# Time Series Analysis of American Candy Production

包亚杰 *10153700321* 熊维佳 *10153700424*

*2018* 年 *6* 月 *18* 日

## 目录

# 1 数据以及描述性统计

数据选取了美国 1972 年 1 月至 2017 年 8 月糖果生产量的月度时间序列数据，数据量 548。从时序图中可以看出序列具有一定趋势性以及周期性。（数据来源：https://www.kaggle.com/rtatman/us-candy-production-by-month/data）

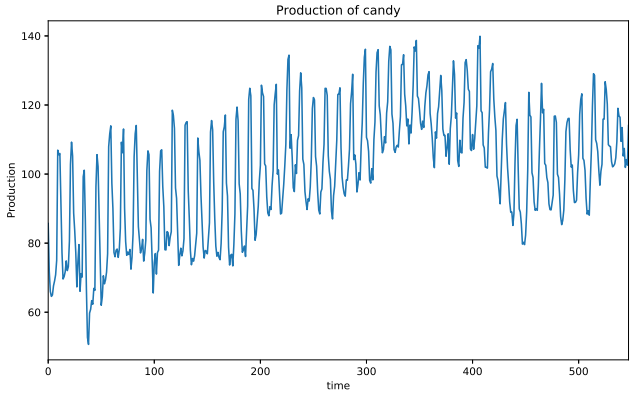



图 1: Time Series Plot of US Candy Production

本文首先对数据进行了描述性统计分析以及相关检验，结果如表 1 所示。结合核密度估计图可以看出，数据偏度小于 0，分布稍微左偏，峰度明显大于 0，相较于正态分布尾更厚。从 JB 检验结果来看，p 值为 0.0016，说明分布为非正态。ADF 单位根检验结果表明序列不存在单位根，Box 检验说明序列非白噪声。

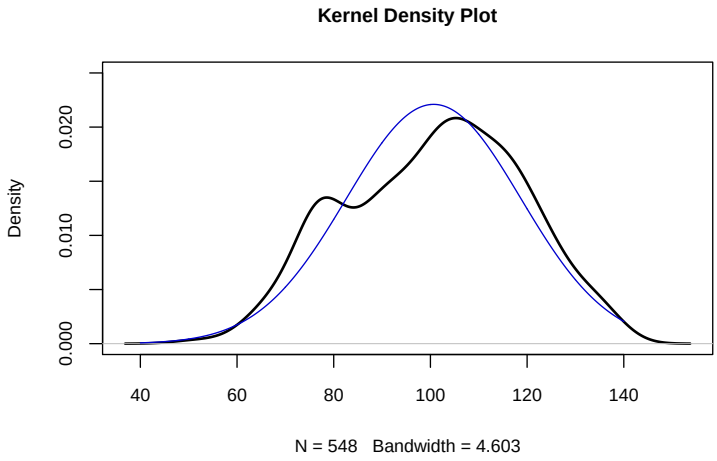



图 2: Kernel Density Plot

本文进一步将序列的季节指数提取出来，月度季节指数的计算方法如下：

$$S_i = \frac{\bar{X}_{i.}}{\bar{X}}, i = 1, 2, ..., 12,$$

表 1: Descriptive Statistics

| | | | |
|---|---|---|---|
| Min | 50.6689 | JB Stat. | 12.8661 |
| Max | 139.9153 | p-value(JB) | 0.0016 |
| Mean | 100.6625 | ADF Stat. | -3.8511 |
| Sd. | 18.0529 | p-value(ADF) | 0.0164 |
| Skewness | -0.1496 | Box Stat. | 418.6124 |
| Kurtosis | 2.3116 | p-value(Box) | 0.0000 |

其中 $\bar{X}_{i\cdot}$ 为第 $i$ 个月的糖果产量的平均值，$\bar{X}$ 为样本序列的均值。序列季节指数如图 3 所示，从图中可以看出，从 12 月到次年的 5 月，美国糖果生产量处于下降阶段，并且在 4 月到 7 月保持一个较低的水平，随后开始上升，在 11 月时达到最高。

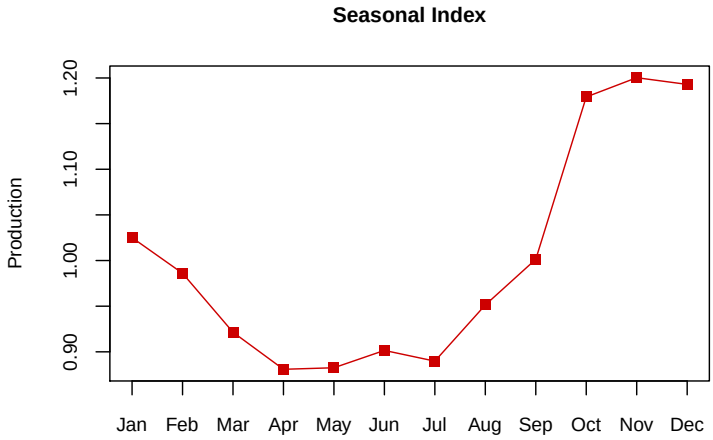



图 3: Seasonal Index Plot

# 2 参数模型

## 2.1 ARMA 模型

图4为原始时间序列的 ACF 以及 PACF 图，从图中可以看出 ACF 值以及 PACF 值在 lag=12 处都很高，可知原数据非平稳且有周期性，我们首先对它做 12 步差分，图5是 12 步差分的后时间序列图。对 12 步差分后的数据进行平稳性检验，结果如表2所示，说明 12 步差分后的序列平稳。图6是 12 步差分后的 ACF 和 PACF。

