Operations Research and Fuzziology 运筹与模糊学, 2015, 5, 30-37
Published Online May 2015 in Hans. http://www.hanspub.org/journal/orf
http://dx.doi.org/10.12677/orf.2015.52005

Hans汉斯

Port Container Throughput Forecasting Based on the Multiplicative Seasonal ARIMA Model

Lili Tao[1], Yan Wang[2]

[1]School of Science, Dalian Ocean University, Dalian Liaoning
[2]Foreign Language School, Dalian Ocean University, Dalian Liaoning
Email: taolili1986@163.com, ivy1900@sina.com

Received: Apr. 20th, 2015; accepted: May 7th, 2015; published: May 19th, 2015

Abstract

Based on the theoretical research of the time series analysis, this paper systematically analyzes the changes rules of the monthly data of container throughput of Shanghai Port from 2002 to 2009 by using MATLAB software. The result shows that the multiplicative seasonal $\mathrm{ARIMA}(0,1,1)\times(0,1,1)_{12}$ model has a high forecasting precision, a reasonable forecasting result and a broad application prospect.

Keywords

Time Series Analysis, Multiplicative Seasonal ARIMA Model, Container Handling Capacity, Forecast

基于ARIMA乘积季节模型的港口集装箱吞吐量预测

陶丽丽[1]，王　艳[2]

[1]大连海洋大学理学院，辽宁 大连
[2]大连海洋大学外国语学院，辽宁 大连
Email: taolili1986@163.com, ivy1900@sina.com

收稿日期：2015年4月20日；录用日期：2015年5月7日；发布日期：2015年5月19日

摘　要

在对时间序列分析理论研究基础上，利用MATLAB软件编写所有算法的程序系统地分析港口集装箱吞吐量月度数据的变化规律，建立的ARIMA乘积季节模型能充分反映港口集装箱吞吐量的时间序列变化规律。以上海港2002~2009年集装箱吞吐量为例，应用MATLAB软件建立了 ARIMA$(0,1,1)\times(0,1,1)_{12}$ 乘积季节模型，结果表明该乘积季节模型的预测精度较高，预测结果更加合理，有着广泛的应用前景。

关键词

时间序列分析，ARIMA乘积季节模型，集装箱吞吐量，预测

1. 引言

中国地处太平洋西海岸，近年来随着经济的高速发展，航运业取得了巨大的发展。于是处于航运业发展核心的港口的发展越显其重要性。其中集装箱吞吐量是衡量一个港口能力的重要指标，是港口发展战略研究的重要内容，对于确定港口发展方向、扩建港口，新建码头，合理制定港口作业计划和港口基本设施规划，提高港口的通过能力和运营效率都具有十分重要的意义，而这些正是保证港口服务能力供给的基础。因此，准确预测港口吞吐量是协调港口服务能力的供给与区域对港口服务需求之间的桥梁，科学合理的预测港口吞吐量对于港口服务供应链内部和外部协同都具有重要的意义。

目前在港口吞吐量预测领域的国内外相关研究中，预测方法比较全面，包括回归分析[1]，支持向量机[2]，神经网络[3]等模型。但现有研究中的主要问题是：预测时段绝大都是关于港口年度吞吐量的预测，而很少研究港口月度吞吐量的变动，这对港口在战术和运作层面上的支持是远远不够的。从月度角度分析，在港口运输中，年末由于一方面港口为了完成全年生产指标都会努力提高港口吞吐量，同时由于圣诞元旦春节假期的临近，对货物的需求旺盛也使得港口吞吐量增加。而在以年为统计单位的港口吞吐量变化时，周期趋势并不显著。年度数据一般只含有增长性和随机因素，而月度数据还要包含周期性规律，这就使得港口年度数据和月度数据所反映的内涵差别较大。ARIMA 乘积季节模型对港口月度吞吐量的变化进行研究，更好地反映了吞吐量的周期性规律，预测精度更高，结果也更加可信。

本文以上海港集装箱吞吐量为例，对其 8 年的历史数据进行系统地分析，建立了 ARIMA$(0,1,1)\times(0,1,1)_{12}$ 乘积季节模型，结果表明该模型的预测精度较高，可采用该模型对上海港 2011 年的集装箱吞吐量进行预测。

