蔬菜价格指数分析与预测

王 倩 倩

2017 年 06 月

分类号 O213.9
UDC 分类号 311

蔬菜价格指数分析与预测

作 者 姓 名　王倩倩
学 院 名 称　数学与统计学院
指 导 教 师　徐兴忠教授
答辩委员会主席　谢田法副教授
申 请 学 位　应用统计专业硕士学位
学 科 专 业　应用统计
学位授予单位　北京理工大学
论文答辩日期　2017 年 6 月 4 日

THE ANALYSIS AND FORECAST OF THE VEGETABLE'S PRICE INDEX

Candidate Name:	Wang Qianqian
School or Department:	School of Mathematics and Statistics
Faculty Mentor:	Prof. Xu XingZhong
Chair, Thesis Committee:	Associate Prof. Xie TianFa
Degree Applied:	Master of Philosophy
Major:	Applied Statistics
Degree by:	Beijing Institute of Technology
The Date of Defence:	June 4th，2017

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签　　名：　　日期：

摘要

近年来农产品价格的频繁波动，已成为政府、社会和百姓的关心重点，农产品价格波动不仅事关农业自身发展，也影响国家整体物价水平和经济发展。根据农产品价格数据的特征，对其进行准确的分析与预测，对农产品市场的应急机制和国家的宏观调控具有重要意义。

随着数据挖掘理论的不断发展，以及时空数据的大规模增长，结合国内外研究现状，本文选用时间序列模型和时空数据模型的统计分析方法，主要是利用 ARIMA(差分自回归滑动平均)模型和 STARMA(时空自回归滑动平均)模型，对蔬菜价格指数进行建模预测分析。文章详细研究了 ARIMA 模型和 STARMA 时空数据模型的基本思想和建模过程,同时考虑了时间相关性和空间相关性，通过蔬菜价格指数的预测验证了两种模型的预测性能，并分析讨论了它们各自的优缺点，解释了蔬菜价格指数的空间变异性，为农业领域的问题研究引入新思路。

关键字：蔬菜价格指数；空间相关性；ARIMA 模型；STARMA 模型

Abstract

In recent years, the frequent fluctuation in agricultural product prices has got more concerns by the government and public. That not only affects the development of agriculture and economy, but also causes the fluctuations of overall price level. An precise analysis and forecast of product's price based on its data feature have its own significance to the response mechanism of agricultural markets and the national micro-control.

With the development of the data-mining methods and the increase of the amount of spatio-temporal data, based on the results of the domestic and foreign researches, in this paper, we use the ARIMA (Auto-regressive Integrated Moving average) model and the STARMA (spatio-temporal Auto-Regressive and Moving Average) model to analyze and forecast the vegetable's price index. In this paper, we discussed the basic theory and the modeling process of the ARIMA model and STARMA model, considering the time correlation and spatial correlation at the same time. We used the forecast results of vegetable's price index to verify the abilities of these two models and compare their advantages and disadvantages. According to the results, we can explain the spatial variability of the vegetable's price index, and create new ideas for the agricultural researches.

Key words: vegetable's price index; spatial correlation; ARIMA model; STARMA model

目录

第一章　绪论

1.1　研究背景和意义

1.1.1　研究背景

农产品是人类赖以生存和发展的基本元素，渗透到社会的各个角落。多年来，农产品市场价格一直是各国政府和全世界重点关注的对象，农产品价格波动所带来的市场风险是关系国计民生和社会稳定的大事。农产品价格不仅事关农业自身发展，而且关系整体物价水平变动和人民的正常生活，对经济社会有序发展至关重要。以蔬菜价格为例，随着我国经济的快速发展，蔬菜价格受到越来越多市场因素的影响，造成蔬菜的市场价格变化趋势很难有规律可循，给农户生产带来很大的风险。然而，蔬菜批发价格指数能够很好地反映一个地区乃至全国的蔬菜市场的整体趋势和水平，对蔬菜价格指数的预测，有助于我们了解蔬菜批发价格变化规律，对农产品市场的宏观调控有指导意义，对促进农业的稳定发展、国家的宏观调控具有重要的理论和现实意义。

农产品价格指数数据是典型的时间序列，分析价格指数需要选择恰当的统计方法，而且由于信息化技术的发展，时间序列预测是一个重要的研究课题，时间序列作为所观测系统的输出,背后往往隐含着研究对象的某种特定规律和特性，故而可以选择时间序列分析方法来辨识和重构所观测的价格指数的行为、性质，从而对价格进行预测和调控。随着计算机、通信、GIS 等科学技术的飞速发展，对地观测技术的建设和空间数据设施的建立，日益累计的时空数据成为当下研究的重点，近年来，时空数据的挖掘分析研究备受关注，随着数据挖掘理论和算法的日渐成熟，时空数据成为数据挖掘的新热点。时空数据也可以看作是时间序列在空间上的扩展，或者是空间序列在时间上的扩展，这些时空数据在时间和空间上均呈现着复杂的相关关系，本文主要以蔬菜价格指数为例，利用时间序列为基础进行时空序列的研究。

1.1.2　研究意义

在研究内容上，论文选题瞄准了当前国家需求，既是现实迫切需要解决的突出问题，也是科学研究亟需攻克的难题。近年来蔬菜市场价格的频繁波动，已成为政府关

心、社会关注和百姓关切的热点与焦点。因此对蔬菜价格指数的准确分析和预测，对促进农业的稳定发展、完善农产品市场应急机制、国家的宏观调控具有重要的理论和现实意义。

在研究方法上，论文尝试在农业方面引入了时空预测的方法，目前在农产品价格预测方面，使用的大部分数据仅仅是时间维度上的价格数据，并没有考虑到地理位置属性的。在一般的经济统计分析、预测、预警等决策中，时间维度的波动已经被认为是事件发展和变化的关键因素，对于其空间位置的因素并没有受到过多的重视。然而，对于日渐累计的时空数据，几乎所有含有地理位置属性的数据，在空间上都会存在空间相关或空间自相关等特征。所以，根据技术的发展和空间领域的逐渐成熟，对于许多数据挖掘等统计分析，在包含地理属性的情况下应当结合一些空间上的性质开展研究，这对于传统的统计分析来说，是很有意义的完善。

1.2 国内外研究现状

在 19 世纪末期的西方主要资本主义国家爆发的经济危机，使经济波动监测和预测成为经济统计学研究的新课题。在经济预测的初期，由于技术的落后，主要以定性预测为主。随着信息技术和理论的飞速发展，进行经济预测的数学方法数不胜数，对于处理数以万计的信息数据处理和经济预测模型的实现提供了很便捷的技术支撑，进行现代化的预测分析、预测的精度和速度均有显著提高。

在农业方面，农产品信息监测预警工作也是越发引起世界各国的重视。早在 20 世纪初期，从一元回归和多元回归模型的方法在价格预测方面应用开始，更加完善准确的数学预测模型和统计分析方法层出不穷，指数平滑法对澳大利亚羊毛价格的短期预测[1]，ARIMA、ARCH、GARCH 等时间序列模型对蔬菜等各类农作物价格的准确预测[2-4]，随着信息技术的不断发展和大数据研究技术的崛起，在智能预测法方面，数据挖掘分析方法得到广泛运用且很受欢迎，小波分析法、支持向量回归技术、灰色神经网络技术、时间序列复合预测模型等广受欢迎，同时也表现出很好的预测效果[5-7]。

自 20 世纪 80 年代起，随着时间序列领域理论的不断完善，有关时间序列预测法的研究兴趣不断上升，结合各种统计分析方法、时间序列模型进行农业方面的分析和预测，各国运用时间序列对农产品的定量预测分析方法成为农产品价格预测的主流。刘峰等应用非平稳时间序列 ARIMA 模型对农产品价格进行预测分析，模型结果难以保

证突然事件引起的转折点的预测准确性，需要引入新的序列值重新拟合模型[8]；傅如南、林丕源等人也运用了非平稳时间序列 ARIMA 模型方法对肉鸡价格进行预测，得到类似的结果[9]。张浩、王勇运用马尔可夫模型对小麦期货价格进行预测，结果表明马尔科夫模型在长期预测方面起到很好的作用[10]。近几年，在农产品价格分析方面各种分析方法不断完善，许多文献资料表明，ARIMA、ARCH、GARCH 等时间序列方法建立的预测模型精确度比较高，Holt-Winters 无季节性模型稳定性最好，此外常用的还有灰色预测、回归曲线拟合等方法[11-12]。

近年来，随着计算机、通信、GIS 等科学技术的飞速发展，对于空间统计分析应用研究开始起步，时空数据的研究在我国也开始得到广泛关注，同时利用时空序列分析空间数据在 GDP 数据、空气质量、气候变化、噪声分析等领域已经得到有效的应用[13-24]。

1.3 研究方法和技术路线

1.3.1 研究方法

在本研究以计量经济学、统计学等学科为基础，采用理论分析与实证研究相结合的研究方法，兼收并蓄传统研究方法和现代方法。本文会结合统计分析方法侧重于定量分析与实证分析，根据研究目的，通过查阅现有的书籍和网络资源，查阅大量国内外时间序列、空间数据挖掘等预测预警研究方法相关文献资料，并对所获取的资料进行分类、汇总、归纳与整理，从而全面地、系统地了解掌握所要研究问题的解决方法，为本文研究问题的提出、研究框架的构建提供了有益的借鉴。同时为了确定模型的有效性，对蔬菜价格指数的研究必须建立在实证分析的基础上，客观的揭示各种时间和空间的价格波动规律。

本文主要研究了利用时间序列 ARIMA 模型进行建模预测，以及时空序列 STARMA 模型进行分析和建模的理论与方法。讨论了时间序列和时空序列的性质，在此基础上介绍常用的时间序列模型和时空序列模型，以及模型的建立。然后会结合我国各个省份地区的蔬菜价格指数进行分析，除了寻找较优的蔬菜价格指数预测模型，还会考虑空间上的因素，对农产品价格在时间空间进行相关性分析，分析地域上是否会有相互影响。运用 ARIMA、STARMA 模型对我国各地区的农产品开展短期预测，在时间序列的基础上向空间扩展，添加地理维度信息建立模型，对数据进行分析。并对这两类模型

预测的效果进行了比较，选择较为理想的预测模型。

1.3.2 技术路线

论文以全国 27 个省、市、自治区的蔬菜价格指数数据为研究对象，结合统计模型 ARIMA 模型、STARMA 模型进行预测分析，利用统计软件 R 软件对数据进行建模分析[25-28]。主要技术路线如下：

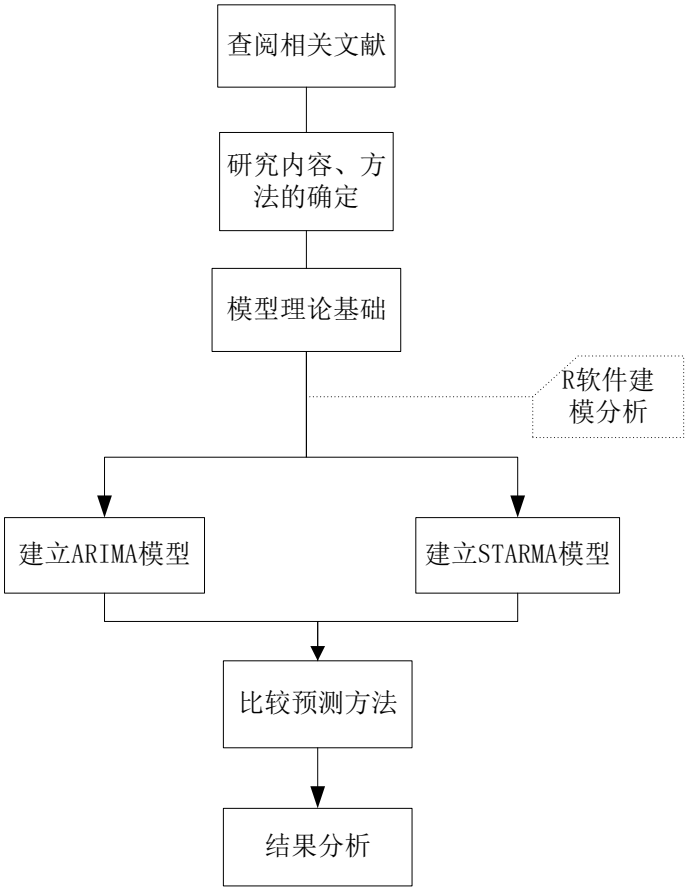

图 1.1 研究技术路线

模型建立过程：

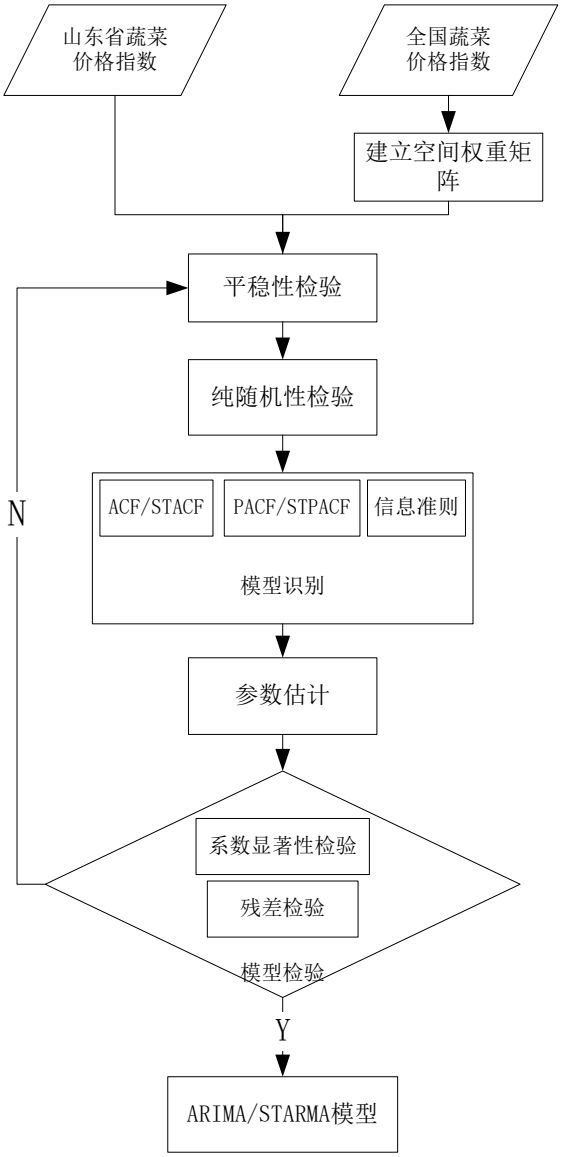

图 1.2　模型建立过程

1.3.3　论文创新点

本文总结了以往的时间序列预测模型，采用 ARIMA 模型在时间维度上进行价格指数的预测，考虑了价格指数影响因素可能会和的地域分布相关，这在以往的农业方面研究中是几乎没有的，在时间维度上向空间维度进行扩展，考虑时空序列建立 STARMA 模型进行预测，这是对农业领域的统计分析的一大补充。

