ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.15, No.19,July. 2019
E-mail:kfyj@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于ARMA模型的房地产销售趋势预测方法研究

邓佳雯,陈继红

(南通大学 信息科学技术学院,江苏 南通 226019)

摘要:数据挖掘是指从大量的、客观的、有噪声的、随机的数据中,通过算法提取隐含在其中的信息。随着房地产行业的不断发展,数据量越来越庞大,将数据挖掘技术应用于房地产行业,对销售数据进行处理和分析,从真正意义上体现数据的价值。本文以江苏省某市房地产数据为案例进行定量研究展示建模过程,基于ARMA模型并结合加权最小二乘法对模型进行改进。研究结果表明加权最小二乘法与ARMA模型的适应性较好,且优化模型在预测精度上有明显的提高。旨在掌握行业目前所处态势,并为房地产行业的开发计划和营销策略提供决策性指导。

关键词:自回归滑动平均模型;加权最小二乘法;单位根检验;房地产销售数据;趋势预测

中图分类号:TP391 文献标识码:A

文章编号:1009-3044(2019)19-0228-05

开放科学(资源服务)标识码(OSID):

Research on Method of Real Estate Sales Trend Forecast Based on ARMA Model

DENG Jia-wen, CHEN Ji-hong

(School of Information Science and Technology, Nantong University, Nantong 226019, China)

Abstract: Data mining refers to the extraction of implicit information from a large number of objective, noisy and random data through algorithms. With the continuous development of the real estate industry, the amount of data is getting larger and larger. The data mining technology is applied to the real estate industry to process and analyze the sales data, so as to truly reflect the value of data. In this paper, a real estate data of a city in jiangsu province is taken as a case to conduct a quantitative study to demonstrate the modeling process. The model is improved based on ARMA model and combined with weighted least square method. The result shows that the weighted least square method and ARMA model have good adaptability, and the optimization model has a significant improvement in the prediction accuracy. It aims to grasp the current situation of the industry and provide decision-making guidance for the development plan and marketing strategy of the real estate industry.

Key words: ARMA model; weighted least square estimate; ADF test;real estate sales data; trend prediction

DOI:10.14004/j.cnki.ckt.2019.2519

国家统计局数据显示,房地产行业在税收中的占比基本保持在百分之十五左右,占国民生产总值的份额也呈逐年递增态势,在国民经济发展中的重要性不言而喻。随着房地产行业的不断发展,房地产行业数据量越来越庞大,由于技术手段的不足,传统的数据处理方式不能对收集的数据进行科学、有效地加工和处理,影响当前数据分析以及未来数据趋势的预测,致使决策失误。众所周知,随着管理体制的不断深化改革,作为典型的数据密集型行业,房地产行业的竞争也越来越激烈。与其他行业相比,房地产行业负载着更多的数据。谁能准确地分析数据,谁就能提供更好的服务,从而获得更多的机会,在市场竞争中取胜。房地产行业的分析样本数据集来源于行业服务支撑系统,包括房源信息,开发信息,销售信息,账务信息,客户信息等,涵盖了所有的业务信息。由于房地产行业的某些数据存在实时性和波动性,如此海量的数据如果没有得到有效的分析和处理,数据的价值将无法体现,因此亟需合理有效的预测手段。

1 房地产预测研究方法

文献[1]中将马尔科夫链的状态转移概率应用到商品房价格预测中,在对2009-2012年北京住房销售数据进行处理后,预测未来6个月的房价走势。

文献[2]通过建立VAR模型选取评价指标对多伦多房价进行预测,评价指标主要有消费价格指数、房价平均增幅的滞后三期、贷款利率等。

文献[3]中分别用人工神经网络、Hedonic回归模型对土耳其的房价进行预测,对比结果表明人工神经网络的预测效果相对较好。

除此之外,张珊玉等人在灰色预测模型的基础上加以改

收稿日期:2019-04-12

作者简介:邓佳雯(1996—),女,江苏南通人,硕士研究生,主要研究方向为数据挖掘;陈继红(1966—),男,江苏南通人,副教授,硕士研究生,主要研究方向为数据挖掘。

