2015-2019 年曹县五里墩奶牛的牛奶产量分析

中文摘要：自 1980 年以来，曹县五里墩村书记通过奶牛养殖业产上的排泄物将大片盐碱地变得肥沃起来，即实现了产业转型也改善了土壤条件。但养殖业的未来发展规划依然是一个难题，奶牛场需要预估下一个时期的牛奶产量来确定给客户的供应量，既不能高估产量导致缺货而失信于客户，也不能过低的估计产量而导致产品浪费。因此搜集了五里墩近几年的奶牛的牛奶产量数据，进行时间序列建模，以企发现数据中的规律，从而建立具有较高精准率的预测模型，用以指导奶牛场管理人员进行未来规划。最终通过趋势分解、季节分解之后，建立起 MA(3)模型，与真实数据对比之后，发现短期预测趋势相对符合以往的规律，但长期预测依然精度不高的问题。

关键词：牛奶产量 时间序列 成分分解 短期预测

一、引言

时间序列分析是数理统计学科中应用性较强的一个分支，在金融经济、气象水文、信号处理、机械振动等众多领域中有着广泛的应用[1]。大量社会经济统计指标都是依年、季、月或日统计其指标值，随着时间的推移，形成了统计指标的时间序列。时间序列分析就是估算和研究某一时间序列在长期变动过程中所存在的统计规律性，以此预测今后的发展和变化[2]。对一个时间序列数据建模，就是建立序列的数学模型，将时间序列过程模型化[3]。

实际背景介绍

当我们走进曹县磐石办事处孙庄养殖场，就可以见到一排排牛舍整齐划一，没个牛舍有 60 多头奶牛或低头吃草，或悠闲地“闭目养神”。据了解，卖牛奶一头牛每年就可赚 3000 元，再加上牛犊、牛粪等，用不了两年就能把成本收回来。靠养殖奶牛发家致富，再利用手中的资金反哺于奶牛产业的发展，如今在曹县，像孙付江这样的奶牛养殖大户不在少数。

此外。该县在进一步深化县情认识的基础上，审时度势，立足农业资源优势，把发展以奶牛为主的畜牧业作为农业产业结构调整的优势产业，实施“规模饲养拉动，品种改良带动，专业协会促动，强化服务推动”的“四动”战略，因地制宜、突出重点，使奶牛业呈现快速健康发展的态势。目前，该县奶牛规模养殖场、小区发展到 8 处，存栏奶牛近 2 万头，年产优质有机奶 6000 万公斤，主要产品有鲜牛奶、酸奶和各种果味奶，产品不仅占领了本地市场，还销往商丘、濮阳、安阳、徐州等 10 多个市区，年奶业总产值 3 亿元。

在养殖模式上，该县主要依托山东银香伟业集团，采取“统一规划设计、统一生产标准、统一防疫程序、统一技术指导、统一收购鲜奶，分别管理”的模式，发展规模饲养场或饲养小区。目前银香伟业高产奶牛存栏 1 万多头，为社会提供就业机会 5000 多个，使五里墩村 1000 多名劳动力全部转化为农业工人，人均年收入 7000 元以上，并带动 18000 农户养牛和种草，使周边农村近 4 万名劳动力得到利用，农民直接或间接受益 8 亿多元。仅 2006 年买收玉米秸秆和全株，就直接支付农民 800 多万元。

在奶牛养殖业的带动下，全县畜牧产业实现了快速发展。目前，全县大牲畜存栏量达 24 万头，规模养殖场 160 处，肉类总产 16 万吨；以银香伟业、鲁牛乳业、雪冠食品、天一皮毛为代表的畜牧加工龙头企业发展到 68 家，实现年产值 14 亿元，带动全县 14 万农户从事畜牧养殖业，优化了农业结构，增强了农民收入。

二、与实例相关的时间序列模型

2.1 一般模型

时间序列是按时间次序排列的随机变量序列。任何时间序列经过合理的函数变换都可以被认为是由三部分叠加而成，这三部分是趋势项部分、季节项部分和随机项部分。

时间序列分析的主要任务就是对时间序列的观测样本建立尽可能合适的统计模型。合理的模型会对所关心的时间序列的预测、控制和诊断提供帮助。大量时间序列的观测样本都表现出趋势性、季节性和随机性，或者只表现出三者中的其二或其一。

这样每个时间序列，或经过适当的函数变换的时间序列，都可以分解成三个部分的叠加：$X_t = T_t + S_t + R_t$，$t = 1,2,\cdots$，$\{T_t\}$是趋势项，$\{S_t\}$是季节项，$\{R_t\}$是随机项。时间序列分析的首要任务是通过对样本的观察分析，把时间序列的趋势项、季节项和随机项分解出来。这项工作被称为时间序列的分解[1]。

2.2 趋势项、季节项和随机项分解

（1）趋势项的分解

趋势项的分解主要是根据时间序列自身发展变化的基本规律和特点，选取适当的趋势模型进行分析和预测。

趋势模型的一般形式是$\hat{Y}_t = f(t)$，$t = 1,2,\cdots$，即为一个实函数，它和季节项一样并不是时间序列内容的核心，因为它们都可以用非随机的函数进行刻画。

趋势模型的具体形式多种多样，常用的模型有：

1．直线模型 $\hat{Y}_t = a + bt$

2．指数曲线 $\hat{Y}_t = ab^t$ 或 $\hat{Y}_t = ae^{bt}$

3．幂函数曲线 $\hat{Y}_t = at^b$

4．对数曲线 $\hat{Y}_t = a + b\ln(t)$

5．多项式 $\hat{Y}_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_k t^k$

6．双曲线 $\hat{Y}_t = L + b/t$

7．*Logistic* 曲线 $\hat{Y}_t = L/(1 + ae^{-bt})$、

此外，根据数据的特点上述模型之间可能产生进一步的组合，比如指数曲线 $\hat{Y}_t = ab^{f(t)}$ 和 *Logistic* 曲线 $\hat{Y}_t = L/(1 + ae^{-bf(t)})$ 中都有 $f(t) = b_0 + b_1 t + b_2 t^2 + \cdots + b_k t^k$。

模型的选择是定性分析和定量分析相结合的分析过程，更常用的方法是绘制曲线图，直观的判断现象大体符合哪种模型。有时数据中不仅包含趋势，还存在周期波动和较强的随机变动，造成趋势识别的困难，这时需要对数据进行预处理，其方法主要包括数据的平滑和周期调整（如季节调整），以便更好的判别数据的趋势[3]。