第二章 理论基础和模型框架

2.1 时空序列性质

2.1.1 平稳性

当时间序列所有的统计性质不随时间的推移而发生变化时，该序列可以被认为是平稳的序列。对于平稳的时间序列$\{Y_t\}$，主要表现在该时间序列均值、方差、自协方差函数等统计量不依赖于时间变量，那么对任意时刻 t 的时间延迟 k、m 都有：

$$\mu_{Y_t} = EY_t = EY_{t+m}$$

$$\sigma_{Y_t}^2 = E\left(Y_t - \mu_{Y_t}\right)^2 = E\left(Y_{t+m} - \mu_{Y_t}\right)^2$$

$$\gamma(\mathrm{k}) = E\left(Y_t - \mu_{Y_t}\right)\left(Y_{t+k} - \mu_{Y_{t+k}}\right) = E\left(Y_{t+s} - \mu_{Y_{t+s}}\right)\left(Y_{t+s+k} - \mu_{Y_{t+s+k}}\right)$$

同理，对于平稳的时空序列$Z_t = (Z(s_1,t), Z(s_2,t), \cdots, Z(s_N,t))$，$t = 1,2,\cdots,T$，其均值、方差、协方差等不随空间位置的不同和时间的推移而发生变化。也就是对于任意时间延迟 k 和空间延迟 h，均值满足：

$$\mu_{Z(s_i,\mathrm{t})} = \mathrm{E}\left(\mathrm{Z}(s_i,\mathrm{t})\right) = \mathrm{E}\left(\mathrm{Z}(s_i+\mathrm{h},\mathrm{t}+\mathrm{k})\right),$$

方差满足：

$$\sigma_{Z(s_i,\mathrm{t})}^2 = \mathrm{E}\left(\mathrm{Z}(s_i,\mathrm{t}) - \mu_{Z(s_i,\mathrm{t})}\right)^2 = \mathrm{E}\left(\mathrm{Z}(s_i+\mathrm{h},\mathrm{t}+\mathrm{k}) - \mu_{Z(s_i,\mathrm{t})}\right)^2,$$

协方差满足：

$$C_{hk} = \mathrm{Cov}\left(\mathrm{Z}(s_i,\mathrm{t}), \mathrm{Z}(s_i+\mathrm{h},\mathrm{t}+\mathrm{k})\right)$$

$$= \mathrm{E}\left(\mathrm{Z}(s_i,\mathrm{t}) - \mu_{Z(s_i,\mathrm{t})}\right)\left(\mathrm{Z}(s_i+\mathrm{h},\mathrm{t}+\mathrm{k}) - \mu_{Z(s_i,\mathrm{t})}\right)$$

其均值、方差和协方差的估计值为：

$$\hat{\mu}_{Z(s_i,\mathrm{t})} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{Z}(s_i,\mathrm{t})}{NT},\ (i = 1,2,\cdots,N; t = 1,2,\cdots,T),$$

$$\hat{\sigma}^2_{Z(s_i,\mathrm{t})}=\frac{1}{kh}\sum_{i=1}^{h}\sum_{t=1}^{k}\left(\mathrm{Z}(s_i,\mathrm{t})-\mu_{Z(s_i,\mathrm{t})}\right)^2,\ (i=1,2,\cdots,N;t=1,2,\cdots,T),$$

$$\hat{C}_{hk}=\frac{1}{kh}\left[\sum_{i=1}^{h}\sum_{t=1}^{k}\left(\mathrm{Z}(s_i,\mathrm{t})-\mu_{Z(s_i,\mathrm{t})}\right)\right]\left[\sum_{i=1}^{h}\sum_{t=1}^{k}\left(\mathrm{Z}(s_i+m,\mathrm{t}+\mathrm{k})-\mu_{Z(s_i,\mathrm{t})}\right)\right],$$

对于时空序列，在理论上满足以上条件即为平稳的时空序列，然而在实际应用中，时空序列很难满足平稳条件，一般只要时空序列的均值和方差满足不随时间和空间的演变发生变化的条件，就可以认为该时空序列为平稳的时空序列[19]。

2.1.2 相关性

时间的自相关性和偏相关性是时间序列的主要特征，利用自相关函数和偏相关函数进行衡量。对于一组观测序列$(Y_1,Y_2,\cdots,Y_n)$，自相关函数的估计值为：

$$r_k=Corr(Y_t,Y_{t-k})=\frac{\sum_{t=k+1}^{n}(Y_t-\bar{Y})(Y_{t-k}-\bar{Y})/(n-k-1)}{\sum_{t=1}^{n}(Y_i-\bar{Y})^2/(n-1)},\mathrm{k}=1,2,\cdots,\mathrm{n},$$

另外，消除中间介入变量$(Y_{t-1},Y_{t-2},Y_{t-3},\cdots,Y_{t-k+1})$的影响后，$Y_t$与$Y_{t-k}$的相关系数称为 k 阶滞后偏自相关函数，利用 Yule-Wolker 方程可以得到偏相关系数为：

$$\phi_{kk}=Corr(Y_t,Y_{t-k}|Y_{t-1},Y_{t-2},\cdots,Y_{t-k+1})=\frac{\begin{vmatrix}\rho_0 & \rho_1 & \cdots & \rho_{k-2} & \rho_1\\ \rho_1 & \rho_0 & \cdots & \rho_{k-3} & \rho_2\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_0 & \rho_k\end{vmatrix}}{\begin{vmatrix}\rho_0 & \rho_1 & \cdots & \rho_{k-1}\\ \rho_1 & \rho_0 & \cdots & \rho_{k-2}\\ \vdots & \vdots & \ddots & \vdots\\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_0\end{vmatrix}}$$

其中，$k=1,2,\cdots$,n

对于空间自相关性最常用的度量方法是 Moran’s I 系数，反应了空间邻近的区域单元属性值的相似程度，是应用最广泛的空间自相关统计量：

$$\mathrm{I}=\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}}\times\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(z_i-\bar{z})(z_j-\bar{z})}{\sum_{i=1}^{n}(z_i-\bar{z})^2},\ \mathrm{I}\epsilon[-1,1]$$

其中，$w_{ij}=\begin{cases}1 & 空间对象i与j相互邻接\\ 0 & 空间对象i与j不邻接\end{cases}$，$\mathrm{W}=(w_{ij})_{n\times n}$为空间权重矩阵，定量测度了空间对象的相关程度,表达了不同空间对象之间邻接关系。若$\mathrm{I}=0$表示空间数据不相关，$\mathrm{I}>0$空间数据正相关，$\mathrm{I}<0$空间数据负相关。I 的绝对值越大，表示空间分布的

相关性越大，说明空间分布上具有聚集分布的现象。I 的绝对值越小，表示空间分布的相关性越小。

同理，对于序列$Z_t = Z(s_i,t) = (Z(s_1,t),Z(s_2,t),\cdots,Z(s_N,t))$，$t = 1,2,\cdots,T$为时间参数，$s_i,i = 1,2,\cdots,N$为空间位置参数，那么时空自相关函数可以定义为：

$$\rho_{h0}(k) = \frac{\gamma_{h0}(k)}{\sigma_h(0)\sigma_0(0)} = \frac{cov\left(W^{(h)}Z_t, W^{(0)}Z_{t+k}\right)}{\sqrt{Var(W^{(h)}Z_t)}\sqrt{Var(W^{(0)}Z_t)}}$$

其中，k为时间延迟系数，Z_{t+k}表示时间延迟为k的样本观测值；h为空间延迟系数，$W^{(h)}$为空间延迟期是h的空间权重矩阵，$W^{(0)}$为空间延迟期是 0 的空间权重矩阵，一般地，$W^{(0)}$为单位阵；$\gamma_{h0}(k)$为空间延迟为h 、时间延迟为k和空间延迟为 0、时间延迟为 k的时空协方差函数，即：

$$\gamma_{h0}(k) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T-k}\left[W^{(h)}Z(s_i,t)\right]\left[W^{(0)}Z(s_i,t+k)\right]}{N(T-k)}$$

时空自相关系数的取值范围为[-1,1]，自相关系数的绝对值越接近 1 说明相关程度越高。结合时间偏相关函数的求法，根据时空自相关系数建立 Yule-Walker 方程组：

$$\rho_h(k) = \sum_{h=0}^{m}\sum_{k=1}^{p}\varphi_{kh}\rho_{h-1}(k)$$

求解方程组就可以得到时空偏相关系数φ_{kh}的表达式，时空偏相关系数能够真实的反应时空序列中两个变量的相关性。时空序列的自相关函数和偏自相关函数可以用来检验时空数据的平稳性、季节性，判断时空模型的时间延迟和空间延迟。

2.2 ARIMA 模型理论和构造介绍

2.2.1 ARIMA 模型理论基础

时间序列模型根据原序列是否平稳以及回归中所含部分的不同，包括移动平均过程(MA)、自回归过程(AR)、自回归移动平均过程(ARMA)以及差分自回归滑动平均(ARIMA)过程。$ARIMA(p,d,q)$建立模型的实质就是对不平稳的时间序列进行差分运算，平稳化之后再建立$ARMA(p,q)$模型进行预测的过程。

因为我们面对的大部分时间序列是非平稳的，结合农产品价格数据的特征，在这

里的模型采用的是$ARIMA(p,d,q)$模型：

$$\begin{cases}\Phi(\mathcal{B})(1-\mathcal{B})^d Y_t=\Theta(\mathcal{B})\varepsilon_t, t=1,2,\cdots\\ E(\varepsilon_t)=0, Var(\varepsilon_t)=\sigma_\varepsilon^2\\ E(\varepsilon_t\varepsilon_s)=0, s\neq t\\ Ex_s\varepsilon_t=0, \forall s<t\end{cases}$$

其中，$\Phi(\mathcal{B})=1-\sum_{j-1}^{p}\varphi_j\mathcal{B}^j$，为平稳可逆$ARIMA(p,d,q)$模型的自回归系数多项式；$\Theta(\mathcal{B})=1-\sum_{j=1}^{q}\theta_j\mathcal{B}^j$，为平稳可逆$ARIMA(p,d,q)$模型的滑动平均系数多项式。

对于$ARIMA(p,d,q)$模型，当$d=0$时，为$ARMA(p,q)$模型：

$$Y_t=\varphi_1Y_{t-1}+\cdots+\varphi_pY_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\cdots-\theta_q\varepsilon_{t-q}$$

当$p=0$时，为$IMA(d,q)$模型：

$$\nabla^d Y_t=\varepsilon_t-\theta_1\varepsilon_{t-1}-\cdots-\theta_q\varepsilon_{t-q}$$

当$q=0$时，为$ARI(p,d)$模型：

$$\nabla^d Y_t=\varphi_1\nabla^d Y_{t-1}+\cdots+\varphi_p\nabla^d Y_{t-p}+\varepsilon_t$$

2.2.2 平稳性检验

从时序图看数据的基本趋势：围绕某直线波动、呈指数上升或下降趋势、显示出季节性或多种趋势的组合等，可以用来判断某个序列是平稳性时间序列还是非平稳性时间序列，如果有明显的趋势性或者周期性，则不是平稳序列。

除时序图之外，单位根检验是判断某一序列是否平稳的标准方法，常见的单位根检验方法有：ADF 检验、DFGLS 检验、PP 检验、KPSS 检验、ERS 检验和 NP 检验。在这里采用的是在实际中最常用的 ADF(Augmented Dickey-Fuller Test)检验法[29]，即增广的 DF 检验方法来进行单位根检验。ADF 检验式有三种：不含常数项和趋势项、只含常数项不含趋势项、包含常数项和时间趋势项。

对于 AR(p)过程：

$$Y_t=\varphi_0+\varphi_1Y_{t-1}+\cdots+\varphi_pY_{t-p}+\varepsilon_t,\ \mathrm{t}=1,2,\cdots$$

其中$\varepsilon_t\sim iid(0,\sigma^2)$，该过程 p 阶序列相关，用 p 阶自回归过程进行修正，上式两端减去y_{t-1}，得到：

$$\nabla y_t = \varphi_0 + \varphi y_{t-1} + \sum_{i-1}^{p-1} \eta_i \nabla Y_{t-i} + \varepsilon_t$$

其中，$\varphi = \sum_{i=1}^{p} \varphi_i - 1, \eta_i = -\sum_{j=i+1}^{p} \varphi_j$。若$\{Y_t\}$为平稳序列，则$\sum_{i=1}^{p} \varphi_i < 1$，即$\varphi < 0$；若$\{Y_t\}$为非平稳的时间序列，则至少存在一个单位根，有$\sum_{i=1}^{p} \varphi_i = 1$，即$\varphi = 0$。那么ADF 检验可以表示为：

$$H_0：\varphi = 0\big(\text{非平稳}\big) \leftrightarrow H_1：\varphi < 0(\text{平稳})$$

ADF 统计量为：$\tau = \hat{\varphi}/S(\hat{\varphi})$，其中$S(\hat{\varphi})$为参数$\hat{\varphi}$的样本标准差。

对于自回归过程 AR(p)，ADF 检验的 3 种基准模型：

(1)不包含常数项和时间趋势项的自回归过程 AR(p)：

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t$$

(2) 包含常数项、不包含时间趋势项的自回归过程 AR(p)：

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t$$

(3) 包含常数项、时间趋势项的自回归过程 AR(p)：

$$Y_t = \varphi_0 + \beta t + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t$$

一般先从模型(3)开始单位根检验，当确定不含有趋势后，继续用模型(2)检验，若存在单位根，继续用模型(1)进行检验。在这个过程中如果发现不存在单位根，则检验结束。

