

Time Series Analysis—Empirical Project

我国国内啤酒销售量的趋势分析及预测初探

2017 年 12 月 27 日

目录

1	研究背景	1
2	描述性统计分析	1
2.1	整体分析	1
2.2	ACF 与 PACF	2
3	研究方法及结果	4
3.1	指数平滑方法	4
3.2	ARIMA 模型	6
3.3	BSTS 模型方法	8
4	研究结论与展望	11

1 研究背景

随着大数据时代的来临，社会经济活动中产生的海量数据对经济发展、生产实践起到了越来越重要的指导作用。例如在商品销售中，若能根据某类商品以往各年度、各季度的历史销售数据，利用数据科学和统计学中的方法，准确的分析出该产业的增长和发展趋势，同时据此对未来年度该商品的销售情况进行有效预测，无疑对政府进行宏观经济管理、分配社会资源，以及对生产企业合理安排生产大有裨益。

基于上述思想和社会生产实践中提出的新课题，本文尝试利用时间序列分析和预测科学中的方法，尝试分析我国 1998-2015 年啤酒销售量的时间序列数据，对我国啤酒销售量数据所呈现出的模式进行了初步的研究与讨论。与此同时，我们还利用三种预测方法，试着对未来 5 个季度的啤酒销量进行了科学预测，取得了较为满意的预测效果，同时也为今后有关研究提供了有益的借鉴。

2 描述性统计分析

2.1 整体分析

本文的数据获取自国家统计局网站。在对数据进行清洗和整理之后，我们对数据进行了描述性分析。数据包含自 1998 年第一季度至 2015 年第四季度的全国啤酒销售量，共 72 条数据。为了便于之后的统计分析和对啤酒销售量进行有效预测，我们首先绘制了 1998-2015 年我国啤酒销售量的趋势图，如下图 1 所示。与此同时，为了更好地识别序列中包含的季节性因素，我们还以年为单位，绘制出了春、夏、秋、冬四季啤酒销量的箱线图，如图 2 所示。

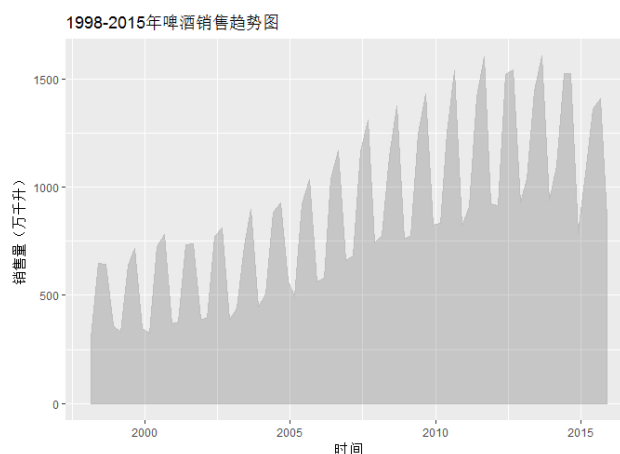


图 1：1998-2015 年我国啤酒销量趋势图

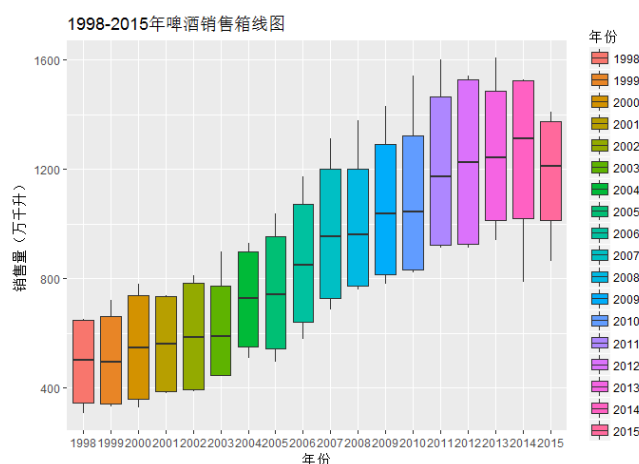


图 2：1998-2015 年分年度啤酒销量箱线图

由以上两图不难看出，中国国内啤酒销售量存在着明显地趋势性和季节性特征：啤酒销量在随时间的增加而增加的同时，季节性因素对啤酒销量也产生了显著地影响。鉴于上述有关啤酒销售量的经验特征，在进行分析时，必须同时从原序列中提取趋势性因素和季节性因素。

