

遼寧大學

毕业论文（设计）



题 目： 上海市二手房价格的评估与预测

学 院： 经济学院

专 业： 经济统计学

姓 名： 刘智珺

指导教师： 郭万山教授

完成日期： 2018 年 5 月 29 日

摘 要

随着房地产市场日趋完善，人们对房地产信息的需求越来越迫切。由于城市可供开发的土地有限，人们对二手房的交易愈发重视，因此对二手房价格的评估可以为消费者、开发商以及政府提供决策依据，同时可以进一步丰富和完善房地产价格评估的相关理论。

以上海市为例，文章采用了数据挖掘的方法，从链家网抓取了上海市2413套二手房源的信息。同时借鉴外国房地产评估的经验，并引入了特征价格理论，从区位特征、建筑特征以及邻里环境这三个方面选取了一些消费者比较关注的特征变量，包括房屋面积、板楼建造年份、是否免营业税、所在区域与是否靠近地铁这五个特征变量。接着，文章在bootstrap抽样的基础上，利用这些数据建立了随机森林回归模型，并基于permutation随机置换的残差均方减小量计算出了每个特征变量的重要性得分，从而对二手房价格进行评估。在模型的评估方面，文章采用了简单交叉验证的方法，将原样本按4:1的比例划分为训练集和测试集。文章将随机森林回归模型与传统的线性回归模型进行了比较，主要是通过计算这两种模型在训练集与测试集上的拟合优度和均方误差以及利用两种模型在测试集的进行单个样本偏差检验。

随机森林回归模型结果表明在所有特征变量中，对房屋价格影响最大的是房屋面积其次是所在地区，其次是板楼建造年份，最后是是否靠近地铁以及是否免营业税。其中后三者对二手房价格的影响不明显。从模型评估结果来看，无论是拟合优度还是均方误差，随机森林回归模型在训练集和测试集上的评估结果都优于传统的线性回归模型。测试集上的单个样本偏差检验结果也表明，用随机森林回归模型预测的房价与实际观测值匹配度更高，绝对误差更小。

最后，文章得出了以下结论：一方面，买房者和卖房者可以利用随机森林回归模型对房价进行评估并参照模型的特征变量重要性排序买卖二手房，使得自身的利益最大化；另一方面，文章进一步证明了随机森林回归模型的拟合效果和预测效果明显优于传统的线性回归模型。

关键词：二手房；价格评估；数据挖掘；特征价格理论；随机森林回归；交叉验证

Abstract

With the development of the real estate market, people demand more and more information about housing price. Because land for people to develop in cities is limited, consumers lay more emphasis on the second-hand house transaction. Consequently, the evaluation of the price of second-hand houses is essential to both consumers and developers, and the evaluation is also beneficial to governments' decision-making. Besides such research can enrich the theory related to evaluation of housing price.

This paper obtained information of 2413 second-hand houses from *Lianjia* Website, with the technique of data mining. Learning from the experience of foreign researchers and hedonic price theory, this paper selected some characteristic variables from three dimensions——Location, Structure and Neighborhoods, such as area, built year, whether subway nearby and so on. Next, based on bootstrap sampling, the paper built random forest regression model with this data, and calculated the score of residual mean square of each characteristic variable by random permutation in order to evaluate the price of second-hand houses. In order to evaluate the regression model, this paper used cross validation to split the sample into a training set accounting for 80 percent and a test set accounting for 20 percent. The paper also compared random forest regression model and lineal regression model by calculating the R-square, mean square error and bias test.

The regression results indicated that area had the maximum influence on the price of second-hand houses, and then location, the built year, whether subway nearby, whether sales tax required. And the last three characteristic variables had little influence on the housing price. The evaluation of two models indicated that random forest regression model was better than traditional linear regression model in all the aspects mentioned above. Besides, the result of bias test showed that random forest regression model could predict more accurately than linear regression model.

Finally, the paper drew two conclusions. For one thing, sellers and buyers can use random forest regression model to evaluate the housing price so that they can maximize their benefits. For another, this paper further demonstrated that random forest regression model had better fitting effect and better ability to predict housing price.

Key words: second-hand house; price evaluation; data mining; hedonic price theory; random forest regression; cross validation

目 录

摘 要	II
Abstract.....	II
一、绪论	1
1. 选题背景.....	1
2. 文献综述.....	1
3. 研究目的与研究意义.....	2
4. 研究方法.....	3
二、相关理论与方法	3
1. 特征价格理论.....	3
2. 随机森林回归模型.....	4
三、描述性统计分析	5
1. 指标选择.....	5
2. 数据来源.....	6
3. 描述统计.....	6
四、实证分析	8
1. 理论模型的设定.....	8
2. 数据来源及处理.....	9
3. 实证结果.....	10
五、研究结论与展望	13
1. 研究结论.....	13
2. 未来展望.....	13
参考文献	15
附录 A	16
附录 B	18
附录 C	19
附录 D	20
致谢	错误!未定义书签。

上海市二手房价格的评估与预测

一、绪论

1. 选题背景

房地产行业是我国国民经济的支柱产业，房地产行业提供的房地产商品在人民生活中有着举足轻重的地位，同时也是其他行业极为重要的生产资料。除此之外，房价的波动也是政府部门十分关注的问题。一方面，房地产行业的平稳发展有利于稳定就业岗位，减轻就业压力。另一方面，房地产的开发与投资能有效带动相关产业的发展，从而对新常态下的国民经济起到支撑作用。

虽然与房地产相关的经济活动越发频繁，但是城市所能提供开发的土地越来越少且成本越来越高，因此二手房交易变得逐渐活跃起来，人们对二手房价格评估的需求也不断增加。二手房的价格的评估有利于政府制定更为合理的政策，如学区房的划分与变动等。无论是从买房者、卖房者或是政府的角度来看，二手房房价的衡量都是个重要的话题，因此二手房价格的评估以及预测成为一项重要的研究课题。

目前，国内学者使用较多的房价评估方法包括市场法、成本法和收益法这三种传统方法。在用传统方法进行评估的过程中有一些弊端：第一，评估需要搜集大量的信息；第二，传统方法难以体现房地产价格与其影响因素之间的非线性关系；第三，传统评估方法的准确性依赖于评估师的个人经验和水平。

因此为了改善房地产评估现状，文章采用数据挖掘的方法抓取了链家网上二手房源的相关信息，并建立随机森林回归模型对二手房价格进行评估和预测，该方法可有效地降低主观随意性。

2. 文献综述

过去，我国主要凭借以往的经验对住宅价格进行评估，近年来相关学者开始逐步将数学模型引入房产评估研究之中。刘晓群（2008）调查了武汉的 19 个楼盘，并用数据建立了特征价格模型，接着用 SAS 软件对其进行主因素分析，得出特征价格模型的线性函数，同时还推断具体楼盘的住宅均价并检验与真实情况的差异。王卓琳（2009）从特征价格理论为框架，利用线性回归模型对北京市住宅特征价格进行分析，并对模型结果作出经济意义上的解释，为政府以及开发商等提供决策依据。曾昭法，唐海滨（2010）在对二手房价格进行评估的研究中，在实例计算的基础上推导出二手

房价格模糊综合评价模型，该模型得到了消费者和开发商的广泛认可。郭志强（2013）利用支持向量机批量评估房地产价格，发现其评估效果明显优于 BP 神经网络的方法。

在房地产价格评估方面，国外已经发展并应用了多种技术与方法，如神经网络、支持向量机、随机森林等方法。Elaine Worzala, Margarita Lenk, Ana Silva（1995）运基于获得的 288 家交易的房屋数据，用神经网络方法来评估房地产价格，并将该方法与传统的多元回归模型进行了对比。Nguyen Nghiep, Cripps Al（2001）用不同规模的数据多方位地对比了神经网络模型和多元线性回归模型，得出了对于中等规模以上的数据来说，运用神经网络模型进行预测的效果要明显好于多元线性回归模型的预测结果。K.C. Lama, C.Y. Yua（2010）使用了支持向量机的方法对房地产进行评估和预测，并与多元回归和人工神经网络模型的预测结果进行对比，发现应用支持向量机建立的房价预测模型效果更好。Evgeny A. Antipov, Elena B. Pokryshevskaya（2012）首次应用随机森林模型对住宅价格进行评估，并通过实证研究发现，该模型的评估表现优于回归树、多元回归以及神经网络模型。

3. 研究目的与研究意义

（1）研究目的

文章在借鉴了外国学者研究成果的基础上并结合了上海市二手房的现状探讨了以下两个问题：一是建立随机森林模型对上海市二手房价格进行评估并确定主要影响因素，二是进一步验证随机森林模型预测能力的准确性。

（2）研究意义

在理论价值方面，文章与以往研究相比最大的创新之处在于引入了随机森林回归模型，对影响二手房价的各种因素进行了分析，丰富了特征价格理论框架。文章以上海市二手房数据为例，进一步验证了随机森林模型的优越性。

在实践价值方面，传统的二手房价格评估方法需消耗大量人力物力，程序复杂，且评估的准确度难以衡量。实证表明，随机森林回归模型在大量数据的基础上建立的二手房房价和特征变量之间的回归模型在模型的拟合效果、回归误差以及预测准确性方面都比传统的线性回归模型有着更好地效果。因此文章中采用的评估方法以及计算出的特征向量重要性对人们的决策具有参考价值。对消费者而言，文章的方法和结论有助于他们挑选到性价比高，且具有投资价值的房源；对于卖房者来说，有利于他们售房时指定合理的价格；对政府而言，文章的研究结果可为其制定政策调控房价出谋

划策。

4. 研究方法

文章主要采取了以下几种研究方法：

（1）实证研究

文章的实证研究实在规范的研究基础上进行的。实证所需要的数据均来自 2018 年 4 月 2 日链家网上的上海市二手房数据。实证研究的过程中，文章通过建立随机森林回归模型计算了各特征变量的重要性得分并对二手房价格进行了预测。文章在实证研究的过程中采用了对比分析的方法，将随机森林回归模型与线性回归模型进行比较，证明了随机森林算法在房价的评估和预测方面具有良好的性能，得到的结论具有应用价值。

