菜鸟-需求预测与分仓规划最终报告

成员：

崔绿叶　2120150981

陈帅　　2120150979

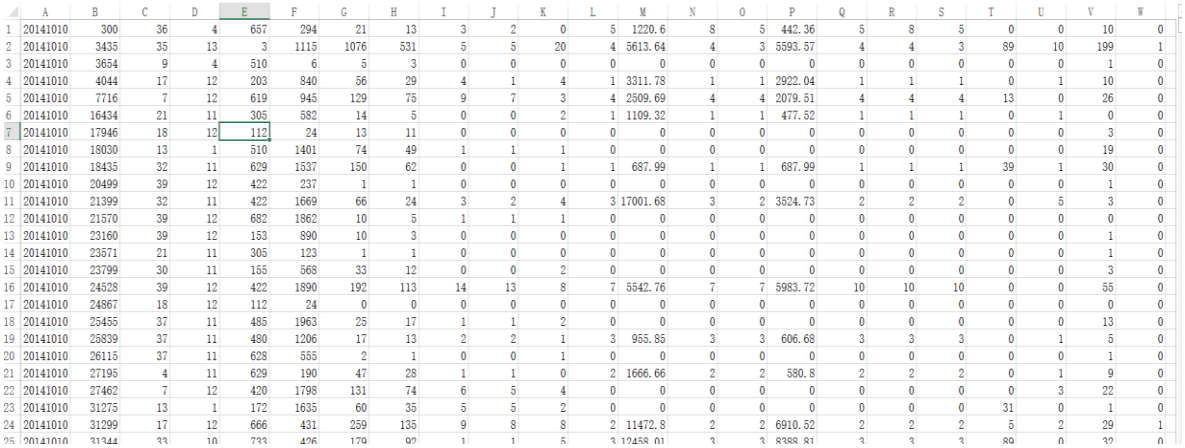
贺辉　　2120150991

1、数据分析：

题目提供了商品从20141010到20151227的全国和区域分仓数据，商品在全国的特征包括商品本身的一些分类：类目、品牌等，还有历史的一些用户行为特征：浏览人数、加购物车人数，购买人数等我们需要预测的未来需求是“非聚划算支付件数”(qty_alipay_njhs)；商品在区域的分仓历史数据其维度跟全国的数据一样，仅有的差别是这些数据表达的是某个仓负责的地理区域内的用户行为，比如qty_alipay_njhs在这里表达的是这个仓负责的区域内的用户的“非聚划算支付件数”。

