Management Science and Engineering 管理科学与工程, 2023, 12(2), 127-135
Published Online March 2023 in Hans. https://www.hanspub.org/journal/mse
https://doi.org/10.12677/mse.2023.122013

Hans汉斯

科技创新绩效影响因素分析及预测

张国强

同济大学经济与管理学院，上海

收稿日期：2023年2月14日；录用日期：2023年3月6日；发布日期：2023年3月17日

摘　要

结合数据分析和挖掘方法建立了一种理论模型，用于评价科技创新绩效。通过对科技创新绩效的众多相关的影响因素进行数据分析，采用Lasso回归方法识别科技创新绩效的关键影响因素，结合灰色预测方法、支持向量机预测模型，建立了科技创新绩效的评估预测模型。以历史数据为实证研究对象，拟合预测了科技创新绩效的期望值和未来发展趋势。

关键词

科技创新绩效，数据分析与挖掘，支持向量机

Analysis and Prediction of Factors Affecting Science and Technology Innovation Performance

Guoqiang Zhang

School of Economics and Management, Tongji University, Shanghai

Received: Feb. 14th, 2023; accepted: Mar. 6th, 2023; published: Mar. 17th, 2023

Abstract

A theoretical model was established for evaluating STI performance by combining data analysis and mining methods. Through data analysis of numerous relevant influencing factors of STI performance, Lasso regression method was used to identify the key influencing factors of STI performance, and combined with gray prediction method and support vector machine prediction model, a prediction model for evaluating STI performance was established. Using historical data as the empirical research object, the expected value and future development trend of science and technology innovation performance were fitted and predicted.

文章引用: 张国强. 科技创新绩效影响因素分析及预测[J]. 管理科学与工程, 2023, 12(2): 127-135.
DOI: 10.12677/mse.2023.122013

Keywords

Science and Technology Innovation Performance, Data Analysis and Mining, Support Vector Machines

1. 引言

自 21 世纪以来，人类便开始逐渐步入数字信息化时代，人们周围始终存在着海量数据，这些海量数据中隐藏着有待发掘的关键信息，对人们进行决策起到一定的支撑作用。在此背景下，数据分析与挖掘技术逐渐地被广泛应用到经济社会发展中，其产生和发展就是为了帮助人们更好地利用这些海量数据，并从中发现并利用数据中潜藏的有用信息。

基于这种背景，本文主要运用数据分析与挖掘技术对科技创新绩效进行分析，挖掘其中隐藏的运行模式，并对未来两年的科技创新绩效进行预测，希望能够帮助政府部门在制定科技创新相关决策时提供理论依据，合理制定科技创新政策，优化创新资源的投入和要素的分配。

2. 研究设计

2.1. 背景与研究目标

在知识经济时代的今天，创新逐渐变成驱动经济发展的重要源泉。现阶段科技创新越发重要，创新主体已从企业发展为政产学研用多主体协同的新阶段。创新能力是推动国家经济社会可持续发展，和经济结构逐步优化的重要支撑，是国家综合竞争力的本质所在[1]。根据相关统计，科技创新在发达国家的经济发展中扮演着决定性作用，大大超过了劳动和资本要素投入的贡献率；相较于发达国家，发展中国家经济发展水平较低，相应的科技创新贡献率比发达国家低[2]。因此，提高科技创新投入、加强科技创新能力、增强科技创新绩效，对于发展中国家而言，是十分必要的。

因此，本项目将采用 2000 年至 2019 年期间由《中国科技统计年鉴》收录的数据资料，进行分析：

1) 对科技创新绩效的相关影响属性进行分析，识别关键因素；

2) 预测 2020 年和 2021 年的科技创新绩效。

2.2. 分析步骤与流程

本项目的数据分析与挖掘方法主要参考 CRISP-DM 数据分析模型流程[3]，该模型将数据处理过程分为 6 个步骤：业务理解(Business Understanding)、数据理解(Data Understanding)、数据准备(Data Preparation)、建立模型(Model Building)、评估(Testing and Evaluation)和部署(Deployment)，具体如图 1 所示。