（2）机器学习方法

随机森林回归模型需要借助计算机编程软件来实现，文章选取 Python 进行编程。首先用 Python 中的 Requests 和 PyQuery 模块编写爬虫获取相关数据。接着，文章将原样本按 4:1 的比例随机分为训练集和测试集。在 bootstrap 抽样的基础上，利用训练集数据建立了随机森林回归模型，并基于 permutation 随机置换的残差均方减小量计算出了每个特征变量的重要性得分，从而对二手房价格进行评估。文章在建立随机森林模型时主要用到的是 sklearn 模块中的 RandomForestRegressor() 函数。

二、相关理论与方法

1. 特征价格理论

Court 于 1939 年首次提出了特征价格理论（Hedonic Pricing Theory），该理论认为商品的价格由一系列影响商品价格的因素决定，即商品价格 $p=F(x)$ ，其中 x 为一列影响商品价格的因素。该理论最先被用于农场土地价格的分析，后来用于分析具体的商品，如汽车、果蔬产品等。1974 年，Rosen 首次将特征价格理论用于房地产领域，并提出了住宅价格特征模型，此后越来越多的学者开始关注该模型在房地产领域的应用。

1982 年，学者 Butler 在已有的住宅价格评估模型基础上加入了影响房价的三类特征变量，分别为区位特征（Location）、建筑特征（Structure）以及邻里环境（Neighborhoods）。区位特征是指住宅所处的城市区域，通常情况下用该区域到市中

心的距离量化这一特征。建筑特征是指住宅本身的特征，如：面积、房龄、装修、户型等。邻里环境是指住宅小区的自然环境、交通情况、人文环境、治安水平等。改进后的住宅价格评估模型的公式可以写成 $P = f(L, S, N)$ ，其中 L 为区位特征， S 为建筑特征， N 为邻里环境。

住宅特征价格理论在以下假设的基础上建立：

第一，市场上的房源具有异质性。所有的房源都可以用相同的特征变量进行衡量，但各自不同的特征导致每个房源都是独一无二的，因而使得每个房源对应的价格有高低。

第二，房源的每个特征对应一个隐含市场，房地产市场可以理解为若干个隐含市场的组合而成。

第三，市场是充分竞争的。市场上有大量的卖房者与买房者，每次交易对市场上总体的供求情况不产生显著影响，且不存在强制性的政策影响买房者与卖房者的交易行为。

第四，市场上的交易信息是充分透明的，且买房者在购房前会收集大量的信息。若市场上的信息不对称，则容易出现价格不合理的情况。

2. 随机森林回归模型

随机森林（Random Forest）是一种基于决策树分类器集成算法的组合预测模型，其中模型中的每一棵决策树都是由随机向量生成的，且所有向量都独立同分布。随后，1984 年 Breiman 等人以分类树的算法为基础，提出了 CART 算法（Classification And Regression Tree），使得决策树的方法可以同时应用于分类和回归问题。

随机森林回归模型通过 bootstrap 抽样技术，并由特征变量组成的随机向量生成若干棵回归树，构成组合模型。该模型将数值型变量作为解释变量，生成多元线性回归随机森林模型。随机森林回归模型原理如图 2.1 所示：

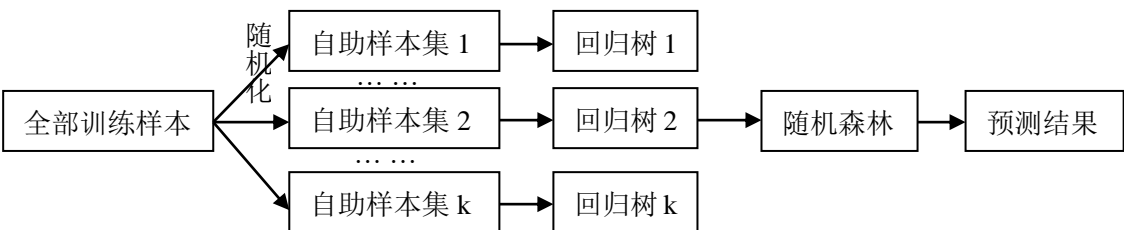


图 2.1 随机森林回归模型原理

相比与其它回归模型，随机森林模型有许多优点。首先，随机森林回归模型通过

随机选取不同样本并组合结果的方式,使得该模型即使在资料缺失的情况下也能有着较高的准确度。其次,将多个模型组合在一起可以有效地克服单个模型过度拟合的问题。第三,随机森林模型无需人工赋予各个特征变量权重,同时该算法可以快速处理大量数据。

三、描述性统计分析

1. 指标选择

文章选择房屋价格(Y)作为因变量,除此之外文章根据特征价格理论从区位特征、建筑特征以及邻里环境这三个方面选取了五个自变量。自变量房屋面积(X₁)、板楼建造年份(X₂)与是否免营业税(X₅)为建筑特征,自变量所在区域(X₃)为区位特征,自变量是否靠近地铁(X₄)为邻里环境特征。

需要注意的是,在所在区域(X₃)这一特征的量化过程中,文章以上海市中心所在区域黄浦区为中心,然后根据距离的远近给每个区划分等级。划分结果为:黄浦区地理位置最佳,其次是静安区、徐汇区和虹口区,然后是长宁区、普陀区和杨浦区,接下来是宝山区、闵行区和浦东区,最后是嘉定区、青浦区和松江区。

对于标注附近有地铁的房源,X₄的取值为1,表明附近有地铁,交通便利;对于未标注地铁信息的房源,X₄取值为0,表明附近无地铁。

对于标注年满两年或年满五年的房源,X₅取值为1,表明免交营业税;其他未标注的房源,X₅取值为0,表明需交营业税。

表 3.1 变量设定

变量类型	变量	变量名称	变量水平
因变量	Y	房屋价格	数值型
自变量	X ₁	房屋面积	数值型
	X ₂	板楼建造年份	数值型
	X ₃	所在区域	1=嘉定区、青浦区、松江区; 2=宝山区、闵行区、浦东区; 3=长宁区、普陀区、杨浦区; 4=静安区、徐汇区、虹口区; 5=黄浦区
	X ₄	是否靠近地铁	0=不靠近地铁; 1=靠近地铁赋
	X ₅	是否免营业税	0=需交营业税; 1=免交营业税

1) 划分所在区域时以黄浦区为中心,根据距离远近划分等级。

2. 数据来源

文章选取的数据均来自 4 月 2 日链家网 (<https://sh.lianjia.com>) 上检索到的上海市二手房信息, 经过分区检索后发现奉贤区、金山区和崇明区的房源数量过少, 因此仅选取上海市其它 13 个区的二手房房源数据作为研究对象。笔者用 Python 中的 Requests 和 PyQuery 模块以及正则表达式和 CSS selector 爬取了 2413 套房源信息, 并储存至 csv 文件中。

网站上的房屋数据主要有: 房源名称、位置、板楼建造年份、房屋价格、房屋均价、房屋面积、附近地铁情况、是否满两年等。笔者从中提取了其中主要的房源信息。

3. 描述统计

文章选取的样本数量较多, 因此在统计建模前对各指标数据进行描述统计可以更加直观的看出数据的分布情况, 加深对建模结果的理解。

从表 3.2 的结果可以看出上海市二手房价格总体来说较高, 均值高达 487.93 万元, 价格分布也较为分散, 极差达到 4380 万元。上海市二手房总体来说面积不大, 平均 84.15 平方米, 中位数为 74.72 万元。上海市二手房的板楼建造年份这一指标的标准差较小, 说明房屋建造年份较为集中, 其中 2000 年左右建造的房屋相对较多。

表 3.2 数值型变量的描述统计

变量	均值	中位数	极差	标准差
房屋价格 (Y)	487.93	380.00	4380.00	365.76
房屋面积 (X ₁)	84.15	74.72	453.74	46.66
板楼建造年份 (X ₂)	1999.21	1999.00	104.00	9.81

文章根据不同区的房屋面积 (X₁) 与房屋价格 (Y) 以及板楼建造年份 (X₂) 与房屋价格 (Y) 作了图 3.1 所示散点图。从房屋面积 (X₁) 与房屋价格 (Y) 的散点图中可以看出在面积相同的情况下地理位置越靠近市中心的二手房价格越高, 对于同一地区的二手房来说面积越大啊的房源价格越高。从板楼建造年份 (X₂) 与房屋价格 (Y) 的散点图中可以看出大多数二手房的板楼建造年份都集中在 1990 年之后, 且在地理位置好的二手房板楼建造年份都较早, 新建的二手房大多数都分布在离市中心较远的地区。

