Statistics and Application 统计学与应用, 2021, 10(6), 981-988
Published Online December 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.106103

Hans汉斯

基于灰色模型和支持向量回归的财政收入预测

贾晓芳[1]，牟唯嫣[1]，李泽妤[2]

[1]北京建筑大学理学院，北京
[2]北京工商大学嘉华学院，北京

收稿日期：2021年11月17日；录用日期：2021年12月1日；发布日期：2021年12月16日

摘　要

随着科学技术的不断发展，大数据应用的越来越普及，已成为提高财政收入的有力工具。本文以1994~2019年数据为依托，借助R统计软件，首先对财政收入、第一产业增加值、工业增加值、建筑业增加值、年末总人口、社会消费品零售总额和受灾面积这六个方面的原始数据进行相关性分析，运用Lasso回归方法识别影响财政收入的关键特征，然后将灰色模型和支持向量回归预测模型相结合，对未来两年的财政收入进行预测，最后对建立的财政收入预测模型进行评价。

关键词

财政收入预测，灰色GM(1,1)，SVR，Lasso回归

Fiscal Revenue Prediction Based on Grey Model and Support Vector Regression

Xiaofang Jia[1], Weiyan Mu[1], Zeyu Li[2]

[1]School of Science, Beijing University of Civil Engineering and Architecture, Beijing
[2]Canvard College, Beijing Technology and Business University, Beijing

Received: Nov. 17th, 2021; accepted: Dec. 1st, 2021; published: Dec. 16th, 2021

Abstract

With the continuous development of science and technology, the application of big data has become more and more popular, and it has become a powerful tool to increase fiscal revenue. Firstly, the relativity of data from 1994 to 2019 among fiscal revenue, primary industry added value, industrial added value, construction industry added value, total population at the end of the year, total retail sales of consumer goods, and disaster-affected area is analyzed by R software in this article. And using the Lasso regression method to choose the key features that affect fiscal revenue. Then we

文章引用:贾晓芳, 牟唯嫣, 李泽妤. 基于灰色模型和支持向量回归的财政收入预测[J]. 统计学与应用, 2021, 10(6): 981-988. DOI: 10.12677/sa.2021.106103

combine the gray model and the support vector regression prediction model to predict the fiscal revenue for the next two years. Finally, the established fiscal revenue forecast model is evaluated.

Keywords

Fiscal Revenue Prediction, Grey GM(1,1), SVR, Lasso Regression

1. 引言

广义的财政收入[1]是指政府为社会提供公共物品与服务、实施公共政策和履行其职能的需要，依据一定的权力原则而筹集的一切资金的总和，它是衡量一国政府财力的重要指标，是实现国家职能的财力保证，在推动经济社会高质量发展和人民群众高品质生活中扮演着至关重要的作用，因此对其进行研究分析很有必要，可为我国财政预算提供一定的理论依据。

近年来，国内外许多学者纷纷运用不同统计方法深入研究分析我国财政收入与地方财政收入。根据预测模型的不同主要分为两类：单一型和组合型。2006 年文献[2]利用 BP 神经网络原理预测税收；文献[3]讨论了组合预测模型在地方财政收入中的应用；文献[4]在 2011 年给出基于主成分分析的支持向量机税收预测模型，通过我国 2001~2004 年税收数据进行验证；2016 年文献[5]提出了灰色 RBF 神经网络的多因素财政收入预测模型，并以安徽省的财政收入数据为例衡量构建的模型；文献[6]在 2018 年指出一种组合预测方法——Lasso-GRNN 神经网络模型对地方财政收入进行分析预测。通过分析发现组合模型在一定程度上可以弥补单一模型的短板，提高结果的可信度。

除此之外，众多学者关于财政收入影响因素的探讨，大部分都是先构建我国财政收入或地方财政收入与待测定的影响要素之间的多元线性回归模型，再使用最小二乘法对所建的回归模型进行系数估计问题，以此来判断彼此之间的关联度，如文献[7] [8] [9]。由于模型的结果对数据具有较高的依赖程度，而且在普通最小二乘估计法下求得的解往往表现为局部最优解，由此一来，对于之后步骤的检验恐怕就会失去本应表达的意义。