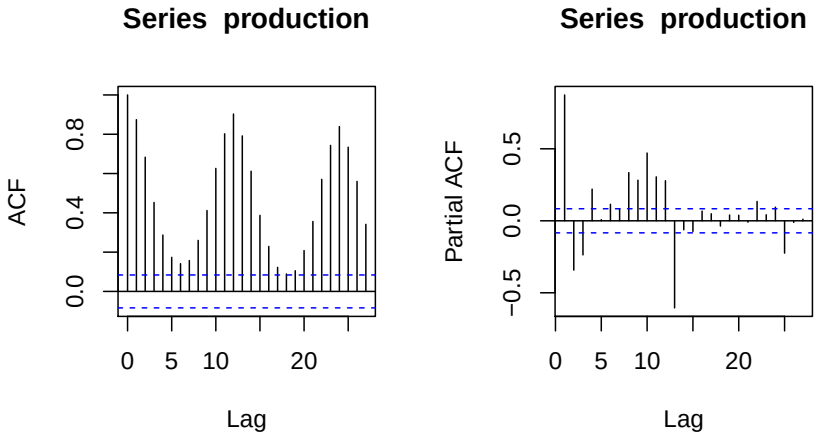



图 4: ACF and PACF of US Candy Production

表 2: Stationary of Prodiff

| Method | p-value |
|---|---|
| Adf | 0.01 |
| pp | 0.01 |
| kpss | 0.1 |

### 2.1.1 建立模型:AR(2)

利用 auto.arima 对 12 步差分后的数据进行自动定阶，得到 ARMA(2,0,0) 模型, 参数见表3，由表3可知，模型参数均显著。

表 3: AR(2) 模型参数

| | $\phi_1$ | $\phi_2$ |
|---|---|---|
| **coef** | 0.6707 | 0.1165 |
| **s.e.** | 0.0430 | 0.0430 |
| **p-value** | 8.749021e-46 | 0.003479314 |

### 2.1.2 AR(2) 模型残差检验

首先对 AR(2) 模型的残差进行了白噪声检验，结果如表6所示，残差是白噪声序列但是残差平方不是白噪声序列，说明 AR(2) 模型残差存在 ARCH 效应。然后对 AR(2) 模型残差进行了 Jarque Bera 正态性检验，得到统计量值为 32.147，p 值为 1.046e-07，说明残差不服从正态分布。

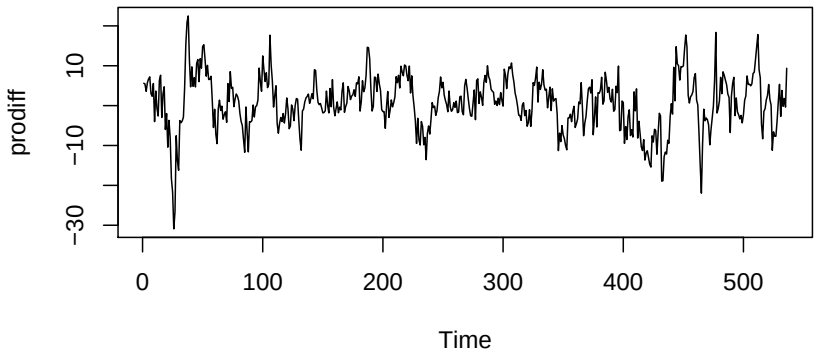



图 5: Time Series Plot of US Candy Production(after differentiating)

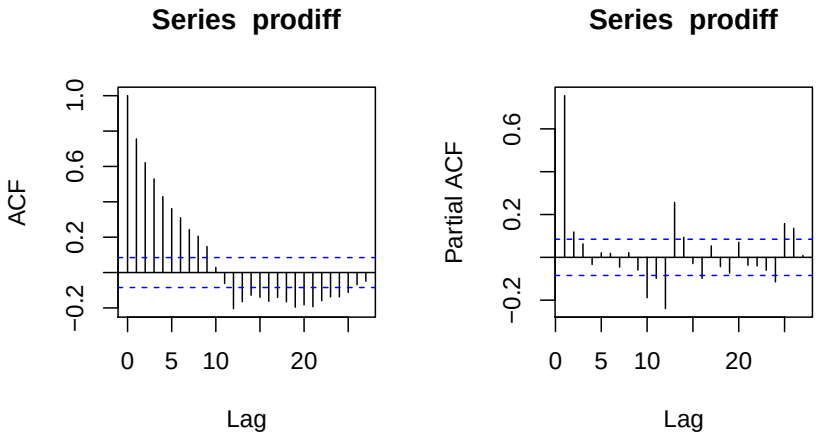



图 6: ACF and PACF of US Candy Production(after differentiating)