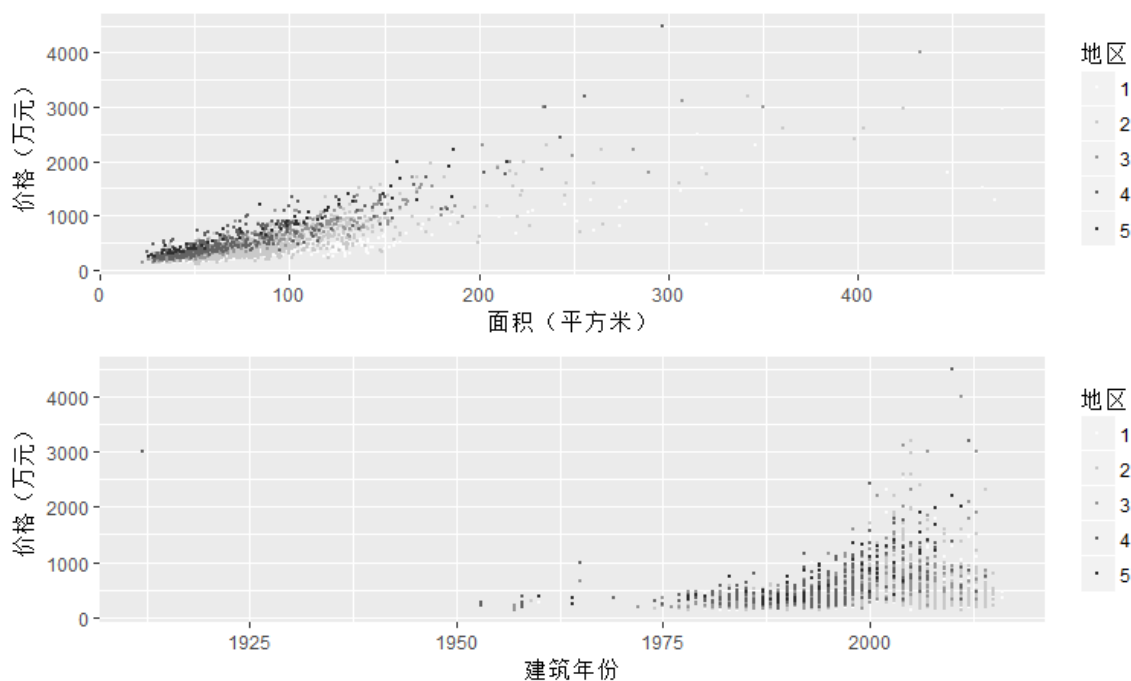


图 3.1 面积与建筑年份散点图

如图 3.2 以及表 3.3 所示，在所有的 2413 套房源中有 1186 套房源标注了附近的地铁信息，1227 套房源未标注地铁信息，交通便利的房源占有所有房源的 49%。在所有房源中仅有 275 套房源标注了年满五年或年满两年，即仅有 11% 的房源可以免交营业税，其余 2138 套房源都需要交营业税。

表 3.3 分类变量的描述统计

变量	均值	标准差
是否靠近地铁 (X_4)	0.49	0.50
是否免营业税 (X_5)	0.11	0.32

从图 3.2 以及表 3.4 的结果可以看出，宝山区、闵行区和浦东区的二手房源数量最多，占有所有房源数量的 47.8%，其它地理位置优越或低于这三个区的地区所在链家网挂牌的二手房数量都较少，并且呈现出地理位置越偏以及地理位置越佳的二手房数量都越小的特点。

表 3.4 所在区域的描述统计

	1	2	3	4	5
频数	399	1154	479	305	76
频率	0.165	0.478	0.199	0.126	0.031

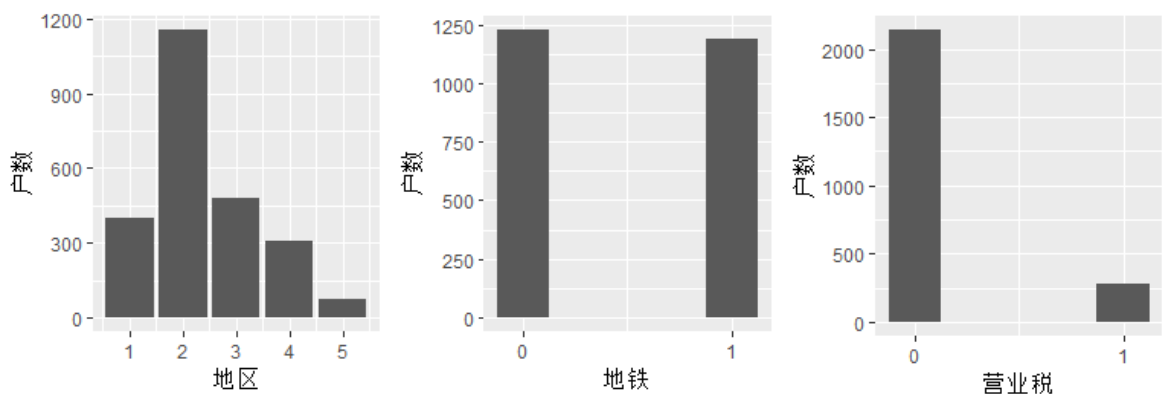


图 3.2 分类变量描述统计

从图 3.3 可以看出，首先地理位置越好的二手房价格整体水平要明显高于地理位置较偏的二手房价格，且嘉定区、青浦区以及松江区的二手房房价分布相对于其他地区的房价较为集中。其次，靠近地铁的二手房价格水平总体高于不靠近地铁的房源，且价格分布更为离散。最后需交营业税的二手房源价格分布比无需交营业税的二手房源价格分布更为离散。

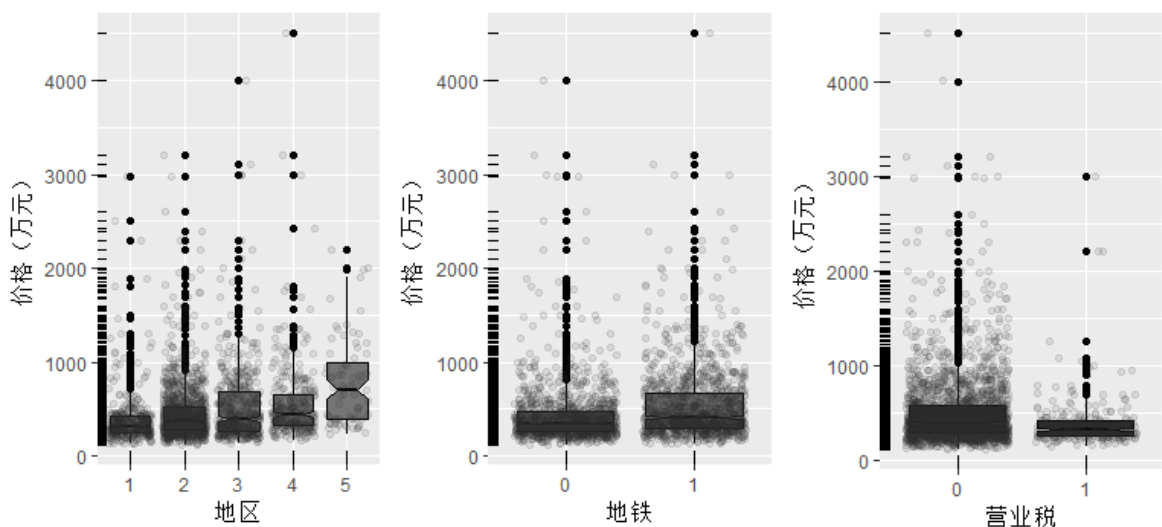


图 3.3 分类变量与房价的箱线图

四、实证分析

1. 理论模型的设定

文章选用的基于 CART 算法的随机森林模型原理如下：

随机森林回归模型是一种非线性回归模型，设模型中共有 K 棵树 $\{T_1(X), T_2(X), \dots, T_K(X)\}$ ，其中 $X = \{x_1, x_2, \dots, x_p\}$ ，是形成森林的 p 维特征向量，每棵树产生一个预测值 $\hat{Y}_i (i = 1, 2, \dots, K)$ 。

(1) 运用 bootstrap 方法随机抽取 K 个样本集用于生成 k 棵树，即采用重复抽样的方法从原样本中随机抽取 K 个子样本。

(2) 在每棵回归树生长过程中，随机选取所有特征变量的一个子集，依据基尼系数 (Gini index) 来选择最优划分属性。基尼系数是用于测算特征变量数据纯度的指标，基尼系数越大说明样本集合的不确定性越大。公式 (4.1) 中假设特征变量共有 M 个类别， P_m 是样本点属于第 m 个类别的概率。因此优先选择基尼系数小的特征变量作为划分属性。

$$Gini(p) = \sum_{m=1}^M p_m(1-p_m) = 1 - \sum_{m=1}^M p_m^2 \quad (4.1)$$

(3) 每棵回归树以损失函数最小为原则构建每棵回归树。

首先，根据公式 (4.2)，选择最优切分点 s 对固定特征变量 j 的阈值进行划分。

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right] \quad (4.2)$$

其次，根据划分结果计算相应的输出值。

$$R_1(j,s) = \{x | x^{(j)} \leq s\}, R_2(j,s) = \{x | x^{(j)} \geq s\} \quad (4.3)$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m(j,s)} y_i, x \in R_m, m = 1, 2 \quad (4.4)$$

最后，重复以上两个步骤进行递归，直至满足设定的最大深度后停止生长。假设已将所有输入空间划分为 M 个单元 R_1, R_2, \dots, R_M ，且每个单元的固定输出值为 c_m ，则模型的输出结果如公式 (4.5) 所示。

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m) \quad (4.5)$$

(4) 将所有回归树预测结果进行简单算术平均得到模型的输出结果，结果如公式 (4.6) 所示。

$$Y = \frac{1}{K} \hat{Y}_i, i = 1, 2, \dots, K \quad (4.6)$$

2. 数据来源及处理

文章选取的数据均来自 4 月 2 日链家网 (<https://sh.lianjia.com>) 上检索到的上海市二手房信息，包括除奉贤区、金山区以及崇明区以外的 13 个区的 2413 套二手房房源信息。选取的指标为：自变量房屋面积 (X_1)、板楼建造年份 (X_2)、所在区域

(X₃)、是否靠近地铁 (X₄)、是否免营业税 (X₅) 以及因变量房屋价格 (Y)。

文章在实证分析过程中采用了简单交叉验证的方法,即将原样本随机分为两部分,一部分作为训练集,另一部分作为测试集。其中训练集包含 1930 个样本,占总体样本的 80%,测试集共有 483 个样本,占总体样本的 20%。此方法可以利用测试集进一步比较和评价模型的效果。

3. 实证结果

(1) 特征变量重要性得分

在建立随机森林回归模型时,文章设定模型的抽样方法为 bootstrap 方法,建立树的棵数为 400 棵,最大特征变量数量为 5,树的最大深度为 50,模型的其他参数都设置为默认值。