2.2 ACF 与 PACF

此外，为了进一步研究我国啤酒销售量序列中所包含的结构性和相依性，我们接下来绘制了原序列的 ACF 图和 PACF 图，如图 3 所示。

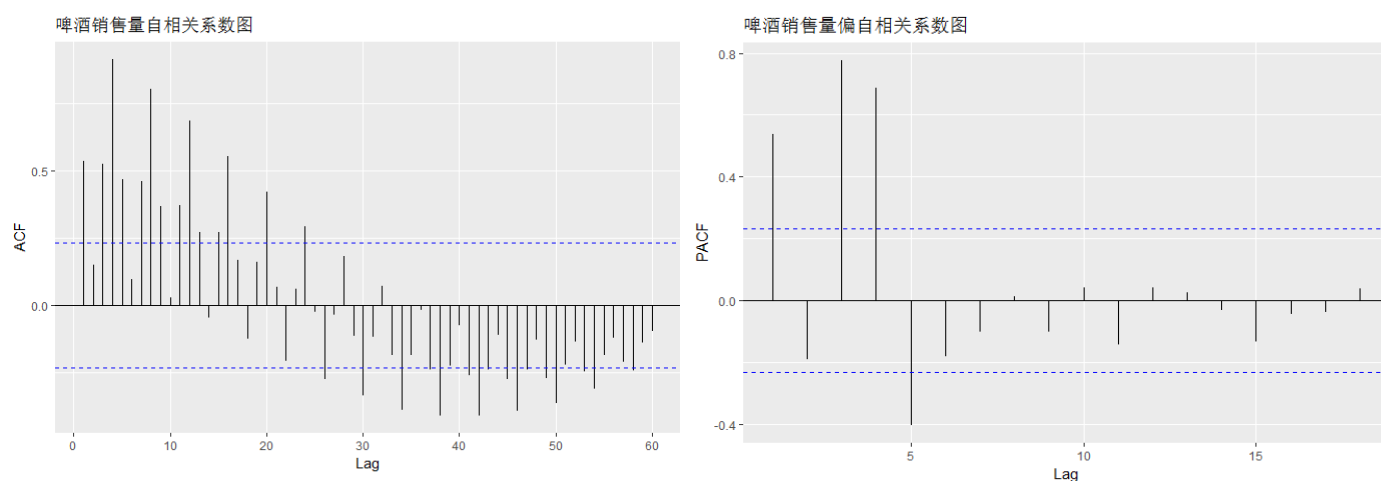


图 3：啤酒销量 ACF 图、PACF 图

由上图可以明显看出，啤酒销量数据明显是非平稳的，而非平稳的序列显然无法直接进行统计分析，于是我们便对原序列 $\{y_t\}$ 进行了一阶差分，得到序列 $\{\Delta(y_t)\}$ ，如图 4 所示。