（2）季节项的分解

在一年之内，由于季节的变动，会使某些社会经济现象（一定的时间序列）产生规律性的变化，这种规律性变化通常称之为季节变动[7]。季节项反映的是具有季节变动规律的部分。引起季节变动的首要因素是四季更迭，随着冬去夏来年复一年发生的周期变动。循环变动同季节变动类似，与季节变动不同的是循环变动是指周期为数年的变动，通常指经济周期。季节模型通常需要利用连续 3—5 年的月度数据或季节数据。

当时间序列数据不包含循环变动适宜采用乘法模型，季节因子亦称为季节指数。某期的实际季节指数（ SI ）=该期的实际值（ Y ）/该期趋势值（ T ）

季节变动的一般规律，可以由同月的实际季节指数的平均数描述，即：季节指数 i_m =同月（季）实际季节指数的合计/计算年数，其中 m 是月份（或季），

$m=1,2,\cdots,12$（或$m=1,2,3,4$）。

序列中存在季节波动常会妨碍对某些问题的认识，尤其在确定序列的趋势类型时，通常需要事先排除季节因素的干扰，然后再判断序列的趋势[8]。

（3）随机项

分离出趋势项和季节项后的时间序列往往表现出某种平稳波动性。如果随机项是平稳序列，那么可以利用 *ARMA* 模型进行模拟。

ARMA 模型是一类常用的随机时序模型，由博克斯(*Box*)、詹金斯(*Jenkins*)创立，亦称$B-J$方法。它是一种精度较高的时序短期预测方法，其基本思想是：某些时间序列是依赖于时间t的一族随机变量，构成该时序的单个序列值虽然具有不确定性，但整个时序的变化却有一定的规律性，可以用相应的数学模型近似描述。通过对该数学模型的分析研究，能够更本质地认识时间序列的结构与特征，达到最小方差意义下的最优预测[3]。

2.3 *ARMA* 模型

（1）平稳时间序列及白噪声

1．平稳序列的定义和统计特性

如果时间序列$\{X_t\}=\{X_t:t\in N\}$满足（1）对任何$t\in N$，$EX_t^2\prec\infty$；（2）对任何$t\in N$，$EX_t=\mu$；（3）对任何$t,s\in N$，$E[(X_t-\mu)(X_s-\mu)]=\gamma_{t-s}$就称$\{X_t\}$是平稳时间序列，简称平稳序列。称实数列$\{\gamma_t\}$为$\{X_t\}$自协方差函数。

设平稳序列$\{X_t\}$的标准化序列是$\{Y_t\}$，$\{Y_t\}$的自协方差函数$\rho_k=\gamma_k/\gamma_0$称为平稳序列$\{X_t\}$的自相关系数[1]。

应用$B-J$方法研究时间序列，最重要的工具是自相关和偏自相关。自相关系数ρ_k表示时间序列中相隔k期的观测值之间的相关程度，而滞后期为 K 的偏自相关函数值是指去掉Y_{t+1}，Y_{t+2}，Y_{t+3}，……Y_{t+k-2}，Y_{t+k-1}的影响之后，反映观测值Y_t和Y_{t+k}之间相关关系的数值[9].

2．白噪声

白噪声是最简单的平稳序列，它在时间序列分析中有特殊的重要地位。

设 $\{\varepsilon_t\}$ 是一个平稳序列，如果对任何 $s,t\in N$，$E\varepsilon_t=\mu$，$\mathrm{cov}(\varepsilon_t,\varepsilon_s)=\begin{cases}\sigma^2,t=s\\0,t\neq s\end{cases}$，就称 $\{\varepsilon_t\}$ 是一个白噪声，记做 $WN(\mu,\sigma^2)$。

当 $\mu=0$ 时，称 $\{\varepsilon_t\}$ 为零均值白噪声，记为 $WN(0,\sigma^2)$；

当 $\mu=0,\sigma^2=1$ 时，称 $\{\varepsilon_t\}$ 为标准白噪声，记为 $WN(0,1)$。

（2）自回归滑动平均 $ARMA(p,q)$ 模型

设 $\{\varepsilon_t\}$ 是 $WN(0,\sigma^2)$，实系数多项式 $A(z)$ 和 $B(z)$ 没有公共根，满足 $b_0=1,a_pb_q\neq0$ 和 $A(z)=1-\sum_{j=1}^{p}a_j z^j\neq0,|z|\leq1$，$B(z)=\sum_{j=0}^{q}b_j z^j\neq0,|z|\prec1$

我们称差分方程 $X_t=\sum_{j=1}^{p}a_jX_{t-j}+\sum_{j=0}^{q}b_j\varepsilon_{t-j},t\in Z$ 是一个自回归滑动平均模型。简称 $ARMA(p,q)$ 模型，称满足上述差分方程的平稳序列 $\{X_t\}$ 为平稳解或 $ARMA(p,q)$ 序列。

利用推移算子可以将上式改写成 $A(\mathrm{B})X_t=B(\mathrm{B})\varepsilon_t$，$t\in Z$。

（3）求和 $ARIMA(p,d,q)$ 模型

对于给定的一组观测值 $\{X_t,t=1,2,\cdots n\}$，如果数据满足（1）与平稳性无显著差异；（2）有迅速下降的自相关函数，则可寻求一合适的 $ARMA$ 模型来表示零均值化的数据。否则，首先要对数据进行预处理，使经过处理后的序列具有上面性质（1）和（2），这可由差分得到[11]。

求和 $ARIMA(p,d,q)$ 模型：设 d 是一个正整数，如果 $Y_t = (1-\mathrm{B})^d X_t = \sum_{k=0}^{d} C_d^k (-1)^k X_{t-k}$，$t \in Z$ 是一个 $ARMA(p,q)$ 序列，就称 $\{X_t\}$ 是一个求和 $ARIMA(p,d,q)$ 序列，其中 C_n^k 是二项式系数。于是 $ARIMA(p,d,q)$ 序列满足的模型是 $A(\mathrm{B})(1-\mathrm{B})^d X_t = B(\mathrm{B})\varepsilon_t$，$t \in Z$，其中实系数多项式 $A(z)$、$B(z)$ 同上述 $ARMA(p,q)$ 序列中的 $A(z)$、$B(z)$ 满足的条件相同。