如公式 (4.7) 所示,该模型通过计算基于 permutation 随机置换的残差均方减小量来衡量每个特征变量的重要性得分。其中 $score_i$ 是第 i 个特征变量的重要性得分, MSE_j 是用已知的随机森林模型估计第 j 个袋外样本的残差均方, MSE_{ij} 是用已知随机森林模型估计 X_i 变量随即置换后的第 j 个袋外样本的残差均方, S_E 是特征变量的标准误。因为每次 bootstrap 抽样都存在样本未被抽中 (即袋外数据),所有对于 K 棵回归树的随机森林模型来说共有 K 个袋外样本。

$$score_i = (\sum_{j=1}^K (MSE_j - MSE_{ij}) / k) / S_E, (1 \leq i \leq p) \quad (4.7)$$

所有特征变量的重要性得分如表 4.1 所示。

表 4.1 特征变量重要性得分

特征变量	重要性得分
房屋面积 (X ₁)	0.78450577
所在地区 (X ₃)	0.12335229
板楼建造年份 (X ₂)	0.06114649
是否靠近地铁 (X ₄)	0.0263354
是否免营业税 (X ₅)	0.00466005

可以看出得分最高的特征变量是房屋面积 (X₁),重要性高达 0.785。其次是所在地区 (X₃),重要性为 0.123。其余特征变量板楼建造年份 (X₂)、是否靠近地铁 (X₄)

以及是否免营业税（ X_5 ）对房价的影响相对较小。

（2）模型评估结果

文章在对模型进行评估的过程中，将随机森林回归模型与线性回归模型进行比较，比较了两者在训练集和测试集的拟合优度和均方误差，并对测试集的被解释变量进行了单个样本的偏差检验。

①拟合优度检验

拟合优度检验是用于检验回归结果对样本观测值的拟合程度，即解释变量对被解释变量的解释程度，其中 $0 \leq R^2 \leq 1$ ，当拟合优度的值越接近 0 时表明回归模型的拟合效果越不好，当拟合优度的值越接近 1 时表明该模型的拟合效果越好。

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} \quad (4.8)$$

表 4.2 模型拟合优度比较

模型选择	训练集拟合优度	测试集拟合优度
随机森林回归模型	0.9776813027083896	0.9751792608807823
线性回归模型	0.7900883876567274	0.7263693290141914

从拟合优度结果来看，随机森林回归模型的拟合优度更接近于 1，且效果明显好于线性回归模型。此外，两个模型测试集的拟合优度都略低于训练集的拟合优度，但随机森林回归模型的拟合优度减幅要远小于线性回归模型的减幅。

随机森林模型在测试集上的拟合优度仍保持着较高的水平，排除了过度拟合的可能性。

②均方误差

均方误差（MSE）是用于衡量平均误差的一种指标，是观测值与估计值差的平方的均值，计算方法如公式（4.9）所示。通常情况下均方误差越小，模型的效果就越好。

$$MSE = \frac{1}{N} \sum_{i=1}^N (y - \hat{y})^2 \quad (4.9)$$

从表 4.3 的结果可以看出无论是训练集还是测试集，线性回归模型的均方误差都是随机森林回归模型均方误差的 10 倍以上，可见随机森林回归模型的效果要优于线性回归模型。

表 4.3 模型均方误差比较

模型选择	训练集均方误差	测试集均方误差
随机森林回归模型	2836.671362837982	3975.9020484275416
线性回归模型	26679.436155314466	43831.44031527819

③单个样本的偏差检验

对测试集的房价估计值进行单个样本的偏差检验，通常可以用到匹配度与绝对误差这两个指标。匹配度的计算方法如公式（4.10）所示，匹配度的值越接近 1，表明模型的预测结果越小。绝对误差的计算方法如公式（4.11）所示，绝对误差越小，表明模型的预测结果越好。

$$MD = \frac{\hat{y}}{y} \quad (4.10)$$

$$AE = |y - \hat{y}| \quad (4.11)$$

表 4.4 和表 4.5 是两个模型在部分测试集上的单个样本偏差检验结果，全部测试集上房价的检验结果由附录 D 可见。随机森林回归模型比传统的线性回归模型准确性高。

表 4.4 随机森林模型测试集单个样本偏差检验结果（部分）

实际值	预测结果	匹配度	绝对误差	实际值	预测结果	匹配度	绝对误差
840	755.439	0.899	84.561	325	329.633	1.014	4.632
265	262.060	0.989	2.940	240	244.124	1.017	4.124
445	421.033	0.946	23.968	428	398.850	0.932	29.150
365	435.848	1.194	70.848	440	435.363	0.989	4.637
380	377.411	0.993	2.589	285	279.463	0.981	5.538
305	354.893	1.164	49.893	1500	1193.115	0.795	306.885
535	556.075	1.039	21.075	290	298.000	1.028	8.000
900	909.163	1.010	9.163	360	368.735	1.024	8.735
430	453.353	1.054	23.353	350	359.113	1.026	9.113
370	342.548	0.926	27.452	410	421.558	1.028	11.558

表 4.5 线性回归模型测试集单个样本偏差检验结果（部分）

实际值	预测结果	匹配度	绝对误差	实际值	预测结果	匹配度	绝对误差
840	446.891	0.532	393.109	325	527.539	1.623	202.539
265	222.315	0.839	42.685	240	326.076	1.359	86.076
445	513.860	1.155	68.860	428	318.418	0.744	109.582
365	563.868	1.545	198.868	440	441.229	1.003	1.229
380	346.606	0.912	33.394	285	272.988	0.958	12.012

305	465.967	1.528	160.967	1500	757.695	0.505	742.305
535	603.624	1.128	68.624	290	241.249	0.832	48.751
900	1046.449	1.163	146.449	360	304.752	0.847	55.248
430	431.616	1.004	1.616	350	383.611	1.096	33.611
370	295.252	0.798	74.748	410	593.326	1.447	183.326

五、研究结论与展望

1. 研究结论

本文在实证研究中发现,对上海市二手房价格贡献较大的变量按重要性排序依次为房屋面积、所在地区、板楼建造年份、是否靠近地铁以及是否免营业税,投资者在购买上海市二手房时可以具体参照这些指标进行选择,如在同等价位的情况下挑选离市中心距离适中且面积在 80 平米(两室一厅)以上的二手房,以达到投资回报最大化。买房者也可以参照这些变量合理定价,使得在自己房源尽快出售的同时最大化自身的利益。

随机森林模型不同于其他回归模型,该模型是一种非参数回归模型,可以适用于复杂的数据,如线性数据、非线性数据以及具有交互性作用的数据等。该模型的拟合效果在质量和稳定性上都明显高于传统的线性回归模型,且预测的准确性也较好,值得推广。

2. 未来展望

虽然文章通过建立随机森林模型探究了二手房价格的影响因素,并证明了该模型的优越性,但还存在诸多不足之处。

首先,文章获取的资料有一定的局限性。文章研究的是上海市二手房的房价,但研究对象仅限于链家网上挂牌出售的房源,其他平台出售的二手房未被纳入文章的研究范围之内。

第二,文章在指标的选取上存在一些不足,人们在购买二手房时往往还会参考其他方面的因素,如楼层、是否有电梯、房源朝向等。这些指标并未在所有房源中都标注出来,因此将其列入指标体系之中必定会导致样本量的大幅减少。另一方面每套房源的特征信息我们无法考证其真实性,只能默认链家网的工作人员对平台上的每套房源都进行了严格的审核。

第二,部分地区的样本量太少,笔者发现奉贤区、金山区和崇明区二手房数量严

重不足，因此未将这些地区纳入研究范围之中，对模型结论的准确性造成影响。

第三，某些特征变量的部分类别样本量太少。例如在是否免交营业税（ X_5 ）这一特征变量中，仅有 11% 的房源可以免交营业税。过少的样本量可能会对该指标的衡量成偏差。

第四，在与其他回归模型进行比较时，文章只选取了传统的线性回归模型与随机森林回归模型进行比较，得出了随机森林模型的效果要远优于线性回归模型的结论。文章还需将该模型与其它非线性回归模型比较，才能更好地证明该模型的优越性。

最后，随机森林本身也存在一定的缺点，如在建模过程中无法得出明确的函数表达式，因而导致我们不能直观地看出各个特征因素对二手房价格影响的具体效果，而这一点是许多线性和非线性回归都可以做到的。

以上的一些问题都需要在后续的研究中不断改进。

参考文献

- [1] 刘晓群.特征价格模型在武汉市商品房定价中的应用研究[D].武汉理工大学,2008.
- [2] 王卓琳.北京市住宅特征价格研究[D].北京工业大学,2009
- [3] 曾昭法,唐海滨,王毅等.长沙二手房价格模糊综合评估模型商业研究[J].商业研究,2010,(3):139-142.
- [4] 郭志强.基于支持向量机回归的房地产批量估价模型研究[D].暨南大学,2013.
- [5] E Worzala, M Lenk, A Silva. An exploration of neural networks and its application to real estate valuation[J]. Journal of Real Estate Research, 1995.
- [6] N Nghiep, C Al. Predicting housing value: A comparison of multiple regression analysis and artificial neural networks[J]. Journal of Real Estate Research, 2001.
- [7] EA Antipov, EB Pokryshevskaya. Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics[J]. Expert Systems With Applications, 2012.
- [8] Rosen, herwin. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition[J]. The Journal of Political Economy, Vol. 82, No. 1. (Jan. - Feb., 1974), pp. 34-55.
- [9] 温海珍,贾生华.住宅的特征与特征的价格——基于特征价格模型的分析[J].浙江大学学报(工学版),2004(10):101-105+112.
- [10] 周志华.机器学习[M].北京:清华大学出版社,2016.
- [11] Breiman L, Friedman J, Stone C. Classification and Regression Trees[M]. Wadsworth, 1984.
- [12] LEO BREIMAN.Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.
- [13] 李航.统计学习方法[M].北京:清华大学出版社,2012:55-75.