2. 知识准备

2.1. 时间序列分析

时间序列分析(Time series analysis)是一种动态数据处理的统计方法。该方法是基于随机过程理论和数理统计学方法，研究随机数据序列所遵从的统计规律，以用于解决实际问题。其基本原理：一是承认事物发展的延续性，应用过去数据，就能推测事物的发展趋势；二是考虑到事物发展的随机性，任何事物发展都可能受偶然因素影响，为此要利用统计分析中加权平均法对历史数据进行处理。根据时间序列分析，可以对未来进行预测，其预测一般反映了趋势性，周期性以及随机性三种实际变化规律。纵观时间序列分析方法的发展历史可将其分为频域分析方法和时域分析方法两大类，本文主要研究的是时域分析方法[4] [5]。

2.2. 数据的预处理

拿到一个观察值序列后，首先应对其平稳性和纯随机性进行检验。通过平稳性检验，序列又可分为平稳序列和非平稳序列两大类。若为平稳序列，还需进一步对其的纯随机性进行检验。一个序列经预处理被识别为平稳非白噪声序列，则说明了该序列是一个蕴含着相关信息的平稳序列，需建立一个线性模型拟合该序列发展，借此提取序列中有用信息。到目前为止，差分方法被认为是一种简便，有效的确定信息的提取的方法，是由 Box 和 Jenkins 提出并用大量的案例证明了。差分运算具有很强大的确定性信息提取能力，许多非平稳序列差分后会显示出平稳序列的性质，这时称该非平稳序列为差分平稳序列。实际生活中，绝大部分序列是非平稳的，当平稳性检验分析结果为非平稳序列，则还需通过有效的手段提取序列中所蕴含的确定性信息，将其化为平稳序列。

2.3. ARIMA 乘积季节模型

时间序列模型有许多种类型，其中有三种是最经典和最重要的，他们是依靠原始时间序列的线性关系[5]-[7]，分别是 AR(自回归)模型、MA(移动平均)模型[8]和 ARIMA (非平稳自回归移动平均)模型。$\mathrm{ARIMA}(p,d,q)$ 模型适用于拟合经差分运算后具有短期相关性的序列，但当经差分运算后的序列还具有季节效应，并且季节效应本身仍具有相关性时，则其季节相关性可采用以周期步长为单位的 $\mathrm{ARIM}(P,Q)$ 模型提取[9]-[11]，这样的序列则适合采用 $\mathrm{ARIMA}(p,d,q)*(P,D,Q)\mathrm{s}$ 乘积季节模型进行拟合，它的完整结构为[12]：

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_t=\theta_q(B)\Theta_Q(B^s)a_t$$

式中

$$\Phi_P(B^s)=1-\Phi_1(B^s)-\Phi_2(B^{2s})-\cdots-\Phi_P(B^{Ps})$$

$$\Theta_Q(B^s)=1+\Theta_1(B^s)+\Theta_2(B^{2s})+\cdots+\Theta_Q(B^{Qs})$$

2.4. 预测

所谓预测就是要利用序列已观测到的样本值对序列在未来某个时刻的取值进行估计。目前对平稳序列最常用的预测方法是线性最小方差预测。线性是指预测值为观测值序列的线性函数，最小方差是使预测方差达到最小。

3. 上海港集装箱吞吐量预测

本文研究 2002 年 1 月至 2009 年 12 月上海港的集装箱吞吐量月度数据，如表 1 所示。

3.1. 数据的分析

数据的分析包括描述性，平稳性和相关性分析。首先画出上海港集装箱吞吐量的时序图，如图 1 所示。

由图 1 可看出，上海港集装箱吞吐量逐年递增，序列具有长期递增趋势，明显属于非平稳序列，并且具有平稳的以年为周期的季节性波动，因此应对原序列作一步差分，提取线性递增趋势，紧接着还应对一阶差分后的序列再进行 12 步同周期差分，提起季节性波动信息。根据图 2 可知，经一阶十二步差分后，序列基本平稳。

3.2. 模型构建

原序列在取一阶十二步差分之后，非零的自相关只是在延迟为 1，11，12 和 13 处。于是笔者确定了

Table 1. Natural Logarthms of monthly container handling capacities (measured in thousands TEU) in Shanghai Port
表 1. 取对数后的上海港集装箱吞吐量(万吨)月度数据