进,对某地区房价进行了预测,同时也证明了灰色预测模型在房地产行业的适用性。

申瑞娜、曹昶基于支持向量机模型建立预测模型,并结合主成份分析法预测了上海市房价,预测结果显示采用该方法构建的模型预测效果较好。

国内外对房价预测的研究方法主要有人工神经网络、各类回归模型以及灰色预测等,表1列举了常用的预测方法,并进行对比分析。

表1 常见预测方法的比较分析

方法名称	适用情况
灰色模型	针对数据量少、信息贫乏且不确定性系统进行预测分析
人工神经网络模型	以神经元的数学模型为基础表述,由网络拓扑、节点特点和学习规则来表示[4],通过不断调整自身权重反映输入和输出的映射关系,网络适应能力较强。
随机理论模型	常用于分析不确定性问题,主要包括随机变量方法和随机过程方法
支持向量回归模型	通过假定预测,将数据分为两部分,一部分用来建模,一部分用来检验
ARMA	基于时间序列分析,且时间序列是平稳的。对异常变化值适应性较强,精确度高数据量较大的情况下运行速度很快。

目前房地产数据在实际应用中还存在着许多不足,主要表现在:一是研究数据源的格式不规范、存在大量数据噪声,难以确保数据的准确性;二是预测和分析技术要适应市场的变化,常规的统计分析方法还远远不够;三是预测维度单一,笼统地分析价格涨幅意义不大。

房地产行业数据预测有这两个趋势,一是将销售数据看作一个时间数列,选用恰当的模型对销售趋势进行预测;二是将房价影响因素建立指标体系,从而构建预测模型。文中采用第一种方式,结合模型适用情况和待预测数据特征,同时为了准确地预测销售趋势,在经过数据预处理操作的基础上采用ARMA模型构建计量经济学模型,以客观真实的预测方法进行多维度分析讨论,排除外在环境、人口、经济发展、季节等周期性因素的影响,并以某市房地产实际销售数据为实例进行预测。本课题涉及的时间序列处理后经检验均具平稳性。因此ARMA模型适用于房地产销售趋势预测,理论上预测精度较高。

2 ARMA模型

自回归-移动平均混合模型(Autoregressive moving average mode,简称ARMA模型)是任何线性时间序列模型的理论方程式[5],是一种常见的随机时间序列模型,由自回归模型和移动平均模型组成的,是对数据进行预测的较为客观科学的计量经济学方法之一[6]。

2.1 ARMA模型的基本思想

基于ARMA模型的房地产销售趋势预测的基本思想是:按时间顺序将房地产销售变化数值视为随机时间序列,其中待预测时间序列中第n个值不仅与第$(n-1)$个值存在关联,且与前$(n-1)$个时刻也存在关联,以此来预测第n个时刻的值[7]。只有预测对象为零均值的平稳随机时间序列,才可以使用ARMA建立预测模型。因此在建模之前,需要对时间序列进行差分平稳化和零均值处理。

ARMA模型由自回归模型(AR模型)和移动平均模型(MA模型)组成。接下来对ARMA模型进行具体描述。

(1)AR模型

AR模型是通过现时的干扰和有限项过去的观测值建立模型预测现时值。其P阶AR模型的数学表达式为[8]:

$$Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p} + e_t \quad (1)$$

式(1)则为时间序列Y_t的AR模型表达式,其自回归系数为θ_i,e_t为随机干扰项,B为定义后移算子。

$$BY_t = Y_{t-1} \quad (2)$$

$$B^k Y_t = Y_{t-k} \quad (3)$$

则公式(3)可表示为:

$$Y_t = (\theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) Y_t + e_t \quad (4)$$