当统计量的 p 值小于临界值时，拒绝原假设，该序列为平稳的时间序列；当 p 值大于临界值时，没有足够的理由拒绝原假设，认为该时间序列是非平稳的。为了模型的需要，此时需要对非平稳的时间序列进行平稳化处理，根据时序图的趋势进行差分处理、对数变换、百分比变动、幂变换等，然后再进行单位根检验，直到该序列平稳。通常情况下，在采用 ARIMA 模型时,一阶差分能够有效地使时间序列平稳，二阶差分也能得到偶尔的使用，更高阶的差分有可能会降低预测的精度，所以进行平稳化处理时，差分的阶数不应该太高。

2.2.3 纯随机性检验

纯随机性检验即白噪声检验，白噪声序列满足下列条件：

$$\begin{cases} \forall t \in T，\ EY_t = \mu \\ \forall t,s \in T，\ \gamma(t,s) = \begin{cases} \sigma^2，\ t = s \\ 0，\ t \neq s \end{cases} \end{cases}$$

白噪声的序列值各项之间没有任何相关关系，没有进一步信息挖掘的意义。对序列进行白噪声检验，如果序列为白噪声序列，说明序列中的数据没有相关关系，则该序列没有建立模型的必要。结合序列值之间的绝对变异性和偶然相关性，白噪声检验可以认为原假设是相互独立的，备择假设具有相关性，即：

$$H_0：\rho_1 = \rho_2 = \cdots = \rho_k = 0，\ \forall k \geq 1$$

$$H_1：\text{至少存在某个}\rho_i \neq 0，\ \forall k \geq 1, i \leq k$$

采用 LB(Liung-Box)检验统计量：

$$\mathrm{T} = n(n+2)\sum_{k=1}^{m}\frac{\rho_k^2}{n-k} \sim \chi^2(m)$$

其中，n 为样本观测期数，m 为制定延迟期数，$\rho_k = \frac{\sum_{i=1}^{n-k}\varepsilon_i\varepsilon_{i+k}}{\sum_{i=1}^{n}\varepsilon_t^2}$，根据统计量和临界值的比较，判断是否有足够的理由拒绝原假设。

2.2.4 模型识别

为了得到最优的预测模型，应选取适当的阶数拟合模型。最常用的方法主要通过平稳性检验确定差分阶数 d，通过时间自相关系数和偏相关系数来确定模型类别和模型的阶数 p 和 q。

对于 AR(p)模型$Y_t = \sum_{i=1}^{p}\varphi_i Y_{t-i} + \varepsilon_t$、MA(q)模型$Y_t = \varepsilon_t - \sum_{j=1}^{q}\theta_j\varepsilon_{t-j}$和 ARMA(p,q)模型 $Y_t = \sum_{i=1}^{p}\varphi_i Y_{t-i} + \varepsilon_t - \sum_{j=1}^{q}\theta_j\varepsilon_{t-j}$，自相关系数和偏自相关系数有不同的表现特征，可以根据此特征初步判断模型类型和模型阶数。若平稳序列的偏相关函数是 p 阶截尾的，而自相关函数是拖尾的，可断定序列适合 AR(p)模型；若平稳序列的偏相关函数是拖尾的，而自相关函数是 q 阶截尾的，则可断定序列适合 MA 模型；若平稳序列的偏相关函数和自相关函数均是拖尾的，则序列适合 ARMA 模型。判段准则如下表所示：

表 2.1　acf/pacf 判断 ARIMA 模型类型

	AR(p)	MA(q)	ARMA(p,q)
自相关系数	指数衰减或衰减正弦波	q 后截尾	拖尾
偏自相关系数	p 后截尾	指数衰减或衰减正弦波	拖尾

选择不同的变量组合可以得到不同的模型，为了识别出拟合效果更优的模型，最常用的方法是赤池信息准则(AIC)和贝叶斯信息准则(BIC)。阶次估计的标准，是建立 AIC、BIC 目标函数，并使目标函数最小化。AIC 准则也就是使下列信息值最小化的模型：

$$\mathrm{AIC} = -2\log(\text{极大似然估计}) + 2\mathrm{k}$$

AIC 信息值刻画了估计模型相对真实模型的信息损失，因为 AIC 是有偏估计，为了消除偏差更好衡量模型的优劣性，在 AIC 中增加另一个非随机的惩罚项：

$$AIC_c = \mathrm{AIC} + \frac{2(k+1)(k+2)}{n-k-2}$$

其中，n 是有效样本容量，k 是除去噪声方差后的总的参数数量。BIC 准则和 AIC 准则同样也是用于选择模型，$\mathrm{BIC} = -2\log(\text{极大似然估计}) + \mathrm{klog(n)}$。

2.2.5　参数估计

识别模型之后需要进行参数估计，常用方法有矩估计、最小二乘估计、极大似然估计等，这里采用包含信息比较完整的极大似然估计。不同于最小二乘仅利用一阶和二阶矩，极大似然法优势在于利用了数据包含所有信息，具有高精度、一致性、渐近正态性和渐进有效性等性质，可在一般条件下得出许多大样本结论。

具体求解极大似然估计的步骤是：一是先求出并计算似然函数，二是求似然函数的最大值。对于 ARIMA 模型，观测值$(Y_1, Y_2, \cdots, Y_n)$，极大似然估计量就是最大化似然函数的参数取值。在使用极大似然估计时，我们要假设白噪声项独立、零均值方差同分布，即$\varepsilon_t \sim N(0, \sigma^2)\ \ i.d.d.$。

对于模型$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$，分布函数通常是未知的，一般假定序列服从多元正态分布。记：

$$\tilde{Y} = (Y_1, Y_2, \cdots, Y_n)$$

$$\tilde{\beta} = (\varphi_1, \cdots, \varphi_p, \theta_1, \cdots, \theta_q)'$$

$$\sum_n = E\left(\tilde{Y}'\tilde{Y}\right) = \Omega\sigma_\varepsilon^2$$

其中，

$$\Omega = \begin{bmatrix} \sum_{i=0}^{\infty} G_i^2 & \cdots & \sum_{i=0}^{\infty} G_i G_{i+n-1} \\ \vdots & \ddots & \vdots \\ \sum_{i=0}^{\infty} G_i G_{i+n-1} & \cdots & \sum_{i=0}^{\infty} G_i^2 \end{bmatrix},\ G_0 = 1, G_k = (\varphi_1 - \theta_1)\varphi_1^{k-1}, k = 1,2,\cdots$$

$\tilde{Y}$的似然函数为：

$$L(\tilde{\beta};\tilde{Y}) = \ (2\pi)^{-\frac{n}{2}}|\sum_n|^{-\frac{1}{2}}\exp\left(-\frac{\tilde{Y}'\sum_n^{-1}\tilde{Y}}{2}\right)$$

$$= (2\pi\sigma_\varepsilon^2)^{-n/2}|\Omega|^{-\frac{1}{2}}\exp\left(-\frac{\tilde{Y}'\Omega^{-1}\tilde{Y}}{2\sigma_\varepsilon^2}\right)$$

对数似然函数为：

$$l\left(\tilde{\beta};\tilde{Y}\right) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma_\varepsilon^2) - \frac{1}{2}ln|\Omega| - \frac{1}{2\sigma_\varepsilon^2}[\tilde{Y}'\Omega^{-1}\tilde{Y}]$$

然后对$l\left(\tilde{\beta};\tilde{Y}\right)$中的未知参数求偏导，求解对数似然方程组，得到相对应的参数估计量。

2.2.6 模型检验

如果模型被正确识别，并且参数估计充分接近真实值，那么残差就应该近似具有正态白噪声的性质。模型的参数检验，包括正态性检验、残差的白噪声检验和参数的显著性检验，通过检验然后判断所建模型是否可取，根据假设检验的结果进行模型优化和修正。

对模型的参数进行显著性检验，则将不显著的参数对应的变量从模型中删掉，重新估计新模型参数，使模型能够精简化。参数的显著性检验为：

$$H_0\colon\ \beta_i = 0 \ \leftrightarrow\ H_1\colon\ \beta_i \neq 0,\ \forall 1 \le i \le k$$

检验统计量为：

$$\mathrm{T} = \sqrt{n-k}\,\frac{\hat{\beta}_i - \beta_i}{\sqrt{a_{ii}Q(\hat{\beta})}} \sim t(n-k)$$

其中$\frac{a_{ii}Q(\hat{\beta})}{n-k}$是$\hat{\beta}_i$的方差的估计[30]，根据统计量和临界值的比较$|\mathrm{T}| \geq t_{1-\alpha}(n-k)$，判断是否有足够的理由拒绝原假设。

对残差的正态性检验常用的方法有 shapiro-wilk 检验、Kolmogorov-Smirnov 检验，其中 shapiro-wilk 检验适合小样本，这里的数据量相对较多，所以采用 Kolmogorov-Smirnov 统计量进行检验。KS 检验基于累积分布函数，用以检验一个经验分布是否符合某种理论分布，即：

$$H_0：e_t \sim N(0,\sigma^2)$$

$$H_1：e_t \text{不服从正态分布}$$

对残差进行白噪声检验，如果残差序列为白噪声序列,白噪声序列就没有可提取的信息，则认为模型的拟合效果很好；若残差检验为非白噪声序列，则说明模型的拟合需要进一步优化。

$$H_0：\rho_1 = \rho_2 = \cdots = \rho_k = 0，\ \forall k \geq 1$$

$$H_1：\text{至少存在某个}\rho_i \neq 0，\ \forall k \geq 1, i \leq k$$

采用 LB(Liung-Box)检验统计量：

$$\mathrm{T} = n(n+2)\sum_{i=1}^{k}\frac{\rho_i^2}{n-i} \sim \chi^2(k-p-q)$$

其中，$\rho_k = \frac{\sum_{i=1}^{n-k}\varepsilon_i\varepsilon_{i+k}}{\sum_{i=1}^{n}\varepsilon_t^2}$，p、q 为模型的阶数，根据统计量和临界值的比较，判断是否有足够的理由拒绝原假设。

2.2.7 模型预测

由序列$\{X_t\}$的模型 ARIMA(p,d,q)：

$$\Phi(\mathcal{B})(1-\mathcal{B})^d X_t = \Theta(\mathcal{B})\varepsilon_t，\ \varepsilon_t \sim WN(0,\sigma^2)\ i.d.d.，\ \mathrm{t} \in \mathrm{N},$$

可知$Y_t = (1-\mathcal{B})^d X_t，\ t = d+1, d+2, \cdots, n$满足 ARMA(p,q)模型：

$$\Phi(\mathcal{B})Y_t = \Theta(\mathcal{B})\varepsilon_t，\ \mathrm{t} \in \mathrm{N}.$$

对于$\{X_t\}$进行向前预测的问题，可以利用 ARMA(p,q)预测方法先得到$\{Y_t\}$的向前 k 步预

测为：

$$\tilde{Y}_{n+j} = L\left(Y_{n+j} \middle| Y_{d+1}, Y_{d+2}, \cdots, Y_n\right)，\ j = 1,2,\cdots,k$$

再通过差分公式变换得到$\{X_t\}$的向前 k 步预测：

$$\hat{X}_{n+k} = \tilde{Y}_{n+k} - \sum_{j=1}^{d} C_d^j (-1)^j \hat{X}_{n+k-j}，\ k \geq 1$$

其中，$\hat{X}_{n-j} = X_{n-j}$，当$j \geq 0$时。

2.3 STARMA 时空序列模型理论综述

2.3.1 STARMA 模型理论基础

时空数据是在空间上有相关关系的时间序列集合，包括时间、位置、属性值 3 个基本特征的数据，又可以称为时空序列。时空序列具有定位、定性和时态特性。定位是指在已知的坐标系里空间对象都具有唯一的空间位置,定性是指与空间对象有关的属性,时态是指空间对象随时间的变化信息。时空序列建模是指寻找一种时空数据分析方法，对未观测时空位置属性进行建模和预测的过程。时空建模相当于在时间模型上的空间扩展，STARMA 模型充分考虑时间、空间因素对属性值序列的影响，对时间自相关移动平均模型进行了空间扩展，比较适合于各种时空序列数据的建模。STARMA 模型建立的方法同 ARMA 模型建立的步骤基本相同，构建空间权重矩阵，通过判别平稳性对模型进行平稳化处理，然后进行识别、建立、检验模型等。

时空序列$Z(t) = (Z(s_1,t), Z(s_2,t), \cdots, Z(s_N,t))$，$t = 1,2,\cdots,T$，$i = 1,2,\cdots,N$，为 t 时刻每个空间位置$s_i$的观测值，那么 STARMA(p,q)模型可以表示为：

$$Z(t) = \sum_{k=1}^{p} \sum_{l=0}^{\lambda_k} \varphi_{kl} W^{(l)} Z(t-k) - \sum_{k=1}^{q} \sum_{l=0}^{m_k} \theta_{kl} W^{(l)} \varepsilon(t-k) + \varepsilon(t)$$

其中，p 为时间自相关阶数，q 是时间移动平均阶数，$W^{(l)}$是空间权重矩阵，λ_k是第 k 个时间自相关项的空间阶数，m_k是第 k 个时间移动平均项的空间阶数，φ_{kl}、θ_{kl}为对应的模型参数，$\varepsilon(t)$为服从正态分布的随机误差向量，满足：

$$E\left(\varepsilon(t)\right) = 0$$

$$E(\varepsilon(t)\varepsilon(t+s)^T)=\begin{cases}\sigma^2 I_N & s=0\\ 0 & \text{其他}\end{cases}$$

对于 STARMA(p,q)模型，当$q=0$时，为时空自回归 STAR(p)模型：

$$Z(t)=\sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\varphi_{kl}W^{(l)}Z(t-k)+\varepsilon(t)$$

当$p=0$时，为时空滑动平均 STMA(q)模型：

$$Z(t)=\varepsilon(t)-\sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}W^{(l)}\varepsilon(t-k)$$

STAR(p)模型和 STMA(q)模型均为时空自回归滑动平均模型 STARMA(p,q)模型的一种类型。

2.3.2 空间权重矩阵的建立

空间权重矩阵定量测度了空间对象的相关程度,表达了不同空间对象之间邻接关系，是描述空间邻近性的定量化测度。根据空间未知的特征，建立空间权重矩阵的方法有邻接关系、重心距离等。这里采用邻接法建立空间权重矩阵，那么一阶空间权重矩阵是指与目标空间位置距离最近的空间单元形成的权矩阵，用$\mathrm{W}^1=(w_{ij})_{n\times n}$表示：

$$\mathrm{W}^1=\begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1n}\\ w_{21} & w_{22} & \cdots & w_{2n}\\ \vdots & \vdots & & \vdots\\ w_{n1} & w_{n2} & \cdots & w_{nn}\end{bmatrix}$$

