编号：A0746

基于北京城区发展的LASSO-ARIMA房价预测

目录

图目录

表目录

摘要

长久以来，房地产市场都是北京上至政府、下至百姓均极为关注的问题。在我国特色社会主义制度下，房价的涨幅由市场运作和政府的宏观调控共同决定。

北京市幅员辽阔，区域间发展存在差异，房地产市场的吸引性不尽相同。本文选取西城、东城（主城区），朝阳、海淀（近郊区），通州（远郊区）作为主要研究对象。通过官方数据库和相关网站，搜集房价及与城市发展房地产方面的影响因素，并利用 Eviews 软件对残缺数据进行插补。

首先，基于房价的月度数据，用灰度预测 GM(1,1)模型和 ARIMA 模型对各城区房价进行预测。在两模型的均通过显著性检验的前提下，得到灰度预测模型的 MSE 值是 ARIMA 模型对应指标的一千倍，后者体现出对时间序列更优的预测精度。

为了进一步探究房价与城区发展的关联，分析北京市房价受环境变化的波动效应，本文选取若干潜在的房地产影响因素。通过房价与变量的相关性检验，筛选得到八个影响因素并用卡尔曼滤波做降噪处理。考察变量间的皮尔逊相关系数，发现影响因素间的相关度与城区的分类有关，如城区中施工、销售面积有关，而远郊区影响因素为房地产施工面积与房地产开发投资密切相关——城区的发展状态或影响房价影响因素变量间的关系构成。

最终，建立基于变量的 LASSO-ARIMA 集成模型。本文创新性地将“滞后性”理念引入模型，使用当月影响因素数据对照下个月房价，而非本月房价进行 LASSO 回归，实现了可以基于已知数据预测未来房价的效果。同时，本文选取对波动的时间序列有更优预测效果的 ARIMA 模型预测残差序列，并将其加入到 LASSO 模型的误差项中，得到集合了两模型优点的集成模型。通过预测各城区 2021 年的月度房价数据，得到相对误差分别为 3.86%、0.94%、0.23%、

4.85%、6.32%，大致在 5%以内。作为对照模型，多元线性回归的相对误差为 8.22%、6.33%、16.15%、11.50%、5.78%，在 10%左右。上述研究证明本文的集成模型相比于传统统计模型更为优越，数据可参考性较强。结合模型，给出了对北京市城区发展的分析及政策建议。

总之，在风云变幻的房地产市场上，房价始终牵动着全民的心。北京，作为首都，向世界展示中国风貌的窗口之一，理应由政府引导塑造更健康稳定的房地产环境。本文揭示了城区发展的政策对房地产市场的影响，并定量化的给出精度较高的集成预测模型，希望给城区发展予以参考指引，注入新动能。

关键词：北京市城区房价；城区政策分析；ARIMA 模型；LASSO 模型

ABSTRACT

For a long time, the real estate market has been an issue of great concern to Beijing from the government to the people. Under the socialist system with Chinese characteristics, the increase in housing prices is jointly determined by market operations and the government's macro-control.

Beijing has a vast territory, different developments between regions, and the attractiveness of the real estate market is not the same. This article selects Xicheng, Dongcheng (the main urban area), Chaoyang, Haidian (near suburbs), and Tongzhou (outer suburbs) as the main research objects. Through official databases and related websites, collect housing prices and the influencing factors of urban development and real estate, and use Eviews software to interpolate the incomplete data.

First, based on the monthly data of housing prices, the gray-scale forecasting GM(1,1) model and ARIMA model are used to predict housing prices in various urban areas. On the premise that both models pass the significance test, the MSE value of the gray-scale prediction model is one thousand times that of the corresponding index of the ARIMA model, and it reflects the better prediction accuracy of the time series.

In order to further explore the relationship between housing prices and urban development, and analyze the volatility effects of Beijing housing prices affected by environmental changes, this paper selects several potential real estate influence factors. Through the correlation test between housing prices and variables, eight influencing factors are screened out and Kalman filter is used for noise reduction. Examining the Pearson correlation coefficient between variables, it is found that the correlation between the influencing factors is related to the classification of urban areas, such as

the construction and sales area in the urban area, while the influencing factors in the outer suburbs are that the real estate construction area is closely related to the real estate development investment-the urban area The state of development or the composition of the relationship among the variables that affect the housing price.

Finally, a variable-based LASSO-ARIMA integrated model is established. Various factors have a lagging effect on real estate price prediction. This article innovatively uses the current month's influencing factor data to compare next month's house prices instead of this month's house prices for LASSO regression, achieving the effect of predicting future house prices based on existing data. At the same time, combined with research experience, this paper selects the ARIMA model prediction residual sequence that has a better prediction effect on the time series, and adds it to the error term of the LASSO model to obtain an integrated model. By predicting the monthly housing price data of each urban area in 2021, the relative errors are 3.86%, 0.94%, 0.23%, 4.85%, and 6.32%, which are roughly within 5%. As a control model, the relative errors of multiple linear regression are 8.22%, 6.33%, 16.15%, 11.50%, 5.78%, which are about 10%. The above research proves that the ensemble model of this paper is superior to the traditional statistical model, and the data can be referenced.

In short, in the ever-changing real estate market, housing prices have always affected the hearts of the people. This article reveals the impact of urban development policies on the real estate market, and quantitatively gives a high-precision integrated forecasting model, hoping to guide the urban development and inject new momentum.

Keywords: Beijing urban housing prices; urban policy analysis; ARIMA model; LASSO model

一、 引言

(一) 选题背景及其意义

买房安居，是中国文化中土地情结的间接表达，也是现代人的生活刚需。中国人买房热情随着社会的蓬勃发展持续高涨，然而目前遇到了买房需求与远超自身购买力的矛盾。中国住建部指出：住房供求矛盾突出、房价上涨压力大的城市要合理增加住宅用地，特别是普通商品住房用地供应规模。

多年来，房地产业对 GDP 增长的贡献度一直维持在 10%以上。作为衡量我国宏观经济状况的重要指标，房地产市场与居民生活紧密相关。另一方面，房地产产业虽在一定程度上促进了经济发展，但是许多问题也接踵而至：人民的工资不匹配高昂的房价、房地产投资达到了过热的程度而房屋的空置率却不断增长、贫富差距进一步拉大等一系列社会问题。因此，能够准确分析房地产价格波动的因素，通过政府政策读出其中扰动房价的关键信息，从而准确的预测出房价，据此提出房地产产业的相关建设性建议，现已经成为一个具有重大现实意义的研究方向和课题[1]。

过去的十几年中，北京等一线城市纷纷出台了稳定房价的方案，例如：限购限贷等。房价作为民生的焦点问题再次受到公众、政府和金融部门多方的高度关注。故而，厘清房价的走势以及其内在发展规律，既能为市民的购房提供量化指引，一定程度减少经济压力；又可以为政府的政策制定提供参考，引导房地产市场，提高城市居民幸福指数。

(二) 研究现状

2011 年以来，虽然我国经济发展进入了新时代，但在亟待解决的重要民生问题中，住房问题依旧是个“ 顽疾 ”。在挑战中发现机遇，房地产业同样也是中国经济增长的关键点，处于非常重要的地位。其中，“ 学区房 ”、“ 买房投

资”、“包租户”等关键词一度成为全民的关注重点，成为全民聚焦的讨论对象。尽管西方学者对房地产市场多有研究，但是由于中外政治体制、市场运作方式等等方面的不同，国外的研究结论不能很好的体现中国国情，参考性较弱。

从已有文献的研究可以得出，国外房地产市场及其研究都较为成熟，我国房地产发展较晚，数据及研究体系有待完备，尚存在较大的研究缺口。张望舒[2]运用行政加权法及随机森林方法建立在特殊经济环境下二手房屋特征价格的评估模型。该回归模型在实际应用中，若欲“预测”某月数据，则需代入当月的相关数据，而此时已经可以得到确切的房价数据，预测的时效性不佳。崔庆岳和赵国瑞[3]采用灰色 GM(1,1)模型预测广州市的短期房价并引用二阶弱化缓冲算子有效弥补了灰色 GM(1,1)模型自身在预测随机波动大数据时的缺陷，最终得到商品房在广州市的需求仍属于刚性需求，商品房销售价格仍有稳步上升的空间的结论。但未考虑到政府政策、重大卫生事件等一系列因素的影响，因此在有因素扰动时误差会增大。且最终预测出的数据仍为年度商品房平均价，忽略了房价在一年中的波动变化，不够细化，参考性不强。王景行[4]将 LASSO 回归和 XGBoost 机器学习算法集成并融合 stacking 模型来对比用单一方法预测房价的效果，得到用集成模型来预测房价比用单一模型预测效果更加显著的结论。

综上所述，本文探究了一种更为优化的集成模型来对北京市各区域房地产市场的变化规律及其波动特征进行研究及分析。

二、 模型构建思路及创新

(一) LASSO-ARIMA 模型构建思路

本文的构建思路如下：

第一部分是数据预处理和相关性分析：选取聚汇数据网，中经网数据库和

前瞻数据库上 2010 年-2020 年的房价和因素指标数据，通过 R 软件画出了各因素的相关系数热力图；

第二部分是基于预处理之后的数据对北京市各城区房价进行灰度预测和 ARIMA(p,d,q)预测探究，展现 ARIMA 模型相比灰色 GM(1,1)模型在预测时间序列方面对数据波动的适应性更合理和精确。

第三部分是房价相关因素指标的相关性研究，核心因素提取（确定自变量）：格兰杰因果检验，多变量间相似程度的量化分析，相关系数基础，Pearson 相关系数，Kendall 相关系数，Pearman 秩相关序数，相关指标的卡尔曼滤波分析数据的科学性与可靠性。