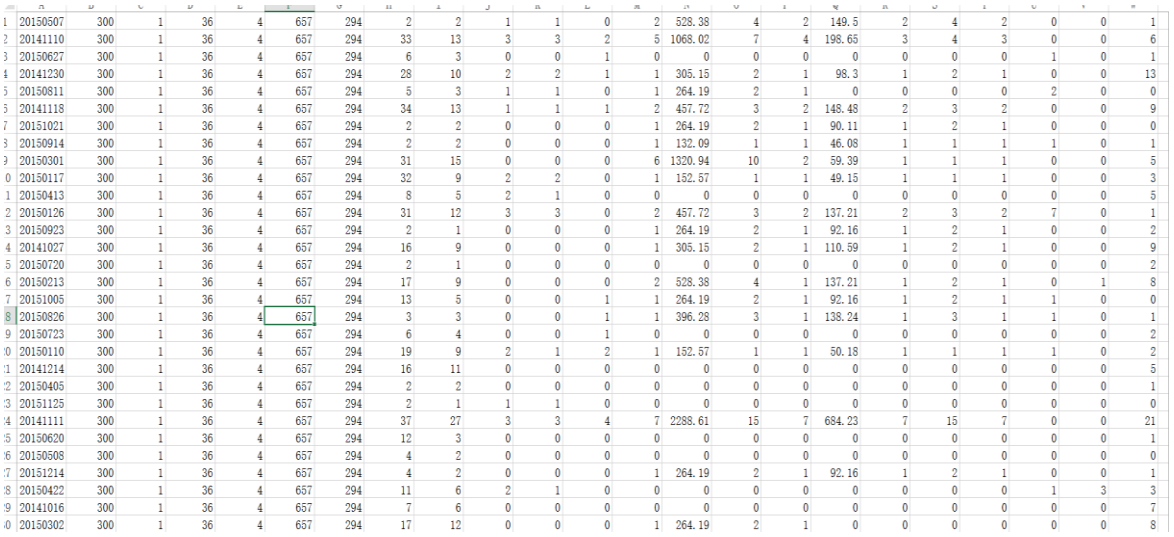
2、数据清理：

关于数据特征，我们使用了商品在全国和地区的“非聚划算支付件数”(qty_alipay_njhs)这一特征；对数据进行处理时，我们首先进行了数据库的建立、链接操作，这样对于数据中所存在的一些问题，比如重复数据等，我们直接使用SQL语言对其进行处理；经过这样的处理后，可以保证我们得到的数据更加有效，其清理具体过程在阶段报告中已经详细说明，在此不再赘述，其清洗前与清洗后的结果如下图所示：



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	20141010	300	36	4	657	294	21	13	3	2	0	5	1220.6	8	5	442.36	5	8	5	0	0	10	0
2	20141010	3435	35	13	3	1115	1076	531	5	5	20	4	5613.64	4	3	5593.57	4	4	3	89	10	199	1
3	20141010	3654	9	4	510	6	5	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	20141010	4044	17	12	203	840	56	29	4	1	4	1	3311.78	1	1	2922.04	1	1	1	0	1	10	0
5	20141010	7716	7	12	619	945	129	75	9	7	3	4	2509.69	4	4	2079.51	4	4	4	13	0	26	0
6	20141010	16434	21	11	305	582	14	5	0	0	2	1	1109.32	1	1	477.52	1	1	1	0	1	0	0
7	20141010	17946	18	12	112	24	13	11	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
8	20141010	18030	13	1	510	1401	74	49	1	1	1	0	0	0	0	0	0	0	0	0	0	19	0
9	20141010	18435	32	11	629	1537	150	62	0	0	1	1	687.99	1	1	687.99	1	1	1	39	1	30	0
10	20141010	20499	39	12	422	237	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
11	20141010	21399	32	11	422	1669	66	24	3	2	4	3	17001.68	3	2	3524.73	2	2	2	0	5	3	0
12	20141010	21570	39	12	682	1862	10	5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
13	20141010	23160	39	12	153	890	10	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
14	20141010	23571	21	11	305	123	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
15	20141010	23799	30	11	155	568	33	12	0	0	2	0	0	0	0	0	0	0	0	0	0	3	0
16	20141010	24528	39	12	422	1890	192	113	14	13	8	7	5542.76	7	7	5983.72	10	10	10	0	0	55	0
17	20141010	24867	18	12	112	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	20141010	25455	37	11	485	1963	25	17	1	1	2	0	0	0	0	0	0	0	0	0	0	13	0
19	20141010	25839	37	11	480	1206	17	13	2	2	1	3	955.85	3	3	606.68	3	3	3	0	1	5	0
20	20141010	26115	37	11	628	555	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
21	20141010	27195	4	11	629	190	47	28	1	1	0	2	1666.66	2	2	580.8	2	2	2	0	1	9	0
22	20141010	27462	7	12	420	1798	131	74	6	5	4	0	0	0	0	0	0	0	0	0	3	22	0
23	20141010	31275	13	1	172	1635	60	35	5	5	2	0	0	0	0	0	0	0	0	31	0	1	0
24	20141010	31299	17	12	666	431	259	135	9	8	8	2	11472.8	2	2	6910.52	2	2	2	5	2	29	1
25	20141010	31344	33	10	733	426	179	92	1	1	5	3	12458.01	3	3	8388.81	3	3	3	89	0	32	0