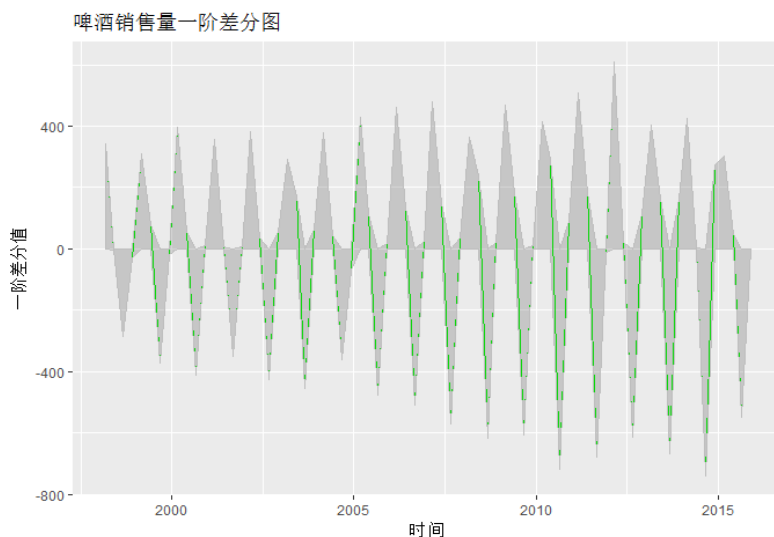
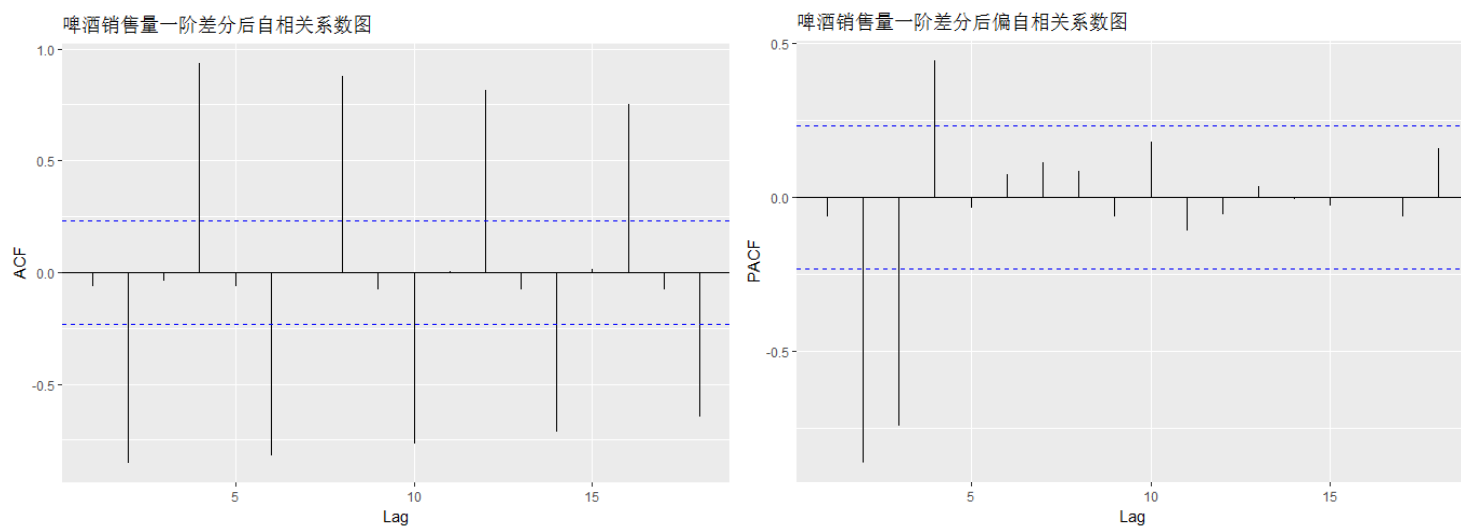