(2)MA模型

MA模型的预测原理是用现时干扰及过去的干扰有限项来预测模型的现时值。其q阶MA模型的数学表达式为[9]:

$$Y_t = (1 - w_1 B - w_2 B^2 - \cdots - w_q B^q) e_t \quad (5)$$

式(5)即为时间序列Y_t的MA模型,w_j为待定系数,e_t为随机干扰项[10]。

(3)ARMA模型

由AR模型和MA模型结合成的ARMA模型来描述平稳随机时间序列的自回归移动平均模型[11]。

AR(p)+MA(q)=ARMA(p,q),ARMA模型的数学表达式为:

$$\Theta_p(B) Y_t = W_q(B) e_t \quad (6)$$

式中e_t为随机干扰项,$\Theta_i(1,2,\cdots,p)$和$w_i(1,2,\cdots,q)$是模型的相关系数。

$\Theta_p(B)$和$w_q(B)$的关系式为:

$$W_p(B) = 1 - w_1 B - w_2 B^2 - \cdots - w_q B^q \quad (7)$$

建模过程中值得注意的是,房地产销售趋势变化存在一定的增长趋势和周期性,是一种非平稳随机过程。为了能用ARMA模型来描述,应该定义一个差分算子为:

$$\Delta = 1 - B \quad (8)$$

$$\Delta Y_t = Y_t - Y_{t-1} \quad (9)$$

$$\Delta^m = (1 - B)^m \quad (10)$$

即可用ARMA模型来描述该时间序列。即为:

$$\Theta_p(B) \Delta^m Y_t = W_q(B) e_t \quad (11)$$

综上所述,得到应用于房地产销售趋势预测的ARMA模型即为公式(11)。

2.2 数据平稳性检验

平稳化的目的是为了建立ARMA模型,从图1中可以看出,该时间序列呈现出非周期性的特征。

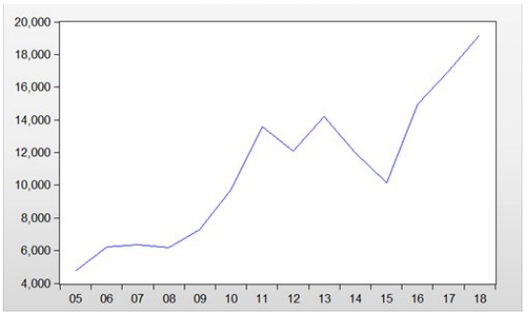

图1　2005－2018年城区房价趋势图

如图2为该时间序列的自相关函数计算结果。结果表明置信区间外有自相关函数且不趋向零[12],因此该序列为非平稳性时间序列。

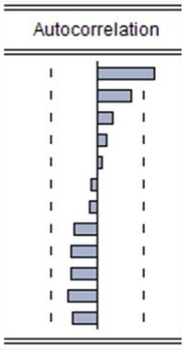

图2　自相关函数计算结果

基于差分分析方法对待预测样本作平稳化处理,用S_{ij}表示一年中第j个月第i天的数据值,其中$j=1,2,\cdots,12$;$i=1,2,\cdots,31$;由于在数据进入排队系统时,均值会发生变化,导致方差也会随之改变。此时为非标准正态随机变量,通过取对数的方法减少标准差,并把lnS_{ij}划分为四个部分为:

$$\ln S_{ij} = \bar{x} + T_i + U_j + y_{ij} \tag{12}$$

其中,$\bar{x}$表示年平均值,T_i表示在第i天的平均值与总平均值$\bar{x}$的偏差,U_j是第j个月平均值与总平均值$\bar{x}$的偏差,y_{ij}为残余值,且满足:

$$\sum_i T_i = 0, \sum_j U_j = 0 \tag{13}$$

经过上述处理后,得到序列y_{ij}。通过ADF检验验证其平稳性,ADF检验的数学表达式为:

$$D_{yt} = a + dy_{t-1} + \sum_{j=1}^{p} al_i Dy_{t-1} + u_t \tag{14}$$

ADF检验结果如表2所示,结果显示在5%、10%的显著性水平下,临界值分别为-4.246503和-3.590496,显然上述T检验统计量值-5.387258小于相应的DW临界值,从而拒绝H_0,表明经差分处理后的时间序列是平稳序列。