图 1. 数据清洗前



1	20150507	300	1	36	4	657	294	2	2	1	1	0	2	528.38	4	2	149.5	2	4	2	0	0	1
2	20141110	300	1	36	4	657	294	33	13	3	3	2	5	1068.02	7	4	198.65	3	4	3	0	0	6
3	20150627	300	1	36	4	657	294	6	3	0	0	1	0	0	0	0	0	0	0	0	1	0	1
4	20141230	300	1	36	4	657	294	28	10	2	2	1	1	305.15	2	1	98.3	1	2	1	0	0	13
5	20150811	300	1	36	4	657	294	5	3	1	1	0	1	264.19	2	1	0	0	0	0	2	0	0
6	20141118	300	1	36	4	657	294	34	13	1	1	1	2	457.72	3	2	148.48	2	3	2	0	0	9
7	20151021	300	1	36	4	657	294	2	2	0	0	0	1	264.19	2	1	90.11	1	2	1	0	0	0
8	20150914	300	1	36	4	657	294	2	2	0	0	0	1	132.09	1	1	46.08	1	1	1	1	0	1
9	20150301	300	1	36	4	657	294	31	15	0	0	0	6	1320.94	10	2	59.39	1	1	1	0	0	5
10	20150117	300	1	36	4	657	294	32	9	2	2	0	1	152.57	1	1	49.15	1	1	1	0	0	3
11	20150413	300	1	36	4	657	294	8	5	2	1	0	0	0	0	0	0	0	0	0	0	0	5
12	20150126	300	1	36	4	657	294	31	12	3	3	0	2	457.72	3	2	137.21	2	3	2	7	0	1
13	20150923	300	1	36	4	657	294	2	1	0	0	0	1	264.19	2	1	92.16	1	2	1	0	0	2
14	20141027	300	1	36	4	657	294	16	9	0	0	0	1	305.15	2	1	110.59	1	2	1	0	0	9
15	20150720	300	1	36	4	657	294	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
16	20150213	300	1	36	4	657	294	17	9	0	0	0	2	528.38	4	1	137.21	1	2	1	0	1	8
17	20151005	300	1	36	4	657	294	13	5	0	0	1	1	264.19	2	1	92.16	1	2	1	1	0	0
18	20150826	300	1	36	4	657	294	3	3	0	0	1	1	396.28	3	1	138.24	1	3	1	1	0	1
19	20150723	300	1	36	4	657	294	6	4	0	0	1	0	0	0	0	0	0	0	0	0	0	2
20	20150110	300	1	36	4	657	294	19	9	2	1	2	1	152.57	1	1	50.18	1	1	1	1	0	2
21	20141214	300	1	36	4	657	294	16	11	0	0	0	0	0	0	0	0	0	0	0	0	0	5
22	20150405	300	1	36	4	657	294	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1
23	20151125	300	1	36	4	657	294	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
24	20141111	300	1	36	4	657	294	37	27	3	3	4	7	2288.61	15	7	684.23	7	15	7	0	0	21
25	20150620	300	1	36	4	657	294	12	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1
26	20150508	300	1	36	4	657	294	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	20151214	300	1	36	4	657	294	4	2	0	0	0	1	264.19	2	1	92.16	1	2	1	0	0	1
28	20150422	300	1	36	4	657	294	11	6	2	1	0	0	0	0	0	0	0	0	0	1	3	3
29	20141016	300	1	36	4	657	294	7	6	0	0	0	0	0	0	0	0	0	0	0	0	0	7
30	20150302	300	1	36	4	657	294	17	12	0	0	0	1	264.19	2	1	0	0	0	0	0	0	8

图 2 数据清洗后

3、利用 ARIMA 模型进行预测：

ARIMA 模型是自回归移动平均模型（AutoRegressive Moving Average Models）的简称，作为一种经典时间序列预测方法，其主要思想是将非平稳时间序列转化为平稳时间序列，然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归。其预测过程如下：