附录 A

用 requests 库和正则表达式抓取房源的房屋面积、板楼建造年分、所在区域、和房源 ID，并储存至数据框中，代码如下：

```
import requests
from requests.exceptions import RequestException
import re
import pandas as pd
from pandas import DataFrame

def get_one_page(url):
    try:
        response=requests.get(url)
        if response.status_code == 200:
            return response.text
        return None
    except RequestException:
        return None

def parse_one_page():
    result2=DataFrame()
    pattern = re.compile('<li.*?img.*?housecode="(\d+).*?</a>'
        + '<div.*?houseInfo.*?houseIcon.*?region">(.*)\s</a>\s\D.*?\D\s(\d\d.*?\D)\s\D.*?</div>'
        + '<div.*?positionInfo.*?\D(\d\d\d\d)\D.*?<a.*?target.*?blank\D>(.*)</a></div>',re.S)
    for i in range(101):
        url = 'https://sh.lianjia.com/ershoufang/ pg' + str(i) + ' tt2/'
        html = get_one_page(url)
        items = re.findall(pattern, html)
        for item in items:
            house = { 'area': item[2][:2],
                      'year': item[3],
                      'location': item[4],
                      'name':item[1],
                      'id':item[0] }
            result1 = DataFrame(house, index=pd.Series(item[1]))
            result2 = pd.concat([result2, result1])
    print(result2)
    result2.to_csv('data.csv',index=True,header=True,sep=' ')

def main():
    parse_one_page()
if __name__ == '__main__':
    main()
```

用 PyQuery 库和 CSS selector 抓取房源的房屋价格、是否靠近地铁、是否免营业税和房源 ID，并储存至数据框中，代码如下：

```
from pyquery import PyQuery as pq
import pandas as pd
from pandas import DataFrame

def parse_one_page():
    result0=DataFrame()
    for i in range(1,101):
        url = 'https://sh.lianjia.com/ershoufang /pg' + str(i) + 'tt2/'
        doc = pq(url=url)
        for data in doc('.sellListContent li'):
            house={'name':pq(data).find('.address .houseInfo a').text(),
                    'price':pq(data).find('.priceInfo .totalPrice span').text(),
                    'subway': pq(data).find('.tag .subway').text(),
                    'taxfree': pq(data).find('.tag .taxfree').text(),
                    '2year':pq(data).find('.tag .five').text(),
                    'id':pq(data).find('.info .title a').attr('data-housecode')}
            result1 = DataFrame(house, index=pd.Series(pq(data).find('.address .houseInfo a').text()))
            result0 = pd.concat([result0, result1])
    print(result0)
    result0.to_csv('data1.csv', index=True, header=True, sep=' ')

def main():
    parse_one_page()

if __name__ == '__main__':
    main()
```

附录 B

对各变量进行描述性统计，代码如下：

```
import pandas as pd
from numpy import array
from numpy import mean, median
from numpy import ptp, std

df = pd.read_csv('data.csv', sep=' ', header=0, index_col=0)

price = array(df['price'])
print('price')
print(mean(price), median(price), ptp(price), std(price))

area = array(df['area'])
print('area')
print(mean(area), median(area), ptp(area), std(area))

subway = array(df['subway'])
print('subway')
print(mean(subway), std(subway))

free_tax = array(df['free_tax'])
print('free_tax')
print(mean(free_tax), std(free_tax))

year = array(df['year'])
print('year')
print(mean(year), median(year), ptp(year), std(year))
```

附录 C

用 R 中的 ggplot2 进行作图，代码如下：

```
>>>setwd("C:\\Users\\admin\\Desktop")
>>>dataframe<-read.csv("data1.csv",header=TRUE)
>>>library(gridExtra)
>>>library(ggplot2)

#箱线图
>>>p1 <- ggplot(dataframe, aes(as.character(location), price))
+ geom_boxplot(fill="grey46", color = "black", notch = TRUE)
+ geom_point(position = "jitter", color="grey17", alpha=0.1)
+ geom_rug (color = "black") + xlab("地区") + ylab("价格（万元）")
>>>p2 <- ggplot(dataframe, aes(as.character(subway), price))
+ geom_boxplot(fill = "grey46", color = "black", notch = TRUE)
+ geom_point(position = "jitter", color = "grey17", alpha = 0.1)
+ geom_rug (color = "black") + xlab("地铁") + ylab("价格（万元）")
>>>p3 <- ggplot(dataframe, aes(as.character(free_tax), price))
+ geom_boxplot (fill = "grey17", color = "black", notch = TRUE)
+ geom_point (position = "jitter", color = "grey17", alpha = 0.1)
+ geom_rug (color = "black") + xlab("营业税") + ylab("价格（万元）")
>>>p4 <- grid.arrange(p1,p2,p3,ncol=3,nrow=1)

#散点图
>>>p5 <- ggplot(dataframe, aes(x=area, y=price, color=as.character(location)))
+ geom_point(size=0.05)
+ labs(x="面积（平方米）", y="价格（万元）", colour = "地区")
+ scale_color_manual(values=c("grey99", "grey79", "grey59", "grey39", "grey19"))
>>>p6 <- ggplot(dataframe, aes(x=year, y=price, color=as.character(location)))
+ geom_point(size=0.2) + labs(x="建筑年份", y="价格（万元）", colour = "地区")
+ scale_color_manual(values=c("grey99", "grey79", "grey59", "grey39", "grey19"))
>>>p7 <- grid.arrange(p5,p6,ncol=1,nrow=2)

#频数统计图
>>>p8 <- ggplot(dataframe, aes(x=location)) + geom_bar() + labs(x="地区", y="户数")
>>>p9 <- ggplot(dataframe, aes(x=subway)) + geom_bar(width = 0.25)
+ scale_x_continuous(breaks = seq(0,1))+labs(x="地铁", y="户数")
>>>p10 <- ggplot(dataframe, aes(x=free_tax)) + geom_bar(width = 0.25)
+ scale_x_continuous(breaks = seq(0,1))+labs(x="营业税", y="户数")
>>>p11<-grid.arrange(p8,p9,p10,ncol=3,nrow=1)
```

附录 D

随机森林模型测试集单个样本偏差检验结果：

实际值	预测结果	匹配度	绝对误差	实际值	预测结果	匹配度	绝对误差	实际值	预测结果	匹配度	绝对误差
840	755.439	0.899	84.561	325	329.633	1.014	4.632	370	321.213	0.868	48.787
265	262.060	0.989	2.940	240	244.124	1.017	4.124	335	348.730	1.041	13.730
445	421.033	0.946	23.968	428	398.850	0.932	29.150	410	401.473	0.979	8.527
365	435.848	1.194	70.848	440	435.363	0.989	4.637	215	217.690	1.013	2.690
380	377.411	0.993	2.589	285	279.463	0.981	5.538	340	427.050	1.256	87.050
305	354.893	1.164	49.893	1500	1193.115	0.795	306.885	800	758.142	0.948	41.858
535	556.075	1.039	21.075	290	298.000	1.028	8.000	860	905.138	1.052	45.138
900	909.163	1.010	9.163	360	368.735	1.024	8.735	280	293.970	1.050	13.970
430	453.353	1.054	23.353	350	359.113	1.026	9.113	718	599.535	0.835	118.465
370	342.548	0.926	27.452	410	421.558	1.028	11.558	380	367.050	0.966	12.950
850	876.923	1.032	26.923	570	533.903	0.937	36.098	450	514.295	1.143	64.295
238	244.470	1.027	6.470	230	244.530	1.063	14.530	1250	1270.175	1.016	20.175
380	327.785	0.863	52.215	490	475.288	0.970	14.713	430	421.820	0.981	8.180
225	220.380	0.979	4.620	550	501.950	0.913	48.050	3200	2873.875	0.898	326.125
1320	1251.900	0.948	68.100	430	438.518	1.020	8.517	465	563.010	1.211	98.010
780	779.038	0.999	0.962	1500	1271.525	0.848	228.475	798	719.265	0.901	78.735
320	293.218	0.916	26.783	310	294.339	0.949	15.661	285	319.438	1.121	34.438
190	199.938	1.052	9.938	250	260.128	1.041	10.128	295	305.783	1.037	10.783
315	325.703	1.034	10.703	288	286.668	0.995	1.332	250	272.545	1.090	22.545
290	303.473	1.046	13.473	738	762.435	1.033	24.435	730	800.750	1.097	70.750
890	872.875	0.981	17.125	1100	1025.348	0.932	74.652	310	332.600	1.073	22.600
315	322.793	1.025	7.793	230	219.048	0.952	10.953	390	385.373	0.988	4.628
310	308.900	0.996	1.100	1100	1037.145	0.943	62.855	1160	1066.760	0.920	93.240
180	193.415	1.075	13.415	485	536.858	1.107	51.858	238	265.275	1.115	27.275
510	557.643	1.093	47.643	1200	1149.788	0.958	50.213	320	423.295	1.323	103.295
620	612.828	0.988	7.173	930	824.813	0.887	105.188	380	360.855	0.950	19.145
240	270.713	1.128	30.713	460	419.183	0.911	40.818	739	836.595	1.132	97.595
260	261.555	1.006	1.555	320	337.305	1.054	17.305	410	397.578	0.970	12.423
245	255.223	1.042	10.223	435	416.856	0.958	18.144	880	895.775	1.018	15.775
315	284.073	0.902	30.928	1200	1203.750	1.003	3.750	920	859.653	0.934	60.348
485	558.988	1.153	73.988	450	460.600	1.024	10.600	340	355.610	1.046	15.610
635	634.060	0.999	0.940	250	260.675	1.043	10.675	313	318.010	1.016	5.010
570	513.310	0.901	56.690	330	306.507	0.929	23.493	350	326.965	0.934	23.035
220	223.760	1.017	3.760	400	489.315	1.223	89.315	410	414.758	1.012	4.757
426	419.844	0.986	6.156	410	406.948	0.993	3.053	470	506.704	1.078	36.704
420	401.274	0.955	18.726	195	201.250	1.032	6.250	1700	1546.535	0.910	153.465
650	641.550	0.987	8.450	198	200.748	1.014	2.748	450	445.440	0.990	4.560
180	195.263	1.085	15.263	210	206.933	0.985	3.068	390	354.531	0.909	35.469
285	307.533	1.079	22.533	330	358.075	1.085	28.075	480	489.818	1.020	9.818
880	756.438	0.860	123.563	300	290.978	0.970	9.022	400	388.338	0.971	11.663
330	348.217	1.055	18.217	190	197.105	1.037	7.105	285	287.181	1.008	2.181
435	447.015	1.028	12.015	340	310.708	0.914	29.293	388	389.720	1.004	1.720
550	525.708	0.956	24.293	375	350.600	0.935	24.400	238	244.142	1.026	6.142
800	931.790	1.165	131.790	275	295.068	1.073	20.068	2600	2669.225	1.027	69.225
930	971.025	1.044	41.025	235	253.436	1.078	18.436	140	211.173	1.508	71.173
285	288.528	1.012	3.527	290	285.315	0.984	4.685	440	490.525	1.115	50.525
340	341.422	1.004	1.422	310	338.935	1.093	28.935	500	492.430	0.985	7.570
550	499.363	0.908	50.638	200	215.098	1.075	15.098	640	603.700	0.943	36.300
290	290.035	1.000	0.035	705	907.878	1.288	202.878	238	275.115	1.156	37.115
217	227.425	1.048	10.425	375	348.338	0.929	26.663	168	207.343	1.234	39.343
260	251.625	0.968	8.375	260	257.502	0.990	2.498	235	239.953	1.021	4.952
495	466.805	0.943	28.195	275	277.262	1.008	2.262	380	327.125	0.861	52.875
480	541.278	1.128	61.278	365	383.453	1.051	18.453	435	400.138	0.920	34.863
386	435.105	1.127	49.105	210	251.702	1.199	41.702	560	598.688	1.069	38.688
700	776.325	1.109	76.325	340	346.648	1.020	6.647	380	357.513	0.941	22.488