采用求和 $ARIMA(p,d,q)$ 模型拟合数据的过程，实质上是先对观测数据进行 d 次差分处理，然后再拟合 $ARMA(p,q)$ 模型。

鉴于 G.E.P.BOX 和 G.M.Jenkins 二人对时间序列数据分析的巨大贡献，因此 ARIMA 模型命名为 Box-Jenkins 模型[12]。

（4） $ARMA(p,q)$ 模型的参数估计

$ARMA(p,q)$ 模型的参数估计分为两部分，一部分是阶数 p、q 的估计，一部分是参数 a_j、b_j 的估计。

$ARMA(p,q)$ 模型定阶方法一般有相关分析方法、$F-$检验准则方法、模型的最小最终预报误差（$Final\quad \mathrm{Pr}ediction\quad Erro$）方法和模型的信息准则 AIC 定阶方法[10]。

在实际应用中，假定已有 $ARMA(p,q)$ 模型的阶数 (p,q) 的一个估计 $(k,j)=(\hat{p},\hat{q})$，无论这个估计是怎样得到的，按照参数的最小二乘估计方法可以估计出 $ARMA(k,j)$ 模型的参数。用 $\hat{\sigma}^2 = \hat{\sigma}^2(k,j)$ 表示白噪声方差 σ^2 的估计。一般来讲，希望 σ^2 的取值越小越好。因为 σ^2 越小表示模型拟合的越精确。通常

较小的残差方差σ^2对应于较大的阶数k, j，这样过多的追求拟合的精度或说过分追求较小的残差方差σ^2会导致较大的$\hat{p}$和$\hat{q}$，从而导致较多的待估参数。其结果会使建立的模型关于数据过于敏感，从而降低模型的稳健性。

AIC 定阶准则就是为了克服模型的过度敏感而提出的。如果已知 p 的上界P_0和q的上界Q_0，对于每一对(k, j)，$0 \le k \le P_0$，$0 \le j \le Q_0$，计算 *AIC* 函数 $AIC(k, j) = \ln(\hat{\sigma}^2(k, j)) + \dfrac{2(k+j)}{N}$ 的最小值点 $(\hat{p}, \hat{q})$ 称为 (p, q) 的 *AIC* 定阶。如果最小值不唯一，应先取$k + j$最小的，然后取 j 最小的[1]。

在$ARMA(p, q)$模型的阶数 p 、 q 估计出来之后，就开始对模型参数进行估计。参数估计方法主要有矩估计方法、最小二乘估计法和极大似然估计法。一般情况下，最小二乘估计较为常用。下面着重介绍$ARMA(p, q)$模型参数的最小二乘估计方法。

首先为数据建立 *AR* 模型，取自回归阶数的上界$P_0 = [N]$，采用 *AIC* 定阶方法得到 *AR* 模型的阶数估计 $\hat{p}$ 和自回归系数的估计$(\hat{a}_1, \hat{a}_2 \cdots \hat{a}_p)$。计算残差 $\hat{\varepsilon}_t = x_t - \sum_{j=1}^{\hat{p}} \hat{a}_j x_{t-j}$，$t = \hat{p}+1, \hat{p}+2, \cdots N$。然后写出近似的$ARMA(p, q)$模型：

$x_t = \sum_{j=1}^{p} a_j x_{t-j} + \hat{\varepsilon}_t + \sum_{j=1}^{q} b_j \hat{\varepsilon}_{t-j}$，$t = L+1, L+2, \cdots, N$，这里$L = \max(\hat{p}, p, q)$，$a_j$，$b_k$是待定系数。然后对目标函数$Q(\vec{a}, \vec{b}) = \sum_{t=L+1}^{N} (x_t - \sum_{j=1}^{p} a_j x_{t-j} - \sum_{j=1}^{q} b_j \hat{\varepsilon}_{t-j})^2$极小化，得到最小二乘估计$(\hat{a}_1, \cdots \hat{a}_p, \hat{b}_1, \cdots \hat{b}_q)$。

定义 $X=\begin{bmatrix} x_{L+1} \\ x_{L+2} \\ \vdots \\ x_N \end{bmatrix}$，$\overline{X}=\begin{bmatrix} x_L & x_{L-1} & \cdots & x_{L-p+1} \\ x_{L+1} & x_L & \cdots & x_{L-p+2} \\ \vdots & \vdots & & \vdots \\ x_{N-1} & x_{N-2} & \cdots & x_{N-p} \end{bmatrix}$，

$\vec{\varepsilon}=\begin{bmatrix} \hat{\varepsilon}_L & \hat{\varepsilon}_{L-1} & \cdots & \hat{\varepsilon}_{L-q+1} \\ \hat{\varepsilon}_{L+1} & \hat{\varepsilon}_L & \cdots & \hat{\varepsilon}_{L-q+2} \\ \vdots & \vdots & & \vdots \\ \hat{\varepsilon}_{N-1} & \hat{\varepsilon}_{N-2} & \cdots & \hat{\varepsilon}_{N-q} \end{bmatrix}$，$\vec{\beta}=\begin{bmatrix} \vec{a} \\ \vec{b} \end{bmatrix}$，这样目标函数可以写成 $Q(\vec{a},\vec{b})=\left|X-\overline{X}\vec{a}-\vec{\varepsilon}\vec{b}\right|^2=\left|X-(\overline{X},\vec{\varepsilon})\vec{\beta}\right|^2$，于是最小二乘估计由方程组 $(\overline{X},\vec{\varepsilon})^T\left[X-(\overline{X},\vec{\varepsilon})\vec{\beta}\right]=0$ 决定。在 $(\overline{X},\varepsilon)$ 和 $(\overline{X},\varepsilon)^T$ 满秩的情况下，可以解出最小二乘估计 $\begin{bmatrix} \vec{a} \\ \vec{b} \end{bmatrix}=\left\{(\overline{X},\vec{\varepsilon})^T(\overline{X},\vec{\varepsilon})\right\}^{-1}(\overline{X},\vec{\varepsilon})^T X=\begin{bmatrix} \overline{X}^T\overline{X} & \overline{X}^T\vec{\varepsilon} \\ \vec{\varepsilon}^T\overline{X} & \vec{\varepsilon}^T\vec{\varepsilon} \end{bmatrix}^{-1}\begin{bmatrix} \overline{X}^T X \\ \varepsilon^T X \end{bmatrix}$[1]。

（5） $ARMA(p,q)$ 模型的检验

在得到 $ARMA(p,q)$ 模型的阶的估计 $(\hat{p},\hat{q})$ 和参数估计 $(\hat{a}_1,\hat{a}_2\cdots\hat{a}_p)$ 和 $(\hat{b}_1,\hat{b}_2\cdots\hat{b}_q)$ 后，对模型进行检验是十分必要的。