第四部分 LASSO 模型与 ARIMA 模型的集成与优化：为了得到更好的结果，预测地更加精确，引入 LASSO 进行回归的思想，对数据进行归一化处理，并进一步得到最佳参数λ，再利用 ARIMA 模型对时间序列预测的优越性预测回归后的残差项，称之为“修正项”，再将预测出的数据与修正项进行加和得到修正后的预测数据，以此实现对模型参数的优化，并将优化后的模型取名为 LASSO-ARIMA 模型。后文将这个集成之后的模型统一叫做 LASSO-ARIMA 模型。

第五部分是实证分析：用 LASSO-ARIMA 模型的预测北京五个城区的房价，并据此给出建议与分析。

(二) LASSO-ARIMA 模型创新

1.使用 Eviews 软件运用不同且适合的插值方法将低频（年度）数据转化为高频（月度）数据，可以高频地预测房价，更微观地展现了房价的变化趋势。

2.将下个月的房价数据与本月的因素数据进行对齐并实现 LASSO 回归（若想预测任意n个月后的房价可将滞后阶变为n，但建议滞后阶不要过大：$n \leq 3$）。此

法考虑了房价波动的滞后性以及对于扰动的后效性，并且可以因此利用本月数据直接预测下月的房价，避免了需要预测下个月的因素数据才能代入方程得出下个月房价的多重误差，可将误差降到最低，同时又体现了时间上的滞后性，以一定的理论依据。

3.通过对 LASSO 回归后的残差进行 ARIMA 预测，使最终预测值补上预测残差来修正和代替仅仅将数值代入回归方程预测的方法。

(三) LASSO-ARIMA 模型构建流程图

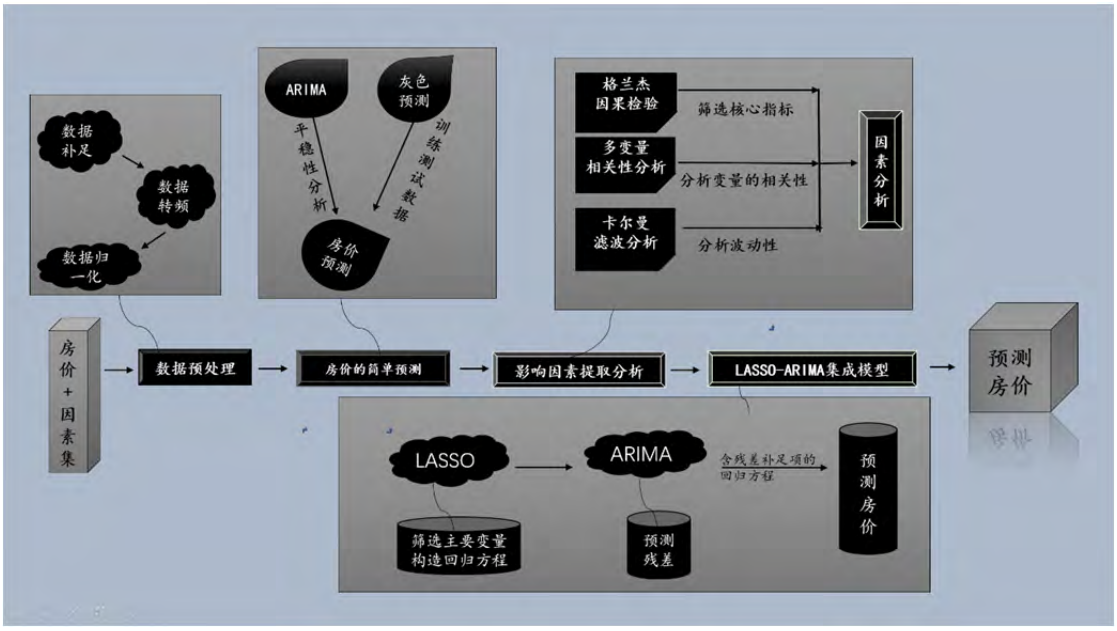


图 1 LASSO-ARIMA 模型构建流程图

三、 变量描述及数据预处理

(一) 变量描述

目前，房价的共识定义为：房屋建筑物，连同其占用土地在特定时间段内房产的市场价值组成房价的总和。这一定义直接反映了房价受其所在地理位置和时间维度而产生了差异性。从地理纬度，按照北京市规划和自然资源委员会的定义：北京市的区划分为城区、近郊区、远郊区，其直接反映各区的组织构成，间接体现各区的发展程度差距。从时间维度，由于时间不可逆转重复，每

一个时刻的房价都在变化、各不相同。

本文为实现对未来任意月份的指定区域房价预测，收集了 2011-2020 年北京五大区域的月度房价数据，以及影响房地产因素各区总 GDP、房地产开发投资、商品房销售面积、房地产 GDP、第三产业 GDP、商品房施工面积、商品房竣工面积以及常住人口。具体内容见表格.

表格 1 城区说明表

序号	城区分类	城区名
1	城区	西城区
2	城区	东城区
3	近郊区	海淀区
4	近郊区	朝阳区
5	远郊区	通州区

(二) 数据预处理

1.数据补足：本文结合 ARIMA 模型和灰度预测模型对残缺数据进行预测补足。

注：由于在 2010 年宣武区与西城区合并，崇文区与东城区合并，因此西城区和东城区在 2010 年的数据产生了突变，不再具有规律性。因此本文在数据检验中便舍弃了这两个区的 2010 年前的数据。

2.数据转频：本文采取 Eviews 软件将最终选定的 2011 年—2020 年北京五个城区前 7 个因素的所有数据进行以 Sum 为基准的二次插值，对第 8 个因素进行以 Average 为基准的二次插值，将所有数据转频成 2011 年 1 月—2020 年 12 月的月度数据。

3.数据归一化：对所有数据作如下的归一化处理

对任意一列指标数据$X_j = (x_{1j}, x_{2j}, \dots, x_{nj}, \dots)$

$$x_{ij}{}' = \frac{x_{ij} - \min\limits_{1\le i\le n} x_{ij}}{\max\limits_{1\le i\le n} x_{ij} - \min\limits_{1\le i\le n} x_{ij}} j = 1,2,\dots \#(1)$$

反之，反归一化为：

$$x_{ij} = x_{ij}{}' \left(\max_{1\le i\le n} x_{ij} - \min_{1\le i\le n} x_{ij}\right) + \min_{1\le i\le n} x_{ij}\, j = 1,2,\dots \#(2)$$

(三) 数据来源

本文主要从北京市各个城区入手进行数据调查分析，主要采用 Python、R、SPSS 等软件进行综合计算分析。为确保数据来源的真实可信基于以下数据来源网站进行相关数据查询。

表格 2 数据说明表

数据名称	变量	单位	时间范围	频度	数据来源
房价	Y	（元/平方米）	2011.1-2020.12	月度	聚汇数据
总 GDP	X1	亿元	2011-2020	年度	前瞻数据库
房地产开发投资	X2	亿元	2011-2020	年度	前瞻数据库
商品房销售面积	X3	万/m^2	2011-2020	年度	前瞻数据库
房地产 GDP	X4	亿元	2011-2020	年度	前瞻数据库
第三产业 GDP	X5	亿元	2011-2020	年度	前瞻数据库
商品房施工面积	X6	万/m^2	2011-2020	年度	前瞻数据库
商品房竣工面积	X7	万/m^2	2011-2020	年度	前瞻数据库
常住人口	X8	万人	2011-2020	年度	前瞻数据库

四、 房价的预测探究

(一) 基于灰度预测 GM(1,1)的北京市各城区未来房价预测

1.GM(1,1)模型的原理

GM(1,1)表示 1 阶的，1 个变量的微分方程。其简历过程如下：

设有原始数据序列（非负序列），记$X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\}$,

灰色系统模型是在生成数列基础上，将时间序列转化为微方程的时间连续模型。$x^{(0)}(k) + az^{(1)}(k) = b$为 GM(1,1)模型的基本形式。

首先对各原始序列进行一次累加，得到生成数列

$$X^{(1)} = \left(x^{(1)}(1), x^{(1)}(2) \cdots x^{(1)}(n)\right) \#(3)$$

其中

$$x^{(1)}(k) = \sum_{j=1}^{i} x^{(0)}(i)\ k = 1,2, \ldots, n \#(4)$$

若$\hat{a} = \begin{pmatrix} a \\ b \end{pmatrix}$为待辩参数序列且参数算式为

$$B = \begin{bmatrix} -\frac{1}{2}\left(x^{(1)}(1) + x^{(1)}(2)\right) & 1 \\ -\frac{1}{2}\left(x^{(1)}(2) + x^{(1)}(3)\right) & 1 \\ \cdots & \cdots \\ -\frac{1}{2}\left(x^{(1)}(n-1) + x^{(1)}(n)\right) & 1 \end{bmatrix} \#(5)$$

$$y_N = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \cdots \\ x^{(0)}(n) \end{bmatrix} \#(6)$$

则 GM(1,1)模型$x^{(0)}(k) + az^{(1)}(k) = b$的最小二乘法估计参数列满足

$$\hat{a} = (B^T B)^{-1} B^T y_N \#(7)$$

这里$\hat{a}$的求法是根据最小二乘法中的定义而来。

2.GM(1,1)模型的后验差检验

记S_1为原始数列标准差，S_2为绝对误差数列标准差，方差比$C = S_2/S_1$，小误差概率为$P = p\left\{\left|\Delta^{(0)}(i) - \overline{\Delta^{(0)}}\right| < 0.6745 \times S_1\right\}$，令

$$e_i = \left|\Delta^{(0)}(i) - \overline{\Delta^{(0)}}\right|,\ S_0 = 0.6745S1,\ 则P = p(e_i < S_0)$$

判断标准如表 3 所示：

表格 3 后验差判断标准

P	C	效果
> 0.95	< 0.35	好
> 0.8	< 0.5	合格
> 0.7	< 0.65	勉强
≤ 0.7	≥ 0.65	不合格

3.GM(1,1)的应用

本文率先尝试着直接用灰度预测来预测未来短期内的房价，本着训练集与测试集之比为 3：1 的原则，将 2011 年—2017 年的所有北京各区的月度房价作为训练集来预测 2018 年 1 月和 2 月的房价，与真实值作对比；再将 2011 年—2018 年的所有北京各区的月度房价作为训练集来预测 2019 年 1 月和 2 月的房价，与真实值作对比；再将 2011 年—2019 年的所有北京各区的月度房价作为训练集来预测 2020 年 1 月和 2 月的房价[7]。

