对数据的存储的环境要进行设定。我们利用 SAS 9.4 对原始数据的存储环境进行了设定。在本案例中，我们设定逻辑库 huifeng。首先建立逻辑库的物理路径为：“F:\重庆理工大学-理学院-赵赞豪\逻辑库”，然后我们再用 LIBNAME 语句在 SAS 环境中创建一个逻辑库。具体结果如下图 1 所示。

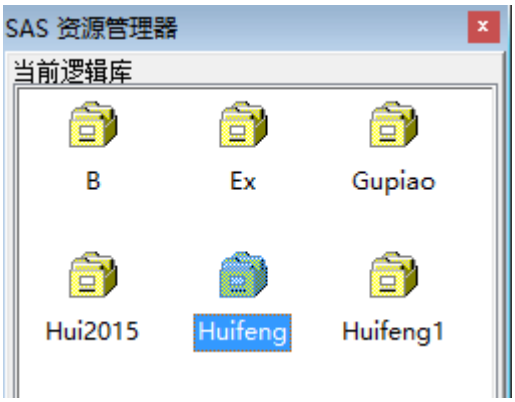


图 1　本案例在 SAS 环境中创建 huifeng 逻辑库

2.2 准备数据

从数据的来源来看，本案例中所使用的数据均为题目中给出的数据，只是我们将题目中给出的案例数据的文档换成了 CSV 格式的数据，方便 SAS 里面对数据进行导入。本文选取了 2013 年 5 月份到 2016 年 5 月份的销售明细里的销售订单数量和与它相关的数据进行数据建模。

2.3 观察并验证数据

原始数据集里面没有经过压缩，没有经过排序。就拿数据集 source_data 而言，它里面有 423214 个观测，有着 26 个变量。这些数据都为以后的建模提供了

数据的基础。

为了确定我们导入的数据和原始数据一样，我们同时打开了原始的 EXCEL 数据文件和 SAS 里的 7bdat 数据文件，两两相互比较一番之后，我们确定了原始数据和导入的数据里面不存在偏差。

3. 数据探索

3.1 数据集的结构探索

将 Excel 表格导入 sas 之后，我们需要对生成的数据集的基本结构进行探索，具体结果如图 2 所示。

CONTENTS PROCEDURE

数据集名	HUIFENG.SOURCE_DATA	观测	423214
成员类型	DATA	变量	26
引擎	V9	索引	0
创建时间	2017-10-03 11:21:12	观测长度	248
上次修改时间	2017-10-03 11:21:12	删除的观测	0
保护		已压缩	NO
数据集类型		已排序	NO
标签			
数据表示法	WINDOWS_64		
编码	euc-cn Simplified Chinese (EUC)		

图 2　销售明细数据集结构图

对应汽车备件销售明细数据集中，共有 423214 个观测，26 个变量。接下来我们就该数据集中的变量分定类与连续两类进行探索。

3.2 通过汇总信息对数据集进行探索

针对数据集中的 26 个变量的定类性与连续性（其结果如下表 2），我们将对它们分别进行分析：

表 2　变量的属性分析表

编号	变量	类型	角色	类型
1	创建日期	数值	输入变量	连续

2	物料编号	字符	输入变量	定类
3	物料描述	字符	输入变量	定类
4	K3 编号	字符	输入变量	定类
5	产品组	字符	输入变量	定类
6	客户物料号	字符	输入变量	定类
7	销售组织	字符	输入变量	定类
8	交货单号	数值	输入变量	连续
9	是否拣配	字符	输入变量	定类
10	交货单创建者	字符	输入变量	定类
11	交货单修改者	字符	输入变量	定类
12	交货单行项目	数值	输入变量	连续
13	销售大区	字符	输入变量	定类
14	销售小区	字符	输入变量	定类
15	发运渠道	字符	输入变量	定类
16	客户代码	数值	输入变量	连续
17	应发库	字符	输入变量	定类
18	采购订单编号	字符	输入变量	连续
19	销售订单号	数值	输入变量	连续
20	销售订单创建者	字符	输入变量	定类
21	订单类型	字符	输入变量	定类
22	销售订单行项目	数值	输入变量	连续
23	销售订单数量	数值	输出变量	连续
24	交货数量	数值	输入变量	连续
25	单位	字符	输入变量	定类
26	币别	字符	输入变量	定类

对于定类型变量，我们对数据进行了频数过程分析，结果如下图 3 和图 4 所示：

FREQ 过程

销售大区	频数	百分比	累积频数	累积百分比
北京大区	16379	3.87	16379	3.87
长春大区	24982	5.91	41361	9.78
成都大区	48513	11.47	89874	21.25
广州大区	22697	5.37	112571	26.62
杭州大区	57702	13.65	170273	40.27
济南大区	43925	10.39	214198	50.65
南京大区	31023	7.34	245221	57.99
上海大区	21928	5.19	267149	63.18
深圳大区	51593	12.20	318742	75.38
天津大区	30045	7.11	348787	82.48
武汉大区	46505	11.00	395292	93.48
西安大区	27578	6.52	422870	100.00
频数缺失 = 344				

图 3　销售大区的定类频数过程分析表

仔细观察本案例中数据集中的变量，我们会发现，其实定类变量有很多，我们运用宏，将它们频数分布批量输出，上面是我们对销售大区的一个频数统计的过程。其结果显示：销售大区包括北京大区、长春大区、成都大区、广州大区、杭州大区、济南大区、南京大区、上海大区、深圳大区、天津大区、武汉大区以及西安大区。还有他们的频数、百分比、累积频数以及累积百分比。

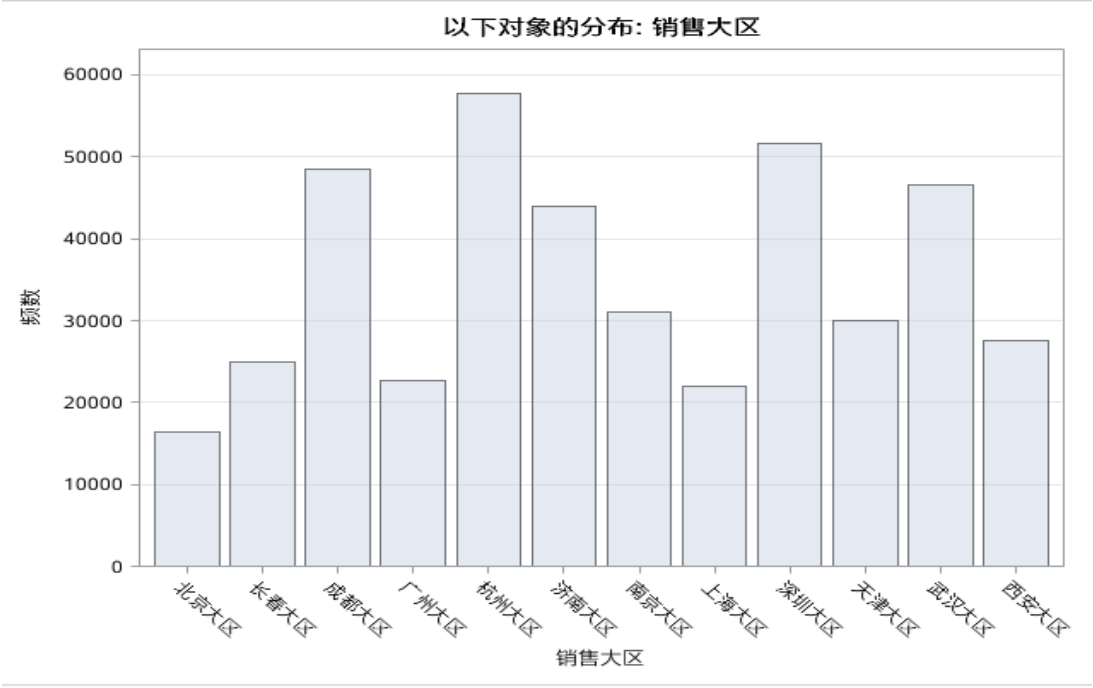


图 4　销售大区的定类频数直方图

对于连续型变量，我们可以应用 MEANS 过程步求出相应的观测个数、缺省个数、均值、中位数、标准差等相应的统计值，进一步了解该汽车备件的销售情况（其统计结果如下图 5 所示），为后面各月份总结销售订单数量的规律进而预测出 2016 年 2 月到 5 月的销售订单数量这一最终目标的实现做出准备。

MEANS PROCEDURE

变量	N	缺失值个数	最小值	均值	中位数	最大值	标准差
创建日期	423213	1	19508.00	20160.79	20213.00	20604.00	302.4413625
交货单号	423041	173	82099258.00	82982041.27	83017917.00	83730506.00	450785.97
交货单行项目	423041	173	1.0000000	83.0713902	18.0000000	4680.00	201.7577687
客户代码	423068	146	14363.00	18036.93	17925.00	80321.00	3068.04
销售订单号	423213	1	11326567.00	11978992.23	11994965.00	20066762.00	337926.62
销售订单行项目	423213	1	1.0000000	83.0728546	18.0000000	4680.00	201.7398489
销售订单数量	423203	11	1.0000000	15.3040881	6.0000000	900.0000000	24.3592597
交货数量	423203	11	0	15.0202716	6.0000000	900.0000000	24.2429810

图 5　连续型变量的 MEANS 过程分析表

3.3 变量的选择

我们选取时间维度为月份（日期数据可以整合成月份数据），产品维度是物料编号，区域维度是应发库，预测变量是销售订单数量，通过建立时间序列模型，对 2016 年 2 月到 5 月的销售订单数量进行预测。

4. 数据加工

4.1 原始数据集的分割

本文在对原始数据集的分割上采用了宏的方法，将物料编号相同的放在一起。具体步骤是：我们首先采取 timeseries 过程步，将原始数据集中的日期数据按照累积的方法变成月度数据，得到数据集 source2；然后利用宏编程将数据集 source2 进行拆分，分成 wuliao1 到 wuliao100 这 100 个数据集。具体拆分结果见下图 6。

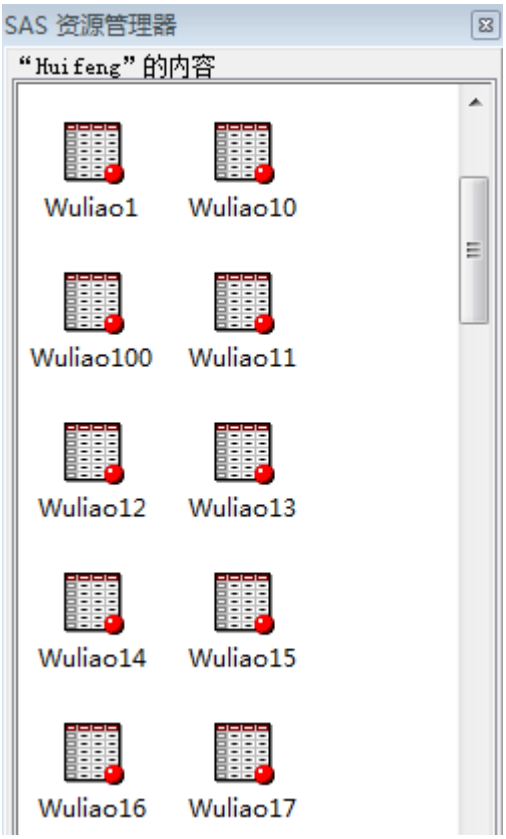


图 6 数据集的拆分的部分结果

在上述的图 7，我们可以看到数据集被拆分后的部分结果图形。被拆分的数据集中，每一个数据集就代表着一种物料编号的信息。其中的数据结构是包括物料编号和应发库两层的数据结构。