其中，$w_{ij}=\begin{cases}1 & \text{空间对象}i\text{与}j\text{相互邻接}\\ 0 & \text{空间对象}i\text{与}j\text{不邻接}\end{cases}$；以此类推，$\mathrm{W}^2$为 2 阶空间权重矩阵，其中$w_{ij}=$

$\begin{cases}1 & i\text{与}k\text{邻接，}j\text{与}k\text{邻接，}i\text{与}j\text{不邻接}\\ 0 & \text{其他}\end{cases}$；$\mathrm{W}^0$是 0 阶空间权重矩阵，为单位阵。在实际应用中，空间权重矩阵的阶数最高取 2 阶即可满足模型要求，过高的阶数容易带来计算的复杂度。

利用邻接关系建立的空间权重矩阵，若两个区域近邻则定义权重为 1，这样的单位权重在时空序列的建模中可能会产生偏差，对建模效果有一定影响。需要对权重进行标准化：

$$w_{ij} = w_{ij}/w_i$$

其中，$w_i = \sum_{j=1}^{N} w_{ij}$。

2.3.3 时空序列平稳性检验

利用时空序列的自相关系数判断序列是否平稳，当自相关系数会在时间延迟k和空间延迟h增加迅速接近于 0，可以认为该时空序列具有平稳性；若自相关系数在时间延迟k和空间延迟h增加不能很快的下降为 0，可以认为该时空序列不具有平稳性。对于非平稳的时空序列通过差分等方法转化为平稳的时空序列。

2.3.4 模型识别

一般情况下，空间延迟取 1 或 2 就可以满足模型需求，类似于时间序列模型，时间延迟的确定和模型类别的判断需要根据时空序列的自相关和偏自相关函数进行分析。

若在所有空间滞后上，自相关函数时间延迟值拖尾呈几何递减，而时空偏相关函数值在空间延迟为h的相邻区域,时间延迟 p 后截尾,那么模型为自回归过程 STAR(p)；若时空自相关函数值在空间延迟为h的相邻区域，时间延 q 后截尾，而偏相关函数在在所有空间滞后上，时间延迟值拖尾呈几何递减，表明模型为时间移动平均过程 STMA(q)；若时空自相关函数和偏相关函数均呈现拖尾趋势，该模型为 STARMA 模型，模型的阶数需要拟合出多个模型根据信息准则进行判断出最优的模型。如下表：

表 2.2 stacf/stpacf 判断 STARMA 模型类型

	STAR(p)	STMA(q)	STARMA(p,q)
自相关系数	几何递减	q 后截尾	几何递减
偏自相关系数	p 后截尾	几何递减	几何递减

2.3.5 参数估计

在参数估计方面，由于时空自相关的存在，用最小二乘法来估计模型容易产生时空自回归参数的偏差和无效估计，一般采用最大似然估计对时空模型进行参数估计。根据 STARMA(p,q)模型可得随机误差项为：

$$\varepsilon(s_i,t)=Z(s_i,t)-\sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\varphi_{kl}W^{(l)}Z(t-k)+\sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}W^{(l)}\varepsilon_i(t-k)$$

又因为模型的随机误差项的分布$\varepsilon\sim N(0,\sigma^2 I_{NT})$，可以得到$Z(s_i,t)$似然函数：

$$L(\varepsilon|\varphi,\theta,\sigma^2)=(2\pi)^{-TN/2}|\sigma^2|^{-TN/2}\exp(-\varepsilon'\varepsilon/2\sigma^2)$$

对数似然函数为：

$$lnL=-\frac{TN}{2}\ln(2\pi)-\frac{TN}{2}\ln(\sigma^2)-\frac{\varepsilon'\varepsilon}{2\sigma^2}$$

对似然函数进行求偏导：

$$\begin{cases}\dfrac{\partial lnL}{\partial\sigma^2}=0\\ \dfrac{\partial lnL}{\partial\theta}=0\\ \dfrac{\partial lnL}{\partial\varphi}=0\end{cases}$$

求解上面方程式即可得到参数的最大似然估计值。

2.3.6 模型检验

估计出时空模型后，应该对时空模型进行检验，检验该模型是否能充分表达时空数据。同时间序列模型检验类似，最重要的方面是对时空模型残差进行检验，也就是对残差进行纯随机性检验和正态性检验。采用时空自相关系数和偏相关系数可以用来检验残差的相关性，因为随机误差过程是序列无关的，所以随机误差过程的时空自相关函数和偏相关函数在相关图中均等于0的水平直线。如果模型的残差为高斯白噪声，则残差的相关系数应近似为0，说明模型的效果比较好。如果拟合残差不是高斯白噪声，那么残差可能具有一定的模式，即在时间和空间上存在相关性或变异性，该模式又可以通过 STRAMA 模型建模表示。识别残差时空模型并将该模型与时空模型结合得到更好的更新时空模型。

第三章　蔬菜价格指数的 ARIMA 预测模型

3.1　数据概况

对蔬菜批发价格指数的预测，可以有效了解我国和各个省市地区蔬菜批发价格的整体水平，记录全国蔬菜农产品批发价格变化动态，揭示蔬菜批发价格变化规律，对农产品市场的宏观调控有指导意义。对此，结合时间序列数据的特点，在这里构建 ARIMA 模型进行蔬菜批发价格指数的预测。

选取 2005 年-2016 年期间月均蔬菜价格指数进行建模分析，为分析模型的预测效果，将数据分为训练集和测试集，把 2005 年-2015 年期间月均蔬菜价格指数作为模型拟合的训练集，把 2016 年 8 个月的数据作为测试集检验模型预测的精确度。

3.2　绘制时序图

选择山东省的蔬菜价格指数数据进行 ARIMA 建模分析，利用统计软件 R 语对数据进行处理，画出时序图：

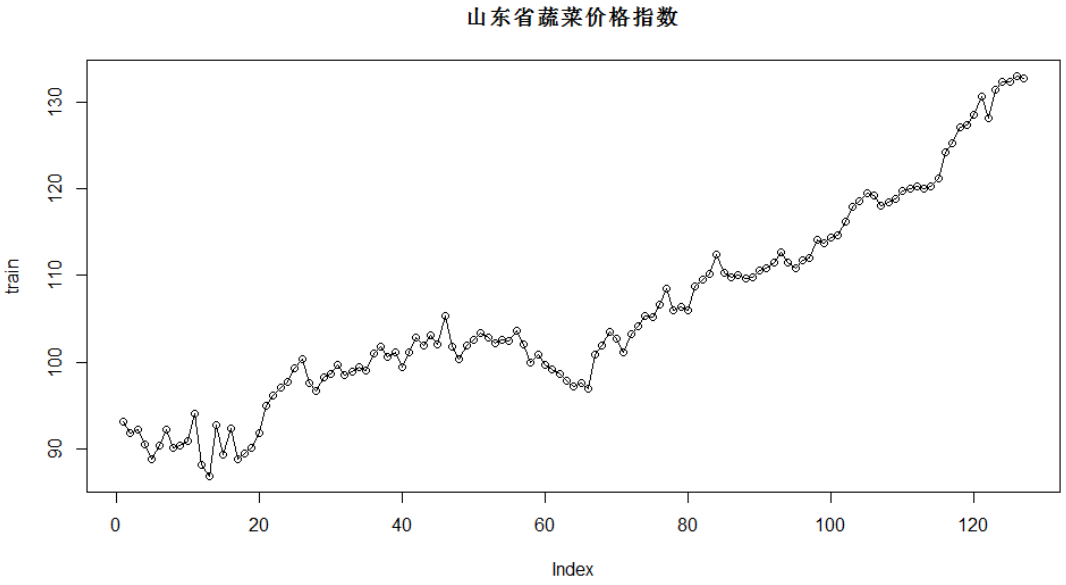

图 3.1　山东省蔬菜价格指数序列

上图为山东省蔬菜市场价格指数的月均数据从 2005.06-2015.12 的趋势，总体呈现不断波动而且逐渐上升的趋势，所以该序列的表现为非平稳的。通常一次或最多两次差

分，可能伴随对数或者其他形式，可以完成平稳性的化简。根据原数据的表现趋势，先进行一次差分处理，查看数据的波动趋势。

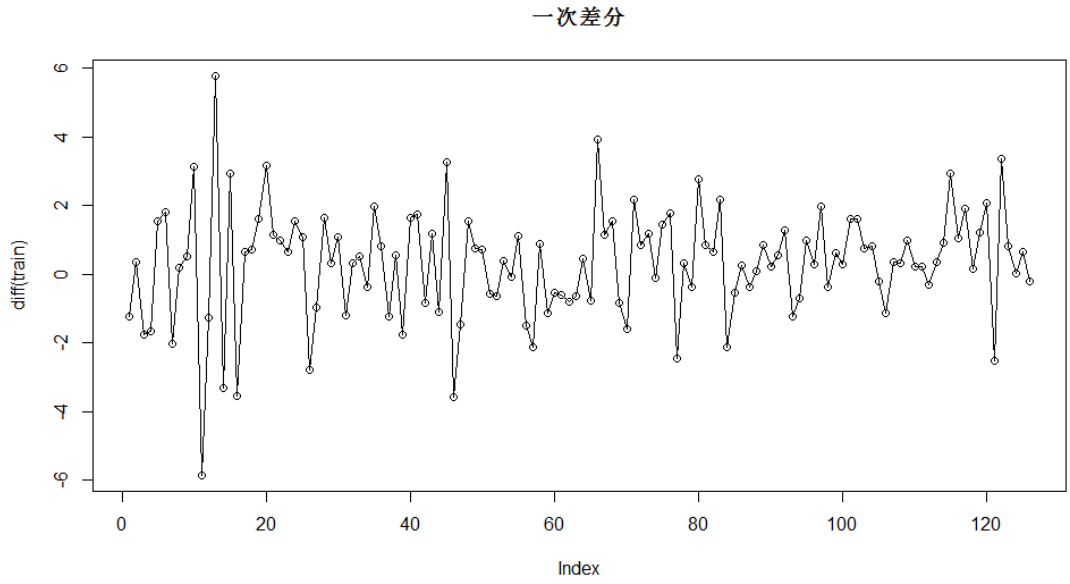

图 3.2　山东省蔬菜价格指数一阶差分序列

从上面的时序图中可以看出，进行一次差分后，该时间序列有良好的平稳性。

3.3　单位根检验

从时序图中我们只可以初步判断出时间序列的大致趋势和波动情形，判断是否平稳性的准确度有待提高，为了更加确定平稳性的判断，应该对数据进行单位根检验，通过假设检验确定该序列是否为平稳序列。根据统计量的 p 值进行判断，当 p 值小于给定置信水平的临界值时则拒绝原假设，当 p 值大于个给定置信水平的临界值则接受原假设。对于原序列进行平稳性检验，利用单位根检验，结果如下：

表 3.1　单位根检验平稳性结果

	无	含常数项	含常数项和趋势项
tau 统计量	2.06	-0.19	-2.47
1%临界值	-2.58	-3.46	-3.99
5%临界值	-1.95	-2.88	-3.43
10%临界值	-1.62	-2.57	-3.13
是否稳定（1/0）(5%)	1	0	0

单位根检验结果显示，该时间序列对于含有常数项和时间趋势项单位根检验中，统计

量的值大于置信水平α为0.05的临界值，是非平稳的序列。结合该时间序列的趋势图，对时间序列进行一次差分，然后对一次差分后的时间序列进行单位根检验，结果如下：

表 3.2　单位根检验平稳性结果

	无	含常数项	含常数项和趋势项
tau 统计量	-2.87	-3.89	-3.85
1%临界值	-2.58	-3.46	-3.99
5%临界值	-1.95	-2.88	-3.43
10%临界值	-1.62	-2.57	-3.13
是否稳定（1/0）(5%)	1	1	1

从检验结果可以看出，差分之后的数据，检验统计量的值小于置信水平α为 0.05 的临界值,说明我们有足够的理由拒绝原假设，故而可以认为一次差分后的时间序列是平稳的。

3.4　白噪声检验

对于一阶差分后的序列进行白噪声检验，用于检验某个时间段内的一系列观测值是不是随机的独立观测值，判断序列是否有进一步挖掘的必要。检验结果如下表，对于延迟期为 6 期、12 期、18 期的 Liung-Box 统计量的 p 值均小于置信水平α=0.05，那么有足够的理由拒绝原假设的完全随机性，可以认为差分后的序列为平稳的非白噪声序列，可以对差分后的序列进行建立模型预测分析。

表 3.3　白噪声检验结果

Box-Pierce test			
lag	X-squared	p-value	是否显著
6	21.153	0.003222	显著
12	30.54	0.004476	显著
18	42.658	0.002381	显著

3.5　拟合 ARIMA 模型

由时序图和单位根检验可以得到，数据需要经过一次差分后可以得到平稳的序列，对差分后的序列计算自相关系数和偏相关系数，并作图：

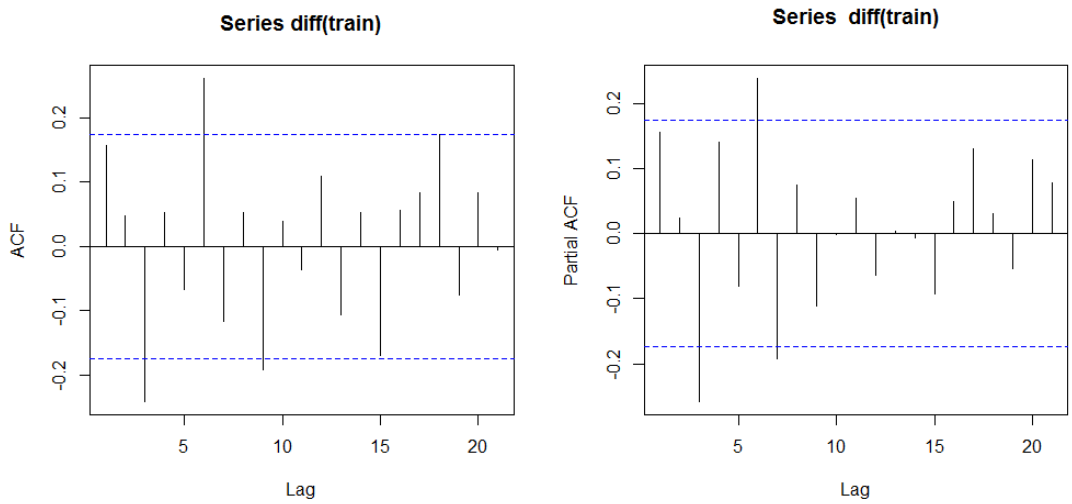

图 3.3　自相关系数和偏相关系数

从差分后序列的自相关和偏相关图可以发现，自相关和偏自相关均表现出拖尾的特征，所以建立 ARIMA 模型。根据判断出的模型类别，对于差分后的序列建立模型，选取其中 AIC 信息准则的大小，确定最优模型。