本文以西城区测试为例，其余结果见附录

表格 4 西城区原始数据和 GM 预测值

时间	原始数据	灰色预测	相对误差	后验差比	MSE
2018.1	112757.0000	127053.2149	12.68%	0.24542	5.83E+09
2018.2	108003.0000	128953.2858	19.40%	0.24542	5.83E+09
2019.1	125402.0000	137860.2069	9.93%	0.26561	6.90E+09
2019.2	126802.0000	139657.2281	10.14%	0.26561	6.90E+09
2020.1	117185.0000	149060.1058	27.20%	0.27218	8.04E+09
2020.2	118899.0000	150796.3183	26.83%	0.27218	8.04E+09

4.灰色预测的实现

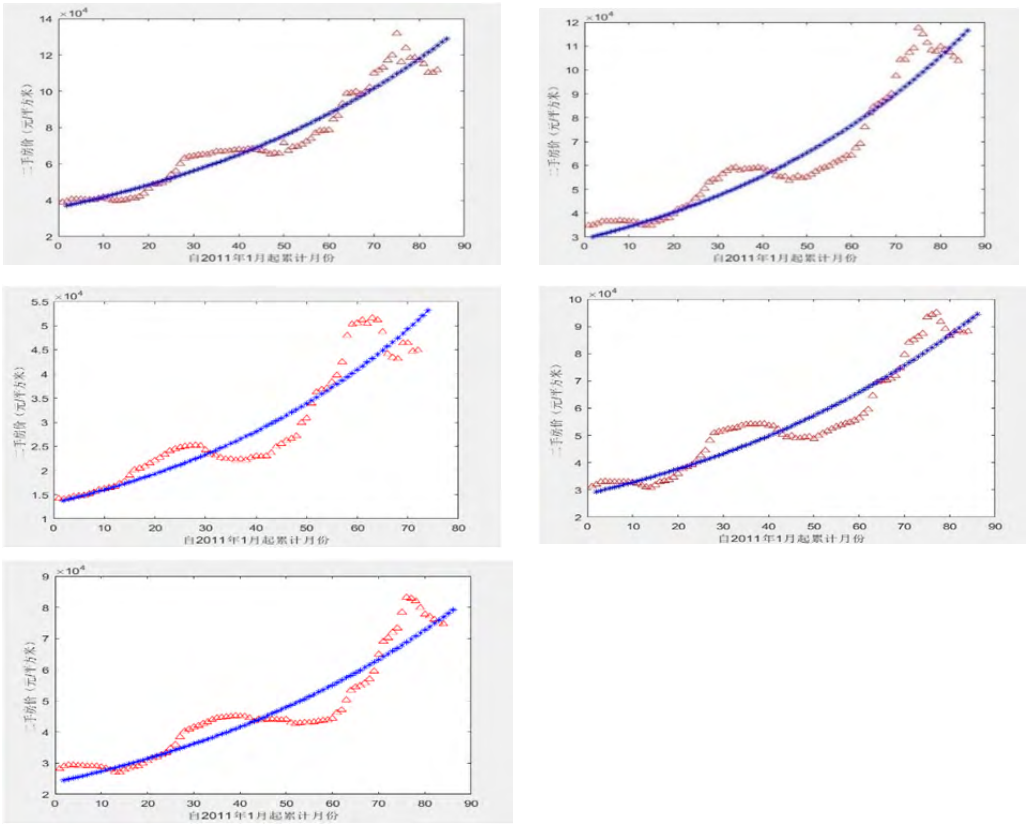


图 2 灰度示意图 a.西城区 b.东城区 c.海淀区 d.朝阳区 e.通州区

由上表可得虽然所有的后验差比值均小于 0.5（可认为灰度预测系统精度合格）但仍与真实值存在较大误差，因此可以说明，直接用灰度预测进行二手房价的预测是不可取的，应采用其他模型或者对上述模型加以改进。

(二) 基于 ARIMA(p,d,q)的北京市各城区未来房价预测

在发现 GM(1,1)灰度预测模型对房价序列的预测精度有限后，为得到对房价——这类波动性时间序列的更好预测效果，本文继而使用 ARIMA 模型对房价预测。

1.ARIMA 模型

ARIMA 模型,是由上世纪 70 年代发明，目前仍广为使用的时间序列预测模

型。该模型的使用前提是输入的变量序列为平稳序列，若原序列非平稳，则需要对其做差分，记其变为平稳序列的差分次数记为 d。其主模型由两个部分组成:AR 是自归,p 为自回归项;MA 为移动平均,q 为移动平均项数。

在 p,q,d 已知的情况下,ARIMA 模型可写为:

$$yt' = \mu + \beta 1 * yt - 1 + \ldots + \beta p * yt - p + \gamma 1 * et - 1 + \ldots + \gamma q * et - qy$$

$$= \mu + \beta 1 * yt - 1 + \cdots + \beta p * yt - p + \gamma 1 * et - 1 + \cdots + \gamma q * et - q\#(8)$$

其中,β为 AR 的系数,γ为 MA 的系数。

算法实现过程如下：

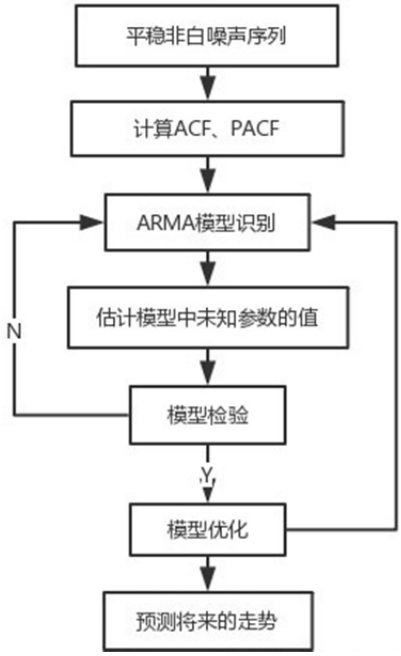


图 3 ARIMA 算法实现步骤

2.ARIMA 模型的实例探究

以下以西城区第一年的数据为例。

为检测数据的平稳性，导入西城区自 2011 年 1 月至 2017 年 12 月的数据,做序列图进行观察。

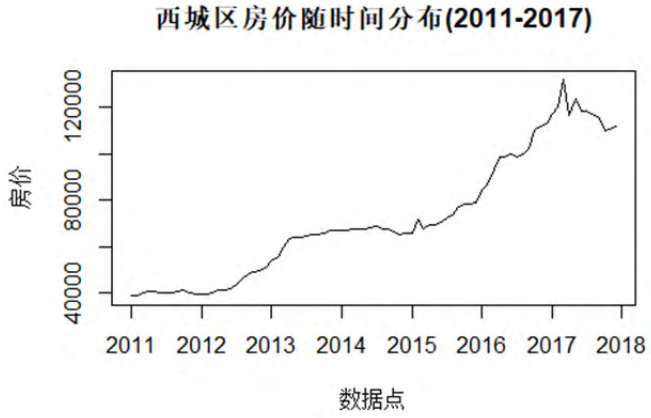


图 4 房价变化趋势（以 2011-2017 西城区为例）

由图所知,房价序列无季节性,随时间推移呈现递增趋势,可直接对数据进行平稳性分析。

平稳序列的定义为：其序列在某一数值附近上下波动，且振幅在一定范围内。将房价序列作一次差分,发现差分序列大致分布在以零为纵坐标的轴的上下两侧,波动程度在可接受范围内,具有较高平稳性。从而确定西城区房价序列的 d 值取为 1。

图 5 房价一阶差分（以 2011-2017 西城区为例）

作出的自相关系数(AC)和偏相关系数(PCA)图标，观察拖尾和截尾的情况可得出自回归项(p)和移动平均项数(q)值分别为 2,0。经 Ljung-Box 残差检验,Q 值为 10.677,P 值为 0.832,P>0.05,由此说明残差为白噪声序列,通过白噪声检验。可以利用其对西城区月度房价进行预测。