	1 月	2 月	3 月	4 月	5 月	6 月
2002	4.0916	3.9232	4.1807	4.2151	4.2515	4.2625
2003	4.4536	4.2151	4.5114	4.5174	4.4898	4.5685
2004	4.6131	4.5444	4.7449	4.7875	4.7908	4.8339
2005	4.9579	4.7131	4.9663	5.0265	5.0193	5.0421
2006	5.0937	4.8331	5.1405	5.1818	5.2051	5.2396
2007	5.3215	5.1985	5.3122	5.3936	5.3982	5.4072
2008	5.4621	5.2241	5.4790	5.4819	5.4702	5.4943
2009	5.2476	5.0291	5.3863	5.2771	5.3428	5.3033
	7 月	8 月	9 月	10 月	11 月	12 月
2002	4.2750	4.3666	4.3906	4.3231	4.4118	4.4694
2003	4.5788	4.6112	4.6161	4.6299	4.6405	4.6161
2004	4.8315	4.8402	4.8744	4.8767	4.8888	4.8881
2005	5.0770	5.0789	5.0531	5.0676	5.0795	5.0490
2006	5.2601	5.2745	5.3003	5.2523	5.2559	5.2617
2007	5.4250	5.4407	5.4165	5.4179	5.4157	5.4385
2008	5.4604	5.5440	5.4613	5.4819	5.4289	5.4112
2009	5.3683	5.3882	5.4040	5.3808	5.3950	5.4798

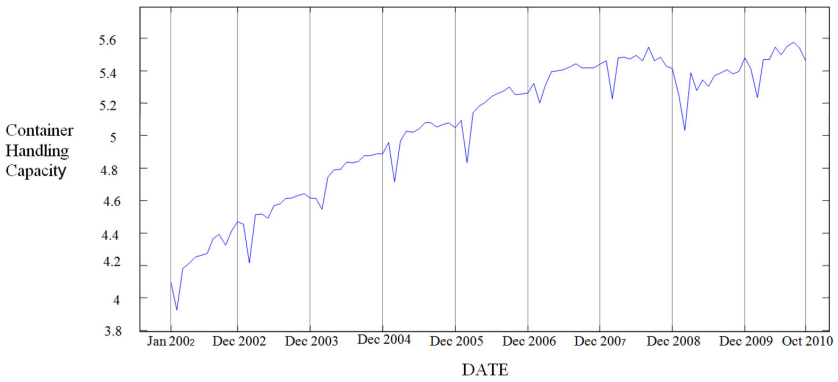

Figure 1. Monthly container handling capacities of Shanghai Port
图 1. 上海港集装箱吞吐量的时序图

上海港的集装箱吞吐量模型为 $\mathrm{SARIMA}(0,1,1)\times(0,1,1)_{12}$：

$$\nabla\nabla_{12}z_t = (1-\theta B)(1-\Theta B^{12})a_t \tag{1}$$

模型为 $(0,1,1)\times(0,1,1)_{12}$ 阶。该模型显然可以写为：

$$z_t - z_{t-1} - z_{t-12} - z_{t-13} = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta\Theta a_{t-13} \tag{2}$$

其中 z_t 表示时间序列值，a_t 表示白噪声序列，B 表示差分算子，θ 描述确定性趋势。这个模型的可逆域由 $(1-\theta B)(1-\Theta B^{12})=0$ 的根在单位圆外这一条件所要求，它由如下不等式定义 $-1<\theta<1$ 和 $-1<\Theta<1$。

注意在式(1)右边的移动平均算子 $(1-\theta B)(1-\Theta B^{12})=1-\theta B-\Theta B^{12}+\theta\Theta B^{13}$ 的阶为 $q+sQ=1+12(1)=13$。

3.3. 参数估计

由式(2)该模型可以视为 $w_t=a_t-\theta a_{t-1}-\Theta a_{t-12}+\theta\Theta a_{t-13}$，这是一个 $w_t=\nabla\nabla_{12}z_t$ 的 13 阶 MA 模型。于是 w_t 的自协方差可以通过如下计算获得[12]：

$$\gamma_0=\left[1+\theta^2+\Theta^2+(\theta\Theta)^2\right]\sigma_a^2$$
$$\gamma_1=\left[-\theta-\Theta(\theta\Theta)\right]\sigma_a^2=-\theta(1+\Theta^2)\sigma_a^2$$
$$\gamma_{11}=\theta\Theta\sigma_a^2$$
$$\gamma_{12}=\left[-\Theta-\theta(\theta\Theta)\right]\sigma_a^2=-\Theta(1+\Theta^2)\sigma_a^2$$
$$\gamma_{13}=\theta\Theta\sigma_a^2.$$