540	502.288	0.930	37.713	350	407.155	1.163	57.155	415	424.750	1.023	9.750
300	294.008	0.980	5.992	290	312.063	1.076	22.063	200	194.005	0.970	5.995
530	681.648	1.286	151.648	700	736.980	1.053	36.980	330	306.278	0.928	23.723
585	567.370	0.970	17.630	400	399.338	0.998	0.663	650	668.118	1.028	18.117
225	268.048	1.191	43.048	250	243.790	0.975	6.210	208	218.755	1.052	10.755
810	807.408	0.997	2.592	480	479.077	0.998	0.923	235	232.178	0.988	2.822
930	891.563	0.959	38.438	690	686.838	0.995	3.163	1900	1813.865	0.955	86.135
480	513.405	1.070	33.405	360	344.087	0.956	15.913	380	364.657	0.960	15.343
595	675.038	1.135	80.038	845	839.020	0.993	5.980	311	318.058	1.023	7.058
750	753.435	1.005	3.435	380	438.268	1.153	58.268	340	343.425	1.010	3.425
500	500.860	1.002	0.860	370	388.913	1.051	18.913	285	282.193	0.990	2.808
288	298.252	1.036	10.252	518	501.423	0.968	16.578	360	351.358	0.976	8.642
500	483.423	0.967	16.578	720	740.305	1.028	20.305	605	596.945	0.987	8.055
338	333.638	0.987	4.363	253	276.638	1.093	23.638	245	226.290	0.924	18.710
195	202.308	1.037	7.308	1050	901.715	0.859	148.285	310	312.136	1.007	2.136
1085	1147.508	1.058	62.507	360	353.218	0.981	6.783	360	348.863	0.969	11.138
540	459.762	0.851	80.238	270	275.855	1.022	5.855	478	483.310	1.011	5.310
585	580.870	0.993	4.130	330	350.998	1.064	20.998	580	600.915	1.036	20.915
560	563.213	1.006	3.212	825	795.378	0.964	29.622	365	357.683	0.980	7.318
500	485.533	0.971	14.468	478	441.745	0.924	36.255	228	206.100	0.904	21.900
320	304.019	0.950	15.981	880	930.658	1.058	50.658	520	422.234	0.812	97.766
370	368.563	0.996	1.438	280	299.755	1.071	19.755	410	348.038	0.849	61.963
530	456.038	0.860	73.963	200	214.553	1.073	14.553	830	968.313	1.167	138.313
960	957.070	0.997	2.930	620	559.938	0.903	60.063	340	357.878	1.053	17.878
1100	1322.465	1.202	222.465	405	452.028	1.116	47.028	190	178.250	0.938	11.750
580	588.665	1.015	8.665	270	257.605	0.954	12.395	505	523.180	1.036	18.180
300	285.470	0.952	14.530	300	300.556	1.002	0.556	720	600.975	0.835	119.025
550	547.890	0.996	2.110	435	454.588	1.045	19.588	310	345.698	1.115	35.698
315	340.278	1.080	25.278	365	355.838	0.975	9.163	800	770.001	0.963	29.999
425	384.533	0.905	40.468	400	391.813	0.980	8.188	610	650.590	1.067	40.590
275	320.565	1.166	45.565	255	286.085	1.122	31.085	327	298.680	0.913	28.320
235	241.603	1.028	6.603	180	212.778	1.182	32.778	325	332.663	1.024	7.663
305	286.688	0.940	18.313	295	300.067	1.017	5.067	280	334.570	1.195	54.570
345	345.503	1.001	0.502	230	208.830	0.908	21.170	230	252.190	1.096	22.190
800	667.003	0.834	132.998	998	962.615	0.965	35.385	700	573.546	0.819	126.454
245	253.878	1.036	8.878	440	427.528	0.972	12.473	335	344.875	1.029	9.875
800	780.630	0.976	19.370	350	362.781	1.037	12.781	298	332.950	1.117	34.950
500	493.433	0.987	6.568	228	228.178	1.001	0.178	400	404.410	1.011	4.410
132	174.445	1.322	42.445	198	227.533	1.149	29.533	135	227.193	1.683	92.193
230	276.448	1.202	46.448	350	355.028	1.014	5.027	470	469.538	0.999	0.462
390	380.198	0.975	9.803	470	512.535	1.091	42.535	225	257.320	1.144	32.320
175	175.595	1.003	0.595	410	401.800	0.980	8.200	480	474.208	0.988	5.793
400	386.191	0.965	13.809	480	438.080	0.913	41.920	776	675.740	0.871	100.260
355	367.455	1.035	12.455	430	439.925	1.023	9.925	340	364.635	1.072	24.635
233	229.265	0.984	3.735	2500	2236.550	0.895	263.450	295	297.893	1.010	2.892
238	268.248	1.127	30.248	335	357.648	1.068	22.648	220	223.798	1.017	3.798
255	245.633	0.963	9.368	434	427.438	0.985	6.563	410	369.730	0.902	40.270
295	346.408	1.174	51.408	705	644.563	0.914	60.438	441	426.255	0.967	14.745
720	882.625	1.226	162.625	380	379.962	1.000	0.038	465	484.050	1.041	19.050
435	451.188	1.037	16.188	460	439.636	0.956	20.364	660	639.500	0.969	20.500
400	413.713	1.034	13.713	222	261.825	1.179	39.825	409	348.777	0.853	60.223
135	141.748	1.050	6.748	189	189.995	1.005	0.995	410	375.533	0.916	34.468
240	227.843	0.949	12.158	260	256.160	0.985	3.840	800	824.875	1.031	24.875
385	479.388	1.245	94.388	330	367.273	1.113	37.273	315	336.745	1.069	21.745
375	397.525	1.060	22.525	730	652.865	0.894	77.135	560	520.413	0.929	39.588
305	275.595	0.904	29.405	235	245.605	1.045	10.605	1100	964.503	0.877	135.498
415	372.440	0.897	42.560	1580	1408.748	0.892	171.253	230	231.350	1.006	1.350
325	310.676	0.956	14.324	390	388.300	0.996	1.700	335	329.485	0.984	5.515
140	156.660	1.119	16.660	320	349.205	1.091	29.205	330	353.540	1.071	23.540
215	301.708	1.403	86.708	285	320.873	1.126	35.873	368	373.743	1.016	5.743
750	804.693	1.073	54.693	300	296.935	0.990	3.065	3100	3005.490	0.970	94.510