4.2 缺失值的处理

因为在题中所给的数据中，我们发现 100 种物料中，并不是每个物料对应的仓库的数量完全相等，存在缺失值，而且机器识别这个过程是录入数据，不能更改源文件中的数据格式。又因为此中缺失值如若直接舍弃会影响最后数据预测的准确性，在此处采用均值替代缺失值。即假如有一个日期这一天的销售数量为缺失，则用该天前一日的销售数量和该天后一日的销售数量的平均值代替这一天的销售数量。

4.3 变量的转化

由于时间排序数目庞大，按照日期来说使得数据处理起来非常费力，所以在此过程中我们采用 timeseries 过程步，将一个月中每天的销售数量加和并为该月的销售量，这样使得数据集中时间间隔为 month, 这样处理起来数据变得较为简单，便于计算统计。具体的变量的转化，见下表 3 和表 4。

表 3　未进行变量转化的结果

	创建日期	物料编号	物料描述
1	2013-05-30	10482258-00	冷媒_R134A_250g/瓶
2	2013-05-30	10261297-00	EG-3703010_蓄电池_M00000
3	2013-05-30	10134443-00	483QA-1017010_机油滤清器_M0000
4	2013-05-30	10195108-00	纽扣电池_CR1632
5	2013-05-30	10261297-00	EG-3703010_蓄电池_M00000
6	2013-05-30	10134443-00	483QA-1017010_机油滤清器_M0000
7	2013-05-30	10746359-00	汽油清净剂_G17_180ml/瓶
8	2013-05-30	10189123-00	BYDQ831A1106_子母扣_M00000
9	2013-05-30	10482258-00	冷媒_R134A_250g/瓶
10	2013-05-30	10534567-00	S6-5206010_前风挡玻璃总成_M000
11	2013-05-30	10134443-00	483QA-1017010_机油滤清器_M0000
12	2013-05-30	10189123-00	BYDQ831A1106_子母扣_M00000
13	2013-05-30	10194030-00	M6-1109302_空滤器滤芯_M00666
14	2013-05-30	10143983-00	LK-4116200A_右前雾灯_M00666
15	2013-05-30	10303138-00	G3-3791400_卡式智能钥匙_M00666
16	2013-05-30	10127489-00	EG-4134200_制动灯开关总成_M000
17	2013-05-30	10551879-00	S6-2905600_右前减阻尼器总成_M0
18	2013-05-30	10201891-00	CQ02_气门嘴总成_含帽盖
19	2013-05-30	10551879-00	S6-2905600_右前减阻尼器总成_M0
20	2013-05-30	10201891-00	CQ02_气门嘴总成_含帽盖

在上表 3 中，处理之前时间间隔为天数，这大大的不利于我们进行月度数据的分析。所以我们第一部要做的工作就是要将日期数据转换为月度数据。我们曾经一度将数据集先不整合成月份数据进行拆分，但是拆分之后，再来进行缺省值的处理时很麻烦，所以在经历很多次的失败之后（第一次用宏进行数据集的拆分，然后用 SQL 过程步对物料编号进行选取，但是到了缺省值的处理时非常难以再进行下去了，然后我们重新选取了方法再进行建模），选用 timeseries 过程步来进行数据的整合、日期数据与月份数据的转化以及变量的择取与转化。

下面是我们用新的方法重新对原始数据集的处理效果，详细请看表 4 的展示。

表 4　进行数据转化后的结果

	物料编号	应发库	创建日期	销售订单数量
1	10062683-00	北京	JUN2013	690
2	10062683-00	北京	JUL2013	570
3	10062683-00	北京	AUG2013	732
4	10062683-00	北京	SEP2013	1584
5	10062683-00	北京	OCT2013	1974
6	10062683-00	北京	NOV2013	2016
7	10062683-00	北京	DEC2013	2676
8	10062683-00	北京	JAN2014	1356
9	10062683-00	北京	FEB2014	324
10	10062683-00	北京	MAR2014	486
11	10062683-00	北京	APR2014	480
12	10062683-00	北京	MAY2014	510
13	10062683-00	北京	JUN2014	510
14	10062683-00	北京	JUL2014	612
15	10062683-00	北京	AUG2014	696
16	10062683-00	北京	SEP2014	1098
17	10062683-00	北京	OCT2014	1656
18	10062683-00	北京	NOV2014	2004
19	10062683-00	北京	DEC2014	2532
20	10062683-00	北京	JAN2015	1806

处理之后时间间隔为月份，然后数据集中，包括了所有的物料编号、所有的应发库、以及我们进行时间序列所要的时间和预测变量——销售订单数量。在数据集的拆分过程中，我们把物料编号相同的放到一个数据集中。一般情况，一个物料编号对应着 9 个应发库，还有我们的订单创建日期是从 2013 年 6 月份到 2016 年 5 月份。

4.4 聚类分析

聚类分析就是根据“物以类聚”的道理，对样本或指标进行分类的一种多元统计分析方法。讨论的对象是许多样本，合理地按各个样本的特性进行分类，而且没有任何模式可以参考。聚类分析首先根据样本的多个观测指标，具体找出可以度量样品或者指标间相似程度的一些统计量，然后利用这些统计量把样品或指标归类。聚类分析的目的就是把分类对象按照一定的规则划分成若干类，对类的数目和结构没有必要做任何假设。

下面我们运用层次法将产品有层次法聚成了几类。下面详见我们用 CLUSTER 语句进行的聚类分析的报表图 4.4.1：

CLUSTER 过程
Ward 离差平方和聚类分析

协方差矩阵的特征值				
	特征值	差分	比例	累积
1	2.85821E10		1.0000	1.0000

根均方总样本标准差	169062.5
观测之间的根均方距离	239090.5

图 4. 4. 1

聚类历史						
聚类数	连接聚类		频数	半偏 R 方	R 方	结值
99	11180297-00	11357755-00	2	0.0000	1.00	T
98	10499305-00	10746359-00	2	0.0000	1.00	T
97	10250530-00	10534447-00	2	0.0000	1.00	T
96	10527531-00	11102031-00	2	0.0000	1.00	T
95	10933037-00	11114099-00	2	0.0000	1.00	T
94	10798958-00	11043400-00	2	0.0000	1.00	T
93	10574838-00	10912463-00	2	0.0000	1.00	T
92	11025495-00	11316899-00	2	0.0000	1.00	T
91	11003344-00	11172268-00	2	0.0000	1.00	T
90	10968344-00	11209757-00	2	0.0000	1.00	T
89	10302013-00	CL94	3	0.0000	1.00	T
88	11065883-00	11328514-00	2	0.0000	1.00	T
.....						
12	CL61	CL22	4	0.0004	.998	
11	10194030-00	11188448-00	2	0.0005	.998	
10	CL21	CL13	71	0.0012	.996	
9	CL15	CL18	4	0.0014	.995	
8	CL98	11188450-00	3	0.0022	.993	
7	CL16	CL11	5	0.0032	.989	
6	CL14	CL12	16	0.0047	.985	
5	CL9	CL7	9	0.0154	.969	
4	CL10	CL6	87	0.0168	.953	
3	10134443-00	CL8	4	0.0635	.889	
2	CL5	CL3	13	0.1958	.693	
1	CL2	CL4	100	0.6934	.000	

图 4.4.2 聚类历史结果图

下面，SAS 系统会输出一个聚类的树状图，我们可以从树状图中知道，我们的产品（物料编号）可以被聚类分成三大种。

聚类的树状图详见图 4.4.3 所示。我们可以从中知道物料编号，根据它们每种的总销售额来进行的分类。

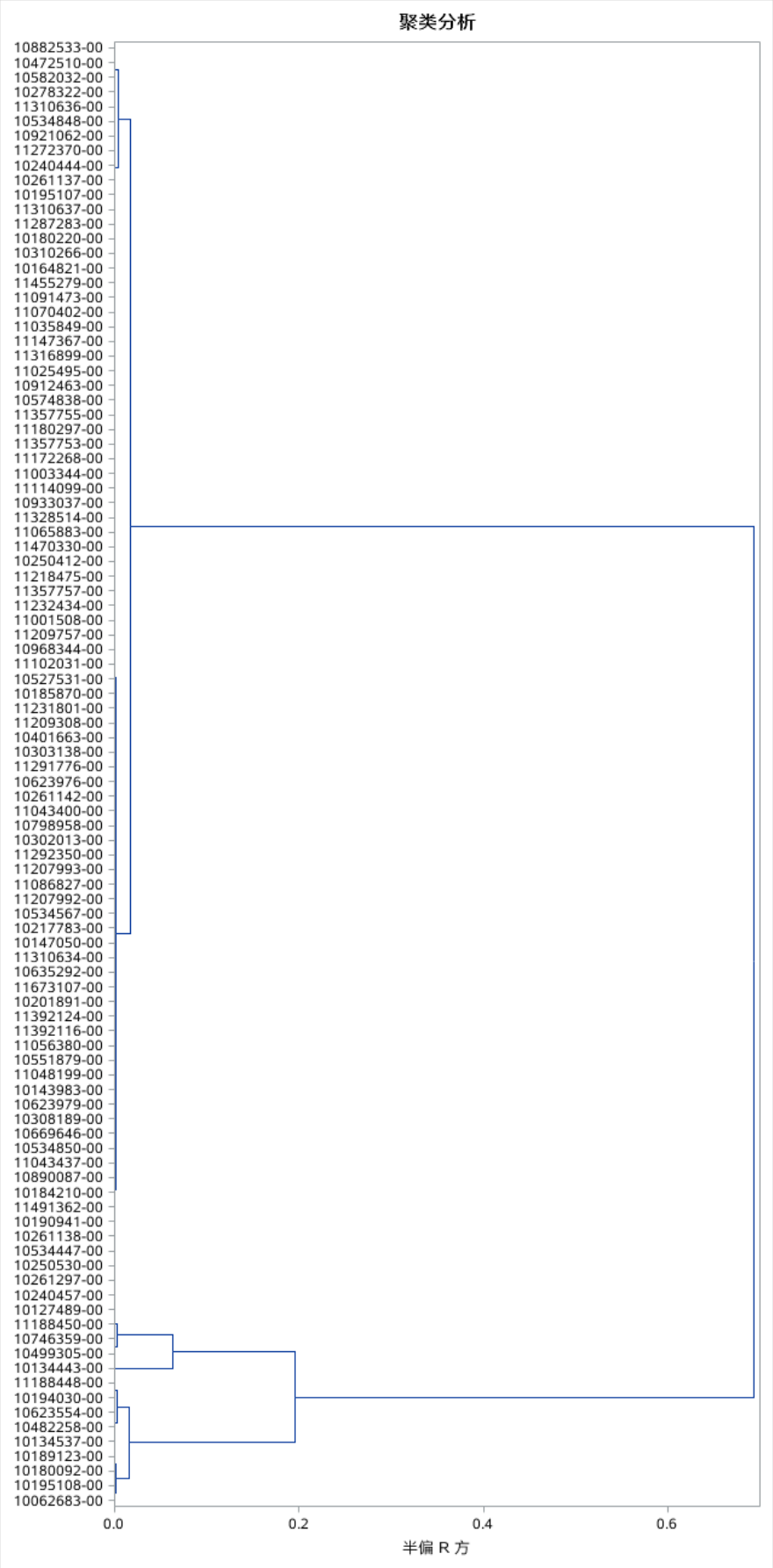

聚类分析
10882533-00
10472510-00
10582032-00
10278322-00
11310636-00
10534848-00
10921062-00
11272370-00
10240444-00
10261137-00
10195107-00
11310637-00
11287283-00
10180220-00
10310266-00
10164821-00
11455279-00
11091473-00
11070402-00
11035849-00
11147367-00
11316899-00
11025495-00
10912463-00
10574838-00
11357755-00
11180297-00
11357753-00
11172268-00
11003344-00
11114099-00
10933037-00
11328514-00
11065883-00
11470330-00
10250412-00
11218475-00
11357757-00
11232434-00
11001508-00
11209757-00
10968344-00
11102031-00
10527531-00
10185870-00
11231801-00
11209308-00
10401663-00
10303138-00
11291776-00
10623976-00
10261142-00
11043400-00
10798958-00
10302013-00
11292350-00
11207993-00
11086827-00
11207992-00
10534567-00
10217783-00
10147050-00
11310634-00
10635292-00
11673107-00
10201891-00
11392124-00
11392116-00
11056380-00
10551879-00
11048199-00
10143983-00
10623979-00
10308189-00
10669646-00
10534850-00
11043437-00
10890087-00
10184210-00
11491362-00
10190941-00
10261138-00
10534447-00
10250530-00
10261297-00
10240457-00
10127489-00
11188450-00
10746359-00
10499305-00
10134443-00
11188448-00
10194030-00
10623554-00
10482258-00
10134537-00
10189123-00
10180092-00
10195108-00
10062683-00
0.0
0.2
0.4
0.6
半偏 R 方