特别地，这些表达式蕴含 $\rho_1=-\theta/(1+\theta^2)$，$\rho_{12}=-\Theta/(1+\Theta^2)$，因此，$\rho_1$ 值不受模型(1)中存在的 MA 季节因子 $(1-\Theta B^{12})$ 的影响，而 ρ_{12} 值不受模型(1)中的非季节 MA 因子 $(1=\theta B)$ 的影响。经过一阶十二步差分后的序列的自协方差估计值如表 2 所示。

接下来，令观察的相关函数等于它们的期望值，可以得到参数 θ 和 Θ 的近似值。再将样本估计值 $r_1=-0.4183$ 和 $r_{12}=-0.3319$ 作为 ρ_1 和 ρ_{12} 的近似值带入下面表达式：

$$\rho_1=-\theta/(1+\theta^2),\quad \rho_{12}=-\Theta/(1+\Theta^2)$$

我们得到估计值 $\hat{\theta}=0.5405$，$\hat{\Theta}=0.3798$，$\sigma_a^2=0.0032$。

3.4. 模型诊断检验(累积周期图检验)

在季节时间序列拟合中，恐怕很可能会未充分考虑到序列的周期特性，因此，我们应注意残差中的周期性。自相关函数对于这类随机状态的偏离并不能给出灵敏的指示，因为周期效应本身常常融汇在自相关之中。而另一方面，周期图就是为检验在白噪声背景下周期波形的模式而设计的。

一个时间序列 a_t，$t=1,2,\cdots,n$ 的周期图是[12]

$$I(f_i)=\frac{2}{n}\left[\left(\sum_{t=1}^{n}a_t\cos(2\pi f_i t)\right)^2+\left(\sum_{t=1}^{n}a_t\sin(2\pi f_i t)\right)^2\right]$$

其中 $f_i=i/n$ 为频率。因此周期图是把 a_t 和不同频率的正弦和余弦波相联系的一种工具。在残差中若含给定的频率 f_i，则自该频率上响应的正弦或余弦波会使波形增强从而产生大的 $I(f_i)$ 值。

白噪声的功率谱 $p(f)$ 在 0~5 周的频率区域上都具有常值 $2\sigma_a^2$。因此，白噪声的累积功率谱

$$P(f)=\int_0^f p(g)\mathrm{d}g$$

对 f 作图就是从 $(0,0)$ 到 $(0.5,\sigma_a^2)$ 的直线，即 $p(f)$ 是一条从 $(0,0)$ 到 $(0.5,1)$ 的直线。

$I(f_i)$ 给出了频率 f 处功率谱的估计。事实上，对于白噪声有 $E[I(f_i)]=2\sigma_a^2$，因此估计是无偏的。故 $(1/n)\sum_{i=1}^{j}I(f_i)$ 给出了积分功率谱 $P(f_j)$ 的无偏估计，且

$$C(f_j)=\frac{\sum_{i=1}^{j}I(f_j)}{ns^2}$$

是 $P(f_j)/\sigma_a^2$，其中 s^2 是 σ_a^2 的估计。我们称 $C(f_j)$ 为归一化的累积周期图。

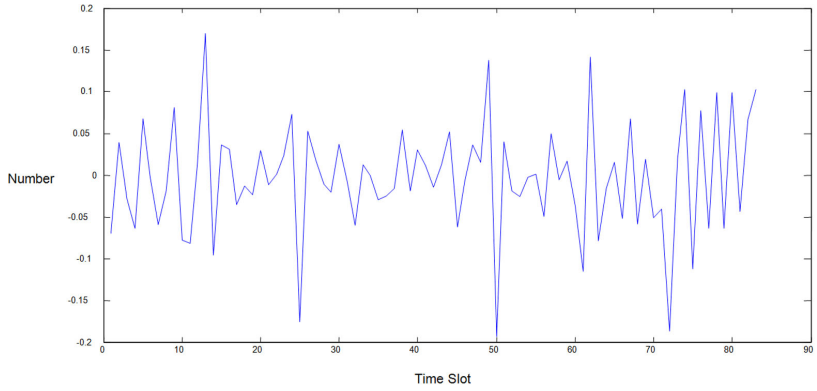

Figure 2. Seasonal differenced container handling capacities of Shanghai Port
图 2. 上海港集装箱吞吐量差分后序列时序图