首先要检验模型的平稳性和合理性。即要检验估计的参数满足 $A(z)B(z)\neq 0,|z|\leq 1$。*Eviews* 软件给出的是滞后多项式 $A(z^{-1})=0$ 和 $B(z^{-1})=0$ 的倒数根，只有这些值在单位圆内时，过程才是平稳的[8]。

其次要检验模型的适应性。模型适应性检验是指一个 *ARMA* 模型已经基本上解释了系统的动态性，从而模型中的残差序列应是独立的[13]。即要对残差序列进行白噪声检验。因为如果残差序列不是白噪声序列，意味着残差序列还存在有用的信息没被提取，模型需进一步改进。对取定初值 $x_0=x_{-1}=\cdots=x_{-p+1}=\hat{\varepsilon}_0=\cdots=\hat{\varepsilon}_{-q+1}=0$，递推计算模型的残差 $\hat{\varepsilon}_t=x_t-\sum_{l=1}^{p}\hat{a}_l x_{t-l}+\sum_{j=1}^{q}\hat{b}_j\hat{\varepsilon}_{t-j},t=1,2,\cdots.$

取 $m = O(N^{\frac{1}{3}})$ 和 $m \succ \max(p,q)$。如果残差 $\hat{\varepsilon}_t, t = m, m+1, \cdots N$ 可以通过白噪声检验，就认为模型合适，否则寻找其他的模型。

在实际工作中，参数 (p,q) 是未知的，但是根据数据的性质有时可以知道阶数的大约范围，可以在这个范围内对每一对 (p,q) 建立 $ARMA(p,q)$ 模型，如果一个模型可以通过检验，就把这个模型留做备用。如果不能确定阶数的范围，可以采用从 $p+q=1, p+q=2, \cdots$，开始由低阶到高阶的依次搜寻的方法，然后在所有备用的模型中选出 $p+q$ 最小的一个模型。如果 $p+q$ 不能唯一决定 (p,q)，可以取 p 较大的一个[14]。

（6）预测

时间序列预测方法的基本思想是：预测一个现象的未来变化时，用该现象的过去行为来预测未来。即通过时间序列的历史数据揭示现象随时间变化的规律，将这种规律延伸到未来，从而对该现象的未来作出预测[15]。

设 $\{\varepsilon_t\}$ 是 $WN(0,\sigma^2)$，实系数多项式 $A(z) = 1 - \sum_{j=1}^{p} a_j z^j$ 满足最小相位条件，$B(z) = \sum_{j=0}^{q} b_j z^j$ 在单位圆内无零点。对于满足 $ARMA(p,q)$ 模型 $A(\mathrm{B})X_t = B(\mathrm{B})\varepsilon_t, t \in Z$，的 $ARMA$ 序列 $\{X_t\}$，定义 $m = \max(p,q)$ 和 $Y_t = \begin{cases} X_t / \sigma, t = 1,2,\cdots,m, \\ A(\mathrm{B})X_t / \sigma, t = m+1, \cdots, \end{cases}$，则 $\{Y_t\}$ 由 $ARMA(p,q)$ 模型的参数 $\vec{\beta} = (a_1, a_2, \cdots a_p, b_1, b_2 \cdots b_q)^T$ 和标准白噪声 $\{\varepsilon_t / \sigma\}$ 决定，从而不依赖于 σ [16]。再令 $W_1 = Y_1$，$W_t = Y_t - L(Y_t | \vec{Y}_{t-1})$ 是 $\{Y_t\}$ 的样本新息，$Z_1 = X_1$，$Z_t = X_t - L(X_t | \vec{X}_{t-1})$ 为 $\{X_t\}$ 的逐步预测误差。经推倒可得出 Y_{n+1} 由

$W_1, W_2 \cdots W_n$ 线形表示，令其系数为 θ，即 $\hat{Y}_{n+1} = L(Y_{n+1} | \vec{Y}_n) = \sum_{j=1}^{n} \theta_{n,j} W_{n+1-j}$ 。

经过推导最终得出 $X_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{n,j} Z_{n+1-j}, 1 \le n \prec m, \\ \sum_{j=1}^{p} a_{j_j} X_{n+1-j} + \sum_{j=1}^{q} \theta_{n,j} Z_{n+1-j}, n \ge m \end{cases}$

三、实例分析

3.1 数据的采集

3.1.1 问题研究背景

1986 年，菏泽市曹县五里墩还是一个只有大片盐碱地的贫困村，没钱改善土壤条件，土壤条件得不到改善，五里墩的经济发展就更受限制。后来，村支书王银香带领村民进行产业转型，实现了土壤改良带动区域经济的发展。她带领去难题村民组建畜牧养殖企业，实现农牧有机结合，农产品来养殖，牲畜排泄物来肥沃土壤，畜牧产品来发展经济。而如今，五里墩的奶牛养殖也已经实现了全自动化管理。

曹县五里墩可以说是奶牛之乡，中国第一头克隆牛便诞生与此，牛奶是五里墩人民的重要产业，他的兴衰关系着数千上万人的生计，也影响这曹县地区的牛奶供应链。据了解，曹县的大多数社区、小学、初中、高中都与五里墩奶牛场签订了长期的供应协议，我在高中时也是饮用此种牛奶，奶质较纯，且由于不需要长途运输而使得牛奶新鲜，价格低廉。因此了解曹县五里墩奶牛场的未来发展计划是极其重要的，不仅关乎人民的生活，更能推动五里墩奶制品行业的快速发展。