## 2.2 AR(2)-GARCH(1,1) 模型

### 2.2.1 AR(2)-GARCH(1,1) 模型

由上一部分的残差检验可知，AR(2) 模型残差有 ARCH 效应，下面我们尝试建立如下 GARCH(1,1) 模型：

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t$$

$$\alpha_t = \sigma_t \varepsilon_t$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1} + \beta \sigma_{t-1}^2$$

其中 $\varepsilon_t$ 服从标准 $t$ 分布，自由度为 $\nu$，$X_t$ 为 12 步差分后的时间序列。

表5为 AR(2)—GARCH(1,1) 模型参数，从中可以看出除了常数项 $\alpha_0$ 不显著外，其他参数均显著，此外 $\alpha_1 + \beta = 0.9836 < 1$。由此可得到序列的无条件波动率：

表 4: Box Test

| | **X-squared** | **p-value** |
|---|---|---|
| $residual$ | 4.215 | 0.6476 |
| $residual^2$ | 26.922 | 0.0001498 |

$$Var(X_t) = \frac{1}{1-\alpha_1-\beta} = 10.5263.$$

表 5: Estimated Parameters of AR(2)-GARCH(1,1) Model

| | **Estimate** | **Std. Error** | **t value** |
|---|---|---|---|
| $\phi_1$ | 0.6571 | 0.0437 | 15.0467*** |
| $\phi_2$ | 0.1263 | 0.0431 | 2.9298*** |
| $\alpha_0$ | 0.3115 | 0.2672 | 1.1656 |
| $\alpha_1$ | 0.0393 | 0.0180 | 2.1793** |
| $\beta$ | 0.9443 | 0.0266 | 35.4044*** |
| $\nu$ | 9.8910 | 3.9996 | 2.4730** |

### 2.2.2 AR(2)-GARCH(1,1) 模型检验

表 6: Box Test

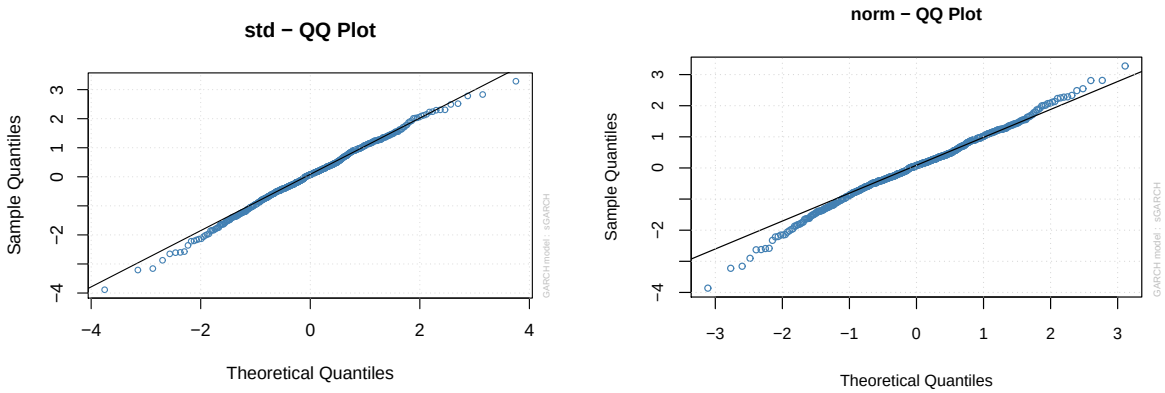
| | Lag | Stat. | p-value |
|---|---|---|---|
| | 1 | 0.07699 | 0.7814 |
| $Residual$ | 5 | 2.9528 | 0.5018 |
| | 9 | 6.7359 | 0.1555 |
| | 1 | 0.4371 | 0.5085 |
| $Residual^2$ | 5 | 2.9528 | 0.7898 |
| | 9 | 1.2962 | 0.7605 |

表6中 p 值均大于 0.05，说明残差以及残差平方序列均为白噪声序列。GARCH 模型消除了残差的 ARCH 效应，模型显著。下面将分析为什么选择 t 分布作为残差分布。

图7是分别选择正态分布以及 t 分布作为误差分布的残差 Q-Q 图, 从图中可以看出，t 分布的 Q-Q 图更接近直线。图8给出了选择 t 分布作为残差分布的模型残差直方图，而且其更接近标准 t 分布的密度曲线，所以我们假定残差近似服从 t 分布。

### 2.2.3 模型拟合以及样本内预测

在拟合差分数据后我们还原到原始时间序列数据，图9是 AR(2)-GARC(1,1) 模型对原始数据的拟合。然后我们对原始数据前 536 个数据建立 AR(2)-GARCH(1,1) 模型，预测后面 12 个数据，得到了如**??**的预测图。