因此，本项目主要包括以下步骤：

1) 采用探索性分析方法处理原始数据，厘清原始属性彼此的相关性；

2) 利用 Lasso 特征选择模型提取影响因素中的关键属性；

3) 综合灰色预测方法、支持向量机预测模型，建立组合预测模型；

4) 运用建立的预测模型，拟合 2020 年和 2021 年科技创新绩效的期望预测值。

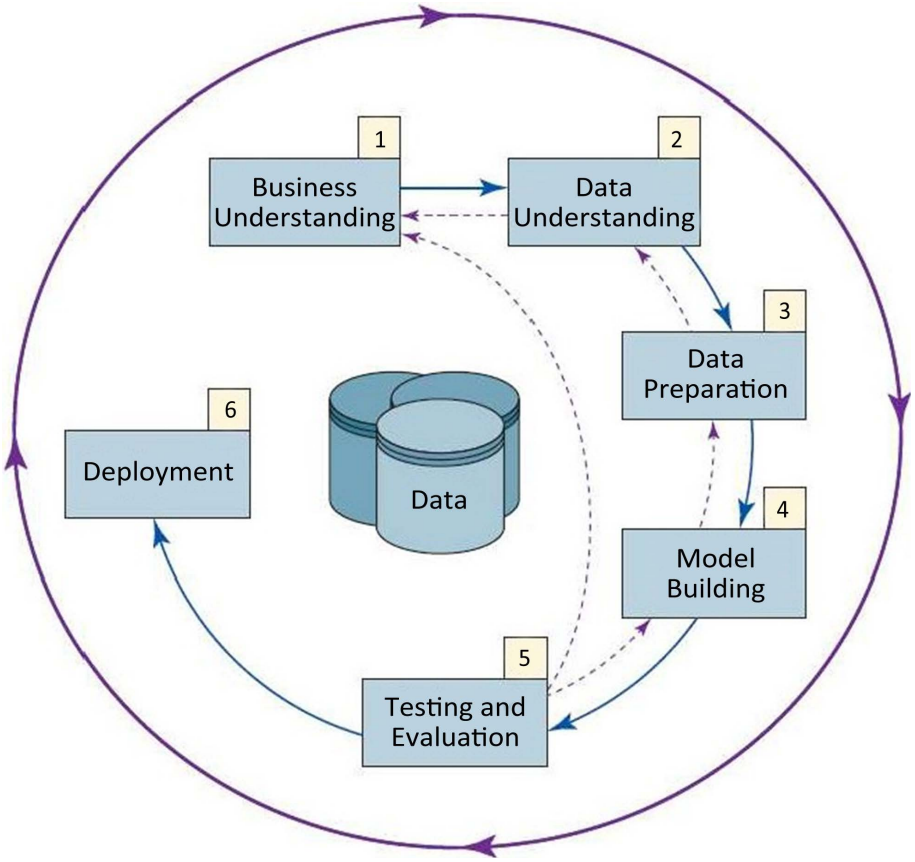

Figure 1. CRISP-DM data analysis process
图 1. CRISP-DM 数据分析过程

3. 科技创新绩效影响因素分析

3.1. 变量解释

影响科技创新绩效(y)的因素有很多，在阅读大量的相关文献[4]-[10]后，通过创新管理理论对创新绩效的解释以及对实际情况的观察，考虑了一些与资源消耗关系密切且有相关关系的因素，初步选取以下属性为自变量，分析他们之间的关系各项属性名称及属性说明如下表 1 所示。

Table 1. Property name and description
表 1. 属性名称和说明

属性名称	属性说明
$x1$：研究与试验发展(R&D)人员全时当量	指报告期内实际从事研发活动时间占系统内工作时间 90%及以上的人员，以 1 人年为单位计算其全时当量
$x2$：研究与试验发展(R&D)经费内部支出	研发经费内部支出是指为实施研发活动，调查单位内部在报告期内发生的所有实际支出
$x3$：科学技术财政支出占公共财政支出的比重	指支持科技创新的政府公共财政的投入力度
$x4$：税收	指国家依照法律规定，强制和无偿获得财政收入的一种规范形式，以向社会提供公共产品、满足社会共同需要、参与社会产品的分配
$x5$：按技术合同构成分全国技术市场成交合同金额	指有关技术合同在技术市场中的成交总金额