3.1.2 数据集的描述

我们搜集了五里墩“银香伟业”一万多头奶牛近五年的月度平均牛奶产量的数据，共 **60** 项，数据以 **kg** 为单位，它反映了单头奶牛的生产能力。

3.2 模型的建立

以下将通过两种方法：加法模型以及乘法模型进行数据的建模。在建模之前通过作图先来观察一下原始数据的分布规律并从实际意义解释数据的分布规律。

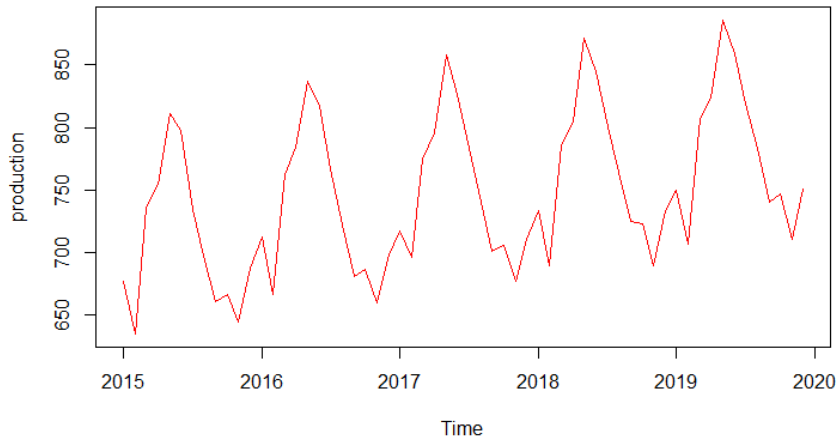

Fig1：数据的原始分布，纵轴为平均每头奶牛的月度牛奶产量

很显然数据有很强的季节性，我试着从实际的角度区分析产生这种规律的原因。

据了解，奶牛的产量主要受到以下几种因素影响[1]：①品种；②个体特征，即使是同一品种的奶牛，其体貌特征（体重，乳用体型，采食性，反应灵敏性等）的不同也会造成产奶量的差异；③生理因素，主要指年龄与胎次，一头奶牛在其第四胎（6~9 岁）时，产奶量会达到最高峰。本例中所有奶牛属于同一品种，且所取的数据为所有奶牛的平均产奶量，所以①②③这三个因素不在考虑之列。我认为产生如图所示的规律的主要原因是受到外部因素的因素：④饲养管理因素，充足的运动，饮水多次份饮等；⑤环境因素，有研究表明温度、湿度、光照、风速、气体等这几个因素对奶牛的产奶量影响最大。

通过对因素④⑤的分析，美国密苏里大学研究表明，温度为 10℃时产奶量为 100%，温度上升到 21.1，26.7，29.4 和 38.9℃时，产奶量分别下降到 89.3%，75.2%，69.6%，和 26.9%。另有研究表明，当温度低于-4℃时，奶牛产奶量会下将，并且会影响牛奶的成分组成。Sharam 报道，荷兰牛在温度为 24.7℃，太阳辐射高于 301 辐射单位或低于 301 辐射单位，增加相对湿度的条件下，产奶量增加。

由以上信息可以得知，春夏季时奶牛的高产时期，产量分布会产生这样的规律也就解释的通了。

此外，数据具有不太显著的长期趋势，即产量是逐年上升的，这一方面是管理技术的进步，另一方面是饲养条件的改良，这是容易解释的；至于上升的区是不太显著的原因是，平均奶牛产量不像工厂作业一样，随着技术的革新，生产效率可以数倍数十倍的提升，奶牛是有生产上限的，我们的技术也只能尽可能地通过改善条件来提高这个上限，但肯定不会再短期内产生显著的影响。

3.2.1 加法模型趋势项的分解

对数据分别进行一次曲线与二次曲线拟合，最终的拟合效果如下图所示

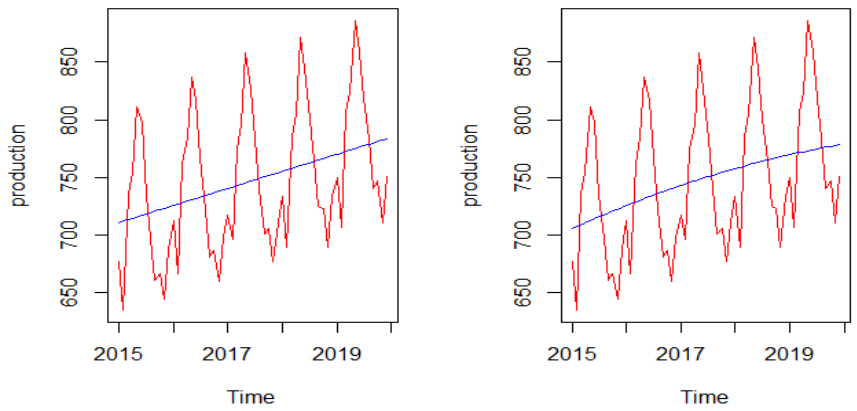

Fig2：蓝色线条为不同曲线的拟合情况（左图为一次函数拟合，右图为二次函数拟合），红色曲线为原始数据的分布

二者几乎没有任何差异，由奥卡姆剃刀理论，我们应该选择使用简单的一次曲线进行数据拟合。但无论使用哪种曲线拟合，其误差都比较大，因为数据受到显著的季节因素影响。

在剔除趋势项之后，数据只包含季节因素和随机因素，其上升趋势不会出现在数据的分布当中，重新做出数据的分布图如下

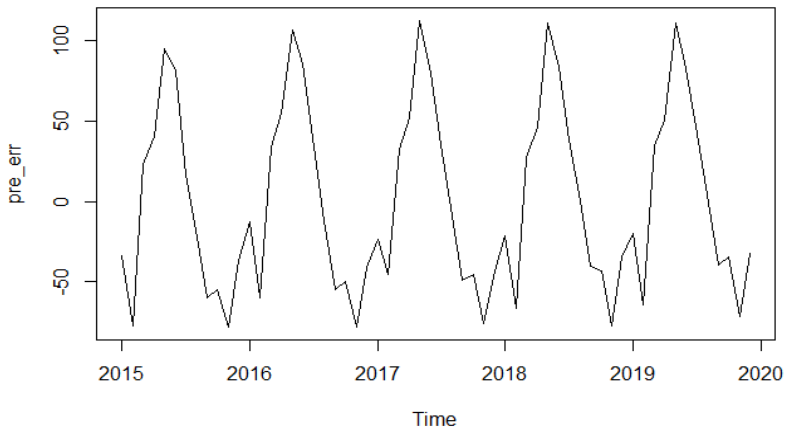

Fig3：除去长期趋势项之后的数据分布

3.2.2 乘法模型的趋势项分解

在本小结，使用乘法模型对数据进行趋势项分解，先来观察用乘法模型分解出的趋势分布是怎么样的

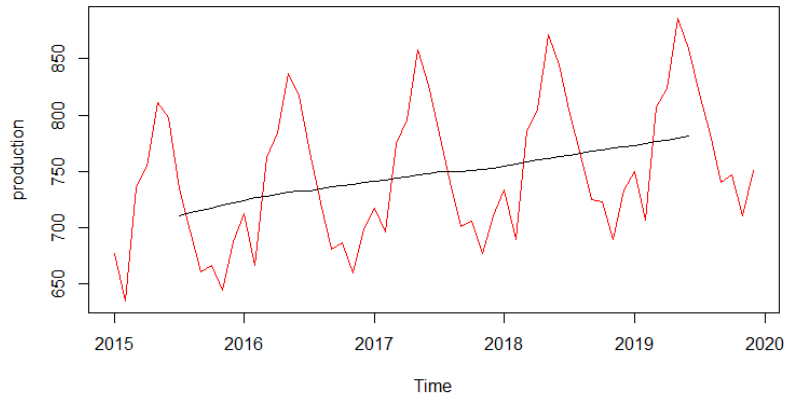

Fig4：乘法模型中趋势项的分布与原始数据的分布对比

注意到，乘法模型中趋势项是没有前 6 期和后 6 期数据的，这是由乘法模型的性质决定的。在乘法模型中我们总会损失 12 期数据（以月度为单位）或损失 4 期数据（以季度为单位），因为在计算中需要使用 12 项中心移动平均法，而前六期和后六期的数据是无法构成使用 12 项中心移动平均的条件的，因此无法计算它们的值。同样的，在乘法模型中剔除趋势项之后，剩余的数据也会缺少 12 期数据（前六期和后六期）。剔除趋势项的数据分布如下图所示