(a) std (b) norm

图 7: std Q-Q plot and Density plot

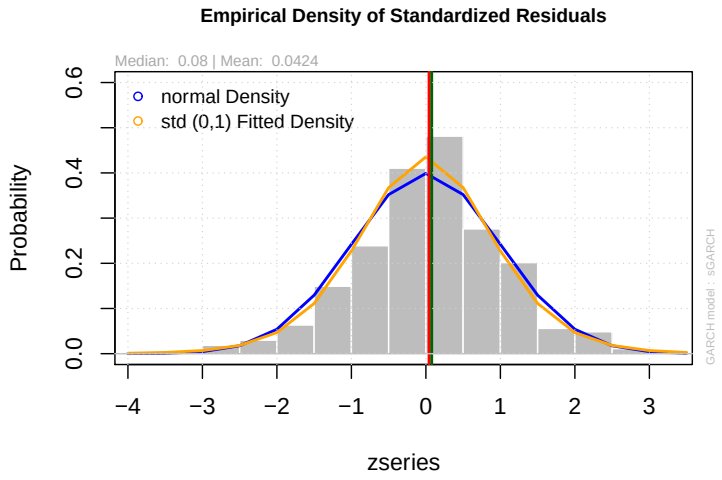


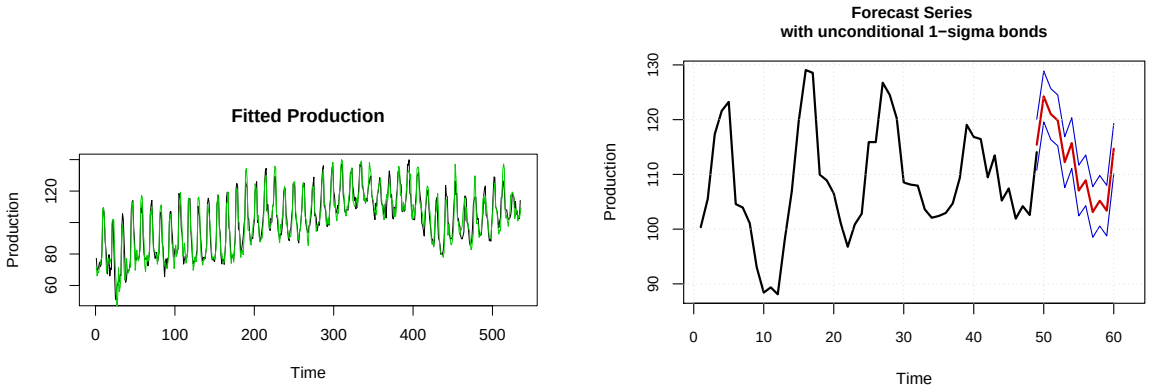

图 8: Emperical Density

# 3 非参数模型

上一部分通过 ARMA+GARCH 模型参数方法对序列建模，这一部分选择了 Holt-Winters 模型和奇异谱分解（SSA）非参数方法对原始时间序列建模。

## 3.1 Holt-Winters 模型

本文首先使用 Holt-Winters 模型建模，从时序图中可以看出序列周期性与趋势性有交互作用，所以选取了如下乘法模型：




(a) Fitted (b) Forecast

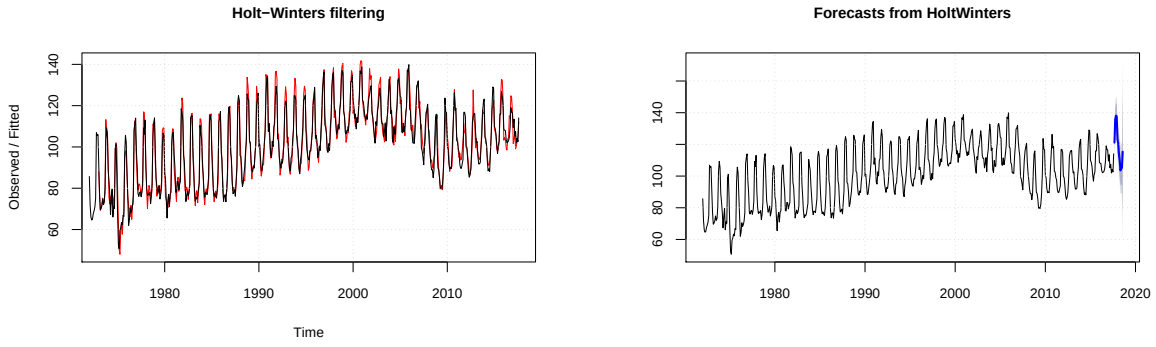
图 9: Fitted and Forecast of AR-GARCH

$$S_t = \alpha(x_t/p_{t-k}) + (1-\alpha)(S_{t-1} + T_{t-1})$$
$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$
$$p_t = \gamma(x_t/S_t) + (1-\gamma)p_{t-k}.$$