综合考虑，为得到更为准确的预测结果，本文在已有研究的基础上，借由 R 统计软件进行编程，以 1994~2019 年我国财政收入和相关特征数据为实例，在 Lasso 特征选择的基础上，集中讨论灰色 GM(1,1)模型和支持向量回归(Support Vector Regression, SVR)的组合预测模型。

2. 研究方法

首先，对于多维数据的分析与处理，通常的做法是进行变量的选取工作，筛选变量的方法通常有主成分分析法、最小二乘回归法、逐步回归法、岭回归、Lasso 回归等多种方法。Lasso (Least Absolute Shrinkage and Selection Operator)回归方法是 Robert Tibshirani [10]在 1996 年提出的一种新型的变量选择方法。这种技术是一种收缩估计方法，通过构造惩罚函数压缩模型特征系数使得模型稀疏表达，达到特征选择的目的，进而更好处理变量间的多重共线性问题。Lasso 方法既结合了子集选择法的优势又囊括了岭回归的优点，相比于传统的变量筛选方法，它能够很好的克服传统方法在变量选择问题中存在的短板，所以该技术在统计学、经济学、医药卫生等领域受到越来越多的关注与重视，详细可参考文献[11] [12] [13] [14] [15]。文献[16]对这几种常用方法进行了探讨，并得出 Lasso 回归在预测准确性和模型可解释性上更

优于其他方法结论，所以本文利用 Lasso 方法来做变量筛选，降低变量的个数。

其次，灰色预测法是一种对既包括已知信息又包含不确定信息的系统进行有效预测的方法，在小样本数据集上表现优良。灰色预测是基于灰色模型展开预测分析的，在诸多灰色模型中，应用最广泛的是 GM(1,1)模型，它能依据已知的微量信息进行模型的设计和预测，进而挖掘系统中隐含的运行变化趋势。灰色预测方法在信息技术、农业科技、电力工业、文化经济等领域都有很广泛的应用背景，可参考文献[17] [18] [19] [20]。另外，支持向量回归在时间序列预测上具有很多明显优势，被广泛应用，具体详见文献[21] [22]。

综上，本文研究的主要思路与方法如下：首先利用函数计算财政收入、第一产业增加值，工业增加值，建筑业增加值，年末总人口，社会消费品零售总额和受灾面积这七个特征间的 Pearson 相关系数矩阵；运用 Lasso 回归法筛选重要特征；接着对单个特征构建灰色预测模型，获得未来两年的预测值；然后再对 2019 年以前的训练样本构建支持向量回归预测模型，得到 2020 年和 2021 年的财政收入预测值；最后给出模型评价与分析。

3. 研究过程与分析

3.1. 收集数据

因为在 1994 年我国的财政体制发生了重大转变，出现了“分税制”财政体制，这就破坏了财政收入相关数据的关联性，1994 年前后的数据不能统一作比较，目前并没有恰当的方法来调整数据发生的变动，因此本文仅收集 1994~2019 年我国财政收入和相关特征数据，如附表 1 所示。其中，数据均来源中华人民共和国国家统计局：《中国统计年鉴(1999~2020)》[23]。为方便研究，将各个特征命名如表 1 所示。

Table 1. The name of features
表 1. 特征命名

命名	y	x_1	x_2	x_3	x_4	x_5	x_6
特征	财政收入(亿元)	第一产业增加值(亿元)	工业增加值(亿元)	建筑业增加值(亿元)	社会消费品零售总额(亿元)	年末总人口(万人)	受灾面积(万公顷)

3.2. 相关性分析选取关键特征

相关性分析是指对两个或若干个彼此具有关联性的特征元素进行分析，以此作为判断两个特征因素关联程度的度量标准。在统计学中，往往通过相关函数计算 Pearson 相关系数来进行相关性分析。表 2 展示了财政收入 7 个特征间的 Pearson 相关系数矩阵。