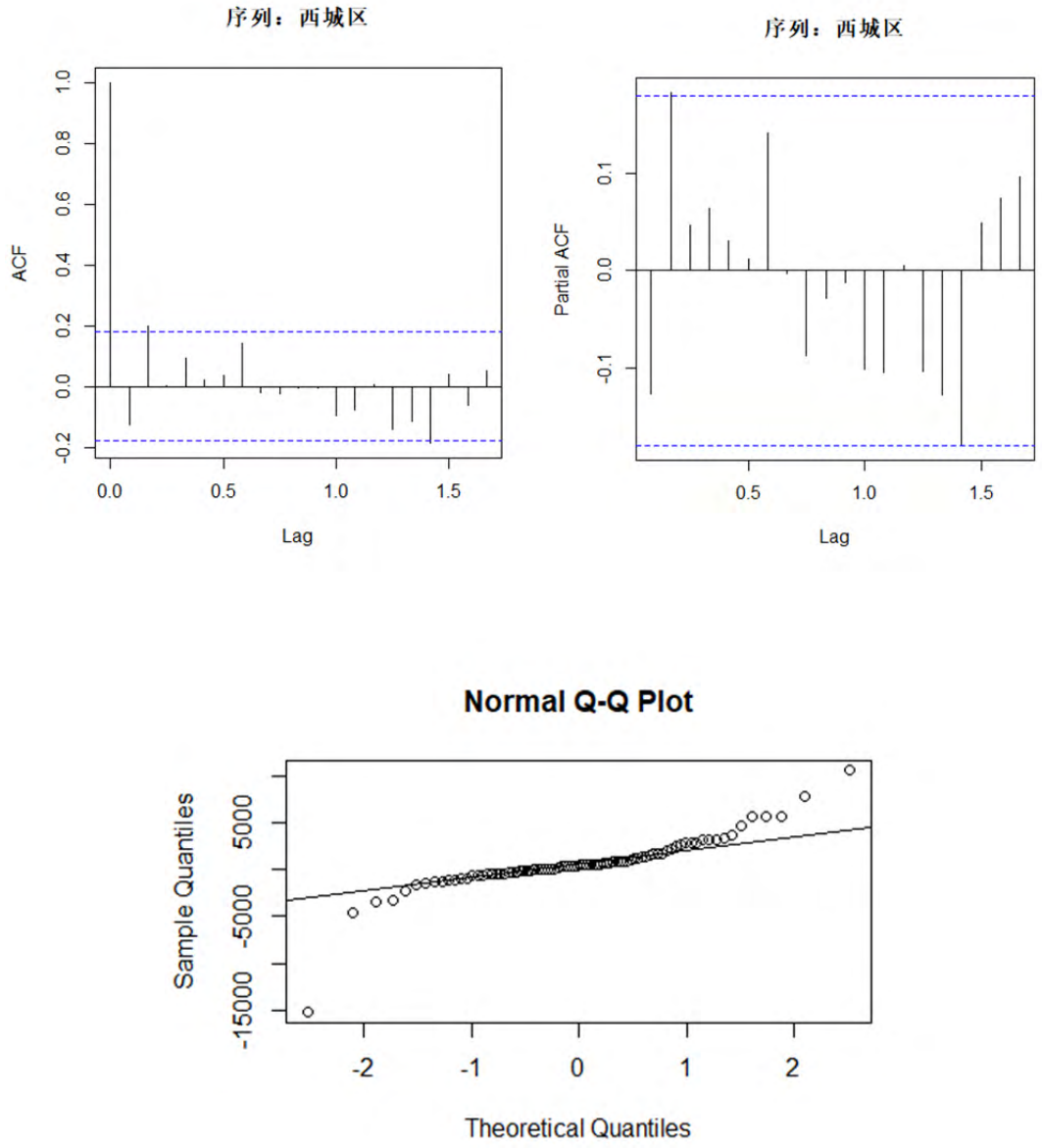


图 6 房价残差检测图（西城区月度）

以下为对西城区未来三年的房价数据预测示意图。

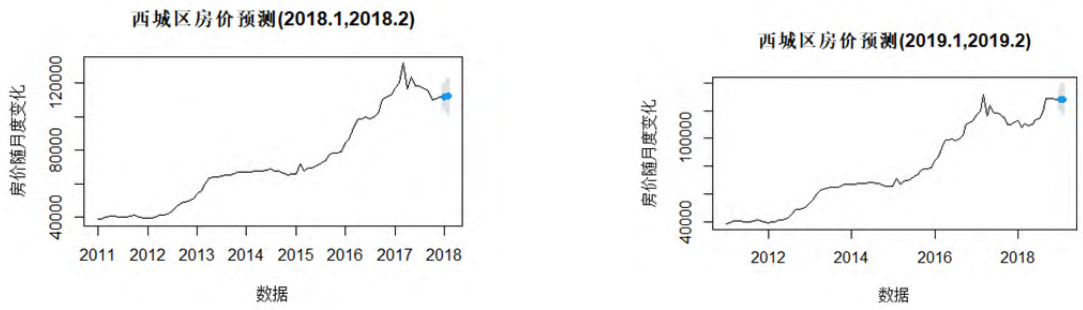


图 7 西城区未来三年的房价数据预测示意图

对于拥有不同特征的时间序列，ARIMA 模型所匹配的 p,d,q 参数值有所不同。重复上述步骤，对所有城区的房价进行预测，并计算精度指标（相对误差、MSE）。本文以西城区测试为例，其余结果见附录

表格 5 西城区原始数据与 ARIMA 预测值

时间	原始数据	灰色预测	相对误差	后验差比	MSE
2018.1	112757.0	111451.1	1.16%	>0.05	8.62E+06
2018.2	108003.0	112470.0	4.14%	>0.05	8.62E+06
2019.1	125402.0	127280.9	1.50%	>0.05	9.28E+06
2019.2	126802.0	127070.2	0.21%	>0.05	9.28E+06
2020.1	117185.0	121525.6	3.70%	>0.05	9.77E+06
2020.2	118899.0	118519.4	0.32%	>0.05	9.77E+06

3.ARIMA 与灰度预测模型的对比

由预测数据对比可知，ARIMA 模型精度相比于灰度预测模型 GM(1,1)高约 5 个（待确认）百分点，大部分房价预测的相对误差可被控制在 5%内。但是，由于其模型特性，不能够以较高精度在对短期的波动进行预测，使得出现相对误差超过 10%的情况，这不能适应房地产市场错综复杂因素的相互作用，以及对突发事件响应能力较弱。综上，本文将选用复合多元回归方法，以寻求更高精度、更高准确性的房价预测模型。

五、 房价相关因素指标的相关性研究

(一) 基于房价相关因素指标的格兰杰因果检验初探

本文查阅数据库，选用了北京各个城区的年度总 GDP、房地产开发投资、商品房销售面积、房地产业 GDP、第三产业 GDP、商品房施工面积、商品房竣工面积、常住人口自 2008 年—2020 年的数据，并对其做了上述数据处理。（注：在用灰度预测填补时的后验差比值均<0.5，所以可以说填补数据具有可靠性）

由于进行格兰杰检验的前提是数据是平稳的，本文接下来用 Eviews 软件将房价真实值的月度数据通过算术平均值整合成年度数据与每一个上述指标先进行了单位根检验，发现大多数是不平稳的，之后继续进行协整性检验，发现也存在因素指标不具有协整性，所以无法对所有因素指标进行格兰杰因果检验。因此本文舍弃格兰杰因果检验来选择合适的因素指标，转而寻找其它更合适的方法。

(二) 多变量间相似程度的量化分析

通过各影响因子和房价因素的统计学三大相关性系数，即 Pearson 系数，Kendall 系数，Spearman 系数来计算各因素数据与房价的相关性。一般来说，三大相关性系数都可以用来表示两个变量之间变化趋势的方向以及它们之间相关性程度大小。

1.相关系数的计算结果

以西城区为例计算各相关系数

表格 6 相关系数表（以西城区为例）

影响因子	Pearson	Kendall	Spearman
总 GDP	0.932	0.834	0.950
房地产开发投资	-0.158	-0.256	-0.310
商品房销售面积	-0.918	-0.583	-0.793

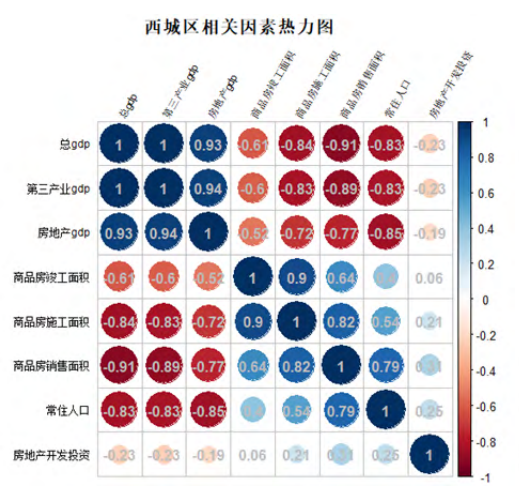
房地产 GDP	0.801	0.805	0.946
第三产业 GDP	0.913	0.824	0.943
商品房施工面积	-0.898	-0.654	-0.863

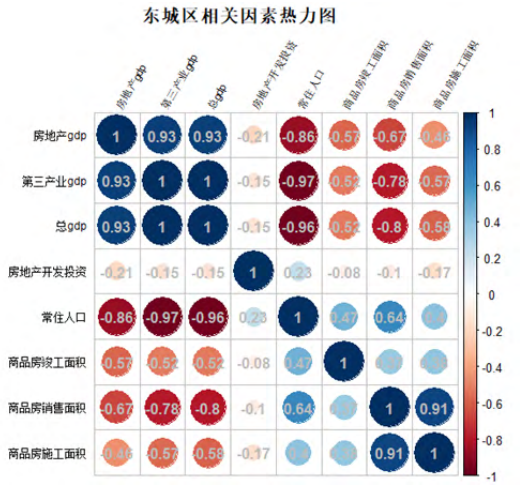
上表为各影响因素和房价的相关系数。各相关系数的显著性水平值 P 均以 0.01 为界判断，显著性水平值 P 越小，两变量间相关性的判别可信度越高。

其中，假设检验验证变量相关性时，设 H_0：两变量无关，P 值即为 H_0 成立的概率；当 P 大于某一值α（选定的显著性水平，一般为 0.05 或 0.01）时，接受原假设，即认为两变量无关；当 P 小于α时，接受其备择假设，即认为两变量存在相关性。在本文中，选取$\alpha = 0.05$。