表 3.4　模型拟合过程

模型	是否含有漂移项	AIC 信息
ARIMA(2,1,2)	with drift	344.9409
ARIMA(1,1,2)	with drift	343.4257
ARIMA(1,1,1)	with drift	362.223
ARIMA(1,1,3)	with drift	344.692
ARIMA(2,1,2)		343.527
ARIMA(1,1,2)		349.2949
ARIMA(1,1,0)	with drift	359.2467
ARIMA(0,1,1)	with drift	359.6258
ARIMA(1,1,2)	with drift	339.9686
ARIMA(1,1,1)	with drift	361.3755
ARIMA(0,1,2)	with drift	359.8277

Best model: ARIMA(1,1,2) with drift

在上表中一阶差分拟合多个 ARIMA 模型，其中 AIC 信息值最小的为最优模型，即 ARIMA(1,1,2)差分自回归滑动平均模型是适合该组数据的最优模型：

$$W_t = \varphi_0 + \varphi_1 W_t + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}，\text{其中} W_t = Y_t - Y_{t-1}$$

可以转化为：

$$Y_t = \varphi_0 + (1+\varphi_1)Y_{t-1} + \varphi_1 Y_{t-2} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

3.6 模型估计与检验

由上一步可以确定试用模型 ARIMA(1,1,2)，对该模型中的参数进行估计，这里采用包含信息比较完整的极大似然估计，参数估计结果如下：

表 3.5 ARIMA（1，1，2）模型拟合结果

Series: train ARIMA(1,1,2) with drift
Coefficients: ar1 ma1 ma2 drift -0.9808 1.3749 0.4727 0.2399 s.e. 0.0246 0.0889 0.0922 0.1205
sigma^2 estimated as 0.8713: log likelihood=-164.73 AIC=339.45 AICc=339.97 BIC=353.47
Training set error measures: ME RMSE MAE MPE MAPE MASE ACF1 -0.0034 0.9247 0.7796 - 0.0091 0.7294 0.8865 -0.0446

对上面得到的系数进行显著性检验，所对应的统计量结果如下：

表 3.6 模型系数显著性检验

ar1	ma1	ma2	drift	5%临界值
-39.9177	15.4699	5.1288	1.9903	1.98

从结果可以看出检验统计量的绝对值均大于 95%的临界值，在这个模型中所有系数的估计值都是显著的，因此可以拒绝显著性检验的原假设，所以 ARIMA(1,1，2)模型可以表示为：

$$Y_t = 0.0192Y_{t-1} - 0.9808Y_{t-2} + e_t + 1.3749e_{t-1} + 0.4727e_{t-2} + 0.2399$$

对于一个拟合效果较优的模型来说，模型的残差应该是具有正态性的完全随机序列，auto.arima 得到的结果（即 AIC 准则结果）也会出现不准确的可能。所以需要对残差进行白噪声检验，只要残差序列为纯随机序列，就可以认为残差中已经没有可利用的信息，模型已经建立成功。纯随机性检验结果如下：

表 3.7　残差的白噪声检验

Box-Pierce test					
lag	p+q	df	X-squared	p-value	是否显著
6	3	3	4.992	0.1724	不显著
12	3	9	9.8381	0.3637	不显著
18	3	15	14.934	0.4562	不显著

由上表中检验结果可以看出，对于延迟期为 6、12、18 期的 Liung-Box 统计量的 p 值均大于置信水平α=0.05，没有足够大的理由拒绝原假设，所以认为残差序列为白噪声序列，该模型拟合效果很好。

同时，根据模型可以得到残差的分布图和正态分位数-分位数图(QQ 图)，从 QQ 图的特点可以看出残差基本上服从正态分布。

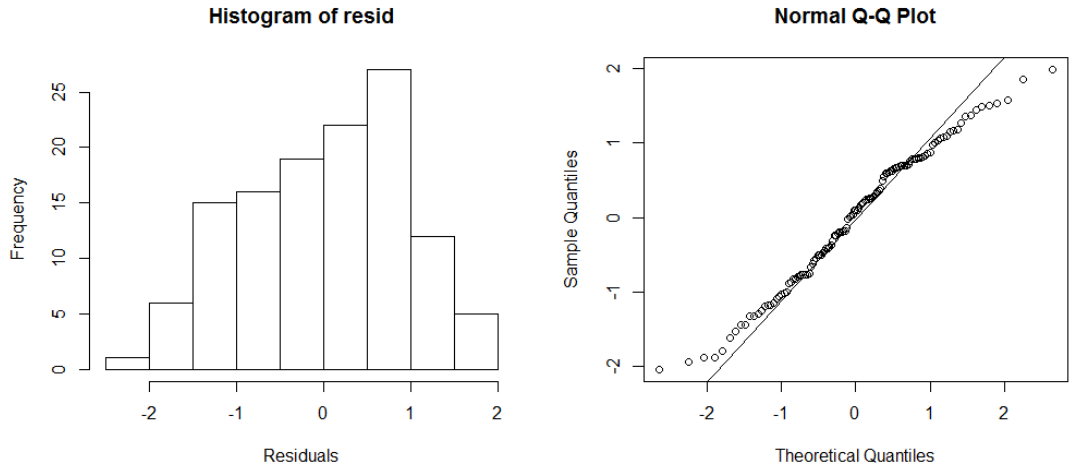

图 3.4　残差分布图和 QQ 图

因为样本数量比较大，不适合使用 Shapiro-Wilk 检验残差的正态性，这里采用 Kolmogorov-Smirnov 单样本检验。检验结果如下图所示，p-value = 0.2772 大于置信水平α=0.05，不能拒绝原假设，我们认为模型的残差符合正态分布，说明模型的拟合效果比较好。

表 3.8　KS 检验

One-sample Kolmogorov-Smirnov test
data: resid D = 0.08957, p-value = 0.2772 alternative hypothesis: two-sided

3.7 预测与分析

根据上面得到的 ARIMA(1,1,2)模型，可以得到拟合值和真实值的趋势，拟合值在真实值附近波动，说明模型的拟合效果较好，但是从图中可以看出，拟合值相对于真实值存在着滞后现象，如下图所示：

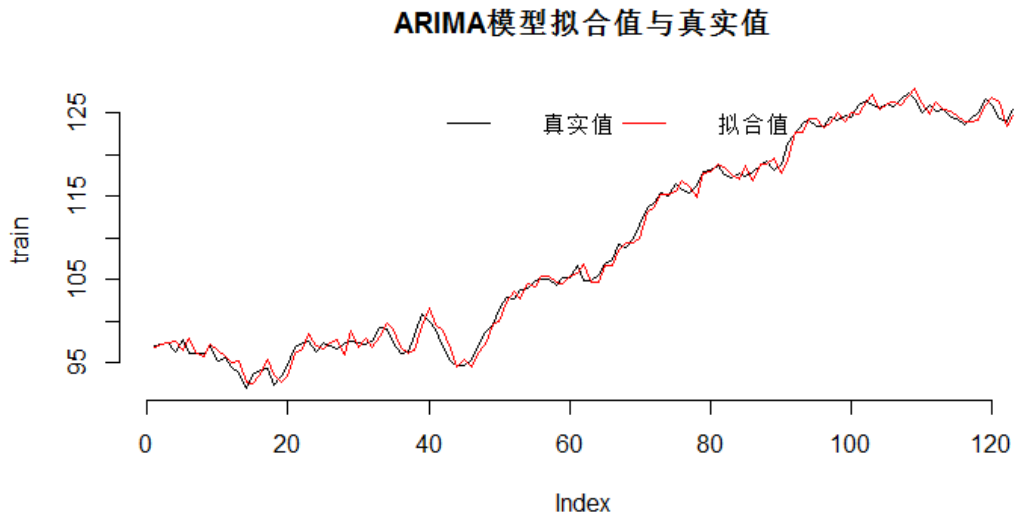

图 3.5　山东省蔬菜价格指数预测值和真实值比较

根据 2005 年-2015 年期间数据的训练样本集得到的预测模型，进行预测 2016 的月均蔬菜价格指数。预测结果均落在 95%的置信区间内，真实值在预测值附近波动，拟合效果较好，如下图：

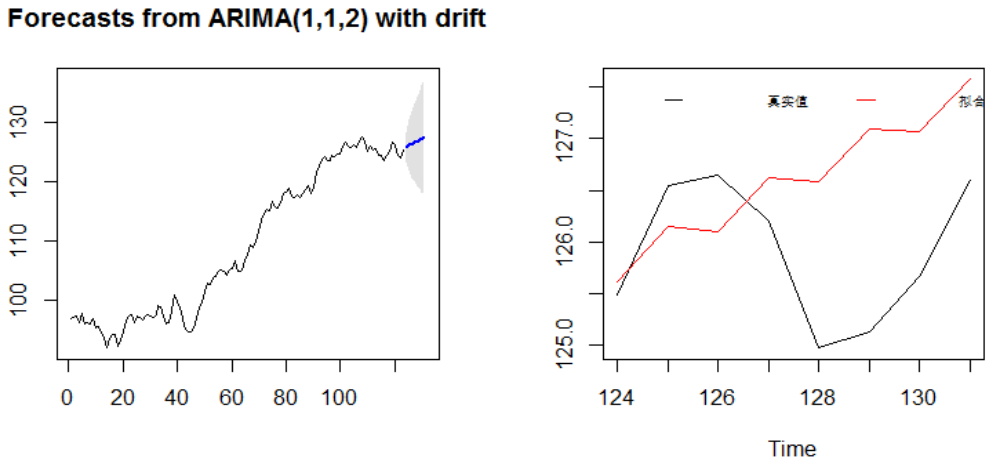

图 3.6　预测结果

跟据上述建模步骤，同理可得到天津市的蔬菜价格指数模型 ARIMA(1,1,1)，从拟合值与真实值的趋势上可以看出，拟合值相对于真实值同样存在着时间滞后性，如下图：

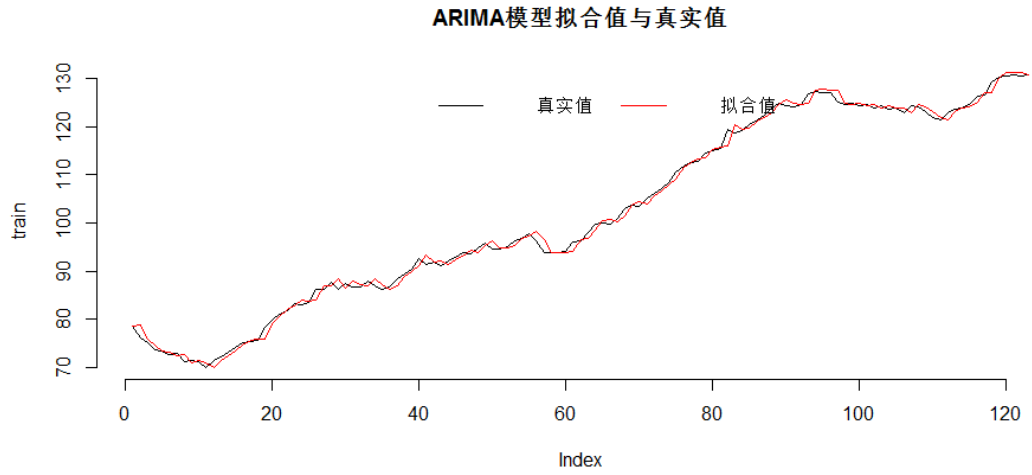

图 3.7　天津市蔬菜价格指数预测值和真实值比较

第四章　蔬菜价格指数 STARMA 模型实例分析

4.1　数据概况

蔬菜价格指数数据既有明显的时间属性，也有隐含的地理位置信息，每个地域的农产品价格彼此存在差异，又相互影响。利用空间统计方法分析农产品价格数据，不仅能够充分利用价格的空间信息，有效地处理数值型的价格数据，挖掘价格指数的分布格局，并且借助 GIS 的空间数据可视化功能，能够直观地展示分析结果，具有独特的优势。例如，如果该地区的价格指数比较高，而周围也出现高值，或是该地区价格指数较低而周围也为低值，称为空间正相关，表明价格指数具有空间集聚的特性；如果价格指数高的地方周围低，价格指数低的地方周围高，则称为空间负相关。

表 4.1　时空相关性表示方法及内涵

区域类型	空间自相关性	邻近地区空间自相关性	区域特征	内涵
类型 1	高	高	热点地区	该地区和其邻近地区价格都易受外部市场价格影响
类型 2	高	低	温点地区	该地区易受外部价格影响，但其邻近地区对外部市场价格变动不敏感
类型 3	低	高	温点地区	该地区不易受外部价格影响，但其邻近地区易受外部价格影响
类型 4	低	低	冷点地区	该地区与其邻近地区价格都不易受外部价格影响

建模所用数据是 27 个省、直辖市、自治区近 10 年来的月均蔬菜价格指数数据，我们通过计算蔬菜价格指数全局和局部的 Moran’s I 指标来度量蔬菜价格指数在各省上的空间模式。