Table 2. Estimated autocorrelations of various differences of the logged container handling capacity data
表 2. 取对数后的集装箱吞吐量自相关系数

延迟	自相关系数					
1~6	−0.4183	0.0121	0.0416	0.0741	−0.0364	−0.1075
7~12	0.1495	−0.0725	0.0205	−0.1090	0.2537	−0.3319
13~18	−0.0003	0.0861	−0.0380	0.0407	−0.1426	0.2086
19~24	−0.1165	−0.0372	0.1074	0.0125	−0.1390	−0.1065
25~30	0.2660	−0.1581	0.0751	−0.0903	0.1115	−0.1741
31~36	0.1357	−0.0854	−0.0283	0.0301	−0.0023	0.2262

现在，如果模型是恰当的且参数精确地已知，那么，a_t 就可以从数据算出，并得到一个白噪声序列。对于白噪声序列来说，$C\left(f_j\right)$ 对于 f 的图就将会散布在连接点 $(0,0)$ 和 $(0.5,1)$ 的直线附近。另一方面，模型不恰当将会产生非随机的 a_t，累积周期图就会表现出对上述直线的系统偏离。

对于真正随机序列或白噪声序列，将会以时间的比例 ε 被越过。他们画在理论值线、下方的距离为 $\pm K_\varepsilon/\sqrt{q}$ 处，其中，若 n 为偶数，$q=(n-2)/2$，若 n 为奇数，$q=(n-1)/2$。

$$\left(\varepsilon=0.01,\ K_\varepsilon=1.63;\ \varepsilon=0.05,\ K_\varepsilon=1.36;\ \varepsilon=0.10,\ K_\varepsilon=1.22;\ \varepsilon=0.25,\ K_\varepsilon=1.02\right)$$

在我们的研究中，周企图检验结果如图 3 所示。从图 3 可以看到累积周期图的点紧密地聚集在期望直线附近，所以可以确定我们之前估计的参数值充分地符合该乘积季节模型。

3.5. 模型评价

在这一节，我们要应用已经构建出的乘积季节模型去预测 2010 年 2 月至 10 月上海港集装箱吞吐量，并与实际的数据进行比较从而证明该模型的适用性和准确性。

直接由差分方程本身来计算预测值是最好的办法。因此，由于

$$z_{t+l}=z_{t+l-1}+z_{t+l-12}-z_{t+l-13}+a_{t+l}-\theta a_{t+l-1}-\Theta a_{t+l-12}+\theta\Theta a_{t+l-13}$$

在令 $\theta=0.5405$，$\Theta=0.3798$ 后，原点 t 提前 l 期最小均方误差预测立刻给出为

$$\widehat{z_t(l)}=\left[z_{t+l-1}+z_{t+l-12}-z_{t+l-13}+a_{t+l}-0.4a_{t+l-1}-0.6a_{t+l-12}+0.24a_{t+l-13}\right]$$

我们称，$\left[z_{t+l}\right]=E\left[z_{t+l}\middle|\theta,\Theta,z_t,z_{t-1},\cdots\right]$ 为 z_{t+l} 在原点 t 所取的条件期望。在上面表达式中，假设参数

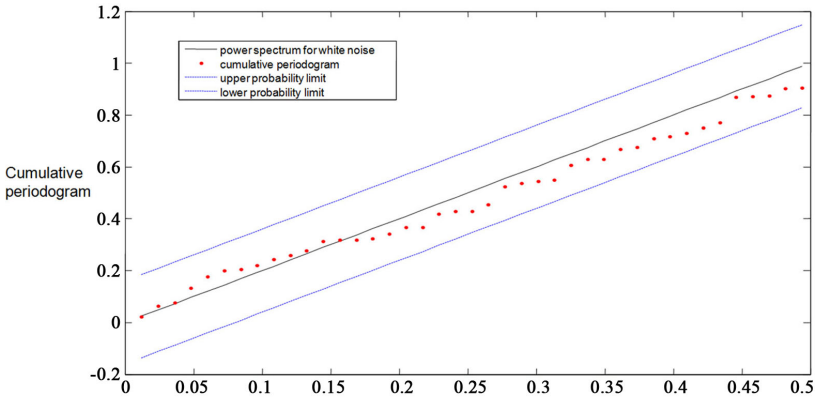

Figure 3. Cumulative perodogram check of the model fitted to the series of container handling capacity
图 3. 集装箱吞吐量序列的累积周期图检验