（1）将收集到的数据以某个时间点为界限分割为训练集和验证

集。对训练集绘制时序图，以判别其是否具有平稳性和季节性特征；

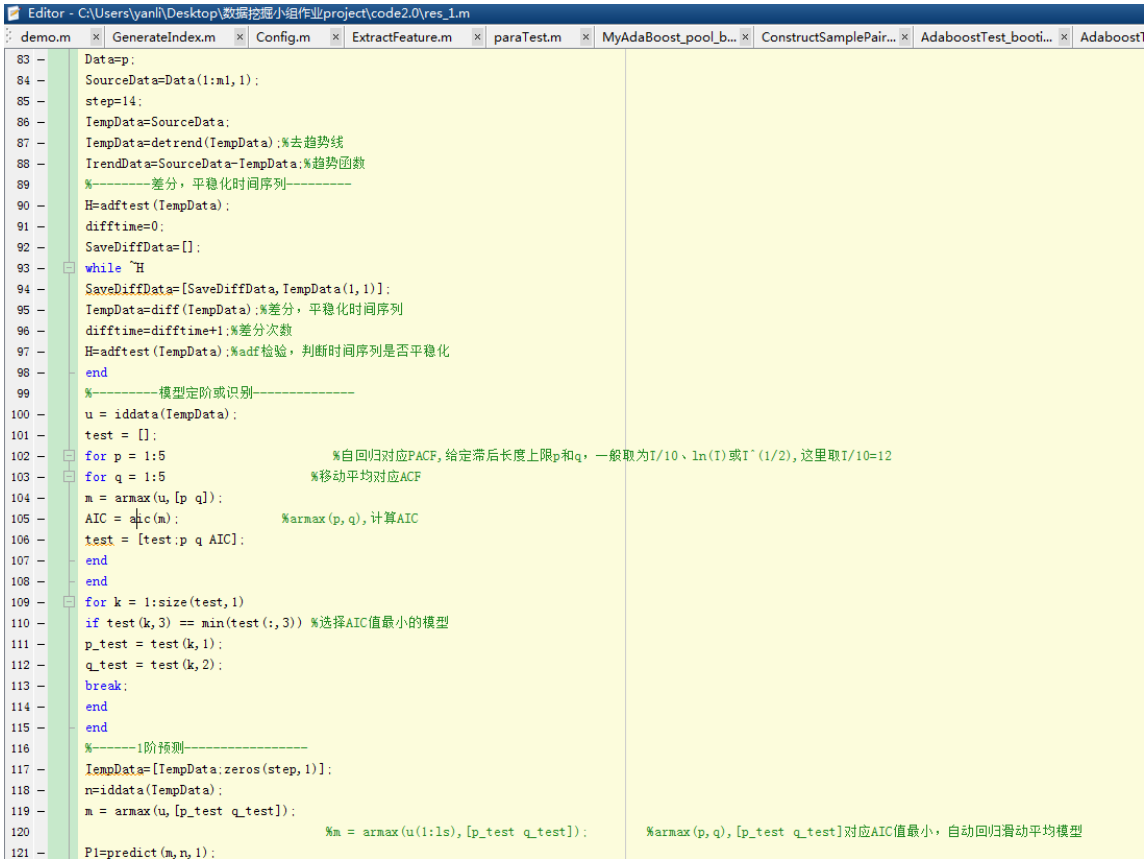
（2）对非平稳序列进行平稳化处理，一般采用差分处理，且处理后的数据能够通过单位根检验；

（3）根据时间序列模型的特征进行模型识别和参数估计，建立相应的模型；

（4）对模型进行假设检验，采用 Ljung-Box 检验残差序列是否为白噪声；

（5）利用已通过检验的模型进行预测模型分析并和验证集进行对比，评估模型的拟合精度。

其相应地代码如下：



```
%------1阶预测-----------------
IempData=[IempData;zeros(step,1)];
n=iddata(IempData);
m = armax(u,[p_test q_test]);
                                %m = armax(u(1:ls),[p_test q_test]);      %armax(p,q),[p_test q_test]对应AIC值最小，自动回归滑动平均模型
P1=predict(m,n,1);
PreR=P1.OutputData;
PreR=PreR';
%----------还原差分-----------------
if size(SaveDiffData,2)~=0
for index=size(SaveDiffData,2):-1:1
PreR=cumsum([SaveDiffData(index),PreR]);
end
end
  %--------------------预测趋势并返回结果----------------
mp1=polyfit([1:size(IrendData',2)],IrendData',1);
xt=[];
for j=1:step
xt=[xt,size(IrendData',2)+j];
end
IrendResult=polyval(mp1,xt);
PreData=IrendResult+PreR(size(SourceData',2)+1:size(PreR,2));
tempx=[IrendData',IrendResult]+PreR;    % tempx为预测结果
```

根据这个预测模型，我们可以得到相应的预测结果。由于 arima 模型要求输入数据满足平顺性，在试验中我们使用 adftest 函数对输入数据进行平顺性检测，如下图 1 展示的是一个 ID300 的商品的不平滑数据。