图 4.4.3　聚类树状图

5. 模型的建立

5.1 平稳性检验与纯随机性检验

平稳性检验有两种方法，一种是图检验法，即根据时序图和自相关系数图显示的特征做出判断；一种是根据单位根检验法，即构造检验统计量进行假设检验的方法。

下面我们首先介绍图检验法。

图检验方法是一种操作简便、运用广泛的平稳性判别方法，但是判别结论有一定的主观性，所以最好是和统计检验方法一起判断。

根据平稳时间序列均值和方差为常数的性质可知，平稳时间序列的时序图应该显示出该序列始终在一个常数值附近随机波动，而且波动的范围有明显相似的特点。如果时间序列图显示出该序列有明显的趋势性和周期性，那么它通常不是平稳序列。根据此性质，很多非平稳序列通过查看时序图就可以被识别出来。

自相关函数是用来描述时间序列中不同观测之间的线性相关程度的，可以证明平稳时间序列通常都具有短期相关性，具体描述就是随着延迟期数 k 的增加，平稳序列的自相关系数就会很快衰减为 0。反之，非平稳序列的自相关系数衰减为 0 的速度通常比较慢。这就是我们利用自相关图进行平稳性判别的标准。自相关图（ACF 图），横坐标表示延迟期数（也称滞后期数），纵坐标表示自相关系数的取值，图中每一个柱子都代表了某延迟期数对应的自相关系数的取值。

然后我们介绍单位根检验法。

单位根检验方法就是用来判断序列是否需要进行差分的方法，换言之，就是用来判断序列是否平稳。在单位根检验法中，最常用的是 DF 检验法。但是 DF 检验法只适用于 1 阶自回归过程的平稳性检验，但是实际绝大多数的时间序列不会是一个简单的 AR(1)过程。我们对 DF 检验进行一定的修正，得到增广的 DF 检验。ADF 检验可以用于如下三种类型的平稳性检验。

① 无常数均值、无趋势的 p 阶自回归过程。

② 有常数均值、无趋势的 p 阶自回归过程。

③ 既有常数均值、又有线性趋势的 p 阶自回归过程。

在 ARIMA 过程中，使用选项 stationary 可以指定进行 ADF 检验。

接下来，我们将介绍白噪声检验。

并不是所有的时间序列都值得建立模型，只有那些序列值之间具有相互依赖性，历史数据对未来有一定影响的序列，才值得建立模型，建模是为了预测序列

未来的发展。如果序列是白噪声序列，那么过去的行为对将来没有影响，从统计分析的角度来说，是没有建模的价值的。

在软件的实现过程中，如果是非白噪声序列，那么“白噪声的自相关检查”的“卡方”（也就是代表着 LB 统计量），会陡增。如果序列是白噪声序列，那么会几乎不变。

下面我们将用一种物料编号下的一个应发库的数据做个示范。我们选取第一种物料编号、应发库为北京的销售订单数量进行平稳性检验和纯随机性检验。

我们绘制出它的时序图，如下图 7 所示。

图 7　选取的数据绘制出的时间序列图

从上图 7，我们可以看出，该时间序列有明显的确定性周期的趋势，所以我们大致判断它不是平稳的。但是这种判断是不准的，所以我们需要用它的自相关系数的图像来进行判断。自相关系数的结果图如下图 8 所示。

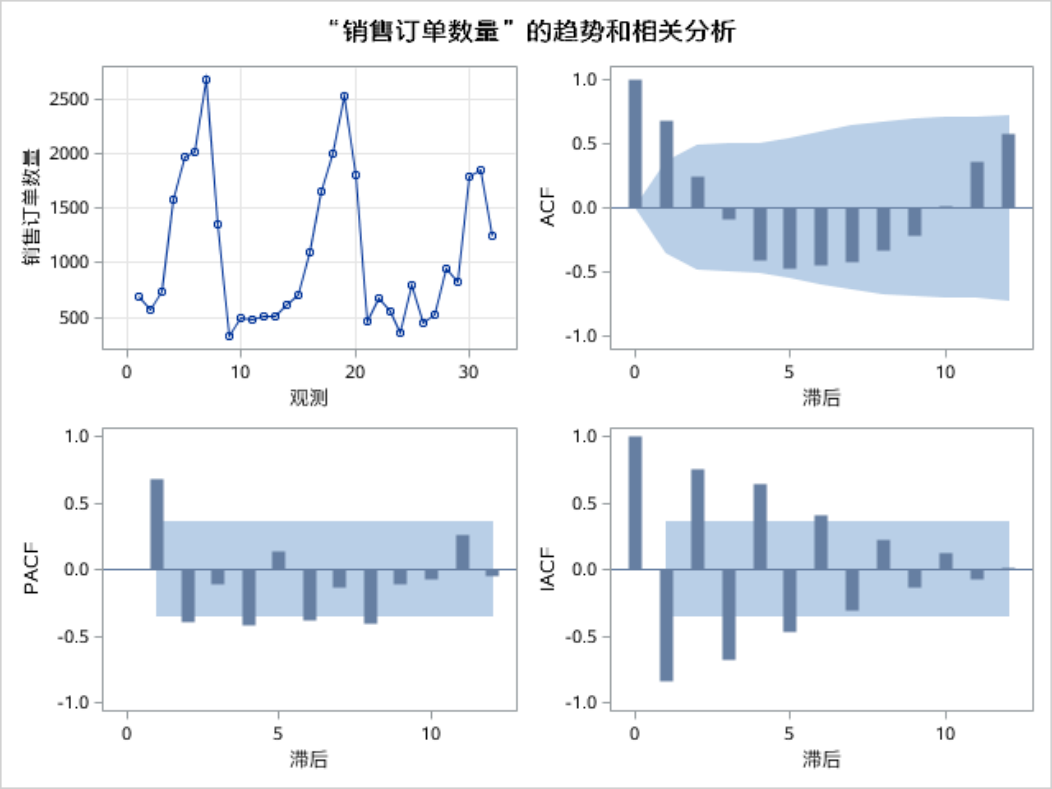


图 8　自相关系系数及其他结果

我们从上图 8 的结果可知：在滞后一阶到滞后十二阶的范围内，自相关系数并没有迅速缩减到 0，而是缓慢地拖尾，并有较大的正弦波摆动。这说明，该时间序列不是平稳序列，而是非平稳的时间序列。

接着我们继续进行单位根检验，如下图 9 所示。

增广 Dickey-Fuller 单位根检验							
类型	滞后	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
零均值	0	-2.5184	0.2697	-1.06	0.2553		
	1	-4.1542	0.1527	-1.32	0.1681		
单均值	0	-10.0759	0.1020	-2.39	0.1516	2.88	0.3582
	1	-22.8828	0.0012	-3.23	0.0274	5.24	0.0398
趋势	0	-10.1492	0.3574	-2.37	0.3875	2.81	0.6276
	1	-23.9562	0.0069	-3.26	0.0917	5.37	0.1453

图 9　单位根检验结果

从上述图形中的 Rho 和 Tau 统计量可知，本时间序列是非平稳的。

与此同时，我们判断它的纯随机性。用 SAS 对此时间序列执行 identify 语句之后，我们会得到它的白噪声自相关检查的结果。具体如下图 10 所示。

白噪声的自相关检查									
至滞后	卡方	自由度	Pr > 卡方	自相关					
6	42.62	6	<.0001	0.674	0.242	-0.085	-0.411	-0.477	-0.452
12	82.30	12	<.0001	-0.427	-0.335	-0.215	0.020	0.358	0.572

图 10　纯随机性检验结果

在上图 10 的纯随机性检验结果中，我们可以知道，“卡方”是会骤增，那么也就是说该序列不是一个白噪声序列，它具有建模的价值。

我们在用 ARIMA 过程步对时间序列进行检验的同时，也可以用 timeseries 过程步对它进行检验，同时也可以输出自相关系数图、偏相关系数图、逆相关系数图以及各自的标准化的图形，详细情况见附录三。

5.2　非平稳时间序列转化为平稳时间序列

首先，在进行非平稳序列转化为平稳时间序列前，我们用 timeseries 过程步对序列的趋势和季节成分进行分解，观察出序列中是否存在明显的趋势或季节成分。详细结果见下图 11 和图 12。