图 4: 啤酒销售量一阶差分图

同时, 为了研究 $\{\Delta(y_t)\}$ 的平稳性, 我们还绘制了 1 阶差分后序列 $\{\Delta(y_t)\}$ 的

ACF 及 PACF 图, 如图 5 所示。

图 5: $\{\Delta(y_t)\}$ 序列 ACF PACF 图

考虑到数据含有明显的季节性特征, 我们还绘制了 4 阶季节性差分后的图以及其 ACF、PACF 图。

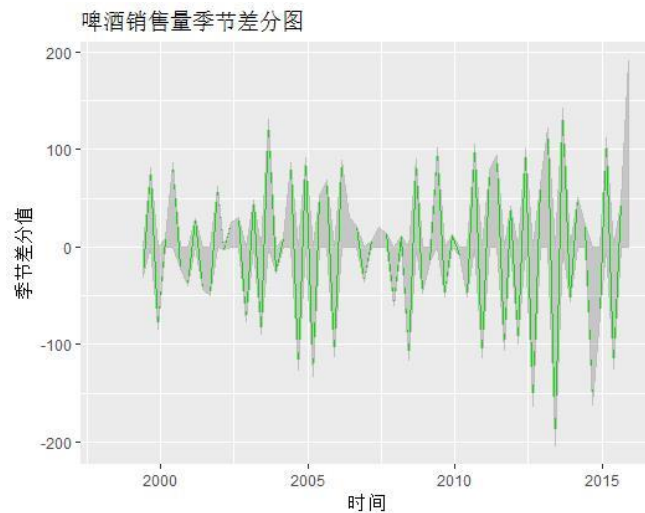


图 6: 四阶季节差分图

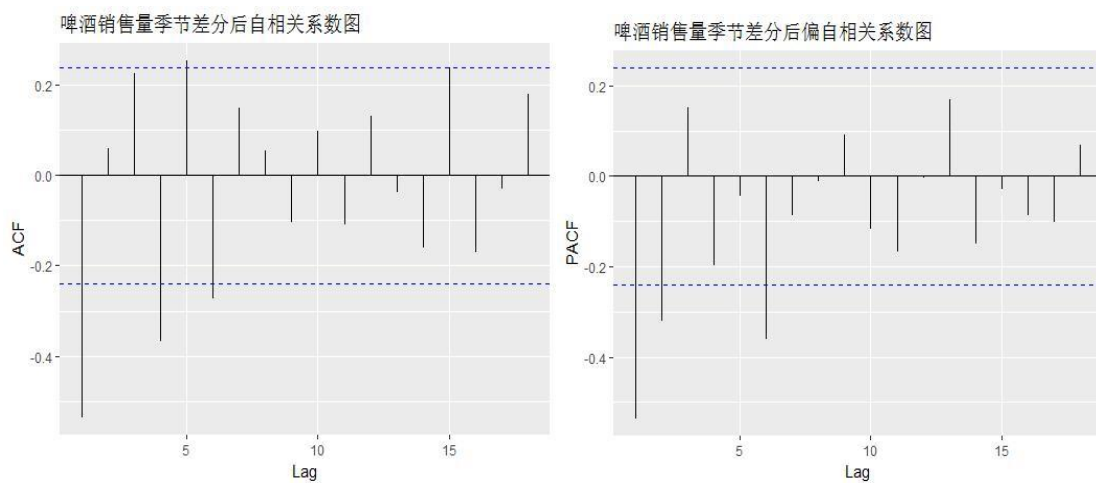


图 7: 4 阶差分 ACF、PACF 图

3 研究方法及结果

3.1 指数平滑方法

指数平滑法是移动平均法中的一种,其特点在于给过去的观测值不一样的权重,对不同的数据给予不同的权重,即近期观测值给予较大的权重,比较远期观测值给予较小的权重,预测值是以前观测值的加权和。

为了采用指数平滑法进行预测,我们首先将 1998 年第一季度至 2014 年第三季度的数据作为训练集进行建模,并用 2014 年第四季度至 2015 年第四季度数据作为预测集。

指数平滑方法将数据分为长期趋势、季节性因素以及随机误差项。

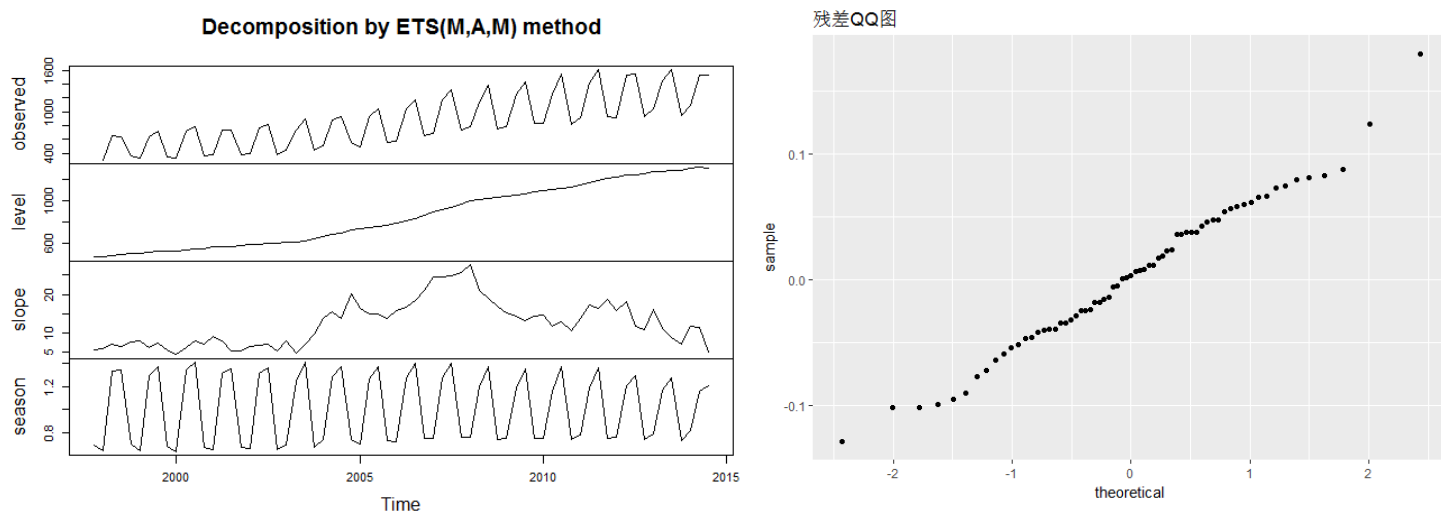


图 8：指数平滑分解与残差 QQ 图

由残差 Q-Q 图可见：利用差分后的序列 $\{\Delta(y_t)\}$ 绘制的残差 Q-Q 图，数据点基本在对角线附近，因此正态性要求大体上得到了满足，因此，我们便可着手进行接下来的预测工作。