其预测方程为：

$$\hat{x}_{t+h} = (S_t + hT_t) * p_{t-k+h}.$$

Holt-Winters 模型参数估计结果为：$\alpha = 0.6209, \beta = 0.0030, \gamma = 0.8597$，其拟合结果如图10(a) 所示。然后进行了 12 步样本外预测，预测结果以及置信区间如图10(b) 所示。




(a) Fitted (b) Forecast

图 10: Fitted and Forecasted values of Holt-Winters Model

## 3.2 奇异谱分析（Singular Spectrum Analysis）

### 3.2.1 SSA 模型原理

奇异谱分析（SSA）是一种分解时间序列的非参数方法，SSA 利用矩阵奇异值分解的方法将时间序列分解为不同的主成分序列（component series），然后通过度量主成分序列之间的相关性进行分组，不同的组反映原始时间序列的趋势性以及周期性等特征。

设原始时间序列为 $x_t, t = 1, 2, ..., N$，SSA 的第一步是将原始时间序列映射到由延迟阶数的观测组成的向量。

给定每个向量的长度 $L$，$L = Window\ length, 2 \leq L \leq N/2$，则有，

$$\begin{aligned} F_1 &= (x_1, x_2, \cdots, x_L)^T \\ F_2 &= (x_2, x_3, \cdots, x_{L+1})^T \\ F_3 &= (x_3, x_4, \cdots, x_{L+2})^T \\ &\cdots \\ F_{N-L+1} &= (x_{N-L+1}, x_{N-L+2}, \cdots, x_N)^T. \end{aligned}$$

这些列向量组成了 $L$-trajectory 矩阵, $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_{N-L+1} \\ x_2 & x_3 & x_4 & \cdots & x_{N-L+2} \\ x_3 & x_4 & x_5 & \cdots & x_{N-L+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_N \end{pmatrix},$$

注意到 $\boldsymbol{X}$ 是反对角矩阵（反对角线元素相同）。

SSA 第二步是对 trajectory 矩阵 $\boldsymbol{X}$ 进行奇异值分解（Singular Value Decomposition），

$$\boldsymbol{X} = U\Sigma V^T = \sum_{i=1}^{d} \sigma_i U_i V_i^T = \sum_{i=1}^{d} X_i,$$

其中，$X_i = \sigma_i U_i V_i^T$ 为 $\boldsymbol{X}$ 的第 $i$ 个元素矩阵。

由于通过以上方法得到的矩阵 $X_j$ 并不一定是反对角矩阵，所以第三步是对得到的元素矩阵 $X_j$ 进行对角化平均（diagonal averaging）变为反对角矩阵 $\tilde{X}_j$ 从而得到序列 $\tilde{F}^j$，即对矩阵 $(x_{ij})_{L\times(N-L+1)}$，$\tilde{X}_j$ 的元素 $\tilde{x}_{m,n}$ 为：

$$\tilde{x}_{m,n} = \begin{cases} = \frac{1}{s-1}\sum_{l=1}^{s-1} x_{l,s-l} & 1 \leq s \leq L \\ = \frac{1}{L}\sum_{l=1}^{L} x_{l,s-l} & L \leq s \leq K-1 \\ = \frac{1}{K+L-s-1}\sum_{l=s-K+1}^{L}\sum_{l=s-K+1}^{L} x_{s,s-l} & K \leq s \leq K+L-2 \end{cases}$$

其中 $K = N - L + 1$。

第四步是测度得到成分序列 $\tilde{F}_j$ 之间的相关性然后进行分组。本文选用加权相关性（*weighted correlation*）来测度相关性，首先定义加权内积（*weighted inner product*）：

$$(\tilde{F}_i, \tilde{F}_j)_w = \sum_{k=1} N w_k \tilde{f}_{i,k} \tilde{f}_{j,k},$$

其中，

$$w_k = \begin{cases} = k+1 & 1 \le k \le L \\ = L & L+1 \le k \le K \\ = N-k & K+1 \le k \le N \end{cases}$$

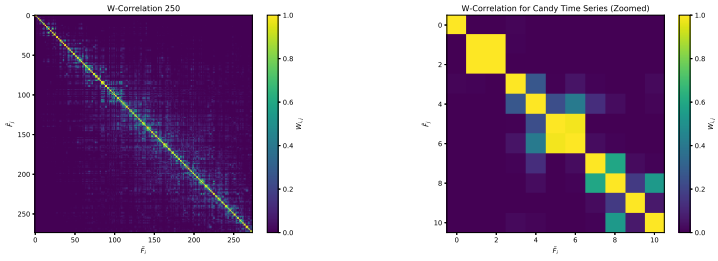
注意到，$w_k$ 反映了 $\tilde{f}_{i,k}$ 和 $\tilde{f}_{j,k}$ 在矩阵 $\tilde{X}_i$ 和 $\tilde{X}_j$ 中出现的次数。若 $(\tilde{F}_i, \tilde{F}_j)_w = 0$，$\tilde{F}_i$ 和 $\tilde{F}_j$ 是 $w-$ 正交的，则这两个序列完全可分。实际中引入加权相关矩阵 $W_{corr}$：

$$W_{i,j} = \frac{(\tilde{F}_i, \tilde{F}_j)_w}{||\tilde{F}_i||_w ||\tilde{F}_j||_w},$$