Table 2. Pearson correlation coefficient matrix
表 2. Pearson 相关系数矩阵

	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1.00	1.00	0.36	0.99	0.99	0.92	−0.93
x_1	1.00	1.00	0.37	0.99	0.99	0.93	−0.94
x_2	0.36	0.37	1.00	0.36	0.35	0.42	−0.22
x_3	0.99	0.99	0.36	1.00	1.00	0.90	−0.92
x_4	0.99	0.99	0.35	1.00	1.00	0.90	−0.93
x_5	0.92	0.93	0.42	0.90	0.90	1.00	−0.86
x_6	−0.93	−0.94	−0.22	−0.92	−0.93	−0.86	1.00

由表 2 可知，受灾面积(x_6)与财政收入(y)的线性关系不显著，呈现负相关。其余特征均与财政收入呈现高度的正相关关系，按照相关性大小排列依次是 x_1，x_3，x_4，x_5 和 x_2。与此同时，各个特征之间存在严重的多重共线性，例如特征 x_1 与 x_3，x_4 存在严重的共线性，x_5 与除了 x_2 和 x_6 以外的其他特征有严重的共线性，x_6 与其他五个特征的共线性不明显。除此之外，x_3 和 x_4 之间存在完全的共线性。

3.3. 选取关键特征

Lasso 回归方法以降阶为主要思想，对特征的系数进行压缩估计并使某些系数变为 0，从而达到筛选特征的目的，是一种常用的正则化方法。由表 2 可知财政收入与各个变量存在严重的多重共线性，这里借用 Lasso 原理和方法实现关键特征识别是恰当的，Lasso 回归系数结果如表 3 所示。从表 3 可看出，影响财政收入的关键影响因素是第一产业增加值(x_1)，建筑业增加值(x_3)和社会消费品零售总额(x_4)。

Table 3. The coefficient table of Lasso regression
表 3. Lasso 回归系数表

系数	x_1	x_2	x_3	x_4	x_5
值	1.7145	0	0.2124	0.1304	0

3.4. 构建预测模型

基于 GM(1,1)预测模型，首先对按照 Lasso 回归法选取出的三个重要特征：第一产业增加值(x_1)，建筑业增加值(x_3)和社会消费品零售总额(x_4)构建灰色预测模型，得到三个特征在 2020 年和 2021 年的预测值和后验差检验判别模型精度的结果如表 4 所示。其中，2020 年第一产业增加值，建筑业增加值和社会消费品零售总额预测值分别为 82,954.97 亿元、102,828.20 亿元和 578,391.60 亿元；2021 年第一产业增加值，建筑业增加值和社会消费品零售总额对应的预测值依次是 89,800.96 亿元、116,399.20 亿元和 652,459.30 亿元。接着将表 4 的预测结果代入财政收入所构建的支持向量回归预测模型，可得到 1994~2021 年财政收入的预测值，如表 5 所示。将财政收入真实值与预测值进行对比，结果如图 1 所示。

表 4 显示第一产业增加值，建筑业增加值和社会消费品零售总额这三个特征通过灰色预测模型输出的预测精度等级良好，由图 1 可观察出建立的预测模型很好地拟合了这 26 年的财政收入的变化情况，都说明构建的模型具备可行性与可靠性。

Table 4. The results of the grey forecasting model
表 4. 灰色预测模型结果

	x_1	x_3	x_4
2020 年预测值	82,954.97	102,828.20	578,391.60
2021 年预测值	89,800.96	116,399.20	652,459.30
预测精度等级	好	好	好

Table 5. The value of revenue forecast from 1994 to 2021
表 5. 1994~2021 年财政收入的预测值

年份	真实值	预测值	年份	真实值	预测值
1994	5218.1	8215.242	2008	61,330.4	58,435.739
1995	6242.2	11,192.990	2009	68,518.3	65,721.819
1996	7408.0	13,849.532	2010	83,101.5	79,469.990

Continued

1997	8651.1	14,790.181	2011	103,874.4	97,424.230
1998	9876.0	15,713.670	2012	117,253.5	111,143.228
1999	11,444.1	16,156.034	2013	129,209.6	125,294.884
2000	13,395.2	17,126.316	2014	140,370.0	137,518.197
2001	16,386.0	18,754.307	2015	152,269.2	146,709.549
2002	18,903.6	20,607.325	2016	159,605.0	157,621.870
2003	21,715.3	22,890.646	2017	172,592.8	170,235.019
2004	26,396.5	28,867.003	2018	183,359.8	181,818.079
2005	31,649.3	32,567.917	2019	190,390.1	196,831.631
2006	38,760.2	37,531.066	2020	NA	237,477.637
2007	51,321.8	46,894.896	2021	NA	244,818.212