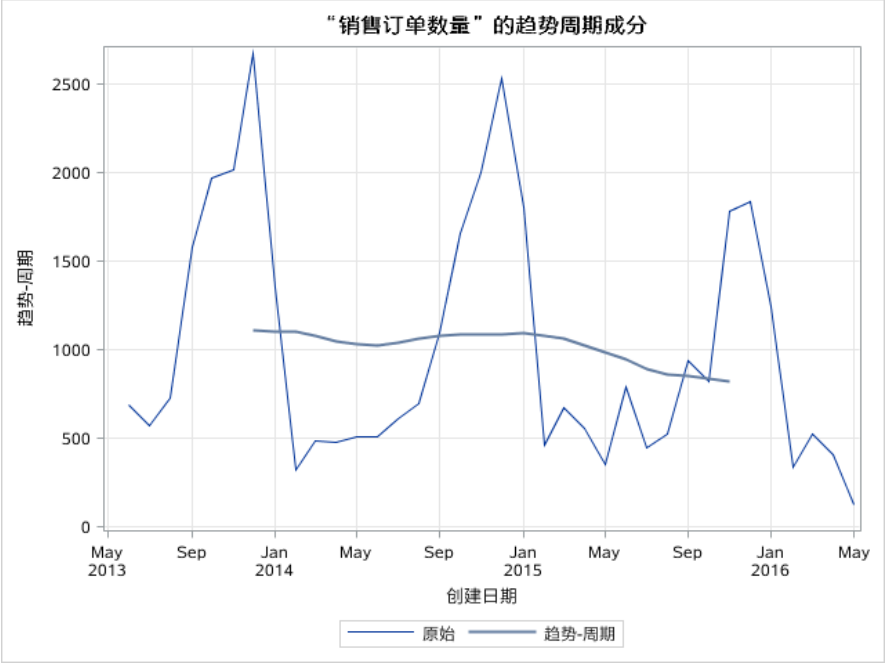


图 11　序列的趋势周期成分结果

从上述的图 11，我们可以看出该序列有较为明显的线性趋势，但是斜率不太，可以将其忽略。

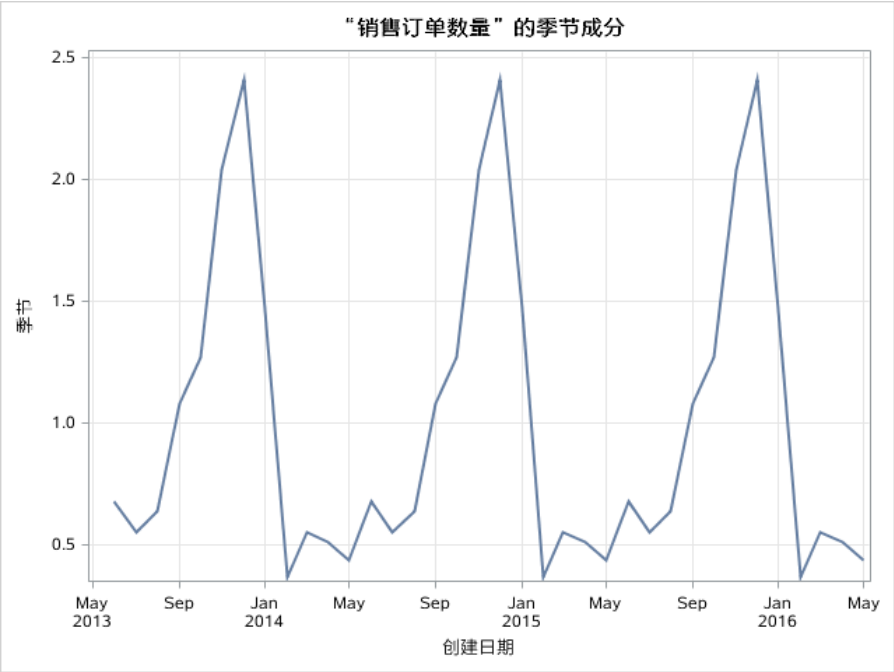


图 12　序列的季节成分结果

从上述的图 12，我们可以知道，该序列有明显的季节性的成分。而且我们观察到每一年的年末和下一年的年初都有很强的季节性趋势，而且周期为 12 个月。

5.3 模型识别

在模型的识别中，我们运用 identify 语句，对我们的时间序列进行识别。在其中，我们会输出自相关的系数（如图 10 中的内容），而且还会输出时间序列的标准化的自相关系数的分析图。具体图形如下：

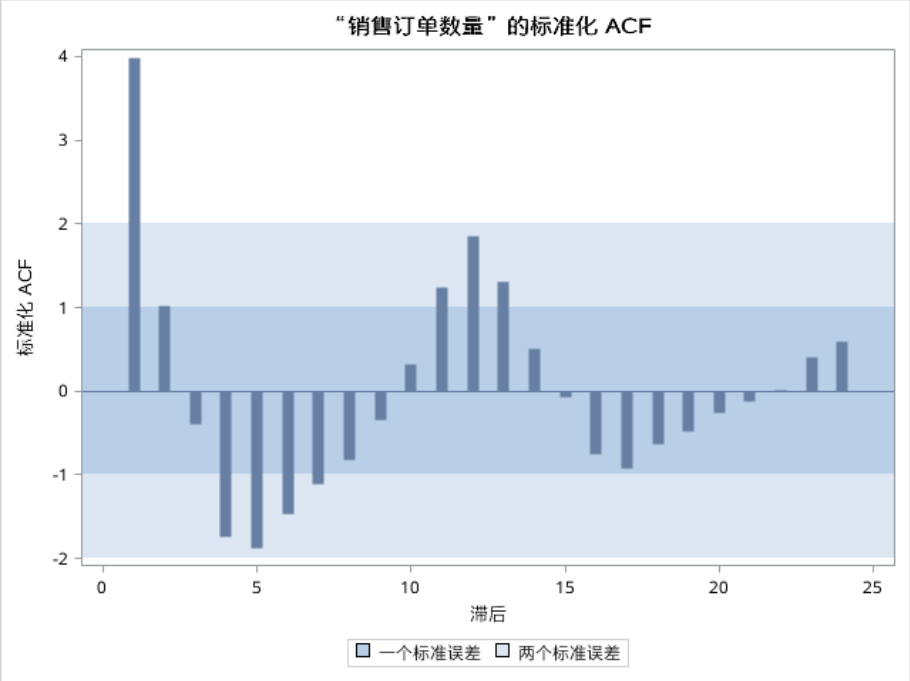


图 13　标准化的自相关系数分析图

从上面的标准化的自相关系数图 13，我们可以知道滞后 1 阶到滞后 24 阶的自相关系数是一个正弦波波动，而且 ACF 图像是拖尾的。

在参数识别的过程中，有三种自动辨别方法：一是，ESACF 延伸自相关系数法；二是，SCAN 最小典型相关法；三是，MINIC 最小信息法则。

下面是三种识别方法的最优定阶的显示结果。

Minimum Information Criterion

Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8	MA 9	MA 10	MA 11	MA 12
AR 0	12.08267	11.97049	11.96718	11.96379	11.57505	11.20789	-23.8572	.	.	.	-30.2819	.	.
AR 1	11.47483	11.5787	11.61769	11.69302	11.09947	8.487651	-24.0556	.	.	.	.	.	.
AR 2	11.25088	10.62353	10.38949	8.263188	7.658341	.	.	.	.	.	.	.	.
AR 3	11.29599	9.891893	9.895728	8.140488	7.117505	.	.	.	.	.	.	.	.
AR 4	10.536	7.967046	8.053618	7.824311	.	.	.	.	.	.	.	.	.
AR 5	10.58181	7.248571	.	-21.212	-21.7207	.	.	.	.	.	.	.	.
AR 6	-21.6293	-21.5175	-20.1593	.	.	.	-24.865	.	.	.	.	.	.
AR 7	.	-19.6119	-21.2858	-22.3016	-26.6248	.	.	.	.	.	.	.	.
AR 8	-24.3482	-23.3584	-24.2016	.	.	.	.	.	.	.	.	.	.
AR 9	.	-23.1815	.	.	.	.	.	.	.	.	.	.	.
AR 10	.	.	.	.	.	.	.	.	.	.	.	.	.
AR 11	.	.	.	.	.	.	.	.	.	.	.	.	.
AR 12	.	.	.	.	.	.	.	.	.	.	.	.	.

图 14　最小信息法则辨别结果

上图 14 是在模型预测中，相对最优识别的输出结果的矩阵。在上面的那个矩阵中，最佳的阶数就是矩阵中数值最小的那个所对应的阶数。

SCAN Chi-Square[1] Probability Values													
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8	MA 9	MA 10	MA 11	MA 12
AR 0	<.0001	0.2190	0.9286	0.1775	0.0948	0.0772	0.0490	0.1688	0.3930	0.9633	0.2226	0.0536	0.1446
AR 1	0.0548	0.2065	0.0574	0.0986	0.5676	0.9661	0.2387	0.4025	0.2145	0.2342	0.0261	0.0110	0.4250
AR 2	0.4586	0.2674	0.0577	0.1056	0.8561	0.5865	0.3094	0.3092	0.7409	0.3318	0.7309	0.3311	0.8714
AR 3	0.0267	0.0685	0.4370	0.1527	0.2426	0.2830	0.8317	0.7388	0.4852	0.5280	0.3962	0.5993	0.4893
AR 4	0.5469	0.2231	0.1261	0.6918	0.5083	0.2458	0.8182	0.7852	0.5438	0.8331	0.9048	0.4825	0.9967
AR 5	0.0534	0.1022	0.2901	0.4657	0.4707	0.6600	0.2671	0.4449	0.5825	0.9473	0.8669	0.3861	0.6719
AR 6	0.1437	0.9654	0.7981	0.5951	0.9321	0.4602	0.9626	0.5868	0.6638	0.6780	0.8408	0.6565	0.8945
AR 7	0.3436	0.9610	0.0002	0.5653	0.2781	0.8721	0.3961	0.8595	0.7538	0.6823	0.6057	0.5506	0.6977
AR 8	0.2976	0.4503	0.6629	0.8523	0.9632	0.7916	0.6417	0.6165	0.9192	0.8920	0.9969	0.5048	.
AR 9	0.3319	0.2217	0.3336	0.7570	0.8527	0.7431	0.8544	0.8832	0.6648	0.8485	0.8748	0.5397	.
AR 10	0.0723	0.4335	0.9293	.	.	.	.	.	.	.	.	.	.
AR 11	.	.	.	.	.	.	.	.	.	.	.	.	.
AR 12	.	.	.	.	.	.	.	.	.	.	.	.	.

图 15　最小典型相关法辨别结果

ESACF Probability Values													
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8	MA 9	MA 10	MA 11	MA 12
AR 0	0.0001	0.3218	0.7341	0.1037	0.0799	0.1284	0.1801	0.3187	0.5340	0.9539	0.3066	0.1132	0.2118
AR 1	0.0260	0.3658	0.7372	0.0362	0.4898	0.9560	0.3910	0.6596	0.4344	0.9065	0.2921	0.0561	0.2681
AR 2	0.0839	0.3302	0.0267	0.0383	0.2839	0.8157	0.3539	0.4529	0.8440	0.8456	0.9801	0.0795	0.3440
AR 3	0.1890	0.0470	0.2487	0.0305	0.3862	0.7063	0.9410	0.6211	0.7128	0.9800	0.6120	0.1259	0.6690
AR 4	0.0914	0.0302	0.4253	0.2937	0.7270	0.2581	0.6963	0.6818	0.3203	0.3817	0.6201	0.3442	0.6786
AR 5	0.1904	0.0701	0.3234	0.3252	0.8708	0.6067	0.7056	0.8720	0.4884	0.8686	0.5348	0.2215	0.3068
AR 6	0.0418	0.0094	0.6073	0.2198	0.5065	0.7257	0.9859	0.3321	0.7948	0.8724	0.4757	0.2141	0.5608
AR 7	0.0445	0.0028	0.6998	0.3020	0.7307	0.7540	0.9318	0.3370	0.6772	0.7538	0.6803	0.2752	0.3418
AR 8	0.0008	0.3420	0.4366	0.9096	0.8749	0.8140	0.2644	0.3406	0.8812	0.8704	0.8576	0.2661	0.9661
AR 9	0.0003	0.3070	0.3084	0.8444	0.7878	0.0674	0.2184	0.3721	0.9212	0.7987	0.5581	0.2331	0.3863
AR 10	0.0723	0.2375	0.7983	0.2659	0.5156	0.3188	0.3339	0.3842	0.7763	0.8544	0.6343	0.3013	0.3970
AR 11	0.0027	0.2067	0.9949	0.0951	0.5887	0.2876	0.2019	0.3372	0.7305	0.9447	0.4985	0.3153	0.5246
AR 12	0.0007	0.7778	0.9540	0.2706	0.4143	0.3713	0.3732	0.4790	0.4103	0.9119	0.5037	0.3903	.