2.以 Pearson 相关系数为例画出各城区的相关系数热力图

本文在建模的时候为了避免多重共线性一般都会分析变量之间的相关性。衡量变量相关性时，一般选用计算变量两两之间的皮尔逊相关系数的方法。





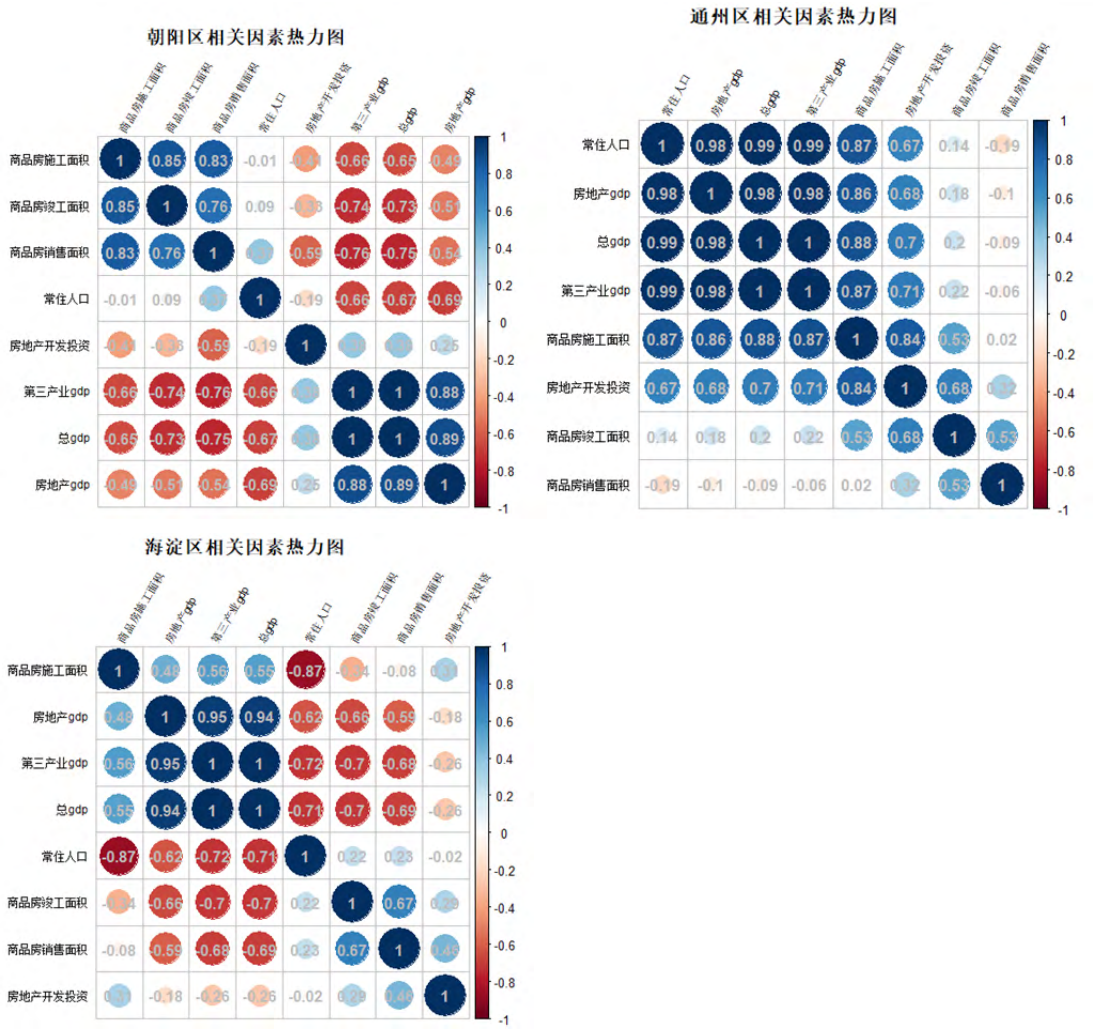


图 8 相关因素指标的 Pearson 热力图

由图所知，影响因素间的相关度与城区的分类有关，如城区中施工、销售面积有关，而远郊区影响因素为房地产施工面积与房地产开发投资密切相关——城区的发展状态或影响房价影响因素变量间的关系构成。

(三) 相关指标的卡尔曼滤波分析

卡尔曼滤波本质上是一种将数据融合在一起算法，它将相同测量目的、不同传感器以及不同单位的数据融合，从而得到一个更精确的测量值。卡尔曼滤波当中的滤波其实是指通过一种算法排除可能的随机干扰以提高检测精度的方法或手段。

基本上用过卡尔曼滤波法的人都知道著名的“ 黄金五条 ”公式，且通过

“预测”与“更新”两个过程来对系统的状态进行最优估计。这里不加证明地直接给出经典的离散系统卡尔曼滤波公式，包括 5 方程如下所示:

状态预测值一步预测方程——黄金一条：

$$\hat{x}_k = A\hat{x}_{k-1} \#(9)$$

协方差矩阵一步预测方程——黄金二条：

$$P_k = AP_{k-1}A' + Q \#(10)$$

最优估计条件下卡尔曼滤波增益方程(权重)——黄金三条：

$$H_k = AP_{k-1}A' + Q \#(11)$$

状态最优估计值(k 时刻的最优值)可由状态更新方程——黄金四条：

$$\hat{x}_k = \hat{x}_k^- + H_k z_k - H\hat{x}_k \#(12)$$

估计误差方差阵方程——黄金五条：

$$P_k = (I - K_k H)P_k^- \#(13)$$

式中，x 表示状态真实值，z 为观测值，A 为状态转移矩阵，H 为协方差矩阵，P 为误差协方差阵，Q 为系统噪声协方差矩阵，K 为卡尔曼增益矩阵，R 为观测噪声协方差矩阵[5]。

通过关联度分析，年度总 GDP、房地产开发投资、商品房销售面积、房地产业 GDP、第三产业 GDP、商品房施工面积、商品房竣工面积、常住人口密度八个指标，基于多元时间序列的卡尔曼滤波算法，通过每期拟合情况分析预测效果如下：

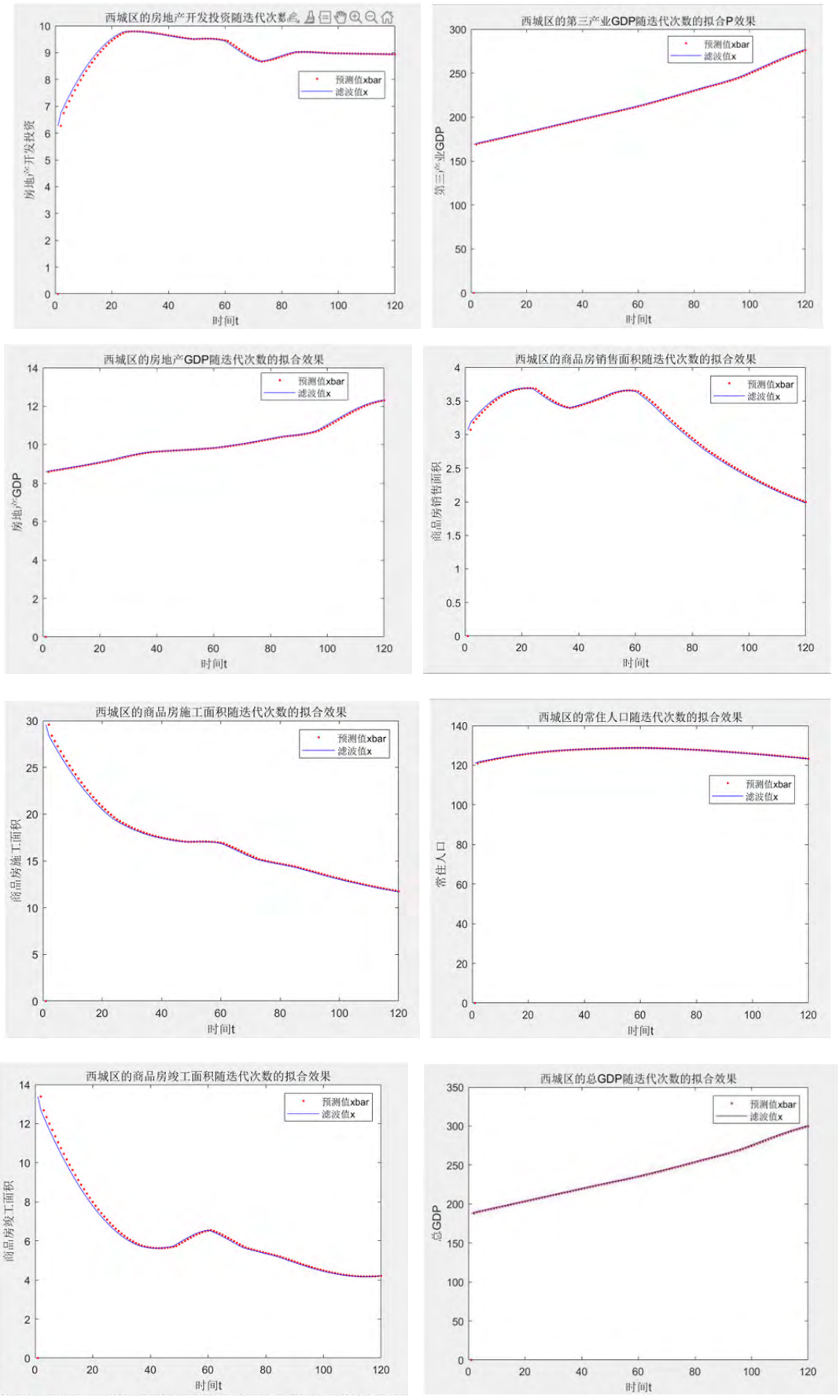


图 9 相关因素指标时序图

由图所示，8 个图分别对应总 GDP，房地产开发投资，商品房销售面积，房地产 GDP，第三产业 GDP，商品房施工面积，商品房竣工面积，常住人口等 8 个指标，从图中本文可以看到随着迭代次数的进行拟合效果越来越好，因此本文可以考虑对指标做进一步分析。

六、 模型优化：LASSO-ARIMA 预测分析房价

(一) LASSO 模型

LASSO 回归本质上是一种压缩估计系数的方法。它通过构造一个惩罚函数得到一个相较于其它模型来说更为精炼的模型，通过不断训练压缩一些系数为零，以此来获得最优参数，并通过这些最优参数找出回归的系数，从而构建方程。因此 LASSO 回归不仅继承了子集收缩的优点，是一种处理具有复共线性数据的有偏估计的优良方法，数学表达式如下:

$$B_{LASSO} = arg_B min\left\{|Y - \sum_{j=1}^{p} X_j B_j|\right\}$$

$$s.t \sum_{j=1}^{p} |B_j| \leq t \#(14)$$

其中 t 为调整参数，通过调整参数 t 可以实现对总体回归系数的压缩。可以利用交叉验证法来估计 t 值。上述数学表达式还可以转换为最小化下述惩罚最小二乘法:

$$B_{LASSO} = arg_B min\left\{\left|Y - \sum_{j=1}^{p} X_j B_j\right|^2 + a\sum_{j=1}^{p} |B_j|\right\} \#(15)$$