Continued

*x*6：研究与开发机构机构数	指开展 R&D 活动的组织或机构
*x*7：国家科技奖项	包括国家科学技术进步奖、国家技术发明奖、国家自然科学奖
*x*8：高技术进出口贸易额合计	指有关高技术产业的进口和出口贸易总额
*x*9：第三产业与第二产业产值比	指第三产业与第二产业的产值比例
*x*10：国际科技合作项目	指中国与其他国家的科技合作项目
*x*11：平均每万名职工中专业技术人员	指每万名职工中，专业技术技能的人员的平均数量
*x*12：污染治理投资总额占 GDP 比重	指创新或经济发展带来的污染治理总额相较于 GDP 的比重
*x*13：商标注册	指商标的注册数量

3.2. 描述性统计分析

首先，对各个属性对象的描述性统计分析结果见表 2 所示。其中科技创新绩效(y)的均值和标准差分别是 912,541.95 和 796,991.97，这说明中国各年份科技创新绩效存在巨大差异；自 2011 年后，各年份科技创新绩效提升幅度比较大。

Table 2. Descriptive statistics for each attribute
表 2. 各个属性的描述性统计

属性	Min	Max	Mean	STD
*x*1	92.21	480.08	254	129.54
*x*2	895.66	22,144.58	8324.94	6816.12
*x*3	3.62	4.67	4.14	0.29
*x*4	65.66	852.56	340.22	251.58
*x*5	6,507,519	223,983,900	60,976,663.93	61,509,776.75
*x*6	8890	24284	14071.3	4880.49
*x*7	223	374	306.2	38.96
*x*8	89,550.08	1,502,104	790,620.11	465,047.79
*x*9	1.03	1.91	1.42	0.25
*x*10	22,524	182,879	69,942.5	41,345.69
*x*11	4233.75	6774.53	5535.23	734.14
*x*12	1.13	1.86	1.37	0.19
*x*13	317,150	12,811,680	2,755,719.3	3,373,422.32
y	105,345	2,591,607	912,541.95	796,991.97

3.3. 相关性分析

通过运用 Pearson 相关系数方法求出原始创新绩效数据的 Pearson 相关系数矩阵。因此，对原始数据经过相关性数据分析后得到所有属性之间相关系数矩阵如下表 3 所示。

Table 3. Pearson correlation coefficient matrix for each variable
表 3. 各变量 Pearson 相关系数矩阵

	*x*1	*x*2	*x*3	*x*4	*x*5	*x*6	*x*7	*x*8	*x*9	*x*10	*x*11	*x*12	*x*13	*y*
*x*1	1.00	0.98	0.65	0.96	0.92	0.97	0.10	0.98	0.91	0.93	0.86	0.35	0.83	0.97
*x*2	0.98	1.00	0.56	0.99	0.97	1.00	−0.03	0.94	0.91	0.94	0.89	0.23	0.90	0.99
*x*3	0.65	0.56	1.00	0.50	0.47	0.54	0.65	0.74	0.70	0.58	0.46	0.82	0.44	0.55
*x*4	0.96	0.99	0.50	1.00	0.97	0.99	−0.08	0.91	0.90	0.92	0.92	0.15	0.90	0.98
*x*5	0.92	0.97	0.47	0.97	1.00	0.97	−0.13	0.86	0.87	0.90	0.87	0.14	0.97	0.98
*x*6	0.97	1.00	0.54	0.99	0.97	1.00	−0.05	0.92	0.89	0.93	0.86	0.22	0.91	0.99
*x*7	−0.10	0.03	−0.65	−0.08	−0.13	0.05	1.00	0.25	0.26	−0.06	0.02	0.67	−0.14	−0.06
*x*8	0.98	0.94	0.74	0.91	0.86	0.92	0.25	1.00	0.93	0.91	0.84	0.46	0.76	0.92
*x*9	0.91	0.91	0.70	0.90	0.87	0.89	0.26	0.93	1.00	0.90	0.91	0.36	0.82	0.89
*x*10	0.93	0.94	0.58	0.92	0.90	0.93	0.06	0.91	0.90	1.00	0.84	0.30	0.86	0.95
*x*11	0.86	0.89	0.46	0.92	0.87	0.86	−0.02	0.84	0.91	0.84	1.00	0.08	0.79	0.87
*x*12	0.35	0.23	0.82	0.15	0.14	0.22	0.67	0.46	0.36	0.30	0.08	1.00	0.15	0.23
*x*13	0.83	0.90	0.44	0.90	0.97	0.91	−0.14	0.76	0.82	0.86	0.79	0.15	1.00	0.92
y	0.97	0.99	0.55	0.98	0.98	0.99	−0.06	0.92	0.89	0.95	0.87	0.23	0.92	1.00