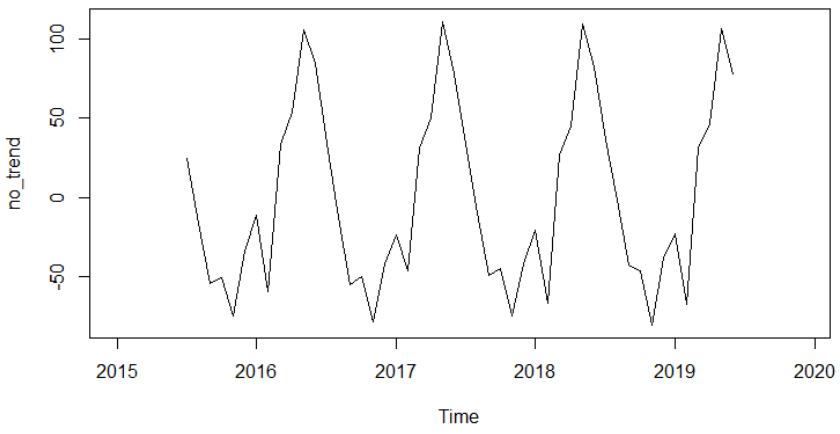

Fig5：乘法模型中剔除趋势项之后的数据分布

我们把加法模型去除趋势项的数据分布与乘法模型去除趋势项的数据分布放在一起进行对比

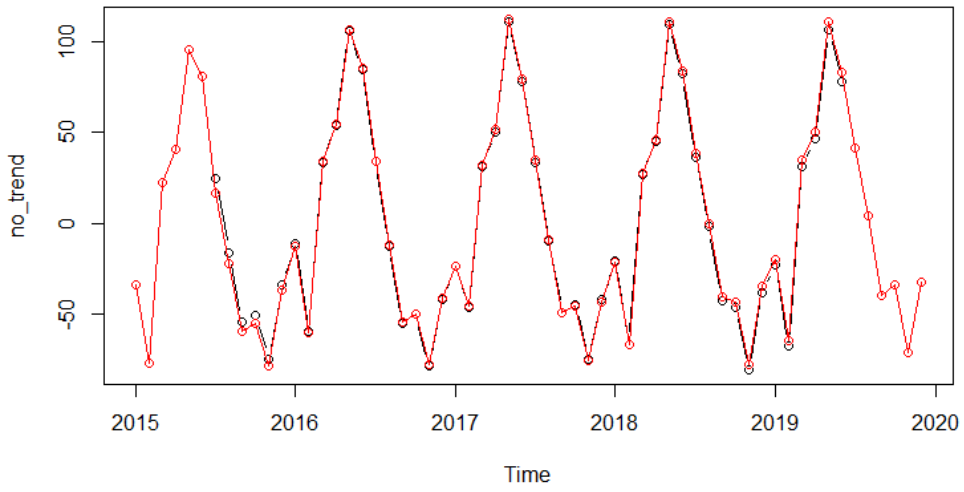

Fig6：分别用加法模型与乘法模型去除趋势项的效果对比，红色实线为加法模型，黑色虚线为乘法模型

两种方法的效果实际上几乎没有差别，也就是说，就去除趋势成分而言，我们既可以使用加法模型也可以使用乘法模型。

3.3 季节项的分解

通过加法模型进行计算，得到各个月份的季节指数如下表所示（已保留小数点后三位小数）

Month	Jan	Feb	Mar	Apr	May	Jun
Seasonal	-19.369	-59.640	30.892	49.235	108.464	80.871
Month	Jul	Aug	Sep	Oct	Nov	Dec
Seasonal	32.517	-9.744	-49.984	-47.692	-76.942	-38.609

Table1：加法模型下各个月份的季节指数

绘制季节指数的分布图如下所示

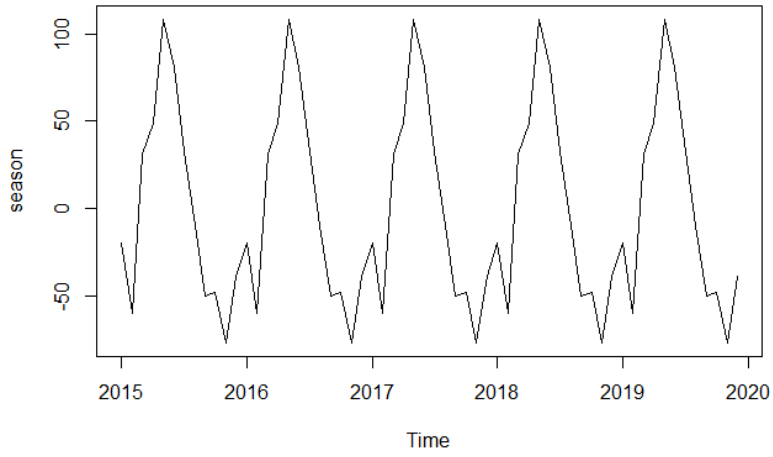

Fig7：加法模型的季节项分布

这与我们在最初分析的影响奶牛产量的因素分析中描述一致，在春夏季节，由于环境比较适宜，从而奶牛的产量要高于其他时期，也即春夏两季的季节因子为正值。并且最佳生产时节在五月份，此时温度适中，且太阳接近北回归线，光照条件充足。在曹县的五月份，空气湿度相对较为舒适，且风速较为适中，这就使得牛棚内具有较好的通风条件，从而减少了疾病的发生，这为奶牛的高产提供了保障。

同样，我们再次观察乘法模型中的季节因子分解的效果，在此模型下的各月份的季节因子的值如下表（已保留小数点后三位小数）：

Month	Jan	Feb	Mar	Apr	May	Jun
Seasonal	0.975	0.921	1.042	1.066	1.144	1.108
Month	Jul	Aug	Sep	Oct	Nov	Dec
Seasonal	1.044	0.987	0.933	0.936	0.897	0.949