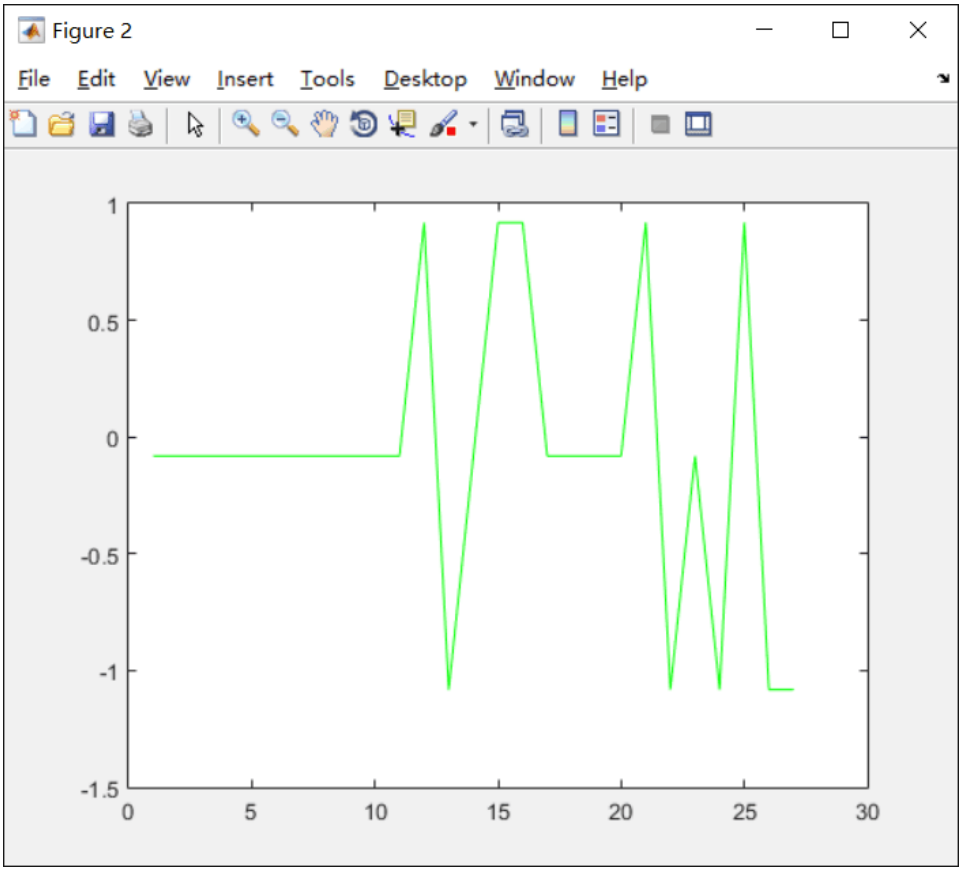


图 1 未差分结果

对于不满足平顺性的数据我们需要做一介差分，使数据满足模型的要求。如下图 2 为数据做一介差分的处理后结果：

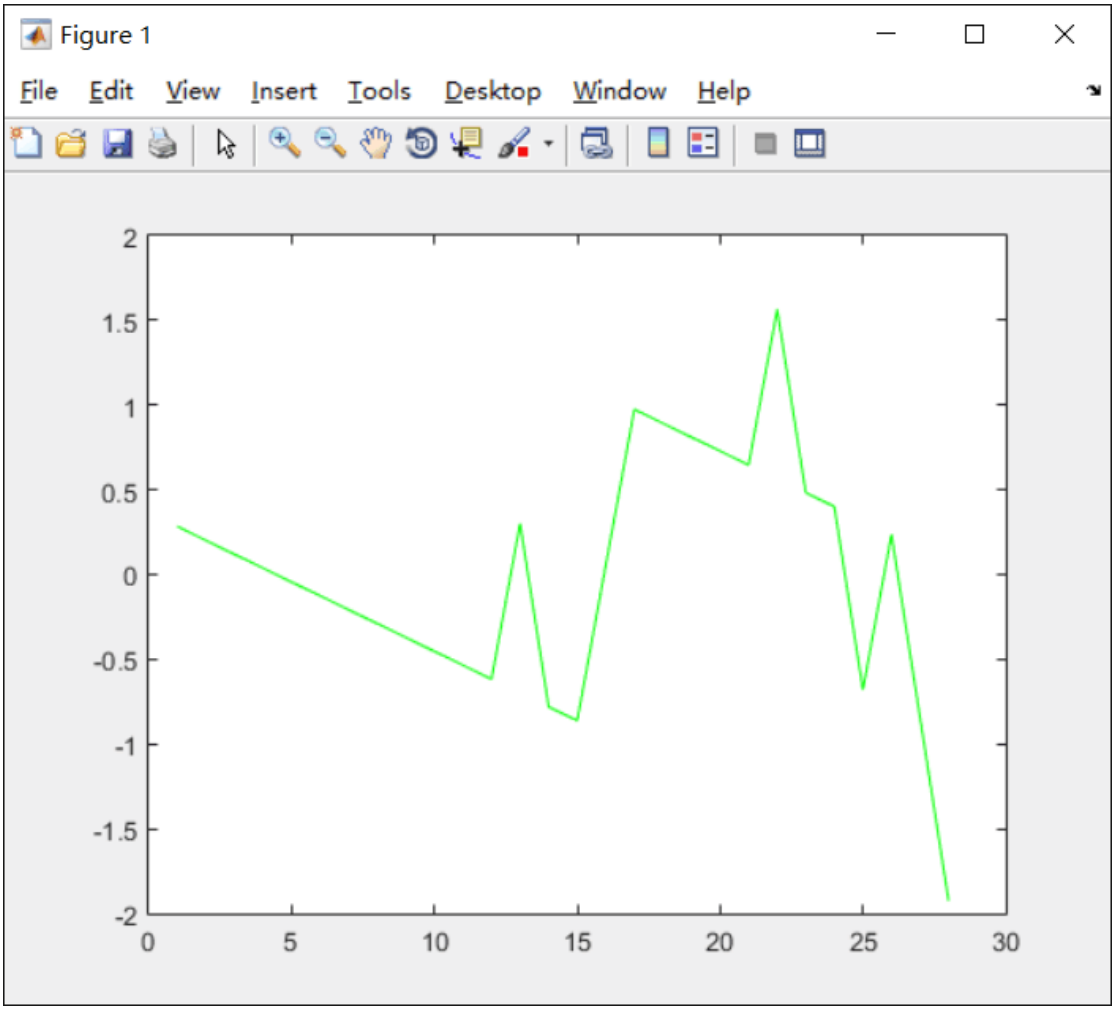


图 2 数据差分结果

满足了数据要求之后，我们使用代码自动调整 armax 的参数 P，Q，并选择出最优的参数。并用此模型参数进行进一步的数据预测。数据预测结果如下图 3 所示，可见该模型对于数据量较小，且不满足平顺性的数据而言表现出较差的性能：