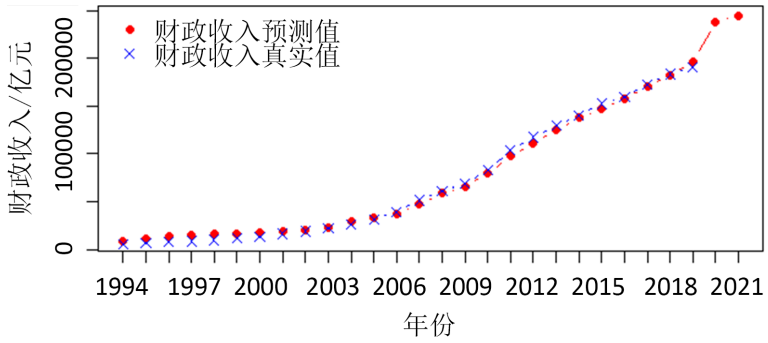

Figure 1. Result of comparing the predicted values to actual values in fiscal revenue
图 1. 财政收入真实值与预测值对比图

3.5. 评价预测模型

使用 R^2 值、调整的 R^2 值、平均百分比误差对模型进行检验，检验结果如表 6 所示。由此表可以看出，平均百分比误差较小，具体值为 0.1984144，R^2 值(0.9957771)与调整的 R^2 值(0.9954099)都特别接近 1，由此说明上述建立的支持向量回归模型拟合效果表现优良，可用于对财政收入的预测分析问题中。

Table 6. The results of the model evaluation indicator
表 6. 模型评价指标结果

指标名称	指标结果
R^2 值	0.9957771
调整的 R^2 值	0.9954099
平均百分比误差	0.1984144

4. 结论与评价

因为影响我国财政收入的因素多种多样，而且这些因素彼此之间总是存在多重共线性问题，本文运用 Lasso 回归方法选择影响财政收入的关键性指标，从 Lasso 的参数估计系数可以看出影响我国财政收入的六个变量(第一产业增加值，工业增加值，建筑业增加值，年末总人口，社会消费品零售总额和受灾面

积)中，第一产业增加值、建筑业增加值，和社会消费品零售总额是表现最为重要的因素，其中第一产业增加值的系数值最大，可见第一产业增加值是影响一个国家经济的最主要原因，因为我国自古以来就是农业大国。农业是人类的生存之本，我国国情决定农业是国民经济的基础与保障，是经济社会发展的“压舱石”。因此，在十四五规划的开局之年，更要持续推进农业绿色、高质、高效发展。

其次建筑业增加值对我国财政收入的影响次之，表明建筑业是我国国民经济的支柱产业，对社会经济发展做出了卓越贡献，因此，在信息飞速发展的时代，要借助数字化、智能化建造技术，促进建筑业转型升级，实现高质量发展，打造“中国建造”品牌。

在变量筛选的基础上，把灰色 GM(1,1)模型和支持向量回归模型相结合，以 1994~2019 年相关数据为例对我国的财政收入做预测分析，通过真实值与预测值的对比图以及 R^2 值、调整的 R^2 值和平均百分比误差验证了所建预测模型的优越性和可靠性。

通过对我国财政收入的预测分析，建议我国政府要着重调整对第一产业、建筑业以及社会消费品零售这几个方面的鼓励政策，加大对第一产业和建筑业的投入力度，努力做好财源建设的优化。具体建议如下：

第一，“农，天下之大业也。”要大力推动中国特色社会主义乡村振兴。优先解决“三农”问题，优先发展农业农村，提高农民合作经济组织的发展；促进多方资源下沉基层，建立健全乡村人才引进制度，加快乡村创新创业队伍建设；缩小城乡区域发展差距，加快农业农村现代化建设步伐，为实现全面建设社会主义现代化国家增添活力！