图 16　延伸自相关系数辨别结果

综合上面的最小典型相关法与延伸自相关系数的方法，我们可以看出，当 p=1,q=0 时，满足最优定阶原则，即可原则 AR(1)模型对该汽车备件销售订单的时序的发展进行拟合。当然这并不符合最小信息法则的辨别结果，但我们可以在尝试 AR（1）模型之后针对这一法则优化模型。

5.4　参数估计

在本次时间序列的建模过程中，我们用到了最大似然估计（ML）。因为这种方法是时间序列建模预测的较为准确的估计。在一般的商业预测中，最大似然估计的方法用的比较多，也比较地准确。

在下图 17 的结果图中，第一个报表是最大似然估计之后的参数估计的报表；第二、三个报表是部分诊断检验的输出报表。

最大似然估计

参数	估计	标准误差	t 值	近似 Pr > |t|	滞后
MU	-160.04180	104.17897	-1.54	0.1245	0
AR1,1	0.35843	0.23009	1.56	0.1193	1

常数估计	-102.678
方差估计	93960.53
标准误差估计	306.5298
AIC	287.8004
SBC	289.7919
残差数	20

参数估计相关性

参数	MU	AR1,1
MU	1.000	-0.103
AR1,1	-0.103	1.000

图 17　最大似然估计的输出结果

对该时序应用 AR（1）模型进行拟合，其参数估计结果如上所图 17 所示，因此，该拟合模型的具体形式为

$$X_t = -160.04180 + 0.35843 * X_{t-1}.$$

5.5 模型检验

在时间序列的模型建立好，拟合之后，我们对模型进行检验。模型诊断检验过程将通过计算多个诊断性统计量来衡量模型的拟合优度与准确度，并对模型的残差序列进行相关性检验和正态性检验。

衡量模型拟合优度的准则一般有两个。第一个是 AIC 准则，又称赤池信息量准则。它建立在熵的概念基础之上，可以权衡所估计模型的复杂度和该模型的拟合数据的优良性。AIC 准则鼓励数据拟合的优良性，但是也表示应尽量避免出现过度拟合的情况。在上述的图 17 输出了 AR（1）模型的拟合优度报表。

下面我们就来讲述此次时间序列建模后的残差分析。下面的图 18 是残差的自相关检查的报表。

残差的自相关检查									
至滞后	卡方	自由度	Pr > 卡方	自相关					
6	10.13	5	0.0716	-0.099	0.278	0.242	-0.308	0.283	-0.221
12	19.67	11	0.0501	-0.074	-0.008	-0.404	0.065	-0.225	-0.055
18	21.91	17	0.1880	0.088	-0.072	0.038	0.061	-0.071	-0.021

图 18　残差自相关检查报表

在上述的图 18 中，我们可以知道：利用 AR（1）模型对该时序进行拟合之后得到一组残差序列，然后我们对该残差序列进行了延后 6 阶、12 阶、18 阶的显著性检验，检验的 p 值大于显著性水平 0.05，该拟合模型大体上还是成立的，但并不明显，该残差序列中可能存在某些相关的成分，需要后期进行优化。

下面的图 19 是残差分析中的时间序列中的销售订单数量进行 12 阶差分之后进行的残差相关性检验。图 20 是残差正态诊断的结果图。

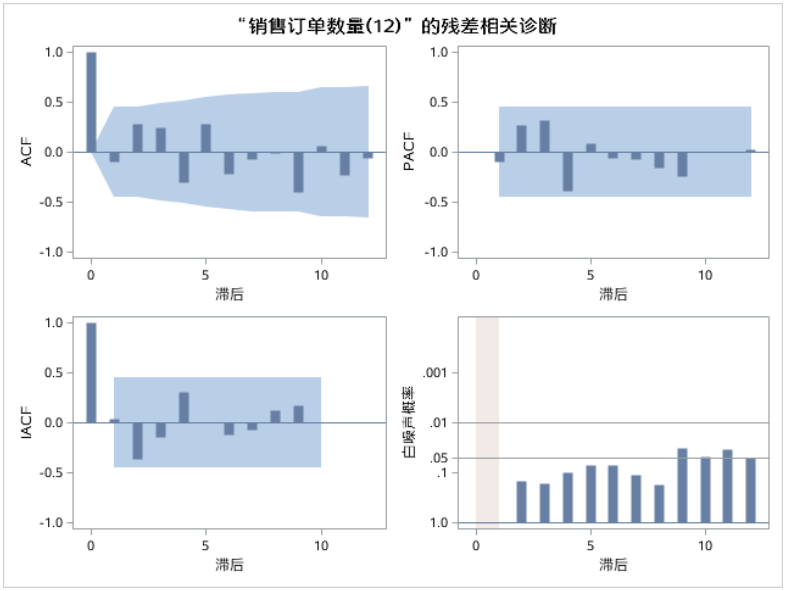


图 19　十二阶差分后的残差相关诊断的结果图

从上述的图 19，我们可以知道，时间序列经过十二阶差分后的残差序列的 ACF 图形在一阶滞后慢慢地趋于平稳，PACF 图像慢慢地减缓直至 0，并且在它的附近震荡。

在白噪声的概率中，我们可以发现所有的滞后的白噪声基本上都是在 0.05 的范围内，这说明我们拟合的序列还算是比较好的。但是，最后几阶可能说明我们提取的信息是不够多的，那么我们还是需要再对模型进行修正的。

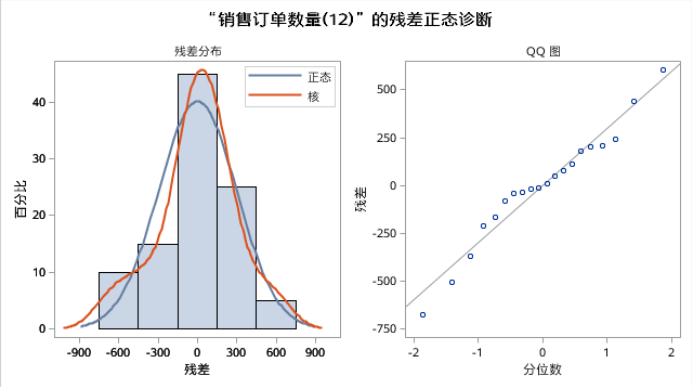


图 20　十二阶差分后的残差正态诊断的结果图

在上述的图 20 中，我们可以看到时间序列经过十二阶的差分之后，残差的分布基本上与正态分布是基本拟合的，但是拟合的程度不是很高。与此同时，我们也输出了数据的正态分布的残差检验的 QQ 图。该图显示数据基本上拟合了正态分布，但是还是会有些偏差。

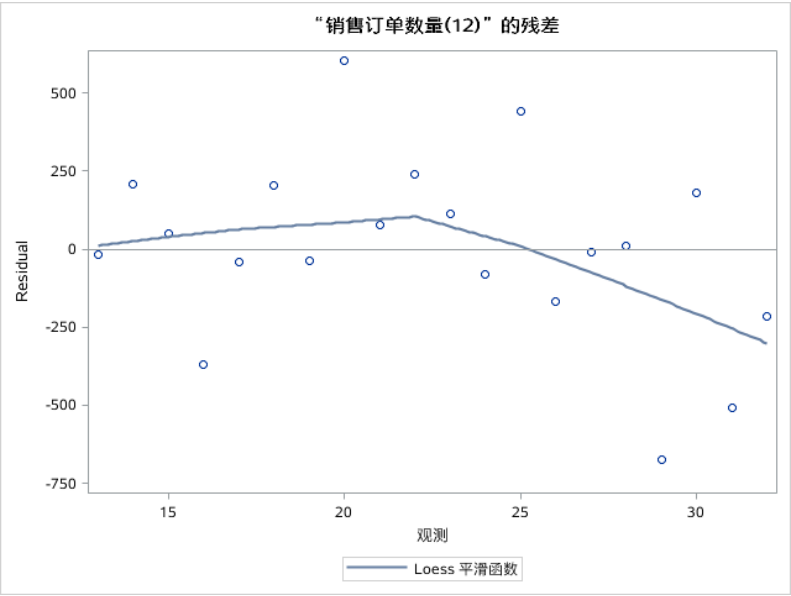


图 21　十二阶差分后的残差分布图

在上述的图 21 中，残差显示是随机的，中间的蓝色的实线是 Loess 平滑函数。这就说明，我们需要对序列做一些相应的平滑处理才能使得预测值更加精确。

5.6　模型优化

模型的优化是靠异常点的检验。下面我们将介绍一下异常点的检验，提高准确度。我们经过使用 ARIMA 里面的语句进行异常点的检验。

5.7 序列预测

最后，我们应用上述优化后的模型对该公司汽车备件的销售订单数量的时序列进行拟合并预测未来四个月的订单数量，其结果如下图 24 所示。

在输出时间序列的外推预测图之前，我们首先是得到时间序列的预测报表图。详情见图 22。那么我们可以知道 2016 年 2 月份到 5 月份，物料编号是“10062683-00”的汽车零备件产品的北京应发库的销售订单的数量是 157 件、460 件、379 件以及 187 件。虽然，外推四个月的预测值可以通过时间序列预测出来，但是本案例中销售订单的数量是件数。所以，我们只好对预测值进行取整。

变量“销售订单数量”的模型	
估计均值	-160.042
差分期间	12

自回归因子	
因子 1:	1 - 0.35843 B**(1)

以下变量的预测:销售订单数量				
观测	预测	标准误差	95% 置信限	
33	157.1679	306.5298	-443.6195	757.9553
34	460.0612	325.6252	-178.1524	1098.2748
35	379.3568	327.9978	-263.5071	1022.2207
36	187.2909	328.3014	-456.1679	830.7498

图 22　销售订单数量的预测值

根据上面的预测数据，我们同时也找着了本时序的真实值。336、552、408、126，它们分别是 2 月份到 5 月份的销售订单数量的真实值。我们可以根据预测值与真实值的差距来判断模型的拟合准确度。

我们提出：相对总离差平方和，将真实值减去预测值的绝对值再除以对应的真实值的加总。那么根据已知的数据，我们可以得出该时序的相对总离差平方和为：1.25。

下面 SAS 中会自动输出一个外推的局部的放大图形，具体结果如下图 23 所示。在下面的图 23 中，蓝色的阴影部分是预测的置信限。置信限的范围越小，那么拟合的准确度也相对来说比较高。在这里，我们的置信限的范围是比较小的。

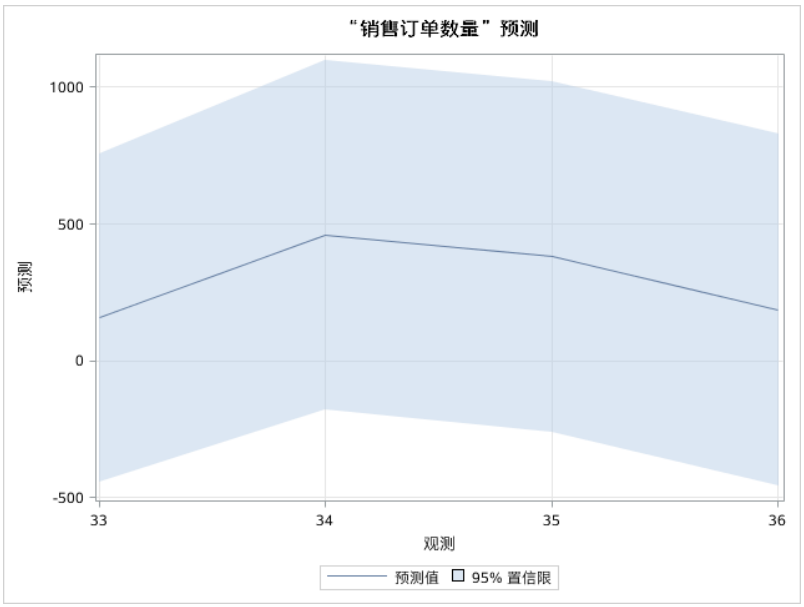


图 23　具体外推的预测图

紧接着就输出销售订单数量的总预测图，详见下图 22 所示。在下图中，我们可以明显的看到时间序列分析的拟合线，两侧的置信限和外推四个月的预测值。

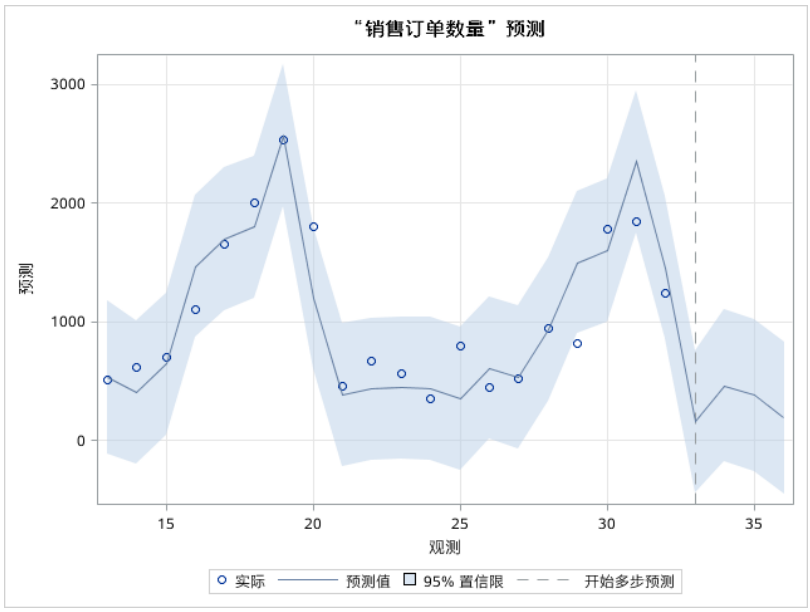


图 22 销售订单数量预测图

如上图所示，该模型对该汽车公司的备件销售订单数量的原始数据记录拟合得很好，因此对未来四个月的预测数据的可信度还是比较高，该公司可以将它作为生产计划制定的参考，也可以从其他的应发库调出产品或者将产品调入其他的应发库。

参考文献

［1］朱世武. SAS编程技术教程［M］. 北京:清华大学出版社, 2013.