根据样本数据和指数平滑模型，我们得出了以下预测值，具体见表 1 所示：

表 1：ETS 模型预测值

季度	2014Q4	2015Q1	2015Q2	2015Q3	2015Q4
真实值	787.8	1060.9	1361.9	1409.6	861.3
预测值	969.12	1083.60	1536.52	1608.28	984.80

同时，我们由建立的 ETS 模型，绘制出了预测值与真实值的比较图，如图所示。从该图中可以看出，在趋势性与季节性兼具的啤酒销售量数据中，ETS 模型具有相当理想的预测效果，不仅捕捉到了序列的趋势性特征，还较为准确的刻画了啤酒销量序列数据的季节性因素。

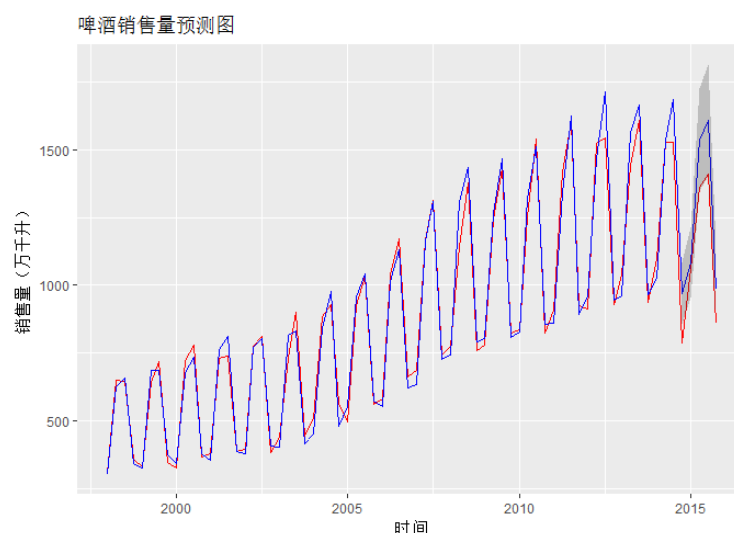


图 9: ETS 模型预测图

此外,我们还计算了 ETS 模型的预测误差,得出:

$$\text{平均误差 } ME = -140.16$$

$$\text{平均均方误差 } RMSE = 154.02$$

$$\text{平均绝对误差 } MAE = 140.16$$

通过与原序列的数量级进行比较,可以看出,指数平滑模型对同时包含趋势性因素和季节性因素的啤酒销量序列具有较为良好的预测效果。

通过指数平滑预测,了解到了其优缺点:优点:(1)对不同时间的数据的非等权处理较符合实际情况。(2)实用中仅需选择一个模型参数,即可进行预测,简便易行。(3)具有适应性,也就是预测模型能自动识别数据模式的变化而加以调整。缺点是:(1)对数据的转折点缺乏鉴别能力。(2)长期预测的效果较差,故多用于短期预测。

3.2 ARIMA 模型

ARIMA 模型也叫做自回归移动平均模型,全称是(Autoregressive Integrated Moving Average Model),是统计中最常见的一种用来进行时间序列预测的模型。ARIMA 模型有三个参数:p,d,q。模型表示为 ARIMA(p,d,q)

- p--代表预测模型中采用的时序数据本身的滞后数(lags),也叫做 AR/Auto-Regressive 项
- d--代表时序数据需要进行几阶差分化,才是稳定的,也叫 Integrated 项。
- q--代表预测模型中采用的预测误差的滞后数(lags),也叫做 MA/Moving

Average 项

接下来，我们尝试利用刻画非平稳时间序列的 ARIMA 模型来对我国的啤酒销售量数据额进行建模。具体而言，我们采用 R 语言中的 `auto.arima()`函数建立 ARIMA 模型，得到 ARIMA(1, 1, 1)(0, 1, 0)模型。其 `ar` 系数为 0.4762，`ma` 系数为-0.8445（修改），然后绘制了关于 ARIMA 模型残差的 Q-Q 图。