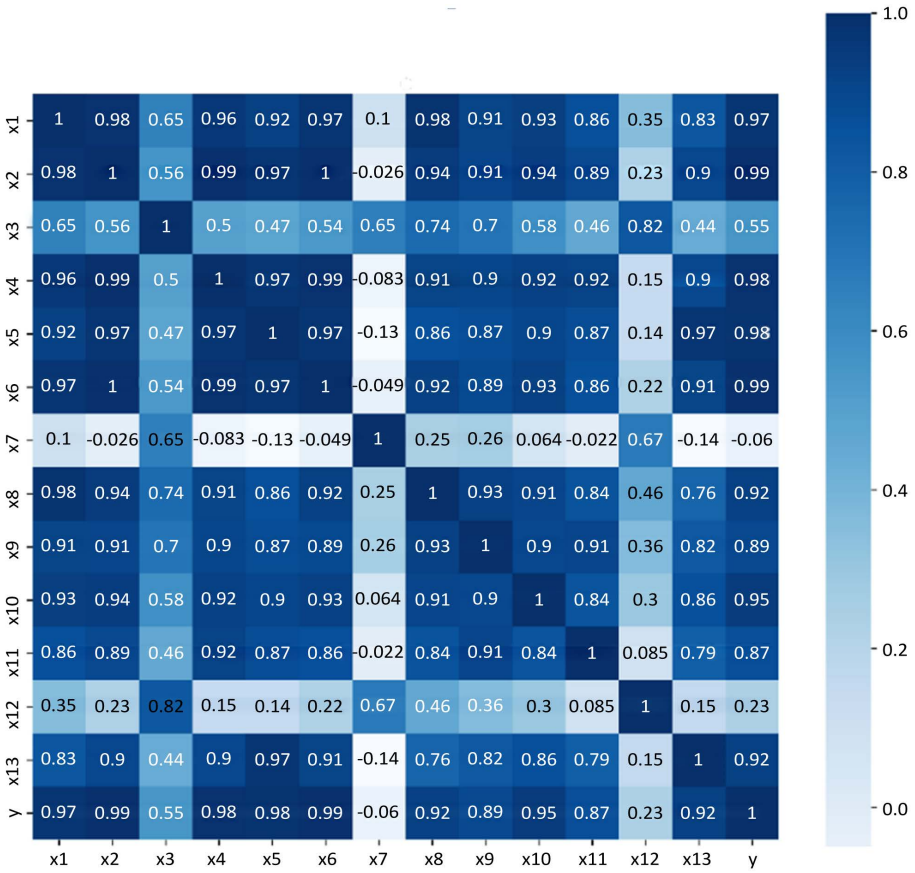

Figure 2. Correlation heat map of each variable
图 2. 各变量相关性热力图

根据表 3 可知，国家科技奖项(x7)与科技创新绩效(y)的线性关系不显著，呈现负相关关系。其余变量与科技创新绩效呈现正相关关系，按照相关系数大小排列，依次是 x2、x6、x4、x5、x1、x10、x8、x13、x9、x11、x3 和 x12。同时，各变量彼此之间的多重共线性现象较为严重，例如，属性 x2 和 x6 之间保持着完全一致的共线性；x1 与除了 x3、x7、x12 外的其他变量存在着严重的多重共线性；x7 与各个变量的共现性不明显；x4 与除了 x7、x12 之外的其他变量有严重的共线性。

通过上述分析可知，选取的各个变量除了 x7 外，其他变量与 y 的相关性较强，可以用作科技创新绩效预测分析的关键变量，但这些变量之间存在着一定的信息重复，需要对变量再做更深一步的筛选分析。如图 2 所示，根据颜色深浅的程度可看出各个变量除了 x7 与 y 为负相关之外，其他变量与 y 存在着一定的相关性。