其中 $0 \le W_{i,j} \le 1$，一般当 $W_{ij} \ge 0.3$ 时，需要把 $\tilde{F}_i$ 和 $\tilde{F}_j$ 归到一组。因为少量的成分的序列就包含了原始时间序列的大部分信息，所以 SSA 的残差序列之间的相关性很高。

### 3.2.2 SSA 模型结果

然后本文对原始序列进行 SSA，窗口长度 $L = 274$。图11中显示了计算得到的各个成分序列的加权相关系数，其中 (b) 为前 10 个成分序列的加权相关系数。从图中可以看出，$\tilde{F}_1$ 与其他成分序列相关性基本为 0，所以将其单独归为一组，$\tilde{F}_2$ 与 $\tilde{F}_3$ 的相关性、$\tilde{F}_4$ 与 $\tilde{F}_5$ 的相关性和 $\tilde{F}_6$ 与 $\tilde{F}_7$ 的相关性的加权相关系数均大于 0.3，所以分别归为三组，剩余成分序列暂时归为残差。序列分组：$\tilde{F}^{(0)} = \tilde{F}_1$,$\tilde{F}^{(1)} = \tilde{F}_2 + \tilde{F}_3$,$\tilde{F}^{(2)} = \tilde{F}_4 + \tilde{F}_5$ 以及 $\tilde{F}^{(3)} = \tilde{F}_6 + \tilde{F}_7$。




(a) 250 elements　　(b) 10 elements

图 11: W-Correlation Plot of Elementary Matrices

根据以上分组，我们对原始时间序列进行分解得到图12，从图中可以看出，$\tilde{F}^{(0)}$ 以及 $\tilde{F}^{(2)}$ 体现了原始时间序列的变化趋势，而 $\tilde{F}^{(1)}$ 以及 $\tilde{F}^{(3)}$ 则体现出原始时间序列的周期性。

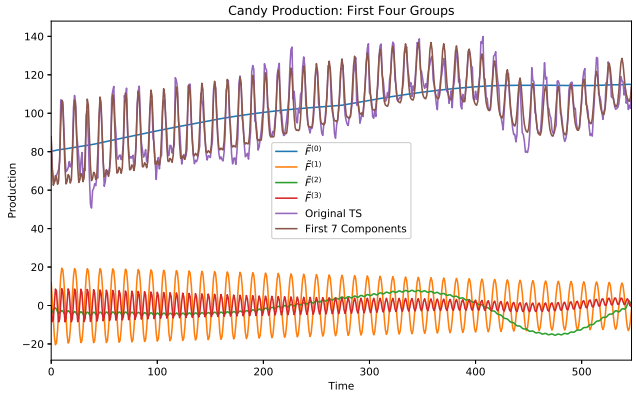
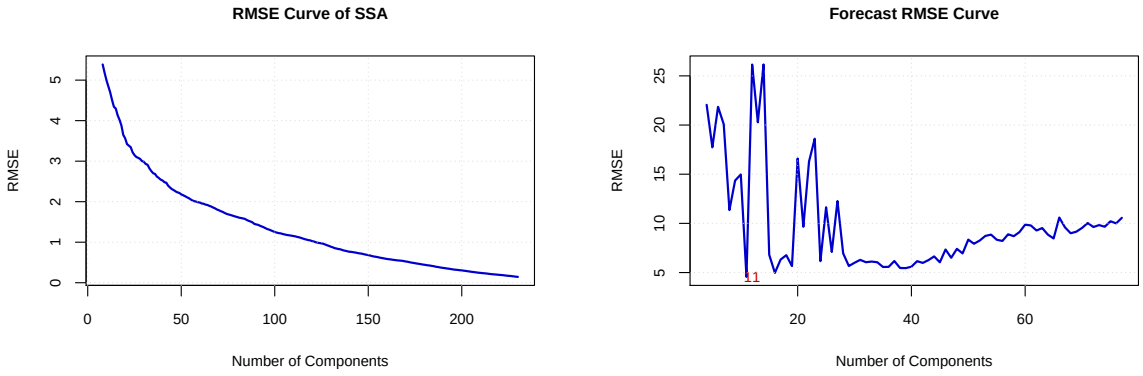



图 12: Decomposition With Fisrt 7 Components

接下来我们研究了成分序列数量对拟合均方误差根（RMSE）以及向外 12 步预测 RMSE 的影响。在作向外 12 步预测时，我对将样本的前 536 个观测建模预测后 12 个数据，然后计算 RMSE。图13可以看出随着模型使用成分序列数量的增加，拟合 RMSE 不断减小到接近 0，但是成分序列数量过大时，预测 RMSE 增大出现过拟合的现象。当模型使用前 11 个成分序列向外进行预测时，RMSE 最小，图**??**是合并前 11 个成分序列的拟合值以及残差图。