其中 a 和 t 是一一对应的。LASSO 回归的主要优势在于对参数估计较大的变量压缩较小的同时将参数估计较小的变量压缩成 0，并且 LASSO 分析得到的参数估计具有连续性,比较适用于高维数据的模型来筛选变量。

同时，为了提高 LASSO 方法的相合性和准确性,目前常用自适应的 LASSO 方法，其把 LASSO 中的惩罚项修正为：

$$P_a(B) = \sum_{j=1}^{p} \frac{1}{|B_j|} |B_j| \#(16)$$

其中B_j是最小二乘估计系数。自适应 LASSO 分析的重要意义在于当样本量趋于无穷且变量个数维持不变时，其估计结果具有相合性，并且这些参数估计的结果与事先给定的非零变量位置的最小二乘法得到的参数估计的分布渐进相同[6]。直接将自适应 LASSO 的想法应用到水平压缩方差分析中，其数学表达式如下:

$$B = arg_B\{|Y - XB|\}$$

$$s.t \sum_{k=1}^{p_j} B_{jk} = 0, l = 1, \dots J$$

$$\sum_{j=1}^{J} \sum_{1 \le k < m \le p_j} w^{(km)} \left| B_{jk} - B_{jm} \right| \le t \#(17)$$

本文将基于以上原理进行计算。

(二) LASSO-ARIMA 模型算法实现步骤

1.选取 2011 年—2020 年北京西城区、东城区、海淀区、朝阳区、通州区的最终选定的 8 个因素数据来分别进行数据补足、转频和归一化处理。

2.用 Lasso 回归对处理后的数据进行回归预测，用本月的各个因素的数据与下个月的房价数据对齐来进行回归（由于房价波动对于政府政策或重大事件等扰动的滞后性）。由此得到每个区对这 8 个因素的回归方程：

$$\tilde{Y}_{i+1} = b_0 + \sum_{j=1}^{8} b_j x_{ij}$$

$$= b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} + b_6 x_{i6} + b_7 x_{i7} + b_8 x_{i8} \#(18)$$

3.将 2011 年—2020 年的所有月度因素数据代入回归方程并反归一化，得到 119 个模型预测值$\tilde{Y}_i, i = 2,3, \dots,120$与 119 个真实值$Y_i, i = 2,3, \dots,120$比较，计算

残差$\varepsilon_i = Y_i - \tilde{Y}_i, i = 2,3,\ldots,120$。

4.参考上述对时间序列预测模型的精度分析，选用 ARIMA 模型对残差$\{\varepsilon_i\}$序列进行预测，预测出下个月的残差值ε_{i+1}作为修正项。

5.修正回归方程为

$$\tilde{Y}_{i+1} = (b_0 + \sum_{j=1}^{8} b_j x_{ij})_{反归一化} + \varepsilon_{i+1}$$

$$= (b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} + b_6 x_{i6} + b_7 x_{i7} + b_8 x_{i8})_{反归一化} + \varepsilon_{i+1} \#(19)$$

6.当$i = 120$时可利用已知的

$X_{120} = (x_{120,1}, x_{120,2}, x_{120,3}, x_{120,4}, x_{120,5}, x_{120,6}, x_{120,7}, x_{120,8})$数据代入式(19)来预测出未知的下个月的房价。即将此数据代入修正后的回归方程中得到：

$$\tilde{Y}_{121} = (b_0 + \sum_{j=1}^{8} b_j x_{ij})_{反归一化} + \varepsilon_{121}$$

$$= (b_0 + b_1 x_{120,1} + b_2 x_{120,2} + b_3 x_{120,3} + b_4 x_{120,4} + b_5 x_{120,5} + b_6 x_{120,6} + b_7 x_{120,7} + b_8 x_{120,8})_{反归一化} + \varepsilon_{121} (20)$$

7.若已知真实值为Y_{i+1}，可与真实值计算相对误差$\eta = \left|\frac{\tilde{Y}_{i+1} - Y_{i+1}}{Y_{i+1}}\right| \times 100\%$，若$\eta < 5\%$可认为预测合格，若$\eta < 3\%$可认为预测精确。

(三) LASSO-ARIMA 模型实证分析

基于以上的 LASSO-ARIMA 集成模型的构建过程，对所有城区的房价月度预测公式进行建立，包括影响因素系数的确立以及残差项的计算，得到 2021 年 1 月的各区房价预测值。本部分将展示各区的预测公式及残差预测曲线，进行模型与对应区发展状况的综合分析。集成模型的稳健性、与传统回归模型相比的优势则将在下部分论述。

1.西城区

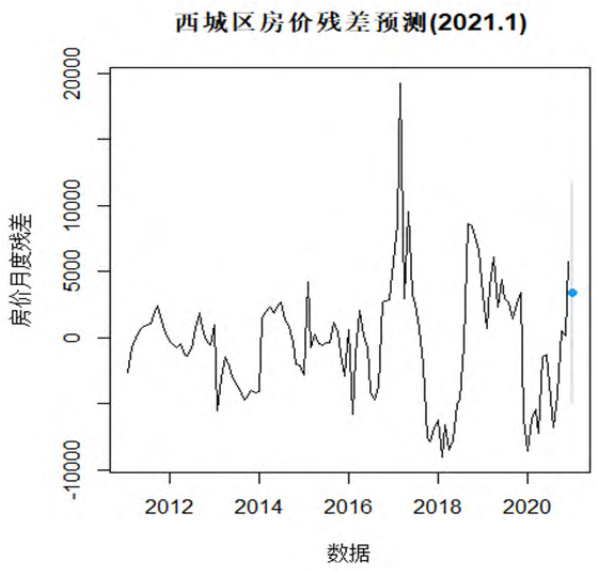


图 10 西城区残差预测图

月度房价回归表达式：

$$\tilde{Y}_{i+1} = [-0.294502 + 2.088061x_{i1} + (-0.44571)x_{i2} + (-0.498963)x_{i3} + (-0.303252)x_{i4} + (-0.647294)x_{i5} + 0.450982x_{i6} + (-0.105685)x_{i7} + 0.549736x_{i8}]_{反归一化} + 4228.695$$

西城区平均房价位于北京市之首，且仍有增长趋势。西城区区内有金融街及众多政府职能部门，区域 GDP 较高，人员流动量较大。同时房地产 GDP 及商品房竣工面积较为靠前。结合西城区城市发展规划，近十年来西城区拆迁了众多老旧胡同，同时加强交通疏堵工程，治理自行车道，新建立体停车位、增设绿地公园等举措提高了西城区的宜居性，使得西城区房价一直较为平稳地增长。

2.东城区

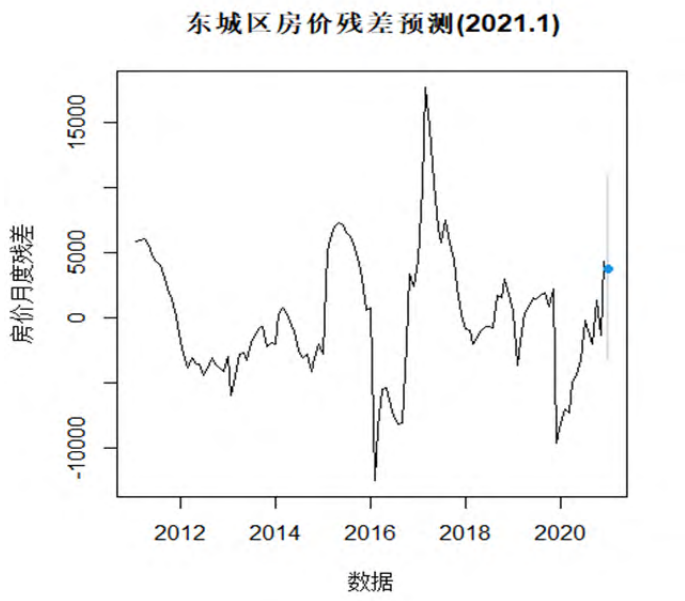


图 11 东城区残差预测图

月度房价回归表达式：

$$\tilde{Y}_{i+1} = [4.693629 + 0x_{i1} + (-0.359523)x_{i2} + (-1.783741)x_{i3} + 1.297679x_{i4} + (-4.464629)x_{i5} + (-0.072201)x_{i6} + (-0.049422)x_{i7} + (-2.832476)x_{i8}]_{反归一化} + 3758.96$$

东城区是平均房价较高的城区，和西城一样作为中心城区有着地理位置优越、历史文化悠久的特点。商品销售面积作为主要影响因素之一，很好的反应地理特点。由于东城区与西城区同为三环内城区且不会向外扩张，加之根据房地产租赁网站的发布信息显示，其城区住房开发基本饱和。故而，东城区未来的房地产交易将以二手房为主。优越的区域 GDP 为东城区吸引来大量投资及人才，加之北京市经济中心的东移趋势，东城区房价仍有很大的升值空间。

3.海淀区

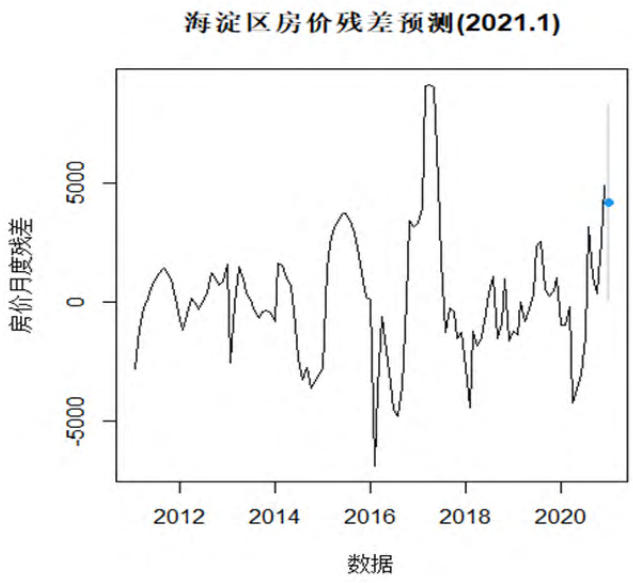


图 12 海淀区残差预测图

月度房价回归表达式：

$$\tilde{Y}_{i+1} = [-0.06890 + 0.436406x_{i1} + (-0.045598)x_{i2} + (-0.599826)x_{i3} + 0.113146x_{i4} + 0x_{i5} + 0.572578x_{i6} + (-0.144654)x_{i7} + 0.474559x_{i8}]_{\text{反归一化}} + 4228.695$$