4.2　空间权重矩阵的建立

在构建时空滑动自回归 STARMA 模型之前，需要建立空间权重矩阵，空间权重矩

阵的选择对任何空间统计分析的结果而言是一个重要的决定因素。27 个省、直辖市、自治区的一阶和二阶区域邻接图为：

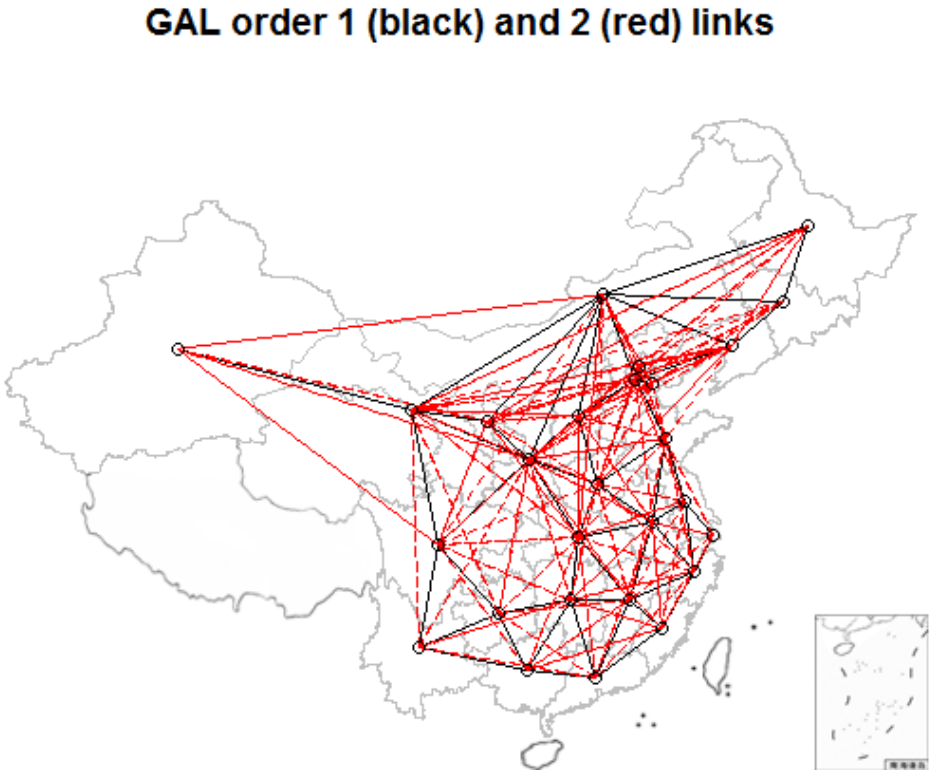

图 4.1　一阶和二阶区域邻接关系

确定了空间目标的位置邻近关系之后，就可以由邻接关系确定空间权重矩阵，一阶空间权重矩阵W^1为：

表 4.2　一阶空间权重矩阵

	黑龙	内蒙	新疆	吉林	辽宁	甘肃	河北	北京	山西	天津	陕西	宁夏	山东	河南	江苏	安徽	四川	湖北	上海	浙江	湖南	江西	云南	贵州	福建	广西	广东
黑龙	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
内蒙	1	0	0	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
新疆	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
吉林	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
辽宁	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
甘肃	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
河北	0	1	0	0	1	0	0	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
北京	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
山西	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
天津	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
陕西	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0
宁夏	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
山东	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
河南	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
江苏	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
安徽	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	1	0	1	0	0	0	0	0
四川	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
湖北	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0
上海	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
浙江	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0
湖南	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	1	1
江西	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	0	0	0	1	0	1
云南	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0
贵州	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	1	0
福建	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1
广西	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1
广东	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0

同理，根据定义可以得到二阶空间权重矩阵W^2为：

表 4.3　二阶空间权重矩阵

	黑龙	内蒙	新疆	吉林	辽宁	甘肃	河北	北京	山西	天津	陕西	宁夏	山东	河南	江苏	安徽	四川	湖北	上海	浙江	湖南	江西	云南	贵州	福建	广西	广东
黑龙	0	0	0	0	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
内蒙	0	0	1	0	0	0	0	1	0	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0
新疆	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
吉林	0	0	0	0	0	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
辽宁	1	0	0	0	0	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
甘肃	1	0	0	1	1	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0
河北	1	0	0	1	0	1	0	0	0	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
北京	0	1	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
山西	1	0	0	1	1	1	0	1	0	1	0	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0
天津	0	1	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
陕西	1	0	1	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	1	1	1	0	0	0
宁夏	1	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0
山东	0	1	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
河南	0	1	0	0	1	1	0	1	0	1	0	1	0	0	1	0	1	0	0	1	1	1	0	0	0	0	0
江苏	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0
安徽	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1
四川	0	1	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0
湖北	0	1	0	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0	0	1	0	0	0	1	1	1	1
上海	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0
浙江	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0	1
湖南	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	0	0	1	0	1	0	0
江西	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	1	0	0	0	0	1	0	1	0
云南	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
贵州	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1
福建	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	1	0
广西	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0
广东	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1	0	0	0

此外，0 阶空间权重矩阵是单位矩阵$W^0 = I_{27\times27}$。

4.3　数据的平稳性检验

在建立模型之前，需要对模型进行平稳性检验，时空数据的自相关系数结果和自相关函数图如下：

表 4.4　时空自相关函数(ST-ACF)

延迟期	slag0	slag1	slag2
tlag1	0.9994	0.9825	0.9838
tlag2	0.9989	0.9822	0.9835
tlag3	0.9984	0.9819	0.9832
tlag4	0.9981	0.9817	0.9831
tlag5	0.9977	0.9816	0.9830
tlag6	0.9975	0.9815	0.9829
tlag7	0.9972	0.9815	0.9829
tlag8	0.9971	0.9815	0.9829
tlag9	0.9969	0.9815	0.9830
Tlag10	0.9968	0.9816	0.9831

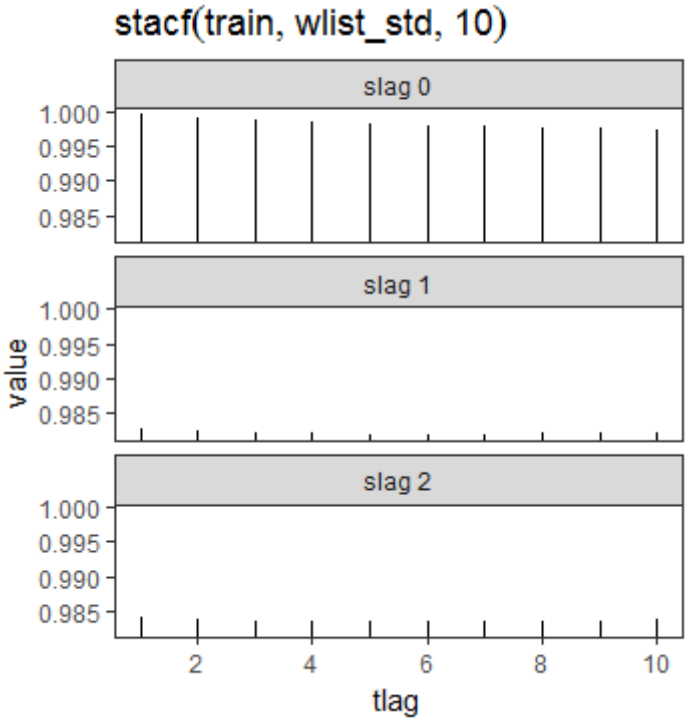

图 4.2　空间自相关系数

从图中可以看出自相关系数虽然在所有空间延迟上时间延迟逐渐递减，但并没有在 0 的附近波动，所以数据可能在时间上存在非平稳性。根据时空数据初步探索过程中了解到的趋势特征，需要对非平稳的蔬菜价格指数时空数据进行平稳化处理，一般情况下，一阶差分可以去除线性趋势，二阶差分可以去除二次曲线趋势。经过一次差分之后，自相关函数图如下：

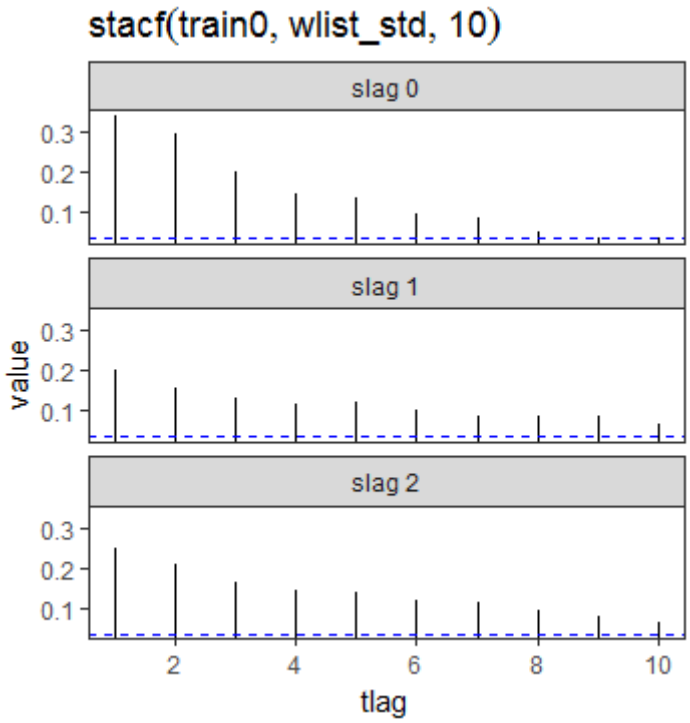

图 4.3　一阶差分空间自相关系数

从差分后数据的时空数据自相关函数可以看出，在一定时间延迟后，自相关函数截尾，说明数据呈现出良好的平稳性，经检验差分后的时空数据为平稳的数据。

对差分后的数据进行相关性检验，p 值远远小于显著性水平α，所以拒绝原假设的不相关性，可以进一步进行建立模型分析，检验结果如下：

表 4.5　相关性检验

```
        Multivariate Box-Pierce Non Correlation Test
        ----------------------------------------


  X.squared df p.value
1  2430.688 63       0

Decision: Non Correlation Hypothesis should be rejected.
```

4.4　模型识别

在实际计算中，空间延迟期最高为 2 即可满足模型需求，根据公式可以计算出样本的自相关函数和偏自相关函数如下表：

表 4.6　时空自相关函数(ST-ACF)

延迟期	slag0	slag1	slag2
tlag1	0.3401	0.1974	0.2459
tlag2	0.2968	0.1547	0.2070
tlag3	0.2037	0.1323	0.1650
tlag4	0.1487	0.1138	0.1449
tlag5	0.1365	0.1210	0.1395
tlag6	0.0946	0.1004	0.1178
tlag7	0.0858	0.0846	0.1128
tlag8	0.0492	0.0862	0.0936
tlag9	0.0370	0.0834	0.0801
tlag10	0.0375	0.0656	0.0650
tlag11	0.0347	0.0831	0.0846
tlag12	0.0390	0.0648	0.0692

表 4.7　时空偏相关函数(ST-PACF)

延迟期	slag0	slag1	slag2
tlag1	0.3401	0.1875	0.2438
tlag2	0.1811	-0.0021	0.0427
tlag3	0.0489	-0.0018	-0.0223
tlag4	0.0154	0.0157	0.0175
tlag5	0.0363	0.0518	0.0427
tlag6	0.0067	0.0015	-0.0064
tlag7	0.0067	-0.0156	0.0254
tlag8	0.0198	0.0311	0.0015
tlag9	0.0131	0.0358	-0.0028
tlag10	0.0024	-0.0082	-0.0189
tlag11	0.0085	0.0581	0.0707
tlag12	0.0055	-0.0083	-0.0171

同时可以得到自相关系数和偏相关系数图：

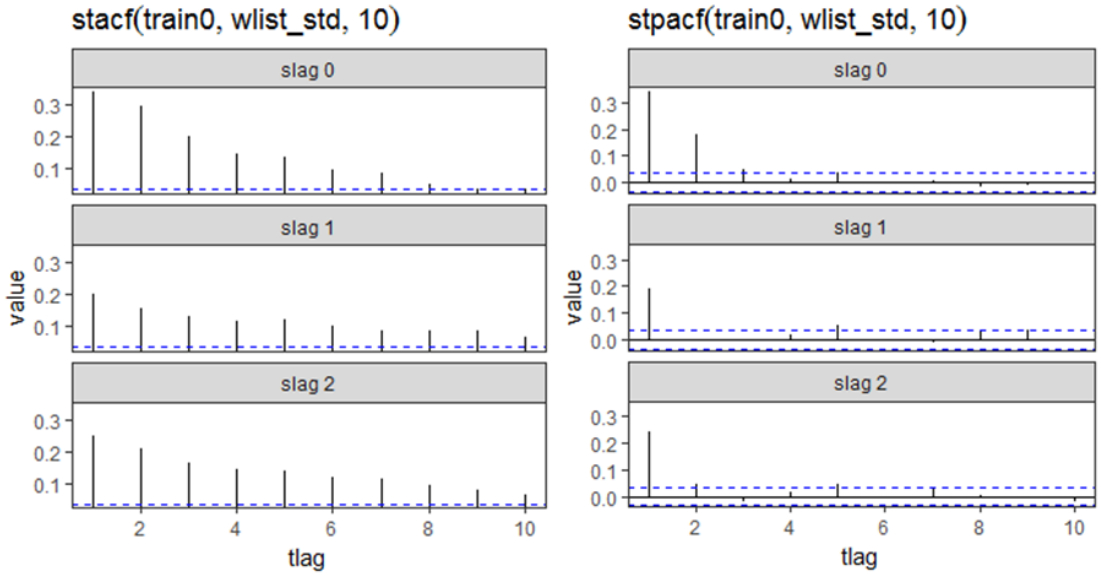

图 4.4　空间自相关系数和空间偏相关系数

从上面数据和图形可以看出，时空自相关函数和时空偏相关函数值在所有空间延迟期上呈现出拖尾的趋势，为STARMA(p,q)模型，不能根据 acf 和 pacf 判断出模型的阶数，需要根据 BIC 信息准则进行判断，通过拟合多个模型计算 BIC 信息值得到，当模型为 STARMA(1，1)时 BIC 信息最小，模型的方程为：

$$Z(t)=\sum_{k=1}^{1}\sum_{l=0}^{2}\varphi_{kl}W^{(l)}Z(t-k)-\sum_{k=1}^{1}\sum_{l=0}^{0}\theta_{kl}W^{(l)}\varepsilon(t-k)+\varepsilon(t)$$

$$= (\varphi_{10}W^0 + \varphi_{11}W^1 + \varphi_{12}W^2)Z(t-1) + \theta_{10}W^0\varepsilon(t-1) + \varepsilon(t)$$