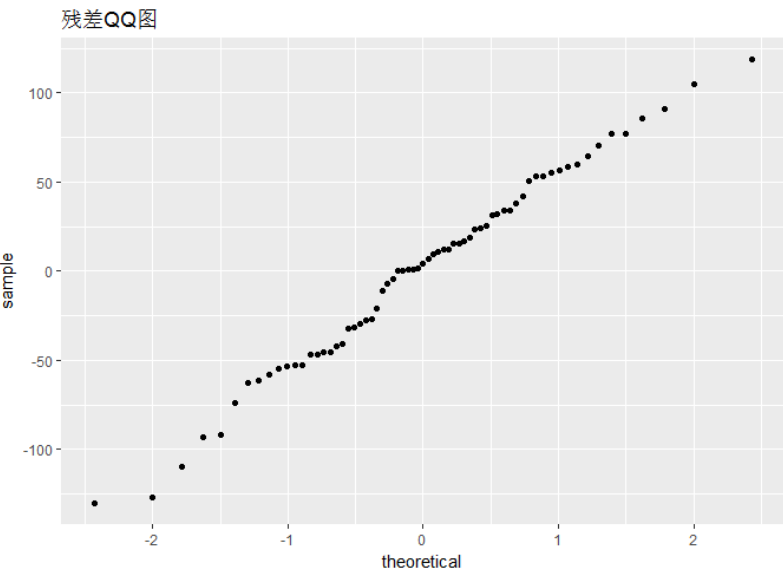


图 10: 残差 Q-Q 图

从 ARIMA 模型残差的 Q-Q 图可见，其数据点基本上分布在 45 度对角线附近，因此，我们有理由认为建立的 ARIMA 是充分的。

接下来，我们便利用建立的 ARIMA 模型，对全国啤酒销售量进行了预测，结果如下表 2 所示。

表 2: ARIMA 模型预测值

季度	2014Q4	2015Q1	2015Q2	2015Q3	2015Q4
真实值	787.8	1060.9	1361.9	1409.6	861.3
预测值	987.57	1148.77	1575.27	1576.47	1037.24

同时，我们由建立的 ARIMA 模型，绘制出了预测值与真实值的比较图，如下图所示。从该图中可以看出，ARIMA 模型也取得了较为理想的预测效果，

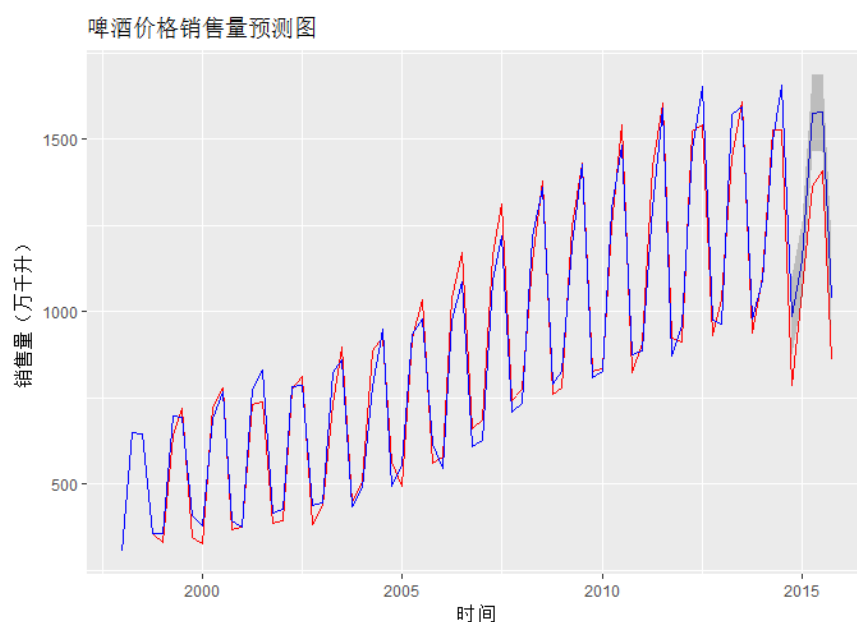


图 11: ARIMA 模型预测图

同时，ARIMA 模型就预测误差指标方面，我们得出如下结果：

$$\text{平均误差 } ME = -168.76$$

$$\text{平均均方误差 } RMSE = 174.33$$

$$\text{平均绝对误差 } MAE = 168.77$$

通过 ARIMA 预测，我们了解了其优缺点。优点是：模型十分简单，只需要内生变量而不需要借助其他外生变量。缺点是：1. 要求时序数据是稳定的（stationary），或者是通过差分化(differencing)后是稳定的。2. 本质上只能捕捉线性关系，而不能捕捉非线性关系。