表2　ADF检验结果

Test critical values	t-Statistic
Augment Dickey-Fuller test statistic	-5.387258
1% level	-5.835186
5% level	-4.246503
10% level	-3.590496

2.3 步骤

基于ARMA模型对房地产销售趋势进行预测的步骤如下:

(1)对原数据进行数据清洗;
(2)时间序列平稳性检验;
(3)确定模型阶数和参数;
(4)检验预测模型。

3 以某市房地产数据进行案例分析

按照上述方法以某市房地产销售数据为案例进行分析。选取2005－2018年的房屋每平方米销售价格作为时间序列X_t $(t=1,2,\cdots,14)$,以此为例,说明预测过程。

3.1 ARMA预测模型

(1)数据清洗

借助SQL Sever平台对数据进行清洗,剔除安置房、低保房、车库等特殊房型的销售数据,尽可能减少对预测结果的影响。

(2)建立预测模型

ARMA模型的建立包括阶数的确定和参数的估计。

通过计算时间序列$Y_t(t=1,2,\cdots,14)$的自相关和偏相关函数来确定ARMA预测模型的阶数,即p、q的值,涉及的数学表达式如下。

自相关函数:

$$\rho_k = \frac{\sum_{t=1}^{n-k} X_t \circ X_t + t}{\sum_{t=1}^{n} X_t^2} \tag{15}$$

偏相关函数:

$$\varphi_{kk} = \begin{cases} \rho_1 & k = 1 \\ \dfrac{\rho_k - \sum_{j=1}^{k-1} \varphi_{k-1,j} \circ \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \varphi_{j,j} \circ \rho_k} & k = 2,3,\cdots \end{cases} \tag{16}$$

取$\varphi_{k,j} = \varphi_{k-1,j} - \varphi_{kk} \circ \varphi_{k-1,k-j}$。

自相关计算分析结果如图3所示。

图3　时间序列y_{ij}自相关分析图

由上图可知时间序列Y_t所有自相关函数都在置信区间内,且数值趋向于0,因此X_t具有随机性和平稳性的特征,验证了上文差分处理操作的正确性。除此之外,还得知该序列的偏相关函数具有截尾特性,可建立AR(p)模型。对时间序列AR(2),AR(3),AR(4),AR(5)基于SPSS22.0工具进行参数估计,结果如表3所示。

表3　AR(p)模型参数拟合结果

模型	参数						AIC	BIC
	φ_1	φ_2	φ_3	φ_4	φ_5	常数		
AR(2)	-0.089	0.067				0.002	59.82	63.47

AR(3)	−0.083	0.061	0.053			−0.004	52.34	59.82
AR(4)	−0.072	0.082	0.061	−0.042		0.007	57.72	62.78
AR(5)	−0.064	0.075	0.034	−0.028	0.059	−0.007	60.62	64.74

根据AIC和BIC最优准则，阶数均应为3，因此得到最优模型为：

$$X_t = -0.083X_{t-1} + 0.061X_{t-2} + 0.053X_{t-3} - 0.004 \quad (17)$$

预测模型式(17)为原时间序列经差分处理后序列y_{ij}的预测模型，差分逆推可得原销售价格数据的预测模型为：

$$X_t = -1.0886X_{t-1} + 1.0673X_{t-2} + 1.5253X_{t-3} - 35.68 \quad (18)$$

3.2 改进的ARMA预测模型

上述方法为ARMA建模的一般方法，借助SPSS分析软件，再结合AIC和BIC准则得到模型的阶数和参数。为了提高预测模型的精确度，本文对基于ARMA的房地产预测模型进行改进。

(1)ARMA模型阶数确定若AR模型的偏相关函数为截尾，对于P阶的AR模型的偏相关函数等于0，因此当$\varphi_{kk}=0$时，则可确定模型的阶数$p=k-1$。

若MA模型的自相关函数也是截尾的，而q阶的MA模型的自相关函数也等于0，则可以确定模型的阶数$q=k-1$。

(2)ARMA模型参数的估计

确定线性回归模型参数的一般方法为最小二乘法。在这种方法的基础上，本文使用加权最小二乘法对参数估计进行优化，其数学描述如下[13]：

对于回归模型

$$Y^{(P)} = \beta_1 X_1^{(P)} + \cdots + \beta_N X_N^{(P)}$$
$$P = 1,2,\cdots M \quad (19)$$