(a) Fitted RMSE (b) Forecast RMSE

图 13: Fitting and Forecast RMSE Curve of SSA

下面我们选择前 11 个成分序列进行预测并且利用 bootstrap 的方法构造置信区间并与 Holt-Winters 模型的预测值以及置信区间进行比较。如图14所示，Holt-Winters 的预测值整体高于 SSA 的预测值，而且置信区间宽度大于 SSA。

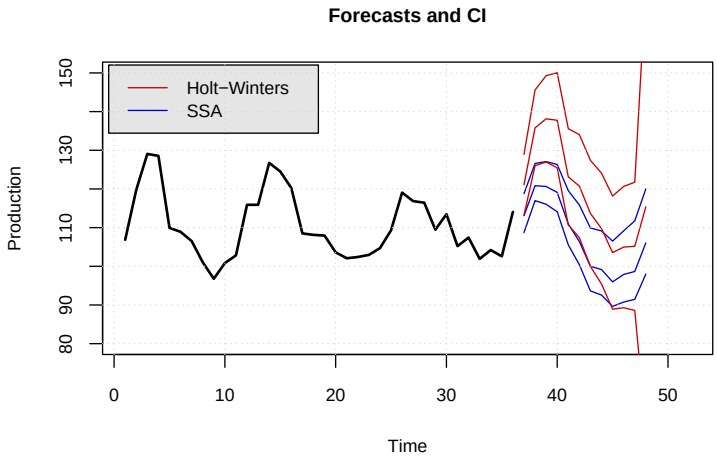



图 14: Forecasts and Confidence Intervals

# 4 模型表现比较

## 4.1 样本内拟合以及预测表现

本文计算了三个模型在全部样本内的拟合 RMSE，然后分别在样本的前 536 个数据内建模然后预测后面 12 个数据并计算 RMSE，结果如表**??**所示，从表中可以看出，Holt-Winters 模型的拟合 RMSE 最小，AR-GARCH 的预测 RMSE 最小。

表 7: In-Sample Fitting and Forecast RMSEs of Models

| **Model** | **AR+GARCH** | **Holt-Winters** | **SSA** |
|---|---|---|---|
| **Fitting** | 4.3940 | 4.2492 | 4.8492 |
| **Forecast** | 4.4631 | 7.6501 | 4.5571 |

图15展示了 3 个模型的样本内 12 步预测值与实际值的比较，可以看出 Holt-Winters 模型在实际值较高时的预测远大于其他两个模型，说明 Holt-Winters 模型对于峰值较为敏感。

## 4.2 Rolling-Sample 预测 RMSE 比较

Rolling-Sample 是在样本内多次构造样本然后窗口向外预测的一种检验模型表现的方法，向外一步预测实施步骤是：　　- 确定窗口长度 $L$，对 $\{x_1, x_2, ..., x_L\}$ 建模向外预测 $\hat{x}_{L+1}$；　　- 对 $\{x_2, x_3, ..., x_{L+1}\}$ 建模向外预测 $\hat{x}_{L+2}$；　　- …… 　　- 对 $\{x_{N-L}, x_{N-L+1}, ..., x_{N-1}\}$ 建模向外预测 $\hat{x}_N$。由此可以得到预测序列 $\{\hat{x}_{L+1}, \hat{x}_{L+2}, ..., \hat{x}_N\}$，从而可以计算 RMSE。向外 12 步预测方法类似。

取窗口长度为 500，并预测样本内最后 36 个数据。图16给出了使用不同数量成分序列预测的 SSA 模型的预测 RMSE 以及 AR-GARCH 模型、Holt-Winters 模型的 RMSE，从中可以看出，当 SSA 模型使用 8 个成分序列时，1 步预测和 12 步预测的 RMSE 最小且均小于另外两个模型的预测 RMSE。　　## Wilcoxon 符号秩和检验