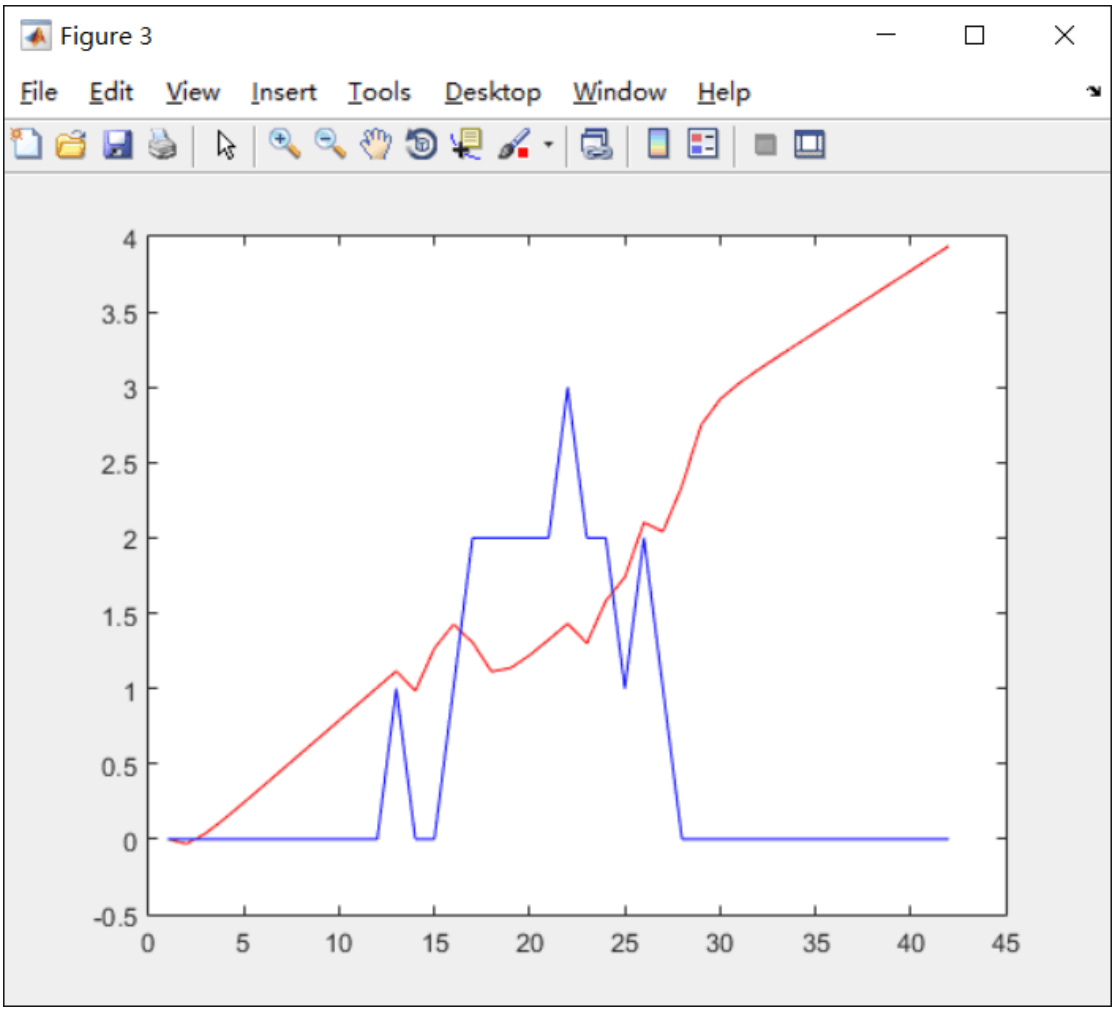


图 3 数据预测结果

其中，图中蓝色线表示实际数据，红色线表示预测数据。

为做进一步的对比，我们使用 id400 的商品数据进行预测分析，该商品数据满足平顺性，且具有一定的数据规模。原始数据如下图 4 所示：

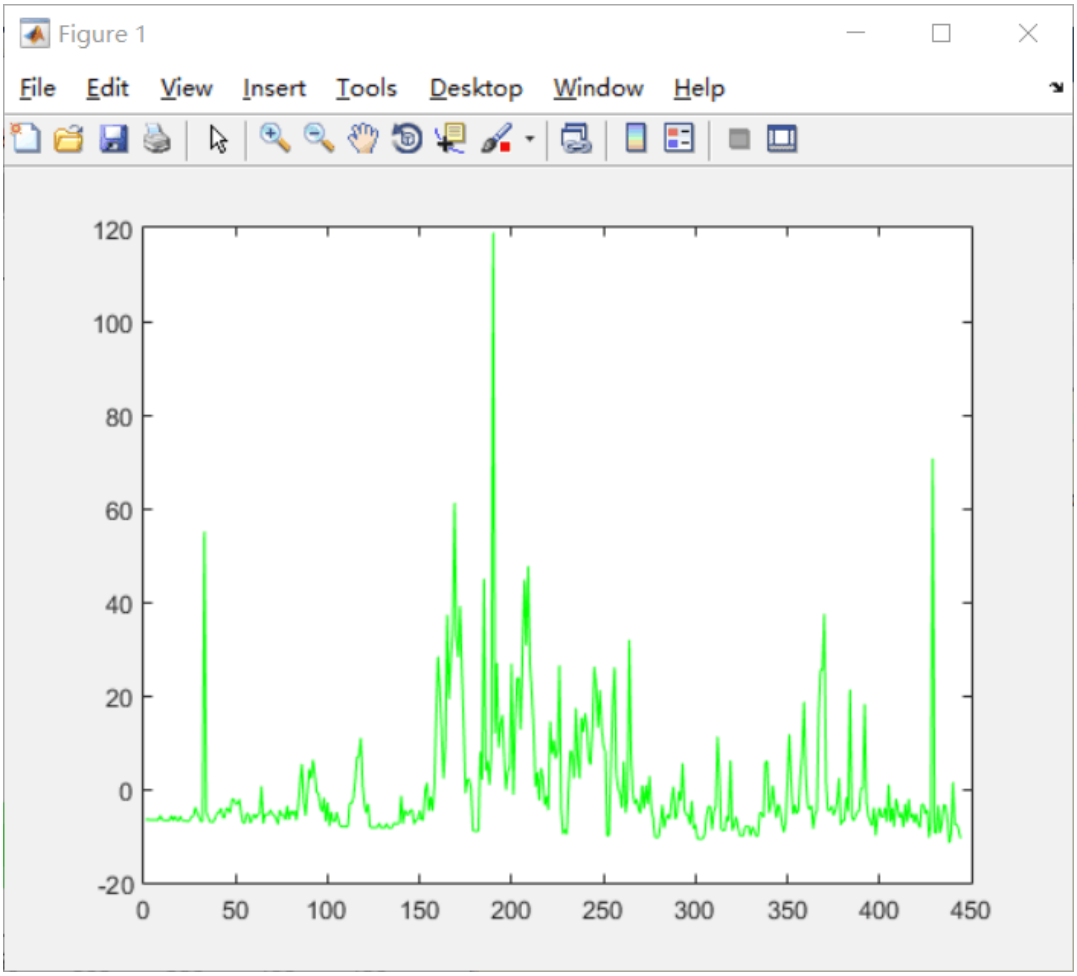


图 4 满足数据平顺性的原始数据