该模型的残差为

$$e^{(P)} = Y^{(P)} - (\beta_1 X_1^{(P)} + \cdots + \beta_N X_N^{(P)})$$
$$P = 1,2,\cdots M \quad (20)$$

其中M为样本集的数目，N为模型阶数。$X_1^{(P)}\cdots X_N^{(P)}$为干扰预测的因变量，$Y^{(P)}$为预测样本集，$\beta_1\cdots\beta_N$为回归系数。

$$\text{设}\beta = \left[\beta_1\cdots\beta_N\right]^T, X = \begin{bmatrix} X_1^{(1)} & \cdots & X_N^{(1)} \\ \cdots & \cdots & \cdots \\ X_1^{(M)} & \cdots & X_N^{(M)} \end{bmatrix}$$
$$, Y = \left[Y^1\cdots Y^{[M]}\right]^T, E = \left[e^{(1)}\cdots e^{(M)}\right]^T$$

相应目标函数为

$$\frac{1}{2}E^TE = \text{Min} \quad (21)$$

按照目标函数进行参数优化，得到最小二乘估计参数为：

$$\beta = (X'X)^{-1}X'Y \quad (22)$$

普通二乘法在待预测数据存在异常值时，会出现运算错误。结合加权最小二乘方法进行优化以解决上述运算错误问题，将式(21)转换为：

$$\frac{1}{2}(H^{-1}E)^T(H^{-1}E) = \text{Min} \quad (23)$$

$$H = \text{Cov}(E) \quad (24)$$

求得参数估计值：

$$\beta = (X^TH^{-1}X)^{-1}(X^TH^{-1}Y) \quad (25)$$

根据上述方法求得原时间序列的ARMA优化预测模型为：

$$X_t = -0.9876X_{t-1} + 1.0042X_{t-2} + 1.2048X_{t-3} - 50.87 \quad (26)$$

将原模型与优化模型进行对比，最直观的检验方法是将预测值与真实值进行误差对比分析。按照2005-2012年已知的城区房屋销售单价对其他年份进行预测，得到结论如表4所示。

表4　预测误差分析统计

年份	实际值	普通最小二乘法预测值	加权最小二乘法预测值	普通最小二乘法误差	加权最小二乘法误差
2013	14208	11617	15690	−18.23%	+10.43%
2014	11990	13129	13245	+9.50%	+10.47%
2015	10129	12184	12017	+20.29%	+19.17%
2016	14947	12744	14144	−14.74%	−5.37%
2017	16977	20521	19627	20.88%	13.51%

由分析结果可得，优化后的模型在房地产销售数据预测上取得了较好的效果。优化模型的精确度和适应性明显优于原模型，因此结合加权最小二乘法的ARMA模型在房地产销售数据预测上的应用具有一定可靠性。

传统的模型检验为DW统计量检验，使用该检验方法需满足三个条件：一阶自相关、回归中有截距项、回归因子无滞后项，而本文中涉及的预测模型大部分都有滞后项且非一阶自相关，因此采用残缺自相关系数求和的方式进行检验，数据序列为白噪声，即通过检验[14]。

此处利用Ljung-Box Q统计量进行检验，其数学表达式为[15]

$$Q_{LB} = T(T+2)\sum_{j-1}^{p} a\frac{r_j^2}{T-j} \quad (27)$$

其结果为0.589大于0.05，表明该残缺序列为白噪声，通过检验。经过先验知识和前人的研究，满足ARMA模型下的时间序列，基本都能通过该检验。

4 结束语

本文研究了建立房地产销售趋势预测的ARMA模型，除了上述对某市未来三年房地产行业销量预测的案例分析，按照本文研究方法，对房价、购房人群年龄结构、各房屋用途量等也进行了预测。预测结果如表5、表6、图3和图4所示。