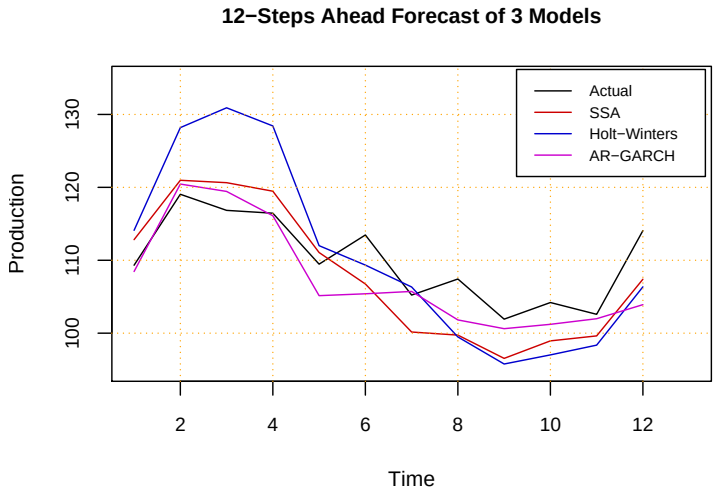
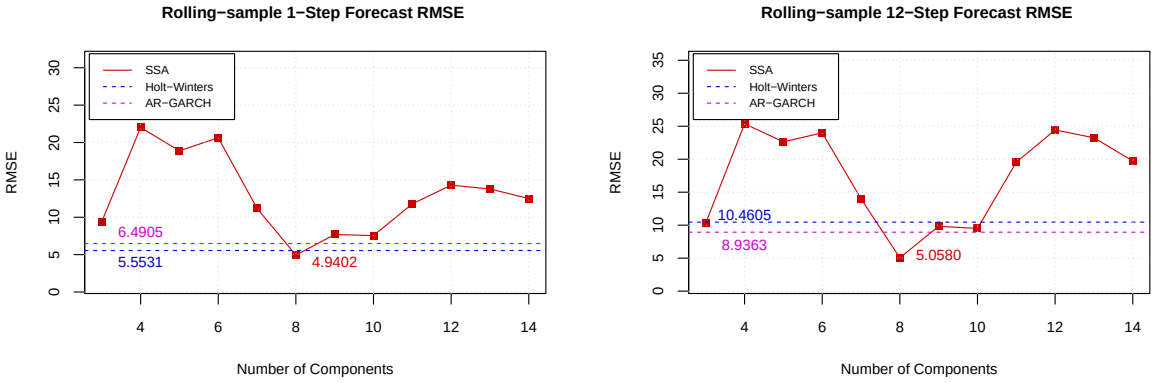



图 15: 12-Step Forecast Values




(a) 1-Step (b) 12-Step

图 16: Rolling Window Forecast RMSE

令 $s_{1i}$ 为 SSA 模型使用 8 个成分序列预测的误差平方序列，$s_{2i}$ 为 Holt-Winters 模型的预测的误差平方序列，$s_{3i}$ 是 AR-GARCH 模型的预测的误差平方序列。则 $(s_{2i}, s_{1i}), i = 1, 2, ..., 36$ 以及 $(s_{3i}, s_i), i = 1, 2, ..., 36$ 为成对数据，下面对 $D_i = (s_{2i} - s_{1i})$ and $M_i = (s_{3i} - s_{1i})$ 进行 Wilcoxon 符号秩和检验如下假设：

$$H_0 : \bar{D} = 0 \; vs \; H_1 : \bar{D} > 0$$

$$H_0 : \bar{M} = 0 \; vs \; H_1 : \bar{M} > 0$$

检验结果如表8所示，对 1 步预测，$p$ 值均大于 0.05，对于 12 步预测，$p$ 值均小于 0.05，说明 SSA 相比其他两个模型长期预测的表现更优，短期预测表现与其他两个模型的差异不显著。

表 8: P-value of Wilcoxon Signed Rank Test

| Model | AR+GARCH vs SSA | Holt-Winters vs SSA |
|---|---|---|
| **1-Step** | 0.8934 | 0.0627 |
| **12-Step** | 0.0306 | 0.001 |

# 5 结论

对美国糖果产量月度时间序列数据，本文分别采用了参数模型 AR(2)+GARCH(1,1) 以及非参数模型 Holt-Winters 和 SSA。参数模型部分，考虑到原始序列的周期性我们对 12 步差分后的序列建立 GARCH(1,1) 模型，与其他两个模型相比该模型的样本内测试集的预测 RMSE 最小。非参数模型部分，Holt-Winters 模型的样本内拟合均方误差根 RMSE 小于 AR(2)+GARCH(1,1)，但是其测试集的预测 RMSE 确是最大的；与其他两个模型在测试集内的预测值相比，Holt-Winters 模型对于峰值的预测较为敏感，即在高峰处的预测值远高于其他两个模型。

相比其他两个模型，SSA 可以通过选择不同数量的成分序列来调整模型。通过 Rolling Sample 的方法检验三个模型的预测性能，SSA 1 步向外预测与 12 步向外预测的 RMSE 均小于另外两个模型，说明 AR-GARCH(1,1) 模型预测准确度易受样本量大小影响。最后通过 Wilcoxon 符号秩和检验，12 步预测 SSA 表现显著优于其他两个模型，而 1 步预测表现差异不显著，说明相比其他两个模型，SSA 更适合中长期预测。

# 参考文献


[1] Blazsek, Szabolcs and Marco Villatoro,Is Beta-t-EGARCH(1,1) Superior to GARCH(1,1)?,Applied Economics,2015

[2] Nina Golyandina and Anton Korobeynikov,Basic Singular Spectrum Analysis and Forecasting with R,Computational Statistics and Data Analysis,2013

[3] RUEY S. TSAY,Analysis of Financial Time Series,2005