3.3 BSTS 模型方法

接下来本文将采用 r 包 bsts 方法来进行建模。BSTS 是一个模型集合体，BSTS 结合了三种统计方法进入一个集体系统中（两个部分：趋势部分+回归部分）。用基础的状态空间结构模型做趋势和季节拟合（模型的趋势部分）；用卡尔曼滤波用来估计先验 p 的分布再用 spike-slab 回归做变量选择（模型的回归部分）；用贝叶斯模型平均法（Bayesian model averaging）做预测；就估计方法而言，BSTS 模型方法用 MCMC 模拟后验分布生成的样本集，来进行估计。

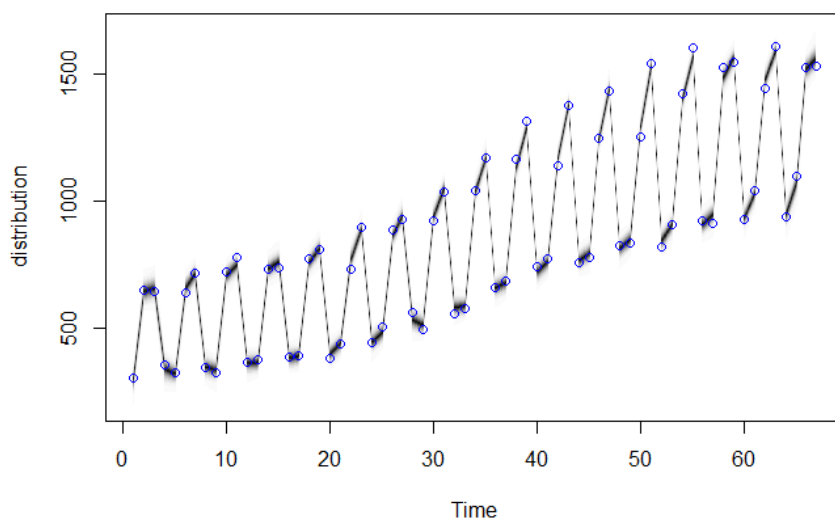


图 12: 预测拟合图

同时, 该模型将数据分为两部分, 一个时间趋势一个季节性趋势, 下图分别表示利用 BSTS 模型得到的啤酒销量数据中的时间趋势和季节性趋势:

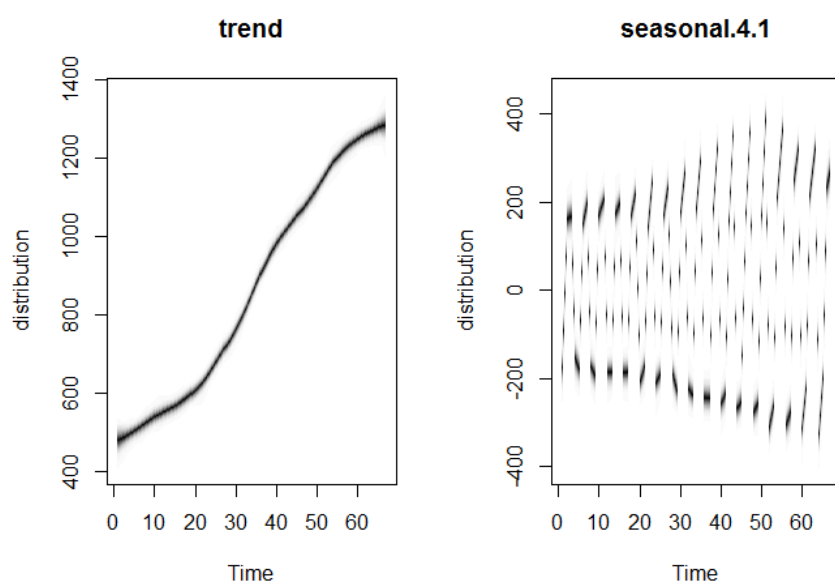


图 12 BSTS 模型的趋势分解

同时, 我们还得到了模型的残差图:

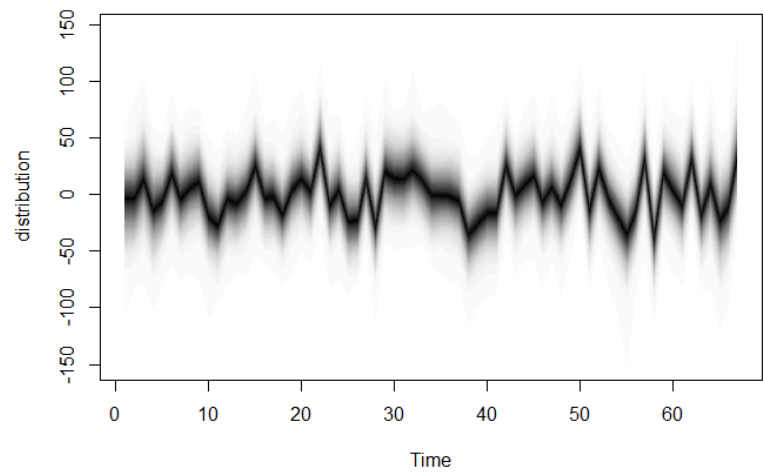


图 13：残差图

接下来，采用 BSTS 模型进行预测，得到的预测结果如下表所示：

表 3：BSTS 模型预测值

季度	2014Q4	2015Q1	2015Q2	2015Q3	2015Q4
真实值	787.8	1060.9	1361.9	1409.6	861.3
预测值	981.4270	1093.4720	1531.0268	1579.7386	999.0906

同时，我们将 BSTS 模型得到的预测值与真实值绘制在同一张图中，不难看出，利用 BSTS 模型进行预测，可以得到很好的预测结果，同时预测的置信区间也保持在一个相对较小的范围内。

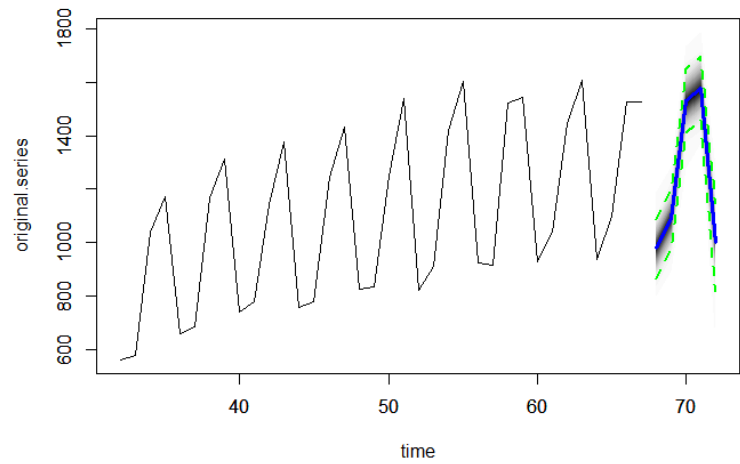


图 14：预测趋势图

进一步地，就预测误差指标而言，我们计算出如下结果：

$$\text{平均误差 } ME = -137.57$$

$$\text{平均均方误差 } RMSE = 149.11$$

$$\text{平均绝对误差 } MAE = 137.57$$

4 研究结论与展望

本文利用我国 1998-2015 年的全国啤酒销售量数据，利用描述性统计和时间序列分析当中有关的方法，成功地分离出我国啤酒销售量历史数据中所呈现出的趋势因素与季节性因素，而正确的识别和分离这两个成分，对于广大啤酒加工生产企业合理分配资源、科学组织生产都具有重要的指导作用。

此外，本文还利用我国啤酒销售量的历史数据，构建了（1）指数平滑、（2）ARIMA 和（3）BSTS 三种不同的预测模型，对我国啤酒的年销售量进行了科学而有效的预测，具有重要的理论意义和实践价值。

值得注意的是，本文所构建的三种模型得出的预测值相对于真实值而言，均呈现出高估的特点，而产生这一现象的原因和机制目前尚不明朗，可能与我国啤酒销量历史数据中所蕴含的一些结构性特征有关，亦或是与我们所选择的三种预测模型的特性有关，这无疑为今后对该问题进行更加深入透彻的研究提出了一个具体的现实问题，也为今后的研究指明了一个可行的方向。