4.5 参数估计

对 STARMA(1,1)模型进行参数估计，给出以下结果：

表 4.8 STARMA 模型拟合结果

```
Call:
starma.default(data = train0, wlist = wlist_std, ar = arlist[[4]],
               ma = malist[[2]])

        Estimate Std..Error  t.value    p.value
phi10    0.681192   0.047210 14.4290 < 2.2e-16 ***
phi11    0.051428   0.028279  1.8186 0.0690646 .
phi12    0.134862   0.036962  3.6487 0.0002675 ***
theta10 -0.459021   0.049487 -9.2755 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

从上表中可以看出，系数的估计值的p.value $< \alpha = 0.1$，具有良好的显著性。故该模型为自回归项的空间延迟为 2、滑动平均项的空间延迟为 0 的 STARMA(1,1)模型，那么 STARMA(1,1)模型可以表示为：

$$\nabla Y(t) = (\varphi_{10}W^0 + \varphi_{11}W^1 + \varphi_{12}W^2)\nabla Y(t-1) + \theta_{10}W^0\varepsilon(t-1) + \varepsilon(t)$$

$$\begin{aligned} Y(t) = {} & (1 + 0.6812W^0 + 0.0514W^1 + 0.1349W^2)Y(t-1) \\ & - (0.6812W^0 + 0.0514W^1 + 0.1349W^2)Y(t-2) - 0.459W^0\varepsilon(t-1) \\ & + \varepsilon(t) \end{aligned}$$

4.6 模型检验

在对时空数据进行拟合之后，要进行模型检验。对于系数的显著性检验从模型估计部分的结果可以看出，参数的 p 值均小于显著性水平$\alpha = 0.001$，具有高度的显著性。主要是对模型残差进行检验，即检验残差序列是否为随机白噪声序列，从而来判断时空模型是否能够准确的表达时空数据。通过时空自相关函数来判断，取空间延迟为 0、1、2，计算残差序列的自相关系数如下：

表 4.9　残差时空自相关函数(ST-ACF)

延迟期	slag0	slag1	slag2
tlag1	-0.00632	0.01465	0.01573
tlag2	0.04430	-0.03406	-0.03003
tlag3	0.00692	-0.01159	-0.02053
tlag4	-0.00618	-0.01063	-0.00402
tlag5	0.02638	0.01841	0.00862
tlag6	-0.01431	-0.00012	-0.01379
tlag7	0.01704	-0.01653	0.00480
tlag8	-0.02578	-0.00404	-0.01114
tlag9	-0.01642	0.00589	-0.00584
tlag10	0.00828	0.01836	0.00615
tlag11	-0.01495	0.01801	0.00761
tlag12	0.00593	0.00480	0.00411
tlag13	-0.01624	-0.00373	-0.00678
tlag14	0.00031	-0.00220	0.04134
tlag15	0.02638	0.03171	-0.00042

同时可以得到残差的时空自相关函数图如下：

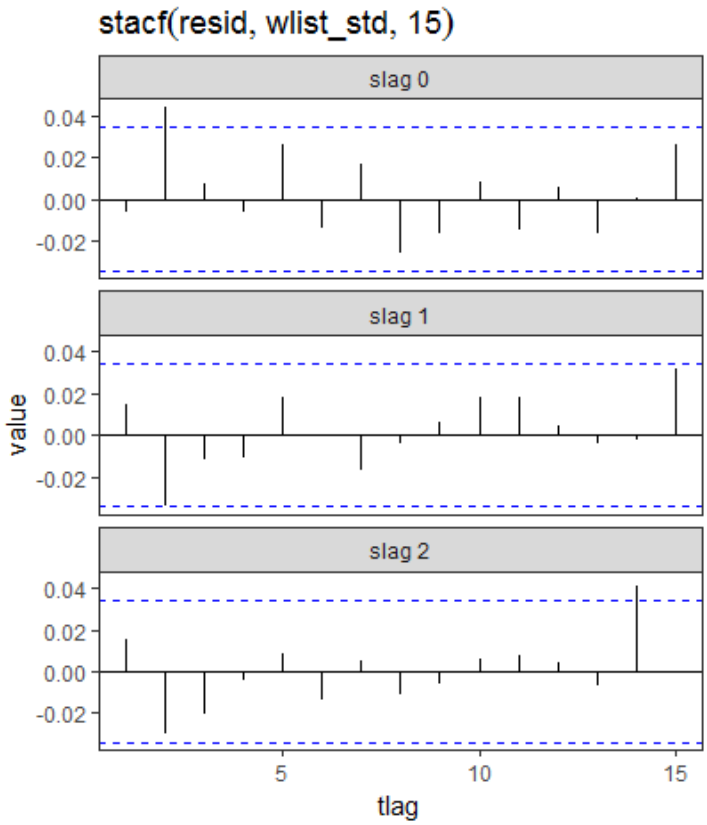

图 4.5　残差空间自相关系数

从残差的时空自相关函数表和图形中可以看出，残差在空间延迟和时间延迟上的自相关系数都接近 0，为平稳的时空序列。经过 Box-Pierce 检验，结果如下：

表 4.10　残差相关性检验

```
        Multivariate Box-Pierce Non Correlation Test
        ----------------------------------------