Table2：乘法模型下各个月份的季节指数

绘制季节指数的分布图如下所示

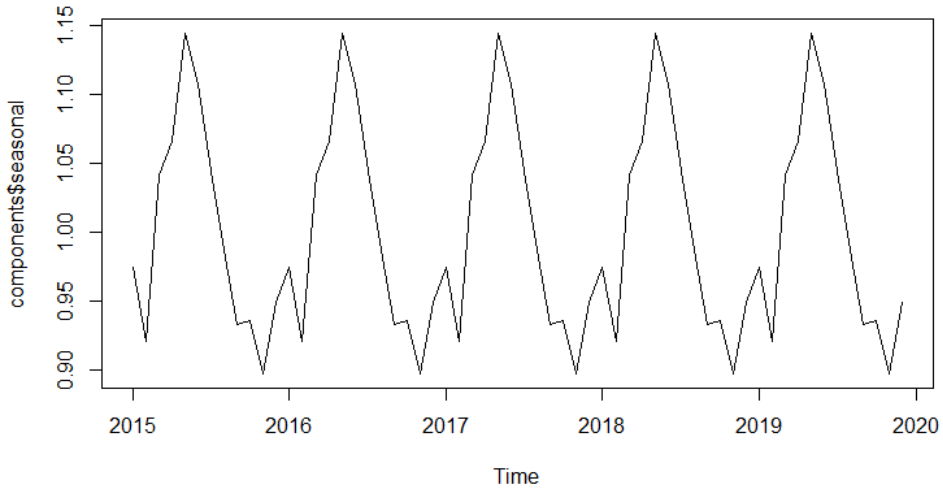

Fig8：乘法法模型的季节项分布

仅从图示来看，二者表达的规律基本一致，但实际上二者有着很大差别，加法模型所揭示的是当月份的产量值相对于平均产量水平的增加量或减少量，他是一个量值上的指标，因此也会受到数据本身水平的影响，即若数据本身的量级或差异值非常大，那此时使用加法模型得到结果就会不准确；相比之下，乘法模型所揭示的是，当月产量相对于平均值水平的增加或减少的比例，它与量纲无关，因此避免了由于数据本身的水平带来的影响。

3.4 随机项的 *ARMA* 模型参数估计

在加法模型中，分离出季节项、趋势项之后就只剩下随机项了，随机项的分布如下图所示

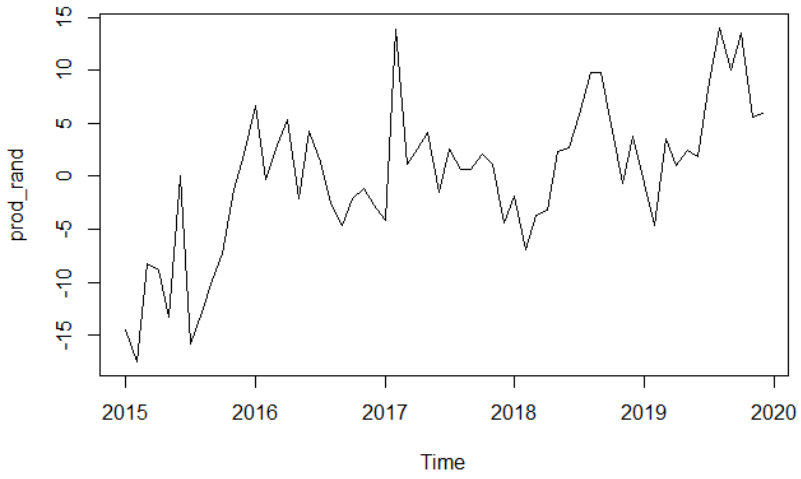

Fig9：随机项分布

直接从随机项的分布图中难以得到有用的信息，因此绘制其自相关图与偏自相关图，以此来估计 arma(p,q)模型的各项参数，图示如下

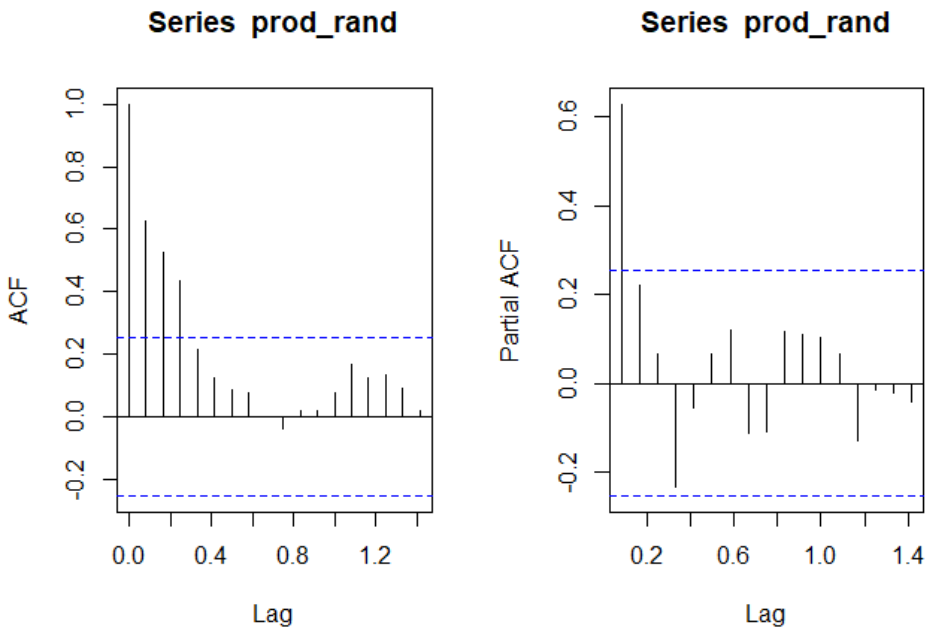

Fig10：自相关系数图（左）与偏自相关系数图（右）

由 Fig10 可知自相关在 Lag=4 处截尾，偏自相关在 Lag=1 处截尾，经检验，在建立 arma(1,3)模型时，ar 模型的参数并不显著不为 0，也就是说，我们不必考虑 ar 模型的成分。因此可依此对我们的序列建立 arma(0,3)模型，最终得到模型的各项参数如下表所示：

	ma1	ma2	ma3	intercept
Coefficients	0.8848	0.9703	0.4263	746.4515
s.e.	0.1601	0.1769	0.1628	14.9467
Z-score	7.1802	5.9771	5.0000	——
sigma^2=1293	log likelihood = -301.18		aic = 612.35	