［2］何晓群. 多元统计分析［M］. 北京:中国人民大学出版社, 2015.

［3］夏坤庄. 深入解析SAS:数据处理、分析优化与商业应用［M］. 北京：机械工业出版社，2015.

［4］谷鸿秋. SAS编程演义［M］. 北京:清华大学出版社，2017.

［5］王燕. 应用时间序列分析［M］. 北京:中国人民大学出版社，2016.

［6］肖枝洪. 时间序列分析与SAS应用［M］. 武汉:武汉大学出版社，2012.

［7］Institute S. SAS 9.4 graph template language reference, third edition［M］. SAS Institute, 2014.

［8］SASInstitute. SAS 9.4 macro language: Reference, second edition［M］. SAS Institute, Incorporated, 2014.

［9］SASInstitute. SAS 9.4 ods graphics: Procedures guide, third edition［M］. SAS Institute, 2014.

［10］SASInstitute. SAS 9.4 output delivery system: User’s guide, third edition［M］. SAS Institute, 2014.

［11］SASInstitute. SAS 9.4 sql procedure user’s guide, third edition［M］. SAS Institute, Incorporated, 2015.

附录

附录一：模型预测结果

附录二：本案例所使用的程序命令

```
/*******************************************/
/*                                         */
/*            2017年SAS汇丰杯案例1          */
/*                                         */
/*     某大型汽车制造厂备件销售预测案例      */
/*                                         */
/*                                         */
/*******************************************/

/*******************************************/
/*                                         */
/*           1.确定业务问题                */
/*                                         */
/*******************************************/

/*************************************************************************/
/*=== 1.1 理解业务现状 ===*/
/*===
【背景】：某汽车备件销售规模不断增大，需建立企业级的备件销量预测管理平台，提高人员工作效率的同时，也能提升预测精准度，优化库存。
===*/
/*=== 基本了解业务现状 ===;
/*************************************************************************/
/*=== 1.2 提出业务问题 ===*/
/*===
【问题】：此案例的主要需求是实现每个备件外推4个月的月度销量的批量预测和分层预测，并给出预测的准确度。
1、将源数据导入SAS并进行清洗，如缺失值和异常值处理、产品替换等；
2、探索数据的特性，据此对产品进行分类，进一步构建相关的特征变量并检验特征变量的合理性。
3、利用sas分层批量预测每个配件未来4个月的销售订单数量，时间粒度是【月度】，产品维度是【物料编号】，区域维度是【应发库】，预测变量是【销售订单数量】，
   可进一步添加事件或者自定义模型优化预测效果。
===*/
/*=== 业务问题明了 ===;
/*************************************************************************/
/*=== 1.3 检查数据的可行性 ===*/
/*===
【源数据】：源数据主要包含两张Excel表--“销售明细”和“主数据”，前者给出了详细的
```

```
销售字段及内容，后者给出了备件的相关属性和替代信息。
历史数据：2013年6月-2016年1月
预测目标：备件销售订单数量
预测层级：备件+应发库
预测期间：2016年2月-2016年5月
===*/
/*=== 数据可行;
/*************************************************************************/

/*****************************************/
/*                                       */
/*              2.数据的准备              */
/*                                       */
/*****************************************/

/*************************************************************************/
/*=== 创建数据挖掘环境、准备数据、数据导入、观察数据 ===*/

/*拓展SAS里面的语言*/
options validvarname=any
        validmemname=extend;
run;

/*在本案例中，我们先新建了逻辑库huifeng*/
libname huifeng "F:\重庆理工大学-理学院-赵赞豪\逻辑库";
/*然后将案例1的数据另存为CSV格式，再进行数据导入*/

/*导入“销售明细”到数据集huifeng.source_data中*/
PROC IMPORT OUT= HUIFENG.source_data
            DATAFILE= "F:\重庆理工大学-理学院-赵赞豪\数据\销售明细.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

/*导入“主数据”到数据集huifeng.main_data中*/
PROC IMPORT OUT= HUIFENG.main_data
            DATAFILE= "F:\重庆理工大学-理学院-赵赞豪\数据\主数据.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

/*导入“仓库”到数据集huifeng.cangku中*/
```

```
PROC IMPORT OUT= HUIFENG.cangku
            DATAFILE= "F:\重庆理工大学-理学院-赵赞豪\数据\仓库.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
/*************************************************************************/

/*****************************************/
/*                                       */
/*            3.数据的探索               */
/*                                       */
/*****************************************/

/*************************************************************************/
/*查看数据集文件huifeng1.source_data、huifeng1.cangku、huifeng1.main_data
的属性信息*/
proc contents data=huifeng.source_data;
run;
proc contents data=huifeng.main_data;
run;
proc contents data=huifeng.cangku;
run;

/*每个变量是否含有缺失值，如有，获得缺失值的个数*/
proc means data=huifeng.source_data  N NMISS;
var _numeric_;
run;
proc means data=huifeng.main_data  N NMISS;/*可以忽略该过程步*/
var _numeric_;
run;
proc means data=huifeng.cangku  N NMISS;/*可以忽略该过程步*/
var _numeric_;
run;

/*对定类数据进行探索*/
%macro FreqBar(ds,varname);
proc freq data=&ds;
tables &varname / plots(only)=freqplot;
run;
%mend;
%FreqBar(huifeng.source_data,物料编号);
%FreqBar(huifeng.source_data,物料描述);
%FreqBar(huifeng.source_data,K3编号);
```

```
%FreqBar(huifeng.source_data,产品组);
%FreqBar(huifeng.source_data,客户物料号);
%FreqBar(huifeng.source_data,销售组织);
%FreqBar(huifeng.source_data,是否拣配);
%FreqBar(huifeng.source_data,交货单创建者);
%FreqBar(huifeng.source_data,交货单修改者);
%FreqBar(huifeng.source_data,销售大区);
%FreqBar(huifeng.source_data,销售小区);
%FreqBar(huifeng.source_data,发运渠道);
%FreqBar(huifeng.source_data,应发库);
%FreqBar(huifeng.source_data,销售订单创建者);
%FreqBar(huifeng.source_data,订单类型);
%FreqBar(huifeng.source_data,单位);
%FreqBar(huifeng.source_data,币别);

 /*对连续变量进行探索*/
proc means data=huifeng.source_data N nmiss min mean median max std;
var 创建日期 交货单号 交货单行项目 客户代码 销售订单号 销售订单行项目 销售订单数量
交货数量;
run;
/*****************************************************************/

/*******************************************/
/*                                         */
/*            4.数据的加工                 */
/*                                         */
/*******************************************/

/*****************************************************************/
/*将数据进行排序*/
proc sort data=huifeng.source_data out=huifeng.source;
by 物料编号 应发库 创建日期;
run;
/*将产品进行分层处理，进行缺失值处理，方便之后的模型建立*/
proc timeseries data=huifeng.source out=huifeng.source1;
by 物料编号 应发库;
id 创建日期 interval=month accumulate=total setmissing=average
          start='06jun2013'd end='30may2016'd;
var 销售订单数量;
run;
/*销售订单数量的清洗*/
proc sql;
create table huifeng.source2 as select * from huifeng.source1 where 应
发库~='零售';
```

```
quit;
/*虽然不需要变量的转化，但是需要对数据集进行分割，数据集的分割如下*/
%macro wuliao(n,k);
proc sql;
create table huifeng.wuliao&k as select * from huifeng.source2 where 物料编号=&n;
quit;
%mend;
/****************/
/*                */
/*     宏调用     */
/*                */
/****************/
%wuliao('10062683-00',1);
%wuliao('10127489-00',2);
%wuliao('10134443-00',3);
%wuliao('10134537-00',4);
%wuliao('10143983-00',5);
%wuliao('10147050-00',6);
%wuliao('10164821-00',7);
%wuliao('10180092-00',8);
%wuliao('10180220-00',9);
%wuliao('10184210-00',10);
%wuliao('10185870-00',11);
%wuliao('10189123-00',12);
%wuliao('10190941-00',13);
%wuliao('10194030-00',14);
%wuliao('10195107-00',15);
%wuliao('10195108-00',16);
%wuliao('10201891-00',17);
%wuliao('10217783-00',18);
%wuliao('10240444-00',19);
%wuliao('10240457-00',20);
%wuliao('10250412-00',21);
%wuliao('10250530-00',22);
%wuliao('10261137-00',23);
%wuliao('10261138-00',24);
%wuliao('10261142-00',25);
%wuliao('10261297-00',26);
%wuliao('10278322-00',27);
%wuliao('10302013-00',28);
%wuliao('10303138-00',29);
%wuliao('10308189-00',30);
%wuliao('10310266-00',31);
```

```
%wuliao('10401663-00',32);
%wuliao('10472510-00',33);
%wuliao('10482258-00',34);
%wuliao('10499305-00',35);
%wuliao('10527531-00',36);
%wuliao('10534447-00',37);
%wuliao('10534567-00',38);
%wuliao('10534848-00',39);
%wuliao('10534850-00',40);
%wuliao('10551879-00',41);
%wuliao('10574838-00',42);
%wuliao('10582032-00',43);
%wuliao('10623554-00',44);
%wuliao('10623976-00',45);
%wuliao('10623979-00',46);
%wuliao('10635292-00',47);
%wuliao('10669646-00',48);
%wuliao('10746359-00',49);
%wuliao('10798958-00',50);
%wuliao('10882533-00',51);
%wuliao('10890087-00',52);
%wuliao('10912463-00',53);
%wuliao('10921062-00',54);
%wuliao('10933037-00',55);
%wuliao('10968344-00',56);
%wuliao('11001508-00',57);
%wuliao('11003344-00',58);
%wuliao('11025495-00',59);
%wuliao('11035849-00',60);
%wuliao('11043400-00',61);
%wuliao('11043437-00',62);
%wuliao('11048199-00',63);
%wuliao('11056380-00',64);
%wuliao('11065883-00',65);
%wuliao('11070402-00',66);
%wuliao('11086827-00',67);
%wuliao('11091473-00',68);
%wuliao('11102031-00',69);
%wuliao('11114099-00',70);
%wuliao('11147367-00',71);
%wuliao('11172268-00',72);
%wuliao('11180297-00',73);
%wuliao('11188448-00',74);
%wuliao('11188450-00',75);
```

```
%wuliao('11207992-00',76);
%wuliao('11207993-00',77);
%wuliao('11209308-00',78);
%wuliao('11209757-00',79);
%wuliao('11218475-00',80);
%wuliao('11231801-00',81);
%wuliao('11232434-00',82);
%wuliao('11272370-00',83);
%wuliao('11287283-00',84);
%wuliao('11291776-00',85);
%wuliao('11292350-00',86);
%wuliao('11310634-00',87);
%wuliao('11310636-00',88);
%wuliao('11310637-00',89);
%wuliao('11316899-00',90);
%wuliao('11328514-00',91);
%wuliao('11357753-00',92);
%wuliao('11357755-00',93);
%wuliao('11357757-00',94);
%wuliao('11392116-00',95);
%wuliao('11392124-00',96);
%wuliao('11455279-00',97);
%wuliao('11470330-00',98);
%wuliao('11491362-00',99);
%wuliao('11673107-00',100);
run;

/*进行聚类分析，将物料编号进行分类*/
proc sql;
create table huifeng.julei as select 物料编号,sum(销售订单数量) as 物料订单
总数 from huifeng.source2 group by 物料编号;
quit;

proc cluster data=huifeng.julei
             outtree=work.tree
             method=war;
     var 物料订单总数;
     id 物料编号;
run;
```