海淀区是主要的科技产业园区与教育园区集中地，但是，由于其面积较广，距离市中心距离不及主城区，其房价维持在中高段。由于海淀区内有众多知名学府，它的房价主要体现在围绕重点学区阶梯型分布的特点。“学区房”是海淀区房地产市场的关键词，许多外区人涌入海淀区，增大其人口居住密度，也由此哄抬提高了房价到近乎非正常水平。提高海淀区人文层面的影响，使该区全面发展吸引高素质人才，是政府亟待考虑之处。

4.朝阳区

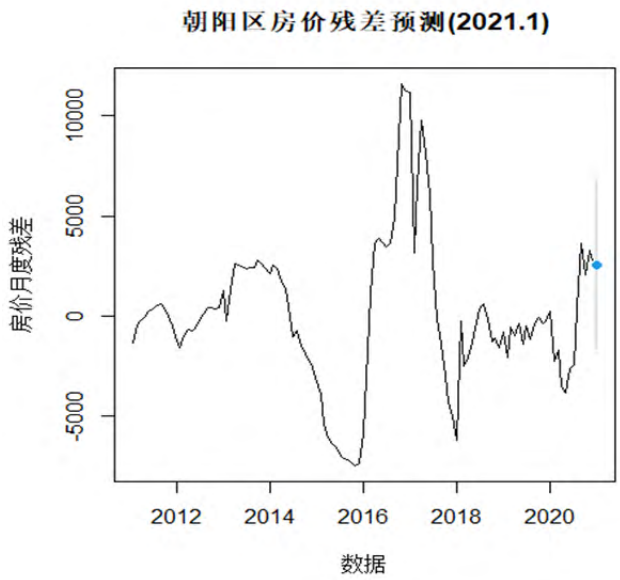


图 13 朝阳区残差预测图

月度房价回归表达式：

$$\tilde{Y}_{i+1} = [0.747174 + (-0.179866)x_{i1} + 0.410958x_{i2} + (-0.236948)x_{i3} + 0.241610x_{i4} + 0x_{i5} + 0.146239x_{i6} + (-0.933596)x_{i7} + (-0.222547)x_{i8}]_{反归一化} + 2553.787$$

朝阳区是一个年轻的城区，承担着北京市的科技产业与创新创业孵化职能，被誉为未来的北京硅谷。过去，按照朝阳区的规划，朝阳的双井、大望路、望京等商圈随着城市规划逐渐承担起向西连接首都功能核心区，向东连接城市副中心的职能。显著的 GDP 增长为朝阳区吸引了大量人员涌入，从而使房价在 2012、2015 年等小范围跨越式提升。朝阳区以望京为首的从“睡城”转型，吸引了大量房地产开发商。

5.通州区

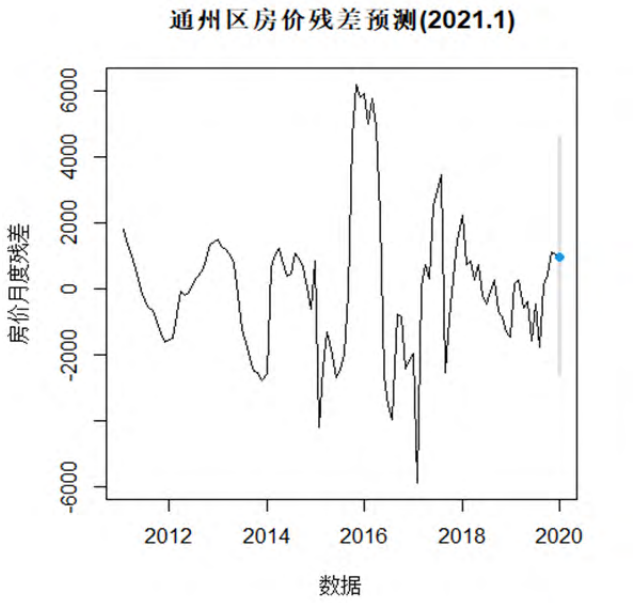


图 14 通州区残差预测

月度房价回归表达式：

$$\tilde{Y}_{i+1} = [0.619573 + 0x_{i1} + (-0.033136)x_{i2} + (-0.530798)x_{i3} + 0.015354x_{i4} + 1.008851x_{i5} + 1.340547x_{i6} + (-1.063319)x_{i7} + (-1.191870)x_{i8}]_{\text{反归一化}} + 976.3933$$

通州区在未来的过去十年里，从一个北京市郊区转型为承担成为北京市城市副中心的城区。政策对房价有着积极引导——对于通州区，商品房销售面积是房价最重要的影响因素。许多城中村被拆迁改建为商品房售卖。未来，北京市政府部门将逐渐向通州区转移，房价仍有很大的上升空间。

6.总体分析

由以上对北京市五个城区月度房价的观察及预测可以从中总结出各区发展的普遍规律：

①北京市各区房价在总体上保持着增长趋势。作为首都，始终源源不断地吸引着外来人才。北京市房价在 2010-2015 年增长较为平缓，2019 年之前处于上涨阶段且前期速度较快，中后期速度减缓,但总体呈现上升趋势。

②北京市房价成交价格呈现明显的圈层城区结构。区位、经济、文化等是城区吸引力的代表。区位方面，离天安门中心区域越远房价越低。经济方面，朝阳与通州的崛起，引发了北侧房价高于南侧，西侧高于东侧的现象。炒房现象层出不穷，在此趋势下,房地产形势愈发严峻，在此情况下的阶层贫富差距也由此产生，应当引发关注。

③北京市住房价格在 2020 年经历新冠疫情重大卫生事件后，房价涨跌起伏不定，波动性较强。此趋势下,房地产形势也愈发严峻：可能在某个方面（例如政府的政策）的一个小扰动会给房价带来很大的影响，给预测工作也带来了极大的挑战。

总之，房地产对于国家经济来说，牵一发动全身，具体体现在其不易描述的波动性和不可预知的滞后性上。对于北京市的总体发展，本文提出几点建议：第一，规范房地产产业，力求创造健康稳定的房地产市场，提高居民幸福感。第二,加大经济适用房建设的扶持力度,缓解住房压力，吸引优秀人才进入公务员团队为城市建设注入活力。第三,科学地制定限购等相关政策，既为降低房地产泡沫风险，也控制“富人越富，穷人越穷”的阶级分化现象，营造和谐良好地社会氛围。

(四)　模型的实证及优越性分析

本文选用传统模型多元线性回归与 LASSO-ARIMA 模型对比，将两者预测出的 2021 年 1 月房价数据进行精度的考察，以相对误差为考察依据。得到下图

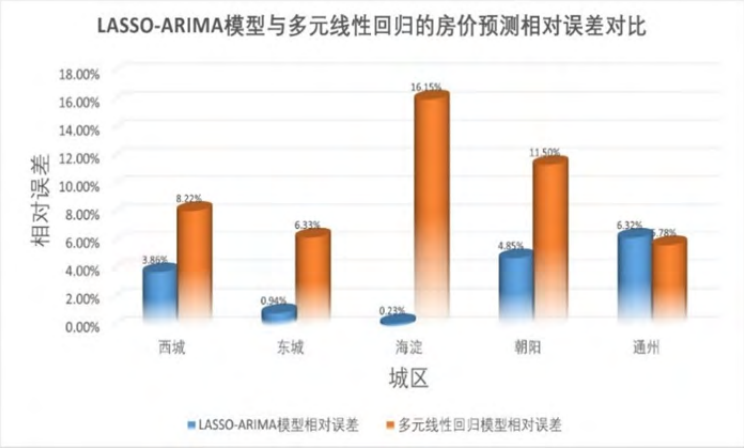


图 15　LASSO-ARIMA 预测和多元线性回归误差对比

通过预测各城区 2021 年的月度房价数据，得到集成模型的相对误差分别为 3.86%、0.94%、0.23%、4.85%、6.32%，大致在 5%以内。作为对照模型，多元线性回归的相对误差为 8.22%、6.33%、16.15%、11.50%、5.78%，在 10%左右。

LASSO-ARIMA 模型结合了 LASSO 和 ARIMA 模型的优势。LASSO 模型可以在迭代的过程中筛选变量，将影响市场的环境因素考虑其中，得到回归方程。但是，由于其本质是多元线性回归模型，对房价这类具有波动性特点的序列预测精度有待提高。ARIMA 模型能够提取波动时间序列的信息，作为残差补足项，它有效减小了回归方程得到的预测值与真实值的误差，提高预测精度。

最为关键的是，与传统回归方程相比，本模型引入“滞后”这一概念，即用当月变量数据与下月因变量相回归，真正实现了房价的提前预测。

综上所述，本文的集成模型相比于传统统计模型在准确性、实时性方面均表现优越，其预测的月度房价数据可信度较高。

七、 模型评价反思

本模型是基于时间序列预测和基于多因素回归预测的结合，对于描述房价的时间波动性与时序性优于仅仅用多因素回归的预测；对于描述房价对于重大事件的突变性优于仅仅用时间序列的预测，达到了时序与因素的合二为一。

当经济稳步发展时，房价的时间波动性大于突变性；当出现突发事件，或对于某因素有扰动时，房价的突变性大于时间波动性。因此，灵活运用本模型才是预测房价的上佳选择。

本模型不足之处在于，对于基于时序和多因素回归预测的误差没有把控，多次预测的误差是否处在一个可以控制的区间内。本文并没有给出一个误差的范围。并且通过 ARIMA 模型补足的数据也不够精确。

由于搜集的信息有限，不能获得房价相关影响因素的详细月度数据，造成一定误差。因此，只选取了 2011 年—2020 年的数据进行模型的构建且仅预测了 2021 年 1 月的房价数据，统计学评价指标选取较少，模型说服力有待加强。

参考文献

引文文献

[1]纪昀瑛.北京房价的思考研究与总结[J].全国流通经济,2017(08):63-64.