4. 预测模型构建

4.1. 科技创新绩效关键因素识别

Lasso 回归法属于压缩估计的正则化方法之一。它通过构建惩罚函数得到一个更精炼的模型，使其在设定某些系数为零的同时压缩某些系数，保留了子集收缩的优点，是一种有偏估计范畴的处理复共线性数据的方法。

当多重共线性存在于原始数据特征中时，Lasso 回归可以被视为处理共线性的有效方法，可以有效地筛选多重共线性存在的数据属性。面对海量的数据，信息降维可以用尽可能少的数据来解决问题，用 Lasso 模型选择特征属性也是信息降维的一种有效方法。Lasso 模型在理论上对数据类型的限制并不多，任何类型的数据都可以采用，同时一般无需对数据进行标准化处理。

Lasso 回归法的优点在于能够很好地弥补最小二乘法和逐步回归方法对于局部最优估计的缺失，可以很好地识别特征数据，同时有效解决多重共线性存在于多个个特征之间的问题。缺点是若有一组相关度较高的数据出现时，Lasso 回归倾向于在其中选择一个显著的特征数据而忽略了所有其他数据，这样的情形会造成结果的不稳定。虽然 Lasso 回归方法有如此弊端，但在合适的场景下依然能展现满意的效果。本文研究数据中同样存在一定的多重共线性，运用 Lasso 回归方法识别关键特征属性是必要的步骤。

使用 Lasso 回归进行关键属性选取，结果见表 4 所示，得到各个属性的系数。由表 4 可知，利用 Lasso 回归方法识别影响科技创新绩效的关键影响因素是研究与试验发展(R&D)人员全时当量(x1)、研究与试验发展(R&D)经费内部支出(x2)、科技拨款占公共财政支出的比重(x3)、税收(x4)、按技术合同构成分全国技术市场成交合同金额(x5)、研究与开发机构机构数(x6)、国家科技奖项(x7)、高技术进出口贸易总额合计(x8)、国际科技合作项目(x10)、平均每万名职工中专业技术人员(x11)、商标注册(x13)。

Table 4. Coefficients of each variable
表 4. 各变量系数表

*x*1	*x*2	*x*3	*x*4	*x*5	*x*6	*x*7
340.351	79.427	104,225.926	−161.864	0.004	−13.299	−1195.842
*x*8	*x*9	*x*10	*x*11	*x*12	*x*13	
0.040	0.000	2.789	−97.022	0.000	−0.002	

4.2. 建立预测模型

4.2.1. 灰色预测模型

灰色预测法是对包含不确定因素的复杂系统的拟合预测的一种方法。在适用灰色预测模型前，需要

对原始数据序列进行数据变换处理，处理后的数据序列称为生成列。累加和累减两种数据处理方式是灰色预测常用的处理方式。

灰色预测以灰色模型为基础，GM(1,1)模型是众多灰色模型中使用频率最高的。其检验模型精度的后验差标准如表 5 所示。

Table 5. Reference table for post-test difference test
表 5. 后验差检验参照表

P	C	模型精度
>0.95	<0.35	好
>0.80	<0.5	合格
>0.70	<0.65	勉强合格
<0.70	>0.65	不合格

灰色预测模型具有很强的通用性，适用于多数的时间序列场景，拟合表现不错，特别适用于对数据产生机理不明确且规律性较差的情况。该模型的优点是预测精度高，模型精度可验证，参数估计方法简单，对小数据集的预测效果很好；缺点是对原始序列的数据平滑度要求较高，灰色预测模型在原始序列的平滑度不佳的情况下，拟合预测精度不高，甚至无法通过方差检验，导致只能放弃使用灰色预测模型。

4.2.2. 支持向量机回归模型

SVR (Support Vector Regression，支持向量回归)是在做拟合时，采用了支持向量的思想，来对数据进行回归分析。给定训练数据集 $T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n y_n)\}$，其中 $\boldsymbol{x}_1=\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(n)}\right)^{\mathrm{T}}\in R^n, y_i\in R,$ $i=1,2,\cdots,n$。对于样本 $(\boldsymbol{x}_i,y_i)$ 通常根据模型计算的 $f(\boldsymbol{x}_i)$ 与实际值 y_i 的差异度来衡量损失，当且仅当 $f(\boldsymbol{x}_i)=y_i$ 时计算损失为零。SVR 的基本思路可以描述为：允许 $f(\boldsymbol{x}_i)$ 与 y_i 至多存在 ε 的偏差。只有当 $f(\boldsymbol{x}_i)=y_i|>\varepsilon$ 时，才认为有损失。当 $|f(\boldsymbol{x}_i)=y_i|\le\varepsilon$ 时，认为预测准确。