295	282.590	0.958	12.410	500	498.555	0.997	1.445	660	582.743	0.883	77.258
600	603.265	1.005	3.265	480	457.893	0.954	22.108	420	448.575	1.068	28.575
1600	1566.325	0.979	33.675	620	671.920	1.084	51.920	460	438.348	0.953	21.653
550	516.393	0.939	33.608	250	228.743	0.915	21.258	260	261.045	1.004	1.045
400	368.538	0.921	31.463	390	427.238	1.095	37.238	990	978.575	0.988	11.425
200	270.518	1.353	70.518	745	693.858	0.931	51.143	208	197.005	0.947	10.995
305	355.050	1.164	50.050	425	404.038	0.951	20.963	660	677.875	1.027	17.875
203	222.393	1.096	19.393	315	375.705	1.193	60.705	500	477.888	0.956	22.113
335	330.630	0.987	4.370	450	402.288	0.894	47.713	700	726.838	1.038	26.838
245	235.758	0.962	9.243	220	230.850	1.049	10.850	1180	1224.978	1.038	44.978
650	636.850	0.980	13.150	200	199.815	0.999	0.185	230	231.708	1.007	1.708
780	709.363	0.909	70.638	340	312.778	0.920	27.222	230	235.080	1.022	5.080
485	478.845	0.987	6.155	580	592.438	1.021	12.438	215	302.508	1.407	87.508
1350	1226.063	0.908	123.938	1030	1025.053	0.995	4.947	1120	1057.750	0.944	62.250
230	251.818	1.095	21.818	700	624.888	0.893	75.113	740	768.223	1.038	28.223
720	758.965	1.054	38.965	390	389.290	0.998	0.710	430	420.400	0.978	9.600
1400	1326.225	0.947	73.775	630	550.268	0.873	79.733	265	282.273	1.065	17.273
240	238.533	0.994	1.468	465	457.945	0.985	7.055	400	378.894	0.947	21.106
800	960.230	1.200	160.230	265	306.984	1.158	41.984	550	619.450	1.126	69.450
950	824.168	0.868	125.833	530	506.195	0.955	23.805	320	321.048	1.003	1.048
380	366.838	0.965	13.163	600	668.548	1.114	68.548	345	323.453	0.938	21.548
213	208.418	0.978	4.583	210	249.640	1.189	39.640	310	276.768	0.893	33.233
365	397.313	1.089	32.313	380	379.989	1.000	0.011	285	273.200	0.959	11.800
880	709.808	0.807	170.193	1100	877.338	0.798	222.663	950	1000.800	1.053	50.800
4500	3858.715	0.857	641.285	780	904.050	1.159	124.050	420	355.463	0.846	64.538
330	339.840	1.030	9.840	860	754.158	0.877	105.843	870	928.380	1.067	58.380
370	374.940	1.013	4.940	230	231.628	1.007	1.628	235	232.323	0.989	2.678
290	329.203	1.135	39.203	253	234.343	0.926	18.658	300	304.450	1.015	4.450
460	521.835	1.134	61.835	240	284.235	1.184	44.235	800	729.380	0.912	70.620
1800	1847.060	1.026	47.060	498	444.118	0.892	53.883	330	412.545	1.250	82.545
336	351.413	1.046	15.413	290	271.213	0.935	18.788	1100	986.775	0.897	113.225
498	549.120	1.103	51.120	540	466.535	0.864	73.465	345	351.808	1.020	6.808
1400	1199.100	0.857	200.900	310	286.168	0.923	23.833	670	610.413	0.911	59.588
300	289.498	0.965	10.503	750	796.918	1.063	46.918	650	651.155	1.002	1.155
320	284.025	0.888	35.975	1030	937.153	0.910	92.848	640	566.315	0.885	73.685
240	247.358	1.031	7.357	850	886.553	1.043	36.553	270	407.538	1.509	137.538
230	239.405	1.041	9.405	345	358.650	1.040	13.650	345	335.470	0.972	9.530
269	275.410	1.024	6.410	345	364.188	1.056	19.188	360	375.438	1.043	15.438
980	979.443	0.999	0.558	230	236.048	1.026	6.048	1050	979.660	0.933	70.340
394	510.228	1.295	116.228	325	335.745	1.033	10.745	285	304.068	1.067	19.068
460	428.675	0.932	31.325	330	329.945	1.000	0.055	364	357.508	0.982	6.493
222	240.725	1.084	18.725	340	325.875	0.958	14.125	230	242.719	1.055	12.719
370	357.866	0.967	12.134	390	401.555	1.030	11.555	518	480.363	0.927	37.638
390	404.738	1.038	14.738	550	579.563	1.054	29.563	1230	1186.545	0.965	43.455
800	937.685	1.172	137.685	460	425.898	0.926	34.103	525	539.893	1.028	14.893

线性回归模型测试集单个样本偏差检验结果：

实际值	预测结果	匹配度	绝对误差	实际值	预测结果	匹配度	绝对误差	实际值	预测结果	匹配度	绝对误差
840	446.891	0.532	393.109	325	527.539	1.623	202.539	370	233.533	0.631	136.467
265	222.315	0.839	42.685	240	326.076	1.359	86.076	335	375.466	1.121	40.466
445	513.860	1.155	68.860	428	318.418	0.744	109.582	410	352.204	0.859	57.796
365	563.868	1.545	198.868	440	441.229	1.003	1.229	215	180.924	0.842	34.076
380	346.606	0.912	33.394	285	272.988	0.958	12.012	340	817.095	2.403	477.095
305	465.967	1.528	160.967	1500	757.695	0.505	742.305	800	557.983	0.697	242.017
535	603.624	1.128	68.624	290	241.249	0.832	48.751	860	1082.214	1.258	222.214
900	1046.449	1.163	146.449	360	304.752	0.847	55.248	280	301.882	1.078	21.882
430	431.616	1.004	1.616	350	383.611	1.096	33.611	718	816.911	1.138	98.911
370	295.252	0.798	74.748	410	593.326	1.447	183.326	380	327.299	0.861	52.701
850	973.265	1.145	123.265	570	474.967	0.833	95.033	450	591.081	1.314	141.081
238	235.360	0.989	2.640	230	236.777	1.029	6.777	1250	1132.670	0.906	117.330
380	308.995	0.813	71.005	490	426.671	0.871	63.329	430	404.882	0.942	25.118
225	151.074	0.671	73.926	550	609.688	1.109	59.688	3200	1841.858	0.576	1358.142
1320	1260.599	0.955	59.401	430	446.358	1.038	16.358	465	578.957	1.245	113.957
780	691.789	0.887	88.211	1500	805.581	0.537	694.419	798	623.503	0.781	174.497
320	186.198	0.582	133.802	310	190.536	0.615	119.464	285	322.127	1.130	37.127
190	112.657	0.593	77.343	250	227.529	0.910	22.471	295	264.742	0.897	30.258
315	287.842	0.914	27.158	288	332.078	1.153	44.078	250	236.131	0.945	13.869
290	577.106	1.990	287.106	738	758.977	1.028	20.977	730	853.624	1.169	123.624
890	752.780	0.846	137.220	1100	858.368	0.780	241.632	310	369.303	1.191	59.303
315	361.702	1.148	46.702	230	113.398	0.493	116.602	390	420.806	1.079	30.806
310	282.777	0.912	27.223	1100	1253.167	1.139	153.167	1160	872.628	0.752	287.372
180	226.136	1.256	46.136	485	662.265	1.365	177.265	238	247.755	1.041	9.755
510	414.214	0.812	95.786	1200	947.351	0.789	252.649	320	585.495	1.830	265.495
620	537.964	0.868	82.036	930	831.510	0.894	98.490	380	404.390	1.064	24.390
240	292.289	1.218	52.289	460	434.508	0.945	25.492	739	947.409	1.282	208.409
260	355.452	1.367	95.452	320	320.084	1.000	0.084	410	446.131	1.088	36.131
245	213.600	0.872	31.400	435	449.610	1.034	14.610	880	866.163	0.984	13.837
315	183.583	0.583	131.417	1200	1011.463	0.843	188.537	920	796.778	0.866	123.222
485	646.959	1.334	161.959	450	489.444	1.088	39.444	340	351.504	1.034	11.504
635	540.059	0.850	94.941	250	217.585	0.870	32.415	313	317.557	1.015	4.557
570	311.555	0.547	258.445	330	380.165	1.152	50.165	350	177.861	0.508	172.139
220	125.787	0.572	94.213	400	610.994	1.527	210.994	410	502.171	1.225	92.171
426	569.174	1.336	143.174	410	360.112	0.878	49.888	470	703.372	1.497	233.372
420	379.194	0.903	40.806	195	166.380	0.853	28.620	1700	951.061	0.559	748.939
650	863.118	1.328	213.118	198	237.517	1.200	39.517	450	498.912	1.109	48.912
180	231.870	1.288	51.870	210	65.470	0.312	144.530	390	297.115	0.762	92.885
285	273.184	0.959	11.816	330	507.223	1.537	177.223	480	691.800	1.441	211.800
880	594.021	0.675	285.979	300	388.694	1.296	88.694	400	474.682	1.187	74.682
330	340.531	1.032	10.531	190	126.780	0.667	63.220	285	245.268	0.861	39.732
435	519.987	1.195	84.987	340	265.383	0.781	74.617	388	372.116	0.959	15.884
550	473.982	0.862	76.018	375	237.503	0.633	137.497	238	229.196	0.963	8.804
800	759.001	0.949	40.999	275	292.687	1.064	17.687	2600	2490.705	0.958	109.295
930	935.885	1.006	5.885	235	269.128	1.145	34.128	140	164.457	1.175	24.457
285	354.959	1.245	69.959	290	245.054	0.845	44.946	440	646.805	1.470	206.805
340	354.715	1.043	14.715	310	470.301	1.517	160.301	500	553.792	1.108	53.792
550	464.864	0.845	85.136	200	140.221	0.701	59.779	640	636.718	0.995	3.282
290	262.308	0.905	27.692	705	1076.449	1.527	371.449	238	385.884	1.621	147.884
217	310.243	1.430	93.243	375	323.225	0.862	51.775	168	215.723	1.284	47.723
260	199.928	0.769	60.072	260	331.799	1.276	71.799	235	193.029	0.821	41.971
495	538.659	1.088	43.659	275	191.776	0.697	83.224	380	213.389	0.562	166.611
480	566.496	1.180	86.496	365	425.728	1.166	60.728	435	405.239	0.932	29.761
386	358.012	0.927	27.988	210	447.158	2.129	237.158	560	670.174	1.197	110.174
700	732.392	1.046	32.392	340	367.338	1.080	27.338	380	317.556	0.836	62.444
540	335.843	0.622	204.157	350	537.407	1.535	187.407	415	512.441	1.235	97.441
300	344.618	1.149	44.618	290	430.709	1.485	140.709	200	73.550	0.368	126.450
530	453.030	0.855	76.970	700	704.042	1.006	4.042	330	301.557	0.914	28.443
585	600.408	1.026	15.408	400	390.694	0.977	9.306	650	655.752	1.009	5.752