[2]张望舒,马立平.城市二手房价格评估方法研究——基于 Lasso-GM-RF 组合模型对北京市二手房价格的分析[J].价格理论与实践,2020(09):172-175+180.

[3]崔庆岳,赵国瑞.基于灰色 GM(1,1)模型的商品房销售价格预测[J].哈尔滨商业大学学报(自然科学版),2019,35(02):253-256.

[4]王景行.基于回归的房价预测模型研究[J].全国流通经济,2020(19):120-122.

[5]徐征,洪明月,宋玉.基于“人才引进”政策的西安楼市现状探究及房价预测[A].中国统计教育学会、教育部高等学校统计学类专业教学指导委员会、全国应用统计专业学位研究生教育指导委员会.2019 年（第六届）全国大学生统计建模大赛优秀论文集[C].中国统计教育学会、教育部高等学校统计学类专业教学指导委员会、全国应用统计专业学位研究生教育指导委员会:中国统计教育学会,2019:56.

[6]张望舒,马立平.城市二手房价格评估方法研究——基于 Lasso-GM-RF 组合模型对北京市二手房价格的分析[J].价格理论与实践,2020(09):172-175+180.

[7]李彬生,梁禹涵.二孩政策下深圳市房价影响因素分析——基于灰色关联度[J].韶关学院学报,2020,41(04):59-64.

[8]王顺钢,李文蝶,杨佳亦. 基于特征工程的 SVM 北京雾霾成因分析及预测[A]. 中国统计教育学会、教育部高等学校统计学类专业教学指导委员会、全国应用统计专业学位研究生教育指导委员会.2019 年（第六届）全国大学

生统计建模大赛优秀论文集[C].中国统计教育学会、教育部高等学校统计学类专业教学指导委员会、全国应用统计专业学位研究生教育指导委员会:中国统计教育学会,2019:42.

[9]黄诗琦.基于 GARCH 模型族的房价波动研究——以北京市为例[J].商讯,2019(12):1-2+8.

阅读型文献

[10]郑永坤,刘春.基于 ARIMA 模型的二手房价格预测[J].计算机与现代化,2018.

[11]赵泰,迟建英.灰色 GM(1,1)模型在商品房销售价格预测中的应用[J].价值工程,2019(23):76-78.

[12] 李志超, 刘升. 基于 ARIMA 模型,灰色模型和回归模型的预测比较[J].统计与决策, 2019, v.35;No.539(23):40-43.

[13]吴承业,沈逸珺,汪慰,曹远寿,周婉茹,孟庆欣.基于 ARMA 模型的杭州市房价研究与预测——以杭州市上城区和下城区为例[J].湖州师范学院学报,2020,42(08):19-26+33.

[14]李广胜,郭欢.基于 GM(1,1)模型的南京市房价预测研究[J].江汉大学学报(自然科学版),2020,48(02):10-13.

[15]朱青,周红.基于灰色预测 GM(1,1) 模型的河南省房价预测[J].周口师范学院学报,2020,37(02):37-40.

[16]田润泽.基于多种机器学习算法的波士顿房价预测[J].中国新通信,2019,21(11):228-230.

[17]任梓铭.基于灰色预测和回归模型的北京城区房价预测研究[J].现代商贸工业,2019,40(10):101-102.

[18]王蕾,刘佳杰.基于 ARIMA 模型的保定市商品房价格预测研究[J].产业

与科技论坛,2019,18(09):96-98.

[19]刘美辰.基于卡尔曼滤波法的房价波动影响因素分析——来自于山东省2000—2014 年的数据[J].现代经济信息,2017(11):479-481+483.

[20]张砚博.基于多元线性回归分析的西安市房价预测分析[J].西部皮革,2020,42(10):71.

[21]郑永坤,刘春.基于 ARIMA 模型的二手房价格预测[J].计算机与现代化,2018(04):122-126..

[22]张家棋,杜金.基于 XGBoost 与多种机器学习方法的房价预测模型[J].现代信息科技,2020,4(10):15-18.

[13]汪静,罗维平,陈永恒.基于神经网络的房价预测与分析[J].襄阳职业技术学院学报,2021,20(02):112-115+140.

[24]曾婷婷.基于机器学习的房价预测模型研究[D].西南科技大学,2020.

附录

1．GM(1,1)的月度区域房价预测误差

表格 东城区原始数据和 GM 预测值

时间	原始数据	灰色预测	相对误差	后验差比值	MSE
2018.1	102918.0000	114598.6187	11.35%	0.29114	4.45E+09
2018.2	102018.0000	116456.1659	14.15%	0.29114	4.45E+09
2019.1	105853.0000	123375.7275	16.55%	0.31599	5.33E+09
2019.2	103503.0000	125061.1457	20.83%	0.31599	5.33E+09
2020.1	97810.0000	128603.4450	31.48%	0.35691	6.05E+09
2020.2	98334.0000	130080.4724	32.28%	0.35691	6.05E+09

表格 海淀区原始数据和 GM 预测值

时间	原始数据	灰色预测	相对误差	后验差比值	MSE
2018.1	86897.0000	93182.9148	7.23%	0.30595	3.30E+09
2018.2	84181.0000	94489.2949	12.25%	0.30595	3.30E+09
2019.1	87951.0000	101097.8989	14.95%	0.30196	3.38E+09
2019.2	87442.0000	102340.8677	17.04%	0.30196	3.38E+09
2020.1	86855.0000	105135.4588	21.05%	0.34048	4.35E+09
2020.2	86748.0000	106232.7267	22.46%	0.34048	4.35E+09

表格 朝阳区原始数据和 GM 预测值

时间	原始数据	灰色预测	相对误差	后验差比值	MSE
2018.1	73419.0000	78066.3880	6.33%	0.34071	2.31E+09
2018.2	74927.0000	79165.2608	5.66%		
2019.1	72573.0000	84416.0771	16.32%	0.33612	2.72E+09
2019.2	71613.0000	85449.0488	19.32%		
2020.1	71355.0000	86374.5012	21.05%	0.38969	3.00E+09
2020.2	68542.0000	87247.8850	27.29%		

表格 通州区原始数据和 GM 预测值

时间	原始数据	灰色预测	相对误差	后验差比值	MSE
2018.1	44963.0000	52135.1320	15.95%	0.35858	9.20E+08
2018.2	38666.0000	53121.0143	37.38%	0.35858	9.20E+09
2019.1	46431.0000	55524.7238	19.59%	0.39106	1.11E+09
2019.2	46520.0000	56361.8019	21.16%	0.39106	1.11E+09
2020.1	44313.0000	56551.1803	27.62%	0.44313	1.24E+09
2020.2	45739.0000	57229.0542	25.12%	0.44313	1.24E+09

2．ARIMA 的月度区域房价预测

表格 东城区原始数据与 ARIMA 预测值

时间	原始数据	灰色预测	相对误差	后验差比值	MSE
2018.1	102918.0	100000.0	2.84%	>0.05	3.44E+06
2018.2	102018.0	100000.0	1.98%	>0.05	3.44E+06
2019.1	105853.0	107323.0	1.39%	>0.05	2.66E+06
2019.2	103503.0	107189.0	3.56%	>0.05	2.66E+06
2020.1	97810.0	92669.0	5.26%	>0.05	3.81E+06
2020.2	98334.0	88038.0	10.47%	>0.05	3.81E+06

表格 海淀区原始数据与 ARIMA 预测值

时间	原始数据	灰色预测	相对误差	后验差比值	MSE
2018.1	86897.0	89186.0	2.63%	>0.05	1.46E+06
2018.2	84181.0	89673.0	6.52%	>0.05	1.46E+06
2019.1	87951.0	86233.0	1.95%	>0.05	2.18E+06
2019.2	87442.0	85567.0	2.14%	>0.05	2.18E+06
2020.1	86855.0	89146.0	2.64%	>0.05	2.10E+06
2020.2	86748.0	89236.0	2.87%	>0.05	2.10E+06

表格 朝阳区原始数据与 ARIMA 预测值

时间	原始数据	灰色预测	相对误差	后验差比值	MSE
2018.1	73419.0	74281.6	1.17%	>0.05	9.63E+05
2018.2	74927.0	74017.5	1.21%	>0.05	9.63E+06
2019.1	72573.0	71526.1	1.44%	>0.05	1.11E+06
2019.2	71613.0	71247.5	0.51%	>0.05	1.11E+06
2020.1	71355.0	71100.8	0.36%	>0.05	1.14E+06
2020.2	68542.0	71049.6	3.66%	>0.05	1.14E+06

表格 通州区原始数据与 ARIMA 预测值

时间	原始数据	灰色预测	相对误差	后验差比值	MSE
2018.1	44963.0	44896.0	0.15%	>0.05	2.93E+06
2018.2	38666.0	44896.0	16.11%	>0.05	2.93E+06
2019.1	46431.0	45543.0	1.91%	>0.05	2.56E+06
2019.2	46520.0	45543.0	2.10%	>0.05	2.56E+06
2020.1	44313.0	44594.0	0.63%	>0.05	2.40E+06
2020.2	45739.0	44594.0	2.50%	>0.05	2.40E+06

致谢

感谢大赛组委会举办此次大赛，感谢学校学院的倾力组织，指导老师和学姐学长的悉心指导，给予小组宝贵的实践机会，学以致用，挖掘出数据中的新动能。