用该数据进行预测的结果如下图 5 所示，可见该模型在这种数据中的预测中表现出了良好的性能。

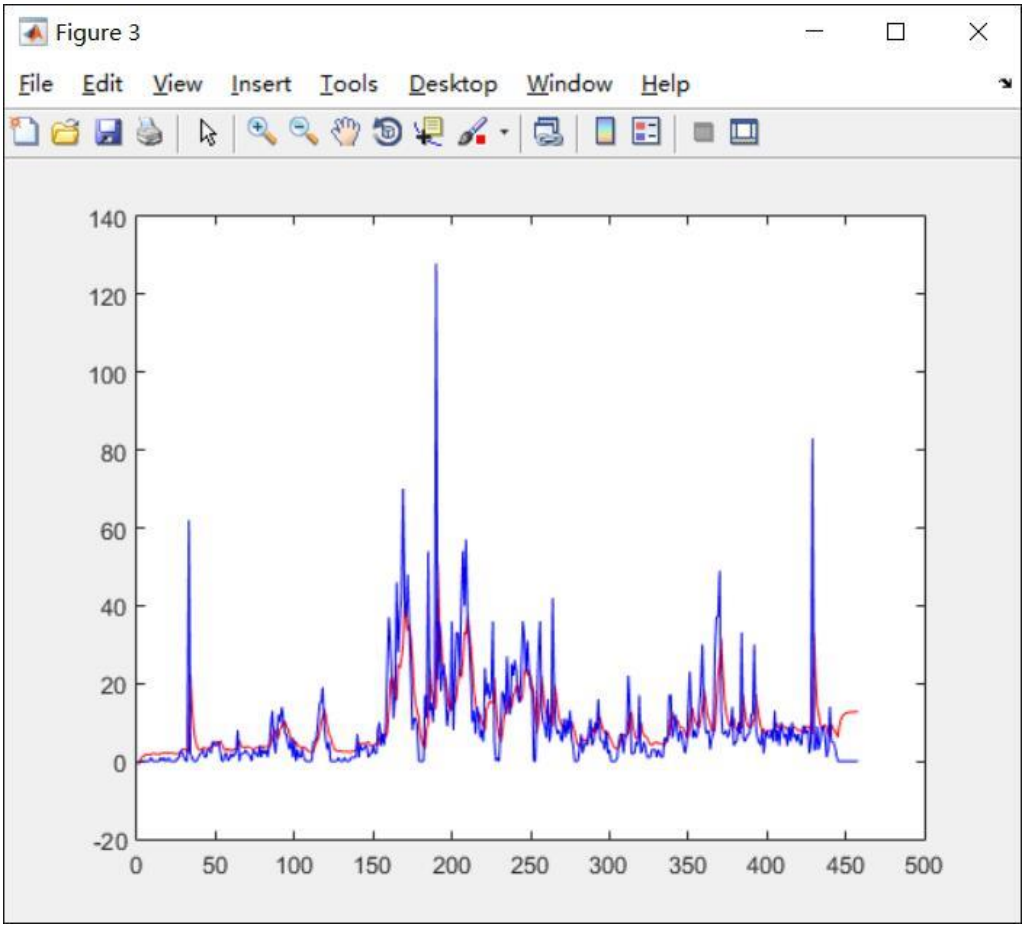


Figure 3
File Edit View Insert Tools Desktop Window Help
140
120
100
80
60
40
20
0
-20
0 50 100 150 200 250 300 350 400 450 500

图 5 数据预测结果