由于支持向量机具有相对完善的理论基础以及良好的特性，在分类、回归、聚类、时间序列分析、异常点检测等众多应用方向，人们对该模型的研究和应用都非常广泛。具体研究包括统计学习的理论基础，建立各种模型，改进模型对应的优化算法，以及实际应用等方面的内容。

相比较于其他方法，支持向量回归的好处在于：既适合线性模型，又能很好地把握非线性关系的数据与特征；避免局部极小化问题，提高泛化性能，解决高维度问题，而无需担心多重共线性问题；虽然过程中不会直接排除异常值，但会让异常引发较小的偏差。缺点是在面对数据资料的数据量巨大时，计算复杂度较高，且耗时较长。

5. 预测模型的实证研究

基于 Lasso 回归，选取研究与试验发展(R&D)人员全时当量(x1)、研究与试验发展(R&D)经费内部支出(x2)、科技拨款占公共财政支出的比重(x3)、税收(x4)、按技术合同构成分全国技术市场成交合同金额(x5)、研究与开发机构机构数(x6)、国家科技奖项(x7)、高技术进出口贸易总额合计(x8)、国际科技合作项目(x10)、平均每万名职工中专业技术人员(x11)、商标注册(x13)变量，通过灰色预测模型得出这些变量 2020 年和 2021 年的预测期望值，见表 6。

基于表 6 的预测结果，将其代入科技创新绩效的支持向量机回归预测模型，得到 2020 年至 2021 年科技创新的预测值，如下表 7 和图 3 所示，其中 y_pred 表示科技创新绩效的预测值。

Table 6. Predicted values of each variable
表 6. 各变量预测值

	2020 年预测值	2021 年预测值	预测精度
*x*1	561.89	610.6	好
*x*2	31,705.28	36,400.93	好
*x*3	4.47	4.5	合格
*x*4	1052.28	1197.04	好
*x*5	240,877,000	290,696,700	好
*x*6	24,933.51	26,527.97	好
*x*7	315.69	316.56	勉强合格
*x*8	1,879,519.86	2,046,673.4	好
*x*10	167,521.98	184,104.76	合格
*x*11	6773.6	6907.4	好
*x*13	16,790,178.98	21,408,026.22	不合格

Table 7. Projections of STI from 2000 to 2021
表 7. 2000 年至 2021 年科技创新的预测值

年份	*y*	*y*_pred	年份	*y*	*y*_pred
2000	105,345	101,046.9	2011	860,513	896,746.6
2001	114,251	262,917.6	2012	1,255,138	1,192,184
2002	132,399	136,221.7	2013	1,313,000	1,313,729
2003	182,226	177,261.6	2014	1,302,687	1,441,474
2004	190,238	154,372.3	2015	1,518,192	1,517,898
2005	219,003	188,814.7	2016	1,753,763	1,749,754
2006	268,002	269,937.1	2017	1,836,434	1,897,466
2007	351,782	339,980	2018	2,447,460	2,447,460
2008	411,982	411,668.2	2019	2,591,607	2,590,852
2009	581,992	587,031.5	2020		3,212,352
2010	814,825	814,852	2021		3,700,227