  X.squared  df    p.value
1  59.76518 61 0.5207568

Decision: Can't reject Non Correlation Hypothesis.
```

从检验结果中可以看出，残差的相关性检验 p 值大于显著性水平α，没有足够的理由拒绝原假设，接受原假设的完全随机性，认为模型能够完全解释时空数据。

4.7　预测结果与评估

通过上面建模过程得到 STARMA 模型，从模型的结果来看，蔬菜价格指数数据在空间上确实存在着相关性，相邻地区的价格指数会相互影响。利用模型对各省、直辖市、自治区的蔬菜价格指数进行时空预测，以 2016 年几个月份的月均蔬菜价格指数作为验证数据，其真实值与预测值的结果如下：

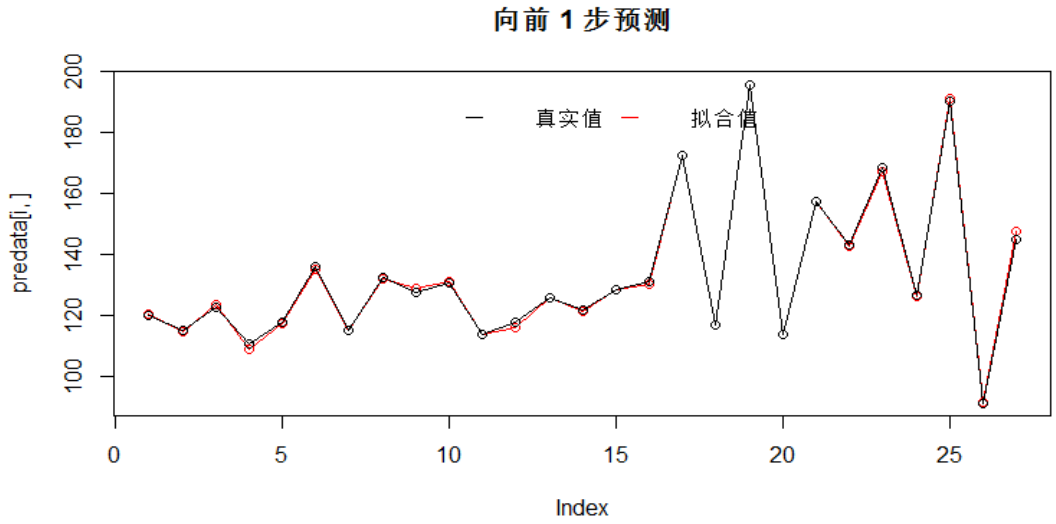

图 4.6　STARMA 模型向前 1 步预测

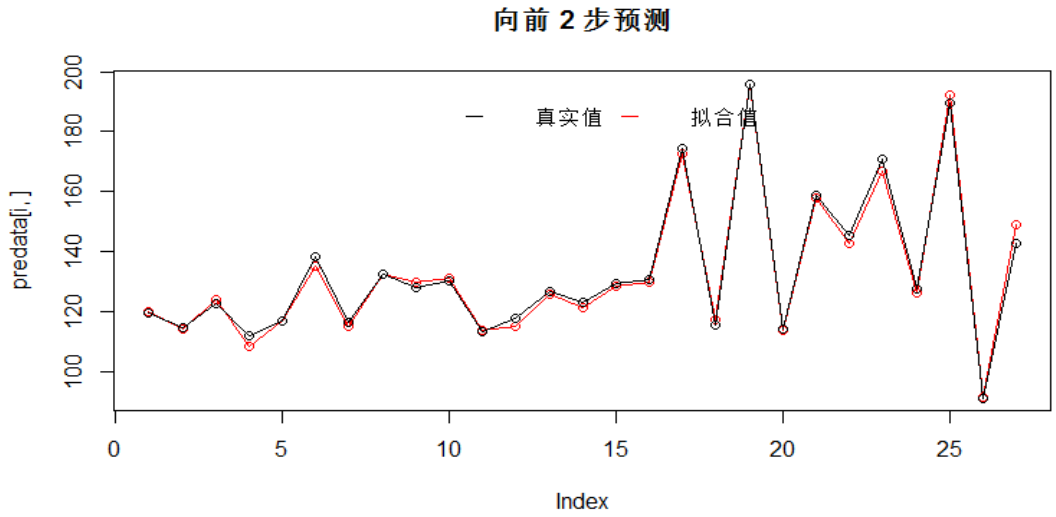

图 4.7　STARMA 模型向前 2 步预测

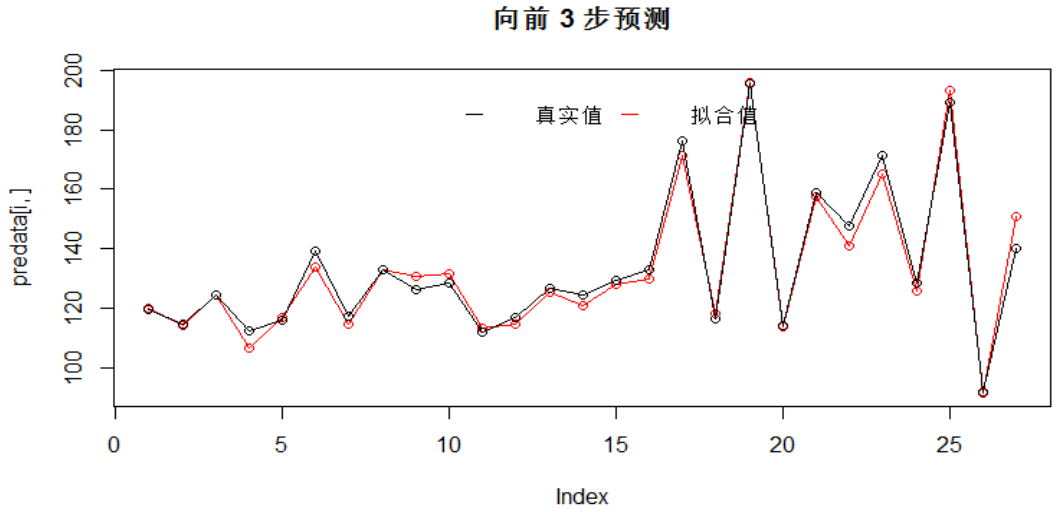

图 4.8　STARMA 模型向前 3 步预测

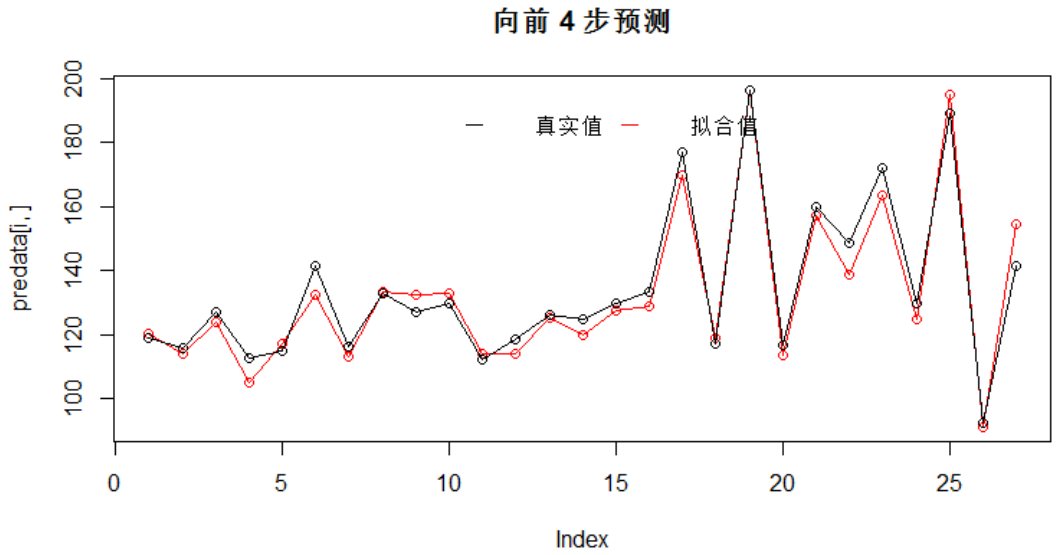

图 4.9　STARMA 模型向前 4 步预测

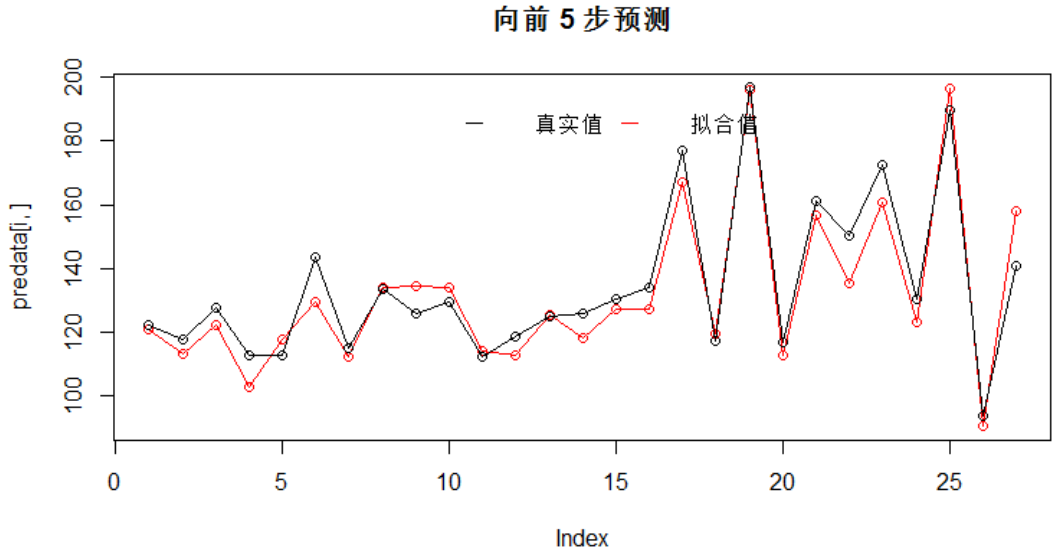

图 4.10　STARMA 模型向前 5 步预测

从上面 5 个月份对于各省、直辖市、自治区的拟合值和真实值对比来看，STARMA 模型在短期内预测的拟合值与实际数据较为接近，能够很好地达到预测效果，及时预测出波动状况，而且很好的解释了空间变异性，说明 STARMA 时空预测模型有较好的拟合效果，对于蔬菜价格指数的预测在时间维度上引入空间性质，对模型预测效果有很大的帮助。

以山东省的预测效果为例，将 ARIMA 模型和 STARMA 模型对山东省蔬菜价格指数的预测结果与真实值进行对比，结果展示如下图：

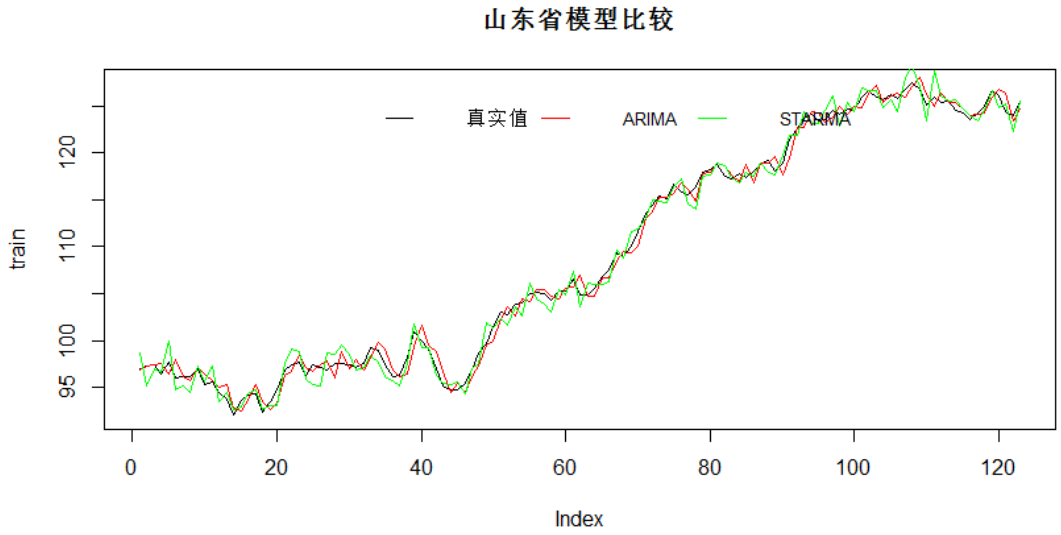

图 4.11　ARIMA、STARMA 拟合值与真实值比较

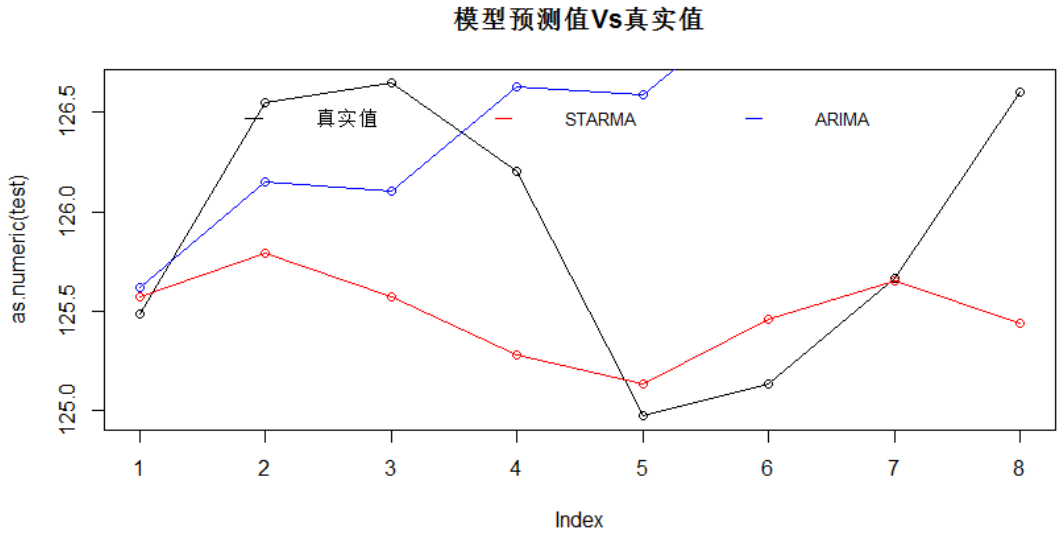

图 4.12　ARIMA、STARMA 预测值与真实值比较

在 ARIMA 模型和 STARMA 模型得到的预测结果对比来看，两种预测各有优势，但是 ARIMA 模型在数据的波动状况中表现出滞后性，而 STARMA 模型得到的数据能够很好的预测出价格指数的波动状况，STRAMA 模型也表明了蔬菜价格指数在空间地理位置上确实存在着相互影响的作用，该模型对解决农业领域补充了新方法，能够很好的解释空间变异性。

第五章　总结

农产品的价格指数和各类农产品的价格都有密切关系，比如蔬菜价格指数就是由各类蔬菜农产品价格得到，代表着整个分类的市场状况，通过分析蔬菜价格指数形势，可以有效了解我国和各个省市地区蔬菜批发价格的整体水平，记录全国蔬菜农产品批发价格变化动态，揭示蔬菜批发价格变化规律，对农产品市场的宏观调控有指导意义。本文主要讨论了蔬菜价格指数预测的方法，运用了时间序列 ARIMA 模型和时空序列 STARMA 模型对蔬菜价格指数进行预测分析，样本数据驱动下的模型能够很好的解释模型之间的优劣，通过不同模型的预测值与实际值的对比，ARIMA 模型和 STARMA 模型都能够很好的对蔬菜价格指数有很好的预测效果。本文从理论分析和实证分析的过程得出，其中 STARMA 模型对蔬菜价格指数的预测得到的误差相比较更小、模型的预测精确度更高，这对于涉及到空间地理信息的时间序列问题来说，提供了更加可行、有效的解决方案，对农业领域的问题研究也不失为一个有效的方法。

在本文使用的预测模型仍需要进一步优化，以本文研究内容为例。首先本文涉及的 ARIMA 模型对于解决非平稳的时间序列，可以充分提取样本中的信息，有效的提高拟合精度，对于仅仅是时间维度上的预测问题，无论是从理论上还是实践上都是很具有实用性的模型，使用 ARIMA 模型时也要考虑到季节性等问题。其次，对于涉及时间和空间地理信息的研究对象，使用时空序列模型进行预测，虽然能够解释空间变异性，但是对于空间地理面积过大的数据往往会使空间相关性减弱，不能有效的解释出空间变异性，会影响预测结果。所以，对于小领域的样本数据能给出更有效的预测效果，过于广泛的空间地理位置或许不能达到理想效果。在本文中，如果数据为某省份的地级市或者某个城市的区县农产品价格数据，模型的预测效果会更理想。

对于农产品价格预测方面的几点建议：

（1）政府的宏观调控。从模型的建立过程中可以看出，蔬菜价格指数的变化并非平稳的序列，而且呈现着逐渐增长的趋势。这和我国居民消费水平的提升和 GDP 的增长息息相关，当然也和农作物的种植情况密不可分。农产品市场价格的快速增长，对我国农也的发展和农产品市场的平衡会造成一定影响，根据模型预测的结果，政府应当予以一定的干预保证农产品的生产和市场平衡，控制农产品市场价格的快速增长和下滑。

（2）稳定物价水平。物价水平是价格是衡量一个国家居民消费水平的重要指标，控制物价水平同我国的发展相适应，这有助于稳定农产品的种植情况的相对平衡，以免出现供不应求或供过于求的状况。

（3）考虑空间相关性的农产品价格预测。对于很多情形下的农产品价格预测，不仅要考虑经济发展水平、农产品种植情况等因素，也可以从空间地理区域方面考虑。通过本文的研究，单一的时间维度上的预测研究可能会产生一些时滞性的结果，加上空间维度上的相关性能够及时的给出预测变化趋势，这对于农业领域的预测研究来说是一个很大的帮助。

参考文献

[1] Jarrett F G.SHORT TERM FORECASTING OF AUSTRALIAN WOOL PRICES*[J]. Australian Economic Papers. l965.4(1-2):93-102.

[2] U.S. Census Bureau, 2002,X-12-ARIMA Reference Manual[M],Version 0.2.10.

[3] Bao Rong Chang. A study of non-periodic short-term random walk forecasting based on RBFNN，ARMA，or SVR-GM approach. In IEEE Conf.on Systems，Man and Cybernetics,2003.254-259.

[4] 唐江桥，雷娜.中国鸡蛋价格波动预警研究[J].西部论坛，2011,21(06):44-49.

[5] Coskun H. Improving Artificial Neural Networks' Performance in Seasonal Time Series Forecasting [J]. Information Sciences, 2008, 178(23): 4550-4559.

[6] 马孝斌，王婷，董霞等.向量自回归法在生猪价格预测中的应用[J].中国畜牧杂志，2007,43(23):4-6.

[7] 李干琼. SV 因子分析框架下的农产品市场短期预测[D].北京：中国农业科学院，2012.

[8] 刘峰，王儒敬，李传席.ARIMA 模型在农产品价格预测中的应用[J].计算机工程与应用，2009,45(25)：238-239.

[9] 傅如南，林丕源等. 基于 ARIMA 的肉鸡价格预测建模与应用[J].中国畜牧杂志，2008,44(20):17-21.

[10] 王勇，张浩.小麦期货价格预测的马尔可夫模型[J].安徽：安徽农业科学，2008,36(05):1721.

[11] Cai Z W. Regression quantiles for time series [J].Econometric Theory, 2002,(1):169-192.

[12] 王舒鸿.灰色预测模型在鸡蛋价格预测中的应用[J].中国禽业导刊，2008,25(15):48-50.

[13] Cliff A D, Ord J K. Space-time modelling with an application to regional forecasting [J].Transactions of the Institute of British Geographers, 1975: 119-128.

[14] Bilonick R A. The space-time distribution of sulfate deposition in the northeastern United States [J]. Atmospheric Environment (1967), 1985, 19(11): 1829-1845.

[15] Cressie N, Wikle C K. Statistics for Spatio-Temporal Data [M]. John Wiley & Sons, 2011.

[16] 李序颖，顾岚. 空间自回归模型及其估计[J].统计研究，2004,(6):48-51.

[17] 张英.时空数据模型的建模研究与应用[D].青岛：青岛大学.2007.

[18] 李晶晶.时空数据挖掘在环境保护中的应用研究[D].长沙：中南大学.2008.

[19] 王佳璆.时空序列数据分析和建模[D].广州：中山大学.2008.

[20] 林艳.基于地统计学与 GIS 的土壤重金属污染评价与预测[D].长沙：中山大学.2009.

[21] 王尚北.基于多噪声监测点机场噪声时空序列预测模型研究[D].南京：南京航空航天大学.2014.

[22] 许克平.案事件时空数据挖掘研究[D].福州：福州大学.2014.

[23] 刘权芳.两类时空数据模型及其应用研究[D].西安：长安大学，2015.

[24] 魏媛. 基于时间与空间关联分析的城市供水管网水质异常检测方法研究[D].宁波：浙江大学.2015.

[25] Jonathan D Cryer,Kung-Sik Chan(潘红宇 等[译]).时间序列分析及应用：R 语言[M].北京：机械工业出版社，2011.

[26] Roger S B, Edzer J P, Virgilio G R (徐爱萍，舒红[译]). Applied Spatial Data Analysis with [M].北京：清华大学出版社.2013.

[27] 杨中庆.基于 R 语言的空间统计分析研究与应用[D].广州：暨南大学.2006.

[28] 薛毅.统计建模与 R 软件[M].北京：清华大学出版社，2007.

[29] 聂巧平，张晓峒. ADF 单位根检验中联合检验 F 统计量研究[J].统计研究，2007,24(2):73-80.

[30] 王燕. 应用时间序列分析[M]. 北京：中国人民大学出版社.2013.

致谢

时光荏苒，岁月如梭，硕士的求学时光转眼就要结束了。在这两年期间，不仅提升了专业水平、专业技能，也拓展了知识面、学到了自己感兴趣的东西，还认识了很多朋友，这是一个收获颇多的两年。

非常感谢导师徐兴忠教授，不论是在学习上还是生活上对我们都是关怀备至、悉心指导，老师严谨的教学方法和实事求是、一丝不苟的学术精神都深深影响着我。尤其是在做毕业设计期间，帮助我拓展思路、明确研究方向，将相关知识、经验和方法传授给我，让我能够独立的完成论文的研究。在此向徐老师致以深深的感谢和崇高的敬意。

在这临近毕业之际，还要衷心感谢这两年间所有的授课老师，特别是杨国孝老师、黄宝胜老师，非常感谢各位老师的无私奉献，严谨的治学态度以及对每一个学生孜孜不倦的教诲，对我在以后的学习和工作中都受益匪浅。

感谢有缘能够在一起学习的每一位 2015 级应用统计的同学，尤其是同在一个实验室的兄弟姐妹们，谢谢你们给了我很多帮助和关心，很高兴能够结识你们，我会珍惜我们曾经在一起的每一天。

感谢博士学长对我毕业论文的指导和帮助。

感谢我最亲爱的室友们和最交心的朋友们，向我们最珍贵的友谊致敬。

衷心感谢在百忙之中抽出宝贵时间评阅本论文和参加答辩的各位专家和学者，对相关内容提出宝贵意见。

特别感谢我的父母，对我一如既往的支持和爱护，让我能够不畏困难勇敢前进。

最后，祝愿老师们身体健康、事事顺利！祝愿同学们在今后的道路上实现自己的梦想！祝愿我的亲人、朋友们永远健康、幸福！