第二，我国建筑业企业数量繁多，要促进建筑业精益化、智能化、绿色化、工业化“四化”融合发展，升级产业链，提高科技创新能力，努力实现“中国建造”向“中国制造”及“中国创造”的大阔步迈进；借助智能、云端计算等新手段，实现人机合力，提高劳动生产率和产业利用率，推动智能化建筑业领域的发展；立足全球，打造中国建造特色品牌，全面促进建筑业全球化发展，提高中国建筑品牌的国际形象。

第三，在当前疫情的大背景下，优先巩固疫情防控，促进产业转型与升级，刺激消费市场回暖，促进国民经济持续高质量发展。对于政府来说，适当出台减租免租政策，减少零售企业这些中小微企业的资金压力，降低新冠疫情导致的不良影响；对于企业来说，合理发放优惠券，积极发展夜间经济，刺激居民消费；积极响应政府政策，抓住时代机遇，转型升级产业，提高改进技术，顺应消费潮流。

基金项目

特别感谢北京建筑大学 2021 年度研究生创新项目(项目编号：PG2021018)对本文的资助。

参考文献

[1] 陈共. 财政学(第五版) [M]. 北京: 中国人民大学出版社, 2007.

[2] 张绍秋, 胡跃明. 基于 BP 神经网络的税收预测模型[J]. 华南理工大学学报(自然科学版), 2006(6): 55-58.

[3] 范敏, 石为人, 梁勇林, 华海玉. 组合预测模型在地方财政收入预测中的应用[J]. 重庆大学学报, 2008(5): 536-540.

[4] 张玉, 尹腾飞. 支持向量机在税收预测中的应用研究[J]. 计算机仿真, 2011, 28(9): 357-360.

[5] 赵海华. 基于灰色 RBF 神经网络的多因素财政收入预测模型[J]. 统计与决策, 2016(13): 79-81.

[6] 蒋锋, 张婷, 周琰玲. 基于 Lasso-GRNN 神经网络模型的地方财政收入预测[J]. 统计与决策, 2018, 34(19): 91-94.

[7] 白萍. 影响我国财政收入的多元线性回归模型[J]. 统计与决策, 2005(10): 92-94.

[8] 纪跃芝, 邓波, 王继新. 影响财政收入增长的相关因素分析[J]. 统计与决策, 2009(19): 110-112.

[9] 周忠辉, 丁建勋, 王丽丽. 我国财政收入影响因素的实证研究[J]. 当代经济, 2011(8): 84-85.

[10] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[11] 贺平, 兰伟, 丁月. 我国股票市场可以预测吗?——基于组合 LASSO-logistic 方法的视角[J]. 统计研究, 2021, 38(5): 82-96.

[12] 冯亚枝, 胡彦蓉, 刘洪久. 临安山核桃产量的 Lasso-灰色预测模型研究[J]. 林业资源管理, 2021(1): 94-102.

[13] 李翼, 张本慧, 郭宇燕. 改进粒子群算法优化下的 Lasso-Lssvm 预测模型[J]. 统计与决策, 2021, 37(13): 45-49.

[14] 刘妍琛, 张晓曙, 崔旭东, 金娜, 赵祥凯, 赵昕, 郑洪淼, 李娟生, 申希平, 孟蕾, 任晓卫. 基于 Group LASSO Logistic 回归分析模型分析流行性乙型脑炎早期临床症状与预后的关联[J]. 中华疾病控制杂志, 2021, 25(8): 891-897+934.

[15] 丰逸轩, 杨会生, 房繁恭, 刘保延, 勾明会, 刘思雨, 郝鸣昭. 基于病例注册登记研究平台和 LASSO-Cox 模型的针灸对早发性卵巢功能不全患者妊娠结局影响的临床预测模型构建[J]. 中华中医药杂志, 2021, 36(4): 1979-1983.

[16] 曹芳, 朱永忠. 基于多重共线性的 Lasso 方法[J]. 江南大学学报(自然科学版), 2012, 11(1): 87-90.

[17] 杜智涛, 谢新洲. 利用灰色预测与模式识别方法构建网络舆情预测与预警模型[J]. 图书情报工作, 2013, 57(15): 27-33.