Table 3. Forecasts and actual values of container handling capacity for 9 months ahead from Feb. 2010
表 3. 从 2010 年 2 月开始往后 9 个月的集装箱吞吐量的实际值和预测值

2010 年	2 月	3 月	4 月	5 月	6 月
实际值	5.2311	5.4689	5.4676	5.5460	5.4980
预测值	5.1941	5.5513	5.4421	5.5078	5.4683
2010 年	7 月	8 月	9 月	10 月	
实际值	5.5491	5.5763	5.5373	5.4592	
预测值	5.5333	5.5531	5.5690	5.5458	

确切地已知，并假设序列 z_t, z_{t-1}, …的信息一直延伸到遥远的过去。为了得到预测值，我们简单地用预测值来代替未知的 z，而用 0 来代替未知的 a。已知 a 当然是已计算出的提前 1 期外推预测误差，即 $a_t = z_t - \widehat{z_{t-1}}(1)$。

应用该预测方法，得出 2010 年 2 月至 10 月上海港集装箱吞吐量的预测值如表 3 所示。

对比真实值和模型的预测值(如表 3 所示)，可看出，预测值和真实值十分接近，相对误差较小，接下来，我们使用均方误差对预测值进行评价。均方误差的计算公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=k+1}^{n} \left(\text{实际值}(t) - \text{预测值}(t)\right)^2}{n-k}}$$

通过上式得到该乘积季节模型的预测值的均方误差为 0.048，从而说明了我们构建的 ARMA $(0,1,1)\times(0,1,1)_{12}$ 乘积模型的拟合效果较好，预测精度较高，可用来预测未来几个月的上海港集装箱吞吐量。

4. 结语

本文通过系统地分析上海港集装箱吞吐量，建立的 ARIMA $(0,1,1)\times(0,1,1)_{12}$ 乘积模型能够很好地拟合实际数据，具有较高的预测精度。因此，对于港口吞吐量这样的数据，既含有季节效应又含有长期趋势效应，并且相互之间有着复杂的先关纠缠关系，最好要采用乘积季节模型进行预测，这样可以得到比较精确的结果。

参考文献 (References)

[1] 陈秀瑛, 古浩 (2010) 灰色线性回归模型在港口吞吐量预测中的应用. *水运工程*, 5.

[2] 高尚, 梅亮 (2007) 基于支持向量机的港口吞吐量预测. *水运工程*, 5.

[3] 程蓉, 吴国付, 张玉洁 (2004) 改进的 RBF 神经网络在港口集装箱吞吐量预测中的应用. *水运工程*, 8.

[4] 安鸿志 (1992) 时间序列分析. 华东师范大学出版社, 上海.

[5] George, E.P.B., Gwilym, M.J. and Reinsel, C.G. (1994) Time series analysis: Forecasting & control. Prentice Hall.

[6] Hosking, J.M.R. (1984) Modeling presistence in hydrological time series using fractional differencing. *Water Resources Research*, **20**, 1898-1908.

[7] Tiao, G.C. and Tsay, R.S. (1994) Some advances in non-linear and adaptive modeling in time series. *Journal of Forecasting*, **13**, 109-131.

[8] Zhang, Y., Bi, P. and Hiller, J.E. (2010) Meteorological variables and malaria in a Chinese temperate city: A twenty-year time-series data analysis. *Environment International*, **36**, 439-445.

[9] Mohan, S. and Vedula, S. (1995) Multiplicative seasonal Arima model for longterm forecasting of inflows. *Water Resources Management*, **9**, 115-126.

[10] 李勇 (2005) 基于乘积 ARIMA 模型的产品不确定性需求预测. *系统工程与电子技术*, **1**, 60-62.

[11] 梁鑫 (2006) 乘积季节模型在商品房市场中的应用研究. *广西师范学院学报*, **2**, 8-12.

[12] 乔治·博克斯, 格威利姆·詹金斯, 格雷戈里·莱因泽尔 (2011) 时间序列分析: 预测与控制. 机械工业出版社, 上海.