225	265.104	1.178	40.104	250	199.809	0.799	50.191	208	195.102	0.938	12.898
810	683.959	0.844	126.041	480	448.250	0.934	31.750	235	297.134	1.264	62.134
930	699.145	0.752	230.855	690	652.831	0.946	37.169	1900	1328.976	0.699	571.024
480	465.438	0.970	14.562	360	340.943	0.947	19.057	380	361.027	0.950	18.973
595	587.358	0.987	7.642	845	731.678	0.866	113.322	311	187.651	0.603	123.349
750	736.269	0.982	13.731	380	564.270	1.485	184.270	340	316.873	0.932	23.127
500	611.456	1.223	111.456	370	484.835	1.310	114.835	285	376.166	1.320	91.166
288	311.982	1.083	23.982	518	490.108	0.946	27.892	360	425.441	1.182	65.441
500	462.316	0.925	37.684	720	718.132	0.997	1.868	605	647.370	1.070	42.370
338	487.467	1.442	149.467	253	280.107	1.107	27.107	245	221.703	0.905	23.297
195	152.742	0.783	42.258	1050	715.437	0.681	334.563	310	344.564	1.111	34.564
1085	1158.823	1.068	73.823	360	366.617	1.018	6.617	360	366.131	1.017	6.131
540	536.545	0.994	3.455	270	252.761	0.936	17.239	478	474.183	0.992	3.817
585	612.763	1.047	27.763	330	501.647	1.520	171.647	580	640.376	1.104	60.376
560	644.400	1.151	84.400	825	676.744	0.820	148.256	365	310.805	0.852	54.195
500	522.125	1.044	22.125	478	357.695	0.748	120.305	228	6.566	0.029	221.434
320	295.157	0.922	24.843	880	855.738	0.972	24.262	520	385.716	0.742	134.284
370	383.350	1.036	13.350	280	374.894	1.339	94.894	410	165.526	0.404	244.474
530	396.446	0.748	133.554	200	196.205	0.981	3.795	830	1050.036	1.265	220.036
960	970.444	1.011	10.444	620	446.981	0.721	173.019	340	314.914	0.926	25.086
1100	2079.740	1.891	979.740	405	481.006	1.188	76.006	190	32.366	0.170	157.634
580	717.378	1.237	137.378	270	254.471	0.942	15.529	505	771.084	1.527	266.084
300	213.866	0.713	86.134	300	225.372	0.751	74.628	720	487.250	0.677	232.750
550	505.465	0.919	44.535	435	458.279	1.054	23.279	310	357.886	1.154	47.886
315	421.429	1.338	106.429	365	305.291	0.836	59.709	800	446.695	0.558	353.305
425	500.408	1.177	75.408	400	456.824	1.142	56.824	610	729.888	1.197	119.888
275	311.507	1.133	36.507	255	164.426	0.645	90.574	327	362.926	1.110	35.926
235	186.670	0.794	48.330	180	316.105	1.756	136.105	325	440.552	1.356	115.552
305	247.982	0.813	57.018	295	292.817	0.993	2.183	280	511.189	1.826	231.189
345	404.199	1.172	59.199	230	199.397	0.867	30.603	230	203.573	0.885	26.427
800	569.197	0.711	230.803	998	978.133	0.980	19.867	700	458.859	0.656	241.141
245	360.522	1.472	115.522	440	508.298	1.155	68.298	335	408.028	1.218	73.028
800	787.808	0.985	12.192	350	347.075	0.992	2.925	298	406.005	1.362	108.005
500	544.329	1.089	44.329	228	268.987	1.180	40.987	400	340.555	0.851	59.445
132	127.577	0.966	4.423	198	288.073	1.455	90.073	135	187.729	1.391	52.729
230	330.765	1.438	100.765	350	414.708	1.185	64.708	470	415.640	0.884	54.360
390	351.164	0.900	38.836	470	504.325	1.073	34.325	225	412.167	1.832	187.167
175	77.403	0.442	97.597	410	436.976	1.066	26.976	480	344.844	0.718	135.156
400	473.086	1.183	73.086	480	468.871	0.977	11.129	776	607.645	0.783	168.355
355	413.955	1.166	58.955	430	426.088	0.991	3.912	340	459.360	1.351	119.360
233	170.503	0.732	62.497	2500	1813.488	0.725	686.512	295	377.014	1.278	82.014
238	380.114	1.597	142.114	335	409.572	1.223	74.572	220	240.318	1.092	20.318
255	210.365	0.825	44.635	434	505.139	1.164	71.139	410	285.337	0.696	124.663
295	284.328	0.964	10.672	705	556.404	0.789	148.596	441	490.070	1.111	49.070
720	1469.751	2.041	749.751	380	329.768	0.868	50.232	465	452.407	0.973	12.593
435	659.570	1.516	224.570	460	417.416	0.907	42.584	660	645.018	0.977	14.982
400	477.871	1.195	77.871	222	305.064	1.374	83.064	409	396.968	0.971	12.032
135	-52.775	-0.391	187.775	189	159.113	0.842	29.887	410	430.287	1.049	20.287
240	260.361	1.085	20.361	260	146.542	0.564	113.458	800	904.134	1.130	104.134
385	754.744	1.960	369.744	330	467.819	1.418	137.819	315	356.686	1.132	41.686
375	556.392	1.484	181.392	730	529.729	0.726	200.271	560	464.150	0.829	95.850
305	197.613	0.648	107.387	235	320.164	1.362	85.164	1100	725.209	0.659	374.791
415	294.724	0.710	120.276	1580	1161.074	0.735	418.926	230	319.866	1.391	89.866
325	186.279	0.573	138.721	390	378.369	0.970	11.631	335	428.251	1.278	93.251
140	82.463	0.589	57.537	320	479.183	1.497	159.183	330	645.341	1.956	315.341
215	321.734	1.496	106.734	285	370.833	1.301	85.833	368	388.729	1.056	20.729
750	920.402	1.227	170.402	300	236.653	0.789	63.347	3100	2064.251	0.666	1035.749
295	291.452	0.988	3.548	500	498.021	0.996	1.979	660	643.267	0.975	16.733
600	503.435	0.839	96.565	480	500.911	1.044	20.911	420	507.877	1.209	87.877
1600	1572.104	0.983	27.896	620	616.925	0.995	3.075	460	386.300	0.840	73.700
550	674.993	1.227	124.993	250	125.812	0.503	124.188	260	372.340	1.432	112.340

400	347.479	0.869	52.521	390	369.783	0.948	20.217	990	1007.266	1.017	17.266
200	452.839	2.264	252.839	745	669.942	0.899	75.058	208	79.498	0.382	128.502
305	682.321	2.237	377.321	425	341.132	0.803	83.868	660	703.604	1.066	43.604
203	344.035	1.695	141.035	315	448.092	1.423	133.092	500	447.853	0.896	52.147
335	419.131	1.251	84.131	450	346.859	0.771	103.141	700	696.249	0.995	3.751
245	169.001	0.690	75.999	220	217.627	0.989	2.373	1180	1190.930	1.009	10.930
650	653.187	1.005	3.187	200	202.891	1.014	2.891	230	301.444	1.311	71.444
780	611.422	0.784	168.578	340	300.957	0.885	39.043	230	192.796	0.838	37.204
485	450.774	0.929	34.226	580	621.093	1.071	41.093	215	492.564	2.291	277.564
1350	977.906	0.724	372.094	1030	1128.646	1.096	98.646	1120	994.127	0.888	125.873
230	392.783	1.708	162.783	700	673.200	0.962	26.800	740	707.000	0.955	33.000
720	723.748	1.005	3.748	390	531.185	1.362	141.185	430	342.800	0.797	87.200
1400	1114.461	0.796	285.539	630	467.676	0.742	162.324	265	287.787	1.086	22.787
240	112.906	0.470	127.094	465	482.936	1.039	17.936	400	335.832	0.840	64.168
800	1553.456	1.942	753.456	265	344.986	1.302	79.986	550	553.007	1.005	3.007
950	711.927	0.749	238.073	530	324.955	0.613	205.045	320	286.013	0.894	33.987
380	401.178	1.056	21.178	600	819.462	1.366	219.462	345	270.821	0.785	74.179
213	150.326	0.706	62.674	210	299.271	1.425	89.271	310	221.959	0.716	88.041
365	454.794	1.246	89.794	380	407.175	1.072	27.175	285	207.837	0.729	77.163
880	550.045	0.625	329.955	1100	552.782	0.503	547.218	950	846.618	0.891	103.382
4500	2108.266	0.469	2391.734	780	913.563	1.171	133.563	420	174.971	0.417	245.029
330	261.381	0.792	68.619	860	413.573	0.481	446.427	870	819.207	0.942	50.793
370	423.745	1.145	53.745	230	343.267	1.492	113.267	235	272.666	1.160	37.666
290	364.589	1.257	74.589	253	203.183	0.803	49.817	300	282.581	0.942	17.419
460	619.677	1.347	159.677	240	450.120	1.876	210.120	800	732.451	0.916	67.549
1800	1482.717	0.824	317.283	498	585.964	1.177	87.964	330	510.547	1.547	180.547
336	397.194	1.182	61.194	290	264.107	0.911	25.893	1100	783.307	0.712	316.693
498	790.325	1.587	292.325	540	263.983	0.489	276.017	345	374.279	1.085	29.279
1400	990.678	0.708	409.322	310	233.876	0.754	76.124	670	632.029	0.943	37.971
300	277.702	0.926	22.298	750	781.199	1.042	31.199	650	667.365	1.027	17.365
320	168.053	0.525	151.947	1030	705.367	0.685	324.633	640	435.553	0.681	204.447
240	300.316	1.251	60.316	850	735.330	0.865	114.670	270	435.189	1.612	165.189
230	191.522	0.833	38.478	345	643.802	1.866	298.802	345	278.135	0.806	66.865
269	347.525	1.292	78.525	345	481.202	1.395	136.202	360	378.618	1.052	18.618
980	931.581	0.951	48.419	230	224.359	0.975	5.641	1050	824.961	0.786	225.039
394	587.693	1.492	193.693	325	303.112	0.933	21.888	285	174.440	0.612	110.560
460	459.050	0.998	0.950	330	368.544	1.117	38.544	364	364.463	1.001	0.463
222	237.026	1.068	15.026	340	311.482	0.916	28.518	230	315.639	1.372	85.639
370	370.792	1.002	0.792	390	375.614	0.963	14.386	518	504.391	0.974	13.609
390	571.109	1.464	181.109	550	607.289	1.104	57.289	1230	1016.135	0.826	213.865
800	898.870	1.124	98.870	460	486.391	1.057	26.391	525	634.677	1.209	109.677