Table3：arma(0,3)模型的各项系数，s.e.为估计的标准误差，Z-score 是其检验统计量的值，sigma^2 为模型的估计方差，aic 为在 AIC 评估准则下的值

由上表可知，若我们取显著性水平 α =0.05，此时统计量的临界值为 1.96，系数都是显著不为 0 的。

3.5 随机项的 *ARMA* 模型检验

在下一步工作之前，需要求出自相关系数与偏自相关系数的值，如下表所示：

Lag	Acf	Pacf
1	1.000000000	0.62777658
2	0.627776576	0.22106654
3	0.528046886	0.06684879
4	0.435517337	-0.23278362
5	0.214418114	-0.05397433
6	0.124324051	0.06529326
7	0.086305313	0.11952942
8	0.074946357	-0.11042330
9	0.001700081	-0.10953677
10	-0.036188253	0.11531823
11	0.019273793	0.10965893
12	0.020268032	0.10204961
13	0.078830243	0.06456177
14	0.169250286	-0.12883559
15	0.126771488	-0.01297483
16	0.131970724	-0.02100689
17	0.089320829	-0.0396349

Table4：上表记录了各期自相关系数与偏自相关系数

在建立起模型之后，我们需要进行 0 均值检验，即检验随机项是否是 0 均值的，若其不通过 0 均值检验，那么后续的预测就没有意义了。

先观察残差项的分布

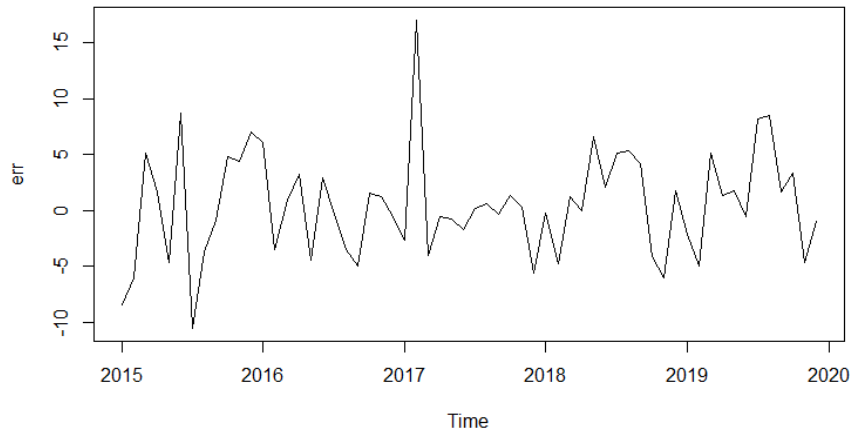

Fig11：残差项分布

下表展示了随机项的检验结果

Mean	S	95% Confident interval
-0.805304	4.163483	[-4.163483，4.163483]

Table5：随机项的均值，统计量方差以及 95%置信区间

显然均值位于 95%置信区间内，即随机项满足 0 均值假设，于是可以进行下一步的预测。

3.6 模型预测

通过已经建立起来的arima(3,1)模型进行为期6个月的预测，其预测结果如下：

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2020	739.5732	694.7699	784.3766	671.0525	808.0940
Feb 2020	757.1335	702.3675	811.8995	673.3761	840.8909
Mar 2020	754.5120	690.0009	819.0231	655.8507	853.1733
Apr 2020	758.7843	692.6171	824.9515	657.5904	859.9783
May 2020	754.5359	688.3227	820.7492	653.2715	855.8003
Jun 2020	752.0327	684.7134	819.3520	649.0767	854.9886

Table4：未来各期的点预测值，Lo80 为 80%预测区间下限，Hi80 为上限，Lo95 为 95%预测区间的下限，Hi95 为上限

将预测结果趋势绘制在往期数据分布上，其效果如下图所示

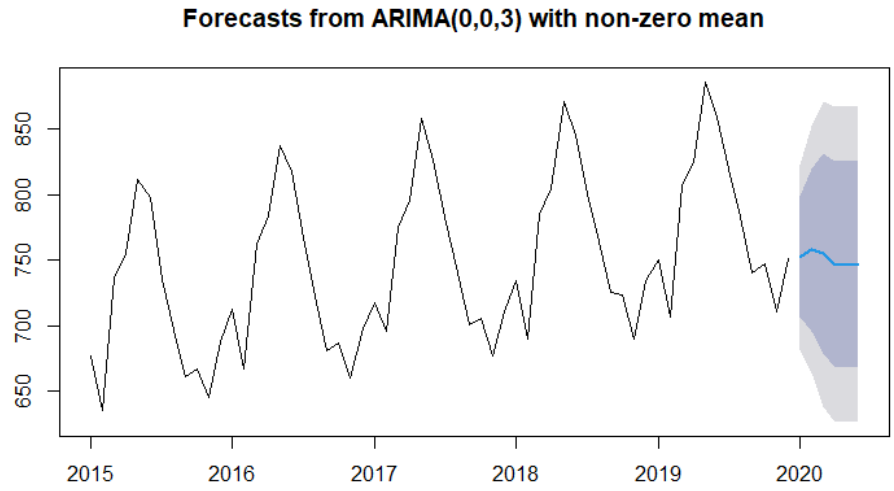

Fig12：蓝色线条为点预测值，深色部分为 80%预测区间，
浅色部分覆盖 95%的预测区间

从图上可知，预测的前两期的结果相对来说比较符合数据以往的观测规律，但此后的预测趋势已经产生较大的偏差，并且预测区间的宽度也越来越大，说明结果的精确性已经很难保证。因此，时间序列建模只适合短期预测，而不能用于长期预测。

四、总结

通过对山东省菏泽市曹县五里墩奶牛场平均每头奶牛的产量进行分析，我们最终建立起 arma(0,3)模型，也即 ma(3)，但是实际上优于模型较为简单，因此预测效果并不理想，如果要建立更为精确的预测模型，可以通过 R 语言中的 auto.arima 函数建立模型，经过测试，预测效果有实际数据的差值已经非常微小，即预测的非常精确。

对此数据的分析结果，并不能进行推广用来指导五里墩奶牛场的未来规划，还需要寻找更为精确的模型来进行分析。

参考文献

[1]王丽芳, 杨健, 支秀平. 影响奶牛生产性能的因素[C]// 2007 年乳业发展国际论坛. 中国奶业协会;中国乳制品工业协会, 2007.