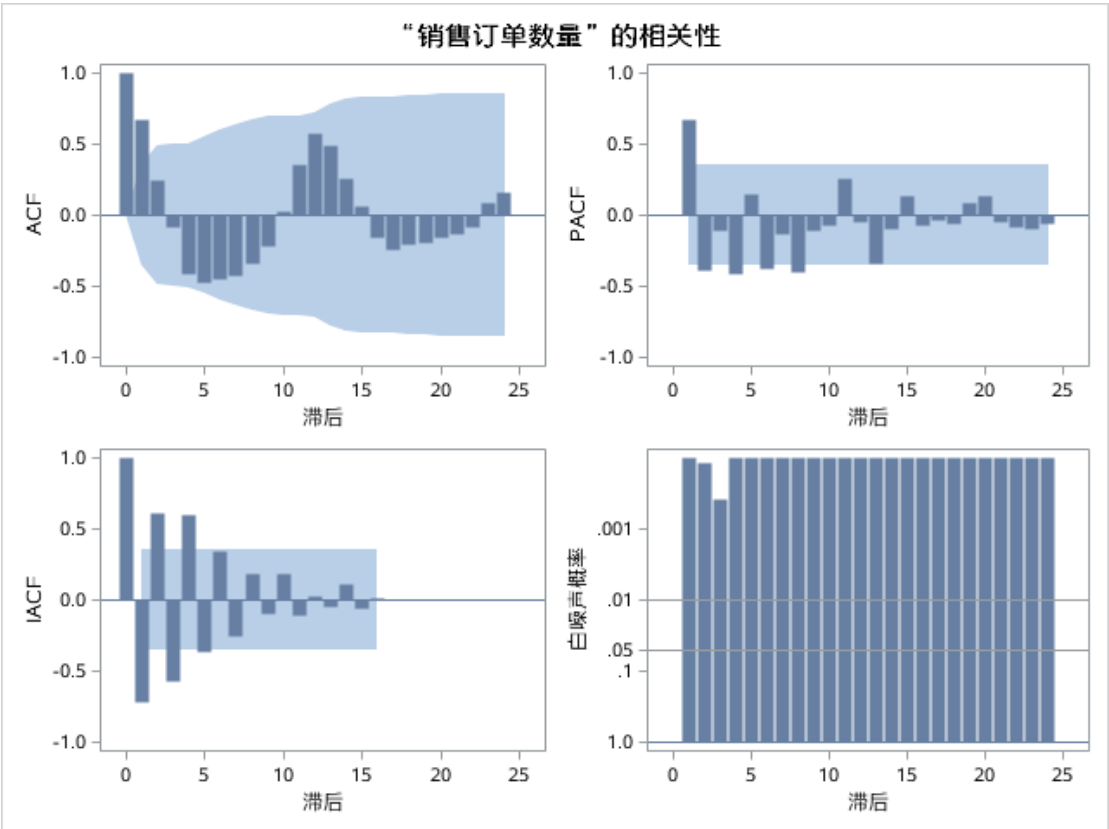
```
/*****************************************/
/*                                       */
/*            5.模型的建立               */
/*                                       */
/*****************************************/

/*************************************************************************/
/*绘制时序图，观察时序图的特征*/
/*定义宏*/
%macro shibie(wu,city);
proc gplot data=huifeng.wuliao&wu(where=(应发库=&city));
title1" 第 &wu 种物料编号";
       title2" 绘制 " &city " 应发库的销售订单情况";
       where 创建日期<='31jan2016'd;
       plot 销售订单数量 * 创建日期=1;
       symbol1 c=red i=join v=star;
run;
quit;
%mend;
/*循环调用宏*/
%macro test;
%do wu=1 %to 100;
%shibie(&wu,'北京');
%shibie(&wu,'长沙');
%shibie(&wu,'成都');
%shibie(&wu,'济南');
%shibie(&wu,'昆明');
%shibie(&wu,'上海');
%shibie(&wu,'深圳');
%shibie(&wu,'沈阳');
%shibie(&wu,'西安');
run;
%end;
%mend;
%test;
/*************************************************************************/
```