表5　按区域销售价格预测表

地区	2019	2020	2021
港闸区	13265.79	14357.82	14556.73
观音山区	15341.24	15772.65	16837.31
开发区	13085.36	14368.25	15824.92
新城区	20211.46	20762.52	22461.32
主城区	16782.31	17323.47	16782.71

表6 按房屋用途销售价格预测表

房屋用途	2019	2020	2021
办公	9977.40	9049.23	9463.43
商业	10892.45	11324.75	11352.86
住宅	16423.15	16829.72	17362.10
其他	11110.71	13249.26	12084.59

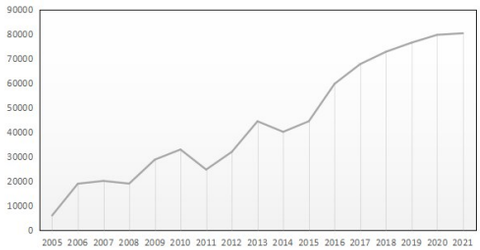

图4 销量预测

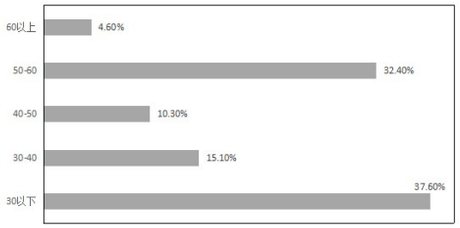

图5 购房人群结构预测

从论文分析和案例应用中可以看出，预测精度较高，结合加权最小二乘法对ARMA模型进行优化，在房地产销售数据上的应用取得了令人满意的预测效果。

在ARMA模型的适应性上可以做进一步的研究与改进，以适应不同的样本数据集合，拓宽应用范围。

参考文献：

[1] 连星. 太原市商品住宅价格预测研究[D]. 太原：山西财经大学, 2017.

[2] 刘扬. 哈尔滨市松北区商品住宅价格预测研究[D]. 哈尔滨：东北林业大学, 2016.

[3] 叶桂芳. 基于国房景气指数的我国房地产市场发展趋势研究[D]. 广州：暨南大学, 2015.

[4] 和湘, 刘晟, 姜吉国. 基于机器学习的入侵检测方法对比研究[J]. 信息网络安全, 2018, 209(05):7-17.

[5] 章晨, 郑循刚, 龚沁. 基于ARMA模型的我国房地产价格预测分析[J]. 生产力研究, 2012(2):27.

[6] Paulo Teles,Paulo S. A. Sousa. The effect of temporal aggregation on the estimation accuracy of ARMA models[J]. Communications in Statistics - Simulation and Computation, 2018, 47(10):2865-2885.

[7] 李瑞莹, 康锐. 基于ARMA模型的故障率预测方法研究[J]. 系统工程与电子技术, 2008, 30(8):1588-1591.

[8] 吕福琴. 基于自回归和神经网络算法加权组合的负荷预测[J]. 广东电力, 2011, 24(5):69-72.

[9] 叶瑰昀, 罗耀华, 刘勇. 基于ARMA模型的电力系统负荷预测方法研究[J]. 信息技术, 2002(6):74-76.

[10] Jongoh Nam, Seonghyun Sim. Forecast accuracy of abalone producer prices by shell size in the Republic of Korea: Modified Diebold - Mariano tests of selected autoregressive models [J]. Aquaculture Economics & Management, 2018, 22(4): 474-489.

[11] 赵彦艳. 随机时间序列模型在煤炭价格预测中的应用[D]. 济南：山东大学, 2010.

[12] 张俊民. 基于特征融合的ARMA短时睡眠状态分析研究[D]. 上海：华东理工大学, 2016.

[13] Wu X , Kumar V , Quinlan J R , et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1):1-37.

[14] Berkhin P. A Survey of Clustering Data Mining Techniques [J]. Grouping Multidimensional Data, 2006, 43(1):25-71.

[15] Rhodes D R, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform[J]. Neoplasia, 2004, 6(1):1-6.

【通联编辑：梁书】