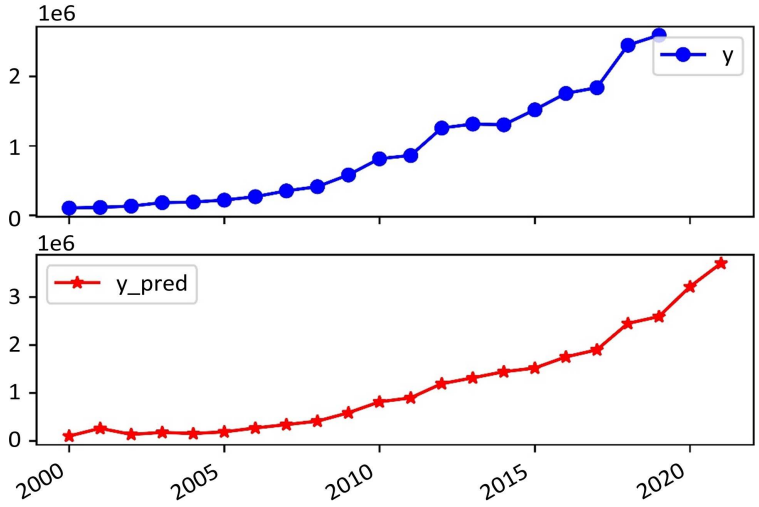

Figure 3. Comparison of real and predicted values of STI performance
图 3. 科技创新绩效真实值与预测值的对比

根据灰色预测模型和支持向量机模型的预测结果，可以得出，该模型的组合预测方法对科技创新绩效的拟合与预测表现较好，模型的精度较高，能够一定程度上反映科技创新绩效的发展方向。

6. 结论

从科技创新绩效的发展历程和预测结果来看，我国科技创新绩效发展较为稳健，自 2005 年开始保持高水平的增长态势，科技创新活力十足。2010 年以后，科技创新绩效波动较大，在科技创新发展步伐放缓的阶段，科技创新投入的力度加大和科技创新的高质量转型，是推动科技创新能力升级的主要动力源。结合 2020 年和 2021 年的预测结果，未来我国科技创新绩效继续较高速率的增长，整体发展形式稳中向好，因此应当不断增强政府的政策引导和策动作用，激发科技创新动力，拓宽科技创新空间，为科技创新能力的发展提供坚实的基础。

本研究的主要成果和不足：通过分析我国科技创新绩效的影响因素，构建识别和预测模型，对我国科技创新的发展进行量化分析，预测其未来发展趋势。但由于数据统计的滞后性，对长远阶段的预测精度不够，后续可以通过分析区域性、季度性数据，提高模型对于未来期望的预测和指导力度。

参考文献

[1] 张浩, 霍国庆, 汪明月, 等. 科技成果转化的战略绩效评价——基于国家科学技术进步奖成果的实证研究[J]. 科学学与科学技术管理, 2020, 41(8): 7-25.

[2] Cooke. (1998) Regional Innovation Systems: The Role of Governances in a Globalized World. UCL Press, London.

[3] 韩宝国, 张良均. R 语言商务数据分析实战[M]. 北京: 人民邮电出版社, 2019.

[4] 王彩明, 李健. 中国区域绿色创新绩效评价及其时空差异分析——基于 2005-2015 年的省际工业企业面板数据[J]. 科研管理, 2019, 40(6): 29-42. https://doi.org/10.19571/j.cnki.1000-2995.2019.06.004

[5] 张家峰, 李佳楠, 陈红喜, 等. 长三角高校科研创新绩效评价及影响因素研究——基于 DEA-Malmquist-Tobit 模型[J]. 科技管理研究, 2020, 40(9): 80-87.

[6] 孙丽文, 李跃. 京津冀区域创新生态系统生态位适宜度评价[J]. 科技进步与对策, 2017, 34(4): 47-53.

[7] Lee, J.-D. and Park, C. (2006) Research and Development Linkages in a National Innovation System: Factors Affecting Success and Failure in Korea. *Technovation*, **26**, 1045-1054. https://doi.org/10.1016/j.technovation.2005.09.004

[8] 柴玮, 申万, 毛亚林. 基于 DEA 的我国资源型企业科技创新绩效评价研究[J]. 科研管理, 2015, 36(10): 28-34. https://doi.org/10.19571/j.cnki.1000-2995.2015.10.004

[9] 许敏, 王慧敏, 钱一奇. 高校科技创新绩效评价及协同创新机制研究——以长三角区域 82 所高校样本比较分析为例[J]. 中国高校科技, 2021(10): 44-49. https://doi.org/10.16209/j.cnki.cust.2021.10.008

[10] Teng, T.W. and Chen, J.Y. (2019) The Performance Space Measurement of Regional Innovation System Based on Neuropsychology. *Cognitive Systems Research*, **56**, 159-166. https://doi.org/10.1016/j.cogsys.2018.10.034