[18] 秦琳琳, 马国旗, 储著东, 吴刚. 基于灰色预测模型的温室温湿度系统建模与控制[J]. 农业工程学报, 2016, 32(S1): 233-241.

[19] 刘运花, 黎雄, 刘志雄, 孙元章, 周歧林. 基于灰色预测的广域电力系统稳定器分布延时补偿设计[J]. 电力系统自动化, 2015, 39(12): 44-49.

[20] 段杰, 张娟. 基于灰色预测的深圳文化创意产业发展对经济增长贡献研究[J]. 中国人口·资源与环境, 2014, 24(S1): 457-460.

[21] 刘兰苓, 孙德山, 张文政. 基于支持向量机与 BP 神经网络的税收收入预测模型[J]. 江苏商论, 2019(2): 131-133.

[22] 杨金芳, 翟永杰, 王东风, 徐大平. 基于支持向量回归的时间序列预测[J]. 中国电机工程学报, 2005(17): 110-114.

[23] 国家统计局. 中国统计年鉴[J]. 北京: 中国统计出版社, 1999-2020.

附 录

Table 1. The data of fiscal revenue from 1994 to 2019
附表 1. 1994~2019 年中国财政收入相关数据

年份	财政收入(亿元)	第一产业增加值(亿元)	工业增加值(亿元)	建筑业增加值(亿元)	社会消费品零售总额(亿元)	年末总人口(万人)	受灾面积(万公顷)
1994	5218.1	9471.8	19,546.3	2968.8	18,622.9	119,850	5504.6
1995	6242.2	12,020.5	25,023.2	3733.7	23,613.8	121,121	4582.4
1996	7408.0	13,878.3	29,528.9	4393.0	28,360.2	122,389	4699.1
1997	8651.1	14,265.2	33,022.6	4628.3	31,252.9	123,626	5342.7
1998	9876.0	14,618.7	34,133.9	4993.0	33,378.1	124,761	5014.5
1999	11,444.1	14,549.0	36,014.4	5180.9	35,647.9	125,786	4998.0
2000	13,395.2	14,717.4	40,258.5	5534.0	39,105.7	126,743	5468.8
2001	16,386.0	15,502.5	43,854.3	5945.5	43,055.4	127,627	5221.5
2002	18,903.6	16,190.2	47,774.9	6482.1	48,135.9	128,453	4694.6
2003	21,715.3	16,970.2	55,362.2	7510.8	52,516.3	129,227	5450.6
2004	26,396.5	20,904.3	65,774.9	8720.5	59,501.0	129,988	3710.6
2005	31,649.3	21,806.7	77,958.3	10,400.5	68,352.6	130,756	3881.8
2006	38,760.2	23,317.0	92,235.8	12,450.1	79,145.2	131,448	4109.1
2007	51,321.8	27,674.1	111,690.8	15,348.0	93,571.6	132,129	4899.2
2008	61,330.4	32,464.1	131,724.0	18,807.6	114,830.1	132,802	3999.0
2009	68,518.3	33,583.8	1,380,926.0	22,681.5	132,678.4	133,450	4721.4
2010	83,101.5	38,430.8	165,123.1	27,259.3	156,998.4	134,091	3742.6
2011	103,874.4	44,781.5	195,139.1	32,926.5	183,918.6	134,735	3247.1
2012	117,253.5	49,084.6	208,901.4	36,896.1	210,307.0	135,404	2496.0
2013	129,209.6	53,028.1	222,333.2	40,896.8	242,842.8	136,072	3135.0
2014	140,370.0	55,626.3	233,197.4	45,401.7	271,896.1	136,782	2489.1
2015	152,269.2	57,774.6	234,968.9	47,761.3	300,930.8	137,462	2177.0
2016	159,605.0	60,139.2	245,406.4	51,498.9	332,316.3	138,271	2622.1
2017	172,592.8	62,099.5	275,119.3	57,905.6	366,261.6	139,008	1847.8
2018	183,359.8	64,745.2	301,089.3	65,493.0	377,783.1	139,538	2081.4
2019	190,390.1	70,466.7	317,108.7	70,904.3	408,017.2	140,005	1925.7