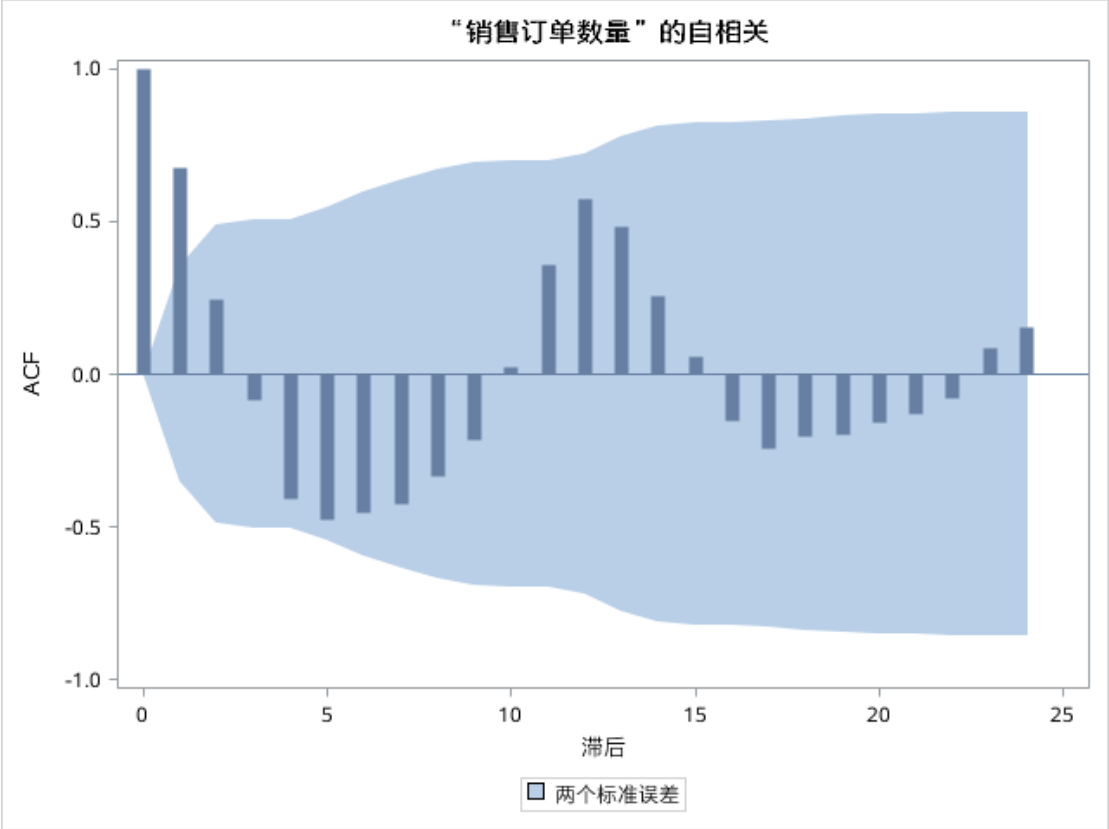
附录三：timeseries 过程步输出的图形

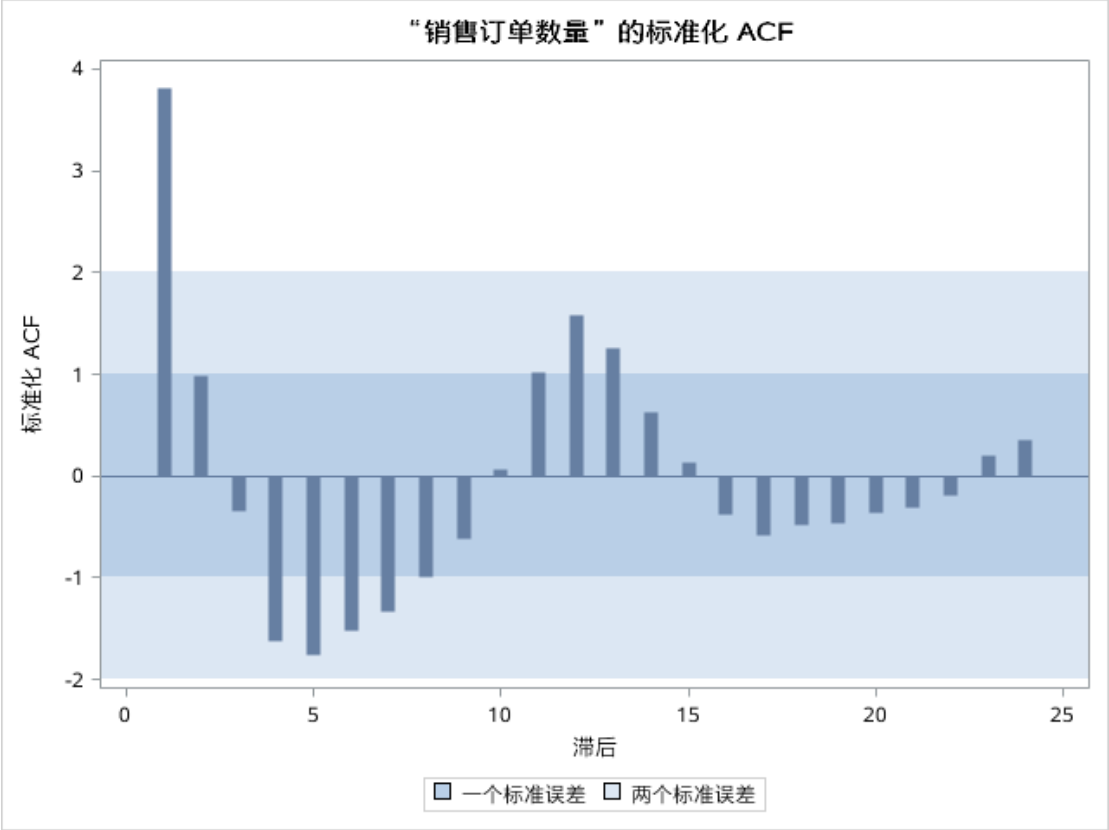
TIMESERIES 过程

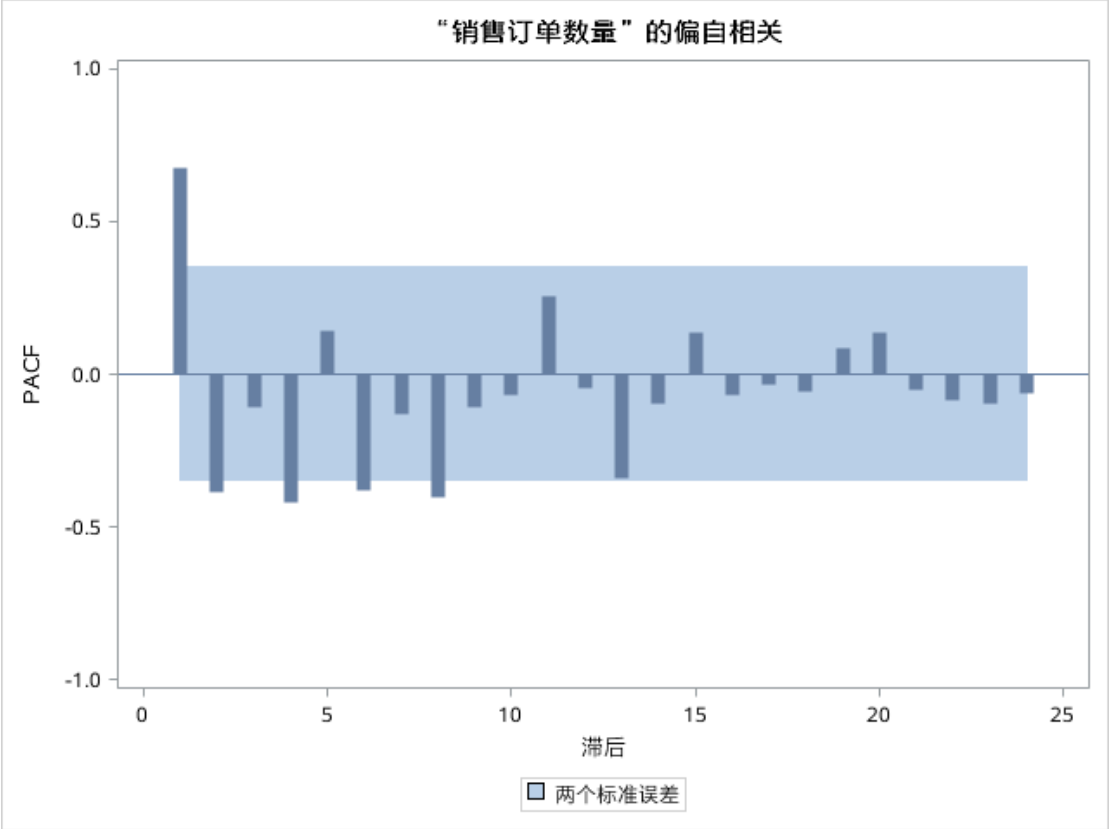
输入数据集	
名称	HUIFENG.WULIAO1
标签	
时间 ID 变量	创建日期
时间间隔	MONTH
季节周期的长度	12

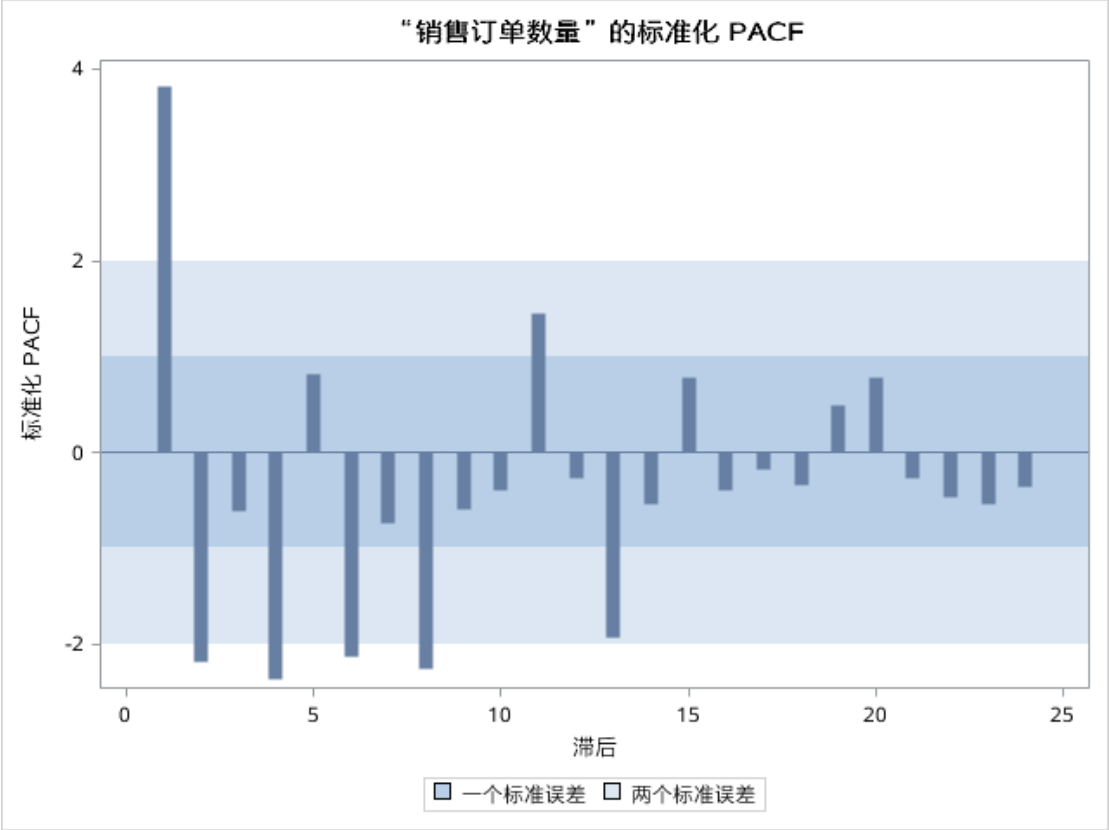
变量信息	
名称	销售订单数量
标签	
第一个	JUN2013
最后一个	JAN2016
读取的观测数	32

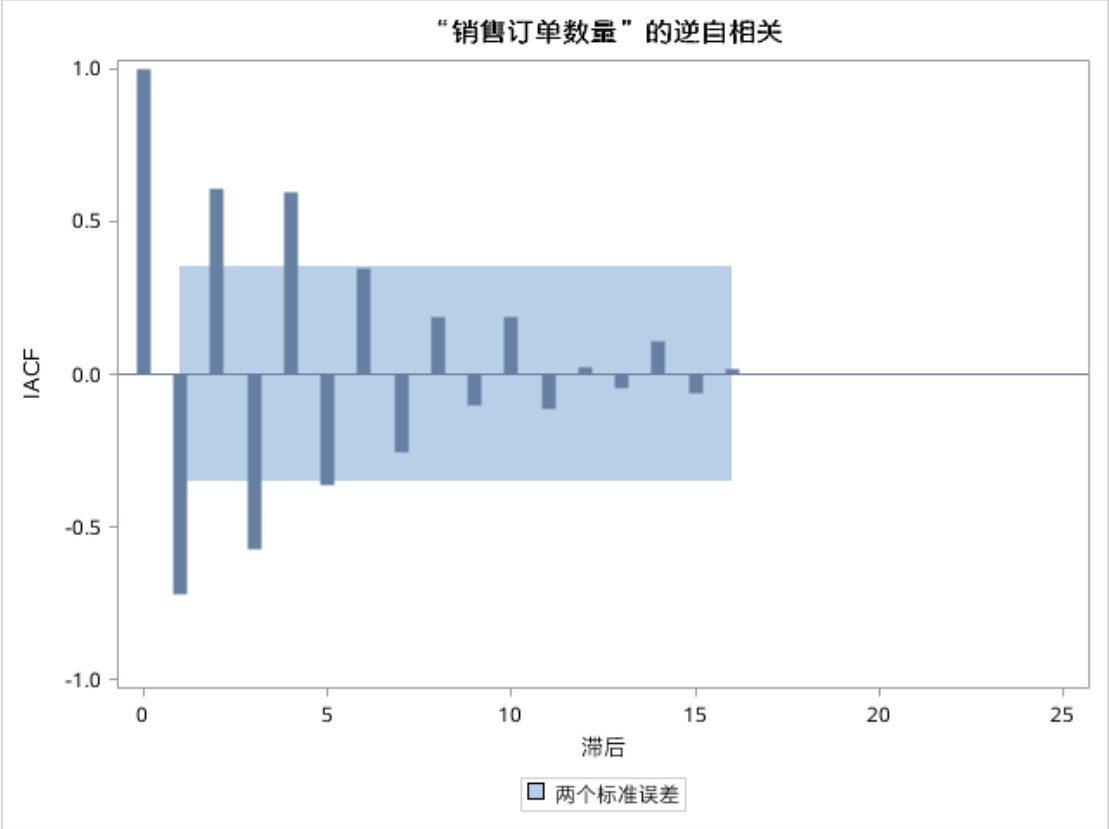


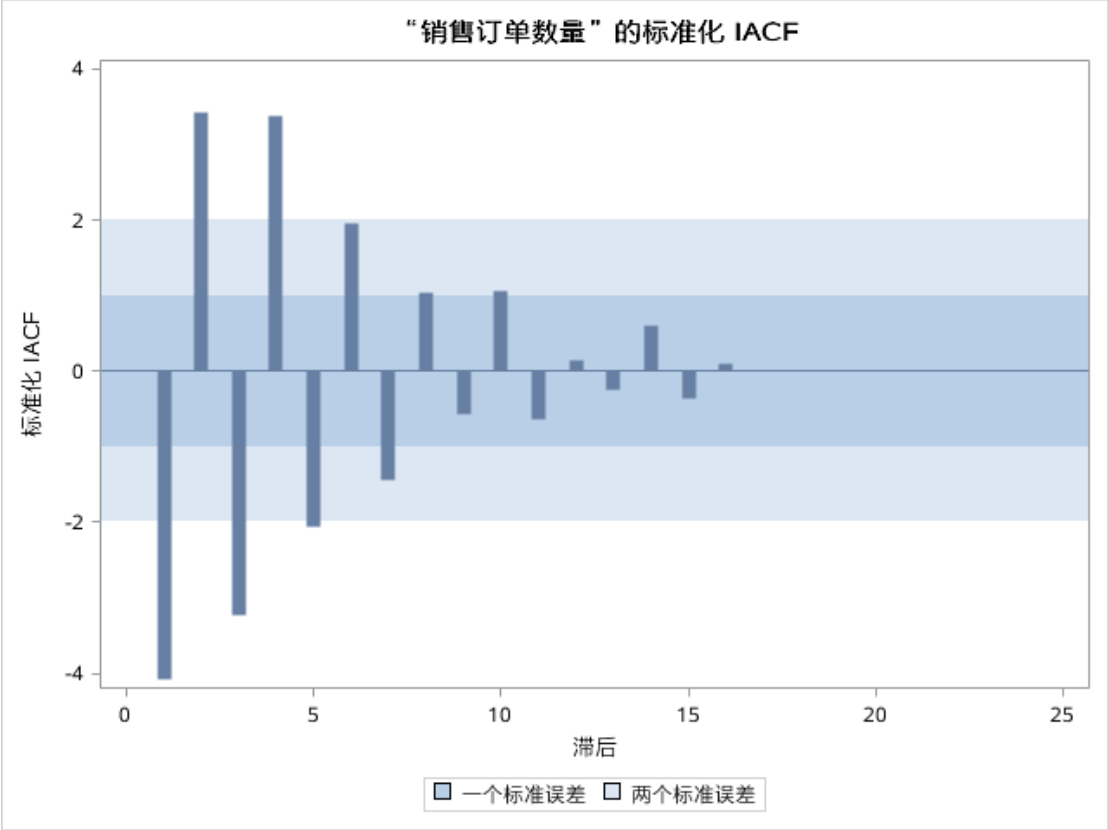

“销售订单数量”的自相关
ACF
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
20
25
滞后
两个标准误差


“销售订单数量”的标准化 ACF
标准化 ACF
4
3
2
1
0
-1
-2
0
5
10
15
20
25
滞后
一个标准误差
两个标准误差


“销售订单数量”的偏自相关
PACF
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
20
25
滞后
两个标准误差


“销售订单数量”的标准化 PACF
标准化 PACF
4
2
0
-2
0
5
10
15
20
25
滞后
一个标准误差
两个标准误差


“销售订单数量”的逆自相关
IACF
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
20
25
滞后
两个标准误差


“销售订单数量”的标准化 IACF
标准化 IACF
4
2
0
-2
-4
0
5
10
15
20
25
滞后
一个标准误差
两个标准误差